MÉTODOS NUMÉRICOS II

Universidad de Valladolid

11 de febrero de 2011

2010-11

Índice General

Lista de Figuras **7**

I PRELIMINARES **9**

1 El sistema de coma flotante **11**
1.1 Cambios de base de numeración 11
1.2 Sistemas numéricos de coma flotante 15
1.3 Errores en un sistema de coma flotante normalizado 18
1.4 Aritmética de coma flotante 19
1.5 Cuestiones y problemas 23

2 Normas vectoriales y matriciales **25**
2.1 Significado numérico de la norma 25
2.2 Normas vectoriales usuales 26
2.3 Normas matriciales 27
2.4 Expresiones de ciertas normas matriciales usuales 28
2.5 Relación entre radio espectral y norma de una matriz 30
2.6 Cuestiones y problemas 31

3 Errores en la resolución numérica de sistemas lineales **33**
3.1 Introducción 33
3.2 Acondicionamiento de un sistema lineal 34
3.3 Estudio cuantitativo del acondicionamiento 35
3.4 Análisis del error en la eliminación Gaussiana 38
3.5 Cuestiones y problemas 42

II INTERPOLACIÓN **45**

4 Polinomios de Chebyshev **47**
4.1 Elección óptima de los nodos de interpolación 47
4.2 Los polinomios de Chebyshev 48
4.3 Cambio de intervalo 51
4.4 Cuestiones y problemas 53

5 La interpolación de Hermite u osculatoria **57**
5.1 El problema de Hermite . . . 57
5.2 Construcción del interpolante en forma de Newton . . . 57
5.3 Diferencias divididas con argumentos repetidos . . . 60
5.4 Caso a trozos. Cúbicas de Hermite segmentarias . . . 62
5.5 Cuestiones y problemas . . . 65

6 Splines **69**
6.1 Definición y construcción de splines cúbicos . . . 69
6.2 B-splines y otras bases . . . 75
6.3 Cuestiones y problemas . . . 76

7 Transformada rápida de Fourier **81**
7.1 La transformada de Fourier discreta . . . 81
7.2 El algoritmo FFT . . . 84
7.3 Aplicación a la interpolación trigonométrica . . . 90
7.4 Cuestiones y problemas . . . 92

III APROXIMACIÓN **95**

8 Introducción a la aproximación **97**
8.1 Conceptos generales sobre aproximación . . . 97
8.2 Ajuste . . . 98
8.3 Aproximación óptima: existencia y unicidad . . . 100
8.4 Convergencia de las mejores aproximaciones. Teorema de Weierstrass . . . 102
8.5 Cuestiones y problemas . . . 104

9 Problemas de mínimos cuadrados **107**
9.1 Aproximación en un espacio con producto interno . . . 107
9.2 Aproximación en subespacios de dimensión finita . . . 110
9.3 Sistemas ortogonales . . . 114
9.4 Cuestiones y problemas . . . 116

10 Polinomios ortogonales **121**
10.1 Funciones peso . . . 121
10.2 Polinomios ortogonales . . . 122
10.3 Sistemas clásicos . . . 124
10.4 Convergencia de los desarrollos ortogonales . . . 127
10.5 Cuadratura Gaussiana . . . 129
10.6 Cuestiones y problemas . . . 131

11 Aproximación funcional discreta **135**
11.1 El efecto del '*sampling*'-'*aliasing*' . . . 135
11.2 Aproximación trigonométrica . . . 138
11.3 Polinomios de Gram . . . 141
11.4 Polinomios de Chebyshev . . . 146
11.5 Cuestiones y problemas . . . 149

IV ÁLGEBRA LINEAL NUMÉRICA **153**

12 Factorizaciones de una matriz **155**
12.1 El caso simétrico 155
12.2 La factorización LDL^T 158
12.3 Cuestiones y problemas 159

13 El problema lineal de mínimos cuadrados **161**
13.1 Solución por mínimos cuadrados 162
13.2 Transformaciones de Householder 163
13.3 Ortonormalización de Gram-Schmidt 165
13.4 Pseudo-inversa de una matriz 167
13.5 Cuestiones y problemas 171

2010-11

Lista de Figuras

2.1 La desigualdad triangular . . . 26
2.2 Superficies de las bolas unidad . . . 27

4.1 Polinomios de Chebyschev de grados 0 a 5 . . . 49
4.2 Puntos de interpolación de Chebyschev para $n = 6$ . . . 50
4.3 Interpolaciones sucesivas en 5, 9, 13 y 17 puntos equiespaciados . . . 52
4.4 Interpolaciones sucesivas en 5, 9, 13 y 17 puntos de Chebyshev . . . 53

5.1 Elementos Φ_i y Θ_i en un punto interior . . . 65

6.1 La cúbica de Hermite a trozos y los *splines* completo y natural . . . 73
6.2 El B-spline [1 3 4 6 9] y sus polinomios componentes. . . . 77

7.1 Raíces octavas de la unidad en el plano complejo . . . 84
7.2 Esquema del algoritmo de Cooley y Tuckey . . . 87
7.3 Esquema del algoritmo de Sande y Tuckey . . . 89
7.4 Interpolación trigonométrica de una función real . . . 92

8.1 Aproximantes de grados 5 a 10 para la función de Runge . . . 104

9.1 Ilustración geométrica de los mínimos cuadrados . . . 109

10.1 Polinomios de Legendre de grados 0 a 4 . . . 126
10.2 Polinomios de Laguerre de grados 0 a 4 (para $\alpha = 0$) . . . 126
10.3 Polinomios de Hermite de grados 0 a 4 . . . 127

11.1 Funciones coseno idénticas sobre una red equiespaciada . . . 136
11.2 Funciones seno idénticas sobre una red equiespaciada . . . 136
11.3 Polinomios de Chebyshev *aliados* en las raíces de otro . . . 137
11.4 Aproximación trigonométrica por mínimos cuadrados . . . 141
11.5 Polinomios de Gram hasta grado 4 en 5 puntos . . . 142
11.6 Polinomios de Gram hasta grado 4 en 9 puntos . . . 143
11.7 Función peso para los polinomios de Chebyshev . . . 146

13.1 $A\mathbf{x}$ es la proyección de $\mathbf{b}$ en el subespacio Im A . . . 162
13.2 $P(\mathbf{u})\mathbf{x}$ como reflejo de $\mathbf{x}$ en el subespacio ortogonal a $\mathbf{u}$ . . . 164
13.3 Como obtener $\mathbf{x}_{LS}$ a partir de alguna solución por mínimos cuadrados . . . 169
13.4 Ilustración gráfica para el caso de una matriz 2×2 de rango 1 . . . 170

2010-11

CAPÍTULO I

PRELIMINARES

2010-11

Lección 1

El sistema de coma flotante

En contra de lo que a primera vista pueda parecer, la forma en la que los ordenadores trabajan con los números es bastante diferente de lo que las leyes aritméticas que hemos estudiado exigen. Cuando nos sentamos ante una terminal, para hacer algún tipo sencillo de cálculo, raramente encontramos alguna diferencia entre lo que esperamos y lo que vemos en la pantalla. Sin embargo, el proceso seguido es realmente complejo. En primer lugar, en la vida ordinaria utilizamos números basados en el sistema decimal mientras que el ordenador sólo puede trabajar con números en base 2, aunque nos haga el favor de entenderse con nosotros en base 10, para lo que tiene que realizar una doble transformación: primero al recoger los datos del teclado y después al escribirlos en la pantalla.

En segundo lugar, el conjunto de números que tenemos a nuestra disposición constituyen un continuo infinito sin más limitaciones que nuestra imaginación para nombrarlos y representarlos, mientras que el ordenador sólo puede trabajar con una cantidad finita de puntos. Se trata de una limitación fundamental, pues por elevada que sea esta cantidad, siempre habrá magnitudes que sea incapaz de representar de forma exacta. Una vez producido uno de estos errores, su control a lo largo de las sucesivas operaciones que se realicen es francamente complicado y requiere la implementación de una aritmética adecuada, que difiere notablemente de la habitual.

Analicemos estos aspectos, para tratar de comprender el cálculo computacional desde dentro del ordenador, lo que resulta de gran ayuda a la hora de diseñar y analizar algoritmos y métodos numéricos.

1.1 Cambios de base de numeración

Comencemos por la parte más fácil y veamos como se representan en diferentes bases los números enteros, más concretamente los naturales. Por ejemplo, si escribimos 37294, por sí mismo, no representa nada, pero si decimos que esta en base 10, ya sabemos que nos estamos refiriendo a

$$37294_{10} = 3 \cdot 10^4 + 7 \cdot 10^3 + 2 \cdot 10^2 + 9 \cdot 10^1 + 4 \cdot 10^0$$

Naturalmente que nunca escribimos el subíndice 10, puesto que es el valor normal y que suponemos *por defecto*; es decir, cuando no se pone nada. En general, usando cifras a_i, tales que $0 \leq a_i \leq 9$ tendremos que

$$N = (a_n a_{n-1} \cdots a_1 a_0)_{10} = \sum_{k=0}^{n} a_k \cdot 10^k$$

pero no hay ninguna razón que nos obligue a utilizar esta base, y de hecho se sabe de otras civilizaciones que utilizaron otras bases, e incluso en nuestros días hay magnitudes que se expresan en términos de *docenas*, que indica base 12. Tampoco parece haber ninguna razón que nos invite a cambiar de base para la vida cotidiana dado lo arraigado del sistema decimal.

Pero, si trabajamos con computadores la cosa cambia porque ellos están obligados a trabajar en base 2, un sistema binario de *sí o no*, de 0 ó 1, porque la corriente pasa o no pasa. En consecuencia, si nos queremos adaptar a su forma de hacer, debemos aprender a manejarnos en base 2 y/o a cambiar de forma eficiente entre dicha base y la decimal que nosotros dominamos.

El paso *directo* (de otra base a la decimal) es francamente sencillo, pues si se trata de trabajar en una base β, sabemos que con dígitos a_i, $0 \le a_i < \beta$, un número

$$N = (a_n a_{n-1} \cdots a_1 a_0)_\beta = a_n \cdot \beta^n + a_{n-1} \cdot \beta^{n-1} + \cdots + a_1 \cdot \beta^1 + a_0 \cdot \beta^0$$

equivale en decimal a evaluar un polinomio con coeficientes a_i en el punto β para lo que disponemos del bien conocido algoritmo de Horner, que actúa en la forma

$$a_n \cdot \beta^n + a_{n-1} \cdot \beta^{n-1} + \cdots + a_1 \cdot \beta^1 + a_0 \cdot \beta^0 = a_0 + \beta[a_1 + \cdots + \beta(a_{n-1} + \beta \cdot a_n)] \quad (1.1)$$

y requiere un número de multiplicaciones igual al grado del polinomio; es decir, tantos dígitos tenga el número menos uno y realizadas todas ellas en la base de llegada, es decir en la 10 en este caso, lo que es sumamente deseable al ser ésta en la que dominamos las tablas para operar.

Aunque realmente todas las bases son equivalentes en cuanto a su operatividad, como ya hemos dicho, para comprender las operaciones computacionales nos basta, además de la 2 por las razones ya mencionadas, con la 8 y la 16 que al ser potencias de 2 nos permiten cambiar de forma inmediata con la binaria y manejar números con una cantidad muy inferior de dígitos.

Veamos algunos ejemplos de cambios de bases. Primero de base 8 a 10

$$(21467)_8 = 7 + 6 \cdot 8 + 4 \cdot 8^2 + 1 \cdot 8^3 + 2 \cdot 8^4 = 9015$$

o aplicando el algoritmo de Horner en la forma de la regla de Ruffini

	2	1	4	6	7
8		16	136	1120	9008
	2	17	140	1126	9015

Ahora un cambio que habrá que realizar frecuentemente de base 2 a 10

$$(1101)_2 = 1 + 1 \cdot 2^2 + 1 \cdot 2^3 = 13$$

Más complejo es el caso en que la base β es mayor que 10, pues se necesitarán símbolos adicionales para los 'dígitos' $10, 11, 12, \ldots, \beta - 1$. Para el caso de base 16 suelen utilizarse

las letras A, B, C, D, E y F para los valores entre 10 y 15, por lo que el lector no tendrá inconveniente en comprobar que

$$(2BED)_{16} = 11245$$

Para hacer el cambio de base *inverso* (de base 10 a otra cualquiera), aunque en principio el proceso a realizar es el mismo al no haber ninguna diferencia de prioridad entre bases, esta simetría se rompe a la hora de operar, pues como hemos visto los cálculos se han de realizar en la base de llegada. Por ejemplo para pasar el número 3781 (en base 10 si no se dice nada) a base 2, aplicando el algoritmo de Horner, tendremos

$$\begin{aligned} 3781 &= 1 + 10(8 + 10(7 + 10 \cdot 3)) = \\ &= 1_2 + (1010)_2[(1000)_2 + (1010)_2\{(111)_2 + (1010)_2(11)_2\}] = \\ &= (111011000101)_2 \end{aligned}$$

donde el lector tendrá ocasión de comprobar su conocimiento de las tablas de sumar y multiplicar en base 2, y lo engorroso de esta forma de proceder en relación con el anterior cambio, debido únicamente a nuestra familiaridad con las operaciones decimales, pero difícil de solventar.

Resulta conveniente, pues, buscar una forma de realizar este cambio inverso operando tambien en base 10. El método resulta evidente si observamos detenidamente la ecuación (1.1), pues es inmediato darse cuenta de que el dígito menos significativo en la base β, a_0 es el resto de dividir (en base 10) por β el número decimal que queremos convertir. El cociente será todo lo contenido en el corchete, que obviamente nos va a permitir calcular a_1 por el mismo procedimiento, y así sucesivamente. Un esquema habitual y con cierta similitud a la regla de Ruffini nos proporciona la solución

2												
3781	1890	945	472	236	118	59	29	14	7	3	1	0
1	0	1	0	0	0	1	1	0	1	1	1	

con la única salvedad de que las cifras van apareciendo de una en una, y al revés de su posición en el número.

El problema de cambiar las expresiones entre dos bases cualesquiera, se puede resolver con facilidad haciendo el tránsito por la decimal, aunque cuando las bases son unas potencias de las otras se pueden realizar cambios triviales, que el lector descubrirá inmediatamente. Por ejemplo:

$$(101101001)_2 = (551)_8 = (169)_{16}$$

¿Sería igual de sencillo pasar de base 8 a 16 sin conocer la expresión en base 2? ¿Por qué? ¿A partir de que expresión es más eficiente el cálculo del valor decimal?.

1.1.1 Números fraccionarios. Nos centraremos ahora en los números reales y positivos comprendidos entre 0 y 1; es decir, a los que se ha privado de la parte entera. Su tratamiento es ligeramente más complejo que para los enteros; en primer lugar, porque hemos de tratar con potencias negativas y en segundo porque las expresiones tienen a veces infinitos dígitos. Veremos aquí los mecanismos para cambiar de base trabajando

en base 10, dejando que el lector descubra la forma de proceder en la otra base (véase el ejercicio 1.5.1). Partiendo de la obviedad de que, por ejemplo

$$\begin{aligned} .7215 &= \frac{7}{10} + \frac{2}{100} + \frac{1}{1000} + \frac{5}{10000} = \\ &= 7 \cdot 10^{-1} + 2 \cdot 10^{-2} + 1 \cdot 10^{-3} + 5 \cdot 10^{-4} \end{aligned}$$

es fácil generalizar a la expresión para una fracción cualquiera, donde recordamos que se debe verificar $0 \leq b_i < \beta$,

$$(.b_1 b_2 \ldots b_n \ldots)_\beta = \sum_{k=1}^{\infty} b_k \cdot \beta^{-k}$$

Veamos en primer lugar el cambio de una fracción en base 8 a una decimal

$$\begin{aligned} (.36207)_8 &= 3 \cdot 8^{-1} + 6 \cdot 8^{-2} + 2 \cdot 8^{-3} + 7 \cdot 8^{-5} = \\ &= 8^{-5}(3 \cdot 8^4 + 6 \cdot 8^3 + 2 \cdot 8^2 + 7) = \\ &= \frac{(36207)_8}{8^5} = \frac{15495}{32768} = .472869873046875 \end{aligned}$$

donde, en el fondo, todo se reduce a pasar el número entero en base β (8 aquí) que resulta de prescindir del punto a decimal y dividir por una potencia de la base (todo ello en base 10) y utilizando, por tanto, las técnicas vistas anteriormente.

Este método parece ser inadecuado cuando la cantidad de cifras es infinita; pero si lo pensamos bien, en la práctica computacional nunca dispondremos más que de un número finito de ellas o de una repetición periódica, y en este caso, el lector será capaz de encontrar el procedimiento eficaz (véase el ejercicio 1.5.2).

Hay que tener en cuenta que la condición de tener infinitas cifras depende estrechamente de la base; así, es evidente que

$$.333\ldots = .\overline{3} = \frac{1}{3} = (.1)_3$$

aunque, en general, en estos casos en una u otra base hay que realizar el proceso de calcular la fracción generatriz. Veamos el proceso de pasar de una fracción decimal a otra base, trabajando en base 10, y esto nos permitirá ver las dificultades y cómo resolverlas. Si, por ejemplo, tenemos el decimal exacto $x = .372$ y lo queremos pasar a base β, procedemos por una técnica de coeficientes indeterminados, y supuesto que

$$x = (.b_1 b_2 b_3 \ldots)_\beta = \sum_{k=1}^{\infty} b_k \cdot \beta^{-k}$$

resulta que multiplicando ambos miembros por β,

$$\beta x = (b_1 . b_2 b_3 \ldots)_\beta$$

y la primera cifra resulta ser la parte entera de este producto, quedando la segunda como primera de la nueva parte fraccionaria, por lo que una vez extraída b_1 podremos seguir con el proceso hasta extraer toda la cadena de cifras; eso sí una a una como en el caso de los enteros y teniendo que transformar su expresión cuando la base sea mayor que 10. Los ejemplos siguientes aclaran esta situación.

Pasemos primero .372 a base 2

$$\begin{array}{lll} & d_0 = .372 & \\ 2\ d_0 = .744 & d_1 = .744 & b_1 = 0 \\ 2\ d_1 = 1.488 & d_2 = .488 & b_2 = 1 \\ 2\ d_2 = .976 & d_3 = .976 & b_3 = 0 \\ 2\ d_3 = 1.952 & d_4 = .952 & b_4 = 1 \\ 2\ d_4 = 1.904 & d_5 = .904 & b_5 = 1 \end{array}$$

con lo que concluímos que $.372 = (.010111\ldots)_2$, sin que de momento se vea hay alguna repetición periódica de cifras. ¿La habrá?. Tratemos ahora de pasar a base 16

$$\begin{array}{lll} & d_0 = .372 & \\ 16\ d_0 = 5.952 & d_1 = .952 & b_1 = 5 \\ 16\ d_1 = 15.232 & d_2 = .232 & b_2 = 15\ (\mathrm{F}) \\ 16\ d_2 = 3.712 & d_3 = .712 & b_3 = 3 \\ 16\ d_3 = 11.392 & d_4 = .392 & b_4 = 11\ (\mathrm{B}) \\ 16\ d_4 = 6.272 & d_5 = .272 & b_5 = 6 \end{array}$$

es decir, que $.372 = (.5F3B6\ldots)_{16}$. ¿Podemos asegurar a partir de esta expresión que $.372 = (.01011111001110110110\ldots)_2$? ¿Por qué?

Como los números tienen, normalmente, parte entera y parte fraccionaria, se debe considerar también los cambios de base de expresiones de la forma general

$$(a_n a_{n-1} \ldots a_1 a_0 . b_1 b_2 \ldots\ldots)_\beta = \sum_{k=0}^{n} a_k \cdot \beta^k + \sum_{k=1}^{\infty} b_k \cdot \beta^{-k}$$

pero puede hacerse perfectamente separando ambas componentes, y merece la pena que nos centremos en el principal aspecto del sistema numérico de los ordenadores: sólo manejan una cantidad finita de dígitos, que se traduce en la práctica en una cantidad finita de puntos o números.

1.2 Sistemas numéricos de coma flotante

Una primera opción es mantener el punto fijo, es decir, exigir que la cantidad de cifras después del punto decimal sea fija, lo que resulta obligado para disponer de una aritmética '*standard*' manejable. Entonces, la cantidad de puntos es especialmente pequeña. Por ejemplo, si nos limitamos a ocho dígitos con cinco de ellos correspondientes a la parte fraccionaria, el número más grande que podemos representar es 999.99999, y el más pequeño no nulo .00001; o sea, que tenemos 100 millones de números positivos en el intervalo [0,1000), equidistando entre sí una cienmilésima. Es obvio que alcanzar cantidades elevadas nos exigiría la utilización de muchas cifras, por lo que estos sistemas de punto fijo no resultan adecuados. Si hemos de utilizar ocho dígitos significativos para nuestros números hay formas mucho más eficaces de hacerlo.

1.2.1 Como primera medida, si dejamos que el punto pueda ocupar cualquier posición, si se le dejar 'flotar' a lo largo de la cadena de cifras ya se puede alcanzar 99999999 por arriba y .00000001 por abajo, aunque naturalmente hemos perdido la equidistancia, que es la única virtud de los sistemas de punto fijo. Los puntos (¿son la misma cantidad que

con la coma fija?) están tanto más separados cuanto mayores son, y en consecuencia los errores que se cometerán al utilizarlos guardarán una cierta proporción con su magnitud.

Llegados a este punto, lo que se hace para extender más aún el rango de valores del sistema es utilizar alguno de estos dígitos como exponentes de 10. En el caso que venimos comentando, si dedicamos dos cifras a este menester, el rango se podrá alargar, con pequeños detalles técnicos que veremos a continuación, a números entre 10^{-99} y 10^{99}. Estamos ante un sistema de punto (coma es el término más habitual) flotante, cuya principal ventaja es que puede manejar números de magnitudes muy diferentes y cuyos ingredientes básicos son:

- Una base β. (2, 8, 10 y 16 son las más frecuentes)
- Una cantidad t de dígitos d_i en base β dedicados a representar el valor absoluto del número con el que se va a operar en la práctica (significando s) y que forman la denominada *mantisa*.

  t puede ser muy variable de unas máquinas a otras. En general es variable, de forma que se pueda operar al menos en dos longitudes (precisiones) distintas.
- Un cierto rango de variación para los *exponentes* enteros e; por ejemplo, $-m \leq e \leq M$, donde m y M son muy variables de unos sistemas a otros, verificando en general que $m = M$.
- Un signo, lógicamente $\pm$.

Entonces, un número de un sistema de coma flotante tiene la forma

$$\pm.d_1 d_2 \ldots d_t \times \beta^e = s \times \beta^e$$

donde $d_1, d_2, \ldots, d_t, e$ son enteros que satisfacen

$$\begin{aligned} 1 \leq d_1 &\leq \beta - 1 \\ 0 \leq d_i &\leq \beta - 1 \qquad i = 2, 3, \ldots, t \\ -m \leq e &\leq M \qquad m > 0 \end{aligned}$$

La primera condición es la característica de los sistemas *normalizados* que utilizan la mayor parte de los ordenadores y que se concreta en que el valor absoluto del significando es tal que $\beta^{-1} \leq |s| < 1$.

1.2.2 Ahora vamos a ver, en un caso concreto, cuántos puntos tiene uno de estos sistemas y cómo se distribuyen. Comenzamos con el siguiente resultado:

TEOREMA

Sea $s = \pm.d_1 d_2 \ldots d_t$ un significando con mantisa normalizada en un sistema de coma flotante de base β, entonces existe un entero positivo N tal que

$$|s| = N \times \beta^{-t} \qquad y \qquad \beta^{t-1} \leq N < \beta^t$$

Demostración. En efecto, puesto que

$$|s| = .d_1 d_2 \dots d_t, \qquad \text{siendo } d_i \text{ dígitos en base } \beta$$

esto significa que

$$|s| = \frac{d_1}{\beta} + \frac{d_2}{\beta^2} + \dots + \frac{d_t}{\beta^t}$$

luego $|s| \times \beta^t = d_1\beta^{t-1} + d_2\beta^{t-2} + \dots + d_t = N$, un entero positivo. Evidentemente se verifica $N \geq \beta^{t-1}$, pues $d_1 \geq 1$ por hipótesis y por otra parte

$$N \leq (\beta - 1)(1 + \beta + \beta^2 + \dots + \beta^{t-1}) = (\beta - 1)\frac{(\beta^t - 1)}{(\beta - 1)} = \beta^t - 1.$$

valor este último que se alcanza cuando $d_1 = d_2 = \ldots = d_t = \beta - 1$. □

Si tenemos en cuenta que el entero más pequeño del rango, β^{t-1} también se alcanza cuando $d_1 = 1$ y $d_2 = d_3 = \ldots = d_t = 0$, tenemos una cantidad de enteros $\beta^t - \beta^{t-1} = \beta^{t-1}(\beta - 1)$, idéntica a la de mantisas existentes (¿por qué?) y no resulta difícil establecer la correspondencia biunívoca entre las mantisas del sistema y los enteros positivos en el rango establecido por el teorema anterior.

Puesto que los exponentes pueden variar entre $-m$ y M, ambos inclusive, tenemos que se dispone de $M + m + 1$ diferentes (incluyendo el cero), con lo que el montante total de números en el sistema en cuestión vendrá dado por la fórmula $\beta^{t-1}(\beta-1)\times(M+m+1)$, si nos atenemos a los positivos. Por tanto, considerando el doble signo y el cero alcanzamos la cantidad total, en función de β, t, m y M

$$F = F(\beta, t, m, M) = 2[\beta^{t-1}(\beta - 1)(M + m + 1)] + 1$$

La expresión del cero difiere de unos sistemas a otros, pero en general es simplemente 0 ó +0, pero en algunas ocasiones es $0 \times \beta^{-m}$.

1.2.3 Ilustremos ahora con un ejemplo como se distribuyen estos puntos en el rango de recta real que son capaces de representar. Tomamos los valores siguientes: $\beta = 2, t = 3, m = 1, M = 2$. Tenemos que $F(2, 3, 1, 2) = 33$ puntos, de los que vamos a representar el cero y los 16 positivos, en cuatro bloques de cuatro asociados con cada uno de los exponentes -1, 0, 1 y 2. Los números naturales con que se corresponden las respectivas mantisas, que deben ser cuatro también para que todo cuadre, son los comprendidos entre $2^2 = 4$ y $2^3 - 1 = 7$, como se ve en la siguiente tabla

N	Mantisa	$N \times 2^{-3}$ = Abscisa
4	$(.100)_2$	0'5
5	$(.101)_2$	0'625
6	$(.110)_2$	0'750
7	$(.111)_2$	0'875

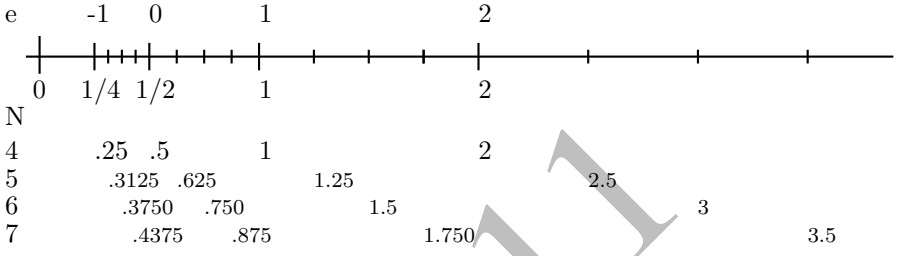
En el gráfico inferior se puede observar la asociación de cada número del sistema con la pareja que forman el entero correspondiente N y el exponente e. Concretamente, si y es uno de estos números (positivo), resultará que

$$y = s \times \beta^e = N \times \beta^{-t}\beta^e = N \times \beta^{e-t}$$

y, en particular los elementos más pequeños de los bloques pertenecientes a cada exponente, resultan ser $\beta^{t-1} \times \beta^{e-t} = \beta^{e-1}$. Esta expresión nos resulta útil, pues es fácil darse cuenta de que los números que corresponden al mismo exponente forman un conjunto equiespaciado de puntos y el intervalo entre ellos será la diferencia entre los principios de cada bloque dividido por el número de mantisas que existen. Por tanto, los puntos de exponente e distan entre sí

$$\frac{\beta^e - \beta^{e-1}}{\beta^{t-1}(\beta - 1)} = \beta^{e-t}$$

relación que resulta fundamental para la acotación de los errores relativos que se cometen al utilizar el sistema de coma flotante.



1.3 Errores en un sistema de coma flotante normalizado

Visto que para un sistema de coma flotante (β, t, m, M) disponemos de una cantidad finita de puntos distribuidos a lo largo de la recta real de una forma irregular pero sistemática, con ellos tenemos que representar en el ordenador cualquier número real; y como este conjunto es infinito, para la mayor parte de ellos sólo podremos tener una aproximación. Parece lógico, a falta de ideas mejores, tomar el punto del sistema más cercano. Cuando hay dos equidistantes la práctica habitual es elegir el de mayor valor absoluto. Si x es un número real cualquiera, la aproximación así obtenida se denomina valor *redondeado*, x_R y la cantidad $|x_R - x|$ es el error de redondeo.

Resulta de sumo interés establecer una cota para estos errores de redondeo, para conocer la precisión de nuestros resultados. La simple observación del gráfico anterior nos muestra que estos errores son muy variables de tamaño, puesto que los puntos están muy desigualmente espaciados. Sin embargo, si podemos establecer cotas en función del exponente que corresponde a x_R en el sistema, ya que en esos tramos los intervalos tienen una longitud β^{e-t}, con lo que el error de redondeo estará acotado por la mitad de esta cantidad.

Pero una cota variable es siempre poco eficiente, y por consiguiente hemos de buscar algún tipo de cota que dependa sólo del sistema. A la vista de la distribución de los puntos (más separados cuanto mayores son), podemos esperar que tal vez el error relativo pueda cumplir este requisito. En efecto, teniendo en cuenta que el error relativo es el cociente entre el absoluto y el verdadero valor, y que entre todos los puntos del sistema que tienen exponente e, el más pequeño es β^{e-1}, tendremos que

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

que vemos depende únicamente de β y de t, es decir del sistema de coma flotante que estamos utilizando.

Una forma alternativa de tomar una aproximación a un número real x es la denominada *truncación*, x_C (C por la palabra inglesa '*chopping*') que es el elemento más próximo y de módulo inferior a x que pertenece al sistema de coma flotante con el que estamos trabajando. Resulta evidente que en este caso las cotas para los errores absoluto y relativo toman un valor doble

1.3.1 Unidad de redondeo. De hecho estas cotas para el error relativo nos permiten dar una expresión sistemática para los valores aproximados de los números reales x, que resulta muy útil para el análisis de los errores.

Si escribimos $fl(x)$ el resultado de redondear o truncar, tendremos que

$$fl(x) = x(1+\delta), \qquad \text{donde } |\delta| \leq \begin{cases} \frac{1}{2}\beta^{1-t} & \text{redondeo} \\ \beta^{1-t} & \text{truncación} \end{cases} \tag{1.2}$$

La notación se unifica escribiendo $|\delta| < u$, donde u es un número, denominado *unidad de redondeo* (aún en el caso de que se usase la truncación), que es característico de la máquina (del sistema de coma flotante, en definitiva) en la que estamos trabajando. De ahí que también se le denomine *epsilon de la máquina*.

La verdad es que la cota es con frecuencia desahogada, cuando los valores x están próximos a números del sistema, pero también es cierto que es la mejor que se puede obtener con validez general. Es un ejercicio interesante tratar de demostrar esto y ver si se alcanza en algún punto. Cuando se trabaja en base dos, la cota toma el aspecto especialmente simple de 2^{-t} para el redondeo y 2^{1-t} para el truncamiento. Es interesante observar que si escribimos esta última cantidad en la forma $2^{-(t-1)}$ resulta ser la cota para el redondeo con un dígito (bit en este caso) significativo menos. ¿Qué significado se le puede dar a esta circunstancia?

1.3.2 Los errores de los que hemos hablado hasta ahora son consecuencia directa de la limitación de cifras en la mantisa. Pero, obviamente, la limitación del rango del exponente (en general también a un cierto número de cifras) también acarrea dificultades al sistema de coma flotante. No parece lógico representar en un sistema números cuyo aproximado sistemático no pertenezca al mismo. Así, cualquier número al que al aplicar la estrategia de aproximación necesite un exponente superior al valor de M, produce el fenómeno de *owerflow*; y si, en las mismas circunstancias, exige un exponente inferior a $-m$ estamos ante un *underflow* (hemos conservado los términos ingleses por su implantación en la jerga informática y por la dificultad de encontrar una palabra castellana que resulte realmente equivalente). Es decir, existe un intervalo, fuera del cuál el sistema de coma flotante carece de significado y de la utilidad para el que se le ha construído.

Es fácil entender que resulta menos controlable el *overflow*, de ahí la mayor difusión del concepto y el distinto tratamiento que recibe de las diferentes aritméticas.

1.4 Aritmética de coma flotante

El problema de un sistema de coma flotante es su aritmética; es decir, la forma en que vamos a representar el resultado de realizar operaciones en el sistema. Aún centrándonos

en las operaciones elementales (sumar, restar, multiplicar y dividir) es evidente que no son internas en el sistema. En la gráfica del sistema $F(2,3,1,2)$ es fácil ver que la suma de dos de sus puntos no tiene por qué recaer en otro punto del sistema, y todos sabemos que el producto de dos números con cinco dígitos necesita, en general, diez para su representación.

Por tanto, aún partiendo de datos exactos, el resultado de estas operaciones sólo se puede representar en el sistema de forma aproximada. Por supuesto, sería deseable que el resultado de la computación fuese el correcto en el sistema; es decir, el elemento del sistema que aproxima al verdadero resultado en los términos ya repetidamente explicados.

Concretamente, si $\circ$ representa una de las operaciones básicas ($+$, -, $\times$, $\div$), y tenemos dos números reales x e y, esperamos que si no hay problemas con los exponentes (*overflow* y/o *underflow*)

$$fl(x \circ y) = (x \circ y)(1+\delta), \qquad \text{con } |\delta| \leq u \tag{1.3}$$

La mayoria de los sistemas de coma flotante lo verifican, al menos para el producto/división; y con mayores dificultades, con frecuencia insalvables, para la suma/diferencia. Son los conocidos fenómenos de *cancelación* y *alineación* (véase los ejercicios 1.5.5 y 1.5.6).

Pero tratemos de aclarar esto, porque el asunto no es trivial. En primer lugar, hemos de tener en cuenta que el ordenador normalmente no trabaja con x e y, sino con sus *representados* $fl(x)$, $fl(y)$. Así, en un sistema decimal con tres dígitos significativos, si queremos sumar los números .1234 y .5674, la máquina los tomará como .123 y .567 respectivamente. Al sumarlos, el ordenador obtiene .690 (exacto esta vez porque recae en otro elemento del sistema al no producirse arrastre). Por otra parte, la suma exacta de los datos iniciales es .6908, cantidad que el ordenador convertirá en .691, que se diferencia en el último dígito y es diferente en el sistema. Para conseguir que se verifique (1.3) se necesita utilizar los denominados *dígitos de guarda* que consiste en que el ordenador trabaje internamente con más precisión de la que muestra externamente (véase el ejercicio 1.5.9).

La cuestión es que, al trabajar con datos aproximados, debemos de considerar el comportamiento de los errores en la aritmética exacta. Recordemos que para la suma(diferencia) los errores *absolutos* se suman, mientras que para la multiplicación(división) son los *relativos* los que se acumulan (de forma muy aproximada). Puesto que un sistema de coma flotante se basa en que la cota de los errores relativos sea constante, resulta explicable que las cosas vayan mejor para el producto que para la suma.

Vamos a analizar teóricamente lo que ocurre. Nuestros dos operandos x e y se representan como $x(1+\delta_1)$ e $y(1+\delta_2)$ respectivamente. El producto exacto es $xy(1+\delta_1+\delta_2+\delta_1\delta_2)$ que nos sugiere una cota de error relativo doble de la cometida en los datos, al ser el producto de las *deltas* despreciable frente a su suma (¿por qué?). Tomando pues una unidad de redondeo interna que sea la mitad que la que percibe el usuario se conseguirá que (1.3) se cumpla para la multiplicación.

Más complejo es el tema para la suma, en vista de que

$$x(1+\delta_1)+y(1+\delta_2) = x+y+x\delta_1+y\delta_2 = (x+y)\left(1+\frac{x\delta_1+y\delta_2}{x+y}\right)$$

es evidente que la cota para el error relativo puede tomar valores extremadamente gran-

des si el denominador $x + y$ tiende a ser muy pequeño en virtud de los fenómenos de cancelación, alineación y otros.

También es fácil ver que ni siquiera se verifica la más elemental de las propiedades de la suma (la asociatividad), por ejemplo en un sistema con seis dígitos significativos, el cálculo $.111113 \times 10^6 - .111111 \times 10^6 + .551111 \times 10^1$ da distinto resultado operando primero con los dos primeros números y después con el tercero, que si operamos el primero con el resultado obtenido de los dos últimos (comprobarlo detenidamente).

Así pues, la expresión $fl(x + y + z)$ es ambigua pues $fl(fl(x + y) + z)$ es en general diferente de $fl(x + fl(y + z))$. Cualquiera de las dos opciones que tomemos es igualmente válida pues nos llevan a resultados erróneos, y el objetivo es estimar cotas del error utilizando la variante que más nos facilite esta labor. Para el tipo de análisis que vamos a explicar a continuación vamos a utilizar la primera; es decir, que empezaremos a sumar por el principio.

En vista de todos los factores que influyen, es fácil comprender lo difícil que resulta predecir los errores que se van a producir en una computación más o menos compleja. Pero es una necesidad tan esencial del cálculo numérico, que se han sugerido múltiples soluciones para controlar estos errores o al menos conocer una cota. De entre ellas consideraremos detenidamente la técnica denominada *análisis regresivo del error* que se basa en el siguiente paradigma: "El resultado de cualquier operación efectuada con error, se puede obtener de forma exacta a partir de unos datos erróneos". Este principio no resulta difícil de aceptar y, en su concepción más general es incluso una obviedad que dicho resultado se puede obtener de una infinidad de maneras. Se trata de ver que una de esas posibilidades consiste en realizar los cálculos con datos que difieren poco de los exactos.

1.4.1 Análisis regresivo del error. La idea básica del análisis regresivo del error es hacer recaer todo el posible error en el resultado de efectuar una serie de operaciones sobre errores en los datos. Se considera por tanto que el susodicho resultado se obtiene operando con aritmética exacta sobre unos datos ligeramente perturbados. La fuerza del método radica en el hecho de que con frecuencia los datos ya vienen con errores derivados de mediciones o a resultas de operaciones anteriores frente a las que estas ligeras perturbaciones son insignificantes.

De hecho el análisis regresivo nos sirve para liberar al ordenador (a su sistema de coma flotante, por supuesto) de culpa en los errores, a veces inaceptables, que se producen. Si lo que obtenemos es el resultado exacto de una ligera perturbación, el método utilizado es numéricamente *estable* y los resultados claramente insatisfactorios serán debidos más al propio problema que a la aritmética, como se verá claramente en las próximas lecciones.

1.4.2 Veamos con un ejemplo, con aplicación directa en las mismas, el mecanismo del análisis. Consideramos el producto escalar de dos vectores n-dimensionales $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ e $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, en los que podemos pensar como una fila y una columna de algunas matrices que estemos multiplicando. Sabemos que

$$\mathbf{x} \cdot \mathbf{y} = \sum_{i=1}^{n} x_i \times y_i$$

un valor que no obtendremos exactamente. Para facilitar el analisis, vamos a denominar

$$p_j = fl(p_{j-1} + fl(x_j \times y_j)), \qquad j = 2, 3, \ldots, n, \qquad p_1 = fl(x_1 \times y_1)$$

y vamos a suponer que los productos básicos son exactos en el sistema, de forma que se verifica $fl(x_i \times y_i) = x_i y_i(1 + \delta_i), |\delta_i| \leq u$ para todos los índices.

Entonces, y teniendo en cuenta que el resultado de nuestro ordenador será p_n, tenemos

$$\begin{aligned} p_2 &= fl(fl(x_1 \times y_1) + fl(x_2 \times y_2)) = fl(x_1 y_1(1+\delta_1) + x_2 y_2(1+\delta_2)) \\ &= (x_1 y_1(1+\delta_1) + x_2 y_2(1+\delta_2))(1+\epsilon_1) \\ &= x_1 y_1(1+\delta_1)(1+\epsilon_1) + \\ &\quad x_2 y_2(1+\delta_2)(1+\epsilon_1) \end{aligned}$$

donde $|\epsilon_1| \leq u$. De forma similar

$$\begin{aligned} p_3 &= fl(p_2 + fl(x_3 \times y_3)) = (p_2 + x_3 y_3(1+\delta_3))(1+\epsilon_2) \\ &= x_1 y_1(1+\delta_1)(1+\epsilon_1)(1+\epsilon_2) + \\ &\quad x_2 y_2(1+\delta_2)(1+\epsilon_1)(1+\epsilon_2) + \\ &\quad x_3 y_3(1+\delta_3)(1+\epsilon_2) \end{aligned}$$

y continuando hasta el final, siempre con $|\epsilon_i| \leq u$, resulta que

$$\begin{aligned} p_n &= (p_{n-1} + x_n y_n(1+\delta_n))(1+\epsilon_{n-1}) \\ &= x_1 y_1(1+\delta_1)(1+\epsilon_1)(1+\epsilon_2)\cdots(1+\epsilon_{n-1}) + \\ &\quad x_2 y_2(1+\delta_2)(1+\epsilon_1)(1+\epsilon_2)\cdots(1+\epsilon_{n-1}) + \\ &\quad x_3 y_3(1+\delta_3)(1+\epsilon_2)\cdots(1+\epsilon_{n-1}) + \\ &\quad \ldots \\ &\quad x_{n-1} y_{n-1}(1+\delta_{n-1})(1+\epsilon_{n-2})(1+\epsilon_{n-1}) + \\ &\quad x_n y_n(1+\delta_n)(1+\epsilon_{n-1}) \end{aligned}$$

Si definimos las cantidades η_i de forma que

$$\begin{aligned} 1+\eta_1 &= (1+\delta_1)(1+\epsilon_1)(1+\epsilon_2)\cdots(1+\epsilon_{n-1}) \\ 1+\eta_2 &= (1+\delta_2)(1+\epsilon_1)(1+\epsilon_2)\cdots(1+\epsilon_{n-1}) \\ 1+\eta_3 &= (1+\delta_3)(1+\epsilon_2)\cdots(1+\epsilon_{n-1}) \\ &\ldots \\ 1+\eta_{n-1} &= (1+\delta_{n-1})(1+\epsilon_{n-2})(1+\epsilon_{n-1}) \\ 1+\eta_n &= (1+\delta_n)(1+\epsilon_{n-1}) \end{aligned}$$

podemos escribir la relación más expresiva y coherente con lo que buscamos

$$p_n = x_1 y_1(1+\eta_1) + x_2 y_2(1+\eta_2) + \ldots + x_{n-1} y_{n-1}(1+\eta_{n-1}) + x_n y_n(1+\eta_n)$$

nuestra solución como suma *exacta* de perturbaciones de los sumandos. Para medir (acotar) el tamaño de estas perturbaciones, tenemos el siguiente resultado que enunciamos sin demostración

PROPOSICIÓN

Si $nu \leq .01$ y $|\delta_i| \leq u$ $(i = 1, 2, \ldots, n)$, entonces

$$1 - nu \leq \prod_{i=1}^{n}(1 + \delta_i) \leq 1 + 1.01nu \tag{1.4}$$

Esto quiere decir que $|\eta_i| \leq 1.01nu$ y de esta forma se consigue acotar efectivamente el error final en función de las cotas de los errores en los datos. La condición $nu \leq .01$ no es excesivamente exigente, pues es normal trabajar con valores de $u = 10^{-15}$ lo que permite unos valores no habituales de n. Relajando dicha exigencia a $nu \leq .1$, la cota aumenta hasta $|\eta_i| \leq 1.06nu$

1.5 Cuestiones y problemas

1.5.1 Calcular el valor del número $(.36207)_8$ en base 10, pero trabajando siempre en base 8. Así mismo, calcular la expresión del número decimal 0.1 en base 8, trabajando en esta última base. Contrastar los resultados con los obtenidos trabajando en base 10.

1.5.2 Calcular la expresión decimal de los números $(.\overline{0011})_2$ y $(.\overline{1})_2$. Encontrar una relación con el ejercicio anterior.

1.5.3 Hallar un número binario que aproxime π hasta 10^{-3}.

1.5.4 Como es sabido, la suma de la serie armónica $1 + 1/2 + 1/3 + \ldots$ se va haciendo infinita. Pero si se calcula con precisión finita por un ordenador, la suma existe, en cierto sentido, ya que los términos se van haciendo tan pequeños que no contribuyen a la suma si se les añade de uno en uno. Por ejemplo, supongamos que calculamos la suma redondeada a una cifra decimal; entonces tenemos $1 + 0.5 + 0.3 + 0.3 + 0.2 + 0.2 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 + 0.1 = 3.9$.

Más precisamente, sea $r_n(x)$ el número x redondeado a n cifras decimales; definimos $r_n(x) = \lfloor 10^n x + 1/2 \rfloor / 10^n$. Entonces queremos hallar

$$S_n = r_n(1) + r_n\left(\frac{1}{2}\right) + r_n\left(\frac{1}{3}\right) + \cdots;$$

sabemos que $S_1 = 3.9$, y el problema consiste en escribir un programa que calcule e imprima S_n para $n = 2$, 3, 4 y 5.

Nota: Existe una manera mucho más rápida de hacer esto en vez de ir sumando $r_n(1/m)$, de uno en uno, hasta que $r_n(1/m)$ sea cero. (Por ejemplo, tenemos $r_5(1/m) = 0.00001$ para todos los m desde 66667 hasta 200000. Es una buena idea evitar calcular $1/m$ 133334 veces !) Sería mejor utilizar el siguiente algoritmo:

A. Empezar con $m_h = 1$ y $S = 1$.

B. Hacer $m_e = m_h + 1$ y calcular $r_n(1/m_e) = r$.

C. Hallar m_h el mayor m para el que $r_n(1/m) = r$.

D. Añadir $(m_h - m_e + 1)r$ a S y volver a B.

1.5.5 Efectuar las siguientes operaciones de coma flotante, y ver lo que ocurre con los errores relativos de las respectivas soluciones:

a) $.76545421 \times 10^1 - .76544200 \times 10^1$ con 8 dígitos significativos.

b) $1 - .999999$ con seis dígitos significativos.

c) $1 - .999999$ con siete dígitos significativos.

1.5.6 Para algunos valores de x, la función $f(x) = \sqrt{x^2 + 1} - x$ no se puede computar exactamente usando esta expresión. Explicarlo, y encontrar un método de superar esta dificultad.

1.5.7 En general, al sumar una lista de números en un sistema de coma flotante, se producir un error más pequeño si los sumandos se añaden en orden de magnitud creciente. Dar algunos ejemplos para ilustrar este principio.

Sin embargo, esto no es siempre cierto. Considerar un sistema de 2 dígitos significativos, y demostrar que los cuatro números 0.25, 0.0034, 0.00051 y 0.061 se pueden sumar con menos error en un orden distinto del ascendente. ¿Por qué?

1.5.8 Considérese una computadora donde los números de coma flotante tengan mantisas de 8 dígitos, seguidas por exponentes de 2 dígitos, en la cual los exponentes están expresados en la forma de 50 en exceso. Considérese que dos números tales son:

$$4735821653\ (.47358216 \times 10^3)$$
$$3294175251\ (.32941752 \times 10^1)$$

a) Sume estos dos números y ponga el resultado en la misma forma de coma flotante.

b) Si ambos números están dados en forma exacta, ¿cuál es el error de la suma?

c) Si ambas mantisas están correctamente redondeadas como se muestra, calcular una cota para el error en la suma determinada en el inciso a).

1.5.9 Supongamos un ordenador que externamente (de cara al usuario) trabaja en un sistema de coma flotante $F(10, 9, 38, 38)$, e internamente (en lenguaje de máquina) en otro diferente $F(2, 32, 127, 127)$.

a) ¿Cuál de los dos sistemas tiene mayor número de puntos? ¿Está uno de ellos contenido en el otro?

b) Calcular los correspondientes rangos y unidades de error de truncación, y decir cuál de los dos sistemas permite mayor precisión.

c) ¿Podría presentar el ordenador un mayor número de dígitos decimales exactos? ¿Podría prescindir de alguna de las cifras binarias para sus cálculos, si se limita a las nueve decimales? En caso afirmativo, decir de cuántas, y qué sentido puede tener el conservar todas.

1.5.10 Probar que

$$\lg x \approx \ln x + \log x,$$

con error menor que 1 %! (Por tanto, puede usarse una tabla de logaritmos naturales y una de logaritmos vulgares para obtener aproximadamente los logaritmos binarios.)

Lección 2

Normas vectoriales y matriciales

2.1 Significado numérico de la norma

Hemos visto en la lección anterior como analizar los errores que se producen al operar en un sistema de coma flotante normalizado. Allí nos referíamos a operaciones con escalares reales, donde el concepto de valor absoluto está perfectamente establecido. Ahora bien, en la práctica de los métodos numéricos casi siempre tendremos que trabajar con varios números reales que tienen un significado conjunto, y para los que no es tan obvia una idea global tamaño.

Así, en el ámbito del algebra lineal, las operaciones con vectores y matrices realizadas en un sistema numérico de como flotante producen errores que el cálculo numérico se encarga de analizar. Por ejemplo, es de gran interés conocer en que medida los errores en los coeficientes perturban el resultado de un sistema de ecuaciones lineales.

Naturalmente, una posibilidad sería contemplar los errores de forma individual para cada una de las componentes de los datos y su influencia en cada componente de los resultados; pero, además de resultar poco práctico por la gran cantidad de elementos implicados (téngase en cuenta que una matriz cuadrada $n \times n$ alberga n^2 elementos), la interpretación se hace prácticamente imposible. Por tanto, se tiende a concretizar la medida de vectores o matrices en un sólo número real, llamado *norma*, que nos permita realizar un análisis de errores similar al utilizado en los escalares.

2.1.1 Una *norma vectorial* es una aplicación $\|\cdot\| : \mathbb{R}^n \longrightarrow \mathbb{R}$ que satisface las siguientes propiedades:

1. $\mathbf{x} \neq 0 \Longrightarrow \|\mathbf{x}\| > 0$,
2. $\|\alpha\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$,
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

La primera condición nos dice que el tamaño de un vector no nulo es positivo (es evidente que el vector nulo tiene tamaño 0, ¿por qué?). La segunda que al multiplicar un vector por un escalar su tamaño cambia en la misma proporción. La tercera es una generalización de la popular propiedad de los triángulos que afirma que uno cualquiera de sus lados es siempre menor que la suma de los otros dos. De hecho se denomina

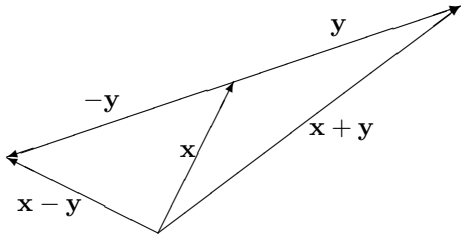


Figura 2.1: La desigualdad triangular

desigualdad triangular. Una variante útil de esta propiedad es la siguiente desigualdad (véase la figura 2.1)

$$\|\mathbf{x}-\mathbf{y}\| \geq \|\mathbf{x}\| - \|\mathbf{y}\|$$

2.1.2 La función valor absoluto definida en la recta real, verifica las condiciones impuestas a una norma vectorial y por tanto representa una norma sobre $\mathbb{R}^1$. De igual forma, la función módulo de un número complejo también es una norma sobre el espacio vectorial $\mathbb{C}^1$. De hecho, generalmente consideraremos que estamos trabajando con escalares complejos; es decir, que nuestros vectores y matrices tienen sus elementos en el cuerpo $\mathbb{C}$.

2.2 Normas vectoriales usuales

Aunque existen infinitas normas vectoriales sobre $\mathbb{C}^n$ (o $\mathbb{R}^n$), las que aparecen más frecuentemente por su utilidad y sentido geométrico son las denominadas *uno*, *dos* e *infinito*. Se definen de la siguiente manera:

1. $\|\mathbf{x}\|_1 = \sum_{i=1}^{n} |x_i|,$

2. $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^{n} |x_i|^2},$

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

2.2.1 En $\mathbb{R}^3$ la norma *dos* de un vector es su longitud *euclídea*, por lo que dicha norma se denomina también *norma euclídea*. La norma *infinito* recibe también el nombre de *norma del máximo* por razones obvias. Las tres son muy sencillas de computar, y satisfacen diversas relaciones entre las que podemos destacar

$$\|\mathbf{x}\|_1 \geq \|\mathbf{x}\|_2 \geq \|\mathbf{x}\|_\infty$$

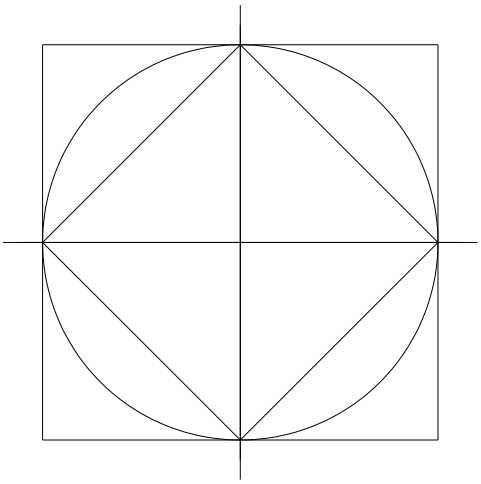


Figura 2.2: Superficies de las bolas unidad

que admite una interpretación dual en términos del tamaño de la bola unidad (conjunto de vectores cuya norma es menor o igual que la unidad) para cada una de las normas. Lógicamente la bola será mayor cuanto menor sea la respectiva norma. En la figura 2.2 se pueden ver las respectivas superficies para las tres normas en el espacio $\mathbb{R}^2$. Obviamente el disco corresponde a la norma *dos*, ¿a cuál de las otras dos normas corresponden cada uno de los cuadrados y por qué?

2.2.2 A veces puede dar la sensación de que es una complicación gratuita considerar múltiples normas en un mismo espacio, teniendo en cuenta además que, desde el punto de vista matemático y sobre espacios de dimensión finita, todas ellas resultan ser equivalentes en el sentido topológico. Pero, aparte de que determinados resultados sean más fáciles de probar en una u otra norma, hay que tener en cuenta que muchas de ellas tienen un significado propio (como hemos visto para la norma euclídea), que no comparten las demás.

Por ejemplo, ¿qué vector es de más tamaño $(1,1,1,1,1,1)^T$ o $(2,0,0,0,0,0)^T$? Si las componentes de los vectores representan, en miles de euros, los ingresos de una familia en los meses de enero a junio de 1990, es claro que el primer vector es de más tamaño que el segundo. Si las componentes de los vectores representan, en decimas de milímetro, los errores de fabricación en los diámetros de seis piezas de un reloj y las piezas con error superior a 0.15 milímetros son inaceptables, entonces el primer vector debe considerarse de menor tamaño que el segundo.

2.3 Normas matriciales

Ya hemos visto cómo el Análisis Numérico necesita dotar de normas a los espacios vectoriales que maneja. Cuando se trabaja con el espacio M_n de las matrices complejas $n \times n$

sería posible identificar dicho espacio con $\mathbb{C}^m$, $m = n^2$ a través de

$$A = (a_{ij})_{1 \leq i,j \leq n} \rightarrow [a_{11}, \ldots, a_{1n}, a_{21}, a_{22}, \ldots, a_{2n}, \ldots, a_{nn}]^T,$$

para trasladar a M_n las normas usuales en $\mathbb{C}^m$. Así de la norma euclídea en $\mathbb{C}^m$ construiríamos la norma matricial

$$\|A\| = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2} \tag{2.1}$$

llamada de Frobenius, y habitualmente denotada como $||.||_F$. ¿Qué normas se obtienen en M_n al transportar las normas uno e infinito de $\mathbb{C}^m$?

2.3.1 Sin embargo, las normas que se obtienen por este procedimiento *tienen poco o nulo interés*, toda vez que no toman en consideración la interpretación de los elementos de M_n como operadores lineales en $\mathbb{C}^n$. Veamos a continuación el proceso a seguir para construir normas en M_n que sean útiles.

2.3.2 Sea $||.||$ una norma cualquiera en $\mathbb{C}^n$. Tomemos una matriz $A \in M_n$. Para cada vector $\mathbf{x} \neq \mathbf{0}$ formemos el cociente $||A\mathbf{x}||/||\mathbf{x}||$ entre la longitud del vector después y antes de transformarse según A (dilatación sufrida por $\mathbf{x}$). Claramente tal dilatación es la misma para $\mathbf{x}$ y cada vector $\lambda\mathbf{x}$ con λ escalar no nulo. En otras palabras la dilatación depende de la *dirección* de $\mathbf{x}$ y no de su longitud. Cuando $\mathbf{x}$ recorre la esfera unidad $||\mathbf{x}|| = 1$ la dilatación $||A\mathbf{x}||/||\mathbf{x}|| = ||A\mathbf{x}||$ ha de tener un máximo no negativo, ya que dicha esfera es compacta. Pues bien, es inmediato comprobar que la aplicación que asocia a cada $A \in M_n$ la máxima dilatación

$$A \rightarrow \max\{||A\mathbf{x}|| : ||\mathbf{x}|| = 1\} = \max\{||A\mathbf{x}||/||\mathbf{x}|| : \mathbf{x} \neq \mathbf{0}\}$$

define una norma en M_n. (¡Compruébelo!) La norma así construida se llama matricial asociada a la vectorial de partida o deducida de la vectorial de partida. Suele denotarse por el mismo símbolo $||.||$, toda vez que no puede haber ambigüedad.

2.3.3 En lo sucesivo sólo manejaremos normas matriciales que hayan sido deducidas de una vectorial por el proceso anterior. Tales normas no sólo verifican las propiedades

$$||A|| > 0 \text{ si } A \neq 0, ||\lambda A|| = |\lambda|||A||, ||A + B|| \leq ||A|| + ||B||$$

sino también las tres siguientes

$$\text{si } I \text{ es la matriz identidad } ||I|| = 1, \tag{2.2}$$

$$\text{si } A \in M_n, \mathbf{x} \in \mathbb{C}^n \text{ entonces } ||A\mathbf{x}|| \leq ||A||||\mathbf{x}||, \tag{2.3}$$

$$\text{si } A, B \in M_n \text{ entonces } ||AB|| \leq ||A||||B||. \tag{2.4}$$

2.4 Expresiones de ciertas normas matriciales usuales

Aunque la dilatación máxima sirve para definir $||A||$ no es útil para evaluar $||A||$ en la práctica. Veremos en esta sección como hallar $||A||$ en ciertos casos importantes.

2.4.1 La norma en M_n deducida de la del máximo en $\mathbb{C}^n$ está dada por

$$\|A\|_\infty = \max_j \sum_{k=1}^{n} |a_{jk}| \tag{2.5}$$

(para cada fila sumar valores absolutos de los elementos y retener la suma de la fila que la tenga mayor).

Para probar (2.5), sea $\mathbf{x} \neq \mathbf{0}$

$$\begin{aligned} \|A\mathbf{x}\|_\infty &= \max_j \left| \sum_{k=1}^{n} a_{jk} x_k \right| \leq \max_j \sum_{k=1}^{n} |a_{jk}|\,|x_k| \\ &\leq \max_j \sum_{k=1}^{n} |a_{jk}|\,\|\mathbf{x}\|_\infty = \|\mathbf{x}\|_\infty \max_j \sum_{k=1}^{n} |a_{jk}|. \end{aligned}$$

Por tanto $\mathbf{x}$ se dilata en no más del segundo miembro de (2.5). Como $\mathbf{x}$ es arbitrario hemos probado que

$$\|A\|_\infty \leq \max_j \sum_{k=1}^{n} |a_{jk}| \tag{2.6}$$

Recíprocamente, denotemos por J el índice de la fila en que se alcanza el máximo del segundo miembro de (2.5). Formemos el vector $\mathbf{y}$ de componentes, $1 \leq k \leq n$

$$y_k = \begin{cases} a^*_{Jk}/|a_{Jk}|, & \text{si} \quad a_{Jk} \neq 0 \\ 0, & \text{si} \quad a_{Jk} = 0 \end{cases}$$

(la estrella denota complejo conjugado). Claramente $||\mathbf{y}||_\infty = 1$, a menos que $a_{jk} = 0$, $1 \leq k \leq n$ en cuyo caso debe ser $A = 0$ (¿por qué?) y (2.5) es obvio. Dejando este caso trivial, podremos escribir

$$\begin{aligned} \|A\mathbf{y}\|_\infty &= \max_j \left| \sum_{k=1}^{n} a_{jk} y_k \right| \geq \left| \sum_{k=1}^{n} a_{Jk} y_k \right| = \sum_{k=1}^{n} |a_{Jk}| \\ &= \max_j \sum_{k=1}^{n} |a_{jk}| = \max_j \sum_{k=1}^{n} |a_{jk}|\,\|\mathbf{y}\|_\infty , \end{aligned}$$

lo cual establece la desigualdad contraria de la (2.6).

2.4.2 Para la norma en M_n definida de la norma $||.||_1$ en $\mathbb{C}^n$ puede demostrarse que

$$\|A\|_1 = \max_k \sum_{j=1}^{n} |a_{jk}|, \tag{2.7}$$

(para cada columna sumar valores absolutos de los elementos y retener la mayor suma).

2.4.3 La norma más interesante es la deducida de la euclídea (no la confunda con (2.1)). Si $A \in M_n$, su conjugada hermítica A^H es la matriz que se obtiene conjugando cada elemento de la traspuesta A^T . Si $A \in M_n$, se define su *radio espectral* $\rho(A)$ por

$$\rho(A) = \max_s |\lambda_s|, \quad \lambda_s \text{ autovalor de } A. \tag{2.8}$$

Con estas notaciones podemos probar que

$$||A||_2 = \rho(A^H A)^{\frac{1}{2}}. \tag{2.9}$$

En efecto, sea $\mathbf{x}$ distinto de $\mathbf{0}$. Al ser $A^H A$ hermítica (es decir $(A^H A)^H = A^H A$) tiene una base ortonormal de autovectores $\mathbf{u}_j$, $1 \leq j \leq n : A^H A\mathbf{u}_j = \lambda_j \mathbf{u}_j$ (los λ_j son reales). Si

$$\mathbf{x} = \sum_{j=1}^{n} \alpha_j \mathbf{u}_j$$

es la expresión de $\mathbf{x}$ en la base $\{\mathbf{u}_j\}$, podremos escribir

$$\begin{aligned} \|A\mathbf{x}\|_2^2 &= (A\mathbf{x})^H A\mathbf{x} = \mathbf{x}^H A^H A\mathbf{x} = \left(\sum_{j=1}^{n} \alpha_j^* \mathbf{u}_j^H\right) A^H A \left(\sum_{j=1}^{n} \alpha_j \mathbf{u}_j\right) \\ &= \sum_{j,k=1}^{n} \alpha_j^* \alpha_k \mathbf{u}_j^H A^H A\mathbf{u}_k = \sum_{j,k=1}^{n} \alpha_j^* \alpha_k \mathbf{u}_j^H \lambda_k \mathbf{u}_k \\ &= \sum_{j=1}^{n} |\alpha_j|^2 \lambda_j \leq \rho(A^H A) \sum_{j=1}^{n} |\alpha_j|^2 = \rho(A^H A)\|\mathbf{x}\|_2^2, \end{aligned}$$

de modo que

$$||A||_2^2 \leq \rho(A^H A). \tag{2.10}$$

Para probar la desigualdad contraria, sea $\mathbf{u} \neq \mathbf{0}$ un autovector de $A^H A$ para el que el módulo del autovalor correspondiente λ coincida con $\rho(A^H A)$. Tendremos

$$||A\mathbf{u}||_2^2 = \mathbf{u}^H A^H A\mathbf{u} = \mathbf{u}^H \lambda \mathbf{u} = \lambda ||\mathbf{u}||_2^2$$

Claramente λ es ≥ 0 como cociente de cantidades no negativas. Por ello

$$||A\mathbf{u}||_2^2 = \rho(A^H A)||\mathbf{u}||_2^2,$$

lo que, junto con (2.10), implica (2.9).

2.5 Relación entre radio espectral y norma de una matriz

Sea $||.||$ una norma matricial deducida de una vectorial. Si $A \in M_n$ y $\lambda \in Spec(A)$ (es decir λ es autovalor de A) existe un vector $\mathbf{x} \neq \mathbf{0}$ para el que $A\mathbf{x} = \lambda\mathbf{x}$.

Así

$$\|A\| \geq \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \frac{|\lambda|\|\mathbf{x}\|}{\|\mathbf{x}\|} = |\lambda|$$

y en conclusión:

$$\rho(A) \leq ||A||. \tag{2.11}$$

Por otro lado en el caso *muy particular* que $||.||$ sea la norma euclídea y A hermítica $A^H = A$, tendremos

$$||A||_2 = \rho(A^H A)^{1/2} = \rho(A^2)^{1/2} = \rho(A),$$

ya que es bien conocido que los autovalores de A^2 son los cuadrados de los autovalores de A. Es decir, en ese caso particular, norma y radio espectral coinciden mientras que en general sólo vale la relación (2.11).

2.6 Cuestiones y problemas

2.6.1 Demuestre las propiedades (2.2) - (2.4).

2.6.2 ¿Hay alguna norma vectorial en $\mathbb{C}^n$ para la cual la matricial asociada sea la de Frobenius (2.1)?

2.6.3 Sea $||.||$ una norma en $\mathbb{C}^n$, B la bola unidad $\{||\mathbf{x}|| \leq 1\}$, A una matriz compleja $n \times n$ y AB la imagen de B por la transformación lineal inducida por A. ¿Cuál es el radio de la mínima bola $\{||\mathbf{x}|| \leq r\}$ que contiene a AB?

2.6.4 Pruebe la relación (2.7).

2.6.5 Dé ejemplos de matrices 2×2 para las que la norma euclídea no coincida con el radio espectral. Dé ejemplos en los que $\rho(A) = 0$, $||A||$ =1000.

2.6.6 Pruebe que para cada matriz cuadrada $||A||_2 \leq (||A||_1 ||A||_\infty)^{1/2}$. (*Indicación:* use (2.11).)

2.6.7 Demostrar que si $||A|| < 1$ en alguna norma matricial, $\ker(I - A) = \{\mathbf{0}\}$ y, en consecuencia, $I - A$ es regular.

2010-11

Lección 3

Errores en la resolución numérica de sistemas lineales

3.1 Introducción

Consideraremos el problema de hallar la solución del sistema

$$A\mathbf{x} = \mathbf{b}, \tag{3.1}$$

donde $\mathbf{b}$ es un vector conocido en $\mathbb{R}^n$, A una matriz real conocida $n \times n$ y $\mathbf{x}$ un vector en $\mathbb{R}^n$ desconocido.

La resolución de (3.1) es sin duda el problema más frecuente en la práctica de la computación, y por consiguiente también el más importante:

(i) Muchas veces nos encontramos con ecuaciones diferenciales o en derivadas parciales cuyas incógnitas son funciones de una o varias variables. Para poder tratar numéricamente dichas problemas hay que *discretizarlos*, es decir sustituirlos por otros cuya incógnita sea un vector $\mathbf{x}$ con un número finito de componentes: los ordenadores son máquinas finitas.

(ii) Si bien los sistemas que involucran un vector $\mathbf{x}$ pueden ser no lineales $\mathbf{F}(\mathbf{x}) = \mathbf{b}$, la resolución numérica de sistemas no lineales suele llevarse a cabo por *linealización*: es decir, resolviendo sistemas lineales que aproximen al dado.

(iii) La aproximación de mínimos cuadrados lleva aparejada, cuando no se dispone de una base ortogonal, la resolución de un sistema, cuya matriz es simétrica y definida positiva.

En conclusión, bien porque el problema propuesto sea de entrada de la forma (3.1) (lineal y discreto), bien porque problemas continuos se reduzcan a dicha forma por discretización y/o linealización, o bien porque se llegue a planteamientos de este tipo por otros conductos, es usual que hayamos de resolver un sistema como (3.1).

3.1.1 Además de la importancia que tiene el estudio numérico de (3.1) a consecuencia de lo frecuente de su aparición, es de destacar que el estudio del Algebra Lineal Numérica es de elevado valor metodológico para el analista numérico, ya que ilustra claramente las peculiaridades del enfoque "numérico" de los problemas.

En efecto, desde el punto de vista *teórico*, (3.1) es de tratamiento fácil: hay solución única si y sólo si A es no singular, extremo que puede comprobarse con el criterio $\det(A) \neq 0$ (el determinante es un polinomio en los elementos de A, de expresión bien conocida). Cuando hay solución, ésta se escribe $\mathbf{x} = A^{-1}\mathbf{b}$. Los libros de álgebra lineal dan clásicamente fórmulas cerradas para los elementos de A^{-1} que involucran determinantes. Expresando A^{-1} de ese modo y operando en $A^{-1}\mathbf{b}$ se obtienen las fórmulas de Cramer que proporcionan las componentes de $\mathbf{x}$, en función de los elementos de A y los de $\mathbf{b}$. Sin embargo, desde el punto de vista numérico las cosas son mucho más complejas y sutiles de lo que acabamos de señalar, hasta el punto que la resolución de (3.1) sigue siendo objeto de múltiples investigaciones. Chocamos, al menos, con dos escollos:

(i) El costo en operaciones de la resolución puede ser prohibitivo si la dimensión n del problema es elevada, o si, aún siendo n pequeño, se escoge un mal algoritmo. Por ejemplo, el costo de la regla de Cramer crece como $n!$. Cuando $n = 20$, la regla de Cramer implementada en un ordenador que efectuase un millón de multiplicaciones por segundo, tardaría más de un millón de años en completar los cálculos.

(ii) El alto número de operaciones necesario y/o ciertas peculiaridades del sistema a resolver pueden hacer que los errores de redondeo, acumulándose, lleguen a ser tan grandes que la solución computada carezca de utilidad.

La primera dificultad está resuelta por el método de Gauss, que es en cierta forma óptimo en cuanto al número de operaciones. Además, sus variantes de *pivotaje* contribuyen a estabilizar el comportamiento de los errores de redondeo.

3.2 Acondicionamiento de un sistema lineal

Sin embargo, hay sistemas con condiciones específicas que dificultan su resolución numérica. Se dice de ellos que están *mal acondicionados*, queriendo significar que cambios *relativamente* pequeños en los datos, pueden provocar cambios *relativamente* grandes en la solución; o al menos pueden hacerlo si se utilizan los algoritmos habituales y/o no se toman las precauciones debidas.

Comencemos por un ejemplo. El sistema

$$\begin{array}{rcrcl} 2x & + & 6y & = & 8 \\ 2x & + & 6.00001y & = & 8.00001 \end{array} \tag{3.2}$$

tiene solución $x = 1$, $y = 1$. Supongamos que los coeficientes y segundo miembro se conocen sólo aproximadamente, por resultar de mediciones experimentales o haber sido obtenidos por alguna técnica numérica que exija aproximación, redondeo etc.. Si en vez de (3.2) obtuviésemos

$$\begin{array}{rcrcl} 2x & + & 6y & = & 8 \\ 2x & + & 5.99999y & = & 8.00002 \end{array} \tag{3.3}$$

equivocándonos en menos de una milésima por ciento, tendríamos como solución: $x = 10$, $y = -2$.

Es evidente el caracter de mal acondicionado del sistema (3.2). Ciertamente, el sistema (3.3) es diferente, y no debe extrañarnos que tenga una solución diferente; pero la enorme diferencia (*relativa*) entre las soluciones debe de alertarnos ante la posibilidad de

que las soluciones que calculemos difieran de las verdaderas en una proporción mucho mayor que los pequeños errores (controlados) que estamos dispuestos a asumir en los datos y/o en las operaciones de redondeo.

Téngase en cuenta que esta es una situación altamente indeseable, ya que no tenemos ninguna posibilidad de estimar, acotar o controlar directamente el error de la solución, al desconocer la verdadera. Sólo podremos conocer los efectos de esa imprecisión.

3.2.1 El mal acondicionamiento de (3.2) es fácil de interpretar geométricamente: Resolver

$$\begin{array}{rcrcl} \alpha x & + & \beta y & = & a \\ \gamma x & + & \delta y & = & b \end{array}$$

con $|\alpha|+|\beta|, |\gamma|+|\delta| > 0$, es encontrar en el plano $0xy$ la intersección de dos rectas. Tal intersección existe y es única si y sólo si las rectas no son paralelas, es decir $\alpha\delta - \beta\gamma \neq 0$. Sin embargo si $\alpha\delta - \beta\gamma$ es pequeño *en relación a* los coeficientes α, β, γ y δ, las rectas son casi paralelas (como en (3.2)). A causa de este *casi paralelismo*, el punto de intersección, que existe y es único, puede cambiar enormemente al cambiar un poco las rectas (es decir cambiar los coeficientes de las ecuaciones y/o los segundos miembros).

Aparece así una realidad más variada que la estudiada en los cursos de álgebra lineal. Allí sólo hay dos casos: a) las rectas tienen pendientes distintas (tal vez muy próximas entre sí, pero distintas) y hay solución única; b) las rectas tienen la misma pendiente y entonces hay infinitas soluciones (si coinciden) o ninguna (si no lo hacen). Desde el punto de vista numérico tendríamos tres situaciones:

(i) las pendientes son "muy distintas": hay solución única y se puede calcular con razonable aproximación (buen acondicionamiento);

(ii) las pendientes son distintas "pero no demasiado" (mal acondicionamiento): hay solución única pero muy sensible a errores;

(iii) las pendientes son iguales: no hay solución o hay infinitas.

La frontera entre el buen y mal acondicionamiento no es nítida sino fluida: hay sistemas que en cualquier circunstancia clasificaríamos como mal acondicionados, otros que siempre clasificaríamos como bien acondicionados (por ejemplo los que tienen por matriz la identidad) y otros que quedan en medio. Es un término, pues, que encierra una cierta relatividad como iremos aclarando en los apartados siguientes. Concretamente, puede depender de la norma usada, del escalado de los datos, etc.

Aunque el ejemplo (3.2) tiene dos ecuaciones con dos incógnitas, es obvio que la idea de acondicionamiento se aplica en cualquier dimensión (problema 3.5.1).

3.3 Estudio cuantitativo del acondicionamiento

Estudiemos cuantitativamente el acondicionamiento, viendo como repercute en la *bondad* de nuestra solución. Supongamos que en (3.1) la matriz A es regular.

Pasemos del segundo miembro $\mathbf{b}$ a un segundo miembro perturbado $\mathbf{b} + \delta\mathbf{b}$, sin introducir perturbaciones en A. Sea $\mathbf{x} + \delta\mathbf{x}$ la nueva solución, es decir

$$A(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}. \tag{3.4}$$

Restando (3.1) de (3.4)

$$\delta\mathbf{x} = A^{-1}\delta\mathbf{b}$$

y usando una norma vectorial y la matricial deducida de ella

$$||\delta\mathbf{x}|| \leq ||A^{-1}||||\delta\mathbf{b}||, \tag{3.5}$$

desigualdad que acota el cambio absoluto de la solución en función del cambio absoluto del segundo miembro. Para los cambios *relativos*, notemos que de (3.1)

$$||\mathbf{b}|| \leq ||A||||\mathbf{x}|| \tag{3.6}$$

relación que combinada con (3.5) da, suponiendo que $\mathbf{b}$ (y por tanto $\mathbf{x}$) sean no nulos, la llamada desigualdad de Turing

$$\frac{||\delta\mathbf{x}||}{||\mathbf{x}||} \leq (||A||||A^{-1}||)\frac{||\delta\mathbf{b}||}{||\mathbf{b}||}. \tag{3.7}$$

El número $\mu(A) = ||A||||A^{-1}||$ se llama *número de condición* de A (en la norma en cuestión). Cuando $\mu(A)$ es pequeño (próximo a 1, pues es evidente que $\mu(A) \geq 1$, ¿por qué?), *errores* relativamente pequeños de $\mathbf{b}$ causan *errores* de $\mathbf{x}$ también relativamente pequeños y la matriz A se llama *bien acondicionada*.

A partir de las relaciones $\delta\mathbf{b} = A\delta\mathbf{x}$ y $\mathbf{x} = A^{-1}\mathbf{b}$, y procediendo de la misma forma, se obtiene una acotación *por debajo* para el error relativo de la solución. Concretamente

$$\frac{||\delta\mathbf{x}||}{||\mathbf{x}||} \geq \frac{1}{\mu(A)}\frac{||\delta\mathbf{b}||}{||\mathbf{b}||}. \tag{3.8}$$

que ratifica el importante significado del valor del número de condición, tanto cuando es pequeño como cuando es muy grande.

3.3.1 Algunas propiedades del número de condición. Además del hecho de ser siempre mayor que la unidad (lo que indica que en la resolución de un sistema lineal los errores nuncan disminuyen al pasar de los datos a la solución), es preciso considerar otras interesantes propiedades:

1. El número de condición depende de la norma; lo que, sin duda, es un grave inconveniente para su interpretación, teniendo en cuenta que puede tomar valores muy dispares. En cualquier caso, cuando queramos precisar este aspecto, escribiremos

   $$\mu_p(A) = ||A||_p||A^{-1}||_p, \qquad 1 \leq p \leq \infty$$

2. $\mu(A)$ es la mejor acotación posible en (3.7); es decir, es un valor que se alcanza eligiendo convenientemente las *direcciones* $\mathbf{b}$ y $\delta\mathbf{b}$.

   En efecto, sabemos que existen vectores $\mathbf{x}$ tales que $||A\mathbf{x}|| = ||A||||\mathbf{x}||$; entonces si $\mathbf{b} = A\mathbf{x}$ y tomamos $\delta\mathbf{b}$ tal que $||A^{-1}\delta\mathbf{b}|| = ||A^{-1}||||\delta\mathbf{b}||$, resulta que (3.7) se convierte en una igualdad, y si $\mu(A)$ es grande, el riesgo de magnificación de los errores relativos de los datos es real.

3. Es útil tener presente que $\mu(A) = \mu(A^{-1})$ y que $\mu(\alpha A) = \mu(A)$, cuando $\alpha \neq 0$

3.3.2 Si consideramos también perturbaciones de la matriz, tendremos

$$(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b} + \delta\mathbf{b}, \tag{3.9}$$

y la primera cuestión que se nos plantea es la posible *no invertibilidad* de la matriz $A+\delta A$. El siguiente resultado, que damos sin demostración, nos facilita condiciones suficientes para que dicha situación no se produzca:

LEMA

Si en alguna norma matricial deducida de una vectorial $||A|| < 1$, entonces $I - A$ es no singular y

$$\frac{1}{(1 + ||A||)} \leq ||(I - A)^{-1}|| \leq \frac{1}{(1 - ||A||)}. \tag{3.10}$$

Ahora podemos demostrar el siguiente resultado:

TEOREMA

Si A es una matriz real e invertible, y en alguna norma matricial deducida de una vectorial $||A^{-1}||||\delta A|| < 1$, entonces $A + \delta A$ es invertible, y se verifica que

$$||(A + \delta A)^{-1}|| \leq \frac{||A^{-1}||}{1 - ||A^{-1}||||\delta A||} \tag{3.11}$$

Demostración. Basta seguir los siguientes pasos:

- $I + A^{-1}\delta A = A^{-1}(A + \delta A)$ es invertible según el lema anterior, y
- al ser invertible A tambien ha de serlo $A + \delta A$ (¿por qué?).

Entonces tenemos que

$$||(A + \delta A)^{-1}|| = ||[A(I + A^{-1}\delta A)]^{-1}|| \leq ||A^{-1}||||(I + A^{-1}\delta A)^{-1}||$$

y en base al lema anterior, la cadena continua

$$\leq \frac{||A^{-1}||}{1 - ||A^{-1}\delta A||} \leq \frac{||A^{-1}||}{1 - ||A^{-1}||||\delta A||}$$

lo que termina la demostración. □

Nota: Para la invertibilidad de $A + \delta A$, basta la hipótesis menos fuerte de que $||A^{-1}\delta A|| < 1$, pero dicha condición no nos garantiza la última desigualdad (¿por qué?), que nos es imprescincible en lo que sigue.

Volviendo pues a la relación (3.9) y supuestas las hipótesis del teorema (que se cumplirán para δA *suficientemente pequeña*), tras unas sencillas operaciones, y teniendo en cuenta (3.1), podemos escribir

$$\begin{aligned} \delta\mathbf{x} &= (A+\delta A)^{-1}(\delta\mathbf{b}-\delta A\mathbf{x}) \\ ||\delta\mathbf{x}|| &\leq ||(A+\delta A)^{-1}||(||\delta\mathbf{b}||+||\delta A||||\mathbf{x}||) \end{aligned}$$

$$\frac{||\delta\mathbf{x}||}{||\mathbf{x}||} \leq \frac{||A^{-1}||}{1-||A^{-1}||||\delta A||}\left(\frac{||\delta\mathbf{b}||}{||\mathbf{x}||}+||\delta A||\right).$$

Usando finalmente $||A^{-1}|| = \mu(A)/||A||$ y que $||\mathbf{b}|| \leq ||A||||\mathbf{x}||$ tendremos

$$\frac{||\delta\mathbf{x}||}{||\mathbf{x}||} \leq \frac{\mu(A)}{1-\mu(A)\frac{||\delta A||}{||A||}}\left(\frac{||\delta\mathbf{b}||}{||\mathbf{b}||}+\frac{||\delta A||}{||A||}\right). \tag{3.12}$$

fórmula que relaciona las perturbaciones relativas de $\mathbf{x}$, A y $\mathbf{b}$. Recordamos que es válida siempre y cuando $||\delta A||||A^{-1}|| < 1$.

3.4 Análisis del error en la eliminación Gaussiana

Sentadas las bases teóricas de la relación entre las soluciones de dos sistemas lineales uno de los cuales tiene como datos *pequeñas* perturbaciones de los del otro, pasamos a estudiar el efecto de los errores de redondeo en la eliminación gaussiana, por el método de análisis regresivo de los errores, que como vimos en 1.4 es el más apropiado y casi el único posible para este tipo de análisis. En este caso tendremos que plantear el problema en los términos de considerar la solución (aproximada) que va a calcular el ordenador, al resolver el sistema $A\mathbf{x} = \mathbf{b}$ en sus sistema de coma flotante, como una solución exacta de un sistema ligeramente perturbado y obtener de esa forma cotas para el error en nuestro resultado.

Concretamente, veremos que la solución computada $\mathbf{x}+\delta\mathbf{x}$ es al solución exacta del sistema lineal

$$(A+\delta A)(\mathbf{x}+\delta\mathbf{x}) = \mathbf{b}, \tag{3.13}$$

donde δA es una *pequeña* matriz, pequeñez medida en términos de la unidad de redondeo utilizada (y de la propia matriz como es natural). Entonces, se puede aplicar la desigualdad (3.12) con $\delta\mathbf{b} = \mathbf{0}$ para acotar el error relativo cometido en la solución calculada (véase también el ejercicio 3.5.5). Utilizaremos la norma *infinito* o del supremo, tanto para vectores como para matrices, porque sus propiedades (especialmente la monotonía) hacen mucho más sencillo el análisis. A pesar de todo, aquí nos limitaremos a enunciar e interpretar los resultados, sin demostraciones (el lector interesado en las mismas, puede encontrarlas en el capítulo 21 del clásico libro de Forsythe y Moler.

Como es sabido, el proceso completo de resolución de un sistema lineal por el método de Gauss consta de dos etapas perfectamente diferenciadas. La primera y más compleja (en el sentido de computacionalemente costosa y también por las diferentes opciones entre las que elegir) es la factorización de la matriz A como producto de una matriz triangular inferior L y una triangular superior U. La segunda consiste en la resolución de dos sistemas triangulares: $L\mathbf{y} = \mathbf{b}$, como paso intermedio para mediante $U\mathbf{x} = \mathbf{y}$

obtener la solución. Lógicamente en ambas etapas se cometerán errores que repercutirán en el resultado final.

Por ejemplo, en la primera etapa, calcularemos unos factores L y U tales que en aritmética exacta verificarán una relación diferente de la esperada, pues su producto no será exactamente A sino una aproximación $A + E$, donde E es una matriz error. Pero para no perder el sentido del análisis regresivo pensaremos que lo que hemos computado es la factorización exacta de una matriz ligeramente perturbada de la A, como muestra el siguiente ejemplo para matrices 2×2.

Supongamos que vamos a realizar la factorización LU (primera etapa del método de Gauss) de la siguiente matriz

$$A = \begin{pmatrix} 3.000 & 2.000 \\ 1.000 & 4.000 \end{pmatrix}$$

en una aritmética con cuatro dígitos significativos (unidad de redondeo $u = .5 \times 10^{-3}$. El primer y único multiplicador que tenemos que calcular es

$$m_{21} = \text{fl}\left(\frac{a_{21}}{a_{11}}\right) = \text{fl}\left(\frac{1.000}{3.000}\right) = .3333$$

Si definimos $\tilde{a}_{21} = .9999$, resulta que

$$m_{21} = \frac{\tilde{a}_{21}}{a_{11}} = \frac{0.9999}{3.000} = .3333$$

exactamente. Es decir, que computamos el mismo elemento que se obtiene con aritmética exacta de una matriz con el elemento $(2, 1)$ ligeramente modificado. Continuando con la factorización, ahora tenemos que calcular el elemento $(2, 2)$ reducido, que pertenecerá a U

$$a'_{22} = \text{fl}(a_{22} - m_{21}a_{12}) = \text{fl}(4.000 - 0.3333 \times 2.000) = \text{fl}(4.000 - 0.6666) = 3.333$$

valor que también se obtiene si reemplazamos a_{22} por $\tilde{a}_{22} = 3.9996$ y operamos en forma exacta

$$a'_{22} = \tilde{a}_{22} - m_{21}a_{12} = 3.9996 - 0.3333 \times 2.000 = 3.9996 - 0.6666 = 3.333$$

En consecuencia, que si tomamos la matriz

$$\tilde{A} = \begin{pmatrix} 3.000 & 2.000 \\ 0.9999 & 3.9996 \end{pmatrix}$$

y la factorizamos en aritmética exacta, obtenemos los mismos factores L y U que con nuestro sistema de coma flotante y cuatro dígitos. Dicho de otra forma estos factores, son tales que la siguiente expresión resulta exacta para ellos

$$LU = A + \begin{pmatrix} 0 & 0 \\ -.0001 & -.0004 \end{pmatrix}$$

Generalizando este proceso se puede demostrar, mediante un proceso largo y tedioso, el siguiente resultado:

3.4.1 TEOREMA

Las matrices L y U computadas por eliminación gaussiana con pivotaje, usando aritmética de punto flotante con unidad de redondeo u, satisface

$$LU = A + E$$

con $\|E\|_\infty \leq n^2 \rho \|A\|_\infty u$, donde n es el tamaño de la matriz y

$$\rho = \max_{i,j,k} \frac{|a_{ij}^k|}{\|A\|_\infty}$$

Los números a_{ij}^k son los sucesivos valores que en el proceso de factorización van tomando los elementos a_{ij} de la matriz en la etapa k-ésima. A la hora de analizar el anterior resultado, conviene hacer dos consideraciones importantes. En primer lugar, en su demostración juega un papel importante el hecho de que los multiplicadores

$$m_{ik} = \text{fl}\left(\frac{a_{ij}^k}{a_{kk}^k}\right)$$

verifiquen la acotación $|m_{ik}| \leq 1$, para todo i, k. Es decir, que la hipótesis de pivotaje no puede suprimirse, si bien basta que sea parcial. En segundo lugar que no hay una buena cota 'a priori' para ρ aunque en el proceso del cálculo, se puede estimar una de forma relativamentes sencilla. Pero, a menos que sea muy grande, vemos que las perturbaciones son del orden de la unidad de redondeo.

El análisis regresivo de error para la resolución de un sistema triangular $R\mathbf{y} = \mathbf{c}$, nos aporta el siguiente resultado:

TEOREMA

El vector $\mathbf{y} + \delta\mathbf{y}$ computado al resolver el sistema triangular, es la solución exacta de un sistema triangular perturbado $(R + \delta R)(\mathbf{y} + \delta\mathbf{y}) = \mathbf{c}$, donde δR satisface

$$\|\delta R\|_\infty \leq \frac{n(n+1)}{2} \cdot 1.01u \max_{i,j} |r_{ij}|$$

La demostración utiliza, via (1.4), la desigualdad $n \cdot u \leq 0.01$, que es poco exigente en la práctica habitual. Es interesante observar que aunque δR depende del $\mathbf{c}$, dicha dependencia no aparece explícitamente en la cota.

Aplicando el anterior resultado a la resolución de los dos sistemas triangulares que completan el proceso de eliminación gaussiana, obtendremos las siguientes relaciones que nos van a permitir enunciar el resultado global del análisis regresivo que estamos efectuando. Por una parte, tendremos que el sistema

$$(L + \delta L)(U + \delta U)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}$$

que tras desarrollar, y considerando que $LU = A + E$, resulta

$$(A + E + (\delta L)U + L(\delta U) + \delta L \delta U)(\mathbf{x} + \delta \mathbf{x}) = \mathbf{b} \tag{3.14}$$

es el que resolveríamos exactamente para obtener la solución computacional aproximada. Si tenemos en cuenta que $|l_{ij}| \leq 1$ debido a la estrategia de pivotamiento, que $|u_{ij}| \leq \rho \|A\|_\infty$ por la propia definición de ρ y que en las aplicaciones normalmente $n^2 u \ll 1$, nos encontremos con la siguiente cadena de desigualdades

$$\begin{aligned}
\|L\|_\infty &\leq n \\
\|U\|_\infty &\leq n\rho\|A\|_\infty \\
\|\delta L\|_\infty &\leq \frac{n(n+1)}{2} 1.01u \\
\|\delta U\|_\infty &\leq \frac{n(n+1)}{2} 1.01\rho\|A\|_\infty u \\
\|\delta L\|_\infty \|\delta U\|_\infty &\leq n^2\rho\|A\|_\infty u, \text{tomando } \frac{(n+1)^2}{4} 1.01^2 u \leq 1 \\
\|E\|_\infty &\leq n^2\rho\|A\|_\infty u, \text{del teorema 3.4.1}
\end{aligned}$$

La comparación de las ecuaciones (3.13) y (3.14), nos permite obtener una expresión para δA y, en el supuesto de que se cumplan todos los requisitos que hemos ido mencionando, enunciar el resultado final:

3.4.2 TEOREMA

La solución computada por eliminación gaussiana con pivotamiento satisface exactamente la ecuación

$$(A + \delta A)(\mathbf{x} + \delta \mathbf{x}) = \mathbf{b}$$

donde

$$\delta A = E + (\delta L)U + L(\delta U) + \delta L \delta U$$

Además, se verifica que

$$\|\delta A\|_\infty \leq 1.01(n^3 + 3n^2)\rho\|A\|_\infty u$$

La demostración es evidente, siguiendo la cadena de desigualdades

$$\begin{aligned}
\|\delta A\|_\infty &\leq \|E\|_\infty + \|\delta L\|_\infty \|U\|_\infty + \|L\|_\infty \|\delta U\|_\infty + \|\delta L\|_\infty \|\delta U\|_\infty \\
&\leq 2n^2\rho\|A\|_\infty u + 2\frac{n^2(n+1)}{2}\rho\|A\|_\infty 1.01u \\
&\leq 1.01(n^3 + 3n^2)\rho\|A\|_\infty u
\end{aligned}$$

Pero lo más importante es concluir que la solución obtenida satisface exactamente un sistema con una matriz *ligeramente* perturbada, lo que certifica la estabilidad numérica del método. De hecho, vemos que la norma de la perturbación está acotada por una cantidad computable de veces la norma de la matriz original, y que salvo valores descabellados para ρ y/o de la dimensión n de la matriz, el tamaño relativo de la perturbación es del orden de la unidad de redondeo u del sistema. Pero, de hecho, dichas cotas son inalcanzables en los casos prácticos, pues tendría que darse el peor de los casos en todas las etapas. De hecho, en el libro de Wilkinson se establece que *raramente* no se cumple la siguiente cota

$$\|\delta A\|_\infty \leq n\|A\|_\infty u \tag{3.15}$$

que aceptaremos como cota más realista. Resulta curiosa la desaparición de ρ, pero lo cierto es que en las matrices que se presentan en la práctica el crecimiento de los valores de $|a_{ij}^k|$ es muy raro y Wilkinson afirma que, si partimos de valores $|a_{ij}| \leq 1$, casi nunca superan el valor de 8!. Sin embargo, existen matrices en las que dichos valores alcanzan el máximo valor teóricamente posible.

3.5 Cuestiones y problemas

3.5.1 Estudie geométricamente el acondicionamiento en el caso de dimensión 3.

3.5.2

a) Para el sistema (3.2), perturbado como en (3.3), y trabajando en la norma ∞, compruebe la desigualdad (3.12) y explique lo que ocurre.

b) ¿Y si tomamos como segunda ecuación la perturbación $2x + 6.00000875y = 8.000035$, trabajando con la misma norma?

c) Construya sistemas de dos ecuaciones/incógnitas en los que la primera ecuación sea la de (3.2) y estén mucho peor acondicionados que (3.2)

3.5.3 ¿Cuál es el número de condición euclídeo de una matriz ortogonal? Dé una interpretación geométrica para $n = 2$ del valor obtenido, considerando la cota (3.7).

3.5.4 Demuestre que el número de condición de una matriz hermítica en la norma euclídea es el cociente de los módulos máximo y mínimo de los autovalores.

3.5.5 De la desigualdad general (3.12) se deduce que no existe una relación similar a (3.7) entre los errores relativos de la solución $\mathbf{x}$ y la matriz del sistema A. Sin embargo, se pueden demostrar dos acotaciones muy parecidas:

a) Una desigualdad exacta, pero donde el primer miembro no es realmente el error relativo de la solución

$$\frac{||\delta\mathbf{x}||}{||\mathbf{x}+\delta\mathbf{x}||} \leq \mu(A)\frac{||\delta A||}{||A||}.$$

b) Una desigualdad exacta, pero donde el segundo miembro no es solamente el error relativo de la matriz

$$\frac{||\delta \mathbf{x}||}{||\mathbf{x}||} \leq \mu(A) \frac{||\delta A||}{||A||}(1 + O(||\delta A||))$$

pero que puede interpretarse como una verdadera desigualdad entre errores relativos cuando $||\delta A||$ es suficientemente pequeño.

2010-11

CAPÍTULO II

INTERPOLACIÓN

2010-11

Lección 4

Polinomios de Chebyshev

4.1 Elección óptima de los nodos de interpolación

Supongamos fijados un intervalo $[a, b]$, una función real f definida en él y un número entero $n \geq 1$. ¿Cómo elegir $n+1$ nodos distintos $x_0, \ldots, x_n$ de suerte que, al efectuar la interpolación lagrangiana de f en $[a, b]$, el error tenga el menor tamaño posible? Sabemos que medir errores exactamente es muy difícil o imposible. Conviene entonces sustituir la cuestión anterior por otra más fácilmente resoluble ¿Cómo elegir $n+1$ nodos distintos $x_0, \ldots, x_n$ de suerte que al efectuar la interpolación lagrangiana de f en $[a, b]$ la *cota de error*

$$\frac{|W(x)|}{(n+1)!} M_{n+1}$$

tenga el menor tamaño posible? Evidentemente la solución de este nuevo problema es, en cierta forma, independiente de la función f, y se reduce a encontrar los nodos para que el *tamaño* de $|W(x)|$ sea el menor posible. Recordemos que $W(x) = (x - x_0) \cdots (x - x_n)$ y M_{n+1} es una cota de la derivada $n+1$ de f, que suponemos existe.

4.1.1 El concepto de norma para funciones Para poder abordar esta tarea es necesario, ante todo, definir precisamente qué entendemos por tamaño de una función. No hay duda sobre qué debemos entender por tamaño de un número real o complejo: su módulo. En la lección 2 establecimos el concepto para vectores reales o complejos n-dimensionales, evidenciando que existen infinitas posibilidades con significados diversos.

Pero ¿y para funciones?, ¿es la función real constantemente igual a 1 de mayor o menor tamaño que la función real $100\, e^{-x^2}$ (dibuje las gráficas)? La solución, una vez más, está en normar el correspondiente espacio vectorial.

4.1.2 Para funciones $v = v(x)$, reales o complejas, definidas en un intervalo $[a, b]$ son usuales las siguientes normas (similares a las definidas en la lección anterior para vectores):

$$\textit{Norma del supremo o norma infinito} \qquad ||v||_\infty = \sup_{a \leq x \leq b} |v(x)|$$

$$\textit{Norma dos} \qquad ||v||_2 = \left(\int_a^b |v(x)|^2 dx \right)^{1/2}$$

Norma uno $$||v||_1 = \int_a^b |v(x)|dx$$

Nótese que $||.||_\infty$ está definida no para todas las funciones sino tan solo para las acotadas. Análogamente las normas *dos* ó *uno* sólo pueden definirse para funciones cuyo cuadrado del módulo o cuyo módulo, respectivamente, sea integrable.

4.2 Los polinomios de Chebyshev

Hechas estas observaciones sobre la norma, podemos formular de modo preciso la cuestión que planteabamos al principio:

¿Cómo elegir los nodos de interpolación para que $||W||_\infty$ sea lo menor posible? La respuesta a esta pregunta fue encontrada por el matemático ruso Chebyshev (1821-1894) en términos de unos polinomios que hoy llevan su nombre.

PROPOSICIÓN

Para cada entero $n \geq 0$ existe un único polinomio T_n, llamado n-ésimo polinomio de Chebyshev, tal que para cada θ real

$$T_n(\cos\theta) = \cos n\theta. \tag{4.1}$$

T_n tiene grado exactamente n. Si $n \geq 1$, el coeficiente de x^n en él es 2^{n-1}. Se tiene, además, para $n \geq 2$, la identidad

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x). \tag{4.2}$$

Demostración. Si hubiese dos polinomios que satisficiesen (4.1) coincidirían para los infinitos valores $-1 \leq x \leq 1$ de su variable, lo cual les fuerza a coincidir idénticamente. Las demás afirmaciones (existencia, grado, coeficiente director y (4.2)) se prueban por inducción. Para $n = 0$, para satisfacer (4.1), será $T_0 = 1$, de grado 0. Cuando $n = 1$ será $T_1(x) = x$, de grado 1, coeficiente director 2^0. Para $n = 2$, de $\cos 2\theta = 2\cos^2\theta - 1$ deducimos que $T_2(x) = 2x^2 - 1$, de grado 2, coeficiente director 2^1. Claramente (4.2) se satisface. Supongamos que las propiedades se verifican para todo entero menor o igual que n. Entonces, en virtud de las fórmulas de transformación del producto de cosenos en suma de cosenos, se tiene $\cos n\theta = 2\cos\theta\cos(n-1)\theta - cos(n-2)\theta$, es decir $\cos n\theta = 2\cos\theta T_{n-1}(\cos\theta) - T_{n-2}(\cos\theta)$, quedando probado que $\cos n\theta$ es un polinomio T_n en el $\cos\theta$. Claramente T_n satisface las propiedades anunciadas.□

4.2.1 Observemos que para $-1 \leq x \leq 1$ se tiene

$$T_n(x) = \cos n \arccos x, \tag{4.3}$$

siendo arccos una determinación cualquiera del arco coseno, digamos la determinación que toma valores en $-\pi \leq \theta \leq 0$. Cuando x decrece monótonamente desde 1 hasta -1, el arco $\alpha = n\arccos x = n\theta$ decrece monótonamente desde 0 hasta $-n\pi$ (gira n medias vueltas en el sentido de las agujas del reloj); y por tanto, la función $T_n(x) = \cos\alpha$ oscila

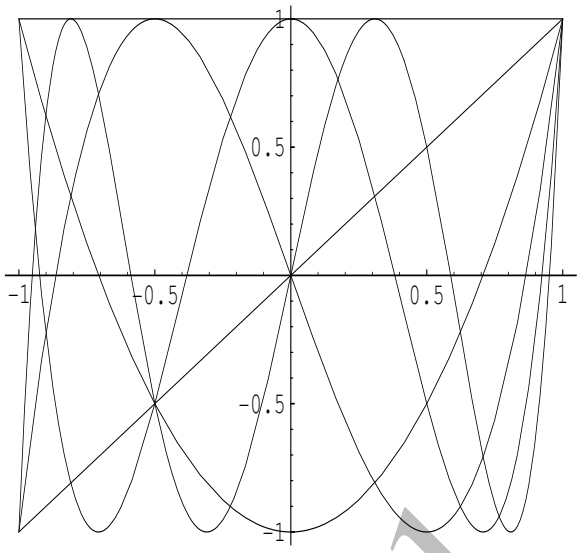


Figura 4.1: Polinomios de Chebyschev de grados 0 a 5

las mismas veces entre 1 y -1 alcanzando el valor que toma la función coseno en α en la abscisa $x = \cos\frac{\alpha}{n}$.

Así, el coseno parte del valor 1 en $\alpha = 0$ ($x = 0$), va decreciendo hasta valer 0 en $\alpha = -\frac{\pi}{2}$ ($x = \cos\frac{\pi}{2n}$), sigue decreciendo hasta valer -1 en $\alpha = -\pi$ ($x = \cos\frac{\pi}{n}$), luego crece, anulándose en $\alpha = -\frac{3\pi}{2}$ ($x = \cos\frac{3\pi}{2n}$), etc... De esta forma, se prueba:

PROPOSICION

Los n ceros de T_n son los puntos

$$\eta_{nk} = \cos\frac{(2k-1)\pi}{2n}, \qquad k = 1, \ldots, n.$$

Para $-1 \leq x \leq 1$, T_n toma valores entre -1 y 1. Estos valores extremos se alcanzan precisamente en los puntos

$$\zeta_{nk} = \cos\frac{2k\pi}{2n}, \qquad k = 0, 1, \ldots, n$$

y en ellos $T_n(\zeta_{nk}) = (-1)^k$.

En la figura 4.1, se representan en $-1 \leq x \leq 1$, los polinomios de Chebyshev de grados cero a cinco, y pueden comprobarse estos extremos y ceros.

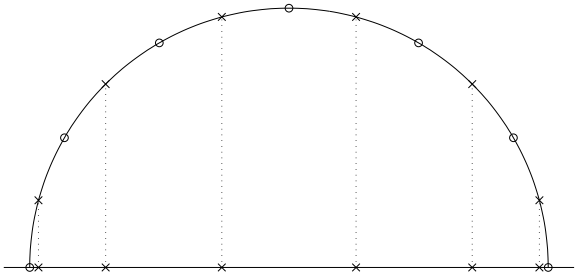


Figura 4.2: Puntos de interpolación de Chebyschev para $n = 6$

4.2.2 Consecuencia inmediata de la última conclusión de la proposición anterior es el siguiente teorema, fundamental para el objetivo que perseguimos:

TEOREMA

El n-ésimo polinomio de Chebyshev tiene norma del supremo en [-1,1] no mayor que cualquier otro polinomio de su mismo grado y coeficiente director.

Demostración. Si existiese un polinomio P del mismo grado y coeficiente director que T_n pero con norma del supremo más pequeña, la diferencia $T_n - P$ tendría grado $\leq n-1$ y sería positiva en $\zeta_{n0} = 1$ (pues ahí T_n vale 1 y $|P| < 1$), negativa en ζ_{n1} (pues ahí T_n vale -1 y $|P| < 1$), positiva en $\zeta_{n2}, \ldots$ Así se encuentran n cambios de signo para un polinomio de grado $\leq n-1$ y no idénticamente nulo, lo cual es absurdo. □

4.2.3 Corolario 1. El polinomio $T_n/2^{n-1}$ tiene norma del supremo en [-1,1] no mayor que cualquier otro polinomio de grado n y coeficiente director la unidad.

4.2.4 Corolario 2. Para la norma del supremo en el intervalo [-1,1], la cantidad $||(x - x_0) \cdots (x - x_n)||$ toma su valor mínimo posible frente a todas las elecciones de números reales $x_0, \ldots, x_n$ (dentro o fuera del intervalo, distintos o no) cuando los x_k se eligen como los $n + 1$ ceros $\eta_{n+1,k}$ del $(n + 1)$-ésimo polinomio de Chebyshev. El valor de dicha norma es en este caso $\frac{1}{2^n}$

Demostración. En efecto, para cada elección de x_i, el producto $(x - x_0) \cdots (x - x_n)$ es un polinomio de grado $n + 1$ y coeficiente director unidad y por el corolario 1 (con $n + 1$ en lugar de n) su norma del supremo no es menor que la de $T_{n+1}/2^n = (x - \eta_{n+1,1}) \cdots (x - \eta_{n+1,n+1})$. El valor de la norma es evidente teniendo en cuenta que la del propio polinomio es 1. □

De este modo la elección de nodos de interpolación más favorable para minimizar la cota del error en la norma infinito de [-1,1] consiste en tomar como abscisas los ceros del polinomio de Chebyshev correspondiente. La figura 4.2 nos muestra un proceso algorítmico para calcular dichos puntos; se trata de dividir la semicircunferencia que abarca el intervalo en n partes iguales, y entonces las abscisas de los puntos medios de los arcos son los puntos de interpolación buscados.

4.3 Cambio de intervalo

El caso de otro intervalo acotado $[a, b]$ se reduce al precedente por medio de *un cambio lineal de variable*. Si denominamos $\tilde{x}$ a la variable en $[-1, 1]$, es fácil ver que la transformación debe ser

$$\tilde{x} = -1 + 2\frac{x-a}{b-a} = \frac{2}{b-a}x - \frac{a+b}{b-a} \tag{4.4}$$

donde se advierten los procesos de traslación y dilatación del intervalo.

Es interesante profundizar en este problema, con el que nos encontraremos más veces, y ver el tratamiento numérico que se le puede dar. Es evidente que los polinomios serán diferentes en el nuevo intervalo. Un proceso simbólico para obtener los nuevos polinomios consiste en la simple sustitución de la x en los polinomios básicos por esta $\tilde{x}$ para obtener nuevas expresiones en x. Es fácil ver que la nueva familia de polinomios es también *triangular* (¿por qué?) y de hecho en el intervalo $[a, b]$ cada uno de los nuevos polinomios toma exactamente los mismos valores que el original en los puntos transformados.

Esta observación nos va a servir para implementar un método numérico que nos permita calcular los coeficientes de los polinomios en $[a, b]$ sin recurrir a la sustitución algebraica. Supongamos que estamos construyendo los polinomios de Chebyschev hasta grado n en el citado intervalo. Es evidente que nos bastará con conocer sus valores en $n+1$ puntos (¿por qué?). Una forma trivial y económica de conseguirlo, es elegir dichos puntos en la forma que mejor nos parezca (equidistantes, por ejemplo), conseguir sus transformados en el intervalo $[-1, 1]$ mediante la fórmula (4.4), y la evaluación de los polinomios originales nos proporcionara los valores buscados.

En términos matriciales, y teniendo en cuenta que evaluar un polinomio de grado n en un punto $\tilde{x}$, equivale a realizar el producto escalar del vector de coeficientes $(a_0, a_1, \ldots, a_n)$ por el de valores de las potencias de x en dicho punto $(1, \tilde{x}, \tilde{x}^2, \ldots, \tilde{x}^n)$, evaluar los $n+1$ primeros polinomios de Chebyschev en $n+1$ puntos equivale a realizar el siguiente producto de matrices

$$T \cdot \tilde{X} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ -1 & 0 & 2 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ T_{n0} & T_{n1} & T_{n2} & \ldots & T_{nn} \end{pmatrix} \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ \tilde{x}_0 & \tilde{x}_1 & \tilde{x}_2 & \ldots & \tilde{x}_n \\ \tilde{x}_0^2 & \tilde{x}_1^2 & \tilde{x}_2^2 & \ldots & \tilde{x}_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \tilde{x}_0^n & \tilde{x}_1^n & \tilde{x}_2^n & \ldots & \tilde{x}_n^n \end{pmatrix}$$

donde T es la matriz (triangular inferior) con los coeficientes de los polinomios de Chebyschev en $[-1, 1]$. Este producto tiene que ser idéntico a

$$C \cdot \begin{pmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_0 & x_1 & x_2 & \ldots & x_n \\ x_0^2 & x_1^2 & x_2^2 & \ldots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & x_2^n & \ldots & x_n^n \end{pmatrix}$$

donde C es la matriz buscada con los coeficientes de los polinomios en el intervalo de trabajo $[a, b]$. De la relación $T \cdot \tilde{X} = C \cdot X$, resulta evidente que

$$C = T \cdot \tilde{X} \cdot X^{-1} \tag{4.5}$$

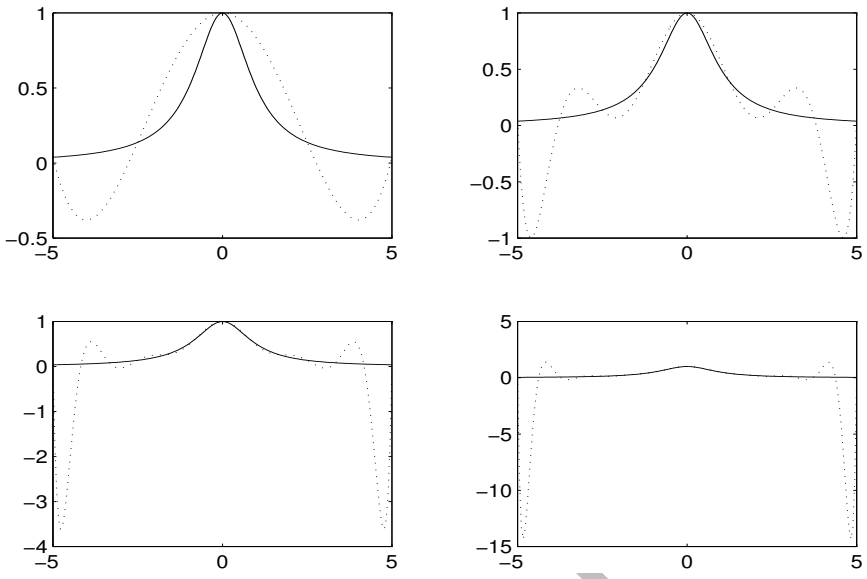


Figura 4.3: Interpolaciones sucesivas en 5, 9, 13 y 17 puntos equiespaciados

La existencia de la matriz inversa de X está garantizada siempre que tomemos puntos distintos (¿por qué?).

4.3.1 El ejemplo de Runge. La mejora en el error cometido cuando se toman los nodos de Chebyschev, es notable en algunos casos. Por ejemplo, en el célebre ejemplo de Runge en que se considera la funcion

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

sobre el intervalo $[-5, 5]$, las figuras 4.3 y 4.4, nos muestran la calidad de la aproximación del polinomio interpolador para distintos valores de n, cuando se utilizan puntos equiespaciados y abscisas de Chebyshev respectivamente.

Aunque en este caso la convergencia parece evidente, existen funciones para las que la interpolación en más y más puntos de Chebyshev no converge. Así pues, se hace necesaria la búsqueda de interpolantes aún mejores que veremos en las próximas lecciones.

4.3.2 Aplicación. Economización de Chebyshev. Pero las aplicaciones de los polinomios de Chebyshev son muchas en el ámbito de la aproximación de funciones debido a sus notables propiedades. Supongamos que se desea aproximar en $-0.5 \leq x \leq 0.5$ la función $\exp(x)$ por un polinomio de segundo grado. Un aproximante obvio es el polinomio de Taylor $p_2 = 1 + x + x^2/2$. Para valores próximos a 0 no puede haber mejor elección de polinomio de segundo grado; pues, cuando x tiende a cero, p_2 difiere de la función en términos de tercer orden en x y no hay otro polinomio cuadrático con esta propiedad. Sin

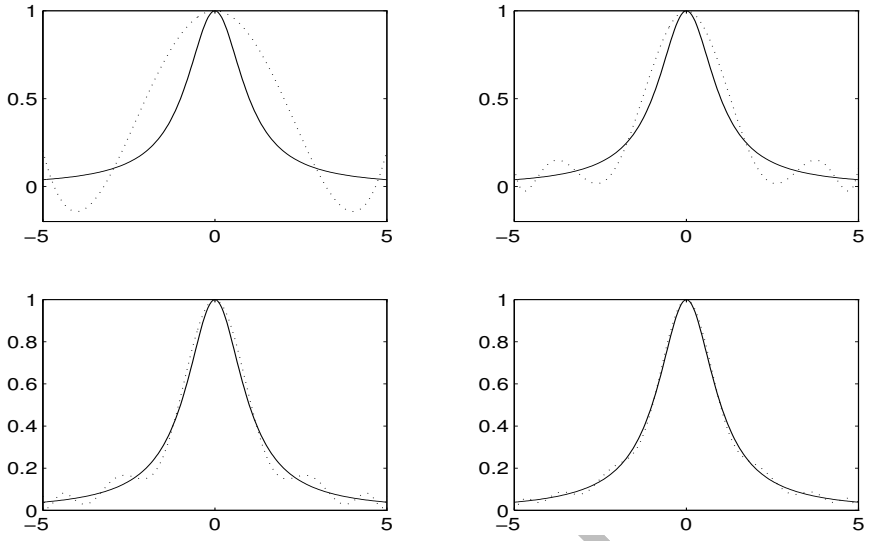


Figura 4.4: Interpolaciones sucesivas en 5, 9, 13 y 17 puntos de Chebyshev

embargo la aproximación dada por p_2 se degrada cuando x está próximo a los extremos del intervalo. Por ejemplo en $x = 0.5$, $p_2 = 1.625$ mientras que $\exp(0.5) = 1.649$.

Mostremos seguidamente cómo obtener un aproximante de segundo grado por un proceso llamado economización de Chebyshev. Para ello necesitamos partir de un aproximante polinómico de grado una unidad superior al aproximante buscado, en nuestro caso 3. El polinomio de Taylor $p_3(x) = 1 + x + x^2/2 + x^3/6$ nos servirá. Ahora sustituiremos p_3 por el polinomio P de grado 2 que haga la diferencia $Q = p_3 - P$ de tamaño más pequeño posible, más precisamente tal que $||Q||_\infty$ sea lo menor posible. Cuando P recorre todos los polinomios de segundo grado, Q recorre todos los polinomios cúbicos de coeficiente director $1/6$ y de entre éstos el de menor norma del supremo será el múltiplo escalar del polinomio de Chebyshev cúbico cuyo coeficiente director sea $1/6$.

El polinomio de Chebyshev cúbico es, por (4.2), $T_3(x) = 4x^3 - 3x$. Aquí estamos tratando con el intervalo [-0.5, 0.5] con lo que el polinomio escalado es $32x^3 - 6x$ y su múltiplo de coeficiente director $1/6$ vale $Q = (1/6)x^3 - (1/32)x$. Esto conduce a $P = p_3 - Q = 1 + (33/32)x + x^2/2$. Este aproximante cuadrático, obtenido por economización en p_3, es más eficaz que p_2, por ejemplo en $x = 0.5$, vale 1.641.

Ahora podríamos todavía economizar P para obtener un polinomio de primer grado $(33x + 34)/32$.

4.4 Cuestiones y problemas

4.4.1 Use la relación de recurrencia para hallar T_4, T_5. Halle sus ceros sin utilizar la fórmula dada en esta lección. Use entonces la fórmula para validar sus cálculos. Como segunda comprobación lea los ceros en la figura 4.1, usando una regla milimetrada.

4.4.2 Paridad de los polinomios de Chebyshev. Pruebe que T_n es un polinomio par o impar según lo sea n.

4.4.3 Ecuación diferencial satisfecha por los polinomios de Chebyshev. Pruebe que T_n satisface la ecuación diferencial

$$(1-x^2)T_n'' - xT_n' + n^2 T_n = 0.$$

4.4.4 *Función generatriz de los polinomios de Chebyshev.* Pruebe que si $-1 < t < 1$ y $-1 \le x \le 1$ entonces

$$\sum_{n=0}^{\infty} t^n T_n(x) = \frac{1-tx}{1-2tx+t^2}$$

La función en el segundo miembro se llama *generatriz* de los polinomios de Chebyshev.

4.4.5 Polinomio interpolador de Lagrange en los ceros de T_n. Pruebe que la solución del problema de interpolación de Lagrange basado en los n ceros de T_n como nodos es

$$n^{-1} \sum_{k=1}^{n} f(x_k) \frac{T_n(x)(-1)^{k-1}\sin(\theta_k)}{x - x_k}$$

donde $\theta_k = (2k-1)\pi/(2n)$, $x_k = \cos\theta_k$, $k = 1, \ldots, n$.

4.4.6 Demuestre que $T_n(x) = x^n - C_{n,2}x^{n-2}(1-x^2) + C_{n,4}x^{n-4}(1-x^2)^2 - \ldots$, siendo $C_{n,k}$ el número combinatorio n sobre k (número de subconjuntos con k objetos de un conjunto de n).

4.4.7 Cambio de intervalo. A la vista de lo afirmado en la sección 4.3, la matriz triangular inferior (¿por qué?) $\tilde{X} \cdot X^{-1}$ de la ecuación (4.4), que podríamos denominar matriz de *cambio de intervalo* no depende de los puntos elegidos para construir cada uno de los factores, sino de la transformación lineal efectuada. Por consiguiente, debe ser posible encontrar una expresión directa para esta matriz de cambio de coeficientes, en función de los extremos del intervalo $[a, b]$. (Indicación: Piense primero en un cambio escrito en la forma $\tilde{x} = mx + n$, y después sustituya en la matriz resultante.)

4.4.8 Expresión de un polinomio como combinación lineal de polinomios de Chebyshev u otras familias 'triangulares'. Consideremos los polinomios $Q_0, Q_1, \ldots, Q_n$ tales que $Q_0(x) = 1$, $Q_1(x) = c_1x - a_1$, $Q_k(x) = (c_kx - a_k)Q_{k-1}(x) - b_kQ_{k-2}(x)$, $k = 2, \ldots, n$, siendo $c_1, a_1, c_2, a_2, b_2, \ldots, c_n, a_n, b_n$ constantes conocidas, con cada c_k no nula. (Un ejemplo lo constituyen los polinomios de Chebyshev ¿por qué?)

Pruebe que Q_k tiene grado exactamente k, y por tanto que cada polinomio P de grado $\le n$ tiene una única expresión $P(x) = \alpha_0 Q_0(x) + \ldots + \alpha_n Q_n(x)$.

4.4.9 Evaluación de un polinomio expresado como combinación lineal de una familia 'triangular' de polinomios. Pruebe que, conocida la expresión del ejercicio anterior, el valor de P en un punto dado x^* puede hallarse mediante el algoritmo $d_{n+2} = 0$; $d_{n+1} = 0$; para $k = n, n-1, \ldots, 0$: $d_k = \alpha_k + (c_{k+1}x^* - a_{k+1})d_{k+1} - b_{k+2}d_{k+2}$; $P(x^*) = d_0$. (Indicación: sustituya en el desarrollo de P cada α_k por su expresión en términos de x^* y los a, b, c, d.)

¿Cuántas operaciones son así necesarias para evaluar P?

Este algoritmo contiene como casos particulares otros estudiados anteriormente ¿cuáles?

4.4.10 Sea $\alpha_0 T_0(x) + \ldots + \alpha_n T_n(x)$ la expresión de un polinomio en la base de Chebyshev. ¿Cuánto vale la suma de todas las α_k?

2010-11

Lección 5

La interpolación de Hermite u osculatoria

5.1 El problema de Hermite

En el problema de interpolación lagrangiana se determina un polinomio de grado $\leq n$ por sus valores en $n+1$ nodos, mientras que en el de Taylor hay un solo nodo pero además del valor de la función hay que reproducir los de las n primeras derivadas. Estos dos problemas son casos particulares extremos de uno más general, llamado de Hermite u osculatorio, donde se contemplan $r+1$ $(0 \leq r \leq n)$ nodos x_i y en cada nodo se pide reproducir la función y sus $m_i \geq 0$ primeras derivadas. Naturalmente se debe cumplir que el número total de condiciones iguale al número $n+1$ de parámetros libres en el polinomio es decir

$$\sum_{0 \leq i \leq r} (1 + m_i) = n + 1. \tag{5.1}$$

5.1.1 Formalmente, el problema que se plantea Hermite es el siguiente:

Dados una función real f definida en $[a, b]$, $r+1$ puntos distintos $x_0, \ldots, x_r$ en dicho intervalo y enteros no negativos $m_0, \ldots, m_r$, n de suerte que se satisfaga (5.1) y que las derivadas de f que vamos a escribir existan, **determinar un polinomio** P_n de grado menor o igual que n tal que

$$\begin{array}{llll} P_n(x_0) = f(x_0), & P_n'(x_0) = f'(x_0), & \ldots, & P_n^{(m_0)}(x_0) = f^{(m_0)}(x_0); \\ P_n(x_1) = f(x_1), & P_n'(x_1) = f'(x_1), & \ldots, & P_n^{(m_1)}(x_1) = f^{(m_1)}(x_1); \\ \vdots & \vdots & \ddots & \vdots \\ P_n(x_r) = f(x_r), & P_n'(x_r) = f'(x_r), & \ldots, & P_n^{(m_r)}(x_r) = f^{(m_r)}(x_r). \end{array} \tag{5.2}$$

5.2 Construcción del interpolante en forma de Newton

Observemos que en un problema de Hermite *no se puede dar como dato la derivada k-ésima en un nodo si no se han dado la función y todas las derivadas hasta la $k-1$.*

5.2.1 TEOREMA

El problema de interpolación de Hermite tiene solución única.

Demostración. El camino usado en ocasiones anteriores (buscar P_n por coeficientes indeterminados en potencias de x) tiene el inconveniente de que la matriz del sistema resultante es de estructura complicada. Vamos a proceder en analogía a lo hecho en la forma de Newton, es decir buscar primero la constante P_0 que satisface la primera condición, luego la recta P_1 que satisface las dos primeras, una parábola P_2 determinada por las tres primeras, y sucesivos polinomios P_d de grado $\leq d$ que satisfaciendo las $d+1$ primeras, y así hasta P_n. Supuesto hallado P_d denotamos por Q_{d+1} el polinomio que hay que sumarle para obtener P_{d+1}.

Antes de comenzar con la demostración notemos que, por definición, z es una raíz de multiplicidad $k \geq 1$ de un polinomio $Q(x)$ si $Q(x)$ es divisible entre $(x-z)^k$. Una condición necesaria y suficiente para que ello ocurra es que z anule no sólo a Q sino también a sus $k-1$ primeras derivadas (demostración inmediata, hágala).

Es obvio que P_0 está unívocamente determinado. Q_1 ha de ser de grado a lo sumo 1 y anularse en x_0, luego de la forma $c_1(x-x_0)$. Imponiendo la segunda condición de la primera línea de (5.2) se determina unívocamente c_1 (¿quién es?) y por tanto P_1. Ahora Q_2 debe tener grado 2 y anularse en x_0 junto con su derivada. Por ello es de la forma $c_2(x-x_0)^2$. La tercera condición de esta primera línea de (5.2) determina $c_2 = \frac{f''(x_0)}{2}$ y por tanto P_2. Continuando en esta forma alcanzamos P_{m_0} (que naturalmente no es otra cosa que el polinomio de Taylor de grado m_0 de f en x_0, expresado en potencias de $(x-x_0)$).

Ahora Q_{m_0+1} es de grado m_0+1 y debe ser nulo con sus m_0 primeras derivadas en x_0, luego de la forma $c_{m_0+1}(x-x_0)^{m_0+1}$. La primera condición de la segunda línea de (5.2) determina la constante. El siguiente término a añadir es de la forma $c_{m_0+2}(x-x_0)^{m_0+1}(x-x_1)$ ya que debe ser nulo con sus m_0 primeras derivadas en x_0 y anularse también en x_1.

Prosiguiendo esta marcha se llega a P_n satisfaciendo los requerimientos del problema escrito como combinación lineal de los $n+1$ polinomios

$$\begin{array}{cccc} 1, & (x-x_0), & \dots, & (x-x_0)^{m_0}, \\ (x-x_0)^{m_0+1}, & (x-x_0)^{m_0+1}(x-x_1), & \dots, & (x-x_0)^{m_0+1}(x-x_1)^{m_1}, \\ \vdots & \vdots & \ddots & \vdots \\ \dots & (x-x_0)^{m_0+1}(x-x_1)^{m_1+1}\cdots(x-x_r)^{m_r}. & & \end{array} \tag{5.3}$$

Para la unicidad, es fácil demostrar que si hay dos polinomios que satisfagan las condiciones impuestas, su diferencia es un polinomio idénticamente nulo. □

5.2.2 Notas

1. El proceso da como casos particulares la forma de Newton del polinomio interpolador de Lagrange (si cada m_i es cero) y el polinomio de Taylor (si $r=0$).

2. No es necesario tomar las condiciones (5.2) en el orden en que lo hemos hecho. Otra posibilidad es empezar por la primera para x_0, seguida de la primera para x_1, hasta la primera para x_r, seguir con la segunda (si la hay) de x_0, de la segunda (si la hay) de x_1, etc. Se comprueba inmediatamente la validez de un proceso análogo al de la demostración

del teorema, usando ahora la base (que debe leerse por columnas)

$$\begin{array}{llll} 1, & (x-x_0)\cdots(x-x_r), & \ldots, & \vdots \\ (x-x_0), & (x-x_0)^2\cdots(x-x_r), & \ldots, & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ (x-x_0)\cdots(x-x_{r-1}), & \ldots (x-x_0)^{m_0'}(x-x_1)^{m_1'}\cdots(x-x_r)^{m_r'}. & & \end{array} \tag{5.4}$$

donde el último término no se puede precisar más, pero los m_i' son siempre m_i+1, excepto para el punto que tenga mayor número de derivadas con valor prefijado (y si hay varios con idéntica cantidad, el de mayor índice), y en esa fila estará el susodicho término.

3. Más generalmente es posible tomar las condiciones en (5.2) en cualquier orden, con tal que al tomar una de una fila hayamos ya satisfecho todas las anteriores de esa fila, es decir al tratar de reproducir una derivada de orden superior en un punto hayamos ya tomado en consideración todas las derivadas de orden más bajo en ese mismo punto, lo cual garantiza que cada subproblema de los que vamos resolviendo sea también un problema de Hermite.

5.2.3 Ejemplo. El polinomio de Hermite $P(x)$ de grado ≤ 4 para una función y con los siguientes valores de y, y', y'' (tomados de la función $\sin x$)

$$\begin{array}{lll} y(0)=0, & y'(0)=1, & y''(0)=0; \\ y(\pi)=0, & y'(\pi)=-1; & \end{array}$$

es el siguiente. Construído como en la demostración del teorema, con la siguiente base de polinómios entrando por filas con su correspondiente condición

$$\begin{array}{lll} 1, & x, & x^2; \\ x^3, & x^3(x-\pi); & \end{array}$$

resulta

$$P(x) = 0 + x + 0x^2 - (1/\pi^2)x^3 + (1/\pi^3)x^3(x-\pi).$$

Con el orden que propicia la base de tipo (5.4), incorporándose ahora por columnas para satisfacer la condición respectiva

$$\begin{array}{lll} 1, & x(x-\pi), & x^2(x-\pi)^2; \\ x, & x^2(x-\pi); & \end{array}$$

el *mismo* polinomio se escribe

$$P(x) = 0 + 0x - (1/\pi)x(x-\pi) + 0x^2(x-\pi) + (1/\pi^3)x^2(x-\pi)^2.$$

Y así hasta un total de 10 expresiones diferentes (escríbalas).

5.2.4 Para el error de interpolación tenemos el siguiente resultado que se demuestra como los correspondientes de los casos de Lagrange y Taylor.

TEOREMA

Con las notaciones del problema de Hermite, si f es de clase C_n en $[a,b]$ y tiene derivada $(n+1)$-ésima en (a,b) entonces para cada x en $[a,b]$ existe un punto ξ en el interior del menor intervalo cerrado que contenga a x y a cada x_i, $0 \leq i \leq r$ tal que si $P_n(x)$ es la solución del problema entonces

$$f(x) - P_n(x) = \frac{(x-x_0)^{m_0+1} \cdots (x-x_r)^{m_r+1}}{(n+1)!} f^{(n+1)}(\xi).$$

Por ejemplo, para el polinomio cúbico que reproduce a f y a f' en x_0, x_1 si nos limitamos a evaluar el interpolante entre x_0, x_1 se tiene

$$|f(x) - P_3(x)| \leq \frac{|x_1 - x_0|^4}{384} M_4, \tag{5.5}$$

siendo M_4 una cota de la derivada cuarta de f en el intervalo cerrado de extremos x_0, x_1. Basta observar que $(x-x_0)(x-x_1)$ toma el valor máximo $\frac{|x_1-x_0|^2}{4}$ en el punto medio del intervalo.

5.3 Diferencias divididas con argumentos repetidos

En la forma de Newton para el problema de Lagrange en los $n+1$ nodos distintos $x_0^*, \ldots, x_n^*$ los polinomios $n+1$ de base son

$$1, (x-x_0^*), (x-x_0^*)(x-x_1^*), \ldots, (x-x_0^*)\cdots(x-x_{n-1}^*). \tag{5.6}$$

La base (5.3) del problema de Hermite sería de la forma (5.6) si autorizásemos a los x_i^* a no ser distintos y tomásemos los m_0+1 primeros nodos $x_0^*, \ldots, x_{m_0}^*$ todos iguales entre sí e iguales a x_0, luego un bloque de m_1+1 nodos iguales entre sí $x_{(m_0+1)}^*, \ldots, x_{(m_0+1)+m_1}^*$ e iguales a x_1, etc. (En el caso del ejemplo 5.2.3 anterior la sucesión de nodos para la primera expresión sería 0, 0, 0, π , π .)

También la base (5.4) es de la forma (5.6) si se permiten nodos repetidos (para la segunda expresión del mismo ejemplo tenemos 0, π , 0, π, 0). Más generalmente, cada base asociada a un orden admisible de tomar las condiciones (5.2) da origen a una sucesión *ordenada* de $n+1$ elementos escogidos de entre el conjunto de $r+1$ nodos distintos, figurando m_i+1 veces cada nodo x_i.

En consecuencia, adoptaremos el *convenio* de que un polinomio P_n interpola a una función f en una sucesión de nodos $x_0^*, \ldots, x_n^*$, *no necesariamente distintos*, si f y P_n coinciden en x_j^* y lo mismo les sucede a sus m_i primeras derivadas cada vez que el valor x_j^* aparece m_i+1 veces en la lista de nodos. En el ejemplo 5.2.3 nuestro polinomio interpola a la función $\operatorname{sin} x$ en las sucesiones nodos 0, 0, 0, π, π o en cualquier permutación (con repetición) de sus elementos.

Con esta nomenclatura el polinomio interpolador en una sucesión $x_0^*, \ldots, x_n^*$ se escribe en la forma

$$c_0 + c_1(x - x_0^*) + c_2(x - x_0^*)(x - x_1^*) + \ldots + c_n(x - x_0^*)\cdots(x - x_{n-1}^*) \tag{5.7}$$

siendo los c_i constantes adecuadas, que ya hemos visto cómo determinar por recurrencia. Claramente c_i, $0 \leq i \leq n$ es el coeficiente de x^i en el polinomio interpolador de grado $\leq i$ en $x_0^*, \ldots, x_i^*$ y por tanto depende sólo de f y $x_0^*, \ldots, x_i^*$. Cuando los x_i^* eran distintos hemos escrito $c_i = f[x_0^*, \ldots x_i^*]$ y, tras la consideración anterior, es lícito extender esta notación al caso en que los nodos pueden repetirse. También seguimos diciendo que c_i es una diferencia dividida de orden i de f.

Observemos que, para poder definir las diferencias divididas con algún nodo repetido, la función f tiene que tener derivadas en él (tantas como veces aparezca el nodo menos una).

5.3.1 Para calcular diferencias divididas conviene notar las siguientes reglas.

(i) Las diferencias divididas no dependen del orden en que se escriban sus argumentos. (Ilustración: para los ejemplos 5.2.3 anteriores $\sin[0,0,0,\pi,\pi] = \sin[0,\pi,0,\pi,0] = 1/\pi^3$.)

(ii) Cuando todos los argumentos toman un mismo valor x_0 la diferencia dividida i-ésima de f vale $f^{(i)}(x_0)/i!$.

(iii) Si entre los argumentos de $f[x_0^*, \ldots x_i^*]$ hay dos con valores distintos (que según (i) podremos suponer son el primero y el último) entonces tal diferencia dividida puede calcularse a partir de dos diferencias divididas de orden $i-1$ de acuerdo con

$$f[x_0^*, \ldots x_i^*] = \frac{f[x_1^*, \ldots x_i^*] - f[x_0^*, \ldots x_{i-1}^*]}{x_i^* - x_0^*}. \tag{5.8}$$

La validez de (i) y (iii) se demuestra exactamente igual que en el caso de nodos distintos, mientras que (ii) es una consecuencia de la conocida forma del polinomio de Taylor.

Con las tres reglas anteriores es fácil construir una tabla de diferencias divididas en cuya diagonal figuren los coeficientes necesarios para escribir el polinomio osculador en la base (5.6). Para el ejemplo 5.2.3 tendremos

x	y	$\triangle$	$\triangle^2$	$\triangle^3$	$\triangle^4$
0	$\underline{0}$				
0	$\underline{0}$	$\underline{1}$			
0	$\underline{0}$	$\underline{1}$	$\underline{0/2}$		
π	$\underline{0}$	0	$-1/\pi$	$-1/\pi^2$	
π	$\underline{0}$	$\underline{\text{-}1}$	$-1/\pi$	0	$1/\pi^3$

Los números que aparecen subrayados se escriben directamente de los valores conocidos de la función y sus derivadas, *teniendo cuidado de escalar con los factoriales*. Los restantes elementos se obtienen por la fórmula (5.8). En esta tabla están también los coeficientes de la otra expresión del ejemplo, ¿dónde? y ¿por qué?

5.3.2 Cálculo infinitesimal y cálculo en diferencias finitas. Escribamos, en forma de Newton, la recta que interpola (lagrangianamente) a f en dos puntos

$$f(x_0) + f[x_0, x_1](x - x_0) \tag{5.9}$$

y consideremos fijo el valor de x_0, tomando x_1 como un parámetro. Para el valor $x_1 = x_0$, la diferencia dividida en (5.9) carece de sentido al ser un cociente ($\frac{f(x_1)-f(x_0)}{x_1-x_0} = \frac{0}{0}$) indeterminado. Esto está de acuerdo con el hecho de que no podemos interpolar lagrangianamente sino en puntos distintos. Sin embargo, si f es derivable en x_0 existe el límite de tal cociente cuando x_1 tiende a confundirse x_0 y vale $f'(x_0)$. Así (5.9) tiende a la recta tangente

$$f(x_0) + f'(x_0)(x - x_0), \tag{5.10}$$

de la que por tanto parece razonable decir que interpola a f en dos puntos confundidos. Esto concuerda con el *convenio* que introdujimos en la sección previa, porque (5.10) es la única recta cuyo valor y el de su derivada coinciden con los de f en x_0. Por otro lado estas observaciones muestran que, para x_0 fijo $f[x_0, x_1]$ es una función continua de x_1 y que podíamos también haber definido el valor de la diferencia dividida en nodos coincidentes ($f[x_0, x_0] = f'(x_0)$), como límite de valores $f[x_0, x_1]$ en nodos distintos.

De hecho, fijada f de clase C^i, la diferencia dividida i-ésima es una función continua de sus $i+1$ variables. Por tanto la extensión hecha en la sección precedente del concepto de *diferencia dividida* para incorporar valores repetidos es la única extensión continua del concepto original de diferencia dividida relativo a puntos distintos. Además, fijado un valor de la variable x, el valor del interpolante de Hermite $P_n(x)$ es función continua de los nodos, lo cual prueba que un polinomio osculador es límite de polinomios de Lagrange. Ninguno de estos resultados debe sorprendernos: justamente el concepto de derivada se ha introducido en el análisis para que las cosas resulten de este modo.

5.4 Caso a trozos. Cúbicas de Hermite segmentarias

Como ya es bien conocido, interpolar por polinomios de grado alto es, en general, muy poco recomendable. Los elementos de los espacios de funciones continuas lineales a trozos $M_0^1(\Delta)$, de los espacios de funciones continuas cuadráticas a trozos $M_0^2(\Delta)$, etc., suministran interpolantes mucho más ventajosos. Sin embargo éstos nuevos interpolantes son a su vez inutilizables en muchas aplicaciones debido a que *no son derivables* en los puntos de la partición Δ, en cuya vecindad el interpolante queda definido por expresiones analíticas distintas a izquierda y derecha. En esta lección consideraremos interpolantes polinómicos a trozos más regulares que los hasta ahora estudiados.

5.4.1 Definición. Si Δ es una partición fijada $a = x_0 < x_1 < \ldots < x_n = b$ del intervalo $[a, b]$ en que estemos interesados, denotamos por $M_1^3(\Delta)$ el espacio formado por las funciones reales H de clase C^1 en $[a, b]$ (esto es, continuas con derivada continua) que restringidas a cada subintervalo (x_{i-1}, x_i) de la partición coinciden con un polinomio $H^{(i)}$ de grado ≤ 3, para $i = 1, \ldots, n$.

Aclaremos el concepto con un ejemplo. Si la partición origina dos únicos subintervalos (-1,0), (0,1), la función q que en [-1,0] coincide con $2+x$ y en [0,1] coincide con $2+x+x^3$ está en el correspondiente M_1^3. Basta observar dos cosas.

(i) Que para $x = 0$ se obtiene $q(0) = 2$ tanto usando la expresión relativa a [-1,0] como usando la expresión relativa a [0,1], lo cual muestra que q está bien definida y es continua en 0 al existir los límites laterales y coincidir.

(ii) Que $q'(0) = 1$ usando cualquiera de las expresiones lo cual implica (¡demuéstrelo!) que q es derivable en 0 con derivada 1 y por tanto q continuamente diferenciable en [-1,1].

5.4.2 PROPOSICION

Si f es una función definida en $[a, b]$ y derivable en los puntos de Δ, hay un único elemento H en $M_1^3(\Delta)$ tal que

$$f(x_i) = H(x_i), f'(x_i) = H'(x_i), \qquad i = 0, 1, \ldots, n. \tag{5.11}$$

Demostración: En efecto, en cada $[x_{i-1}, x_i]$ tenemos cuatro datos $f(x_{i-1})$, $f(x_i)$, $f'(x_{i-1})$, $f'(x_i)$ que determinan unívocamente un polinomio cúbico $H^{(i)}$ (¿por qué?). Los n segmentos polinómicos definen en $[a, b]$ una función H, que es continuamente diferenciable como fácilmente se demuestra utilizando el mismo argumento que en el ejemplo anterior. □

Este resultado implica en particular que la dimensión de $M_1^3(\Delta)$ es $2n+2$ (¿por qué?). El interpolante H se llama *cúbico de Hermite a trozos.*

5.4.3 Para el error, y como consecuencia inmediata de 5.5

$$||f(x) - H(x)||_\infty \leq \frac{h^4}{384} M_4,$$

donde la norma es la del supremo en $[a, b]$, h denota el diámetro de la partición y M_4 es una cota en $[a, b]$ de la derivada cuarta de f. Se concluye que una función de clase C^4 se puede aproximar tanto como se desee por cúbicas de Hermite a trozos, con sólo refinar convenientemente la partición. Además el interpolante converge a la función con orden cuatro.

Por otra parte, aunque para la cota utilizamos una constante M_4 que mayora a la derivada cuarta de la función en todo el intervalo, es evidente que en cada uno de los trozos se puede tomar una cota de la citada derivada en el correspondiente subintervalo y la longitud del mismo como h. De modo que si estamos en situación de poder elegir los puntos de interpolación siempre podemos espaciarlos de modo que los intervalos sean menores en las zonas en que la derivada (cuarta en este caso) sea muy grande y más grandes en aquellas otras zonas en que la derivada sea menor. De esta forma se pueden obtener errores relativamente pequeños con pocos puntos de interpolación adecuadamente distribuidos.

5.4.4 Para manejar una función H de $M_1^3(\Delta)$ (sea interpolante que satisface las condiciones (5.11) para una función dada f o no), es usual guardar las ecuaciones de sus n segmentos cúbicos $H^{(i)}$. Si empleamos la representación tipo Newton, tendremos en $[x_{i-1}, x_i]$, $1 \leq i \leq n$

$$\begin{aligned} H(x) = H^{(i)}[x_{i-1}] \quad & +H^{(i)}[x_{i-1}, x_{i-1}](x - x_{i-1}) \\ & +H^{(i)}[x_{i-1}, x_{i-1}, x_i](x - x_{i-1})^2 \\ & +H^{(i)}[x_{i-1}, x_{i-1}, x_i, x_i](x - x_{i-1})^2(x - x_i), \end{aligned} \tag{5.12}$$

en la que las necesarias diferencias divididas, que mantendríamos en memoria, se habrían obtenido de la diagonal de la tabla

$$\begin{array}{llll} x_{i-1} & H_{i-1} & & \\ x_{i-1} & H_{i-1} & H'_{i-1} & \\ x_i & H_i & H[x_{i-1}, x_i] & \dfrac{H[x_{i-1}, x_i] - H'_{i-1}}{(x_i - x_{i-1})} \\ x_i & H_i & H'_i & \dfrac{H'_i - H[x_{i-1}, x_i]}{(x_i - x_{i-1})} \qquad \dfrac{H'_{i-1} - 2H[x_{i-1}, x_i] + H'_i}{(x_i - x_{i-1})^2} \end{array} \tag{5.13}$$

donde por claridad no hemos escrito el desarrollo explícito de $H[x_{i-1}, x_i]$ y hemos empleado abreviaturas como $H_{i-1} = H(x_{i-1})$, $H'_{i-1} = H'(x_{i-1})$, etc... Así hemos obtenido la representación del segmento cúbico en términos de los valores de H y H' en x_{i-1} y x_i, valores que en el problema de interpolación (5.1) son datos.

5.4.5 Por otro lado, podemos usar la representación en potencias de $x - x_{i-1}$ (que es más conveniente que la (5.12), sobre todo si hay que evaluar no sólo H sino también sus derivadas). Entonces para $x_{i-1} \leq x \leq x_i$,

$$\begin{aligned} H(x) = H^{(i)}[x_{i-1}] \quad & +H^{(i)}[x_{i-1}, x_{i-1}](x - x_{i-1}) \\ & +H^{(i)}[x_{i-1}, x_{i-1}, x_{i-1}](x - x_{i-1})^2 \\ & +H^{(i)}[x_{i-1}, x_{i-1}, x_{i-1}, x_{i-1}](x - x_{i-1})^3 \end{aligned} \tag{5.14}$$

Los coeficientes a almacenar se obtienen a partir de los valores de H y H' en x_{i-1} y x_i, utilizando la tabla (5.13), la definición recursiva de las diferencias divididas y teniendo en cuenta que las diferencias terceras de una cúbica son constantes. Concretamente

$$\begin{aligned} H^{(i)}[x_{i-1}, x_{i-1}, x_{i-1}, x_{i-1}] &= \frac{H'_{i-1} - 2H[x_{i-1}, x_i] + H'_i}{(x_i - x_{i-1})^2} \\ H^{(i)}[x_{i-1}, x_{i-1}, x_{i-1}] &= H^{(i)}[x_{i-1}, x_{i-1}, x_i] \\ &\quad -(x_i - x_{i-1})H^{(i)}[x_{i-1}, x_{i-1}, x_{i-1}, x_i] \\ &= \frac{3H[x_{i-1}, x_i] - 2H'_{i-1} - H'_i}{(x_i - x_{i-1})} \\ H^{(i)}[x_{i-1}, x_{i-1}] &= H'_{i-1} \\ H^{(i)}[x_{i-1}] &= H_{i-1}, \end{aligned} \tag{5.15}$$

Obsérvese que las derivadas segunda y tercera en un punto x_i son, en general, diferentes por la izquierda y la derecha.

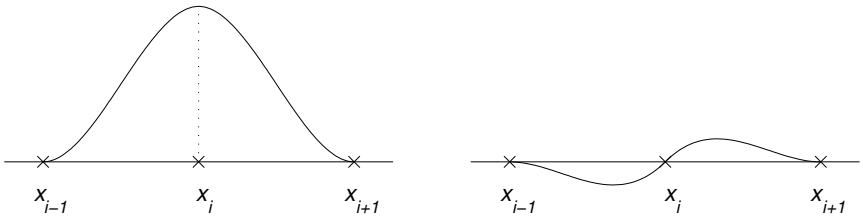


Figura 5.1: Elementos Φ_i y Θ_i en un punto interior

5.4.6 En cualquiera de las representaciones (5.12) y (5.14), útiles en la práctica, se almacenan $4n$ coeficientes, lo que contrasta con la dimensión $2n+2$ del espacio (¿cuál es la explicación?). En ciertas ocasiones conviene representar H en términos de una base. La base más usual es (véase la figura 5.1) la formada por las funciones $\Phi_i(x)$ y $\Theta_i(x)$ de $M_1^3(\Delta)$, $0 \leq i \leq n$, tales que $\Phi_i(x_j) = \delta_{ij}$, $\Phi_i'(x_j) = 0$, $\Theta_i(x_j) = 0$, $\Theta_i'(x_j) = \delta_{ij}$.

En este base se tiene (¿por qué?)

$$H(x) = \sum_{0 \leq i \leq n} \left[H(x_i)\Phi_i(x) + H'(x_i)\Theta_i(x) \right]. \tag{5.16}$$

Fácilmente se comprueba que las funciones de esta base poseen soporte local (a lo sumo dos subintervalos), lo cual muestra que la interpolación cúbica de Hermite a trozos posee carácter local.

5.5 Cuestiones y problemas

5.5.1 Caracterización analítica del polinomio osculador. Con las notaciones del problema de Hermite pruebe que P_n es el único polinomio de grado $\leq n$ tal que $f(x) - P_n(x) = \circ((x - x_i)^{m_i})$ cuando $x \to x_i$, $1 \leq i \leq r$. Así las gráficas de f y P_n se acercan entre sí suavemente en cada nodo x_i con $m_i > 0$. Este es el origen etimológico de osculador (ósculo = beso).

5.5.2 Pruebe que los ceros del polinomio de Chebyshev T_{n+1} dan las $n+1$ posiciones óptimas de interpolación en [-1, 1] incluso si se autoriza a que esas posiciones puedan ser coincidentes. ¿Cuál es la peor elección?

5.5.3 Pruebe, sin utilizar resultados de esta lección, que para funciones regulares el polinomio de Taylor en x_0 es límite de polinomios de Lagrange relativo a $n+1$ puntos que tienden a x_0. ¿Es el límite uniforme?

5.5.4 Interpolación de Hermite en sentido estricto. El caso del problema de Hermite en que cada m_i es 1 (es decir se reproducen funciones y derivadas primeras) se llama interpolación de Hermite en sentido estricto. Sean x_i, $0 \leq i \leq r$ los nodos dos a dos distintos y denotemos por L_i los correspondientes polinomios de base de Lagrange

(es decir cada L_i posee grado $\leq r$ y además $L_i(x_j) = \delta_{ij}$, $0 \leq i, j \leq r$). Pruebe que los $2r+2$ polinomios

$$[1 - 2(x - x_i)L_i'(x_i)]L_i(x)^2, (x - x_i)L_i(x)^2$$

dan una base en la que el polinomio solución del problema de Hermite estricto tiene por coeficientes los valores de f y de f' en los nodos.

5.5.5 Evaluación de derivadas en la forma de Newton. Sea $p(x)$ el polinomio

$$c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + \ldots + c_n(x - x_0)\cdots(x - x_{n-1})$$

donde no se supone que los x_i sean dos a dos distintos. Pruebe que cuando se evalúa $p(x^*)$ por el algoritmo de Horner, es decir se toman

$$b_n = c_n, b_{k-1} = c_{k-1} + (x^* - x_{k-1})b_k, \qquad k = n, n-1, \ldots, 1$$

resulta $p(x^*) = b_0$. Los resultados intermedios b_k son útiles porque se puede escribir

$$p(x) = b_0 + b_1(x - x^*) + b_2(x - x^*)(x - x_0) + \ldots + b_n(x - x^*)(x - x_0)\cdots(x - x_{n-2}),$$

representación análoga a la de partida donde x^* ha desplazado a x_{n-1}. Si itera el algoritmo con el mismo valor para x^* acaba obteniendo el desarrollo de $p(x)$ en potencias de $(x - x^*)$, de donde conoce inmediatamente las derivadas sucesivas de p en x^*.

Precaución: Si el p dado es el polinomio interpolador de una cierta función f en $x_0, \ldots, x_n$ entonces la nueva representación **NO** es el polinomio interpolador de f en $x^*, x_0, \ldots, x_{n-1}$. Basta pensar que en general $f(x^*) \neq p(x^*)$. Lo que en realidad tenemos es un polinomio del mismo grado que el anterior y que le interpola en $n+1$ puntos, por lo que ambos han de coincidir.

5.5.6 Consideremos un polinomio de tercer grado que en las abscisas 1, 2, 3 y 4 toma los valores 1, 7, 25 y 61 respectivamente. Calcular su valor y el de sus derivadas hasta el tercer orden en los puntos 0 y 5. Los cálculos deben hacerse numéricamente y utilizando únicamente dos vectores de datos: uno guardará abscisas y el otro valores diversos del polinomio y sus diferencias divididas.

5.5.7 La interpolación lineal desde un punto de vista abstracto. Compruebe que todos los casos de interpolación estudiados hasta ahora son casos particulares del siguiente problema abstracto. Dado un espacio vectorial X de dimensión finita $n+1$, $n+1$ formas lineales L_i sobre X y $n+1$ escalares f_i, $0 \leq i \leq n$ encontrar un elemento p (el interpolante) en X tal que $L_i(p) = f_i$ para cada i, $0 \leq i \leq n$.

5.5.8 Problemas de interpolación lineal unisolventes. Pruebe que con las notaciones de la cuestión precedente, el problema abstracto tiene solución única para cada elección de f_i si y sólo si las formas L_i son independientes, o también si y sólo si el problema sólo tiene la solución nula cuando los datos f_i son todos nulos. Cuando se verifican estas condiciones necesarias y suficientes se dice que el problema es unisolvente.

5.5.9 Como aplicación de la cuestión anterior, dé una prueba breve del teorema 5.2.1.

5.5.10 Compare las cotas de error para la interpolación cúbica a trozos de Hermite y para la interpolación cúbica a trozos de Lagrange (donde recordemos que el interpolante coincide con f en cada nodo de la partición y además en cada uno de los puntos que sirven para dividir los subintervalos en cuatro partes iguales). ¿Qué ventajas e inconvenientes cree que tiene cada tipo de interpolación?

5.5.11 Interpolación quíntica con continuidad C^2. Considere el espacio $M_2^5(\Delta)$ de funciones de clase C^2 que restringidas a cada subintervalo de la partición de la sección 5.4 coinciden con un polinomio de grado menor o igual que cinco. Reproduzca para este espacio todas las consideraciones hechas en dicha sección para $M_1^3(\Delta)$.

2010-11

Lección 6

Splines

6.1 Definición y construcción de splines cúbicos

El espacio $M_1^3(\Delta)$ es claramente un subespacio del $M_0^3(\Delta)$, donde a los segmentos cúbicos sólo se les exige coincidir en los nodos. Dar n segmentos cúbicos equivale a dar n cuaternas de números reales, y por ello puede hacerse con $4n$ grados de libertad. No todas de esas $4n$ elecciones conducen a elementos de $M_0^3(\Delta)$, puesto que para que un grupo de n cuaternas defina una función continua en $[a, b]$ han de imponérsele las $n-1$ condiciones independientes que expresan la continuidad en cada nodo interno $x_2, \ldots, x_{n-1}$. Con $4n$ parámetros y $n-1$ condiciones, restan $3n+1$ parámetros libres en $M_0^3(\Delta)$, que, como sabemos, se pueden usar, por ejemplo, para reproducir los valores de una función dada en los $n+1$ nodos de la partición y los $2n$ puntos que sirven para dividir los subintervalos de la misma en tres partes iguales.

Ahora para que un grupo de n cuaternas defina una función de $M_1^3(\Delta)$ habremos de imponerle las $n-1$ condiciones anteriores de continuidad en los nodos internos y otras $n-1$ nuevas condiciones para la continuidad de la derivada. Así $M_1^3(\Delta)$ tiene, como vimos, dimensión $4n-2(n-1) = 2n+2$, número igual al de condiciones en (5.11). En resumen al pasar de $M_0^3(\Delta)$ a $M_1^3(\Delta)$ *se incrementa la regularidad de los interpolantes, a costa de sacrificar el número de grados de libertad*; es decir, de condiciones que se pueden exigir a los mismos.

6.1.1 Siguiendo en la línea anterior es todavía posible considerar el espacio $M_2^3(\Delta)$, llamado de splines (*léase splains*) cúbicos, formado por las funciones de clase C^2 en $[a, b]$ que restringidas a cada subintervalo de la partición coinciden con un polinomio cúbico. Supongamos que tenemos un elemento de $M_1^3(\Delta)$ dado por sus segmentos polinómicos en la forma (5.14), con los coeficientes obtenidos, vía (5.15), de los $2n+2$ parámetros libres H_i, H_i', $i = 0, \ldots, n$. Dado que, con frecuencia los valores de $f'(x_i)$ son desconocidos y es preciso aplicar unos valores apróximados a los mismos como condiciones H_i' para establecer el interpolante de Hermite, se aprovecha esta libertad de elección para imponer que la segunda derivada del interpolador a trozos resulte también continua. Las H_i' resultarán ahora incógnitas de un sistema lineal de ecuaciones.

En efecto, para que tal elemento esté en $M_2^3(\Delta)$ se necesita y basta con que coincidan las evaluaciones en x_i, $i = 1, \ldots n-1$ de la derivada segunda de los segmentos polinómicos $H^{(i)}$, $H^{(i+1)}$ correspondientes a $[x_{i-1}, x_i]$ y $[x_i, x_{i+1}]$. Por (5.14), la evaluación en el

segmento a la izquierda de x_i conduce a

$$2H^{(i)}[x_{i-1}, x_{i-1}, x_{i-1}] + 6H^{(i)}[x_{i-1}, x_{i-1}, x_{i-1}, x_{i-1}](x_i - x_{i-1}),$$

que usando (5.15) equivale a

$$\frac{-6H[x_{i-1}, x_i] + 2H'_{i-1} + 4H'_i}{x_i - x_{i-1}}. \tag{6.1}$$

Por otro lado la derivada segunda en x_i utilizando el segmento a la derecha es claramente $2H^{(i+1)}[x_i, x_i, x_i]$, o por (5.15) (con i en vez de $i-1$)

$$\frac{6H[x_i, x_{i+1}] - 4H'_i - 2H'_{i+1}}{x_{i+1} - x_i}. \tag{6.2}$$

Igualando (6.1) y (6.2) para $i = 1, \ldots, n-1$, se obtiene el siguiente sistema de ecuaciones que una función cúbica de Hermite a trozos satisface si y sólo si es un spline cúbico

$$\begin{gathered}\frac{1}{x_i - x_{i-1}} H'_{i-1} + \left(\frac{2}{x_i - x_{i-1}} + \frac{2}{x_{i+1} - x_i}\right) H'_i + \frac{1}{x_{i+1} - x_i} H'_{i+1} = \\ 3\left(\frac{H_i - H_{i-1}}{(x_i - x_{i-1})^2} + \frac{H_{i+1} - H_i}{(x_{i+1} - x_i)^2}\right)\end{gathered} \tag{6.3}$$

Observamos que el número de incógnitas es $n+1$ por lo que el sistema admite infinitas soluciones. Concretamente, disponemos de dos grados de libertad para generar una familia doblemente infinita de *splines*. A la misma conclusión se puede llegar razonando sobre las dimensiones de los espacios, pues si a los $2n+2$ grados de libertad que nos quedaban, les restamos las $n-1$ nuevas condiciones impuestas, resultan $n+3$ (la dimensión del espacio de *splines* cúbicos) y como sólo disponemos de los $n+1$ valores en los puntos para determinarlos, aún disponemos de dos grados de libertad para exigir alguna otra cualidad del *spline* solución.

La elección de una u otra de estas soluciones puede hacerse en base a distintos principios: bien añadiendo nuevas condiciones que se traducen en un aumento del número de ecuaciones para el sistema (véase el problema 6.3.2), bien eliminando alguna de las incógnitas mediante la adjudicación de un valor a las mismas. El caso más frecuente es el siguiente:

6.1.2 Supongamos que además de H_i, $i = 0, \ldots, n$, también conocemos H'_0, H'_n. Entonces (6.3) es el sistema lineal de $n-1$ ecuaciones para las $n-1$ pendientes incógnitas H'_i, $i = 1, \ldots, n-1$ con matriz

$$\begin{pmatrix}
\frac{2}{h_1} + \frac{2}{h_2} & \frac{1}{h_2} & 0 & \cdots & \cdots & 0 & 0 \\
\frac{1}{h_2} & \frac{2}{h_2} + \frac{2}{h_3} & \frac{1}{h_3} & \cdots & \cdots & 0 & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots & \vdots & \vdots \\
0 & \cdots & \frac{1}{h_i} & \frac{2}{h_i} + \frac{2}{h_{i+1}} & \frac{1}{h_{i+1}} & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & \cdots & \frac{1}{h_{n-2}} & \frac{2}{h_{n-2}} + \frac{2}{h_{n-1}} & \frac{1}{h_{n-1}} \\
0 & 0 & \cdots & \cdots & 0 & \frac{1}{h_{n-1}} & \frac{2}{h_{n-1}} + \frac{2}{h_n}
\end{pmatrix} \tag{6.4}$$

y término independiente el vector

$$\begin{pmatrix} 3\left(\frac{d_1}{h_1}+\frac{d_2}{h_2}\right)-\frac{1}{h_1}H'_0 \\ 3\left(\frac{d_2}{h_2}+\frac{d_3}{h_3}\right) \\ \vdots \\ 3\left(\frac{d_i}{h_i}+\frac{d_{i+1}}{h_{i+1}}\right) \\ \vdots \\ 3\left(\frac{d_{n-2}}{h_{n-2}}+\frac{d_{n-1}}{h_{n-1}}\right) \\ 3\left(\frac{d_{n-1}}{h_{n-1}}+\frac{d_n}{h_n}\right)-\frac{1}{h_n}H'_n \end{pmatrix} \tag{6.5}$$

donde de nuevo hemos abreviado en la escritura, siendo

$$h_i = x_i - x_{i-1} \quad \text{y} \quad d_i = H[x_{i-1}, x_i] = \frac{H_i - H_{i-1}}{h_i}$$

.

La correspondiente matriz se llama *tridiagonal*, porque sus elementos no nulos sólo se encuentran en la diagonal principal y las dos a ella adyacentes y *estrictamente diagonalmente dominante* porque en cada fila el término diagonal excede a la suma de los módulos de los restantes elementos. Se demuestra (cuestión 6.3.4) que A es regular, con lo que (6.3) tiene solución única y hemos probado:

PROPOSICION

Dada una función f en $[a, b]$, derivable en a y en b, existe un único elemento S de $M_2^3(\Delta)$ tal que

$$S(x_i) = f(x_i), \qquad i = 0, \ldots, n, \qquad S'(a) = f'(a), S'(b) = f'(b).$$

Este *spline* S se llama el interpolante *completo* de f. En síntesis, para manejar S ante todo resolveremos el sistema (6.4), (6.5) para hallar las pendientes del spline en los nodos (¡que NO coincidirán en general con las de f!), luego iremos a (5.15) para hallar los $4n$ coeficientes de los segmentos cúbicos, que quedarán almacenados. Cada vez que sea preciso evaluar S en un punto x^* comenzaremos por determinar en qué subintervalo se halla x^* y luego evaluaremos la correspondiente cúbica (5.14) por medio del algoritmo de Horner.

Para el *spline* interpolante completo S de f se demuestra que, para x en $[a, b]$

$$|f(x) - S(x)| \leq (5/384)h^4 M_4,$$

donde, como siempre, h es el diámetro de la partición y M_4 una cota superior de la derivada cuarta de f.

6.1.3 Ejemplo Comsideremos un ejemplo muy sencillo que nos ayude a fijar todas estas ideas. A partir de la siguiente tabla de valores

x	1	2	3	4
$f(x)$	3	5	4	7
$f'(x)$	1	-1	2	3

calcularemos el interpolante cúbico de Hermite a trozos $H(x)$ y dos *splines*: el interpolante completo $S_1(x)$ en base a las derivadas conocidas en los extremos y el *spline* natural $S_2(x)$ para el supuesto en que estas fuesen desconocidas (véase el ejercicio 6.3.3).

Comencemos por el interpolante de Hermite. El primer trozo $H^{(1)}(x)$ es muy fácil de escribir en base a la siguiente tabla de diferencias divididas con puntos repetidos

x	y	$\triangle$	$\triangle^2$	$\triangle^3$
1	3			
1	3	1		
2	5	2	1	
2	5	-1	-3	-4

donde el significado del subrayado ya es conocido por el lector. Resulta pues

$$H^{(1)}(x) = 3 + (x-1) + (x-1)^2 - 4(x-1)^2(x-2)$$

que puede transformarse en múltiples expresiones de potencias según las necesidades y el uso que se le vaya a dar. Aunque es evaluable de manera muy eficiente en esta representación de Newton, mediante la regla de Horner, en general aporta más información un desarrollo de Taylor en el extremo inferior del intervalo. Una única aplicación del problema 5.5.5 con $x^* = 1$ nos permite obtener el dato que nos falta: la segunda derivada en dicho punto (o más exactamente su mitad). Así resulta finalmente

$$H^{(1)}(x) = 3 + (x-1) + 5(x-1)^2 - 4(x-1)^3$$

La expresión de los otros dos trozos en la misma versión es

$$H^{(2)}(x) = 5 - (x-2) - 3(x-2)^2 + 3(x-2)^3$$

y

$$H^{(3)}(x) = 4 + 2(x-3) + 2(x-3)^2 - (x-3)^3$$

y en su conjunto aparece en la figura 6.1 como una línea continua. Tabulamos los valores de sus derivadas a la izquierda y a la derecha de los nodos, para comparar posteriormente con lo que sucede en los *splines* que construiremos

x	1	2	3	4
$f(x)$	3	5	4	7
$f'(x)$	1	-1	2	3
$f''_-(x)$		-14	12	-2
$f''_+(x)$	10	-6	4	
$f'''_-(x)$		-24	18	-6
$f'''_+(x)$	-24	18	-6	

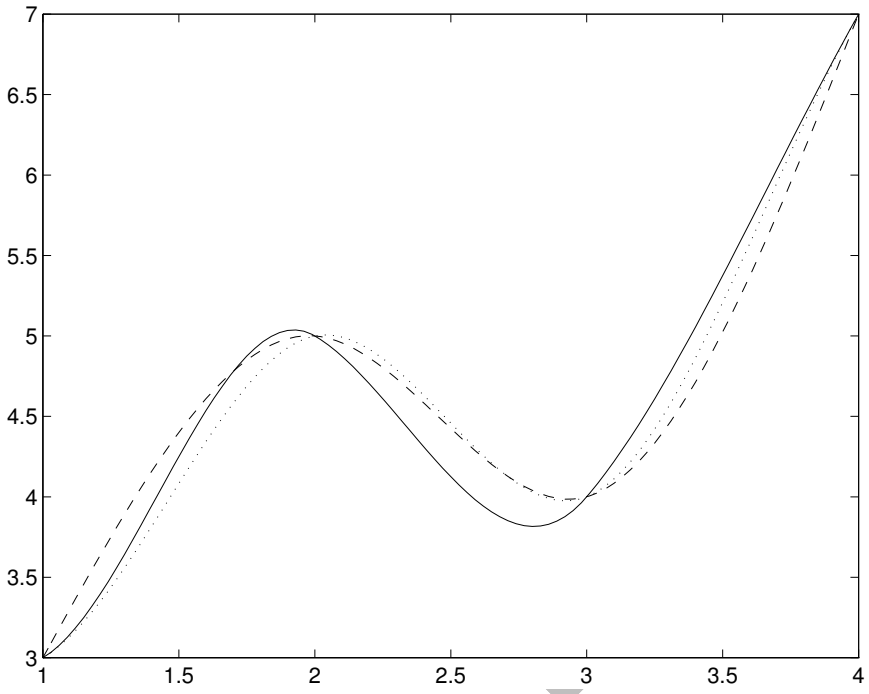


Figura 6.1: La cúbica de Hermite a trozos y los *splines* completo y natural

Para el cálculo del interpolante completo, el sistema que resulta es el siguiente

$$\begin{pmatrix} 4 & 1 \\ 1 & 4 \end{pmatrix} \begin{pmatrix} H_1' \\ H_2' \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \end{pmatrix}$$

cuya solucción obvia es $\frac{1}{3}$ y $\frac{2}{3}$ respectivamente. En consecuencia, una vez conocidas todas las pendientes, los polinomios de cada segmento se pueden escribir en la misma forma que antes, resultando:

$$S_1^{(1)}(x) = 3 + (x-1) + \frac{11}{3}(x-1)^2 - \frac{8}{3}(x-1)^3$$

$$S_1^{(2)}(x) = 5 + \frac{1}{3}(x-2) - \frac{13}{3}(x-2)^2 + 3(x-2)^3$$

$$S_1^{(3)}(x) = 4 + \frac{2}{3}(x-3) + \frac{14}{3}(x-3)^2 - \frac{7}{3}(x-3)^3$$

En su conjunto, el *spline* aparece en la figura 6.1 como una línea de puntos, y la tabla

correspondiente a sus valores y derivadas laterales es

x	1	2	3	4
$f(x)$	3	5	4	7
$f'(x)$	1	$\frac{1}{3}$	$\frac{2}{3}$	3
$f''(x)$	$\frac{22}{3}$	$-\frac{26}{3}$	$\frac{28}{3}$	$-\frac{14}{3}$
$f'''_-(x)$		-16	18	-14
$f'''_+(x)$	-16	18	-14	

El *spline* natural se obtiene resolviendo un caso particular del sistema ampliado por la incorporacion de dos nuevas ecuaciones, las que resultan de expresar las derivadas segundas en los extremos. De (6.2) se obtiene que

$$\frac{6d_1 - 4H'_0 - 2H'_1}{h_1} = H''_0, \quad \text{es decir,} \quad 4\frac{H'_0}{h_1} + 2\frac{H'_1}{h_1} = 6\frac{d_1}{h_1} - H''_0$$

y de (6.1) que

$$\frac{-6d_n + 2H'_{n-1} + 4H'_n}{h_n} = H''_n, \quad \text{es decir,} \quad 2\frac{H'_{n-1}}{h_n} + 4\frac{H'_n}{h_n} = 6\frac{d_n}{h_n} + H''_n$$

lo que traducido a nuestro caso particular ($H''_0 = H''_n = 0$), y teniendo en cuenta que $d_1 = 2$ y $d_n = 3$, desemboca en el sistema de cuatro ecuaciones y cuatro incógnitas que figura a continuación

$$\begin{pmatrix} 4 & 2 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 2 & 4 \end{pmatrix} \begin{pmatrix} H'_0 \\ H'_1 \\ H'_2 \\ H'_3 \end{pmatrix} = \begin{pmatrix} 12 \\ 3 \\ 6 \\ 18 \end{pmatrix}$$

cuya solucción es $H'_0 = \frac{46}{15}$, $H'_1 = -\frac{2}{15}$, $H'_2 = \frac{7}{15}$ y $H'_4 = \frac{64}{15}$. Por tanto, procediendo como en el caso anterior , calculamos los polinomios de cada segmento, que resultan ser:

$$S_2^{(1)}(x) = 3 + \frac{46}{15}(x-1) - \frac{16}{15}(x-1)^3$$
$$S_2^{(2)}(x) = 5 - \frac{2}{15}(x-2) - 3.2(x-2)^2 + \frac{7}{3}(x-2)^3$$
$$S_2^{(3)}(x) = 4 + \frac{7}{15}(x-3) + 3.8(x-3)^2 - \frac{19}{15}(x-3)^3$$

En su conjunto, el *spline* aparece en la figura 6.1 como una línea de trazos, y la tabla correspondiente a sus valores y derivadas laterales es

x	1	2	3	4
$f(x)$	3	5	4	7
$f'(x)$	$\frac{46}{15}$	$-\frac{2}{15}$	$\frac{7}{15}$	$\frac{64}{15}$
$f''(x)$	0	-6.4	7.6	0
$f'''_-(x)$		$-\frac{32}{5}$	14	-7.6
$f'''_+(x)$	$-\frac{32}{5}$	14	-7.6	

6.2 B-splines y otras bases

El manejo de los *splines* en la forma descrita, tiene la ventaja de resultar muy eficiente cuando hay que evaluar la función en un determinado punto, pues basta determinar el subintervalo al que pertenece y utilizar el algoritmo de Horner en alguna de las variantes. Sin embargo, hemos visto que su construcción es un tanto compleja, pues hay que resolver un sistema de ecuaciones para determinar las pendientes y después calcular los coeficientes en cada uno de los subintervalos. Existen otras formas más eficientes de cosntruir los *splines*, aunque también tienen sus inconvenientes.

Lo ideal, a la vista de lo que ocurre con otros tipos de interpolación, sería encontrar una base $\{\Phi_j\}_{j=1}^{n+3}$ que nos permita expresar cualquier *spline* como una combinación lineal

$$S(x) = \sum_{j=1}^{n+3} \lambda_j \Phi_j(x)$$

La primera base en que se piensa, teniendo en cuenta que estamos en un problema de interpolación, es en la formada por los *splines* denominados 'cardinales'. El elemento C_i sería tal que tomase el valor 1 en el punto x_i y 0 en todos los demás. La idea es conseguir, como en el caso de Laplace o de Hermite, que los coeficientes fuesen los valores de interpolación. Pero vemos que de esta forma tendríamos sólo $n+1$ *splines* que no pueden ser una base del espacio M_2^3 (¿por qué?). De hecho resulta difícil completar esta base con dos elementos adicionales que nos permitiesen incorporar las habituales condiciones frontera tantas veces mencionadas. Además son funciones cuyo soporte no es local y para la evaluación del *spline* en cualquier punto se necesita disponer de todos los cardinales a la vez y este valor se ve influído por los cambios que se producen en cualquier punto por alejado que esté. Pero no hay que preocuparse, porque existen otras muchas opciones para construir una base.

Por ejemplo, resulta bastante sencillo construir una base utilizando la función x_+^n que, por definición vale x^n si $x \geq 0$ y 0 en otro caso. Dicho de otra forma, $(x - x_0)_+ = \max(0, x - x_0)$.

Entonces, si Δ es la típica partición $a = x_0 < x_1 < \ldots < x_n = b$ del intervalo $[a, b]$, se consideran por una parte las cuatro funciones

$$1, (x - x_0), (x - x_0)^2, (x - x_0)^3$$

y por otro, las $n - 1$ siguientes

$$(x - x_1)_+^3, (x - x_2)_+^3, \ldots, (x - x_{n-1})_+^3,$$

Es evidente que todas ellas son *splines* (¿por qué?), y que cualquier otro en $[a, b]$ puede escribirse en la forma

$$c_0 + c_1(x - x_0) + c_2(x - x_0)^2 + c_3(x - x_0)^3 + \sum_{j=1}^{n-1} d_j (x - x_j)_+^3$$

pues lo que hacemos es: a partir de un polinomio inicial (de tercer grado) en el intervalo $[x_0, x_1]$, vamos añadiendo en todos los nodos interiores unos 'semipolinomios' (cúbicos

también) que corrigen el anterior para que cumpla las condiciones de interpolación, respetando la continuidad hasta la segunda derivada. Es evidente que la tercera derivada sufre un salto de $6d_j$ en el punto x_j para $j = 1, 2, \ldots, n-1$. El cálculo de los coeficientes depende del tipo de condiciones que se impongan (además de las $n+1$ interpolatorias) y se presta a multiples opciones como hemos visto. La imposición de ambas condiciones en el punto x_0 haría especialmente sencillo el proceso, pero aún en el caso habitual de una condición en cada extremo el 'sistema' de ecuaciones que tenemos que resolver no tiene la dificultad de (6.4), (6.5). En realidad, se trata simplemente de arrastrar un parámetro hasta imponer la condición del extremo final.

Pero esta base de tan sencilla construcción presenta graves inconvenientes para la evaluación del *spline* resultante, no sólo por tener que calcular muchos términos cuando hay muchos subintervalos (ya habrá notado el lector que los elementos de esta base tampoco tienen carácter local), sino porque presenta graves pérdidas de precisión por la cancelación que se produce al tener que evaluar polinomios cúbicos en puntos muy alejados del nodo raíz, cuando trabajamos con intervalos amplios. Por otra parte, esta clara la asimetría de las funciones de la base.

Entre las bases del espacio $M_2^3(\Delta)$, la más útil la constituyen ciertos splines cúbicos llamados B-splines. El objetivo que se persigue con esta base es que sus elementos tengan las propiedades de que adolecen las que hemos mencionado: un soporte lo más pequeño posible, es decir que sean distintos de cero en el menor número de subintervalos que permita su condición de *splines* y que sean fáciles de calcular los coeficientes dadas las condiciones interpolatorias y 'frontera'.

Para definirlos, comencemos eligiendo 6 nodos suplementarios $x_{-3} < x_{-2} < x_{-1} < a$, $x_{n+3} > x_{n+2} > x_{n+1} > b$. También consideraremos la función de dos variables $(t-x)_+^3$, que, por definición vale $(t-x)^3$ si $t \geq x$, y 0 en otro caso. Ahora para cada $i = -1, 0, 1, \ldots, n, n+1$ definimos una función real de variable real (B-spline) $B_i(x)$ como sigue:

$$B_i(x) = (x_{i+2} - x_{i-2})\{(.-x)_+^3[x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}]\}.$$

Aquí $(.-x)_+^3[x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2}]$ representa la diferencia dividida cuarta, con respecto a los puntos x_{i-2}, x_{i-1}, x_i, x_{i+1}, x_{i+2} de la función de la variable real t dada por $(t-x)_+^3$ (donde la x se considera como un parámetro). El valor de tal diferencia es un número real que depende del valor x en que se haya fijado el parámetro, y, justamente, $B_i(x)$ es el producto $(x_{i+2} - x_{i-2})$ por la diferencia. Note hay $n+3$ B-splines para una partición de $n+1$ puntos. Estos $n+3$ B-splines (restringidos a $[a, b]$) forman una base de $M_2^3(\Delta)$, y además una base de soporte local, pues el soporte de cada uno consta, cuando más, de cuatro subintervalos de la partición. (Cuestiones 6.3.17-6.3.18)

6.3 Cuestiones y problemas

6.3.1 *Caracterización variacional del spline interpolante completo. Significado de la palabra spline.* Demuestre que el spline cúbico interpolante completo de una función g es, de entre todas las funciones f de clase C^2 que coinciden con g en los nodos y satisfacen $f'(a) = g'(a)$, $f'(b) = g'(b)$ la única que hace mínima la integral en $[a, b]$ del cuadrado de f''. (Indicación: inspírese en lo hecho para la caracterización variacional del interpolante lineal a trozos.)

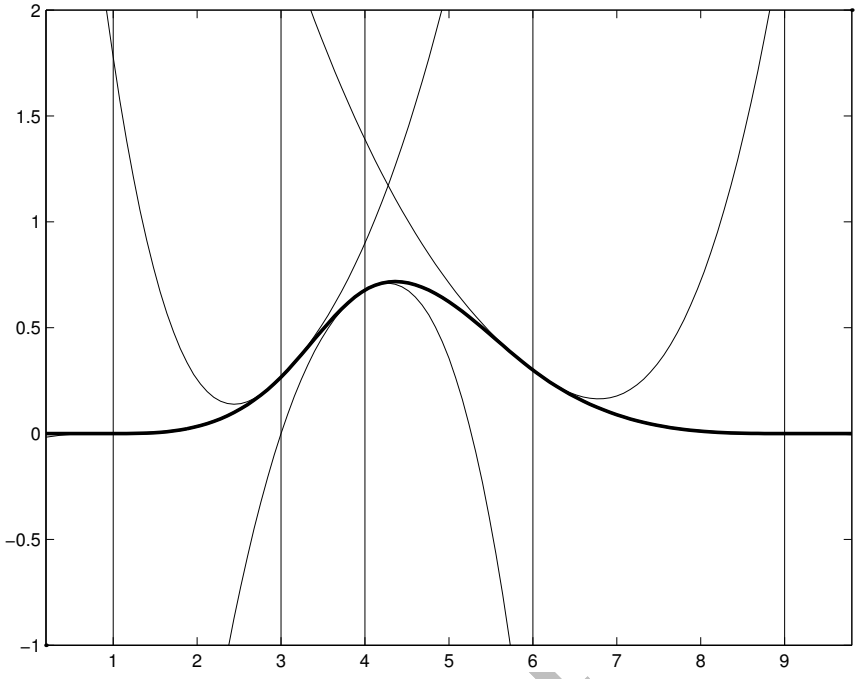


Figura 6.2: El B-spline [1 3 4 6 9] y sus polinomios componentes.

Como la integral en cuestión mide, en alguna manera, la curvatura de f (sería 0 si y solo si f fuese una recta), vemos que el interpolante completo es de entre todas las funciones que pasan por los puntos $(x_i, g(x_i))$ y tienen pendientes dadas en a y en b, la que menos se curva. En dibujo se usaban, para trazar una curva que se quería que pasase por puntos dados del plano, unas varillas de madera flexible fijadas en dichos puntos. En inglés esas varillas se denominaban splines. Como las varillas adoptan la posición de curvatura mínima (energía potencial mínima), se comprende que el nombre se extendiese a la función matemática que hoy conocemos como spline.

6.3.2 *Otras condiciones frontera para el spline cúbico.*

(i) Suponga que desconoce los valores de f' en a y b, pero conoce los de f''. Demuestre que, dada una partición, hay un único *spline* cúbico S en ella que coincide con f en cada nodo y tal que además $f'' = S''$ en $x = a, b$.

(ii) Suponga que no dispone ni de f' ni de f'' en los extremos del intervalo. Entonces, de entre la familia biparamétrica de splines cúbicos en una partición dada que coinciden con f en los nodos, conviene quedarse con el único que tiene derivada tercera continua en x_1, x_{n-1}. Pruebe que tal spline realmente existe y es único.

6.3.3 *Spline natural.* El spline cúbico s que coincide con una función dada g en los

nodos de la partición y satisface $s''(a) = s''(b) = 0$ se llama spline natural de g. Dé una caracterización variacional del mismo.

6.3.4 *Matrices estrictamente diagonalmente dominantes.* Demuestre que la matriz A del sistema (6.3), como cualquier otra matriz estrictamente diagonalmente dominante, es regular. (Indicación: reducción al absurdo; si A no es regular existe un vector no nulo $\mathbf{v}$ tal que $A\mathbf{v} = \mathbf{0}$; si $\mathbf{v}$ tiene su componente de módulo máximo en el lugar i, tome la i-ésima componente de $A\mathbf{v}$ y vea que no puede ser nula.)

6.3.5 Sea S un *spline* cúbico con nodos $t_0 < t_1 < \cdots < t_n$. Supongamos que en los dos intervalos $[t_1, t_2]$ y $[t_3, t_4]$, S se reduce a polinomios lineales. ¿Qué se puede decir de S sobre $[t_2, t_3]$?

6.3.6 Encontrar, calculando manualmente si es posible, el *spline* cúbico natural que interpola la siguiente tabla de valores

x	1	2	3	4	5
y	0	1	0	1	0

6.3.7 Describir explícitamente el *spline* cúbico natural que interpola una tabla con sólo dos entradas:

x	t_1	t_2
y	y_1	y_2

donde t_1 y t_2 son los nodos. Dar una fórmula para el *spline*.

6.3.8 Supongamos que $f(0) = 0$, $f(1) = 1.1752$, $f'(0) = 1$, $f'(1) = 1.5431$. Determinar el interpolante polinomial cúbico $p_3(x)$ para estos datos. ¿Es un *spline* natural este interpolante?

6.3.9

a) Encontrar la matriz de coeficientes y el vector de términos independientes para calcular el *spline* cúbico determinado por los datos siguientes y la condición de "sin nodo" en los extremos.

x	0.15	0.76	0.89	1.07	1.73	2.11
y	0.3945	0.2989	0.2685	0.2251	0.0893	0.0431

b) Resolver las ecuaciones por cualquier sistema, y determinar después las constantes de los diversos segmentos cúbicos. (Los datos son ordenadas de la función normal de probabilidad. Comparar algunos valores interpolados con los valores tabulados o calculados de la función, digamos en $x = 0.30, 0.80, 1.50$ y 2.00.

6.3.10 Considerar la siguiente tabla de datos (los cuales pertenecen obviamente a $f(x) = 1/x$)

x	1.0	1.5	2.0	2.5	3.0
y	1.000	0.667	0.500	0.400	0.333

Encontrar valores de $f'(x)$ y $f''(x)$ en $x = 1.5, 2.0$ y 2.5 del *spline* cúbico natural que aproxima a $f(x)$. Comparar con los valores analíticos de $f'(x)$ y $f''(x)$ para determinar sus errores.

6.3.11 Un *spline* cúbico *periódico* con nodos $t_0, t_1, \cdots, t_n$ se define como un *spline* cúbico $S(t)$ tal que $S(t_0) = S(t_n)$, $S'(t_0) = S'(t_n)$ y $S''(t_0) = S''(t_n)$. Se usa principalmente para ajustar datos de los que se conoce que se comportan de forma periódica. Desarrollar el análisis necesario para obtener un *spline* cúbico periódico para los datos de una tabla

x	t_0	t_1	$\ldots$	t_n
y	y_0	y_1	$\cdots$	y_n

donde se supone que $y_n = y_0$.

6.3.12 Encontrar un *spline* cúbico sobre los nodos $-1, 0, 1$ tal que $S''(-1) = S''(1) = 0$, $S(-1) = S(1) = 0$, y $S(0) = 1$.

6.3.13 Determinar a, b y c para que la siguiente función sea un *spline* cúbico.

$$S(x) = \begin{cases} x^3, & 0 \le x \le 1 \\ \frac{1}{2}(x-1)^3 + a(x-1)^2 + b(x-1) + c, & 1 \le x \le 3 \end{cases}$$

6.3.14 ¿Existe una elección de los coeficientes para los que la siguiente función es un *spline* cúbico *natural*?

$$S(x) = \begin{cases} x + 1, & -2 \le x \le -1 \\ ax^3 + bx^2 + cx + d, & -1 \le x \le 1 \\ x - 1, & 1 \le x \le 2 \end{cases}$$

6.3.15 Determinar los coeficientes en la función

$$S(x) = \begin{cases} x^3 - 1, & -9 \le x \le 0 \\ ax^3 + bx^2 + cx + d, & 0 \le x \le 5 \end{cases}$$

de manera que sea un *spline* cúbico tomando el valor 2 cuando $x = 1$.

6.3.16 Determinar los coeficientes para que la función

$$S(x) = \begin{cases} x^2 + x^3, & 0 \le x \le 1 \\ a + bx + cx^2 + dx^3, & 1 \le x \le 2 \end{cases}$$

sea un *spline* cúbico con la propiedad de que $S'''(x) = 12$.

6.3.17 *Soporte de los B-splines.* Pruebe que los $n + 3$ funciones B_i definidas en la sección 6.2, restringidas a $[x_{-3}, x_{n+3}]$ son elementos de $M_2^3\{x_{-3}, x_{-2}, ..., x_{n+3}\}$. Pruebe, sin efectuar ningún cálculo, que B_i es nula fuera del intervalo (x_{i-2}, x_{i+2}), $i = -1(1)n + 1$. Utilice esto para ver que los $n + 3$ B_i son linealmente independientes en $M_2^3\{x_{-3}, x_{-2}, \ldots, x_{n+3}\}$. ¿Hay algún elemento de $M_2^3\{x_{-3}, x_{-2}, \ldots, x_{n+3}\}$ que sea nulo fuera de (x_{i-2}, x_{i+1}), $i = -1(1)n + 2$? (Indicación si lo hubiese sería una combinación lineal de las funciones $(x - x_{i-2})_+^3$, $(x - x_{i-1})_+^3$, $(x - x_i)_+^3$, nula junto con sus dos primeras derivadas en x_{i+1}. Concluya que los B_i tienen por soporte $[x_{i-2}, x_{i+2}]$ (cuatro subintervalos contiguos de la partición) y que no hay splines con soporte más pequeño.

6.3.18 *Base de B-splines.* Pruebe que los B_i son una base del espacio de splines $M_2^3(\Delta)$. (Indicación: según hemos visto en el problema anterior los B_i son libres en $M_2^3\{x_{-3}, x_{-2}, \ldots, x_{n+3}\}$, muestre que una combinación lineal de los B_i que fuese nula en $[a, b]$ de hecho lo sería en $[x_{-3}, x_{n+3}]$, con lo cual también son libres en $M_2^3(\Delta)$.)

6.3.19 *Suma de los B-splines.* Pruebe que la suma de todos los B_i, evaluada en cada punto de $[a, b]$ es 1. Justamente para lograr esta propiedad se introduce el factor de escala (x_{i-2}, x_{i+2}) al definir B_i .

6.3.20 *B-splines en una partición uniforme.* Suponga que la partición Δ es uniforme de paso h, y que los puntos suplementarios x_{-3}, x_{-2}, x_{-1}, x_{n+1}, x_{n+2}, x_{n+3} también se eligen con espaciado h. Pruebe que los B_i se obtienen transladando uno cualquiera de ellos, es decir $B_i(x) = B_j(x + (j - i)h)$. Halle explícitamente los B_i.

Lección 7

Transformada rápida de Fourier

En esta lección vamos a presentar una de las herramientas fundamentales de la matemática aplicada moderna: la transformada de Fourier discreta. Esta transformación se implementa mediante un algoritmo llamado transformada rápida de Fourier (aunque tal vez fuese más apropiado decir transformada de Fourier rápida) y se aplica en campos tan dispares como, por ejemplo, el tratamiento de señales en telecomunicación, la resolución de ecuaciones en derivadas parciales o la multiplicación de números largos en computación. De hecho, entre sus numerosas aplicaciones se incluyen métodos numéricos de interpolación y aproximación, y esta es la razón de que se incorpore en este momento del texto.

La idea subyacente a la transformación que vamos a definir es que ciertas operaciones importantes entre vectores $\mathbf{x}$, tienen una formulación mucho más simple cuando se expresan en términos de sus transformados. Un caso análogo se plantea con los logaritmos de los números positivos: multiplicar x e y corresponde, en lenguaje logarítmico a sumar $\log x$ y $\log y$, una operación mucho más sencilla en los tiempos anteriores a las calculadoras o mucho más económica de computar en la actualidad. El proceso sería pues

1. Se transforma $x \longrightarrow X = \log x$, $y \longrightarrow Y = \log y$.
2. Se opera con los transformados $Z = X + Y$.
3. Se deshace la transformación, pasando del número Z a otro z tal que $\log z = Z$.

Es esencial, por tanto, para realizar el último paso y cerrar el proceso, conocer la transforamción inversa, que en este caso es el *antilogaritmo*.

Aquí vamos a presentar la transformada de Fourier como una transformación lineal de un espacio N-dimensional.

7.1 La transformada de Fourier discreta

Vamos a trabajar en el espacio $\mathbb{C}^N$, cuyos elementos son n-uplas de números complejos $\mathbf{x} = (x_0, x_1, \ldots, x_{N-1})^T$, donde la secuencia de índices irá siempre de 0 a $N-1$ para trabajar con mayor sencillez en la notación. La T significa *transpuesta*, es decir, que cada vez que escribamos $\mathbf{x}$ nos referimos a un vector columna.

7.1.1 Definición. La transformada de Fourier discreta (N-dimensional) es la aplicación lineal de $\mathbb{C}^N$ en sí mismo dada por $\mathbf{x} \longrightarrow \mathbf{X} = F_N\mathbf{x}$, donde F_N es una matriz compleja $N \times N$ cuyo elemento $(F_N)_{jk}$ $(j,k = 0,1,\ldots,N-1)$ es ω_N^{jk}, la potencia jk-ésima del número complejo

$$\omega_N = e^{-\frac{2\pi}{N}i} = \cos\frac{2\pi}{N} - i\,\mathrm{sen}\,\frac{2\pi}{N}$$

Veamos algunos ejemplos. Para $N = 2$, $\omega_2 = e^{-\pi i} = -1$ y

$$F_2 = \begin{pmatrix} \omega_2^0 & \omega_2^0 \\ \omega_2^0 & \omega_2^1 \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix};$$

para $N = 3$, $\omega_3 = e^{-\frac{2\pi}{3}i} = \cos\frac{2\pi}{3} - i\,\mathrm{sen}\,\frac{2\pi}{3} = -\frac{1}{2} - i\frac{\sqrt{3}}{2}$ y

$$F_3 = \begin{pmatrix} \omega_3^0 & \omega_3^0 & \omega_3^0 \\ \omega_3^0 & \omega_3^1 & \omega_3^2 \\ \omega_3^0 & \omega_3^2 & \omega_3^4 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & e^{-\frac{2\pi}{3}i} & e^{-\frac{4\pi}{3}i} \\ 1 & e^{-\frac{4\pi}{3}i} & e^{-\frac{8\pi}{3}i} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & -\frac{1}{2} - i\frac{\sqrt{3}}{2} & -\frac{1}{2} + i\frac{\sqrt{3}}{2} \\ 1 & -\frac{1}{2} + i\frac{\sqrt{3}}{2} & -\frac{1}{2} - i\frac{\sqrt{3}}{2} \end{pmatrix}$$

y para $N = 8$, donde $\omega_8 = e^{-\frac{\pi}{4}i} = \cos\frac{\pi}{4} - i\,\mathrm{sen}\,\frac{\pi}{4} = \frac{\sqrt{2}}{2} - i\frac{\sqrt{2}}{2}$, escribimos solamente la expresión de la matriz que muestra más claramente la estructura de la misma

$$F_8 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & e^{-\frac{\pi}{4}i} & -i & e^{-\frac{3\pi}{4}i} & -1 & e^{\frac{3\pi}{4}i} & i & e^{\frac{\pi}{4}i} \\ 1 & -i & -1 & i & 1 & -i & -1 & i \\ 1 & e^{-\frac{3\pi}{4}i} & i & e^{-\frac{\pi}{4}i} & -1 & e^{\frac{\pi}{4}i} & -i & e^{\frac{3\pi}{4}i} \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & e^{\frac{3\pi}{4}i} & -i & e^{\frac{\pi}{4}i} & -1 & e^{-\frac{\pi}{4}i} & i & e^{-\frac{3\pi}{4}i} \\ 1 & i & -1 & -i & 1 & i & -1 & -i \\ 1 & e^{\frac{\pi}{4}i} & i & e^{\frac{3\pi}{4}i} & -1 & e^{-\frac{3\pi}{4}i} & -i & e^{-\frac{\pi}{4}i} \end{pmatrix} \tag{7.1}$$

Es importante observar que en todos los casos $\omega_N^N = e^{-2\pi i} = 1$, porque esto nos permite calcular todos los elementos de la matriz sin que los exponentes crezcan por encima de $N-1$. Basta tener en cuenta que, en base a dicha propiedad $\omega_N^{jk} = \omega_N^{jk \bmod N}$, siendo la función mod la que nos da el resto al dividir por N.

Además, desde el punto de vista teórico, el hecho de que ω_N sea una raíz N-ésima de la unidad, concretamente la raíz principal (véase el problema 7.4.1) de argumento negativo más pequeño, tiene como consecuencia que el producto de ω_N por su conjugado es siempre la unidad, y lo mismo ocurrirá con todos los elementos de la correspondiente matriz (¿por qué?). Como, por otra parte, es evidente que las matrices F_N son simétricas, resulta realmente fácil establecer cuál es la transformación inversa de la de Fourier discreta, que como ya hemos comentado es imprescincible para recuperar en los vectores $\mathbf{x}$ los posibles resultados obtenidos en sus transformados $\mathbf{X}$.

7.1.2 TEOREMA

La matriz F_N es invertible y $F_N^{-1} = \frac{1}{N}\overline{F_N}$, donde $\overline{F_N}$ es la matriz que resulta al conjugar cada elemento de F_N. Por tanto,

$$N(F_N^{-1})_{jk} = \overline{\omega_N^{jk}} = \overline{\omega_N}^{jk}, \quad \text{siendo} \quad {\omega_N}^{jk} = e^{\frac{2\pi}{N}i} = \cos\frac{2\pi}{N} + i\,\mathrm{sen}\,\frac{2\pi}{N}$$

para $j, k = 0, 1, \ldots, N-1$.

Demostración. Basta ver que $F_N\overline{F_N} = NI$. El elemento (j,k) de dicha matriz es

$$\begin{aligned}(F_N\overline{F_N})_{jk} &= \sum_{l=0}^{N-1}(F_N)_{jl}\overline{(F_N)}_{lk} = \sum_{l=0}^{N-1}\omega_N^{jl}\overline{\omega_N}^{lk} \\ &= \sum_{l=0}^{N-1}\omega_N^{jl}\omega_N^{-lk} = \sum_{l=0}^{N-1}(\omega_N^{j-k})^l,\end{aligned}$$

donde hemos utilizado que $\overline{\omega_N} = \omega_N^{-1}$ (¿por qué?). Ahora, si $j = k$

$$(F_N\overline{F_N})_{jk} = \sum_{l=0}^{N-1} 1^l = N,$$

y si $j \neq k$ hay que sumar una progresión geométrica de razón $\omega_N^{j-k} \neq 1$, resultando

$$\begin{aligned}(F_N\overline{F_N})_{jk} &= \frac{(\omega_N^{j-k})^N - 1}{\omega_N^{j-k} - 1} \\ &= \frac{(\omega_N^N)^{j-k} - 1}{\omega_N^{j-k} - 1} = \frac{1^{j-k} - 1}{\omega_N^{j-k} - 1} = 0.\end{aligned}$$

Para el caso $N = 8$, la matriz de la transformada inversa de la discreta de Fourier es la siguiente

$$F_8^{-1} = \frac{1}{8}\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & e^{\frac{\pi}{4}i} & i & e^{\frac{3\pi}{4}i} & -1 & e^{-\frac{3\pi}{4}i} & -i & e^{-\frac{\pi}{4}i} \\ 1 & i & -1 & -i & 1 & i & -1 & -i \\ 1 & e^{\frac{3\pi}{4}i} & -i & e^{\frac{\pi}{4}i} & -1 & e^{-\frac{\pi}{4}i} & i & e^{-\frac{3\pi}{4}i} \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & e^{-\frac{3\pi}{4}i} & i & e^{-\frac{\pi}{4}i} & -1 & e^{\frac{\pi}{4}i} & -i & e^{\frac{3\pi}{4}i} \\ 1 & -i & -1 & i & 1 & -i & -1 & i \\ 1 & e^{-\frac{\pi}{4}i} & -i & e^{-\frac{3\pi}{4}i} & -1 & e^{\frac{3\pi}{4}i} & i & e^{\frac{\pi}{4}i} \end{pmatrix} \tag{7.2}$$

donde puede observarse que la conjugación equivale a una permutación de las columnas (o de las filas) de F_8 y que el producto escalar complejo de dos filas (o columnas) distintas es cero. La interpretación de estos fenómenos nos ayudará a completar una visión del significado de la transformada de Fourier, a medida que vayamos analizando sus diversas utilidades.

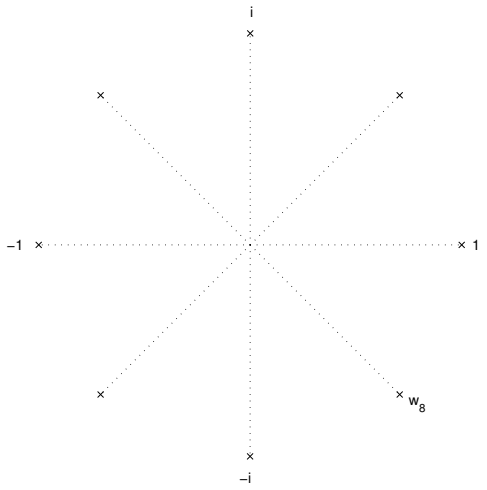


Figura 7.1: Raíces octavas de la unidad en el plano complejo

7.2 El algoritmo FFT

Si $\mathbf{x} \in \mathbb{C}^N$, el cálculo directo de $F_N(\mathbf{x})$ requiere obviamente N^2 multiplicaciones de números complejos (cada elemento de F_N se multiplica por uno de $\mathbf{x}$). El algoritmo de la transformada de Fourier rápida, conocido por las siglas FFT (Fast Fourier Transform) permite reducir drásticamente este costo operativo hasta $\frac{N}{2} \lg \frac{N}{2}$ (donde lg representa el logaritmo en base 2). El ahorro es ciertamente importante, pues para $N = 2^{10} = 1024$ se pasa de un millón de operaciones a menos de 5000. Es decir, un factor de más de 200. Y no se trata de una cantidad exagerada de puntos, pues en la práctica es frecuente encontrarse con transformadas mucho mayores aún.

El algoritmo se debe esencialmente a Runge y König, que en 1924 lo tenían desarrollado, pero hasta que no se vislumbra la utilización masiva de los ordenadores su aplicación práctica era limitada, como luego comprenderemos. Su popularidad llega cuando en 1965 es reformulado y generalizado por Cooley y Tuckey. En 1966 se publica una variante esencial por parte de Sande y Tuckey para la diezmación en frecuencias, y desde entonces infinitas variaciones sobre el mismo tema en función de la cantidad de puntos, de si los elementos son reales o complejos, de la arquitectura de las nuevas máquinas paralelas, etc.

La versión que vamos a exponer aquí es la más habitual y corresponde al caso en que $N = 2^m$, es decir, que la dimensión del espacio es una potencia de 2. Es además el caso más sencillo de comprender, para lo que puede ayudar la figura 7.1 que nos muestra las raíces octavas de la unidad en el plano complejo. Trate el lector de ubicar en la misma los distintos elementos de la matriz F_8 por filas, así como los complejos básicos ω_1, ω_2, ω_4, ω_8 y profundizar en la relaciones entre ellos y sus potencias esquematizadas en el siguiente cuadro, donde debe entenderse que los elementos en la misma columna son iguales.

$$\begin{array}{cccccccc} w_8^0 & w_8^1 & w_8^2 & w_8^3 & w_8^4 & w_8^5 & w_8^6 & w_8^7 \\ w_4^0 & & w_4^1 & & w_4^2 & & w_4^3 & \\ w_2^0 & & & & w_2^1 & & & \\ w_1^0 & & & & & & & \end{array}$$

7.2.1 Algoritmo de Cooley y Tuckey. La idea básica es aprovechar la gran cantidad de operaciones que se repiten en los cálculos y aplicar una técnica de diseño de algoritmos que vulgarmente se conoce como *divide y vencerás* y más técnicamente, en este caso, de diezmación en tiempos.

Se comienza por dividir el vector $\mathbf{x}$ de longitud $N = 2^m$ (y por consiguiente par) en dos vectores de tamaño mitad

$$\mathbf{x}^P = (x_0, x_2, \ldots, x_{N-2})^T, \quad \text{y} \quad \mathbf{x}^I = (x_1, x_3, \ldots, x_{N-1})^T$$

donde los superíndices P e I indican que el vector contiene las componentes pares e impares del original. A continuación se calculan las transformadas (¡de tamaño mitad!) de ambos vectores

$$\mathbf{X}^P = F_{\frac{N}{2}}\mathbf{x}^P, \qquad \mathbf{X}^I = F_{\frac{N}{2}}\mathbf{x}^I$$

y finalmente hemos de reconstruir el vector transformado $\mathbf{X}$ a partir de estos dos. El siguiente resultado nos demuestra de forma constructiva como hacerlo:

PROPOSICIÓN

En la situación anterior las $\frac{N}{2}$ primeras y las $\frac{N}{2}$ últimas componentes de $\mathbf{X}$ están dadas por

$$X_k = X_k^P + \omega_N^k X_k^I, \quad k = 0, 1, \ldots \frac{N}{2} - 1 \tag{7.3}$$

$$X_{\frac{N}{2}+k} = X_k^P - \omega_N^k X_k^I, \quad k = 0, 1, \ldots \frac{N}{2} - 1 \tag{7.4}$$

Demostración. Para $k = 0, 1, \ldots N-1$, a partir de la definición matricial que hemos dada, se deduce que

$$X_k = \sum_{j=0}^{N-1} \omega_N^{jk} x_j$$

y agrupando por separado en este sumatorio los términos con j par e impar

$$X_k = \sum_{j=0}^{\frac{N}{2}-1} \omega_N^{2jk} x_{2j} + \sum_{j=0}^{\frac{N}{2}-1} \omega_N^{(2j+1)k} x_{2j+1},$$

y como $\omega_N^2 = \omega_{\frac{N}{2}}$ (¿por qué?), resulta que

$$X_k = \sum_{j=0}^{\frac{N}{2}-1} \omega_{\frac{N}{2}}^{jk} x_j^P + \omega_N^k \sum_{j=0}^{\frac{N}{2}-1} \omega_{\frac{N}{2}}^{jk} x_j^I,$$

Esta igualdad prueba (7.3), pues las expresiones del segundo miembro son totalmente válidas para $k = 0, 1, \ldots, \frac{N}{2} - 1$, y evidentemente los sumatorios coinciden con las componentes k-ésimas de $\mathbf{X}^P$ y $\mathbf{X}^I$ respectivamente. Para los índices $\frac{N}{2}, \frac{N}{2}+1, \ldots, N-1$, que escribiremos en la forma $\frac{N}{2}+k$ con $k = 0, 1, \ldots, \frac{N}{2}-1$, siguiendo el mismo proceso, se obtiene

$$X_{\frac{N}{2}+k} = \sum_{j=0}^{\frac{N}{2}-1} \omega_{\frac{N}{2}}^{jk} x_j^P + \omega_N^{\frac{N}{2}} \omega_N^k \sum_{j=0}^{\frac{N}{2}-1} \omega_{\frac{N}{2}}^{jk} x_j^I,$$

y considerando que $\omega_{\frac{N}{2}}^{\frac{N}{2}} = 1$, y que $\omega_N^{\frac{N}{2}} = -1$ (¿por qué?), el resultado (7.4) se deduce de forma inmediata. □

Si lo pensamos detenidamente, la esencia del algoritmo termina aquí. En efecto, acabamos de ver como reemplazar la tarea de calcular una transformada N-dimensional (cuando N es par) por la de dos de dimensión $\frac{N}{2}$. Como $\frac{N}{2}$ también es par (en el supuesto en que nos estamos moviendo de que $N = 2^m$), podemos calcular cada una de ellas a partir de dos transformadas $\frac{N}{4}$-dimensionales, y de forma recurrente reduciremos el cálculo de F_N al de calcular un buen número de transformadas de dos elementos (¿cuántas exactamente?), que son triviales y no necesitan hacer ninguna multiplicación.

7.2.2 Costo operativo. Si escribimos como $M(m)$ el número de multiplicaciones de números complejos necesarias para realizar la transformada de Fourier discreta de $N = 2^m$ elementos mediante el algoritmo FFT. Tomamos como condición inicial del cálculo recursivo que $M(1) = 0$ (pues sabemos que si $\mathbf{x} = (x_0, x_1)^T$, $F_2(\mathbf{x}) = (x_0 + x_1, x_0 - x_1)^T$).

Entonces, es evidente a partir de las relaciones (7.3) y (7.4) que

$$M(m) = 2M(m-1) + 2^{m-1} \tag{7.5}$$

teniendo en cuenta que el producto que aparece en el segundo término de ambas ecuaciones es el mismo, y no es preciso por tanto repetirmo para el cálculo de la segunda mitad de los términos. Una sola aplicación de este proceso de división y recomposición ya reduce el costo operativo desde las N^2 multiplicaciones, que se realizan con el producto crudo de matriz por vector, hasta $2\left(\frac{N}{2}\right)^2 + \frac{N}{2} = \frac{N^2}{2} + \frac{N}{2}$, es decir poco más de la mitad cuando N es grande. Pero es la iteración del proceso lo que permite la reducción espectacular de que hablamos al principio. En efecto, en base a la recurrencia (7.5), tenemos que

$$\begin{aligned} M(m) &= 2[2M(m-2) + 2^{m-2}] + 2^{m-1} = 2^2 M(m-2) + 2 \cdot 2^{m-1} \\ &= 2^i M(m-i) + i2^{m-1} \qquad i = 3, \ldots, m-2 \\ &= 2^{m-1} M(1) + (m-1)2^{m-1} = (m-1)2^{m-1} \end{aligned}$$

lo que en términos de N resulta ser $\frac{N}{2}(\lg N - 1) = \frac{N}{2} \lg \frac{N}{2}$ como queriamos demostrar.

7.2.3 Implementación práctica. La implementación de algoritmos en forma recursiva tiene graves inconvenientes incluso en los lenguajes más favorablemente enfocados a este tipo de procesos. El esquema de la figura (7.2) pretende ser autoexplicativo respecto a la forma de implementar el algoritmo de Cooley y Tuckey para el cálculo de la FFT de una sucesión con un número finito (una potencia de 2) de complejos, pero de una forma iterativa, es decir comenzando por los ejemplares más pequeños, los de dimensión

dos, que como hemos visto son los primeros a calcular para después ir reconstruyendo los ejemplares mayores mediante (7.3) y (7.4). Así pues, lo primero que hacemos es reordenar los elementos a transformar de manera que queden juntos por parejas, aquellos que van a ser objeto de una transformación bidimensional. La segunda línea del esquema muestra la tarea de agrupar los pares y los impares por separado, y la tercera lo mismo pero dentro de cada uno de estos grupos, con lo que nos quedan cuatro parejas que podiamos denominar pares-pares, pares-impares, impares-pares e impares-impares.

En este momento hay que empezar a operar realizando las cuatro transformadas indicadas por las *mariposas* de tamaño 1, que son perfectamente realizables en paralelo (es decir, de forma simultánea si tenemos un hardware o un software apropiado). El elemento que figura a la izquierda, ω_2^0 (que es uno en este caso concreto) nos indica el factor que hemos de aplicar a la punta superior derecha de la *mariposa* antes de sumarle y restarle de la punta superior izquierda, para obtener las puntas inferiores izquerda y derecha respectivamente. De esta forma obtenemos los valores $\mathbf{X}^{PP,PI,IP,II}$, que podemos guardar en las mismas posiciones de memoria que ocupaban $\mathbf{x}^{PP,PI,IP,II}$, pues estos ya nunca más serán necesarios.

La siguiente línea de *mariposas*, nos presenta dos parejas entrecruzadas y dos factores

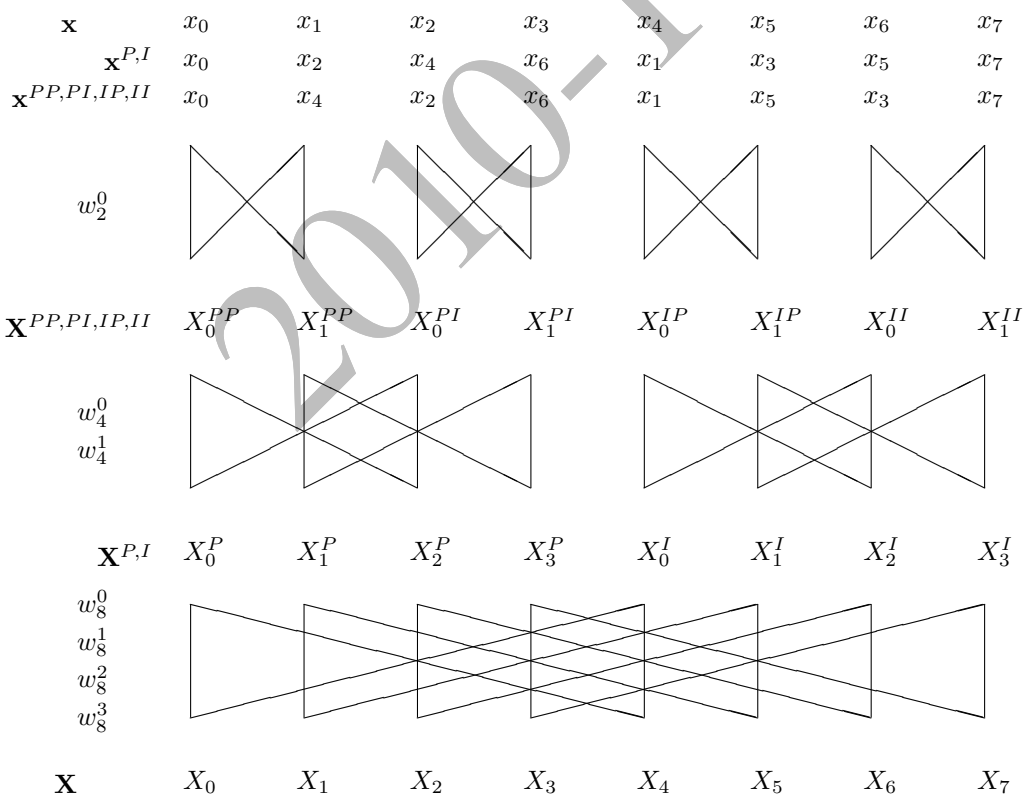


Figura 7.2: Esquema del algoritmo de Cooley y Tuckey

a la izquierda. El proceso es el mismo, pero el primero de los factores se utiliza con la primera *mariposa* de cada par y el segundo con la segunda. Al final tendremos en las posiciones de memoria originales la transformada de dimensión 4 del vector $\mathbf{x}^P$ ocupando las cuatro primeras posiciones, y la del vector $\mathbf{x}^I$ ocupando las cuatro últimas.

La última fila de *mariposas* ya no necesita ninguna explicación ni debe dejar dudas respecto a que los lugares de memoria (o elementos de una matriz) inicialmente ocupados por los elementos de $\mathbf{x}$, lo están ahora por los de su transformada de Fourier discreta.

Hay que tener en cuenta dos detalles fundamentales de este algoritmo que hemos descrito:

1. La matriz F_N no aparece explícitamente en ningún momento, aunque analizando conjuntamente el esquema (7.2) y la matriz, se observa como por ejemplo x_0 y x_4 solamente aparecen sumados o restados (para F_8) y nunca de forma independiente, etc.

2. El costo operativo en forma de multiplicaciones de complejos, se puede calcular ahora teniendo en cuenta que el número de filas de *mariposas* es $\lg N - 1$ (teniendo en cuenta que la primera no exige multiplicar), y en cada una de ellas hay que multiplicar los $\frac{N}{2}$ elementos que están en las esquinas superiores derechas por elementos que pueden estar precalculados o tabulados. El total es pues idéntico al estimado para la transformada de Fourier rápida.

7.2.4 Bit-inversión. La parte más compleja en la implementación del anterior algoritmo es la permutación que se necesita hacer con los elementos del vector original antes de comenzar el proceso de cálculo. Habitualmente se la denomina *bit-inversión*, y con un ejemplo, veremos que es fácil entender el por qué de esta denominación y el camino para implementarla (que en determinados lemguajes es especialmente sencillo).

Consideremos los índices de las componentes de nuestro vector $\mathbf{x}$ escritos en base 2, pero completando todos los *bits* aunque sean ceros iniciales

Índice	Cadena de *bits*	Selección 1		Selección 2		Cadena invertida
0	000	0	000	0	000	000
1	001	2	010	4	100	001
2	010	4	100	2	010	010
3	011	6	110	6	110	011
4	100	1	001	1	001	100
5	101	3	011	5	101	101
6	110	5	101	3	011	110
7	111	7	111	7	111	111

El proceso de selección de los elementos, comienza tomando los pares, es decir ¡los que terminan por cero!, que pasaran a la cabeza, quedando al final el subgrupo impar de los elementos cuyo índice termina en 1 cuando se le escribe en notación binaria. En el siguiente paso se hace lo mismo dentro de cada uno de estos subgrupos, luego en el grupo de los pares, pasarán a los primeros lugares aquellos que terminando por 0, tengan

también otro cero en al anterior posición, mientras que en el de los impares quedarán al final los que no sólo terminan por 1, sino que su anterior dígito binario es otro uno.

Si tenemos en cuenta que esto es exactamente lo contrario de lo que ocurre en el orden habitual (ceros en cabeza al principio y unos al final a nivel de cada dígito), hemos de concluir que el orden en que es preciso colocar las componentes del vector original, antes de proceder a su transformación por parejas para iniciar el camino ascendente del algoritmo, es el que le correspondería ordenando de forma natural las cadenas de sus *bits* escritas en orden inverso. A la vista de la tabla anterior, es evidente que el orden en que han de tomarse los elementos $(0, 4, 2, 6, 1, 5, 3, 7)$ es el orden previsto.

Es interesante observar que esta transformación de índices es recíproca y que muchos de ellos permanecen invariantes. Esto deriva en que resulta posible efectuar esta permutación sin necesidad de memoria adiccional y con una cantidad relativamente pequeña de intercambios.

7.2.5 Algoritmo de Sande-Tuckey. Para tratar de eludir este mecanismo de reordenar 'a priori' los elementos, el algoritmo de Sande y Tuckey, que ya hemos mencionado, se basa en un mecanismo en cierta forma dual. El principio de *diezmación en frecuencias* se

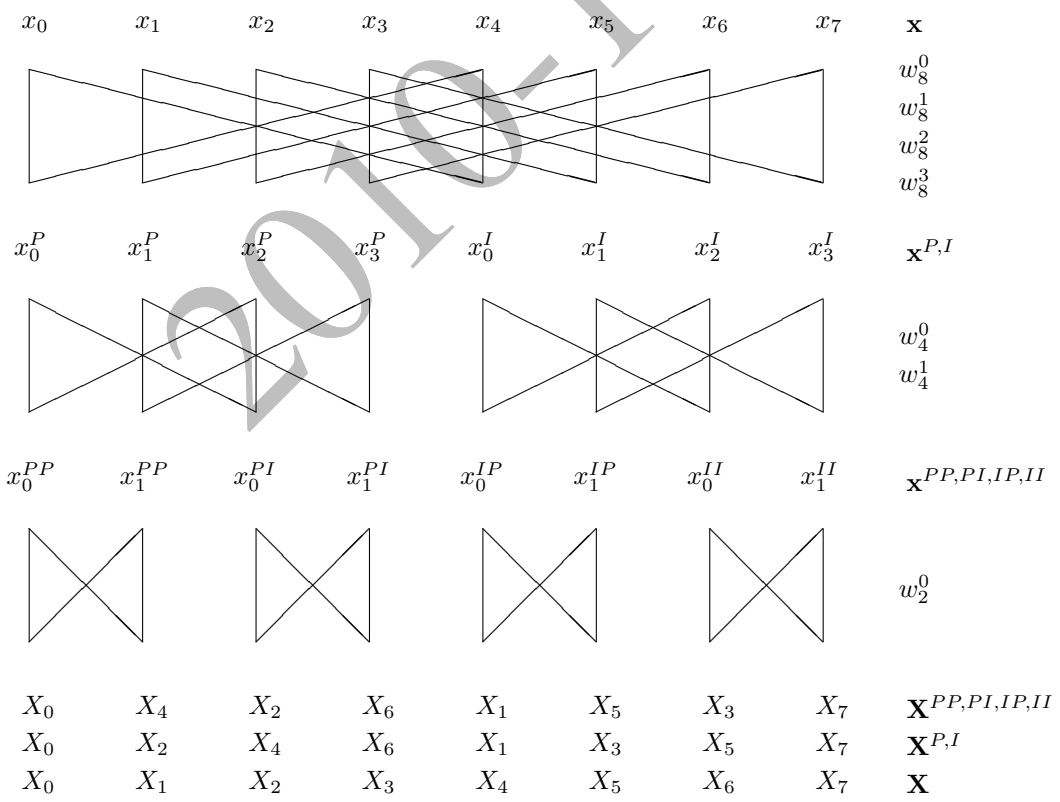


Figura 7.3: Esquema del algoritmo de Sande y Tuckey

traduce en el cálculo de los elementos pares e impares de la transformada por separado. Resulta fácil seguir la siguiente cadena de cálculos después de lo ya visto:

$$X_{2k} = \sum_{j=0}^{N-1} \omega_N^{2kj} x_j, \qquad \text{para } k = 0, 1, \ldots, \frac{N}{2}$$

y agrupando por separado en este sumatorio la primera y segunda mitad de los términos

$$X_{2k} = \sum_{j=0}^{\frac{N}{2}-1} \omega_N^{2kj} x_j + \sum_{j=0}^{\frac{N}{2}-1} \omega_N^{2k(j+\frac{N}{2})} x_{j+\frac{N}{2}},$$

y como ya sabemos que $\omega_N^2 = \omega_{\frac{N}{2}}$, resulta que

$$X_{2k} = \sum_{j=0}^{\frac{N}{2}-1} (x_j + x_{j+\frac{N}{2}}) \omega_{\frac{N}{2}}^{kj},$$

Esta igualdad prueba los elementos pares de la transformada final son la transformada $\frac{N}{2}$-dimensional de una sucesión construída a partir de la original de una forma sumamente sencilla: sumando a los $N-1$ primeros los que difieren en índice $\frac{N}{2}$. Por otra parte es igual de fácil de ver que los elementos impares son también la transformada de una sucesión de $\frac{N}{2}$ puntos obtenidos como muestra la siguiente expresión (basta aplicar que $\omega_N^{\frac{N}{2}} = -1$). Para $k = 0, 1, \ldots, \frac{N}{2}$:

$$X_{2k+1} = \sum_{j=0}^{\frac{N}{2}-1} \left((x_j - x_{j+\frac{N}{2}}) \omega_N^j \right) \omega_{\frac{N}{2}}^{kj},$$

De nuevo, podemos actuar de forma recurrente para calcular estas transformadas y obtener los elementos pares de esta sucesión de coeficientes pares con una transformada de $\frac{N}{4}$ elementos adecuadamente calculados. La figura (7.3) es autoexplicativo en el mismo sentido que el anterior esquema, y teniendo en cuenta que si ahora ponemos los factores que debemos aplicar a cada *mariposa* a la derecha, es para significar que el producto por dicho número complejo se realiza después de restar los elementos implicados y no antes, como ocurría en el algoritmo de Cooley y Tuckey. Sin embargo, el algoritmo también funciona sin necesitar espacio de almacenamiento adicional y con el mismo costo operativo, por lo que es una verdadera variante de FFT.

7.3 Aplicación a la interpolación trigonométrica

La relación de la transformada de Fourier discreta (y por consiguiente del algoritmo FFT) con la interpolación trigonométrica aparecera de forma natural cuando en la lección 12 estudiemos la aproximación de funciones por series trigonométricas. Pero en este punto, podemos aprovechar para establecer una relación con la interpolación polinomial y mostrar una utilidad práctica de la transformada de Fourier. Ya conocemos perfectamente que podemos determinar un polinomio de grado n de distintas maneras: dando los n ceros y uno de los coeficientes, dando todos los $(n+1)$ coeficientes, pero también nos

basta conocer su valor en $n+1$ puntos cualesquiera e interpolar por cualquiera de las estrategias estudiadas.

Si observamos la expresión de la transformada discreta de Fourier de una sucesión finita $\mathbf{c} = (c_0, c_1, \ldots, c_{N-1})^T$ de números complejos, veremos que los elementos de la misma

$$C_k = \sum_{j=0}^{N-1} c_j \omega_N^{kj}, \qquad k = 0, 1, \ldots, N-1$$

se pueden considerar evaluaciones del polinomio

$$P(z) = c_0 + c_1 z + c_2 z^2 + \cdots + c_{N-1} z^{N-1}$$

en los puntos $z_k = \omega_N^k$, $k = 0, 1, \ldots, N-1$, que como sabemos son las N raíces N-ésimas de la unidad (complejas). La transformada inversa, es entonces una interpolación en toda regla, pues nos permite calcular los coeficientes de este polinomio dando su valor en las mencionados números complejos. No es frecuente, sin embargo, que nos resulte de interés ni de utilidad interpolar una función real en unos cuantos números complejos, que además siempre tienen que ser los mismos. (Sería bueno que se practicase con algunos ejemplos sencillos, y ver lo que pasa).

Una perspectiva más interesante se presenta si consideramos las evaluaciones de la función de variable real

$$f(x) = c_0 + c_1 e^{ix} + c_2 e^{2ix} + \cdots + c_{N-1} e^{(N-1)ix}$$

en los puntos $x_k = \frac{2\pi}{N} k$, $k = 0, 1, \ldots, N-1$, que forman un conjunto de N puntos equidistantes en el intervalo $[0, 2\pi)$. En realidad, si $\mathbf{x} = \{x_0, x_1, \ldots, x_{N-1}\}$ vemos que $f(\mathbf{x}) = \overline{F_N}\mathbf{c} = N F_N^{-1}\mathbf{c}$. En estas circunstancias, si que puede ser de interés para la interpolación de una función real cuyos valores se conocen en dichos puntos. El proceso de calcular la transformada de Fourier discreta (dividida por N) de los valores de la función nos proporcionaría los coeficientes, ¡pero no de un polinomio!, sino de una base de funciones que toman valores complejos. Los propios coeficientes $c_k = c_k^R + i c_k^I, k = 0, 1, \ldots, N-1$ son complejos.

Si separamos las partes real e imaginaria, tendremos (para x real)

$$\begin{aligned} f^R(x) = c_0^R \quad &+ \quad c_1^R \cos x + c_2^R \cos 2x + \cdots + c_{N-1}^R \cos(N-1)x \\ &- \quad (c_1^I \operatorname{sen} x + c_2^I \operatorname{sen} 2x + \cdots + c_{N-1}^I \operatorname{sen}(N-1)x) \end{aligned}$$

y

$$\begin{aligned} f^I(x) = c_0^I \quad &+ \quad c_1^I \cos x + c_2^I \cos 2x + \cdots + c_{N-1}^I \cos(N-1)x \\ &+ \quad (c_1^R \operatorname{sen} x + c_2^R \operatorname{sen} 2x + \cdots + c_{N-1}^R \operatorname{sen}(N-1)x) \end{aligned}$$

es decir, hemos interpolado tanto la parte real como la imaginaria, por los razonablemente denominados *polinomios trigonométricos*. La expresión de los segundos miembros puede hacerse mucho más sencilla, como se desprende de los problemas de la lección. Hemos de esperar por otra parte que la componente imaginaria se anule en los nodos de interpolación, en el caso habitual de que f sea real.

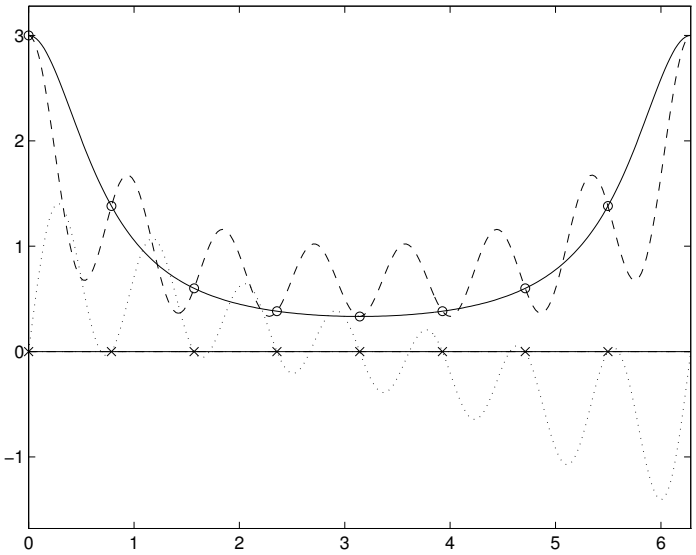


Figura 7.4: Interpolación trigonométrica de una función real

7.3.1 Ejemplo. Consideremos la función $f(x) = \frac{3}{(5-4\cos x)}$ en el intervalo $[0, 2\pi]$, y el conjunto de 8 abscisas equidistantes $\mathbf{x} = \{\frac{2\pi}{8}k | k = 0, 1, \ldots, 7\}$. Es habitual escribir como $\mathbf{X}(f)$ el vector (columna) con los valores de la función f en la red de puntos $\mathbf{x}$. En nuestro caso $\mathbf{X}(f) = (3, 1.3815, .6, .3832, .3333, .3832, .6, 1.3815)^T$. El vector de coeficientes de la interpolación trigonométrica será

$$\mathbf{c} = \frac{1}{8} F_8 \mathbf{X}(f)$$

que resulta ser $\mathbf{c} = (1.0078, .5098, .2667, .1569, .1255, .1569, .2667, .5098)^T$ (¿encuentra el lector alguna explicación para que todos resulten reales?). En la figura 7.4 hemos dibujado las partes real (línea de trazos) e imaginaria (línea de puntos). Puede observarse que las cosas efectivamente ocurren como se esperaba.

7.4 Cuestiones y problemas

7.4.1 Se dice que w es una raíz principal n-ésima de la unidad si cumple las tres condiciones siguientes:

a) $w \neq 1$ (salvo si $n = 1$).

b) $w^n = 1$.

c) $\sum_{j=0}^{n-1} w^{pj} = 0$ para todo $1 \leq p < n$.

Cuando w es una raíz principal n-ésima de la unidad, demostrar que:

1) La sucesión $\{w^j, j = 0, 1, \ldots, n-1\}$ contiene todas las raices n-ésimas de la unidad.

2) w^{-1} es también una raíz principal n-ésima de la unidad.

3) Si n es par $w^{n/2} = -1$, y en consecuencia $w^{j+n/2} = -w^j$.

4) Si n es par w^2 es una raíz principal $\frac{n}{2}$-ésima de la unidad.

7.4.2 Dado el polinomio $P(x) = p_{n-1}x^{n-1} + \ldots + p_1x + p_0$ con n par y $q = n/2 - 1$, una variación de la regla de Horner establece que

$$\begin{array}{rcl} P(x) & = & (\ldots(p_{2q}x^2 + p_{2q-2})x^2 + \ldots)x^2 + p_0 \\ & + & ((\ldots(p_{2q-1}x^2 + p_{2q-3})x^2 + \ldots)x^2 + p_1)x \end{array}$$

Demostrar que esta fórmula permite calcular económicamente el valor P en un punto y su opuesto.

Derivar de esta expresión el algoritmo de Cooley-Tuckey.

7.4.3 Como es sabido, el valor numérico de un polinomio $P(x)$ en un punto a coincide con el resto de dividir P entre $(x - a)$. Por otra parte es evidente que se pueden agrupar las raices de la unidad de forma que sean siempre raices de polinomios de la forma $x^t - c$.

Demostrar que si $P(x) = p_{2t-1}x^{2t-1} + \ldots + p_1x + p_0$, el resto de $P(x)/(x^t - c)$ es la suma de $(p_j + cp_{t+j})x^j$ para $j = 0, \ldots, t-1$.

Derivar de estas consideraciones el algoritmo de Sande-Tuckey.

7.4.4 Dado un polinomio $Q(x)$, diseñar un algoritmo que compute los coeficientes de el polinomio $Q(x + c)$ para una constante c.

7.4.5 Supongamos que el polinomio $Q(x)$ tiene coeficientes reales, pero lo queremos evaluar en un número complejo $z = u + iv$, con u y v reales. Desarrollar un algoritmo para esto.

7.4.6 Dar una interpretación de la transformada inversa de Fourier en términos polinomiales.

7.4.7 La transformada de Fourier se puede generalizar a k dimensiones. Por ejemplo, la transformada bidimensional toma la matriz $a(0 : n-1, 0 : n-1)$ y genera la matriz transformada

$$A(i,j) = \sum_{k=0}^{n-1}\sum_{l=0}^{n-1} a_{(k,l)}w^{(ik+jl)/n}$$

La transformación inversa es

$$a(i,j) = \frac{1}{n^2}\sum_{k=0}^{n-1}\sum_{l=0}^{n-1} A_{(i,j)}w^{-(ik+jl)/n}$$

Generalizar el algoritmo FFT a este caso.

7.4.8 Demostrar que la transformada inversa de Fourier se puede calcular con los mismos algoritmos que nos sirven para la transformada directa más unas pequeñas manipulaciones en los datos y los resultados de la misma.

CAPÍTULO III

APROXIMACIÓN

2010-11

Lección 8

Introducción a la aproximación

8.1 Conceptos generales sobre aproximación

Los procesos de *interpolación* estudiados anteriormente (véase el problema 5.5.7) enseñan cómo asociar a una función f, perteneciente a un espacio vectorial real X, un interpolante p. Este interpolante se busca en un subespacio S de X, previamente elegido, de modo que los valores de $n+1$ formas lineales en f y p coincidan ($n+1$ es la dimensión de S). (¿Quiénes son X, S y las formas lineales en el problema polinómico de Lagrange? ¿Y en el de Taylor? ¿Y en el caso de spline cúbico completo?). Naturalmente, el objeto de construir p radica en la idea de manejarlo en vez de f, digamos a efectos de evaluarlo en un punto, hallar su integral en un cierto intervalo, etc... Es típico de los problemas de interpolación que se disponga de tantas condiciones como parámetros libres hemos de calcular. Pero no siempre es posible plantear la sustitución de una función por otra más sencilla en estos términos.

En la teoría de *aproximación*, que ahora comenzamos, asociaremos a $f \in X$, el elemento $p \in S$ que haga la diferencia $f - p$ "lo menor posible". Naturalmente para dar sentido a la expresión "lo menor posible" es menester que el espacio X esté dotado de una norma. El elemento p así encontrado, cuando existe, se denomina mejor aproximación a f (por elementos de S). Son pues dos los elementos esenciales de esta teoría: la *FORMA* del aproximante, que decidimos al elegir el conjunto S de los mismos; y la *NORMA* que determina cuál es el mejor de los posibles e incluso su propia existencia y/o unicidad.

Veamos seguidamente algunos ejemplos típicos de situaciones que conducen a problemas de aproximación.

8.1.1 Aproximación funcional. Se desea reemplazar, a efectos de evaluarla, la función $\operatorname{sen} x$, $0 \leq x \leq \pi/2$, por un polinomio p de grado ≤ 10, de suerte que $||\operatorname{sen} x - p(x)||_\infty = \sup_{0\leq x\leq \pi/2} |\operatorname{sen} x - p(x)|$ sea lo más pequeño posible. Nótese que una formulación alternativa es encontrar $(a_0, a_1, \ldots, a_{10})^T \in \mathbb{R}^{11}$ para que la función $F(a_0, a_1, \ldots, a_{10})$ sea mínima, con

$$F(a_0, a_1, \ldots, a_{10}) = \sup\{|\operatorname{sen} x - a_0 - a_1 x - \ldots - a_{10}x^{10}| : 0 \leq x \leq \frac{\pi}{2}\}$$

Es un problema en que habitualmente la *NORMA* está preestablecida (la norma del *supremo*) y lo que se decide en cada caso es la *FORMA* del posible aproximante, bien

sea polinomios, funciones trigonométricas, exponenciales, *splines*, etc., pero siempre con la pretensión de que la aproximación sea uniforme.

Pero este problema de mínimos presenta una gran dificultad, y es que no se puede resolver por las técnicas usuales para el cálculo de extremos del cálculo infinitesimal, basadas en igualar a 0 las derivadas. ¿Por qué? De hecho, ni siquiera existen métodos directos de cálculo, que nos proporcionen una solución *exacta*. Entonces, es frecuente renunciar a este tipo de aproximación, en beneficio de otras normas que nos permitan calcular el elemento *óptimo* de forma más eficiente (aunque naturalmente es distinto). Se trata de busca un equilibrio entre la precisión deseada y el costo requerido.

¿Qué función real definida en $\mathbb{R}^{11}$ hay que minimizar si trabajamos con la norma L^2 :

$$\|f\|_2 = \left(\int_0^{\pi/2} f(x)^2 dx \right)^{1/2} \quad ?$$

¿Se puede hallar con esta nueva norma la solución anulando derivadas?

8.1.2 Sistemas sobredeterminados o incompatibles. Sea A una matriz real con m filas y n columnas. Tratemos de resolver el sistema $A\mathbf{x} = \mathbf{b}$ donde $\mathbf{b}$ es un vector conocido con m componentes y $\mathbf{x}$, desconocido, tiene n componentes. Sea S un subespacio de $X = \mathbb{R}^m$ generado por los n vectores columna de A. Supongamos que S es un subespacio propio, lo cual ocurre:

a) Siempre que $m > n$, es decir haya más ecuaciones (condiciones impuestas) que incógnitas (parámetros libres).

b) Si $m \leq n$ pero entre las n columnas de A no hay m independientes, es decir haya un número de incógnitas mayor o igual que el de ecuaciones, pero la matriz A no tenga rango máximo.

Si $\mathbf{b} \in \mathbb{R}^m$ no está en S, el sistema carece de solución. Dado que, cuando $\mathbf{x}$ recorre $\mathbb{R}^n$, $A\mathbf{x}$ nunca coincide con $\mathbf{b}$, podemos preguntarnos para qué $\mathbf{x}$ es $||\mathbf{b} - A\mathbf{x}||$ lo menor posible ($||.||$ denota una norma en $\mathbb{R}^m$ previamente elegida). Tales $\mathbf{x}$ juegan el papel de "soluciones" del sistema en un sentido generalizado.

Observemos que su determinación comprende dos etapas:

1. Hallar la mejor aproximación $\mathbf{b}^*$ a $\mathbf{b}$ por elementos de la imagen S. O las mejores, porque en función de la norma utilizada, la solución puede no ser única.

2. Resolver el sistema (o los sistemas) $A\mathbf{x} = \mathbf{b}^*$ en sentido convencional.

En este caso es la *FORMA* del aproximante lo que está totalmente definido (pues es la imagen de la aplicación lineal A), y la *NORMA* lo que hemos de decidir en función de la comodidad y eficiencia de los métodos a emplear, salvo que nos venga impuesta por su significado en el problema.

8.2 Ajuste

Es sin duda el caso más interesante de aproximación desde el punto de vista de los métodos numéricos, y por eso le vamos a dedicar una mayor atención en esta lección

introductoria. En la práctica, constituye una aproximación funcional discreta, y casi siempre desemboca en la resolución de un sistema sobredeterminado. Nos apoyamos en el siguiente supuesto práctico:

La evolución de la masa m de una muestra radiactiva que se está desintegrando sigue la ley

$$m(t) = \left(\frac{1}{2}\right)^{\frac{t}{r}} m(0) \left(= m(0)e^{-\frac{t}{r}\ln 2}\right), \tag{8.1}$$

donde t es el tiempo y r una constante, que depende del tipo de átomo que se está desintegrando y se llama periodo de semidesintegración o vida media (¿por qué?). Para una muestra de Radio 224 se han efectuado las siguiente medidas

t(días)	0	1	3
m(gramos)	1.00	0.84	0.57

Se desea conocer r.

8.2.1 Método erróneo de solución. Evidentemente $m(0) = 1$, luego poniendo $t = 1$ en (8.1)

$$0.84 = e^{(-\ln 2/r)},$$
$$\ln 0.84 = -\ln 2/r,$$
$$r = -\ln 2/\ln 0.84 = 3.97;$$

y nuestra *respuesta* sería $r = 3.97$ días.

Si ahora pusiéramos $t = 3$

$$0.57 = e^{(-3\ln 2/r)}, \qquad r = 3.70 \text{ días.}$$

Tenemos dos resultados discrepantes ¿Cuál preferir? ¿Cuál es la razón de la discrepancia?

Quizás (8.1) no sea una ley exacta sino sólo aproximada. Quizá haya habido errores al pesar la muestra o al medir el intervalo entre pesadas. Podríamos preferir la solución 3.97 a la solución 3.70 si supiésemos que los números $t = 0$, $t = 1$, $m = 1.00$, $m = 0.84$ no están afectados de errores de medida y que la ley es exacta de $t = 0$ a $t = 1$. Pero si fuesen las medidas en $t = 0$ y $t = 3$ las exactas habría que preferir 3.70. Naturalmente no hay motivo para preferir unas a otras medidas y en realidad probablemente *todas* las medidas de la tabla serán erróneas (incluyendo el dato $m(0) = 1$ usado en ambas soluciones).

8.2.2 Método correcto: el ajuste. Encontremos m_0 y r tales que

$$\left\| \begin{pmatrix} 1.00 \\ 0.84 \\ 0.57 \end{pmatrix} - \begin{pmatrix} m_0 \\ m_0 e^{(-\ln 2/r)} \\ m_0 e^{(-3\ln 2/r)} \end{pmatrix} \right\|$$

sea lo menor posible. Este es un problema de mejores aproximaciones ¿Quién es X, S, f? Se dice que se ha ajustado m_0 y r en el modelo (8.1).

8.2.3 Observación: En la formulación anterior S no es un subespacio. Para *linealizar* el problema, observamos que, de (8.1)

$$\ln m(t) = \ln m(0) - \frac{t}{r}\ln 2 \tag{8.2}$$

Considerando como nuevos parámetros $R = 1/r$, $M = \ln m_0$ llegamos a la formulación: hacer mínimo

$$\left\| \begin{pmatrix} \ln 1.00 \\ \ln 0.84 \\ \ln 0.57 \end{pmatrix} - \begin{pmatrix} M \\ M - R\ln 2 \\ M - 3R\ln 2 \end{pmatrix} \right\|$$

Así estamos resolviendo el sistema lineal incompatible

$$\begin{pmatrix} 0 & 1 \\ -\ln 2 & 1 \\ -3\ln 2 & 1 \end{pmatrix} \begin{pmatrix} R \\ M \end{pmatrix} = \begin{pmatrix} \ln 1.00 \\ \ln 0.84 \\ \ln 0.57 \end{pmatrix}$$

Resuelto éste, $r = 1/R$.

8.2.4 Relación con la interpolación. Dejemos la formulación (8.2) y volvamos a la (8.1). Disponemos de la familia biparamétrica F de funciones de t de la forma $F(t) = m_0 e^{-\frac{t}{r}\ln 2}$ y hemos estado tratando de encontrar el elemento φ de esta familia tal que $\varphi(0) = 1.00$, $\varphi(1) = 0.84$, $\varphi(2) = 0.57$. Como F es biparamétrica y hay tres condiciones, no cabe esperar exista $\varphi \in F$ que las satisfaga (y de hecho hemos visto que para los valores que la tabla maneja φ no existe). Por tanto *no podemos interpolar*. Convienen entonces, más que satisfacer dos de las tres condiciones, tratar de violar *las tres* lo menos posible: esto es el ajuste. La curva solución $(t, \varphi(t))$ no pasará en general por los puntos (t_j, m_j) que representan los datos, pero se acercará a ellos lo más posible.

En la práctica el número de datos suele ser mucho mayor que el de parámetros a determinar. Los datos forman una *nube de puntos* en el plano (t, m) y la función ajustada se ciñe a dicha nube. Ahora es fácil entender que algunas veces se conozca la interpolación como *aproximación exacta*.

8.2.5 Generalizaciones. En nuestro modelo $m = \varphi(t, m(0), r)$ hay una función real m, una variable independiente real y dos parámetros $m(0)$, r. Más generalmente podremos considerar $\mathbf{y} = \boldsymbol{\varphi}(\mathbf{x}, \boldsymbol{\lambda})$, $\mathbf{y} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{R}^n$, $\boldsymbol{\lambda} \in \mathbb{R}^{\boldsymbol{p}}$ y una tabla de valores $(\mathbf{x}_i, \mathbf{y}_i)$, $i = 1, \ldots, N$. Así hay $n \times m$ condiciones a satisfacer con los p parámetros libres.

Notemos por último que, si bien en el ejemplo conocíamos 'a priori' la expresión funcional $m = \varphi(t)$, tal cosa no es necesarias para ajustar. Así, verbigracia, podremos describir una tabla (x_i, y_i), $i = 1, \ldots, 100$, $x_i, y_i \in \mathbb{R}$ ajustándola por una recta $y = Ax + B$ si los puntos (x_i, y_i) en el plano están *grosso modo* alineados, aunque no haya una ley que exprese que y es función lineal de x. Es decir, en el problema genérico de ajuste ni la *FORMA* ni la *NORMA* están predeterminados, siendo muchas veces la intuición o la repetida experimentación lo que nos permite dar con una solución convincente.

8.3 Aproximación óptima: existencia y unicidad

Establezcamos de forma precisa el concepto de aproximación (si no es óptima no es una verdadera aproximación):

8.3.1 Definición. Dados un espacio normado real X, un subconjunto (no necesariamente subespacio) S y un elemento f de X decimos que $p^* \in S$ es una mejor aproximación a f (por elementos de S) si, para cada $p \in S$, $||f - p^*|| \leq ||f - p||$, es decir

$$||f - p^*|| = \inf\{||f - p|| : p \in S\}.$$

8.3.2 Ejemplo. Si X es el espacio de funciones reales continuas definidas en [-1,1] con la norma del supremo $||f||_\infty = \sup\{|f(x)| : -1 \leq x \leq 1\}$, S conjunto de polinomios mónicos de grado n y f es la función idéntica nula, la mejor aproximación es el polinomio de Chebyshev escalado $T_n/2^{n-1}$ y es única. (Corolario 4.2.3)

8.3.3 Ejemplo. Sean $X = \mathbb{R}^2$ con la norma usual $||(x_1, x_2)^T|| = \sqrt{(x_1^2 + x_2^2)}$, $S = \{(x, 1)^T : x \in \mathbb{R}\}$, $f = (0, 0)^T$. El conjunto $\{\mathbf{x} \in \mathbb{R}^2 | \ ||\mathbf{x} - f|| = r\}, r > 0$ es la circunferencia de radio r centrada en el origen. Cuando $r < 1$ no interseca a S. Para $r = 1$ la intersección es el punto $(0, 1)^T$. Por consiguiente $(0, 1)^T$ es la mejor aproximación; no hay elementos de S que disten de f una cantidad $r < 1$.

8.3.4 Ejemplo. Como en el ejemplo 8.3.3 salvo que ahora usaremos la norma $||(x_1, x_2)^T|| = \max\{|x_1|, |x_2|\}$. El conjunto $\{\mathbf{x} \in \mathbb{R}^2 | \ ||\mathbf{x} - f|| = r\}$ es un cuadrado centrado en el origen y de lado $2r$. Para $r < 1$ no interseca a S. Para $r = 1$ la intersección es el segmento $\{(x, 1)^T : -1 \leq x \leq 1\}$. Cada punto del segmento es *una* mejor aproximación a f. Note como p^* depende de $||.||$.

8.3.5 Ejemplo. $X = \mathbb{R}^2$ con su norma usual $S = \{(x, 0) : x \in \mathbb{R}\}$, $f = (0, -1)^T$. Aquí S y f resultan de los del ejemplo 8.3.3, tras efectuar una translación del vector $(0, -1)^T$. La solución ahora es $(0, 0)^T$, trasladada de la solución del ejemplo 8.3.3 ¿por qué?

La teoría de la aproximación se ocupa de la existencia, unicidad y construcción de las mejores aproximaciones. Se tienen los siguientes resultados básicos:

8.3.6 Proposición

Si X es un espacio normado, S un subconjunto no vacío y compacto de X y f un elemento de X, entonces f tiene, al menos, una mejor aproximación por elementos de S.

Demostración. La función real $||f - p||$ definida en S es continua y acotada inferiormente por 0. Por consiguiente alcanza su mínimo. □

8.3.7 Proposición

Si X es un espacio normado, S un subespacio vectorial *de dimensión finita* y f un elemento de X, entonces f tiene, al menos, una mejor aproximación por elementos de S.

Demostración. La aproximación óptima, si existe, debe pertenecer al conjunto $S^* = \{p \in S : ||f - 0|| \geq ||f - p||\}$, ya que si p está en S pero no en S^* no es el elemento buscado pues aproxima a f peor que $0 \in S$. Así, basta probar que existe una mejor aproximación a f por elementos de S^*. Ahora bien, S^* no vacío, cerrado y acotado en un finito-dimensional, y se puede aplicar la proposición precedente. □

Veremos mas adelante que la hipótesis sobre la dimensión de S no puede suprimirse.

8.4 Convergencia de las mejores aproximaciones. Teorema de Weierstrass

Hasta ahora hemos considerado problemas de aproximación en los que intervenían un espacio normado X, un elemento $f \in X$ y un subconjunto S de X. Un caso importante es aquel en que S es un subespacio vectorial, y muy frecuentemente S es uno de los términos de una sucesión de subespacios $S_0, S_1, S_2, \ldots$ expansiva (es decir donde cada subespacio contiene a los anteriores). Así, por ejemplo, S es a menudo el espacio Π_n de polinomios de grado $\leq n$, un miembro de la cadena $\Pi_0, \Pi_1, \Pi_2, \ldots$.

Supongamos pues dados X, una cadena expansiva de subespacios $S_0, S_1, S_2, \ldots$ y un elemento $f \in X$, de forma que existan las aproximaciones óptimas p_i a f por elementos de S_i, $i = 0, 1, 2, \ldots$ Obviamente $||f - p_0|| \geq ||f - p_1|| \geq \ldots \geq 0$. Nos vamos a plantear bajo que circunstancias $\lim_i ||f - p_i|| = 0$, es decir, $f = \lim_i p_i$ en X.

Consideremos ante todo el caso en que $X = C[a, b]$, funciones continuas en el intervalo acotado $[a, b]$ con la norma del supremo, y $S_n = \Pi_n$. Si $f \in C[a, b]$ su polinomio $p_n \in \Pi_n$ de mejor aproximación existe, y vamos a demostrar que, para la norma del supremo en $C[a, b]$

$$\lim p_n = f, \quad n \to \infty, \tag{8.3}$$

es decir, f es límite uniforme de sus polinomios de mejor aproximación.

El resultado (8.3) es consecuencia del siguiente teorema, de importancia notable en todo el Análisis Matemático (véase el problema 8.5.4).

8.4.1 TEOREMA DE WEIERSTRASS (1885).

Si f es una función real continua en un intervalo compacto $[a, b]$, dado $\varepsilon > 0$ existe un polinomio P tal que $|f(x) - P(x)| \leq \varepsilon$ para cada x en $[a, b]$.

Demostración. El teorema posee muchas demostraciones. La que damos aquí se debe a Bernstein (1812) y es constructiva (aunque no práctica, vea problema 8.5.5).

Se supone, sin perder generalidad, que $[a, b] = [0,1]$, pues cualquier intervalo $[a, b]$ puede transformarse en el [0,1] mediante un cambio lineal de variable.

Para $n = 1, 2, \ldots$ se define el n-ésimo polinomio de Bernstein B_n (relativo a f) mediante la fórmula

$$B_n(x) = \sum_{k=0}^{n} \binom{n}{k} x^k (1 - x)^{n-k} f\left(\frac{k}{n}\right)$$

Demostraremos que los B_n convergen a f, uniformemente en [0,1].

Probemos ante todo las relaciones

$$\sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} = 1 \tag{8.4}$$

$$\sum_{k=o}^{n} \binom{n}{k} x^k (1-x)^{n-k} \frac{k}{n} = x \tag{8.5}$$

$$\sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-x} \left(\frac{k}{n}\right)^2 = \left(1 - \frac{1}{n}\right) x^2 + \frac{1}{n} x \tag{8.6}$$

que calculan los polinomios de Bernstein de las funciones 1, x, x^2. Escribiendo

$$(p+q)^n = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} \tag{8.7}$$

y tomando $p = x$, $q = 1 - x$ resulta (8.4). Derivando (8.7) respecto a p y poniendo en el resultado $p = x$, $q = 1 - x$, obtenemos (8.5). Dos derivaciones respecto a p en (8.7) conducen a (8.6).

La identidad

$$\sum_{k=0}^{n} \left(\frac{k}{n} - x\right)^2 \binom{n}{k} x^k (1-x)^{n-k} = \frac{x(1-x)}{n} \tag{8.8}$$

es consecuencia de (8.4) - (8.6), como se ve al desarrollar el cuadrado $(k/n - x)^2$.

Tras estos preliminares, multiplicamos (8.6) por $f(x)$ y restamos $B_n(x)$ al resultado para tener

$$f(x) - B_n(x) = \sum_{k=0}^{n} [f(x) - f(k/n)] \binom{n}{k} x^k (1-x)^{n-k} \tag{8.9}$$

Al ser f continua en un compacto es uniformemente continua y acotada, y existen $\delta > 0$, $M > 0$ tales que $|f(x) - f(y)| < \varepsilon/2$ si $|x - y| < \delta$, $x, y \in [0,1]$ y además $|f(x)| < M$ para cada $x \in [0,1]$. Fijados x en [0,1] y n, dividamos los índices k en (8.9) en dos subconjuntos A y B como sigue: k está en A si $|k/n - x| < \delta$ y k está en B en caso contrario. (Nótese que A, B dependen de x, n y δ ; a su vez δ depende de ε y f .)

Ahora

$$\left| \sum_{A} [f(x) - f(k/n)] \binom{n}{k} x^k (1-x)^{n-k} \right| \leq \frac{\varepsilon}{2} \sum_{A} \binom{n}{k} x^k (1-x)^{n-k} \leq \frac{\varepsilon}{2} \tag{8.10}$$

y por otro lado, por las elecciones de M, B y por (8.8):

$$\begin{aligned} \left| \sum_{B} [f(x) - f(k/n)] \binom{n}{k} x^k (1-x)^{n-k} \right| &\leq 2M \sum_{B} \binom{n}{k} x^k (1-x)^{n-k} \\ = 2M \sum_{B} \frac{(k/n - x)^2}{(k/n - x)^2} \binom{n}{k} x^k (1-x)^{n-k} &\leq \frac{2M}{\delta^2} \frac{x(1-x)}{n} \leq \frac{M}{2n\delta^2} \end{aligned} \tag{8.11}$$

(hemos usado que $x(1-x) \leq \frac{1}{4}$ para $0 \leq x \leq 1$). Tomemos n suficientemente avanzado para que $\frac{M}{(2n\delta^2)} \leq \frac{\varepsilon}{2}$. Entonces, llevando (8.10) y (8.11) a (8.9) es $|f(x) - B_n(x)| \leq \varepsilon$ y el teorema está probado. □

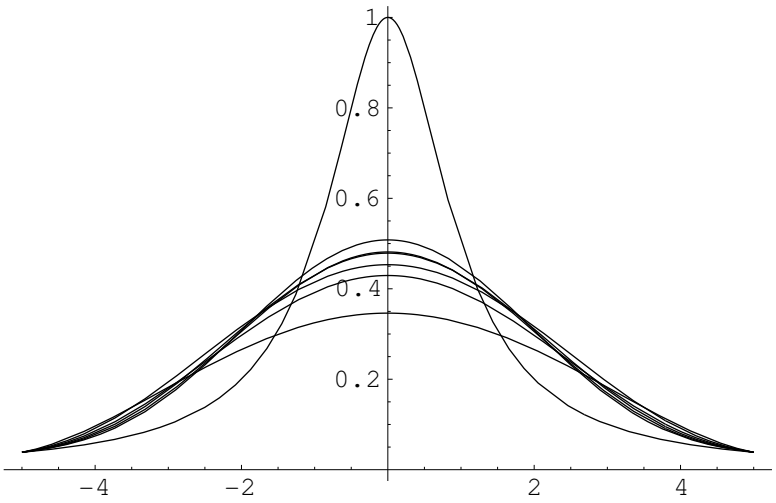


Figura 8.1: Aproximantes de grados 5 a 10 para la función de Runge

8.5 Cuestiones y problemas

8.5.1 Sea X el espacio de funciones f reales definidas en $-\infty < a \leq x \leq b < +\infty$ tales que

$$\int_a^b |f|^2 \, dx < \infty$$

En X se usa la norma

$$\|f\| = \left(\int_a^b |f|^2 \, dx \right)^{1/2}$$

Demuestre que, si $f \in X$, su mejor aproximación por una constante es su valor medio

$$\int_a^b f(x)dx/(b-a).$$

8.5.2 Sea ahora X el espacio de funciones reales acotadas definidas en $-\infty \leq a \leq x \leq b \leq +\infty$ con la norma $||f|| = \sup\{|f(x)| : a \leq x \leq b\}$ ¿Cuál es la mejor aproximación a f por una constante?

8.5.3 Si $y = f(x)$ es una función real definida, estrictamente creciente y continua en [0,1] ¿Cuál es la constante que mejor la aproxima en la norma L_1 (integral del módulo)? (Indicación: la distancia de f a una constante dada corresponde al área de cierta región del plano (x, y); exprese tal área como una integral definida en que la variable de integración sea y.)

8.5.4 Use el teorema de Weierstrass para demostrar la fórmula (8.3) de la lección.

8.5.5 ¿Qué debe valer n para que, cuando $f(x) = x^2$, $||B_n f - f||_\infty < 10^{-8}$? ¿Cree que los polinomios de Bernstein proporcionan un medio *práctico* de obtener polinomios de aproximación? Aquí y más abajo $B_n f$ denota el polinomio de Bernstein de grado $\leq n$ asociado a f.

8.5.6 Probar que $||B_n f||_\infty \leq ||f||_\infty$.

8.5.7 Si f es un polinomio de grado $\leq k$, pruebe que $B_n f$ lo es para cada $n = 0, 1, 2, \ldots$

8.5.8 Pruebe que B_n es un operador monótono, es decir que $B_n f \leq B_n g$ si $f \leq g$.

8.5.9 Sea f convexa en [0,1]. Pruebe que para $n = 2, 3, \ldots$, $0 < x < 1$ se tiene $B_{n-1} f(x) \geq B_n f(x)$. Si f es continua la desigualdad es estricta, a menos que f sea lineal en cada subintervalo

$$\left[\frac{k-1}{n-1}, \frac{k}{n-1}\right], \qquad k = 1, \ldots, n-1$$

en cuyo caso $B_{n-1} f = B_n f$.

8.5.10 Obtenga explícitamente $B_n f$, si $f(x) = x^3$. Pruebe que $\lim_n n(B_n f - f) = 3x^2(1-x)$.

8.5.11 Si f y su derivada son continuas en $[a, b]$, $a, b \in \mathbb{R}$, entonces para cada $\varepsilon > 0$, existe un polinomio tal que $||f - p||_\infty \leq \varepsilon$, $||f' - p'|| \leq \varepsilon$.

2010-11

Lección 9

Problemas de mínimos cuadrados

9.1 Aproximación en un espacio con producto interno

El grupo más importante de problemas de aproximación se refiere a situaciones donde el espacio ambiente X es un espacio con producto interno. Tales problemas se llaman, a veces, de mínimos cuadrados, por razones que serán obvias más tarde.

Si X es un espacio vectorial real, un *producto interno* en X es una aplicación definida en $X \times X$ y con valores reales $(f, g) \to \langle f, g\rangle$, tal que sea

(i) *Bilineal:* Para cada f_1, f_2, g en X y cada α_1, α_2 reales

$$\langle \alpha_1 f_1 + \alpha_2 f_2, g\rangle = \alpha_1\langle f_1, g\rangle + \alpha_2\langle f_2, g\rangle,$$
$$\langle g, \alpha_1 f_1 + \alpha_2 f_2\rangle = \alpha_1\langle g, f_1\rangle + \alpha_2\langle g, f_2\rangle.$$

(ii) *Simétrica:* $\langle f, g\rangle = \langle g, f\rangle$ para cada $f, g \in X$.

(iii) *Definida positiva:* para cada $f \neq 0$, $\langle f, f\rangle > 0$.

9.1.1 Ejemplo. En $\mathbb{R}^n$

$$\langle \mathbf{x}, \mathbf{y}\rangle = \mathbf{x}^T\mathbf{y} = \sum_{i=1}^{n} x_i y_i \tag{9.1}$$

define el producto interno usado habitualmente. Más generalmente, a cada matriz real $n \times n$ simétrica y definida positiva A le corresponde un producto interno dado por

$$\mathbf{x}^T A\mathbf{y} = \sum_{i,j=1}^{n} x_i a_{ij} y_i$$

Todos los productos internos en $\mathbb{R}^n$ son de esta forma (¿por qué?). Concretamente si A es una matriz diagonal de elementos positivos (normalmente denominados $\omega_i, i = 1, \dots n$) estamos ante un producto escalar ponderado,

$$\langle \mathbf{x}, \mathbf{y}\rangle_\omega = \sum_{i=1}^{n} \omega_i x_i y_i$$

donde cada componente tiene un peso diferente en el resultado final.

9.1.2 Ejemplo. En el conjunto de las funciones reales continuas en $0 \leq x \leq 1$, la expresión

$$\int_0^1 f(x)g(x)dx$$

define un producto interno. Demostrar que esto es cierto, y tratar de definir otros productos internos en espacios de funciones. Por ejemplo, un equivalente de los productos ponderados discretos para el espacio de las funciones continuas en $-1 \leq x \leq 1$ son productos internos de la forma

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

donde $\omega(x)$ es una función estrictamente positiva en todo el intervalo (véase el apartado 11.1).

9.1.3 Un resultado clave en el estudio de los espacios con producto interno es la llamada desigualdad de Cauchy - Schwarz que afirma que, cualesquiera que sean f y g en X

$$|\langle f, g\rangle| \leq \sqrt{\langle f, f\rangle}\sqrt{\langle g, g\rangle}$$

(véase ejercicio 9.4.2). Como consecuencia se tiene que

$$\sqrt{\langle f+g, f+g\rangle} \leq \sqrt{\langle f.f\rangle} + \sqrt{\langle g.g\rangle},$$

desigualdad de Minkowski (véase el ejercicio 9.4.3). A su vez esta nueva desigualdad permite probar de modo inmediato que la aplicación $\sqrt{\langle f, f\rangle}$ define una norma en X (ejercicio 9.4.4). Por consiguiente, todo espacio con producto interno es, de modo natural, un espacio normado, y tiene sentido en él hablar de distancias. En particular, podremos plantear en él problemas de aproximación.

Por otro lado, en los espacios con producto interno hay una gama de conceptos que carecen de sentido en espacios normados generales. El concepto específico más destacado es el de *ortogonalidad*: dos vectores de X se dicen *ortogonales* si su producto interno es nulo. Claramente esta relación es simétrica. Un subconjunto S de X se dice *ortogonal* si sus vectores son dos a dos ortogonales. Un subconjunto *ortonormal* es un subconjunto *ortogonal* en el que todos los vectores tienen norma unidad.

El concepto de ortogonalidad generaliza a espacios como el del ejemplo 9.1.2 la noción elemental de perpendicularidad del espacio común (ejercicio 9.4.5), lo que permite utilizar algunos resultados útiles basados en dicho concepto, como por ejemplo la relación de Pitágoras (una versión analítica del famoso teorema), que nos asegura que si dos vectores $\mathbf{x}$ e $\mathbf{y}$ son ortogonales, se verifica

$$\|\mathbf{x}+\mathbf{y}\|^2 = \|\mathbf{x}\|^2 + \|\mathbf{y}\|^2$$

que sería bueno que el lector tratase de probar.

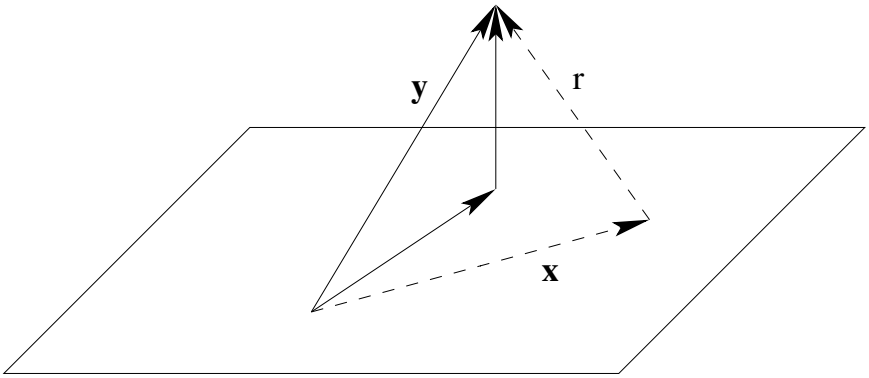


Figura 9.1: Ilustración geométrica de los mínimos cuadrados

9.1.4 En el caso de que la norma derive de un producto interno es fácil dar una *caracterización* de la mejor aproximación, es decir obtener un criterio que nos permita reconocer si un elemento dado es la mejor aproximación o no.

TEOREMA

Sea X un espacio con producto interno, S un subespacio de X, f un elemento de X. Si existe p^*, aproximación óptima a f por elementos de S, entonces tal mejor aproximación es única y satisface:

$$f - p* \text{ es ortogonal a cada } p \text{ de } S \tag{9.2}$$

Recíprocamente si un elemento p^* satisface la condición anterior, es la mejor aproximación.

Demostración. Sea p^* la mejor aproximación. Fijemos p en S. Para cada α real

$$\begin{aligned} ||f - p^*||^2 \leq ||f - (p^* + \alpha p)||^2 &= \langle (f - p^*) - \alpha p, (f - p^*) - \alpha p \rangle \\ &= ||f - p^*||^2 - 2\alpha \langle f - p^*, p \rangle + \alpha^2 ||p||^2. \end{aligned}$$

Así el 'trinomio' $\alpha \to \alpha^2||p||^2 - 2\alpha\langle f - p^*, p\rangle$ toma sólo valores no negativos. Su discriminante $4\langle f - p^*, p\rangle^2$ será ≤ 0. Como también es ≥ 0 será nulo y por ello $f - p^*$ ortogonal a p.

Recíprocamente, si vale la condición (9.2) y $p \in S$, $p \neq p^*$,

$$\begin{aligned} ||f - p||^2 &= ||(f - p^*) + (p^* - p)||^2 \\ &= ||f - p^*||^2 + 2\langle f - p^*, p^* - p \rangle + ||p^* - p||^2 \\ &= ||f - p^*||^2 + ||p^* - p||^2 > ||f - p^*||^2 \end{aligned}$$

luego p^* es una mejor aproximación y no puede haber otra. □

9.2 Aproximación en subespacios de dimensión finita

Si S es de *dimensión finita* la relación (9.2) da un medio práctico para calcular p^* (que existe, proposición 8.3.7). En efecto, tomemos una base $g_0, g_1, \ldots, g_n$ de S. Encontrar p^* es encontrar los escalares α_i, $i = 0, 1, \ldots, n$ tales que

$$p^* = \sum_{i=0}^{n} \alpha_i g_i$$

Para que se verifique (9.2) se necesita y basta que $f - p^*$ sea ortogonal a cada g_i, $i = 0, 1, ..., n$ (ejercicio 9.4.7). Por tanto los α_i quedan caracterizados por

$$\langle f - \sum_{i=0}^{n} \alpha_i g_i, g_j \rangle = 0, \qquad j = 0, 1, ..., n$$

ó

$$\sum_{i=0}^{n} \langle g_i, g_j \rangle \alpha_i = \langle f, g_j \rangle, \qquad j = 0, 1, ..., n \tag{9.3}$$

Las ecuaciones (9.3) se llaman *ecuaciones normales* del problema. La matriz del sistema $\{\langle g_i, g_j \rangle\}_{i,j=0,1,...,n}$ (que es la misma para todas las $f \in X$), se llama matriz de Gram de los vectores g_i. Puesto que (9.3) tiene solución única (¿por qué?) la matriz de Gram ha de resultar *regular*.

9.2.1 Ejemplo. Hallemos, en $0 \leq x \leq 1$ la mejor aproximación a la función exponencial por un polinomio de grado ≤ 2, respecto del producto interno usual (el del ejemplo 9.1.2). La mejor aproximación es $a + bx + cx^2$, se determina imponiendo que $e^x - (a + bx + cx^2)$ sea ortogonal a los elementos de la base 1, x, x^2, es decir imponiendo que

$$\int_0^1 \left[e^x - (a + bx + cx^2)\right] x^j dx = 0, \qquad j = 0, 1, 2.$$

Esto conduce al sistema de *ecuaciones normales*

$$\begin{array}{ccccccc} a & + & \frac{1}{2}b & + & \frac{1}{3}c & = & e - 1 \\ \frac{1}{2}a & + & \frac{1}{3}b & + & \frac{1}{4}c & = & 1 \\ \frac{1}{3}a & + & \frac{1}{4}b & + & \frac{1}{5}c & = & e - 2 \end{array}$$

con solución $a = 39e - 105$, $b = -216e + 588$, $c = 210e - 570$. Dando los coeficientes con cuatro cifras decimales tras la coma, la mejor aproximación es $1.0130 + 0.8511x + 0.8392x^2$. Por ejemplo en $x = 1$ vale 2.7033, mientras que $e = 2.7183$; en este punto el polinomio cuadrático de Taylor $1 + x + 0.5x^2$ vale 2.5000.

9.2.2 Aplicación a la aproximación en variedades afines. Supongamos que deseamos aproximar la exponencial (en el mismo intervalo con el mismo producto interno que en el ejemplo anterior por elementos del conjunto S^* de polinomios de grado ≤ 2, que coincidan con la exponencial en $x = 0$ y $x = 1$.

Vemos que S^* no es un subespacio vectorial; sin embargo, si p y q están en S^*, la diferencia $p - q$ está en el conjunto S de polinomios de grado a lo sumo 2 *nulos* en $x = 0$ y $x = 1$, y S ya es un subespacio vectorial. Por consiguiente S^* es una variedad afín; y, fijado un elemento p_0 de S^*, los elementos de S^* son justamente los de la forma $p_0 + q$ con q recorriendo el subespacio S.

La distancia de $p_0 + q$ a la exponencial es la misma que la de q a $e^x - p_0$ (¿por qué?). Concluimos que la solución del problema es $p_0 + q^*$ siendo q^* la aproximación mejor a $e^x - p_0$ por elementos de S.

Para p_0 podemos tomar cualquier elemento de S^*, pero lo más cómodo será tomar la *recta* de Newton $1 + (e - 1)x$. La dimensión de S es evidentemente 1, y una base obvia la constituye el polinomio $x(x - 1)$.

Si $q^* = Ax(x - 1)$ es la mejor aproximación a $e^x - p_0$ en S, A se halla imponiendo que $e^x - p_0 - Ax(x - 1)$ sea ortogonal a $x(x - 1)$. Esto da $A = (130e - 350)/4$, de manera que la aproximación buscada es

$$1 + (e - 1)x + \frac{130e - 350}{4}x(x - 1) = 1 + 0.8741x + 0.8442x^2$$

9.2.3 Aplicación a la aproximación funcional discreta. En el caso de que el problema a resolver sea de aproximación funcional discreta (es decir, aproximar una función real en un conjunto finito de puntos $x_k, k = 0, 1, \ldots, m$) el espacio X de todas las *funciones* es de dimensión finita (¿cuál será la dimensión exacta?) y en consecuencia lo mismo ocurre con todos los subespacios, y siempre se puede utilizar esta técnica constructiva de la solución.

En concreto, si el subespacio S tiene una base formada por los *vectores* $\mathbf{g}_j, j = 0, 1, \ldots, n$ con $n \leq m$, y $\mathbf{f}$ es el *vector* a aproximar (cualquier función en este contexto es simplemente un vector), el *error* a minimizar admite en esta situación la siguiente expresión

$$||\mathbf{f} - \sum_{i=0}^{n} c_i \mathbf{g}_i||^2 = \langle \mathbf{f} - \sum_{i=0}^{n} c_i \mathbf{g}_i, \mathbf{f} - \sum_{i=0}^{n} c_i \mathbf{g}_i \rangle = \sum_{k=0}^{m} \left[f(x_k) - \sum_{i=0}^{n} c_i g_i(x_k) \right]^2$$

según (9.1), y que justifica perfectamente el nombre de "mínimos cuadrados". De hecho utilizando los procedimientos de cálculo infinitesimal para la localización de extremos, al imponer la condición de que las derivadas parciales respecto de los $c_j, j = 0, 1, \ldots, n$ sean nulos, se obtienen las relaciones

$$2 \sum_{k=0}^{m} \left[f(x_k) - \sum_{i=0}^{n} c_i g_i(x_k) \right] g_j(x_k) = 0, \quad j = 0, 1, \ldots, n$$

es decir,

$$\sum_{k=0}^{m} f(x_k) g_j(x_k) = \sum_{i=0}^{n} c_i \left[\sum_{k=0}^{m} g_i(x_k) g_j(x_k) \right], \quad j = 0, 1, \ldots, n$$

que en términos del producto interno (9.1), resulta

$$\langle \mathbf{f}, \mathbf{g}_j \rangle = \sum_{i=0}^{n} c_i \langle \mathbf{g}_i, \mathbf{g}_j \rangle, \quad j = 0, 1, \ldots, n$$

que esencialmente son las mismas *ecuaciones normales* (9.3), pero donde al ser X de dimensión finita, todos los elementos implicados aparecen como vectores de $\mathbb{R}^{m+1}$.

Esta circunstancia permite que se pueda hacer un planteamiento matricial del problema discreto de "mínimos cuadrados". Si consideramos la matriz A, de dimensión $(m+1) \times (n+1)$ que tiene por columnas los vectores $\mathbf{g}_j, j = 0, 1, \ldots, n$ y $\mathbf{c} = (c_0, c_1, \ldots, c_n)^T$, las ecuaciones normales resultan ser

$$A^T A \mathbf{c} = A^T \mathbf{f}$$

teniendo en cuenta que el producto de la fila i-ésima de A^T por la columna j-ésima de A no es otra cosa que el producto escalar interno (9.1) de los vectores $\mathbf{g}_i$ y $\mathbf{g}_j$, que es precisamente el elemento ij-ésimo de la matriz de Gram.

Observaciones: Conviene notar que aunque la matriz A pueda ser rectangular (si m y n son diferentes), $A^T A$ es cuadrada y regular (¿por qué?) de dimensión $n+1$.

Si A fuese cuadrada (y regular), esto significa que el subespacio S tiene la misma dimensión que todo el espacio, por lo que evidentemente vamos a encontrar unos coeficientes $c_i, i = 0, 1, \ldots, m$, tales que $\mathbf{f} = \sum_{i=0}^{n} c_i \mathbf{g}_i$, en cuyo caso la suma de los cuadrados es cero porque en los puntos de la red coincide con el aproximante, que es de hecho un interpolante en el sentido que vimos en los capítulos anteriores. Por eso la interpolación se denomina a veces *aproximación exacta*, y aquí hemos visto un enfoque general para abordarla, con independencia de que tipo de función sean los interpolantes.

En este apartado hemos cambiado la notación de los coeficientes de α_i a c_i, porque aún cuando la función a aproximar y el subespacio sean descritos inicialmente de la misma manera, estos coeficientes varian en función de los puntos de discretización, pues la presunta identidad de los datos es sólo aparente debido al fenómeno de *'sampling'* o muestreo, que nos obliga a trabajar con los valores de las funciones en unos pocos puntos, sin que influya para nada su comportamiento en el resto del intervalo.

9.2.4 Ejemplo. Hallemos, en $0 \le x \le 1$ la mejor aproximación discreta, en la norma euclídea basada en los puntos $\{0, .25, .5, .75, 1\}$, a la función exponencial por un polinomio de grado ≤ 2. La mejor aproximación será de nuevo de la forma $a + bx + cx^2$, pero para el cálculo de los coeficientes, los *polinomios* 1, x, x^2 de la base son en realidad los vectores $(1,1,1,1,1)^T$, $(0, .25, .5, .75, 1)^T$ y $(0, .0625, .25, .5625, 1)^T$, mientras que la *función* a aproximar es el vector $(1, 1.2840, 1.6487, 2.1170, 2.7183)^T$ de los valores de la función exponencial en la red.

Utilizando la mecánica matricial que acabamos de ver, resulta que el sistema de ecuaciones normales resulta ser

$$\begin{pmatrix} 5 & 2.5 & 1.875 \\ 2.5 & 1.875 & 1.5625 \\ 1.875 & 1.5625 & 1.3828 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 8.7680 \\ 5.4514 \\ 4.4015 \end{pmatrix}$$

cuya solución, con cuatro cifras decimales tras la coma, nos proporciona ahora como mejor aproximación $1.0051 + 0.8643x + 0.8435x^2$, que en $x = 1$ vale 2.7130. Como se ve los resultados son diferentes al caso continuo.

9.2.5 Aplicación a los sistemas lineales sobredeterminados. Consideremos el sistema lineal $A\mathbf{x} = \mathbf{b}$ donde A tiene m filas y n columnas. De acuerdo con las consideraciones del apartado 9.2.1, tratamos de hallar $\mathbf{x}$ que minimice la cantidad $||\mathbf{b} - A\mathbf{x}||$ para alguna norma en $\mathbb{R}^m$. Se hablaba allí de hallar en una primera etapa un elemento $\mathbf{b}^*$ de la forma $A\mathbf{x}$, es decir combinación lineal de los vectores columna de A, que mejor aproxime a $\mathbf{b}$, y en una segunda resolver el sistema resultante.

Pero, cuando en $\mathbb{R}^m$ se toma la *norma euclídea usual* recaemos en un problema de mínimos cuadrados. Como trivialmente, los mencionados vectores columna $\mathbf{a}_i$ generan el espacio de sus combinaciones lineales, que es la imagen de la aplicación lineal A, la solución $\mathbf{x}$ del problema existe (¿por qué?), y viene caracterizada por:

$$\langle \mathbf{b} - A\mathbf{x}, \mathbf{a}_i \rangle = 0, \qquad i = 1, 2, \ldots, n; \tag{9.4}$$

o equivalentemente

$$\langle A\mathbf{x}, \mathbf{a}_i \rangle = \langle \mathbf{a}_i, \mathbf{b} \rangle, \qquad i = 1, 2, ..., n,$$

es decir

$$A^T A\mathbf{x} = A^T \mathbf{b}. \tag{9.5}$$

expresión que nos permite calcular $\mathbf{x}$ directamente, sin necesidad de tener previamente $\mathbf{b}^*$.

Observaciones: Notemos que aunque $A\mathbf{x}$ está únicamente definido (teorema 9.1.4) pudiera muy bien darse que para $\mathbf{x}_1 \neq \mathbf{x}_2$, $A\mathbf{x}_1 = A\mathbf{x}_2 = \mathbf{b}^*$. La solución $\mathbf{x}$ será única si y solo si $\mathbf{b}^*$ sólo puede escribirse de una manera como combinación lineal de las columnas $\mathbf{a}_i$, esto es *si las n columnas de A son linealmente independientes* (en $\mathbb{R}^m$), lo que sólo puede ocurrir cuando $m \geq n$, número de incógnitas no superior al de ecuaciones.

Como vemos la aproximación discreta de mínimos cuadrados, ocupa un lugar intermedio entre la aproximación funcional continua (de mínimos cuadrados) y la solución euclídea de los sistemas sobredeterminados. Por una parte, podemos escribir directamente la matriz de Gram de la base elegida (única opción posible en la aproximación continua), pero por otra se puede plantear la solución como una combinación lineal de vectores que optimize la aproximación, desembocando en un sistema sobredeterminado y sacar provecho de todas las técnicas matriciales. Veremos la utilidad de este mecanismo a la hora de discretizar una aproximación continua, pues tenemos la posibilidad de seleccionar los puntos del *'sampling'*. En otras ocasiones, cuando la función esta tabulada, la discreta es la única opción posible para aproximar.

9.2.6 Ejemplo. En el apartado 9.3.1, dejamos planteado el problema de resolver el sistema:

$$\begin{pmatrix} 0 & 1 \\ -\ln 2 & 1 \\ -3\ln 2 & 1 \end{pmatrix} \begin{pmatrix} R \\ M \end{pmatrix} = \begin{pmatrix} 0 \\ \ln 0.84 \\ \ln 0.57 \end{pmatrix}$$

La "solución" mejor (en norma euclídea), satisfará:

$$\begin{pmatrix} 0 & -\ln 2 & -3\ln 2 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ -\ln 2 & 1 \\ -3\ln 2 & 1 \end{pmatrix} \begin{pmatrix} R \\ M \end{pmatrix} = \begin{pmatrix} 0 & -\ln 2 & -3\ln 2 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 0 \\ \ln 0.84 \\ \ln 0.57 \end{pmatrix}$$

es decir

$$\begin{pmatrix} 10(\ln 2)^2 & -4\ln 2 \\ -4\ln 2 & 3 \end{pmatrix} \begin{pmatrix} R \\ M \end{pmatrix} = \begin{pmatrix} -\ln 2 \ln 0.84 - 3\ln 2 \ln 0.57 \\ \ln 0.84 + \ln 0.57 \end{pmatrix}$$

Multiplicando por tres la primera ecuación, por $4\ln 2$ la segunda y sumando, se tiene

$$R = \frac{\ln 0.84 - 5\ln 0.57}{14\ln 2}.$$

Así la estimación para $r = 1/R$ (= vida media del Ra 224) será $14\ln 2/(\ln 0.84 - 5\ln 0.57) = 3.68$ (días). (El verdadero valor es 3.64 días.)

9.3 Sistemas ortogonales

Un caso particularmente sencillo en (9.3) es el cuando los elementos g_i forman una base ortogonal: entonces las ecuaciones toman la forma

$$||g_j||^2 \alpha_j = \langle f, g_j \rangle, \qquad j = 0, 1, \ldots, n;$$

con lo que

$$p^* = \sum_{j=0}^{n} \frac{\langle f, g_j \rangle}{||g_j||^2} g_j. \tag{9.6}$$

Todavía, a efectos teóricos, se suelen normalizar los elementos de la base, multiplicándolos por el inverso de su norma para que tengan norma unidad. Si los g_i son una base ortonormal, (9.6) se simplifica, $p^* = \sum \langle f, g_i \rangle g_i$. En el caso particular en que X es de dimensión finita y S es todo el espacio, se tiene evidentemente que, para cada f de X, $p^* = f$, y por ello $f = \sum \langle f, g_i \rangle g_i$; expresión bien conocida de un vector en una base ortonormal. Con todo, el usar bases ortonormales en vez de simplemente ortogonales suele tener más bien interés teórico; para el cálculo efectivo no es ni necesario ni útil normalizar los elementos de las bases ortogonales.

Observaciones: Las ventajas de utilizar bases ortogonales son grandes. Como acabamos de ver, la primera de ellas es que no hemos de resolver ningún sistema lineal para tener la solución de las ecuaciones normales, ya que la matriz de Gram es diagonal.

Esta misma condición, hace que sea muy económico el cálculo de la propia matriz (de hecho no hace falta especificarla como tal), y si bien es cierto que para cada problema de aproximación sólo hemos de calcularla una vez, la reducción de n^2 productos escalares a n es muy notable, especialmente en el caso de la aproximación continua donde estos productos requieren operaciones analíticas y no sólo numéricas.

Pero la consecuencia más importante del uso de bases ortogonales en el subespacio aproximante es la permanencia de la aproximación ya calculada cuando se quiere mejorar la aproximación. Esto quiere decir que si la aproximación expresada en (9.6) no es suficientemente buena para nuestras necesidades, para mejorarla basta añadir el término siguiente correspondiente al siguiente elemento ortogonal de la base (que amplia el espacio de aproximantes) sin tener que modificar para nada los coeficientes ya calculados. Cuando la base no es ortogonal, esto es imposible: la mejora de la aproximación mediante la ampliación del conjunto de aproximantes, exige recalcular todos los coeficientes, el trabajo hecho anteriormente se pierde.

9.3.1 Ejemplo. Volviendo al ejemplo 9.2.1, si hubiesemos calculado las mejores aproximaciones de grados 0 y 1, los resultados hubiesen sido $a = e-1 = 1.7183$ para el primer caso, y el resultado de resolver el sistema 2×2

$$\begin{array}{rcrcl} a & + & \frac{1}{2}b & = & e-1 \\ \frac{1}{2}a & + & \frac{1}{3}b & = & 1 \end{array}$$

para el segundo (¿por qué?). Es decir $a = 0.8731$ y $b = 1.6903$. Como vemos el valor de a cambia en las dos ocasiones que cambiamos de orden de aproximación y el de b en la única que puede. Es evidente que el sistema a resolver es distinto cada vez.

En cambio, si para formar la base del espacio de polinomios de primer grado tomamos $1-2x$ (junto al constante 1), resulta que el sistema de ecuaciones normales, ahora quedará (omitiendo los ceros procedentes de la ortogonalidad)

$$\begin{array}{rcrcl} a & + & & = & e-1 \\ & + & \frac{1}{3}b & = & e-3 \end{array}$$

cuyo resultado ahora es $b = 3*(e-3) = -0.8452$ (con a el mismo que para la aproximación de grado 0). En resumen el polinomio de grado 1, mejor aproximante en el sentido de los mínimos cuadrados, se puede excribir como

$$1.7183 - 0.8452(1 - 2x)$$

que es el mismo obtenido anteriormente, pero donde gracias a la ortogonalidad de los elementos de la base, sólo hemos tenido que calcular el coeficiente b, en lugar de tener que recalcular ambos mediante un nuevo sistema de ecuaciones.

Si ahora conseguimos encontrar un polinomio de segundo grado ortogonal a los dos utilizados (ejercicio 9.4.5), bastará obtener un único coeficiente, y al añadir el término correspondiente al polinomio de grado 1 que ya tenemos. Volveremos a obtener la solución del ejemplo 9.2.1, pero de una forma progresiva. Este es el sentido y la ventaja de la *permanencia* cuando se usan bases ortogonales. La dificultad está en calcularlas.

9.3.2 Sistema trigonométrico. Aunque es posible construir de forma sistemática conjuntos ortogonales para cualquier problema de aproximación que se nos presente, resulta fundamental desde el punto de vista práctico disponer de familias ortogonales explícitas y sencillas de manejar.

Un ejemplo muy importante por su utilidad práctica lo constituye la familia de funciones

$$1, \cos x, \operatorname{sen} x, \cos 2x, \operatorname{sen} 2x, \ldots, \cos kx, \operatorname{sen} kx, \ldots$$

que son ortogonales en $[-\pi, \pi]$ con el producto escalar habitual

$$(f, g) = \int_{-\pi}^{\pi} f(x)g(x)dx$$

En efecto, es inmediato probar (se deja la demostración como un ejercicio) que

$$\begin{aligned}
\int_{-\pi}^{\pi} \cos jx \cos kx dx &= \begin{cases} 0 & j \neq k \\ \pi & j = k \neq 0 \\ 2\pi & j = k = 0 \end{cases} \\
\int_{-\pi}^{\pi} \operatorname{sen} jx \operatorname{sen} kx dx &= \begin{cases} 0 & j \neq k \\ \pi & j = k = 1, 2, \ldots \end{cases} \\
\int_{-\pi}^{\pi} \cos jx \operatorname{sen} kx dx &= 0 \qquad j, k = 0, 1, 2, \ldots
\end{aligned}$$

donde además se observa el valor de las normas (al cuadrado) de las funciones. Todas valen todas π, excepto para la constante que vale 2π.

En virtud de (9.6), los coeficientes del polinomio trigonométrico de grado n

$$S_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n} (a_k \cos kx + b_k \operatorname{sen} kx)$$

que representa la mejor aproximación en mínimos cuadrados dentro del subespacio de dimensión $2n + 1$ generado por las funciones implicadas, serán

$$a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos kx dx, \qquad k \geq 0 \tag{9.7}$$

$$b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \operatorname{sen} kx dx, \qquad k \geq 1 \tag{9.8}$$

teniendo en cuenta que el coeficiente de la función constante 1, tiene que ser $\frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)dx$ que es claramente el *promedio* de la función en el intervalo (¿por qué?). Para mejorar la aproximación basta con añadir más términos al polinomio sin modificar los anteriores.

Observaciones: Los coeficientes a_k y b_k pueden calcularse para cualquier función, y por consiguiente se puede construir una suma infinita denominada serie de Fourier correspondiente o asociada a la función

$$f(x) \sim \frac{a_0}{2} + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \operatorname{sen} kx) \tag{9.9}$$

La forma en que esta serie converge a la función es un interesante problema de análisis en el que no entraremos de momento. Es evidente que en (9.9) no es posible establecer la igualdad en todo caso, pues el segundo miembro es periódico aunque no lo sea el primero. Sí debemos saber que la serie de Fourier converge en norma cuadrática a la función (sea ésta periódica o no, pero perteneciente a $L^2[-\pi, \pi]$). Además, que toda función periódica en $[-\pi, \pi]$ y continua se puede aproximar uniformemente por polinomios trigonométricos, y que si la función tiene derivada continua, la serie de Fourier converge uniformemente.

9.4 Cuestiones y problemas

9.4.1 Describa geométricamente el subconjunto de $\mathbb{R}^n$, $\{\mathbf{x} : ||\mathbf{x}||_A = 1\}$, siendo $||.||_A$ la norma inducida por una matriz real, simétrica, definida positiva A (cf. ejemplo 9.1.1).

9.4.2 Pruebe la desigualdad de Cauchy-Schwarz. (Indicación: fije f, g; la función real de variable real $F(\alpha) = \langle f + \alpha g, f + \alpha g\rangle$ sólo toma valores no negativos; muestre que F es un trinomio de segundo grado en α; su discriminante no puede ser positivo.)

9.4.3 Pruebe la desigualdad de Minkowski.

9.4.4 Pruebe que si $\langle .,. \rangle$ es un producto interno en X, entonces $\sqrt{\langle f, f\rangle}$ define una norma en X.

9.4.5 En el espacio del ejemplo 9.1.2, construya tres polinomios, de grados 1,2,3 respectivamente y que sean dos a dos ortogonales.

9.4.6 Pruebe que todo subconjunto ortogonal que no contenga al vector nulo es libre. En particular los conjuntos ortonormales son libres.

9.4.7 Pruebe que un vector es ortogonal a cada vector de un subconjunto C si y solo si lo es a cada vector del subespacio que C genera.

9.4.8 Matriz de Gram. Sean $g_0, g_1, \ldots, g_n$ vectores cualesquiera de un espacio con producto interno. Por definición su matriz de Gram es la matriz G de elementos $\{\langle g_i, g_j\rangle\}$. Pruebe, sin utilizar resultados de teoría de la aproximación, que G es regular si y sólo si los g_i son linealmente independientes.

9.4.9 Obtener el sistema (9.3) por procedimientos del cálculo infinitesimal, imponiendo las condiciones de minimización a la función real $F(\alpha_0, \ldots, \alpha_n) = ||f - \sum \alpha_i g_i||_2^2$.

9.4.10 Matriz de Hilbert. En el espacio del ejemplo 9.1.2, halle la matriz de Gram de los elementos $1, x, \ldots, x^n$. Tal matriz, llamada de Hilbert, es el ejemplo clásico de matriz mal acondicionada.

9.4.11

a) Encontrar la recta que mejor ajusta a los datos de la siguiente tabla en el sentido de los mínimos cuadrados, suponiendo que los valores de x están libres de error:

x	1	2	3	4	5	6
y	2.04	4.12	5.64	7.18	9.20	12.04

(Los datos están tabulados a partir de la ecuación $y = 2x$, con perturbaciones obtenidas de una tabla de números aleatorios).

b) Demostrar que el punto cuya coordenada x es el promedio de todos los valores de x y cuya coordenada y el de todos los valores de y, pertenece a la recta de ajuste. ¿Ocurre ésto en todos los casos?

c) La ocurrencia, debida al azar, de tres desviaciones consecutivas del mismo signo, permite pensar que tal vez una curva ajuste mejor estos datos. Probar con un ajuste cuadrático y calcular las sumas de las desviaciones en los dos casos.

9.4.12 Si representasemos los datos de la siguiente tabla en un papel semi-logarítmico, encontrariamos que los puntos están casi alineados.

x	77	100	185	239	285
y	2.4	3.4	7.0	11.1	19.6

Esto sugiere una relación del tipo $y = ae^{bx}$. Determinar las constantes a y b que mejor ajuste los valores por mínimos cuadrados. Utilizar la relación $\ln y = \ln a + bx$.

9.4.13 En $[a, b]$ se considera una partición uniforme Δ dada por $x_i = a + ih$, $i = 0, 1, \ldots, n$; $h = (b - a)/n$. Respecto del producto interno,

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

calcule la matriz de Gram de la base usual del espacio $M_0^1(\Delta)$ (cf. sección 5.4). Haga a continuación el caso no uniforme.

9.4.14 En la situación del ejercicio anterior, con $a = 0$, $b = 1$, encuentre la mejor aproximación a e^x por elementos de $M_0(\Delta)$, $\Delta = \{0, 1/2, 1\}$. Calcule la distancia a e^x de la mejor aproximación hallada. ¿Cuál es la distancia a e^x de su interpolante en $M_0^1(\Delta)$?

9.4.15 Pruebe que si A es una matriz real $m \times n$, la matriz A^TA es una matriz real $n \times n$, simétrica y semidefinida positiva. Pruebe, utilizando sólo álgebra lineal, que A^TA es regular (equivale a decir definida positiva) si y sólo si las columnas de A son linealmente independientes. Pruebe, utilizando solo algebra lineal, que todo sistema de la forma $A^TAx = A^Tb$ es *compatible*. Estos resultados permiten discutir la existencia y unicidad de la solución del sistema (9.5) sin usar teoría de la aproximación.

9.4.16 Cociente de Rayleigh. Sea A una matriz $m \times m$ y $\mathbf{x} \neq \mathbf{0}$ un vector de $\mathbb{R}^m$ que aproxima un autovector de A. Para aproximar el correspondiente autovalor λ, se resuelve por mínimos cuadrados el sistema de m ecuaciones en una incógnita $A\mathbf{x} = \lambda\mathbf{x}$. Pruebe que la solución es el cociente de Rayleigh $\mathbf{x}^TA\mathbf{x}/(\mathbf{x}^T\mathbf{x})$.

9.4.17 Determinar el desarrollo en serie de Fourier de la función periódica definida en un período por

$$f(t) = \begin{cases} 0 & -\pi < t < 0 \\ \operatorname{sen} t & 0 < t < \pi \end{cases}$$

Demostrar como aplicación que

$$\frac{1}{1 \cdot 3} - \frac{1}{3 \cdot 5} + \frac{1}{5 \cdot 7} - \frac{1}{7 \cdot 9} + \cdots = \frac{\pi - 2}{4}$$

9.4.18 Determinar el desarrollo en serie de Fourier de la función periódica definida en un período por

$$f(t) = \begin{cases} 1 & 0 < t < \pi \\ -1 & \pi < t < 2\pi \end{cases}$$

9.4.19 Determinar el desarrollo en serie de Fourier de la función periódica definida en un período por

$$f(t) = \begin{cases} -t & -3 < t < 0 \\ t & 0 < t < 3 \end{cases}$$

9.4.20 Escribir el desarrollo de Fourier de la siguiente función. A continuación, escribir también un desarrollo en serie de senos solamente y otro en serie unicamente de cosenos

$$f(t) = t - t^2 \qquad 0 < t < 1$$

9.4.21 Determinar el desarrollo en serie de Fourier de la función periódica definida en un período por

$$f(t) = \begin{cases} 0 & -\pi < t < 0 \\ t^2 & 0 < t < \pi \end{cases}$$

y como aplicación deducir las siguientes sumas de series

$$\begin{aligned} 1 + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots &= \frac{\pi^2}{6} \\ 1 - \frac{1}{2^2} + \frac{1}{3^2} - \frac{1}{4^2} + \cdots &= \frac{\pi^2}{12} \\ 1 + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots &= \frac{\pi^2}{8} \end{aligned}$$

9.4.22 Determinar el desarrollo en serie de Fourier de las funciones periódicas definidas en un período por

a) $f(t) = \operatorname{sen} \dfrac{t}{2} \qquad -\pi < t < \pi$

b) $f(t) = \cos t \qquad \dfrac{-\pi}{2} < t < \dfrac{\pi}{2}$

c) $f(t) = \begin{cases} \cos t & -\pi < t < 0 \\ \operatorname{sen} t & 0 < t < \pi \end{cases}$

d) $f(t) = e^t \qquad -1 < t < 1$

9.4.23 Determinar el desarrollo en serie de senos y de cosenos (por separado) de las funciones siguientes:

a) $f(t) = e^t \qquad 0 < t < 1$

b) $f(t) = \cos t \qquad 0 < t < 2\pi$

c) $f(t) = \operatorname{sen} t \qquad 0 < t < 2\pi$

d) $f(t) = \begin{cases} t^2 & 0 < t < 1 \\ 2 - t & 1 < t < 2 \end{cases}$

9.4.24 ¿Cuál es la amplitud del término resultante de frecuencia $\frac{n\pi}{p}$ en las series de Fourier de la funció cuya definición en un período se indica a continuación? ¿Cuál es la fase de cada uno de estos términos con relación a $\cos \frac{n\pi t}{p}$? ¿Y con relación a $\operatorname{sen} \frac{n\pi t}{p}$?

$$f(t) = \begin{cases} 1 & 0 < t < 1 \\ -1 & 1 < t < 2 \\ 0 & 2 < t < 4 \end{cases}$$

9.4.25 Encontrar la forma compleja de la serie de Fourier correspondiente a la función cuya definición en un período es

$$f(t) = e^{-t} \qquad -1 < t < 1$$

Comprobar a partir de dicha expresión que el desarrollo real es

$$f(t) = \sinh 1 \; - \; 2\sinh\left(\frac{\cos \pi t}{1+\pi^2} - \frac{\cos 2\pi t}{1+4\pi^2} + \frac{\cos 3\pi t}{1+9\pi^2} - \cdots\right)$$
$$- \; 2\pi \sinh 1 \left(\frac{\sin \pi t}{1+\pi^2} - \frac{2\sin 2\pi t}{1+4\pi^2} + \frac{3\sin 3\pi t}{1+9\pi^2} - \cdots\right)$$

teniendo en cuenta que $\sinh x = \frac{e^x - e^{-x}}{2}$.

¿Sirve aquí el algoritmo de la transformada rápida (FFT)?

9.4.26 Determinar la forma compleja de las series de Fourier de las funciones periódicas cuyas definiciones en un período son

a) $f(t) = t \qquad 0 < t < 1$

b) $f(t) = t \qquad -1 < t < 1$

c) $f(t) = \operatorname{sen} \dfrac{t}{2} \qquad -\pi < t < \pi$

d) $f(t) = \cos t \qquad \dfrac{-\pi}{2} < t < \dfrac{\pi}{2}$

Lección 10

Polinomios ortogonales

10.1 Funciones peso

Si (a, b) es un intervalo de la recta real, acotado o no, *una función peso* w en (a, b) es, por definición, una función real definida en (a, b), continua, positiva excepto quizá en un conjunto finito de puntos (en los que es nula), y tal que para cada $n = 0, 1, 2, \ldots$ las integrales

$$\int_a^b x^n w(x) dx \tag{10.1}$$

existan.

10.1.1 Ejemplos

a) Supongamos primero que el intervalo (a, b) es acotado. Tras un cambio lineal de variable podemos tomar $(a, b) =$ (-1,1). La función

$$w(x) = (1 - x)^\alpha (1 + x)^\beta$$

α, β reales, satisface todas las condiciones de una función peso, salvo la existencia de las integrales (10.1). Estas existen si, y sólo si, $\alpha > -1$, $\beta > -1$ ¿por qué?

b) Supongamos que (a, b) es semi-infinito y que, tras un cambio lineal de variable, $(a, b) = (0, \infty)$. La expresión

$$x^\alpha e^{-x}, \qquad \alpha > -1$$

define una función peso.

c) Si (a, b) es toda la recta, e^{-x^2} es una función peso.

10.1.2 Dada una función peso w en (a, b) la integral

$$\int_a^b f(x) g(x) w(x) dx \tag{10.2}$$

define un producto interno en el espacio $L^2_w(a,b)$ de las funciones f para las cuales

$$\int_a^b |f(x)|^2 w(x)dx < \infty$$

En la lección anterior habíamos considerado el caso particular de (10.2) en que (a,b) era acotado, y casi siempre usamos $w \equiv 1$. Al introducir funciones peso no idénticamente unidad se hace posible satisfacer (10.1) en un intervalo no acotado. Notemos además que el efecto de la función peso en los problemas de aproximación es que al calcular la distancia

$$||f-p|| = \left(\int_a^b |f(x)-p(x)|^2 w(x)dx \right)^{1/2}$$

las desviaciones $f(x)-p(x)$ correspondientes a los x en que w es mayor contribuyen más que las correspondientes a los x en que w es pequeña.

En esta lección estudiaremos problemas de aproximación en que $X = L^2_w(a,b)$, y $S = \Pi_n = \{$polinomios de grado $\leq n$ $\}$ (por la condición en (10.1), $L^2_w(a,b)$ contiene todos los polinomios). Como observamos en la lección precedente, la obtención efectiva de la mejor aproximación se simplifica notablemente si en Π_n se elige una base ortogonal. La construcción de tales bases se lleva a cabo en el punto siguiente.

10.2 Polinomios ortogonales

Dada una función peso w en (a,b), una sucesión de polinomios ortogonales $\{Q_n\}_{n=0}^{\infty}$ es aquella en que Q_n es un polinomio de grado exactamente n, $n = 0,1,2,\ldots$ y además $\langle Q_n, Q_m \rangle = 0$ si $n \neq m$, $n,m = 0,1,\ldots$

Notemos que, fijada w, si $\{Q_n\}$ y $\{R_n\}$ son dos sucesiones de polinomios ortogonales, entonces $Q_n = \alpha_n R_n$, $n = 0,1,2,\ldots$ donde α_n es un número real no nulo. En efecto,

$$Q_n(x) = q_n x^n - Q_n^*(x), \qquad R_n = r_n x^n - R_n^*(x),$$

donde q_n, $r_n \neq 0$ y Q_n^*, R_n^* tienen grado $\leq n-1$. Como $q_n x^n - Q_n^*(x)$ y $(q_n/r_n)(r_n x^n - R_n^*(x)) = q_n x^n - (q_n/r_n)R_n^*(x)$ son ortogonales a todo polinomio de grado $\leq n-1$ (¿por qué?), resulta que $Q_n^*(x)$ y $(q_n/r_n)R_n^*(x)$ son mejores aproximaciones a $q_n x^n$ por polinomios de grado $\leq n-1$. Por la unicidad, $Q_n^*(x) = (q_n/r_n)R_n^*(x)$ y por tanto $q_n/r_n R_n = Q_n$. En resumen los polinomios ortogonales están definidos salvo una normalización que determine el coeficiente director (o alternativamente el valor en un punto que no sea un cero, etc...)

Una propiedad fundamental es que los polinomios ortogonales se pueden generar por recurrencia.

10.2.1 TEOREMA

Si $\{Q_n\}_{n=0}^{\infty}$ es una sucesión de polinomios ortogonales entonces existen constantes c_n, a_n, b_n tales que

$$Q_n(x) = (c_n x - a_n)Q_{n-1}(x) - b_n Q_{n-2}(x), \quad n = 2, 3, 4, \ldots \tag{10.3}$$

Recíprocamente, definiendo

$$\begin{aligned}
&Q_0(x) = 1\\
&Q_1(x) = x - a_1\\
&a_n = \langle xQ_{n-1}, Q_{n-1}\rangle / \langle Q_{n-1}, Q_{n-1}\rangle, \quad n = 1, 2, 3, \ldots\\
&b_n = \langle xQ_{n-1}, Q_{n-2}\rangle / \langle Q_{n-2}, Q_{n-2}\rangle, \quad n = 2, 3, \ldots\\
&Q_n = (x - a_n)Q_{n-1} - b_n Q_{n-2}, \qquad n = 2, 3, \ldots
\end{aligned}$$

se genera la sucesión de polinomios ortogonales mónicos.

Demostración. Demostramos sólo el teorema directo, ya que el recíproco es análogo. Como xQ_{n-1} tiene grado exactamente n, se expresa como

$$xQ_{n-1} = \alpha_0 Q_0 + \alpha_1 Q_1 + \ldots + \alpha_{n-1}Q_{n-1} + \alpha_n Q_n, \quad \alpha_n \neq 0$$

(¿Por qué los Q_i son una base?). Por consiguiente, dividiendo por α_n

$$Q_n = \beta_0 Q_0 + \ldots + \beta_{n-2}Q_{n-2} + \beta_{n-1}Q_{n-1} + \beta_n xQ_{n-1},$$

para ciertos β_i, y sólo hay que probar que $\beta_i = 0$ si $i < n-2$. Ahora bien, para $i < n-2$,

$$\begin{aligned}
\langle Q_n, Q_i\rangle \;=\; &\beta_0\langle Q_0, Q_i\rangle + \ldots +\\
&\beta_{n-2}\langle Q_{n-2}, Q_i\rangle + \beta_{n-1}\langle Q_{n-1}, Q_i\rangle + \beta_n\langle xQ_{n-1}, Q_i\rangle.
\end{aligned}$$

El primer miembro es nulo por la hipótesis de ortogonalidad y lo mismo ocurre con todos los $\langle Q_j, Q_i\rangle$, cuando $i \neq j$. Por tanto,

$$0 = \beta_i\langle Q_i, Q_i\rangle + \beta_n\langle xQ_{n-1}, Q_i\rangle.$$

Ahora bien, $\langle xQ_{n-1}, Q_i\rangle = \langle Q_{n-1}, xQ_i\rangle$ (¿por qué?) y Q_{n-1} es ortogonal a xQ_i, al ser el grado de este $< n-1$. En definitiva $\beta_i\langle Q_i, Q_i\rangle = 0$ ó $\beta_i = 0$, y resulta

$$Q_n = (\beta_n x + \beta_{n-1})Q_{n-1} + \beta_{n-2}Q_{n-2}$$

En consecuencia, $c_n = \beta_n$ y

$$a_n = -\beta_{n-1} = \beta_n \frac{\langle xQ_{n-1}, Q_{n-1}\rangle}{\langle Q_{n-1}, Q_{n-1}\rangle} \qquad b_n = -\beta_{n-2} = \beta_n \frac{\langle xQ_{n-1}, Q_{n-2}\rangle}{\langle Q_{n-2}, Q_{n-2}\rangle}$$

□

Vemos que β_n es el cociente entre coeficientes principales y β_{n-2} se puede escribir como

$$-\frac{\beta_n\langle Q_{n-1}, Q_{n-1}\rangle}{\beta_{n-1}\langle Q_{n-2}, Q_{n-2}\rangle}$$

pues $\langle Q_n, Q_n\rangle = \beta_n\langle xQ_{n-1}, Q_n\rangle = \beta_n\langle xQ_n, Q_{n-1}\rangle$, para todo n. Esto evita algunos cálculos.

10.2.2 La relación (10.3) se llama "relación de recurrencia de tres términos". Gracias a ella es posible evaluar eficientemente polinomios P que vengan expresados en la forma $P = \alpha_0 Q_0 + \ldots + \alpha_n Q_n$ (ejercicio 4.4.9).

Por completitud, notemos que si $\{Q_n\}_0^\infty$ es una sucesión de polinomios ortogonales, entonces $Q_0, Q_1, \ldots, Q_n$ son una base de Π_n y que la mejor aproximación en Π_n a una función $f \in L_w^2$ se escribe

$$p^* = \sum_{i=0}^{n} \frac{\langle f, Q_i \rangle}{\langle Q_i, Q_i \rangle} Q_i. \tag{10.4}$$

10.2.3 Una propiedad importante de los polinomios ortogonales es la siguiente:

TEOREMA

Si $\{Q_n\}$ es una sucesión de polinomios ortogonales y $f \in L_w^2 \cap C(a,b)$ es ortogonal a $Q_0, \ldots, Q_{n-1}$ entonces ó f es idénticamente nula ó hay n puntos r_i en (a,b) en los que f cambia de signo (es decir, hay un entorno de r_i en que ó bien $f > 0$ para $x > r_i$, y $f < 0$ para $x < r_i$, ó bien $f > 0$, para $x < r_i$, y $f > 0$ para $x < r_i$).

Demostración. La condición $\langle f, Q_0 \rangle = 0$ significa

$$\int_a^b f(x) w(x) dx = 0$$

luego si f no es idénticamente nula, toma valores positivos y negativos y, siendo continua, hay, cuando menos, un punto en el que cambia el signo. Supongamos que cambie el signo sólo $k < n$ veces, y sean $r_1 < r_2 < \ldots < r_k$ los puntos en que lo hace. Entonces

$$\int_a^b f(x)(x - r_1) \cdots (x - r_k) w(x) dx \neq 0$$

pues el integrando no cambia de signo (¿por qué?). Pero esto es absurdo, pues f debe ser ortogonal a $(x - r_1) \cdots (x - r_k) \in \Pi_{n-1}$. □

10.2.4 Corolario. Q_n tiene sus n raíces reales, simples y en el intervalo (a,b).

10.3 Sistemas clásicos

10.3.1 Polinomios de Chebyshev. Los definimos en un capítulo anterior como $T_n(\cos\theta) = \cos n\theta$. Demostraremos ahora que $\{T_n\}$ es una sucesión de polinomios ortogonales en (-1,1) para la función peso

$$\frac{1}{\sqrt{1 - x^2}}$$

(Obsérvese que este es el caso $\alpha = \beta = -1/2$ del ejemplo a) de 10.1.1). En efecto, si n, m son enteros no negativos.

$$\begin{aligned}\langle T_n, T_m\rangle &= \int_{-1}^{1} T_n(x)T_m(x)\frac{dx}{\sqrt{1-x^2}} = \int_{\pi}^{0} T_n(\cos\theta)T_m(\cos\theta)\frac{-\operatorname{sen}\theta d\theta}{\operatorname{sen}\theta}\\ &= \int_0^{\pi} \cos n\theta \cos m\theta d\theta = \frac{1}{2}\int_0^{\pi}[\cos(n+m)x + \cos(n-m)x]dx\\ &= \begin{cases} 0 & \text{si} \quad n \neq m\\ \pi/2 & \text{si} \quad n = m \neq 0\\ \pi & \text{si} \quad n = m = 0\end{cases}\end{aligned}$$

Las relaciones $\langle T_n, T_m\rangle = 0$, $n \neq m$, n, m enteros no negativos, junto con $T_n(1) = 1$, $n = 0, 1, 2, \ldots$ caracterizan a los polinomios de Chebyshev (¿por qué?).

Particularizando la fórmula (10.4) al caso presente vemos que si definimos, para $f \in L^2_w$

$$A_n = \frac{2}{\pi}\int_{-1}^{1} f(x)T_n(x)\frac{dx}{\sqrt{1-x^2}}, \quad n = 0, 1, 2, \ldots \tag{10.5}$$

entonces la aproximación mejor a f por un polinomios de grado $\leq m$, respecto de la función peso $1/\sqrt{1-x^2}$ es

$$\frac{A_0}{2} + \sum_{i=1}^{n} A_i T_i(x) \tag{10.6}$$

10.3.2 Polinomios de Legendre. Se pueden definir por

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^n, \quad n = 0, 1, 2, \ldots \tag{10.7}$$

(ejercicio 10.6.3). Son ortogonales en (-1,1) para $w(x) \equiv 1$, caso particular $\alpha = \beta = 0$ en el ejemplo a) de 10.1.1. La ortogonalidad se establece como sigue. Si Q es un polinomio de grado $< n$, $n = 1, 2, \ldots$

$$\int_{-1}^{1} Q(x)\frac{d^n}{dx^n}(x^2-1)^n dx = Q(x)\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n\Big|_{-1}^{1} - \int_{-1}^{1} Q'(x)\frac{d^{n-1}}{dx^{n-1}}(x^2-1)^n dx$$

el término integrado es nulo, ya que ± 1 son ceros de orden n de $(x^2-1)^n$. Reiterando el proceso llegamos a

$$= (-1)^n \int_{-1}^{1} Q^{(n)}(x)(x^2-1)^n dx$$

que es nulo por serlo $Q^{(n)}(x)$ idénticamente.

Los $\{P_n\}$ se caracterizan por la ortogonalidad y por la propiedad $P_n(1) = 1$ (ejercicio 10.6.3), por lo que también es posible definir los P_n como los polinomios ortogonales en (-1,1) que toman en 1 el valor 1. Si se adopta tal definición, (10.7) pasa a ser un teorema, llamado de Rodrigues.

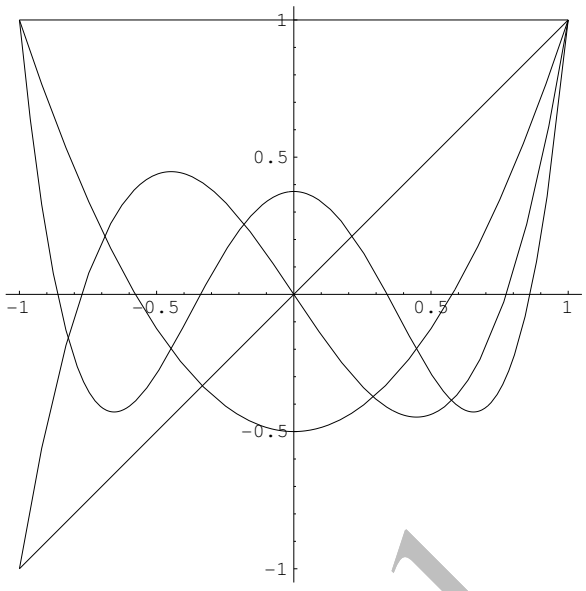


Figura 10.1: Polinomios de Legendre de grados 0 a 4

10.3.3 Polinomios de Laguerre. Se definen por

$$L_n(x) = e^x \frac{d^n}{dx^n}(x^n e^{-x}), \quad n = 0, 1, 2, \ldots \tag{10.8}$$

son ortogonales en $(0, \infty)$ para el peso e^{-x}, caso particular $\alpha = 0$ del ejemplo b) de 10.1.1 . Véase el ejercicio 10.6.4.

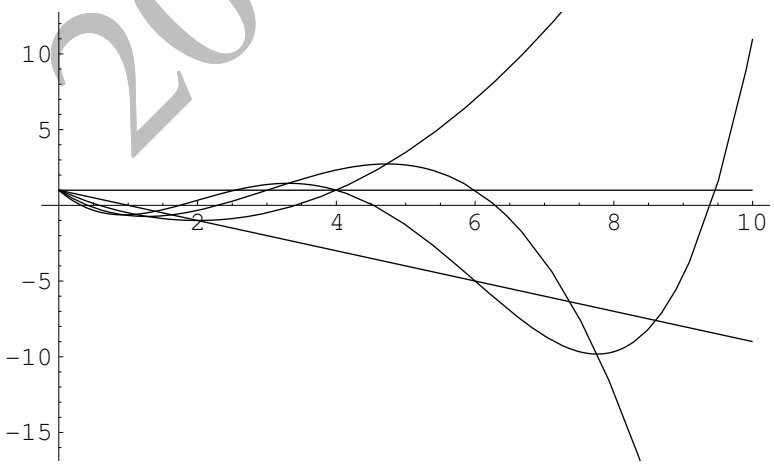


Figura 10.2: Polinomios de Laguerre de grados 0 a 4 (para $\alpha = 0$)

10.3.4 Polinomios de Hermite. Se definen por

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}(e^{-x^2}), \quad n = 0, 1, 2, \ldots \tag{10.9}$$

son ortogonales en $(-\infty, \infty)$ para el peso e^{-x^2}, ejemplo c) de 10.1.1. Véase el ejercicio 10.6.5.

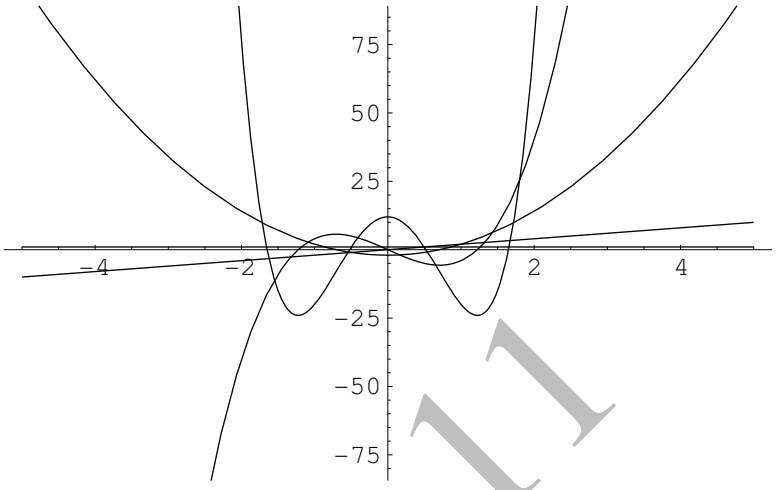


Figura 10.3: Polinomios de Hermite de grados 0 a 4

10.4 Convergencia de los desarrollos ortogonales

Supongamos ahora que X es uno de los espacios $L^2_w(a,b)$ considerados en el apartado 10.1.2 y que $S_n = \Pi_n$. Si tomamos una sucesión de polinomios ortogonales $\{Q_n\}_{n=0}^{\infty}$ en $L^2_w(a,b)$, la mejor aproximación p_n a $f \in L^2_w$ por polinomios de Π_n, se representa explícitamente como

$$p_n = \sum_{i=0}^{n} \frac{\langle f, Q_i\rangle}{\langle Q_i, Q_i\rangle} Q_i$$

y por tanto la cuestión de si $p_n \to f$ en $L^2_w(a,b)$ se reduce a examinar si

$$f = \sum_{i=0}^{\infty} \frac{\langle f, Q_i\rangle}{\langle Q_i, Q_i\rangle} Q_i \tag{10.10}$$

donde la suma de la serie se entiende en el sentido de la norma de $L^2_w(a,b)$. Cuando (10.10) vale se dice que se ha desarrollado f en *serie de polinomios ortogonales*. Nos limitaremos aquí al caso en que $(a,b) = (-1,1)$, $w(x) = 1/\sqrt{1-x^2}$ (desarrollos en serie de Chebyshev).

El desarrollo de $f \in L^2_w(a,b)$ es

$$\frac{1}{2}A_0 + \sum_{k=1}^{\infty} A_k T_k(x) \tag{10.11}$$

donde los coeficientes A_k están dados por la fórmula (10.5) .

10.4.1 El siguiente resultado no tiene las hipótesis más débiles posibles y es fácil de demostrar.

TEOREMA

Si $f \in C([-1,1])$, entonces su desarrollo de Chebyshev converge a f en la norma de L^2_w, $w(x) = 1/\sqrt{1-x^2}$.

Demostración. Observemos ante todo que $C([-1,1])$ está contenido L^2_w (¿por qué?). Basta probar que f es límite en L^2_w de polinomios. Por el teorema de Weierstrass, dado $\varepsilon > 0$ hay un polinomio P con $||f-p||_\infty \leq \varepsilon$, pero entonces, en L^2_w

$$||f-p|| = \left(\int_{-1}^{1} [f(x)-p(x)]^2 \frac{dx}{\sqrt{1-x^2}} \right)^{1/2} \leq \varepsilon \left(\int_{-1}^{1} \frac{dx}{\sqrt{1-x^2}} \right)^{1/2} = \varepsilon \pi^{1/2}$$

lo que demuestra que podemos aproximar f por polinomios tanto como deseemos.□

10.4.2 Bajo hipótesis mas restrictivas la serie (10.11) converge hacia f *uniformemente*. (Note que la convergencia uniforme implica la convergencia en L^2_w , pero no al revés.)

TEOREMA

Si $f \in C^2([-1,1])$, su desarrollo en serie de Chebyshev converge uniformemente hacia f.

Demostración. Cambiando de variable $x = \cos\theta$ en la expresión de los A_k

$$A_k = \frac{2}{\pi} \int_0^\pi f(\cos\theta) \cos k\theta d\theta$$

Poniendo $f(\cos\theta) = g(\theta)$, dos integraciones por partes dan

$$A_k = -\frac{2}{\pi k^2} \int_0^\pi \cos k\theta g''(\theta) d\theta$$

(¿por qué?) . Así $|A_k| = 0(k^{-2})$, luego por el criterio M de Weierstrass, la serie (10.11) converge uniformemente hacia cierta función continua F. Pero el teorema anterior muestra que la serie converge a f en L^2_w. Por la unicidad del límite en L^2_w, $f = F$ y el teorema está probado.□

10.4.3 Notas:

a) Si en el teorema $f \in C^n([-1,1])$, $n \geq 2$, se puede iterar la integración por partes y $|A_k| = 0(k^{-n})$: *La regularidad de la función se traduce en la rapidez de decaimiento de los coeficientes del desarrollo de Chebyshev y por ello en la rapidez en la convergencia de la serie de Chebyshev.*

b) Según el teorema $f \in C^2$ basta para garantizar la convergencia uniforme del desarrollo de Chebyshev. Sin embargo C^∞ no garantiza ni la convergencia puntual del desarrollo en serie de Taylor!

10.5 Cuadratura Gaussiana

Hasta ahora, suponíamos que, en la fórmula de cuadratura

$$I_{n+1}(f) = \sum_{j=0}^{n} \alpha_j f(x_j) \tag{10.12}$$

los x_j habían sido elegidos de antemano y veíamos que los $n+1$ parámetros α_j podían determinarse únicamente para que (10.12) fuese exacta en las $n+1$ funciones $1, x, \ldots, x^n$. Supongamos ahora que en (10.12) podemos elegir también los nodos x_j. Disponemos de $2n+2$ parámetros y parece plausible esperar que (10.12) pueda hacerse exacta para las $2n+2$ funciones $1, x, \ldots, x^{2n+1}$. Imponer esa exactitud conduce (método directo) a un sistema con $2n+2$ ecuaciones para $2n+2$ incógnitas. Sin embargo la solubilidad de tal sistema no puede ser decidida por métodos algebraicos ya que las ecuaciones no son lineales.

Las fórmulas de cuadratura que eligen los nodos para lograr el mayor grado de exactitud posible se llaman gaussianas en honor a su inventor. A pesar de su antigüedad, su popularización ha sido reciente: en la época del cálculo manual, los datos, al estar tabulados, imponían los nodos.

10.5.1 Fórmulas gaussianas Tratemos de determinar $\{\alpha_j\}$, $\{x_j\}$ en (10.12) para lograr el mayor grado de precisión posible. Como tal fórmula ha de ser interpolatoria (¿por qué?)

$$\int_a^b f(x)dx - I_{n+1}(f) = \int_a^b W(x) f[x_0, \ldots, x_n, x]dx$$

con

$$W(x) = (x - x_0) \cdots (x - x_n) \tag{10.13}$$

De éste modo, (10.12) es exacta para f si y sólo si $f[x_0, \ldots, x_n, x]$ es *ortogonal* a W (en $[a,b]$ con peso unidad). Si f es un polinomio de grado $n+k+1$, $k = 0, 1, 2, \ldots$ su diferencia $f[x_0, x_1, \ldots, x_n, x]$ es un polinomio en x de grado k. Más precisamente cuando f recorre Π_{n+k+1}, su diferencia recorre *todo* Π_k (¿por qué?). Así W es exacta en Π_{n+k+1} si y sólo si $\{x_j\}$ son tales que W es ortogonal a todo polinomio de grado $\leq k$. No es entonces posible tomar $k = n+1$, ya que W tiene grado $n+1$. Para $k = n$ la ortogonalidad $W \perp \Pi_k$ se produce si y sólo si los $\{x_j\}$ se eligen como las $n+1$ raíces

distintas del polinomio ortogonal de grado $n+1$ (en $[a,b]$ con peso 1) (note que tales raíces están en (a,b)). Podemos concluir entonces:

TEOREMA

Una fórmula del tipo (10.12) con $n+1$ nodos puede tener a lo sumo grado $2n+1$. Tal máximo se alcanza únicamente en la fórmula interpolatoria llamada gaussiana basada en los ceros del polinomio mónico de grado $n+1$ ortogonal a los de grado n en (a,b) para el peso unidad.

10.5.2 Ejemplo. Construyamos la regla gaussiana de 3 nodos en [-1,1]. Según el teorema los nodos son los ceros de uno cualquiera de los polinomios P de grado tres que son ortogonales a todos los de grado a lo sumo dos. Estos P son multiplos escalares del polinomio de Legendre, que de acuerdo con la fórmula 10.7 es $(5x^3-3)/2$. Así los nodos son $\pm\sqrt{3/5}$, 0. Los pesos se determinan imponiendo que la regla sea interpolatoria, o equivalentemente que sea exacta para $1, x, x^2$. Así se obtiene (efectúe los cálculos) que el nodo 0 tiene peso 8/9 y los dos restantes peso 5/9 cada uno. (Vea problema 10.6.11.) Esta fórmula integra exactamente todos los polinomios hasta el grado 5 inclusive.

10.5.3 Errores En relación con los errores de redondeo, una propiedad interesante de las reglas gaussianas es la siguiente:

TEOREMA

Los coeficientes α_j, $j=0,\ldots,n$ de la regla gaussiana de $n+1$ nodos son todos positivos.

Demostración. Si Q_j denota el polinomio, de grado $2n$, $W^2/(x-x_j)^2$, podemos escribir, a la vista del grado de las fórmulas de Gauss

$$\int_a^b Q_j(x)dx = \sum_{k=0}^{n} \alpha_k Q_j(x_k)$$

La integral es positiva. En la suma, los términos con $k \neq j$ tienen $Q_j(x_k)=0$, mientras que $Q_j(x_j) = (x_j-x_0)^2\cdots(x_j-x_n)^2 > 0$. Es claro entonces que α_j es positivo. □

Para el error de interpolación tendremos:

TEOREMA

Si f tiene $2n+2$ derivadas continuas en $[a,b]$, el error de cuadratura $E_{n+1}(f)$ de la regla de cuadratura gaussiana de n+1 nodos viene dado por

$$E_{n+1}(f) = \frac{f^{(2n+2)}(\xi)}{(2n+2)!}\int_a^b W^2(t)dt$$

siendo ξ un punto de $[a,b]$ (que depende de f) y W la función en (10.13).

Demostración. Denotemos por P el único polinomio de grado $\leq 2n+1$, tal que $P(x_j) = f(x_j)$, $P'(x_j) = f'(x_j)$, $0 \leq j \leq n$.

$$E_{n+1}(f) = \int_a^b f(x)dx - \sum_{j=0}^{n} \alpha_j f(x_j) = \int_a^b f(x)dx - \sum_{j=0}^{n} \alpha_j P(x_j)$$

Invocando la exactitud de la regla gaussiana para P

$$E_{n+1}(f) = \int_a^b [f(x) - P(x)]dx$$

bastando entonces para concluir la prueba usar la forma del error en la interpolación de Hermite y el teorema del valor medio del cálculo integral.□

10.6 Cuestiones y problemas

10.6.1 ¿Existe una función peso en un intervalo apropiado para la que $1 + x^2$ sea un miembro de una sucesión de polinomios ortogonales? ¿Y para la que x^2 lo sea?

10.6.2 Pruebe que si f es una función continua de $L^2_w(a, b)$ su mejor aproximación p^*_n por un polinomio de grado $\leq n$ la interpola en $n+1$ puntos distintos de (a, b). Así p^*_n es de hecho un polinomio interpolador de Lagrange. Este resultado no puede usarse para construir p^*_n pues los nodos de la interpolación, que dependen de f, son desconocidos *a priori* .

10.6.3 Polinomios de Legendre Demostrar las siguientes propiedades:

a) La expresión (10.7) define un polinomio de grado exactamente n.

b) $P_n(1) = 1$, $n = 0, 1, 2, \ldots$. Justamente la constante $2^n n!$ en (10.7) se elige para que valga esta propiedad.

c) P_n es un polinomio par o impar de acuerdo con la paridad de n.

10.6.4 Pruebe que la expresión (10.8) define un polinomio de grado $\leq n$. Pruebe la ortogonalidad de tales polinomios. Pruebe que los $\{L_n\}$ se caracterizan por la ortogonalidad y por tener coeficiente director $(-1)^n/n!$

10.6.5 Pruebe que la expresión (10.9) define un polinomio de grado $\leq n$. Pruebe la ortogonalidad de tales polinomios. Pruebe que los $\{H_n\}$ se caracterizan por la ortogonalidad y por tener coeficiente director 2^n.

10.6.6 Demostrar las siguientes fórmulas de recurrencia:

a) $P_0 = 1$, $P_1 = x$,

$$P_{n+1} = \frac{2n+1}{n+1} x P_n(x) - \frac{n}{n+1} P_{n-1}(x) \qquad n \geq 1$$

b) $L_0 = 1$, $L_1 = -x + 1$,

$$L_{n+1} = (2n + 1 - x)L_n(x) - n^2 L_{n-1}(x) \qquad n \geq 1$$

c) $H_0 = 1$, $H_1 = 2x$,

$$H_{n+1} = 2xH_n(x) - 2nH_{n-1}(x) \qquad n \geq 1$$

10.6.7 Polinomios de Jacobi $P_n^{(\alpha,\beta)}(x)$. Para α, β parámetros reales > -1, la sucesión $\{P_n^{(\alpha,\beta)}\}_{n=0,1,2,\ldots}$ se define por ser ortogonal en $(-1, 1)$ para el peso $(1-x)^\alpha(1+x)^\beta$ junto con la condición

$$P_n^{(\alpha,\beta)}(1) = \frac{\Gamma(n + \alpha + 1)}{\Gamma(n + 1)\Gamma(\alpha + 1)}$$

Demuestre que cuando $\alpha = \beta = 0$ los polinomios de Jacobi son los de Legendre. Estudie la relación entre $P_n^{(-1/2,-1/2)}$ y T_n.

10.6.8 Polinomios de Chebyshev de segunda especie. Pruebe que para cada $n = 0, 1, 2, \ldots$ hay un único polinomio U_n tal que, para cada θ real, $U_n(\cos\theta)\operatorname{sen}\theta = \operatorname{sen}(n+1)\theta$. Pruebe que U_n (Polinomio de Chebyshev de segunda especie) tiene grado n. Pruebe la relación de recurrencia $U_{n+2}(x) = 2xU_{n+1}(x) - U_n(x)$, $n = 1, 2$. Pruebe que los U_n son ortogonales en $(-1, 1)$ para el peso $\sqrt{1 - x^2}$. Halle una fórmula de Rodrigues que exprese U_n como múltiplo escalar de la derivada n-ésima de una función.

10.6.9 Pruebe que si una función f continua en [0,1] es tal que

$$\int_0^1 f(x)x^n dx = 0,$$

para $n = 0, 1, 2, \ldots$ entonces f es idénticamente nula.

10.6.10 Desarrolle en serie de Chebyshev las funciones

$$\sqrt{\frac{1+x}{2}}, \sqrt{1 - x^2}, \arccos x$$

10.6.11 En [-1,1] demuestre que la regla gaussiana es simétrica en el sentido de que si x es un nodo de tal regla, $-x$ también lo es y además x y $-x$ entran con el mismo peso. (Utilice que la regla gaussiana es la única con su grado de precisión.)

10.6.12 Cuadratura gaussiana respecto de una función peso. En un intervalo abierto (a, b) no necesariamente acotado, sea w una función peso (véase 10.1). Se trata de aproximar integrales de la forma

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

por una combinación lineal de valores de f (no de valores del integrando) en $n+1$ puntos distintos fijados (independientes de f) x_i. Los pesos de la combinación lineal también se suponen independientes de f. Pruebe, que dados los x_i, hay una única elección de pesos que haga la fórmula exacta cuando f es un polinomio de grado $\leq n$. Con esos únicos pesos la fórmula se llama interpolatoria ¿por qué? Pruebe que hay una única elección de nodos y pesos para los que la fórmula integre exactamente todos los polinomios de grado $\leq 2n+1$, y que tal fórmula no integra exactamente todos los polinomios de grado $2n+2$. ¿Son los pesos positivos? Dé una expresión para el error de cuadratura.

10.6.13 Cuadratura de Lobatto. Se va a aproximar

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

por una regla del tipo $\alpha_0 f(a) + \alpha_1 f(x_1) + \ldots + \alpha_{n-1} f(x_{n-1}) + \alpha_n f(b)$, donde $\alpha_0, \alpha_1, \ldots, \alpha_n$; $x_1, \ldots, x_{n-1}$ son parámetros a determinar ($n \geq 1$). Demuestre que hay una única fórmula (llamada de Lobatto) para la que el grado de precisión sea $2n-1$. ¿Qué ventajas puede tener, en uso compuesto, la regla de Lobatto sobre la gaussiana con el mismo número de nodos? ¿Cuál es la fórmula de Lobatto con $n=1$? ¿Y con $n=2$?

10.6.14 Abscisas de la fórmula de Lobatto. Pruebe que las abscisas de la fórmula de Lobatto de $n+1$ nodos son a, b y los $n-1$ ceros de la derivada del polinomio ortogonal de grado n en (a,b) para el peso 1.

10.6.15 Cuadratura de Radau. Se va a aproximar

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

por una regla del tipo $\alpha_0 f(a)+\alpha_1 f(x_1)+\ldots+\alpha_n f(x_n)$, donde $\alpha_0, \alpha_1, \ldots, \alpha_n$; $x_1, \ldots, x_n$ son parámetros libres. Demuestre que hay una única fórmula (llamada de Radau) para la que el grado de precisión sea $2n$.

2010-11

Lección 11

Aproximación funcional discreta

11.1 El efecto del '*sampling*'-'*aliasing*'

La construcción de aproximaciones a una función (óptimas en distintos normas *euclídeas* que hemos ido construyendo), lleva aparejada en general la necesidad de realizar cálculos analíticos para obtener los coeficientes. Como ejemplo, recordemos las expresiones (10.8) y (10.9), o las (11.5). Estos cálculos son con frecuencia complejos, a veces imposibles y desde luego siempre ajenos a los métodos numéricos que nos ocupan. Naturalmente, siempre podemos evaluar dichas integrales numéricamente, pero tiene interés abordar directamente el problema.

Ya hemos visto como se puede resolver el problema de ajuste para una serie de datos tabulados por un procedimiento que desemboca en un problema lineal de mínimos cuadrados, siendo la técnica de resolución idéntica al caso continuo pero utilizando productos escalares discretos, donde las integrales se sustituyen por sumas finitas y se evita, en consecuencia, la necesidad de cálculos simbólicos. La idea consiste en aprovechar estas técnicas para calcular aproximaciones funcionales *discretas*; es decir, basadas en un conjunto finito de puntos, tratando de que sean lo más parecidas posibles a los aproximantes continuos.

Tenemos el aliciente de que al disponer de la función podemos elegir la cantidad y ubicación de los puntos que deseemos, con vista a que se verifiquen propiedades útiles desde el punto de vista computacional, como la ortogonalidad de las *funciones* (serán siempre vectores de un espacio de dimensión finita) base, que hemos visto es la piedra angular en que se basa una resolución eficiente del problema de mínimos cuadrados.

Sin embargo, hay dos fenómenos, estrechamente asociados, que surgen en la aproximación funcional discreta. Mantenemos las palabras inglesas para denominarlos por lo extendidas que están en la literatura. El primero es el *sampling* o muestreo. Desde el momento que seleccionamos un conjunto de puntos la función deja de serlo y pasa a ser un vector, es como si sólo viésemos esos puntos a través de una rejilla y ni podemos utilizar los puntos ignorados, ni podemos hacer nada porque la aproximación en ellos sea mejor de lo que los datos utilizados permite.

El *aliasing* es una consecuencia inmediata. Dos funciones muy diferentes pueden ser idénticas sobre la red elegida, en nuestra *rejilla* las vemos iguales y, de hecho, una vez seleccionados los puntos son iguales. Las figuras 11.1 y 11.2 nos muestran como sobre un conjunto equidistante de puntos en el intervalo $[0, 2\pi]$, las funciones $\cos(x), \cos(7x)$ y $\cos(9x)$ coinciden en la primera de ellas, mientras que $\operatorname{sen}(x), -\operatorname{sen}(7x)$ y $\operatorname{sen}(9x)$ son

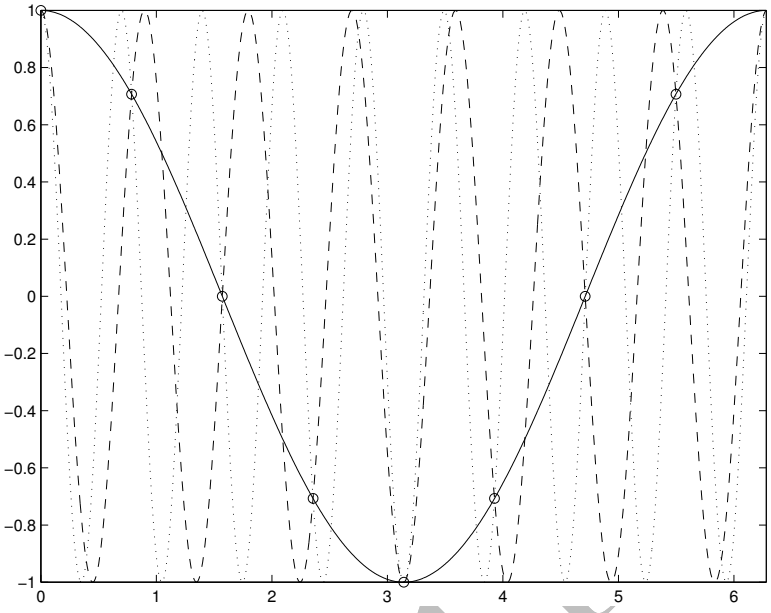


Figura 11.1: Funciones coseno idénticas sobre una red equiespaciada

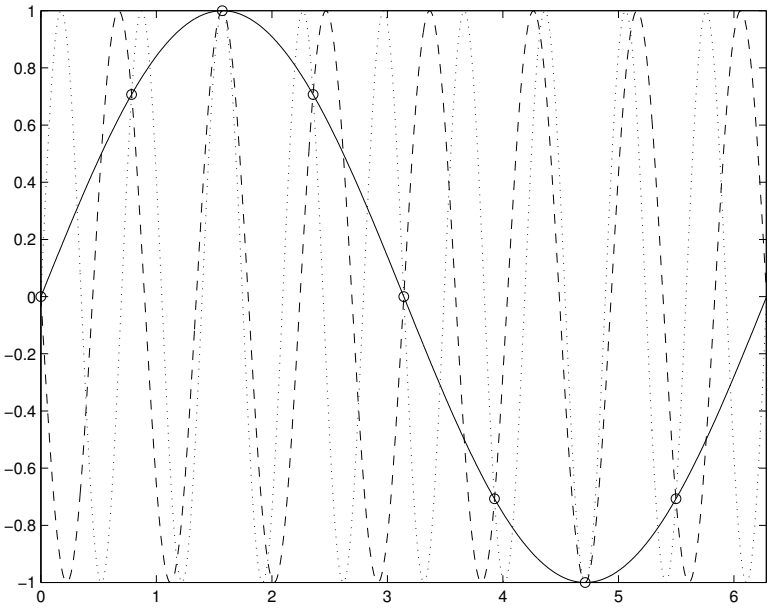


Figura 11.2: Funciones seno idénticas sobre una red equiespaciada

las que lo hacen en la segunda. Podría decirse que sobre esos puntos cada una de las funciones es una *alias* de las otras. Es fácil ver que hay infinitas de estas funciones indistinguibles en el marco de trabajo en que nos hemos situado. Busque el lector unas cuantas dentro de las mismas familias.

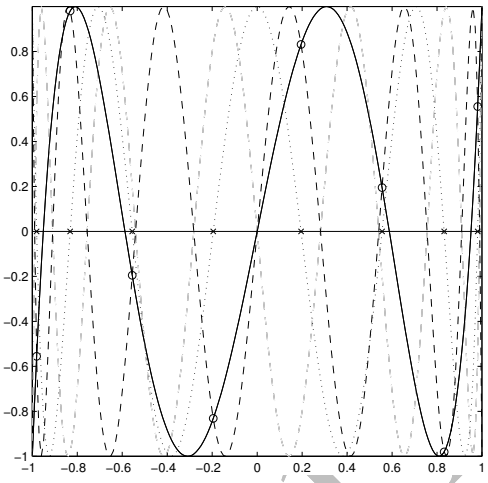


Figura 11.3: Polinomios de Chebyshev *aliados* en las raíces de otro

Otro caso de interés y gran utilidad nos lo proporcionan, como no, los polinomios de Chebyshev. La relación (demuéstrela el lector)

$$T_{N+m}(x) + T_{N-m}(x) = 2T_N(x)T_m(x)$$

válida para todo m y N, nos permite deducir que sobre la rejilla formada por las raíces de T_N para un N fijado, $T_{N+m}(x)$ y $T_{N-m}(x)$ son alias uno del otro, concretamente son opuestos. En la figura 11.3 vemos como T_{11} y $-T_5$ coinciden sobre los ceros de T_8.

Este fenómeno, que en principio puede causar cierta confusión tiene sin embargo una gran cantidad de aplicaciones, y en el caso concreto del tema de esta lección nos permite establecer interesantes resultados.

El *aliasing* de polinomios ocurre de hecho, aunque no sea reconocido como tal, en cualquier proceso de aproximación discreta por polinomios. Cuando la aproximación está basada en los puntos $x_i, i = 1, 2, \ldots, n$, el polinomio

$$P(x) = (x - x_1)(x - x_2)\cdots(x - x_n)$$

es esencial en todo el proceso. Es bien sabido que si $P_m(x)$ es un polinomio de grado $m \geq n$ y le dividimos por $P(x)$, resulta que obtenemos como resto un polinomio $R(x)$ de grado inferior a n (si $m < n$, es evidente que $R(x) = P_m(x)$), pero que sobre la rejilla de trabajo, toma los mismos valores que $P_m(x)$, el decir es un *alias* de él.

Mas aún, si realizamos este cociente para $x^m, m = 0, 1, \ldots$ y denominamos $R_m(x)$ al correspondiente resto, tenemos que a partir del desarrollo de Taylor de una función

analítica cualquiera $f(x)$, en los puntos del *sampling* se verifica

$$f(x) = \sum_{m=0}^{\infty} a_m x^m \to \sum_{m=0}^{\infty} a_m R_m(x)$$

donde el último miembro es un polinomio de grado $n-1$ (a lo sumo), que coincide con la función $f(x)$ en los puntos $x_i, i = 1, 2, \ldots, n$. En otras palabras es el polinomio interpolador de la función en esos puntos, donde cada uno de los coeficientes aparecera como una suma infinita de multiplos de los a_m.

11.2 Aproximación trigonométrica

Comenzamos por este caso que es el más ilustrativo de las ideas expuestas en el apartado anterior. El problema es encontrar de manera eficiente aproximaciones numéricas a los sucesivos coeficientes de la serie de Fourier (9.9), sin tener que recurrir a una integración numérica. La idea es resolver un problema de mínimos cuadrados sobre una red de puntos lo suficientemente densa para que la aproximación discreta nos de la precisión requerida.

En este caso, la familia de funciones trigonométricas tomada como base, tiene la interesante propiedad de ser tambien ortogonal sobre un conjunto discreto de puntos fáciles de construir y manipular. Concretamente, si consideramos la red

$$x_m = x_0 + \frac{2\pi}{N} m, \qquad m = 0, 1, \ldots, N-1, \qquad x_0 \in \left[0, \frac{2\pi}{N}\right]$$

de N puntos equidistantes en el intervalo $[0, 2\pi]$, veremos que es fácil extraer un subconjunto de N elementos de la familia que sean ortogonales sobre la red en el sentido discreto

$$\langle f, g \rangle = \sum_{m=0}^{N-1} f(x_m) g(x_m) \tag{11.1}$$

Observesé que nunca pueden pertenecer a la red los dos extremos del intervalo $[0, 2\pi]$. Lo que se hace es dividir el intervalo en N partes iguales, y elegir un punto de cada una de ellas. Esto nos permitirá suponer que la función discretizada es periódica, pues bastaría reasignar los valores en los extremos, o al menos en uno de ellos que siempre va a ser superfluo para el problema discreto. De cualquier forma esto es imprescincible para la ortogonalidad que deseamos, puesto que al ser las funciones de la base periódicas, si consideramos los dos extremos el valor en dichos puntos entraría dos veces en la suma del producto escalar.

Para simplificar, tomamos $x_0 = 0$ (véase problema 11.5.1). Además, vamos a tomar N par, puesto que simplifica un poco el desarrollo y es el caso más frecuente. Concretamente, se puede probar que si $0 \le j, k \le \frac{N}{2}$

$$\begin{aligned}
\sum_{m=0}^{N-1} \cos jx_m \operatorname{sen} kx_m &= 0, \qquad \text{para todo } j, k \\
\sum_{m=0}^{N-1} \cos jx_m \cos kx_m &= \begin{cases} 0, & \text{si } j \neq k \\ \frac{N}{2}, & \text{si } j = k \neq 0, \frac{N}{2} \\ N, & \text{si } j = k = 0, \frac{N}{2} \end{cases}
\end{aligned} \tag{11.2}$$

$$\sum_{m=0}^{N-1} \operatorname{sen} jx_m \operatorname{sen} kx_m \quad = \quad \begin{cases} 0, & \text{si } j \neq k \\ \frac{N}{2}, & \text{si } j = k \neq 0, \frac{N}{2} \end{cases}$$

Comencemos por observar que de la expresión $e^{ix} = \cos x + i \operatorname{sen} x$, cuando i es la unidad imaginaria, se deduce sin dificultad que

$$\sum_{m=0}^{N-1} e^{ijx_m} = \sum_{m=0}^{N-1} e^{ij(\frac{2\pi}{N}m)} = \sum_{m=0}^{N-1} (e^{ij\frac{2\pi}{N}})^m$$

El primer miembro es

$$\sum_{m=0}^{N-1} \cos jx_m + i \sum_{m=0}^{N-1} \operatorname{sin} jx_m$$

y el segundo es la suma de N términos de una progresión geométrica de razón $r = e^{ij\frac{2\pi}{N}}$, cuyo valor es, teniendo en cuenta que el primer término es 1

$$\begin{cases} N & \text{si } r = 1, \text{ es decir, si } j \text{ es un múltiplo entero de } N \\ \frac{r^N-1}{r-1} = \frac{e^{2\pi ij}-1}{r-1} = 0 & \text{si } r \neq 1 \end{cases}$$

puesto que, como es bien sabido $e^{2\pi i} - 1 = 0$ (¿por qué?).

En consecuencia,

$$\sum_{m=0}^{N-1} \cos jx_m \quad = \quad \begin{cases} 0 & \text{si } j \neq 0, \pm N, \pm 2N, \ldots \\ N & \text{si } j = 0, \pm N, \pm 2N, \ldots \end{cases}$$

$$\sum_{m=0}^{N-1} \operatorname{sen} jx_m \quad = \quad 0 \quad \text{para todo } j$$

Ahora, a partir de conocidas relaciones trigonométricas, se obtienen las siguientes relaciones, que contienen a las (11.3) como casos particulares. Para la primera de ellas, que por otra parte es obvia al tratarse del producto de una función par y una impar en el intervalo de trabajo, nos basamos en

$$\sum_{m=0}^{N-1} \cos jx_m \operatorname{sen} kx_m = \frac{1}{2} \sum_{m=0}^{N-1} [\operatorname{sen}(j+k)x_m - \operatorname{sen}(j-k)x_m] = 0$$

Para la segunda, la secuencia es como sigue

$$\sum_{m=0}^{N-1} \cos jx_m \cos kx_m = \frac{1}{2} \sum_{m=0}^{N-1} [\cos(j+k)x_m + \cos(j-k)x_m] =$$

$$= \begin{cases} 0, & \text{si } j+k \text{ y } j-k \text{ no son } 0, \pm N, \pm 2N, \ldots \\ \frac{N}{2}, & \text{si } j+k \text{ ó } j-k \text{ son } 0, \pm N, \pm 2N, \ldots \text{ (pero no ambos)} \\ N, & \text{si } j+k \text{ y } j-k \text{ son a la vez de la forma } 0, \pm N, \pm 2N, \ldots \end{cases} \tag{11.3}$$

Y, finalmente, para la tercera tenemos

$$\sum_{m=0}^{N-1} \operatorname{sen} jx_m \operatorname{sen} kx_m = \frac{1}{2} \sum_{m=0}^{N-1} [\cos(j-k)x_m + \cos(j+k)x_m] =$$

$$= \begin{cases} 0, & \text{si } j+k \text{ y } j-k \text{ no son } 0, \pm N, \pm 2N, \ldots \text{ o ambos lo son} \\ \frac{N}{2}, & \text{si } j-k = 0, \pm N, \pm 2N, \ldots \\ -\frac{N}{2}, & \text{si } j+k = 0, \pm N, \pm 2N, \ldots \end{cases} \tag{11.4}$$

De cara al futuro, es bueno tener en cuenta que $j+k$ es un múltiplo de N si y sólo si $k = Nl - j, l = 0, \pm 1, \pm 2, \ldots$; $j-k$ lo es cuando y sólo cuando $k = Nl + j, l = 0, \pm 1, \pm 2, \ldots$; y, para que ambas cantidades lo sean, es necesario y suficiente que j y k sean simultáneamente la mitad de un múltiplo entero de N.

11.2.1 Base ortogonal. Como resumen de todo lo anterior, podemos seleccionar un conjunto de N funciones que son ortogonales respecto al producto escalar discreto (11.1), lo que implica que los vectores N-dimensionales que forman sus valores sobre la red de puntos lo son respecto al producto escalar ordinario y, en consecuencia, son independientes. Por una parte, tenemos las $\frac{N}{2}+1$ funciones

$$1, \cos x, \cos 2x, \ldots, \cos \frac{N}{2} x$$

y no más funciones coseno, ya que $\cos(\frac{N}{2}+1)x$ no es ortogonal a $\cos(\frac{N}{2}-1)x$, como de deduce de (11.3). De hecho ambas funciones son 'alias' sobre la red de puntos que estamos utilizando (¿de acuerdo?). Lo mismo ocurre para cualquier otra función coseno que no aparece en la lista, siempre tiene un 'alias' dentro de ella, al que en consecuencia no puede ser ortogonal en el sentido discreto que estamos tratando en este capítulo, aunque lo sea como sabemos en la forma continua.

Para completar una base de un espacio que tiene dimensión N necesitamos aún otras $\frac{N}{2}-1$. Resulta ya muy fácil comprobar que nos sirve el conjunto

$$\operatorname{sen} x, \operatorname{sen} 2x, \ldots, \operatorname{sen} \left(\frac{N}{2} - 1\right) x$$

donde la ausencia de sen $\frac{N}{2}x$ se debe a que es idénticamente nula en la red de puntos, y las siguientes funciones seno ya disponen de un 'alias' entre las seleccionadas.

Por tanto, hemos resuelto el problema discreto de mínimos cuadrados, que planteamos como aproximación numérica al continuo. La mejor aproximación sobre el conjunto $\{x_m\}, m = 0, 1, \ldots, N-1$ a la función f se escribe en la forma

$$\frac{A_0}{2} + \sum_{k=1}^{\frac{N}{2}-1} (A_k \cos kx + B_k \operatorname{sen} kx) + \frac{A_{\frac{N}{2}}}{2} \cos \frac{N}{2} x \tag{11.5}$$

siendo

$$A_k = \frac{1}{\frac{N}{2}} \sum_{m=0}^{N-1} f(x_m) \cos kx_m \qquad k = 0, 1, \ldots, \frac{N}{2} \tag{11.6}$$

$$B_k = \frac{1}{\frac{N}{2}} \sum_{m=0}^{N-1} f(x_m) \operatorname{sen} kx_m \qquad k = 1, \ldots, \frac{N}{2} - 1 \tag{11.7}$$

los coeficientes de Fourier correspondientes a esta base, donde los $\frac{N}{2}$ del denominador son las normas (al cuadrado) de los elementos ortogonales. Las normas de $1 = (1, 1, \ldots, 1)$ y de $\cos \frac{N}{2}x = (1, -1, 1, -1, \ldots, -1)$ que valen N están disfrazadas con el 2 del denominador de los coeficientes.

Obsérvese que ya no tenemos que integrar como en el caso continuo, sino solamente multiplicar una determinada matriz (que contiene en sus filas los valores de la base en la red) por el vector columna de los valores de f en los mismos puntos. Recuérdese que el error cuadrático en este caso va a ser 0, pues se trata de una interpolación.

11.2.2 Ejemplo. Tratemos de aproximar la misma función que en apartado 7.3 utilizando la base ortogonal sobre el conjunto de los puntos $x_k = \frac{2\pi}{8}k$, $k = 0, 1, \ldots, 7$ que allí utilizabamos para interpolar. La base estará formada por cinco funciones coseno (incluyendo la constante 1) y tres funciones seno. El mejor aproximante, en la norma derivada del producto escalar, se obtendrá calculando los coeficientes según las fórmulas (11.6) y (11.7) respectivamente. Si utilizamos toda la base para construir el polinomio trigonométrico aproximante, nos encontraremos con una interpolación (¿por qué?). Resulta interesante comparar este interpolante con el que muestra la figura 7.4. Unos sencillos cálculos nos llevan a obtener los valores $A_0 = 2.0157, A_1 = 1.0196, A_2 = 0.5333, A_3 = 0.3137, A_4 = 0.2510$ y $B_1 = B_2 = B_3 = 0$, que sustituídos en (11.5) y evaluando en el intervalo $[0, 2\pi)$ nos da el resultado que muestra la figura 11.4.

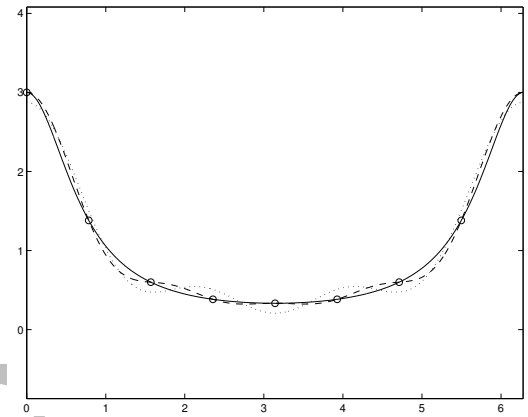


Figura 11.4: Aproximación trigonométrica por mínimos cuadrados

Es evidente que el interpolante es como aproximación muy superior al obtenido en 7.4 donde sólo atendiamos a la interpolación, mientras que ahora es un caso extremo de un proceso de aproximación funcional (aunque discreta en este caso). Se ha dibujado también con línea de puntos el polinomio trigonométrico con dos términos menos y vemos que, aunque lógicamente no interpola, es una aproximación excelente.

11.3 Polinomios de Gram

En el caso de que queramos realizar una aproximación funcional discreta por polinomios, no se conocen familias ortogonales, que lo sean también sobre un conjunto de puntos equidistantes y por tanto nos vemos obligados a generarlas de forma específica o de escoger adecuadamente los puntos en los que aproximar (interpolar como sabemos en el

caso extremo de que lleguemos al máximo grado posible en función de la dimensión del espacio establecida por el número de puntos elegidos).

Cuando se trata de puntos equidistantes, tienen especial interés los polinomios de Gram $\{P_{n,m}\}_{n=0}^{m}$. Son la versión discreta de los polinomios de Legendre, es decir, familias de polinomios ortogonales en los puntos $x_k = -1 + \frac{2k}{m}$, $k = 0, 1, \ldots, m$ con el producto escalar

$$\langle f, g \rangle = \sum_{k=0}^{m} f(x_k) g(x_k) \tag{11.8}$$

Esta claro que lo más que podemos tener son $m+1$ polinomios (de grados 0 a m), y para evitar que su norma crezca de forma incontrolada, el grado de libertad de que se dispone se utiliza para imponer la norma 1. La relación de tres términos que nos permite calcular esta familias es la siguiente (demuéstrelo el lector):

$$P_{n+1,m}(x) = \alpha_{n,m} x P_{n,m}(x) - \gamma_{n,m} P_{n-1,m}(x), \qquad n = 0, 1, \ldots, m-1$$

donde

$$\alpha_{n,m} = \frac{m}{n+1} \left(\frac{4(n+1)^2 - 1}{(m+1)^2 - (n+1)^2} \right)^{\frac{1}{2}}, \qquad n = 0, 1, \ldots, m-1$$

y

$$\gamma_{n,m} = \frac{\alpha_{n,m}}{\alpha_{n-1,m}}, \qquad n = 1, \ldots, m-1$$

y tomando como condiciones iniciales

$$P_{0,m}(x) = \frac{1}{\sqrt{(m+1)}} \quad \text{y} \quad P_{-1,m}(x) = 0$$

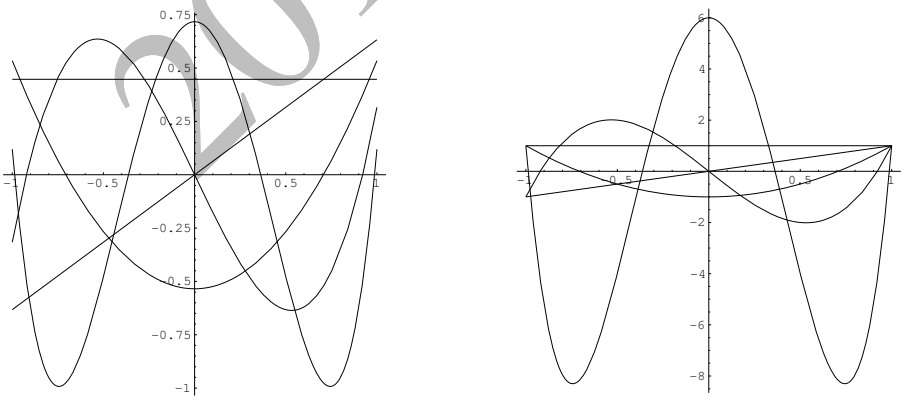


Figura 11.5: Polinomios de Gram hasta grado 4 en 5 puntos

Los polinomios dependen evidentemente de m y sólo tienen cierta similitud con los polinomios de Legendre cuando n es inferior a la raíz cuadrada de m. Cuando supera

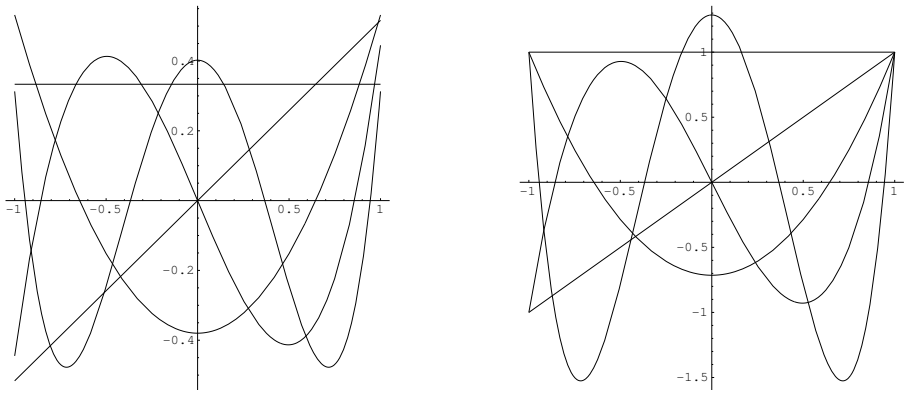


Figura 11.6: Polinomios de Gram hasta grado 4 en 9 puntos

claramente dicho valor, los polinomios presentan enormes oscilaciones entre los puntos de la red y se obtienen aproximantes muy pobres aún cuando aumentemos el número de puntos. De hecho, en una aproximación polinomial discreta en puntos *equidistantes* nunca debería tomarse n mayor que $2\sqrt{m}$. Las figuras 11.5 y 11.6 ilustran claramente estos hechos, al compararlas con los polinomios de Legendre.

Estos fenómenos limitan bastante el uso de estos polinomios. Dado que el costo de su cálculo es proporcional al número de puntos, se tiende a reducir este con el riesgo de que no obtengamos una aproximación satisfactoria, puesto que a partir de un determinado grado del polinomio, hemos visto que lejos de mejorar se degrada. Si decidimos aumentar entonces el número de puntos, nos vemos obligados a recalcular de nuevo toda la familia, sin poder aprovechar ninguno de los resultados obtenidos.

11.3.1 Ejemplo. Consideramos el conjunto de cinco puntos equidistantes en el intervalo. Es decir los puntos de abscisas $\{-1, -.5, 0, .5, 1\}$. Pese a que existirá un polinomio de grado cuatro que interpole a la función en estos cinco puntos, ni siguiera el polinomio de Legendre de grado cinco verifica esta propiedad, lo cual evidencia que el aproximante óptimo continuo, no lo es discreto. O vuelto por pasiva, el mejor aproximante discreto para un conjunto finito de puntos no es el mejor aproximante continuo, y por tanto cuando tratemos de resolver este problema, no basta con discretizar por muchos puntos que tomemos.

La siguiente tabla ilustra estos hechos, utilizando la función exponencial. En ella aparecen además de las abscisas de los puntos y el valor de la función en ellos, las evaluaciones de los polinomios mejores aproximantes continuos hasta el grado 5. En las últimas líneas aparecen el error en la norma del supremo y el error cuadrático en el caso continuo. A continuación, los mismos valores sobre el conjunto discreto de puntos, y finalmente el error cuadrático dividido por el número de puntos para comparar de forma normalizada cuando tengamos varias discretizaciones. La igualdad de los errores en la norma del supremo no debe engañarnos, se trata de una particularidad de esta función que toma los errores máximos en el punto 1.

```
x     exp(x)

-1   0.367879  1.17520  0.07156  0.42937  0.35892  0.36888  0.36778
-.5  0.606530  1.17520  0.62338  0.57865  0.60947  0.60659  0.60650
0    1.        1.17520  1.17520  0.99629  0.99629  1.00003  1.00003
.5   1.648721  1.17520  1.72702  1.68229  1.65146  1.64858  1.64868
1    2.718281  1.17520  2.27883  2.63665  2.70710  2.71707  2.71817

norm0          1.54308  0.43944  0.08162  0.01117  0.00120  0.00010
norm2          0.92987  0.22946  0.03795  0.00472  0.00047  0.00003

norm0          1.54308  0.43944  0.08162  0.01117  0.00120  0.00010
norm2          3.61116  0.31802  0.01236  0.00023  2.49E-6  2.31E-8
norm2          0.72223  0.06360  0.00247  0.00004  4.99E-7  4.62E-9
```

A continuación comprobamos la no ortogonalidad de los polinomios de Legendre sobre este conjunto discreto de puntos. Primero escribimos la matriz que tiene como columnas los valores de los mismos sobre la red (un buen ejercicio sería comprobarlo). Los cinco primeros de estos vectores constituirán, si son independientes, una base del espacio en el que vamos a aproximar (o interpolar en el caso límite)

```
-1    1.    -1.      1.      -1.       1.          -1.
-.5   1.    -0.5    -0.125    0.4375  -0.2890625  -0.08984375
0     1.     0      -0.5      0        0.375       0
.5    1.     0.5    -0.125   -0.4375  -0.2890625   0.08984375
1     1.     1.      1.       1.       1.          1.
```

El producto de la transpuesta de esta matriz por ella misma, nos dara la matriz de Gram (debidamente ampliada en este caso)

```
5.        0.        1.25      0.        1.796875  0.
0.        2.5       0.        1.5625    0.        2.089843
1.25      0.        2.28125   0.        1.884765  0.
0.        1.5625    0.        2.382812  0.        1.921386
1.796875  0.        1.884765  0.        2.307739  0.
0.        2.089843  0.        1.921386  0.        2.016143
```

Como puede observarse estos vectores no son ortogonales, y por supuesto tampoco son independientes al superar la dimensión del espacio. Si lo son en cambio los cinco primeros, y los menores de esta matriz nos pueden servir tan bien como los de cualquier otra para ir calculando los mejores aproximantes discretos (la matriz es por supuesto mejor condicionada que la de Hilbert).

Pero obviamente, el mejor camino es utilizar los polinomios ortogonales sobre este conjunto discreto de puntos. Es decir los correspondientes polinomios de Gram. Los cinco primeros, normalizados tal y como se ven en la parte izquierda de la figura 11.5 son los siguientes:

$$
\begin{array}{lllll}
0.447213 & & & & \\
 & 0.632455\,x & & & \\
-0.534522 & & +1.069044\,x^2 & & \\
 & -1.791957\,x & & +2.108185\,x^3 & \\
0.717137 & & -6.175347\,x^2 & & +5.577733\,x^4
\end{array}
$$

Los vectores de la base, son los valores de estos polinomios en la red. Sus valores no tiene en realidad gran importancia, pues la matriz de Gram es en este caso la unidad (¿por que?), pero si se utilizan para la generación de los términos independientes que serán a su vez los coeficientes del desarrollo.

```
2.835966197742235
1.816094601021502
0.512358671754581
0.084123278820002
0.007787338454465
```

En consecuencia, la sucesión de aproximantes óptimos queda de la siguiente manera.

$$
\begin{array}{llllll}
1.268282 & & & & & \\
1.268382 & +1.148599\,x & & & & \\
0.994415 & +1.148599\,x & +0.547734\,x^2 & & & \\
0.994415 & +0.997853\,x & +0.547734\,x^2 & +0.177347\,x^3 & & \\
.999999 & +0.997853\,x & +0.499644\,x^2 & +0.177347\,x^3 & +0.0434356\,x^4
\end{array}
$$

Es de sobra conocido que el de grado cuatro ya interpola, pero merecen un análisis los valores de estos polinomios en la red, los errores discretos y continuos y compararlos con los de los polinomios de Legendre. La tabla siguiente esta contruída de igual forma que la anterior, pero ahora figuran primero las tres líneas de errores discreto y después las dos de continuos

```
x exp(x)

-1      0.367879    1.268282  0.119683  0.393550  0.366948  0.367879
-.5     0.606530    1.268282  0.693983  0.557049  0.610253  0.606530
0       1.          1.268282  1.268282  0.994415  0.994415  1.
.5      1.648721    1.268282  1.842582  1.705648  1.652444  1.648721
1       2.718281    1.268282  2.416881  2.690748  2.717351  2.718281

norm0               1.449999  0.301400  0.056927  0.005584  8.88E-15
norm2               3.567848  0.269648  0.007137  0.000060  1.70E-30
norm2               0.713569  0.053929  0.001427  0.000012  3.40E-31
```

```
norm0              1.449999  0.301400  0.058822  0.008730  0.001123
norm2              0.939144  0.267074  0.053067  0.007113  0.000744
```

Los resultados de haber discretizado puede considerarse como muy satisfactorios si comparamos detenidamente los errores, y nos hemos evitado la integración. Sin embargo, como ya hemos dicho, un intento de mejora en la aproximación, nos llevaria a recalcular todos los polinomios en un nuevo conjunto de puntos y a la obtención de unos nuevos coeficientes, sin que se pueda aprovechar tampoco nada del trabajo realizado.

11.4 Polinomios de Chebyshev

Una alternativa a los polinomios de Gram, y que permite soslayar esta dificultad, nos la proporcionan una vez más los polinomios de Chebyshev. En efecto, es posible encontrar una rejilla de puntos para la cuál estos polinomios son ortogonales en la mayor cantidad que dicho conjunto de abscisas permite. Además estos puntos son precisamente ceros de otro polinomio de Chebyshev, lo que hace innecesario decir que no son equiespaciados.

Concretamente, si consideramos las raíces del N-ésimo polinomio $T_N(x)$, resulta que los polinomios $T_0, T_1, \ldots, T_{N-1}$ son ortogonales respecto del producto escalar ordinario (11.3).

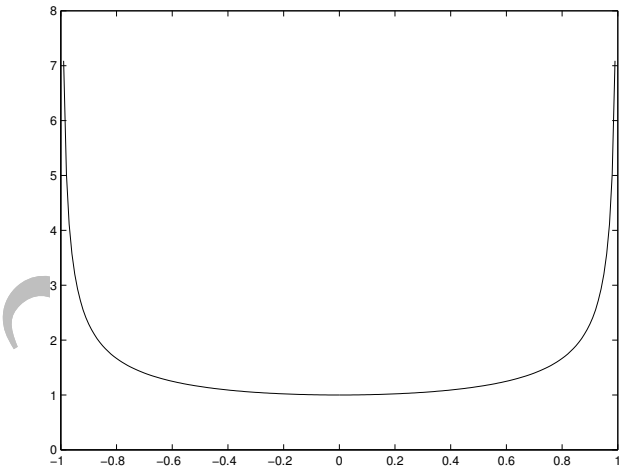


Figura 11.7: Función peso para los polinomios de Chebyshev

Para evitar confusiones, es preciso recordar (apartado 10.3.1) que dichos polinomios son ortogonales respecto del producto escalar continuo ponderado con

$$w(x) = \frac{1}{\sqrt{1-x^2}}$$

cuya representación gráfica evidencia una mayor participación de los extremos en el producto escalar. En el caso discreto se ve compensada por la desigual distribución de los nodos, más abundantes en los extremos del intervalo.

La demostración resulta sencilla, teniendo en cuenta que los ceros del citado polinomio son los puntos

$$x_m = \cos\frac{(2m-1)\pi}{2N}, \qquad m = 1, 2, \ldots, N.$$

es decir, cosenos de una familia de puntos equidistantes entre 0 y π como se evidencia en la figura 4.3. En consecuencia, y teniendo en cuenta la definición de los polinomios de Chebyshev, resulta que para $0 \leq j, k < N$

$$\begin{aligned}\sum_{m=1}^{N} T_j(x_m)T_k(x_m) &= \sum_{m=1}^{N} T_j(\cos\frac{(2m-1)\pi}{2N})T_k(\cos\frac{(2m-1)\pi}{2N}) \\ &= \sum_{m=1}^{N} \cos j\frac{(2m-1)\pi}{2N}\cos k\frac{(2m-1)\pi}{2N} \\ &= \begin{cases} 0 & \text{si } j \neq k \\ \frac{N}{2} & \text{si } j = k \neq 0 \\ N & \text{si } j = k = 0 \end{cases} \end{aligned} \tag{11.9}$$

como podrá comprobar el lector sin dificultad utilizando ligeras variantes de las técnicas utilizadas en la sección 11.2 (véase el problema 11.5.4).

Así pues, partiendo de una cantidad suficientemente alto de puntos en la rejilla (es decir ceros de un polinomio de grado bastante alto, cuya expresión se conoce sin recurrir a ningún cálculo), y teniendo en cuenta que no tenemos que construir los polinomios, pues son los mismos continuos de Chebyshev, fáciles de generar por su relación recursiva, podemos ir obteniendo aproximaciones funcionales discretas tan buenas como deseemos con solo añadir nuevos términos al desarrollo, pero sin tener que recalcular los términos anteriores.

Si además de todo lo dicho, consideramos las buenas condiciones de aproximación uniforme de los desarrollos en polinomios de Chebyshev y su capacidad de *aliasing* mencionada en el apartado 11.1, no es extraño que en muchas ocasiones sea la elección más acertada y que esta sea la familia triangular de polinomios más importante.

11.4.1 Ejemplo. En base a la misma función del ejemplo 11.3.1 vamos a tratar de aclarar los conceptos mencionados en el apartado anterior.

Los polinomios de Chebyshev en el intervalo $[-1, 1]$ hasta el orden 5 son

$$\begin{array}{ccccccccc} 1 & & & & & & & & \\ & x & & & & & & & \\ -1 & & +2x^2 & & & & & & \\ & -3x & & +4x^3 & & & & & \\ 1 & & -8x^2 & & +8x^4 & & & & \\ & 5x & & -20x^3 & & +16x^5 & & & \end{array}$$

y los cuadrados de sus normas respectivas

$$\left\{\pi, \frac{\pi}{2}, \frac{\pi}{2}, \frac{\pi}{2}, \frac{\pi}{2}, \frac{\pi}{2}\right\}$$

La sucesión de aproximantes óptimos resultantes al resolver el correspondiente sistema normal es, para la función $y = e^x$:

$$
\begin{array}{lllll}
1.266065 & & & & \\
1.266065 & +1.130318\,x & & & \\
0.994570 & +1.130318\,x & +0.542990\,x^2 & & \\
0.994570 & +0.997307\,x & +0.542990\,x^2 & +0.177347\,x^3 & \\
1.000044 & +0.997307\,x & +0.499196\,x^2 & +0.177347\,x^3 & +0.0437939\,x^4 \\
1.000044 & +1.000022\,x & +0.499196\,x^2 & +0.166488\,x^3 & \\
 & & & +0.0437939\,x^4 & +0.00868682\,x^5
\end{array}
$$

Sus coeficientes son, lógicamente, parecidos a los que resultaban de los desarrollos de Gram (un poco menos a los de Legendre, véase el ejercicio 11.5.3), lo que ya permite pensar en una buena aproximación discreta cuando los puntos se elijan adecuadamente.

Las raices del polinomio T_5, y los valores de los T_0, T_1, T_2, T_3, T_4 en estos puntos, aprecen en la tabla siguiente como columnas.

```
x

-0.95105      1.  -0.95105   0.80901  -0.58778   0.30901
-0.58778      1.  -0.58778  -0.30901   0.95105  -0.80901
 0.           1.   0.        -1.       0.         1.
 0.58778      1.   0.58778  -0.30901  -0.95105  -0.80901
 0.95105      1.   0.95105   0.80901   0.58778   0.30901
```

Al premultiplicar esta matriz de valores por su transpuesta, obtenemos la prueba patente de la ortogonalidad discreta (y euclidea!!!). La diagonal nos informa también de cuál va a ser la norma discreta.

```
5.        0        0        0        0
0         2.5      0        0        0
0         0        2.4999   0        0
0         0        0        2.4999   0
0         0        0        0        2.5
```

Si ahora nos planteamos hacer un ajuste discreto en estos puntos, calculando aproximantes hasta llegar a la interpolación, obtendremos estos resultados de forma muy económica. Los términos independientes del sistema, y los consiguientes coeficientes de Fourier, tan simplemente relacionados con ellos, aparecen a continuación

```
Independientes       Fourier              Factor

6.330329386007094    1.266065877201419    5
2.825795492308047    1.130318196923219    2.5
0.678737850801393    0.271495140320557    2.5
0.110834128530406    0.044333651412162    2.5
0.013573157797844    0.005429263119137    2.5
```

Los polinomios de ajuste, obtenidos sumando a cada uno el correspondiente polinomio de Chebyshev por el coeficiente calculado, son prácticamente exactos a los continuos, excepto el último (que como sabemos es interpolador) se diferencia claramente

$$
\begin{array}{lllll}
1.266065 & & & & \\
1.266065 & +1.130318\,x & & & \\
0.994570 & +1.130318\,x & +0.542990\,x^2 & & \\
0.994570 & +0.997317\,x & +0.542990\,x^2 & +0.1773346\,x^3 & \\
0.999999 & +0.997317\,x & +0.499556\,x^2 & +0.1773346\,x^3 & +0.0434341\,x^4
\end{array}
$$

El la tabla adjunta, aparecen junto a las abscisas, los valores de la función exponencial en las mismas, y los valores de los sucesivos aproximantes, así como las normas habituales

```
x          exp(x)

-0.95105   0.38633  1.26606  0.19106  0.41071  0.38465  0.38633
-0.58778   0.55555  1.26606  0.60168  0.51778  0.55994  0.55555
 0.        1.       1.26606  1.26606  0.99457  0.99457  0.99999
 0.58778   1.79999  1.26606  1.93045  1.84655  1.80438  1.79999
 0.95105   2.58844  1.26606  2.34106  2.56070  2.58676  2.58844

norma0              1.32237  0.26606  0.04655  0.00542  1.7E-15
norma2              3.38330  0.18926  0.00498  0.00007  4.9E-30
norma2              0.82259  0.19455  0.03158  0.00383  9.9E-16

norma0              1.45221  0.32189  0.05040  0.00606  0.00063
norma2              0.93871  0.26389  0.04384  0.00544  0.00054
```


Sin embargo, ninguno de los polinomios que han aparecido hasta ahora representa el mejor aproximante en la norma del supremo, aunque resulta evidente que el ahora calculado es mucho mejor que el obtenido con la base de Gram (¿por qué?). El cálculo de este aproximante mini-max es un problema no lineal y fuera del contexto de este curso.

11.5 Cuestiones y problemas

11.5.1 Indicar como se transforman las relaciones (11.3) cuando $x_0 \neq 0$. En concreto, estudiar lo que ocurre cuando $x_0 = \frac{\pi}{N}$

11.5.2 Aliasing. Supongamos que f es una función definida en el intervalo $[0, 2\pi]$ y con serie de Fourier convergente a ella, es decir, tal que

$$f(x) = \frac{a_0}{2} + \sum_{k=1}^{\infty}(a_k \cos kx + b_k \operatorname{sen} kx)$$

Si consideramos una red de un número par N de puntos equidistantes en dicho intervalo, obtenemos el interpolante (11.5)

$$\frac{A_0}{2} + \sum_{k=1}^{\frac{N}{2}-1} (A_k \cos kx + B_k \operatorname{sen} kx) + \frac{A_{\frac{N}{2}}}{2} \cos \frac{N}{2} x$$

Es evidente que los coeficientes en mayúsculas no coinciden (salvo casos excepcionales, ¿cuales?) con los otros; es decir, que el último sumatorio NO es una suma parcial del primero. Entre otras razones, las sumas parciales del desarrollo infinito no coinciden con la función en los puntos de la red (en general en ningún punto del intervalo), mientras que el segundo sumatorio sí que lo hace por construcción (¿por qué?). Sin embargo, el hecho de que los elementos de la base sean esencialmente los mismos, hace pensar en una posible relación entre ambas familias de coeficientes. Demostrar que:

$$\begin{aligned} A_j &= a_j + \sum_{l=1}^{\infty} (a_{Nl-j} + a_{Nl+j}) \qquad j = 0, 1, 2, \ldots, \frac{N}{2} \\ B_j &= b_j + \sum_{l=1}^{\infty} (-b_{Nl-j} + b_{Nl+j}) \qquad j = 1, 2, \ldots, \frac{N}{2} - 1 \end{aligned}$$

lo que de hecho significa que cada uno de los términos de la expresión discreta aglutina en su coeficiente los de todos los 'alias' del desarrollo infinito. Escribir la expresión detallada de A_0 y de $A_{\frac{N}{2}}$

11.5.3 Calcular los aproximantes óptimos de grados 0 a 5, en la norma euclídea continua en el intervalo $[-1, 1]$, para la función e^x, expresados como combinación lineal de los polinomios de Legendre, y comparar los coeficientes con los que aparecen en el ejemplo 11.3.1. Repetir después la aproximación discreta con 9 y 17 puntos y ver lo que ocurre con estos coeficientes.

11.5.4 Demostrar las relaciones (11.9)

11.5.5 Demostrar que si consideramos los $N + 1$ puntos

$$x_m = \cos \frac{m\pi}{N}, \qquad m = 0, 1, 2, \ldots, N.$$

los polinomios de Chebyshev de grados 0 a N son ortogonales en la rejilla que forman siempre que se modifique el producto escalar en la forma

$$\langle f, g \rangle = \frac{1}{2} f(x_0) g(x_0) + \sum_{m=1}^{N-1} f(x_m) g(x_m) + \frac{1}{2} f(x_N) g(x_N)$$

y calcular las normas correspondientes.

11.5.6 Con los datos

x_i	-1.00	$-.75$	$-.50$	$-.25$	0	$.25$	$.50$	$.75$	1.00
$\bar{f}_i$	$-.2209$	$.3295$	$.8826$	1.4392	2.0003	2.5645	3.1334	3.7061	4.2836

a) Calcular los coeficientes en las ecuaciones normales para $m = 3, 4, 5$ usando como base las funciones $\phi_j(x) = x^j$.

b) Si en el conjunto de puntos $\{x_i\}$, éstos son simétricos respecto a 0, demuestre que el sistema de ecuaciones normales puede ser dividido en dos conjuntos de ecuaciones con $(m+1)/2$ ecuaciones en cada conjunto si m es impar y, con $m/2$ y $(m/2)+1$ ecuaciones en los dos conjuntos si m es par.

c) Escriba estos dos conjuntos de ecuaciones para los datos del problema.

d) Usando las ecuaciones obtenidas en uno cualquiera de los apartados a) ó c), calcule los coeficientes de las aproximaciones por mínimos cuadrados para el caso que nos ocupa. Use cualquier técnica para resolver las ecuaciones normales, y estime cúal de las aproximaciones a) ó c) es m ás efectiva.

11.5.7 Aplicando desarrollos discretos de Fourier (normalmente FFT), determinar el desarrollo del arco circular

$$y = \sqrt{2\pi x - x^2} \qquad 0 \leq x \leq 2\pi$$

hasta los términos a_6 y b_6 inclusive.

Utilizar diversas cantidades de puntos (que sean potencia de 2), y explicar lo que ocurre con los sucesivos coeficientes.

a) Tratar de encontrar los coeficientes para el desarrollo continuo de Fourier, y comparar con los que van apareciendo en los desarrollos discretos.

b) Hacer la aproximación continua por polinomios de Chebyshev hasta un mismo orden y comparar las gráficas resultantes.

2010-11

CAPÍTULO IV

ÁLGEBRA LINEAL NUMÉRICA

2010-11

Lección 12

Factorizaciones de una matriz

Hemos visto que una forma de encontrar la mejor aproximación por mínimos cuadrados en el caso general, es plantear las ecuaciones normales y resolver el correspondiente sistema lineal, cuya matriz es simétrica y definida positiva. De hecho, esta es la única posibilidad en el caso de la aproximación funcional continua, puesto que ya nos viene dada la matriz de Gram.

Sin embargo, en el caso discreto (bien sea voluntariamente porque hemos decidido muestrear la función, o involuntariamente porque sólo conocemos una tabla de valores), siempre tenemos la oportunidad de plantearlo como un problema lineal de mínimos cuadrados, afrontando la resolución de un sistema lineal incompatible y/o indeterminado. Dejaremos esta tema para el siguiente capítulo y abordamos en este la resolución del sistema de ecuaciones normales.

La mejor forma de resolver un sistema lineal $A\mathbf{x} = \mathbf{b}$ es la utilización de alguna de las variantes del método de Gauss. Es bien conocido que la etapa clave del algoritmo de eliminación gaussiana es la factorización de la matriz A del sistema dado en producto de dos factores L y U, triangulares inferior y superior respectivamente, de suerte que los sistemas $L\mathbf{c} = \mathbf{b}$ y $U\mathbf{x} = \mathbf{c}$ son fácilmente resolubles.

12.1 El caso simétrico

Cuando A es simétrica, como es el caso que nos ocupa en este capítulo, parece razonable esperar que podamos sacar ciertas ventajas de dicha estructura tanto en necesidades de almacenamiento como en costo computacional.

12.1.1 Factorización de Cholesky En concreto, es de esperar una factorización *simétrica* de la forma $A = RR^T$, donde R es triangular inferior. Sin embargo no todas las matrices simétricas admiten una tal factorización, como muestra el siguiente resultado:

12.1.2 TEOREMA

Si A es una matriz real y regular, entonces existe G triangular inferior tal que $A = GG^T$ si y sólo si A es simétrica y definida positiva. Si exigimos elementos diagonales positivos en G, ésta es única (la factorización resultante se denomina de Cholesky).

Demostración. Si $A = GG^T$ entonces claramente A es simétrica puesto que

$$A^T = (GG^T)^T = GG^T = A$$

Además, $\det(G) = \sqrt{\det(A)} \neq 0$ al ser A regular, y G es no singular. Por tanto, si $\mathbf{x} \neq 0$

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T GG^T\mathbf{x} = (G^T\mathbf{x})^T(G^T\mathbf{x}) > 0, \qquad \text{pues} \qquad G^T\mathbf{x} \neq 0$$

que es precisamente la definición de matriz *definida positiva*.

Recíprocamente, sea A simétrica y definida positiva y factoricémosla en la forma $A = LU$. Pongamos $D = \text{diag}(u_{11}, u_{22}, \ldots, u_{nn})$ donde los elementos diagonales de U son estrictamente positivos (¿por qué?). Entonces, dado que $u_{ii} > 0$, $1 \leq i \leq n$, podemos dividir por ellos y escribir

$$A = L \begin{pmatrix} u_{11} & & & \\ & u_{22} & & \\ & & \ddots & \\ & & & u_{nn} \end{pmatrix} \begin{pmatrix} 1 & \frac{u_{21}}{u_{11}} & \ldots & \frac{u_{1n}}{u_{11}} \\ 0 & 1 & \ldots & \frac{u_{n2}}{u_{22}} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{pmatrix} = LDM$$

lo que implica que $A = A^T = M^T(DL^T)$, y como la factorización LU es única y M tiene evidentemente unos en la diagonal, resulta que $M^T = L$ y tenemos $A = LDL^T$. Tomando $G = L\sqrt{D}$, se obtiene la descomposición buscada, entendiendo que se toman las raíces positivas de los elementos de D .

La unicidad se demuestra por reducción al absurdo, pues si hubiese dos de estas factorizaciones $GG^T = HH^T$, premultiplicando esta igualdad por H^{-1} y postmultiplicando por $(G^T)^{-1}$ resulta la siguiente cadena

$$H^{-1}G = H^T(G^T)^{-1} = [G^T(H^T)^{-1}]^{-1} = [(H^{-1}G)^T]^{-1}$$

y en consecuencia $H^{-1}G$ resulta ser una matriz diagonal (¿por qué?) y ortogonal al mismo tiempo, por lo que sus elementos en la diagonal tienen que ser 1 ó -1; y dado que estos elementos son precisamente el cociente entre los respectivos de G y los de H, si todos ellos son positivos como estamos suponiendo, sólo cabe la primera posibilidad y $H^{-1}G = I$ que es lo que se trata de probar. □

12.1.3 Implementación y costo operativo. Veamos en que se traduce la esperada economía de recursos. De momento solo hemos de almacenar una matriz triangular ingerior G cuyos elementos diagonales no son fijos, y esto equivale a $(n^2 + n)/2$ números reales, exactamente la misma necesitada para almacenar la matriz inicial, pero además y pese a que hemos presentado esta factorización de Cholesky como una prolongación de la de Gauss, es posible implementar algoritmos 'in situ' que nos permiten calcular el factor G con un costo computacional muy inferior al requerido por la eliminación gaussiana; aproximadamente la mitad, como veremos y es lógico si pensamos que estamos calculando uno de los factores en vez de los dos de LU.

El siguiente algoritmo *compacto* nos muestra claramente este fenómeno si le comparamos con el correspondiente gaussiano. Escribamos

$$\begin{pmatrix} a_{11} & a_{21} & \dots & a_{n1} \\ a_{21} & a_{22} & \dots & a_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{pmatrix} = \begin{pmatrix} g_{11} & 0 & \dots & 0 \\ g_{21} & g_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \dots & g_{nn} \end{pmatrix} \begin{pmatrix} g_{11} & g_{21} & \dots & g_{n1} \\ 0 & g_{22} & \dots & g_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & g_{nn} \end{pmatrix}$$

tenemos que para $j \leq i$,

$$a_{ij} = \sum_{k=1}^{n} g_{ik} g_{jk} = \sum_{k=1}^{j} g_{ik} g_{jk}$$

luego, actuando con los elementos de G como indeterminados, tenemos que para la primera columna ($j = 1$) los cálculos necesarios son

$$a_{11} = g_{11}^2, \qquad \text{con lo que} \qquad g_{11} = \sqrt{a_{11}} > 0$$

para el elemento diagonal, y para los $n-1$ restantes $i = 2, 3, \dots, n$

$$a_{i1} = g_{i1} g_{11}, \qquad \text{es decir} \qquad g_{i1} = \frac{a_{i1}}{g_{11}}$$

Para la segunda columna tendremos respectivamente para el elemento diagonal y los siguientes

$$a_{22} = g_{21}^2 + g_{22}^2, \qquad \text{con} \qquad g_{22} = \sqrt{a_{22} - g_{21}^2} > 0$$

y, para $i = 3, \dots, n$

$$a_{i1} = g_{i1} g_{21} + g_{i2} g_{22}, \qquad \text{y} \qquad g_{i2} = \frac{a_{i2} - g_{i1} g_{21}}{g_{22}}$$

Continuando con el proceso, cuando ya tengamos $p-1$ columnas de G calculadas, las operaciones a realizar para la p-ésima son

$$\begin{aligned} g_{pp} &= \sqrt{a_{pp} - \sum_{k=1}^{p-1} g_{pk}^2} \quad > 0 \\ g_{ip} &= \frac{a_{ip} - \sum_{k=1}^{p-1} g_{ik} g_{pk}}{g_{pp}}, \quad \text{para} \quad i > p \end{aligned}$$

El conteo ahora es inmediato, además de una raíz cuadrada por cada columna, deberemos de realizar para la p-ésima una cantidad de $n-p$ divisiones y $(p-1)(n-p+1)$ multiplicaciones. Sumando las multiplicaciones/divisiones y considerando un costo acotado para la raíz cuadrada, totalizamos

$$\sum_{k=1}^{n-1} k(n-k) + \sum_{k=1}^{n-1} (n-k) + 0(n) = \frac{n^3 + 3n^2 + 14n}{6} + 0(n) = \frac{n^3}{6} + 0(n^2)$$

frente a los $\frac{n^3}{3}$ del método de Gauss. La posterior solución de los sistemas triangulares $G\mathbf{c} = \mathbf{b}$ y $G\mathbf{x} = \mathbf{c}$, requieren n divisiones más que en el gaussiano (¿por qué?), y totalizan $n^2 + n$ multiplicaciones/divisiones, en cualquier caso de orden inferior a la factorización.

12.2 La factorización LDL^T

Hemos visto que si A es (real) simétrica y definida positiva se puede efectuar la factorización de Cholesky. La demostración constructiva que hemos visto, se basa en realizar la factorización LU sin pivotamiento (para no perder las condiciones de simetría) y efectuar algunas modificaciones en los factores. En matrices definidas positivas se puede llevar a cabo la eliminación gaussiana sin pivotar porque nunca aparecen pivots nulos. Más aún, se puede demostrar que el pivotaje tampoco es necesario para evitar el crecimiento de los errores de redondeo.

Por tanto, el método de Cholesky presenta la ventaja de construir los factores triangulares óptimos de una forma eficiente, pero tiene el inconveniente de tener que ejecutar raíces cuadradas, lo que a veces choca con algunos tipos de aritmética, y además la resolución con nuevos términos independientes una vez factorizada la matriz es más costosa que en el caso LU. Por eso tiene interés la alternativa LDL^T, que recoge las ventajas de ambas factorizaciones y elude algunos de los inconvenientes.

12.2.1 Tratamos de factorizar una matriz A dada en la forma $A = LDL^T$ con L real, triangular inferior con unos en la diagonal y D matriz diagonal. Conocida tal factorización, resolver $A\mathbf{x} = \mathbf{b}$ es hacer una sustitución progresiva con matriz L, resolver un sistema con matriz D (es decir hacer n divisiones) y hacer una sustitución regresiva con matriz L^T. Esto requiere

$$(n^2 - n)/2 + n + (n^2 - n)/2 = n^2$$

multiplicaciones / divisiones, igual que en el caso LU.

Notemos también que almacenar los factores requiere guardar

$$(n^2 - n)/2 + n = (n^2 + n)/2$$

números, pues evidentemente, disponiendo de L, no es preciso conservar L^T. Pero hay que tener en cuenta que los elementos diagonales, no forman un todo con los elementos del triángulo inferior como en el caso de Cholesky.

Por otra parte, no se necesita que la matriz sea definida positiva para asegurar la existencia de una factorización de este tipo. El siguiente resultado, que damos sin demostración nos generaliza el teorema 12.1.2:

12.2.2 TEOREMA

Si A es una matriz regular, todos cuyos menores principales son también regulares. Entonces $A = LDL^T$ (L triangular inferior con unos en la diagonal, D diagonal) si y sólo si A es simétrica.

12.2.3 Habiendo probado la existencia de la factorización $A = LDL^T$ para A simétrica, conviene buscar los factores de forma compacta. De

$$A = LDL^T \tag{12.1}$$

tendremos, al considerar el elemento (1,1), $d_{11} = a_{11}$; al considerar los elementos (1,2), ..., (1,n)

$$a_{12} = d_{11}l_{21}, a_{13} = d_{11}l_{31}, \ldots, a_{1n} = d_{11}l_{n1} \tag{12.2}$$

relaciones que permiten hallar $l_{21}, \ldots, l_{n1}$. Si ahora identificamos los elementos $(2,1), \ldots, (n,1)$ en (12.1), volvemos a caer en las relaciones (12.2). Justamente esta redundancia se produce por la simetría de A y la de la factorización, y nos llevará a que el coste computacional sea la mitad que en forma compacto para LU. Es decir, será similar al método de Cholesky pero no necesitamos extraer raíces.

En efecto, si seguimos identificando, el elemento (2,2) en (12.1) será

$$a_{22} = l_{21}d_{11}l_{21} + d_{22} \qquad \text{y por tanto} \qquad d_{22} = a_{22} - d_{11}l_{21}^2$$

y tenemos d_{22}. Seguidamente identificamos los elementos $(2,3), \ldots, (2,n)$ en (12.1) para obtener $l_{32}, \ldots, l_{n2}$. Una expresión general vendrá dada para $p = 2, 3, \ldots, n$ por

$$\begin{aligned} d_{pp} &= a_{pp} - \sum_{k=1}^{p-1} d_{kk}l_{pk}^2 \\ l_{ip} &= \frac{a_{ip} - \sum_{k=1}^{p-1} d_{kk}l_{ik}l_{pk}}{d_{pp}}, \quad \text{para} \quad i > p \end{aligned}$$

12.3 Cuestiones y problemas

12.3.1 ¿Cuál es el costo operativo del algoritmo compacto de factorización de la sección 12.2? Compárelo con el de hallar los factores L y U.

12.3.2 ¿Es única la factorización LDL^T?

12.3.3 Programe una subrutina que encuentre los factores L, D de la factorización LDL^T de una matriz simétrica definida positiva A. Luego programe una subrutina que, conocidos los factores y el segundo miembro $\mathbf{b}$, halle la solución de $A\mathbf{x} = \mathbf{b}$. Use el menor almacenamiento posible.

12.3.4 Escriba un programa de resolución de sistemas con matriz tridiagonal simétrica definida positiva, tratando de usar tan poco almacenamiento y tan pocas operaciones aritméticas como le sea posible. Cuente el número las operaciones necesarias y compárelo con los correspondientes a (i) matrices tridiagonales no simétricas, (ii) simétricas no tridiagonales y (iii) generales.

12.3.5 Resuelva la cuestión anterior cambiando tridiagonal por pentadiagonal.

12.3.6 *Descomposición en suma de cuadrados de una forma cuadrática.* Con las notaciones de la igualdad (12.1) de la lección, pongamos $\mathbf{y} = L^T\mathbf{x}$, de modo que la j-ésima componente y_j de $\mathbf{y}$ es una combinación lineal de las componentes j a n de $\mathbf{x}$, en la que x_j entra con coeficiente unidad. La forma cuadrática $Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ queda escrita como combinación lineal $d_{11}(y_1)^2 + \ldots + d_{nn}(y_n)^2$ de cuadrados de formas y_j lineales en las componentes de $\mathbf{x}$. El algoritmo del apartado 12.2.3 da, pues, una manera de descomponer una forma cuadrática en suma de cuadrados (de hecho no es otra cosa que el bien conocido algoritmo de Gauss para tal descomposición). Descomponga en suma de cuadrados la forma cuadrática asociada a la matriz

$$\begin{pmatrix} 4 & 2 & -2 \\ 2 & 2 & -3 \\ -2 & -3 & 14 \end{pmatrix}$$

12.3.7 *Unicidad de la factorización de Choleski.* Si en el teorema 12.1.2 no hubiésemos pedido que G tuviese elementos diagonales positivos ¿cuántas factorizaciones G^TG tendría una matriz real simétrica, definida positiva?

12.3.8 Programe una subrutina que encuentre la factorización de Cholesky de una matriz simétrica definida positiva A. Luego programe una subrutina que, conocida la factorización y el segundo miembro $\mathbf{b}$, halle la solución de $A\mathbf{x} = \mathbf{b}$. Use el menor almacenamiento posible.

Lección 13

El problema lineal de mínimos cuadrados

En esta lección, abordaremos todos los casos de sistemas de ecuaciones lineales distintos del clásico en que la matriz A es invertible y tiene solución única. Veremos tanto el caso en que el sistema no tiene solución y tenemos que buscar un vector $\mathbf{x}$ tal que la norma residual $\|A\mathbf{x}-\mathbf{b}\|_2$ sea mínima, es decir, cuya imagen se acerque *lo más posible* al término independiente $\mathbf{b}$, como el caso en que las soluciones sean infinitas y se trata de buscar la de norma más pequeña. También se analizará el supuesto en que se dan ambas circunstancias, y veremos que todos las situaciones anteriores son casos particulares de este problema general.

Dada una matriz $A \in \mathbb{R}^{m\times n}$, es bien conocido que para que el sistema $A\mathbf{x} = \mathbf{b}$, donde $\mathbf{b} \in \mathbb{R}^m$ tenga al menos una solución se requiere que el término independiente $\mathbf{b}$ sea una combinación lineal de los vectores columna de la matriz A, y para que ésta sea única dichos vectores han de ser además independientes; es decir, una base del subespacio imagen. Dando una interpretación geométrica de la matriz como una aplicación lineal de $\mathbb{R}^n$ en $\mathbb{R}^m$, donde las columnas de A son las imágenes de los elementos de la base canónica del espacio de partida expresados en la base canónica del de llegada, la existencia para todo $\mathbf{b} \in \mathbb{R}^m$ exige que la tal aplicación sea suprayectiva, y la unicidad que sea inyectiva, es decir que su núcleo sea el subespacio trivial consistente en el vector nulo.

Así pues, dado que el espacio de llegada tiene dimensión m y que las columnas de A generan el subespacio imagen de la aplicación, el máximo número de ellas que pueden ser independientes es precisamente m, luego solamente es posible la unicidad si $n \leq m$ (es decir, el número de ecuaciones tiene que ser al menos igual que el número de incognitas) y sólo estará garantizada si, además, las columnas de A son vectores independientes. En cuanto a la existencia, solamente se será posible cuando el vector objetivo $\mathbf{b}$ pertenezca a la citada imagen, lo que sólo estará garantizado cuando dicho subespacio llene todo el espacio de llegada; es decir, cuando su dimensión sea m, lo que implica que necesariamente $n \geq m$.

En consecuencia, para tener existencia y unicidad es imprescindible que $m = n$, pero esta condición no es suficiente, como sabemos. Además, el determinante de A debe ser distinto de cero (lo que es equivalente a que el rango de la matriz A sea el máximo, es decir la aplicación sea suprayectiva y, por tanto inyectiva) y entonces ya conocemos los métodos teóricos y numéricos para calcularlos.

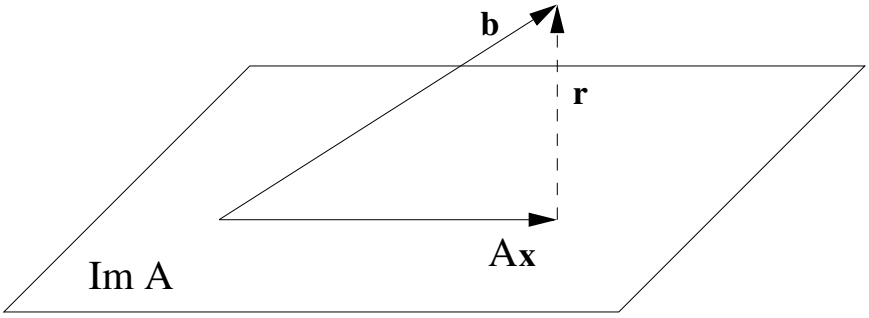


Figura 13.1: $A\mathbf{x}$ es la proyección de $\mathbf{b}$ en el subespacio Im A

13.1 Solución por mínimos cuadrados

Ahora nos ocuparemos de las otras situaciones que se pueden presentar. Primero abordaremos el caso en que la existencia no está garantizada, pero si la unicidad en caso de darse; es decir, el caso en que $m > n$ y las columnas de A son linealmente independientes. Estamos ante el denominado problema lineal de mínimos cuadrados.

Como quiera que, aún en el caso de existir una solución, los métodos vistos anteriormente no nos sirven, pues sólo son válidos para matrices cuadradas, enfocaremos el problema de la existencia desde un punto de vista constructivo. Se trata de desarrollar un método tal que si el sistema tiene solución, el algoritmo la encuentre.

La idea no es nueva, ya en la sección (8.1.2) hablamos de los sistemas sobredeterminados o incompatibles como uno de los problemas típicos de la aproximación. Más tarde, en la sección (9.2.5) vimos cómo se podía aplicar el método de los mínimos cuadrados para resolver este tipo de sistemas, si lo que buscábamos era la aproximación óptima en la norma euclídea. Allí se demuestra que los vectores $\mathbf{x}$ que resuelven el problema de aproximación

$$\min_{\mathbf{x}} \|A\mathbf{x} - \mathbf{b}\|_2 \tag{13.1}$$

son las soluciones de un sistema lineal $n \times n$, con matriz simétrica $A^T A$. En la lección anterior hemos visto la forma numérica de resolverlo. Además, se ha insistido en que esta solución será única, si y sólo si, los vectores columna de A son independientes, que es la hipótesis bajo la que estamos trabajando en esta sección. Pero, seguimos sin saber si el sistema original $A\mathbf{x} = \mathbf{b}$ tiene solución o no, puesto que en ningún momento del proceso llegamos a saber si el valor mínimo en (13.1) es nulo, con lo que $\mathbf{x}$ sería solución del sistema, o extrictamente positivo, en cuyo caso $A\mathbf{x}$ no sería igual a $\mathbf{b}$, sino simplemente su mejor aproximación en el subespacio Im A (en norma euclídea). La figura 13.1 ilustra gráficamente el hecho, ya conocido, de que esta mejor aproximación es la proyección ortogonal de $\mathbf{b}$ en el subespacio imagen.

En ésta situación, y dado que la norma euclídea es invariante por transformaciones ortogonales (¿por qué?), resulta que el problema

$$\min_{\mathbf{x}} \|(Q^T A)\mathbf{x} - (Q^T \mathbf{b})\|_2$$

tiene la misma solución que el primitivo, siempre que Q sea una matriz ortogonal $m \times m$.

Se trata pues de encontrar la matriz ortogonal más conveniente, en el sentido de que sea fácil de calcular y nos deje al mismo tiempo un sistema sencillo de resolver en el nuevo problema de mínimos cuadrados. Por ejemplo, una situación muy favorable se presenta cuando $Q^T A$ resulta ser *triangular* en un sentido amplio, pues de hecho puede ser una matriz no cuadrada. Concretamente, supongamos que hemos computado una matriz ortogonal Q tal que

$$Q^T A = R = \begin{pmatrix} R_1 \\ 0 \end{pmatrix} \begin{matrix} n \\ m-n \end{matrix} \qquad \text{y} \qquad Q^T \mathbf{b} = \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \begin{matrix} n \\ m-n \end{matrix}$$

donde R_1 es triangular superior (ahora si cuadrada $n \times n$) y regular (su rango es el de A y estamos trabajando bajo el supuesto de independencia de sus columnas). Entonces se verifica (¿por qué?) que

$$\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|(Q^T A)\mathbf{x} - (Q^T \mathbf{b})\|_2^2 = \|R_1\mathbf{x} - \mathbf{c}\|_2^2 + \|\mathbf{d}\|_2^2$$

y como el último sumando no depende de $\mathbf{x}$, el mínimo de esta expresión se alcanzará para el vector que minimice el otro sumando. Pero el sistema $R_1\mathbf{x} = \mathbf{c}$ tiene solución única, que anula el citado sumando. Basta pues resolver este último sistema triangular para resolver el problema discreto de mínimos cuadrados. Además, sabemos la magnitud del error cometido, pues anulado dicho término, resulta que $\|A\mathbf{x} - \mathbf{b}\|_2^2 = \|\mathbf{d}\|_2^2$.

Dado que $Q^T = Q^{-1}$, resulta que en realidad tenemos una factorización $A = QR$, que es con el nombre con que se conoce este método. Hay distintos caminos para conseguirla y algunas variantes interesantes que veremos a continuación.

13.2 Transformaciones de Householder

Dado un vector no nulo $\mathbf{u} \in \mathbb{R}^m$, la aplicación lineal de matriz

$$P(\mathbf{u}) = I - \frac{2}{\mathbf{u}^T\mathbf{u}}\mathbf{u}\mathbf{u}^T$$

se denomina transformación de Householder. Si tenemos en cuenta que $\mathbf{u}^T\mathbf{u} = \|\mathbf{u}\|_2^2$, resulta que

$$P(\mathbf{u}) = I - 2\frac{\mathbf{u}}{\|\mathbf{u}\|_2}\frac{\mathbf{u}^T}{\|\mathbf{u}\|_2} = P(\frac{\mathbf{u}}{\|\mathbf{u}\|_2})$$

y la transformación depende únicamente de la dirección del vector y no de su módulo, por lo que supondremos en lo que sigue que siempre se verifica $\|\mathbf{u}\|_2 = 1$ y en consecuencia $P(\mathbf{u}) = I - 2\mathbf{u}\mathbf{u}^T$, salvo que se diga explícitamente lo contrario.

Estas transformaciones son también conocidas como los *reflectores* de Householder, debido a que la aplicación de $\mathbb{R}^m$ en sí mismo

$$\mathbf{x} \to P(\mathbf{u})\mathbf{x} = \mathbf{x} - 2(\mathbf{u}^T\mathbf{x})\mathbf{u} \tag{13.2}$$

es de hecho una reflexión respecto del subespacio

$$\mathbf{u}^\perp = \{\mathbf{v} \in \mathbb{R}^m | \mathbf{u}^T\mathbf{v} = 0\}$$

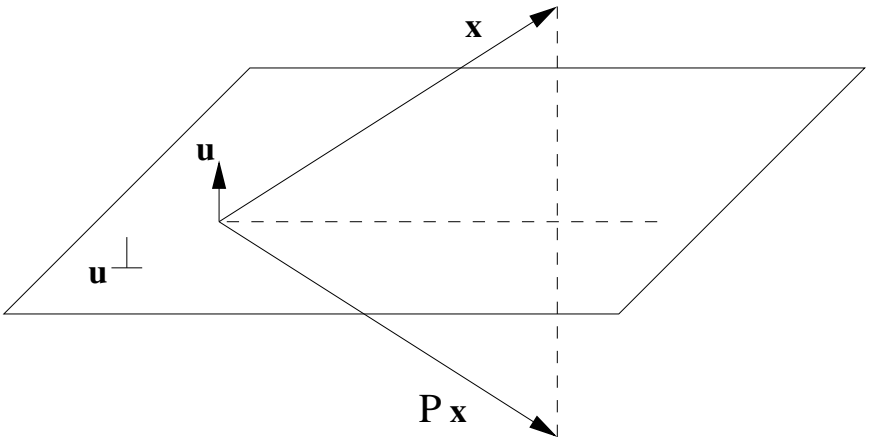


Figura 13.2: $P(\mathbf{u})\mathbf{x}$ como reflejo de $\mathbf{x}$ en el subespacio ortogonal a $\mathbf{u}$

de los vectores ortogonales a $\mathbf{u}$. En efecto, la expresión (13.2) nos dice que la imagen de un vector cualquiera se obtiene restándole el doble de su proyección sobre $\mathbf{u}$. En particular $P(\mathbf{u})\mathbf{u} = -\mathbf{u}$ y $P(\mathbf{u})\mathbf{v} = \mathbf{v}$ para cada vector $\mathbf{v} \in \mathbf{u}^\perp$. La figure 13.2 muestra claramente el papel de *espejo* que juega el subespacio $\mathbf{u}^\perp$ en la transformación de Householder.

De esta interpretación geométrica se derivan unas cuantas propiedades muy interesantes de los *reflectores*. Concretamente $P^2 = I$ por lo que $P^{-1} = P$, y como evidentemente P es simétrica, resulta que es ortogonal y las transformaciones de Householder conservan la norma euclídea. Pero no sólo eso, sino que también se verifica el siguiente resultado de sumo interés para el objetivo que perseguimos.

13.2.1 TEOREMA

Si $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 \neq 0$, resulta que

$$P(\mathbf{a} - \mathbf{b})\mathbf{a} = \mathbf{b}$$

Demostración. Busquemos $\mathbf{u}$ tal que $\mathbf{b} = P(\mathbf{u})\mathbf{a}$, es decir que $\mathbf{b} = \mathbf{a} - 2(\mathbf{u}^T\mathbf{a})\mathbf{u}$, lo que implica que $2(\mathbf{u}^T\mathbf{a})\mathbf{u} = \mathbf{a} - \mathbf{b}$, y esta ha de ser la dirección de cualquier vector cuyo *reflector* transforme $\mathbf{a}$ en $\mathbf{b}$. □

La consecuencia inmediata es que dado un vector cualquiera, no nulo, $\mathbf{x} \in \mathbb{R}^k$, siempre podemos encontrar una matriz ortogonal Q tal que $Q\mathbf{x}$ sea de la forma $\mathbf{y} = (y, 0, 0, \ldots, 0)^T$ con $y = \sqrt{x_1^2 + x_2^2 + \ldots + x_k^2}$. Bastará construir el *reflector* correspondiente al vector $\mathbf{x} - \mathbf{y} = (x_1 - y, x_2, \ldots, x_k)$.

13.2.2 Estamos ya en condiciones de encontrar de forma constructiva la factorización QR de una matriz real $A_{m\times n}$. Primero tomamos la matriz ortogonal Q_1 de dimensión

m y tal que

$$Q_1A = \begin{pmatrix} \times & \times & \times & \dots & \times \\ 0 & \times & \times & \dots & \times \\ 0 & \times & \times & \dots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \times & \times & \dots & \times \end{pmatrix}$$

A continuación tomamos Q_2' de dimensión $m-1$ tal que nos haga nulos los elementos de la segunda columna a partir del tercero, y tomando como Q_2 la matriz $m \times m$ resultante de orlar la anterior con una primera fila y una primera columna de ceros, excepto el elemento (1,1) que será la unidad, tendremos que

$$Q_2Q_1A = \left(\begin{array}{c|ccc} 1 & 0 & \dots & 0 \\ \hline 0 & & & \\ \vdots & & Q_2' & \\ 0 & & & \end{array}\right) Q_1A = \begin{pmatrix} \times & \times & \times & \dots & \times \\ 0 & \times & \times & \dots & \times \\ 0 & 0 & \times & \dots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \times & \dots & \times \end{pmatrix}$$

y así sucesivamente vamos construyendo Q_k que nos aniquila los elementos $k+1, \dots, n$ de la k-ésima columna y nos deja invariantes las $k-1$ primeras, hasta llegar a

$$Q_nQ_{n-1}\cdots Q_2Q_1A = \begin{pmatrix} \times & \times & \times & \dots & \times \\ 0 & \times & \times & \dots & \times \\ 0 & 0 & \times & \dots & \times \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \times \\ 0 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix}$$

que nos proporciona una factorización QR, tomado como $Q = Q_1Q_2\cdots Q_{n-1}Q_n$, y tras calcular $Q^T\mathbf{b}$ podemos proceder a la resolución del problema lineal de mínimos cuadrados. Si exigimos que los elementos *diagonales* de R sean positivos (para lo que basta tomar la raíz cuadrada positiva en la construcción de los sucesivos vectores $\mathbf{y}$ del apartado anterior), dicha factorización es única.

13.3 Ortonormalización de Gram-Schmidt

Una vez obtenida la factorización $A_{m\times n} = Q_{m\times m}R_{m\times n}$, observamos que si denominamos por $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_n$ a las columnas de A y por $\mathbf{q}_1, \mathbf{q}_2, \dots, \mathbf{q}_m$ a las de Q ocurre que el

subespacio generado por las k primeras columnas de A coincide con el generado por las k primeras de Q, para $k = 1, 2, \ldots, n$. En efecto, resulta evidente que

$$\begin{aligned} \mathbf{a}_1 &= r_{11}\mathbf{q}_1 && (13.3)\\ \mathbf{a}_2 &= r_{12}\mathbf{q}_1 + r_{22}\mathbf{q}_2 && (13.4)\\ \ldots &= \ldots && (13.5)\\ \mathbf{a}_n &= r_{1n}\mathbf{q}_1 + r_{2n}\mathbf{q}_2 + \ldots + r_{nn}\mathbf{q}_n && (13.6) \end{aligned}$$

A partir de estas igualdades resulta evidente que podemos calcular los vectores unitarios $\mathbf{q}_i, i = 1, 2, \ldots, n$ y la matriz de paso R_1 sin utilizar los reflectores de Householer. Si tenemos en cuenta que con frecuencia m es mucho más grande que n (por ejemplo en los polinomios de Gram) resulta interesante implementar métodos eficientes de realizar estos cálculos.

El más conocido es el algoritmo denominado *clásico* de Gram-Schmidt, que se base en una ortonormalización secuencial de las columnas de A, en el sentido de que el subespacio generado por $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_k$ coincida con el generado por $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k$ para $k = 1, 2, \ldots, n$ y sean vectores ortonormales. Veamos su implementación.

Evidentemente $\mathbf{q}_1$ será simplemente $\mathbf{a}_1$ dividido por su norma euclídea, y por tanto $r_{11} = \|\mathbf{a}_1\|_2$, $\mathbf{q}_2$ será el normalizado de $\mathbf{a}_2$ menos su proyección ortogonal sobre $\mathbf{q}_1$, que como sabemos es el mejor aproximante a $\mathbf{a}_2$ en el subespacio generado por $\mathbf{q}_1$, y nos basta calcular el coeficiente de Fourier, es decir

$$\mathbf{q'}_2 = \mathbf{a}_2 - \langle \mathbf{a}_2, \mathbf{q}_1 \rangle \mathbf{q}_1$$

y normalizando, tendremos $\mathbf{q}_2 = \frac{\mathbf{q'}_2}{r_{22}}$ con $r_{22} = \|\mathbf{q'}_2\|_2$, $r_{12} = \langle \mathbf{a}_2, \mathbf{q}_1 \rangle$, y así sucesivamente proyectando cada $\mathbf{a}_k$ sobre el subespacio generado por los $k-1$ anteriores $\mathbf{q}$'s (o $\mathbf{a}$'s, pero utilizamos los $\mathbf{q}$'s para aprovechar tanto su ortogonalidad como el hecho de que tengan norma unitaria), vamos obteniendo los sucesivos elementos de

$$r_{jk} = \langle \mathbf{a}_k, \mathbf{q}_j \rangle, \qquad \text{para} \qquad j = 1, 2, \ldots, k-1$$

$$\mathbf{q'}_k = \mathbf{a}_k - \sum_{j=1}^{k-1} r_{jk}\mathbf{q}_j, \qquad r_{kk} = \|\mathbf{q'}_k\|_2, \qquad \mathbf{q}_k = \frac{\mathbf{q'}_k}{r_{kk}}$$

Este algoritmo, cuyo costo es del orden de mn^2 multiplicaciones/divisiones, tiene un comportamiento numérico inestable, sobre todo por la perdida de ortogonalidad de los vectores $\mathbf{q}$ resultantes. Existe una variante más estable denominada método *modificado* de Gram-Schmidt con el mismo coste y un comportamiento algo mejor (véase el ejercicio 13.5.1).

Denominando Q' a la submatriz $m \times n$ generada, que coincide con las n primeras columnas de la Q del método de Householder en el supuesto de que hayamos tomado las raíces positivas en la construcción de los sucesivos reflectores (¿por qué?), podemos escribir una nueva factorización $A = Q'R_1$, que si bien no nos va a permitir conocer de forma inmediata la norma del error, si contiene información suficiente para resolver el problema de mínimos cuadrados, basta resolver el sistema $R_1\mathbf{x} = Q'^T\mathbf{b}$. Además es más eficiente que el método de Householder, que además de una mayor necesidad de almacenamiento requiere al menos el doble de operaciones.

13.4 Pseudo-inversa de una matriz

Es evidente que en el caso de que las columnas de A no sean independientes, no se puede aplicar ninguno de los dos métodos de factorización que hemos explicado (lo que no quiere decir que no exista una tal factorización $A = QR$, pero entonces R tendría el mismo rango que A y no sería posible utilizarla para calcular la solución, al no ser regular). La cuestión está en que la dimensión de la imagen o rango de A (número de columnas independientes de A) es estrictamente menor que la dimensión n del espacio de partida, lo que implica la existencia de un núcleo (conjunto de vectores cuya imagenes el vector nulo) no trivial y, en consecuencia, de infinitas soluciones en el caso de que el sistema sea compatible, o de infinitas mejores aproximaciones en el sentido de los mínimos cuadrados en el caso de que no lo sea.

Es evidente, que si a una solución del sistema le sumamos un vector cualquiera del núcleo obtenemos una nueva solución. Es decir, el conjunto de soluciones es una variedad lineal afín paralela al núcleo y, por tanto, es fácil demostrar que hay un único vector en ella que tenga norma euclídea mínima, el más próximo al vector nulo. Este vector, denominado $\mathbf{x}_{LS}$, es la solución de norma euclídea mínima (en el sentido de mínimos cuadrados en el caso que el sistema sea incompatible) y además su imagen por A es el vector que produce el error residual $\rho_{LS} = \|A\mathbf{x}_{LS} - \mathbf{b}\|_2$ más pequeño (nulo en el caso de que el sistema tenga solución). El método para calcular esta solución de norma (euclídea) mínima por mínimos cuadrados se basa en el cálculo de la denominada pseudo-inversa de la matriz A. Para entender la forma de proceder, pensemos que esta solución es de hecho el único vector solución ortogonal al núcleo de la aplicación, y en consecuencia se puede obtener restando a una cualquiera de las soluciones su proyección sobre el núcleo. Este proceder es en cierta forma dual del utilizado para encontrar la mejor aproximación (en el sentido de los mínimos cuadrados) a un vector del espacio de llegada, que consiste en proyectarlo sobre el subespacio imagen de la aplicación. Estas ideas geométricas, pueden guiarnos en la compresión de los resultados siguientes.

13.4.1 Descomposición en valores singulares. El resultado fundamental, que daremos sin demostración, es la denominada descomposición en valores singulares (SVD) que resulta razonablemente sencillo de entender después de lo escrito anteriormente. Vamos a suponer por sencillez y coherencia con el contenido de la lección que $m \geq n$, pero el siguiente teorema es válido para cualesquiera valores.

TEOREMA

Dada una matriz arbitraria $A \in \mathbb{R}^{m\times n}$, existen matrices ortogonales

$$U = [\mathbf{u}_1, \ldots, \mathbf{u}_m] \in \mathbb{R}^{m\times m} \qquad \text{y} \qquad V = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{R}^{n\times n}$$

tales que

$$U^T A V = \Sigma,$$

siendo $\Sigma \in \mathbb{R}^{m\times n}$ la matriz

$$\Sigma = \begin{bmatrix} \Sigma_0 \\ 0 \end{bmatrix},$$

con $\Sigma_0 = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$, donde $\sigma_1 \geq \sigma_2 \geq \cdots \sigma_n \geq 0$, son los valores singulares de A.

Como se ve, el resultado no exige que todos los valores σ_i sean extrictamente positivos, y de hecho sólo lo serán en número igual al rango de la matriz A, como es fácil de deducir de la ortogonalidad de U y de V, que en consecuencia conservan el rango. El siguiente corolario del teorema, nos ofrece una visión detallada de cada una de las partes implicadas.

13.4.2 Corolario. Si la descomposición en valores singulares de la matriz A es tal que

$$\sigma_1 \geq \sigma_2 \geq \ldots \sigma_r > \sigma_{r+1} = \sigma_{r+2} = \sigma_n = 0,$$

se verifica que:

- el rango de A es r
- el núcleo de A es el subespacio $\langle \mathbf{v}_{r+1}, \ldots, \mathbf{v}_n \rangle$ de dimensión $n - r$.
- la imagen de A es el subespacio $\langle \mathbf{u}_1, \ldots, \mathbf{u}_r \rangle$ de dimensión r.
- la matriz
$$A = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^T = U_r \Sigma_r V_r^T$$
donde
$$U_r = [\mathbf{u}_1, \ldots, \mathbf{u}_r], V_r = [\mathbf{v}_1, \ldots, \mathbf{v}_r] \text{ y } \Sigma_r = \text{ diag}(\sigma_1 \ldots \sigma_r)$$
son matrices de dimensiones $m \times r$, $n \times r$ y $r \times r$ respectivamente, pero todas ellas con rango r,

La demostración se deja como un ejercicio, pero lo más importante es pensar sobre su significado geométrico.

13.4.3 Solución de norma mínima en el sentido de los mínimos cuadrados. En base a estos resultados, y teniendo en cuenta que todas las soluciones al problema de aproximación planteado se pueden obtener sumando a una cualquiera de ellas un vector del núcleo, ahora resulta fácil dar una expresión directa de la mejor solución :

TEOREMA

El vector

$$\mathbf{x}_{LS} = \sum_{i=1}^{r} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i \tag{13.7}$$

es ortogonal al núcleo de A y su imagen es la proyección ortogonal de $\mathbf{b}$ en la imagen de A. Se trata, por tanto, de la solución de norma mínima al problema lineal de mínimos cuadrados.

Demostración. Teniendo en cuenta que para cualquier vector $\mathbf{x} \in \mathbb{R}^n$, tenemos que

$$\begin{aligned} \|A\mathbf{x} - \mathbf{b}\|^2 &= \|(U^T A V)(V^T \mathbf{x}) - U^T \mathbf{b}\|_2^2 = \|\Sigma \mathbf{w} - U^T \mathbf{b}\|_2^2 \\ &= \sum_{i=1}^{r} (\sigma_i w_i - \mathbf{u}_i^T \mathbf{b})^2 + \sum_{i=r+1}^{m} (\mathbf{u}_i^T \mathbf{b})^2, \end{aligned}$$

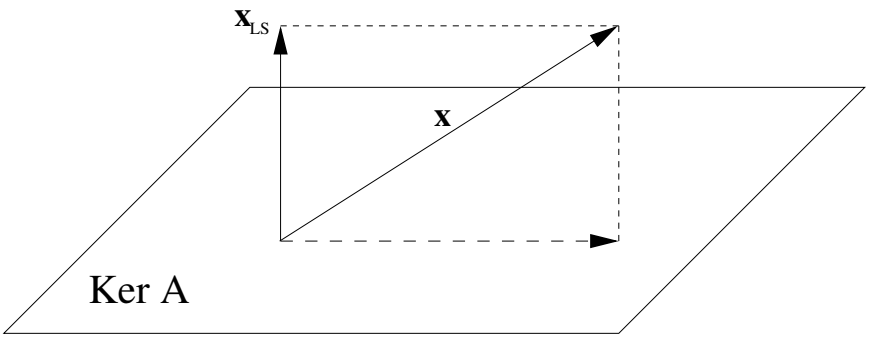


Figura 13.3: Como obtener $\mathbf{x}_{LS}$ a partir de alguna solución por mínimos cuadrados

donde $\mathbf{w} = V^T\mathbf{x} \in \mathbb{R}^n$. Claramente si hacemos $w_i = (\mathbf{u}_i^T\mathbf{b}/\sigma_i)$ para $i = 1, \ldots, r$, el primer sumando de expresión anterior se anula, y como el segundo no depende para nada de $\mathbf{w}$, resulta que cualquier $\mathbf{w}$ con estas r primeras coordenadas es una solución del sistema en el sentido de los mínimos cuadrados, siendo el segundo sumando el cuadrado del error residual ρ_{LS}, que será nulo si, y sólo si, $\mathbf{b}$ pertenece a la imagen de A; es decir si es ortogonal a los vectores $\mathbf{u}_{r+1}, \ldots, \mathbf{u}_m$ (¿por qué?).

Ahora bien, como $\mathbf{x} = V\mathbf{w}$ y tiene la misma norma euclídea por ser V ortogonal, dicha norma será mínima cuando lo sea la de $\mathbf{w}$, vector de dimensión n, del que sólo conocemos las primeras r componentes, lo que es suficiente para garantizar que la imagen $A\mathbf{x}$ sea el vector con menor error residual, con independencia de quienes sean las restantes componentes $w_{r+1}, \ldots, w_n$. Está claro que cuando todas estas componentes sean nulas, y sólo entonces, la norma de $\mathbf{w}$, y en consecuencia la de $\mathbf{x}$, será mínima entre todas las soluciones del sistema en el sentido de los mínimos cuadrados.

Geométricamente esto es trivial, pues la anulación de estas componentes para cualquier vector $\mathbf{w}$, lo único que hace es restar a su imagen por V su proyección sobre el núcleo de A (¿por qué?), y en consecuencia obtener un vector $\mathbf{x}$ ortogonal al mismo sin cambiar su imagen por A, y que obviamente es de mínima norma entre todos los que comparten esta imagen común. □

Como resultado de todo lo anterior, tenemos que la matriz $A^+ = V\Sigma^+U^T$, donde

$$\Sigma_+ = \mathrm{diag}(\frac{1}{\sigma_1}, \frac{1}{\sigma_2}, \ldots, \frac{1}{\sigma_r}, 0, \ldots, 0) \in \mathbb{R}^{n\times m}$$

es tal que

$$\mathbf{x}_{LS} = A^+\mathbf{b} \qquad \text{y} \qquad \rho_{LS} = \|(I - AA^+)\mathbf{b}\|_2$$

como se puede fácilmente demostrar. Por eso en la literatura se denomina a A^+ la matriz *pseudo-inversa* de A.

Ni que decir tiene que este método para calcular la solución de norma mínima en el sentido de los mínimos cuadrados cuando un sistema carece de existencia y unicidad, engloba como casos particulares cualquier situación más favorable; es decir, si el sistema es incompatible pero de rango máximo, nos permite calcular la solución única en el sentido de los mínimos cuadrados, si el compatible pero con infinitas soluciones, encuentra aquella

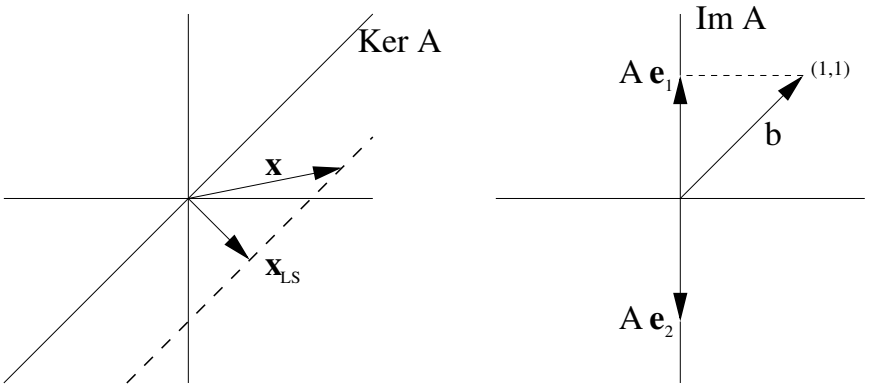


Figura 13.4: Ilustración gráfica para el caso de una matriz 2×2 de rango 1

que tiene norma mínima y, finalmente, si es determinado nos proporciona un nuevo método de resolverlo y de hecho la *pseudo-inversa* coincide con A^{-1}. Sería interesante estudiar su costo computacional.

13.4.4 Cálculos efectivo de la pseudo-inversa de una matriz Consideremos un ejemplo 2×2 ilustrado geométricamente en la figura 13.4. Se trata del resolver el sistema $A\mathbf{x} = \mathbf{b}$

$$\begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$ entendido como una aplicación lineal $\mathbb{R}^2 \to \mathbb{R}^2$

que es evidente que no tiene solución. Por tanto, tenemos que resolverlo en el sentido de los mínimos cuadrados tomando como término independiente el vector $(0\ 1)^T$ que es la proyección del $\mathbf{b}$ original sobre el subespacio Im A. Tenemos pues el sistema

$$\begin{pmatrix} 0 & 0 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

que ahora tiene infinitas soluciones, para todos los valores de x e y tales que $x - y = 1$, es decir la recta $y = x - 1$.

Para todos estos vectores $\mathbf{v} = (x, y)$, claramente se verifica que la norma residual $\|A\mathbf{v} - \mathbf{b}\|_2 = 1$, y es también evidente que el de menor norma euclídea de todos ellos, por ser el más cercano al origen, es $\mathbf{x}_{LS} = (1/2, -1/2)$, cuya norma es $\|\mathbf{x}_{LS}\| = \sqrt{2}/2$.

Se trata ahora de comprobar que efectivamente $\mathbf{x}_{LS} = A^+\mathbf{b}$, tras calcular la matriz pseudo-inversa de A, $A^+ = V\Sigma^+U^T$. Está claro, tras calcular AA^T o A^TA y sus autovalores, que deben de coincidir en este caso (¿por qué?), que los valores singulares de A son $\sqrt{2}$ y 0, por lo que $\Sigma^+ = \text{diag}(1/\sqrt{2}, 0)$. Por otra parte, será fácil para el lector calcular, con el apoyo del ejercicio 13.5.3, las matrices V y U, y verificar que

$$A^+ = V\Sigma^+U^T = \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} =$$

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

y que efectivamente

$$A^+\mathbf{b} = \begin{pmatrix} 0 & \frac{1}{2} \\ 0 & \frac{-1}{2} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{-1}{2} \end{pmatrix} = \mathbf{x}_{LS}$$

13.5 Cuestiones y problemas

13.5.1 Método *modificado* de Gram-Schmidt. La variante consiste en resolver el esquema compacto (13.6) por columnas en vez de por filas como hace el clásico. Así, calcularemos primero la norma r_{11} y hacemos $\mathbf{a}_1 = \frac{\mathbf{a}_1}{r_{11}}$; es decir, sobreescribimos el $\mathbf{q}_1$ *clásico* sobre la primera columna de A. Inmediatamente calculamos toda la primera columna de r's (que son la primera fila de la matriz R_1)

$$r_{1j} = \langle \mathbf{a}_j, \mathbf{a}_1 \rangle = \sum_{i=1}^{m} a_{ij} a_{i1}$$

y modificamos todos los vectores $\mathbf{a}$ a partir del segundo

$$\mathbf{a}_j = \mathbf{a}_j - r_{1j}\mathbf{a}_1, \qquad j = 2, 3, \ldots, n$$

que como vemos se sobrescriben sobre los anteriores. Terminar el algoritmo y demostrar que efectivamente es equivalente al clásico en cuanto al costo de operaciones y requiere menos memoria. Además, como ya dijimos es más estable numéricamente.

13.5.2 Demostrar el teorema de descomposición en valores singulares.

13.5.3 Demostrar que con las notaciones del teorema de la sección 13.4.1, se verifica que la matriz $n \times n$

$$V^T(A^TA)V = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{pmatrix} \quad \text{con } \sigma_i > 0, i = 1, \ldots, r \text{ y } \sigma_i = 0, i = r+1, \ldots, n$$

y que la matriz $m \times m$

$$U^T(AA^T)U = \begin{pmatrix} \sigma_1^2 & & & & & \\ & \ddots & & & & \\ & & \sigma_n^2 & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

con las mismas condiciones para los σ_i, que son los valores singulares.

13.5.4 Demostrar que los valores singulares son las raíces cuadradas positivas de los autovalores (necesariamente no negativos) de la matriz $A^T A$, que es simétrica y definida (o semi-definida si el rango r de A es menor que $\min\{m, n\}$) positiva.

13.5.5 Demostrar que si $(\sigma_1, \sigma_2, \ldots, \sigma_r)$ son los valores singulares no nulos de una matriz $A \in \mathbb{R}^{m \times n}$, su norma de Frobenius es $\|A\|_F = \sigma_1^2 + \ldots + \sigma_r^2$ y que su norma asociada a la vectorial euclídea $\|A\|_2 = \sigma_1$

13.5.6 Demostrar que en las circunstancias de la sección 13.4, se cumple que $\mathbf{x}_{LS} = A^+\mathbf{b}$ y $\rho_{LS} = \|(I - AA^+)\mathbf{b}\|_2$

13.5.7 Demostrar que la pseudo-inversa de una matriz es la única solución del problema de mínimos en norma de Frobenius

$$\min_{X \in \mathbb{R}^{n \times m}} \|AX - I_m\|_F$$