UNED

Introducción al cálculo numérico

Carlos Moreno González



Introducción al Cálculo Numérico

CARLOS MORENO GONZÁLEZ

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

INTRODUCCIÓN AL CÁLCULO NUMÉRICO

Quedan rigurosamente prohibidas, sin la autorización escrita de los titulares del Copyright, bajo las sanciones establecidas en las leyes, la reproducción total o parcial de esta obra por cualquier medio o procedimiento, comprendidos la reprografía y el tratamiento informático, y la distribución de ejemplares de ella mediante alquiler o préstamo públicos.

© *Universidad Nacional de Educación a Distancia*
Madrid 2021

www.uned.es/publicaciones

© *Carlos Moreno González*

ISBN electrónico: 978-84-362-7799-9

Edición digital: mayo de 2021

Prólogo

Modern numerical analysis can be credibly said to begin with the 1947 paper by John von Neumann and Herman Goldstine, "Numerical Inverting of Matrices of High Order"(Bulletin of the AMS, Nov. 1947). It is one of the first papers to study rounding error and include discussion of what today is called scientific computing. Although numerical analysis has a longer and richer history, "modern"numerical analysis, as used here, is characterized by the synergy of the programmable electronic computer, mathematical analysis, and the opportunity and need to solve large and complex problems in applications. The need for advances in applications, such as ballistics prediction, neutron transport, and nonsteady, multidimensional fluid dynamics drove the development of the computer and depended strongly on advances in numerical analysis and mathematical modeling.
SIAM History of Numerical Analysis and Scientific Computing Project 2007.

Las matemáticas suministran un lenguaje que se utiliza en las ciencias e ingenierías para describir con rigor modelos que pretenden representar fenómenos reales. En ellos, se estudian aspectos cualitativos de esos fenómenos y se plantean problemas cuya resolución conduce en último término al desarrollo de tecnologías aplicadas.

Las dificultades que surgen en el estudio de estos problemas, han estimulado el desarrollo de las matemáticas generando la necesidad de teorías abstractas y contribuyendo de este modo a construir el pensamiento matemático actual. Pero, junto a cada nuevo concepto que permite profundizar en la estructura del conocimiento matemático, surge la necesidad de desarrollar técnicas que permitan utilizar esos conceptos de un modo cuantitativo preciso.

En las matemáticas de la Edad Antigua, los problemas eran de naturaleza aritmética o geométrica. Entonces, una vez comprendido el concepto

de longitud, área y volumen, los métodos *exhaustivos* permitían resolverlos utilizando unidades de medida cada vez más fraccionadas. Posteriormente, el Cálculo Diferencial de Leibniz permitió precisar muchos de estos conceptos y simultáneamente desarrolló los métodos analíticos que permitieron su evaluación exacta. No obstante, las ecuaciones que aparecen en los modelos comienzan a ser más complicadas, ya que junto a las ecuaciones numéricas (es decir, aquellas cuyas incógnitas son valores numéricos), empiezan a considerarse ecuaciones cuyas incógnitas son funciones y en ellas aparecen operadores distintos de los algebraicos tales como los integrales y diferenciales.

En la primera parte del siglo XX es ya muy manifiesta la insuficiencia del cálculo simbólico para resolver ecuaciones de la complejidad con la que se presentaban en los modelos de las ciencias e ingenierías. Durante la primera mitad de este siglo permanece un sentimiento de incapacidad por la imposibilidad material de llevar a cabo los cálculos por procedimientos numéricos aproximados. Este sentimiento se mitiga en la segunda mitad del siglo con el desarrollo exponencial de las capacidades de calcular automáticas. Este hecho condiciona de un modo radical el modo de calcular y las técnicas comienzan a adaptarse a este nuevo medio de cálculo.

Los medios de cálculo electrónico permiten realizar una elevada cantidad de operaciones algebraicas en periodos muy reducidos de tiempo. En ese momento surge la necesidad de diseñar algoritmos capaces de aproximar las soluciones a los problemas, mediante operaciones elementales. Los sistemas de ecuaciones numéricas lineales, aun cuando el número de variables fuese muy elevado, podían ya directamente ser resueltos por estas técnicas. También, las ecuaciones diferenciales, que tenían funciones como incógnitas y que involucraban operadores diferenciales, podían ser resueltas eficientemente, si bien necesitaban de un paso intermedio que permitiera convertirlas en ecuaciones numéricas.

El modo más común de pasar de lo continuo a lo discreto es a través de los métodos de aproximación e interpolación que permiten aproximar funciones mediante polinomios u otras funciones discretas. Es bastante razonable pensar que en casi todo el Análisis Numérico, más tarde o más temprano, cualquier análisis de un método constructivo acaba siempre convirtiéndose en una cuestión relacionada con la teoría de la aproximación de funciones. No es de extrañar que los nombres de los que han fundamentado el Análisis Matemático clásico, como Newton, Fourier, Gauss, Legendre, Euler, Chebyshev y otros, se utilicen para denominar métodos y procedimientos del cálculo numérico.

Los objetivos primordiales que pretende cubrir este texto son:

1. La resolución de sistemas de ecuaciones lineales numéricas.

2. El estudio de las técnicas de aproximación e interpolación y su aplicación a la integración aproximada.

3. La resolución de sistemas de ecuaciones no-lineales numéricas.

4. La resolución de problemas de valor inicial y de contorno para ecuaciones diferenciales.

Los dos primeros apartados corresponden al curso *Análisis numérico matricial e interpolación* y los dos siguientes, al curso *Resolución numérica de ecuaciones*, ambos del grado de Matemáticas de la Universidad Nacional de Educación a Distancia (UNED). Deseo expresar mi agradecimiento a aquellas personas que han contribuido con sus observaciones a la mejora de este texto, y particularmente a Alfredo Cano y Juan Bódalo, alumnos del Master de Matemáticas Avanzadas de la UNED, por su contribución a eliminar algunos errores que contenía en sus primeras versiones.

Carlos Moreno
Madrid, marzo de 2021

Índice general

Parte I: Análisis numérico matricial e interpolación

1. **Estabilidad y errores en el cálculo numérico**
   1.1. Introducción
   1.2. Representación de números en un computador
   1.3. Aritmética en un sistema de representación finito 1.4. Estabilidad en los algoritmos numéricos
   1.5. Ejercicios

2. **Sistemas de ecuaciones numéricas lineales**
   2.1. Introducción
   2.2. Norma matricial subordinada a una norma vectorial
   2.3. Estabilidad de un sistema de ecuaciones lineales
   2.4. Sistemas lineales de gran dimensión
   2.5. Matrices dispersas
   2.6. Método de eliminación de Gauss
   2.7. Métodos especiales para matrices simétricas
   2.8. Factorización QR
      2.8.1. Método de ortogonalización de Gram-Schmidt
      2.8.2. Método de Householder
   2.9. Métodos iterativos
   2.10. Métodos iterativos clásicos
   2.11. Ejercicios

3. Aproximación de autovalores
3.1. Autovalores y vectores propios
3.2. Sucesiones de Krylov
3.3. Método de la potencia iterada
3.4. Método QR
3.5. Ejercicios

4. Aproximación de funciones
4.1. Introducción
4.2. Evaluación de polinomios
4.3. Aproximación de funciones
4.4. Aproximación por mínimos cuadrados
4.5. Aproximación discreta por mínimos cuadrados
4.6. Polinomios de Chebyshev
4.7. Aproximación trigonométrica
4.8. Aproximación uniforme
4.9. Ejercicios

5. Interpolación de funciones
5.1. Introducción
5.2. Interpolación de Lagrange
5.3. Método de Newton
5.4. Error en la interpolación de Lagrange
5.5. Algoritmos de Aitken y Neville
5.6. Interpolación compuesta
5.7. Interpolación de Hermite.
5.8. Interpolación por esplines cúbicos
5.9. Ejercicios

6. Derivación e integración numérica
6.1. Introducción
6.2. Fórmulas de derivación numérica
6.3. Método de Extrapolación de Richardson
6.4. Cuadratura basada en la interpolación
6.5. Fórmulas cerradas de Newton-Cotes
6.6. Cuadratura compuesta
6.7. Fórmulas de Gauss
6.8. Ejercicios

Parte II: Resolución numérica de ecuaciones

7. **Resolución numérica de ecuaciones no-lineales escalares**
   7.1. Introducción
   7.2. Método de dicotomía o bisección
   7.3. Métodos de punto fijo
   7.4. Velocidad de convergencia
   7.5. Método de la secante
   7.6. Método de Müller
   7.7. Método de Newton
   7.8. Método de Newton para raíces múltiples
   7.9. Raíces de ecuaciones polinómicas
   7.10. Ejercicios

8. **Resolución de sistemas de ecuaciones no lineales**
   8.1. Introducción
   8.2. Métodos de punto fijo
   8.3. Método de Newton
   8.4. Método de Broyden
   8.5. Raíces complejas de un polinomio
   8.6. Optimización sin restricciones
   8.7. Ejercicios

9. **Ecuaciones en diferencias finitas**
   9.1. Introducción
   9.2. Ecuaciones lineales homogéneas en diferencias con coeficientes constantes
   9.3. Ecuaciones lineales en diferencias con coeficientes constantes
   9.4. Estabilidad
   9.5. Ejercicios

10. **Problemas de valor inicial para ecuaciones diferenciales**
   10.1. Introducción
   10.2. Método de Euler
   10.3. Esquemas lineales multipaso
   10.4. Estabilidad de los métodos multipaso
   10.5. Convergencia de los métodos multipaso
   10.6. Estabilidad en intervalos que no están acotados
   10.7. Ejercicios

11. Problemas de contorno para ecuaciones diferenciales
11.1. Introducción
11.2. Métodos de diferencias finitas
11.3. Análisis de la convergencia
11.4. Estabilidad, consistencia y convergencia
11.5. Otras condiciones de contorno
11.6. Ejercicios

Parte I: Análisis numérico matricial e interpolación

CAPÍTULO 1

Estabilidad y errores en el cálculo numérico

1.1 Introducción

El saber contar y realizar operaciones aritméticas es el resultado de una larga evolución en el pensamiento humano. En todo este proceso, el hombre ha tratado de ayudarse en la realización de los cálculos mediante artilugios que simplificaran su labor. Primero fueron simples guijarros (calculi en latín), después fueron instrumentos mecánicos simples como el ábaco, instrumentos mecánicos articulados como las máquinas de Pascal o de Schickard y finalmente los computadores electrónicos. También, desde la programación en cilindros rotatorios o tarjetas perforadas hasta los desarrollos actuales en *software* basados en los lenguajes modernos de programación, se ha recorrido un largo camino. Todo este conocimiento y tecnología permite actualmente realizar cálculos de gran complejidad.

Sin embargo, un computador produce resultados en respuesta a cálculos programados que posiblemente difieren ligeramente de los valores exactos esperados. Ello es consecuencia de que trabajan con una aritmética discreta que no coincide plenamente con la aritmética exacta de los números enteros o reales. Todavía el ser humano no alcanza a realizar por medios físicos, todos los cálculos que su mente puede concebir, ni siguiera a representar en la memoria de un ordenador más que un subconjunto finito del conjunto de todos los números que puede manejar.

El propósito de este capítulo es estudiar las representaciones más comunes de números en un computador y las aritméticas que usan en sus cálculos. Una vez establecido que los errores respecto a la aritmética real pueden es-

tar presentes, se introduce el concepto de estabilidad numérica que permite seleccionar los algoritmos que son válidos para esta clase de cálculos.

1.2 Representación de números en un computador

La diferencia (en valor absoluto) entre el valor exacto x y el valor obtenido en un cálculo $\bar{x}$, determina el error absoluto cometido $E(x) = |x - \bar{x}|$. Una indicación más precisa del error cometido en el cálculo la da el error relativo, que se define por la siguiente expresión

$$E_R(x) = \frac{|x - \bar{x}|}{|x|}.$$

Si un determinado cálculo produce como resultado $\bar{x} = 1,2345 \times 10^{12}$ siendo el valor exacto esperado $x = 1.2331 \times 10^{12}$, el error absoluto sería $E(x) = 1.4 \times 10^9$. Este error parece muy alto si no se considera el orden de magnitud de los números con los que se realiza el cálculo. Por el contrario, el error relativo

$$E_R(x) = \frac{1.4 \times 10^9}{1.2331 \times 10^{12}} = 1.1353 \times 10^{-3}$$

da una idea más proporcionada de la imprecisión cometida.

Los computadores representan los números en sus memorias asignándoles una cantidad fija de posiciones a las que puede acceder. Esta cantidad puede variar de un computador a otro, de acuerdo con el estándar de representación que se haya atribuido a ese tipo de número. Es obvio que el modo de representación interna de un número en una memoria no tiene que coincidir necesariamente con el modo en que este número se muestra en una pantalla u otro dispositivo terminal.

Con el fin de comprender cómo usan los ordenadores la aritmética discreta, se recuerda el concepto de representación posicional de los números enteros y reales. Se considera el conjunto de números $\mathcal{M}$ que pueden ser representados en la forma

$$\pm 0.d_1 d_2 \cdots d_n \times b^e$$

donde $0 \leq d_i < b$ para $i = 1, 2, \cdots, n$. El número entero positivo b es fijo y corresponde a la base de la representación. El entero e corresponde al exponente y puede variar en un determinado rango. Finalmente, el entero positivo fijo n controla la precisión de la representación. La parte fraccionaria es frecuentemente llamada mantisa. El valor numérico real asignado a esta representación es

$$x = \pm(d_1 b^{-1} + d_2 b^{-2} + \cdots + d_n b^{-n}) \times b^e.$$

Por ejemplo, a la representación 0.1011×2^3 en base 2 le corresponde el valor decimal

$$x = 4 + 1 + \frac{1}{2} = 5.5$$

Idealmente, se puede ampliar el sistema admitiendo que n sea infinito y de este modo, se podrían incluir todos los números reales en esta representación. Obviamente, para un sistema de representación posicional que se pretenda poner en práctica en un computador, se requiere que n sea un número entero fijo y que se ajuste a la capacidad física de su memoria accesible. Por esta razón, todo sistema de representación que utilice un computador, tendrá un conjunto finito de números-máquina.

Habitualmente, para números enteros se emplea una representación en la que el exponente e es fijo e igual a n. De este modo, a la representación $\pm d_1 d_2 \cdots d_n$ le corresponde el valor entero

$$x = \pm(d_1 b^{n-1} + d_2 b^{n-2} + \cdots + d_n).$$

▪ Ejemplo **1** En un sistema binario ($b = 2$), con un número fijo de dígitos $n = 3$, el conjunto de números (que no son negativos) que pueden ser representados de este modo, son

Representación	000	001	010	011	100	101	110	111
Valor real	0	1	2	3	4	5	6	7

Un aspecto relevante de este sistema de representación es que puede producirse un desbordamiento en las operaciones aritméticas debido a que el conjunto es acotado. Por ejemplo, la suma de 3 y 5 produce un número que no puede ser representado en este sistema. ◇

Si el exponente e es un entero variable en un determinado rango, el conjunto de números que pueden ser representados de este modo (llamado de punto flotante) es más amplio, como muestra el siguiente ejemplo

▪ Ejemplo **2** En un sistema binario $b = 2$, si el número de dígitos es $n = 2$, el conjunto de números (que no son negativos) que pueden ser representados en punto flotante con el rango $-2 \leq e \leq 2$, es el siguiente

Representación	$.00 \times 2^e$	$.01 \times 2^e$	$.10 \times 2^e$	$.11 \times 2^e$
Valor	0	$\frac{1}{4} \times 2^e$	$\frac{1}{2} \times 2^e$	$\frac{3}{4} \times 2^e$

En definitiva, el conjunto de números que no son negativos y pueden ser representados de este modo son

$$\{0, \frac{1}{16}, \frac{1}{8}, \frac{3}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{2}, \frac{3}{4}, 1, \frac{3}{2}, 2, 3\}.$$

Los aspectos más relevantes de este sistema de representación son los siguientes:

- Los números representados no son equidistantes y se observa una mayor densidad en las proximidades de 0.
- Un número puede tener representaciones distintas. Por ejemplo, el número $\frac{1}{8}$ se puede representar por $.01 \times 2^{-1}$ o por $.10 \times 2^{-2}$.
- Se puede producir un desbordamiento en las operaciones aritméticas debido a que el conjunto es acotado. Por ejemplo, la suma de 1 y 3 produce un número que no puede ser representado en este sistema.
- El resultado de algunas operaciones aritméticas reales con números de este conjunto está fuera del conjunto aunque no se haya producido un desbordamiento del rango. Por ejemplo, la suma real de 2 y $\frac{1}{8}$ no pertenece al conjunto. $\diamond$

Una representación en punto flotante está normalizada si el primer dígito en la parte fraccionaria es necesariamente distinto de 0. De este modo, se evita que un número pueda tener representaciones distintas en un mismo sistema. Actualmente, la representación que usan la mayoría de computadores es la llamada IEEE Standard 754 en punto flotante. Está basada en los tres elementos mencionados anteriormente, el signo, la mantisa y el exponente.

Si la representación es binaria, el primer dígito es 1 (dígito principal) y en la mayoría de las puestas en práctica de este estándar, no es almacenado en memoria (dígito principal implícito). El estándar de representación en punto flotante de IEEE que corresponde a lo que se conoce como precisión simple, utiliza 4 bytes de memoria (32 bits) de los cuales 8 bits son para el exponente E, 23 para la parte fraccionaria F y uno para el signo S. Cada bit almacena un 0 o un 1. El valor asignado a una representación es

$$(-1)^S \times 2^{E-127} \times 1.F$$

El campo del exponente necesita representar a la vez exponentes positivos y negativos. Para conseguirlo se añade un sesgo al valor positivo almacenado para lograr el exponente deseado. En el estándar IEEE 754 para precisión

simple, este valor es 127. De este modo, un valor almacenado E representa un valor real $E - 127$. Para doble precisión el campo del exponente tiene 11 bits y un sesgo de 1023. El bit del signo es 0 para positivos y 1 para negativos.

Por ejemplo, si en memoria está la siguiente información almacenada de acuerdo al estándar IEEE 754 con dígito principal implícito

$$\underbrace{1}_{-} \ \underbrace{01010011}_{\text{exponente}} \ \underbrace{10011110001010000101000}_{\text{mantisa}}$$

el número representado es

$$-\left(1+\frac{1}{2}+\frac{1}{2^4}+\frac{1}{2^5}+\frac{1}{2^6}+\frac{1}{2^7}+\frac{1}{2^{11}}+\frac{1}{2^{13}}+\frac{1}{2^{18}}+\frac{1}{2^{20}}\right)\times 2^{83-127}$$

ya que $01010011_2 = 83$.

1.3 Aritmética en un sistema de representación finito

En general, cuando se utiliza un sistema discreto $\mathcal{M}$, inevitablemente aparecen errores al introducir los números reales convirtiéndolos a números-máquina pero también como el resultado de una operación entre números-máquinas que en general no coincide con el resultado que se obtendría en la aritmética real. Conceptualmente, los números reales pueden aproximarse por números del sistema discreto de dos modos ligeramente distintos que se conocen como truncamiento y redondeo.

La puesta en práctica de métodos de truncamiento está basada en la siguiente idea: Un numero real puede ser aproximado por un número de punto flotante con parte fraccionaria infinita en la base b

$$\pm .d_1 d_2 \cdots \times b^e.$$

Consecuentemente, si se trunca esta serie con n dígitos se obtiene una aproximación en el sistema discreto disponible en el computador. Sin embargo, es más usado el llamado método de redondeo que consiste en aproximar un número real positivo por el número-máquina más próximo. Este procedimiento sería el que nos proporcionaría mayor precisión. Para precisar estas ideas se representa la función parte entera de un número real por un corchete [ ] y se define para un número real positivo representado por $x = 0.m \times b^e$ donde m es una cifra con posiblemente una infinidad de dígitos, las siguientes aproximaciones por números máquina de n dígitos

Truncamiento	Redondeo
$\mathcal{T}(x) = [b^n \times 0.m] \times b^{e-n}$	$\mathcal{R}(x) = [b^n \times 0.m + 0.5] \times b^{e-n}$
Ejemplo (3 dígitos, base 10):	Ejemplo(3 dígitos, base 10):
$\mathcal{T}(1/3) = 10^{-3}[10^3 \times 0.333...] = 0.333$	$\mathcal{R}(1/3) = 10^{-3}[10^3 \times 0.333... + 0.5] = 0.333$
$\mathcal{T}(2/3) = 10^{-3}[10^3 \times 0.666...] = 0.666$	$\mathcal{R}(2/3) = 10^{-3}[10^3 \times 0.666... + 0.5] = 0.667$

En lo que sigue se centrará la atención en el método de redondeo. Si $x < 0$ entonces se define

$$\mathcal{R}(x) = -\mathcal{R}(-x).$$

De la definición de redondeo se deduce que

$$|x - \mathcal{R}(x)| \leq \frac{1}{2} b^{e-n}.$$

El número $\frac{1}{2}b^{e-n}$ se conoce como unidad de redondeo o precisión de la máquina.

Una aritmética para un sistema de representación finita estaría disponible si se consigue asignar como resultado de una operación el que corresponde al redondeo de la operación aritmética exacta.

EJERCICIO 1 *Sea $x > 0$ un número que se representa exacto en una aritmética finita de punto flotante. Analizar el comportamiento del cociente*

$$\frac{\operatorname{sen}(x+h) - \operatorname{sen} x}{h}$$

cuando se consideran valores de h próximos a 0 y los cálculos se realizan en una aritmética de punto flotante.

Solución: Para h suficientemente pequeño el resultado de redondear $x + h$ coincide con x. Por ello, el numerador es nulo y el valor asignado al cociente es 0. ◇

EJERCICIO 2 *Si se usa una aritmética con redondeo, en un sistema de representación decimal y de 3 dígitos de precisión, para realizar la siguiente la operación*

$$\frac{a \times b - c}{b + 2 \times c}, \quad a = 1.34,\ b = 0.712,\ c = -0.355,$$

determinar el error cometido en relación con la aritmética real.

Solución: En la representación finita los números dados son

$$a = 0.134 \times 10^1,\ b = 0,712 \times 10^0,\ c = -0.355 \times 10^0.$$

Consecuentemente, el resultado de las operaciones es el siguiente:

- $a \times b = \mathcal{R}(0.95408) = 0.954 \times 10^0$.
- $a \times b - c = \mathcal{R}(0.954 + 0.355) = 0.131 \times 10^1$.
- $2 \times c = -0.710$.
- $b + 2 \times c = 0.2 \times 10^{-2}$.
- $\frac{a \times b + c}{b + 2 \times c} = 0.655 \times 10^3$.

El resultado exacto es 0.65454×10^3 y consecuentemente el error relativo es

$$E_R = \frac{|655 - 654.54|}{654.54} = 0.00070278\ . \qquad \diamond$$

1.4 Estabilidad en los algoritmos numéricos

Una vez que un entorno de cálculo dispone de una aritmética finita que le permite realizar las operaciones elementales dentro del rango numérico que puede representar, se hace necesario complementarlo con otras funciones elementales que permitan realizar cálculos más complejos. La dificultad está en que estas funciones implican un número elevado de operaciones elementales y puesto que los errores de redondeo respecto a la aritmética exacta son inevitables, el resultado final puede estar muy deteriorado en relación con el exacto.

▪ Ejemplo **3** Si en un entorno que usa el estándar IEEE Double Precision, se realiza el cálculo de e^{-10} mediante el truncamiento de la serie de potencias

$$e^x \approx 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \cdots \frac{x^{30}}{30!}$$

el resultado aproximado es $9.703415796025504 \times 10^{-4}$. Desafortunadamente, este valor se aleja del valor exacto $4.539992976248485 \times 10^{-5}$. El error de truncamiento de la serie está dado

$$E(-10) = \frac{e^\xi}{31!} < \frac{1}{31!}$$

para algún valor $-10 < \xi < 0$. Consecuentemente, si el cálculo se realizase en una aritmética exacta, el truncamiento de la serie no podría producir un error de ese tamaño. El motivo de la inexactitud está en que los pequeños errores causados por la aritmética finita, han sido magnificados a lo largo del cálculo.

Por otra parte, si se calcula $e^{-10} = \frac{1}{e^{10}}$ usando sólo el desarrollo en serie del denominador, el resultado es $4.539993338712231 \times 10^{-5}$ que es muy próximo al valor exacto. $\diamond$

La razón por la que un procedimiento produce resultados más imprecisos que otro se atribuye a su inestabilidad. Es decir, un error en alguna etapa del procedimiento se propaga de un modo creciente en las siguientes. La necesidad de utilizar algoritmos estables para evaluar una función es obligada por la inevitable presencia de errores de redondeo debidos al uso de aritméticas finitas en los entornos de cálculo automático.

1.5 Ejercicios

▪ Ejemplo **4** En este ejemplo se muestra cómo convertir el número decimal $x = 324.65$ a base 2. Se separa la parte entera 324 de la parte decimal 0.65. Para la parte entera se procede de modo sigue:

$$\begin{aligned} 324 &= 162 \times 2 + 0 \\ 162 &= 81 \times 2 + 0 \\ 81 &= 40 \times 2 + 1 \\ 40 &= 20 \times 2 + 0 \\ 20 &= 10 \times 2 + 0 \\ 10 &= 5 \times 2 + 0 \\ 5 &= 2 \times 2 + 1 \\ 2 &= 1 \times 2 + 0 \end{aligned}$$

En consecuencia se tiene que $324 = 101000100_2$. El dígito más significativo siempre es 1 y los siguientes son los restos de la divisiones por 2 comenzando desde la última división.

En la parte decimal, se procede como sigue

$$\begin{array}{rcl} 0.65 & \times & 2 = 1.3 \\ 0.3 & \times & 2 = 0.6 \\ 0.6 & \times & 2 = 1.2 \\ 0.2 & \times & 2 = 0.4 \\ 0.4 & \times & 2 = 0.8 \\ 0.8 & \times & 2 = 1.6 \\ 0.6 & \times & 2 = 1.2 \\ \dots & \dots & \dots \end{array}$$

En consecuencia se tiene que $0.65 = 0.1010011\cdots_2$. Los dígitos son las partes enteras de las sucesivas multiplicaciones por 2 comenzando por la primera. ◇

EJERCICIO 3 *Hallar la representación en punto flotante (sin dígito principal implícito) de*

1. $2/7$ *en un sistema* $b = 10$, $n = 6$.
2. 327 *en un sistema* $b = 2$, $n = 6$.
3. $-\frac{4}{3}$ *en un sistema* $b = 2$, $n = 6$.

por redondeo.

Solución:

1. $\mathcal{R}(\frac{2}{7}) = 0.285714 \times 10^0$.
2. Puesto que $327 = 2^8 + 2^6 + 2^2 + 2 + 1 = 0.101000111_2 \times 2^9$ entonces se tiene que

$$\mathcal{R}(327) = [2^6 \times 0.101000111_2 + 0.5] \times 2^3 = 0.101001_2 \times 2^9 = 328.$$

3. Puesto que $\frac{4}{3} = 0.10101010...{}_2 \times 2$ entonces se tiene que

$$\mathcal{R}(\frac{4}{3}) = [2^6 \times 0.1010101010_2 + 0.5] \times 2^{-5} = 0.101011_2 \times 2.$$

Consecuentemente $-\frac{4}{3} = -0.101011_2 \times 2$ ◇.

EJERCICIO 4 *Hallar el resultado con la aritmética finita de un sistema de representación $b = 2$, $n = 6$ y $-3 \leq e \leq 3$, sin dígito principal implícito, de las siguientes operaciones*

1. $\frac{1}{3} + \frac{7}{13}$
2. $47 + 347$

Solución:

1. Puesto que $\mathcal{R}(1/3) = 0.101011_2 \times 2^{-1}$ y $\mathcal{R}(7/13) = 0.100010_2$ entonces

$$\begin{aligned} \mathcal{R}(\mathcal{R}(1/3) + \mathcal{R}(7/13)) &= \mathcal{R}(0.1101111_2) = 2^{-6}[110111.1_2 + 0.1_2] \\ &= 2^{-6} \times 111000_2 = 0.111000_2 \ . \end{aligned}$$

2. El valor real máximo que se puede representar en este sistema es

$$0.111111_2 \times 2^3 = 4 + 2 + 1 + 0.5 + 0.25 + 0.125 = 7.875.$$

Consecuentemente, las representaciones de 47 y 347 producen un desbordamiento (overflow) que habitualmente es notificado . ◇

EJERCICIO 5 *Se pretende aproximar el valor de $e^{0.1}$ usando la serie de potencias*

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots + \frac{x^n}{n!} + \cdots$$

en una aritmética basada en el redondeo en base 10 y 7 dígitos de precisión. Si se usan cinco sumandos de la serie de potencias, calcular el error relativo que se comete al realizar la aproximación en la aritmética finita respecto de la que corresponde a la aritmética exacta.

Solución: Los cálculos exactos de los cinco primeros términos de la serie dan las siguientes resultados

n	1	2	3	4	5	$p_5(0.1)$
$\frac{x^n}{n!}$	1	0.1	0.005	$0,0001\hat{6}$	$0.0000041\hat{6}$	$1.1051708\hat{3}$
$\mathcal{R}\left(\frac{x^n}{n!}\right)$	1	0.1	0.005	0.0001667	0.0000042	1.105171

donde p_5 representan los valores de la serie truncada, calculados en aritmética exacta y finita. El sombrero representa el decimal periódico en la expresión decimal.

$$\mathcal{R}(x) = [10^7 \times 0.110517083\hat{} + 0.5] \times b^{-6} = 1.105171$$

En este caso, el resultado no depende del orden en el que se realicen las sumas. El error relativo es

$$E_R(0.1) = \frac{1.105171 - 1.10517083\hat{}}{1.10517083\hat{}}. \qquad \diamond$$

CAPÍTULO 2

Sistemas de ecuaciones numéricas lineales

2.1 Introducción

La resolución eficiente de sistemas de ecuaciones numéricas lineales es una necesidad presente en muchos problemas del cálculo científico. El estudio de estructuras elásticas discretas, circuitos eléctricos complejos u otros modelos discretos de la ciencia y de la ingeniería pueden conducir directamente a sistemas lineales de ecuaciones numéricas de gran dimensión. Pero, también la necesidad de resolver grandes sistemas de ecuaciones lineales puede surgir como una etapa intermedia en un procedimiento más amplio como puede ser la resolución numérica de ecuaciones diferenciales o de problemas formulados en modelos estadísticos que usan cantidades importantes de datos. Se debe tener presente que las ecuaciones diferenciales forman parte ineludible de muchos de los modelos matemáticos que se construyen no sólo en Física e Ingeniería sino también en Biología, en Economía y en otras Ciencias Sociales. Desafortunadamente, algunos de los métodos elementales de resolución que se introducen en los cursos básicos de Álgebra Lineal pueden resultar ineficaces cuando el número de incógnitas y ecuaciones crece considerablemente. El objetivo de este capítulo es el estudio de métodos de resolución de sistemas lineales que sean eficientes en problemas reales con media o alta dimensión.

Una breve introducción en las primeras secciones de algunos conceptos básicos del análisis matricial, permitirá posteriormente analizar la estabilidad y la convergencia de los métodos considerados en este capítulo.

Aunque la elección del método más adecuado para resolver un sistema lineal depende fundamentalmente de las propiedades de la matriz de coeficien-

tes, para un sistema lineal sin características especiales, uno de los métodos más eficientes es el clásico método de eliminación de Gauss o alguna de sus variantes. Estos métodos pertenecen a la clase de los métodos directos que conducen a la solución exacta en un número finito de operaciones aunque ello sólo desde un punto de vista teórico, debido a la presencia de errores de redondeo cuando los cálculos se realizan en un ordenador. En otra categoría de métodos, los iterativos permiten generar una sucesión de soluciones aproximadas que convergen a la solución exacta. En este capítulo se estudian métodos de ambas categorías analizando sus propiedades y el modo de ponerlos en práctica.

2.2 Norma matricial subordinada a una norma vectorial

La norma euclídea de un vector (real o complejo) $\mathbf{x}$ de n componentes, cuya componente $i-$ésima es x_i, está definida por la siguiente expresión

$$\| \mathbf{x} \| = \sqrt{\mathbf{x} \cdot \mathbf{x}} = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}}.$$

Más general es la norma p que está definida por

$$\| \mathbf{x} \|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}$$

para cualquier número positivo p. En el límite, la norma ∞ está definida por

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

Una matriz puede representar a una aplicación lineal entre dos espacios vectoriales. Por ello, existen normas matriciales que están definidas en base a esta representación y cuya definición está relacionada con las normas euclídeas en ambos espacios. Así, una norma matricial subordinada a una norma vectorial está definida por

$$\| \mathrm{A} \| = \max_{\|\mathbf{x}\|=1} \| \mathrm{A}\mathbf{x} \|$$

para cualquier matriz A. Obviamente, si la matriz A tiene n filas y m columnas, dos normas vectoriales entrarían en juego. Aunque muchos de los conceptos y resultados que se exponen en este capítulo son extensibles a esta situación, por limitación de objetivos en este texto, se supondrá siempre,

salvo mención en sentido contrario, que la matriz A es cuadrada y la elección de norma para el dominio e imagen es la misma.

En los casos más simples se pueden obtener expresiones que permiten determinar el valor de una norma subordinada de una matriz en términos de sus coeficientes. Algunas de estas situaciones son las siguientes (véase [8, Quarteroni, Sacco, Saleri], páginas 23-24)

- Norma subordinada 1

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

si se considera la norma vectorial

$$\| \mathbf{x} \|_1 = \sum_{j=1}^{n} |x_j|$$

y a_{ij} representa el coeficiente de la matriz A correspondiente a la fila i y la columna j.

- Norma subordinada ∞

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

si se considera la norma vectorial

$$\| \mathbf{x} \|_\infty = \max_{1 \leq j \leq n} |x_j|.$$

- Norma subordinada 2

$$\| \mathrm{A} \|_2 = \rho(\mathrm{A}^t \mathrm{A})^{\frac{1}{2}},$$

si se considera la norma vectorial

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

y A^t representa la matriz traspuesta de A y ρ el radio espectral de la matriz (es decir, el máximo de los módulos de los autovalores). En particular, si A es simétrica entonces se cumple que $\| \mathrm{A} \|_2 = \rho(A)$.

La norma subordinada euclídea no coincide con la norma matricial, asociada al producto escalar matricial

$$\mathrm{A} : \mathrm{B} = \mathrm{tr}\ \mathrm{A}^{\mathrm{t}}\mathrm{B},$$

donde tr representa el operador que asocia a una matriz su traza, y que se conoce como norma de Frobenius

$$\| \mathrm{A} \|_F = (\mathrm{A} : \mathrm{A})^{\frac{1}{2}} = \left(\sum_{i,j=1}^{n} a_{ij}^2 \right)^{\frac{1}{2}}.$$

Mientras que la norma subordinada euclídea está relacionada con el radio espectral del producto $\mathrm{A}^t\mathrm{A}$, la norma de Frobenius lo está con su traza (es decir, la suma de los elementos de la diagonal principal). Por otra parte, para la matriz identidad I todas las normas matriciales toman el valor 1 pero la norma de Frobenius toma el valor $\sqrt{n}$.

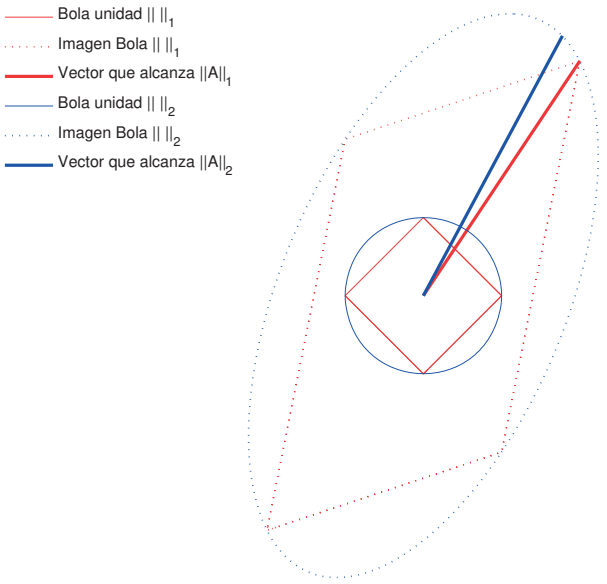


Figura 2.1: Normas subordinadas

▪ Ejemplo **5** En la figura 2.1 se muestran las bolas unidad correspondientes a las normas $\| \ \|_1$ y $\| \ \|_2$ y sus imágenes por la transformación definida por la matriz

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

También se muestran algunos de los vectores en los que se alcanza el máximo del siguiente conjunto

$$\{\|\mathrm{A}\mathbf{x}\| : \|\mathbf{x}\| = 1\}$$

para cada una de ambas normas. ◇

EJERCICIO 6 *Determinar la norma subordinada a las normas vectoriales* $\| \ \|_1$, $\| \ \|_2$ *y* $\| \ \|_\infty$ *de la matriz*

$$\mathrm{A} = \begin{pmatrix} 2 & 1 \\ 5 & 4 \end{pmatrix}$$

así como su norma de Frobenius.

Solución: Directamente de la definición se obtiene que

$$\| \mathrm{A} \|_1 = \max\{2+5, 1+4\} = 7,$$

$$\| \mathrm{A} \|_\infty = \max\{2+1, 5+4\} = 9.$$

El producto

$$\mathrm{A}^t\mathrm{A} = \begin{pmatrix} 29 & 22 \\ 22 & 17 \end{pmatrix}$$

tiene como autovalores $\lambda_1 = 0.1965$ y $\lambda_2 = 45.8035$ y como traza $29+17 = 46$. Consecuentemente se tiene que

$$\| \mathrm{A} \|_2 = \sqrt{45.8035} = 6.7678,$$

$$\| \mathrm{A} \|_F = \sqrt{46} = 6.7823 \ .$$

En general, el radio espectral de A no coincide con la norma subordinada a la norma euclídea. De hecho, en el ejemplo de este ejercicio, la matriz A tiene como autovalores $\lambda_1 = 0.5505$ y $\lambda_2 = 5.4495$ y por ello, su radio espectral es 5.4495; valor que no coincide con el de la norma 2. ◇

Una consecuencia inmediata de la definición de norma subordinada es la siguiente desigualdad

$$\| \mathrm{AB} \| \leq \| \mathrm{A} \| \| \mathrm{B} \|$$

válida para dos matrices cualesquiera. Como consecuencia de esta desigualdad se tiene que para toda matriz A

$$\| \mathrm{A}^q \| \leq \| \mathrm{A} \|^q,$$

lo que implica en particular, que la sucesión de potencias sucesivas de una matriz converge a cero si para alguna de sus normas subordinadas se cumple que $\|\mathrm{A}\| < 1$.

Si $\mathbf{x}$ es un vector propio asociado a un autovalor λ de la matriz A, $\mathbf{x}$ verifica que $\mathrm{A}\mathbf{x} = \lambda\mathbf{x}$. De la definición de norma subordinada se deduce que

$$\|\mathrm{A}\mathbf{x}\| \leq \|\mathrm{A}\|\|\mathbf{x}\|$$

En consecuencia

$$|\lambda|\|\mathbf{x}\| \leq \|\mathrm{A}\|\|\mathbf{x}\|$$

y

$$|\lambda| \leq \|\mathrm{A}\|.$$

Puesto que esta desigualdad se verifica para todo autovalor, se concluye que

$$\rho(\mathrm{A}) \leq \|\mathrm{A}\|.$$

El siguiente teorema, cuya demostración se omite (véase [11, Varga] o [8, Quarteroni, Sacco, Saleri], páginas 25-27), establece de modo más preciso la relación del radio espectral de una matriz con las normas subordinadas

Teorema 1 *Para cualquier matriz cuadrada* A *se cumple*

$$\rho(\mathrm{A}) = \inf\{\|\mathrm{A}\| : \|\ \| \text{ es una norma subordinada}\}.$$

Es importante tener en cuenta que si A es simétrica, el ínfimo se alcanza en la norma subordinada a la norma euclídea, ya que

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

Otra consecuencia del teorema anterior es el siguiente

Corolario 1 *Para una matriz cuadrada arbitraria, las tres afirmaciones siguientes, referidas a las potencias de* A *y el radio espectral, son equivalentes*

1. $\lim_{n\to\infty} \mathrm{A}^n = \mathrm{O}$,
2. $\lim_{n\to\infty} \|\mathrm{A}^n\| = 0$, *para alguna norma subordinada,*
3. $\rho(\mathrm{A}) < 1$.

Es conveniente tener en cuenta que la segunda afirmación podría ampliarse a cualquier norma subordinada, ya que todas ellas son equivalentes por ser finita la dimensión de los espacios vectoriales de matrices.

▪ EJEMPLO **6** La existencia de una norma subordinada para la que $\|\mathrm{A}\| > 1$ no impide que $\lim_{n\to\infty} \mathrm{A}^n = \mathrm{O}$. En efecto, para la matriz

$$\mathrm{A} = \begin{pmatrix} \frac{1}{2} & \frac{3}{10} \\ 0 & \frac{4}{5} \end{pmatrix}$$

la norma $\|\mathrm{A}\|_1 = \frac{11}{10} > 1$ pero $\rho(\mathrm{A}) = 0.8 < 1$, lo que garantiza que las potencias de A convergen a O. ◇

2.3 Estabilidad de un sistema de ecuaciones lineales

Cuando se pretende resolver un sistema de ecuaciones numéricas lineales es importante conocer previamente si el sistema es sensible a pequeñas modificaciones en los datos iniciales o a los errores de redondeo que se producen en el curso del cálculo. Por el momento, se analizará la cuestión de si una pequeña modificación de los datos del sistema $\mathrm{A}\mathbf{x} = \mathbf{b}$ puede generar un importante cambio en la solución. Para ello se considera el sistema perturbado

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

donde $\delta\mathrm{A}$, $\delta\mathbf{b}$ son elementos arbitrarios de la misma naturaleza que A y $\mathbf{b}$. El error que introduce esta perturbación está dado por

$$\delta\mathbf{x} = (\mathrm{A} + \delta\mathrm{A})^{-1}(\delta\mathbf{b} - \delta\mathrm{A}\mathbf{x}).$$

Si se toman normas en ambos miembros y se usa para las matrices una norma subordinada, se obtiene que

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

Por otra parte, puesto que $\mathbf{x}$ es solución de la ecuación, se cumple que

$$\|b\| \leq \|\mathrm{A}\|\|\mathbf{x}\|.$$

De las dos últimas desigualdades se deduce que

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

Para completar la estimación del error relativo se usará el siguiente lema de Banach

Lema 1 *Si* D *es una matriz cuadrada tal que* $\|\mathrm{D}\| < 1$ *para una determinada norma subordinada a una norma vectorial entonces* $\mathrm{I}+\mathrm{D}$ *es no-singular y se cumple que*

$$\|(\mathrm{I}+\mathrm{D})^{-1}\| \leq \frac{1}{1-\|\mathrm{D}\|}.$$

Demostración: Para todo vector $\mathbf{x}$ no nulo se tiene que

$$\|(\mathrm{I}+\mathrm{D})\mathbf{x}\| \geq \|\mathbf{x}\| - \|\mathrm{D}\mathbf{x}\| \geq \|\mathbf{x}\|(1-\|\mathrm{D}\|) > 0.$$

Consecuentemente, $\mathrm{I}+\mathrm{D}$ no es singular. Por otra parte, se cumple la siguiente desigualdad

$$\begin{aligned} 1 &= \|\mathrm{I}\| = \|(\mathrm{I}+\mathrm{D})(\mathrm{I}+\mathrm{D})^{-1}\| \\ &\geq \|(\mathrm{I}+\mathrm{D})^{-1}\| - \|\mathrm{D}\|\|(\mathrm{I}+\mathrm{D})^{-1}\| = (1-\|\mathrm{D}\|)\|(\mathrm{I}+\mathrm{D})^{-1}\|, \end{aligned}$$

lo que prueba el resultado. $\diamond$

Volviendo a la estimación del error, si se usa el lema de Banach se obtiene la siguiente acotación

$$\|(\mathrm{A}+\delta\mathrm{A})^{-1}\| = \|(\mathrm{I}+\mathrm{A}^{-1}\delta\mathrm{A})^{-1}\mathrm{A}^{-1}\| \leq \|\mathrm{A}^{-1}\|\frac{1}{1-\|\mathrm{A}^{-1}\|\|\delta\mathrm{A}\|},$$

si se cumple que $\|\mathrm{A}^{-1}\|\|\delta\mathrm{A}\| < 1$. Si se usa esta acotación en la estimación del error relativo de la solución del sistema perturbado, se obtiene que

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

Esta estimación sugiere que una medida de esta sensibilidad es el número de condición de A, definido por

$$\mathrm{cond}(\mathrm{A}) = \|\mathrm{A}\|\|\mathrm{A}^{-1}\|$$

para toda matriz cuadrada no-singular A. De este modo, si $\mathrm{cond}(\mathrm{A})\frac{\|\delta\mathrm{A}\|}{\|\mathrm{A}\|} < 1$ se cumple que

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

Además, de la relación

$$1 = \|\mathrm{I}\| = \|\mathrm{A}\mathrm{A}^{-1}\| \leq \|\mathrm{A}\|\|\mathrm{A}^{-1}\| = \mathrm{cond}(\mathrm{A})$$

se deduce que el número de condición de una matriz es siempre mayor o igual que 1. Si el número de condición es relativamente pequeño los errores que se generarían en la solución serían pequeños. Por el contrario, para sistemas lineales cuya matriz de coeficientes tenga un número de condición muy alto, no hay garantía de que la solución calculada con una aritmética finita se mantenga próxima a la solución exacta.

▪ EJEMPLO **7** Se considera la matriz

$$\mathrm{A} = \begin{pmatrix} 1+\frac{1}{\epsilon} & -1+\frac{1}{\epsilon} \\ -1+\frac{1}{\epsilon} & 1+\frac{1}{\epsilon} \end{pmatrix}$$

donde $\epsilon > 0$ es un número destinado a tender a 0. La matriz inversa de A es

$$\mathrm{A}^{-1} = \frac{1}{4}\begin{pmatrix} \epsilon+1 & \epsilon-1 \\ \epsilon-1 & \epsilon+1 \end{pmatrix}.$$

Fácilmente se comprueba que

$$\|\mathrm{A}\|_2 = \rho(\mathrm{A}^t\mathrm{A})^{\frac{1}{2}} = \frac{2}{\epsilon}; \quad \|\mathrm{A}^{-1}\|_2 = \rho(\mathrm{A}^{-t}\mathrm{A}^{-1})^{\frac{1}{2}} = \frac{1}{2}$$

y de ello se deduce que $\mathrm{cond}(\mathrm{A}) = \frac{1}{\epsilon}$. La matriz A está mal condicionada cuando $\epsilon \to 0$. ◇

Una matriz está perfectamente condicionada si su número de condición es 1. En particular, las matrices ortogonales, que verifican que $\mathrm{A}^t\mathrm{A} = \mathrm{I}$, están perfectamente condicionadas, ya que

$$\|\mathrm{A}\|_2 = \rho(\mathrm{A}^t\mathrm{A})^{\frac{1}{2}} = \rho(\mathrm{I})^{\frac{1}{2}} = 1$$

y

$$\|\mathrm{A}^{-1}\|_2 = \rho(\mathrm{A}\mathrm{A}^t)^{\frac{1}{2}} = \rho(\mathrm{I})^{\frac{1}{2}} = 1.$$

Por otra parte, si A es una matriz simétrica definida positiva entonces se cumple que

$$\|\mathrm{A}\|_2 = \lambda_{max}, \quad \|\mathrm{A}^{-1}\|_2 = \frac{1}{\lambda_{min}}$$

donde λ_{min} and λ_{max} representan respectivamente el mínimo y el máximo autovalor de A. De este modo, se tiene que

$$\mathrm{cond}_2(\mathrm{A}) = \frac{\lambda_{max}}{\lambda_{min}}.$$