

Computación Científica I,¹ 2007
Guía N^o 2
Valparaíso–Santiago, 18 de junio de 2007²

1. PROBLEMAS DE CUADRADOS MÍNIMOS.

1. [5, cf. p. 79, Example 11.2] 1) *El problema clásico de ajuste de cuadrados mínimos.* Sea \mathfrak{P}_{n-1} la colección de los polinomios complejos de grado *exactamente* $n - 1$ en la variable compleja ζ . Considere la data $(\zeta_k, y_k) \in \mathbb{C}^2$, $k = 1 : m$, $m \geq n$, con $\zeta_k \neq \zeta_\ell$ para todo $k \neq \ell$. Se dice que el polinomio $p(\zeta) = x_1 + x_2\zeta + \dots + x_n\zeta^{n-1} \in \mathfrak{P}_{n-1}$ (note que los coeficientes de $p(\zeta)$ se llaman x_k y la variable independiente ζ) ajusta la data en el sentido de los cuadrados mínimos ssi $p(\zeta)$ realiza el mínimo $\min_{q \in \mathfrak{P}_{n-1}} \sum_{k=1}^m |p(\zeta_k) - y_k|^2$, i.e., ssi:

$$\sum_{k=1}^m |p(\zeta_k) - y_k|^2 = \min_{q \in \mathfrak{P}_{n-1}} \sum_{k=1}^m |q(\zeta_k) - y_k|^2.$$

El polinomio optimal $p(\zeta)$ queda caracterizado, entonces, por el vector $x = [x_1, x_2, x_3, \dots, x_n]^T \in \mathbb{R}^n$ que minimiza la norma cuadrática:

$$\left\| \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_m \end{bmatrix} - \begin{bmatrix} 1 & \zeta_1 & \zeta_1^2 & \dots & \zeta_1^{n-1} \\ 1 & \zeta_2 & \zeta_2^2 & \dots & \zeta_2^{n-1} \\ 1 & \zeta_3 & \zeta_3^2 & \dots & \zeta_3^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \zeta_m & \zeta_m^2 & \dots & \zeta_m^{n-1} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix} \right\|_2$$

Puesto que m puede ser mayor que n , se observa que no cabe esperar que el mínimo precedente sea cero. En particular, el problema no puede ser resuelto por simple “inversión” de la matriz de Vandermonde $[\zeta_k^\ell]_{m \times n}$ (esta matriz no es cuadrada en general!).

Evidentemente, el polinomio optimal $p \in \mathfrak{P}_{n-1}$ se corresponde biunívocamente con el vector optimal $x = [x_1, x_2, x_3, \dots, x_n]^T \in \mathbb{R}^n$ cuyas componentes son los coeficientes del polinomio optimal. Por su parte, el vector optimal $x \in \mathbb{R}^n$ del problema de cuadrados mínimos discutido queda caracterizado, más generalmente, por el siguiente teorema:

Teorema 1. Sean $m \geq n$, $A \in M(m \times n, \mathbb{C})$, $b \in \mathbb{C}^m$, $r(u) := b - Au$ para todo $u \in \mathbb{C}^n$. Entonces las siguientes propiedades de un vector $x \in \mathbb{C}^n$ son equivalentes:

(a) x minimiza $\|r(u)\|_2$ sobre la colección de todos los vectores $u \in \mathbb{C}^n$, i.e.:

$$\|r(x)\|_2 = \|b - Ax\|_2 = \min_{u \in \mathbb{C}^n} \|b - Au\|_2.$$

(b) $r(x) \perp \text{Range}(A)$, i.e., $(b - Ax) \perp Au$ para todo $u \in \mathbb{C}^n$.

(c) $A^*r(x) = 0$, i.e. $A^*(b - Ax) = 0$.

(d) $A^*Ax = A^*b$; esta es la llamada *ecuación normal del problema de cuadrados mínimos*.

(e) $Ax = Pb$, donde $P \in M(m \times m, \mathbb{C})$ es el *proyector ortogonal de \mathbb{C}^m sobre $\text{Range}(A)$*

Zusatz: Si, además, A es una matriz de rango completo, entonces el vector $x \in \mathbb{C}^n$ caracterizado por las propiedades equivalentes (a)-(e), es único y viene dado por $x = (A^*A)^{-1}A^*b$. Este vector x se denomina entonces la *solución del problema de cuadrados mínimos*.

¹Proyecto UTFSM-DGIP 2006-2007 Desarrollo de Textos Docentes.

²© Luis Salinas Carrasco, Valparaíso, 18 de junio de 2007. De antemano se agradece toda corrección, crítica o comentario que el amable lector tenga a bien hacer llegar a luis.salinas@usm.cl.

TAREA: Demuestre este teorema.

2) *Matriz pseudo-inversa o inversa de Moore Penrose.* En el caso en que A es de rango completo se define la matriz *pseudo-inversa* A^+ de A , o *inversa de Moore-Penrose*, como $A^+ = (A^*A)^{-1}A^* \in M(n \times m, \mathbb{C})$. Así, la solución del problema de cuadrados mínimos, en este caso, viene dada simplemente por $x = A^+b$.

TAREA: Verifique que cuando A es de rango completo y $m = n$, se tiene $A^+ = A^{-1}$.

3) *Cuadrados mínimos mediante el sistema de ecuaciones normales.* El **método clásico** para resolver los problemas de cuadrados mínimos consiste en resolver el sistema de ecuaciones normales. Si A es de rango completo, la matriz de este sistema es cuadrada, Hermitiana, positiva definida. El correspondiente sistema tiene n ecuaciones lineales y n incógnitas.

TAREA: Verifique estos últimos hechos.

El método estándar para resolver sistemas de ecuaciones lineales en que la matriz del sistema tiene las propiedades mencionadas más arriba, recurre a la *factorización de Cholesky* que factoriza A^*A en la forma $A^*A = R^*R$, donde R es una matriz triangular superior y reduce la ecuación normal al sistema

$$R^*Rx = A^*b.$$

El correspondiente algoritmo es bastante obvio:

Cuadrados mínimos mediante las ecuaciones normales.

1. Formar la matriz A^*A y el vector A^*b .
2. Calcular la factorización de Cholesky $A^*A = R^*R$.
3. Resolver el sistema triangular inferior $R^*w = A^*b$ para w .
4. Resolver el sistema triangular superior $Rx = w$ para x .

TAREA: Programe este algoritmo y verifique su correcta operación mediante algunos ejemplos sencillos. Compruebe que su complejidad es del orden de $mn^2 + \frac{1}{3}n^3$ "flops".

4) *Cuadrados mínimos mediante la factorización QR.* El **método "clásico moderno"** para resolver los problemas de cuadrados mínimos, que se ha hecho popular a partir de 1960 aproximadamente, recurre a la factorización QR. Aplicando la ortogonalización de Gram-Schmidt o, más a menudo, la triangularización de Householder, se calcula una factorización $A = \widehat{Q}\widehat{R}$. Como se sabe \widehat{Q} es una matriz, en general, rectangular de $m \times n$ cuyas *columnas* son ortonormales, y \widehat{R} es una matriz triangular superior cuadrada de $n \times n$. Luego, si A es de rango completo (i.e., $\text{rank}(A) = n = \min\{m, n\}$), la matriz \widehat{R} es invertible. Por consiguiente, en este caso, el proyector ortogonal P sobre $\text{Range}(A)$, se escribe simplemente como:

$$P = A(A^*A)^{-1}A^* = \widehat{Q}\widehat{R} \left(\widehat{R}^*\widehat{Q}^*\widehat{Q}\widehat{R} \right)^{-1} \widehat{R}^*\widehat{Q}^* = \widehat{Q}\widehat{R}\widehat{R}^{-1}(\widehat{R}^*)^{-1}\widehat{R}^*\widehat{Q}^* = \widehat{Q}\widehat{Q}^*.$$

De este modo se tiene:

$$y = Pb = \widehat{Q}\widehat{Q}^*b.$$

Como $y \in \text{Range}(A)$, el sistema $Ax = y$, esto es:

$$\widehat{Q}\widehat{R}x = \widehat{Q}\widehat{Q}^*b,$$

tiene una solución exacta, a saber:

$$x = \widehat{R}^{-1}\widehat{Q}^*\widehat{Q}\widehat{Q}^*b = \widehat{R}^{-1}\widehat{Q}^*b.$$

Como es usual, en la práctica nunca se calcula explícitamente \widehat{R}^{-1} sino que el sistema correspondiente se resuelve mediante sustitución inversa ("back-substitution") pues \widehat{R} es una matriz triangular superior cuadrada. En vista de la solución general $x = A^+b$, el resultado precedente muestra, de paso, que cuando A es de rango completo se tiene:

$$A^+ = \widehat{R}^{-1}\widehat{Q}^*.$$

El algoritmo discutido puede resumirse del siguiente modo:

Cuadrados mínimos mediante factorización QR.

1. Calcular la factorización QR reducida $A = \widehat{Q}\widehat{R}$.
2. Calcular el vector \widehat{Q}^*b .
3. Resolver el sistema triangular superior $\widehat{R}x = \widehat{Q}^*b$ para x .

TAREA: Programe este algoritmo y verifique su correcta operación mediante algunos ejemplos sencillos. Compruebe que cuando se aplican las reflexiones de Householder para obtener la factorización QR, la complejidad de este algoritmo es del orden de $2mn^2 - \frac{2}{3}n^3$ "flops".

5) *Cuadrados mínimos mediante la factorización SVD.* Si se dispone de una factorización SVD reducida de A , $A = \widehat{U}\widehat{\Sigma}V^*$, donde las columnas de U son ortonormales, Σ es una matriz diagonal cuadrada y V es una matriz unitaria, se puede diseñar un algoritmo análogo al precedente. En efecto, cuando A es de rango completo, Σ y V son invertibles y el proyector ortogonal P sobre $\text{Range}(A)$ adopta entonces la forma:

$$\begin{aligned} P &= A(A^*A)^{-1}A^* = \widehat{U}\widehat{\Sigma}V^* \left(V\widehat{\Sigma}\widehat{U}^*\widehat{U}\widehat{\Sigma}V^* \right)^{-1} V\widehat{\Sigma}\widehat{U}^* \\ &= \widehat{U}\widehat{\Sigma}V^*(V^*)^{-1}\widehat{\Sigma}^{-1}\widehat{\Sigma}^{-1}V^{-1}V\widehat{\Sigma}\widehat{U}^* = \widehat{U}\widehat{U}^*. \end{aligned}$$

La ecuación $Ax = Pb$ queda ahora:

$$\widehat{U}\widehat{\Sigma}V^*x = \widehat{U}\widehat{U}^*b,$$

de donde resulta:

$$x = V\widehat{\Sigma}^{-1}\widehat{U}^*b,$$

y, consecuentemente,

$$A^+ = V\widehat{\Sigma}^{-1}\widehat{U}^*,$$

El algoritmo vía SVD se resume del siguiente modo:

Cuadrados mínimos mediante factorización SVD.

1. Calcular la factorización SVD reducida $A = \widehat{U}\widehat{\Sigma}V^*$.
2. Calcular el vector \widehat{U}^*b .
3. Resolver el sistema diagonal $\widehat{\Sigma}w = \widehat{U}^*b$ para w .
4. Calcular $x = Vw$.

TAREA: Programe este algoritmo y verifique su correcta operación mediante algunos ejemplos sencillos.

2. [5, p. 85, exc. 11.1] Suponga que la matriz $A \in M(m \times n, \mathbb{C})$ tiene la forma $A = \begin{bmatrix} A_1 \\ A_2 \end{bmatrix}$,

donde $A_1 \in M(n \times n, \mathbb{C})$ es una matriz no singular y $A_2 \in M((m - n) \times n, \mathbb{C})$. Demuestre que $\|A^+\|_2 \leq \|A_1^{-1}\|_2$, donde A^+ denota el (pseudo-) inverso de Moore-Penrose de A y

$$\|A\|_2 = \sup_{0 \neq x \in \mathbb{C}^n} \frac{\|Ax\|_2}{\|x\|_2} \text{ con } \|x\|_2 = \left(\sum_{k=1}^n |x_k|^2 \right)^{1/2}, \text{ como es usual.}$$

Nota: Como el lector sabe, el (pseudo-) inverso de Moore-Penrose de A , cuando A es de rango completo (i.e., $\text{rank } A = \min\{m, n\}$), se define como $A^+ = (A^*A)^{-1}A^* \in M(n \times m, \mathbb{C})$.

3. [5, p. 85, exc. 11.2] (a) En el sentido de la norma L^2 sobre el intervalo $[1, 2] \subset \mathbb{R}$, ¿cuál es la mejor aproximación a la función $f(x) = x^{-1}$ que se puede lograr mediante una combinación lineal de las funciones e^x , $\sin x$ y $\Gamma(x)$?

Nota: Discuta el problema tanto desde el punto de vista teórico como computacional. La función "gama" $\Gamma(x)$, o *integral de Euler de segunda especie*, que Ud. conoce de sus cursos

de Matemática, se define como la integral $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$ que converge para $x > 0$. Mayor información acerca de esta función puede Ud. obtener en [1]. La función $\Gamma(x)$ ya está implementada en MATLAB, de modo que no es necesario que Ud. la implemente. Escriba un programa que responda esta pregunta garantizando un error relativo de, a lo más, 3%. Utilice una discretización apropiada del intervalo $[1, 2]$ y una versión discreta del problema de los cuadrados mínimos (por ejemplo, aproximando las integrales involucradas mediante el conocido método de la regla de Simpson). Su programa debe entregar una estimación de la respuesta, los coeficientes de la combinación lineal optimal, y un gráfico computacional que ilustre esa aproximación lineal optimal.

(b) Repita todo lo anterior pero ahora trabaje sobre el intervalo $[0, 1]$ en lugar de $[1, 2]$. La siguiente relación puede serle útil: si $g(x) = 1/\Gamma(x)$, entonces $g'(0) = 1$.

4. [3, p. 99, Theor. 4.64] Sea $A \in M(m \times n, \mathbb{R})$, $1 \leq n \leq m$, una matriz dada con rango maximal n , y sea $A = QS$ una factorización de A tal que $Q \in M(m \times m, \mathbb{R})$ es una matriz cuadrada ortogonal y $S \in M(m \times n, \mathbb{R})$ es una matriz (rectangular!) triangular superior:

$$S = \begin{bmatrix} R \\ Z \end{bmatrix} \in M(m \times n, \mathbb{R}), \quad R = \begin{bmatrix} r_{11} & r_{12} & \dots & r_{1,n-1} & r_{1,n} \\ 0 & r_{22} & \dots & r_{2,n-1} & r_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \dots & r_{n,n-1} & r_{n,n} \\ 0 & 0 & \dots & 0 & r_{n,n} \end{bmatrix} \in M(n \times n, \mathbb{R}),$$

y $Z \in M(m - n \times n, \mathbb{R})$ es una matriz cuyos coeficientes son todos nulos. (Apreciado lector, ¿cómo obtendría Ud. las matrices Q, R, S mencionadas?)

Para un vector dado $b \in \mathbb{R}^m$ considere el vector $Q^T \cdot b$ particionado en la forma:

$$Q^T \cdot b = \begin{bmatrix} c \\ d \end{bmatrix} \in M(m \times n, \mathbb{R}), \quad c \in \mathbb{R}^n, \quad d \in \mathbb{R}^{m-n}.$$

Entonces para $x^* \in \mathbb{R}^n$ es válida la equivalencia:

$$\|A \cdot x^* - b\|_2 = \min_{x \in \mathbb{R}^n} \|A \cdot x - b\|_2 \Leftrightarrow R \cdot x^* = c. \tag{1}$$

Tarea. Demuestre (1). Estudie el caso $A \in M(m \times n, \mathbb{C})$, $b \in \mathbb{C}^m$. Construya un ejemplo computacional no trivial que ilustre la situación descrita.

2. ARITMÉTICA DE PUNTO FLOTANTE.

1. Hurgando en la *web* dimos con la siguiente interesante contribución de Steve Hollasch. Estúdiela críticamente, así como el texto de Stallings en el cual se basa.

IEEE Standard 754 Floating Point Numbers. By Steve Hollasch (2003.Jan.02).

[http:// research.microsoft.com/~hollasch/cgindex/coding/ieeefloat.html](http://research.microsoft.com/~hollasch/cgindex/coding/ieeefloat.html)

<http://standards.ieee.org/>

IEEE Standard 754 floating point is the most common representation today for real numbers on computers, including Intel-based PC's, Macintoshes, and most Unix platforms. This article gives a brief overview of IEEE floating point and its representation. Discussion of arithmetic implementation may be found in the book mentioned at the bottom of this article.

What Are Floating Point Numbers? There are several ways to represent real numbers on computers. Fixed point places a radix point somewhere in the middle of the digits, and is equivalent to using integers that represent portions of some unit. For example, one might

represent 1/100ths of a unit; if you have four decimal digits, you could represent 10,82, or 00,01. Another approach is to use rationals, and represent every number as the ratio of two integers.

Floating-point representation –the most common solution– basically represents reals in scientific notation. Scientific notation represents numbers as a base number and an exponent. For example, 123,456 could be represented as $1,23456 \times 10^2$. In hexadecimal, the number 123.abc might be represented as $1,23abc \times 16^2$.

Floating-point solves a number of representation problems. Fixed-point has a fixed window of representation, which limits it from representing very large or very small numbers. Also, fixed-point is prone to a loss of precision when two large numbers are divided.

Floating-point, on the other hand, employs a sort of “sliding window” of precision appropriate to the scale of the number. This allows it to represent numbers from 1,000,000,000,000 to 0.0000000000000001 with ease.

Storage Layout. IEEE floating point numbers have three basic components: the sign, the exponent, and the mantissa. The mantissa is composed of the fraction and an implicit leading digit (explained below). The exponent base (2) is implicit and need not be stored.

The following figure shows the layout for single (32-bit) and double (64-bit) precision floating-point values. The number of bits for each field are shown (bit ranges are in square brackets):

	Sign	Exponent	Fraction	Bias
Single Precision	1 [31]	8 [30-23]	23 [22-00]	127
Double Precision	1 [63]	11 [62-52]	52 [51-00]	1023

The Sign Bit. The sign bit is as simple as it gets. 0 denotes a positive number; 1 denotes a negative number. Flipping the value of this bit flips the sign of the number.

The Exponent. The exponent field needs to represent both positive and negative exponents. To do this, a bias is added to the actual exponent in order to get the stored exponent. For IEEE single-precision floats, this value is 127. Thus, an exponent of zero means that 127 is stored in the exponent field. A stored value of 200 indicates an exponent of (200-127), or 73. For reasons discussed later, exponents of -127 (all 0s) and +128 (all 1s) are reserved for special numbers. For double precision, the exponent field is 11 bits, and has a bias of 1023.

The Mantissa. The mantissa, also known as the significand, represents the precision bits of the number. It is composed of an implicit leading bit and the fraction bits.

To find out the value of the implicit leading bit, consider that any number can be expressed in scientific notation in many different ways. For example, the number five can be represented as any of these:

$$5,00 \times 10^0 \qquad 0,05 \times 10^2 \qquad 5000 \times 10^{-3}$$

In order to maximize the quantity of representable numbers, floating-point numbers are typically stored in normalized form. This basically puts the radix point after the first non-zero digit. In normalized form, five is represented as $5,0 \times 10^0$.

A nice little optimization is available to us in base two, since the only possible non-zero digit is 1. Thus, we can just assume a leading digit of 1, and don't need to represent it explicitly. As a result, the mantissa has effectively 24 bits of resolution, by way of 23 fraction bits.

Putting it All Together. So, to sum up:

- (1) The sign bit is 0 for positive, 1 for negative.
- (2) The exponent's base is two.
- (3) The exponent field contains 127 plus the true exponent for single-precision, or 1023 plus the true exponent for double precision.
- (4) The first bit of the mantissa is typically assumed to be 1. f , where f is the field of fraction bits.

Ranges of Floating-Point Numbers. Let's consider single-precision floats for a second. Note that we're taking essentially a 32-bit number and re-jiggering the fields to cover a much broader range. Something has to give, and it's precision. For example, regular 32-bit integers, with all precision centered around zero, can precisely store integers with 32-bits of resolution. Single-precision floating-point, on the other hand, is unable to match this resolution with its 24 bits. It does, however, approximate this value by effectively truncating from the lower end. For example:

$$\begin{array}{ll}
 11110000\ 11001100\ 10101010\ 00001111 & \text{32-bit integer} \\
 = +1.1110000\ 11001100\ 10101010 \times 2^{31} & \text{Single-precision float} \\
 = 11110000\ 11001100\ 10101010\ 00000000 & \text{Corresponding value}
 \end{array}$$

This approximates the 32-bit value, but doesn't yield an exact representation. On the other hand, besides the ability to represent fractional components (which integers lack completely), the floating-point value can represent numbers around 2^{127} , compared to 32-bit integers maximum value around 2^{32} .

The range of positive floating point numbers can be split into normalized numbers (which preserve the full precision of the mantissa), and denormalized numbers (discussed later) which use only a portion of the fractions's precision.

	Single Precision	Double Precision
Denormalized	$\pm 2^{-149}$ to $(1 - 2^{-23}) \times 2^{-126}$	$\pm 2^{-1074}$ to $(1 - 2^{-52}) \times 2^{-1022}$
Normalized	$\pm 2^{-126}$ to $(2 - 2^{-23}) \times 2^{127}$	$\pm 2^{-1022}$ to $(2 - 2^{-52}) \times 2^{1023}$
Approximate Decimal	$\sim \pm 10^{-44,85}$ to $\sim 10^{38,53}$	$\sim \pm 10^{-323,3}$ to $\sim 10^{308,3}$

Since the sign of floating point numbers is given by a special leading bit, the range for negative numbers is given by the negation of the above values.

There are five distinct numerical ranges that single-precision floating-point numbers are not able to represent:

- (1) Negative numbers less than $-(2 - 2^{-23}) \times 2^{127}$ (negative overflow)
- (2) Negative numbers greater than -2^{-149} (negative underflow)
- (3) Zero
- (4) Positive numbers less than 2^{-149} (positive underflow)
- (5) Positive numbers greater than $(2 - 2^{-23}) \times 2^{127}$ (positive overflow)

Overflow means that values have grown too large for the representation, much in the same way that you can overflow integers. Underflow is a less serious problem because it just denotes a loss of precision, which is guaranteed to be closely approximated by zero.

Here's a table of the effective range (excluding infinite values) of IEEE floating-point numbers:

Operation	Result
$n / \pm \infty$	0
$\pm \infty \times \pm \infty$	$\pm \infty$
$\pm \text{nonzero} / 0$	$\pm \infty$
$\infty + \infty$	∞
$\pm 0 / \pm 0$	NaN
$\infty - \infty$	NaN
$\pm \infty / \pm \infty$	NaN
$\pm \infty \times 0$	NaN

FIGURA 1. Table on operations on special numbers.

	Binary	Decimal
Single	$\pm(2 - 2^{-23})^{127}$	$\sim \pm 10^{38,53}$
Double	$\pm(2 - 2^{-52})^{1023}$	$\sim \pm 10^{308,25}$

Note that the extreme values occur (regardless of sign) when the exponent is at the maximum value for finite numbers (2^{127} for single-precision, 2^{1023} for double), and the mantissa is filled with 1's (including the normalizing 1 bit).

Special Values. IEEE reserves exponent field values of all 0's and all 1's to denote special values in the floating-point scheme.

Zero. As mentioned above, zero is not directly representable in the straight format, due to the assumption of a leading 1 (we'd need to specify a true zero mantissa to yield a value of zero). Zero is a special value denoted with an exponent field of zero and a fraction field of zero. Note that -0 and $+0$ are distinct values, though they both compare as equal.

Denormalized. If the exponent is all 0's, but the fraction is non-zero (else it would be interpreted as zero), then the value is a denormalized number, which does not have an assumed leading 1 before the binary point. Thus, this represents a number $(-1)^s \times 0.f \times 2^{-126}$, where s is the sign bit and f is the fraction. For double precision, denormalized numbers are of the form $(-1)^s \times 0.f \times 2^{-1022}$. From this you can interpret zero as a special type of denormalized number.

Infinity. The values $+\infty$ and $-\infty$ are denoted with an exponent of all 1's and a fraction of all 0's. The sign bit distinguishes between negative infinity and positive infinity. Being able to denote infinity as a specific value is useful because it allows operations to continue past overflow situations. Operations with infinite values are well defined in IEEE floating point.

Not A Number. The value NaN (*Not a Number*) is used to represent a value that does not represent a real number. NaN's are represented by a bit pattern with an exponent of all 1's and a non-zero fraction. There are two categories of NaN's: QNaN (*Quiet NaN*) and SNaN (*Signalling NaN*).

A QNaN is a NaN with the most significant fraction bit set. QNaN's propagate freely through most arithmetic operations. These values pop out of an operation when the result is not mathematically defined.

Float Values ($b = \text{bias}$)			
Sign	Exponent (e)	Fraction (f)	Value
0	00...00	00...00	+0
0	00...00	00...01 ⋮ 11...11	Positive Denormalized Real $0.f \times 2^{(-b+1)}$
0	00...01 ⋮ 11...10	XX...XX	Positive Normalized Real $1.f \times 2^{(e-b)}$
0	11...11	00...00	$+\infty$
0	11...11	00...01 ⋮ 01...11	SNaN
0	11...11	10...00 ⋮ 11...11	QNaN
1	00...00	00...00	-0
1	00...00	00...01 ⋮ 11...11	Negative Denormalized Real $-0.f \times 2^{(-b+1)}$
1	00...01 ⋮ 11...10	XX...XX	Negative Normalized Real $-1.f \times 2^{(e-b)}$
1	11...11	00...00	$-\infty$
1	11...11	00...01 ... 01...11	SNaN
1	11...11	10...00 ⋮ 11...11	QNaN

FIGURA 2. Float Values ($b = \text{bias}$).

An SNaN is a NaN with the most significant fraction bit clear. It is used to signal an exception when used in operations. SNaN's can be handy to assign to uninitialized variables to trap premature usage.

Semantically, QNaN's denote indeterminate operations, while SNaN's denote invalid operations.

Special Operations. Operations on special numbers are well-defined by IEEE. In the simplest case, any operation with a NaN yields a NaN result. Other operations are presented in Figure 1.

Summary. To sum up, the corresponding values for a given representation are presented in Figure 2.

References. A lot of this stuff was observed from small programs I (Hollash) wrote to go back and forth between hex and floating point (printf-style), and to examine the results of various operations. The bulk of this material, however, was lifted from Stallings' book.

- [1] William Stallings. *Computer Organization and Architecture*. Macmillan Publishing Company, ISBN 0-02-415480-6. Pp. 222–234.
Versión en Castellano: *Organización y Arquitectura de Computadores*. Quinta edición. Traducción: Antonio Cañas Vargas, Julio Ortega Lopera, Francisco José Pelayo Valle, Beatriz Prieto Campos; coordinación y revisión técnica: Alberto Prieto Espinosa; todos del Departamento de arquitectura y tecnología de computadores de la Universidad de Granada. Pearson Educación S.A./Prentice Hall, Madrid, 2000.
- [2] IEEE Computer Society (1985). *IEEE Standard for Binary Floating-Point Arithmetic*. IEEE Std 754-1985.
- [3] Intel Architecture Software Developer's Manual, Volume 1: Basic Architecture (a PDF document downloaded from intel.com).
- [4] IEEE Standards Site

©2003, Steve Hollasch

2. [5, p. 101, exc. 13.1] En el sistema aritmético IEEE, ¿cuántos números de doble precisión hay entre dos números no nulos *adyacentes* de precisión simple? Recuerde que el sistema *idealizado* de números de punto flotante se define como:

$$\mathbb{F} = \{0\} \cup \{\pm(m/\beta^t)\beta^e : m \in \mathbb{N}, \beta^{t-1} \leq m \leq \beta^t - 1, e \in \mathbb{Z}\},$$

donde la base β es típicamente 2 (en general, $\beta \geq 2$) y la precisión t , en el sistema aritmético IEEE, es 24 para precisión simple y 53 para precisión doble.

3. [5, p. 101, exc. 13.2] (slightly modified!). El sistema de punto flotante \mathbb{F} contiene muchos enteros pero no todos ellos. (a) Obtenga una fórmula exacta para el menor entero positivo n que no pertenece a \mathbb{F} .

(b) Calcule n para las aritméticas de precisión IEEE simple y doble, respectivamente.

(c) Determine $\sum_{0 < k \in \mathbb{F}} 1/k$ y $\sum_{0 < k \in \mathbb{Z} \setminus \mathbb{F}} 1/k$.

(d) Sea $\{n_k\}_{k \in \mathbb{N}}$ la sucesión (estrictamente creciente) de *todos* los números naturales con las propiedades $n_k - 1 \in \mathbb{F}$ y $n_k \notin \mathbb{F}$. Determine $\sum_{k \in \mathbb{N}} 1/n_k$.

4. [5, p. 101, exc. 13.3] (slightly modified!). Obtenga sendas estimaciones L y U para el *infimum* y el *supremum* de la expresión

$$A = (x - 2)^9 - x^9 + 18x^8 - 144x^7 + 672x^6 - 2016x^5 + 4032x^4 - 5376x^3 + 4608x^2 - 2304x + 512,$$

sobre el intervalo $[-1, 920; 2, 080]$ computando mediante un sistema aritmético IEEE de precisión simple. ¿Podría disminuir el valor de la diferencia $U - L$ si calcula el polinomio expandido mediante el esquema de Horner?

5. [2, Ch. 1, Sec. 2, pp. 17–21] Los errores aritméticos están inevitablemente presentes en los cálculos efectuados mediante computadores digitales. La llamada *aritmética de punto flotante* es una de las formas estándar de la computación numérica. Aunque los computadores digitales usualmente representan los números en forma binaria, (la mayor parte de) los humanos pensamos (hoy día) en términos de una representación decimal. Por esta razón, el presente ejercicio está planteado en términos “decimales”. No obstante, el (la) estudiante *despierto(a)* no dejará pasar la oportunidad de: (i) investigar qué pasa si se sustituye la base decimal $\beta = 10$ por una base cualquiera $2 \leq \beta \in \mathbb{N}$, (ii) investigar más detenidamente la norma

del IEEE relativa a la representación interna de números en los computadores digitales de la presente generación, en particular lo concerniente a representaciones de precisión simple (24 bits) y precisión doble (53 bits).

Si $a \neq 0$ tiene una representación decimal exacta:

$$a = \pm 0.d_1d_2d_3 \cdots \times 10^q \quad q \in \mathbb{Z}, \quad d_i \in \mathbb{N}_0, \quad 0 \leq d_i \leq 9, \quad d_1 \neq 0, \quad (2)$$

entonces la *representación decimal de punto flotante de a con t dígitos* o, brevemente, el *flotante* del a tiene la forma:

$$\mathbf{fl}(a) = \pm 0.\delta_1\delta_2\delta_3 \dots \delta_t \times 10^q \quad q \in \mathbb{Z}, \quad \delta_i \in \mathbb{N}_0, \quad 0 \leq \delta_i \leq 9, \quad \delta_1 \neq 0 \quad t \in \mathbb{N}. \quad (3)$$

El número $0.\delta_1\delta_2\delta_3 \dots \delta_t$ se llama *mantisa* y q *exponente* de $\mathbf{fl}(a)$. En discusiones teóricas se supone a menudo que el exponente q puede adoptar cualquier valor entero, aunque esto, obviamente, no es muy realista. En la práctica, usualmente, q satisface una restricción de la forma:

$$-N \leq q \leq M, \quad N, M \in \mathbb{N} \quad \text{“grandes”}. \quad (4)$$

Si para un número $a \neq 0$ se tiene $q \notin [-N, M]$, entonces el flotante de a no está definido. Si en el curso de una computación se obtiene $q < -N$, se habla de “*underflow*” y si $q > M$ se dice que ha ocurrido un “*overflow*”. Hay dos maneras populares de obtener los dígitos flotantes δ_i a partir de los dígitos exactos d_i : *truncamiento* y *redondeo*. Para el *truncamiento* se tiene:

$$\delta_i = d_i, \quad i = 1, 2, \dots, t, \quad (5)$$

es decir, simplemente se omiten todos los dígitos con índice mayor que t . Para el *redondeo* se tiene:

$$\delta_1\delta_2\delta_3 \dots \delta_t = [d_1d_2d_3 \dots d_t . d_{t+1} + 0,5], \quad (6)$$

donde $[x]$ denota la parte entera de $x \in \mathbb{R}$ (note la ubicación del punto decimal en el miembro derecho de (6)). En ambos casos se puede obtener una cota para el error introducido. En efecto:

Lema 1. *El error (absoluto) en la representación de punto flotante de un número decimal $a \neq 0$ con mantisa de t dígitos viene acotado por:*

$$|a - \mathbf{fl}(a)| \leq \begin{cases} 5|a|10^{-t}, & \text{caso de redondeo,} \\ |a|10^{-t+1}, & \text{caso de truncamiento.} \end{cases} \quad (7)$$

(a) Demuestre (o corrija) el Lema 1.

Supongamos ahora que la unidad aritmética del computador ideal con el que trabajamos, es capaz de realizar *internamente* las operaciones básicas con $2t$ dígitos correctos y luego los transfiere al *buffer* de salida como un número de punto flotante de t dígitos, ya sea redondeando o bien truncando. Entonces se tiene:

$$\left. \begin{aligned} \mathbf{fl}(a \pm b) &= (a \pm b)(1 + \eta 10^{-t}), \\ \mathbf{fl}(ab) &= ab(1 + \eta 10^{-t}), \\ \mathbf{fl}\left(\frac{a}{b}\right) &= \frac{a}{b}(1 + \eta 10^{-t}), \end{aligned} \right\}, \quad \text{donde} \quad \begin{cases} 0 \leq |\eta| \leq 5, & \text{caso de redondeo,} \\ 0 \leq |\eta| \leq 10, & \text{caso de truncamiento,} \end{cases} \quad (8)$$

(b) Demuestre (o corrija) (8).

En muchos cálculos, particularmente en los concernientes a sistemas lineales se requiere la acumulación de productos parciales (e.g., en el cálculo del producto interno de dos vectores). Si se supone que tras cada multiplicación ocurre redondeo o truncamiento, al igual que después de

cada suma sucesiva, entonces el *flotante* del producto interno de dos vectores $a = (a_1, \dots, a_n)$, $b = (b_1, \dots, b_n)$, se puede expresar recursivamente mediante la fórmula:

$$\mathbf{fl} \left(\sum_{i=1}^n a_i b_i \right) = \mathbf{fl} \left[\mathbf{fl} \left(\sum_{i=1}^{n-1} a_i b_i \right) + \mathbf{fl}(a_n b_n) \right]. \quad (9)$$

El resultado de tales cálculos se puede representar como un producto interno exacto alterando levemente, por ejemplo, los a_i 's:

Lema 2. *Supóngase que la dimensión n de los vectores y el número t de dígitos de las mantisas en nuestro computador ideal satisfacen:*

$$n 10^{1-t} \leq 1. \quad (10)$$

Entonces, para el producto interno de punto flotante (9) computado mediante redondeo sistemático en todas las operaciones se tiene:

$$\mathbf{fl} \left(\sum_{i=1}^n a_i b_i \right) = \sum_{i=1}^n (a_i + \delta a_i) b_i, \quad (11)$$

donde

$$|\delta a_i| \leq n |a_i| 10^{1-t}, \quad |\delta a_i| \leq (n - i + 2) |a_i| 10^{1-t}, \quad i = 2, 3, \dots, n. \quad (12)$$

(c) Demuestre (o corrija) (11), (12). Investigue qué ocurre si se en el Lema 2 se sustituye “redondeo” por “truncamiento”.

6. [2, Ch. 1, Sec. 2, p. 21, probl. 2]

(a) Obtenga una representación (¡fórmula!) para $\mathbf{fl} \left(\sum_{i=1}^n c_i \right)$ (en términos de los $\mathbf{fl}(c_i)$).

(b) Si $c_1 > c_2 > \dots > c_n > 0$, ¿en qué orden habría que calcular $\mathbf{fl} \left(\sum_{i=1}^n c_i \right)$ para minimizar los efectos de redondeo (resp., truncamiento)?

3. ESTABILIDAD Y CONDICIONAMIENTO.

1. Consideremos el sistema de ecuaciones lineales:

$$A.x = b, \quad A \in M(n \times n, \mathbb{C}), \quad \text{vector incógnitas: } x \in \mathbb{C}^n, \quad \text{vector datos: } b \in \mathbb{C}^n \quad (13)$$

con solución única, i.e., con $\det A \neq 0$. Queremos estudiar la sensibilidad de la solución x de (13) con respecto a pequeñas perturbaciones en el vector de datos b . Con este objetivo, consideremos el sistema perturbado:

$$A.\tilde{x} = b + \delta b, \quad \text{perturbación: } \delta b \in \mathbb{C}^n, \quad (14)$$

que se puede escribir en la forma:

$$A.\delta x = \delta b, \quad \text{o bien } \delta x = A^{-1}.\delta b, \quad \text{donde } \delta x := \tilde{x} - x \in \mathbb{C}^n. \quad (15)$$

El vector de *error absoluto* $\delta x \in \mathbb{C}^n$ no es necesariamente es unitario. Tomando normas en (15) se obtiene:

$$\|\delta b\| = \|A.\delta x\| \leq \|A\| \|\delta x\|, \quad \|\delta x\| = \|A^{-1}.\delta b\| \leq \|A^{-1}\| \|\delta b\|. \quad (16)$$

De (16) se obtienen sendas cotas inferior y superior para el error relativo de $\|\delta x\|/\|x\|$ de x :

$$\frac{\|\delta b\|}{\|A\| \|x\|} \leq \frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\| \|\delta b\|}{\|x\|}. \quad (17)$$

Análogamente, tomando normas en (13) se obtiene:

$$\|b\| = \|A.x\| \leq \|A\| \|x\|, \quad \|x\| = \|A^{-1}.b\| \leq \|A^{-1}\| \|b\|. \quad (18)$$

Combinando (17) con (18) se obtiene:

$$\frac{1}{\|A\| \|A^{-1}\|} \frac{\|\delta b\|}{\|b\|} \leq \frac{\|\delta x\|}{\|x\|} \leq \|A\| \|A^{-1}\| \frac{\|\delta b\|}{\|b\|}. \quad (19)$$

El número:

$$\kappa(A) := \|A\| \|A^{-1}\| \quad (20)$$

se denomina *número de condición* de la matriz A .

Nótese que en la definición (20) el número de condición depende de la norma matricial utilizada. A veces se define un número de condición independiente de toda norma matricial en términos del *radio espectral* $\rho(A)$ de la matriz A , mediante la expresión:

$$\kappa_*(A) := \rho(A) \rho(A^{-1}) = \frac{\max_{\lambda \in \sigma(A)} |\lambda|}{\min_{\lambda \in \sigma(A)} |\lambda|} \leq \kappa(A) \quad (21)$$

donde

$$\sigma(A) := \{\lambda \in \mathbb{C} : \lambda \text{ es valor propio de } A\} \quad (22)$$

es el llamado *spectrum* de la matriz A .

OBSERVACIÓN. 1) El número de condición $\kappa(A)$ está acotado inferiormente por 1:

$$1 = \|I\| = \|A.A^{-1}\| \leq \|A\| \|A^{-1}\| = \kappa(A).$$

2) De (19) se observa que si $\kappa(a)$ es un número muy próximo a 1, entonces perturbaciones pequeñas $\delta b \in \mathbb{C}^n$ en $b \in \mathbb{C}^n$ conducen a pequeñas perturbaciones $\delta x \in \mathbb{C}^n$ en $x \in \mathbb{C}^n$. Por el contrario, si $\kappa(a) \gg 1$, entonces pequeñas perturbaciones en b pueden producir grandes variaciones en x .

3) Como el número de condición $\kappa(A)$ varía con la norma matricial utilizada, a veces es necesario especificarla explícitamente para evitar confusiones.

4) Los números de condición $\kappa(A)$ son buenos predictores del mal condicionamiento de un sistema de ecuaciones lineales $A.x = b$ cuando los coeficientes de la matriz A no varían mucho en magnitud. Por el contrario, si, por ejemplo, $A = \begin{bmatrix} 1 & 0 \\ 0 & 10^{10} \end{bmatrix}$ con $A^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 10^{-10} \end{bmatrix}$, entonces $\kappa_1(A) = \|A\|_1 \|A^{-1}\|_1 = 10^{10}$. Pero de aquí no puede concluirse que el sistema esté mal condicionado ya que la matriz A es diagonal y tales sistemas nunca pueden estar mal condicionados. \square

EJEMPLO. Consideremos el sistema $A.x = b$:

$$\begin{cases} 7x_1 + 10x_2 = 1 \\ 5x_1 + 7x_2 = 0,7 \end{cases}, \quad \text{de modo que } A = \begin{bmatrix} 7 & 10 \\ 5 & 7 \end{bmatrix}, \quad A^{-1} = \begin{bmatrix} -7 & 10 \\ 5 & -7 \end{bmatrix}.$$

La solución exacta de este sistema es $x_1 = 0, x_2 = 0,1$, y se tiene:

$$\begin{cases} \kappa_1(A) := \|A\|_1 \|A^{-1}\|_1 = \max\{12, 17\} \cdot \max\{12, 17\} = 289, \\ \kappa_2(A) := \|A\|_2 \|A^{-1}\|_2 = \sqrt{7^2 + 10^2 + 5^2 + 7^2} \cdot \sqrt{(-7)^2 + 10^2 + 5^2 + (-7)^2} = 223, \end{cases}$$

donde, en la primera ecuación se ha utilizado la norma matricial l_1 y en la segunda, la norma l_2 . En ambos casos se tiene $\kappa(A) \gg 1$, de modo que el sistema está mal condicionado.

Consideremos ahora el sistema perturbado:

$$\begin{cases} 7\tilde{x}_1 + 10\tilde{x}_2 = 1,01, \\ 5\tilde{x}_1 + 7\tilde{x}_2 = 0,69, \end{cases}$$

cuya solución es $\tilde{x}_1 = -0,17, \tilde{x}_2 = 0,22$, lo que corresponde a una perturbación relativa de x dada por:

$$\frac{\|\delta x\|_1}{\|x\|_1} = \frac{\|\tilde{x} - x\|_1}{\|x\|_1} = \frac{|-0,17 - 0,00| + |0,22 - 0,10|}{|0,00| + |0,10|} = \frac{0,29}{0,10} = 2,9,$$

que es bastante grande comparado con la correspondiente perturbación en el vector de datos b :

$$\frac{\|\delta b\|_1}{\|b\|_1} = \frac{\|\tilde{b} - b\|_1}{\|b\|_1} = \frac{|1,01 - 1,00| + |0,69 - 0,70|}{|1,00| + |0,70|} = \frac{0,02}{1,70} = 0,011764705.$$

De este modo, la perturbación en x representa una amplificación de $\frac{2,9}{0,011764705} = 246,50$ veces la perturbación en b . El sistema planteado, en consecuencia, está mal condicionado. \square

Teorema 2. Sea $A \in M(n \times n, \mathbb{C})$ una matriz no singular. Entonces:

$$\frac{1}{\kappa(A)} = \inf \left\{ \frac{\|A - B\|}{\|A\|} : B \in M(n \times n, \mathbb{C}) \text{ es singular} \right\}. \quad (23)$$

Demostración. Sea $B \in M(n \times n, \mathbb{C})$ una matriz singular. Entonces existe $x \in \mathbb{C}^n, x \neq 0 \in \mathbb{C}^n$, tal que $B.x = 0$. Luego, para este x se tiene:

$$\begin{cases} \|(A - B).x\| = \|A.x\| = \frac{\|A^{-1}\| \|A.x\|}{\|A^{-1}\|} \geq \frac{\|A^{-1} A.x\|}{\|A^{-1}\|} = \frac{\|x\|}{\|A^{-1}\|}, \\ \|(A - B).x\| \leq \|(A - B)\| \|x\|. \end{cases}$$

De aquí resulta:

$$\frac{\|x\|}{\|A^{-1}\|} \leq \|A - B\| \|x\|, \quad \text{i.e.,} \quad \frac{1}{\|A^{-1}\|} \leq \|A - B\|, \quad \text{pues } x \neq 0.$$

Por consiguiente:

$$\frac{1}{\kappa(A)} = \frac{1}{\|A^{-1}\| \|A\|} \leq \frac{\|A - B\|}{\|A\|} \quad \text{para toda matriz } B \in M(n \times n, \mathbb{C}) \text{ singular.}$$

lo que demuestra la desigualdad “ \leq ” en (23). Ahora sólo resta probar la desigualdad “ \geq ” en (23), lo que queda de ejercicio para el (la) amable lector(a). \square

TAREA. Complete la demostración del Teorema 2.

OBSERVACIÓN. El Teorema 2 pone de manifiesto que un valor grande de $\kappa(A)$ es equivalente a la posibilidad de aproximar bien una matriz no singular dada A mediante una matriz singular B . Nótese que desde el punto de vista expuesto, el número de condición de una matriz singular es ∞ .

ESTIMACIÓN DEL NÚMERO DE CONDICIÓN. La estimación del número de condición de una matriz requiere estimar la norma de esa matriz y su inversa. Si se utiliza cualquiera de las normas matriciales que se expresan en términos de los coeficientes de las matrices (como la norma de Frobenius-Schur-Hilbert, por ejemplo), el cálculo del correspondiente $\kappa(A)$ es directo y no ofrece complicación.

Consideremos, entonces, el problema general de estimar la norma de una matriz dada *en cuanto operador lineal*. Para simplificar plantaremos la discusión para matrices reales y consideraremos la norma l_2 en \mathbb{R} . Sea, entonces, $A \in M(n \times n, \mathbb{R})$. En vista de la definición de la norma de A en cuanto operador lineal:

$$\|A\| = \sup_{0 \neq x \in \mathbb{R}^n} \frac{\|A.x\|}{\|x\|} = \sup_{x \in \mathbb{R}^n, \|x\|=1} \|A.x\|, \quad (24)$$

es claro que el problema se puede plantear como un problema de optimización con condición subsidiaria:

$$\text{maximizar } y = \|A.x\| \quad \text{bajo la restricción } \|x\| = 1. \quad (25)$$

Evidentemente, (25) es equivalente a:

$$\text{maximizar } z = \|A.x\|^2 = x^T A^T A x \quad \text{bajo la restricción } x^T x = 1. \quad (26)$$

Sea $B = A^T A = [b_{ij}]_{n \times n}$. Evidentemente, la forma cuadrática $z = \sum_{i,j} b_{ij} x_i x_j$ es positiva definida (¡es una norma!). El problema (26) es un ejercicio “standard” de multiplicadores de Lagrange (que conduce a un interesante problema de valores propios que no discutiremos aquí) pero también podría abordarse mediante el siguiente procedimiento algorítmico:

ALGORITMO. Un algoritmo para resolver (26).

- (i) Elegir un $x_0 \in \mathbb{R}^n$ y hacer $j \leftarrow -1$.
- (ii) Hacer $j \leftarrow j + 1$.
- (iii) Calcular $\mathbf{grad} z(x_j)$.
- (iv) Calcular $t_0 \in \mathbb{R}$ tal que maximice la función de una sola variable real:

$$\varphi(t) := \frac{\|A.(x_j + t \mathbf{grad} z(x_j))\|}{\|x_j + t \mathbf{grad} z(x_j)\|}, \quad t \in \mathbb{R}.$$

- (v) Hacer $x_{j+1} \leftarrow \frac{x_j + t_0 \mathbf{grad} z(x_j)}{\|x_j + t_0 \mathbf{grad} z(x_j)\|}$.
- (vi) Calcular $\|A.x_{j+1}\|$ y $\alpha := \left| \frac{\|A.x_{j+1}\| - \|A.x_j\|}{\|A.x_j\|} \right|$.
- (vii) Si $\alpha > \mathbf{eps}$, donde \mathbf{eps} representa un criterio de detención, volver al paso (ii) y repetir el bucle. Si $\alpha < \mathbf{eps}$, interrumpir el bucle e imprimir $\|A.x_{j+1}\|$.

El cálculo del número de condición de una matriz invertible $A \in M(n \times n, \mathbb{R})$ ahora no ofrece ninguna dificultad.

TAREA. (a) Resuelva (26) mediante multiplicadores de Lagrange.

(b) Implemente computacionalmente el algoritmo descrito más arriba para calcular la norma de una matriz en cuanto operador lineal. El algoritmo discutido es simplemente el cálculo numérico de un máximo con restricciones. Pero, ¿es seguro que el máximo que entrega corresponde a un máximo *global*?

(c) Utilizando el algoritmo precedente, implemente computacionalmente un procedimiento para calcular el número de condición de una matriz dada.

Teorema 3. Sea $A \in M(n \times n, \mathbb{C})$ tal que $\|A\| < 1$ para alguna norma $\|\cdot\|$. Entonces $I - A$ es invertible y $\|(I - A)^{-1}\| = 1/(1 - \|A\|)$.

Demostración. (1) Consideremos el problema:

$$x = A.x + b, \quad (27)$$

donde $b \in \mathbb{C}^n$ es un vector de datos fijo dado. Los problemas del tipo (27) se llaman problemas de punto fijo. La ecuación (27) sugiere la iteración:

$$x^{(k+1)} = A.x^{(k)} + b, \quad k \in \mathbb{N}_0, \tag{28}$$

que por simple resta de dos términos sucesivos conduce a:

$$x^{(k+1)} - x^{(k)} = A.(x^{(k)} - x^{(k-1)}) = A^2.(x^{(k-1)} - x^{(k-2)}) = \dots = A^k.(x^{(1)} - x^{(0)}),$$

de modo que:

$$\|x^{(k+1)} - x^{(k)}\| \leq \|A^k\| \|x^{(1)} - x^{(0)}\|, \quad k \in \mathbb{N}. \tag{29}$$

Luego, en vista de que por hipótesis $\|A\| < 1$, la sucesión $\{x^{(k)}\}_{k \in \mathbb{N}_0}$ es una sucesión de Cauchy en \mathbb{C}^n (el lector interesado no trepidará en examinar detalladamente la demostración de este hecho).

Existe, por lo tanto, un *único* vector $x \in \mathbb{C}^n$ tal que $\lim_{k \rightarrow \infty} x^{(k)} = x$ (aquí juega un papel decisivo la noción de *completitud* de \mathbb{C}^n ; el lector interesado no perderá la ocasión de fundamentar la *unicidad* del vector x). Tal x es, evidentemente, *la* solución de (27), lo que operacionalmente se expresa mediante la ecuación:

$$(I - A).x = b. \tag{30}$$

Dado que no se ha impuesto ninguna condición especial sobre el vector de datos b , la ecuación (30) admite solución *para todo* $b \in \mathbb{C}^n$. Por consiguiente, la matriz $I - A$ es invertible y tiene sentido escribir $(I - A)^{-1}$, de modo que la solución de (30) puede expresarse mediante:

$$x = (I - A)^{-1}.b. \tag{31}$$

(2) Nuestro objetivo ahora es obtener una expresión para la inversa $(I - A)^{-1}$ de la matriz $I - A$. Iterando (28) resulta:

$$\begin{aligned} x^{(1)} &= A.x^{(0)} + b \\ x^{(2)} &= A.x^{(1)} + b = A(A.x^{(0)} + b) + b = A^2.x^{(0)} + A.b + b \\ x^{(3)} &= A.x^{(2)} + b = A(A^2.x^{(0)} + A.b + b) + b = A^3.x^{(0)} + A^2.b + A.b + b \\ &\dots \quad \dots \quad \dots \end{aligned}$$

De este modo se tiene, en general:

$$x^{(k)} = A^k.x^{(0)} + \left(\sum_{j=0}^{k-1} A^j \right).b, \quad k \in \mathbb{N}, \tag{32}$$

de modo que:

$$\left\| x^{(k)} - \left(\sum_{j=0}^{k-1} A^j \right).b \right\| = \|A^k.x^{(0)}\|, \quad k \in \mathbb{N}. \tag{33}$$

En vista de que $\|A\| < 1$, se tiene $A^k.x^{(0)} \rightarrow 0$ cuando $k \rightarrow \infty$ y, por consiguiente:

$$\left\| x - \left(\sum_{j=0}^{\infty} A^j \right).b \right\| = 0,$$

esto es:

$$x = \left(\sum_{j=0}^{\infty} A^j \right).b. \tag{34}$$

Comparando (31) y (34) y notando que ambas ecuaciones son válidas *para todo* $b \in \mathbb{C}$, concluimos que:

$$(I - A)^{-1} = \lim_{k \rightarrow \infty} \sum_{j=0}^{k-1} A^j = \sum_{j=0}^{\infty} A^j. \quad (35)$$

(3) Finalmente, sea $S_k = \sum_{j=0}^k A^j$. Entonces se tiene:

$$\|S_k\| \leq \sum_{j=0}^k \|A^j\| \leq \sum_{j=0}^k \|A\|^j = \frac{1 - \|A\|^{k+1}}{1 - \|A\|} \leq \frac{1}{1 - \|A\|}.$$

Luego, en vista de (35), se tiene:

$$\|(I - A)^{-1}\| = \lim_{k \rightarrow \infty} \|S_k\| \leq \frac{1}{1 - \|A\|}, \quad (36)$$

lo que concluye la demostración. ■

Teorema 4. Sean $A, B \in M(n \times n, \mathbb{C})$ tales que A es no singular y $\|A - B\| < \frac{1}{\|A^{-1}\|}$.

Entonces:

(i) B es no singular.

(ii) $\|B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|A - B\|}.$

(iii) $\|A^{-1} - B^{-1}\| \leq \frac{\|A^{-1}\|^2 \|A - B\|}{1 - \|A^{-1}\| \|A - B\|}.$

Demostración. (1) Primerament observamos que:

$$B = A - (A - B) = A [I - A^{-1}(A - B)]. \quad (37)$$

Por otro lado, por hipótesis se tiene:

$$\|A^{-1}(A - B)\| \leq \|A^{-1}\| \|A - B\| < \|A^{-1}\| \frac{1}{\|A^{-1}\|} < 1.$$

Luego, por el Teorema 3, $I - A^{-1}(A - B)$ es no singular. Por (37), B es no singular pues es un producto de matrices no singulares.

(2) En vista de (37) se tiene $B^{-1} = [I - A^{-1}(A - B)]^{-1} A^{-1}$ y, por consiguiente, por el Teorema 3:

$$\|B^{-1}\| \leq \left\| [I - A^{-1}(A - B)]^{-1} \right\| \|A^{-1}\| \leq \frac{1}{1 - \|A^{-1}(A - B)\|} \|A^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|A - B\|},$$

donde se ha utilizado la desigualdad $\|A^{-1}(A - B)\| \leq \|A^{-1}\| \|A - B\| < 1$ en la última desigualdad.

(3) En vista de la identidad $A^{-1} - B^{-1} = -A^{-1}(A - B)B^{-1}$ se tiene:

$$\|A^{-1} - B^{-1}\| \leq \|A^{-1}\| \|A - B\| \|B^{-1}\| \leq \|A^{-1}\| \|A - B\| \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|A - B\|} = \frac{\|A^{-1}\|^2 \|A - B\|}{1 - \|A^{-1}\| \|A - B\|},$$

lo que concluye la demostración. ■

Teorema 5. Sea $A \in M(n \times n, \mathbb{C})$ no singular y consideremos el sistema de ecuaciones lineales $A.x = b$ con vector de incógnitas $x \in \mathbb{C}^n$ y vector de datos $b \in \mathbb{C}^n$. Sean δA y δb perturbaciones de A y b respectivamente, y supongamos que $\|\delta A\| < \|A^{-1}\|^{-1}$. Entonces $A + \delta A$ es no singular y la perturbación δx resultante en x , i.e., aquella definida implícitamente por $(A + \delta A).(x + \delta x) = b + \delta b$, satisface:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A) \|\delta A\|/\|A\|} \left(\frac{\|\delta A\|}{\|A\|} + \frac{\|\delta b\|}{\|b\|} \right). \quad (38)$$

Demostración. (1) Escribiendo $B := A + \delta A$, notamos por medio del Teorema 4 $A + \delta A$ es no singular.

(2) Análogamente, aplicando el Teorema 4, (ii), se tiene:

$$\|(A + \delta A)^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\delta A\|}. \quad (39)$$

Por otro lado, desarrollando la expresión $(A + \delta A).(x + \delta x) = b + \delta b$, teniendo en cuenta que $A.x = b$, se obtiene:

$$\delta x = (A + \delta A)^{-1} [\delta b - (\delta A).x],$$

lo que conduce a:

$$\begin{aligned} \|\delta x\| &\leq \|(A + \delta A)^{-1}\| (\|\delta b\| + \|\delta A\| \|x\|) \\ &\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \|\delta A\|} (\|\delta b\| + \|\delta A\| \|x\|) \\ &= \frac{\|A^{-1}\| \|A\|}{1 - \|A^{-1}\| \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta b\|}{\|A\|} + \frac{\|\delta A\|}{\|A\|} \|x\| \right) \\ &= \frac{\text{cond}(A)}{1 - \text{cond}(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta b\|}{\|A\|} + \frac{\|\delta A\|}{\|A\|} \|x\| \right). \end{aligned}$$

(3) Dividiendo por $\|x\|$ y teniendo en cuenta que $\|A\| \|x\| \geq \|b\|$, resulta finalmente:

$$\begin{aligned} \frac{\|\delta x\|}{\|x\|} &\leq \frac{\text{cond}(A)}{1 - \text{cond}(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta b\|}{\|A\| \|x\|} + \frac{\|\delta A\|}{\|A\|} \right) \\ &\leq \frac{\text{cond}(A)}{1 - \text{cond}(A) \frac{\|\delta A\|}{\|A\|}} \left(\frac{\|\delta b\|}{\|b\|} + \frac{\|\delta A\|}{\|A\|} \right), \end{aligned}$$

lo que concluye la demostración. ■

Teorema 6. (Cotas “a posteriori”) Sea $A \in M(n \times n, \mathbb{C})$ no singular, sea C la inversa computada (o estimada por algún método de aproximación) de A . Sean:

$$R := I - CA, \quad \delta x = \tilde{x} - x, \quad A.\tilde{x} = \tilde{b}, \quad \delta b = \tilde{b} - b = A.\tilde{x} - b. \quad (40)$$

Entonces, si $\|R\| < 1$, la matriz C es no singular y se tiene:

$$\frac{\|C.\delta b\|}{\|x\| (1 + \|R\|)} \leq \frac{\|\delta x\|}{\|x\|} \leq \frac{\|C.\delta b\|}{\|x\| (1 - \|R\|)} \quad (41)$$

Demostración. (1) En vista de que $\|R\| < 1$, por el Teorema 3 se tiene que $I - R$ es no singular y $\|(I - R)^{-1}\| \leq 1/(1 - \|R\|)$. Puesto que por hipótesis $CA = I - R$, se tiene:

$$\det(C) \det(A) = \det(CA) = \det(I - R) \neq 0.$$

Luego, $\det(C) \neq 0$ y, por lo tanto, C es no singular.

(2) *La segunda desigualdad en (41).* En vista de:

$$\delta b = A.\tilde{x} - b = A.\tilde{x} - A.x = A.(\tilde{x} - x) = A.\delta x = C^{-1}(I - R).\delta x,$$

se tiene

$$\delta x = (I - R)^{-1}C.\delta b,$$

de donde resulta:

$$\|\delta x\| \leq \|(I - R)^{-1}\| \|C.\delta b\|,$$

y finalmente:

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|C.\delta b\|}{(1 - \|R\|)\|x\|},$$

lo que demuestra la segunda desigualdad en (41).

(3) *La primera desigualdad en (41).* ¡TAREA! ■

2. Sea A una matriz de Hilbert de 5×5 . Sea $b = [1, -1/2, 1/3, -1/4, 1/5]^T$.

(a) Calcule los números de condición de A con respecto a las normas $\|\cdot\|_1$, $\|\cdot\|_2$ (norma espectral), $\|\cdot\|_\infty$ y $\|\cdot\|_F$ (norma de Frobenius-Hilbert-Schmidt-Schur).

Sea δA una perturbación de la matriz A tal que $\|\delta A\|_1$ es menor que el 3% de $\|A\|_1$. Sea δb una perturbación del vector b tal que $\|\delta b\|_1$ es menor que el 7% de $\|A\|_1$.

(b) ¿Es $B := A + \delta A$ una matriz no singular?

(c) Si $B := A + \delta A$ es no singular, obtenga estimaciones para $\|B^{-1}\|_1$ y $\|A^{-1} - B^{-1}\|_1$.

(d) Obtenga una estimación (cota superior) para la razón $\|\delta x\|/\|x\|$, donde $x + \delta x$ es la solución del sistema lineal $(A + \delta A).u = b + \delta b$, esto es, $x + \delta x$ satisface $(A + \delta A).(x + \delta x) = b + \delta b$.

(e) Para poder responder las preguntas precedentes, ¿necesita Ud. conocer exactamente los coeficientes de la matriz perturbación δA ?

3. [5, p. 96, exc. 12.1] Suponga que A es una matriz de 202×202 con $\|A\|_2 = 100$ y $\|A\|_F = 101$. Determine la mejor cota inferior para el número de condición $\kappa(A) \equiv \kappa_2(A)$ con respecto a la norma $\|\cdot\|_2$.

4. [5, p. 96, exc. 12.2] Como se sabe, la interpolación polinomial sobre puntos de apoyo equiespaciados es un problema mal condicionado. Para ilustrar este fenómeno, considere un sistema de n , $n \in \mathbb{N}$, puntos de apoyo *equiespaciados* $\{(x_k, y_k)\}_{k=1:n}$ con $-1 \leq x_1 < x_2 < \dots < x_n \leq 1$ e $y_k \in \mathbb{R}$. Considere, además, el sistema de m , $m \in \mathbb{N}$, abscisas equiespaciadas $-1 \leq \xi_1 < \xi_2 < \dots < \xi_m \leq 1$. Sea $p(x)$ el polinomio de grado $n - 1$ que interpola la data equiespaciada $\{(x_k, y_k)\}_{k=1:n}$.

(a) Obtenga una fórmula para la matriz A de $m \times n$ que lleva un n -vector de *data* $[y_1, \dots, y_n]^T$ a un m -vector de valores muestrados $[p(\xi_1), \dots, p(\xi_m)]^T$, donde $p(x)$ es el polinomio interpolador ya mencionado.

Hint. Como se sabe, hay muchas maneras de obtener el polinomio interpolador $p(x) = \sum_{k=0}^{n-1} c_k x^k$. Una de estas vías, bastante obvia pero no necesariamente la más eficiente, consiste

en considerar el sistema de ecuaciones *lineales*:

$$\sum_{k=0}^{n-1} c_k x_i^k = p(x_i) = y_i, \quad i = 1 : n, \quad (42)$$

para las *incógnitas* c_k . Matricialmente (42) se escribe en la forma $V[x_1, \dots, x_k].c = y$, donde:

$$c = \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ \dots \\ c_{n-1} \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix}, \quad \text{y} \quad V[x_1, \dots, x_k] = \begin{bmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{n-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{n-1} \\ 1 & x_3 & x_3^2 & \dots & x_3^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^{n-1} \end{bmatrix}, \quad (43)$$

es la matriz de Vandermonde. Como se sabe, $\det V[x_1, \dots, x_k] \neq 0$ cuando los x_i son todos distintos, como en el caso que nos ocupa (de paso, desafiamos al amable lector a calcular explícitamente el valor de $\det V[x_1, \dots, x_k] \neq 0$). De este modo, el sistema $V[x_1, \dots, x_k].c = y$ tiene una única solución y el polinomio interpolador está bien definido.

Ahora consideramos la transformación:

$$\begin{bmatrix} p(x_1) \\ p(x_2) \\ p(x_3) \\ \dots \\ p(x_n) \end{bmatrix} \equiv \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_n \end{bmatrix} \mapsto \begin{bmatrix} p(\xi_1) \\ p(\xi_2) \\ p(\xi_3) \\ \dots \\ p(\xi_m) \end{bmatrix}. \quad (44)$$

Esta transformación no necesariamente es lineal en términos de las abscisas x_i y ξ_j . No obstante, para abscisas x_i, ξ_j fijas, determinemos una matriz que lleva $[y_1, \dots, y_n]^T$ a $[p(\xi_1), \dots, p(\xi_m)]^T$. Es fácil ver que hay muchas de tales matrices pero hay una que tiene un interés especial para el presente ejercicio. Consideremos la ecuación matricial $[p(\xi_1), p(\xi_2), \dots, p(\xi_m)]^T = A. [p(x_1), p(x_2), \dots, p(x_n)]^T$, i.e:

$$\begin{bmatrix} p(\xi_1) \\ p(\xi_2) \\ \dots \\ p(\xi_m) \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1,n} \\ a_{21} & a_{22} & \dots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m,1} & a_{m,2} & \dots & a_{m,n} \end{bmatrix} \begin{bmatrix} p(x_1) \\ p(x_2) \\ \dots \\ p(x_n) \end{bmatrix}, \quad (45)$$

Queremos determinar una matriz de $A = [a_{ij}]_{m \times n}$, lo más general posible, que satisfaga esta ecuación. Consideremos la i -ésima ecuación del sistema (45). En vista de la forma que tienen los $p(x_j)$, $p(x_j) = \sum_{k=0}^{n-1} c_k x_j^k$, donde los c_k son conocidos, el problema puede interpretarse del siguiente modo: hallar constantes $a_{i,1}, \dots, a_{i,n}$ tales que:

$$\sum_{k=0}^{n-1} c_k \xi_i^k = p(\xi_i) \stackrel{\downarrow}{=} \sum_{j=1}^n a_{ij} p(x_j) = \sum_{j=1}^n a_{ij} \sum_{k=0}^{n-1} c_k x_j^k = \sum_{k=0}^{n-1} c_k \left(\sum_{j=1}^n x_j^k a_{ij} \right), \quad i = 1 : m. \quad (46)$$

Una solución para este problems (ciertamente no la única) consiste en hacer:

$$\sum_{j=1}^n x_j^k a_{ij} = \xi_i^k, \quad k = 0 : n - 1, \quad (47)$$

lo que matricialmente se escribe:

$$\begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_n \\ x_1^2 & x_2^2 & x_3^2 & \dots & x_n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_1^{n-1} & x_2^{n-1} & x_3^{n-1} & \dots & x_n^{n-1} \end{bmatrix} \begin{bmatrix} a_{i,1} \\ a_{i,2} \\ a_{i,3} \\ \vdots \\ a_{i,n} \end{bmatrix} = \begin{bmatrix} 1 \\ \xi_i \\ \xi_i^2 \\ \vdots \\ \xi_i^{n-1} \end{bmatrix}, \tag{48}$$

esto es:

$$V[x_1, \dots, x_n]^T \cdot A_i^T = [1, \xi_i, \xi_i^2, \dots, \xi_i^{n-1}]^T, \tag{49}$$

donde A_i denota la i -ésima fila de la matriz A . En vista de que la matriz de Vandermonde tiene determinante no nulo en nuestro caso, la ecuación (49) es soluble y determina unívocamente los coeficientes $a_{i,j}$, $j = 1 : n$. Repitiendo el proceso para $i = 1 : m$ se obtiene finalmente la matriz A para cada sistema de abscisas $x_1, \dots, x_n, \xi_1, \dots, \xi_m$. Los experimentos computacionales que siguen se refieren a estas matrices $A = A[x_1, \dots, x_n; \xi_1, \dots, \xi_m]$.

(b) Escriba un programa para calcular A y grafique $\|A\|_\infty$ en una escala semi-logarítmica para $n = 1 : 30$, con $m = 2n - 1$. En el límite continuo $m \rightarrow \infty$, los números $\|A\|_\infty$ se conocen como las *constantes de Lebesgue* para interpolación equiespaciada. Estos números son asintóticos a $\frac{2^n}{e(n-1) \log n}$ para $n \rightarrow \infty$.

(c) Para $n = 1 : 30$ y $m = 2n - 1$, ¿cuál es el número de condición κ_∞ , i.e. con respecto a la norma $\|\cdot\|_\infty$, del problema de interpolación de la constante 1?

5. [5, p. 107, excs. 14.1/2] (con leves modificaciones). En el Análisis Asintótico, el Análisis de Errores, la Complejidad de Algoritmos, etc., aparecen frecuentemente las funciones “*O mayúscula*” y “*o minúscula*” introducidas por Paul Bachmann,³ Edmund Landau,⁴ y otros, para describir el comportamiento asintótico de las funciones de interés en un determinado contexto. Se dice que una función $\varphi(x)$ queda dominada asintóticamente por otra función $\psi(x)$, o que $\varphi(x)$ es de tipo *O*-mayúscula con respecto a $\psi(x)$, para la convergencia $x \rightarrow x_0$ en un conjunto numérico apropiado (por ejemplo, $\mathbb{R} \cup \{\pm\infty\}$, $\mathbb{Z} \cup \{\pm\infty\}$ o similares), si y sólo si:

$$\text{existe } C \in \mathbb{R}^+ \text{ tal que } |\varphi(x)| \leq C|\psi(x)| \text{ cuando } x \rightarrow x_0. \tag{50}$$

Cuando (50) se cumple se anota:

$$\varphi(x) = O(\psi(x)) \text{ para } x \rightarrow x_0. \tag{51}$$

Análogamente, se dice que una función $\varphi(x)$ es de tipo *o*-minúscula con respecto a $\psi(x)$ para la convergencia $x \rightarrow x_0$ en un conjunto numérico apropiado, si y sólo si:

$$\frac{|\varphi(x)|}{|\psi(x)|} \rightarrow 0 \text{ cuando } x \rightarrow x_0. \tag{52}$$

Cuando (52) se cumple se anota:

$$\varphi(x) = o(\psi(x)) \text{ para } x \rightarrow x_0. \tag{53}$$

Con estos antecedentes, determine si las siguientes proposiciones son verdaderas o falsas ¡Justifique sus respuestas!

³Paul Bachmann, “*Die analytische Zahlentheorie.*” Teubner, Leipzig, 1894.

⁴Edmund Landau, “*Handbuch der Lehre von der Verteilung der Primzahlen.*” Teubner, Leipzig, 1909. *Breve biografía:* Edmund Landau, 1877-1938. Landau fit ses études à Berlin, sa ville natale, y soutint une thèse, en 1899, et se destina à l’enseignement. A partir de 1909 il fut professeur à Göttingen jusqu’à ce qu’en 1933 le régime national-socialiste le forçât d’abandoner sa chaire. Landau était membre de nombreuses académies.

- (a) $\sin x = O(1)$ cuando $x \rightarrow \infty$.
- (b) $(\sin x)/x = O(1)$ cuando $x \rightarrow 0$.
- (c) $\log x = O(x^{1/100})$ cuando $x \rightarrow \infty$.
- (d) $n! = O((n/e)^n)$ cuando $n \rightarrow \infty$.
- (e) $A = O(V^{2/3})$ cuando $V \rightarrow \infty$, donde A es el área (medida en leguas cuadradas) de una esfera y V es el volumen (medido en micrones cúbicos) encerrado por esa superficie.
- (f) $\text{fl}(\pi) - \pi = O(\varepsilon_{\text{machine}})$. Notar que en este caso (usualmente) no se escribe “cuando $\varepsilon_{\text{machine}} \rightarrow 0$ ” pues ésto se subentiende en el contexto del Análisis de Errores.
- (g) $\text{fl}(n\pi) - n\pi = O(\varepsilon_{\text{machine}})$ uniformemente para todos los enteros n . Aquí $n\pi$ representa la cantidad matemática exacta y no el resultado de una computación de punto flotante.
- (h) $[1 + O(\varepsilon_{\text{machine}})]^s = 1 + O(\varepsilon_{\text{machine}})$ para todo $s \in \mathbb{R}$.
- (i) [Difícil. Consulte un buen texto de Teoría de Números.] $\pi(x)/\log(x) = O(1)$ cuando $x \rightarrow \infty$, donde $\pi(x)$ es el número de números primos menores o iguales que x , $x \in \mathbb{R}$.

6. [4, p. 33, ex. 4] Demuestre (teórica y prácticamente) cómo calcular las siguientes expresiones de manera estable, resp., inversamente estable:

$$\frac{1}{1+2x} - \frac{1-x}{1+x} \quad (|x| \ll 1), \quad \sqrt{x + \frac{1}{x}} - \sqrt{x - \frac{1}{x}} \quad (x \gg 1), \quad \frac{1 - \cos x}{x} \quad (|x| \ll 1).$$

7. [4, p. 33, ex. 6] Para un z dado, el valor de $\text{tg}(z/2)$ se puede calcular mediante la fórmula

$$\text{tg}(z/2) = \pm \sqrt{\frac{1 - \cos x}{1 + \cos x}}.$$

- (a) Demuestre (o refute y entonces corrija) la fórmula dada. ¿Cómo ha de elegirse el signo?
- (b) ¿Es estable (resp., inversamente estable) este método para $z \approx 0$, $z \approx \pi/2$? Si no lo es, proponga un algoritmo alternativo que sí lo sea.

8. [5, p. 112, Exe. 15.1] Cada uno de los siguientes problemas describe un algoritmo implementado en un sistema computacional que satisface los axiomas (13.5) de [5, p. 99]:

$$\forall x \in \mathbb{R} \quad \exists \epsilon \text{ con } |\epsilon| \leq \epsilon_{\text{machine}} \quad \text{tal que} \quad \mathbf{fl}(x) = x(1 + \epsilon), \quad (\text{Axioma (13.5)})$$

y (13.7), también de [5, p. 99],

$$\forall x, y \in \mathbb{R} \quad \exists \epsilon \text{ con } |\epsilon| \leq \epsilon_{\text{machine}} \quad \text{tal que} \quad x \otimes y = (x * y)(1 + \epsilon), \quad (\text{Axioma (13.5)})$$

donde “*” denota cualquiera de las operaciones aritméticas elementales y “⊗” la correspondiente implementación en el sistema computacional disponible.

Determine si el algoritmo correspondiente es *inversamente estable* (“backward stable”), *estable pero no inversamente estable* (“stable but not backward stable”), o *inestable* (“unstable”). En cada caso, dé demostraciones válidas o, al menos, argumentos razonablemente convincentes. Asegúrese de seguir fielmente las definiciones.

- (a) Data: $x \in \mathbb{C}$. Solución: $2x$, computada como $x \oplus x$.
- (b) Data: $x \in \mathbb{C}$. Solución: x^2 , computada como $x \otimes x$.
- (c) Data: $x \in \mathbb{C} \setminus \{0\}$. Solución: 1, computada como $x \oslash x$.
N.B.: (i) “ \oslash ” denota la división aritmética implementada en el sistema computacional.
 (ii) Un sistema computacional que satisfaga el axioma (13.6) “ $x \otimes y = (x * y)(1 + \epsilon)$ ” de [5, p. 99], dará la respuesta exacta (*Why?*), pero las definiciones con base en la condición más débil (13.7) no necesariamente conducen a la respuesta exacta. (*Why?*)
- (d) Data: $x \in \mathbb{C}$. Solución: 0, computada como $x \ominus x$.
N.B.: Nuevamente, un sistema computacional real puede dar mejores resultados que los obtenidos con base en la condición más débil (13.7).

(e) Data: ninguna. Solución: e , computado como

$$\sum_{k=0}^{\infty} \frac{1}{k!} = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots$$

sumado de izquierda a derecha usando “ \otimes ” y “ \oplus ”, para detenerse cuando se llega al primer sumando con magnitud $< \epsilon_{\text{machine}}$.

(f) Data: ninguna. Solución: e , computado igual que en (e), excepto que la serie se suma de derecha a izquierda.

(g) Data: ninguna. Solución: π , computado mediante búsqueda exhaustiva para encontrar el menor número de coma flotante x en el intervalo $[3, 4]$ tal que $s(x) \otimes s(x') \leq 0$, donde $s(x)$ es un algoritmo que calcula $\sin(x)$ en forma estable en el intervalo dado y x' denota el número de coma flotante que sigue inmediatamente a x en el sistema de números de coma flotante.

9. [5, p. 113, Exe. 15.2] Considere un algoritmo para calcular la factorización SVD *completa* de una matriz dada. La *data* de este problema es una matriz A y la *solución* consiste de tres matrices, U (unitaria), Σ (diagonal), y V (unitaria), tales que $A = U\Sigma V^*$. U, Σ, V son matrices explícitas, no representaciones implícitas como productos o reflectores.

(a) Explique qué significaría declarar que este algoritmos es inversamente estable (“backward stable”).

(b) Demuestre que este algoritmo no puede ser inversamente estable (“backward stable”).

(c) Explique qué significaría declarar que este algoritmos es estable (“stable”).

10. [5, p. 119, Exe. 16.1] (a) Sean $Q_1, \dots, Q_k \in M(m \times m, C)$ matrices unitarias fijas y $A \in M(m \times n, C)$ una matriz dada. Considere el problema de calcular $B = Q_k \cdots Q_1 A$. Supóngase que los cálculos se efectúan de derecha a izquierda mediante operaciones de coma flotante directas en un sistema computacional que satisface el Axioma (13.5) y el Axioma (13.5). Demuestre que este algoritmo es inversamente estable. En este ejercicio, la matriz A se considera como una data que puede ser perturbada. Las matrices Q_k , en cambio, están fijas y no pueden ser perturbadas.

(b) Dé un ejemplo que muestre que este resultado no es válido si las matrices unitarias Q_j se reemplazan por matrices arbitrarias $X_j \in M(m \times m, C)$.

11. En el texto [5] hay 3 capítulos dedicados a la discusión de la “accuracy”, la estabilidad y la estabilidad inversa de algunos algoritmos famosos en la Computación Científica. Le invitamos a estudiar esos capítulos a cabalidad y a resolver todos los ejercicios propuestos en el texto en esos capítulos (desde luego, siempre que Ud. pueda conseguir el mentado texto [5]).

12. [3, p. 97, Sect. 4.8.4] SOLUCIÓN ESTABLE DE SISTEMAS LINEALES $A.x = b$ MAL CONDICIONADOS. Para resolver de manera *estable* el sistema de ecuaciones $A.x = b$, cuando la matriz $A \in M(N \times N, \mathbb{R})$ es *regular* (= no singular = invertible), se puede aplicar una factorización de la forma $A = QR$, donde $Q \in M(N \times N, \mathbb{R})$ es ortogonal y $R \in M(N \times N, \mathbb{R})$ es triangular superior. Más precisamente, dado un vector $b \in \mathbb{R}^N$, el sistema lineal de ecuaciones $A.x = b$ es equivalente al sistema *triangular* de ecuaciones $R.x = Q^T.b$, donde el *número de condición* de la matriz R no es mayor que el *número de condición* de la matriz A con respecto a la norma $\|\cdot\|_2$ y, además, la norma $\|\cdot\|_2$ del vector $Q^T.b$ no es mayor que la norma $\|\cdot\|_2$ del vector b , i.e.:

$$\kappa_2(R) = \kappa_2(Q^T A) = \kappa_2(A), \quad \|Q^T.b\|_2 = \|b\|_2.$$

Tarea. Demuestre las declaraciones precedentes. Estudie el caso $A \in M(N \times N, \mathbb{C})$, $b \in \mathbb{C}^N$. Construya un ejemplo computacional no trivial que ilustre la situación descrita.

13. [3, p. 101, Theor. 4.65] Sea $A \in M(m \times m, \mathbb{R})$ una matriz no singular. Considere la *SVD* de la matriz A : $A = U\Sigma V^T$, con $U, V \in M(m \times m, \mathbb{R})$ ortonormales y $\Sigma = \text{diag}\{\sigma_1, \dots, \sigma_m\} \in M(m \times m, \mathbb{R})$, $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_m$.

(a) Encuentre una expresión para $\kappa_2(A)$ en términos de los valores singulares σ_k de A .

(b) Considere la ecuación $A.(x + \Delta x) = b + \Delta b$. Demuestre las estimaciones:

$$\|b\|_2 \leq \|A\|_2 \|x\|_2, \quad \|\Delta x\|_2 \leq \|A^{-1}\|_2 \|\Delta b\|_2, \quad \frac{\|\Delta x\|_2}{\|x\|_2} \leq \kappa_2(A) \frac{\|\Delta b\|_2}{\|b\|_2}. \quad (54)$$

donde $\kappa_2(A)$ es el número de condición de la matriz A con respecto a la norma $\|\cdot\|_2$.

(c) Determine los vectores $b \in \mathbb{R}^m$, respectivamente $\Delta b \in \mathbb{R}^m$, en función de la matriz U , para los cuales se cumple la igualdad en cada una de las desigualdades (54) consideradas separadamente.

(d) Encuentre condiciones para $b \in \mathbb{R}^m$ de modo que se cumpla:

$$\frac{\|\Delta x\|_2}{\|x\|_2} \leq \frac{\|\Delta b\|_2}{\|b\|_2} \quad \forall \Delta b \in \mathbb{R}^m.$$

REFERENCIAS

- [1] M. Abramowitz and I.A. Stegun (eds.). *Handbook of Mathematical Functions*. Dover, New York, 1965.
- [2] E. Isaacson and H.B. Keller. *Analysis of Numerical Methods*. John Wiley and Sons, New York, 1966. There is a more recent edition by Dover, New York, 1994.
- [3] R. Plato. *Concise Numerical Mathematics*. American Mathematical Society, Providence, Rhode Island 02904-2294, 2003. Originally published in German language by Friedr. Vieweg & Sohn Verlagsgesellschaft mbH, D-65189 Wiesbaden, under the title Robert Plato: Numerische Mathematik kompakt. 1. Auflage". Translated from the original German edition by Richard Le Borne and Sabine Le Borne.
- [4] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Springer Verlag, Berlin-Heidelberg, 1993. A rather advanced but very good book indeed!
- [5] L.N. Trefethen and D. Bau III. *Numerical Linear Algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1997. A very good book indeed! More advanced than Strang's book.

LSC/lsc, Valparaíso, 18 de junio de 2007