Ciencia en Desarrollo, Vol. 10 No. 2
ISSN 0121 - 7488 - Julio a Diciembre de 2019

Refinamiento iterativo del método de Gauss-Jordan, en sistemas mal condicionados.

Iterative refinement of the Gauss-Jordan method, in ill conditioned systems.

A. Y. Mesa Juya [a*]
A. Calvache Archila [b]

Fecha de Recepción: 03.02.2018
Fecha de Aceptación: 23.05.2019
DOI: https://doi.org/10.19053/01217488.v10.n2.2019.8761

Resumen

En este Artículo, se construye un algoritmo iterativo para mejorar la solución de un sistema de ecuaciones lineales, de la forma $Ax = b$, cuando se resuelve utilizando el método de Gauss-Jordan y utilizando aritmética finita. Comprender el funcionamiento del algoritmo, mostrar su alcance y analizar cómo se dedujo, se logra a través del concepto de norma matricial, junto con algunas de sus propiedades. Se introduce el concepto del número de condición de una matriz, y se le encuentran cotas mediante el uso de las normas matriciales. Finalmente, se expone el algoritmo iterativo del Refinamiento, que muestra el poder de éste, al resolver un sistema de Ecuaciones lineales mal condicionadas.

Palabras clave*:* Norma matricial, números de condición, método de Refinamiento.

Abstract

In this paper, an iterative algorithm is constructed to improve the solution of a system of linear equations, of the form $Ax = b$, when it is solved using the Gauss-Jordan method and by using finite arithmetic. Understanding the functioning of the algorithm, showing its scope and analyzing how it is deduced, is achieved through the concept of matrix norm, together with some of its properties. The concept of the condition number of a matrix is introduced, and are found bounds for it by using the matrix norms. Finally, it is explained the iterative algorithm of the Refinement, showing the power of this one, when it is solved a system of linear equations ill conditioned.

Key words*:* Matrix norm, condition numbers, Refinement Method.

a INEM Carlos Arturo Torres
* Correo electrónico: yetsen044@gmail.com
b Universidad Pedagógica y Tecnológica de Colombia. Escuela de Matemáticas y Estadística
* Correo electrónico: alvaro.calvache@uptc.edu.co

1. INTRODUCCIÓN

Un resultado muy conocido del álgebra lineal indica que si A es una matriz cuadrada invertible de orden n con entradas complejas, entonces, el sistema de ecuaciones $Ax = b$ tiene solución única, sin embargo, encontrar dicha solución no siempre es una tarea fácil. La dificultad en el cálculo de una solución para un problema de este tipo se da, o porque la "magnitud" de la matriz o de su inversa podría ser muy grande o porque las entradas de la matriz podrían ser cantidades extremadamente grandes o extremadamente pequeñas, lo que causa que hallar la solución sea una tarea larga, tediosa y poco precisa cuando se trabaja con una aritmética finita, y en algunos casos casi imposible de realizar. Cuando se menciona la "magnitud" de una matriz, se hace referencia al tamaño dado por alguna de las normas matriciales.

Como una posible solución a este inconveniente, se empezó a pensar en el desarrollo de una máquina que hiciera cálculos matemáticos de forma rápida y precisa, uno de los primeros en iniciar la construcción de dicha máquina fue Charles Babbage, (1791-1891) quien es conocido como el "padre de la computación", por el diseño, (no construcción), de lo que podría ser la primera máquina analítica que ejecutó programas de tabulación o computación. Posteriormente, en el marco de la segunda guerra mundial, Alan Turing (1912 - 1954) y John von Neumann (1903 - 1957) contribuyeron en el desarrollo y construcción de las primeras computadoras digitales, como herramientas que facilitaron y agilizaron los cálculos numéricos, como se establece en Turing [10] y en Díaz [5].

Turing es conocido por ser uno de los pioneros de la computación moderna, sus trabajos permitieron la creación de las primeras calculadoras digitales y sus aportes en el proceso para la creación de máquinas que facilitaran los cálculos numéricos fueron significativos. Fue además, uno de los primeros en hablar acerca de lo que ahora se conoce como normas matriciales, sin embargo, él las definió como la "medida de la magnitud de una matriz". Turing en [10] menciona que hay diferentes formas de medir la magnitud de una matriz por medio de un número real, y estas incluyen :

(a) La norma de una matriz A, denotada por $N(A)$ y calculada por:

$$N(A) = (Traza(A^*A))^{1/2} = \left(\sum_{i,j} |a_{ij}|^2\right)^{1/2}$$

(b) La expansión máxima, denotada por $B(A)$ y calculada por:

$$B(A) = \max_{x\neq 0} \frac{|Ax|}{|x|} = \max_{x\neq 0} \frac{(Ax,Ax)^{1/2}}{(x,x)^{1/2}}$$

(c) El coeficiente máximo, denotado por $M(A)$ y calculado por:

$$M(A) = \max_{ij} |a_{ij}| .$$

Turing además estableció las siguientes desigualdades:

$$\begin{aligned}
M(X+Y) &\leq M(X)+M(Y),\\
M(XY) &\leq n\, M(X)M(Y),\\
B(X+Y) &\leq B(X)+B(Y),\\
B(XY) &\leq B(X)B(Y),\\
N(X+Y) &\leq N(X)+N(Y),\\
N(XY) &\leq N(X)N(Y),\\
N(X) &\leq n\, M(X),\\
M(X) &\leq N(X),\\
M(X) &\leq B(X),\\
B(X) &\leq \sqrt{n}\, M(X),\\
B(X) &\leq N(X),\\
N(X) &\leq \sqrt{n}\, B(X).
\end{aligned}$$

Von Neumann y Goldstine, por su parte, presentan en [8] un estudio acerca de la precisión y estabilidad del método de eliminación, como herramienta para hallar la inversa de una matriz de orden n, cuando n es grande. En éste, se presenta el siguiente listado con las posibles fuentes de error:

(a) La formulación matemática que se elige para representar el problema subyacente, ya que ésta puede representarlo sólo con ciertas idealizaciones, simplificaciones y negaciones.

(b) Asumiendo que la formulación matemática no presente errores, la descripción de acuerdo a dicho modelo puede implicar parámetros cuyos valores se deriven directa o indirectamente (es decir, a través de otras teorías o cálculos) de las observaciones. Estos parámetros se verán afectados por errores y estos errores subyacentes causarán errores en el resultado final.

(c) La formulación matemática del modelo puede implicar el uso de funciones trascendentes y operaciones como diferenciación o integración, entre otras, que para ser abordados mediante cálculos numéricos deben reemplazarse por procesos elementales y definiciones explícitas que corresponden a un procedimiento finito y constructivo que se resuelve en una secuencia lineal de pasos. Todos estos reemplazos son aproximados y constituyen por tanto, una ter- cera fuente de error.

(d) Asumiendo que se puedan superar los tres posibles errores mencionados antes, aún queda una limitación por superar y es el hecho de que ningún dispositivo de cálculo puede realizar todas las operaciones elementales de manera rigurosa y precisa.

Por todo lo antes mencionado, resulta de gran interés estudiar la sensibilidad en la solución de un sistema de ecuaciones lineales cuando se hacen pequeñas perturbaciones en las entradas de los datos. Una herramienta poderosa para controlar dicha sensibilidad es el número de condición de la matriz de coeficientes, éste se puede acotar usando normas matriciales y ha sido objeto de estudio por muchos años. Von Newmann y Turing fueron algunos de los primeros en estudiar los números de condición, ellos centraron sus esfuerzos en buscar cotas superiores para los números de condición. En trabajos más recientes como el presentado por Pyzara et al. en [9], se muestra la relación que tienen los números de condición con la convergencia de algunos métodos iterativos, usados para resolver sistemas de ecuaciones lineales.

Este Artículo está organizado de la siguiente forma: en la Sección 2 se presenta la notación que se utilizará y algunas generalidades concernientes a las normas matriciales. Las propiedades y ejemplos, se enuncian sin sus demostraciones, ya que son resulta- dos ampliamente conocidos en el área del Análisis Matricial, y pueden ser consultados en textos como el de Hörner [6]. La Sección 3, hace referencia al concepto de condicionamiento de una matriz y se deducen algunas de las cotas que se utilizarán cuando se aborde el problema del refinamiento de soluciones de sistemas lineales. En la sección 4, se presenta la base teórica necesaria para mostrar y desarrollar un método que permite perfeccionar de manera iterativa, los resultados que se obtienen mediante el método de Gauss-Jordan, cuando se trabaja con una aritmética fija con un número prederminado de dígitos. Este tipo de métodos han sido usado recientemente en la solución de sistemas de ecuaciones, tal como se puede observar en los artículos de Arunachalam y Dharmaraja [2], de Calvache et al. [4] y en el de Arun et al. [1], en donde sistemas de ecuaciones integrales, por medio de transformadas de Laplace, se modifican a sistemas lineales de ecuaciones, cuyas primeras soluciones no son tan precisas, pero trabajándolo reiterativamente se encuentran soluciones más precisas, las cuales mediante transformadas inversas de Laplace, conllevan a soluciones aproximadas de los sistemas de ecuaciones integrales. Al algoritmo deducido se le denominará método de Refinamiento. En la Sección 5 se ilustra el método mediante un ejemplo numérico. Por último se enuncian unas conclusiones.

2. NORMAS MATRICIALES

Para empezar se establecen las siguientes notaciones: $\mathcal{M}m,n$ y $\mathcal{M}n$ se usan para denotar respectivamen- te, al conjunto de matrices rectangulares de m filas y n columnas con entradas en $\mathbb{C}$ y al conjunto de matrices cuadradas con n filas y n columnas, con entradas en $\mathbb{C}$. Se representa por $\mathbb{C}^n$ al conjunto de vectores columna, de tamaño n, cuyas entradas son números complejos, es decir $\mathbb{C}^n = \mathcal{M}_{n,1}$. Dada $A \in \mathcal{M}_{m,n}$, A^t es la matriz transpuesta de A. La matriz nula se representa por $\mathbf{O}$ y en cada caso se asume que tiene el tamaño indicado para que las operaciones a realizar estén bien definidas. La matriz identidad de orden n, se nota por I_n, y cuando no haya lugar a confusión se representa simplemente por I. $\sigma(A)$ simboliza el conjunto de todos los valores propios de la matriz A y es llamado el espectro de A, además $\rho(A)$

representa el radio espectral de A, que es el mayor valor absoluto de los elementos de $\sigma(A)$.

Para el posterior desarrollo del método de Refinamiento, se hace necesario controlarlo de acuerdo con el número de condición de una matriz, y por esta razón se abordará primero el concepto de norma matricial. A continuación se presenta la definición de norma matricial, seguida de algunos ejemplos y propiedades.

Definición 1. *Considerando* $(\mathscr{M}_n,+,\cdot)$ *como espacio vectorial sobre* $\mathbb{C}$*, una función,*

$$|||\cdot||| : \mathscr{M}_n \to \mathbb{R},$$

es una norma matricial sobre $\mathscr{M}_n$*, si para cada par de matrices* $A, B \in \mathscr{M}_n$ *y cada* $\lambda \in \mathbb{C}$ *se satisfacen las siguientes propiedades:*

1. $|||A||| \geq 0$ *y* $|||A||| = 0$*, si y sólo si,* $A = \mathbf{O}$.
2. $|||\lambda A||| = |\lambda| \ |||A|||$.
3. $|||A + B||| \leq |||A||| + |||B|||$.
4. $|||AB||| \leq |||A||| \ |||B|||$.

Las funciones que satisfacen únicamente las propiedades 1 a 3, son llamadas normas vectoriales y debido a ésto, es válido afirmar que toda norma matricial es una norma vectorial, sin embargo, el recíproco no es cierto.

Ejemplo 2. Dada $A \in \mathscr{M}_n$, las funciones definidas como siguen, son normas matriciales sobre $\mathbb{C}$.

(i) $\|A\|_1 = \sum_{i,j=1}^{n} |a_{ij}|$.

(ii) $\|A\|_2 = \left(\sum_{i,j=1}^{n} |a_{ij}|^2 \right)^{1/2}$.

(iii) $\|A\|_\infty = n \|A\|_\infty$, donde

$$\|A\|_\infty = \max_{i,j=1,\ldots,n} |a_{ij}|.$$

Nótese que $|||\cdot|||_\infty$, es una norma vectorial, pero no es norma matricial.

(iv) $|||A|||^{(1)} = \max_{j=1,\cdots,n} \sum_{i=1}^{n} |a_{ij}|$.

(v) $|||A|||^{\infty} = \max_{i=1,\cdots,n} \sum_{j=1}^{n} |a_{ij}|$.

(vi) $|||A|||^{(2)} = \max_{\|x\|_2=1} \|Ax\|_2$,

donde, $x = (x_1, x_2, \ldots, x_n)^t \in \mathbb{C}^n$ y

$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2}$ es la norma usual en $\mathbb{C}^n$.

Propiedad 3. *Sean* $A \in \mathscr{M}_n$ *y* $|||\cdot|||$ *una norma matricial definida sobre* $\mathscr{M}_n$*, entonces,*

(i) $|||A^2||| \leq |||A|||^2$.

(ii) *Para cada* $k \in \mathbb{Z}^+$, $|||A^k||| \leq |||A|||^k$.

(iii) *Si* A *es idempotente y no nula, entonces* $|||A||| \geq 1$.

(iv) $|||I_n||| \geq 1$.

(v) *Si* A *es invertible,* $|||A^{-1}||| \geq \dfrac{1}{|||A|||}$.

A partir de las normas vectoriales de $\mathbb{C}^n$ es posible construir normas sobre $\mathscr{M}_n$.

Teorema 4. *Sea* $\|\cdot\|$ *una norma vectorial en* $\mathbb{C}^n$*. Se define la función* $|||\cdot|||$ *de* $\mathscr{M}_n$ *en* $\mathbb{R}$*, como sigue:*

$$|||A||| = \max_{\|x\|=1} \|Ax\|.$$

Entonces $|||\cdot|||$ *es una norma matricial, llamada norma matricial inducida por la norma vectorial* $\|\cdot\|$.

Teorema 5. *Sean* $A \in \mathscr{M}_n$*,* $\|\cdot\|$ *una norma vectorial de* $\mathbb{C}^n$ *y* $|||\cdot|||$ *su norma matricial inducida, entonces,*

(i) $|||I_n||| = 1$.

(ii) *Para cada* $x \in \mathbb{C}^n$, $\|Ax\| \leq |||A||| \ \|x\|$.

(iii) $|||A||| = \max_{\|x\|\leq 1} \|Ax\| = \max_{x\neq 0} \dfrac{\|Ax\|}{\|x\|} = \max_{\|x\|_\alpha=1} \dfrac{\|Ax\|}{\|x\|}$,

en donde, $\|\cdot\|_\alpha$ *es una norma vectorial de* $\mathbb{C}^n$.

En el siguiente ejemplo se presenta una interpretación gráfica de la norma matricial inducida por $\|\cdot\|_\infty$, cuando ésta se trabaja sobre el conjunto de las matrices de orden dos, con entradas reales.

Ejemplo 6. Si en $\mathbb{R}^2$ se toma la norma vectorial $\|\cdot\|_\infty$, definida por

$$\|(x, y)^t\|_\infty = \max\{|x|, |y|\},$$

y se considera la matriz,

$$M = \begin{bmatrix} 1 & 2 \\ 1 & 4 \end{bmatrix},$$

a ésta le corresponde la transformación lineal,

$$T_M : \begin{array}{ccc} \mathbb{R}^2 & \longrightarrow & \mathbb{R}^2 \\ \begin{bmatrix} x \\ y \end{bmatrix} & \longmapsto & M\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} x+2y \\ x+4y \end{bmatrix} \end{array}.$$

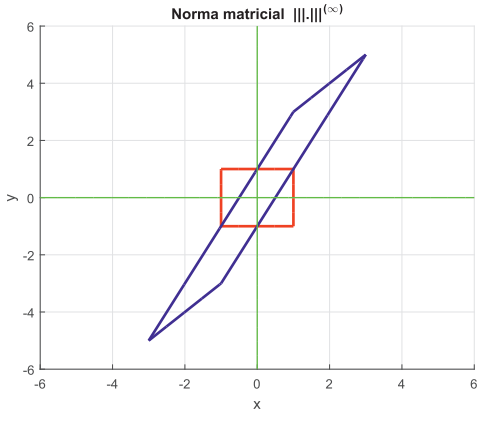

Figura 1. Gráfica de $L = \{x \in \mathbb{R}^2 \mid \|x\|_\infty = 1\}$ y de su imagen bajo M.

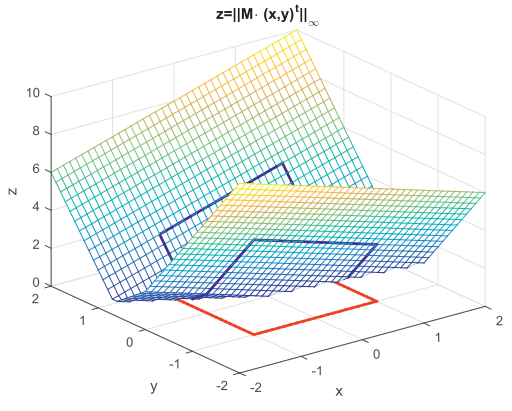

Figura 2. Gráfica de $z = |Mx|\infty$ y de la curva $\alpha(t)$.

Ahora bien, en la Figura 1, se observan dos gráficas. En color rojo se representan los vectores unitarios con la norma $\|\cdot\|_\infty$, y en color azul se representan los vectores $Z = MX$, en donde X es unitario; es decir, la gráfica en color azul representa la imagen bajo M de los vectores de la gráfica en color rojo. Ahora, si para cada vector Z del gráfico en azul, se calcula $\|Z\|_\infty$ y se toma la máxima de estas normas, es posible ver, que para el ejemplo se tiene que la norma máxima es igual a 6, y se concluye que $|||M||| = 6$, donde $|||\cdot|||$ es la norma inducida por $\|\cdot\|_\infty$.

Por otra parte, en la Figura 2 se aprecian tres elementos. El primero está en el plano xy, con color rojo y representa los vectores U de $\mathbb{R}^2$, tales que, $\|U\|_\infty = 1$; el segundo elemento es la superficie,

$$Z(X) = \|MX\|_\infty,$$

siendo X un vector de $\mathbb{R}^2$; finalmente con color azul y sobre la superficie se muestra la curva paramétrica

$$\alpha(t) = \begin{bmatrix} MV(t) \\ Z(V(t)) \end{bmatrix},$$

donde $V(t)$ es una parametrización de la curva representada en color rojo. También se aprecia que la mayor altura de la superficie $z = Z(X)$, condicionada a $\|X\|_\infty = 1$, es igual a 6; es decir, también se concluye que la norma inducida por $\|\cdot\|_\infty$, evaluada en M es 6.

A partir del Ejemplo 6 es posible conjeturar la existencia de una relación entre $|||\cdot|||^\infty$ y la norma matricial inducida por la norma vectorial $\|\cdot\|_\infty$. Esto es en efecto cierto y se enuncia formalmente en la siguiente proposición.

Proposición 7. *Dada* $A \in \mathcal{M}_n$,

(i) $|||A|||^{(1)} = \max_{\|x\|_1=1} \|Ax\|_1$.

(ii) $|||A|||^\infty = \max_{\|x\|_\infty=1} \|Ax\|_\infty$.

Adicionalmente, se usará el siguiente resultado, el cual es ampliamente conocido:

Proposición 8. *Dada* $A \in \mathcal{M}_n$,

$\rho(A) = \inf \{|||A||| \mid |||\cdot||| \text{ es norma matricial en } \mathcal{M}_n\}$.

3. CONDICIONAMIENTO DE MATRICES

En matemáticas se acepta la existencia de números con un número infinito de dígitos, sin embargo, en las computadoras ésto no es posible y por tanto este tipo de números es representado con una aproximación que sólo tiene un número finito de dígitos, es decir, las calculadoras y computadoras usan un subconjunto relativamente pequeño de números racionales para representar

a todos los números reales, por esta razón, los números irracionales e infinitos racionales son representados por una aproximación que está lo suficientemente cercana al valor exacto como para que los resultados de muchas operaciones sean aceptables, pero en algunos casos, más comunes de lo que se cree, esto causa diferencias considerablemente grandes entre el resultado obtenido por la máquina y el resultado preciso. El error causado al usar estas aproximaciones para resolver problemas numéricos, fue considerado por Neumann y Goldstine en [8] y ahora es llamado error de redondeo.

Por ejemplo, al solucionar sistemas de ecuaciones lineales, aun cuando éstos tengan pocas variables y pocas ecuaciones, se pueden obtener errores de redondeo, tal como se puede apreciar en el siguiente ejemplo, en el que se presenta un sistema de ecuaciones lineales con dos incógnitas x e y:

Ejemplo 9.

$$\begin{aligned} 73184x + 29515y &= 41599573, \\ 16189x + 6529y &= 9202223. \end{aligned} \qquad (1)$$

Las soluciones exactas de este sistema son $x_v = 272$ e $y_v = 735$, como puede comprobarse fácilmente al reemplazar estos valores por x e y en (1).

Ahora se resolverá este sistema, usando aritmética de diez dígitos significativos (esta es la aritmética que se utiliza en una gran variedad de calculadoras científicas), trabajando con el método de Cramer.

El sistema (1) se puede escribir matricialmente, como:

$$Aw = b, \qquad (2)$$

donde.

$$A = \begin{bmatrix} 73184 & 29515 \\ 16189 & 6529 \end{bmatrix}, \qquad w = \begin{bmatrix} x \\ y \end{bmatrix}$$

y

$$b = \begin{bmatrix} 41599573 \\ 9202223 \end{bmatrix}.$$

Ahora, se calcula:

$$\Delta = \det(A) = 1.$$

$$\begin{aligned} \Delta_x &= \begin{vmatrix} 41599573 & 29515 \\ 9202223 & 6529 \end{vmatrix} \\ &\approxeq 0.000000003 \times 10^{11} \\ &= 300. \end{aligned}$$

$$\begin{aligned} \Delta_y &= \begin{vmatrix} 73184 & 41599573 \\ 16189 & 9202223 \end{vmatrix} \\ &\approxeq 0.000000007 \times 10^{11} \\ &= 700. \end{aligned}$$

Por tanto los valores aproximados de x e y, obtenidos por este método son:

$$x_a = \frac{\Delta_x}{\Delta} = \frac{300}{1} = 300,$$

$$y_a = \frac{\Delta_y}{\Delta} = \frac{700}{1} = 700.$$

Comparando con las respuestas verdaderas, se obtienen los siguientes errores relativos:

$$\begin{aligned} \text{Error relativo en } x &= \left| \frac{x_a - x_v}{x_v} \right| \\ &= \left| \frac{300 - 272}{272} \right| \approxeq 10.29\,\%, \end{aligned}$$

$$\begin{aligned} \text{Error relativo en } y &= \left| \frac{y_a - y_v}{y_v} \right| \\ &= \left| \frac{700 - 735}{735} \right| \approxeq 4.76\,\%. \end{aligned}$$

Como puede observarse son errores relativos muy grandes.

Ahora se considera el sistema:

$$\begin{aligned} 73184\,x + 29515\,y &= 41599573, \\ 3679\,x - 81235\,y &= -58707037. \end{aligned} \qquad (3)$$

Puede comprobarse que las verdaderas respuestas de este sistema son $x_v = 272$ e $y_v = 735$.

Procediendo en forma similar a como se resolvió el sistema (1), se obtiene la solución aproximada:

$$x_a = \frac{\Delta_x}{\Delta} = \frac{-1.646\,603\,116 \times 10^{12}}{-6\,053\,687\,925} = 272.000\,000\,1$$

$$y_a = \frac{\Delta_y}{\Delta} = \frac{-4.449\,460\,625 \times 10^{12}}{-6\,053\,687\,925} = 735.$$

Comparando con las respuestas verdaderas, se obtienen los siguientes errores relativos:

$$\text{Error relativo en } x = \left|\frac{x_a - x_v}{x_v}\right| = \left|\frac{272.000\,000\,1 - 272}{272}\right| \approxeq 3.68 \times 10^{-11}.$$

$$\text{Error relativo en } y = \left|\frac{y_a - y_v}{y_v}\right| = \left|\frac{735 - 735}{735}\right| = 0.$$

Estos ejemplos muestran que si $\widehat{x}$ es una solución aproximada del sistema $Ax=b$, $r=b-A\widehat{x}$ es el vector residual y $\|r\|$ es pequeña, no es necesariamente cierto que el error absoluto, $\|x-\widehat{x}\|$ sea también pequeño. Adicionalmente, es posible asumir que la matriz de coeficientes B, del sistema (3), debe cumplir alguna condición que no cumple la matriz A del sistema (1) y que por esa razón el error de redondeo en las solución del sistema (1) es grande, mientras que el error de redondeo en la solución de (3) es muy pequeño.

Sean $\|\cdot\|$ una norma matricial dada y $A \in \mathcal{M}_n$ una matriz invertible, si se quiere calcular la matriz inversa de A, se trabajará con la matriz $B = A + \Delta A$, en donde, ΔA es una perturbación de la matriz A.

$$|||A^{-1}\Delta A||| < 1, \qquad (4)$$

y ya que $B = A(I + A^{-1}\Delta A)$ y

$$\rho(A^{-1}\Delta A) \leq |||A^{-1}\Delta A|||, \qquad (5)$$

entonces, (4) y (5) implican que,

$$-1 \notin / \sigma\,(A^{-1}\Delta A),$$

y por tanto $0 \notin / \sigma\,(B)$, es decir, B es invertible.

También se tiene que,

$$\begin{aligned} A^{-1}(\Delta A)B^{-1} &= A^{-1}(B-A)B^{-1} \\ &= A^{-1}BB^{-1} - A^{-1}AB^{-1} \\ &= A^{-1} - B^{-1} \end{aligned}$$

y así,

$$\begin{aligned} |||A^{-1} - B^{-1}||| &= |||A^{-1}(\Delta A)B^{-1}||| \\ &\leq |||A^{-1}\Delta A|||\ \ |||B^{-1}||| \end{aligned} \qquad (6)$$

Además, $B^{-1} = A^{-1} - A^{-1}(\Delta A)B^{-1}$, de donde,

$$\begin{aligned} |||B^{-1}||| &\leq |||A^{-1}||| + |||A^{-1}(\Delta A)B^{-1}||| \\ &\leq |||A^{-1}||| + |||A^{-1}\Delta A|||\ |||B^{-1}|||, \end{aligned}$$

lo que es equivalente a tener,

$$|||B^{-1}||| = |||(A+\Delta A)^{-1}||| \leq \frac{|||A^{-1}|||}{1 - |||A^{-1}\Delta A|||}. \qquad (7)$$

Combinando (6) y (7) se obtiene,

$$\begin{aligned} |||A^{-1} - B^{-1}||| &\leq |||A^{-1}\Delta A|||\ \ |||B^{-1}||| \\ &\leq \frac{|||A^{-1}|||\ |||A^{-1}\Delta A|||}{1 - |||A^{-1}\Delta A|||} \\ &\leq \frac{|||A^{-1}|||\ \ |||A^{-1}|||\ \ |||\Delta A|||}{1 - |||A^{-1}\Delta A|||}, \end{aligned}$$

y entonces,

$$\begin{aligned} \frac{|||A^{-1} - (A+\Delta A)^{-1}|||}{|||A^{-1}|||} &= \frac{|||A^{-1} - B^{-1}|||}{|||A^{-1}|||} \\ &\leq \frac{|||A^{-1}|||\ \ |||A^{-1}|||\ \ |||\Delta A|||}{(1 - |||A^{-1}\Delta A|||)\,|||A^{-1}|||} \\ &= \frac{|||A^{-1}|||\ \ |||\Delta A|||}{1 - |||A^{-1}\Delta A|||} \\ &= \frac{|||A^{-1}|||\ \ |||A|||\ \ |||\Delta A|||}{(1 - |||A^{-1}\Delta A|||)\,|||A|||} \\ &= \frac{|||A^{-1}|||\ \ |||A|||}{1 - |||A^{-1}\Delta A|||}\ \frac{|||\Delta A|||}{|||A|||}. \end{aligned}$$

Así, una cota superior para el error relativo en el cálculo de la inversa es:

$$\frac{|||A^{-1} - (A+\Delta A)^{-1}|||}{|||A^{-1}|||} \leq \frac{|||A^{-1}|||\ \ |||A|||}{1 - |||A^{-1}\Delta A|||}\ \frac{|||\Delta A|||}{|||A|||},$$

Con base en esta argumentación, se presenta la siguiente definición.

Definición 10. *Sean A una matriz en $\mathscr{M}_n$ y $|||\cdot|||$ una norma matricial dada, el número de condición de la matriz A con respecto a la norma matricial $|||\cdot|||$, se define y denota por:*

$$k(A) = \begin{cases} |||A^{-1}|||\ |||A|||, & \text{si } A \text{ es invertible}, \\ \infty, & \text{si } A \text{ no es invertible}. \end{cases}$$

Hasta el momento se ha encontrado la cota,

$$\frac{|||A^{-1} - (A+\Delta A)^{-1}|||}{|||A^{-1}|||} \leq \frac{k(A)}{1 - |||A^{-1}\Delta A|||} \frac{|||\Delta A|||}{|||A|||}. \quad (8)$$

Obsérvese que,

$$|||A^{-1}|||\ |||\Delta A||| = |||A^{-1}|||\ |||A|||\ \frac{|||\Delta A|||}{|||A|||} = k(A)\frac{|||\Delta A|||}{|||A|||}.$$

Si se fortalece la suposición hecha en (4) y se asume que,

$$|||A^{-1}|||\ \ |||\Delta A||| < 1, \quad (9)$$

entonces,

$$\frac{|||A^{-1} - (A+\Delta A)^{-1}|||}{|||A^{-1}|||} \leq \frac{k(A)}{1 - k(A)\dfrac{|||\Delta A|||}{|||A|||}} \frac{|||\Delta A|||}{|||A|||}, \quad (10)$$

es decir, se tiene una cota superior para el error relativo en el cálculo de la inversa de una matriz, en términos del error relativo de los datos y del número de condición, tal cota es llamada, cota a priori.

Si $|||A^{-1}|||\ \ |||\Delta A|||$ no sólo es menor que 1, sino que es mucho más pequeña que 1, entonces, la expresión del lado derecho en (10), sería muy cercana al valor,

$$k(A)\frac{|||\Delta(A)|||}{|||A|||}$$

y por tanto es válido pensar que el error relativo en la inversa, tiene el mismo orden que el error relativo en los datos, concluyéndose así que $k(A)$ no es muy grande. Por esta razón, se establece que,

(i) Si $k(A)$ es grande, se dirá que A está mal condicionada.

(ii) Si $k(A)$ es pequeño y cercano a 1, se dirá que A está bien condicionada.

(iii) Si $k(A) = 1$, se dirá que A está perfectamente bien condicionada.

Todo esto con respecto a una norma matricial específica.

Ejemplo 11. Si nuevamente se toman las matrices A de (1) y B de (3) y la norma matricial $|||\cdot|||^{\infty}$, se tiene que,

$$k(A) = 9178517{,}727 \qquad \text{y} \qquad k(B) \cong 1.88.$$

Es decir, A está mal condicionada, mientras que B está bien condicionada.

4. MÉTODO DE REFINAMIENTO

En este momento se comenzarán a deducir diferentes cotas de error, que permitirán establecer un método iterativo, en el cual en cada repetición se perfecciona cada vez más, la solución inicial dada por el método de Gauss-Jordan. Por supuesto, que este algoritmo que se va a derivar, será útil en caso que la matriz del sistema esté mal condicionada, pues de lo contrario, el método de Gauss-Jordan, dará como solución una muy buena aproximación a la solución exacta del sistema.

Supóngase que se quiere resolver el sistema lineal $Ax = b$, con A una matriz invertible de $\mathscr{M}_n$ y b un vector no nulo de $\mathbb{C}^n$. Sin embargo por los errores computacionales, en realidad se va a resolver el sistema perturbado

$$(A + \Delta A)\widehat{x} = b + \Delta b, \quad (11)$$

donde $A,\ \Delta A \in \mathscr{M}_n$; $b,\ \Delta b \in \mathbb{C}^n$ y $\widehat{x} = x + \Delta x$. Para saber qué tan grande podría ser Δx es posible usar algunas normas matriciales para obtener una cota para el error relativo en la solución, en términos del error relativo de los datos y el número de condición de A.

Sea $|||\cdot|||$ una norma matricial de $\mathcal{M}_n$, inducida por la norma vectorial $\|\cdot\|$ de $\mathbb{C}^n$ y nuevamente asúmase lo dicho en (4), entonces,

$$\begin{aligned}(A+\Delta A)\hat{x} &= (A+\Delta A)(x+\Delta x)\\ &= Ax + (\Delta A)x + (A+\Delta A)\Delta x\\ &= b + (\Delta A)x + (A+\Delta A)\Delta x\end{aligned}$$

y por (11)

$$(\Delta A)x + (A+\Delta A)\Delta x = \Delta b,$$

por tanto, $\Delta x = (A+\Delta A)^{-1}(\Delta b - (\Delta A)x)$ y

$$\begin{aligned}\|\Delta x\| &= \|(A+\Delta A)^{-1}(\Delta b - \Delta A x)\|\\ &\le |||(A+\Delta A)^{-1}|||\ \ (\|\Delta b\| + \|(\Delta A)x\|).\end{aligned}$$

Usando (7), se tiene que,

$$\|\Delta x\| \le \frac{|||A^{-1}|||}{1-|||A^{-1}\Delta A|||}\left(\|\Delta b\| + |||\Delta A|||\,\|x\|\right),$$

por lo tanto,

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{|||A|||\,|||A^{-1}|||}{1-|||A^{-1}\Delta A|||}\left(\frac{\|\Delta b\|}{|||A|||\,\|x\|} + \frac{|||\Delta A|||}{|||A|||}\right).$$

Tomando en consideración la definición de $k(A)$ y $\|b\| = \|Ax\| \le |||A|||\,\|x\|$, se concluye que,

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{k(A)}{1-|||A^{-1}\Delta A|||}\left(\frac{\|\Delta b\|}{\|b\|} + \frac{|||\Delta A|||}{|||A|||}\right). \quad (12)$$

Si nuevamente se asume $|||A^{-1}|||\ \ |||\Delta A||| < 1$, entonces,

$$\frac{\|\Delta x\|}{\|x\|} \le \frac{k(A)}{1-k(A)\dfrac{|||\Delta A|||}{|||A|||}}\left(\frac{\|\Delta b\|}{\|b\|} + \frac{|||\Delta A|||}{|||A|||}\right), \quad (13)$$

la cual es una cota más débil pero más fácil de calcular. Además, esta cota tiene las mismas características y consecuencias que (10), y si la matriz de coeficientes del sistema de ecuaciones lineales está bien condicionada, entonces, el error relativo en la solución tiene el mismo orden que el error relativo de los datos.

Si se conoce una solución aproximada $\hat{x}$ de $Ax = b$, ésta puede ser usada para determinar una cota a posteriori. Sean $\|\cdot\|$ una norma en $\mathbb{C}^n$, $|||\cdot|||$ la norma inducida por $\|\cdot\|$, s la solución exacta de $Ax = b$ y $r = b - A\hat{x}$ el vector residual. Como

$$A^{-1}r = A^{-1}(b - A\hat{x}) = A^{-1}b - \hat{x} = s - \hat{x},$$

entonces,

$$\|s-\hat{x}\| = \|A^{-1}r\| \le |||A^{-1}|||\,\|r\|$$

y

$$\|b\| = \|As\| \le |||A|||\ \ \|s\|,$$

es decir,

$$1 \le \frac{|||A|||\,\|s\|}{\|b\|},$$

entonces,

$$\begin{aligned}\|s-\hat{x}\| \le |||A^{-1}|||\,\|r\| &\le \frac{|||A|||\,\|s\|}{\|b\|}\,|||A^{-1}|||\,\|r\|\\ &= |||A|||\,|||A^{-1}|||\,\frac{\|r\|}{\|b\|}\,\|s\|.\end{aligned}$$

Así, el error relativo entre la solución calculada y la solución exacta está acotado como sigue,

$$\frac{\|s-\hat{x}\|}{\|s\|} \le k(A)\frac{\|r\|}{\|b\|}, \quad (14)$$

en donde la norma matricial usada para calcular el número de condición es la inducida por la norma vectorial $\|\cdot\|$. En un problema bien condicionado el error relativo en la solución es del mismo orden que el tamaño relativo del residuo; en un problema mal condicionado, sin embargo, una solución calculada que produce un residuo pequeño puede todavía estar lejos de la solución exacta.

Por tanto, el número de condición de una matriz dependerá únicamente de las normas de la matriz y de su inversa, pero el cálculo de la inversa está sujeto a errores de redondeo y éstos dependen de la precisión con la que se hacen los cálculos. Si las operaciones se hacen con aritmética de t dígitos de precisión, entonces, la aproximación al número de condición de la matriz A, es el producto de la norma de A con la norma de la aproximación de la inversa de A, obtenido con aritmética de t dígitos. En efecto, este número también depende del método usado para calcular la inversa de A. Por tal razón es conveniente poder estimar el número de condición sin necesidad de calcular directamente la inversa.

Si se asume que la solución aproximada del sistema $Ax = b$ se determina usando aritmética de t dígitos y eliminación gaussiana, en IEEE Computer Society, 2008 [7], demuestran que el vector residual r de la aproximación $\widehat{x}$ satisface,

$$\|r\| \approx 10^{-t} \; |||A||| \, \|\widehat{x}\|, \qquad (15)$$

con esta aproximación es posible estimar el número de condición, con aritmética de t dígitos, sin tener que invertir la matriz A.

La aproximación de $k(A)$ con t dígitos significativos viene de considerar el sistema de ecuaciones lineales

$$Ay = r,$$

este sistema puede ser resuelto usando en el mismo orden las mismas matrices elementales usadas para resolver $Ax = b$. Si $\widehat{y}$ es la solución aproximada de

$Ay = r$, entonces,

$$\begin{aligned} \widehat{y} \approx A^{-1}r \; &= A^{-1}(b - A\widehat{x}) \\ &= A^{-1}b - A^{-1}A\widehat{x} \\ &= s - \widehat{x} \end{aligned} \qquad (16)$$

y se obtiene una nueva aproximación de la solución dada por

$$\widehat{s} = \widehat{x} + \widehat{y},$$

es decir, $\widehat{y}$ es una estimación del error que se produce cuando $\widehat{x}$ es la aproximación de la solución del sistema original.

De las ecuaciones (15) y (16) se obtiene que,

$$\begin{aligned} \|\widehat{y}\| \approx \|s - \widehat{x}\| = \left\|A^{-1}r\right\| &\leq \left|||A^{-1}|||\,\|r\| \\ &\approx \left|||A^{-1}|||\,\left(10^{-t}\,|||A|||\,\|\widehat{x}\|\right) \\ &= 10^{-t}\,\|\widehat{x}\|\,k(A). \end{aligned}$$

En consecuencia,

$$k(A) \approx \frac{\|\widehat{y}\|}{\|\widehat{x}\|} 10^{t}, \qquad (17)$$

es una aproximación para el número de condición asociado al sistema $Ax = b$, usando eliminación gaussiana y aritmética de t dígitos.

A la repetición sistemática de este proceso se le llama refinamiento iterativo, si el sistema está bien condicionado, una o dos iteraciones serán suficientes para indicar la solución correcta, además existe la posibilidad de una mejora significativa en sistemas mal condicionados, a no ser que la matriz esté tan mal condicionada que $k(A) > 10^{t}$, como se demuestra en IEEE Computer Society [7].

Además, en [6] se advierte que si A^{-1} tiene algunas entradas relativamente grandes, entonces, algunas entradas de la solución $\widehat{s}$ pueden tener una gran e inevitable sensibilidad respecto a las perturbaciones que se hagan en algunas de las entradas de b y de A.

4.1 Descripción del algoritmo

Todo lo que se acaba de deducir y justificar, se puede sintetizar en el siguiente algoritmo, que sirve para mejorar la aproximación a la solución de un sistema de la forma $Ax = b$, especialmente cuando la matriz A está mal condicionada. Antes de iniciar la aplicación del algotirtmo se deben elegir, una norma vectorial $\|\cdot\|$ y su norma matricial inducida $|||\cdot|||$.

Inicio

Lectura de información

1. Se entra la información de A, b.
2. Se lee n, que es el número máximo de iteraciones.
3. Se lee ET , que es la cota para el error interno de redondeo, de la solución.
4. Se lee d, que es el número de dígitos de la aritmética con la que se va a trabajar.

Proceso

1. Se inicializa $k \leftarrow 0$, donde k es una variable que cuenta el número de iteraciones del algoritmo.
2. Se resuelve el sistema $Ax = b$, usando el método de Gauss-Jordan, y la respuesta aproximada se guarda en la variable *res*. Se actualiza $k \leftarrow k + 1$.
3. Se pregunta si k es menor a n.
   a. Si la respuesta es no, se ordena que realice la subrutina **Imprimir**.
   b. Si la respuesta es sí, entonces:

i. Se calcula la variable $r \leftarrow b - A \cdot res$.
ii. Se resuelve el sistema $Ax = r$, usando exactamente los mismos pasos que se realizaron en el ítem 2, y la solución se guarda en la variable y.
iii. Se calcula la variable $xx \leftarrow res + y$.
iv. Se calculan $\|y\|$, $\|xx\|$.
   • Si $k = 1$, entonces

$$cond \leftarrow \frac{\|y\|}{\|xx\|} \cdot 10^{d}.$$

v. Se pregunta si $|y| < ET$.
   • Si la respuesta es sí, entonces $k \leftarrow n$ e $ind \leftarrow true$. Esta última variable *ind* indica si ya se obtuvo una solución satisfactoria.
vi. Se actualizan $k \leftarrow k + 1$ y $res \leftarrow xx$.
vii. Se inicia de nuevo el ítem 3.

Imprimir

Si se observa que la variable indicadora *ind* es falsa, se escribe un mensaje de que no se obtuvo una solución adecuada en n pasos. En caso de que *ind* sea verdadera, se muestra la solución refinada *res*.

5. ILUSTRACIÓN NUMÉRICA

Para ilustrar la potencia del algoritmo de Refinamiento Iterativo se va a resolver el siguiente sistema de ecuaciones lineales, el cual está escrito en la forma, $Ax = b$.

$$\begin{bmatrix} \frac{4000}{3001} & \frac{7}{5} & -\frac{2}{9} \\ \frac{7}{6} & \frac{49}{40} & -\frac{1000}{5143} \\ -\frac{13}{6} & -\frac{91}{40} & \frac{13}{36} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -\frac{276272}{135045} \\ -\frac{46001}{25715} \\ \frac{299}{90} \end{bmatrix}.$$

Se puede comprobar que la solución exacta del sistema es $s = [3, -4, 2]^t$. Al aplicar el algoritmo de Refinamiento, tomando en consideración un error de tolerancia de 1×10^{-12}, la norma vectorial $\|\cdot\|_\infty$ y su norma matricial inducida $|||\cdot|||^\infty$, se observa que según (17), el número de condición aproximado de la matriz de coeficientes es 1.1427×10^{16}, esto significa que la matriz está mal condicionanda, sin embargo, el algoritmo encuentra en doce pasos, una aproximación $\hat{x}$ de la solución, para la cual el error absoluto es menor a 1×10^{-12}. Algunos de los resultados obtenidos en los doce pasos, son:

Primera iteración

Se recuerda que en esta primera iteración, lo que realmente se usa es el método de Gauss-Jordan, con una aritmética de cinco dígitos y se obtiene la aproximación,

$$\hat{x} = [3.1079,\ -7.5024,\ -19.418]^t,$$

obsérvese que el tamaño del error cometido esta dado por

$$|||s - \hat{x}|||^\infty = 17.418,$$

y por ser este resultado tan grande, se aprecia que la respuesta obtenida en este paso, es desastrosa. En las siguientes iteraciones del algoritmo, se pretende disminuir paulatinamente el error de aproximación.

Segunda iteración

1. Se toma la aproximación inicial

$$\hat{x} = [3.1079,\ -7.5024,\ -19.418]^t.$$

2. Se encuentra el vector residual,

$$r = b - A\hat{x} \approx [0.625,\ 7.92,\ 6.77]^t \times 10^{-5}$$

3. Se resuelve el sistema $Ax = r$, tomando exactamente los mismos pasos que se usaron para hallar la primera aproximación, $\hat{x}$. La solución de este nuevo sistema se guarda en una variable llamada y, en este paso,

$$y = [0.0040555,\ -0.13061,\ -0.79879]^t.$$

4. Se calcula $\|y\|_\infty = 0.79879$ y como este valor es mayor al error de tolerancia, se debe realizar una nueva iteración, actualizando el valor de $\hat{x}$ y aplicando nuevamente los pasos 1 a 3. Para actualizar el valor de $\hat{x}$, se suma y al valor de $\hat{x}$ que se obtuvo en el paso anterior. Es decir,

$$\hat{x} \leftarrow \hat{x} + y = [2.9961,\ -3.874,\ 2.7708]^t.$$

Cuarta iteración

1. $\widehat{x} = [3.0001,\ 4.0046,\ 1.972]^t$.
2. $r \approx [0.701,\ 0.578,\ -1.176]^t \times 10^{-7}$.
3. $y \approx [5.34,\ -171.91,\ -1051.4]^t \times 10^{-6}$.
4. $\|y\|_\infty = 0.0010514$, como este valor es mayor al error de tolerancia, se continua con el proceso iterativo. En este punto,

$$\widehat{x} = [2.9961,\ 3.874,\ 2.7708]^t.$$

Décimo segunda iteración

1. $\widehat{x} = [3,\ -4,\ 2]^t$.
2. $r = [0,\ 0,\ 0]^t$.
3. $y = [0,\ 0,\ 0]^t$.
4. $\|y\|_\infty = 0$, como este valor es menor al error tolerancia, se detiene el proceso y se acepta como solución, el vector $\widehat{x} = [3,\ -4,\ 2]^t$.

Al observar detalladamente los resultados presentados en esta última iteración, es natural preguntar, ¿por qué el proceso no se detuvo en la undécima iteración? El algoritmo no se detiene en una iteración anterior, pues a pesar de que el algoritmo, en algunos casos, imprime el valor $\widehat{x} = [3,\ -4,\ 2]^t$ la realidad es que en memoria el valor asignado para $\widehat{x}$ es diferente y por ejemplo en la undécima iteración, el valor $||y||_\infty = 9.8814 \times 10^{-8}$ es mayor al error de tolerancia. Por esa razón es necesario hacer la déci- mosegunda iteración, en donde sí se puede dar por terminado el algoritmo.

CONCLUSIONES

1. Para determinar si una matriz, está o no, bien condicionada, se pueden usar las propiedades de las normas matriciales para acotar el número de condición.

2. Si se usa aritmética finita de t dígitos para hallar la solución de un sistema de ecuaciones lineales mal condicionado, la aproximación ini- cial va a estar muy lejos del valor exacto, sin embargo, si $k(A) < 10^t$, es posible usar el método de Refinamiento presentado en la Sección 4, para mejorar la aproximación inicial y así, obtener una solución con un error suficientemente pequeño para ser aceptada.

REFERENCIAS

[1] V. Arunachalam y A. Calvache, "Approximation of the Bivariate Renewal Function", *Communications in Statistics - Simulation and Computation*, vol 44. , no. 1, pp. 154-167, 2015.

[2] V. Arunachalam y S. Dharmajara, "Fluid Queue Driven by Finite Markov Processes", *Revista Ciencia en Desarrollo*, vol. 5, no. 2, pp. 79-86, 2015.

[3] R. L. Burden y J. D. Faires, "Numerical Analysis", *BROOKS/COLE*, ed. 9, United States, 2011.

[4] A. Calvache, "The Transient and Asymptotic Moments for the Random Mission Time of a System", *Revista Ciencia en Desarrollo*, vol. 7, no. 2, pp. 109-124, 2016.

[5] J.I. Díaz. "John von Neumann: de la matemática pura a la matemática aplicada", *Boletín de la Sociedad Española de Matemática Aplicada*, vol. 32, pp. 149-169, 2005.

[6 R. A. Horn y C. R. Johnson, "Matrix Analysis", *Cambridge University Press*, ed. 2, New York, 2013.

[7] IEEE Computer Society, "IEEE Standard for Floating Point Arithmetic", *IEEE Std*, 2008.

[8] J. Neumann y H. H. Goldstine, "Numerical Inverting of Matrices of High Order", *Bull. American Mathematical Society*, no. 53, pp. 1021- 1099, 1947.

[9] A. Pyzzara, B. Bylina y J. Bylina, "The Influence of a Matrix Condition Number on Iterative Methods Convergence", *Conference on Computer Science and Information Systems*, pp.459- 464, 2017.

[10] A. M. Turing, "Rounding-off Errors in Matrix Processes", *Quartely J. Mech. Appl. Math*, no.1, pp. 287-308, 1948.