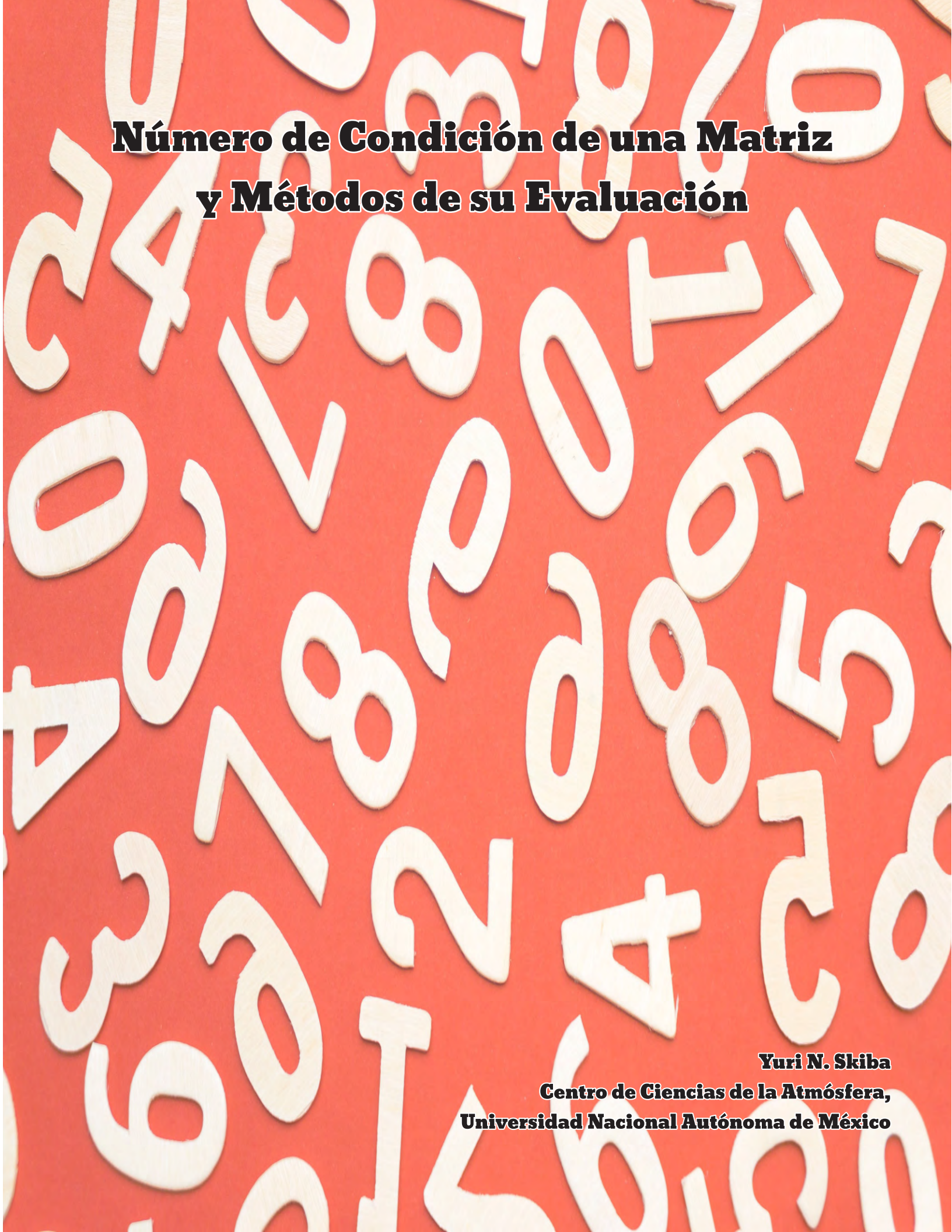
Número de Condición de una Matriz y Métodos de su Evaluación



Yuri N. Skiba
Centro de Ciencias de la Atmósfera,
Universidad Nacional Autónoma de México

Abstract

A system of linear algebraic equations $A\vec{x} = \vec{b}$ is considered. The condition number of the matrix A is introduced. It is shown to be a very important matrix characteristic, since it is the main indicator of the stability of the solution of the system, a measure of the sensitivity of the system with respect to errors in the elements of matrix A and components of vector $\vec{b}$. The greater the condition number, the stronger this effect will be and the more unstable will be the process of finding the solution of the linear system. It is shown that the condition number of a matrix does not depend on the value of its determinant. Methods are proposed to calculate or estimate this important characteristic.

Resumen

Se considera un sistema de ecuaciones algebraicas lineales $A\vec{x} = \vec{b}$. Se introduce el número de condición de la matriz A. Se muestra que es una característica matricial muy importante, ya que es el principal indicador de la estabilidad de la solución del sistema, una medida de la sensibilidad del sistema con respecto a los errores en las entradas de la matriz A y componentes del vector $\vec{b}$. Cuanto mayor sea el número de condición, más fuerte será este efecto y más inestable será el proceso de encontrar la solución del sistema lineal. Se muestra que el número de condición de una matriz no depende del valor de su determinante. Se proponen métodos para calcular o estimar esta característica importante.

Palabras clave:
Sistema de ecuaciones lineales, Número de condición de una matriz, Estabilidad de la solución de un sistema, Evaluación del número de condición

Key words: System of linear equations, Condition number of a matrix, Stability of the solution of a system, Evaluation of the condition number

1. Introducción

La condicionalidad de un sistema de ecuaciones lineales

Sea A una matriz de $n \times n$. Es bien conocido que el determinante de una matriz A es una de sus características más importantes. En efecto, si la matriz A es singular, es decir, si $\det A = 0$ entonces el sistema de ecuaciones de algebra lineal

$$A\vec{x} = \vec{b} \qquad (1)$$

tiene un conjunto infinito de soluciones o no tiene ninguna solución. Los n eigenvalores (valores propios) λ_i de la matriz A también pueden dar información valiosa sobre sus propiedades. Son soluciones del problema de eigenvalores

$$A\vec{u}_i = \lambda_i \vec{u}_i$$

donde $\vec{u}_i$ son vectores propios que corresponden a los valores propios λ_i ($i = 1,...,n$). En este trabajo se introduce otra característica muy importante de la matriz A del sistema (1).

Definición 1. El sistema de ecuaciones lineales (1) se dice *bien condicionado* si pequeños cambios en los coeficientes de la matriz A o en el lado derecho (es decir, en $\vec{b}$) causan pequeños cambios en la solución. El sistema (1) se dice *mal condicionado* cuando pequeñas perturbaciones en A o/y en $\vec{b}$ producen cambios relativamente grandes en la solución exacta $\vec{x}$.

En la siguiente sección mostraremos que la condicionalidad (buena o mala) del sistema (1) depende de la condicionalidad de su matriz A. Incluso en el caso de sistemas simples cuyas matrices son del orden de dos, las soluciones pueden ser muy sensibles a errores causados por varias fuentes (errores de redondeo, errores en métodos numéricos, etc.). De hecho, los investigadores a menudo encuentran este

problema en la solución numérica de los sistemas de ecuaciones lineales.

Ejemplo 1. Sea

$$\begin{bmatrix} 2 & 4 \\ 4 & 7.998 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} 8 \\ 15.998 \end{bmatrix} \qquad (2)$$

Su solución es $(x, y) = (2, 1)$. Haremos pequeños cambios en la primera componente del vector $\vec{b}$ del sistema (2):

$$\begin{bmatrix} 2 & 4 \\ 4 & 7.998 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} 8.001 \\ 15.998 \end{bmatrix}$$

A pesar de que el error cometido es 0.001, la solución del sistema perturbado es fundamentalmente diferente de la solución del sistema (2): $(x, y) = (0.0005, 2)$. Ahora introducimos el mismo error en un elemento de la matriz del sistema (2):

$$\begin{bmatrix} 2 & 4 \\ 4 & 7.999 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} 8 \\ 15.998 \end{bmatrix}$$

La solución del último sistema es otra vez fundamentalmente diferente de la solución del sistema (2): $(x, y) = (0, 2)$. Obviamente, el sistema (2) es mal condicionado, ya que pequeños cambios realizados en los coeficientes de la matriz o en el lado derecho, resultaron en grandes cambios en la solución del sistema. ■

Ejemplo 2. Consideremos el sistema lineal (1) con

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} 4 \\ 7 \end{bmatrix} \qquad (3)$$

Su solución es $(x,y) = (2,1)$. Después de hacer pequeños cambios en el lado derecho del sistema (3),

$$\begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} 4.001 \\ 7.001 \end{bmatrix},$$

obtenemos la solución $(x, y) = (1.999, 1.001)$. Ahora hacemos pequeños cambios en los elementos matriciales del sistema (3):

$$\begin{bmatrix} 1.001 & 2.001 \\ 2.001 & 3.001 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix}=\begin{bmatrix} 4 \\ 7 \end{bmatrix}$$

La solución del último sistema es $(x, y) = (2.003, 0.997)$. Como se puede ver, el sistema (3) es bien condicionado, ya que pequeños cambios realizados en la matriz o en el lado derecho resultaron en pequeños cambios en la solución del sistema. ■

2. Número de condición de una matriz y estabilidad de la solución

En esta sección, veremos más ejemplos de sistemas de ecuaciones lineales. Introduciremos el número de condición de la matriz del sistema y mostraremos que este número es muy importante si se usan métodos numéricos (aproximados) para resolver el sistema. En general, la precisión de los cálculos está garantizada solo en el caso de una matriz bien condicionada. Y si la matriz es mal condicionada, entonces los cálculos se realizan sin ningún control y pequeños errores en los elementos de la matriz A y/o en el vector $\vec{b}$ pueden causar errores grandes en la solución del sistema.

Ejemplo 3 (Kahan, 1966). Sea

$$A=\begin{bmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{bmatrix} \quad \text{y} \quad \vec{b}=\begin{bmatrix} 0.8642 \\ 0.1440 \end{bmatrix} \qquad (4)$$

la matriz y la parte derecha del sistema (1). Denotemos *el término residual* $\vec{r} = \vec{b} - A\vec{y}$, donde $\vec{y}$ es una solución aproximada. Ya que $\vec{r} = 0$ para la solución exacta $\vec{x} = A^{-1}\vec{b}$, es natural suponer que $\vec{y}$ es buena aproximación de la solución exacta cuando el término residual $\vec{r}$ es muy pequeño. Sin embargo, esto no es siempre una buena idea. Por ejemplo, para la

matriz (4) esta suposición no es cierta. En efecto, elegimos $\vec{y} = (0.9911, -0.4870)^T$ (índice superior "T" denota la operación transpuesta). En este caso el vector residual es $\vec{r} = (-10^{-8}, 10^{-8})^T$, es decir, muy pequeño. No obstante, el vector $\vec{y}$ queda lejos de la solución exacta $\vec{x} = (2, -2)^T$. ■

Ejemplo 4. Consideremos el sistema (1) con

$$A = \begin{bmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{bmatrix} \quad \text{y} \quad \vec{b} = \begin{bmatrix} 0.217 \\ 0.254 \end{bmatrix} \qquad (5)$$

Si elegimos $\vec{y}_1 = (0.341, -0.087)^T$ como una solución aproximada, entonces el término residual es $\vec{r}_1 = (10^{-6}, 0)^T$. Y si elegimos $\vec{y}_2 = (0.999, -1.001)^T$ como otra solución aproximada, entonces el término residual es $\vec{r}_2 = (0.0013\ldots, -0.0015\ldots)^T$. Al comparar $\vec{r}_1$ con $\vec{r}_2$ concluimos que el vector $\vec{y}_1$ es una mejor aproximación a la solución exacta $\vec{x}$ que el vector $\vec{y}_2$. No obstante, la solución exacta del sistema (5) es $(1,-1)^T$ y, en realidad, el vector $\vec{y}_2$ es la mejor aproximación entre dos vectores. ■

Surge la pregunta: "¿Por qué un sistema mal condicionado es tan inestable ?" Es fácil visualizar que ocurre en un sistema mal condicionado, en el caso de dos ecuaciones. Geométricamente, las soluciones de cada ecuación se representan por una línea directa sobre el plano, y el punto de intersección de dos líneas es la solución del sistema. Dos líneas directas que corresponden a un sistema mal condicionado son casi paralelas. En este caso, si la inclinación de una de las líneas se cambia ligeramente (por ejemplo, por pequeños errores en $\vec{b}$), entonces el punto de intersección se altera drásticamente (Fig.1).

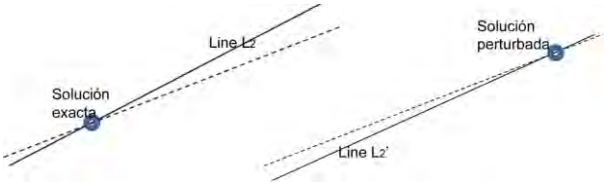


Fig.1. Impacto de una perturbación pequeña en b.

Con el fin de demostrar que un término residual muy pequeño $\vec{r} = \vec{b} - A\vec{y}$ no siempre garantiza la proximidad de la solución aproximada $\vec{y}$ a la solución exacta $\vec{x} = A^{-1}\vec{b}$, consideraremos un ejemplo más.

Ejemplo 5 (Maubach, 2005). Demostramos que dos soluciones aproximadas pueden ser muy distintas a pesar de que sus términos residuales son iguales. El sistema (1) es 2×2 con la matriz

$$A = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \end{bmatrix}$$

donde $\varepsilon > 0$ es un número muy pequeño.

Sea $\vec{w} = \vec{y} - \vec{x}$. Tenemos

$$\left\| \vec{Ay} - \vec{b} \right\|^2 = \left\| A(\vec{y} - \vec{x}) \right\|^2 = \vec{w}^*(A^*A)\vec{w} = w_1^2 + \varepsilon^2 w_2^2$$

$$= (y_1 - x_1)^2 + \varepsilon^2 (y_2 - x_2)^2$$

donde

$\left\| \vec{w} \right\| = \left| \vec{w} \right| = (\vec{w}^* \vec{w})^{1/2} = \langle w, w \rangle^{1/2} = (|w_1|^2 + |w_2|^2)^{1/2}$

es la norma euclidiana del vector $\vec{w}$ con dos componentes w_1 y w_2; $\vec{w}^*$ es el adjunto (transpuesto y complejo conjugado) del vector $\vec{w}$; $\langle \vec{x}, \vec{y} \rangle = y^* x$ es el producto escalar de dos vectores; y $A^* = \overline{A}^T = \{\overline{a}_{ji}\}$ es la adjunta (transpuesta y compleja conjugada) de la matriz $A = \{a_{ij}\}$. Por lo tanto, $\left\| \vec{Ay} - \vec{b} \right\| = a$ implica

$$\frac{(y_1 - x_1)^2}{a^2} + \frac{(y_2 - x_2)^2}{a^2 \varepsilon^{-2}} = 1,$$

es decir, todas las soluciones aproximadas $\vec{y}$, cuyas términos residuales son $\left\|A\vec{y}-\vec{b}\right\|=a$, pertenecen a un elipse con los radios a y a/ε. Por ejemplo, dos vectores $\vec{y}_1=(x_1+a,x_2)^T$ y $\vec{y}_2=(x_1,x_2+a\varepsilon^{-1})^T$ corresponden al mismo término residual $\left\|A\vec{y}-\vec{b}\right\|=a$, pero

$$\left\|\vec{y}_1-\vec{x}\right\|=a << a\varepsilon^{-1}=\left\|\vec{y}_2-\vec{x}\right\|.$$

De un lado, $\det A=\varepsilon<<1$. Por otro lado, veremos más adelante que nuestra matriz A ese mal condicionada, ya que su número de condición en la norma espectral es enorme: $\nu_2(A)=\varepsilon^{-1}$. ■

2.1. Control de los cálculos. Ahora explicaremos la inestabilidad de la solución de un sistema mal condicionado (Forsythe et al., 1977; Ciarlet, 1995). Supongamos que la matriz del sistema (1) es no singular ($\det A\neq 0$) y $\vec{b}\neq 0$. En este caso, el sistema tiene una sola solución $\vec{x}\neq 0$. Analicemos ahora un sistema perturbado

$$A(\vec{x}+\vec{\varepsilon})=\vec{b}+\vec{\delta} \qquad (6)$$

donde $\vec{\varepsilon}$ es el error absoluto cometido en el cálculo de la solución $\vec{x}$ del problema (1), y $\vec{\delta}$ es el *error absoluto* cometido en vector $\vec{b}$ (supongamos que no hay errores en la matriz A). Claro que

$$A\vec{\varepsilon}=\vec{\delta}, \quad \text{y} \quad \vec{\varepsilon}=A^{-1}\vec{\delta}. \qquad (7)$$

Dividiendo el error relativo $\|\vec{\varepsilon}\|/\|\vec{x}\|$ cometido en $\vec{x}$ por el error relativo $\left\|\vec{\delta}\right\|/\left\|\vec{b}\right\|$ cometido en $\vec{b}$, y usando (1) y (7) obtenemos

$$\frac{\|\vec{\varepsilon}\|/\|\vec{x}\|}{\left\|\vec{\delta}\right\|/\left\|\vec{b}\right\|}=\frac{\left\|\vec{b}\right\|}{\|\vec{x}\|}\cdot\frac{\|\vec{\varepsilon}\|}{\left\|\vec{\delta}\right\|}=\frac{\|A\vec{x}\|}{\|\vec{x}\|}\cdot\frac{\left\|A^{-1}\vec{\delta}\right\|}{\left\|\vec{\delta}\right\|}\le\|A\|\left\|A^{-1}\right\| \qquad (8)$$

Definición 2. Sea A una matriz. El número

$$\nu(A)\equiv cond\,A=\begin{cases}\|A\|\left\|A^{-1}\right\|, & \text{si } A \text{ no es singular}\\ \infty, & \text{si } A \text{ es singular}\end{cases} \qquad (9)$$

se denomina número de condición de la matriz A. ■

Se deduce de (8) y (9) que

$$\frac{\|\vec{\varepsilon}\|}{\|\vec{x}\|}\le\nu(A)\frac{\left\|\vec{\delta}\right\|}{\left\|\vec{b}\right\|}, \qquad (10)$$

es decir, el error relativo cometido en la solución $\vec{x}$ del problema (1) se estima mediante el error relativo cometido en el vector $\vec{b}$ multiplicado por el número de condición de la matriz. Por eso, cuando $\nu(A)$ es pequeño o moderado, el error $\|\vec{\varepsilon}\|/\|\vec{x}\|$ en la solución del problema (1) está acotado y depende continuamente del error $\left\|\vec{\delta}\right\|/\left\|\vec{b}\right\|$ en $\vec{b}$ (en el sentido de que $\|\vec{\varepsilon}\|/\|\vec{x}\|$ tiende a cero junto con $\left\|\vec{\delta}\right\|/\left\|\vec{b}\right\|$). En esta situación, la matriz A (y por consiguiente, el sistema (1)) se llama *bien condicionada* (véase Ejemplo 2). Sin embargo, si el número de condición de la matriz A es muy grande entonces el error en la solución $\|\vec{\varepsilon}\|/\|\vec{x}\|$ ya no es controlable a pesar de que el error $\left\|\vec{\delta}\right\|/\left\|\vec{b}\right\|$ es muy pequeño. En efecto, si $\nu(A)$ es 10^{12} y $\left\|\vec{\delta}\right\|/\left\|\vec{b}\right\|$ es 10^{-10} obtenemos

$$\frac{\|\vec{\varepsilon}\|}{\|\vec{x}\|}\le 100,$$

y no hay control sobre la precisión de los cálculos. En la última situación, el sistema (1) y su matriz A se llaman *mal condicionados*, y pueden aparecer errores graves durante la solución numérica del problema (véase Ejemplo 1).

El análisis permite contestar la pregunta sobre el extraño comportamiento de las soluciones en

los ejemplos 3 y 4. En efecto, lo que pasa en dichos ejemplos se debe a la condicionalidad mala de las matrices (4) y (5), cuando la estimación (10) no controla los cálculos, y un error pequeño en el vector $\vec{b}$ produce un error bastante grande en la solución $\vec{x}$.

Ahora mostraremos que el número de condiciones (9) también es una característica importante al evaluar la respuesta del sistema (1) a los errores cometidos en los elementos de la matriz A. De hecho, supongamos que el vector $\vec{b}$ es exacto, pero A contiene un error δA:

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

Así, en lugar de la solución exacta $\vec{x}=A^{-1}\vec{b}$, tenemos una solución aproximada $\vec{x}+\vec{\varepsilon}=\left(A+\delta A\right)^{-1}\vec{b}$, o $\vec{\varepsilon}=\{\left(A+\delta A\right)^{-1}-A^{-1}\}\vec{b}$. Sustituyendo $B=A+\delta A$ en la identidad $B^{-1}-A^{-1}=A^{-1}(A-B)B^{-1}$, se obtiene

$$\vec{\varepsilon}=-A^{-1}\,\delta A\left(A+\delta A\right)^{-1}\vec{b}=-A^{-1}\delta A(\vec{x}+\vec{\varepsilon})$$

Por lo tanto, $\|\vec{\varepsilon}\|\le\|A^{-1}\|\,\|\delta A\|\,\|\vec{x}+\vec{\varepsilon}\|$. Se deduce que

$$\frac{\|\vec{\varepsilon}\|}{\|\vec{x}+\vec{\varepsilon}\|}\le\nu(A)\frac{\|\delta A\|}{\|A\|}$$

y si $\nu(A)\dfrac{\|\delta A\|}{\|A\|}<1$ entonces

$$\frac{\|\vec{\varepsilon}\|}{\|\vec{x}\|}\le\nu(A)\frac{\|\delta A\|}{\|A\|}\left(1-\nu(A)\frac{\|\delta A\|}{\|A\|}\right)^{-1} \qquad 11)$$

Por lo tanto, si $\nu(A)\dfrac{\|\delta A\|}{\|A\|}\square\ 1$ entonces

$\left(1-\nu(A)\dfrac{\|\delta A\|}{\|A\|}\right)^{-1}$ es cerca a uno, y el error relativo en la solución está nuevamente limitado por el error relativo en la matriz A multiplicado por el número de condición (9).

Es fácil demostrar que en el caso general cuando se presentan ambos tipos de errores (δA y δ) y $\nu(A)\dfrac{\|\delta A\|}{\|A\|}<1$, la estimación (11) acepta la forma

$$\frac{\|\vec{\varepsilon}\|}{\|\vec{x}\|}\le\nu(A)\left(\frac{\|\delta\|}{\|\vec{b}\|}+\frac{\|\delta A\|}{\|A\|}\right)\left(1-\nu(A)\frac{\|\delta A\|}{\|A\|}\right)^{-1} \qquad (12)$$

Es obvio que, (10) y (11) se deducen de (12) en los casos particulares cuando $\delta=0$ y $\delta A=0$, respectivamente. Así, el número de condición $\nu(A)$ es una medida de la sensibilidad del sistema de ecuaciones lineales, determinada por su matriz, a los errores en las entradas de la matriz y el vector del lado derecho de la ecuación. Además, cuanto mayor sea el número de condición, más fuerte será este efecto y más inestable será el proceso de encontrar una solución para un sistema lineal.

2.2. Límite inferior para el número de condición. Según (9), el número de condición $\nu(A)$ depende de una norma matricial $\|A\|$ elegida

$$\nu(A)=\|A\|\,\|A^{-1}\| \qquad (13)$$

Notemos que

$$\nu(A)=\|A\|\,\|A^{-1}\|\ge\|AA^{-1}\|=\|E\|\ge\|E\|_2=1 \qquad (14)$$

donde *E* es la matriz identidad, y

$$\|A\|_2=\max_{\|\vec{x}\|=1}\langle\vec{Ax},\vec{Ax}\rangle^{1/2} \qquad (15)$$

se llama norma espectral de A.

Así, $\nu(A)\ge1$ y $\nu(A)$ no puede ser menor que uno en ninguna norma matricial. Para una computadora específica, también puede especificar el límite superior, cuyo exceso puede llevar a decisiones deliberadamente

falsas: la solución se considera no confiable si $\nu(A) \geq \delta^{-1}$ o incluso $\nu(A) \geq \delta^{-1/2}$, donde δ es un único error de redondeo (precisión de la computador a). Es importante tener en cuenta que el escalamiento de la matriz A multiplicándola por un escalar no cambia su número de condición.

Ejemplo 6. Especificamos el número de condición de una matriz simétrica, no singular A usando la norma espectral (15). Ya que A es simétrica, (15) se convierte en

$$\|A\|_2 = \max_{1 \leq i \leq n} |\lambda_i(A)|$$

donde $\lambda_i(A)$ es un eigenvalor de la matriz A. Como $(A^{-1})^T = (A^T)^{-1} = A^{-1}$, la matriz inversa también es simétrica. Además,

$$\|A^{-1}\|_2^2 = \max_{\|\vec{x}\|=1} \langle A^{-1}\vec{x}, A^{-1}\vec{x} \rangle = \left\{ \min_{\|\vec{x}\|=1} \langle A^{-1}\vec{x}, A^{-1}\vec{x} \rangle \right\}^{-1}$$

$$= \left\{ \min_{\|\vec{y}\|=1} \langle A\vec{y}, A\vec{y} \rangle \right\}^{-1} = \left\{ \min_{\|\vec{y}\|} \langle A^2\vec{y}, \vec{y} \rangle \right\}^{-1} = \frac{1}{\min_{1 \leq i \leq n} |\lambda_i(A)|^2}$$

Así, el número de condición de una matriz simétrica A en la norma espectral es

$$\nu_2(A) = \max_{1 \leq i \leq n} |\lambda_i(A)| / \min_{1 \leq i \leq n} |\lambda_i(A)| \qquad (16)$$

y se llama *número de condición espectral* de A. La fórmula (16) es válida para cualquier matriz normal ($AA^* = A^*A$) no singular ($\det A \neq 0$). En el caso particular cuando $A = \text{diag}\{d_1, d_2, ..., d_n\}$ es una matriz diagonal,

$$\nu_2(A) = \max_{1 \leq i \leq n} |d_i| / \min_{1 \leq i \leq n} |d_i| \,. \blacksquare$$

Ejemplo 7. Demostramos que cualquier matriz unitaria $n \times n$ es perfectamente condicionada en la norma espectral. En efecto, sea U una matriz unitaria, es decir, $UU^{-1} = E$. Es bien conocido que en la norma euclidiana,

$$\|\vec{x}\| \equiv |\vec{x}| = |U\vec{x}| \equiv \|U\vec{x}\|$$

para cualesquier $\vec{x}$ y matriz unitaria U (Skiba, 2005). Por lo tanto, la norma espectral de U y de su matriz inversa $U^{-1} = U^*$ son iguales a uno. Así, el número de condición espectral de U es

$$\nu(U) \equiv \nu_2(U) = \|U\|_2 \|U^{-1}\|_2 = 1 \qquad (17)$$

Sin embargo, si $\nu(U)$ de una matriz unitaria U de orden n se estima en la norma de Frobenius,

$$\|U\|_F = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |u_{ij}|^2 \right)^{1/2},$$

entonces el número de condición $\nu_F(U)$ ya depende del orden de U:

$$\nu_F(U) = \|U\|_F \|U^{-1}\|_F = n \qquad (18)$$

En efecto (Skiba, 2018),

$$\|U\|_F = \|UE\|_F = \|E\|_F = \sqrt{n}$$

y $\|U^{-1}\|_F = \|U^{-1}E\|_F = \|E\|_F = \sqrt{n}$. $\blacksquare$

Se puede demostrar que para la norma espectral, la igualdad $\nu_2(A) = 1$ se cumple si y solo si $A = \alpha Q$ o $A = \alpha U$, donde α es un número, Q es una matriz ortogonal y U es una matriz unitaria.

Sean Q y U dos matrices unitarias u ortogonales. Se deduce de las fórmulas

$$\|A\|_2 = \|QAU\|_2 \quad \text{y} \quad \|A\|_F = \|QAU\|_F$$

que

$$\nu_2(A) = \nu_2(QAU) \quad \text{y} \quad \nu_F(A) = \nu_F(QAU) \qquad (19)$$

donde $\nu_2(A)$ y $\nu_F(A)$ son los números de condición de una matriz A, calculados usando la norma espectral y la norma de Frobenius, respectivamente.

Las siguientes desigualdades tienen lugar (Voevodin y Kuznetzov, 1984):

$$\max\{\frac{\nu(A)}{\nu(B)},\frac{\nu(B)}{\nu(A)}\}\le\nu(AB)\le\nu(A)\,\nu(B) \qquad (20)$$

Ejemplo 8. Consideremos la matriz de Hilbert

$$H_n=\left[h_{ij}\right]=\begin{bmatrix}1 & \frac{1}{2} & \cdots & \frac{1}{n}\\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1}\\ \cdots & \cdots & \cdots & \cdots\\ \frac{1}{n} & \frac{1}{n+1} & \cdots & \frac{1}{2n-1}\end{bmatrix},\quad h_{ij}=\frac{1}{i+j-1}. \qquad (21)$$

Aparece al minimizar el error e de aproximación de una función $f(x)$ en el intervalo $0\le x\le 1$ por un polinomio algebraico $\sum_{i=1}^{n}c_i x^{i-1}$ (Skiba, 2018):

$$e=\int_0^1\{\sum_{i=1}^{n}c_i x^{i-1}-f(x)\}^2dx$$

Horn y Johnson (1999) mencionan que el número de condición de H_n empeora cuando la dimensión n aumenta y asintóticamente coincide con la función exponencial e^{cn}, donde la constante c es aproximadamente igual a 3.5. Por ejemplo, $\nu_2(H_n)$ crece como $e^{3.5n}$:

$$\nu_2(H_3)\approx 5\cdot 10^2,\quad \nu_2(H_6)\approx 1.5\cdot 10^7,\ \nu_2(H_8)\approx 1.5\cdot 10^{10},$$

Notemos que la norma de Frobenius $\|H_n\|_F$ tiende a infinito cuando n aumenta, ya que

$$\|H_n\|_F^2\ge\sum_{k=1}^{n}\frac{1}{k},$$

y la serie del lado derecho diverge. Sin embargo, para el radio espectral $\rho(H_n)=\max\limits_{1\le i\le n}|\lambda_i(A)|$ de la matriz de Hilbert es válida la estimación

$$\rho(H_n)=\pi+O\left(\frac{1}{\log n}\right)\quad\text{cuando}\quad n\to\infty.$$

Así, a pesar de que los elementos de la matriz (21) son uniformemente acotados en n, y su radio espectral no es grande, la matriz es mal condicionada cuando n es grande. El hecho es que H_n es simétrica y, según (16), su número de condición espectral es

$$\nu_2(H_n)=\rho(H_n)/\min_i|\lambda_i(H_n)|$$

Entonces el valor propio del módulo más pequeño $\min\limits_{1\le i\le n}|\lambda_i(H_n)|$ de la matriz de Hilbert tiende a cero cuando $n\to\infty$. ■

2.3. Equivalencia de los números de condición. En virtud de la equivalencia de dos normas matriciales arbitrarias (Skiba, 2005), se cumplen desigualdades

$$C\|A\|_p\le\|A\|_q\le K\|A\|_p \qquad (22)$$

para cualquier matriz *A*, donde C *y K* son dos constantes universales positivas que dependen solo de las normas elegidas $\|\cdot\|_p$ y $\|\cdot\|_q$, y no dependen de *A*. Se deduce de (22) la equivalencia de los números de condición

$$C^2\nu_p(A)\le\nu_q(A)\le K^2\nu_p(A) \qquad (23)$$

donde C *y K* son las constantes de (22). Así, los números de condición de una matriz A calculados en dos normas diferentes, también son equivalentes, es decir, si A es bien (o mal) condicionada en una norma y las constantes C y K no son enormes o muy pequeñas, entonces, según (23), A también es bien (mal) condicionada en otra norma.

3. Estimación del número de condición

Como sabemos ahora, el número de condición $\nu(A)=\|A\|\|A^{-1}\|$ de la matriz *A* es un indicador importante de la estabilidad de la solución del sistema (1). A pesar de que es muy útil conocer esta característica, su cálculo no es trivial, ya que el factor $\|A^{-1}\|$ está desconocido, ya que la matriz inversa A^{-1} es desconocida. De hecho, el cálculo de la matriz inversa A^{-1} resuelve el problema (1), ya que formalmente $\vec{x}=A^{-1}\vec{b}$. Por lo tanto, el cálculo del número de condición

utilizando su definición, requeriría mucho más trabajo que el cálculo de la solución en sí misma, cuya precisión debe evaluarse. En este sentido, en la práctica, el número de condición se estima económicamente como un subproducto del proceso de solución del sistema (1).

En virtud de lo anterior, es deseable contar con un conjunto de métodos para la estimación aproximada del número de condición. Ahora consideramos algunos de estos métodos.

Ejemplo 9. Calculemos el número de condición de la matriz triangular de Toeplitz

$$T = \begin{bmatrix} 2 & -1 & \cdots & 0 & 0 \\ -1 & 2 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 2 & -1 \\ 0 & 0 & \cdots & -1 & 2 \end{bmatrix} \qquad (24)$$

de orden *n*. La matriz es simétrica y definida positiva, es decir, todos sus eigenvalores son positivos. Es bien conocido que los eigenvalores de *T* se hallan mediante la fórmula

$$\lambda_k(T) = 2(1+\cos\frac{k\pi}{n+1}) = 2(1+\cos kh) \qquad (25)$$

donde $h = \pi/(n+1)$ (Smith, 1978). Ya que $\cos(n+1)h = \cos\pi = -1$, y, por lo tanto, $\cos nh = \cos n\pi/(n+1) = -\cos h$. Usando (25) obtenemos

$$\begin{aligned} \min_{1\le i\le n}|\lambda_i(T)| &= \lambda_n(T) = 2(1-\cos h) \\ \max_{1\le i\le n}|\lambda_i(T)| &= \lambda_1(T) = 2(1+\cos h) \end{aligned} \qquad (26)$$

Según (16), tenemos

$$\nu_2(T) = \frac{1+\cos h}{1-\cos h} \qquad (27)$$

Si h es pequeño, entonces $\cos h \approx 1 - h^2/2$, y

$$\nu_2(T) = \frac{4-h^2}{h^2} = O(h^{-2}), \qquad (28)$$

es decir, la matriz de Toeplitz (24) es bien condicionada si, por ejemplo, $h = 10^{-3}$ y el error de redondeo de la computadora es 10^{-10}. Matrices de Toeplitz surgen a menudo al aproximar el problema unidimensional de contorno para el operador de Laplace. ■

3.1. Determinante y el número de condición. Es preciso notar que el determinante y el número de condición son dos características matriciales bastante independientes. Los siguientes dos ejemplos muestran que $\nu(A)$ es un mejor criterio para estimar la degeneración de matrices cuadradas que el determinante.

Ejemplo 10. Consideremos la matriz diagonal $D_n = \text{diag}\,(10^{-1}, 10^{-1}, \ldots, 10^{-1})$ del orden *n*. Es perfectamente condicionada, ya que, según (16), $\nu_2(D_n) = 1$ para cualquier *n*. Sin embargo, $\det(D_n) = 10^{-n}$, es decir, el determinante tiende a cero al aumentar *n*. Así, una matriz casi singular puede ser perfectamente condicionada. ■

Ejemplo 11. Sea aQ una matriz ortogonal de orden *n*, donde a es un número. Entonces

$$\det(aQ) = a^n \det Q = \pm a^n,$$

es decir, el determinante de la matriz aQ puede ser arbitrariamente pequeño (si $|a| < 1$) o grande (si $|a| > 1$), aunque la matriz aQ es perfectamente condicionada. ■

3.2. Estimación del número $\nu(A)$ **desde abajo**. Ahora escribimos un método que usa la desigualdad (10) y estima el número de condición $\nu(A)$ de la matriz A desde abajo, es decir, a veces permite demostrar que este número es enorme.

Ejemplo 12. Consideremos la matriz

$$A=\begin{bmatrix} 1 & -1 & -1 & \cdots & -1 & -1 \\ 0 & 1 & -1 & \cdots & -1 & -1 \\ 0 & 0 & 1 & \cdots & -1 & -1 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 & -1 \\ 0 & 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \qquad (29)$$

del orden n, cuyo determinante es uno. Demostremos que A es mal condicionada. Examinemos el sistema (1) con la matriz (29) y el vector columna $\vec{b}=(-1,-1,\ldots,-1,1)^T$ con todas sus componentes iguales a -1, excepto la última componente que es uno. En una forma más detallada, este sistema tiene el aspecto siguiente:

$$\begin{aligned} x_1 - x_2 - x_3 - \ldots - x_n &= -1 \\ x_2 - x_3 - \ldots - x_n &= -1 \\ \cdots\cdots\cdots\cdots & \\ x_{n-1} - x_n &= -1 \\ x_n &= 1 \end{aligned} \qquad (30)$$

El sistema (30) tiene la única solución $\vec{x}=(0,0,\ldots,0,1)^T$ que se puede obtener fácilmente mediante la sustitución regresiva.

Supongamos ahora que en la sustitución regresiva usada para resolver el sistema (30) se ha cometido un solo error: en lugar del número $b_n=1$ se ingresó el número $b_n=1+\delta$, donde $\delta>0$ es muy pequeño en comparación con la unidad. Entonces, en vez de la solución exacta $\vec{x}=(0,0,\ldots,0,1)^T$ del sistema (30) obtendremos la solución perturbada $\vec{x}+\vec{\varepsilon}$ del sistema $A(\vec{x}+\vec{\varepsilon})=\vec{b}+\vec{\delta}$, donde $\vec{\delta}=(0,0,\ldots,0,\delta)^T$ y el error $\vec{\varepsilon}=(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_n)^T$ satisface el sistema lineal

$$\begin{aligned} \varepsilon_1 - \varepsilon_2 - \varepsilon_3 - \ldots - \varepsilon_n &= 0 \\ \varepsilon_2 - \varepsilon_3 - \ldots - \varepsilon_n &= 0 \\ \cdots\cdots\cdots\cdots & \\ \varepsilon_{n-1} - \varepsilon_n &= 0 \\ \varepsilon_n &= \delta \end{aligned} \qquad (31)$$

De aquí obtenemos

$\varepsilon_n=\delta,\ \varepsilon_{n-1}=\delta,\ \varepsilon_{n-2}=2\delta,\ldots,\ \varepsilon_{n-k}=2^{k-1}\delta,\ \ldots,\varepsilon_1=2^{n-2}\delta$

Usando la ∞-norma vectorial, a saber, $\|\vec{x}\|_\infty=\max\limits_{1\le i\le n}|x_i|$, obtenemos

$$\|\vec{\varepsilon}\|_\infty=2^{n-2}\delta,\ \ \|\vec{x}\|_\infty=1,\ \ \|\vec{\delta}\|_\infty=\delta,\ \ \|\vec{b}\|_\infty=1 \qquad (32)$$

y según (10),

$$\nu_\infty(A)\equiv\|A\|_\infty\left\|A^{-1}\right\|_\infty\ge\frac{\|\vec{\varepsilon}\|_\infty/\|\vec{x}\|_\infty}{\|\vec{\delta}\|_\infty/\|\vec{b}\|_\infty}=2^{n-2} \qquad (33)$$

Por ejemplo, si $n=102$ (el orden de la matriz es bastante pequeño), (33) lleva a $\nu_\infty(A)\ge 2^{100}>10^{30}$, es decir, la matriz es mal condicionada. Además, según (32), $\|\vec{\varepsilon}\|_\infty=2^{100}\delta>10^{30}\delta$.

Particularmente, supongamos que el único error cometido en la sustitución regresiva es muy pequeño: $\delta=10^{-15}$. Sin embargo, el error cometido en la solución es enorme: $\|\vec{\varepsilon}\|_\infty>10^{15}$ (en comparación con $\|\vec{x}\|_\infty=1$).■

3.3. Estimación del número $\nu(A)$ desde arriba. Consideremos ahora otro método que tiene la aplicación limitada, pero permite fácilmente evaluar el número de condición $\nu(A)=\|A\|\left\|A^{-1}\right\|$ desde arriba y, por lo tanto, es muy útil para demostrar que la matriz de un sistema es bien condicionada.

Sea $A\vec{x}=\vec{b}$ un sistema para resolver. Introducimos otra matriz, $B=E-A$, y representamos el sistema original como

$$\vec{x}=B\vec{x}+\vec{b} \qquad (34)$$

El método que describimos ahora es válido sólo para un grupo de las matrices A que satisfacen la condición

$$\|B\| = \|E - A\| < 1 \qquad (35)$$

en una norma matricial. A condición de que (35) se cumple, el problema (34) tiene una solución única $\vec{x}_*$ y

$$\|\vec{x}_*\| \equiv \left\|A^{-1}\vec{b}\right\| \le \frac{\|\vec{b}\|}{1-\|B\|} \qquad (36)$$

para cualquier vector $\vec{b}$. El denominador en (36) es positivo debido a (35). Al usar la norma matricial subordinada se obtiene

$$\left\|A^{-1}\right\| = \max_{\vec{b}\neq 0} \frac{\left\|A^{-1}\vec{b}\right\|}{\|\vec{b}\|} \le \frac{1}{1-\|B\|}$$

Por otro lado,

$\|A\| = \|E - B\| \le \|E\| + \|B\| < 1 + \|E\|$. Entonces,

según (9),

$$\nu(A) = \|A\|\left\|A^{-1}\right\| \le \frac{1+\|E\|}{1-\|B\|} \qquad (37)$$

Ejemplo 13. Evaluamos el número de condición de la matriz $A = E - B$, donde

$$b_{ij} = \frac{0.8}{n}\cdot(-1)^{i+j}, \quad 1 \le i, j \le n \qquad (38)$$

son los elementos de la matriz B. Tenemos

$$\|B\|_\infty \equiv \max_i \sum_{j=1}^{n} \left|b_{ij}\right| = \sum_{j=1}^{n} \left|\frac{0.8}{n}\right| = 0.8$$

y también,

$$\|B\|_2 \le \|B\|_F \equiv \left(\sum_{j=1}^{n} b_{ij}^2\right)^{1/2} = 0.8$$

Por lo tanto, la condición (35) se cumple tanto en la norma espectral como en la norma de Frobenius. Entonces, de acuerdo con la fórmula (37),

$$\nu(A) = \nu(E - B) \le \frac{1+1}{1-0.8} = 10$$

Así, el número de condición de la matriz A es pequeño, es decir, A es bien condicionada. ■

3.4. Simetrización de un sistema de ecuaciones. Consideremos el sistema (1) con una matriz normal no singular A. Tratando de mejorar la estructura de la matriz del sistema, se puede transformar (1) al sistema

$$A^* A\vec{x} = A^*\vec{b} \qquad (39)$$

con la matriz hermitiana A^*A. Sin embargo, es válida la siguiente afirmación:

Teorema 1. La simetrización de una matriz A del sistema (1) sólo aumenta el número de condición de la matriz A^*A del sistema nuevo (39):

$$\nu_2(A^*A) \ge \nu_2(A) \qquad (40)$$

Demostración (véase, por ejemplo, Skiba, 2018). Ya que A es normal entonces, según el teorema general de factorización, existen matrices unitarias U, V y una matriz diagonal $D = \text{diag}\{\mu_1, \mu_2, ..., \mu_n\}$ tales que

$$A = VDU^* = VDU^{-1}$$

donde $\mu_i = \sqrt{\lambda_i(A^*A)} \ge 0$ son los números singulares de la matriz A. Por lo tanto,

$$A^* = UDV^*, \quad A^{-1} = UD^{-1}V^*$$
$$\text{y} \quad (A^*)^{-1} = VD^{-1}U^*.$$

Así pues,

$$A^*A = UDV^*VDU^* = UD^2U^*,$$
$$(A^*A)^{-1} = A^{-1}(A^*)^{-1} = UD^{-1}V^*VD^{-1}U^* = UD^{-2}U^*$$

Las transformaciones unitarias no cambian la norma espectral de una matriz y, por lo tanto,

$$\nu_2(A) = \|A\|_2 \|A^{-1}\|_2 = \|VDU^*\|_2 \|UD^{-1}V^*\|_2 = \|D\|_2 \|D^{-1}\|_2$$

.

Al tomar en cuenta las ecuaciones

$$\nu_2(A^*A) = \|A^*A\|_2 \|(A^*A)^{-1}\|_2 = \|UD^2U^*\|_2 \|UD^{-2}U^*\|_2$$

$$= \|D^2\|_2 \|D^{-2}\|_2 = \|D\|_2^2 \|D^{-1}\|_2^2$$

y la desigualdad $\nu_2(A) \geq 1$ (véase (14)) se obtiene

$$\nu_2(A^*A) = \|D\|_2^2 \|D^{-1}\|_2^2 = \nu_2^2(A) \geq \nu_2(A) \text{ .}\blacksquare$$

En la demostración de la desigualdad (40) se usa la norma espectral de las matrices. Pero, en la realidad, solo se usa la propiedad de que las transformaciones unitarias no cambian la norma espectral. Ya que las transformaciones unitarias tampoco cambian la norma de Frobenius, se obtiene

$$\nu_F(A^*A) \geq \nu_F(A). \qquad (41)$$

Así, la simetrización de la matriz A del sistema (1) sólo aumenta el número de condición de la matriz A^*A del sistema nuevo (39), lo que hace que su solución sea aún más sensible a los errores en los elementos de la matriz A y las componentes del vector $\vec{b}$, como se muestra el siguiente ejemplo.

Ejemplo 14. Consideremos el sistema

$$\begin{bmatrix} 1 & 1 \\ 2 & 2.01 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 4 \end{bmatrix} \qquad (42)$$

cuya matriz no es simétrica. La solución es $(x, y) = (2, 0)$. Perturbamos ahora la última componente del vector $\vec{b}$ del sistema (42) con un error pequeño 0.01:

$$\begin{bmatrix} 1 & 1 \\ 2 & 2.01 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 2 \\ 4.01 \end{bmatrix} \qquad (43)$$

La solución del sistema perturbado (43) cambia drásticamente: $(x, y) = (1, 1)$, es decir, la matriz del sistema (42) es mal condicionada, y su solución es inestable. Después de la simetrización el sistema (42) acepta la forma

$$\begin{bmatrix} 5 & 5.02 \\ 5.02 & 5.0401 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 10 \\ 10.04 \end{bmatrix} \qquad (44)$$

Por supuesto, tiene la misma solución que (42): $(x, y) = (2, 0)$. Ahora introducimos el mismo error (0.01 en el sistema (43)) en la parte derecha del sistema (44):

$$\begin{bmatrix} 5 & 5.02 \\ 5.02 & 5.0401 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 10 \\ 10.05 \end{bmatrix} \qquad (45)$$

La solución del sistema perturbado (45) ya es $(x, y) = (-500, 500)$. La comparación de las soluciones $(x, y) = (1, 1)$ y $(x, y) = (-500, 500)$ de dos sistemas perturbados (43) y (45) con la solución exacta $(x, y) = (2, 0)$ muestra que el sistema simetrizado (44) está mucho más inestable que el sistema original (42). ■

4. Conclusiones

Este artículo analiza una característica importante de un sistema de ecuaciones algebraicas lineales $A\vec{x} = \vec{b}$. La característica está asociada con la matriz A del sistema y se denomina número de condición de la matriz. Este es el principal indicador de la estabilidad de la solución del sistema con respecto a pequeños errores en la matriz A y el vector $\vec{b}$. Dicha inestabilidad crece junto con el número de condición de la matriz. Si la matriz A es mal condicionada, entonces es prácticamente imposible encontrar su solución con suficiente precisión, siempre que los cálculos se realicen en presencia de errores (al menos errores de

redondeo). Se muestra que el número de condición no depende del valor del determinante de A. También se demuestra que la simetrización de la matriz A solo puede empeorar su número de condición. En el trabajo, se consideran métodos para calcular o evaluar esta importante característica del sistema.

Agradecimientos

Esta investigación fue apoyada por la beca 14539 del Sistema Nacional de Investigadores (SNI-CONACyT, México).

Referencias

Ciarlet, P.G., *Introduction to Numerical Linear Algebra and Optimization*. Cambridge, Cambridge University Press, 1995.

Forsythe, G.E., Malcolm, M.A. and Moler, C.B., *Computer Methods for Mathematical Computations*. Prentice-Hall, Englewood Cliffs, N.J., 1977.

Horn, R.A. and Johnson, Ch.R., *Matrix Analysis.* Cambridge, Cambridge University Press, 1999.

Kahan, W., Numerical linear algebra, Canadian Math. Bulletin, 9, pp. 757-801, 1966.

Maubach J.M., *Numerical Methods in Scientific Computing.* University of Pittsburgh, 2005.

Skiba, Yu.N., *Métodos y Esquemas Numéricos. Un Análisis Computacional*. México, Dirección General de Publicaciones y Fomento Editorial, La Universidad Nacional Autónoma de México, México, 2005.

Skiba, Yu.N., *Fundamentos de los Métodos Computacionales en Álgebra Lineal.* Dirección General de Publicaciones y Fomento Editorial, La Universidad Nacional Autónoma de México, México, 2018.

Voevodin, V.V. and Kuznetzov, Yu.A., *Matrices and Calculations*. Moscow, Nauka, 1984 (en ruso).