¿Es fácil resolver en la computadora un sistema de ecuaciones lineales 2×2?

6 de marzo de 2012

Resumen

Resolver un sistema de ecuaciones lineales en la computadora, aparentemente no implica mayor dificultad; en teoría basta con diseñar un pequeño algoritmo basado en la eliminación gaussiana para lograrlo. No obstante, como veremos, sin las medidas adecuadas, la resolución de un sistema de ecuaciones simultáneas utilizando la computadora puede generar resultados absurdos o erróneos. Este trabajo destaca la necesidad de modificar el método de resolución con el fin de obtener un buena aproximación de la solución por medios computacionales; también se analizan algunos casos excepcionales de sistemas 2×2 que nos permitirán establecer la importancia y necesidad de construir un estimador adecuado para medir el mal condicionamiento de los sistemas.

1 Eliminación gaussiana en la computadora

A la pregunta ¿es fácil resolver un sistema de ecuaciones lineales de 2×2 en la computadora? le siguen dos probables respuestas, y casi invariablemente la respuesta es afirmativa.

1.1 ¿Realmente es fácil?

Evitemos las adivinanzas, veamos qué tan fácil es dicha actividad en la computadora. En general un sistema 2×2 es de la forma

$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2$$

Supongamos que el sistema tiene solución única y $a_{11} \neq 0$. Por eliminación gaussiana

$$\begin{aligned} a_{11}x_1 + \quad & a_{12}x_2 = b_1 \\ & (a_{22} - ma_{12})x_2 = b_2 - mb_1 \end{aligned}$$

donde $m = a_{21}/a_{11}$. Despejando x_2 de la segunda ecuación obtenemos

$$x_2 = \frac{b_2 - mb_1}{a_{22} - ma_{12}}$$

Sustituyéndola en la primera ecuación y despejando x_1 tenemos

$$x_1 = \frac{b_1 - a_{12}x_2}{a_{11}}$$

Ahora veamos qué pasa al utilizar las fórmulas de x_2 y x_1 para resolver algunos sistemas de ecuaciones lineales utilizando la computadora, en particular usamos una hoja de cálculo y sugerimos al lector realizar la misma actividad utilizando el *software* computacional de su preferencia. El primer sistema es

$$2x_1 + 5x_2 = 3$$
$$3x_1 - 2x_2 = -5$$

cuya solución exacta es $x_1 = -1, x_2 = 1$. En la siguiente tabla se muestra el resultado de los cálculos hechos por la computadora

x_1	x_2
-1	1

efectivamente, el resultado es correcto. Ahora resolvamos el siguiente sistema

$$2x_1 - 4x_2 = 3$$
$$-x_1 - x_2 = 0$$

según la computadora $x_1 = 0.5$ y $x_2 = -0.5$, al sustituirlos en el sistema comprobamos que son la solución buscada.

1.2 La primera dificultad

Seguros de que nuestro algoritmo funciona correctamente, apliquémoslo nuevamente al siguiente sistema

$$\begin{aligned} 10^{-n}x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned} \tag{1}$$

para diversos valores enteros positivos de n, una observación antes de ver los resultados al aplicar el algoritmo es que si n es grande entonces x_1 y x_2 son cercanos a 1. La siguiente tabla muestra los valores calculados por la computadora para x_1 y x_2:

n	x_1	x_2
1	1.1111111111111100	0.8888888888888980
5	1.0000100000962000	0.9999899998999990
10	1.0000000827403700	0.9999999999000000
11	1.0000000827403700	0.9999999999900000
12	0.9999778782798780	0.9999999999990000
13	1.0003109451872700	0.9999999999999000
14	0.9992007221626410	0.9999999999999900
15	0.9992007221626410	0.9999999999999990
16	2.2204460492503100	1.0000000000000000
17	0.0000000000000000	1.0000000000000000
18	0.0000000000000000	1.0000000000000000

Según la tabla anterior, para $n \geq 17$ tenemos que $x_1 = 0$ y $x_2 = 1$; basta con sustituir estos valores en la segunda ecuación del sistema para concluir que dichos valores no son solución ¿qué sucedió? ¿qué falló?

Si observamos la tabla, es posible detectar sucesos excepcionales, tal vez indicios:

1. x_1 oscila alrededor del valor 1, desde n igual a 12 hasta 15.
2. $x_2 < 1$ hasta $n = 15$, para los valores subsecuentes de n, $x_2 = 1$.

Una pregunta pertinente podría ser ¿el comportamiento qué estamos observando de x_2 y x_1 es real? veamos, de acuerdo con los fórmulas de m, x_1 y x_2 para el sistema (1) tenemos que

$$m = 10^n, \quad x_2 = \frac{2 - 10^n}{1 - 10^n}, \quad x_1 = \frac{1 - x_2}{10^{-n}}$$

Claramente para $n \geq 1$ la variable $x_2 < 1$, ya que $2 - 10^n > 1 - 10^n$ por lo tanto $\frac{2-10^n}{1-10^n} < 1$. En consecuencia $x_1 > 1$, ya que $1 - x_2 > 10^{-n}$ por lo tanto $\frac{1-x_2}{10^{-n}} > 1$.

Lo anterior muestra que tanto la oscilación como los supuestos valores de cero de x_1 a partir de $n \geq 17$, son teoricamente inexistentes. Este fenómeno se debe principalmente a la aritmética de la computadora. En este sentido, al ir creciendo n entonces m cuyo valor en este caso es 10^n es claramente cada vez más y más grande, de forma tal que al sumarle o restarle una cantidad pequeña como 1 ó 2, en algún momento para la computadora deja de tener efecto, generando pérdida de información, haciendo que $x_2 = 1$ y en consecuencia que $x_1 = 0$.

Como hemos visto, resolver en la computadora un sistema de ecuaciones lineales 2×2 no es cosa fácil ¡imaginemos sistemas más grandes!

Ahora, permutemos las ecuaciones del sistema (1), que como sabemos nos genera un sistema equivalente al original, es decir, tienen la misma solución

$$\begin{aligned} x_1 + x_2 &= 2 \\ 10^{-n} x_1 + x_2 &= 1 \end{aligned} \tag{2}$$

y aplicamos nuestro pequeño algoritmo para algunos valores de n, como lo muestra la siguiente tabla:

n	x_1	x_2
1	1.1111111111111100	0.8888888888888890
5	1.0000100001000000	0.9999899998999990
10	1.0000000001000000	0.9999999999000000
12	1.0000000000010000	0.9999999999990000
14	1.0000000000000100	0.9999999999999900
15	1.0000000000000000	0.9999999999999990
16	1.0000000000000000	1.0000000000000000
17	1.0000000000000000	1.0000000000000000
18	1.0000000000000000	1.0000000000000000

De acuerdo con la tabla, para $n \geq 16$, $x_1 = 1$ y $x_2 = 1$. Esta aproximación a la solución del sistema es adecuada, ya que:

$$(1) + (1) = 2$$
$$10^{-n}(1) + (1) = 10^{-n} + 1 \approx 1$$

De hecho, entre más grande sea n estos valores terminan por ser la solución del sistema. Y como el sistema (2) es equivalente al (1) entonces ($x_1 = 1, x_2 = 1$) es la solución aproximada del sistema (1).

Seguramente el lector atraviesa por cierta confusión, en el siguiente apartado se aclarará el panorama.

2 Aritmética Computacional

Los sucesos anteriores están intrínsecamente relacionados con la aritmética de la computadora. Con el propósito de aclarar esta afirmación comencemos analizando el caso del sistema (1).

2.1 Explicación a la primera dificultad

Fijemos nuestra atención en x_2 cuando $n = 16$, en aritmética exacta $x_2 = \frac{2-10^{16}}{1-10^{16}}$ es claramente menor y cercano a 1, además este cociente da como resultado un decimal periódico puro, sin embargo para la computadora el cociente anterior es 1 ¿por qué sucede esto? Veamos, si calculamos exactamente el dividendo y el divisor del cociente que representa a x_2 obtenemos

$$2 - 10^{16} = -\underbrace{99\cdots998}_{16}$$

$$1 - 10^{16} = -\underbrace{99\cdots999}_{16}$$

por lo tanto $x_2 = \overline{.9999999999999998}$. Al realizar estos mismos cálculos en la computadora[1] resulta, como ya se ha visto que

$$x_2 = \frac{2 - 10^{16}}{1 - 10^{16}} = 1$$

[1]Bajo una aritmética con una representación a 32 bits

y esto pasa por lo siguiente: la aritmética de la computadora es finita y discreta. Finita porque los números a utilizar se encuentran dentro de un intervalo donde los extremos son cantidades humanamente manipulables.

Discreta porque los números se conforman por una determinada cantidad de dígitos, comunmente se habla de dígitos de precisión o dígitos significativos. Si los números rebasan la precisión establecida entonces se aplica la siguiente acción: si el dígito siguiente al último dígito significativo es mayor o igual que 5 entonces se suma un 1 al último dígito significativo, y son convertidos en cero todos los dígitos que se encuentran después del último dígito significativo. En este sentido, si aplicamos lo anterior a

$$x_2 = .\overline{9999999999999998}$$

con una precisión de 15 dígitos, tenemos que el dígito que se encuentra en la posición decimosexta es 8 por lo tanto sumamos 1 al 9 de la decimoquinta posición, es decir

$$.9999999999999990 + .0000000000000010 = 1$$

y por esta razón para la computadora $x_2 = 1$. Lo que se ha hecho es lo que usualmente se llama *redondeo*[1].

Ahora bien, la ligera explicación anterior sobre la aritmética computacional ha sido sólo el entremés para explicar el verdadero problema al resolver el sistema (1). El valor de x_1 se calcula por medio de la fórmula

$$\frac{1 - x_2}{10^{-n}}$$

en aritmética exacta $x_2 < 1$ o bien $1 - x_2 > 0$ entonces $x_1 > 0$ para cualquier $n > 0$; no obstante, como $x_2 \approx 1$ para valores de n grandes y debido a los dígitos de precisión de la arítmetica computacional, para algún n la resta $1 - x_2 = 0$. Por ejemplo, para $n = 17$ en aritmética exacta

$$x_2 = .\underbrace{\overline{999999...998}}_{17}$$

y $1 - x_2 = .\overline{000000...001}$ que bajo una aritmética a 15 dígitos significativos tenemos que $1 - x_2 = 0$, cuya consecuencia catastrófica resulta en que $x_1 = 0$, situación que sucede en la resolución del sistema (1). Este fenómeno se denomina *cancelación catastrófica*[2].

Queda demostrada la sutil diferencia entre la aritmética exacta y la aritmética computacional, con certeza reafirmamos que resolver un sistema de ecuaciones lineales en la computadora no es fácil.

Con estos hechos, es natural buscar una forma conveniente de evitarlos y garantizar una buena aproximación a la solución de un sistema de ecuaciones lineales utilizando la computadora; la respuesta la encontraremos en el análisis del sistema (2). Para este sistema

$$x_2 = \frac{1 - 2(10^{-n})}{1 - 10^{-n}}$$

y cuyo comportamiento para valores de n es el mismo que en el caso del sistema (1), lo cual no debe asombrarnos por ser sistemas equivalentes. En el caso de x_1 tenemos que

$$x_1 = 2 - x_2$$

en aritmética computacional $x_2 \leq 1$ ($x_2 = 1$ cuando $n = 16$) entonces $x_1 \geq 1$, que corresponde a su comportamiento real, y no se observa por ninguna parte la cancelación catastrófica. De esto surge una primera conclusión: **la operación elemental de permutar ecuaciones en un sistema lineal puede modificar sustancialmente la solución bajo una aritmética computacional**.

Una pregunta que se antoja natural es ¿Cuándo permutar ecuaciones con el propósito de obtener una buena aproximación a la solución de un sistema de ecuaciones lineales? De los sitemas (1) y (2), claramente el segundo conlleva al mejor resultado, por ello conviene observar el comportamiento del multiplicador m y las incógnitas para ambos sistemas de ecuerdo con los cálculos computacionales realizados

Sistema(1)	Sistema(2)
$m > 1$	$m < 1$
$x_2 \leq 1$	$x_2 \leq 1$
$0 \leq x_2 < 3$	$x_1 \geq 1$

De acuerdo con la tabla, cuando $m > 1$ los resultados de la resolución fueron catastróficos; por otra parte, cuando $m < 1$ se obtuvo un buen resultado, es decir, si el multiplicador es menor que 1, hay mayor posibilidad de éxito. Bajo esta idea, $m < 1$ implica que $a_{21} < a_{11}$, donde a_{11} es el pivote, entonces ¡el pivote debe ser el mayor de los coeficientes correspondientes a la misma columna! En conclusión obtendriamos el siguiente criterio:**permutar**

las ecuaciones de un sistema de ecuaciones lineales con el fin de obtener como pivote al mayor de los coeficientes asociados a la variable que se quiere eliminar.

En el subsecuente apartado enunciaremos un método adecuado para resolver computacionalmente un sistema de ecuaciones lineales bajo lo hechos anteriores.

3 Modificando el Método de resolución

El método de Gauss es el algoritmo tradicionalmente usado para resolver un sistema de ecuaciones lineales. La idea central se basa en la aplicación de las operaciones elementales entre ecuaciones lineales con el fin de transformar el sistema original en un sistema equivalente facilmente de resolver. De las tres operaciones elementales, la más utilizada es la de sumar un múltiplo de una ecuación a otra ecuación con el fin de eliminar incógnitas, desde luego, la eliminación tiene un orden, por ejemplo, para un sistema de $n \times n$ se inicia por sustraer múltiplos de la primera ecuación a las $n-1$ ecuaciones restantes a fin de eliminar de ellas la primera incognita. Desde luego se sustraen múltiplos de la segunda ecuación a las $n-2$ ecuaciones restantes a fin de eliminar de ellas la segunda incógnita; este proceso se repite hasta obtener una ecuación con una incógnita.

Las ecuaciones de las que se van obteniendo múltiplos usualmente se les dice ecuaciones pivote, y se llama pivote al coeficiente de la incógnita que será eliminada de las ecuaciones restantes. La resolución de un sistema $n \times n$ implica el cálculo de $\frac{n^2-n}{2}$ multiplicadores de la forma $\frac{a}{b}$ con $b \neq 0$, el pivote en turno.

El resumen anterior del también llamado método de eliminación muestra una única restricción para que un coeficiente sea pivote, ser distinto de cero. Como ya vimos en el apartado anterior, bajo una aritmética a k dígitos significativos, no es suficiente con elegir un coeficiente distinto de cero como pivote, también se necesita que el pivote sea el mayor en valor absoluto de los coeficientes candidatos. Por ejemplo, en el sistema

$$\begin{aligned} 30^{-11}x_1 + x_2 &= 1 \\ -x_1 + x_2 &= 2 \end{aligned} \tag{3}$$

tenemos dos cantidatos para ser pivotes 30^{-11} y -1, como $|-1| > |30^{-11}|$ entonces elejimos como pivote al coeficiente cuyo valor es -1. Debido a que el

pivote elegido es un coeficiente de la segunda ecuación, se necesita permutar las ecuaciones para llevar a cabo la eliminación como se establece en el método de Gauss. Con esto último el sistema (3) es transformado en el siguiente sistema equivalente

$$\begin{aligned} -x_1 + x_2 &= 2 \\ 30^{-11}x_1 + x_2 &= 1 \end{aligned}$$

procedemos a la eliminación de x_1 de la segunda ecuación, tal como se hizo en la introducción para encontrar las fórmulas del multiplicador y de las incógnitas. El lector puede verificar que la solución aproximada del sistema es $x_1 = -1$ y $x_2 = 1$; también se recomienda resolver el sistema (3) sin permutar las ecuaciones.

Recapitulando, la modificación del método de eliminación tiene como propósito obtener una buena aproximación computacional de la solución de un sistema de ecuaciones lineales. De los fenómenos analizados se concluye es conveniente agregar dos indicaciones en el proceso de resolución:

1. Elegir de entre los coeficientes correspondientes, el mayor en valor absoluto para ser el pivote.

2. Si la ecuación pivote no se encuentra en la posición que le corresponde, aplicar la permutación conveniente.

A este algoritmo se le llama comunmente eliminación gaussina con *Pivoteo Parcial*[3]. Una vez resuelta la situación anterior tiene sentido preguntarse ¿cómo asegurarnos que la solución obtenida computacionalmente es la correcta? Hay distintas maneras de medir la bondad de la solución aproximada, la más natural se encuentra al sustituir la solución aproximada en el sistema de ecuaciones lineales para determinar qué tan satisfactoria es, de esto y otras cosas hablaremos con detalle en el siguiente apartado.

4 Residuo y Error

Como hemos visto, al resolver un sistema lineal en la computadora obtenemos aproximaciones a la solución del sistema. En teoría una solución debe satisfacer el sistema, pero en el caso de una solución computacional bastaría con un "casi" satisfacerlo. Por ejemplo, supongamos que dos personas encuentran por distintos métodos las soluciones

$$\hat{x} = \begin{bmatrix} 2.015 \\ -.592 \end{bmatrix}; \quad \bar{x} = \begin{bmatrix} 1.998 \\ -1.314 \end{bmatrix}$$

como aproximación a la solución del sistema

$$\begin{aligned} .334x_1 - 1.091x_2 &= 1.759 \\ 3.001x_1 + .998x_2 &= 5.004 \end{aligned} \tag{4}$$

¿Cómo saber cuál de las dos aproximaciones propuestas es la mejor? Una primera idea es sustituirlas en el sistema. Al hacerlo con $\hat{x}$ obtenemos

$$\begin{aligned} .334(2.015) - 1.091(-.592) &= 1.318882 \\ 3.001(2.015) + .998(-.592) &= 5.456199 \end{aligned}$$

denotamos por

$$\hat{b} = \begin{bmatrix} 1.318882 \\ 5.456199 \end{bmatrix}$$

a simple vista se puede observar la diferencia entre $\hat{b}$ y b ¿cuánta? hagamos la resta

$$\hat{b} - b = \begin{bmatrix} 1.318882 \\ 5.456199 \end{bmatrix} - \begin{bmatrix} 1.759 \\ 5.004 \end{bmatrix} = \begin{bmatrix} -.440118 \\ .452199 \end{bmatrix}$$

con el mismo juego para $\bar{x}$ tenemos que

$$\bar{b} = \begin{bmatrix} 2.100906 \\ 4.684626 \end{bmatrix}$$

y la diferencia entre $\bar{b}$ y b es

$$\bar{b} - b = \begin{bmatrix} .341906 \\ -.319374 \end{bmatrix}$$

ante estos resultados, la solución que "mejor" satisface al sistema (4) es $\bar{x}$ porque sus elementos están más cercanos a cero. A las diferencias entre $\hat{b}$ y $\bar{b}$ con respecto a b se les llama *vectores residuales*. En general, si x^* es la solución computacional, el vector residual es

$$r = Ax^* - b$$

donde A es la matriz de coeficientes del sistema de ecuaciones lineales y b el vector columna de términos independientes. En el caso de un sistema 2×2, si el vector residual es $\approx (0, 0)$ entonces la solución computacional es una buena aproximación a la solución exacta del sistema; y ya que hablamos de cercanía

entre el vector residual y el vector cero, es posible calcular su proximidad en términos de la distancia entre ellos, regularmente llamada *residuo*

$$\|Ax^* - b\| = \|r\| = \sqrt{{r_1}^2 + {r_2}^2}$$

con $r = (r_1, r_2)$. Por ejemplo, el residuo para $\tilde{x}$ es .6310 y para $\bar{x}$ es .4679, bajo la comparación de estos datos, la conclusión es la misma que se obtuvo con el vector residual, $\bar{x}$ es la mejor solución.

Alternativamente podemos comparar una solución computacional respesto a la solución exacta; claramente es necesario conocer la solución exacta para llevar a cabo la comparación, información que en realidad se desconoce. Sin embargo vale la pena ahondar en el tema, por ello, supongamos que se conoce la solución exacta de un sistema lineal y una solución computacional, el *vector error* es la diferencia entre ambas. Por ejemplo, la solución exacta del sistema (4) es

$$x = \begin{bmatrix} 2 \\ -1 \end{bmatrix}$$

y el vector error para $\hat{x}$ es

$$\hat{e} = x - \hat{x} = \begin{bmatrix} -.015 \\ -.408 \end{bmatrix}$$

para la pimera entrada de $\hat{x}$ el error respecto a la primera entrada de x es $< 10^{-1}$ y para la segunda es < 1 ¿$\hat{x}$ es buena aproximación? En el caso de $\bar{x}$

$$\bar{e} = x - \bar{x} = \begin{bmatrix} .002 \\ .314 \end{bmatrix}$$

donde el error para la primera entrada de $\bar{x}$ con respecto a x es $< 10^{-2}$ y para la segunda es < 1. De acuerdo con los datos del vector error, $\bar{x}$ es la mejor aproximación a la solución que $\hat{x}$, esta misma conclusión la obtuvimos con el residuo. De igual forma se puede calcular la distancia entre una solución computacional y la solución exacta, una vez calculado el vector error $e = (e_1, e_2)$, el *error* es

$$\|e\| = \sqrt{{e_1}^2 + {e_2}^2}$$

Dejamos al lector calcular el error para las soluciones aproximadas $\hat{x}$ y $\bar{x}$ respecto a x. Ahora consideremos el siguiente sistema

$$\begin{aligned} 1.2969x_1 + .8648x_2 &= .8642 \\ .2161x_1 + .1441x_2 &= .1440 \end{aligned} \tag{5}$$

cuya solución exacta es $x_1 = 2$ y $x_2 = -2$. Resolviendo el sistema en la computadora con el algoritmo de eliminación gaussiana, la solución aproximada que obtenemos es

$$x_1 = 2.00000000359962$$
$$x_2 = -2.00000000240030$$

y el residuo es $< 10^{-14}$; no hay duda, es una buena aproximación.

A manera de experimento modifiquemos muy poquito el lado derecho del sistema (5), por ejemplo, cambiemos b_2 por .1441 para obtener el siguiente sistema

$$\begin{aligned} 1.2969x_1 + .8648x_2 &= .8642 \\ .2161x_1 + .1441x_2 &= .1441 \end{aligned} \tag{6}$$

Sugerimos al lector tomarse un minuto para pensar ¿cómo debe ser la solución de este sistema?

Una vez más resolvemos este sistema con la computadora ¿ya tienes el resultado? Será este

$$x_1 = -8646.00001344409$$
$$x_2 = 12967.0000201615$$

¿es el que esperabas? Quizá se piense increible el resultado, sobre todo si creías que la solución sería aproximadamente (2, -2), lo cual parecería razonable a consecuencia de adicionar a b_2 del sistema (5) una cantidad cercana a cero como 10^{-4}. Conviene calcular el residuo para dar certeza al resultado anterior, este es $< 10^{-10}$ ¡pequeño!

Hagamos otro movimiento, ahora cambiemos b_2 por .1439

$$\begin{aligned} 1.2969x_1 + .8648x_2 &= .8642 \\ .2161x_1 + .1441x_2 &= .1439 \end{aligned} \tag{7}$$

la solución computacional de este sistema es

$$x_1 = 8650.00001344649$$
$$x_2 = -12971.0000201651$$

el lector puede verificar que el residuo es $< 10^{-10}$ ¿qué está pasando?

Las soluciones computacionales de los sistemas (5), (6) y (7) son totalmente diferentes y los cambiós en b_2 son pequeños como lo muestra resumidamente la siguiente tabla

b_2	x_1	x_2
.1440	2.00000000359962	-2.00000000240030
.1441	-8646.00001344409	12967.0000201615
.1439	8650.00001344649	-12971.0000201651

¿Porqué sucede esto? ¿Qué tienen de especial los sistemas (5), (6) y (7)? Evidentemente hay una explicación para este fenómeno, de esto trata el siguiente apartado.

5 Mal Condicionamiento

Imaginen se recopila información de cierto fenómeno, el modelo matemático que resulta es un sistema de ecuaciones lineales como el siguiente

$$\begin{aligned} .010x_1 + .009x_2 &= .019 \\ .012x_1 + .011x_2 &= .023 \end{aligned} \tag{8}$$

Después de un tiempo, nuevamente se recopila información sobre el mismo fenómeno, originando este sistema

$$\begin{aligned} .010x_1 + .009x_2 &= .019 \\ .012x_1 + .011x_2 &= .024 \end{aligned} \tag{9}$$

la única diferencia visible se encuentra en el lado derecho de los sistemas, particulamente en el término independiente asociado a b_2, y hay que notar que ésta es pequeña. La lógica nos sugiere resolver uno de los sistemas, tomar su solución como solución del otro bajo el argumento que cambios pequeños en los valores del lado derecho de un sistema le corresponden cambios pequeños en la solución. Sin embargo, en la última parte del apartado anterior nos percatamos que esto puede resultar falso y por lo tanto un falso argumento, ya que un cambio pequeño en el lado derecho de un sistema puede generar soluciones muy diferentes, a este tipo de sistemas les llamaremos *Mal Condicionados*.

Para el caso de los sistemas lineales anteriores, la solución exacta del (8) es $(1, 1)$ y la del (9) es $(-3.5, 6)$, evidentemente son bastante diferentes respecto a la pequea diferencia entre los sistemas. Dada la existencia de sistemas de ecuaciones lienales inestables, nos planteamos dos cuestiones que

consideramos importantes para comenzar con el análisis de este tipo de sistemas: porqué tal comportamiento y cómo saber, sin que quepa la duda, si un sistema está mal condicionado.

Contestaremos las preguntas en el orden en que han sido escritas. Veamos las gráficas de los sistemas (8) y (9) para encontrar información sobre su comportamiento

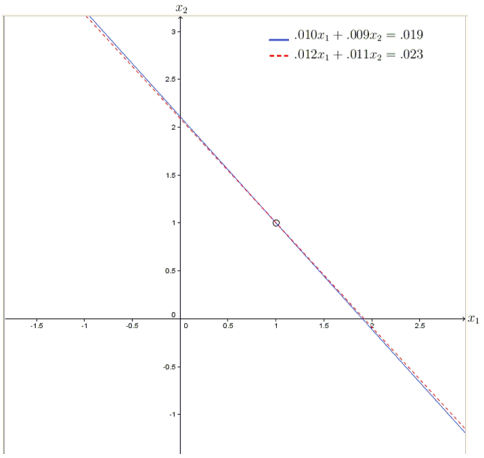

Figura 1: Sistema (8)

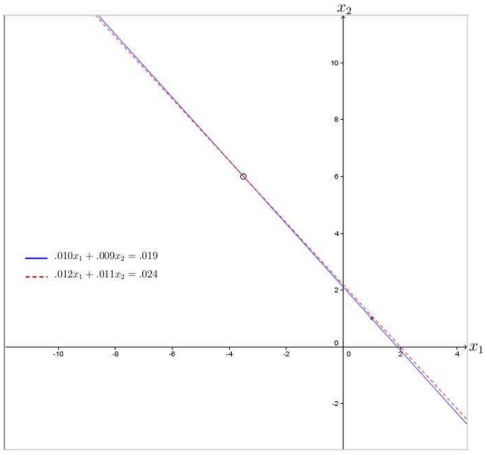

Figura 2: Sistema (9)

A primera vista, la gráficas muestran una característica peculiar e impor-

tante de los sistemas: las rectas casi se empalman o están muy pegadas, es decir, tienen casi la misma inclinación o pendiente.

Lo anterior explica por que pequeños cambios en el lado derecho del sistema provocan cambios sustanciales en la solución, ya que cambiar el lado derecho del sistema equivale graficamente a subir o bajar las rectas, desde luego las pendientes no cambian; ese desplazamiento mueve el punto de intersección inicial, como sucede entre los sistemas (8) y (9). Además, en cuanto más pegadas estén las rectas o se parezcan sus pendientes, la intersección tendrá saltos considerables, a veces abismales como en el caso de los sistemas (5), (6) y (7); dejamos al lector que grafique los mencionados sistemas y verifique que son indistingibles las rectas.

Para continuar consideramos conveniente utilizar la notación matricial de un sistema de ecuaciones lineales

$$\begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

denotamos por A la matriz de coeficientes, por $\mathbf{x}$ el vector columna de incógnitas y por $\mathbf{b}$ el vector columna de términos independientes. Recordemos que un sistema tiene solución única si su matriz de coeficientes tiene determinate distinto de cero

$$\det A = a_{11}a_{22} - a_{12}a_{21} \neq 0$$

una matriz con determinante igual a cero se dice *singular*.

Regresando a lo gráficamente observado, la característica del mal condicionamiento de un sistema 2×2 está relacionada con la casi misma inclinación de las rectas; analíticamente, las pendientes de las rectas son casi iguales, digamos que m_1 y m_2 son las pendientes de dichas rectas, entonces

$$m_1 \approx m_2$$

Si escribimos en la forma simplificada las ecuaciones de las rectas que representa el sistema obtendríamos que $\frac{a_{11}}{a_{12}} = m_1$ y $\frac{a_{21}}{a_{22}} = m_2$, y como las pendientes son casi iguales entonces

$$\frac{a_{11}}{a_{12}} \approx \frac{a_{21}}{a_{22}}$$

y con un poco de álgebra nos percatamos que

$$a_{11}a_{22} - a_{12}a_{21} \approx 0$$

es decir, el determinate de la matriz de coeficientes es casi cero o la matriz es casi singular. Esto parece ser una buena forma de saber cuando un sistema de ecuaciones lienales está mal condicionado, veamos si es así. El sistema (8) en su forma matricial es

$$\begin{pmatrix} .010 & .009 \\ .012 & .011 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} .019 \\ .023 \end{pmatrix}$$

el que sabemos está mal condicionado, el determinante de su matriz de coeficientes es

$$.000002$$

por lo tanto es casi singular; aparentemente el determinante de la matriz de coeficientes es un buen estimador del mal condicionamiento de un sistema de ecuaciones lienales. Hagamos una prueba más, determinemos si el siguiente sistema está mal condicionado

$$\begin{aligned} 10000x_1 + 9000x_2 &= 19000 \\ .012x_1 + .011x_2 &= .023 \end{aligned} \tag{10}$$

lo que es fácil porque basta con calcular el determinate de

$$\begin{pmatrix} 10000 & 9000 \\ .012 & .011 \end{pmatrix}$$

que resulta ser 2 ¡dista de ser cero! Podemos concluir que el sistema (10) no está mal condicionado ¿estamos seguros? Desconfiando un poco del resultado, gráficamos nuestro sistema para liberarnos de la duda

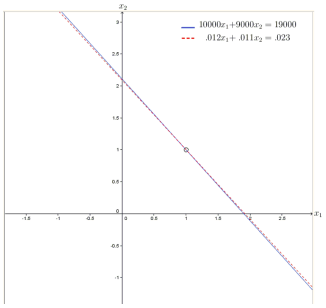

Figura 3: Sistema (10)

¿Se parecen las gráficas de la figura 2 y 3? Sin duda, no hay diferencia entre ellas, de hecho son iguales. La figura 3 muestra ineludiblemente que el sistema (10) está mal condicionado, esto contradice la conclusión obtenida del determinante de su matriz y que evidentemente necesitamos aclarar. El sistema (10) es equivalente al sistema (8), básicamente se multiplicó la primera ecuación de (8) por 1000000, por lo tanto el determinante de la matriz de coeficientes del sistema (10) es un múltiplo de la matriz de coeficientes del sistema (8)

$$\det \begin{pmatrix} 10000 & 9000 \\ .012 & .011 \end{pmatrix} = (1000000) \det \begin{pmatrix} .010 & .009 \\ .012 & .011 \end{pmatrix} = (1000000)(.000002) = 2$$

esto revela que el determinante no es buen estimador del mal condicionamiento de un sistema de ecuaciones lineales, ya que basta con multiplicar por un escalar cualquiera de las ecuaciones para modificar el determinate pero el sistema sigue siendo mal condicionado.

Estamos casi como iniciamos, sin un estimador del mal condicionamiento, sin embargo nos hace falta explorar otros caminos. Para empezar notemos que si hay mal condicionamiento

$$\begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} \approx k \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}$$

es decir, una columna de la matriz de coeficientes del sistema es casi un múltiplo de la otra o son casi depende una columna de la otra. Gráficamente tenemos un par de vectores bastante juntos cuando el sistema está mal condicionado como el caso del sistema (8), cuya gráfica de sus vecores columna es

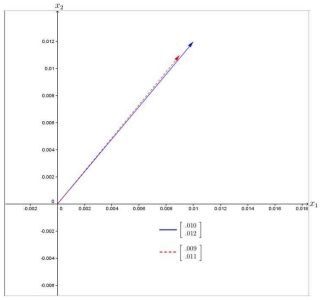

Figura 4: Vectores Columna de Matriz de Coeficientes del Sistema (8)

Como se puede notar, las gráficas de las figuras 1 y 4 comparten una configuración peculiar. Las rectas como los vectores en las respectivas figuras son casi paralelos, indudablemente es la matriz de coeficientes quien determina dicha configuración gráfica, y por lo tanto la que contiene información trascendental del sistema de ecuaciones lineales. En la siguiente sección buscaremos la forma de aprehender a la matriz de coeficientes, interrogarle hasta hacerle revelar los secretos necesarios para establecer un estimador adecuado para medir el mal condicionamiento.

6 Índice de Sensibilidad

Las gráficas han mostrado que la sensibilidad o mal condicionamiento de un sistema 2×2 está fuertemente asociado al casi paralelismo, ya sea de las rectas o de los vectores columna de la matriz de coeficientes; también descubrimos que los vectores columna de la matriz de coeficientes son casi múltiplos pero caminamos en una dirección equivocada. Lo que nos lleva a formular otra pregunta ¿qué tratamiento dar a esta información para medir la sensibilidad?

6.1 Obertura para medir la sensibilidad de un sistema

Multiplicar una matriz y un vector, siempre que sea posible, genera un nuevo vector. Si la matriz A es 2×2 y x es un vector columna en $\mathbb{R}^2$, entonces el producto Ax es un vector columna en $\mathbb{R}^2$. Por ejemplo, dado un vector en el plano, digamos

$$e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

y una matriz A

$$\begin{pmatrix} 1 & 4 \\ 2 & 1 \end{pmatrix}$$

al muliplicarlos se suscitaría claramente un cambio. La matriz A transformaría el vector e_1 en el vector

$$\begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

la siguiente figura muestra la transformación

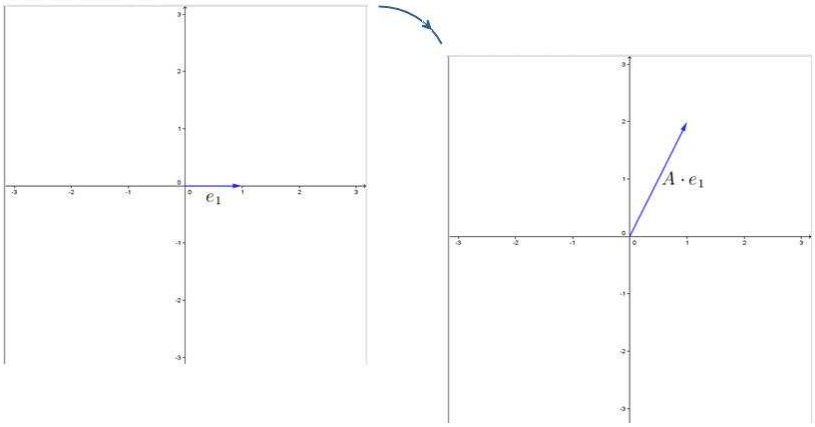

Figura 5: Transformación

Esto mismo sucede con el vector

$$e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

donde $A \cdot e_2$ es

$$\begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

Seguramente ya se alcanzó a observar que $A \cdot e_1$ es la primera columna de A y $A \cdot e_2$ es la segunda columna de A; a partir de este punto, denotamos por C_1 y C_2 las columnas de A. En la idea de encontrar una configuración conveniente para extraer información relevante, notemos que los vectores e_1 y e_2 son perpendiculares, al multiplicarles A se transforman en C_1 y C_2, en el caso de que el sistema este mal condicionado, los vectores serían casi paralelos.

Ahora pensemos que en el plano tenemos un cuadrado cuyos vertices son $(1,1)$, $(-1,1)$, $(-1,-1)$ y $(1,-1)$ o bien e_1+e_2, $-e_1+e_2$, $-e_1-e_2$ y e_1-e_2, respectivamente. En qué figura se transforma dicho cuadrado al multiplicarle cualquier matriz a sus puntos?

Supongamos que la matriz es

$$\begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$$

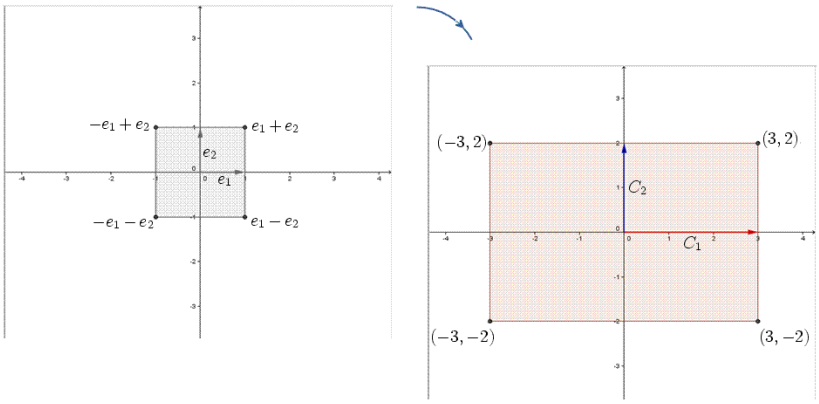

Figura 6: Transformación de un cuadrado

entonces la transformación del cuadrado sería un rectángulo con vértices $(3,2)$, $(-3,2)$, $(-3,-2)$ y $(3,-2)$

Evidentemente para este caso $C_1 = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$ y $C_2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$, si esta matriz fuera la de coeficientes de un sistema de ecuaciones lineales, podriamos hablar de buen condicionamiento del sistema lineal cómo estimarlo sin recurrir a la graficación?

Hay que observar una vez más la figura 6 ¿qué relación tienen las dimensiones del rectangulo con el problema? ¿A caso la proporción entre el largo y el ancho del rectángulo nos puede ser útil? Antes de contestar este par de preguntas conviene mostrar de forma un tanto sencilla que un cuadrado unitario se convierte en un paralelogramo.

Como ya vimos, los vertices del cuadrado unitario son $e_1 + e_2$, $-e_1 + e_2$, $-e_1 - e_2$ y $e_1 - e_2$, dada cualquier matriz A con columnas C_1 y C_2, si multiplicamos A a los vértices del cuadrado tenemos que

$$A(e_1 + e_2) = A \cdot e_1 + A \cdot e_2 = C_1 + C_2$$
$$A(-e_1 + e_2) = -A \cdot e_1 + A \cdot e_2 = -C_1 + C_2$$
$$A(-e_1 - e_2) = -A \cdot e_1 - A \cdot e_2 = -C_1 - C_2$$
$$A(e_1 - e_2) = A \cdot e_1 - A \cdot e_2 = C_1 - C_2$$

donde $C_1 + C_2$, $-C_1 + C_2$, $-C_1 - C_2$ y $C_1 - C_2$, son los vértices de un paralelogramo. El ancho y largo están estrachamente relacionados con las columnas de la matriz A.

6.2 Estimando la sensibilidad

Retomando la tranformación del cuadrado unitario por la aplicación de la matriz

$$\begin{pmatrix} 3 & 0 \\ 0 & 2 \end{pmatrix}$$

que resultó en un rectángulo. Un sistema de ecuaciones lineales asociado a dicha matriz no estaría de ninguna manera mal condicionado; el rectángulo de la transformación está basado en la posición de los vectores columna de la matriz, de hecho su ancho es

$$\|2C_2\| = 2(\sqrt{0^2 + 2^2}) = 4$$

y su largo es

$$\|2C_1\| = 2(\sqrt{3^2 + 0^2}) = 6$$

Con esta infomación ¿es posible medir la sensibilidad? Dado que un sistema de ecuaciones lineales es más sensible en cuanto el paralelogramo que le representa tenga una anchura cada vez menor. Entonces una manera que parece adecuada para medir la sensibilidad es la razón entre el largo y el ancho del paralelogramo; digamos que el índice de sensibilidad de un sistema de ecuaciones lineales es

$$IS = \frac{Largo \quad del \quad Paralelogramo \quad de \quad A}{Ancho \quad de \quad paralelogramo \quad de \quad A}$$

Hay que notar, que el mínimo valor que puede tomar el IS es 1. Denotemos por $\acute{s}$ al índice de sensibilidad; de este modo, $\acute{s}(A) = \frac{6}{4} = 1.5$.

Ahora pensemos en otra situación posible, que los vectores columna de una matriz 2×2 no sean necesariamente ortogonales ¿cómo calcular $\acute{s}$? Por ejemplo, la matriz

$$B = \begin{pmatrix} 1 & 4 \\ 3 & 3 \end{pmatrix}$$

cuyos vectores columnas representan al paralelogramo de la figura 7 ¿Cuál es el ancho? Resulta ser

$$\|\vec{w} - \vec{u}\|$$

donde $\vec{u}$ es el vector columna con menor longitud y $\vec{w}$ es la proyección ortogonal de $\vec{u}$ sobre el vector columna de mayor longitud. Por lo tanto,

$$\acute{s}(B) = \frac{\|v\|}{\|\vec{w} - \vec{u}\|}$$

con v el vector columna de B con mayor longitud.

De este modo para los vectores columna $\vec{a}_1 = \begin{pmatrix}1\\3\end{pmatrix}$ y $\vec{a}_2 = \begin{pmatrix}4\\3\end{pmatrix}$ tenemos

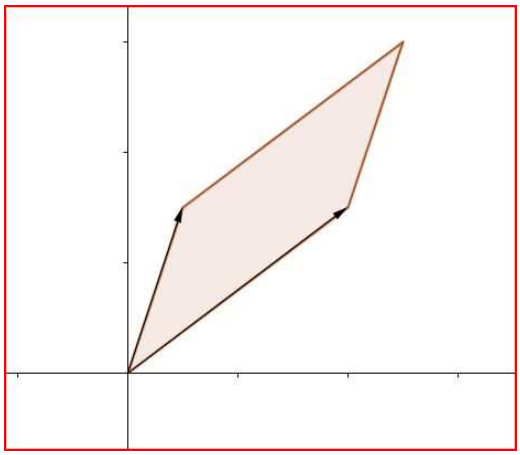

Figura 7: Vectores no ortogonales.

que $\|\vec{a}_1\| = \sqrt{10} \approx 1.778$ y $\|\vec{a}_2\| = \sqrt{25} = 5$ respectivamente. Por lo que necesitamos encontrar la proyección ortogonal de $\vec{a}_1$ sobre $\vec{a}_2$, digamos $\vec{w} = k\vec{a}_2$ con $k = \frac{\vec{a}_1^t \cdot \vec{a}_2}{\|\vec{a}_2\|^2}$; para este caso $k = \frac{13}{25} \approx 0.52$ por lo tanto

$$\vec{w} = \frac{13}{25}\begin{bmatrix}4\\3\end{bmatrix} \approx \begin{bmatrix}2.08\\1.56\end{bmatrix}$$

de aquí que $\|\vec{w} - \vec{a}_1\| = \sqrt{(2.08-1)^2 + (1.56-3)^2} = 1.8$, por lo que concluimos que

$$\acute{s}(B) = \frac{5}{1.8} \approx 2.778$$

es decir, la matriz B no está mal condicionada. Ahora apliquemos esta idea a la matriz de coeficientes del sistema (8), la cual es

$$A = \begin{pmatrix}0.010 & 0.009\\0.012 & 0.011\end{pmatrix}$$

realizando los correspondientes cálculos llegamos a que

$$\acute{s}(A) \approx 122$$

bajo esta estimación, aseguramos que la matriz esta mal condicionada, así tambin se puede observar en su gráfica, a una escala de 0.002 por marca (ver figura 8).

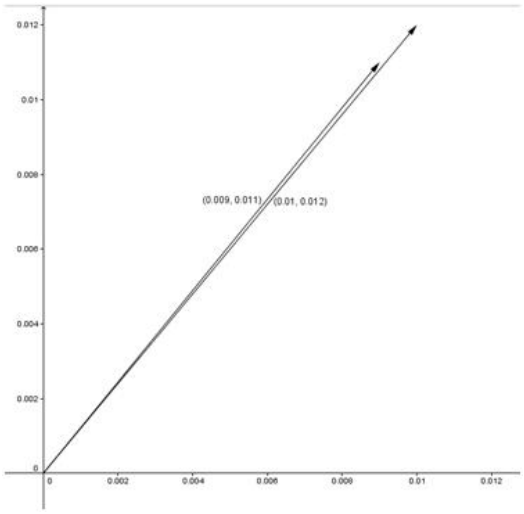

Figura 8: Gráfica de matriz mal condicionada

Esta misma matriz tiene, de acuerdo con Matlab, un número condición igual a 222.9955. Sorprendentemente para la matriz

$$\begin{pmatrix} 1.2969 & 0.8648 \\ 0.2161 & 0.1441 \end{pmatrix}$$

el número condicón es 2.4973×10^8 y el índice de sensibilidad es aproximádamente 1.7286×10^8, en ambos casos concluimos que la matriz esta mal condicionada.

6.3 Índice de sensilidad para culquier matriz cuadrada

El índice de sensibilidad de una matriz A de $n \times n$ se calcula bajo una idea similar a la del apartado anterior. Primero determinamos cuál es el vector columna de menor longitud, es decir, para $\vec{a}_1, \vec{a}_2, ..., \vec{a}_n$ tenemos que $\vec{u} = \{a_i : \|a_i\| = min\{\|a_1\|, ..., \|a_n\|\}\}$. Sean $\vec{c}_1, \vec{c}_2, ..., \vec{c}_{n-1}$ los vectores columna de A con mayor o igual longitud que $\vec{u}$ y determinamos las $n-1$ proyecciones ortogonales, digamos, $P_{u1}, P_{u2}, ..., P_{u(n-1)}$ entonces el índice de sensibilidad de la matriz A es

$$\acute{s}(A) = max\{\frac{\|a_i\|}{\|P_{ui} - \vec{u}\|}\}$$

Bibliografía

[1] Cleve B. Moler, Numerical Computing with Matlab, SIAM(2004).

[2] David Kahaner, Cleve Moler y Stephen Nash, Numerical Methods and Software, Prentice-Hall(1989).

[3] John R. Rice, Matrix Computations Mathematical Software, International Student Edition(1985).

[4] Stanley I. Grossman, Álgebra Lineal, Grupo Idetorial Iberoamérica(1987).