Guía de Cálculo Numérico Para Ingenieros

Autor: Prof. Lisbeth Torres
ljtorres@usb.ve

Capítulo 1

Introducción

El análisis numérico es el desarrolo y estudio de procedimientos para resolver problemas con la ayuda de una computadora. Quien se dedica al análisis numérico suele interesarse en determinar cuál de varios algoritmos que resuelven un problema es el más eficiente en cierto sentido. La eficiencia puede medirse mediante el número de iteraciones que realiza el algoritmo, el tiempo y la capacidad de memoria que requiere la computadora entre otras.

Una ventaja fundamental del análisis numérico es que puede obtenerse una respuesta numérica, aún cuando un problema no tenga solución analítica.

Por ejemplo, la siguiente integral proporciona la longitud de arco de la curva $y = sen(x)$, la cual no tiene solución cerrada:

$$\int_0^{\pi} \sqrt{1 + cos^2(x)}dx$$

se puede calcular de forma aproximada la la longitud de esta curva con métodos estandar que se aplican a cualquier integrando, sin necesidad de utilizar sustituciónes especiales ni integración por partes. Más aún las únicas operaciones necesarias son la suma, resta, multiplicación, división y compración.

Es importante destacar que la solución obtenida a través del cálculo numérico siempre será numérica. Los métodos analíticos suelen proporcionar el resultado en términos de funciones matemáticas que luego pueden evaluarse para casos específicos. De manera que hay una ventaja en los métodos analíticos, en el sentido de que las propiedades y el comportamiento de la función a menudo son claros, lo cual no ocurre con los resultados numéricos. No obstante los resultados numéricos pueden graficarse par mostrar algo del comportamiento de la solución.

Chapter 2

Errores y Arimética del Computador

2.1. Fuentes de errores y tipos de errores

2.1.1. Errores de datos originales

Los problemas del mundo real donde una situación física existente o propuesta es modelada por una ecuación matemática, casi siempre presenta coecientes conocidos de manera imperfecta. La razón de esto es que los problemas a menudo dependen de mediciones cuya exactitud no es confiable. El modelo en si puede no representar de manera correcta la situación a resolver.

Nada se puede hacer para reparar este tipo de errores.

2.1.2. Errores computacionales

El error computacional puede ser dividido en dos clases error de truncamiento y de redondeo.

2.1.2.1. Error de Truncamiento

Es la diferencia entre el resultado verdadero y el resultado producido por un algoritmo dado usando aritmética exacta. Se debe a aproximaciones tales como el truncamiento de una serie innita, reemplazar una derivada por el valor del cociente de una diferencia nita, reemplazar una función arbitraria por un polinomio.

2.1.2.2. Error de redondeo

Es la diferencia entre el resultado obtenido por un algoritmo dado usando aritmética exacta y el resultado obtenido por el mismo algoritmo usando ar-

itmética finita. Se debe a la representación de números reales y operaciones aritméticas con estas representaciones.

2.1.2.2.1 **Error Absoluto**

Error Absoluto= Valor aproximado- Valor real. Sea x el valor real (solución del problema) y sea $\hat{x}$ el valor aproximado.

$$e_a = |x - \hat{x}|$$

2.1.2.2.2 **Error relativo**

Error Relativo= Error absoluto/ Valor real Sea x el valor real (solución del problema) y sea $\hat{x}$ el valor aproximado.

$$e_r = \frac{e_a}{|x|} = \left|\frac{x - \hat{x}}{x}\right|$$

Definición: Cifras Significativas

Diremos que el número $\hat{p}$ aproxima a p con t cifras significativas si t es el mayor entero no negativo tal que se cumpla la siguiente relación:

$$\left|\frac{p - \hat{p}}{p}\right| < 5 * 10^{-t}$$

2.2. Sistema de punto flotante

Con el fin de analizar en detalle el error por redondeo es necesario comprender la representación numérica en las computadoras. En casi todos los casos los números se almacenan como cantidades de punto flotante que se parece mucho a la notación científica.

El punto flotante consta de 3 partes: el signo (1 bit), la parte fraccionaria o mantisa y la parte exponencial o característica. Las tres partes de los numeros posee una cantidad total fija que suele ser 32 o 64 bits. La mayor parte de estos bits sirve para la parte fraccionaria, quizás desde 24 hasta 52 bits y esta cantidad determina la precisión de la representación, la parte exponencial maneja desde 7 hasta 11 bits, y esta cantidad determina el intervalo de los valores.

La forma general de un número en punto flotante es la siguiente:

$$\pm.d1d2d3...dp * B^e$$

donde las di son dígitos o bits con valores desde 0 hasta B-1

B representa la base (2,10 o 16)

p es el número de dígitos signicativos (la precisión)

e exponente (número entero) pertenece al intervalo [L, U] (rango del exponente).

Si un número real x no puede ser representado como punto otante entonces deberá ser aproximado por el punto flotante más cercano, denotaremos la aproximación del punto flotante de x como $fl(x)$, este proceso de escogencia del punto flotante más cercano que represente a x se llama redondeo.

Se dice que la representación esta normalizada cuando $d_1 > 0$.

Ejemplo del error en punto flotante

Ahora se considera el error en el cáculo que es atribuible a la longitud de la palabra.

Sean $x = 0,31426 * 10^3$, $y = 0,92577 * 10^5$ utilizaremos operaciones de máquina con un sistema de punto flotante que utiliza 5 dígitos decimales de precisión y base 10.

$x * y = 0,2909324802 * 10^8$

$x + y = 0,9289126 * 10^5$

x-$y = 0,92262740 * 10^5$

$x/y = 0,3394579647 * 10^{-2}$

Note que estos resultados fueron calculados con 10 decimales. La computadora con 5 decimales guardará en forma redondeada lo siguiente:

$fl(xy) = 0,29093 * 10^8$

$fl(x + y) = 0,92891 * 10^5$

$fl(x$-$y) = -0,92263 * 10^5$

$fl(x/y) = 0,33946 * 10^{-2}$

Los errores relativos en estos resultados son$8,510^{-6}$, $2,310^{-6}$, $2;810^{-6}$, $6;010^{-6}$.

2.3. Sensitividad o Condicionamiento de un Problema

La dificultades al resolver un problema no siempre se deben a que la fórmula está mal concebida o que el algoritmo utilizado para hallar la solución sea inexacto e ineficiente, estas dificultades pueden ser inherentes al problema mismo.

Un problema es insensitivo o bien condicionado si cambios pequeños en la entrada del problema produce cambios pequeños en la solución del problema y además grandes cambios en la entrada del problema producen grandes cambios en la solución del mismo. En otro caso se dice que el problema está mal condicionado.

El número de condición de un problema de cálcula como el cociente entre el error relativo de la solución y el error relativo en la entrada, es decir:

$$cond = \left| \frac{(f(\hat{x}) - f(x))/f(x)}{(\hat{x} - x)/x} \right|$$

Si el problema está mal condicionado el número de condición será más grande que 1.

Ejemplo:

Aproximar el valor de $cos(x)$ para valores de x cercanos a $\pi/2$.

Sea $x \approx \pi/2$ y sea $x + h$ una perturbación de x. Calculemos el número de condición de este problema.

$$\left| \frac{(cos(x+h) - cos(x))/cos(x)}{(x+h-x)/x} \right| \approx \left| \frac{-xsen(x)}{cos(x)} \right| = |x\, tan(x)|$$

cuyo valor sabemos que tiende a ∞ cuando x esta cerca de $\pi/2$. Por lo tanto el problema esta mal condicionado, de hecho

$cos(1.57079) = 0.63267949 * 10^{-5}$

$cos(1.57078) = 1.63267949 * 10^{-5}$

se hace evidente que un cambio pequeño en la entrada de la función produce un cambio grande en la solución que se obtiene.

2.4. Otros tipos de errores

A pesar de que el error de redondeo es inevitable y dicil de controlar, existen otros tipos de errores que si pueden estar bajo nuestro control.

2.4.1. Pérdida de dígitos significativos

Para aclarar el prolema se usara un ejemplo. sean $x = 0,3721478693$ y $y = 0,3720230572$ asi tenemos que $x - y = 0,00011248121$ (solución exacta) supondremos que estamos operando en una máquina que trabaja con una mantisa de 5 dígitos, asi que la representación de las cantidades x y y serán las siguientes $fl(x) = 0,37215$, $fl(y) = 0,37202$al hacer las cuentas siguiendo las restricciones de nuestra máquina se obtiene que $fl(x - y) =; 00013$.

2.4.2. Cancelación Catastrófica

Como regla se deben evitar situaciones en las que la exactitud de los resultados se vea comprometida por una resta de cantidades similares, a continuación se ilustra esta situación.

$\sqrt{x^2 + 1} - 1$

cuando x tome valores muy pequeños entonces el correspondiente valor de y sufrira los fenómenos de perdida de signicancia y cancelación, valdría la pena verificar si esta situación la podemos evitar reescribiendo la expresión y de la siguiente manera:

$$\frac{x^2}{\sqrt{x^2 + 1} + 1}$$

para realizar esta verificación se invita al estudiante a realizar las respectivas cuentas en MATLAB.

2.5. Epsilon de la máquina

La exactitud de un sistema de punto flotante es caracterizada por una cantidad a la que llamamos unidad de redondeo o epsilón de la máquina. Esta cantidad es importante porque determina el máximo error relativo posible que se comete al representar un número diferente de cero en punto otante. Caracterizaremos la unidad de redondeo (ε) de la siguiente manera: la unidad de

redondeo es aquel número tal que se cumple que $1 + \varepsilon = 1$. A continuación damos el algoritmo (metalenguaje) para determinar el epsilón de una máquina.

Algoritmo para determinar el epsilon de la máquina
$s = 1$
$t = 1$
$iter = 0$
mientras $t + s > t$
$s = s/2$
$iter = iter + 1$

A pesar de que el error de redondeo es inevitable y difícil de controlar, existen otros tipos de errores que si pueden estar bajo nuestro control.

Chapter 3

Solución Numérica de Sistemas de Ecuaciones Lineales

Los sistemas de ecuaciones lineales surgen en casi cualquier aspecto de las matemáticas aplicadas, también pueden ser el resultado de aproximaciones a ecuaciones no lineales o dela aproximación de ecuaciones diferenciales a través de ecuaciones algebráicas, de manera que la eciencia y la exactitud de la solución de sistemas de ecuaciones lineales tiene gran importancia en los métodos numéricos que resuelven una gran variedad de problemas computacionales. La notación matriz-vector de un sistema de ecuaciones lineales tiene la forma

$$Ax = b$$

donde A es una matriz mxn, b es un vector de longitud m y x es el vector de incognitas de longitud n. Decimos que este sistema tiene solución si el vector b puede ser escrito como combinación lineal de las columnas de A. También puede suceder que este sistema no tenga solución o que tenga infinitas soluciones. Por el momento consideraremos sólo el caso en el que el sistema tiene la misma cantidad de ecuaciones que de incognitas, lo que genera que la matriz aociada sea cuadrada es decir A es nxn.

Recordemos algunos de los conceptos básicos del algebra matricial.

Matriz Singular

Una matriz cuadrada A se dice singular si se cumple una de las siguientes propiedades equivalentes:

1. A no posee inversa (no existe B tal que $AB = BA = I$)
2. det(A)=0

3. rang(A)<n

4. $Az = 0$para algún vector $z \neq 0$

En otro caso diremos que la matriz es no-singular.

Si la matriz A es no-singular entonces el sistema $Ax = b$ tendra solución única y viene dada por $x = A^{-1}b$, por otro lado si a es singular entonces la solución del sistema viene dada por el vector b (puede tener infinitas soluciones o no tener solución).

Ejemplo:

Consideremos el siguiente sistema

$$\begin{cases} 2x_1 + 3x_2 &= 8 \\ 5x_1 + 4x_2 &= 13 \end{cases}$$

en este caso la solución del sistema será $x = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$$\begin{cases} 2x_1 + 3x_2 &= b_1 \\ \frac{8}{3}x_1 + 4x_2 &= b_2 \end{cases}$$

ahora consideremos $b = \begin{bmatrix} 4 \\ 7 \end{bmatrix}$, $b = \begin{bmatrix} 4 \\ 8 \end{bmatrix}$

en el primer caso el sistema no tiene solución y en el segundo caso el sistema tiene infinitas soluciones de la forma $x = \begin{bmatrix} x_1 \\ \frac{4-2x_1}{3} \end{bmatrix}$.

Algunos sistemas sencillos de resolver

1. Consideremos que $A = I$
   en este caso la solución del sistema viene dada por $x = b$.

2. Sea $A = D$ (esto es A matriz diagonal)
   En este caso la solución viene dada por $x_i = b_i/a_{ii}$

3. Sea A una matriz triangular superior
   En este caso la solución viene dada de la siguiente manera:
   $x_n = b_n/a_{nn}$
   $x_i = (b_i - \sum_{j=i+1}^{n} x_j a_{ij})/a_{ii}$

4. Sea A una matriz triangular inferior
   En este caso la solución viene dada de la siguiente manera:
   $x_1 = b_1/a_{11}$
   $x_i = (b_i - \sum_{j=1}^{i-1} x_j a_{ij})/a_{ii}$

Para resolver un sistema lineal la estrategia general sugiere que se debe transformar el sistema en uno cuya solución es la misma que del original pero más fácil de calcular. Una manera de hacerlo es premultiplicar el sistema por una matriz no singular M y hacer esto no afecta la solución, de esta forma la solución del sistema $MAx = Mb$ viene dada por: $x = (MA)^{-1}Mb = A^{-1}M^{-1}Mb = A^{-1}b$ que es la misma solución del sistema original.

Definición: Sistemas equivalentes

Dos sistemas de ecuaciones $Ax = b$ y $Bx = d$ se dicen equivalentes si tienen la misma solución.

Por lo tanto esto nos lleva a pensar que para resolver un sistema de ecuaciones lo transformamos a través de ciertas operaciones elementales en un sistema equivalente pero más sencillo.

Las operaciones elementales por fila (o.e.f) son:

1. $f_i \longleftrightarrow f_j$ o.e.f de tipo 1.
2. $f_i \leftarrow \lambda f_i \quad \lambda \in \mathbb{R}$ o.e.f de tipo 2
3. $f_j \leftarrow \lambda f_i + f_j \quad \lambda \in \mathbb{R}$ o.e.f de tipo 3

Teorema: Si un sistema de ecuaciones $Bx = d$ se obtiene a partir de otro $Ax = b$ aplicando una cantidad finita de operaciones elementales por fila entonces estos sistemas son equivalentes.

Demostración:

Aplicar una operación elemental por fila a la matriz A es equivalente a premultiplicar A por una matriz elemental E (Las matrices elementales son aquellas que se obtienen a partir de una única operación elemental de matrices sobre la matriz identidad además se puede asegurar que las matrices elementales siempre tienen inversa).

Luego como $Bx = d$ se obtuvo aplicando una cantidad finita de operaciones elementales por fila al sistema $Ax = b$ entonces se puede escribir $Bx = d$ de la siguiente manera:

$E_nE_{n-1}\cdots E_2E_1Ax = E_nE_{n-1}\cdots E_2E_1b \hookrightarrow x = (E_nE_{n-1}\cdots E_2E_1A)^{-1}E_nE_{n-1}\cdots E_2E_1b \hookrightarrow x = A^{-1}b$ por lo tanto ambos sistemas tienen la misma solución, lo que nos indica que son equivalentes.

Definición: Matriz Inversa

Sea A una matriz nxn . Se dice que B es la inversa de A si y solo si $AB = BA = I$.

Definición: Matriz Positiva Definida

Se dice que la matriz A es positiva definida si

$$x^tAx > 0, \forall x \neq 0$$

.

Ejemplo: Sea $A = \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix}$ verificamos si es o no positiva definida siendo $x^t = \begin{bmatrix} x_1 & x_2 \end{bmatrix}$tal que $x \neq 0$.

$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 3 & 2 \\ 2 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = [3x_1 + 2x_2\,,\, 2x_1 + 3x_2] \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3x_1^2 + 4x_1x_2 + 3x_2^2 = x_1^2 + (2x_1^2 + 4x_1x_2 + 2x_2^2) + x_2^2 =$

$2(x_1^2 + 2x_1x_2 + x_2^2) + x_1^2 + x_2^2 = 2(x_1 + x_2)^2 + x_1^2 + x_2^2 > 0$

3.1. Métodos Directos

Los métodos directos llevan a la solución del sistema en un número finito de pasos, mientras que los métodos iterativos necesitan (en teoría) un número infinito de pasos.

3.1.1. Eliminación Gaussiana Clásica (sin pivoteo)

En palabras sencillas el Método de Eliminación gaussiana consiste en convertir el sistema $Ax = b$ en un sistema equivalente $Bx = d$ donde la matriz B sea triangular superior, aplicando al sistema original o.e.f del tipo 3, para finalmente aplicar sustitución hacia atrás y así conseguir la solución del sistema.

Los pasos a seguir para completar una iteración del método son lo siguientes:

1. Escoger el pivote (en nuestro caso los pivotes siempre serán los elementos de la diagonal a_{ii} de la matriz asociada al paso). Al determinar el pivote se halla la fila pivote también.

2. Armar los multiplicadores $z_{i+1} = -a_{ij}/a_{ii}$ (los a_{ij}son los elemnetos que se encuentran por debajo del pivote).
   Note que como queremos llegar a un sistema equivalente pero triangular superior entonces la tarea de los multiplicadores será la de eliminar (hacer 0) todas las entradas que esten por debajo del pivote escogido.

3. Aplicar los multiplicadores al sistema y hallar el nuevo sistema equivalente.

Estos pasos se repetiran $n-1$ veces donde n es la dimensión de la matriz. Al final de las $n-1$ iteraciones tendremos un sistema triangular superior, así para hallar los valores de las entradas de x se aplica el método de sustitución hacia atrás.

Ejemplo:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix}$$

Iteración 1:

Pivote: 6, fila pivote: fila 1

$z_2 = -12/6 = -2$, $z_3 = -3/6 = -1/2$, $z_4 = -(-6)/6 = 1$

Aplicamos los multiplicadores al sistema:

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix} \begin{matrix} \\ -2f_1 + f_2 \\ -\frac{1}{2}f_1 + f_3 \\ f_1 + f_4 \end{matrix} \hookrightarrow \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ 21 \\ -26 \end{bmatrix}$$

Iteración 2:

Pivote: -4, fila pivote:fila 2

$z_3 = -(-12)/-4 = -3$, $z_4 = -2/-4 = 1/2$

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ 21 \\ -26 \end{bmatrix} \begin{matrix} \\ \\ -3f_2 + f_3 \\ \frac{1}{2}f_2 + f_4 \end{matrix} \hookrightarrow \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -21 \end{bmatrix}$$

Iteración 3:
Pivote: 2, fila pivote:fila 3
$z_4 = -4/2 = -2$

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -21 \end{bmatrix} \begin{matrix} \\ \\ \\ -2f_3 + f_4 \end{matrix} \hookrightarrow \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ 3 \end{bmatrix}$$

Aplicando sustitución hacia atrás se obtiene:

$$x = \begin{bmatrix} 1 \\ -3 \\ -2 \\ 1 \end{bmatrix}$$

En el paso 3 de la iteración del método se aplican los multiplicadores al sistema, hacer esto es equivalente a pre-multiplicar a ambos lados de la igualdad por una matriz que llamaremos matriz de eliminación.

La forma general de la matriz de eliminación M_k es la siguiente:

$$fila\, k \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -a_{k+1k}/a_{kk} & 1 & \cdots & 0 \\ \vdots & \ddots & & \vdots & \ddots & \vdots \\ 0 & \cdots & -a_{nk}/a_{kk} & 0 & \cdots & 1 \end{bmatrix}$$

Observaciones sobre M_k

1. M_k es una matriz triangular inferior con unos en la diagonal, por lo tanto es invertible.
2. $M_k^{-1} = I + me_k^t$, esto significa que M_k^{-1} es igual a M_k excepto que los signos de los multiplicadores son contrarios.

Finalmente para resolver el sistema $Ax = b$ se pre- multiplica a ambos lados por estas matrices de eliminación de manera que se obtiene:

$$M_{k-1}M_{k-2}\cdots M_2M_1Ax = M_{k-1}M_{k-2}\cdots M_2M_1b$$

el cual es equivalente al sistema original y por lo tanto tiene la misma solución x.

Hallaremos la solución del sistema del ejemplo anterior utilizando ahora matrices de eliminación.

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix} \quad M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ -\frac{1}{2} & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

$M_1Ax = M_1b$

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & -12 & 8 & 1 \\ 0 & 2 & 3 & -14 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ 21 \\ -26 \end{bmatrix} \quad M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -3 & 1 & 0 \\ 0 & -\frac{1}{2} & 0 & 1 \end{bmatrix}$$

$M_2M_1Ax = M_2M_1b$

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ -21 \end{bmatrix} \quad M_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -2 & 1 \end{bmatrix}$$

$$M_3M_2M_1Ax = M_3M_2M_1b \quad \begin{bmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 10 \\ -9 \\ 3 \end{bmatrix}$$

3.1.2. Factorización LU de una matriz A

S urge la idea de expresar a la matriz A como la multiplicación de dos matrices de la forma $A = LU$ donde L es triangular inferior y U es triangular superior.

Supongamos por un momento que A es una matriz $2x2$.

$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} l_{11} & 0 \\ l_{21} & l_{22} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ 0 & u_{22} \end{bmatrix}$

nuestra finalidad entonces seria encontrar los valor de las l_{ij} y de las u_{ij}. para ello nos planteamos el siguiente sistema:

$l_{11}u_{11} = a_{11}$

$l_{11}u_{12} = a_{12}$

$l_{21}u_{11} = a_{21}$

$l_{21}u_{12} + l_{22}u_{22} = a_{22}$

este sistema tiene 6 ecuaciones y 4 incognitas, luego en caso de existir solución serían infinitas. esto nos indica que la factorización LU de una matriz A no es única.

Con el fin de asegurar que se tengan la misma cantidad de incognitas que de ecuaciones se imponen restricciones sobre la escogencia de las entradas de L y de U.

- Se consideran las entradas de la diagonal de L todas iguales a 1, así definimos la factorización de Doolittle.
- Se consideran las entradas de la diagonal de U todas iguales a 1, así se define la factorización de Crout.

Ya sabemos que el proceso de eliminación Gaussiana nos conduce a un sistema $Bx = d$ equivalente al sistema $Ax = b$ donde B es triangular superior. Una manera entonces de escoger U será como la resultante de este último estado de la eliminación. Para determinar L dado que $U = M_{n-1}M_{n-2}\cdots M_2M_1A$ y queremos que $LU = A$ entonces $L = M_1^{-1}M_2^{-1}\cdots M_{n-2}^{-1}M_{n-1}^{-1}$.

Esta escogencia de L y de U nos da como resultado la llamada factorización de Doolittle en la cual la diagonal de la matriz L esta formada por unos.

Teorema: Si en el proceso de eliminación gaussiana todos los pivotes son distintos de cero entonces la matriz A tendrá factorización LU.

Demostración: Kincaid

3.1.3. Eliminación gaussiana con Pivoteo Parcial

Consideremos el siguiente sistema $2x2$:

$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ cuya solución se puede verificar facilmente es $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Tratemos de aplicar eliminación gaussiana.

En la primera iteración observamos que el pivote correspondiente es 0, por lo que no podemos iterar.

Consideremos ahora el siguiente sistema:

$\begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$ donde ε es un número muy cercano al cero y cuya solución es aproximadamente $x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Tratemos de resolver el sistema considerando que somos la máquina y que por supuesto tenemos ciertas restricciones numéricas.

Iteración 1:

Pivote:ε

$z_2 = -1/\varepsilon$

$\begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad -\frac{1}{\varepsilon}f_1 + f_2 \quad \hookrightarrow \begin{bmatrix} \varepsilon & 1 \\ 0 & -\frac{1}{\varepsilon}+1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{1}{\varepsilon}+2 \end{bmatrix} = \begin{bmatrix} \varepsilon & 1 \\ 1 & \frac{-1+\varepsilon}{\varepsilon} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{-1+2\varepsilon}{\varepsilon} \end{bmatrix} \approx \begin{bmatrix} \varepsilon & 1 \\ 0 & \frac{-1}{\varepsilon} \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \frac{-1}{\varepsilon} \end{bmatrix}$

Aplicando sustitución hacia atrás

$x_2 \approx 1$

$\varepsilon x_1 + x_2 = 1 \approx \varepsilon x_1 + 1 \approx 1 \hookrightarrow \varepsilon x_1 \approx 0 \hookrightarrow x_1 \approx 0$

Luego la solución sería aproximadamente $x \approx \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ que está muy lejos de la que sabemos es la verdadera solución aproximada.

El problema que se nos presenta es el siguiente:

Cuando se escoge un pivote en la eliminación se debe evitar que dicho pivote sea muy pequeño o cero. A continuación se plantea una estrategia adecuada para la escogencia del pivote de manera que se pueda evitar que en cualquier iteración se nos pueda presentar este problema.

La escogencia del pivote debe realizarse de la siguiente manera:

Se busca en la columna correspondiente, por debajo del pivote original, el mayor valor a_{ij} en valor absoluto, esto es:

$$pivote = \max_j |a_{ij}|, i > j$$

Una vez hecho esto se intercambian la fila pivote original con aquella fila que tiene al máximo y se prosigue con los siguientes paso de la eliminación Gaussiana (pasos 2 y 3).

Apliquemos esta metodología para resolver de nuevo el ejemplo anterior.

Iteración 1:

$\begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

pivote original ε, nuevo pivote 1, intercambio fila 2 con fila 1

$\begin{bmatrix} 1 & 1 \\ \varepsilon & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$

$z_2 = -\varepsilon$

$\begin{bmatrix} 1 & 1 \\ \varepsilon & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \quad -\varepsilon f_1 + f_2 \hookrightarrow \begin{bmatrix} 1 & 1 \\ 0 & 1-\varepsilon \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1-2\varepsilon \end{bmatrix} \approx \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \hookrightarrow x_2 \approx 1, x_1 + x_2 = 2 \hookrightarrow x_1 \approx 2 - 1 \hookrightarrow x_1 \approx 1$

así la solución aproximada es $x \approx \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

Intercambiar dos filas de un sistema es equivalente a pre-multiplicar a la derecha por una matriz de permutación.

Es decir el sistema $Ax = b$ será transformado aplicando las siguientes multiplicaciones de matrices:

$$M_{n-1}P_{n-1}M_{n-2}P_{n-2}\cdots M_2P_2M_1P_1Ax = M_{n-1}P_{n-1}M_{n-2}P_{n-2}\cdots M_2P_2M_1P_1b$$

Apliquemos este método para resolver el sistema

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix}$$

$$\begin{bmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 12 \\ 34 \\ 27 \\ -38 \end{bmatrix} \quad P_1 = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$P_1Ax = P_1b$

$$\begin{bmatrix} 12 & -8 & 6 & 10 \\ 6 & -2 & 2 & 4 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 34 \\ 12 \\ 27 \\ -38 \end{bmatrix} \quad M_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ -\frac{1}{4} & 0 & 1 & 0 \\ \frac{1}{2} & 0 & 0 & 1 \end{bmatrix}$$

$M_1P_1Ax = M_1P_1b$

$$\begin{bmatrix} 12 & -8 & 6 & 10 \\ 0 & 2 & -1 & -1 \\ 0 & -11 & \frac{15}{2} & \frac{1}{2} \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 34 \\ -5 \\ \frac{37}{2} \\ -21 \end{bmatrix} \quad P_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$P_2M_1P_1Ax = P_2M_1P_1b$

$$\begin{bmatrix} 12 & -8 & 6 & 10 \\ 0 & -11 & \frac{15}{2} & \frac{1}{2} \\ 0 & 2 & -1 & -1 \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 34 \\ \frac{37}{2} \\ -5 \\ -21 \end{bmatrix} \quad M_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{2}{11} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

$M_2P_2M_1P_1Ax = M_2P_2M_1P_1b$

$$\begin{bmatrix} 12 & -8 & 6 & 10 \\ 0 & -11 & \frac{15}{2} & \frac{1}{2} \\ 0 & 0 & \frac{4}{11} & -\frac{10}{11} \\ 0 & 0 & 4 & -13 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 34 \\ \frac{37}{2} \\ -\frac{18}{11} \\ -21 \end{bmatrix} \quad P_3 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$$

$P_3M_2P_2M_1P_1Ax = P_3M_2P_2M_1P_1b$

$$\begin{bmatrix} 12 & -8 & 6 & 10 \\ 0 & -11 & \frac{15}{2} & \frac{1}{2} \\ 0 & 0 & 4 & -13 \\ 0 & 0 & \frac{4}{11} & -\frac{10}{11} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 34 \\ \frac{37}{2} \\ -21 \\ -\frac{18}{11} \end{bmatrix} \quad M_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -\frac{1}{11} & 1 \end{bmatrix}$$

$M_3P_3M_2P_2M_1P_1Ax = M_3P_3M_2P_2M_1P_1b$

$$\begin{bmatrix} 12 & -8 & 6 & 10 \\ 0 & -11 & \frac{15}{2} & \frac{1}{2} \\ 0 & 0 & 4 & -13 \\ 0 & 0 & 0 & \frac{3}{11} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 34 \\ \frac{37}{2} \\ -21 \\ \frac{3}{11} \end{bmatrix}$$

por lo tanto $x = \begin{bmatrix} 1 \\ -3 \\ -2 \\ 1 \end{bmatrix}$

Ahora tratemos de expresar A como $A = LU$ siguiendo la misma idea propuesta anteriormente, tendríamos que:

$\hat{L} = P_1^{-1}M_1^{-1}P_2^{-1}M_2^{-1}P_3^{-1}M_3^{-1}$ por lo tanto

$$\hat{L} = \begin{bmatrix} \frac{1}{2} & -\frac{2}{11} & \frac{1}{11} & 1 \\ 1 & 0 & 0 & 0 \\ \frac{1}{4} & 1 & 0 & 0 \\ -\frac{1}{2} & 0 & 1 & 0 \end{bmatrix}$$

que evidentemente no es triangular inferior. De manera que tenemos $A = \hat{L}U$, donde $\hat{L}$ no es T.I, sin embargo $PA = P\hat{L}U = LU$

$$P\hat{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{4} & 1 & 0 & 0 \\ -\frac{1}{2} & 0 & 1 & 0 \\ \frac{1}{2} & -\frac{2}{11} & \frac{1}{11} & 1 \end{bmatrix}$$

Factorización de Cholesky (Caso especial de factorización LU)

Sea A una matriz con entradas reales, simétrica y positiva definida, entonces tiene una factorización única $A = LL^t$ donde L es triangular inferior con diagonal positiva.

Ejemplo:

$A = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}$

$L = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix}$ por lo tanto $LL^t = \begin{bmatrix} l_{11}^2 & l_{11}l_{21} & l_{11}+l_{31} \\ l_{11}l_{21} & l_{21}^2+l_{22}^2 & l_{31}l_{21}+l_{32}l_{22} \\ l_{11}+l_{31} & l_{31}l_{21}+l_{32}l_{22} & l_{31}^2+l_{32}^2+l_{33}^2 \end{bmatrix}$

de manera que:

$l_{11}^2 = 1$

$l_{11}l_{21} = \frac{1}{2}$

$l_{11}l_{31} = \frac{1}{3}$

$l_{21}^2 + l_{22}^2 = \frac{1}{3}$

$l_{31}l_{21} + l_{32}l_{22} = \frac{1}{4}$

$l_{31}^2 + l_{32}^2 + l_{33}^2 = \frac{1}{5}$

$$\text{por lo tanto } L = \begin{bmatrix} 1 & 0 & 0 \\ \frac{1}{2} & \sqrt{\frac{1}{12}} & 0 \\ \frac{1}{3} & \sqrt{\frac{1}{12}} & \frac{1}{6\sqrt{5}} \end{bmatrix}$$

3.1.4. Eliminación de Gauss-Jordan

El método de eliminación de Gauss-Jordan es una variación del método de Gauss previamente estudiado en el que se reduce la matriz a una diagonal utilizando o.e.f. Las matrices de eliminación de Gauss-Jordan son muy parecidas a las que ya hemos visto y está es su forma general:

$$M_k = \begin{bmatrix} 1 & \cdots & -\frac{a_{1,k}}{a_{k,k}} & 0 & \cdots & 0 \\ 0 & \ddots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & -\frac{a_{k+1,k}}{a_{k,k}} & 1 & \cdots & 0 \\ 0 & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -\frac{a_{n,k}}{a_{k,k}} & 0 & \cdots & 1 \end{bmatrix}$$

Las matrices de eliminación de gauss-Jordan elimina todo lo que esta por encima y por debajo del pivote a_{kk}

3.1.5. Conteo de Operaciones

Definición: Matriz Diagonal Dominante (DD)
Sea dice que la matriz A es DD si y solo si se cumple que

$$|a_{ii}| \quad > \quad \sum_{j=1}^{n} a_{ij} \; j \neq i \; 1 \leq i \leq n$$

Teorema: La eliminación Gaussiana sin pivoteo preserva la dominancia diagonal de la matriz
Teorema: Toda matriz DD es no singular y tiene factorización LU

3.2. Norma de una Matriz y Número de condición

Definición: Norma de un vector
Sea V un espacio vectorial sobre un cuerpo K. Se dice que la función $\|.\|$ que va de V a $\mathbb{R}$ es una norma si se cumplen las siguientes:

1. $\|v\| \geq 0 \; \forall v \in V$ ($\|v\| = 0$ si y sólo si $v = 0$)
2. $\|\alpha v\| = |\alpha| \, \|v\| \;\; \forall \alpha \in K, \forall v \in V$

3. $\|v+w\| \leq \|v\| + \|w\| \ \ \forall v, w \in V$

Definición: Norma p de un vector

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

Definición: Norma de una matriz A

La norma de una matriz A es una función $\|.\| : R^{mxn} \rightarrow R$ tal que:

1. $\|A\| \geq 0 \ \forall A \in R^{mxn}$ ($\|A\| = 0$ si y sólo si $A = 0$)
2. $\|\alpha A\| = |\alpha| \, \|A\| \ \ \forall \alpha \in K, \forall A \in R^{mxn}$
3. $\|A+B\| \leq \|A\| + \|B\| \ \ \forall A, B \in R^{mxn}$

Definición: Norma compatible o consistente

Diremos que la norma matricial $\|.\|$ es compatible o consistente con la norma vectorial si $\|Ax\| \leq \|A\| \, \|x\|$

Definición: Norma submultiplicativa

Diremos que la norma vectorial $\|.\|$ es submultiplicativa si se cumple que: $\|AB\| \leq \|A\| \, \|B\|$

Definición: Sea $\|.\|$ una norma vectorial. La función

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|}$$

se llama norma matricial inducida.

Algunas normas matriciales

Norma 1:

$$\|A\|_1 = \max_j \sum_{i=1}^{m} |a_{ij}|$$

Norma ∞:

$$\|A\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}|$$

Norma 2 o norma espectral:

$$\|A\|_2 = \sqrt{\rho(A^t A)} = \sqrt{\rho(A A^t)}$$

donde $\rho(B) = max\{|\lambda| : det(B - \lambda I) = 0\}$ a $\rho(B)$ se le denomina radio espectral de la matriz B.

Norma Frobenius:

$$\|A\|_F = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2} = traza(AA^t)$$

Teorema: $\rho(A) \leq A$

Demostración:

Sea λ un autovalor de A y $v \neq 0$ su autovector asociado.

$|\lambda| \|v\| = \|\lambda v\| = \|Av\| \leq \|A\| \|v\|$ por lo tanto $|\lambda| \leq \|A\|$. Como esto se cumple para cualquier autovalor en particular se cumple para el mayor de ellos en valor absoluto.

Definición: Número de Condición

El número de condición de una matriz A se define como:

$$K(A) = \|A\| \left\|A^{-1}\right\|$$

donde $\|.\|$ es una norma inducida.

En general el resultado de $K(A)$ depende de la norma utilizada.

Un incremento en el número de condición produce sensitividad en la solución del sistema lineal con respecto a cambios en los datos.

Teorema: El número de condición de una matriz siempre será mayor o igual a la unidad, es to es : $K(A) \geq 1 \, \forall A$

Demostración:

$K(A) = \|A\| \left\|A^{-1}\right\| \geq \|A\, A - 1\| = 1$

Teorema: $K(A^{-1}) = \|K(A)\|$

Demostración:

$K(A^{-1}) = \left\|A^{-1}\right\| \left\|(A^{-1})^{-1}\right\| = \left\|A^{-1}\right\| \|A\| = \|A\| \left\|A^{-1}\right\| = K(A)$

Teorema: $K(\alpha A) = K(A)$ $\alpha \neq 0$

Demostración:

$K(\alpha A) = \|\alpha A\| \left\|(\alpha A)^{-1}\right\| = |\alpha| \|A\| \left|\alpha^{-1}\right| \left\|A^{-1}\right\| = |\alpha| \left|\alpha^{-1}\right| \|A\| \left\|A^{-1}\right\| = \|A\| \left\|A^{-1}\right\| = K(A)$

El número de condición de una matriz singular se establece como infinito.

Teorema: Si A es una matriz ortogonal entonces $K_2(A) = 1$

Demostración:

Como A es ortogonal entonces $A.A^t = I$, o si bien se quiere $A^{-1} = A^t$

$K_2(A) = \sqrt{\rho(A^t A)} = \sqrt{\rho(I)} = 1$

Definamos la distancia relativa de la matriz A al conjunto de las matrices singulares con respecto a la norma p como:

$$dist_p(A) = min \left\{ \frac{\|\delta A\|_p}{\|A\|_p} : det(A + \delta A) = 0 \right\}$$

se puede establecer que $dist_p(A) = \frac{1}{K(A)}$ lo cual sugiere que si la matriz A tiene un número de condición muy grande se puede comportar entonces como una matriz singular de la forma $A + \delta A$.

Exactitud de la solución

Intuitivamente la manera más sencilla de verificar la exactitud de la solucón obtenida es sustituir en el sistema original y calcular el valor del vector residual $r = b - A\hat{x}$. En teoría si A es no singular el error $\|x - \hat{x}\| = 0$ si y sólo si $\|r\| = 0$. En la práctica, sin embargo, estas cantidades no necesariamente son pequeñas al mismo tiempo.

Estimando la exactitud

Supongtamos que $\hat{x}$ cumple que $A\hat{x} = b + \Delta b$. Si definimos $\Delta x = \hat{x} - x$ entonces tenemos

$$b + \Delta b = A\hat{x} = A(x + \Delta x) = Ax + A\Delta x$$

.

Como sabemos $Ax = b$, entonces debe cumplirse que $A\Delta x = \Delta b$, de donde $\Delta x = A^{-1}\Delta b$.

Por otro lado $b = Ax \Longrightarrow \|b\| = \|Ax\| \Longrightarrow \|b\| \leq \|A\| \|x\|(1)$

y como $\Delta x = A^{-1}\Delta b \Longrightarrow \|\Delta x\| = \left\|A^{-1}\Delta b\right\| \Longrightarrow \|\Delta x\| \leq \left\|A^{-1}\right\| \|\Delta b\|(2)$

de (1) y (2) se obtiene

$$\|b\| \|\Delta x\| \leq \|A\| \|x\| \left\|A^{-1}\right\| \|\Delta b\|$$

$$\frac{\|\Delta x\|}{\|x\|} \leq \|A\| \left\|A^{-1}\right\| \frac{\|\Delta b\|}{\|b\|} \Longrightarrow \frac{\|\Delta x\|}{\|x\|} \leq K(A)\frac{\|\Delta b\|}{\|b\|}$$

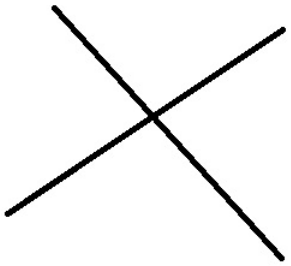
Por lo tanto el número de condición determina el posible cambio relativo en la solución producto de cambios en el lado derecho del sistema.

Se puede obtener un resultado similar si consideramos cambios en las entradas de la matriz A suponiendo que la solución aproximada $\hat{x}$ salisface que

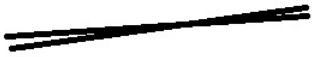
$$(A+E)\hat{x} = b \Longrightarrow x - \hat{x} = A^{-1}(b - A\hat{x}) = A^{-1}E\hat{x} \Longrightarrow \|\Delta x\| \leq \left\|A^{-1}\right\| \|E\| \|\hat{x}\| \Longrightarrow \frac{\|\Delta x\|}{\|\hat{x}\|} \leq K(A)\frac{\|E\|}{\|A\|}$$

Interpretación geométrica

Consideremos el caso de dos dimensiones si las rectas que conforman el sistema son casi paralelas entonces el punto de intersección no queda tan claramente definido y debido a los errores de redondeo es probable que se obtenga una solución al sistema poco precisa. En el caso en que las rectas son casi perpendiculares la intersección de ellas quedará perfectamente definida.



Sistema Bien condicionado



Sistema mal condicionado

Teorema:

3.3. Métodos Iterativos

3.3.1. Mejorando una aproximación. Refinamiento Iterativo

El refinamiento iterativo es una técnica que nos permite mejorar la exactitud de una aproximación obtenida a partir de la aplicación de un método directo.

Partiendo de una solución aproximada $\hat{x}$ y suponiendo que desconocemos la solución exacta x, sabemos que $A\hat{x} \neq b$, definamos $r = b - A\hat{x}$, por supuesto sabemos que $r \neq 0$, lo que deseamos es que r sea lo más pequeño posible y que $\hat{x}$ este lo más cerca posible de la solución exacta.

Supongamos que la variable z es la diferencia que existe entre la solución exacta y la solución aproximada que tenemos, esto es $z = x - \hat{x}$, por lo tanto premultiplicando por A a ambos lados se obtiene que : $Az = Ax - A\hat{x} = b - A\hat{x} = r$, asi que $Az = r$, es decir, puedo estimar esta diferencia resolviendo el sistema anterior. Ahora si tomo la solución aproximada $\hat{x}$ y le sumo z se obtiene que $\hat{x} + z \approx \widehat{x} + x - \hat{x} \approx x$ será una solución aproximada que este más cerca de la solución exacta.

Pasos a realizar para completar una iteración del refinamiento:

1. Calcular el residual $r^i = b - Ax^i$
2. Resolver el sistema $Az = r^i$
3. Calcular mejor aproximación $x^{i+1} = x^i + z$

4. Verificar criterio de parada: si $\|z\| / \left\|x^{i+1}\right\| < tol$ se termina el proceso siendo la mejor solución x^{i+1}

La elimiación Gaussiana es un ejemplo de método directo para resolver sistemas de ecuaciones lineales $Ax = b$, es decir, el método produce la solución exacta (asumiendo aritmética exacta) del sistema de ecuaciones en una cantidad finita de pasos. Un método iterativo comienza con un estimado inicial (x^0) de la solución y sucesivamente la mejora hasta que la solución sea tan exacta como se desee. En teoría se requiere una cantidad infinita de pasos para obtener convergencia a la solución exacta (x^*), pero en la práctica las iteraciones se detienen cuando la medida del error sea lo suficientemente pequeña.

Intuitivamente un método iterativo genera una sucesión $x^{(k)}$ de aproximaciones a la solución, de manera que se cumpla que:

$$\lim_{k\to\infty} x^{(k)} = x^*$$

Consideraremos métodos iterativos con la siguiente estructura:

$x^{(0)}$ iterado inicial.

$x^{(k+1)} = Bx^{(k)} + f$, donde B se llama la matriz de iteración del método, f es un vector que se obtiene a partir de b.

Definición: Métodos Consistentes

Un método iterativo se dice consistente si f y B son tales que $x^* = Bx^* + f$ esto es $f = (I - B)A^{-1}b$.

Teorema: Si $x^{(k+1)} = Bx^{(k)} + f$ es un método consistente entonces la sucesión de vectores $x^{(k)}$ converge a la solución de $Ax = b$ para alguna escogencia de $x^{(0)}$ si $\rho(B) < 1$.

Demostración: (pendiente)

Una manera de generar métodos iterativos consistentes es expresar A como la resta de matrices de la forma $P - N$ donde P es una matriz invertible. Dependiendo de la escogencia de P y N obtendremos diferentes métodos.

Partiendo de $x^{(0)}$ se calcula $x^{(k)}$ para $k \geq 1$ resolviendo el sistema

$Px^{(k+1)} = Nx^{(k)} + b \hookrightarrow x^{(k+1)} = P^{-1}Nx^{(k)} + P^{-1}b$ siendo el residual $r^{(k)} = b - Ax^{(k)}$

renombramos $B = P^{-1}N$ y $f = P^{-1}b$

3.3.2. Método de Richardson

Escogemos $P = I$, $N = I - A$ de manera que la matriz de iteración es $B = I - A$, y $f = b$, así el método queda expresado como:

$$x^{(k+1)} = (I - A)x^{(k)} + b$$

Para los siguientes métodos consideraremos $A = D - (E + F)$, donde D es diagonal (de hecho la diagonal de A), E es triangular inferior (se extrae la triangular inferior de A con signo contrario) y F es triangular superior (se estrae la triangular superior de A con signo contrario) con ceros en la diagonal.

3.3.3. Método de Jacobi

Consideraremos $P = D$, $N = E + F$, en este caso la matriz de iteración es $B = D^{-1}(E + F)$ y $f = D^{-1}b$, la iteración del método de Jacobi se escribe como:

$$x^{(k+1)} = D^{-1}(E + F)x^{(k)} + D^{-1}b$$

El método de Jacobi también se puede escribit de la siguiente forma:

$$x^{(k+1)} = \frac{b_i - \sum_{i=1}^{n} a_{ij}x_j^k}{a_{ii}}$$

3.3.4. Método de Gauss Seidel

Consideraremos $P = D - E$, $N = F$, la matriz de iteración es $B = (D - E)^{-1}F$ y $f = (D - E)^{-1}b$ de manera que la iteración queda expresada de la siguiente manera:

$$x^{(k+1)} = (D - E)^{-1}x^{(k)} + (D - E)^{-1}b$$

El método de Gauss-Seidel también se puede escribir como:

$$x^{(k+1)} = \frac{b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}}{a_{ii}}$$

Teorema: Si A es estrictamente diagonal dominante entonces los métodos de Jacobi y Gauss-Seidel convergen.

Demostración:

$$|a_{ii}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{jj}|$$

Jacobi converge:
$B = D^{-1}(E + F)$

$$\|B\|_\infty = \left\|D^{-1}(E + F)\right\|_\infty = \max_{1 \leq i \leq n} |1/a_{ii}| \sum_{j=1}^{n} |a_{ij}| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}/a_{ii}| < 1$$

ya que por definición de diagonal dominante.

Para demostrar que Gauss-Seidel converge se sigue la misma idea que para Jacobi.

Ejemplo:

Dado el siguiente sistema. Establezca la convergencia del sistema y realice 4 iteraciones del método de Jacobi y 4 iteraciones del método de Gauss-Seidel, en cada iteración calcule la norma del vector de residuos

$$\begin{bmatrix} 2 & -1 & 0 \\ 1 & 6 & -2 \\ 4 & -3 & 8 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 2 \\ -4 \\ 5 \end{bmatrix}$$

La matriz es estrictamente diagonal dominante de manera que tenemos asegurada la convergencia del método de Jacobi

Expreso A de la forma$A = D - (E + F)$

$D = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 6 & 0 \\ 0 & 0 & 8 \end{bmatrix}$, $E = \begin{bmatrix} 0 & 0 & 0 \\ -1 & 0 & 0 \\ -4 & 3 & 0 \end{bmatrix}$, $F = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$

consideremos que partimos de $x^{(0)} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$

Armo la matriz de iteración

$B = D^{-1}(E + F) = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ -\frac{1}{6} & 0 & \frac{1}{3} \\ -\frac{1}{2} & \frac{3}{8} & 0 \end{bmatrix}$

$f = D^{-1}b = \begin{bmatrix} 1 \\ -\frac{2}{3} \\ \frac{5}{8} \end{bmatrix}$

construyo los iterado y verifico si cumple con el criterio de parada determinando

$x^{(1)} = B * x^{(0)} + f = \begin{bmatrix} 1 \\ -\frac{2}{3} \\ \frac{5}{8} \end{bmatrix}$ $\|r_1\| = 6.0421$

$x^{(2)} = B * x^{(1)} + f = \begin{bmatrix} \frac{2}{3} \\ -\frac{5}{8} \\ -\frac{1}{8} \end{bmatrix}$ $\|r_2\| = 1.8680$

$x^{(3)} = B * x^{(2)} + f = \begin{bmatrix} \frac{11}{16} \\ -\frac{59}{72} \\ \frac{11}{192} \end{bmatrix}$ $\|r_3\| = 0.7749$

$x^{(4)} = B * x^{(3)} + f = \begin{bmatrix} \frac{85}{144} \\ -\frac{459}{576} \\ -\frac{5}{192} \end{bmatrix}$ $\|r_4\| = 0.5679$

Usando Gauss-Seidel

$B = (D - E)^{-1}F = \begin{bmatrix} 0 & \frac{1}{2} & 0 \\ 0 & -\frac{1}{12} & \frac{1}{3} \\ 0 & -\frac{9}{32} & \frac{1}{8} \end{bmatrix}$

$f = (D - E)^{-1}b = \begin{bmatrix} 1 \\ -\frac{5}{6} \\ -\frac{3}{16} \end{bmatrix}$

$$x^{(1)} = B * x^{(0)} + f = \begin{bmatrix} 1 \\ -\frac{5}{6} \\ -\frac{3}{16} \end{bmatrix} \quad \|r_1\| = 0.9138$$

$$x^{(2)} = B * x^{(1)} + f = \begin{bmatrix} \frac{7}{12} \\ -\frac{199}{144} \\ \frac{3}{128} \end{bmatrix} \quad \|r_2\| = 0.4219$$

$$x^{(3)} = B * x^{(2)} + f = \begin{bmatrix} \frac{169}{288} \\ -\frac{2615}{3456} \\ -\frac{49}{1024} \end{bmatrix} \quad \|r_3\| = 0.0851$$

$$x^{(4)} = B * x^{(3)} + f = \begin{bmatrix} \frac{327}{526} \\ -\frac{915}{1213} \\ \frac{769}{24576} \end{bmatrix} \quad \|r_4\| = 0.0332$$

Se puede observar claramente que en las iteraciones de Gauss-Seidel la norma del vector de residuos se hace pequeña con mayor rapidez que si usamos Jacobi.

Observaciones:

- En caso de que la matriz no sea diagonal dominante se verifica que el radio espectral de la matriz de iteración sea menor a 1.
- Cuando sucede que convergen tanto Jacobi como Gauss Seidel convergen, la convergencia de Gauss-Seidel será más rápida.