DEPARTAMENTO DE MATEMÁTICA APLICADA

ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

ÁLGEBRA LINEAL
(MÉTODOS NUMÉRICOS)

NOTAS DE CLASE

Profesores de la asignatura

CONTENIDOS

Tema 1

INTRODUCCIÓN AL ANÁLISIS NUMÉRICO.

ERRORES Y OTROS ASPECTOS IMPORTANTES

1.1. INTRODUCCIÓN

El objetivo del Análisis Numérico es proporcionar métodos convenientes para obtener soluciones útiles (resultados numéricos) de ciertos problemas matemáticos.

En ciertos casos, puede ocurrir que aunque el problema tenga solución y ésta sea única, no exista un método analítico que nos la permita calcular (por ejemplo, en el cálculo de ciertas integrales). En otros casos, sí existen métodos analíticos que nos conducen a la solución exacta, pero éstos pueden requerir un gran número de operaciones o no ser admisibles para su implementación en un ordenador; por ejemplo, según se ha visto en la resolución de sistemas lineales, la regla de Cramer conducía a la solución exacta, pero en la práctica, esto no será cierto para sistemas con una cierta dimensión, por el elevado número de operaciones que requiere.

En los casos anteriormente planteados, se buscan soluciones que, sin coincidir plenamente con la solución exacta del problema se aproximen a ella tanto como sea posible.

En definitiva, el Análisis Numérico es una parte de las Matemáticas que estudia los métodos constructivos para la resolución de problemas matemáticos de una forma aproximada. El conjunto de reglas que definen el método numérico recibe el nombre de ***algoritmo***. El fin de todo algoritmo es llegar a un programa, que implementado en un ordenador, proporcione la solución del problema planteado. Interesará, obviamente, elegir algoritmos que produzcan resultados lo más precisos posibles. En este sentido se va a introducir el concepto de error.

1.2. TIPOS DE ERRORES

Se van a considerar únicamente los errores generados en el proceso de resolución numérica de un problema. Conviene mencionar que también se cometen errores en los procesos de modelización matemática de los problemas a resolver; sin embargo, aunque también son importantes, su estudio no constituye una parte del Análisis Numérico.
Los errores generados en el proceso de resolución numérica pueden ser de dos tipos:

a) Errores de truncatura (e_t): Son debidos al truncamiento de un proceso matemático infinito.
Por ejemplo, si para el cálculo de la función elemental f(x) = sen(x) consideramos el polinomio de Taylor de grado 2n+1

$$p_n(x)=x-\frac{x^3}{3!}+\frac{x^5}{5!}-...+(-1)^n\frac{x^{2n+1}}{(2n+1)!}$$

el error de truncatura será $e_t=f(x)-p_n(x)$.

b) Errores de redondeo (e_r): Estos errores son consecuencia de que un ordenador trabaja con un número finito de dígitos.
Así, si en el ejemplo anterior se quiere evaluar el polinomio $p_n(x)$ en $x=\sqrt{3}$, el ordenador sólo admite un valor $x'=1.7...$ que tendrá un número finito de decimales; por tanto, lo que devuelve el ordenador es $p_n(x')$ y tendremos un error de redondeo $e_{r_1}=p_n(x)-p_n(x')$. También puede ocurrir que los coeficientes de $p_n(x)$ tengan un número infinito de decimales; en este caso, el ordenador no evaluará $p_n(x)$, sino otro polinomio $\tilde{p}_n(x)$ y tendremos un segundo error de redondeo $e_{r_2}=p_n(x')-\tilde{p}_n(x')$.
El error total en el proceso quedará

$$e=f(x)-\tilde{p}_n(x')=[f(x)-p_n(x)]+[p_n(x)-p_n(x')]+[p_n(x')-\tilde{p}_n(x')]=e_t+e_{r_1}+e_{r_2}$$

es decir, el error total es la suma de los errores producidos a lo largo de todo el proceso.

Una cuestión importante relacionada con el error es que, como en general no se va a disponer de la solución exacta del problema, tampoco conoceremos valores exactos del

error. Lo que se manejarán serán cotas o aproximaciones de dicho error y de lo que se trata, es de asegurarse que tales cotas o aproximaciones se asemejen lo más posible al error real.

Otros dos conceptos a tener en cuenta en el tema del error son los de error absoluto y error relativo. Estos conceptos se aplican a cualquiera de los errores (redondeo, truncatura y total) vistos anteriormente.

Error absoluto: Es la diferencia entre el valor verdadero y el valor aproximado y, para evitar valores negativos, se utilizan valores absolutos. Luego, llamando x al valor exacto y $\overline{x}$ al valor aproximado, se tiene que

$$e_{abs} = d(x, \overline{x}) = |x - \overline{x}| \tag{1}$$

En el caso en que x y $\overline{x}$ sean vectores de $\mathbb{R}^n$, el error absoluto quedará expresado como

$$e_{abs} = d(x, \overline{x}) = \|x - \overline{x}\| \tag{2}$$

pudiéndose utilizar cualquiera de las normas habituales de $\mathbb{R}^n$ vistas en temas anteriores.

Al no conocer la solución exacta x, en la práctica se calcularán cotas o aproximaciones de esta cantidad.

Error relativo: Es el cociente entre el error absoluto y el valor exacto y por tanto, nos informa del tamaño del error con respecto al tamaño de la solución exacta.

$$e_{rel} = \frac{e_{abs}}{|x|} \tag{3}$$

En el caso en que x y $\overline{x}$ sean vectores de $\mathbb{R}^n$, se volverán a utilizar normas.

Como el valor exacto x no se conocerá, a efectos prácticos se define el error relativo como

$$e_{rel} = \frac{e_{abs}}{|\overline{x}|} \tag{4}$$

En general, el error relativo proporciona una mejor medida del error que el error absoluto, como queda patente en los ejemplos siguientes:

Ejemplo 1: Si x = 0.00006 y su valor aproximado es $\bar{x}$ = 0.00005, el error absoluto cometido es $e_{abs} = |x - \bar{x}| = 10^{-5}$ que es muy pequeño; en cambio $e_{rel} = \frac{e_{abs}}{|\bar{x}|} = \frac{10^{-5}}{5 \cdot 10^{-5}} = 0.2$, es decir el error relativo es del 20%.

Ejemplo 2: Considérese ahora que x = 100500 y que su valor aproximado es $\bar{x}$ = 100000. En este caso, el error absoluto cometido es $e_{abs} = |x - \bar{x}| = 500$ y $e_{rel} = \frac{e_{abs}}{|\bar{x}|} = \frac{500}{100000} = 0.005$, es decir, el error relativo es del 0.5%.

Estos dos ejemplos muestran que porcentualmente hablando, con respecto al tamaño de la solución exacta, es más significativo el error de 10^{-5} cometido en el primer caso que el error de 500 cometido en el segundo.

1.3. ARITMÉTICA DEL ORDENADOR

Como ya se ha comentado anteriormente, las calculadoras y ordenadores sólo trabajan con una cantidad finita de cifras, de modo que los cálculos se realizan muchas veces con representaciones aproximadas de los números que se deberían utilizar. Al conjunto de los números que pueden ser representados de forma exacta en un ordenador, esto es, que coinciden con los contenidos en el ordenador, se les denomina ***números máquina***. Cuando se introduce (bien desde el exterior o como resultado de otra operación) un número x diferente de estos números máquina, el ordenador lo traduce en un número máquina, que en adelante denotaremos por **fl(x)**.

Los cálculos científicos en ordenadores se llevan a cabo en la forma que se conoce como ***representación de punto flotante normalizada***. Aunque el ordenador opera en representación binaria, para el análisis que aquí realizaremos de los problemas computacionales que surgen como consecuencia de utilizar únicamente los números máquina en las operaciones, es suficiente considerar los números representados en base 10. Un número x está escrito en representación de punto flotante normalizada cuando viene dado en la forma

$$x = \pm (0.d_1 d_2 ... d_t ...) \cdot 10^n \qquad d_i \in \{0,1,...,9\} \quad d_1 \neq 0 \tag{5}$$

Ejemplo: La representación en punto flotante para el número 3478.02 es

$$0.347802 \cdot 10^4 = 0.347802\text{E}4$$

y la del número 0.000347802 es

$$0.347802 \cdot 10^{-3} = 0.347802\text{E} - 3$$

Los ordenadores operan con un valor máximo para t. A este valor máximo se le denomina ***número de dígitos significativos*** con los que opera la máquina.

Para obtener el número máquina fl(x) asociado a un número real x, el ordenador realiza lo que se conoce como ***redondeo***, que consiste en que

- si la primera cifra que se desprecia es mayor o igual a 5, se añadirá una unidad a la última cifra.
- si la primera cifra que se desprecia es menor que 5, la última cifra no se modificará.

Ejemplo: Para el número real x=3.6666666…. trabajando con t=4 dígitos significativos, el número máquina asociado es fl(x)=3.667 y para el número real x=0.0036666…. el número máquina asociado es fl(x)=0.003667.

A continuación vamos a calcular las cotas de los errores absolutos y relativos cometidos por el ordenador en el proceso de redondeo. Suponemos que trabajamos con t dígitos significativos y que por tanto, al número real

$$x = a \cdot 10^n \quad \text{con} \quad a = \pm 0.d_1 d_2 ... d_t d_{t+1} ...$$

le corresponde el número máquina

$$fl(x) = a' \cdot 10^n \quad \text{con} \quad a' = \pm 0.d_1 d_2 ... d_t' \quad \text{siendo} \quad d_t' = \begin{cases} d_t & \text{si } d_{t+1} < 5 \\ d_t + 1 & \text{si } d_{t+1} >= 5 \end{cases}$$

El error absoluto está acotado por

$$e_{abs} = |x - fl(x)| = |a - a'| \times 10^n \le 5 \times 10^{-(t+1)} \times 10^n = 5 \times 10^{n-t-1}$$

y la cota del error relativo es

$$e_{rel} = \frac{e_{abs}}{|x|} = \frac{|x - fl(x)|}{|a| \cdot 10^n} \le \frac{5 \times 10^{n-t-1}}{10^{-1} \cdot 10^n} = 5 \times 10^{-t} \qquad (6)$$

Esta cota se conoce con el nombre de ***precisión del ordenador*** y se denota por **eps**.

Otro inconveniente de la representación en punto flotante, está relacionada con el hecho de que el exponente n, debe de estar comprendido entre un valor mínimo m y un valor máximo M, es decir, $m \le n \le M$.

Así, si se supone que m=-300 y M=300, el número $0.5432 \cdot 10^{400}$ es demasiado grande para ser representado mediante un número máquina. Este fenómeno se denomina

overflow y detendrá la ejecución del programa (esto puede ocurrir, por ejemplo, al realizar una división por un número próximo a 0).

Por otro lado, el número $0.5432 \cdot 10^{-400}$ es demasiado pequeño para ser representado por un número máquina distinto de cero. Este fenómeno se denomina ***underflow***.

Además de todo esto, hay que tener en cuenta también, que al utilizar estas representaciones inexactas de los números, las operaciones elementales (suma, resta, multiplicación y división) realizadas por el ordenador pueden perder algunas de las propiedades de las operaciones matemáticas que representan. Así, por ejemplo:

- La suma puede no ser asociativa
- Existen elementos neutros distintos de 0 para la suma

En consecuencia, en Análisis Numérico resulta de gran importancia estudiar cómo se propagan los errores en las operaciones, siendo necesario para ello obtener expresiones de los errores absolutos y relativos cometidos en ellas.

1.4. ASPECTOS A ANALIZAR EN LA ELECCIÓN DE UN ALGORITMO

Teniendo en cuenta los hechos ya comentados de que el Análisis Numérico da sólo una respuesta aproximada a la solución real de problema y que al operar con un ordenador se producen una serie de errores en las operaciones, a la hora de elegir un algoritmo es importante considerar los siguientes aspectos:

1- Convergencia del algoritmo

Muchos algoritmos dan como resultado términos de una sucesión $\{u_n\}_{n\in\mathbb{N}}$ que se aproxima a la solución exacta del problema **u** a medida que n crece. En estos casos, es importante conocer cuáles son las condiciones que debe verificar esta sucesión para poder asegurar que converge y que además lo haga hacia la solución **u** del problema, es decir, $\lim_{n\to\infty} u_n = \mathbf{u}$.

Otro factor importante es el análisis de la velocidad de convergencia del algoritmo, pues podría ocurrir que un algoritmo convergente necesite una gran cantidad de pasos para dar una buena aproximación, con lo cual podría resultar ineficaz.

2- Estabilidad del algoritmo y condicionamiento del problema

El hecho de que un algoritmo conduzca en teoría a la solución del problema no garantiza que en la práctica sea útil, ya que todo algoritmo debe satisfacer un requerimiento más: *el algoritmo debe de ser numéricamente estable***.** En este sentido, un ***algoritmo*** se dice que es ***inestable*****,** cuando los pequeños errores de redondeo que se producen en cada paso se propagan a través de cálculos posteriores con efecto creciente, llegándose a obtener un resultado que, aunque teóricamente debería ser próximo a la solución, en la práctica no tiene nada que ver con ella.

Por otro lado, cuando pequeñas perturbaciones en los datos iniciales conducen a resultados muy diferentes (independientemente del algoritmo utilizado en la resolución), se dice que el ***problema*** está ***mal condicionado*****.** Consecuentemente, un problema se dirá bien condicionado cuando pequeños errores en los datos iniciales den lugar a pequeños cambios en la solución.

Ejemplo: El sistema lineal

$$\begin{cases} 2x+6y=8 \\ 2x+6.00001y=8.00001 \end{cases}$$

tiene como solución exacta x = y = 1, mientras que el sistema

$$\begin{cases} 2x+6y=8 \\ 2x+5.99999y=8.00002 \end{cases}$$

tiene como solución exacta x = 10, y = -2.

En este ejemplo se puede observar cómo una diferencia en los coeficientes del orden de 10^{-5} da lugar a una diferencia en los resultados del orden de 10^{1}. Es, por tanto, un ejemplo de problema mal condicionado.

3- Coste operativo

Se trata de elegir métodos numéricos que disminuyan lo más posible el número de operaciones a realizar en el proceso. Ello es debido a que, de esta forma el problema se resolverá con mayor rapidez y además, al reducir el número de operaciones a realizar disminuiremos la propagación de errores de redondeo, lo que contribuirá a la estabilidad del algoritmo.

A la hora de analizar costes operativos, contabilizaremos sólo multiplicaciones y divisiones. El objetivo de contabilizar costes operativos es el de poder comparar distintos algoritmos entre sí con el fin de quedarnos con el más eficiente y normalmente,

cuando un algoritmo tiene más multiplicaciones y divisiones que otro también suele tener más sumas y restas.

4- Uso de la memoria del ordenador

Hay que tener en cuenta que la capacidad de memoria de un ordenador es limitada. En este sentido, a la hora de elegir un algoritmo es importante analizar sus necesidades de memoria, intentando reducirlas al mínimo en el momento de programarlo (por ejemplo, no utilizar dos matrices si con una puede realizarse todo el proceso).

1.5. TIPOS DE MÉTODOS NUMÉRICOS

A la hora de resolver un problema, existen dos tipos de métodos numéricos:

Métodos directos:

Son métodos que teóricamente proporcionan la solución exacta del problema, pero en la práctica esto no es cierto debido a los errores de redondeo en las operaciones. Estos métodos se usan, por ejemplo, para resolver ciertos sistemas de ecuaciones lineales.

En estos métodos, el problema principal consiste en el control de los errores de redondeo, que está en función del número de operaciones realizadas y de cómo se realicen dichas operaciones.

En esta situación, un método que en teoría sea aplicable puede que en la práctica no lo sea. Esto puede ser debido al elevado número de operaciones requerido, como es el caso de la regla de Cramer, pero también puede ser debido a que como consecuencia de los errores de redondeo la solución obtenida no verifique el problema que se pretendía resolver.

Métodos iterativos:

Estos son los métodos que dan como resultado términos de una sucesión $\{u_n\}_{n\in\mathbb{N}}$ que se aproxima a la solución exacta del problema **u** a medida que n crece. Tal y como ya se ha comentado anteriormente, lo importante en ellos es la convergencia de la sucesión hacia la solución exacta así como su velocidad de convergencia.

En estos métodos los errores de redondeo suelen tener, en general, efectos menos negativos, ya que si la sucesión converge a la solución exacta del problema las aproximaciones obtenidas se van mejorando en cada paso.

Tema 2

SISTEMAS DE ECUACIONES LINEALES

2.1. INTRODUCCIÓN

El problema que se va a estudiar en este tema es el de resolver un sistema de n ecuaciones lineales con n incógnitas de la forma:

$$\begin{cases} a_{1,1}\cdot x_1 + a_{1,2}\cdot x_2 + \ldots\ \ldots + a_{1,n}\cdot x_n = b_1 \\ a_{2,1}\cdot x_1 + a_{2,2}\cdot x_2 + \ldots\ \ldots + a_{2,n}\cdot x_n = b_2 \\ \quad . \qquad\quad . \qquad\ . \qquad . \qquad\quad . \\ \quad . \qquad\quad . \qquad\ . \qquad . \qquad\quad . \\ a_{n,1}\cdot x_1 + a_{n,2}\cdot x_2 + \ldots\ \ldots + a_{n,n}\cdot x_n = b_n \end{cases} \Leftrightarrow \tag{1}$$

$$\Leftrightarrow \begin{pmatrix} a_{1,1} & a_{1,2} & \ldots & \ldots & a_{1,n} \\ a_{2,1} & a_{2,2} & \ldots & \ldots & a_{2,n} \\ . & . & & . & . \\ . & . & & . & . \\ a_{n,1} & a_{n,2} & \ldots & \ldots & a_{n,n} \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ . \\ . \\ b_n \end{pmatrix} \Leftrightarrow A\cdot \underset{\sim}{x} = \underset{\sim}{b}$$

donde

- A es la matriz de los coeficientes, que será regular y de dimensión $n\times n$.
- $\underset{\sim}{x}$ es el vector de dimensión $n\times 1$, que contiene a las incógnitas.
- $\underset{\sim}{b}$ es el vector de los términos independientes.

Estos sistemas surgen en muchos problemas de Ingeniería y Ciencia. Las matrices que aparecen en la práctica con más frecuencia, pueden dividirse en dos categorías:

1 - *Matrices densas de orden moderado:* son matrices de tamaño no excesivamente grande (n<100 ó 200) y que además tienen muchos elementos no nulos (de ahí el nombre de densas).

2 – *Matrices dispersas de orden elevado:* son matrices que tienen muy pocos elementos no nulos (de ahí el nombre de dispersa), situados en general cerca de la diagonal principal, y con dimensiones del orden de 1000x1000 o superior.

Los métodos de resolución del sistema (1) son diferentes para cada uno de estos dos tipos de matrices:

Cuando la matriz es densa de orden moderado se emplean métodos directos. Estos métodos necesitan almacenar toda la matriz en la memoria del ordenador, lo cual daría problemas si la matriz fuese de orden elevado.

Cuando la matriz es dispersa de orden elevado suelen emplearse métodos iterativos. Estos métodos respetan los elementos nulos del sistema, lo cual conduce a que sólo será necesario almacenar en la memoria del ordenador los elementos no nulos, que no serán muchos puesto que la matriz es dispersa.

Observación: Si el sistema (1) se resolviera mediante la regla de Cramer, habría que calcular (n+1) determinantes distintos entre sí y de dimensión $n \times n$, además de realizar n divisiones $\left(x_i = \frac{|A_i|}{|A|} \quad i = 1, ..., n \right)$. Desarrollando por filas o columnas, el coste operativo de un determinante $n \times n$ es mayor que n! y por tanto, el coste operativo total para el sistema será mayor que (n+1)!. Así, por ejemplo, si n = 20 el número de operaciones sería 21! , es decir, aproximadamente $5.109094217 \times 10^{19}$; con un ordenador que realizase 100 millones de operaciones por segundo, el tiempo necesario para la resolución de dicho sistema es del orden de 16200 años. A la vista del ejemplo expuesto parece evidente que el método de Cramer no resulta muy eficaz.

2.2. RESOLUCIÓN DE SISTEMAS DE ECUACIONES LINEALES CON MATRICES TRIANGULARES

Sustitución regresiva

En este apartado se desarrolla un algoritmo para resolver un sistema de ecuaciones lineales $U \cdot \underset{\sim}{x} = \underset{\sim}{c}$, siendo U una matriz triangular superior. El motivo de desarrollar este apartado, es que más adelante veremos que el sistema general (1) se va a transformar en uno de este tipo, equivalente al sistema inicial.

Sea el sistema

$$\begin{cases} u_{1,1}\cdot x_1 + u_{1,2}\cdot x_2 + \ldots\ \ldots + u_{1,n-1}\cdot x_{n-1} + u_{1,n}\cdot x_n = c_1 \\ \quad\ u_{2,2}\cdot x_2 + \ldots\ \ldots + u_{2,n-1}\cdot x_{n-1} + u_{2,n}\cdot x_n = c_2 \\ \qquad\ \ . \qquad\qquad\ \ . \qquad\qquad\ \ . \qquad\ \ . \\ \qquad\ \ \ . \qquad\qquad\ \ . \qquad\qquad\ \ . \qquad\ \ . \\ \qquad\qquad\qquad\ u_{n-1,n-1}\cdot x_{n-1} + u_{n-1,n}\cdot x_n = c_{n-1} \\ \qquad\qquad\qquad\qquad\qquad\ u_{n,n}\cdot x_n = c_n \end{cases} \tag{2}$$

Comenzamos despejando x_n de la última ecuación como $x_n = \frac{c_n}{u_{n,n}}$; a continuación x_{n-1} de la penúltima ecuación como $x_{n-1} = \frac{c_{n-1} - u_{n-1,n}\cdot x_n}{u_{n-1,n-1}}$. Así sucesivamente, moviéndonos hacia arriba y despejando x_i de la i-ésima ecuación, llegaremos hasta despejar x_1 de la primera ecuación. El algoritmo que define este proceso quedará en la forma:

$$\begin{cases} x_n = \dfrac{c_n}{u_{n,n}} \\ x_k = \dfrac{c_k - \sum_{j=k+1}^{n} u_{k,j}\cdot x_j}{u_{k,k}} \quad k = n-1, n-2, \ldots, 2, 1 \end{cases} \tag{3}$$

Ejemplo: Se utilizará el método de sustitución regresiva para resolver el sistema:

$$\begin{cases} x_1 + 2x_2 - x_3 + x_4 = 1 \\ \quad -x_2 + x_3 + 2x_4 = 3 \\ \qquad\ 3x_3 + 13x_4 = 13 \\ \qquad\qquad -13x_4 = -13 \end{cases}$$

Despejando x_4 de la cuarta ecuación, se tiene que $x_4 = \frac{-13}{-13} = 1$; en la tercera ecuación $x_3 = \frac{13 - 13x_4}{3} = 0$; de la segunda ecuación $x_2 = -3 + x_3 + 2x_4 = -1$ y finalmente de la primera ecuación $x_1 = 1 - 2x_2 + x_3 - x_4 = 2$.

Coste operativo

- Para calcular x_n se efectuará 1 división.
- Para calcular x_k, serán necesarias (n - k) multiplicaciones y 1 división, es decir, (n – k + 1) operaciones. Como el proceso se realiza para k = n-1, n-2, …, 2, 1, el número total de operaciones será:

$$N = 1 + \sum_{k=1}^{n-1}(n-k+1) = 1+2+\ldots+n = \frac{n(n+1)}{2} \simeq \frac{n^2}{2}$$

Sustitución progresiva

Al igual que hay sistemas cuya matriz asociada es una matriz triangular superior, existen otros, cuya matriz de coeficientes es triangular inferior, es decir, será un sistema de la forma $L \cdot \underset{\sim}{y} = \underset{\sim}{b}$ que desarrollado queda:

$$\begin{cases} l_{1,1} \cdot y_1 & & & & & & & = & b_1 \\ l_{2,1} \cdot y_1 & + & l_{2,2} \cdot y_2 & & & & & = & b_2 \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot & & \cdot & & \cdot \\ l_{n-1,1} \cdot y_1 & + & l_{n-1,2} \cdot y_2 & + \ldots \ldots + & l_{n-1,n-1} \cdot y_{n-1} & & & = & b_{n-1} \\ l_{n,1} \cdot y_1 & + & l_{n,2} \cdot y_2 & + \ldots \ldots + & l_{n,n-1} \cdot y_{n-1} & + & l_{n,n} \cdot y_n & = & b_n \end{cases} \quad (4)$$

Comenzamos despejando y_1 de la primera ecuación como $y_1 = \frac{b_1}{l_{1,1}}$; a continuación y_2 de la segunda ecuación como $y_2 = \frac{b_2 - l_{2,1} \cdot y_1}{l_{2,2}}$. Así sucesivamente, moviéndonos hacia abajo y despejando y_i de la i-ésima ecuación, llegaremos hasta despejar y_n de la última ecuación. El algoritmo que define este proceso quedará en la forma:

$$\begin{cases} y_1 = \frac{b_1}{l_{1,1}} \\ y_k = \frac{b_k - \sum_{j=1}^{k-1} l_{k,j} \cdot y_j}{l_{k,k}} \quad k = 2,3,\ldots,n \end{cases} \quad (5)$$

Teniendo en cuenta que este algoritmo tiene la misma estructura que el anterior, su coste operativo será el mismo, es decir, del orden de $n^2/2$.

2.3. MÉTODO DE ELIMINACIÓN GAUSSIANA

El método de eliminación gaussiana tiene como objetivo transformar el sistema inicial $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$ en un sistema equivalente $U \cdot \underset{\sim}{x} = \underset{\sim}{c}$, siendo U una matriz triangular superior. Este sistema se resolverá finalmente mediante el algoritmo de sustitución regresiva visto en el apartado anterior.

Para transformar el sistema $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$ se realizarán transformaciones elementales de filas sobre la matriz A y su término independiente $\underset{\sim}{b}$. Un forma cómoda de realizar estas transformaciones, será trabajando en todo momento con la matriz ampliada del sistema.

2.3.1. Descripción del método

Comenzaremos desarrollando un ejemplo, para a continuación, basándonos en el procedimiento visto, escribir las ecuaciones que describen el algoritmo de eliminación gaussiana.

Ejemplo:

Dado el sistema $\begin{cases} x_1 + x_2 + & + 3x_4 = 4 \\ 2x_1 + x_2 - x_3 + & x_4 = 1 \\ 3x_1 - x_2 - x_3 + & 2x_4 = -3 \\ -x_1 + 2x_2 + 3x_3 - & x_4 = 4 \end{cases}$

modificaremos simultáneamente la matriz A y el término independiente; en consecuencia, partiremos en nuestro algoritmo de la matriz ampliada del sistema.

La matriz ampliada del sistema es: $\left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 2 & 1 & -1 & 1 & 1 \\ 3 & -1 & -1 & 2 & -3 \\ -1 & 2 & 3 & -1 & 4 \end{array}\right)$

En el primer paso del algoritmo y que llamaremos paso k=1, usando el elemento $a_{1,1}$ y en consecuencia la fila 1 del sistema, haremos transformaciones elementales de filas que nos permitan obtener ceros en la primera columna por debajo de la diagonal.

En adelante denotaremos por F_i a la fila i del sistema.

En consecuencia, las operaciones que realizaremos en este paso k = 1 serán:

$$\xrightarrow[\substack{F_3 = F_3 - 3F_1 \\ F_4 = F_4 + F_1}]{F_2 = F_2 - 2F_1} \left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & -4 & -1 & -7 & -15 \\ 0 & 3 & 3 & 2 & 8 \end{array}\right)$$

Nótese que la fila 1 no se modifica en este paso. De hecho, tampoco se modificará en pasos posteriores.

En el segundo paso del algoritmo y que llamaremos paso k=2, usando el elemento $a_{2,2}$ (con el valor -1 que tiene en el momento actual) y en consecuencia la fila 2 del sistema, haremos transformaciones elementales de filas que nos permitan obtener ceros en la segunda columna por debajo de la diagonal.

En consecuencia, las operaciones que realizaremos en este paso k = 2 serán:

$$\xrightarrow[F_4=F_4+3F_2]{F_3=F_3-4F_2} \left(\begin{array}{cccc|c} 1 & 1 & 0 & 3 & 4 \\ 0 & -1 & -1 & -5 & -7 \\ 0 & 0 & 3 & 13 & 13 \\ 0 & 0 & 0 & -13 & -13 \end{array}\right)$$

Nótese que en este paso ya no modificamos las filas 1 y 2 del sistema.

La matriz obtenida en este ejemplo es ya una matriz triangular superior, lo que significa que ya habríamos finalizado con la transformación del sistema, pasando a continuación a aplicar el algoritmo de sustitución regresiva.

Analizando los resultados, vemos que el cero que se ha obtenido en la posición (4,3) es casual (en el sentido de que no se ha realizado ninguna operación con el objetivo de conseguirlo). Esto significa que para una matriz cualquiera de orden 4×4, el algoritmo tendría otro paso al que llamaríamos k=3 y en el que haciendo uso del elemento $a_{3,3}$ actual ($a_{3,3}=3$) y en consecuencia de la fila 3, haríamos ceros en la columna 3 por debajo de la diagonal.

Generalizando estos resultados, para un sistema $n\times n$ realizaremos pasos $k=1,2,...,n-1$.

Paso k=1: usando el elemento $a_{1,1}$ y en consecuencia la fila 1 del sistema, haremos transformaciones elementales de filas que nos permitan obtener ceros en la primera columna por debajo de la diagonal. Estas transformaciones serán de la forma

$$F_i = F_i - c\cdot F_1 \quad i=2,3,..,n$$

La constante c de cada fila se elegirá para que el nuevo elemento de la posición (i,1), al que denotaremos por $a_{i,1}^{(nuevo)}$, tome valor cero; esto nos lleva a que:

$$a_{i,1}^{(nuevo)} = a_{i,1}^{(antiguo)} - c\cdot a_{1,1}^{(antiguo)} = 0 \Rightarrow c = \frac{a_{i,1}^{(antiguo)}}{a_{1,1}^{(antiguo)}}$$

Llamaremos $m_{i,1} = \dfrac{a_{i,1}^{(antiguo)}}{a_{1,1}^{(antiguo)}}$

Paso k=2: usando el elemento $a_{2,2}$ con su valor actual y en consecuencia la fila 2 del sistema, haremos transformaciones elementales de filas que nos permitan obtener ceros en la segunda columna por debajo de la diagonal. Estas transformaciones serán de la forma

$$F_i = F_i - c \cdot F_2 \quad i = 3, 4, .., n$$

La constante c de cada fila se elegirá para que el nuevo elemento de la posición (i,2), al que denotaremos por $a_{i,2}^{(nuevo)}$, tome valor cero; esto nos lleva a que:

$$a_{i,2}^{(nuevo)} = a_{i,2}^{(antiguo)} - c \cdot a_{2,2}^{(antiguo)} = 0 \Rightarrow c = \frac{a_{i,2}^{(antiguo)}}{a_{2,2}^{(antiguo)}}$$

Llamaremos $m_{i,2} = \frac{a_{i,2}^{(antiguo)}}{a_{2,2}^{(antiguo)}}$.

Nótese que en este paso se mantienen los ceros obtenidos en el paso anterior para la columna 1 y que las filas 1 y 2 ya no se modifican.

Cuando llegamos a un paso k genérico, ya habremos obtenido ceros en las columnas 1, 2, …, k-1 por debajo de la diagonal. Ahora, usando el elemento $a_{k,k}$ con su valor actual y en consecuencia la fila k del sistema, haremos transformaciones elementales de filas que nos permitan obtener ceros en la columna k por debajo de la diagonal. Estas transformaciones serán de la forma

$$F_i = F_i - c \cdot F_k \quad i = k+1, k+2, .., n$$

La constante c de cada fila se elegirá para que el nuevo elemento de la posición (i,k), al que denotaremos por $a_{i,k}^{(nuevo)}$, tome valor cero; esto nos lleva a que:

$$a_{i,k}^{(nuevo)} = a_{i,k}^{(antiguo)} - c \cdot a_{k,k}^{(antiguo)} = 0 \Rightarrow c = \frac{a_{i,k}^{(antiguo)}}{a_{k,k}^{(antiguo)}}$$

Llamaremos $m_{i,k} = \frac{a_{i,k}^{(antiguo)}}{a_{k,k}^{(antiguo)}}$.

En resumen, podemos decir que el algoritmo se podría escribir de la siguiente forma:

Para los pasos k=1,2, …, n-1, modificamos las filas i = k+1,k+2,…,n mediante la operación

$$F_i = F_i - m_{i,k} \cdot F_k \quad i = k+1, k+2, .., n \quad \text{siendo} \quad m_{i,k} = \frac{a_{i,k}^{(antiguo)}}{a_{k,k}^{(antiguo)}} \qquad (6)$$

Nótese que estas operaciones no afectan a las columnas 1,2, …,k-1 para las que ya se han conseguido los ceros deseados.

Desarrollando estas operaciones elemento a elemento, obtenemos que:

$$a_{i,j}^{(nuevo)} = \begin{cases} a_{i,j}^{(antiguo)} - m_{i,k} \cdot a_{k,j}^{(antiguo)} & si \quad i = k+1, k+2, ..., n \quad j = k+1, k+2, ..., n \\ 0 & si \quad i = k+1, k+2, ..., n \quad j = k \\ a_{i,j}^{(antiguo)} & en \ el \ resto \ de \ los \ casos \end{cases}$$

$$b_i^{(nuevo)} = \begin{cases} b_i^{(antiguo)} - m_{i,k} \cdot b_k^{(antiguo)} & si \quad i = k+1, k+2, ..., n \\ b_i^{(antiguo)} & en \ el \ resto \ de \ los \ casos \end{cases} \qquad (7)$$

Observaciones:

1) El elemento $a_{k,k}^{(antiguo)}$ recibe el nombre de ***pivote***. Podría ocurrir que en algún paso fuese 0. En la práctica, este problema lo solventaremos intercambiando filas, ya que al ser la matriz no singular, siempre existirá algún elemento entre $a_{k+1,k}^{(antiguo)}, a_{k+2,k}^{(antiguo)}, ..., a_{n,k}^{(antiguo)}$ que será distinto de cero. La forma de llevar a cabo esta técnica se verá posteriormente.

   Sin embargo, hay dos tipos especiales de matrices para las que se ha comprobado que el algoritmo de eliminación gaussiana no presenta ningún problema numérico, siendo por tanto un algoritmo estable. Estas matrices son: las matrices simétricas y definidas positivas y las matrices estrictamente diagonal dominantes. Definiremos a continuación este último tipo de matriz.

   Definición: Se dice que una matriz $A = (a_{i,j})$ de orden $n \times n$ es *estrictamente diagonal dominante* si

$$|a_{i,i}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{i,j}| \qquad \forall i = 1, 2, ..., n$$

   Ejemplo: La matriz $A = \begin{pmatrix} 7 & 2 & 0 \\ 3 & 5 & -1 \\ 0 & 5 & -6 \end{pmatrix}$ es estrictamente diagonal dominante pues

$$|a_{1,1}| = |7| > |a_{1,2}| + |a_{1,3}| = |2| + |0| = 2$$
$$|a_{2,2}| = |5| > |a_{2,1}| + |a_{2,3}| = |3| + |-1| = 4$$
$$|a_{3,3}| = |-6| = 6 > |a_{3,1}| + |a_{3,2}| = |0| + |5| = 5$$

2) Las constantes

$$m_{i,k} = \frac{a_{i,k}^{(antiguo)}}{a_{k,k}^{(antiguo)}}$$

reciben el nombre de ***multiplicadores.*** Su valor se calcula una única vez en cada fila i, guardándose para su posterior utilización.

2.3.2. Coste operativo

Para calcular el coste operativo necesario para resolver el sistema (1), se distinguirán tres fases:

a) Coste operativo para transformar la matriz de coeficientes A en la matriz triangular superior final.
b) Coste operativo para modificar el término independiente.
c) Coste operativo de la sustitución regresiva final.

Analicemos los costes de a) y b).

a) En cada etapa k y en cada fila i de dicha etapa, se efectúa una división al calcular el multiplicador $m_{i,k}$. A continuación, se efectúa una multiplicación para cada uno de los elementos $a_{i,j}^{(nuevo)}$ con $j = k+1, k+2, ..., n$. Por tanto, como calculamos n – k elementos diferentes, realizaremos (n – k + 1) operaciones.

Como en esta etapa k, operamos las filas $i = k+1, k+2, ..., n$, esto significa que operamos en n – k filas distintas. Por tanto, el coste operativo para realizar el paso k es de $(n-k+1)(n-k) = ((n-k)+1)(n-k) = (n-k)^2 + (n-k)$.

Como el algoritmo tiene pasos $k = 1, 2, ..., n-1$, el coste operativo total en este apartado será:

$$\sum_{k=1}^{n-1}[(n-k)^2 + (n-k)] = [1^2 + 2^2 + ... + (n-1)^2] + [1 + 2 + ... + (n-1)] =$$

$$= \frac{n(n-1)(2n-1)}{6} + \frac{n(n-1)}{2} \simeq \frac{n^3}{3}$$

En las operaciones anteriores se ha tenido en cuenta que $\begin{cases} \sum_{i=1}^{n} i = \frac{n(n+1)}{2} \\ \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6} \end{cases}$

b) Para contabilizar las operaciones realizadas en la modificación del término independiente, tenemos en cuenta que, en cada paso k del algoritmo calculamos los elementos $b_i^{(nuevo)}$ siendo $i = k+1, k+2, ..., n$ y realizamos una multiplicación en cada

uno de ellos; esto nos lleva a que en este paso k realizamos $(n-k)$ multiplicaciones. Por tanto, el número total de operaciones en la modificación del término independiente es de

$$\sum_{k=1}^{n-1}(n-k)=1+2+\ldots+(n-1)=\frac{(n-1)n}{2}\simeq\frac{n^2}{2}$$

Como conclusión de los apartados a), b) y c) anteriores, tenemos que el coste operativo total del método de eliminación gaussiana a la hora de resolver un sistema es del orden de

$$\frac{n^3}{3}+\frac{n^2}{2}+\frac{n^2}{2}\simeq\frac{n^3}{3}$$

Ejemplo:

Dado el sistema $\begin{cases} x_1 + 2x_2 + 4x_4 = 4 \\ 5x_1 + 6x_2 + 7x_3 + 8x_4 = 3 \\ 9x_1 + x_3 = 2 \\ 3x_1 + 4x_2 + 5x_3 + 6x_4 = 1 \end{cases}$

La matriz ampliada del sistema es: $\left(\begin{array}{cccc|c} 1 & 2 & 0 & 4 & 4 \\ 5 & 6 & 7 & 8 & 3 \\ 9 & 0 & 1 & 0 & 2 \\ 3 & 4 & 5 & 6 & 1 \end{array}\right)$

k=1: $\xrightarrow[\substack{m_{2,1}=5/1\ \ F_2=F_2-m_{2,1}F_1 \\ m_{3,1}=9/1\ \ F_3=F_3-m_{3,1}F_1 \\ m_{4,1}=3/1\ \ F_4=F_4-m_{4,1}F_1}]{\text{pivote } a_{1,1}=1} \left(\begin{array}{cccc|c} 1 & 2 & 0 & 4 & 4 \\ 0 & -4 & 7 & -12 & -17 \\ 0 & -18 & 1 & -36 & -34 \\ 0 & -2 & 5 & -6 & -11 \end{array}\right)$

k=2: $\xrightarrow[\substack{m_{3,2}=(-18)/(-4)=4.5\ \ F_3=F_3-m_{3,2}F_2 \\ m_{4,2}=(-2)/(-4)=0.5\ \ F_4=F_4-m_{4,2}F_2}]{\text{pivote } a_{2,2}=-4} \left(\begin{array}{cccc|c} 1 & 2 & 0 & 4 & 4 \\ 0 & -4 & 7 & -12 & -17 \\ 0 & 0 & -61/2 & 18 & 85/2 \\ 0 & 0 & 3/2 & 0 & -5/2 \end{array}\right)$

k=3: $\xrightarrow[m_{4,3}=(3/2)/(-61/2)=-3/61\ \ F_4=F_4-m_{4,3}F_3]{\text{pivote } a_{3,3}=-61/2} \left(\begin{array}{cccc|c} 1 & 2 & 0 & 4 & 4 \\ 0 & -4 & 7 & -12 & -17 \\ 0 & 0 & -61/2 & 18 & 85/2 \\ 0 & 0 & 0 & 54/61 & -25/61 \end{array}\right)$

Resolviendo por sustitución regresiva, queda que la solución es:

$x_4 = -25/54;\ \ x_3 = -5/3;\ \ x_2 = 49/18;\ \ x_1 = 11/27.$

2.3.3. Interpretación matricial del método de Gauss. Factorización LU

Analizando las transformaciones elementales de filas realizadas en el paso k=1 del algoritmo, que son

$$\begin{cases} F_2 = F_2 - m_{2,1} \cdot F_1 \\ F_3 = F_3 - m_{3,1} \cdot F_1 \\ \quad . \\ \quad . \\ F_n = F_n - m_{n,1} \cdot F_1 \end{cases}$$

tenemos, que la matriz obtenida al aplicar las mismas transformaciones elementales a la matriz unidad I_n es:

$$E_1 = \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & 0 \\ \frac{-a_{2,1}}{a_{1,1}} & 1 & 0 & \dots & \dots & 0 \\ \frac{-a_{3,1}}{a_{1,1}} & 0 & 1 & 0 & \dots & 0 \\ . & & & . & & \\ . & & & & . & \\ \frac{-a_{n,1}}{a_{1,1}} & 0 & \dots & \dots & 0 & 1 \end{pmatrix} \tag{8}$$

Las matrices $A^{(1)} = A$ y $A^{(2)}$ obtenida en el primer paso de la eliminación gaussiana, están relacionadas según la igualdad

$$A^{(2)} = E_1 \cdot A^{(1)}$$

Es interesante observar que la matriz E_1 es una matriz triangular inferior, con unos en la diagonal principal y que en su primera columna, contiene a los multiplicadores $m_{i,1}\ \ i = 2,3,...,n$ calculados en el paso k=1 cambiados de signo.

De la misma forma, la matriz obtenida al realizar sobre la matriz identidad las transformaciones elementales de filas realizadas en el paso k, es:

$$E_k = \begin{pmatrix} 1 & 0 & ... & ... & 0 & 0 & 0 & ... & ... & 0 \\ 0 & 1 & ... & ... & 0 & 0 & 0 & ... & ... & 0 \\ . & . & & & . & . & . & & & \\ . & . & & & . & . & . & & & \\ 0 & 0 & ... & ... & 1 & 0 & 0 & ... & ... & 0 \\ 0 & 0 & ... & ... & 0 & 1 & 0 & ... & ... & 0 \\ 0 & 0 & ... & ... & 0 & -m_{k+1,k} & 1 & & & \\ . & . & & & . & . & . & . & & \\ . & . & & & . & . & & & . & \\ 0 & 0 & ... & ... & 0 & -m_{n,k} & 0 & .. & .. & 1 \end{pmatrix} \quad (9)$$

es decir, la matriz E_k es una matriz triangular inferior, con unos en la diagonal principal y que en su columna k contiene a los multiplicadores $m_{i,k} = \frac{a_{i,k}}{a_{k,k}} \quad i = k+1, k+2, ..., n$ calculados en el paso k cambiados de signo.

Podemos entonces relacionar la matriz triangular superior $A^{(n)} = U$ del final del proceso de eliminación gaussiana con la matriz inicial A, a través de estas matrices $E_1, E_2, ..., E_{n-1}$ en la forma:

$$U = E_{n-1} \cdot E_{n-2} \cdot ... \cdot E_2 \cdot E_1 \cdot A$$

Despejando la matriz A premultiplicando por $E_{n-1}^{-1}, E_{n-2}^{-1}, ..., E_2^{-1}, E_1^{-1}$, se obtiene que:

$$A = E_1^{-1} \cdot E_2^{-1} \cdot ... \cdot E_{n-2}^{-1} \cdot E_{n-1}^{-1} \cdot U$$

Cada E_k^{-1} es la matriz asociada a las transformaciones elementales inversas a las realizadas en el paso k de la eliminación gaussiana, es decir, será la matriz obtenida a partir de la matriz unidad I_n, aplicándole las transformaciones $F_i = F_i + m_{i,k} \cdot F_k \quad i = k+1, k+2, ..., n$, y que por tanto, será una matriz triangular inferior, con unos en la diagonal principal y que en su columna k contiene a los multiplicadores $m_{i,k} \quad i = k+1, k+2, ..., n$ del paso k. Luego:

$$E_k^{-1}=\begin{pmatrix}1&0&\dots&\dots&0&0&0&\dots&\dots&0\\0&1&\dots&\dots&0&0&0&\dots&\dots&0\\.&.&&&.&.&.&&&\\.&.&&&.&.&.&&&\\0&0&\dots&\dots&1&0&0&\dots&\dots&0\\0&0&\dots&\dots&0&1&0&\dots&\dots&0\\0&0&\dots&\dots&0&m_{k+1,k}&1&&&\\.&.&&&.&.&.&.&&\\.&.&&&.&.&&.&&\\0&0&\dots&\dots&0&m_{n,k}&0&..&..&1\end{pmatrix} \tag{10}$$

Denotaremos por $L=E_1^{-1}\cdot E_2^{-1}\cdot\ldots\cdot E_{n-2}^{-1}\cdot E_{n-1}^{-1}$. Como puede comprobarse fácilmente, el resultado de estas multiplicaciones es

$$L=\begin{pmatrix}1&0&0&\dots&\dots&0&0\\m_{2,1}&1&0&\dots&\dots&0&0\\m_{3,1}&m_{3,2}&1&\dots&\dots&0&0\\.&.&.&&&.&.\\.&.&.&&&.&.\\m_{n-1,1}&m_{n-1,2}&m_{n-1,3}&\dots&\dots&1&0\\m_{n,1}&m_{n,2}&m_{n,3}&\dots&\dots&m_{n,n-1}&1\end{pmatrix} \tag{11}$$

Teniéndose, por tanto, que la matriz A admite una factorización en la forma $A=L\cdot U$, siendo U la matriz triangular superior $A^{(n)}$ y L la matriz triangular inferior anterior.

Ejemplo:

Si realizamos el proceso de eliminación gaussiana sobre la matriz $A=\begin{pmatrix}1&1&2\\1&2&1\\2&1&1\end{pmatrix}$, tendremos que

$$k=1:\ \xrightarrow[m_{3,1}=2/1\ \ F_3=F_3-m_{3,1}F_1]{\text{pivote }\ a_{1,1}=1\ \ m_{2,1}=1/1\ \ F_2=F_2-m_{2,1}F_1}\begin{pmatrix}1&1&2\\0&1&-1\\0&-1&-3\end{pmatrix}.\ \text{Luego}\ E_1=\begin{pmatrix}1&0&0\\-1&1&0\\-2&0&1\end{pmatrix}$$

$$k=2:\ \xrightarrow[m_{3,2}=-1/1\ \ F_3=F_3-m_{3,2}F_2]{\text{pivote }\ a_{2,2}=1}\begin{pmatrix}1&1&2\\0&1&-1\\0&0&-4\end{pmatrix}=U.\ \text{Luego}\ E_2=\begin{pmatrix}1&0&0\\0&1&0\\0&1&1\end{pmatrix}$$

Puede observarse que la matriz E_1 es la que se obtiene al aplicar las transformaciones elementales $F_2 - F_1$ y $F_3 - 2F_1$ a la matriz identidad, mientras que la matriz E_2 es la que se obtiene si se le aplica la transformación elemental $F_3 + F_2$ del segundo paso. En este caso tendremos que:

$$E_1^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 0 & 1 \end{pmatrix} \quad \text{y} \quad E_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}$$

siendo E_1^{-1} la matriz que se obtiene al aplicar las transformaciones elementales inversas $F_2 + F_1$ y $F_3 + 2F_1$ a la matriz identidad y, E_2^{-1} la que se obtiene a aplicar la transformación $F_3 - F_2$, inversa de $F_3 + F_2$, a la matriz identidad. Con lo que,

$L = E_1^{-1} \cdot E_2^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & -1 & 0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ m_{2,1} & 1 & 0 \\ m_{3,1} & m_{3,2} & 1 \end{pmatrix}$, es decir, L tiene en las columnas por debajo de la diagonal principal a los multiplicadores de los distintos pasos del proceso de eliminación Gaussiana.

Por tanto, la matriz A queda factorizada en la forma:

$$A = \begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix} = L \cdot U = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & -1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 1 & 1 & 2 \\ 0 & 1 & -1 \\ 0 & 0 & -4 \end{pmatrix}$$

Observaciones:

1) Conocida ya la factorización $A = L \cdot U$ de la matriz A, ésta podría usarse para resolver nuevos sistemas $A \cdot \underset{\sim}{x} = \underset{\sim}{d}$, con términos independientes $\underset{\sim}{d}$ diferentes al inicial, sin más que sustituir A por $L \cdot U$, es decir, usando que

$$A \cdot \underset{\sim}{x} = \underset{\sim}{d} \Leftrightarrow L \cdot U \cdot \underset{\sim}{x} = \underset{\sim}{d}$$

y llamando $\underset{\sim}{y} = U \cdot \underset{\sim}{x}$, se resolverá en primer lugar el sistema triangular inferior $L \cdot \underset{\sim}{y} = \underset{\sim}{d}$ mediante sustitución progresiva y, a continuación, se resolverá el sistema triangular superior $U \cdot \underset{\sim}{x} = \underset{\sim}{y}$ mediante sustitución regresiva.

Ejemplo:

Para resolver los sistemas:

$\begin{cases} x + 2y + z = 6 \\ 2x + y + 2z = 6 \\ x + 2y + 2z = 7 \end{cases}$ y $\begin{cases} x + 2y + z = 2 \\ 2x + y + 2z = 4 \\ x + 2y + 2z = 3 \end{cases}$, que tienen la misma matriz de coeficientes, tenemos dos opciones:

a) Modificar la matriz A ampliada con los 2 términos independientes, mediante las transformaciones del método de eliminación gaussiana

$$\xrightarrow[m_{3,1}=1/1\ \ F_3=F_3-m_{3,1}F_1]{\text{pivote } a_{1,1}=1\ \ m_{2,1}=2/1\ \ F_2=F_2-m_{2,1}F_1} \left(\begin{array}{ccc|cc} 1 & 2 & 1 & 6 & 2 \\ 0 & -3 & 0 & -6 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{array}\right)$$

y a continuación resolver mediante sustitución regresiva los sistemas $\begin{cases} x + 2y + z = 6 \\ -3y = -6 \\ z = 1 \end{cases}$ y

$\begin{cases} x + 2y + z = 2 \\ -3y = 0 \\ z = 1 \end{cases}$, obteniéndose que la solución del primer sistema es x=1, y=2, z=1, mientras que la solución del segundo sistema queda x=1, y=0, z=1.

b) Modificar sólo la matriz A mediante las transformaciones del método de eliminación gaussiana

$$\xrightarrow[m_{3,1}=1/1\ \ F_3=F_3-m_{3,1}F_1]{\text{pivote } a_{1,1}=1\ \ m_{2,1}=2/1\ \ F_2=F_2-m_{2,1}F_1} \begin{pmatrix} 1 & 2 & 1 \\ 0 & -3 & 0 \\ 0 & 0 & 1 \end{pmatrix} = U \text{ (en este caso, } m_{3,2}=0\text{)}.$$

A continuación, realizar la factorización $A = L \cdot U$, siendo $L = \begin{pmatrix} 1 & 0 & 0 \\ m_{2,1} & 1 & 0 \\ m_{3,1} & m_{3,2} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ y $U = \begin{pmatrix} 1 & 2 & 1 \\ 0 & -3 & 0 \\ 0 & 0 & 1 \end{pmatrix}$ de forma que, realizando las correspondientes sustituciones progresivas $L \cdot \underset{\sim}{y} = \underset{\sim}{d}$ y regresivas $U \cdot \underset{\sim}{x} = \underset{\sim}{y}$, llegamos a las mismas soluciones que en la opción anterior.

2) Conocida la factorización $A = L \cdot U$ para calcular el determinante de la matriz A, tenemos que tener en cuenta que

$$|A| = |L \cdot U| = |L| \cdot |U|$$

y que al ser L y U matrices triangulares, sus determinantes son los productos de los elementos de la diagonal principal.

2.4. MÉTODOS DE ELIMINACIÓN COMPACTA

2.4.1. Obtención de factorizaciones. Métodos de Doolittle y Crout

Según acabamos de ver el proceso de eliminación gaussiana conduce a una factorización de la matriz A del sistema en la forma $A = L \cdot U$, siendo L una matriz triangular inferior con unos en la diagonal principal y U una matriz triangular superior. Sin embargo, ésta no es la única forma de factorizar la matriz A como producto de una matriz triangular inferior por otra matriz triangular superior.

Para comprobarlo, sólo tenemos que considerar una matriz diagonal cualquiera D con elementos $d_{i,i}$ $i = 1, 2, ..., n$ en su diagonal principal y tener en cuenta que:

$$A = L \cdot U = L \cdot (D \cdot D^{-1}) \cdot U = (L \cdot D) \cdot (D^{-1} \cdot U) = L_1 \cdot U_1$$

obteniéndose, por tanto, otra factorización $A = L_1 \cdot U_1$ distinta de la inicial.

De la relación anterior, se puede concluir que dos factorizaciones distintas de una matriz A no singular, siempre están relacionadas a través de una matriz diagonal y consecuentemente, si fijamos los valores de los elementos de la diagonal principal en L o U obtendremos factorizaciones únicas.

Las situaciones más usuales suelen ser las dos siguientes:

1.- Fijar que $l_{i,i}=1 \ \forall i$. En este caso se habla de ***factorización de Doolittle.***

2.- Fijar que $u_{i,i}=1 \ \forall i$. En este caso se habla de ***factorización de Crout.***

Para obtener estas factorizaciones, iremos calculando sus elementos multiplicando directamente las filas de la matriz L por las columnas de la matriz U e igualando los resultados a los elementos de A. Para garantizar que este proceso nos lleve a los resultados deseados, los pasos que realizaremos son:

Paso k=1: calcularemos en primer lugar el elemento de la diagonal ($l_{1,1}$ o $u_{1,1}$) y a continuación los elementos $l_{2,1}, l_{3,1}, ..., l_{n,1}$ de la primera columna de L y los elementos $u_{1,2}, u_{1,3}, ..., u_{1,n}$ de la primera fila de U.

Paso k=2: calcularemos en primer lugar el elemento de la diagonal ($l_{2,2}$ o $u_{2,2}$) y a continuación los elementos $l_{3,2}, l_{4,2}, ..., l_{n,2}$ de la segunda columna de L y los elementos $u_{2,3}, u_{2,4}, ..., u_{2,n}$ de la segunda fila de U.

Así sucesivamente, llegaremos al paso k=n en el que sólo tendremos ya que calcular el elemento de la diagonal $l_{n,n}$ o $u_{n,n}$.

Concretamente, en el paso k calcularemos el elemento de la diagonal $l_{k,k}$ o $u_{k,k}$ multiplicando la fila k de L por la columna k de U. A continuación, calcularemos los elementos $l_{i,k}$ $i = k+1, k+2, ..., n$ multiplicando la fila i de L por la columna k de U y finalmente calcularemos los elementos $u_{k,j}$ $j = k+1, k+2, ..., n$ multiplicando la fila k de L por la columna j de U. La situación que tendremos será:

$$\begin{pmatrix} . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ l_{k,1.} & l_{k,2} & ... & ... & l_{k,k} & 0 & 0 & ... & ... & .... & ... & 0 \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ l_{i,1} & l_{i,2} & ... & ... & l_{i,k} & . & . & l_{i,i} & 0 & .... & ... & 0 \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . & . & . & . & . \end{pmatrix} \begin{pmatrix} . & . & . & u_{1,k} & . & . & u_{1,j} & . & . & . \\ . & . & . & u_{2,k} & . & . & u_{2,j} & . & . & . \\ . & . & . & ... & . & . & ... & . & . & . \\ . & . & . & ... & . & . & ... & . & . & . \\ . & . & . & u_{k,k} & . & . & u_{k,j} & . & . & . \\ . & . & . & 0 & . & . & ... & . & . & . \\ . & . & . & 0 & . & . & ... & . & . & . \\ . & . & . & ... & . & . & u_{j,j} & . & . & . \\ . & . & . & ... & . & . & 0 & . & . & . \\ . & . & . & ... & . & . & ... & . & . & . \\ . & . & . & .... & . & . & ... & . & . & . \\ . & . & . & 0 & . & . & 0 & . & . & . \end{pmatrix}$$

Al realizar las multiplicaciones de filas y columnas antes comentadas obtendremos:

$$\begin{cases} a_{k,k} = \sum_{r=1}^{k} l_{k,r} \cdot u_{r,k} = \sum_{r=1}^{k-1} l_{k,r} \cdot u_{r,k} + l_{k,k} \cdot u_{k,k} \Rightarrow l_{k,k} \cdot u_{k,k} = a_{k,k} - \sum_{r=1}^{k-1} l_{k,r} \cdot u_{r,k} \\ a_{i,k} = \sum_{r=1}^{k} l_{i,r} \cdot u_{r,k} = \sum_{r=1}^{k-1} l_{i,r} \cdot u_{r,k} + l_{i,k} \cdot u_{k,k} \Rightarrow l_{i,k} = \frac{1}{u_{k,k}} \left[a_{i,k} - \sum_{r=1}^{k-1} l_{i,r} \cdot u_{r,k} \right] \quad i = k+1, k+2, ..., n \\ a_{k,j} = \sum_{r=1}^{k} l_{k,r} \cdot u_{r,j} = \sum_{r=1}^{k-1} l_{k,r} \cdot u_{r,j} + l_{k,k} \cdot u_{k,j} \Rightarrow u_{k,j} = \frac{1}{l_{k,k}} \left[a_{k,j} - \sum_{r=1}^{k-1} l_{k,r} \cdot u_{r,j} \right] \quad j = k+1, k+2, ..., n \end{cases} \quad (12)$$

Sustituyendo $l_{k,k}$ por 1 en las ecuaciones anteriores obtendremos el algoritmo de Doolittle y, si sustituimos $u_{k,k}$ por 1 obtendremos el algoritmo de Crout.

Observaciones:

1) Se puede comprobar que el coste operativo de estas factorizaciones es el mismo que el de la modificación de la matriz A mediante la eliminación gaussiana, es decir, de orden de $n^3/3$ operaciones. En este sentido, estos algoritmos no suponen ni ventajas ni desventajas frente a la eliminación gaussiana.
2) Se ha dicho antes que una vez fijados los elementos de la diagonal en L o U, la factorización obtenida mediante este método es única. Esto significa que la factorización de Doolittle y la obtenida antes mediante el proceso de eliminación gaussiana serán la misma, ya que ambas tienen unos en la diagonal principal de L. En consecuencia, siempre que la eliminación gaussiana tenga problemas para ser llevada a cabo tal y como la hemos descrito, lo mismo ocurrirá con las factorizaciones descritas en este apartado.

Ejemplo:

Factorizaremos mediante el método de Doolittle la matriz $A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 3 & 1 \\ 1 & -1 & -1 \end{pmatrix}$

$$A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 3 & 1 \\ 1 & -1 & -1 \end{pmatrix} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ l_{2,1} & 1 & 0 \\ l_{3,1} & l_{3,2} & 1 \end{pmatrix} \cdot \begin{pmatrix} u_{1,1} & u_{1,2} & u_{1,3} \\ 0 & u_{2,2} & u_{2,3} \\ 0 & 0 & u_{3,3} \end{pmatrix}$$

k = 1: Comenzamos calculando el elemento de la diagonal de U

$$u_{1,1} = a_{1,1} = 1$$

A continuación, calculamos los elementos correspondientes a la primera columna de L y a la primera fila de U.

$$l_{2,1} \cdot u_{1,1} = a_{2,1} \Rightarrow l_{2,1} = 2$$

$$l_{3,1} \cdot u_{1,1} = a_{3,1} \Rightarrow l_{3,1} = 1$$

$$u_{1,2} = a_{1,2} = 1$$

$$u_{1,3} = a_{1,3} = 1$$

k = 2: Comenzamos calculando el elemento de la diagonal de U

$$l_{2,1} \cdot u_{1,2} + u_{2,2} = 2 \cdot 1 + u_{2,2} = a_{2,2} = 3 \Rightarrow u_{2,2} = 1$$

A continuación, calculamos los elementos correspondientes a la segunda columna de L y a la segunda fila de U.

$$l_{3,1}\cdot u_{1,2}+l_{3,2}\cdot u_{2,2}=1\cdot 1+l_{3,2}\cdot 1=a_{3,2}=-1 \Rightarrow l_{3,2}=-2$$

$$l_{2,1}\cdot u_{1,3}+u_{2,3}=2\cdot 1+u_{2,3}=a_{2,3}=1 \Rightarrow u_{2,3}=-1$$

k = 3: Calculamos únicamente el elemento de la diagonal de U

$$l_{3,1}\cdot u_{1,3}+l_{3,2}\cdot u_{2,3}+u_{3,3}=1\cdot 1+(-2)\cdot(-1)+u_{3,3}=a_{3,3}=-1 \Rightarrow u_{3,3}=-4$$

En consecuencia, las matrices L y U de la factorización quedan en la forma

$$L=\begin{pmatrix}1 & 0 & 0\\ 2 & 1 & 0\\ 1 & -2 & 1\end{pmatrix} \qquad U=\begin{pmatrix}1 & 1 & 1\\ 0 & 1 & -1\\ 0 & 0 & -4\end{pmatrix}$$

A continuación, factorizaremos la misma matriz mediante el método de Crout

$$A=\begin{pmatrix}1 & 1 & 1\\ 2 & 3 & 1\\ 1 & -1 & -1\end{pmatrix}=\begin{pmatrix}a_{1,1} & a_{1,2} & a_{1,3}\\ a_{2,1} & a_{2,2} & a_{2,3}\\ a_{3,1} & a_{3,2} & a_{3,3}\end{pmatrix}=\begin{pmatrix}l_{1,1} & 0 & 0\\ l_{2,1} & l_{2,2} & 0\\ l_{3,1} & l_{3,2} & l_{3,3}\end{pmatrix}\cdot\begin{pmatrix}1 & u_{1,2} & u_{1,3}\\ 0 & 1 & u_{2,3}\\ 0 & 0 & 1\end{pmatrix}$$

k = 1: Comenzamos calculando el elemento de la diagonal de L

$$l_{1,1}=a_{1,1}=1$$

A continuación, calculamos los elementos correspondientes a la primera columna de L y a la primera fila de U.

$$l_{2,1}=a_{2,1}=2$$

$$l_{3,1}=a_{3,1}=1$$

$$l_{1,1}\cdot u_{1,2}=u_{1,2}=a_{1,2}=1$$

$$l_{1,1}\cdot u_{1,3}=u_{1,3}=a_{1,3}=1$$

k = 2: Comenzamos calculando el elemento de la diagonal de L

$$l_{2,1}\cdot u_{1,2}+l_{2,2}=2\cdot 1+l_{2,2}=a_{2,2}=3 \Rightarrow l_{2,2}=1$$

A continuación, calculamos los elementos correspondientes a la segunda columna de L y a la segunda fila de U.

$$l_{3,1}\cdot u_{1,2}+l_{3,2}=1\cdot 1+l_{3,2}=a_{3,2}=-1 \Rightarrow l_{3,2}=-2$$

$$l_{2,1}\cdot u_{1,3}+l_{2,2}\cdot u_{2,3}=2\cdot 1+1\cdot u_{2,3}=a_{2,3}=1 \Rightarrow u_{2,3}=-1$$

k = 3: Calculamos únicamente el elemento de la diagonal de L

$$l_{3,1}\cdot u_{1,3}+l_{3,2}\cdot u_{2,3}+l_{3,3}=1\cdot 1+(-2)\cdot(-1)+l_{3,3}=a_{3,3}=-1 \Rightarrow l_{3,3}=-4$$

En consecuencia, las matrices L y U de la factorización quedan en la forma

$$L=\begin{pmatrix}1 & 0 & 0\\ 2 & 1 & 0\\ 1 & -2 & -4\end{pmatrix} \qquad U=\begin{pmatrix}1 & 1 & 1\\ 0 & 1 & -1\\ 0 & 0 & 1\end{pmatrix}$$

2.4.2. Método de Cholesky

El método de Cholesky es un caso particular de factorización compacta para las matrices simétricas definidas positivas. Ya comentamos anteriormente que para estas matrices el proceso de eliminación gaussiana no presentaba ningún problema numérico, siendo por tanto, un algoritmo estable. Teniendo en cuenta la relación existente entre las factorizaciones compactas y la factorización obtenida en la eliminación gaussiana, podemos afirmar que también ahora obtendremos algoritmos estables.

Recordemos que las matrices simétricas definidas positivas verificaban el criterio de Sylvester, y que por tanto, los determinantes de sus submatrices principales son siempre positivos. En particular, esto significa que en estas matrices el elemento $a_{1,1}>0$ siempre.

El método de Cholesky está basado en el siguiente resultado:

Teorema

Una matriz simétrica A es definida positiva si y sólo si existe una única matriz L triangular inferior, con los términos de la diagonal estrictamente positivos, tal que

$$A=L\cdot L^t$$

Esta factorización recibe el nombre de *factorización de Cholesky*.

Tomando como punto de partida, las ecuaciones (12) de los métodos de eliminación compacta

$$\begin{cases} l_{k,k}\cdot u_{k,k}=a_{k,k}-\sum_{r=1}^{k-1} l_{k,r}\cdot u_{r,k} \\ l_{i,k}=\dfrac{1}{u_{k,k}}\left[a_{i,k}-\sum_{r=1}^{k-1} l_{i,r}\cdot u_{r,k}\right] \quad i=k+1,k+2,...,n \\ u_{k,j}=\dfrac{1}{l_{k,k}}\left[a_{k,j}-\sum_{r=1}^{k-1} l_{k,r}\cdot u_{r,j}\right] \quad j=k+1,k+2,...,n \end{cases}$$

e imponiendo las condiciones $u_{k,k}=l_{k,k}$, $u_{r,k}=l_{k,r}$, $u_{k,j}=l_{j,k}$, se obtiene que ahora el algoritmo será para k=1, 2, …, n:

$$\begin{cases} l_{k,k}^2 = a_{k,k} - \sum_{r=1}^{k-1} l_{k,r}^2 \Rightarrow l_{k,k} = \sqrt{a_{k,k} - \sum_{r=1}^{k-1} l_{k,r}^2} \\ l_{i,k} = \frac{1}{l_{k,k}} \left[a_{i,k} - \sum_{r=1}^{k-1} l_{i,r} \cdot l_{k,r} \right] \quad i = k+1, k+2, \ldots, n \end{cases} \tag{13}$$

Observaciones:

1) El teorema anterior nos dice que este algoritmo se podrá llevar a cabo si y sólo si la matriz simétrica A es definida positiva. Observando las ecuaciones (13), podemos concluir que los cálculos podrán realizarse siempre que las cantidades que aparecen dentro de las raíces cuadradas sean positivas; por tanto, un análisis de la positividad o no de estas cantidades, permitirá en la práctica determinar de una forma rápida si una matriz simétrica y densa es definida positiva. En este sentido, este método se puede considerar un test para saber si una matriz A simétrica y densa es definida positiva o no; de hecho, es la forma práctica de comprobarlo.
2) Teniendo en cuenta que en la factorización de Cholesky sólo tenemos que calcular la matriz L, en vez de la L y la U, el coste operativo para su realización será aproximadamente la mitad del coste operativo de un método compacto genérico; por tanto será de aproximadamente $n^3/6$ operaciones (además de las n raíces cuadradas).

Ejemplo:

Comprobaremos, mediante la aplicación del método de Cholesky que la matriz simétrica $A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 5 & 1 \\ 1 & 1 & 3 \end{pmatrix}$ es definida positiva para a continuación, resolver el sistema

$$\begin{cases} x_1 - x_2 + x_3 = 1 \\ -x_1 + 5x_2 + x_3 = 1 \\ x_1 + x_2 + 3x_3 = 0 \end{cases}.$$

$$A = \begin{pmatrix} 1 & -1 & 1 \\ -1 & 5 & 1 \\ 1 & 1 & 3 \end{pmatrix} = \begin{pmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \\ a_{3,1} & a_{3,2} & a_{3,3} \end{pmatrix} = \begin{pmatrix} l_{1,1} & 0 & 0 \\ l_{2,1} & l_{2,2} & 0 \\ l_{3,1} & l_{3,2} & l_{3,3} \end{pmatrix} \cdot \begin{pmatrix} l_{1,1} & l_{2,1} & l_{3,1} \\ 0 & l_{2,2} & l_{3,2} \\ 0 & 0 & l_{3,3} \end{pmatrix}$$

k = 1: Comenzamos calculando el elemento de la diagonal de L

$$l_{1,1}^2 = a_{1,1} = 1 \Rightarrow l_{1,1} = \sqrt{1} = 1$$

A continuación, calculamos los elementos correspondientes a la primera columna de L.

$$l_{2,1}\cdot l_{1,1}=l_{2,1}=a_{2,1}=-1$$

$$l_{3,1}\cdot l_{1,1}=l_{3,1}=a_{3,1}=1$$

k = 2: Comenzamos calculando el elemento de la diagonal de L

$$l_{2,1}^2+l_{2,2}^2=1+l_{2,2}^2=a_{2,2}=5 \Rightarrow l_{2,2}=\sqrt{4}=2$$

A continuación, calculamos los elementos correspondientes a la segunda columna de L.

$$l_{3,1}\cdot l_{2,1}+l_{3,2}\cdot l_{2,2}=-1+2\cdot l_{3,2}=a_{3,2}=1 \Rightarrow l_{3,2}=1$$

k = 3: Calculamos únicamente el elemento de la diagonal de L

$$l_{3,1}^2+l_{3,2}^2+l_{3,3}^2=1+1+l_{3,3}^2=a_{3,3}=3 \Rightarrow l_{3,3}=\sqrt{1}=1$$

Como hemos podido completar el proceso de factorización, podemos afirmar que A es una matriz definida positiva, siendo

$$L=\begin{pmatrix}1&0&0\\-1&2&0\\1&1&1\end{pmatrix} \qquad L^t=\begin{pmatrix}1&-1&1\\0&2&1\\0&0&1\end{pmatrix}$$

A continuación, resolveremos el sistema, usando la factorización de Cholesky obtenida. Para ello sustituimos A por $L\cdot L^t$ en el sistema $A\cdot \underset{\sim}{x}=\underset{\sim}{b} \Rightarrow L\cdot L^t\cdot \underset{\sim}{x}=\underset{\sim}{b}$ y llamando $\underset{\sim}{y}=L^t\cdot \underset{\sim}{x}$, resolveremos en primer lugar el sistema

$$L\cdot \underset{\sim}{y}=\underset{\sim}{b} \Leftrightarrow \begin{cases} y_1 & & & & = & 1\\ -y_1 & + & 2y_2 & & = & 1\\ y_1 & + & y_2 & + & y_3 & = & 0\end{cases}$$

por sustitución progresiva, obteniendo que $\underset{\sim}{y}=\begin{pmatrix}1\\1\\-2\end{pmatrix}$. A continuación, resolvemos el sistema $L^t\cdot \underset{\sim}{x}=\underset{\sim}{y} \Leftrightarrow \begin{cases} x_1 & - & x_2 & + & x_3 & = & 1\\ & & 2x_2 & + & x_3 & = & 1\\ & & & & x_3 & = & -2\end{cases}$ por sustitución regresiva, obteniendo que la solución del sistema inicial es $\underset{\sim}{x}=\begin{pmatrix}9/2\\3/2\\-2\end{pmatrix}$.

2.5. CÁLCULO DE LA MATRIZ INVERSA

Teniendo en cuenta que la matriz inversa A^{-1} es la única matriz que verifica que

$$A \cdot A^{-1} = A^{-1} \cdot A = I = \begin{pmatrix} 1 & 0 & 0 & ... & ... & 0 \\ 0 & 1 & 0 & ... & ... & 0 \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 0 & 0 & ... & ... & 1 & 0 \\ 0 & 0 & ... & ... & 0 & 1 \end{pmatrix},$$

lo único que tenemos que hacer para calcularla es resolver los sistemas

$$A \cdot \underset{\sim}{x}_j = \underset{\sim}{e}_j \quad j = 1, 2, ..., n \qquad \text{siendo} \quad \underset{\sim}{e}_j = \begin{pmatrix} 0 \\ . \\ . \\ 0 \\ 1 \\ 0 \\ . \\ . \\ 0 \end{pmatrix} \rightarrow \text{posición } j$$

ya que si consideramos la matriz que tiene por columnas a estos vectores $\underset{\sim}{x}_j$, se verifica que $A \cdot (\underset{\sim}{x}_1 \ \underset{\sim}{x}_2 \ ... \ \underset{\sim}{x}_n) = (\underset{\sim}{e}_1 \ \underset{\sim}{e}_2 \ ... \ \underset{\sim}{e}_n) = I$. Consecuentemente, $A^{-1} = (\underset{\sim}{x}_1 \ \underset{\sim}{x}_2 \ ... \ \underset{\sim}{x}_n)$.

Para calcular A^{-1} en el algoritmo de eliminación gaussiana, modificaremos todos los términos independientes $\underset{\sim}{e}_1, \underset{\sim}{e}_2, ..., \underset{\sim}{e}_n$ a la vez.

Ejemplo:

Calcularemos A^{-1} por el método de Gauss siendo $A = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 3 & 1 \\ 1 & -1 & -1 \end{pmatrix}$.

Tenemos que resolver los sistemas $A \cdot \underset{\sim}{x}_j = \underset{\sim}{e}_j \quad j = 1, 2, 3$. Como pretendemos modificar los 3 términos independientes a la vez, partiremos de la matriz ampliada con los 3 términos independientes y a ella le aplicaremos las transformaciones de la eliminación gaussiana.

$$\left(\begin{array}{ccc|ccc} \boxed{1} & 1 & 1 & 1 & 0 & 0 \\ 2 & 3 & 1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 0 & 0 & 1 \end{array}\right) \xrightarrow[\substack{F_2=F_2-2\cdot F_1 \\ F_3=F_3-F_1}]{k=1} \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & \boxed{1} & -1 & -2 & 1 & 0 \\ 0 & -2 & -2 & -1 & 0 & 1 \end{array}\right)$$

$$\xrightarrow[F_3=F_3+2\cdot F_2]{k=2} \left(\begin{array}{ccc|ccc} 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 1 & -1 & -2 & 1 & 0 \\ 0 & 0 & -4 & -5 & 2 & 1 \end{array}\right)$$

Los tres sistemas triangulares que habrá que resolver a continuación mediante sustitución regresiva son de la forma:

$$\left\{\begin{array}{rcrcrcl|l|l} x & + & y & + & z & = & 1 & 0 & 0 \\ & & y & - & z & = & -2 & 1 & 0 \\ & & & & -4z & = & -5 & 2 & 1 \\ & & & & & & (1) & (2) & (3) \end{array}\right.$$

Al resolverlos, se obtiene que $A^{-1}=\left(\begin{array}{ccc} 1/2 & 0 & 1/2 \\ -3/4 & 1/2 & -1/4 \\ 5/4 & -1/2 & -1/4 \\ \text{Solución de (1)} & \text{Solución de (2)} & \text{Solución de (3)} \end{array}\right)$

En los algoritmos de Doolittle y Crout, una vez obtenida la factorización $A=L\cdot U$ de la matriz A, tenemos dos posibilidades equivalentes para calcular la inversa:

1) Sustituyendo A por $L\cdot U$, resolver los sistemas $A\cdot \underset{\sim}{x}_j=\underset{\sim}{e}_j \quad j=1,2,...,n$ llamando $y_j=U\cdot \underset{\sim}{x}_j$. Con esto, tendremos que resolver en primer lugar los sistemas $L\cdot \underset{\sim}{y}_j=\underset{\sim}{e}_j \quad j=1,2,...,n$ mediante sustitución progresiva, para a continuación resolver los sistemas $U\cdot \underset{\sim}{x}_j=\underset{\sim}{y}_j \quad j=1,2,...,n$ por sustitución regresiva.

2) Tener en cuenta que si $A=L\cdot U \Rightarrow A^{-1}=(L\cdot U)^{-1}=U^{-1}\cdot L^{-1}$. Para calcular L^{-1}, resolveremos los sistemas $L\cdot \underset{\sim}{y}_j=\underset{\sim}{e}_j \quad j=1,2,...,n$ por sustitución progresiva y para calcular U^{-1} resolveremos los sistemas $U\cdot \underset{\sim}{z}_j=\underset{\sim}{e}_j \quad j=1,2,...,n$ por sustitución regresiva.

En el caso particular de la factorización de Cholesky, utilizaremos esta segunda posibilidad siempre, ya que con ella se evita la resolución de los n sistemas de sustitución regresiva, dado que se tiene:

$$A=L\cdot L^t \Rightarrow A^{-1}=\left(L\cdot L^t\right)^{-1}=\left(L^t\right)^{-1}\cdot L^{-1}=\left(L^{-1}\right)^t\cdot L^{-1}$$

Es decir, sólo habrá que calcular L^{-1}.

Ejemplo:

Calcularemos mediante el método de Cholesky la inversa de la matriz $A=\begin{pmatrix}1 & -1 & 1\\ -1 & 5 & 1\\ 1 & 1 & 3\end{pmatrix}$. Esta matriz ya se factorizó en el apartado 2.4.2 habiéndose obtenido:

$$A=L\cdot L^t \quad \text{siendo} \quad L=\begin{pmatrix}1 & 0 & 0\\ -1 & 2 & 0\\ 1 & 1 & 1\end{pmatrix}$$

Calcularemos L^{-1} resolviendo los sistemas $L\cdot \underset{\sim}{y}_j=\underset{\sim}{e}_j \quad j=1,2,3$ por sustitución progresiva. Estos sistemas tienen la forma

$$\begin{cases} x & & & = & 1 & \mid 0 & \mid 0\\ -x & + & 2y & = & 0 & \mid 1 & \mid 0 \\ x & + & y & + z = & \underset{(1)}{0} & \mid \underset{(2)}{0} & \mid \underset{(3)}{1} \end{cases} \quad \text{cuyas soluciones son}$$

$$\underset{\sim}{y}_1=\begin{pmatrix}1\\ 1/2\\ -3/2\end{pmatrix} \quad \underset{\sim}{y}_2=\begin{pmatrix}0\\ 1/2\\ -1/2\end{pmatrix}, \underset{\sim}{y}_3=\begin{pmatrix}0\\ 0\\ 1\end{pmatrix}$$

con lo que $L^{-1}=\begin{pmatrix}1 & 0 & 0\\ 1/2 & 1/2 & 0\\ -3/2 & -1/2 & 1\end{pmatrix} \Rightarrow \left(L^{-1}\right)^t=\begin{pmatrix}1 & 1/2 & -3/2\\ 0 & 1/2 & -1/2\\ 0 & 0 & 1\end{pmatrix}$ y por tanto

$$A^{-1}=\left(L^{-1}\right)^t\cdot L^{-1}=\begin{pmatrix}7/2 & 1 & -3/2\\ 1 & 1/2 & -1/2\\ -3/2 & -1/2 & 1\end{pmatrix}$$

2.6. MÉTODO DE GAUSS CON PIVOTAJE PARCIAL Y CAMBIO DE ESCALA

Ya se dijo que el algoritmo de eliminación gaussiana tal y como lo habíamos planteado no podía llevarse a cabo si obteníamos un pivote $a_{k,k}$ nulo, pero la obtención de pivotes nulos no es el único problema de la eliminación gaussiana.

Cuando aparecen pivotes pequeños, la eliminación gaussiana presenta problemas de estabilidad. Veamos esta situación mediante el siguiente ejemplo:

Ejemplo: Considérese el sistema $\begin{cases} 0.003x_1 + 59.14x_2 = 59.17 \\ 5.291x_1 - 6.130x_2 = 46.78 \end{cases}$ cuya solución exacta es $\begin{cases} x_1 = 10 \\ x_2 = 1 \end{cases}$.

Supongamos ahora que dicho sistema se resuelve mediante eliminación gaussiana operando con 4 dígitos significativos. El pivote en este caso es $a_{1,1} = 0.003$ y el multiplicador asociado

$$m_{2,1} = \frac{a_{2,1}}{a_{1,1}} = \frac{5.291}{0.003} = 1763.\widehat{6} \simeq 1764$$

El sistema, después de la transformación $F_2 = F_2 - m_{2,1} \cdot F_1$ queda en la forma

$$\begin{cases} 0.003x_1 + 59.14x_2 = 59.17 \\ -104300x_2 = -104400 \end{cases}$$

cuya resolución mediante sustitución regresiva conduce a

$$\begin{cases} x_1 = -10.00 \\ x_2 = 1.001 \end{cases}$$

Como puede observarse, el pequeño error que tiene el cálculo de x_2, se amplifica al calcular x_1 en la operación $x_1 = \frac{59.17 - 59.14 \cdot x_2}{0.003}$, debido a la pequeña magnitud del pivote. En este ejemplo, esto se solucionaría intercambiando entre sí las filas 1 y 2 del sistema, escribiendo por tanto dicho sistema como

$$\begin{cases} 5.291x_1 - 6.130x_2 = 46.78 \\ 0.003x_1 + 59.14x_2 = 59.17 \end{cases}$$

Ahora el pivote es $a_{1,1} = 5.291$ y $m_{2,1} = \frac{a_{2,1}}{a_{1,1}} = \frac{0.003}{5.291} = 0.0005670$ y la transformación elemental $F_2 = F_2 - m_{2,1} \cdot F_1$ da lugar al sistema

$$\begin{cases} 5.291x_1 - 6.130x_2 = 46.78 \\ 59.14x_2 = 59.14 \end{cases}$$

cuya solución coincide con la exacta.

Lo anterior nos indica que en cada paso k del algoritmo deberíamos elegir como pivote el elemento de mayor valor absoluto que haya en la columna k-ésima a partir del elemento $a_{k,k}$, es decir, deberíamos buscar

$$\max_{i \geq k} | a_{i,k} |$$

y a continuación intercambiar la fila en la que se encuentre este máximo con la fila k. Esta técnica es lo que se conoce con el nombre de ***pivotaje parcial.***

Sin embargo, el problema anterior no es el único problema de estabilidad que se puede presentar en el algoritmo de eliminación gaussiana. Este algoritmo también presenta problemas de estabilidad cuando aparecen pivotes que son muy pequeños con respecto a los elementos de su fila. Veamos esta situación con un nuevo ejemplo.

Ejemplo: Considérese el sistema $\begin{cases} 30.00x_1 + 591400x_2 = 591700 \\ 5.291x_1 - 6.130x_2 = 46.78 \end{cases}$, que según puede observarse coincide con el anterior, salvo que su primera ecuación ha sido multiplicada por 10^4; por tanto, su solución exacta al igual que en el ejemplo anterior es $\begin{cases} x_1 = 10 \\ x_2 = 1 \end{cases}$.

Supongamos de nuevo que dicho sistema se resuelve mediante eliminación gaussiana operando con 4 dígitos significativos. El pivote en este caso $a_{1,1} = 30$, es ya el elemento de mayor valor absoluto en su columna. El multiplicador asociado es

$$m_{2,1} = \frac{a_{2,1}}{a_{1,1}} = \frac{5.291}{30} = 0.176\widehat{3} \simeq 0.1764$$

Y la operación $F_2 = F_2 - m_{2,1} \cdot F_1$ transforma el sistema en

$$\begin{cases} 30.00x_1 + 591400x_2 = 591700 \\ -104300x_2 = -104400 \end{cases}$$

que, mediante sustitución regresiva, da como solución

$$\begin{cases} x_1 = -10.00 \\ x_2 = 1.001 \end{cases}$$

En este caso, el error evidentemente no se produce por tener un pivote pequeño, sino por tener un pivote pequeño con respecto a los restantes elementos de su fila. Este problema se resuelve mediante lo que se denomina ***técnica de escalado.*** Para ello, en primer lugar, se definen los llamados ***factores de escala*** s_i para cada fila como

$$s_i = \max_{1 \leq j \leq n} | a_{i,j} |$$

Una vez obtenidos los factores de escala, deberíamos dividir todos los elementos de la fila i-ésima por el correspondiente s_i (de esta forma todos los elementos de la matriz A tomarían valores absolutos entre 0 y 1 y, por tanto, ya no tendríamos elementos que sean muy pequeños con respecto a los elementos de su fila).

Nota: La técnica de escalado no modifica la solución del sistema original, ya que únicamente transforma el sistema inicial $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$ en otro sistema equivalente.

En la práctica, y para evitar errores de redondeo de partida, el escalado de filas del sistema no se lleva a cabo, sino que se busca el pivote máximo de la estrategia de pivotaje parcial, teniendo en cuenta que la matriz debería de haberse escalado al inicio. Esto conduce a que en el paso k, deberíamos buscar

$$\max_{i \geq k} \frac{|a_{i,k}|}{s_i}$$

Por otro lado, cuando los sistemas de ecuaciones son grandes, el intercambio de filas de la estrategia de pivotaje parcial realmente no se lleva a cabo, y por tanto, sin haber intercambiado realmente las filas se funciona como si estuvieran intercambiadas. Esto se consigue mediante la utilización de un vector $\underset{\sim}{p}$ al que denominaremos ***vector de pivotaje*****;** en cada paso del algoritmo, este vector contendrá la posición que deberían estar ocupando las distintas filas en ese momento**.** Con esta idea, es claro que al inicio del algoritmo este vector tomará valores $\underset{\sim}{p} = \begin{pmatrix} 1 \\ 2 \\ . \\ . \\ n \end{pmatrix}$.

Comenzamos, desarrollando unos ejemplos:

Ejemplo: En el ejemplo con problemas de estabilidad antes analizado, esta técnica nos llevará a empezar calculando los factores de escala

$$s_1 = \max\{|30|, |591400|\} = 591400$$
$$s_2 = \max\{|5.291|, |-6.130|\} = 6.130$$

Para elegir el pivote se considera

$$\max\left\{\frac{|a_{1,1}|}{s_1},\frac{|a_{2,1}|}{s_2}\right\}=\max\left\{\frac{|30|}{591400},\frac{|5.291|}{6.130}\right\}=\max\left\{0.5073\times10^{-4},0.8631\right\}=0.8631,$$ lo que supone que el pivote debe de ser el elemento $a_{2,1}=5.291$ y que por tanto habría que intercambiar las filas 1 y 2 entre sí. La transformación elemental $F_1=F_1-m_{1,1}\cdot F_2$ y la correspondiente sustitución regresiva llevan a la solución $\begin{cases}x_1=10.00\\x_2=1.000\end{cases}$, que coincide con la solución exacta.

Ejemplo: Se trata de resolver el sistema $\begin{cases}x_1+3x_2+5x_3+7x_4=1\\2x_1-x_2+3x_3+5x_4=2\\2x_3+5x_4=3\\-2x_1-6x_2-3x_3+x_4=4\end{cases}$ mediante pivotaje parcial y cambio de escala.

Comenzamos calculando los factores de escala:

$$s_1=\max_{1\le j\le n}|a_{1,j}|=\max\{|1|,|3|,|5|,|7|\}=7;\ s_2=\max_{1\le j\le n}|a_{2,j}|=\max\{|2|,|-1|,|3|,|5|\}=5;$$

$$s_3=\max_{1\le j\le n}|a_{3,j}|=\max\{|0|,|0|,|2|,|5|\}=5;\ s_4=\max_{1\le j\le n}|a_{4,j}|=\max\{|-2|,|-6|,|-3|,|1|\}=6;$$

Inicializamos el vector de pivotaje con valores $\underset{\sim}{p}=\begin{pmatrix}1\\2\\3\\4\end{pmatrix}$ y vamos realizando las siguientes etapas:

k=1: Buscamos el elemento de máximo valor absoluto en la columna 1 como si la matriz se hubiera escalado, es decir, buscamos

$$\max_{i\ge1}\frac{|a_{i,1}|}{s_i}=\max\left\{\frac{|a_{1,1}|}{s_1},\frac{|a_{2,1}|}{s_2},\frac{|a_{3,1}|}{s_3},\frac{|a_{4,1}|}{s_4}\right\}=\max\left\{\frac{|1|}{7},\frac{|2|}{5},\frac{|0|}{5},\frac{|-2|}{6}\right\}=\frac{|2|}{5}=\frac{|a_{2,1}|}{s_2}$$

Luego, el pivote será el elemento $a_{2,1}$ y deberíamos intercambiar las filas 1 y 2 entre sí. En lugar de eso, intercambiamos los valores 1 y 2 en $\underset{\sim}{p}$, obteniendo que el nuevo vector de pivotaje es $\underset{\sim}{p}=\begin{pmatrix}2\\1\\3\\4\end{pmatrix}=\begin{pmatrix}p_1\\p_2\\p_3\\p_4\end{pmatrix}$. Las transformaciones a realizar en este paso serán

$$\begin{pmatrix} 1 & 3 & 5 & 7 & |1 \\ \boxed{2} & -1 & 3 & 5 & |2 \\ 0 & 0 & 2 & 5 & |3 \\ -2 & -6 & -3 & 1 & |4 \end{pmatrix} \xrightarrow[\substack{m_{1,1}=1/2\ \ F_1=F_1-m_{1,1}F_2 \\ m_{3,1}=0/2\ \ F_3=F_3-m_{3,1}F_2 \\ m_{4,1}=-2/2\ \ F_4=F_4-m_{4,1}F_2}]{} \begin{pmatrix} 0 & 7/2 & 7/2 & 9/2 & |0 \\ 2 & -1 & 3 & 5 & |2 \\ 0 & 0 & 2 & 5 & |3 \\ 0 & -7 & 0 & 6 & |6 \end{pmatrix}$$

k=2: Buscamos el elemento de máximo valor absoluto en la columna 2 como si la matriz se hubiera escalado al inicio. Además, buscaremos este pivote en todas las filas salvo en la fila 2, puesto que de ella ya hemos extraído el pivote en el paso anterior. Esto significa que buscamos

$$\max_{i\geq 2} \frac{|a_{p_i,2}|}{s_{p_i}} = \max\left\{\frac{|a_{1,2}|}{s_1}, \frac{|a_{3,2}|}{s_3}, \frac{|a_{4,2}|}{s_4}\right\} = \max\left\{\frac{|7/2|}{7}, \frac{|0|}{5}, \frac{|-7|}{6}\right\} = \frac{|-7|}{6} = \frac{|a_{4,2}|}{s_4} = \frac{|a_{p_4,2}|}{s_{p_4}}$$

Por tanto, deberíamos intercambiar las filas $p_2=1$ y $p_4=4$ entre sí. Intercambiando los valores 1 y 4 en $\underset{\sim}{p}$, obtenemos que el nuevo vector de pivotaje es $\underset{\sim}{p}=\begin{pmatrix}2\\4\\3\\1\end{pmatrix}=\begin{pmatrix}p_1\\p_2\\p_3\\p_4\end{pmatrix}$. El pivote será el elemento $a_{p_2.2}=a_{4,2}$. Las transformaciones a realizar en este paso serán

$$\xrightarrow[\substack{m_{3,2}=0/(-7)\ \ F_3=F_3-m_{3,2}F_4 \\ m_{1,2}=(7/2)/(-7)\ \ F_1=F_1-m_{1,2}F_4}]{} \begin{pmatrix} 0 & 0 & 7/2 & 15/2 & |3 \\ \boxed{2} & -1 & 3 & 5 & |2 \\ 0 & 0 & 2 & 5 & |3 \\ 0 & \boxed{-7} & 0 & 6 & |6 \end{pmatrix}$$

k=3: Buscamos el elemento de máximo valor absoluto en la columna 3 como si la matriz se hubiera escalado al inicio. Además, buscaremos este elemento en las filas $p_3=3$ y $p_4=1$ de las que todavía no hemos extraído ningún pivote. Esto significa que buscamos

$$\max_{i\geq 3} \frac{|a_{p_i,3}|}{s_{p_i}} = \max\left\{\frac{|a_{3,3}|}{s_3}, \frac{|a_{1,3}|}{s_1}\right\} = \max\left\{\frac{|2|}{5}, \frac{|7/2|}{7}\right\} = \frac{1}{2} = \frac{|a_{1,3}|}{s_1} = \frac{|a_{p_4,3}|}{s_{p_4}}$$

Por tanto, deberíamos intercambiar las filas $p_3=3$ y $p_4=1$ entre sí. Intercambiando los valores 3 y 1 en $\underset{\sim}{p}$, obtenemos que el nuevo vector de pivotaje es $\underset{\sim}{p}=\begin{pmatrix}2\\4\\1\\3\end{pmatrix}=\begin{pmatrix}p_1\\p_2\\p_3\\p_4\end{pmatrix}$. El pivote será el elemento $a_{p_3.3}=a_{1,3}$ Las transformaciones a realizar en este paso serán

$$\xrightarrow{m_{3,3}=2/(7/2)=4/7 \quad F_3=F_3-m_{3,3}F_1} \left(\begin{array}{cccc|c} 0 & 0 & \boxed{7/2} & 15/2 & 3 \\ \boxed{2} & -1 & 3 & 5 & 2 \\ 0 & 0 & 0 & 5/7 & 9/7 \\ 0 & \boxed{-7} & 0 & 6 & 6 \end{array}\right)$$

El sistema final resultante después de realizadas las transformaciones es:

$$\begin{cases} & & & (7/2)x_3 & + & (15/2)x_4 & = & 3 \\ 2x_1 & - & x_2 & + \; 3x_3 & + & 5x_4 & = & 2 \\ & & & & & (5/7)x_4 & = & 9/7 \\ & & -7x_2 & & + & 6x_4 & = & 6 \end{cases}$$

Despejando x_4 de la ecuación $3 = p_4$; a continuación x_3 de la ecuación $1 = p_3$, a continuación x_2 de la ecuación $4 = p_2$ y finalmente x_1 de la ecuación $2 = p_1$, obtenemos que la solución del sistema es $x_1 = 47/35;\ x_2 = 24/35; x_3 = -3; x_4 = 9/5$.

Puede observarse que la resolución del sistema final ha seguido el orden regresivo del último vector de pivotaje. Además, reordenando la matriz final del proceso según los valores obtenidos en el último vector de pivotaje, tendremos la matriz triangular superior que define el algoritmo.

Interpretación matricial: Teniendo en cuenta que en este algoritmo de eliminación gaussiana con pivotaje parcial y cambio de escala, las transformaciones elementales que realizamos son del mismo tipo a las que realizábamos en el algoritmo de eliminación gaussiana sin pivotaje, en la interpretación matricial la única diferencia está en el hecho de que ahora debemos tener en cuenta los intercambios de filas que hemos realizado a lo largo del proceso. Por tanto, si llamamos U a la matriz triangular superior que se obtiene al final del proceso de transformaciones y

$$L = \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & 0 & 0 \\ m_{p_2,1} & 1 & 0 & \dots & \dots & 0 & 0 \\ m_{p_3,1} & m_{p_3,2} & 1 & \dots & \dots & 0 & 0 \\ . & . & . & & & . & . \\ . & . & . & & & . & . \\ m_{p_{n-1},1} & m_{p_{n-1},2} & m_{p_{n-1},3} & \dots & \dots & 1 & 0 \\ m_{p_n,1} & m_{p_n,2} & m_{p_n,3} & \dots & \dots & m_{p_n,n-1} & 1 \end{pmatrix} \qquad (14)$$

donde el vector de pivotaje $\underset{\sim}{p}=\begin{pmatrix} p_1 \\ p_2 \\ . \\ . \\ p_n \end{pmatrix}$ sería también el del final del proceso, el producto $L\cdot U$ no será ahora la matriz A, sino la matriz A con sus filas reordenadas según el orden dado por este último vector de pivotaje. Esto puede expresarse considerando la matriz $P=\left(\underset{\sim}{e}_{p_1}\ \underset{\sim}{e}_{p_2}\ ...\underset{\sim}{e}_{p_n}\right)$ y, teniendo en cuenta que la matriz A con sus filas reordenadas según el último vector de pivotaje es realmente la matriz $P^t\cdot A$. Por tanto, podemos decir que

$$L\cdot U=P^t\cdot A$$

Teniendo en cuenta que cada vez que en una matriz se intercambian dos filas entre sí el determinante cambia de signo, podemos concluir que el determinante de la matriz A se calcula como

$$|A|=(-1)^{\text{nº de intercambios}}\ |L\cdot U|=(-1)^{\text{nº de intercambios}}\ |U|$$

Ejemplo: En el ejemplo desarrollado anteriormente, el último vector de pivotaje era $\underset{\sim}{p}=\begin{pmatrix} 2 \\ 4 \\ 1 \\ 3 \end{pmatrix}=\begin{pmatrix} p_1 \\ p_2 \\ p_3 \\ p_4 \end{pmatrix}$, con lo que, reordenando la matriz que se había obtenido al final del proceso según los valores de este vector, tendremos que la matriz U triangular superior es $U=\begin{pmatrix} 2 & -1 & 3 & 5 \\ 0 & -7 & 0 & 6 \\ 0 & 0 & 7/2 & 15/2 \\ 0 & 0 & 0 & 5/7 \end{pmatrix}$

$$|A|=(-1)^3\ |U|=-2\cdot(-7)\cdot(7/2)\cdot 5/7=35$$

$$L=\begin{pmatrix} 1 & 0 & 0 & 0 \\ m_{p_2,1} & 1 & 0 & 0 \\ m_{p_3,1} & m_{p_3,2} & 1 & 0 \\ m_{p_4,1} & m_{p_4,2} & m_{p_4,3} & 1 \end{pmatrix}=\begin{pmatrix} 1 & 0 & 0 & 0 \\ m_{4,1} & 1 & 0 & 0 \\ m_{1,1} & m_{1,2} & 1 & 0 \\ m_{3,1} & m_{3,2} & m_{3,3} & 1 \end{pmatrix}=\begin{pmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 1/2 & -1/2 & 1 & 0 \\ 0 & 0 & 4/7 & 1 \end{pmatrix}$$

$$P=\left(\underset{\sim}{e}_{p_1}\ \underset{\sim}{e}_{p_2}\ \underset{\sim}{e}_{p_3}\ \underset{\sim}{e}_{p_4}\right)=\left(\underset{\sim}{e}_2\ \underset{\sim}{e}_4\ \underset{\sim}{e}_1\ \underset{\sim}{e}_3\right)=\begin{pmatrix}0&0&1&0\\1&0&0&0\\0&0&0&1\\0&1&0&0\end{pmatrix}$$ verificándose que

$$L\cdot U=P^t\cdot A=\begin{pmatrix}2&-1&3&5\\-2&-6&-3&1\\1&3&5&7\\0&0&2&5\end{pmatrix}$$

Observación:

Con este algoritmo de eliminación gaussiana con pivotaje parcial y cambio de escala se solucionan los problemas de estabilidad que aparecían en el algoritmo de eliminación gaussiana planteado al inicio de este tema. Sin embargo, conviene mencionar que existen matrices que también pueden tener problemas de condicionamiento. Las matrices con problemas de condicionamiento son siempre matrices que aun siendo regulares, están muy próximas a una matriz singular. Para este tipo de matrices cualquier algoritmo que se aplique conducirá, en general, a resultados imprecisos.

2.7. MÉTODOS ITERATIVOS PARA RESOLVER SISTEMAS DE ECUACIONES LINEALES

Los métodos iterativos se usan para resolver sistemas lineales con un elevado número de ecuaciones en el caso en que la matriz asociada al sistema sea una matriz dispersa. Esto es debido, a que estos métodos iterativos, a diferencia de los métodos directos estudiados en los apartados anteriores, conservan los ceros de la matriz inicial, por lo que sólo será necesario almacenar en memoria del ordenador los elementos no nulos de la matriz.

Para resolver el sistema lineal (1) $A\cdot\underset{\sim}{x}=\underset{\sim}{b}$ mediante un método iterativo, se transforma dicho sistema en otro equivalente de la forma

$$\underset{\sim}{x}=T\cdot\underset{\sim}{x}+\underset{\sim}{c} \qquad (15)$$

para alguna matriz fija T y para algún vector $\underset{\sim}{c}$. Partiendo de una aproximación inicial $\underset{\sim}{x}^{(0)}$ a la solución del sistema $A\cdot\underset{\sim}{x}=\underset{\sim}{b}$ o a la del sistema equivalente (15), se genera una sucesión de vectores de la forma:

$$\underset{\sim}{x}^{(k+1)} = T \cdot \underset{\sim}{x}^{(k)} + \underset{\sim}{c} \qquad k = 1, 2, ... \tag{16}$$

Si esta sucesión es convergente, es decir, si existe $\lim_{k\to\infty} \underset{\sim}{x}^{(k)} = \underset{\sim}{x}$, el vector límite $\underset{\sim}{x}$ será solución del sistema (15) y consecuentemente del sistema equivalente de partida (1).

En la práctica, por tanto, se construirán unas cuantas aproximaciones al vector solución de (1) utilizando el proceso (16) y nos detendremos cuando se cumpla algún criterio de parada. Ahora bien, el problema estará en determinar si la sucesión generada mediante este proceso es convergente. En este sentido, se tiene un teorema que establece las condiciones bajo las cuales estos métodos generados por una ecuación de la forma (16) son convergentes. Antes de enunciarlo, comenzamos estableciendo la siguiente definición:

Definición. Dada una matriz A cuadrada, se define el *radio espectral de* A y se denota por $\rho(A)$ como $\rho(A) = \max|\lambda|$ siendo λ valor propio de A. Recuérdese que si λ es un autovalor complejo, $\lambda = \alpha + i\beta$, se tiene $|\lambda| = \sqrt{\alpha^2 + \beta^2}$.

Teorema. Dada la ecuación $\underset{\sim}{x} = T \cdot \underset{\sim}{x} + \underset{\sim}{c}$ (15) con $\underset{\sim}{c} \neq \underset{\sim}{0}$, entonces, para cualquier aproximación inicial $\underset{\sim}{x}^{(0)} \in \mathbb{R}^n$ la sucesión $\left\{\underset{\sim}{x}^{(k)}\right\}_{k=0}^{\infty}$ construida mediante el proceso iterativo $\underset{\sim}{x}^{(k+1)} = T \cdot \underset{\sim}{x}^{(k)} + \underset{\sim}{c}$ $k = 1, 2, ...$, converge a la única solución de la ecuación (15) si y sólo si $\rho(T) < 1$.

Además, también se puede demostrar que cuanto menor sea el radio espectral de T, mayor será la velocidad de convergencia.

A continuación, se van a describir dos métodos iterativos clásicos, el ***método de Jacobi*** y el de ***Gauss-Seidel*** y se comentará bajo qué condiciones tales métodos son convergentes. Antes de ello, vamos a ver a través de un ejemplo una posible forma de transformar un sistema lineal $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$ en otro equivalente $\underset{\sim}{x} = T \cdot \underset{\sim}{x} + \underset{\sim}{c}$:

Ejemplo.

Sea el sistema $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$ dado por:

$$\left.\begin{aligned} E_1 &: 10x_1 - x_2 + 2x_3 = 6 \\ E_2 &: -x_1 + 11x_2 - x_3 + 3x_4 = 25 \\ E_3 &: 2x_1 - x_2 + 10x_3 - x_4 = -11 \\ E_4 &: 3x_2 - x_3 + 8x_4 = 15 \end{aligned}\right\}$$ que tiene por solución única $\underset{\sim}{x}=(1,2,-1,1)^t$.

Para transformarlo en uno de la forma $\underset{\sim}{x} = T\cdot\underset{\sim}{x} + \underset{\sim}{c}$, despejamos en cada ecuación E_i la incógnita x_i para i=1, 2, 3, 4; así:

$$\left.\begin{aligned} x_1 &= \frac{1}{10}(6 + x_2 - 2x_3) \\ x_2 &= \frac{1}{11}(25 + x_1 + x_3 - 3x_4) \\ x_3 &= \frac{1}{10}(-11 - 2x_1 + x_2 + x_4) \\ x_4 &= \frac{1}{8}(15 - 3x_2 + x_3) \end{aligned}\right\}$$

con lo que este sistema puede escribirse, ya, en la forma $\underset{\sim}{x} = T\cdot\underset{\sim}{x} + \underset{\sim}{c}$ siendo:

$$T = \begin{pmatrix} 0 & \frac{1}{10} & -\frac{1}{5} & 0 \\ \frac{1}{11} & 0 & \frac{1}{11} & -\frac{3}{11} \\ -\frac{1}{5} & \frac{1}{10} & 0 & \frac{1}{10} \\ 0 & -\frac{3}{8} & \frac{1}{8} & 0 \end{pmatrix} \quad \text{y} \quad \underset{\sim}{c} = \begin{pmatrix} \frac{3}{5} \\ \frac{25}{11} \\ -\frac{11}{10} \\ \frac{15}{8} \end{pmatrix}$$

Entonces, operando con redondeo a 4 decimales y partiendo de una aproximación inicial, por ejemplo $\underset{\sim}{x}^{(0)} = (0,0,0,0)^t$, construimos un vector $\underset{\sim}{x}^{(1)}$ a partir de las ecuaciones anteriores en la forma:

$$\left.\begin{aligned} x_1^{(1)} &= \frac{1}{10}\left(6 + x_2^{(0)} - 2x_3^{(0)}\right) = \frac{6}{10} = 0.6 \\ x_2^{(1)} &= \frac{1}{11}\left(25 + x_1^{(0)} + x_3^{(0)} - 3x_4^{(0)}\right) = \frac{25}{11} \simeq 2.2727 \\ x_3^{(1)} &= \frac{1}{10}\left(-11 - 2x_1^{(0)} + x_2^{(0)} + x_4^{(0)}\right) = -\frac{11}{10} = -1.1 \\ x_4^{(1)} &= \frac{1}{8}\left(15 - 3x_2^{(0)} + x_3^{(0)}\right) = \frac{15}{8} = 1.8750 \end{aligned}\right\}$$

$$\rightarrow \underset{\sim}{x}^{(1)} = (0.6,\ 2.2727,\ -1.1,\ 1.8750)^t$$

Repitiendo este proceso (operando siempre con redondeo a 4 decimales), se van obteniendo sucesivas aproximaciones $\underset{\sim}{x}^{(k)} = (x_1^{(k)},\ x_2^{(k)},\ x_3^{(k)},\ x_4^{(k)})^t$ en la misma forma:

$$
\left.\begin{aligned}
x_1^{(2)} &= \frac{1}{10}\left(6 + \quad x_2^{(1)} - 2x_3^{(1)}\right) = \frac{1}{10}\left(6 + 2.2727 - 2\cdot(-1.1)\right) \simeq 1.0473 \\
x_2^{(2)} &= \frac{1}{11}\left(25 + x_1^{(1)} + \quad x_3^{(1)} - 3x_4^{(1)}\right) = \frac{1}{11}\left(25 + 0.6 + (-1.1) - 3\cdot 1.875\right) \simeq 1.7159 \\
x_3^{(2)} &= \frac{1}{10}\left(-11 - 2x_1^{(1)} + x_2^{(1)} + \quad x_4^{(1)}\right) = \frac{1}{10}\left(-11 - 2\cdot 0.6 + 2.2727 + 1.875\right) \simeq -0.8052 \\
x_4^{(2)} &= \frac{1}{8}\left(15 - \quad 3x_2^{(1)} + x_3^{(1)}\right) = \frac{1}{8}\left(15 - 3\cdot 2.2727 + (-1.1)\right) \simeq 0.8852
\end{aligned}\right\}
$$

$$
\rightarrow \underset{\sim}{x}^{(2)} = (1.0473,\ 1.7159,\ -0.8052,\ 0.8852)^t
$$

Se parará el proceso cuando se cumpla algún cierto criterio de parada. Algunas aproximaciones más se recogen en la tabla siguiente:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$
0	0	0	0	0
1	0.6000	2.2727	-1.1000	1.8750
2	1.0473	1.7159	-0.8052	0.8852
3	0.9326	2.0533	-1.0493	1.1309
4	1.0152	1.9537	-0.9681	0.9739
5	0.9890	2.0114	-1.0103	1.0214
6	1.0032	1.9922	-0.9945	0.9944
7	0.9981	2.0023	-1.0020	1.0036
8	1.0006	1.9987	-0.9990	0.9989
9	0.9997	2.0004	-1.0004	1.0006
10	1.0001	1.9998	-0.9998	0.9998

Este método se conoce como método iterativo de Jacobi y se describe a continuación.

2.7.1. Método de Jacobi

El método de Jacobi consiste en despejar la variable x_i de la i-ésima ecuación del sistema (1) $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$, a condición de que $a_{ii} \neq 0$. El sistema (1) queda, por tanto, en la forma:

$$\left.\begin{aligned} x_1 &= \frac{1}{a_{1,1}}\left(b_1 - a_{1,2}x_2 - a_{1,3}x_3 - \ldots - a_{1,n}x_n\right) \\ x_2 &= \frac{1}{a_{2,2}}\left(b_2 - a_{2,1}x_1 - a_{2,3}x_3 - \ldots - a_{2,n}x_n\right) \\ &\ldots\ldots\ldots\ldots\ldots\ldots \\ x_n &= \frac{1}{a_{n,n}}\left(b_n - a_{n,1}x_1 - a_{n,2}x_2 - \ldots - a_{n,n-1}x_{n-1}\right) \end{aligned}\right\} \Rightarrow x_i = \frac{1}{a_{i,i}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{i,j}x_j\right) \quad i = 1,\ldots n \quad (17)$$

Estas ecuaciones pueden escribirse, ya, en la forma $\underset{\sim}{x} = T \cdot \underset{\sim}{x} + \underset{\sim}{c}$ siendo:

$$T = \begin{pmatrix} 0 & \frac{-a_{1,2}}{a_{1,1}} & \frac{-a_{1,3}}{a_{1,1}} & \ldots & \frac{-a_{1,n-1}}{a_{1,1}} & \frac{-a_{1,n}}{a_{1,1}} \\ \frac{-a_{2,1}}{a_{2,2}} & 0 & \frac{-a_{2,3}}{a_{2,2}} & \ldots & \frac{-a_{2,n-1}}{a_{2,2}} & \frac{-a_{2,n}}{a_{2,2}} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ \frac{-a_{n,1}}{a_{n,n}} & \frac{-a_{n,2}}{a_{n,n}} & \ldots & \ldots & \frac{-a_{n,n-1}}{a_{n,n}} & 0 \end{pmatrix} \quad y \quad \underset{\sim}{c} = \begin{pmatrix} \frac{b_1}{a_{1,1}} \\ \frac{b_2}{a_{2,2}} \\ . \\ . \\ . \\ \frac{b_n}{a_{n,n}} \end{pmatrix} \quad (18)$$

es decir, $T = (t_{i,j})$ con $t_{i,j} = \begin{cases} 0 & \text{si } i = j \\ -\frac{a_{i,j}}{a_{i,i}} & \text{si } i \neq j \end{cases}$

De esta forma se pueden ir obteniendo los valores de las distintas incógnitas x_i i=1,2,..n en iteraciones sucesivas, mediante la expresión general (17), partiendo de una aproximación inicial $\underset{\sim}{x}^{(0)}$, esto es:

$$x_i^{(k+1)} = \frac{1}{a_{i,i}}\left(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{i,j}x_j^{(k)}\right) \quad k = 0,1,2,\ldots \quad (19)$$

Criterio de parada: Si esta sucesión $\left\{\underset{\sim}{x}^{(k)}\right\}_{k=0}^{\infty}$ converge, a partir de un cierto k los términos de la misma estarán muy próximos a la solución y además muy próximos entre sí, por lo que en la práctica, el error absoluto del paso k $e_{abs} = \| \underset{\sim}{x} - \underset{\sim}{x}^{(k)} \|$ se aproximará por la diferencia entre dos iteraciones consecutivas, es decir

$$e_{abs} \simeq \| \underset{\sim}{x}^{(k+1)} - \underset{\sim}{x}^{(k)} \| \quad (20)$$

De manera análoga, el error relativo del paso k se aproximará por

$$e_{rel} \simeq \frac{\| \underset{\sim}{x}^{(k+1)} - \underset{\sim}{x}^{(k)} \|}{\| \underset{\sim}{x}^{(k+1)} \|} \tag{21}$$

Para detener el proceso suele fijarse el error máximo permitido ε y parar de construir aproximaciones cuando se cumpla que:

$$\| \underset{\sim}{x}^{(k+1)} - \underset{\sim}{x}^{(k)} \| \leq \varepsilon \quad \text{o bien} \quad \frac{\| \underset{\sim}{x}^{(k+1)} - \underset{\sim}{x}^{(k)} \|}{\| \underset{\sim}{x}^{(k+1)} \|} \leq \varepsilon$$

En este problema consideraremos la norma del supremo ($\| \underset{\sim}{x} \| = \max\{|x_1|, \ldots |x_n|\}$),ya que

$$\| \underset{\sim}{x}^{(k+1)} - \underset{\sim}{x}^{(k)} \|_\infty = \max_i \{ |x_i^{(k+1)} - x_i^{(k)}| \} \leq \varepsilon \Leftrightarrow |x_i^{(k+1)} - x_i^{(k)}| < \varepsilon \quad \forall i = 1, 2, \ldots, n$$

aunque podría utilizarse cualquier otra norma.

Observaciones:

1) A veces este método suele ser bastante lento, por lo que es conveniente fijar un número máximo de iteraciones a realizar y, si al cabo de ese número de iteraciones no se ha alcanzado la precisión deseada, se concluirá que el método no es eficaz y se utilizará otro.
2) Este método requiere, según se ha visto, que $a_{i,i} \neq 0 \ \forall \ i = 1, \ldots n$. Si éste no es el caso, siempre se puede realizar un reordenamiento de las ecuaciones del sistema de forma que ningún elemento de la diagonal principal sea nulo (lo cual siempre se puede conseguir si el sistema tiene solución).

Ejemplo:

Considérese el sistema de ecuaciones:

$$\begin{cases} 3x - y + z = 4 \\ 2x + 5y - 2z = -6 \\ x - y - 3z = 6 \end{cases}$$ que tiene por solución única x = 1, y = -2, z = -1.

La aplicación del método de Jacobi, según acabamos de ver, consiste en despejar la x, y, z de cada una de las ecuaciones:

$$\begin{cases} x = \frac{1}{3}(4 + y - z) \\ y = \frac{1}{5}(-6 - 2x + 2z) \\ z = -\frac{1}{3}(6 - x + y) \end{cases}$$

Por tanto, las sucesivas aproximaciones se obtienen mediante el proceso:

$$\begin{cases} x^{(k+1)} = \frac{1}{3}(4 + y^{(k)} - z^{(k)}) \\ y^{(k+1)} = \frac{1}{5}(-6 - 2x^{(k)} + 2z^{(k)}) \quad k = 0,1,... \\ z^{(k+1)} = -\frac{1}{3}(6 - x^{(k)} + y^{(k)}) \end{cases}$$

Operando ahora con redondeo a 6 dígitos significativos y comenzando con una aproximación inicial, por ejemplo $\underset{\sim}{x}^{(0)}=(0,0,0)^t$, obtenemos un nuevo vector $\underset{\sim}{x}^{(1)}$ sustituyendo los valores en la expresión anterior:

$$\begin{cases} x^{(1)} = \frac{4}{3} \simeq 1.33333 \\ y^{(1)} = -\frac{6}{5} = -1.2 \quad \rightarrow \underset{\sim}{x}^{(1)} = \begin{pmatrix} 1.33333 \\ -1.2 \\ -2 \end{pmatrix} \\ z^{(1)} = -2 \end{cases}$$

El error relativo cometido en este paso es:

$$e_{rel}^{(1)} = \frac{\| \underset{\sim}{x}^{(1)} - \underset{\sim}{x}^{(0)} \|_\infty}{\| \underset{\sim}{x}^{(1)} \|_\infty} = \frac{\max\{|1.33333|,|-1.2|,|2|\}}{\max\{|1.33333|,|-1.2|,|2|\}} = 1$$

A continuación calculamos $\underset{\sim}{x}^{(2)}$:

$$\begin{cases} x^{(2)} = \frac{1}{3}(4 + y^{(1)} - z^{(1)}) = \frac{1}{3}(4 - 1.2 + 2) = 1.6 \\ y^{(2)} = \frac{1}{5}(-6 - 2x^{(1)} + 2z^{(1)}) = \frac{2}{5}(-6 - 2 \cdot 1.33333 - 4) = -2.53333 \quad \rightarrow \underset{\sim}{x}^{(2)} = \begin{pmatrix} 1.6 \\ -2.53333 \\ -1.15556 \end{pmatrix} \\ z^{(2)} = -\frac{1}{3}(6 - 2x^{(1)} + y^{(1)}) = -\frac{1}{3}(6 - 2 \cdot 1.33333 - 1.2) = -1.15556 \end{cases}$$

El error relativo cometido en este paso es:

$$e_{rel}^{(2)} = \frac{\| \underset{\sim}{x}^{(2)} - \underset{\sim}{x}^{(1)} \|_\infty}{\| \underset{\sim}{x}^{(2)} \|_\infty} = \frac{\max\{|0.2667|,|-1.33333|,|0.84444|\}}{\max\{|1.6|,|-2.53333|,|-1.15536|\}} = \frac{1.33333}{2.53333} = 0.526315$$

De esta forma continuaríamos calculando aproximaciones sucesivas a la solución hasta que el error relativo fuera menor que una cierta cantidad ε. En la siguiente tabla aparecen representadas algunas iteraciones más:

k	$\mathbf{x^{(k)}}$	$\mathbf{y^{(k)}}$	$\mathbf{z^{(k)}}$	$\mathbf{e_{rel}^{(k)}}$
0	0	0	0	
1	1.33333	-1.2	-2	1
2	1.6	-2.53333	-1.15556	0.526315
3	0.874077	-2.30222	-0.622223	0.315314
4	0.773334	-1.79852	-0.941234	0.280064
…	…	…	…	…
27	0.999999	-2	-0.999997	3.06669×10^{-6}

2.7.2. Método de Gauss-Seidel

Un análisis de las ecuaciones del método de Jacobi sugiere una posible mejora del algoritmo. En el método de Jacobi para calcular $x_i^{(k+1)}$ se utilizan los valores de todas las incógnitas de la etapa anterior. Ahora bien, en el **método de Gauss-Seidel** se pretende realizar la siguiente mejora:

Para calcular la primera componente del paso k+1 $x_1^{(k+1)}$, sólo se disponen de los valores de todas las componentes de la etapa k, y son las que se utilizarán. Sin embargo, para calcular la segunda componente $x_2^{(k+1)}$ ya se conoce el valor de $x_1^{(k+1)}$, de modo que si el proceso es convergente, cabe pensar que será una mejor aproximación a la solución final que $x_1^{(k)}$ y por tanto, se utilizará ya tal valor $x_1^{(k+1)}$. Así sucesivamente, para calcular $x_3^{(k+1)}$ se emplearán $x_1^{(k+1)}$ y $x_2^{(k+1)}$. En general, para calcular la componente i-ésima $x_i^{(k+1)}$ con i>1 se utilizarán $x_1^{(k+1)}$, …., $x_{i-1}^{(k+1)}$, que si hay convergencia, deberían ser ya mejores aproximaciones a las componentes $x_1,\ldots,x_{i-1}$ de la solución del sistema. Según este nuevo planteamiento, el algoritmo queda:

Para i=1, 2,…,n:

$$x_i^{(k+1)} = \frac{1}{a_{i,i}}\left(b_i - \sum_{j=1}^{i-1} a_{i,j} x_j^{(k+1)} - \sum_{j=1+i}^{n} a_{i,j} x_j^{(k)} \right) \quad k = 0,1,2,... \tag{22}$$

El criterio de parada es el mismo que en el método de Jacobi.

Ejemplo:

El sistema $\begin{cases} 10x_1 - x_2 + 2x_3 = 6 \\ -x_1 + 11x_2 - x_3 + 3x_4 = 25 \\ 2x_1 - x_2 + 10x_3 - x_4 = -11 \\ 3x_2 - x_3 + 8x_4 = 15 \end{cases}$ se resolvió anteriormente por el método de Jacobi. Al incorporar ahora el algoritmo (22), se obtienen las ecuaciones:

$$\left.\begin{aligned} x_1^{(k+1)} &= \frac{1}{10}\left(6 + x_2^{(k)} - 2x_3^{(k)}\right) \\ x_2^{(k+1)} &= \frac{1}{11}\left(25 + x_1^{(k+1)} + x_3^{(k)} - 3x_4^{(k)}\right) \\ x_3^{(k+1)} &= \frac{1}{10}\left(-11 - 2x_1^{(k+1)} + x_2^{(k+1)} + x_4^{(k)}\right) \\ x_4^{(k+1)} &= \frac{1}{8}\left(15 - 3x_2^{(k+1)} + x_3^{(k+1)}\right) \end{aligned}\right\} \; k = 0,1,2,\ldots$$

Tomando como primera aproximación $\underset{\sim}{x}^{(0)} = (0,0,0,0)^t$ y operando de nuevo con redondeo a 4 decimales, generamos las iteraciones:

$$\left.\begin{aligned} x_1^{(1)} &= \frac{1}{10}\left(6 + x_2^{(0)} - 2x_3^{(0)}\right) = \frac{6}{10} = 0.6 \\ x_2^{(1)} &= \frac{1}{11}\left(25 + x_1^{(1)} + x_3^{(0)} - 3x_4^{(0)}\right) = \frac{1}{11}(25 + 0.6) \simeq 2.3273 \\ x_3^{(1)} &= \frac{1}{10}\left(-11 - 2x_1^{(1)} + x_2^{(1)} + x_4^{(0)}\right) = \frac{1}{10}(-11 - 2\cdot 0.6 + 2.3273) \simeq -0.9873 \\ x_4^{(1)} &= \frac{1}{8}\left(15 - 3x_2^{(1)} + x_3^{(1)}\right) = \frac{1}{8}(15 - 3\cdot 2.3273 - 0.9873) \simeq 0.8789 \end{aligned}\right\}$$

$$\rightarrow \underset{\sim}{x}^{(1)} = (0.6,\ 2.3273,\ -0.9873,\ 0.8789)^t$$

Así sucesivamente, se van construyendo sucesivas aproximaciones $\underset{\sim}{x}^{(k)} = (x_1^{(k)},\ x_2^{(k)},\ x_3^{(k)},\ x_4^{(k)})^t$ de la misma forma:

$$\left.\begin{aligned} x_1^{(2)} &= \frac{1}{10}\left(6 + x_2^{(1)} - 2x_3^{(1)}\right) = \frac{1}{10}\left(6 + 2.3273 - 2\cdot(-0.9873)\right) = 1.0302 \\ x_2^{(2)} &= \frac{1}{11}\left(25 + x_1^{(2)} + x_3^{(1)} - 3x_4^{(1)}\right) = \frac{1}{11}\left(25 + 1.0302 + (-0.9873) - 3\cdot 0.8789\right) = 2.0369 \\ x_3^{(2)} &= \frac{1}{10}\left(-11 - 2x_1^{(2)} + x_2^{(2)} + x_4^{(1)}\right) = \frac{1}{10}\left(-11 - 2\cdot 1.0302 + 2.0369 + 0.8789\right) = -1.0145 \\ x_4^{(2)} &= \frac{1}{8}\left(15 - 3x_2^{(2)} + x_3^{(2)}\right) = \frac{1}{8}\left(15 - 3\cdot 2.0369 + (-1.0145)\right) = 0.9844 \end{aligned}\right\}$$

$$\rightarrow \underset{\sim}{x}^{(1)} = \left(1.0302,\ 2.0369,\ -1.0145,\ 0.9844\right)^t$$

En la siguiente tabla se recogen algunas aproximaciones más:

k	$x_1^{(k)}$	$x_2^{(k)}$	$x_3^{(k)}$	$x_4^{(k)}$
0	0	0	0	0
1	0.6000	2.3273	-0.9873	08789
2	1.0302	2.0369	-1.0145	0.9844
3	1.0066	2.0035	-1.0025	0.9984
4	1.0009	2.0003	-1.0003	0.9999
5	1.0001	2.0000	-1.0000	1.0000

Puede observarse que la aproximación $\underset{\sim}{x}^{(5)}$ verifica que $\frac{\| \underset{\sim}{x}^{(5)} - \underset{\sim}{x}^{(4)} \|_\infty}{\| \underset{\sim}{x}^{(5)} \|_\infty} = \frac{0.0008}{2.0000} = 4 \cdot 10^{-4}$, es decir, si tomamos como aproximación a la solución el vector $x^{(5)} = (1.0001,\ 2.0000, -1.0000, 1.0000)^t$, estaremos cometiendo un error relativo del orden de $4x10^{-4}$. Para este mismo ejemplo, se calcularon también 10 iteraciones empleando el método de Jacobi y puede observarse que la aproximación $\underset{\sim}{x}^{(10)}$ obtenida por ese procedimiento cumple que $\frac{\| \underset{\sim}{x}^{(10)} - \underset{\sim}{x}^{(9)} \|_\infty}{\| \underset{\sim}{x}^{(10)} \|_\infty} = \frac{8 \cdot 10^{-4}}{1.9998} = 4.0004 \cdot 10^{-4}$. Nótese que para alcanzar el mismo grado de exactitud el método de Jacobi ha requerido el doble número de iteraciones.

Expresión matricial del método de Gauss-Seidel

Se trata de transformar el sistema lineal $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$ en el sistema equivalente $\underset{\sim}{x} = T \cdot \underset{\sim}{x} + \underset{\sim}{c}$ usando el método de Gauss-Seidel.

Se multiplican ambos miembros de las ecuaciones (22) del algoritmo por a_{ii} y se dejan en cada lado de la ecuación las componentes correspondientes a una misma etapa, es decir,

$$\left.\begin{array}{ll}
a_{1,1}x_1^{(k+1)} & = b_1 - a_{1,2}x_2^{(k)} - \ a_{1,3}x_3^{(k)} - \ldots - a_{1,n}x_n^{(k)} \\
a_{2,1}x_1^{(k+1)} + a_{2,2}x_2^{(k+1)} & = b_2 - a_{2,3}x_3^{(k)} - \ldots - a_{2,n}x_n^{(k)} \\
a_{3,1}x_1^{(k+1)} + a_{3,2}x_2^{(k+1)} + a_{3,3}x_3^{(k+1)} & = b_3 - a_{3,4}x_4^{(k)} - \ldots - a_{3,n}x_n^{(k)} \\
\ldots\ldots\ldots\ldots\ldots\ldots & \\
a_{n-1,1}x_1^{(k+1)} + a_{n-1,2}x_2^{(k+1)} + \ldots + a_{n-1,n-1}x_{n-1}^{(k+1)} & = b_{n-1} - a_{n-1,n}x_n^{(k)} \\
a_{n,1}x_1^{(k+1)} + a_{n,2}x_2^{(k+1)} + \ldots + a_{n,n-1}x_{n-1}^{(k+1)} + a_{n,n}x_n^{(k+1)} & = b_n
\end{array}\right\}$$

lo que expresado matricialmente queda en la forma:

$$\begin{pmatrix} a_{1,1} & 0 & \ldots & \ldots & 0 & 0 \\ a_{2,1} & a_{2,2} & 0 & \ldots & 0 & 0 \\ a_{3,1} & a_{3,2} & a_{3,3} & 0 & \ldots & 0 \\ . & . & . & . & . & . \\ a_{n-1,1} & a_{n-1,2} & \ldots & \ldots & a_{n-1,n-1} & 0 \\ a_{n,1} & a_{n,2} & \ldots & \ldots & a_{n,n-1} & a_{n,n} \end{pmatrix} \cdot \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ . \\ . \\ x_{n-1}^{(k+1)} \\ x_n^{(k+1)} \end{pmatrix} =$$

$$= \begin{pmatrix} 0 & -a_{1,2} & -a_{1,3} & \ldots & -a_{1,n-1} & -a_{1,n} \\ 0 & 0 & -a_{2,3} & \ldots & -a_{2,n-1} & -a_{2,n} \\ 0 & 0 & 0 & -a_{3,4} & \ldots & -a_{3,n} \\ . & . & . & . & . & . \\ 0 & 0 & \ldots & \ldots & 0 & -a_{n-1,n} \\ 0 & 0 & \ldots & \ldots & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ . \\ . \\ x_{n-1}^{(k)} \\ x_n^{(k)} \end{pmatrix} + \begin{pmatrix} b_1 \\ b_2 \\ . \\ . \\ b_{n-1} \\ b_n \end{pmatrix} \qquad (23)$$

Si descomponemos la matriz original A de coeficientes del sistema como A=D+L+U siendo:

$$D = \begin{pmatrix} a_{1,1} & 0 & 0 & \ldots & 0 \\ 0 & a_{2,2} & 0 & \ldots & 0 \\ 0 & . & . & . & . \\ . & . & . & . & . \\ 0 & 0 & 0 & \ldots & a_{n,n} \end{pmatrix}, L = \begin{pmatrix} 0 & 0 & 0 & \ldots & 0 \\ a_{2,1} & 0 & 0 & \ldots & 0 \\ a_{3,1} & a_{3,2} & 0 & . & . \\ . & . & . & . & . \\ a_{n,1} & a_{n,2} & \ldots & a_{n,n-1} & 0 \end{pmatrix}, U = \begin{pmatrix} 0 & a_{1,2} & a_{1,3} & \ldots & a_{1,n} \\ 0 & 0 & a_{2,3} & \ldots & a_{2,n} \\ . & . & . & . & . \\ 0 & 0 & . & 0 & a_{n-1,n} \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}$$

La relación matricial (23) puede expresarse en la forma:

$$(L + D) \cdot \underset{\sim}{x}^{(k+1)} = -U \cdot \underset{\sim}{x}^{(k)} + \underset{\sim}{b}$$

y despejando $\underset{\sim}{x}^{(k+1)}$, se obtiene:

$$\underset{\sim}{x}^{(k+1)} = -(L + D)^{-1} \cdot U \cdot \underset{\sim}{x}^{(k)} + (L + D)^{-1} \cdot \underset{\sim}{b} \Leftrightarrow \underset{\sim}{x}^{(k+1)} = T \cdot \underset{\sim}{x}^{(k)} + \underset{\sim}{c} \qquad (24)$$

es decir, tenemos ya el sistema $\underset{\sim}{x} = T \cdot \underset{\sim}{x} + \underset{\sim}{c}$ equivalente al original $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$, siendo la matriz $T = -(L + D)^{-1} \cdot U$ y el vector $\underset{\sim}{c} = (L + D)^{-1} \cdot \underset{\sim}{b}$.

2.7.3. Estudio de la convergencia de los métodos de Jacobi y Gauss-Seidel

El teorema enunciado al comienzo del apartado 2.7 establece como condición necesaria y suficiente para que estos métodos iterativos converjan, que el radio espectral de la matriz T que define el método sea menor que uno, siendo, además, la velocidad de convergencia tanto mayor cuanto menor sea el radio espectral. Ahora bien, verificar esta condición para las matrices T_J y T_G asociadas a los métodos de Jacobi y Gauss-Seidel no siempre es una tarea sencilla. En consecuencia, se van a dar a continuación algunas condiciones suficientes para la convergencia de estos métodos, más sencillas de verificar.

Teorema. Si la matriz A es estrictamente diagonal dominante o puede serlo bajo alguna reordenación, entonces con cualquier elección $\underset{\sim}{x}^{(0)} \in \mathbb{R}^n$, tanto el método de Jacobi como el de Gauss-Seidel dan sucesiones $\left\{\underset{\sim}{x}^{(k)}\right\}_{k=0}^{\infty}$ que convergen a la única solución del sistema $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$.

Teorema. Si la matriz A es simétrica definida positiva, entonces para cualquier elección $\underset{\sim}{x}^{(0)} \in \mathbb{R}^n$, el método de Gauss-Seidel da lugar a una sucesión $\left\{\underset{\sim}{x}^{(k)}\right\}_{k=0}^{\infty}$ que converge a la única solución del sistema $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$.

En general, no existen resultados generales que nos digan cuál de los dos métodos, si el de Jacobi o el de Gauss-Seidel, será el más eficaz para un sistema lineal arbitrario (hay sistemas donde el método de Jacobi converge y el de Gauss-Seidel no y viceversa). Sin embargo, en algunos casos especiales sí se conoce cuál de los dos métodos funciona mejor, ya que se cumple el siguiente teorema:

Teorema de Stein-Rosenberg. Sea el sistema $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$. Si la matriz A cumple que sus elementos $a_{i,j} \le 0 \ \forall\ i \neq j$ y $a_{i,i} > 0 \ \forall\ i = 1,\ldots n$, entonces se satisface una y sólo una de las siguientes afirmaciones:

a) $0 < \rho(T_G) < \rho(T_J) < 1$

b) $1 < \rho(T_J) < \rho(T_G)$

c) $\rho(T_J) = \rho(T_G) = 1$

d) $\rho(T_J) = \rho(T_G) = 0$

donde T_G y T_J denotan las matrices de paso T de los métodos de Gauss-Seidel y de Jacobi, respectivamente.

Observación:

Este teorema establece que, en las condiciones enunciadas, los dos métodos convergen o divergen a la vez, (convergen si se satisface a) o d) y divergen si se verifica b) o c)). Además, cuando ambos convergen, el método de Gauss-Seidel lo hace más rápidamente pues el radio espectral de T_G es menor que el de T_J. La situación de este teorema es frecuente en las matrices que surgen al resolver ecuaciones en derivadas parciales por métodos numéricos.

Otro resultado, en este sentido, que también se verifica es el siguiente:

Teorema. Si la matriz A es simétrica, definida positiva y tridiagonal, entonces se cumple:

$$\rho(T_G) = \left(\rho(T_J)\right)^2 < 1$$

Ejemplo 1:

El sistema $A \cdot \underset{\sim}{x} = \underset{\sim}{b}$ dado por:

$$\left.\begin{array}{l} E_1 : 2x_1 - x_2 + 10x_3 - x_4 = -11 \\ E_2 : 10x_1 - x_2 + 2x_3 = 6 \\ E_3 : -x_1 + 11x_2 - x_3 + 3x_4 = 25 \\ E_4 : 3x_2 - x_3 + 8x_4 = 15 \end{array}\right\} \text{ tiene por matriz asociada } A = \begin{pmatrix} 2 & -1 & 10 & -1 \\ 10 & -1 & 2 & 0 \\ -1 & 11 & -1 & 3 \\ 0 & 3 & -1 & 8 \end{pmatrix}$$

que no es ni estrictamente diagonal dominante, ni simétrica definida positiva, por lo que no cumple ninguna de las condiciones suficientes para tener garantizada la convergencia de los métodos de Jacobi o Gauss-Seidel. (De hecho, ambos métodos divergen, porque se puede demostrar que tanto $\rho(T_G)$ como $\rho(T_J)$ son mayores que uno). Ahora bien, si reordenamos las ecuaciones poniendo la E_2 la primera, la E_3 la segunda y la E_1 la tercera, obtenemos el sistema:

$$\left.\begin{array}{l} E_1 : 10x_1 - x_2 + 2x_3 = 6 \\ E_2 : -x_1 + 11x_2 - x_3 + 3x_4 = 25 \\ E_3 : 2x_1 - x_2 + 10x_3 - x_4 = -11 \\ E_4 : 3x_2 - x_3 + 8x_4 = 15 \end{array}\right\} \text{ cuya matriz asociada es } A = \begin{pmatrix} 10 & -1 & 2 & 0 \\ -1 & 11 & -1 & 3 \\ 2 & -1 & 10 & -1 \\ 0 & 3 & -1 & 8 \end{pmatrix} \text{ que}$$

es estrictamente diagonal dominante y también simétrica definida positiva, por tanto, convergerán tanto el método de Jacobi como el de Gauss-Seidel.(Nótese que este es el ejemplo que se ha resuelto por ambos métodos en los apartados anteriores).

Ejemplo 2:

El sistema lineal $\left.\begin{array}{l} x_1 + 2x_2 - 2x_3 = 7 \\ x_1 + x_2 + x_3 = 2 \\ 2x_1 + 2x_2 + x_3 = 5 \end{array}\right\}$ tiene por solución $(1, 2,-1)^t$. Su matriz asociada

$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}$ no es ni estrictamente diagonal dominante ni simétrica definida positiva, por lo que no se cumplen las condiciones suficientes de convergencia anteriormente vistas. Sin embargo, si calculamos la matriz T_J asociada al método de Jacobi

$$T_J = \begin{pmatrix} 0 & -\frac{a_{1,2}}{a_{1,1}} & -\frac{a_{1,3}}{a_{1,1}} \\ -\frac{a_{2,1}}{a_{2,2}} & 0 & -\frac{a_{2,3}}{a_{2,2}} \\ -\frac{a_{3,1}}{a_{3,3}} & -\frac{a_{3,2}}{a_{3,3}} & 0 \end{pmatrix} = \begin{pmatrix} 0 & -2 & 2 \\ -1 & 0 & -1 \\ -2 & -2 & 0 \end{pmatrix}$$

puede observarse que el polinomio característico de T_J es $|T_J - \lambda I| = -\lambda^3 \Rightarrow \lambda = 0$ es autovalor triple de esta matriz, por lo que su radio espectral $\rho(T_J) = 0 < 1$, y por tanto, el método de Jacobi convergerá a la única solución $(1, 2,-1)^t$ del sistema y además lo hará muy rápidamente.

En cambio, si calculamos la matriz T_G asociada al método de Gauss-Seidel, que según hemos visto es $T_G = -(L + D)^{-1} \cdot U$ siendo

$$L = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 2 & 2 & 0 \end{pmatrix} D = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \text{ y } U = \begin{pmatrix} 0 & 2 & -2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \Rightarrow$$

$$L + D = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix} \Rightarrow (L + D)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -2 & 1 \end{pmatrix} \Rightarrow$$

$$T_G = -\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -2 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 & 2 & -2 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{pmatrix}$$

La matriz T_G es una matriz triangular superior, por lo que sus valores propios son $\lambda_1 = 0$ $\lambda_2 = 2$ (doble); su radio espectral es ahora $\rho(T_G) = 2 > 1$, y por lo tanto, el método de Gauss-Seidel será divergente en este caso.

2.8. CÁLCULO DEL VALOR PROPIO DOMINANTE DE UNA MATRIZ. MÉTODO DE LAS POTENCIAS

Según acabamos de ver, para saber si un método iterativo converge hay que calcular el radio espectral de su matriz T asociada, es decir, habrá que determinar el valor propio de T de mayor módulo. Sabemos, por los resultados del Álgebra, que los valores propios de una matriz T son las raíces de su polinomio característico $p(\lambda) = |T - \lambda I| = 0$; ahora bien, en la práctica los autovalores no se obtienen mediante el cálculo directo de las raíces de tal polinomio característico. Esto es debido, por un lado, a que si el orden de la matriz es elevado, el coste operativo de obtener el determinante $|T - \lambda I|$ es muy alto. Por otro lado, también resulta difícil hallar buenas aproximaciones a las raíces de una ecuación polinómica, ya que se trata de un problema mal condicionado, es decir, pequeños errores en los coeficientes del polinomio pueden dar lugar a grandes variaciones en algunas de sus soluciones. En consecuencia, para calcular el valor propio dominante o de mayor módulo de una matriz de orden elevado, que son para las que se emplean principalmente los métodos iterativos, se utiliza, normalmente, el denominado **método de las potencias**, que es una técnica iterativa y que se va a describir a continuación.

Supondremos una matriz A real de orden n, con n valores propios λ_1, λ_2,…, λ_n y tal que A tiene exactamente un autovalor λ_1 de módulo máximo, es decir, $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n| \geq 0$. Supondremos además que asociado a este conjunto de autovalores existe un conjunto de n vectores propios linealmente independientes $\{\underset{\sim}{v}_1, \underset{\sim}{v}_2, ..., \underset{\sim}{v}_n\}$, y que por tanto verifican $A \cdot \underset{\sim}{v}_i = \lambda_i \cdot \underset{\sim}{v}_i \quad \forall\, i = 1, ..., n$. El **método de las potencias**, a partir de un vector $\underset{\sim}{x}^{(0)} \in \mathbb{R}^n$, permite construir una sucesión de vectores $\left\{\underset{\sim}{x}^{(k)}\right\}_{k=1}^{\infty}$ que converge al vector propio asociado al autovalor dominante de A, λ_1. El **algoritmo** es el siguiente:

Se empieza eligiendo un vector $\underset{\sim}{x}^{(0)} \in \mathbb{R}^n$ cuya norma del supremo sea 1, es decir, con $\| \underset{\sim}{x}^{(0)} \|_\infty = 1$. A continuación, en el primer paso del algoritmo, k=1, se realizan las siguientes operaciones:

Se calcula el vector $\underset{\sim}{y}^{(1)} = A \cdot \underset{\sim}{x}^{(0)}$.

Se calcula la componente de $\underset{\sim}{y}^{(1)}$ que conduce a la norma de tal vector $\underset{\sim}{y}^{(1)}$, es decir, se calcula la componente $y_p^{(1)} / | y_p^{(1)} | = \| \underset{\sim}{y}^{(1)} \|_\infty = \max_{1 \le i \le n} | y_i^{(1)} |$ y se hace $\mu_1 = y_p^{(1)}$.

Se normaliza el vector $\underset{\sim}{y}^{(1)}$, definiendo $\underset{\sim}{x}^{(1)} = \frac{1}{\mu^{(1)}} \cdot \underset{\sim}{y}^{(1)}$.

Se repiten de nuevo las operaciones anteriores para pasos k=2, 3, …, generándose de esta forma las sucesiones: $\left\{ \underset{\sim}{y}^{(k)} \right\}_{k=1}^{\infty}$, $\{ \mu_k \}_{k=1}^{\infty}$, $\left\{ \underset{\sim}{x}^{(k)} \right\}_{k=1}^{\infty}$ mediante las relaciones:

$$\begin{aligned} &\underset{\sim}{y}^{(k)} = A \cdot \underset{\sim}{x}^{(k-1)} \\ &\mu_k = y_p^{(k)} \quad / \quad | y_p^{(k)} | = \| \underset{\sim}{y}^{(k)} \|_\infty = \max_{1 \le i \le n} | y_i^{(k)} | \\ &\underset{\sim}{x}^{(k)} = \frac{1}{\mu_k} \cdot \underset{\sim}{y}^{(k)} \end{aligned} \qquad (25)$$

En estas condiciones, se puede demostrar que la sucesión de vectores $\left\{ \underset{\sim}{x}^{(k)} \right\}_{k=1}^{\infty}$ converge hacia un vector propio asociado al autovalor λ_1, con norma del supremo 1 y la sucesión de escalares $\{ \mu_k \}_{k=1}^{\infty}$ converge al valor propio dominante λ_1. El algoritmo finalizará imponiendo como criterio de parada que:

$$\| \underset{\sim}{x}^{(k+1)} - \underset{\sim}{x}^{(k)} \|_\infty = \max_{1 \le i \le n} | x_i^{(k+1)} - x_i^{(k)} | \le \varepsilon \quad \text{o bien que} \quad \frac{\| \underset{\sim}{x}^{(k+1)} - \underset{\sim}{x}^{(k)} \|_\infty}{\| \underset{\sim}{x}^{(k+1)} \|_\infty} \le \varepsilon \quad \text{y/o}$$

$$| \mu_{k+1} - \mu_k | \le \varepsilon \quad \text{o bien que} \quad \frac{| \mu_{k+1} - \mu_k |}{| \mu_{k+1} |} \le \varepsilon \qquad (26)$$

siendo ε una tolerancia fijada. De esta forma, se considerará como aproximación al autovalor λ_1 el valor μ_{k+1} y como aproximación a su vector propio asociado el vector $\underset{\sim}{x}^{(k+1)}$.

Observaciones:

1) El método de la potencia sigue teniendo validez cuando el autovalor dominante λ_1 no es único.

2) Se puede demostrar que la rapidez con la que converge este método es tanto mayor cuanto menor sea el cociente $\frac{|\lambda_2|}{|\lambda_1|}$.

Ejemplo:

La matriz $A=\begin{pmatrix}-4 & 14 & 0\\ -5 & 13 & 0\\ -1 & 0 & 2\end{pmatrix}$ tiene tres autovalores reales $\lambda_1=6$, $\lambda_2=3$, $\lambda_3=2$, por lo que el método de las potencias nos conducirá al autovalor dominante $\lambda_1=6$. Elegimos $\underset{\sim}{x}^{(0)}=(1,1,1)^t$ y calculamos:

$$\underset{\sim}{y}^{(1)}=A\cdot\underset{\sim}{x}^{(0)}=\begin{pmatrix}-4 & 14 & 0\\ -5 & 13 & 0\\ -1 & 0 & 2\end{pmatrix}\cdot\begin{pmatrix}1\\1\\1\end{pmatrix}=\begin{pmatrix}10\\8\\1\end{pmatrix}\rightarrow\|\underset{\sim}{y}^{(1)}\|_\infty=y_1^{(1)}=10\rightarrow$$

$$\mu_1=10\quad\text{y}\quad\underset{\sim}{x}^{(1)}=\frac{1}{10}\cdot(10,8,1)^t=(1,0.8,\ 0.1)^t$$

De manera análoga (operando con redondeo a 6 dígitos significativos), obtenemos:

$$\underset{\sim}{y}^{(2)}=A\cdot\underset{\sim}{x}^{(1)}=\begin{pmatrix}-4 & 14 & 0\\ -5 & 13 & 0\\ -1 & 0 & 2\end{pmatrix}\cdot\begin{pmatrix}1\\0.8\\0.1\end{pmatrix}=\begin{pmatrix}7.2\\5.4\\-0.8\end{pmatrix}\rightarrow\|\underset{\sim}{y}^{(2)}\|_\infty=y_1^{(2)}=7.2\rightarrow$$

$$\mu_2=7.2\quad\text{y}\quad\underset{\sim}{x}^{(2)}=\frac{1}{7.2}\cdot(7.2,5.4,0.8)^t=(1,\ 0.75,\ -0.111111)^t$$

$$\underset{\sim}{y}^{(3)}=A\cdot\underset{\sim}{x}^{(2)}=\begin{pmatrix}-4 & 14 & 0\\ -5 & 13 & 0\\ -1 & 0 & 2\end{pmatrix}\cdot\begin{pmatrix}1\\0.75\\-0.111111\end{pmatrix}=\begin{pmatrix}6.5\\4.75\\-1.22222\end{pmatrix}\rightarrow\|\underset{\sim}{y}^{(3)}\|_\infty=y_1^{(3)}=6.5\rightarrow$$

$$\mu_3=6.5\quad\text{y}\quad\underset{\sim}{x}^{(3)}=\frac{1}{6.5}\cdot(6.5,4.75,-1.22222)^t=(1,\ 0.730769,\ -0.188034)^t$$

$$\underset{\sim}{y}^{(4)}=A\cdot\underset{\sim}{x}^{(3)}=\begin{pmatrix}-4 & 14 & 0\\ -5 & 13 & 0\\ -1 & 0 & 2\end{pmatrix}\cdot\begin{pmatrix}1\\0.730769\\-0.188034\end{pmatrix}=\begin{pmatrix}6.23077\\4.5\\-1.37607\end{pmatrix}\rightarrow\|\underset{\sim}{y}^{(3)}\|_\infty=y_1^{(3)}=6.23077\rightarrow$$

$$\mu_4=6.23077\ \text{y}\ \underset{\sim}{x}^{(4)}=\frac{1}{6.23077}\cdot(6.23077,4.5,-1.37607)^t=(1,\ 0.722222,\ -0.220851)^t$$

Así sucesivamente, vamos calculando aproximaciones $\mu^{(k)}$ al autovalor $\lambda_1=6$ y $\underset{\sim}{x}^{(k)}$ a su autovector propio asociado. La siguiente tabla recoge algunas aproximaciones más:

$\underset{\sim}{y}^{(1)} = A \cdot \underset{\sim}{x}^{(0)}$	$\underset{\sim}{x}^{(1)}$	$\underset{\sim}{y}^{(2)} = A \cdot \underset{\sim}{x}^{(1)}$	$\underset{\sim}{x}^{(2)}$	$\underset{\sim}{y}^{(3)} = A \cdot \underset{\sim}{x}^{(2)}$	$\underset{\sim}{x}^{(3)}$
$\begin{pmatrix} 10 \\ 8 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.8 \\ 0.1 \end{pmatrix}$	$\begin{pmatrix} 7.2 \\ 5.4 \\ -0.8 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.75 \\ -0.111111 \end{pmatrix}$	$\begin{pmatrix} 6.5 \\ 4.75 \\ -1.22222 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.730769 \\ -0.188034 \end{pmatrix}$
$\mu_1 = 10$		$\mu_2 = 7.2$		$\mu_3 = 6.5$	

$\underset{\sim}{y}^{(4)} = A \cdot \underset{\sim}{x}^{(3)}$	$\underset{\sim}{x}^{(4)}$	$\underset{\sim}{y}^{(5)} = A \cdot \underset{\sim}{x}^{(4)}$	$\underset{\sim}{x}^{(5)}$	$\underset{\sim}{y}^{(6)} = A \cdot \underset{\sim}{x}^{(5)}$	$\underset{\sim}{x}^{(6)}$
$\begin{pmatrix} 6.23077 \\ 4.5 \\ -1.37607 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.722222 \\ -0.220851 \end{pmatrix}$	$\begin{pmatrix} 6.11111 \\ 4.38889 \\ -1.44170 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.718182 \\ -0.235915 \end{pmatrix}$	$\begin{pmatrix} 6.05455 \\ 4.33637 \\ -1.47183 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.716217 \\ -0.243095 \end{pmatrix}$
$\mu_4 = 6.23077$		$\mu_5 = 6.11111$		$\mu_6 = 6.05455$	

$\underset{\sim}{y}^{(7)} = A \cdot \underset{\sim}{x}^{(6)}$	$\underset{\sim}{x}^{(7)}$	$\underset{\sim}{y}^{(8)} = A \cdot \underset{\sim}{x}^{(7)}$	$\underset{\sim}{x}^{(8)}$	$\underset{\sim}{y}^{(9)} = A \cdot \underset{\sim}{x}^{(8)}$	$\underset{\sim}{x}^{(9)}$
$\begin{pmatrix} 6.02704 \\ 4.31082 \\ -1.48619 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.715247 \\ -0.246587 \end{pmatrix}$	$\begin{pmatrix} 6.01346 \\ 4.29821 \\ -1.49318 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.714763 \\ -0.248306 \end{pmatrix}$	$\begin{pmatrix} 6.00668 \\ 4.29192 \\ -1.49661 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.714524 \\ -0.249158 \end{pmatrix}$
$\mu_7 = 6.02704$		$\mu_8 = 6.01346$		$\mu_9 = 6.00668$	

$\underset{\sim}{y}^{(10)} = A \cdot \underset{\sim}{x}^{(9)}$	$\underset{\sim}{x}^{(10)}$	$\underset{\sim}{y}^{(11)} = A \cdot \underset{\sim}{x}^{(10)}$	$\underset{\sim}{x}^{(11)}$	$\underset{\sim}{y}^{(12)} = A \cdot \underset{\sim}{x}^{(11)}$	$\underset{\sim}{x}^{(12)}$
$\begin{pmatrix} 6.00334 \\ 4.28881 \\ -1.49832 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.714404 \\ -0.249581 \end{pmatrix}$	$\begin{pmatrix} 6.00166 \\ 4.28725 \\ -1.49916 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.714344 \\ -0.249791 \end{pmatrix}$	$\begin{pmatrix} 6.00082 \\ 4.28647 \\ -1.49958 \end{pmatrix}$	$\begin{pmatrix} 1 \\ 0.714314 \\ -0.249896 \end{pmatrix}$
$\mu_{10} = 6.00334$		$\mu_{11} = 6.00166$		$\mu_{12} = 6.00082$	

En el paso k=12 el error relativo cometido en la aproximación al autovalor es $\frac{|\mu_{k+1}-\mu_k|}{|\mu_{k+1}|} = 1.39981\times10^{-4}$ y el error relativo en el vector propio asociado es $\frac{\|\underset{\sim}{x}^{(k+1)}-\underset{\sim}{x}^{(k)}\|_\infty}{\|\underset{\sim}{x}^{(k+1)}\|_\infty} = 1.05\times10^{-4}$, es decir, en ambos casos es del orden de 10^{-4}.

EJERCICIOS TEMA 2

1.- Resolver los siguientes sistemas mediante eliminación gaussiana

$$\text{a)}\begin{cases} x + y + 2z = 1 \\ x + 2y + z = 1 \\ 2x + y + z = 1 \end{cases} \qquad \text{b)}\begin{cases} x_1 + 2x_2 + x_3 = 6 \;;\; 2 \\ 2x_1 + x_2 + 2x_3 = 6 \;;\; 4 \\ x_1 + 2x_2 + 2x_3 = 7 \;;\; 3 \end{cases}$$

$$\text{c)}\begin{cases} 6x_1 - 2x_2 + 2x_3 + 4x_4 = 10 \\ 12x_1 - 8x_2 + 6x_3 + 10x_4 = 20 \\ 3x_1 - 13x_2 + 9x_3 + 3x_4 = 2 \\ -6x_1 + 4x_2 + x_3 - 18x_4 = -19 \end{cases}$$

2.- Resolver el siguiente sistema, obteniendo previamente la factorización LU de la matriz del sistema (sin modificar el término independiente), mediante eliminación gaussiana.

$$\begin{cases} x + y + 3t = 4 \\ 2x + y - z + t = 1 \\ 3x - y - z + 2t = -3 \\ -x + 2y + 3z - t = 4 \end{cases}$$

3.- Calcular la inversa y el determinante de las siguientes matrices mediante eliminación gaussiana

$$\text{a)}\begin{pmatrix} 1 & 1 & 2 \\ 1 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix} \qquad \text{b)}\begin{pmatrix} 1 & 2 & 1 \\ 2 & 1 & 2 \\ 1 & 2 & 2 \end{pmatrix} \qquad \text{c)}\begin{pmatrix} 1 & 1 & 0 \\ 1 & 4 & 6 \\ 0 & 6 & 14 \end{pmatrix}$$

$$\text{d)}\begin{pmatrix} 1 & 1 & 0 & 3 \\ 2 & 1 & -1 & 1 \\ 3 & -1 & -1 & 2 \\ -1 & 2 & 3 & -1 \end{pmatrix} \qquad \text{e)}\begin{pmatrix} 1 & 2 & 3 & 4 \\ 2 & 1 & 4 & 2 \\ 2 & 5 & 7 & 12 \\ 2 & -2 & 6 & 22 \end{pmatrix} \qquad \text{f)}\begin{pmatrix} 6 & -2 & 2 & 4 \\ 12 & -8 & 6 & 10 \\ 3 & -13 & 9 & 3 \\ -6 & 4 & 1 & -18 \end{pmatrix}$$

4.- Resolver los siguientes sistemas a partir de la factorización de Doolittle:

$$\text{a)}\begin{cases} x_1 + 2x_2 = 3 \\ x_2 + x_3 = 1 \\ -2x_1 - 4x_2 + x_3 = 1 \end{cases} \qquad \text{b)}\begin{cases} x_1 + x_3 = 2 \\ x_2 = 3 \\ -x_1 + x_3 = 0 \end{cases}$$

$$\text{c)}\begin{cases} 6x_1 - 2x_2 + 2x_3 + 4x_4 = 10 \\ 12x_1 - 8x_2 + 6x_3 + 10x_4 = 20 \\ 3x_1 - 13x_2 + 9x_3 + 3x_4 = 2 \\ -6x_1 + 4x_2 + x_3 - 18x_4 = -19 \end{cases}$$

5.- Resolver de nuevo los sistemas del ejercicio anterior utilizando ahora la factorización de Crout.

6.- Obtener la factorización LU de las siguientes matrices utilizando el método de Crout

$$\text{a)}\begin{pmatrix} 1 & -1 & 2 \\ -1 & 5 & 4 \\ 2 & 4 & 14 \end{pmatrix} \qquad \text{b)}\begin{pmatrix} 6 & 2 & 1 & -1 \\ 2 & 4 & 1 & 0 \\ 1 & 1 & 4 & -1 \\ -1 & 0 & -1 & 3 \end{pmatrix}$$

7.- a) Comprobar mediante un método numérico si la siguiente matriz es o no definida positiva

$$A = \begin{pmatrix} 4 & 4 & 0 & 2 \\ 4 & 5 & -1 & 5 \\ 0 & -1 & 5 & 1 \\ 2 & 5 & 1 & 7 \end{pmatrix}$$

b) Haciendo uso de los resultados del apartado anterior, calcular mediante el método más adecuado la inversa de la siguiente matriz:

$$B = \begin{pmatrix} 4 & 4 & 0 \\ 4 & 5 & -1 \\ 0 & -1 & 5 \end{pmatrix}$$

8.- Aplicar el método de Cholesky para resolver los siguientes sistemas:

a) $\begin{cases} x_1 + x_3 = 1 \\ 4x_2 = 2 \\ x_1 + 4x_3 = 4 \end{cases}$ b) $\begin{cases} x_1 + x_2 = 2 \\ x_1 + 4x_2 + 6x_3 = 5 \\ 6x_2 + 14x_3 = 6 \end{cases}$

c) $\begin{cases} 4x + 2y + 6t = 8 \\ 2x + 17y + 8z - t = 20 \\ 8y + 5z + t = 17 \\ 6x - y + z + 28t = 50 \end{cases}$

9.- a) Resolver el sistema mediante el método de Crout.

$$\begin{cases} 3x_1 + 2x_2 = 1 \\ x_1 + 4x_2 + 3x_3 = 2 \\ 4x_2 + 9x_3 + 4x_4 = 3 \\ 9x_3 + 16x_4 = 4 \end{cases}$$

b) Resolver de nuevo el sistema mediante el método de Doolittle obteniendo simultáneamente el algoritmo correspondiente a la factorización para una matriz de dimensión nxn.

10.- Resolver los siguientes sistemas mediante eliminación gaussiana con pivotaje parcial y cambio de escala. Obtener además las matrices L, U y P que dan lugar a la factorización de $P^t \cdot A$, siendo A la matriz del sistema

a) $\begin{cases} x_1 + x_2 + x_3 = 4 \\ 2x_1 + 3x_2 + x_3 = 9 \\ x_1 - x_2 - x_3 = -2 \end{cases}$ b) $\begin{cases} 4x + 5y - z + 5t = 1 \\ 4x + 4y + 2t = 2 \\ 2x + 5y + z + 7t = 2 \\ - y + 5z + t = 1 \end{cases}$

c) $\begin{cases} 3x_1 + 5x_2 + 3x_3 + 7x_4 = -8 \\ 3x_1 + 4x_2 + x_3 + 2x_4 = -3 \\ 3x_1 + 5x_2 + 3x_3 + 5x_4 = -6 \\ 6x_1 + 8x_2 + x_3 + 5x_4 = -8 \end{cases}$

11.- Dada la matriz $A=\begin{pmatrix} -1 & 0 & 1 & 3 \\ 1 & -1 & 0 & 1 \\ 1 & -2 & 1 & 1 \\ 0 & 2 & -1 & 1 \end{pmatrix}$

a) ¿Es posible calcular su inversa mediante el método de eliminación gaussiana? ¿Y mediante el método de Doolittle?

b) Calcular la inversa de la matriz anterior mediante el método de eliminación gaussiana con ténica de pivotaje y escalado.

12.- Calcular la inversa y el determinante de las siguientes matrices, mediante el método de eliminación gaussiana con pivotaje parcial y cambio de escala:

a) $\begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 4 & 2 \end{pmatrix}$ b) $\begin{pmatrix} 1 & 2 & 4 & 1 \\ 2 & 8 & 6 & 4 \\ 3 & 10 & 8 & 8 \\ 4 & 12 & 10 & 6 \end{pmatrix}$

13.- Resolver los siguientes sistemas utilizando el método de Jacobi. Utilizar un criterio de parada de error absoluto y trabajar con redondeo a 6 dígitos significativos.

a) $\begin{cases} 4x_1 + 2x_2 + x_3 = 11 \\ -x_1 + 2x_2 = 3 \\ 2x_1 + x_2 + 4x_3 = 16 \end{cases} \quad \underset{\sim}{x}^0=\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \text{ y } \varepsilon=0.05$

b) $\begin{cases} 4x_2 + 2x_3 = 2 \\ 4x_1 + 2x_2 + 10x_3 = 6 \\ 5x_1 + 4x_2 = 5 \end{cases} \quad \underset{\sim}{x}^0=\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \text{ y } \varepsilon=0.05$

14.- Resolver de nuevo el ejercicio anterior usando ahora el método de Gauss-Seidel.

15.- a) Estudiar la convergencia y velocidad de convergencia de los métodos de Jacobi y de Gauss-Seidel para la resolución del sistema

$$\begin{cases} 4x - y - z = 1 \\ -x + 4y - t = 2 \\ -x + 4z - t = 0 \\ -y - z + 4t = 1 \end{cases}$$

b) Realizar 3 iteraciones del método de Jacobi y otras 3 del método de Gauss-Seidel, tomando como valor inicial $\underset{\sim}{x}^0 = \begin{pmatrix} 0.45 \\ 0.71 \\ 0.21 \\ 0.526 \end{pmatrix}$ y trabajando con redondeo a 5 decimales. Calcular el porcentaje de error cometido en cada caso.

16.- a) Estudiar la convergencia y velocidad de convergencia de los métodos de Jacobi y de Gauss-Seidel para la resolución del sistema

$$\begin{cases} -2x + 5z - 2t = 2 \\ 5x - 2y - 2z = 1 \\ -2y - 2z + 5t = 0 \\ -2x + 5y - 2t = 1 \end{cases}$$

b) Resolverlo mediante el método más rápido con una precisión del 3%, tomando como valor inicial $\underset{\sim}{x}^0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$ y trabajando con redondeo a 5 dígitos significativos.

17.- Dado el sistema $\begin{cases} x + z = 1 \\ 4y = 2 \\ x + 4z = 4 \end{cases}$

Estudiar la convergencia del método de Gauss-Seidel. Utilizando este método efectuar 4 iteraciones partiendo del vector $\underset{\sim}{x}^0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$ y trabajando con redondeo a 4 decimales. Calcular el porcentaje de error cometido.

18.- Dado el sistema

$$\begin{cases} -x + y + 3z = 2 \\ 3x + y + z = 1 \\ 2x + 5y + z = -1 \end{cases}$$

hallar su solución con un precisión del 5% mediante el método de Gauss-Seidel, tomando como valor inicial $\underset{\sim}{x}^0 = \begin{pmatrix} 0 \\ -0.5 \\ 1 \end{pmatrix}$ y operando con redondeo a 5 decimales.

19.- Dados los sistemas

$$1)\begin{cases} 9x_1 - 2x_2 = 5 \\ -2x_1 + 4x_2 - x_3 = 1 \\ -x_2 + x_3 = -5/6 \end{cases} \qquad 2)\begin{cases} -y + 3z = 5 \\ 10x - 2y = 4 \\ -2x + 6y - z = 3 \end{cases}$$

a) ¿Cuál es la relación entre los radios espectrales de las matrices T para los métodos de Jacobi y Gauss-Seidel? ¿Cuál de los dos métodos converge más aprisa?

b) Para el sistema 1), partiendo del vector inicial $\underset{\sim}{x}^0 = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$, calcular la solución mediante el método más rápido con una tolerancia $\varepsilon = 5\times10^{-4}$ y trabajando con redondeo a 4 decimales.

20.- Estudiar la convergencia del método de Gauss-Seidel para los sistemas

$$a)\begin{cases} 3x_1 + x_3 = 0 \\ x_1 + 2x_2 + 3x_3 = 0 \\ 3x_2 + x_3 = 5 \end{cases} \qquad b)\begin{cases} x_1 + x_4 = 2 \\ x_1 + 4x_2 - x_4 = 4 \\ x_1 + x_3 = 2 \\ x_3 + x_4 = 4 \end{cases}$$

21.- Estudiar la convergencia de los métodos de Jacobi y Gauss-Seidel para los sistemas

a) $\begin{cases} 2x_1 - x_2 + x_3 = -1 \\ 2x_1 + 2x_2 + 2x_3 = 4 \\ -x_1 - x_2 + 2x_3 = -5 \end{cases}$ b) $\begin{cases} x_1 + 2x_2 - 2x_3 = 7 \\ x_1 + x_2 + x_3 = 2 \\ 2x_1 + 2x_2 + x_3 = 5 \end{cases}$

22.- Dada la matriz $A = \begin{pmatrix} -5 & -2 & 0 \\ -2 & 3 & -1 \\ 0 & -1 & 1 \end{pmatrix}$, calcular el valor propio de módulo máximo y su vector propio asociado, ambos con una precisión de 10^{-3} . Operar con redondeo a 6 decimales y tomar $\underset{\sim}{x}^0 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$

23.- Calcular el radio espectral de la matriz $A = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}$ con una precisión del 0.01%, tomando como aproximación inicial $\underset{\sim}{x}^0 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix}$ y trabajando con redondeo a 5 dígitos significativos.

24.- Estudiar la convergencia o divergencia del método de Jacobi para el sistema

$$\begin{cases} 6x + 3y + 6z = 1 \\ 3x + 6y + 6z = 5 \\ 6x + 6y + 3z = 2 \end{cases}$$

aplicando el método de las potencias, tomando como valor inicial $\underset{\sim}{x}^0 = \begin{pmatrix} 0.6 \\ 0.6 \\ 1 \end{pmatrix}$ y finalizando dicho método cuando se alcance una precisión del 3%. Operar con redondeo a **5** dígitos significativos.

SOLUCIONES EJERCICIOS TEMA 2

1.- a) $L=\begin{pmatrix}1&0&0\\1&1&0\\2&-1&1\end{pmatrix}$; $U=\begin{pmatrix}1&1&2\\0&1&-1\\0&0&-4\end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix}1/4\\1/4\\1/4\end{pmatrix}$

b) $L=\begin{pmatrix}1&0&0\\2&1&0\\1&0&1\end{pmatrix}$; $U=\begin{pmatrix}1&2&1\\0&-3&0\\0&0&1\end{pmatrix}$; $\underset{\sim}{x}_1=\begin{pmatrix}1\\2\\1\end{pmatrix}$; $\underset{\sim}{x}_2=\begin{pmatrix}1\\0\\1\end{pmatrix}$

c) $L=\begin{pmatrix}1&0&0&0\\2&1&0&0\\1/2&3&1&0\\-1&-1/2&2&1\end{pmatrix}$; $U=\begin{pmatrix}6&-2&2&4\\0&-4&2&2\\0&0&2&-5\\0&0&0&-3\end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix}1\\1\\1\\1\end{pmatrix}$

2.- $L=\begin{pmatrix}1&0&0&0\\2&1&0&0\\3&4&1&0\\-1&-3&0&1\end{pmatrix}$; $U=\begin{pmatrix}1&1&0&3\\0&-1&-1&-5\\0&0&3&13\\0&0&0&-13\end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix}-1\\2\\0\\1\end{pmatrix}$

3.- a) $A^{-1}=\begin{pmatrix}-1/4&-1/4&3/4\\-1/4&3/4&-1/4\\3/4&-1/4&-1/4\end{pmatrix}$; $|A|=-4$

b) $A^{-1}=\begin{pmatrix}2/3&2/3&-1\\2/3&-1/3&0\\-1&0&1\end{pmatrix}$; $|A|=-3$

c) $A^{-1}=\begin{pmatrix}10/3&-7/3&1\\-7/3&7/3&-1\\1&-1&1/2\end{pmatrix}$; $|A|=6$

d) $A^{-1}=\begin{pmatrix}-3/13&8/39&1/3&7/39\\1/13&19/39&-1/3&2/39\\0&-1/3&1/3&1/3\\5/13&-3/13&0&-1/13\end{pmatrix}$; $|A|=39$

e) $A^{-1}=\begin{pmatrix} 183 & -31 & -65 & 5 \\ 40 & -7 & -14 & 1 \\ -110 & 19 & 39 & -3 \\ 17 & -3 & -6 & 1/2 \end{pmatrix}$; $|A|=-2$

f) $A^{-1}=\begin{pmatrix} -251/72 & 155/72 & -25/36 & 11/36 \\ 199/24 & -115/24 & 17/12 & -7/12 \\ 143/12 & -83/12 & 13/6 & -5/6 \\ 11/3 & -13/6 & 2/3 & -1/3 \end{pmatrix}$; $|A|=144$

4.- a) $L=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}$; $U=\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix} 15 \\ -6 \\ 7 \end{pmatrix}$

b) $L=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}$; $U=\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 2 \end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}$

c) $L=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 1/2 & 3 & 1 & 0 \\ -1 & -1/2 & 2 & 1 \end{pmatrix}$; $U=\begin{pmatrix} 6 & -2 & 2 & 4 \\ 0 & -4 & 2 & 2 \\ 0 & 0 & 2 & -5 \\ 0 & 0 & 0 & -3 \end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

5.- a) $L=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -2 & 0 & 1 \end{pmatrix}$; $U=\begin{pmatrix} 1 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix} 15 \\ -6 \\ 7 \end{pmatrix}$

b) $L=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & 0 & 2 \end{pmatrix}$; $U=\begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix} 1 \\ 3 \\ 1 \end{pmatrix}$

c) $L=\begin{pmatrix} 6 & 0 & 0 & 0 \\ 12 & -4 & 0 & 0 \\ 3 & -12 & 2 & 0 \\ -6 & 2 & 4 & -3 \end{pmatrix}$; $U=\begin{pmatrix} 1 & -1/3 & 1/3 & 2/3 \\ 0 & 1 & -1/2 & -1/2 \\ 0 & 0 & 1 & -5/2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}$

6.- a) $L=\begin{pmatrix} 1 & 0 & 0 \\ -1 & 4 & 0 \\ 2 & 6 & 1 \end{pmatrix}$; $U=\begin{pmatrix} 1 & -1 & 2 \\ 0 & 1 & 3/2 \\ 0 & 0 & 1 \end{pmatrix}$

b) $L=\begin{pmatrix} 6 & 0 & 0 & 0 \\ 2 & 10/3 & 0 & 0 \\ 1 & 2/3 & 37/10 & 0 \\ -1 & 1/3 & -9/10 & 191/74 \end{pmatrix}$; $U=\begin{pmatrix} 1 & 1/3 & 1/6 & -1/6 \\ 0 & 1 & 1/5 & 1/10 \\ 0 & 0 & 1 & -9/37 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

7.- a) $L=\begin{pmatrix} 2 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & -1 & 2 & 0 \\ 1 & 3 & 2 & * \end{pmatrix}$; A no es definida positiva pues $l_{4,4}^2=-7$

b) $B^{-1}=\begin{pmatrix} 3/2 & -5/4 & -1/4 \\ -5/4 & 5/4 & 1/4 \\ -1/4 & 1/4 & 1/4 \end{pmatrix}$

8.- a) $L=\begin{pmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 1 & 0 & \sqrt{3} \end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix} 0 \\ 1/2 \\ 1 \end{pmatrix}$

b) $L=\begin{pmatrix} 1 & 0 & 0 \\ 1 & \sqrt{3} & 0 \\ 0 & 2\sqrt{3} & \sqrt{2} \end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix}$

c) $L=\begin{pmatrix} 2 & 0 & 0 & 0 \\ 1 & 4 & 0 & 0 \\ 0 & 2 & 1 & 0 \\ 3 & -1 & 3 & 3 \end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix} -5/24 \\ -7/12 \\ 4 \\ 5/3 \end{pmatrix}$

9.- a) $L=\begin{pmatrix} 3 & 0 & 0 & 0 \\ 1 & 10/3 & 0 & 0 \\ 0 & 4 & 27/5 & 0 \\ 0 & 0 & 9 & 28/3 \end{pmatrix}$; $U=\begin{pmatrix} 1 & 2/3 & 0 & 0 \\ 0 & 1 & 9/10 & 0 \\ 0 & 0 & 1 & 20/27 \\ 0 & 0 & 0 & 1 \end{pmatrix}$; $\underset{\sim}{x}=\begin{pmatrix} 0 \\ 1/2 \\ 0 \\ 1/4 \end{pmatrix}$

b) $L=\begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/3 & 1 & 0 & 0 \\ 0 & 6/5 & 1 & 0 \\ 0 & 0 & 5/3 & 1 \end{pmatrix}$; $U=\begin{pmatrix} 3 & 2 & 0 & 0 \\ 0 & 10/3 & 3 & 0 \\ 0 & 0 & 27/5 & 4 \\ 0 & 0 & 0 & 28/3 \end{pmatrix}$

$$\begin{cases} \alpha_1 = a_1 \\ \begin{cases} \beta_i = b_i / \alpha_{i-1} \\ \alpha_i = a_i - \beta_i c_{i-1} \end{cases} \end{cases} \quad i = 2, 3, \ldots, n$$

10.- a) $L = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & -1/2 & 1 \end{pmatrix}$; $U = \begin{pmatrix} 1 & 1 & 1 \\ 0 & -2 & -2 \\ 0 & 0 & -2 \end{pmatrix}$; $P = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$; $\underset{\sim}{p} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$; $\underset{\sim}{x} = \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix}$

b) $L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 0 \\ 0 & -1/3 & 1 & 0 \\ 1 & 1/3 & -1/4 & 1 \end{pmatrix}$; $U = \begin{pmatrix} 4 & 4 & 0 & 2 \\ 0 & 3 & 1 & 6 \\ 0 & 0 & 16/3 & 3 \\ 0 & 0 & 0 & 7/4 \end{pmatrix}$;

$P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$; $\underset{\sim}{p} = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 1 \end{pmatrix}$; $\underset{\sim}{x} = \begin{pmatrix} -1/2 \\ 9/7 \\ 4/7 \\ -4/7 \end{pmatrix}$

c) $L = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 2 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{pmatrix}$; $U = \begin{pmatrix} 3 & 4 & 1 & 2 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 2 \end{pmatrix}$;

$P = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}$; $\underset{\sim}{p} = \begin{pmatrix} 2 \\ 3 \\ 4 \\ 1 \end{pmatrix}$; $\underset{\sim}{x} = \begin{pmatrix} 2 \\ -2 \\ 1 \\ -1 \end{pmatrix}$

11.- a) No porque el pivote $a_{3,3} = 0$. Este pivote es el elemento $u_{3,3}$ del método de Doolittle.

b) $A^{-1} = \begin{pmatrix} -1/2 & -1 & 3/2 & 1 \\ -1/4 & -3/2 & 5/4 & 1 \\ -1/4 & -5/2 & 9/4 & 1 \\ 1/4 & 1/2 & -1/4 & 0 \end{pmatrix}$; $\underset{\sim}{p} = \begin{pmatrix} 2 \\ 4 \\ 3 \\ 1 \end{pmatrix}$

12.- a) $A^{-1}=\begin{pmatrix} 0 & 1 & -1/2 \\ 1 & -2 & 1 \\ -2 & 4 & -3/2 \end{pmatrix}$; $\underset{\sim}{p}=\begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$; $|A|=-2$

b) $A^{-1}=\begin{pmatrix} -1/2 & -7/4 & 0 & 5/4 \\ -1/4 & 11/24 & -1/6 & -1/24 \\ 1/2 & 1/4 & 0 & -1/4 \\ 0 & -1/6 & 1/3 & -1/6 \end{pmatrix}$; $\underset{\sim}{p}=\begin{pmatrix} 4 \\ 2 \\ 1 \\ 3 \end{pmatrix}$; $|A|=48$

13.- a)

k	x_1^k	x_2^k	x_3^k	e_{abs}
0	1	1	1	
1	2	2	3.25	2.25
2	0.9375	2.5	2.5	1.0625
3	0.875	1.96875	2.90625	0.53125
4	1.03906	1.9375	3.07031	0.16406
5	1.01367	2.01953	2.99610	0.08203
6	0.99121	2.00684	2.98828	0.02246

b)

k	x_1^k	x_2^k	x_3^k	e_{abs}
0	0	0	0	
1	1	0.5	0.6	1
2	0.6	0.2	0.1	0.5
3	0.84	0.45	0.32	0.25
4	0.64	0.34	0.174	0.20
5	0.728	0.413	0.276	0.102
6	0.6696	0.362	0.2262	0.0584
7	0.7104	0.3869	0.25976	0.0408

14.-

k	x_1^k	x_2^k	x_3^k	e_{abs}
0	1	1	1	
1	2	2.5	2.375	1.5
2	0.90625	1.95313	3.05860	1.09375
3	1.00879	2.00440	2.99451	0.10254
4	0.999173	1.99959	3.00052	0.009617

b)

k	x_1^k	x_2^k	x_3^k	e_{abs}
0	0	0	0	
1	1	0.5	0.1	1
2	0.6	0.45	0.27	0.4
3	0.64	0.365	0.271	0.085
4	0.708	0.3645	0.2439	0.068
5	0.7084	0.37805	0.24103	0.01355

15.- a) Matriz estrictamanete diagonal dominante y Teorema de Stein-Rosenberg.

b) **Jacobi**

k	x^k	y^k	z^k	t^k
0	0.45	0.71	0.21	0.526
1	0.48	0.744	0.244	0.48
2	0.497	0.74	0.24	0.497
3	0.495	0.7485	0.2485	0.495

$e_{rel} \times 100 = 1.13560\%$

Gauss-Seidel

k	x^k	y^k	z^k	t^k
0	0.45	0.71	0.21	0.526
1	0.48	0.7515	0.2515	0.50075
2	0.50075	0.75038	0.25038	0.50019
3	0.50019	0.75010	0.25010	0.50005

$e_{rel} \times 100 = 0.07466\%$

16.- a) Reordenando es estrictamente diagonal dominante. Teorema de Stein-Rosenberg.

b)

k	x^k	y^k	z^k	t^k	$e_{rel}\times 100$
0	1	1	1	1	
1	1	1	1.2	0.88	
2	1.08	0.984	1.184	0.8672	6.7568%
3	1.0672	0.97376	1.1738	0.85902	1.0905%

17.- A es simétrica y definida positiva.

k	x^k	y^k	z^k
0		0	0
1	1	0.5	0.75
2	0.25	0.5	0.9375
3	0.0625	0.5	0.9844
4	0.0156	0.5	0.9961

$e_{rel}\times 100 = 4.7084\%$

18.-

k	x^k	y^k	z^k	$e_{rel}\times 100$
0	0	-0.5	1	
1	0.16667	-0.46667	0.87778	
2	0.19630	-0.45408	0.88346	3.35386%

19.- a) $0 < \rho(T_G) = \rho(T_j)^2 < 1$

b)

k	x_1^k	x_2^k	x_3^k	e_{abs}
0	0	0	0	
1	0.5556	0.5278	-0.3055	0.5556
2	0.6728	0.5100	-0.3233	0.1172
3	0.6689	0.5036	-0.3297	3.9×10^{-3}
4	0.6675	0.5013	-0.3320	2.3×10^{-3}
5	0.6670	0.5005	-0.3328	8×10^{-4}
6	0.6668	0.5002	-0.3331	3×10^{-4}

20.- a) Intercambiar $F_2 \leftrightarrow F_3 \quad \rho(T_G)=1/3<1$

b) $\rho(T_G)=1$

21.- a) $\rho(T_J)=\sqrt{5}/2>1; \quad \rho(T_G)=1/2<1$

b) $\rho(T_J)=0<1; \quad \rho(T_G)=2>1$

22.-

$\underset{\sim}{x}^3$	…	$\underset{\sim}{x}^6$	…	$\underset{\sim}{x}^9$	…	$\underset{\sim}{x}^{12}$	…
$\begin{pmatrix}1\\0.313726\\0.013072\end{pmatrix}$	…	$\begin{pmatrix}1\\0.216043\\0.045029\end{pmatrix}$	…	$\begin{pmatrix}1\\0.248384\\0.034373\end{pmatrix}$	…	$\begin{pmatrix}1\\0.237468\\0.03797\end{pmatrix}$	…
$\mu_3=-5.275862$		$\mu_6=-5.550832$		$\mu_9=-5.456996$		$\mu_{12}=-5.488312$	

$\underset{\sim}{x}^{15}$	…	$\underset{\sim}{x}^{18}$	…	$\underset{\sim}{x}^{21}$
$\begin{pmatrix}1\\0.241129\\0.036763\end{pmatrix}$	…	$\begin{pmatrix}1\\0.239899\\0.037169\end{pmatrix}$	..	$\begin{pmatrix}1\\0.240312\\0.037033\end{pmatrix}$
$\mu_{15}=-5.477768$		$\mu_{18}=-5.481308$		$\mu_{21}=-5.480118$

23.- $\mu_1=-4;\mu_2=3.5;\mu_3=3.4286;\mu_4=3.4167;\mu_5=3.4146;\mu_6=3.4143;$

$e_{rel}=0.0087866\%$

24.

$$T_J=\begin{pmatrix}0 & -0.5 & -1\\ -0.5 & 0 & -1\\ -2 & -2 & 0\end{pmatrix}$$

$\mu_1=-2.4;\mu_2=-2.1667;\mu_3=-2.3460;\mu_4=-2.2051;\mu_5=-2.3139;\mu_6=-2.2286;$

$\mu_7=-2.2949; e_{rel}=2.8890\%$

$\rho(T_J)\simeq 2.2949>1$

Tema 3

APROXIMACIÓN MÍNIMO CUADRÁTICA

3.1. PLANTEAMIENTO DEL PROBLEMA Y CARACTERIZACIÓN DEL ELEMENTO MEJOR APROXIMACIÓN

En ocasiones es necesario aproximar elementos de un determinado espacio vectorial por elementos de uno de sus subespacios. Por ejemplo, a veces, se necesita sustituir una función f "poco manejable" por otra f* que le sea cercana en algún sentido, y a la vez, "más manejable". El problema que se plantea es qué elemento se ha de elegir para sustituir al otro, lo cual es un problema de mejor aproximación.

Comenzamos estableciendo un planteamiento general del problema de aproximación:
Sea E un espacio vectorial sobre un cuerpo $\mathbb{K}$ y supongamos que en él está definida una norma $\| \cdot \|$. Sea H un subespacio vectorial de E. Entonces, dado un elemento $\mathbf{f} \in E$ se trata de hallar un elemento $\mathbf{u} \in H$ que "se asemeje lo más posible" a **f**. El criterio que vamos a seguir es tomar como aproximación el elemento de H que minimice la distancia entre **f** y **u**, es decir, $\|\mathbf{f}-\mathbf{u}\|$. A tal elemento **u** si existe, se le llama ***elemento mejor aproximación de f en H con la norma*** $\| \cdot \|$.
De todo esto, se concluye, que el problema de aproximación depende de tres cosas: del espacio vectorial E, del subespacio vectorial H y de la norma utilizada. Dependiendo de estos factores el problema recibe nombres distintos. Así, cuando en el espacio vectorial E existe definido un producto escalar y la norma es la inducida por tal producto escalar, el problema de aproximación recibe el nombre de ***aproximación mínimo-cuadrática***.
Por otro lado, las cuestiones que se nos plantean ante este tipo de problemas son: la *existencia* del elemento mejor aproximación, la *unicidad* del mismo y su *obtención* mediante un algoritmo que nos conduzca a él. Trataremos estas cuestiones para el caso de la aproximación mínimo cuadrática.
El resultado que se cumple es el siguiente:

Teorema. Sea E un espacio vectorial euclídeo y la norma inducida por el producto escalar $\|\cdot\|$, y sea H un subespacio vectorial de dimensión finita de E. Entonces, $\forall \mathbf{f}\in E$ $\exists$ un único vector $\mathbf{u}\in H$ / $\mathbf{u}$ es la mejor aproximación de $\mathbf{f}$ en H, es decir, $\|\mathbf{f}-\mathbf{u}\|\le\|\mathbf{f}-\mathbf{v}\|$ $\forall\ \mathbf{v}\in H$. Además, $\mathbf{u}$ es la mejor aproximación de $\mathbf{f}$ en H $\Leftrightarrow \mathbf{f}-\mathbf{u}\in H^{\perp}$, donde $H^{\perp}$ es el subespacio ortogonal de H.

Demostración:

$\Rightarrow$) Si $\mathbf{u}$ es el elemento de H más próximo a **f,** entonces $\|\mathbf{f}-\mathbf{u}\|\le\|\mathbf{f}-\mathbf{v}\|$ $\forall\ \mathbf{v}\in H$.

Sea $\mathbf{v}$ un elemento cualquiera de H. Entonces $\begin{cases}\mathbf{u}+\lambda\mathbf{v}\in H\\ \mathbf{u}-\lambda\mathbf{v}\in H\end{cases}$ $\forall\lambda>0\ \Rightarrow$

$$\begin{cases}\|\mathbf{f}-\mathbf{u}\|\le\|\mathbf{f}-\mathbf{u}-\lambda\mathbf{v}\|\Leftrightarrow\|\mathbf{f}-\mathbf{u}\|^2\le\|\mathbf{f}-\mathbf{u}-\lambda\mathbf{v}\|^2\\ \|\mathbf{f}-\mathbf{u}\|\le\|\mathbf{f}-\mathbf{u}+\lambda\mathbf{v}\|\Leftrightarrow\|\mathbf{f}-\mathbf{u}\|^2\le\|\mathbf{f}-\mathbf{u}+\lambda\mathbf{v}\|^2\end{cases}\Rightarrow$$

$$\begin{cases}\|\mathbf{f}-\mathbf{u}\|^2\le\langle\mathbf{f}-\mathbf{u}-\lambda\mathbf{v},\mathbf{f}-\mathbf{u}-\lambda\mathbf{v}\rangle=\|\mathbf{f}-\mathbf{u}\|^2+\lambda^2\|\mathbf{v}\|^2-2\lambda\langle\mathbf{f}-\mathbf{u},\mathbf{v}\rangle\\ \|\mathbf{f}-\mathbf{u}\|^2\le\langle\mathbf{f}-\mathbf{u}+\lambda\mathbf{v},\mathbf{f}-\mathbf{u}+\lambda\mathbf{v}\rangle=\|\mathbf{f}-\mathbf{u}\|^2+\lambda^2\|\mathbf{v}\|^2+2\lambda\langle\mathbf{f}-\mathbf{u},\mathbf{v}\rangle\end{cases}\Rightarrow$$

$$\begin{cases}0\le\lambda^2\|\mathbf{v}\|^2-2\lambda\langle\mathbf{f}-\mathbf{u},\mathbf{v}\rangle\\ 0\le\lambda^2\|\mathbf{v}\|^2+2\lambda\langle\mathbf{f}-\mathbf{u},\mathbf{v}\rangle\end{cases}\Rightarrow\begin{cases}2\langle\mathbf{f}-\mathbf{u},\mathbf{v}\rangle\le\lambda\|\mathbf{v}\|^2\\ -\lambda\|\mathbf{v}\|^2\le2\langle\mathbf{f}-\mathbf{u},\mathbf{v}\rangle\end{cases}\quad\forall\lambda>0$$

Haciendo tender $\lambda\to0$, se tiene que $\begin{cases}2\langle\mathbf{f}-\mathbf{u},\mathbf{v}\rangle\le0\\ 0\le2\langle\mathbf{f}-\mathbf{u},\mathbf{v}\rangle\end{cases}$, de donde $\langle\mathbf{f}-\mathbf{u},\mathbf{v}\rangle=0\ \forall\mathbf{v}\in H\Leftrightarrow\mathbf{f}-\mathbf{u}\in H^{\perp}$.

$\Leftarrow$) Se demostrará, a continuación que si $\mathbf{f}-\mathbf{u}\in H^{\perp}$, entonces $\mathbf{u}$ es el elemento de H más próximo a **f**, ya que se cumplirá que $\|\mathbf{f}-\mathbf{u}\|\le\|\mathbf{f}-\mathbf{v}\|$ $\forall\ \mathbf{v}\in H$. En efecto, para cualquier vector $\mathbf{v}\in H$, se puede poner:

$$\|\mathbf{f}-\mathbf{v}\|^2=\|(\mathbf{f}-\mathbf{u})+(\mathbf{u}-\mathbf{v})\|^2=\langle(\mathbf{f}-\mathbf{u})+(\mathbf{u}-\mathbf{v}),(\mathbf{f}-\mathbf{u})+(\mathbf{u}-\mathbf{v})\rangle=$$
$$=\|\mathbf{f}-\mathbf{u}\|^2+\|(\mathbf{u}-\mathbf{v})\|^2+2\langle\mathbf{f}-\mathbf{u},\mathbf{u}-\mathbf{v}\rangle=\|\mathbf{f}-\mathbf{u}\|^2+\|\mathbf{u}-\mathbf{v}\|^2$$

ya que como **f-u** $\in H^{\perp}$ y **u-v** $\in$ H, se tiene que el producto escalar $\langle\mathbf{f}-\mathbf{u},\mathbf{u}-\mathbf{v}\rangle=0$. De la expresión anterior se deduce ya trivialmente que:

$$\|\mathbf{f}-\mathbf{v}\|^2\ge\|\mathbf{f}-\mathbf{u}\|^2\Rightarrow\|\mathbf{f}-\mathbf{u}\|\le\|\mathbf{f}-\mathbf{v}\|\quad\forall\mathbf{v}\in H$$

La unicidad del elemento mejor aproximación es consecuencia de la demostración que hemos realizado. ■

Esta caracterización nos permite además calcular el elemento mejor aproximación, ya que como H es un subespacio vectorial de dimensión finita, tiene una base $B_H = \{\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n\}$ Entonces, dado $\mathbf{f} \in E$, su mejor aproximación **u** en H se expresará como $\mathbf{u} = \lambda_1\mathbf{u}_1 + \lambda_2\mathbf{u}_2 + \ldots + \lambda_n\mathbf{u}_n = \sum_{i=1}^{n} \lambda_i \cdot \mathbf{u_i}$. Basta determinar los coeficientes λ_i; para ello, según acabamos de ver, imponemos que $\mathbf{f} - \mathbf{u} \in H^{\perp}$, lo cual implica que **f-u** será ortogonal a todos los elementos de la base de H, es decir,

$$\langle (\mathbf{f} - \sum_{i=1}^{n} \lambda_i \cdot \mathbf{u_i}), \mathbf{u_j} \rangle = 0 \quad \forall\ j=1,\cdots,n$$

lo cual, por la linealidad del producto escalar $\Rightarrow$

$$\langle \mathbf{f}, \mathbf{u_j} \rangle = \sum_{i=1}^{n} \lambda_i \langle \mathbf{u_i}, \mathbf{u_j} \rangle \quad j=1,\cdots,n$$

lo que desarrollado conduce al sistema lineal siguiente:

$$\begin{cases} \langle \mathbf{f}, \mathbf{u}_1 \rangle = \lambda_1 \langle \mathbf{u}_1, \mathbf{u}_1 \rangle + \lambda_2 \langle \mathbf{u}_2, \mathbf{u}_1 \rangle + \ldots + \lambda_n \langle \mathbf{u}_n, \mathbf{u}_1 \rangle \\ \langle \mathbf{f}, \mathbf{u}_2 \rangle = \lambda_1 \langle \mathbf{u}_1, \mathbf{u}_2 \rangle + \lambda_2 \langle \mathbf{u}_2, \mathbf{u}_2 \rangle + \ldots + \lambda_n \langle \mathbf{u}_n, \mathbf{u}_2 \rangle \\ \qquad\qquad \ldots\ldots\ldots\ldots\ldots \\ \langle \mathbf{f}, \mathbf{u}_n \rangle = \lambda_1 \langle \mathbf{u}_1, \mathbf{u}_n \rangle + \lambda_2 \langle \mathbf{u}_2, \mathbf{u}_n \rangle + \ldots + \lambda_n \langle \mathbf{u}_n, \mathbf{u}_n \rangle \end{cases} \tag{1}$$

y expresado matricialmente queda:

$$\begin{pmatrix} \langle \mathbf{u}_1, \mathbf{u}_1 \rangle & \langle \mathbf{u}_2, \mathbf{u}_1 \rangle & \ldots & \langle \mathbf{u}_n, \mathbf{u}_1 \rangle \\ \langle \mathbf{u}_1, \mathbf{u}_2 \rangle & \langle \mathbf{u}_2, \mathbf{u}_2 \rangle & \ldots & \langle \mathbf{u}_n, \mathbf{u}_2 \rangle \\ \ldots & \ldots & \ldots & \ldots \\ \langle \mathbf{u}_1, \mathbf{u}_n \rangle & \langle \mathbf{u}_2, \mathbf{u}_n \rangle & \ldots & \langle \mathbf{u}_n, \mathbf{u}_n \rangle \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \ldots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} \langle \mathbf{f}, \mathbf{u}_1 \rangle \\ \langle \mathbf{f}, \mathbf{u}_2 \rangle \\ \ldots \\ \langle \mathbf{f}, \mathbf{u}_n \rangle \end{pmatrix} \tag{2}$$

Este sistema lineal de n ecuaciones con n incógnitas $\lambda_1, \lambda_2, \cdots, \lambda_n$ tiene solución única por ser su matriz la matriz asociada al producto escalar en H, que, como sabemos es regular, simétrica y definida positiva . A las ecuaciones del sistema (2) se les llama ***ecuaciones normales*** y su solución nos permite obtener la mejor aproximación de **f** en H.

En el caso particular de que se considere una base de H ortogonal $B'_H = \{\mathbf{u}'_1, \mathbf{u}'_2, \ldots, \mathbf{u}'_n\}$ entonces, la matriz de los coeficientes del sistema (2) será diagonal, con lo que tal sistema queda reducido a:

$$\begin{pmatrix} \langle \mathbf{u'}_1, \mathbf{u'}_1 \rangle & 0 & \dots & 0 \\ 0 & \langle \mathbf{u'}_2, \mathbf{u'}_2 \rangle & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \langle \mathbf{u'}_n, \mathbf{u'}_n \rangle \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \dots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} \langle \mathbf{f}, \mathbf{u'}_1 \rangle \\ \langle \mathbf{f}, \mathbf{u'}_2 \rangle \\ \dots \\ \langle \mathbf{f}, \mathbf{u'}_n \rangle \end{pmatrix} \tag{3}$$

siendo las soluciones del mismo:

$$\lambda_1 = \frac{\langle \mathbf{f}, \mathbf{u'}_1 \rangle}{\langle \mathbf{u'}_1, \mathbf{u'}_1 \rangle} = \frac{\langle \mathbf{f}, \mathbf{u'}_1 \rangle}{\| \mathbf{u'}_1 \|^2}, \quad \lambda_2 = \frac{\langle \mathbf{f}, \mathbf{u'}_2 \rangle}{\langle \mathbf{u'}_2, \mathbf{u'}_2 \rangle} = \frac{\langle \mathbf{f}, \mathbf{u'}_2 \rangle}{\| \mathbf{u}_2 \|^2}, \dots,$$
$$\lambda_n = \frac{\langle \mathbf{f}, \mathbf{u'}_n \rangle}{\langle \mathbf{u'}_n, \mathbf{u'}_n \rangle} = \frac{\langle \mathbf{f}, \mathbf{u'}_n \rangle}{\| \mathbf{u'}_n \|^2} \tag{4}$$

con lo que se ha simplificado notablemente la resolución del problema.

3.2. APROXIMACIÓN MÍNIMO-CUADRÁTICA CONTINUA MEDIANTE POLINOMIOS

Nos encontramos muchas veces ante la conveniencia de aproximar una función f(x) mediante un polinomio de cierto grado con coeficientes reales $p(x) = a_0 \cdot x^n + a_1 \cdot x^{n-1} + \dots + a_{n-1} \cdot x + a_n$. Este problema se puede considerar como un caso particular de aproximación mínimo-cuadrática.

Consideramos como espacio vectorial euclídeo E el espacio de funciones continuas sobre un intervalo I = [a, b], es decir,

$$E = \boldsymbol{C}[a,b] = \{ f : I \longrightarrow \mathbb{R} \ / \ f \text{ continua} \}$$

con el producto escalar:

$$\langle f, g \rangle = \int_a^b w(x) f(x) g(x) dx \tag{5}$$

donde w(x) es una función llamada ***función de peso*** que verifica:

i) w(x) continua $\forall\ x \in I$.

ii) $w(x) > 0\ \forall\ x \in I$.

iii) w(x) es integrable en el intervalo I.

Consideramos como subespacio vectorial H de dimensión finita el espacio de los polinomios de grado menor o igual que n, es decir, $H = \mathbb{P}_n = \text{Span}\{1, x, x^2, \dots, x^n\}$. Entonces, los resultados del apartado anterior garantizan la existencia y unicidad del polinomio de grado menor o igual que n mejor aproximación mínimo-cuadrática de

cualquier función $f \in \boldsymbol{C}[a,b]$. Tal polinomio, se calculará mediante la resolución del sistema de ecuaciones normales (2) que en este caso es de la forma:

$$\begin{pmatrix} \int_a^b w(x)dx & \int_a^b w(x)xdx & \cdots & \int_a^b w(x)x^n dx \\ \int_a^b w(x)xdx & \int_a^b w(x)x^2 dx & \cdots & \int_a^b w(x)x^{n+1} dx \\ \vdots & \vdots & \cdots & \vdots \\ \int_a^b w(x)x^n dx & \int_a^b w(x)x^{n+1} dx & \cdots & \int_a^b w(x)x^{2n} dx \end{pmatrix} \begin{pmatrix} \lambda_0 \\ \lambda_1 \\ \vdots \\ \lambda_n \end{pmatrix} = \begin{pmatrix} \int_a^b w(x)f(x)dx \\ \int_a^b w(x)f(x)xdx \\ \vdots \\ \int_a^b w(x)f(x)x^n dx \end{pmatrix} \tag{6}$$

Observación: En la resolución de este sistema para valores de n no demasiado grandes, e incluso para el caso en que w(x) = 1, aparecen grandes dificultades numéricas. Esto se debe a que la matriz del sistema está mal condicionada. Por ello, para evitar los problemas computacionales, es conveniente utilizar una base de $\mathbb{P}_n$ formada por polinomios ortogonales respecto a la función de peso w(x), para lo cual basta ortogonalizar la base de polinomios $\{1, x, x^2, \cdots, x^n\}$ por el procedimiento de Gram-Schmidt. Como ya vimos en el apartado anterior, con una base ortogonal el sistema de ecuaciones normales se transforma en un sistema diagonal de resolución inmediata. En esta situación, el polinomio mejor aproximación $p_n^*(x)$ de una función $f \in \boldsymbol{C}[a,b]$ puede expresarse en la forma:

$$p_n^*(x) = \lambda_0 \cdot p_0(x) + \lambda_1 \cdot p_1(x) + \cdots + \lambda_n \cdot p_n(x) \tag{7}$$

siendo $p_0(x), p_1(x), \cdots, p_n(x)$ la base de polinomios ortogonales de $\mathbb{P}_n$ con respecto a la función de peso w(x) considerada, y los coeficientes λ_j según la expresión (4) serán:

$$\lambda_j = \frac{\langle f, \mathbf{u}'_j \rangle}{\| \mathbf{u}'_j \|^2} = \frac{\langle f, p_j(x) \rangle}{\| p_j(x) \|^2} = \frac{\int_a^b w(x)f(x)p_j(x)dx}{\int_a^b w(x)p_j^2(x)dx} \quad \forall j = 0,1,\ldots n \tag{8}$$

Si además se eligen los polinomios $p_0(x), p_1(x), \cdots, p_n(x)$ ortonormados, es decir de norma 1, entonces:

$$\lambda_j = \int_a^b w(x)f(x)p_j(x)dx \quad \forall j = 0,1,\ldots n \tag{9}$$

Se puede demostrar que, cuando $n\longrightarrow\infty$, el polinomio mejor aproximación $p_n^*(x)$ converge hacia la función f(x) en la norma inducida por el producto escalar, es decir, se cumple:

$$\lim_{n\to\infty}\| f-p_n^*\|=\lim_{n\to\infty}\left(\int_a^b w(x)\left(f(x)-p_n^*(x)\right)^2 dx\right)^{1/2}=0 \qquad (10)$$

teniéndose que el error cometido en la aproximación es:

$$e_n=\| f-p_n^*\|=\left(\int_a^b w(x)\left(f(x)-p_n^*(x)\right)^2 dx\right)^{1/2} \qquad (11)$$

Esta cantidad recibe el nombre de ***error mínimo-cuadrático*** y tiende hacia cero al aumentar el grado del polinomio.

Ejemplo 1:

Se trata de hallar la aproximación parabólica óptima en el sentido de los mínimos cuadrados de la función e^x en el intervalo [-1,1] para la función de peso w(x)=1, resolviendo el sistema de ecuaciones normales.

El subespacio aproximante H es $\mathbb{P}_2=\text{Span}\{1,x,x^2\}$ y el producto escalar considerado $\langle f,g\rangle=\int_a^b f(x)\,g(x)\,dx$, con lo que el sistema de ecuaciones normales queda:

$$\begin{pmatrix} \int_{-1}^1 dx & \int_{-1}^1 x\,dx & \int_{-1}^1 x^2dx \\ \int_{-1}^1 x\,dx & \int_{-1}^1 x^2dx & \int_{-1}^1 x^3dx \\ \int_{-1}^1 x^2dx & \int_{-1}^1 x^3dx & \int_{-1}^1 x^4dx \end{pmatrix}\begin{pmatrix}\lambda_0\\ \lambda_1\\ \lambda_2\end{pmatrix}=\begin{pmatrix} \int_{-1}^1 e^x dx \\ \int_{-1}^1 xe^x dx \\ \int_{-1}^1 x^2e^x dx\end{pmatrix}\Leftrightarrow\begin{pmatrix} 2 & 0 & \frac{2}{3}\\ 0 & \frac{2}{3} & 0\\ \frac{2}{3} & 0 & \frac{2}{5}\end{pmatrix}\begin{pmatrix}\lambda_0\\ \lambda_1\\ \lambda_2\end{pmatrix}=\begin{pmatrix} e-e^{-1}\\ 2e^{-1}\\ e-5e^{-1}\end{pmatrix}$$

Resolviendo este sistema obtenemos:

$\lambda_0=\dfrac{-3e+33e^{-1}}{4}$, $\lambda_1=3e^{-1}$, $\lambda_2=\dfrac{15}{4}(e-7e^{-1})$, con lo que el polinomio de grado 2 mejor aproximación de e^x es: $p_2^*(x)=\dfrac{-3e+33e^{-1}}{4}+3e^{-1}x+\dfrac{15}{4}(e-7e^{-1})x^2$.

Ejemplo 2:

Se trata de hallar el polinomio de segundo grado mejor aproximación por mínimos cuadrados de la función $f(x)=x^{1/3}$ para la función de peso w(x)=1 en [-1,1], utilizando polinomios ortogonales.

Hay que empezar ortogonalizando la base usual de los polinomios $\{1,x,x^2\}$, para el producto escalar $\langle f,g\rangle=\int_a^b f(x)\,g(x)\,dx$. Llamamos $\{q_0(x),q_1(x),q_2(x)\}$ a la base ortogonal de $\mathbb{P}_2$, la cual, utilizando el procedimiento de Gram-Schmidt es:

$$q_0(x)=1$$

$$q_1(x)=x+\alpha_1 q_0(x)=x+\alpha_1 \text{ tal que}$$

$$\langle q_0(x),q_1(x)\rangle=0=\int_{-1}^{1}(x+\alpha_1)dx=2\alpha_1\Rightarrow\alpha_1=0\Rightarrow q_1(x)=x$$

$$q_2(x)=x^2+\alpha_2 q_1(x)+\alpha_3 q_0(x)=x^2+\alpha_2 x+\alpha_3 \text{ tal que}$$

$$\langle q_2(x),q_0(x)\rangle=0=\int_{-1}^{1}(x^2+\alpha_2 x+\alpha_3)dx=\frac{2}{3}+2\alpha_3\Rightarrow\alpha_3=-\frac{1}{3}$$

$$\langle q_2(x),q_1(x)\rangle=0=\int_{-1}^{1}(x^2+\alpha_2 x+\alpha_3)xdx=\int_{-1}^{1}(x^3+\alpha_2 x^2+\alpha_3 x)dx=$$

$$=\frac{2}{3}\alpha_2\Rightarrow\alpha_2=0\Rightarrow q_2(x)=x^2-\frac{1}{3}$$

Para calcular la base ortonormal $\{p_0(x),p_1(x),p_2(x)\}$ de $\mathbb{P}_2$ basta hacer:

$$p_0(x)=\frac{q_0(x)}{\|q_0(x)\|}=\frac{1}{\left(\int_{-1}^{1}dx\right)^{1/2}}=\frac{1}{\sqrt{2}}$$

$$p_1(x)=\frac{q_1(x)}{\|q_1(x)\|}=\frac{x}{\left(\int_{-1}^{1}x^2dx\right)^{1/2}}=\frac{x}{\sqrt{\frac{2}{3}}}=\sqrt{\frac{3}{2}}x$$

$$p_2(x)=\frac{q_2(x)}{\|q_2(x)\|}=\frac{x^2-\frac{1}{3}}{\left(\int_{-1}^{1}\left(x^2-\frac{1}{3}\right)^2dx\right)^{1/2}}=\frac{x^2-\frac{1}{3}}{\sqrt{\frac{8}{45}}}=\sqrt{\frac{45}{8}}\left(x^2-\frac{1}{3}\right)$$

Luego la base ortonormal es $\left\{\frac{1}{\sqrt{2}},\sqrt{\frac{3}{2}}x,\sqrt{\frac{45}{8}}\left(x^2-\frac{1}{3}\right)\right\}$

Utilizando esta base, el polinomio mejor aproximación de la función $f(x)=x^{1/3}$ es:

$$p_2^*(x)=\lambda_0\frac{1}{\sqrt{2}}+\lambda_1\sqrt{\frac{3}{2}}x+\lambda_2\sqrt{\frac{45}{8}}\left(x^2-\frac{1}{3}\right)$$

con los λ_j dados por la expresión (9), esto es: $\lambda_j=\int_{-1}^{1}x^{1/3}p_j(x)dx\quad\forall j=0,1,2\ \Rightarrow$

$$\lambda_0=\int_{-1}^{1}x^{1/3}\frac{1}{\sqrt{2}}dx=0,\ \lambda_1=\int_{-1}^{1}x^{1/3}\sqrt{\frac{3}{2}}xdx=3\frac{\sqrt{6}}{7},\ \lambda_2=\int_{-1}^{1}x^{1/3}\sqrt{\frac{45}{8}}\left(x^2-\frac{1}{3}\right)dx=0$$

con lo que el polinomio mejor aproximación de $x^{1/3}$ en [-1,1] es:

$$p_2^*(x) = 3\frac{\sqrt{6}}{7}\sqrt{\frac{3}{2}}x = \frac{9}{7}x$$

3.2.1. Bases ortogonales. Polinomios ortogonales

Existen familias de polinomios ortogonales respecto de determinadas funciones de peso que han sido utilizadas y estudiadas ampliamente a lo largo de la literatura de las matemáticas. Pueden obtenerse, según acabamos de ver, ortogonalizando la base usual $\{1, x, x^2, \cdots, x^n\}$ de los polinomios de $\mathbb{P}_n$, por el procedimiento de Gram-Schmidt, pero todos estos sistemas de polinomios ortogonales verifican ciertas relaciones de recurrencia que permiten obtenerlos de una forma más cómoda. Algunos de los polinomios ortogonales más utilizados son los siguientes:

Polinomios de Legendre: Son polinomios ortogonales en el intervalo [-1, 1] respecto a la función de peso $w(x) = 1$. Para determinarlos unívocamente hay que imponer alguna condición de estandarización; la más utilizada es imponer $p_n(1)=1$. Bajo esta condición la relación de recurrencia que verifican estos polinomios es:

$$(n+1)\cdot p_{n+1}(x) = (2n+1)x\cdot p_n(x) - n\cdot p_{n-1}(x)$$

siendo $p_0(x) = 1$, $p_1(x) = x$.

Esta relación puede considerarse como un algoritmo para la obtención de todos los polinomios de Legendre. Así dando valores a n, quedan los polinomios:

$$p_0(x) = 1$$
$$p_1(x) = x$$
$$p_2(x) = \frac{3}{2}\cdot x^2 - \frac{1}{2}$$
$$p_3(x) = \frac{5}{2}\cdot x^3 - \frac{3}{2}\cdot x$$
$$p_4(x) = \frac{35}{8}\cdot x^4 - \frac{15}{4}\cdot x^2 + \frac{3}{8}$$
$$p_5(x) = \frac{63}{8}\cdot x^5 - \frac{35}{4}\cdot x^3 + \frac{15}{8}\cdot x$$

Puede deducirse fácilmente que la norma de ellos es:

$$\| p_n \|^2 = \int_{-1}^{1} p_n^2(x)dx = \frac{2}{2n+1}$$

Polinomios de Chebyshev de primera especie:

Forman un sistema ortogonal para la función de peso $w(x)=\frac{1}{\sqrt{1-x^2}}$ en el intervalo $(-1,1)$. Se pueden expresar en la forma:

$$T_n(x)=\cos\left[n\cdot\arccos(x)\right] \quad n=0,1,\cdots.$$

La relación de recurrencia verificada por esta familia de polinomios es:

$$T_{n+1}(x)=2x\cdot T_n(x)-T_{n-1}(x)$$

siendo $T_0(x)=1, \quad T_1(x)=x.$

Los polinomios que se van obteniendo mediante esta relación de recurrencia son:

$$\begin{aligned}
&T_0(x)=1\\
&T_1(x)=x\\
&T_2(x)=2x^2-1\\
&T_3(x)=4x^3-3x\\
&T_4(x)=8x^4-8x^2+1\\
&T_5(x)=16x^5-20x^3+5x
\end{aligned}$$

Puede comprobarse que su norma es en este caso:

$$\|T_n(x)\|^2=\begin{cases}\pi/2 & \text{si } n\neq 0\\ \pi & \text{si } n=0\end{cases}$$

Polinomios de Laguerre: Forman un sistema ortogonal para la función de peso $w(x)=e^{-x}$ en el intervalo $(0,\infty)$. Verifican la siguiente ley de recurrencia:

$$L_{n+1}(x)=[(2n+1)-x]\cdot L_n(x)-n^2\cdot L_{n-1}(x)$$
$$\text{siendo } L_0(x)=1, \quad L_1(x)=1-x$$

A partir de esta relación de recurrencia los polinomios que se obtienen son:

$$\begin{aligned}
&L_0(x)=1\\
&L_1(x)=1-x\\
&L_2(x)=x^2-4x+2\\
&L_3(x)=-x^3+9x^2-18x+6\\
&L_4(x)=x^4-16x^3+72x^2-96x+24\\
&L_5(x)=-x^5+25x^4-200x^3+600x^2-600x+120
\end{aligned}$$

Además se puede comprobar que su norma es: $\|L_n(x)\|^2=(n!)^2$.

Polinomios de Hermite: Son ortogonales respecto a la función de peso $w(x)=e^{-x^2}$ en el intervalo $(-\infty,\infty)$. Verifican la siguiente ley de recurrencia:

$$H_{n+1}(x)=2x\cdot H_n(x)-2n\cdot H_{n-1}(x)$$

siendo $H_0(x)=1, \quad H_1(x)=2x$.

A partir de esta relación de recurrencia los polinomios que se obtienen son:

$$H_0(x)=1$$
$$H_1(x)=2x$$
$$H_2(x)=4x^2-2$$
$$H_3(x)=8x^3-12x$$
$$H_4(x)=16x^4-48x^2+12$$
$$H_5(x)=32x^5-160x^3+120x$$

Además $\|H_n(x)\|^2=\sqrt{\pi}\cdot 2^n\cdot n!$.

Observación: Si quisiéramos realizar una aproximación en un intervalo [a,b] diferente a los expresados en estos polinomios ortogonales, deberíamos realizar el correspondiente cambio de variable lineal para transformar los intervalos.

Ejemplo.

Se trata de hallar el polinomio de segundo grado mejor aproximación mínimo cuadrática de la función |x| en el intervalo (-1,1) usando como función de peso $w(x)=\dfrac{1}{\sqrt{1-x^2}}$.

Los polinomios ortogonales para la función de peso $w(x)=\dfrac{1}{\sqrt{1-x^2}}$, según acabamos de ver, son los polinomios de Chebyshev de primera especie, los cuales quedaban, a partir de la relación de recurrencia:

$$\left\{T_0(x)=1, \quad T_1(x)=x, \quad T_2(x)=2x^2-1\right\}.$$

Considerando este sistema como base del espacio aproximante, el polinomio mejor aproximación de la función f(x)=|x| será de la forma:

$$p_2^*(x)=\lambda_0T_0(x)+\lambda_1T_1(x)+\lambda_2T_2(x)=\lambda_0+\lambda_1x+\lambda_2(2x^2-1)$$

con los λ_j dados por la expresión (8), esto es:

$$\lambda_j = \frac{\int_a^b w(x)f(x)p_j(x)dx}{\int_a^b w(x)p_j^2(x)dx} = \frac{\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} |x| T_j(x)dx}{\| T_j(x) \|^2} \quad \forall j = 0,1,2 \Rightarrow$$

$$\lambda_0 = \frac{\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} |x| dx}{\| T_0(x) \|^2} = \frac{2}{\pi} \qquad \lambda_1 = \frac{\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} x|x| dx}{\| T_1(x) \|^2} = \frac{0}{\pi/2} = 0$$

$$\lambda_2 = \frac{\int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} (2x^2-1)|x| dx}{\| T_1(x) \|^2} = \frac{2/3}{\pi/2} = \frac{4}{3\pi}$$

Luego el polinomio mejor aproximación es:

$$p_2^*(x) = \frac{2}{\pi} + \frac{4}{3\pi}(2x^2-1)$$

3.3. APROXIMACIÓN MÍNIMO-CUADRÁTICA DISCRETA. PROBLEMAS DE AJUSTE

El problema de la aproximación mínimo-cuadrática discreta o problema de ajuste, consiste en aproximar unos valores previamente dados, que normalmente representarán algún fenómeno de carácter experimental, mediante una función.

Así por ejemplo, si al realizar un experimento obtenemos unos valores $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$ que estén más o menos alineados, podemos pensar en aproximar estos datos por una función lineal de la forma $y = C + D \cdot x$. Si la relación existente entre los datos del experimento es en realidad lineal y no hay error experimental, entonces no habrá problema para estimar C y D. Pero si existe algún error, los datos obtenidos no estarán exactamente sobre una recta. La cuestión entonces es cómo calcular las constantes C y D a partir de tales resultados experimentales para que la recta $y = C + D \cdot x$ esté lo más próxima a ellos. A esta recta en estadística se la conoce con el nombre de *recta de regresión*. Este sería un problema de ajuste lineal. Sin embargo, existen problemas de ajuste que no son lineales. Así, en muchos experimentos no hay razón para esperar una relación lineal, pudiéndose tener, por ejemplo, una relación polinómica de cualquier grado $y = a_0 + a_1 x + a_2 x^2 + \ldots + a_n x^n$, casos que también vamos a

considerar aquí. En definitiva, un problema de ajuste consistirá en resolver un sistema de la forma:

$$\begin{cases} a_0 + a_1 \cdot x_1 + a_2 \cdot x_1^2 + \cdots + a_n \cdot x_1^n = y_1 \\ a_0 + a_1 \cdot x_2 + a_2 \cdot x_2^2 + \cdots + a_n \cdot x_2^n = y_2 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{siendo } m > n+1 \quad \Leftrightarrow \\ a_0 + a_1 \cdot x_m + a_2 \cdot x_m^2 + \cdots + a_n \cdot x_m^n = y_m \end{cases}$$

$$a_0 \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{pmatrix} + a_1 \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_m \end{pmatrix} + a_2 \begin{pmatrix} x_1^2 \\ x_2^2 \\ . \\ . \\ . \\ x_m^2 \end{pmatrix} + \cdots + a_n \begin{pmatrix} x_1^n \\ x_2^n \\ . \\ . \\ . \\ x_m^n \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_m \end{pmatrix}$$

Este es un sistema de m ecuaciones (número de datos experimentales) con n+1 incógnitas a_0, a_1, …, a_n. Si el número de datos experimentales m es mayor que n + 1 siendo n el grado del polinomio con el que queremos aproximar los datos, estamos ante un sistema sobredimensionado y por tanto puede no tener solución. En esta situación, vamos a formular el problema de la aproximación óptima de los coeficientes a_0, a_1, …, a_n en el sentido de los mínimos cuadrados.

Para ello, llamamos:

$$\underset{\sim}{x}_0 = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{pmatrix}, \underset{\sim}{x}_1 = \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_m \end{pmatrix}, \underset{\sim}{x}_2 = \begin{pmatrix} x_1^2 \\ x_2^2 \\ . \\ . \\ . \\ x_m^2 \end{pmatrix}, \cdots \underset{\sim}{x}_n = \begin{pmatrix} x_1^n \\ x_2^n \\ . \\ . \\ . \\ x_m^n \end{pmatrix}$$

que son vectores de $\mathbb{R}^m$ y además $\{\underset{\sim}{x}_0, \underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n\}$ son linealmente independientes por ser su matriz asociada una matriz de Vandermonde cuyo determinante es distinto de cero. Consideramos entonces el espacio vectorial $E=\mathbb{R}^m$ con el producto escalar usual, esto es:

$$\langle \underset{\sim}{f},\underset{\sim}{g}\rangle = \sum_{i=1}^{m} f_i \cdot g_i \quad \text{siendo } \underset{\sim}{f} = \begin{pmatrix} f_1 \\ f_2 \\ . \\ . \\ . \\ f_m \end{pmatrix}, \ \underset{\sim}{g} = \begin{pmatrix} g_1 \\ g_2 \\ . \\ . \\ . \\ g_m \end{pmatrix}$$

y como subespacio del mismo $H = \text{span}\{\underset{\sim}{x}_0, \underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n\}$. Según esto, el problema de aproximación a resolver es el siguiente:

Dado $\underset{\sim}{y} = \begin{pmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_m \end{pmatrix} \in \mathbb{R}^m$ encontrar un elemento $\mathbf{u} \in H /$

$$\mathbf{u} = \sum_{i=0}^{n} a_i \cdot \underset{\sim}{x}_i = a_0 \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{pmatrix} + a_1 \begin{pmatrix} x_1 \\ x_2 \\ . \\ . \\ . \\ x_m \end{pmatrix} + a_2 \begin{pmatrix} x_1^2 \\ x_2^2 \\ . \\ . \\ . \\ x_m^2 \end{pmatrix} + \cdots + a_n \begin{pmatrix} x_1^n \\ x_2^n \\ . \\ . \\ . \\ x_m^n \end{pmatrix} = \begin{pmatrix} a_0 + a_1 \cdot x_1 + a_2 \cdot x_1^2 + \cdots + a_n \cdot x_1^n \\ a_0 + a_1 \cdot x_2 + a_2 \cdot x_2^2 + \cdots + a_n \cdot x_2^n \\ . \\ . \\ . \\ a_0 + a_1 \cdot x_m + a_2 \cdot x_m^2 + \cdots + a_n \cdot x_m^n \end{pmatrix}$$

que minimice $\| \underset{\sim}{y} - \mathbf{u} \|^2 = \sum_{j=1}^{m} \left[y_j - (a_0 + a_1 \cdot x_j + \cdots + a_n \cdot x_j^n) \right]^2$. Por los resultados vistos en el primer apartado, este problema tiene solución única y dicha solución **u** está caracterizada por el hecho de que $\underset{\sim}{y} - \mathbf{u} \in H^{\perp}$, caracterización que como ya vimos nos lleva a resolver el sistema de ecuaciones normales (2), que en este caso queda:

$$\begin{pmatrix} \langle \underset{\sim}{x}_0, \underset{\sim}{x}_0 \rangle & \langle \underset{\sim}{x}_1, \underset{\sim}{x}_0 \rangle & \cdots & \langle \underset{\sim}{x}_n, \underset{\sim}{x}_0 \rangle \\ \langle \underset{\sim}{x}_0, \underset{\sim}{x}_1 \rangle & \langle \underset{\sim}{x}_1, \underset{\sim}{x}_1 \rangle & \cdots & \langle \underset{\sim}{x}_n, \underset{\sim}{x}_1 \rangle \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \langle \underset{\sim}{x}_0, \underset{\sim}{x}_n \rangle & \langle \underset{\sim}{x}_1, \underset{\sim}{x}_n \rangle & \cdots & \langle \underset{\sim}{x}_n, \underset{\sim}{x}_0 \rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ . \\ . \\ . \\ a_n \end{pmatrix} = \begin{pmatrix} \langle \underset{\sim}{y}, \underset{\sim}{x}_0 \rangle \\ \langle \underset{\sim}{y}, \underset{\sim}{x}_1 \rangle \\ . \\ . \\ . \\ \langle \underset{\sim}{y}, \underset{\sim}{x}_n \rangle \end{pmatrix} \Leftrightarrow$$

$$\begin{pmatrix} \sum_{i=1}^{m} 1 & \sum_{i=1}^{m} x_i & \cdots & \sum_{i=1}^{m} x_i^n \\ \sum_{i=1}^{m} x_i & \sum_{i=1}^{m} x_i^2 & \cdots & \sum_{i=1}^{m} x_i^{n+1} \\ \vdots & \vdots & \cdots & \vdots \\ \sum_{i=1}^{m} x_i^n & \sum_{i=1}^{m} x_i^{n+1} & \cdots & \sum_{i=1}^{m} x_i^{2n} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{m} y_i \\ \sum_{i=1}^{m} y_i \cdot x_i \\ \vdots \\ \sum_{i=1}^{m} y_i \cdot x_i^n \end{pmatrix} \tag{12}$$

Este es un sistema de n+1 ecuaciones con n+1 incógnitas, con lo cual hemos sustituido un sistema sobredimensionado por otro que no lo es.

Al igual que ocurría en la aproximación mínimo-cuadrática continua, el sistema de ecuaciones normales (12) está mal condicionado, por lo que en general este sistema no se suele resolver. Para solucionar esta situación, se va a volver a considerar una base de polinomios ortogonales, para lo cual es preciso plantear el problema que acabamos de resolver de la siguiente forma:

Dados los m datos de experimentación $(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)$, vamos a ***identificar*** cada función $f \in C[a,b]$ con el vector $\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \underset{\sim}{y} \in \mathbb{R}^m$, considerando el producto escalar $\langle f, g \rangle = \sum_{i=1}^{m} f(x_i) \cdot g(x_i)$ y la norma inducida por él: $\| f \|^2 = \sum_{i=1}^{m} f(x_i)^2$. Tomamos como subespacio $H = \mathbb{P}_n = \{1, x, x^2, \cdots, x^n\}$ con n<m. Entonces, dada $f \in C[a,b]$ o equivalentemente, dado el vector $\begin{pmatrix} f(x_1) \\ f(x_2) \\ \vdots \\ f(x_m) \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix} = \underset{\sim}{y} \in \mathbb{R}^m$, se trata de hallar el polinomio $p(x) = \sum_{i=0}^{n} a_i x^i \in H = \mathbb{P}_n$ de forma que se minimice

$$\| f-p \|^2 = \sum_{j=1}^{m} \left[f(x_j) - (a_0 + a_1 \cdot x_j + \cdots + a_n \cdot x_j^n) \right]^2 = \sum_{j=1}^{m} \left[y_j - (a_0 + a_1 \cdot x_j + \cdots + a_n \cdot x_j^n) \right]^2$$

Este problema es, claramente, equivalente al anterior, pues estamos identificando funciones con vectores. Entonces, bajo esta idea, en lugar de resolver el sistema de ecuaciones normales (12) se puede ortogonalizar el conjunto de polinomios $\{1, x, x^2, \cdots, x^n\}$ respecto del producto escalar $\langle f, g \rangle = \sum_{i=1}^{m} f(x_i) \cdot g(x_i)$ siendo $x_1, x_2, \ldots, x_m$ las abscisas de los datos experimentales, utilizando el procedimiento de ortogonalización de Gram-Schmidt. Llamando $\{p_0(x), p_1(x), \cdots, p_n(x)\}$ a la base de polinomios ortogonales, la solución del problema según la relación (4) del apartado 3.1 será:

$$p^*(x) = \sum_{j=0}^{n} \lambda_j \cdot p_j(x) \text{ donde } \lambda_j = \frac{\langle f, p_j \rangle}{\| p_j \|^2} = \frac{\sum_{i=1}^{m} f(x_i) p_j(x_i)}{\sum_{i=1}^{m} p_j^2(x_i)} \quad \forall j = 0,1,\cdots,n \qquad (13)$$

Para obtener estos polinomios ortogonales, al igual que comentamos en el caso de la aproximación mínimo cuadrática continua también se podrían utilizar relaciones de recurrencia, que permiten calcular tales familias de polinomios de una forma más eficiente.

Observación: Existen problemas de ajuste que no son lineales, pero que mediante algún cambio de variable pueden convertirse en lineales. Así por ejemplo, sabemos que la cantidad de sustancia radiactiva de un cuerpo se mide por una expresión de la forma $z = C \cdot e^{Dt}$ que no es una función lineal. Sin embargo, tomando logaritmos neperianos en la expresión anterior obtenemos L(z) = L(C) + Dt, que es ya una expresión lineal. De manera ánaloga:

Si queremos buscar una curva de la forma $y = b \cdot x^a$ que se ajuste a unos ciertos datos (x_i, y_i), tomando logaritmos neperianos queda $L(y) = L(b) + a \cdot L(x)$, la cual ya es una expresión lineal. Por tanto, aplicaríamos todo lo visto anteriormente al conjunto de datos $(L(x_i), L(y_i))$.

Si queremos encontrar una curva de la forma $y=\frac{1}{ax+b}$ haríamos el cambio $\overline{y}=\frac{1}{y}=ax+b$ que es ya lineal.

Si la función a determinar fuera de la forma $y=\frac{a}{x}+b$, haríamos el cambio $\overline{x}=\frac{1}{x}$ obteniendo que $y=a\cdot\overline{x}+b$.

Si buscáramos una función $y=\frac{x}{a+bx}$, hacemos los cambios $\begin{cases}\overline{y}=\frac{1}{y}\\ \overline{x}=\frac{1}{x}\end{cases}$ teniéndose entonces que $\overline{y}=a\cdot\overline{x}+b$.

Ejemplo 1:

Nos interesa obtener lo que se llama un circuito lineal equivalente de un sistema eléctrico. Aplicamos un voltaje E y anotamos el flujo de corriente que en correspondencia se origina en el sistema. Las anotaciones hechas son las siguientes:

E ≡ Voltaje en voltios	5	10	20
I ≡ Intensidad en amperios	2	3.9	8.2

El sistema sobredimensionado a resolver sería:

$$\begin{cases}a_0+5a_1=2\\ a_0+10a_1=3.9\\ a_0+20a_1=8.2\end{cases}\rightarrow \underset{\sim}{x}_0=\begin{pmatrix}1\\1\\1\end{pmatrix},\ \underset{\sim}{x}_1=\begin{pmatrix}5\\10\\20\end{pmatrix},\ \underset{\sim}{I}=\begin{pmatrix}2\\3.9\\8.2\end{pmatrix}$$

Como este sistema no tiene solución, aplicando lo que acabamos de ver, se trata de obtener el vector $\underset{\sim}{u}$ que sea la mejor aproximación de $\underset{\sim}{I}$ en $H=\text{Span}\{\underset{\sim}{x}_0,\underset{\sim}{x}_1\}$, para lo cual hay que resolver el sistema de ecuaciones normales (12). Los productos escalares quedan:

$$\langle \underset{\sim}{x}_0, \underset{\sim}{x}_0 \rangle = 1+1+1 = 3 \quad \langle \underset{\sim}{x}_0, \underset{\sim}{x}_1 \rangle = 5+10+20 = 35 \quad \langle \underset{\sim}{x}_1, \underset{\sim}{x}_1 \rangle = 5^2+10^2+20^2 = 525$$
$$\langle \underset{\sim}{I}, \underset{\sim}{x}_0 \rangle = 2+3.9+8.2 = 14.1 \qquad \langle \underset{\sim}{I}, \underset{\sim}{x}_1 \rangle = 2\cdot 5+3.9\cdot 10+8.2\cdot 20 = 213 \Rightarrow$$

El sistema de ecuaciones normales (12) se convierte en:

$$\begin{pmatrix} \langle \underset{\sim}{x}_0, \underset{\sim}{x}_0 \rangle & \langle \underset{\sim}{x}_1, \underset{\sim}{x}_0 \rangle \\ \langle \underset{\sim}{x}_0, \underset{\sim}{x}_1 \rangle & \langle \underset{\sim}{x}_1, \underset{\sim}{x}_1 \rangle \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} \langle \underset{\sim}{I}, \underset{\sim}{x}_0 \rangle \\ \langle \underset{\sim}{I}, \underset{\sim}{x}_1 \rangle \end{pmatrix} \Leftrightarrow \begin{pmatrix} 3 & 35 \\ 35 & 525 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 14.1 \\ 213 \end{pmatrix}$$

Resolviendo este sistema se obtiene $a_0 = -0.15$ $a_1 = 0.42$, es decir, polinomio de grado 1 mejor aproximación mínimo-cuadrática de I es $I \simeq 0.42E - 0.15$.

Ejemplo2:

Se lanza una pelota al aire. Sabiendo que la altura que alcanza está dada por la fórmula $S = S_0 + v_0 \cdot t - \frac{1}{2} g \cdot t^2$ siendo S_0 la altura inicial desde la que se lanzó la pelota, v_0 la velocidad inicial del lanzamiento y g = 9.8 m/s^2, se trata de estimar los valores S_0 y v_0. Para ello se realizaron las siguientes mediciones:

Tiempo transcurrido en segundos	**Altura alcanzada en metros**
1	57
1.5	67
2.5	68
4	9.5

Como la altura viene dada por la fórmula $S = S_0 + v_0 \cdot t - \frac{1}{2} g \cdot t^2$, se tratará de resolver el sistema:

$$\begin{cases} 57 = S_0 + v_0 - 4.9 \\ 67 = S_0 + 1.5v_0 - 11.025 \\ 68 = S_0 + 2.5v_0 - 30.625 \\ 9.5 = S_0 + 4v_0 - 78.4 \end{cases} \rightarrow \begin{cases} 61.9 = S_0 + v_0 \\ 78.025 = S_0 + 1.5v_0 \\ 98.625 = S_0 + 2.5v_0 \\ 87.9 = S_0 + 4v_0 \end{cases} \rightarrow$$

$$\underset{\sim}{x}_0 = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \ \underset{\sim}{x}_1 = \begin{pmatrix} 1 \\ 1.5 \\ 2.5 \\ 4 \end{pmatrix}, \ \underset{\sim}{y} = \begin{pmatrix} 61.9 \\ 78.025 \\ 98.627 \\ 87.9 \end{pmatrix}$$

Como este sistema no tiene solución, se trata de obtener el vector $\underset{\sim}{u} \in H = \text{Span}\{\underset{\sim}{x}_0, \underset{\sim}{x}_1\}$ que sea la mejor aproximación de $\underset{\sim}{y}$ en $H = \text{Span}\{\underset{\sim}{x}_0, \underset{\sim}{x}_1\}$ para lo cual hay que resolver el sistema de ecuaciones normales (12). Los productos escalares quedan:

$$\langle \underset{\sim}{x}_0, \underset{\sim}{x}_0 \rangle = 1+1+1+1 = 4 \quad \langle \underset{\sim}{x}_0, \underset{\sim}{x}_1 \rangle = 1+1.5+2.5+4 = 9$$
$$\langle \underset{\sim}{x}_1, \underset{\sim}{x}_1 \rangle = 1^2 + 1.5^2 + 2.5^2 + 4^2 = 25.5$$
$$\langle \underset{\sim}{y}, \underset{\sim}{x}_0 \rangle = 61.9 + 78.025 + 98.627 + 87.9 = 326.452$$
$$\langle \underset{\sim}{y}, \underset{\sim}{x}_1 \rangle = 61.9 \cdot 1 + 78.025 \cdot 1.5 + 98.627 \cdot 2.5 + 87.9 \cdot 4 = 777.105 \Rightarrow$$

El sistema de ecuaciones normales (12) se convierte en:

$$\begin{pmatrix} \langle \underset{\sim}{x}_0, \underset{\sim}{x}_0 \rangle & \langle \underset{\sim}{x}_1, \underset{\sim}{x}_0 \rangle \\ \langle \underset{\sim}{x}_0, \underset{\sim}{x}_1 \rangle & \langle \underset{\sim}{x}_1, \underset{\sim}{x}_1 \rangle \end{pmatrix} \begin{pmatrix} S_0 \\ v_0 \end{pmatrix} = \begin{pmatrix} \langle \underset{\sim}{S}, \underset{\sim}{x}_0 \rangle \\ \langle \underset{\sim}{S}, \underset{\sim}{x}_1 \rangle \end{pmatrix} \Leftrightarrow \begin{pmatrix} 4 & 9 \\ 9 & 25.5 \end{pmatrix} \begin{pmatrix} S_0 \\ v_0 \end{pmatrix} = \begin{pmatrix} 326.452 \\ 777.105 \end{pmatrix}$$

Resolviendo este sistema se obtiene $S_0 = 63.361$ $v_0 = 8.112$, es decir, la altura inicial desde la que se lanzó la pelota fue de 63.361 metros, la velocidad inicial del lanzamiento fue 8.112 m/s y la función buscada es: $S(t) \simeq 63.361 + 8.112 \cdot t - \frac{1}{2} g \cdot t^2$.

Ejemplo3:

Se trata de obtener el ajuste parabólico óptimo en el sentido de los mínimos cuadrados para los datos siguientes utilizando polinomios ortogonales:

x_i	-1	0	1	2	3
y_i	4	5	4	5	6

En este caso consideramos como subespacio aproximante $H=\text{Span}\{1,x,x^2\}$ y ortonormalizamos su base $\{1, x, x^2\}$ respecto el producto escalar $\langle f, g \rangle = \sum_{i=1}^{m} f(x_i) \cdot g(x_i)$ por el método de Gram-Schmidt. Para ello tomamos:

$$q_0(x)=1$$

$$q_1(x)=x+\alpha_1 q_0(x)=x+\alpha_1 \quad \text{tal que}$$

$$\langle q_0(x),q_1(x)\rangle=0=\sum_{i=1}^{5}(x_i+\alpha_1)\cdot 1=5\alpha_1+(-1+0+1+2+3)=$$

$$=5\alpha_1+5\Rightarrow \alpha_1=-1\Rightarrow q_1(x)=x-1$$

$$q_2(x)=x^2+\alpha_2 q_1(x)+\alpha_3 q_0(x)=x^2+\alpha_2(x-1)+\alpha_3 \text{ tal que}$$

$$\langle q_2(x),q_1(x)\rangle=0=\langle x^2+\alpha_2 q_1(x)+\alpha_3 q_0(x),q_1(x)\rangle=$$

$$=\langle x^2,q_1(x)\rangle+\alpha_2\langle q_1(x),q_1(x)\rangle=$$

$$=\sum_{i=1}^{5}x_i^2\cdot q_1(x_i)+\alpha_2\sum_{i=1}^{5}q_1(x_i)^2=$$

$$=\sum_{i=1}^{5}x_i^2\cdot(x_i-1)+\alpha_2\sum_{i=1}^{5}(x_i-1)^2=$$

$$=20+10\alpha_2\Rightarrow\alpha_2=-2$$

$$\langle q_2(x),q_0(x)\rangle=0=\langle x^2-2q_1(x)+\alpha_3 q_0(x),q_0(x)\rangle=$$

$$=\langle x^2,q_0(x)>+\alpha_3\langle q_0(x),q_0(x)\rangle=$$

$$=\sum_{i=1}^{5}x_i^2+\alpha_3\sum_{i=1}^{5}1^2=15+5\alpha_3\Rightarrow\alpha_3=-3\Rightarrow$$

$$q_2(x)=x^2-2(x-1)-3=x^2-2x-1$$

Para calcular el sistema ortonormal $\{p_0(x),p_1(x),p_2(x)\}$ basta hacer:

$$p_i(x)=\frac{q_i(x)}{\|q_i(x)\|}\ i=0,1,2\ \Rightarrow$$

$$p_0(x)=\frac{1}{\left(\sum_{i=1}^{5}1^2\right)^{1/2}}=\frac{1}{\sqrt{5}},$$

$$p_1(x)=\frac{x-1}{\left(\sum_{i=1}^{5}(x_i-1)^2\right)^{1/2}}=\frac{x-1}{\sqrt{10}},$$

$$p_2(x)=\frac{x^2-2x-1}{\left(\sum_{i=1}^{5}(x_i^2-2x_i-1)^2\right)^{1/2}}=\frac{x^2-2x-1}{\sqrt{14}}$$

Entonces, según lo que acabamos de ver en la relación (13), la solución del problema es:

$$p^*(x)=\sum_{j=0}^{2}\lambda_j\cdot p_j(x)\ \text{ con }\ \lambda_j=\frac{\langle f,p_j\rangle}{\|p_j\|^2}=\sum_{i=1}^{5}y_i\cdot p_j(x_i)\quad\forall\, j=0,1,2\ \Rightarrow$$

$$\lambda_0=\sum_{i=1}^{5}y_i\cdot\frac{1}{\sqrt{5}}=\frac{24}{\sqrt{5}},\ \lambda_1=\sum_{i=1}^{5}y_i\cdot\frac{(x_i-1)}{\sqrt{10}}=\frac{4}{\sqrt{10}},\ \lambda_2=\sum_{i=1}^{5}y_i\cdot\frac{(x_i^2-2x_i-1)}{\sqrt{14}}=\frac{2}{\sqrt{14}}$$

con lo que el ajuste parabólico óptimo queda:

$$p^*(x) = \frac{24}{5} + \frac{4}{10}(x-1) + \frac{2}{14}(x^2 - 2x - 1) = \frac{1}{7}x^2 + \frac{4}{35}x + \frac{149}{35}$$

EJERCICIOS TEMA 3

1.- Aproximar la función $f(x) = L(x)$ por una recta en el intervalo [1, 2], resolviendo el sistema de *ecuaciones normales*. Plantear la expresión del error mínimo cuadrático cometido en la aproximación.

2.- Encontrar, resolviendo el sistema de *ecuaciones normales*, la aproximación polinómica de mínimos cuadrados de grado uno de la función f(x) en el intervalo indicado:

a) $f(x)=x^3-1$ en [0, 2].

b) $f(x)=\cos(\pi x)$ en [0, 1].

3.- Dada la función $f(x)=x^4$ con $x\in[-1,1]$, hallar las aproximaciones de grado uno y dos mediante la técnica de mínimos cuadrados. Realizar primero el problema resolviendo el sistema de *ecuaciones normales* y después, utilizando polinomios de Legendre.

4.- Utilizando polinomios ortogonales de Legendre, aproximar la función $f(x)=1+x^3$ por un polinomio de grado menor o igual que dos en el intervalo [-1, 1].

5.- Aproximar, mediante el método de mínimos cuadrados, las funciones $y_1(t)=t^2$ y $y_2(t)=e^{-t}$ sobre el intervalo [0, 1] mediante una línea recta, utilizando polinomios de Legendre.

6.- Hallar la aproximación a $y(t) = \text{sen}(t)$ mediante mínimos cuadrados por medio de una parábola utilizando polinomios de Legendre en el intervalo $[0, \pi]$.

7.- Hallar el polinomio de segundo grado mejor aproximación mínimo cuadrática de la función $y = |x|$ en el intervalo (-1, 1), usando como función de peso $w(x)=\frac{1}{\sqrt{1-x^2}}$ y utilizando polinomios ortogonales.

8.- Hallar la línea de mínimos cuadrados que mejor se aproxime a $y(t)=t^2$ en el intervalo [0, 1] mediante la función de peso $w(x)=\frac{1}{\sqrt{1-x^2}}$ en el intervalo (-1, 1).

9.- Hallar el polinomio que mejor se ajusta a los datos de la tabla siguiente, utilizando la técnica de mínimos cuadrados (sin utilizar polinomios ortogonales).

x_i	**1**	**2**	**3**	**4**	**5**	**6**	**7**
y_i	4	7	9	10	9	7	4

10.- Encontrar la recta que mejor se ajusta a los datos de la siguiente tabla

x_i	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**
y_i	2	4	3	6	5	7	9	8

11.- Dada la siguiente tabla de datos, obtenida experimentalmente, hallar la constante g que relaciona las variables t y d ($d \simeq g\,t^2/2$)

t_i	**0.2**	**0.4**	**0.6**	**0.8**	**1.0**
d_i	0.1960	0.7850	1.7665	3.1405	4.9075

12.- Ajustar los datos de la tabla siguiente mediante una parábola considerando el método de mínimos cuadrados. Determinar el error mínimo cuadrático

x_i	**0.00**	**0.25**	**0.5**	**0.75**	**1.00**
y_i	1.0000	1.2840	1.6487	2.1170	2.7183

Operar con redondeo a 4 decimales.

13.- Dada la siguiente tabla

x_i	**1**	**2**	**3**
y_i	7	9	11

encontrar una curva por el procedimiento de mínimos cuadrados, que se ajuste a los datos, de la siguiente forma: a) y = ax b) y = ax + b.

14.- Dada la tabla

x_i	**1.00**	**1.25**	**1.50**	**1.75**	**2.00**
y_i	5.10	5.79	6.53	7.45	8.46

utilizar la técnica de mínimos cuadrados para obtener un ajuste aproximado a estos valores. Trabajar con redondeo a 5 decimales.

15.- Hallar la parábola que mejor se ajusta a los siguientes datos

x_i	**1**	**2**	**3**	**4**
y_i	2	4	6	8

16.- a) Encontrar una curva de la forma $y = \frac{1}{A \cdot x + B}$ por el procedimiento de mínimos cuadrados que se ajuste a los datos de la tabla

x_i	**-1**	**0**	**1**	**2**	**3**
y_i	6.62	3.94	2.17	1.35	0.89

b) Obtener el error mínimo cuadrático y el error cometido por el ajuste anterior al aproximar los valores y_i por la curva y.

Nota: Operar con redondeo a 6 decimales.

17.- Repetir el ejercicio anterior, con los datos de la tabla:

x_i	**0**	**1**	**2**	**3**
y_i	1.01	0.34	0.20	0.14

y operando con redondeo a 2 decimales.

18.- Cuando el crecimiento de una población está acotado por un valor constante L sigue una curva logística que tiene la forma $y(x) = \frac{L}{1 + B \cdot e^{Ax}}$. Se pide:

a) Realizar un cambio de variable que transforme la función y(x) en una expresión lineal en x (Tener en cuenta que $\frac{L}{y}-1=B\cdot e^{Ax}$).

b) Utilizar los datos de una población, dados por la tabla

Año	1800	1850	1900	1950
$\mathbf{x_k}$	-10	-5	0	5
$\mathbf{y_k}$ **(millones)**	5.3	23.2	76.1	152.3

para encontrar una curva logística y(x) correspondiente a L = 800 (millones), aplicando mínimos cuadrados sobre la función transformada del apartado a). Calcular el error cometido.

c) Estimar la población en el año 2000.

Nota: Operar con redondeo a 2 decimales.

19.- Dada la tabla de datos

$\mathbf{x_i}$	**10.0**	**10.2**	**10.4**	**10.6**	**10.8**	**11.0**
$\mathbf{f(x_i)}$	0.000	0.004	0.016	0.036	0.064	0.10

encontrar el polinomio de grado 2 que mejor se ajuste a estos datos, utilizando la técnica de mínimos cuadrados y polinomios ortogonales. Operar con redondeo a 7 dígitos significativos.

20.- Dada la tabla de datos

$\mathbf{x_i}$	**-1.0**	**0.0**	**0.5**	**1.0**
$\mathbf{y_i}$	1	3	0	4

encontrar el polinomio de grado 2 que mejor se ajuste a estos datos, utilizando la técnica de mínimos cuadrados y polinomios ortogonales. Operar con redondeo a 7 dígitos significativos.

SOLUCIONES EJERCICIOS TEMA 3

1.- $p_1(x) = -0.637056 + 0.682234 \cdot x$

2.-

a) $p_1(x) = -\frac{13}{5} + \frac{18}{5} \cdot x$

b) $p_1(x) = \frac{12}{\pi^2} \cdot (1 - 2x)$

3.- $p_1(x) = \frac{1}{5}, \quad p_2(x) = -\frac{3}{35} + \frac{6}{7} \cdot x^2 = \frac{1}{5} + \frac{4}{7} \cdot \left(-\frac{1}{2} + \frac{3}{2} \cdot x^2 \right)$

4.- $p_2(x) = 1 + \frac{3}{5} \cdot x$

5.- $p_1(t) = -\frac{1}{6} + t; \quad p_1(t) = \left(\frac{8}{e} - 2 \right) + \left(6 - \frac{18}{e} \right) \cdot t$

6.- $p_2(t) = \frac{2}{\pi} + \frac{5}{\pi} \cdot \left(1 - \frac{12}{\pi^2} \right) \cdot \left[3 \cdot \left(-1 + \frac{2}{\pi} \cdot t \right)^2 - 1 \right]$

7.- $p_2(x) = \frac{2}{\pi} + \frac{4}{3\pi} \cdot \left(-1 + 2x^2 \right) = \frac{2}{3\pi} \cdot \left(1 + 4x^2 \right)$

8.- $p_1(t) = -\frac{1}{8} + t$

9.- $p_2(x) = -\frac{4}{7} + \frac{36}{7} \cdot x - \frac{9}{14} \cdot x^2$

10.- $p_1(x) = \frac{9}{4} + \frac{13}{14} \cdot x$

11.- $g \simeq 9.8146$

12.-

$p_2(x) = 1.0046 + 0.8669x + 0.8413x^2$

Error : 0.0166

13.-

a) $y \simeq \frac{29}{7}x$

b) $y \simeq 2x + 5$

14.- $f(x) = 3.07243 \cdot e^{0.50573 \cdot x}$

15.- $p_2(x) = 2x$

16.- $y \simeq \frac{1}{0.302805 + 0.243201x}$

Error mínimo cuadrático: 0.169364

Error ajuste: 10.183655

17.- $y \simeq \frac{1}{0.94 + 2.05x}$

Error mínimo cuadrático: 0.09539

Error ajuste: 0.0541

18.-

b) $f(x) = \frac{800}{1 + 11.7 \cdot e^{-0.24 \cdot x}}$

Error mínimo cuadrático : 0.35

Error ajuste : 28.01

c) $y(x = 10) \simeq 388.09$

19.-

$p_2(x) = 0.03666667 + 0.1 \cdot (x - 10.5) + 0.09987722 \cdot \left[x^2 - 21x + 110.1333 \right] =$

$= 0.09987722x^2 - 1.997422x + 9.986475$

20.-

$$p_2(x) = 2 + 0.9142857 \cdot (x - 0.125) + 0.8181815 \cdot \left[x^2 + 0.07142857x - 0.5714286 \right] =$$

$$= 0.8181815x^2 + 0.9727272x + 1.418182$$

BIBLIOGRAFÍA

1. **BURDEN, R.L. y FAIRES, J.D.** *"Análisis Numérico"***.** Grupo Editorial Iberoamericana, S.A. de C.V.

2. **CHAPRA, S.C. y CANALE, R.P.** *"Métodos numéricos para ingenieros"*. McGraw-Hill.

3. **CONTE, S.D. y de BOOR, C.** *"Análisis Numérico"*. McGraw-Hill.

4. **GRAU SÁNCHEZ M. y NOGUERA BATLLE M.** *"Cálculo Numérico: teoría y práctica"*. Edicions UPC.

5. **KINCAID, D. y CHENEY, W.** *"Análisis Numérico. Las matemáticas del cálculo científico"***.** Addison-Wesley Iberoamericana.

6. **MATHEWS J. H. y FINK K.D.** *"Métodos numéricos con Matlab"*. Prentice-Hall.