MATEMÁTICA SUPERIOR APLICADA

SOLUCIÓN DE SISTEMAS DE ECUACIONES ALGEBRAICAS LINEALES

Planteo del Problema. Teoremas Básicos

Se trata de resolver *n* ecuaciones lineales simultáneas con *n* incógnitas:

$$\sum_{j=1}^{n} a_{ij} x_j = b_i \ \ con\ i = 1,2,...,n \tag{1}$$

o, en su forma matricial compacta:

$$\underline{\underline{A}} . \underline{x} = \underline{b} \tag{2}$$

donde $\underline{\underline{A}} = [a_{ij}]$ es la matriz de coeficientes (cuadrada, nxn) y $\underline{x}^t = (x_1, ..., x_n)$ es el vector de incógnitas, siendo $\underline{b}^t = (b_1, ..., b_n)$ el vector de términos independientes. Para el caso que nos interesa, tanto la matriz $\underline{\underline{A}}$ como el vector $\underline{b}$ tienen componentes reales. En adelante se obviará la notación transpuesta de la matriz, suponiendo que en las operaciones entre matrices se disponen éstas de tal forma que sean compatibles para las mismas. Si definimos la denominada matriz aumentada como:

$$\underline{\underline{A}}_b = \left[\underline{\underline{A}}\ \underline{b} \right] \ Matriz\ ampliada\ de\ \underline{\underline{A}}\ \left[n\ x\ (n+1) \right] \tag{3}$$

dado que $\underline{b}$ es un vector columna; y recordando la definición de rango de una matriz: $r(\underline{\underline{A}})$, el teorema básico de existencia de solución establece:

1) El sistema de ecuaciones (Ecuación (2)) tiene solución sí y sólo sí: $r(\underline{\underline{A}}) = r(\underline{\underline{A}}_b)$.
2) Si $r(\underline{\underline{A}}) = r(\underline{\underline{A}}_b) = k < n$, luego las $x_1, x_2, ..., x_k$ son variables cuyas columnas son linealmente independientes en $\underline{\underline{A}}$, de modo que las restantes $(n-k)$ variables pueden asignarse arbitrariamente. O dicho de otra forma, hay una familia paramétrica de $(n-k)$ soluciones.
3) Si $r(\underline{\underline{A}}) = r(\underline{\underline{A}}_b) = n$, hay una única solución.

Corolario: Para el caso homogéneo (**$\underline{b} = \underline{0}$**), o sea **$\underline{\underline{A}} . \underline{x} = \underline{0}$** habrá solución no trivial, si y sólo si $r(\underline{\underline{A}}) < n$.

Para estos problemas, cosa que no ocurre en el caso no lineal, existe solución analítica (recordemos la denominada *Regla de Cramer*), pero la dificultad reside principalmente en computar esa solución. La evaluación de determinantes no hace práctico dicho procedimiento analítico; luego, el problema es desarrollar algoritmos computacionales más eficientes, es decir que sean más rápidos, sobre todo en el número de operaciones necesarias y que además

sean robustos de modo que la solución calculada sea lo más precisa posible.

Un punto vital es discutir cómo se espera que sea la matriz de coeficientes. En general, puede encontrarse entre alguna de estas dos categorías:

i. Llena pero no muy grande. Es decir con muy pocos ceros, y en donde *n* no sea mayor que *100*, por ejemplo.
ii. Dispersa y relativamente muy grande, denominadas también ralas. En estos casos, son muy pocos (en relación al orden) los elementos distintos de cero y *n* puede ser mayor a *1000*.

Naturalmente, los métodos desarrollados deben estar dirigidos a resolver alguna de estas dos categorías, y si es posible haciendo uso de sus características para incrementar su eficiencia.

Un problema, cual es la condición del sistema, se puede analizar considerando un vector residual $\underline{r}$, cuando se tiene una solución calculada $\underline{x}^{(c)}$. Esto es:

$$\underline{r} = \underline{b} - \underline{\underline{A}}.\underline{x}^{(c)} \tag{4}$$

se sabe que si $\underline{x}^*$ es la solución exacta del SEAL:

$$\underline{\underline{A}}.\underline{x}^* = \underline{b} \; o \; bien \; \underline{r} = \underline{b} \; - \underline{\underline{A}}.\underline{x}^* = \underline{0} \tag{5}$$

entonces será:

$$\underline{r} = \underline{\underline{A}}.\left(\underline{x}^* - \underline{x}^{(c)}\right)$$

luego,

$$\left(\underline{x}^* - \underline{x}^{(c)}\right) = \underline{\underline{A}}^{-1} \cdot \underline{r} \tag{6}$$

de donde se ve que aunque $\underline{r}$ tenga elementos muy chicos, si $\underline{\underline{A}}^{-1}$ (matriz inversa) contiene coeficientes muy grandes, la diferencia entre $\underline{x}^*$ y $\underline{x}^{(c)}$ puede ser aún muy grande. Esto permite anticipar la importancia de un escalado en los coeficientes de la matriz $\underline{\underline{A}}$ original, ya que aunque el vector residual $\underline{r}$ impuesto sea pequeño, el error encontrado para la solución puede ser muy grande.

Métodos Directos

Un método directo para hallar la solución es uno en el cual, si todos los cálculos (computaciones) fueran llevados a cabo sin error de redondeo conduciría a la solución exacta del sistema dado. Prácticamente todos están basados en la *técnica de eliminación*. El error de truncamiento para estos métodos es intrascendente.

Eliminación Gaussiana

Desarrollando la Ecuación (1), se obtiene el sistema en la siguiente forma:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 = a_{1,n+1} \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \cdots + a_{2n}x_n &= b_2 = a_{2,n+1} \\ \cdots\cdots\cdots & \\ \cdots\cdots\cdots & \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \cdots + a_{nn}x_n &= b_n = a_{n,n+1} \end{aligned} \tag{7}$$

Se supone que la matriz es no singular, y que $a_{11} \neq 0$ de manera de poder dividir la primer columna por a_{11} y así restar para las ecuaciones, donde $i=2, ..., n$. Esto da como resultado:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 = a_{1,n+1} \\ a_{22}^{(1)}x_2 + a_{22}^{(1)}x_3 + \cdots + a_{2n}^{(1)}x_n &= b_2^{(1)} = a_{2,n+1}^{(1)} \\ \cdots\cdots\cdots & \\ \cdots\cdots\cdots & \\ a_{n2}^{(1)}x_2 + \cdots\cdots\cdots\cdots + a_{nn}^{(1)}x_n &= b_n^{(1)} = a_{n,n+1}^{(1)} \end{aligned} \tag{8}$$

siendo los $a^{(1)}_{ij}$, tal que:

$$a_{ij}^{(1)} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j} \; ; \; i = 2, \cdots, n \; ; \; j = 2, \cdots, (n+1) \tag{9}$$

si fuese $a_{11} = 0$ se intercambian columnas, y se opera en consecuencia. Por comodidad se define: $m_{i1} = a_{i1} / a_{11}$ con $i=2, ... , n$.

De igual forma se continúa con el procedimiento haciendo ahora (si $a^{(1)}_{22} \neq 0$) $m_{i2} = a^{(1)}_{i2} / a^{(1)}_{22}$ con $i=3, ..., n$ resultando, al restar, el siguiente sistema:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots + a_{1n}x_n &= b_1 = a_{1,n+1} \\ a_{22}^{(1)}x_2 + a_{22}^{(1)}x_3 + \cdots + a_{2n}^{(1)}x_n &= b_2^{(1)} = a_{2,n+1}^{(1)} \\ a_{32}^{(2)}x_3 + \cdots + a_{3n}^{(2)}x_n &= b_3^{(2)} = a_{3,n+1}^{(2)} \\ \cdots\cdots\cdots & \\ a_{n2}^{(2)}x_3 + \cdots + a_{nn}^{(2)}x_n &= b_n^{(2)} = a_{n,n+1}^{(2)} \end{aligned} \tag{10}$$

donde:

$$a_{ij}^{(2)} = a_{ij}^{(1)} - m_{i2} a_{2j}^{(1)} \; ; \; i = 3, \cdots, n \; ; \; j = 3, \cdots, (n+1)$$

Y, continuando con el procedimiento hasta *(n-1)* pasos llegamos al sistema final:

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \cdots\cdots + a_{1n}x_n &= b_1 = a_{1,n+1} \\
a_{22}^{(1)}x_2 + a_{22}^{(1)}x_3 + \cdots\cdots + a_{2n}^{(1)}x_n &= b_2^{(1)} = a_{2,n+1}^{(1)} \\
+ a_{32}^{(2)}x_3 + \cdots\cdots + a_{3n}^{(2)}x_n &= b_3^{(2)} = a_{3,n+1}^{(2)} \\
\cdots\cdots\cdots\cdots & \\
+ a_{nn}^{(n-1)}x_n &= b_n^{(n-1)} = a_{n,n+1}^{(n-1)}
\end{aligned}
\tag{11}
$$

con los elementos en la diagonal (distintos de cero), tal que:

$$
\begin{aligned}
& a_{ij}^{(k)} = a_{ij}^{(k-1)} - m_{ik}a_{kj}^{(k-1)} \;\; ; \;\; m_{ik} = \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} \\
& k = 1,\cdots,n-1 \\
& j = k+1,\cdots,(n+1) \\
& i = k+1,\cdots,n \\
& a_{ij}^{(0)} = a_{ij}
\end{aligned}
\tag{12}
$$

Luego, la solución es fácilmente calculada por sustitución hacia atrás, al terminar el procedimiento de eliminación. Esto es:

$$
x_i = \frac{1}{a_{ii}^{(i-1)}}\left[a_{1,n+1}^{(i-1)} - \sum_{j=i+1}^{n} a_{ij}^{(i-1)}x_j \right] ; \; i = n,\cdots,1 \tag{13}
$$

Método de Gauss-Jordan

Una variante muy importante es el *procedimiento de reducción o eliminación* de Gauss-Jordan. En este caso se procede a eliminar con los elementos diagonales en toda la columna, dejando sólo ese elemento; y se deriva en un sistema que, comparando con la Ecuación (10), en una primera pasada tiene la forma:

$$
\begin{aligned}
a_{11}x_1 + \;\; 0 \;\; + a_{13}^{(1)}x_3 + \cdots\cdots\cdots\cdots + a_{1n}^{(1)}x_n &= b_1 = a_{1,n+1}^{(1)} \\
a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 + \cdots\cdots\cdots\cdots + a_{2n}^{(1)}x_n &= b_2^{(1)} = a_{2,n+1}^{(1)} \\
a_{33}^{(2)}x_3 + \cdots\cdots\cdots\cdots + a_{3n}^{(2)}x_n &= b_3^{(2)} = a_{3,n+1}^{(2)} \\
\cdots\cdots\cdots\cdots\cdots & \\
a_{n3}^{(2)}x_n + \cdots\cdots\cdots\cdots + a_{nn}^{(2)}x_n &= b_n^{(2)} = a_{n,n+1}^{(2)}
\end{aligned}
\tag{14}
$$

Siguiendo con el procedimiento se hacen cero todos los elementos, excepto los correspondientes a la diagonal, resultando:

$$
\begin{aligned}
a_{11}x_1 + \ 0 \ + 0 \ + \cdots\cdots\cdots\cdots + 0 &= b_1^{(n-1)} = a_{1,n+1}^{(n-1)} \\
a_{22}^{(1)}x_3 + \ 0 \ + \cdots\cdots\cdots\cdots + 0 &= b_2^{(n-1)} = a_{2,n+1}^{(n-1)} \\
a_{33}^{(2)}x_3 + \cdots\cdots\cdots\cdots + 0 &= b_3^{(n-1)} = a_{3,n+1}^{(n-1)} \\
\cdots\cdots & \\
a_{nn}^{(n-1)}x_n &= b_n^{(n-1)} = a_{n,n+1}^{(n-1)}
\end{aligned}
\qquad (15)
$$

de modo que la solución es simplemente:

$$x_i = \frac{a_{i,n+1}^{(n-1)}}{a_{ii}^{(i-1)}} \ ; \ i = 1,2,...,n \qquad (16)$$

A pesar de lo que resulta del procedimiento, la eliminación Gaussiana es la más eficiente de las dos, considerando sólo las multiplicaciones y divisiones (recordar, además, la acumulación del error de redondeo), llegando, para grandes sistemas ($n>>1$), el método Gauss-Jordan a requerir cerca de un *50%* más de operaciones que el de Gauss.

Se ha trabajado mucho para lograr *formas compactas* del método de Gauss, no sólo para ahorrar espacio de almacenamiento (memoria) sino también para mejorar la precisión en los cálculos que más inciden en el resultado. Para ello se han definido matrices especiales en cuanto a la característica de su formulación, las cuales permiten ahorrar tiempo de cálculo una vez aplicadas. No trataremos este punto aquí, remitiendo al lector a la bibliografía recomendada al final del capítulo.

Análisis de Errores

Debido a que generalmente no es posible obtener la solución exacta, se considerarán los posibles errores y sus cotas. Las fuentes de error son variaciones en los elementos de $\underline{\underline{A}}$ y de $\underline{b}$, ya sean originales o debidas al redondeo. Estudiando primero el caso más simple, que considera cambios sólo en $\underline{b}$, será:

$$\underline{\underline{A}}.\underline{x} = \underline{b} + \underline{\Delta b} \qquad (17)$$

si resulta a su vez:

$$\underline{x} = \underline{x}^* + \underline{\Delta x}$$

se ve que:

$$\underline{\underline{A}}.\left(\underline{x}^* + \underline{\Delta x}\right) = \underline{b} + \underline{\Delta b}$$

con lo que:

$$\underline{\underline{A}}.\underline{\Delta x} = \underline{\Delta b}$$

Si $\underline{\underline{A}}$ es no singular, entonces resulta:

$$\underline{\Delta x} = \underline{\underline{A}}^{-1} . \underline{\Delta b} \tag{18}$$

y/o, tomando normas:

$$\left\| \underline{\Delta x} \right\| \leq \left\| \underline{\underline{A}}^{-1} \right\| \left\| \underline{\Delta b} \right\| \tag{19}$$

de igual modo:

$$\left\| \underline{b} \right\| \leq \left\| \underline{\underline{A}} \right\| \left\| \underline{x}^* \right\| \tag{20}$$

Multiplicando miembro a miembro las Desigualdades (19) y (20) se obtiene una cota para el error relativo:

$$\frac{\left\| \underline{\Delta x} \right\|}{\left\| \underline{x} \right\|} \leq \left\| \underline{\underline{A}} \right\| \left\| \underline{\underline{A}}^{-1} \right\| \frac{\left\| \underline{\Delta b} \right\|}{\left\| \underline{b} \right\|} \tag{21}$$

La cantidad $\left\| \underline{\underline{A}} \right\| \left\| \underline{\underline{A}}^{-1} \right\|$ se denomina *número de condición de* $\underline{\underline{A}}$ y se indica $K(\underline{\underline{A}})$; además se verifica que $K(\underline{\underline{A}}) \geq 1$, de modo que:

$$\frac{\left\| \underline{\Delta x} \right\|}{\left\| \underline{x} \right\|} \leq K(\underline{\underline{A}}) \frac{\left\| \underline{\Delta b} \right\|}{\left\| \underline{b} \right\|} \tag{22}$$

Así, si $K(\underline{\underline{A}})$ es cercano a 1, se dice que $\underline{\underline{A}}$ está *bien condicionada*; y si es muy grande nos encontramos frente a un caso *mal condicionado*.

Si la fuente de error fueran los elementos de la matriz de coeficientes, esto es:

$$\left(\underline{\underline{A}} + \underline{\underline{\Delta A}} \right) . \left(\underline{x}^* + \underline{\Delta x} \right) = \underline{b} \tag{23}$$

Por un procedimiento similar al anterior se llega a que:

$$\frac{\left\| \underline{\Delta x} \right\|}{\left\| \underline{x}^* + \underline{\Delta x} \right\|} \leq K\left(\underline{\underline{A}} \right) \frac{\left\| \underline{\underline{\Delta A}} \right\|}{\left\| \underline{\underline{A}} \right\|} \tag{24}$$

con igual connotación que la Ecuación (21). Los efectos del valor $\left\| \underline{\underline{\Delta A}} \right\|$ pueden subsanarse, en parte, trabajando con mayor precisión (mayor retención de cifras significativas durante el cálculo). En general se puede afirmar que $\frac{\left\| \underline{\underline{\Delta A}} \right\|}{\left\| \underline{\underline{A}} \right\|} \approx C(n)\varepsilon$ donde $C(n)$ es una constante del orden de la dimensión de la matriz y ε es el epsilon de la computadora.

Refinamiento Iterativo

Habíamos definido al vector residuo como $\underline{r} = \underline{b} - \underline{\underline{A}} . \underline{x}^{(c)}$ En situaciones en las que nos

interesa una solución muy cercana a la verdadera: $\underline{x}^* = \underline{\underline{A}}^{-1}.\underline{b}$ con un vector residuo $\underline{r}$ que sea pequeño o, en caso en que $\underline{x}^{(c)}$ sea tal que:

$$\underline{e} = \underline{x}^* - \underline{x}^{(c)} \tag{25}$$

sea muy pequeño en términos relativos a $\underline{x}^*$; si $\underline{b} = \underline{\underline{A}}.\underline{x}^*$, entonces resulta:

$$\underline{r} = \underline{\underline{A}}.\underline{x}^* - \underline{\underline{A}}.\underline{x}^{(c)} = \underline{\underline{A}}.\left(x^* - x^{(c)}\right) = \underline{\underline{A}}.\underline{e} \tag{26}$$

Tomando normas, y procediendo como ya se ha visto, se puede probar que:

$$\frac{1}{K\left(\underline{\underline{A}}\right)}\frac{\|\underline{r}\|}{\|\underline{b}\|} \leq \frac{\|\underline{e}\|}{\|\underline{x}^*\|} \leq K\left(\underline{\underline{A}}\right)\frac{\|\underline{r}\|}{\|\underline{b}\|} \tag{27}$$

Por lo que, si el número de condición $K(\underline{\underline{A}})$ es cercano a la unidad, pequeños errores relativos de $\underline{b}$ y $\underline{x}^*$ siguen la misma tendencia. Este hecho no ocurre para sistemas mal condicionados, como se había anticipado.

La respuesta al problema dependerá de la condición de $\underline{\underline{A}}$ y de la precisión de la aritmética (redondeo). Esto lleva a que el cómputo del residuo $\underline{r}$ debe hacerse con doble precisión, por la razón de que $\underline{r}$ suele ser del mismo orden de magnitud que el error de redondeo.

Formalizar un procedimiento que contemple el problema es lo que se denomina *refinamiento iterativo*. En realidad se arma un esquema iterativo con una solución calculada previamente, tal que el residuo para una etapa *m* de cálculo sea:

$$\underline{r}^{(m)} = \underline{b} - \underline{\underline{A}}.\underline{x}^{(m)} \quad ; \quad m=1,2,... \tag{28}$$

y, de acuerdo con la Ecuación (25),

$$\underline{x}^{(m+1)} = \underline{x}^{(m)} + \underline{e}^{(m)} \tag{29}$$

donde $\underline{e}^{(m)} = \underline{\underline{A}}^{-1}.\underline{r}^{(m)}$, según se desprende de la Ecuación (26). Si $\underline{x}^{(m+1)}$ no es satisfactoria se procede a la etapa $(m+1)$ de cálculo. En cambio, si $\left\|\underline{e}^{(m+1)}\right\| / \left\|\underline{e}^{(m)}\right\| < \varepsilon$, esto es, menor que una cierta tolerancia preestablecida aceptamos la solución $\underline{x}^{(m)}$. Así, se conforma un criterio de terminación que funciona muy bien en la práctica. Cada etapa de este proceso, además, es mucho más rápida que la solución del problema original. Se debe recordar que el cómputo de la Ecuación (28) se realiza en doble precisión, lo que implica un aumento en el requerimiento de memoria computacional.

Métodos Iterativos

Su basamento es idéntico al método de aproximaciones sucesivas visto en el caso de

sistemas de ecuaciones no lineales, con lo que, empezando con un vector inicial $\underline{x}^{(0)}$ se genera una sucesión de vectores, $\underline{x}^{(0)}$, $\underline{x}^{(1)}$, $\underline{x}^{(2)}$, ..., $\underline{x}^{(n)}$, tal que:

$$\underline{x}^{(n+1)} = \underline{F}^{(n)}\left(\underline{x}^{(n)},\ \underline{x}^{(n-1)},\ \underline{x}^{(n-2)},\ ....,\underline{x}^{(n-k)}\ \right) \tag{30}$$

Si la función iteradora $\underline{F}^{(n)}$ no depende del nivel de iteración, la recurrencia se la llama estacionaria.

Para la mayoría de las matrices estos métodos requieren más cálculo, para un deseado grado de convergencia, que los métodos directos; pero para matrices ralas (de gran interés en aplicaciones) el esfuerzo computacional es comparable. Además, como hemos mencionado, para estas matrices es posible lograr una mejor utilización de la memoria computacional. Luego para matrices ralas grandes, los métodos iterativos son los más aconsejados por su eficiencia computacional.

Por cuestiones de eficiencia computacional, sólo nos interesarán los *procesos iterativos lineales*. Un procedimiento de iteración *matricial lineal en un punto* tiene la forma general:

$$\underline{x}^{(n+1)} = \underline{B}^{(n)}.\underline{x}^{(n)} + \underline{c}^{(n)} \tag{31}$$

Si $\underline{B}$ y $\underline{c}$ son independientes del nivel *n*, la iteración matricial se denomina *estacionaria*.

El proceso iterativo continúa hasta que la diferencia (según una norma elegida) entre las aproximaciones obtenidas en dos iteraciones sucesivas sea menor que una cota δ prefijada. Esto es, se itera hasta que:

$$\left\|\underline{x}^{(n+1)} - \underline{x}^{(n)}\right\| \le \delta \tag{32}$$

La notación $\|\bullet\|$ indica la norma elegida. Tanto δ como $\|\bullet\|$ son arbitrarias, adecuándose el criterio de error a las características físicas de la simulación que se realiza.

Normas Usualmente Empleadas

a) Norma Euclideana: $\|\bullet\|_2 \le \sqrt{\sum_{i=1}^{m}\left(x_i^{(n+1)} - x_i^{(n)}\right)^2}$

b) Norma de Máximo (Norma ∞): $\|\bullet\|_\infty \le \underset{1\le i\le m}{máx}\left|x_i^{(n+1)} - x_i^{(n)}\right|$

En ambos casos el subíndice indica la i-ésima componente del vector $\underline{x}$.

Método de Aproximaciones Sucesivas

Recordar que en los métodos iterativos basados en aproximaciones sucesivas (o Wegstein) debe explicitarse la variable independiente. A modo de introducción se puede ver dicha estructura sumando miembro a miembro en la Ecuación (2) el vector de incógnitas $\underline{x}$, así:

$$\underline{x} + \underline{\underline{A}}.\underline{x} = \underline{b} + \underline{x}$$

En forma equivalente,

$$\left(\underline{\underline{A}} + \underline{\underline{I}}\right).\underline{x} = \underline{b} + \underline{x}$$

de donde:

$$\underline{x} = \left(\underline{\underline{A}} + \underline{\underline{I}}\right).\underline{x} - \underline{b}$$

De manera que la ley de recurrencia adopta la siguiente forma:

$$\underline{x}^{(n+1)} = \left(\underline{\underline{A}} + \underline{\underline{I}}\right).\underline{x}^{(n)} - \underline{b} \tag{32}$$

que nos da una forma genérica simple de los métodos iterativos estacionarios que son los universalmente usados. De esta forma concluimos, que en este caso, la función iteradora es $\underline{F}\left(\underline{x}\right) = \left(\underline{\underline{A}} + \underline{\underline{I}}\right).\underline{x} - \underline{b}$, esto es, de acuerdo con la estructura de los esquemas iterativos lineales, resulta: $\underline{\underline{B}} = \left(\underline{\underline{A}} + \underline{\underline{I}}\right)$ y $\underline{c} = -\underline{b}$.

Convergencia del Método de Aproximaciones Sucesivas

Función Continua de Lipschitz en un Recinto R

Una función vectorial $\underline{F(x)}$ definida sobre un recinto R es una función continua de Lipschitz si verifica que:

$$\left\|\underline{F}\left(\underline{a}\right) - \underline{F}(\underline{b})\right\| \leq K\left\|\underline{a} - \underline{b}\right\| \;\; ; \;\; con\; K \in (0,1)\; y\; \forall\; \underline{a}\;,\; \underline{b}\; \in\; R \tag{33}$$

Si $\underline{F(x)}$ es una función continua de Lipschitz en el recinto R, entonces:

a) Llamando $\underline{x}^*$ a la raíz de: $\underline{x} = \underline{F}\left(\underline{x}\right)$, luego:

$$\left\|\underline{x}^{(n+1)} - \underline{x}^*\right\| = \left\|\underline{F}\left(\underline{x}^{(n)}\right) - \underline{F}(\underline{x}^*)\right\| \tag{34}$$

Si $\underline{F(x)}$ es una función continua de Lipschitz, la Ecuación (34) se convierte en:

$$\left\|\underline{x}^{(n+1)} - \underline{x}^*\right\| \leq K\left\|\underline{x}^{(n)} - \underline{x}^*\right\| \tag{35}$$

Esto es, el error cometido en una iteración es proporcional al error cometido en la iteración anterior. Se dice que un método presenta *convergencia lineal* si satisface la Ecuación (35).

b) Si se aplican en forma secuencial las Ecuaciones (34) y (35) se obtiene:

$$\left\| \underline{x}^{(n+1)} - \underline{x}^* \right\| \leq K^{(n+1)} \left\| \underline{x}^{(0)} - \underline{x}^* \right\| \tag{36}$$

Si $0 < K < 1$, el error decrecerá en las sucesivas iteraciones y el decrecimiento será mayor cuanto menor sea K, esto es, habrá mayor velocidad de convergencia ya que se requiere un menor número de iteraciones para lograr un error dado.

Convergencia de Funciones Continuas de Lifschitz

Sea R un recinto cerrado que contiene a $\underline{x}^*$, raíz en el recinto del sistema de ecuaciones $\underline{x}=\underline{F}(\underline{x})$. Cualquiera sea la aproximación inicial $\underline{x}^{(0)} \in$ R, la sucesión de vectores generada por la ley de recurrencia:

$$\underline{x}^{(n)} = \underline{F}\left(\underline{x}^{(n-1)}\right) \; con \; n = 1,2,... \tag{37}$$

converge hacia $\underline{x}^*$ siempre que $\underline{F}(\underline{x})$ satisfaga la condición de Lipschitz (Ec. (34)).

Sin embargo, la condición de Lipschitz es difícil de determinar para un problema dado. Una forma alternativa que genera una condición suficiente para la convergencia del método, es trabajar con la matriz Jacobiana, $\underline{\underline{J}}$, cuyas componentes se definen como:

$$J_{ij}\left(\underline{x}\right) = \frac{\partial F_i\left(\underline{x}\right)}{\partial x_j} \;\; \forall i, j \tag{38}$$

Si las componentes de dicha matriz existen y

$$\left\| \underline{\underline{J}}\left(\underline{x}\right) \right\| \leq M < 1 \;\; \forall \underline{x} \in R \tag{39}$$

queda establecida a través de la Ecuación (39) una condición suficiente para la convergencia del Método de Aproximaciones Sucesivas.

En un dominio unidimensional, la condición definida por la Ecuación (39) se reduce a:

$$\left| F'\left(x\right) \right| < 1 \tag{40}$$

Condición Necesaria de Convergencia

Una *condición necesaria* para que el método de aproximaciones sucesivas converja $\forall \; \underline{x}^{(0)} \in R$ es que el radio espectral de $\underline{\underline{J}}(\underline{x}) \leq 1$. Se define el radio espectral de una matriz $\underline{\underline{A}}$ a:

$$r\left(\underline{\underline{A}}\right) = \max_k \left| \lambda_k \left(\underline{\underline{A}}\right) \right| \tag{41}$$

Esto es, el radio espectral de una matriz (cuadrada) es el autovalor de la matriz de máximo valor absoluto.

Método de Jacobi

Escribimos la matriz $\underline{\underline{A}}$ de coeficientes de la siguiente manera:

$$\underline{\underline{A}} = \underline{\underline{D}} + \underline{\underline{L}} + \underline{\underline{U}} \tag{42}$$

siendo $\underline{\underline{D}}$ la matriz diagonal formada con los elementos diagonales de $\underline{\underline{A}}$. $\underline{\underline{L}}$ y $\underline{\underline{U}}$ son matrices triangulares del tipo inferior y superior, respectivamente, formadas con el resto de los elementos de la matriz $\underline{\underline{A}}$ y ceros en su diagonal principal. Luego, reemplazamos la descomposición de la matriz $\underline{\underline{A}}$ (Ec. (42)) en el SEAL (Ecuación (2)), así:

$$\left(\underline{\underline{D}} + \underline{\underline{L}} + \underline{\underline{U}}\right).\underline{x} = \underline{b}$$

$$\underline{\underline{D}}.\underline{x} = -\left(\underline{\underline{L}} + \underline{\underline{U}}\right).\underline{x} + \underline{b}$$

$$\underline{x} = -\underline{\underline{D}}^{-1}.\left(\underline{\underline{L}} + \underline{\underline{U}}\right).\underline{x} + \underline{\underline{D}}^{-1}.\underline{b}$$

Obtenemos la ley de recurrencia:

$$\underline{x}^{(n+1)} = -\underline{\underline{D}}^{-1}.\left(\underline{\underline{L}} + \underline{\underline{U}}\right).\underline{x}^{(n)} + \underline{\underline{D}}^{-1}.\underline{b} = \underline{\underline{D}}^{-1}.\left(\underline{\underline{D}} - \underline{\underline{A}}\right).\underline{x}^{(n)} + \underline{\underline{D}}^{-1}.\underline{b} \tag{43}$$

que supone que la diagonal principal no tiene ceros.

De no ser así, si $\underline{\underline{A}}$ es no singular, se debe permutar filas y columnas para obtener una forma que permita definir $\underline{\underline{D}}$. Es deseable que sea de la forma diagonal dominante, esto es, que los elementos de la diagonal sean lo más grande posible respecto a los demás. Luego, el método de Jacobi puede entonces escribirse como:

$$\underline{x}^{(n+1)} = \underline{\underline{B}}.\underline{x}^{(n)} + \underline{c} \tag{44}$$

donde:

$$\underline{\underline{B}} = -\underline{\underline{D}}^{-1}.\left(\underline{\underline{L}} + \underline{\underline{U}}\right) \tag{45}$$

$$\underline{c} = \underline{\underline{D}}^{-1}.\underline{b} \tag{46}$$

Por componentes:

$$x_i^{(n+1)} = \frac{b_i - \sum_{\substack{k=1 \\ k \neq i}}^{n} a_{ik} x_k^{(n)}}{a_{ii}} \tag{47}$$

Convergencia del Método de Jacobi

Siendo del Método de Jacobi un caso particular del Método de Aproximaciones Sucesivas, la condición necesaria y suficiente para su convergencia es que la función iteradora:

$$\underline{F}\left(\underline{x}\right) = -\underline{\underline{D}}^{-1}.\left(\underline{\underline{L}}+\underline{\underline{U}}\right).\underline{x}+\underline{\underline{D}}^{-1}.\underline{b} = \underline{\underline{B}}.\underline{x}+\underline{c} \tag{48}$$

sea una función continua de Lipschitz, lo que puede ser explicitado exigiendo que:

$$\left\|\underline{\underline{B}}\right\|_{\infty} \leq \underset{1\leq i\leq m}{máx}\sum_{j=1}^{n}\left|b_{ij}\right| \leq K < 1 \tag{49}$$

Lo cual significa sumar todos los elementos de una fila y elegir el máximo entre ellos.

Si se analiza la expresión de $\underline{\underline{B}}$ (Ecuación (48)), se observa que la expresión (49) es equivalente a:

$$\left\|\underline{\underline{B}}\right\|_{\infty} \leq \underset{i}{máx}\sum_{\substack{j=1\\ j\neq i}}^{n}\frac{\left|a_{ij}\right|}{\left|a_{ii}\right|} < 1 \tag{50}$$

Una expresión similar a la (50) puede generarse para el Método Gauss-Seidel.

Se concluye que el método es convergente solamente cuando se aplica a matrices cuya matriz de coeficientes sea o pueda ser transformada mediante intercambio de filas y/o columnas en diagonal dominante, lo que no es habitual en simulación de procesos químicos.

Método de Gauss-Seidel

Este es el método ampliamente usado que siempre converge si converge Jacobi, (y aún en casos en que éste no converge); y en general lo hace más rápido, siendo su implementación más eficiente.

La diferencia fundamental es que a medida que se calcula cada componente de $\underline{x}^{(n+1)}$ se utiliza inmediatamente en la misma iteración. Al proceder de ese modo, la ecuación obtenida para Jacobi, resulta:

$$\underline{x}^{(n+1)} = -\left(\underline{\underline{D}}^{-1}.\underline{\underline{L}}\right).\underline{x}^{(n+1)} - \left(\underline{\underline{D}}^{-1}.\underline{\underline{U}}\right).\underline{x}^{(n)} + \underline{\underline{D}}^{-1}.\underline{b} \tag{51}$$

reordenando y sacando factor común $\underline{x}^{(n+1)}$ se tiene:

$$\left(\underline{\underline{I}}+\underline{\underline{D}}^{-1}.\underline{\underline{L}}\right)\underline{x}^{(n+1)} = -\left(\underline{\underline{D}}^{-1}.\underline{\underline{U}}\right).\underline{x}^{(n)} + \underline{\underline{D}}^{-1}.\underline{b}$$

por lo que, premultiplicando por $\underline{\underline{D}}$, obtenemos:

$$\left(\underline{\underline{D}}+\underline{\underline{L}}\right).\underline{x}^{(n+1)} = -\underline{\underline{U}}.\underline{x}^{(n)} + \underline{b}$$

Luego resulta:

$$\underline{x}^{(n+1)} = -\left(\underline{\underline{D}}+\underline{\underline{L}}\right)^{-1}.\underline{\underline{U}}.\underline{x}^{(n)} + \left(\underline{\underline{D}}+\underline{\underline{L}}\right)^{-1}.\underline{b} \tag{52}$$

entonces la expresión generalizada será:

$$\underline{x}^{(n+1)} = \underline{\underline{B}} . \underline{x}^{(n)} + \underline{c} \tag{53}$$

donde:

$$\underline{\underline{B}} = -\left(\underline{\underline{D}} + \underline{\underline{L}}\right)^{-1} . \underline{\underline{U}} \tag{54}$$

y

$$\underline{c} = \left(\underline{\underline{D}} + \underline{\underline{L}}\right)^{-1} . \underline{b} \tag{55}$$

Por componentes, resulta:

$$x_i^{(n+1)} = \frac{b_i - \sum_{k=1}^{i-1} a_{ik} x_k^{(n+1)} - \sum_{k=i+1}^{n} a_{ik} x_k^{(n)}}{a_{ii}} \tag{56}$$

Existe un teorema que prueba que: *Si una matriz es definida positiva, el procedimiento de Gauss-Seidel converge independientemente del vector inicial propuesto.*

Otro hecho importante es que los métodos matriciales iterativos convergen linealmente. Es natural pensar que el error de redondeo es mayor en un método iterativo que en uno directo; sin embargo como siempre se usa la matriz de coeficientes original, el error por redondeo en que se incurre en un método iterativo es sólo aquel producido en la última iteración. En consecuencia, ambos tipos de métodos tienen un error equivalente, tan serio para unos como para otros.

El error de truncamiento, en el caso de los métodos iterativos, suele ser de un orden mayor de magnitud que el mencionado de redondeo. No obstante, dado que en los métodos iterativos se tiene un criterio de control de error, este es acotado convenientemente.

Método de Sobrerrelajación

Una condición necesaria y suficiente para que el método de Jacobi (y eventualmente el de Gauss-Seidel), planteado a través del algoritmo $\underline{x}^{(n+1)} = \underline{\underline{B}} . \underline{x}^{(n)} + \underline{c}$ converja, es que el máximo autovalor de $\underline{\underline{B}}$ sea en valor absoluto menor que 1 (radio espectral menor que 1).

Cuanto menor sea dicho radio espectral, más rápida será la convergencia ya que se puede demostrar que:

$$\frac{\left\| \underline{x}^{(n)} - \underline{x}^{*} \right\|}{\left\| \underline{x}^{(0)} - \underline{x}^{*} \right\|} \leq \left| \lambda \right|_{máx}^{n} \tag{57}$$

El método de sobrerrelación pretende modificar la matriz B y consecuentemente su

autovalor de manera de acelerar la convergencia. Para ello modifica la Ec. (44) conforme a:

$$\underline{x}^{(n+1)} = \underline{\underline{B}}.\,\underline{x}^{(n)} + \underline{c} \tag{58}$$

Siendo los valores estimados para una nueva iteración:

$$\underline{x}^{(n+1)} = \alpha\,\underline{x}^{(n+1)} + (1-\alpha)\,\underline{x}^{(n)} \tag{59}$$

Si se sustituye la Ec. (58) en la Ec. (59) queda:

$$\underline{x}^{(n+1)} = \left[\alpha\,\underline{\underline{B}} + (1-\alpha)\,\underline{\underline{I}}\right]\underline{x}^{(n)} + \alpha\,\underline{c} \tag{60}$$

Siendo ahora el máximo autovalor en valor absoluto de la matriz:

$$\underline{\underline{B}}' = \alpha\,\underline{\underline{B}} + (1-\alpha)\,\underline{\underline{I}} \tag{61}$$

el que determinará la velocidad de convergencia. Pero, la ecuación característica de $\underline{\underline{B}}'$ vendrá expresada por:

$$det\left(\underline{\underline{B}}' - \lambda'\,\underline{\underline{I}}\right) = det\left\{\alpha\left[\underline{\underline{B}} + \frac{(1-\alpha-\lambda')}{\alpha}\,\underline{\underline{I}}\right]\right\} = \alpha^n\,det\left[\underline{\underline{B}} + \frac{(1-\alpha-\lambda')}{\alpha}\,\underline{\underline{I}}\right] = 0 \tag{62}$$

Si definimos:

$$-\lambda = \frac{1-\alpha-\lambda'}{\alpha} \tag{63}$$

la Ecuación (62) también dará los autovalores de $\underline{\underline{B}}$, entonces:

$$\lambda'_i = 1 + \alpha(\lambda_i - 1) \tag{64}$$

Adoptando $\alpha>1$ para $\lambda_i>0$ es posible reducir $|\lambda'|_{máx}$ y consecuentemente acelerar la convergencia del método.

El método de relajación puede también aplicarse al método de Gauss-Seidel y una discusión del mismo tenor que la precedente puede llevarse a cabo en dicha circunstancia.