Análisis numérico, matricial e interpolación

Nelson

Resumen

Este es un resumen de la parte primera del libro *Introducción al cálculo numérico*, temas del 1 al 6, temario correspondiente a la asignatura Análisis numérico, matricial e interpolación de segundo curso del grado en Matemáticas de la UNED.

El resumen incluye todos los teoremas, definiciones, proposiciones, lemas y textos relevantes del libro respetando su nomenclatura y numeración para facilitar su uso en relación a la fuente original.

El resumen no incluye ejemplos (salvo textos relevantes), demostraciones ni ejercicios.

Puede contener erratas (consulte la fuente original).

Índice

1. Estabilidad y errores en el cálculo numérico

1.1. Introducción

Un ordenador produce resultados en respuesta a cálculos programados que posiblemente difieren ligeramente de los valores exactos esperados. Ello es consecuencia de que trabajan con una aritmética discreta que no coincide plenamente con la aritmética exacta de los números enteros o reales.

El propósito de esta capítulo es etudiar las representaciones más comunes de números en un computador y las aritméticas que usan en sus cálculos. Una vez establecido que los errores respecto a la aritmética real pueden estar presentes, se introduce el concepto de estabilidad numérica que permite seleccionar los algoritmos que son válidos para esta clase de cálculos.

1.2. Representación de números en un computador

La diferencia (en valor absoluto) entre el valor exacto x y el valor obtenido en un cálculo $\overline{x}$, determina el error absoluto cometido $E(x) = |x - \overline{x}|$. Una indicación más precisa del error cometido en un cálculo la da el error relativo, que se define por la siguiente expresión

$$E_R(x) = \frac{|x - \overline{x}|}{|x|}.$$

Los computadores representan los números en sus memorias asignándoles una cantidad fija de posiciones a las que puede acceder. Esta cantidad puede variar de un computador a otro, de acuerdo con el estándar de representación que se haya atribuido a ese tipo de número.

Con el fin de comprender cómo usan los ordenadores la aritmética discreta, se recuerda el concepto de representación posicional de los números enteros y reales. Se considera el conjunto de números $\mathcal{M}$ que pueden ser representados en la forma

$$\pm 0.d_1 d_2 ... d_n \times b^e$$

donde $0 \leqslant d_i < b$ para $i = 1, 2, ..., n$. El número entero positivo b es fijo y corresponde a la base de la representación. El entero e corresponde al exponente y puede variar en un determinado rango. Finalmente, el entero positivo fijo n controla la precisión de la representación. La parte fraccionaria es frecuentemente llamada mantisa. El valor numérico real asignado a esta representación es

$$x = \pm(d_1 b^{-1} + d_2 b^{-2} + ... + d_n b^{-n}) \times b^e$$

Idealmente, se puede ampliar el sistema admitiendo que n sea infinito y de este modo, se podrían incluir todos los números reales en esta representación. Obviamente, para un sistema de representación posicional que se pretenda poner en práctica en un computador, se requiere que n sea un número entero fijo y que se ajuste a la capacidad física de su memoria accesible. Por esta razón, todo sistema de representación que utilice un computador, tendrá un conjunto finito de números-máquina.

Habitualmente, para números enteros se emplea una representación en la que el exponente e es fijo e igual a n. De este modo, a la representación $\pm d_1 d_2 ... d_n$ le corresponde el valor entero

$$x = \pm(d_1 b_{n-1} + d_2 b_{n-2} + ... + d_n)$$

Un aspecto relevante de este sistema de representación es que puede producirse un desbordamiento en las operaciones aritméticas debido a que el conjunto es acotado.

Si el exponente e es un entero variable en un determinado rango, el conjunto de números que pueden ser representados de este modo (llamado punto flotante) es más amplio.

Los aspectos más relevantes de este sistema de representación son los siguientes:

- Los números representados no son equidistantes y se observa una mayor densidad en las proximidades de 0.
- Un número puede tener representaciones distintas.
- Se puede producir un desbordamiento en las operaciones aritméticas debido a que el conjunto es acotado.
- El resultado de algunas operaciones aritméticas reales con números de este conjunto está fuera del conjunto aunque no se haya produciodo un desbordamiento del rango.

Una representación en punto flotante está normalizada si el primer dígito en la parte fraccionaria es necesariamente distinto de 0. De este modo, se evita que un número pueda tener representaciones distintas en un mismo sistema. Actualmente, la representación que usan la mayoría de computadores es la llamada IEEE Standard 754 en punto flotante. Está basada en los tres elementos mencionados anteriormente, el signo, la mantisa y el exponente.

Si la representación es binaria, el primer dígito es 1 (dígito principal) y en la mayoría de las puestas en práctica de este estándar, no es almacenado en la memória (dígito implicito). El estándar de representación en punto flotante IEEE que corresponde a lo que se conoce como precisión simple, utiliza 4 bytes de memoria (32 bits) de los cuales 8 bits son para el exponente E, 23 para la parte fraccionaria F y uno para el signo S. Cada bit almacena un 0 o un 1. El valor asignado a una representación es

$$(-1)^S \times 2^{E-127} \times 1.F$$

El campo del exponente necesita representar a la vez exponentes positivos y negativos. Para conseguirlo se añade un sesgo al valor positivo almacenado para lograr el exponente deseado. En el estándar IEEE 745 para precisión simple, este valor es 127. De este modo, un valor almacenado E representa un valor real $E - 127$. Para doble precisión el campo del exponente tiene 11 bits y un sesgo de 1023. El bit del signo 0 es para positivos y el 1 para negativos.

1.3. Aritmética en un sistema de representación finito

Conceptualmente, los números reales pueden aproximarse por números de un sistema discreto $\mathcal{M}$ de dos modos ligeramente distintos que se conocen como truncamiento y redondeo.

La puesta en práctica de métods de truncamiento está basada en la siguiente idea: Un número real puede ser aproximado por un número de punto flotante con parte fraccionaria infinita en la base b

$$\pm .d_1 d_2 ... \times b^e.$$

Consecuentemente, si se trunca esta serie con n dígitos se obtiene una aproximación en el sistema discreto disponible en el computador. Sin embargo, es más usado el llamado método de redondeo que consiste en aproximar un número real positivo por el número-máquina más próximo. Este procedimiento sería el que nos proporcionaría mayor precisión. Para precisar estas ideas se representa la función parte entera de un número real por un corchete [ ] y se define para un número real positivo representado por $x = 0.m \times b^e$ donde m es una cifra con posiblemente una infinidad de dígitos, las siguientes aproximaciones por números-máquina de n dígitos

- Truncamiento: $\mathcal{T}(x) = [b^n \times 0.m] \times b^{e-n}$
- Redondeo: $\mathcal{R}(x) = [b^n \times 0.m + 0{,}5] \times b^{e-n}$

En lo que sigue se centrará la atención en el método de redondeo. Si $x < 0$ entonces se define

$$\mathcal{R}(x) = -\mathcal{R}(-x).$$

De la definición de redondeo se deduce que

$$|x - \mathcal{R}(x)| \leqslant \frac{1}{2} b^{e-n}.$$

El número $\frac{1}{2} b^{e-n}$ se conoce como unidad de redondeo o precisión de la máquina.

Una aritmética para un sistema de representación finita estaría disponible si se consigue asignar como resultado de una operación el que corresponde al redondeo de la operación aritmética exacta.

1.4. Estabilidad en los algoritmos numéricos

Una vez que un entorno de cálculo dispone de una aritmética finita que permite realizar las operaciones elementales dentro del rango numérico que puede representar, se hace necesario complementarlo con otras funciones elementales que permitan realizar cálculos más complejos. La dificultad está en que estas funciones implican un número elevado de operaciones elementales y puesto que los errores de redondeo respecto a la artimética exacata son inevitables, el resultado final puede estar muy deteriorado en relación con el exacto.

La razón por la que un procedimiento produce resultados más precisos que otro se atribuye a su inestabilidad. Es decir, un error en alguna etapa del procedimiento se

propaga de un modo creciente en las siguientes. La necesidad de utilizar algoritmos estables para evaluar una función es obligada por la inevitable presencia de errores de redondeo debidos al uso de arimtéticas finitas en los entornos de cálculo automático.

2. Sistemas de ecuaciones numéricas lineales

2.1. Introducción

El objetivo de este capítulo es el estudio de métodos de resolución de sistemas lineales que sean eficientes en problemas reales con media o alta dimensión.

Aunque la elección del método más adecuado para resolver un sistema lineal depende fundamentalmente de las propiedades de la matriz de coeficientes, para un sistema lineal sin características especiales, uno de los métodos más eficientes es el clásico método de eliminación de Gauss o alguna de sus variantes.

Estos métodos pertenecen a la clase de los métodos directos que conducen a la solución exacta en un número finito de operaciones aunque ello sólo desde un punto de vista teórico, debido a la presencia de errores de redondeo cuando los cálculos se realizan en un ordenador. En otra categoría de métods, los iterativos permiten generar una sucesión de soluciones aproximadas que convergen a la solución exacta.

2.2. Norma matricial subordinada a una norma vectorial

La norma euclídea de un vector (real o complejo) $\overrightarrow{x}$ de n componentes, cuya componente i-ésima es x_i, está definida por la siguiente expresión

$$\|\overrightarrow{x}\| = \sqrt{\overrightarrow{x}\cdot\overrightarrow{x}} = \left(\sum_{i=1}^{n}|x_i|^2\right)^{1/2}.$$

Más general es la norma p que está definida por

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

para cualquier número positivo p. En el límite, la norma ∞ está definida por

$$\|\overrightarrow{x}\|_\infty = \max_{i=1,\ldots,n}|x_i|.$$

Una matriz puede representar una aplicación lineal entre dos espacios euclídeos. Por ello, existen normas matriciales que están definidas en base a esa representación y cuya definición está relacionada con las normas euclídeas en ambos espacios. Así, una norma matricial subordinada a una norma vectorial está definida por

$$\|A\| = \max_{\|\overrightarrow{x}=1\|}\|A\overrightarrow{x}\|$$

para cualquier matriz A. Obviamente, si la matriz A tiene n filas y m columnas, dos normas vectoriales entrarían en juego. Aunque muchos de los conceptos y resultados de este capítulo son extensibles a esta situación, en este texto se supondrá siempre, salvo mención en sentido contrario, que la matriz A es cuadrada y la elección de norma para el dominio e imagen es la misma.

En los casos más simples se pueden obtener expresiones que permiten determinar el valor de una norma subordinada de una matriz en términos de sus coeficietnes.

- Norma subordinada 1

  $$\|A\|_1 = \max_{1\leqslant j\leqslant n}\sum_{i=1}^{n}|a_{ij}|,$$

  si se considera la norma vectorial

  $$\|\overrightarrow{x}\|_1 = \sum_{j=1}^{n}|x_j|$$

  y a_{ij} representa el coeficiente de la matriz A correspondiente a la fila i y la columna j.

- Norma subordinada ∞

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

  , si se considera la norma vectorial

  $$\|\overrightarrow{x}\|_\infty = \max_{1\leqslant j\leqslant n}|x_j|.$$

- Norma subordinada 2

  $$\|A\|_2 = \rho(A^tA)^{1/2},$$

  si se considera la norma vectorial

  $$\|\overrightarrow{x}\|_2 = \sqrt{\sum_{1}^{n}|x_j|^2},$$

  y ρ representa el radio espectral de la matriz (es decir, el máximo de los módulos de los autovalores). En particular, si A es simétrica entonces se cumple que $\|A\|_2 = \rho(A)$.

La norma subordinada euclídea no coincide con la norma matricial, asociada al producto escalar matricial

$$A : B = \mathrm{tr}A^tB$$

donde tr representa el operador que asocia a una matriz su traza, y que se conoce como norma de Frobenius

$$\|A\|_F = (A:A)^{1/2} = \left(\sum_{i,j=1}^{n}a_{ij}^2\right)^{1/2}.$$

Mientras que la norma subordinada euclídea está relacionada con el radio espectral del producto A^tA, la norma de Frobenius lo está con su traza. Por otra parte, para la matriz identidad I todas las normas matriciales toman el valor 1 pero la norma de Frobenius toma el

valor $\sqrt{n}$.

Una consecuencia inmediata de la definición de norma subordinada es la siguiente desigualdad

$$\|AB\| \leqslant \|A\| \, \|B\|$$

válida para dos matrices cualesquiera. Como consecuencia de esta desigualdad se tiene que para toda matriz A

$$\|A^q\| = \|A\|^q,$$

lo que implica en particular, que la sucesión de potencias sucesivas de una matriz converge a cero si para alguna de sus normas subordinadas se cumple que $\|A\| < 1$.

Si $\overrightarrow{x}$ es un vector propio asociado a un autovalor λ de la matriz A, $\overrightarrow{x}$ verifica que $A\overrightarrow{x} = \lambda\overrightarrow{x}$. De la definción de norma subordinada se deduce que

$$\|A\overrightarrow{x}\| \leqslant \|A\| \, \|\overrightarrow{x}\|$$

En consecuencia

$$|\lambda| \|\overrightarrow{x}\| \leqslant \|A\| \, \|\overrightarrow{x}\|$$

y

$$|\lambda| \leqslant \|A\|.$$

Puesto que esta desigualdad se verifica para todo autovalor, se concluye que

$$\rho(A) \leqslant \|A\|.$$

Teorema 1 Para cualquier matriz cuadrada A se cumple

$$\rho(A) = \text{ínf}\{\|A\| : \| \ \|\}$$

es una norma subordinada.

Es importante tener en cuenta que si A es simétrica, el ínfimo se alcanza en la norma subordinada a la norma euclídea ya que

$$\rho(A) \leqslant \|A\|_2 = \rho(A^2)^{1/2} = \rho(A).$$

Corolario 1 Para una matriz cuadrada arbitraria, las tres afirmaciones siguientes, referidas a las potencias de A y el radio espectral, son equivalentes

1. $\lim_{n \leftarrow \infty} A^n = 0$
2. $\lim_{n \leftarrow \infty} \|A^n\| = 0$, para alguna norma subordinada
3. $\rho(A) < 1$

2.3. Estabilidad de un sistema de ecuaciones lineales

Por el momento, se analizará la cuestió de si una pequeña modificación de los datos del sistema $A\overrightarrow{x} = \overrightarrow{b}$ puede generar un importante cambio en la solución. Para ello se considera el sistema perturbado

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

donde $\delta A, \delta\overrightarrow{b}$ son elementos arbitrarios de la misma naturaleza que A y $\overrightarrow{b}$.

El error que introduce esta perturbación está dado por

$$\delta\overrightarrow{x} = (A + \delta A)^{-1}(\delta\overrightarrow{b} - \delta A\overrightarrow{x}).$$

Si se toman normas en ambos miembros y se usa para las matrices una norma subordinada, se obtiene que

$$\|\delta\overrightarrow{x}\| \leqslant \|(A + \delta A)^{-1}\|(\|\delta\overrightarrow{b}\| + \|\delta A\| \, \|\overrightarrow{x}\|).$$

Por otra parte, puesto que $\overrightarrow{x}$ es solución de la ecuación, se cumple que

$$\|\overrightarrow{b}\| \leqslant \|A\| \, \|\overrightarrow{x}\|.$$

De las dos últimas desigualdades se deduce que

$$\frac{\|\delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|} \leqslant \|(A + \delta A)^{-1}\| \, \|A\| \left(\frac{\|\delta\overrightarrow{b}\|}{\|\overrightarrow{b}\|} + \frac{\|\delta A\|}{\|A\|} \right)$$

Para completar la estimación del error relativo se usará el siguiente lema de Banach

Lema 1 Si D es una matriz cuadrada tal que $\|D\| < 1$ para una determinada norma subordinada a una norma vectorial entonces $I + D$ es no-singular y se cumple que

$$\|(I + D)^{-1}\| \leqslant \frac{1}{1 - \|D\|}.$$

Volviendo a la estimación del error, si se usa el l ema de Banach se obtiene la siguiente acotación

$$\|(A+\delta A)^{-1}\| = \|(I+A^{-1}\delta A)^{-1}A^{-1}\| \leqslant \|A^{-1}\| \frac{1}{1 - \|A\| \, \|\delta A\|},$$

si se cumple que $\|A^{-1}\| \, \|\delta A\| < 1$. Si se usa esta acotación en la estimación del error relativo de la solución del sistema perturbado, se obtiene que

$$\frac{\|\delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|} \leqslant \frac{\|A\| \, \|A^{-1}\|}{1 - \|A^{-1}\| \, \|\delta A\|} \left(\frac{\delta\overrightarrow{b}}{\|\overrightarrow{b}\|} + \frac{\|\delta A\|}{\|A\|} \right).$$

Esta estimación sugiere que una medida de esta sensibilidad es el número de condición de A, definido por

$$\text{cond}(A) = \|A\| \, \|A^{-1}\|$$

para toda matriz cuadrada no-singular A. De este modo, si $\text{cond}(A)\frac{\|\delta A\|}{\|A\|} < 1$ se cumple que

$$\frac{\|\delta\overrightarrow{x}\|}{\|\overrightarrow{x}\|} \leqslant \frac{\text{cond}(A)}{1 - \text{cond}(A) = \|A\| \, \|A^{-1}\|} \left(\frac{\delta\overrightarrow{b}}{\|\overrightarrow{b}\|} + \frac{\|\delta A\|}{\|A\|} \right).$$

Además, de la relación

$$1 = \|I\| = \|AA^{-1}\| \leqslant \|A\| \, \|A^{-1}\| = \text{cond}(A)$$

se deduce que el número de condición de una matriz es siempre mayor o igual que 1. Si el número de condición es relativamente pequeño los errores que se generarían en la solución serían pequeños y viceversa.

Una matriz está perfectamente condicionada si su número de condición es 1. En particular, las matrices ortogonales, que verifican que $A^tA = I$, están perfectamente condicionadas ya que

$$\|A\| = \rho(A^tA)^{1/2} = \rho(I)^{1/2} = 1$$

y

$$\|A^{-1}\| = \rho(AA^t)^{1/2} = \rho(I)^{1/2} = 1$$

Por otra parte, si A es una matriz simétrica definida positiva entonces se cumple que

$$\|A\| = \delta_{\text{máx}}, \quad \|A^{-1}\| = \delta_{\text{mín}}$$

donde $\delta_{\text{mín}}$ y $\delta_{\text{máx}}$ representan respectivamente el mínimo y el máximo autovalor de A. De este modo, se tiene que

$$\text{cond}(A) = \frac{\delta_{max}}{\delta_{\text{mín}}}.$$

2.4. Sistemas lineales de gran dimensión

El cálculo científico tiene en cuenta en la elección de algoritmo de resolución de un sistema lineal, su coste computacional, su estabilidad y cómo se maneja la información de los coeficientes del sistema.

Se puede expresar un sistema lineal compatible y determinado de ecuaciones numéricas en forma matricial como

$$A\overrightarrow{x} = \overrightarrow{b}$$

donde A es una matriz de dimensión $n \times n$ de coeficientes reales o complejos y $\overrightarrow{b}$ un vector columna de n componentes reales o complejas.

Desde un punto de vista teórico, un sistema lineal de ecuaciones, cuya matriz de coeficientes tenga determinante diferente de cero, puede ser resuelto por la fórmula de Cramer.

Desde un punto de vista teórico, las fórmulas de Cramer tienen dos estimables cualidades: son explícitas y tienen validez general para cualquier sistema lineal. Sin embargo, desde el punto de vista científico, la aplicación de estas fórmulas directamente tiene una dificultad que las hace poco prácticas para el cálculo; requieren el cálculo de determinantes y si la evaluación directa de un determinante se realiza mediante la regla de Sarrus se requieren $n!$ sumas, cada una de ellas con $n-1$ productos. La función entera $n!$ alcanza pronto valores muy elevados. Es preciso buscar algoritmos de resolución que reduzcan el coste computacional, es decir, tales que el número de operaciones elementales que deban realizarse, sea reducido.

2.5. Matrices dispersas

De la experiencia con los sistemas lineales de gran tamaño, se conoce que en su mayoría, el número de coeficientes nulos es muy elevado en proporción a los que no son nulos. Esto nos obliga a considerar formas que no son estándares para representar uan matriz en un computador para evitar tener que almacenar y manejar tanta información inútil.

Si la matriz posee una proporción pequeña de coeficientes que no son nulos, lo realmente eficiente es buscar una estructura que almacene únicamente la información útil. Se puede precisar esta idea con el concepto de densidad de una matriz. Concretamente, se define la densidad de una matriz como el cociente entre el número de coeficientes no nulos y el total de coeficientes.

La primera idea para simplificar el almacenamiento de la matriz es tener en cuenta la estructura simétrica tridiagonal. En otras palabras, la información útil de la matriz está situada en la diagonal principal y las dos diagonales adyacentes. Bastaría con almacenar la diagonal principal y una adyacente para poder reconstruir la matriz entera. De este modo, el sistema lineal está descrito por 3 vectores (A_1, A_2, b) que representan las dos diagonales y el término independiente. Existen recursos eficaces para manejar matrices dispersas que almacenan internamente

- Los coeficientes que no son nulos.
- Los índices de filas de los coeficientes que no son nulos.
- Los índices de columnas de los coeficientes que no son nulos.

La eficacioa de un método de resolución está relacionada en general con el hecho de que el método respete o no, estas formas de almacenamiento reducido.

2.6. Método de eliminación de Gauss

Desde el punto de vista tradicional del Álgebra, el método de eliminación de Gauss no es tan grato como la regla de Cramer, ya que no está definido por fórmulas explícitas y además no tiene validez general puesto que falla cuando encuentra un coeficiente nulo en la posición seleccionada de la diagonal principal. Sin embargo, es el que tiene el coste computacional más bajo y se adapta perfectamente a muchas formas de almacenamiento reducido.

Sea un sistema de ecuaciones

$$A\overrightarrow{x} = \overrightarrow{b}$$

Para realizar la eliminación de los coeficientes de la primera columna se considera la siguente matriz

$$E(1) = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ \frac{-a_{21}}{a_{11}} & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{-a_{n-1,1}}{a_{11}} & 0 & \cdots & 1 & 0 \\ \frac{-a_{n,1}}{a_{11}} & 0 & \cdots & 0 & 1 \end{pmatrix}$$

Se introduce la siguiente notación $A(1) = A, A(2) = E(1)A(1)$. La matriz $A(2)$ tiene ceros en las posiciones que están debajo de la diagonal en la primera columna. Además $\det(A) = \det(A(2))$.

Para la eliminación de coeficientes correspondientes a posiciones que están debajo de la diagonal principal, se procede de modo recurrente. Si $A(j)$ representa la matriz

en la que se han eliminado estos coeficientes en las $j-1$ primeras columnas y $E(j-1)$ la matriz de multiplicadores de la $(j-1)$-ésima columna entonces se construye la matriz de multiplicadores para la columna j-ésima como

$$E(j) = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 1 & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{-a^j_{j+1,j}}{a^j_{jj}} & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{-a^j_{n-1,j}}{a^j_{jj}} & \cdots & \cdots & 1 & 0 \\ 0 & 0 & \cdots & \frac{-a^j_{n,j}}{a^j_{jj}} & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

Con esta matriz se eliminan los correspondientes coeficientes en la columna j-ésima de la matriz $A(j)$, obteniendose la matriz

$$A(j+1) = E(j)A(j)$$

que tiene ceros en las posiciones que están debajo de la diagonal principal en las j primeras columnas. Finalmente se obtiene que

$$A(n) = E(n-1)A(n-1) = ... = E(n-1)...E(1)A.$$

Fácilmente se prueba que

$$E(j)^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 1 & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{a^j_{j+1,j}}{a^j_{jj}} & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \frac{a^j_{n-1,j}}{a^j_{jj}} & \cdots & \cdots & 1 & 0 \\ 0 & 0 & \cdots & \frac{a^j_{n,j}}{a^j_{jj}} & \cdots & \cdots & 0 & 1 \end{pmatrix}$$

y que $L = E(1)^{-1} \cdots E(n-1)^{-1}$ está dado por

$$L = \begin{pmatrix} 1 & 0 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ \frac{a^1_{21}}{a^1_{11}} & 1 & \cdots & 0 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & 1 & \vdots & \vdots & \vdots & \vdots \\ \frac{a^1_{j1}}{a^1_{11}} & \frac{a^2_{j2}}{a^2_{22}} & \cdots & \frac{a^j_{j+1,j}}{a^j_{jj}} & 1 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{a^1_{n-1,1}}{a^1_{11}} & \frac{a^2_{n-1,2}}{a^2_{22}} & \cdots & \frac{a^j_{n-1,j}}{a^j_{jj}} & \cdots & \cdots & 1 & 0 \\ \frac{a^1_{n,1}}{a^1_{11}} & \frac{a^2_{n,2}}{a^2_{22}} & \cdots & \frac{a^j_{n,j}}{a^j_{jj}} & \cdots & \cdots & \cdots & 1 \end{pmatrix}$$

Por otra parte, la matriz $U = A(n)$ es una matriz triangular superior y se cumple que $A = LU$.

Si se cuentan las operaciones necesarias para realizar el cálculo de la factorización LU, se obtiene que son del orden de $\frac{n^3}{3}$, contando sumas, productos y divisiones. Así si $n = 200$, el método requiere más de dos millones de operaciones elementales. En estas circunstancias el efecto de los errores de redondeo puede ser considerable. Un modo de reducir el deterioro de la solución es evitar la división por números pequeños. La técnica, llamada de pivote parcial, en la elminicación Gaussiana utiliza esta idea realizando permutaciones de las filas de modo que el pivote sea el de mayor valor absoluto en la columna considerada. A veces, el uso de estrategias como la del pivote parcial es inevitable. Este es el caso que ocurre si en el cálculo de los multiplicadores para la eliminación de los coeficientes que están por debajo de la diagonal, aparece un cero como pivote.

Para una matriz arbitraria, es evidente que la existencia de la factorización depende de la presencia de un pivote nulo en la diagonal principal. Si se utiliza una estrategia que altera dos de las $n-k+1$ filas de la matriz $A(k)$ mediante la multiplicación por una matriz de permutación $P(k)$, el proceso de eliminación Gaussiana sería el siguiente

$$A(n) = E(n-1)P(n-1)E(n-2)P(n-2)\cdots E(1)P(1)A.$$

Sin dificultad se prueba que $P(i)E(j) = \hat{\mathrm{E}}(j)P(i)$ si $i > j$ siendo $\hat{\mathrm{E}}$ la matriz que resulta de intercambiar en $E(j)$ las posiciones (i,j) y (k,j), donde $k \geqslant i > j$ es el índice de la fila que ocupa el nuevo pivote. La razón es que $P(i)$ sólo afecta a filas y columnas posteriores a la j. De ello se deduce que la factorización anterior se puede expresar como

$$U = A(n) = E(n-1)\hat{\mathrm{E}}(n-2)\cdots\hat{\mathrm{E}}(1)PA = L^{-1}PA$$

donde la matriz P representa la permutación

$$P = P(n-1)P(n-2)\cdots P(1)$$

y

$$L = \left(E(n-1)\hat{\mathrm{E}}(n-1)\cdots\hat{\mathrm{E}}(1)\right)^{-1}$$

De este modo se obtiene que

$$PA = LU$$

Es decir, el método de eliminación de Gauss permite obtener una factorización LU de la matriz de A. Si se pretendiese resolver un sistema $A\overrightarrow{x} = \overrightarrow{b}$ usando la factorización con pivote, una vez calculada ésta, se aplicaria la técnica de la sustitución retrógrada al sistema $LU\overrightarrow{x} = P\overrightarrow{b}$.

El producto de dos matrices triangulares inferiores es triangular inferior. En efecto, si A y B son matrices triangulares inferiores y $C = AB$ entonces

$$c_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj} = \sum_{k=j}^{i} a_{ik}b_{kj}.$$

Consecuentemente, si $j > i$ se tiene que $c_{ij} = 0$ como se quería probar. Además, si A y B son matrices triangulares inferiores con unos en la diagonal principal entonces

$c_{ii} = a_{ii}b_{ii} = 1$. Finalmente, se puede probar sin dificultad que la inversa de una matriz triangular inferior (cuyos elementos de a diagonal principal no se anulen) es triangular inferior. Obviamente, se pueden establecer resultados análogos para las matrices triangulares superiores. Con estos resultados se analiza la unicidad de la factorización LU de una matriz. Si se adminte que una matriz tiene dos factorizaciones LU

$$A = LU = \bar{L}\bar{U}$$

sin cambio de pivote, entonces la matriz triangular $\bar{L}^{-1}L$ con unos en la diagonal coincide con la matriz triangular superior $\bar{U}U^{-1}$. Esto solamente ocurre si ambas matrices son la identidad. Es decir, la factorización LU de una matriz, si existe es única.

Si una matriz arbitraria A tiene una factorización $A = LU$ y D representa la matriz diagonal cuyos miembros de la diagonal principal coinciden con los de U entonces la matriz A puede expresarse como $A = LD\bar{U}$ donde $\bar{U} = D^{-1}U$ es una matriz triangular superior con unos en la diagonal principal. Es frecuente usar la siguiente terminología

- $A = LU$, factorización de Doolittle
- $A = \bar{L}\bar{U}$ con $\bar{L} = LD$, factorización de Crout

de modo que la diagonal de unos está en el primer caso, en el primer factor y en el segundo caso, en el segundo.

Puesto que toda matriz de permutación P verifica que $P^tP = I$, realmente la factorización obtenida para la matriz original A es $A = P^tLU$ o lo similar para las restantes factorizaciones.

Teorema 2 Si los menores principales de A no se anulan entonces admite una factorización $A = LU$.

2.7. Métodos especiales para matrices simétricas

Otra posibilidad es factorizar la matriz en la forma $A = \tilde{L}\tilde{U}$ donde $\tilde{L} = LD^{1/2}$ y $\tilde{U} = D^{1/2}U$ (la matriz diagonal $D^{1/2}$ tiene en la diagonal principal las raíces cuadradas de los elementos de D en las mismas posiciones). La matriz $\tilde{L}$ es triangular inferior y $\tilde{U}$ es triangular superior y ambas tienen la misma diagonal principal. Además, de la unicidad de la factorización LU se deduce la unicidad de la factorización $\tilde{L}\tilde{U}$.

Si la matriz A es simétrica entonces $A = \tilde{L}\tilde{U} = \tilde{U}^t\tilde{L}^t$. De la unicidad de la factorización $\tilde{L}\tilde{U}$ se deduce $\tilde{L} = \tilde{U}^t$, es decir, que existe una factorización $A = \tilde{L}\tilde{L}^t$ donde $\tilde{L}$ es una matriz triangular inferior. Esta factorización se conoce como factorización de Cholesky de una matriz simétrica.

Teorema 3 Una matriz simétrica A es definida positiva si y sólo si admite una factorización de Cholesky $A = LL^t$ donde L es una matriz triangular inferior invertible.

Si una matriz A admite una factorización de Cholesky $A = LL^t$ entonces

$$a_{ij} = \sum_{k=1}^{n} l_{ik}l_{jk} = \sum_{k \leqslant \min\{i,j\}} l_{ik}l_{jk}$$

Para $j = 1$ se tiene que $a_{i1} = l_{i1}l_{11}$ para $i = 1, \cdots, n$, de donde se deduce que

$$l_{11} = \sqrt{a_{11}}, \quad l_{i1} = \frac{a_{i1}}{l_{11}}.$$

Para $j > 1$ e $i > j$, se tiene que

$$l_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} l_{jk}^2}, \quad l_{ij} = \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}}{l_{jj}}.$$

Este modo de realizar el cálculo de L se adapta bien a las formas de almacenamiento reducido. Por ejemplo, si la matriz A es tridiagonal, la matriz L es tridiagonal.

2.8. Factoricación QR

Las matrices ortogonales conservan las longitudes de los vectores y están perfectamente condicionadas, de modo que en su uso, no se espera que se produzcan crecientes notables errores de redondeo. El siguiente resultado puede verse como una variante del que establece la factorización LU de una matriz con la modificación de que uno de los factores es una matriz ortogonal

Teorema 4 Toda matriz invertible A admite una factorización $A = QR$ donde Q es una matriz ortogonal y R una matriz triangular superior.

Se puede usar una factorización QR de la matriz A para resolver el sistema lineal $A\overrightarrow{x} = \overrightarrow{b}$. En efecto, una vez calculada la factorización bastaría resolver el sistema triangular $R\overrightarrow{x} = Q^t\overrightarrow{b}$ por sustitución retrógrada. Si bien es ierto que en la norma subordinada a la norma euclídea, el número de condición de A y R coinciden, lo cierto es que evita la sustitución progresiva y consecuentemente limita la generación de errores ed redondeo.

2.8.1. Método de ortogonalización de Gram-Schmidt

La idea básica de este método aplicado a una matriz invertible A es la siguiente:

- Se consideran los n vectores columna $\{\overrightarrow{a}^i : i = 1, 2, \cdots, n\}$de A.
- A cada uno de ellos se le restan las componentes respecto a los vectores que le preceden.
- Finalmente se normalizan esos vectores.

No obstante, se pueden separar las etapas de ortogonalización y normalización del siguiente modo: Se introduce

la nueva base

$$\begin{aligned}\overrightarrow{p}^1 &= \overrightarrow{a}^1\\ \overrightarrow{p}^2 &= \overrightarrow{a}^2 - \frac{\overrightarrow{a}^2\cdot\overrightarrow{p}^1}{\|\overrightarrow{p}^1\|^2}\overrightarrow{p}^1\\ &\cdots\\ \overrightarrow{p}^n &= \overrightarrow{a}^n - \sum_{i=1}^{n-1}\frac{\overrightarrow{a}^n\cdot\overrightarrow{p}^i}{\|\overrightarrow{p}^i\|^2}\overrightarrow{p}^i.\end{aligned}$$

Si P representa la matriz de columnas $\{\overrightarrow{p}^i : i = 1, 2, \cdots, n\}$, las ecuaciones anteriores pueden expresarse en forma matricial como

$$P = A - P\begin{pmatrix} 0 & m_{12} & \cdots & m_{1n}\\ 0 & 0 & \cdots & m_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & m_{n-1,n}\\ 0 & 0 & \cdots & 0\end{pmatrix}$$

con $m_{ij} = \frac{\overrightarrow{a}^j\cdot\overrightarrow{p}^i}{\|\overrightarrow{p}^i\|^2}$ para $j > i$, de donde se deduce que

$$A = P\begin{pmatrix} 1 & m_{12} & \cdots & m_{1,n-1} & m_{1n}\\ 0 & 1 & \cdots & m_{2,n-1} & m_{2n}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & 1 & m_{n-1,n}\\ 0 & 0 & \cdots & 0 & 1\end{pmatrix}$$

Si D es la matriz diagonal definida por $d_{ii} = \|\overrightarrow{p}^i\|$, las matrices $Q = PD^{-1}$ y

$$R = D\begin{pmatrix} 1 & m_{12} & \cdots & m_{1,n-1} & m_{1n}\\ 0 & 1 & \cdots & m_{2,n-1} & m_{2n}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & 0 & \cdots & 1 & m_{n-1,n}\\ 0 & 0 & \cdots & 0 & 1\end{pmatrix}$$

forman una factorización QR de la matriz A.

El proceso de Gram-Schmidt requiere que los vectores columnas sean linealmente independientes ya que si uno de ellos es combinación lineal de los precedentes, el vector resultante es nulo. De hecho el método se vuelve inestable cuando los vectores columnas están proximos a ser linealmente dependientes.

2.8.2. Método de Householder

Una transformación de Householder es una aplicación lineal que lleva un vector en su reflejado respecto a un eje. La matriz asociada a una transformación de Householder tiene la siguiente forma

$$H = I - 2\overrightarrow{e}\,\overrightarrow{e}^t$$

donde $\overrightarrow{e}$ es un vector de norma 1 que representa al eje. Se prueba sin dificultad que H es una matriz simétrica ortogonal.

Siempre se puede encontrar una aplicación lienal que transforme un vector $\overrightarrow{x}$ fijado en otro vector $\overrightarrow{y}$ de la misma norma, también fijado. Para construirla, basta tomar

$$\overrightarrow{e} = \pm\frac{\overrightarrow{x} - \overrightarrow{y}}{\|\overrightarrow{x} - \overrightarrow{y}\|}$$

En efecto, de la identidad

$$\|\overrightarrow{x}\|^2 - \overrightarrow{x}\cdot\overrightarrow{y} = \frac{1}{2}(\|\overrightarrow{x}\|^2 - \|\overrightarrow{y}\|^2) + \frac{1}{2}\|\overrightarrow{x} - \overrightarrow{y}\|^2,$$

teniendo en cuenta que $\|\overrightarrow{x}\| = \|\overrightarrow{y}\|$, se deduce que

$$H\overrightarrow{x} = \overrightarrow{x} - 2\frac{\overrightarrow{x} - \overrightarrow{y}}{\|\overrightarrow{x} - \overrightarrow{y}\|^2}(\|\overrightarrow{x}\|^2 - \overrightarrow{x}\cdot\overrightarrow{y}) = \overrightarrow{x} - (\overrightarrow{x} - \overrightarrow{y}) = \overrightarrow{y}.$$

También es evidente que la transformación de Householder $H^* = I - 2\overrightarrow{e'}\,\overrightarrow{e'}^t$ para

$$\overrightarrow{e'} = \pm\frac{\overrightarrow{x} + \overrightarrow{y}}{\|\overrightarrow{x} + \overrightarrow{y}\|}$$

transforma $\overrightarrow{x}$ en $-\overrightarrow{y}$ y verifica que $H^* = 2\overrightarrow{e}\,\overrightarrow{e}^t - I$.

Se examina ahora cómo pueden usarse las transformaciones de Householder para construir una factorización QR de una matriz A. Si se representan por $\overrightarrow{a}^i$ los vectores columna de la matriz A y P es una matriz arbitraria de las mismas dimensiones que A, entonces

$$PA = (P\overrightarrow{a}^1, P\overrightarrow{a}^2, \cdots, P\overrightarrow{a}^n)$$

donde las comas indican la concatenación de los vectores columna $P\overrightarrow{a}^i$. El método de Householder utiliza la siguiente construcción: En la primera etapa, se construye la transformación de Householder $P(1)$ que transforma $\overrightarrow{a}^1$ en el vector $\|\overrightarrow{a}^1\|(1, 0, \cdots, 0)^t$. La construcción de esta transformación es posible ya que en este caso ambos vectores tienen la misma norma. Si se representa $A(1) = A$ y se define la matriz $A(2) = P(1)A(1)$, esta última matriz tiene ceros debajo del primer elemento de la diagonal principal. Se descompone la matriz $A(2)$ como sigue

$$A(2) = \left(\begin{array}{c|c} a_{11}(2) & A_{12}(2)\\ \hline 0 & A_{22}(2)\end{array}\right)$$

donde $A_{22}(2)$ es una matriz cuadrada de dimensión $n-1$. Se construye una matriz de Householder $P_{22}(2)$ de dimensión $n-1$ que produzca ceros debajo del primer elemento de la diagonal y con ella se construye la matriz de dimensión n

$$P(2) = \left(\begin{array}{c|c} 1 & 0\\ \hline 0 & P_{22}(2)\end{array}\right).$$

Con esta matriz ortogonal se construye $A(3) = P(2)A(2) = P(2)P(1)A$ que tiene ceros debado de la diagonal en las dos primeras columnas. Si se reitera el procedimiento hasta construir la matriz triangular superior

$$R = P(n-1)\cdots P(2)P(1)A.$$

Si se toma

$$Q = P(1)P(2)\cdots P(n-1)$$

se obtiene una matriz ortogonal tal que $A = QR$.

Ahora se analiza la unicidad de la factorización QR. Supongamos que A tiene dos descomposiciones QR,

$$A = QR = \hat{Q}\hat{R}.$$

Se considera la matriz $T = \hat{Q}^t Q = \hat{R}R^{-1}$. En consecuencia $T = \hat{Q}^t Q$ es ortogonal y también $T = \hat{R}R^{-1}$ es triangular superior. Puesto que la inversa de una matriz ortogonal es su traspuesta, la matriz T tiene que ser diagonal con coeficientes en valor absoluto igual a 1. De la igualdad $\hat{R} = TR$ se deduce que si se impone la condición de que los coeficientes de la diagonal principal en R son positivos entonces la factorización QR es única. Por otra parte, de las construcciones de los apartados anteriores se deduce que toda matriz no singular A adminte una única factorización $A = QR$ donde Q es una matriz ortogonal y R una matriz triangular superior con coeficientes positivos en la diagonal principal.

2.9. Métodos iterativos

La idea básica de esta clase de métodos es la de considerar la solución $\overrightarrow{x}$ como un punto fijo de una transformación afín

$$T\overrightarrow{x} = H\overrightarrow{x} + \overrightarrow{c}.$$

La aplicación reiterada de esta transformación

$$\overrightarrow{x}^{(k+1)} = H\overrightarrow{x}^{(k)} + \overrightarrow{c}$$

puede conducir en el límite a la solución si se cumplen determinadas condiciones de convergencia.

Teorema 5 Sean H una matriz cuadrada y $\overrightarrow{c}$ un vector tales que la ecuación $\overrightarrow{x} = H\overrightarrow{x} + \overrightarrow{c}$ tiene una solución única. La sucesión $\{\overrightarrow{x}^{(k)}\}$ generada por la transformación

$$\overrightarrow{x}^{(k+1)} = H\overrightarrow{x}^{(k)} + \overrightarrow{c}$$

con un vector de partida arbitrario $\overrightarrow{x}^{(0)}$, converge a $\overrightarrow{x}$ si y sólo si $\rho(H) < 1$.

2.10. Métodos iterativos clásicos

Hay muchos modos de interpretar la solución de un sistema lineal $A\overrightarrow{x} = \overrightarrow{b}$ como un punto fijo de una transformación $\overrightarrow{x} = H\overrightarrow{x} + \overrightarrow{c}$. No obstante, el llamado método de relajación agrupa a algunos de ellos y puede utilizarse incluso en el caso no-lineal. La idea de este método es la siguente: Se considera una matriz M fácilmente invertible que se conoce como preacondicionador (preconditioner). Con ayuda de esta matriz se construye la ecuación de punto fijo

$$\overrightarrow{x} = \overrightarrow{x} - M^{-1}(A\overrightarrow{x} - \overrightarrow{b})$$

que se acomoda a la situación de la sección anterior si se toma $H = I - M^{-1}A$ y $\overrightarrow{c} = M^{-1}\overrightarrow{b}$. De acuerdo con el teorema 5, la dificultad está en encontrar preacondicionadores que verifiquen

$$\rho(I - M^{-1}A) < 1$$

y que la resolución del sistema lineal

$$M(\overrightarrow{x}^{(k+1)} - \overrightarrow{x}^{(k)}) = \overrightarrow{b} - A\overrightarrow{x}^{(k)}$$

pueda realizarse de un modo muy simple.

Un grupo de métodos iterativos clásicos puede ser incluido en esta clase, si se descompone la matriz A como

$$A = D - L - U$$

donde

- D es una matriz diagonal verificando que
$$d_{ii} = a_{ii} \quad \text{para } i = 1, \cdots, n$$
- L es una matriz triangular inferior verificando que
$$l_{ij} = -a_{ij} \quad \text{para } i, j = 1, \cdots, n,\ i > j$$
- U es una matriz triangular superior verificando que
$$u_{ij} = -a_{ij} \quad \text{para } i, j = 1, \cdots, n,\ i < j$$

Con ayuda de estas matrices se definen los siguientes preacondicionadores para la relajación del sistema

1. Método de Jacobi $M = D$. La iteración resultante es
$$D\overrightarrow{x}^{(k+1)} = (L + U)\overrightarrow{x}^{(k)} + \overrightarrow{b}.$$
2. Método Gauss-Seidel $M = D - L$. La iteración resultante es
$$(D - L)\overrightarrow{x}^{(k+1)} = U\overrightarrow{x}^{(k)} + \overrightarrow{b}.$$
3. Método de sobre-relajación (Successive-Over-Relaxation) $M = \frac{1}{\omega}D - L$. La iteración resultante es
$$\left(\frac{1}{\omega}D - L\right)\overrightarrow{x}^{(k+1)} = \left(\left(\frac{1}{\omega} - 1\right)D + U\right)\overrightarrow{x}^{(k)} + \overrightarrow{b}.$$
El parámetro de relajación ω permite escalar adecuadamente la diagonal del sistema frente a la parte inferior a la diagonal.

En la puesta en práctica del método de Jacobi hay que tener cuidado en no mezclar las componentes $x_i^{(k+1)}$ con las componentes $x_i^{(k)}$. Es necesario no almacenar el valor de $x_i^{(k+1)}$ en la posición que ocupa $x_i^{(k)}$ ya que este valor se utilizará en el cálculo en los nodos adyacentes. Por lo tanto es preciso mantener simultáneamente dos vectores para poder realizar la iteración completa.

El cálculo del radio espectral de H para verificar la condición necesaria y suficiente para la convergencia, no es en general sencillo. Sin embargo, el teorema 5 permite establecer algunas condiciones suficietnes para la convergencia, que pueden ser fácilmente verificables. En concreto, las condiciones

- $\|H\|_\infty < 1$
- $\|H\|_1 < 1$

garantizan la convergencia del método.

A fin de expresar de un modo más simple estas condiciones en el caso de los métodos clásicos se introducen las siguientes definiciones

- A es estrictamente diagonal dominante por filas si $\sum_{i \neq i} |a_{ij}| < |a_{ii}|$ para cada valor de i.
- A es estrictamente diagonal dominantes por columnas si $\sum_{i \neq j} |a_{ij}| < |a_{jj}|$ para cada valor de j.

que permiten fomular el siguiente

Teorema 6 Si A es una matriz estrictamente diagonal dominante por filas entonces los métodos de Jacobi y Gauss-Seidel son convergentes.

Teorema 7 Si A es una matriz simétrica definida positiva, el método de Gauss-Seidel es convergente.

Teorema 8 Si $\omega \notin (0,2)$ el método SOR no es convergente. Si A es una matriz simétrica definida positiva, el método SOR es convergente si y sólo si $0 < \omega < 2$.

En el espacio vectorial de las matrices reales se considera la siguiente relación de orden parcial: una matriz D verifica la relación $D \geqslant 0$ si y sólo si $d_{ij} \geqslant 0$ para $i, j = 1, \cdots, n$. Las siguientes definiciones se apoyan en esta relación de orden

- $A = B - C$ es una descomposición regular de A si $B^{-1} \geqslant 0$ y $C \geqslant 0$.
- A es una M-matriz si $a_{ij} \leqslant 0$ para $i \neq j$ y $A^{-1} \geqslant 0$.

Teorema 9

- Si $A^{-1} \geqslant 0$ y $A = B - C$ es una descomposición regular entonces $\rho(B^{-1}C) < 1$.
- Si A es una M-matriz, entonces las descomposiciones de Jacobi Y GAuss-Seidel son regulares. En este caso, ambos métodos son convergentes.

Una etapa fundamental en la aplicación de un método iterativo es la adecuación previa de la matriz. Un sistema lineal $A\overrightarrow{x} = \overrightarrow{b}$ no modifica sus soluciones si se multiplican ambos miembros de la igualdad por una matriz no-singular. Las elecciones más convenientes de estas matrices corresponden a matrices de permutación que alteran el orden de las ecuaciones o de las incógnitas o matrices diagonales que modifican las escalas de los coeficientes.

3. Aproximación de autovalores

3.1. Autovalores y vectores propios

Un vector propio es un vector que se transforma en un múltiplo de sí mismo por una aplicación lineal y la constante de proporcionalidad del transformado de un vector propio respecto al orignal se conoce como autovalor. De este modo, si λ es un autovalor real de una matriz real A, existe un vector $\overrightarrow{x}$ (vector propio) que no se anula, tal que $A\overrightarrow{x} = \lambda\overrightarrow{x}$. Además de los autovalores reales, una matriz puede tener autovalores complejos cuando se extiende A como transformación lineal de $\mathbb{C}^n$ en $\mathbb{C}^n$.

Salvo en los casos de baja dimensión o del algunas matrices especiales, no existen fórmulas explícitas que permitan calcular de un modo directo los autovalores de una matriz que deben ser aproximados por métodos iterativos. Existen dos clases de métodos para aproximar los autovalores de una matriz:

- Un primer grupo que contempla al autovalor como una raíz de un polinomio y usa técnicas especializadas en resolución de ecuaciones escalares numéricas. La condición de autovalor se puede expresar como

  $$\det(A - \lambda I) = 0$$

  El primer miembro de esta ecuación es el polinomio característico de grado menos o igual que n en la variable λ, que tiene como raíces, n autovalores complejos, si cada uno se cuenta tantas veces como indica su multiplicidad algebraica.
- Un segundo grupo que enfoca el problema como matricial y no implica directamente el concepto de polinomio característico.

Se dedica esta sección a los métodos basados en el análisis matricial.

Si A está representada por una matriz triangular como la siguiente

$$A = \begin{pmatrix} a_{11} & a_{22} & \cdots & a_{1,n-1} & a_{1,n} \\ 0 & a_{22} & \cdots & a_{2,n-1} & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & \cdots & 0 & a_{nn} \end{pmatrix}$$

los autovalores son $\{a_{11}, a_{22}, \cdots, a_{nn}\}$.

Teorema 10 (de Schur) Para toda matriz real cuadrada A existe una matriz ortogonal P tal que $T = P^{-1}AP$ es una matriz triangular por bloques, siendo los bloques, matrices 1×1 ó 2×2. En cada caja 1×1, aparece un autovalor real y en cada caja 2×2, aparecen dos autovalores complejos conjugados.

Puesto que la matriz inversa de una matriz ortogonal coincide con su traspuesta, se tiene que T y $A = PTP^{-1}$ son semejantes y por consiguiente, tienen los mismos autovalores.

3.2. Sucesiones de Krylov

Se introduce el concepto de sucesión de Krylov, definida como la acción de las potencias sucesivas de una matriz sobre un vector fijado. De un modo más preciso, se llama sucesión de Krylov asociada al vector $\overrightarrow{x}$ y a la matriz cuadrada A, a la siguiente sucesión

$$\{\overrightarrow{x}, A\overrightarrow{x}, A^2\overrightarrow{x}, \cdots, A^n\overrightarrow{x}, \cdots\}.$$

Sea $p_n(\lambda) = \sum_{i=0}^{n} a_i\lambda^i$ el polinomio característico de A. De acuerdo con el teorema de Cayley-Hamilton la matriz A es solución de la ecuación característica matricial. Es decir, A verifica que

$$(-1)^n A^n + a_{n-1}A^{n-1} + \cdots + a_1 A + a_0 I = O$$

donde O representa la matriz cero. Si se multiplican ambos miembros por un vector arbitrario $\overrightarrow{x}$ se obtiene la expresión

$$a_{n-1}A^{n-1}\overrightarrow{x} + \cdots + a_1 A\overrightarrow{x} + a_0\overrightarrow{x} = -(-1)^n A^n\overrightarrow{x}$$

que establece que el $(n+1)$-ésimo vector de la sucesión de Krylov es linealmente dependiente de los que le preceden. Esta relación puede ser vista como un sistema lineal de incógnitas $a = (a_0, a_1, \cdots, a_{n-1})^t$

$$(\overrightarrow{x}|A\overrightarrow{x}|A^{n-1}\overrightarrow{x})a = (-1)^{n+1}A^n\overrightarrow{x}$$

que podría ser utilizado para calcular el polinomio característico.

Sin embargo, una dificultad podría estar en que el conjunto de los n primeros vectores de la sucesión de Krylov no fuese un conjunto de vectores independientes, en cuyo caso el sistema lineal podría ser singular. El polinomio q de grado mínimo que verifique la ecuación matricial $q(A) = O$ se conoce como polinomio mínimo de la matriz A. Obviamente, el grado del polinomio mínimo es menor o igual que n.

Teorema 11 El polinomio mínimo de una matriz A verifica las siguientes propiedades:

- El polinomio mínimo existe y es único.
- El polinomio mínimo es invariante frente a transformaciones de semejanza.
- El polinomio mínimo tiene las mismas raíces que el polinomio característico aunque posiblemente con multiplicidad inferior.

Se llama grado de Krylov del vector $\overrightarrow{x}$ respecto a la matriz A al número máximo de vectores linealmente independientes en la sucesión de Krylov asociada a $\overrightarrow{x}$. El grado de Krylov de cualquier vector está entre 1 (este es el caso de un vector propio) y el grado del polinomio mínimo de la matriz.

Una vez determinado el polinomio característico de una matriz, sus autovalores podrían ser calculados como las raíces de este polinomio. Existe la posibilidad de recorrer el camino inverso. Es decir, dado un polinomio $p(x) = \sum_{i=0}^{n} a_i x^i$ de una variable se busca una matriz que tenga ese polinomio como el característico. Obviamente, esta matriz no es única ya que todas las matrices equivalentes tienen el mismo polinomio característico. Una de estas matrices es la llamada matriz de compañia del polinomio, que está definida como

$$C(p) = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ -\frac{a_0}{a_n} & -\frac{a_1}{a_n} & -\frac{a_2}{a_n} & \cdots & -\frac{a_{n-1}}{a_n} \end{pmatrix}$$

o su traspuesta. Más precisamente, la matriz de compañía $C(p)$ tiene como polinomio característico el polinomio $(-1)^n \frac{p(x)}{a_n}$ y consecuentemente el polinomio p tiene como raíces, los autovalores de la matriz $C(p)$.

3.3. Método de la potencia iterada

Sea A una matriz diagonalizable mediante una matriz P y cuyos autovalores puedan ser ordenados como

$$|\lambda_1| > |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$$

de modo que la primera desigualdad sea estricta. La sucesión de Krylob asociada a un vector $\overrightarrow{x}^{(0)}$ arbitrario que no sea nula

$$\overrightarrow{x}^{(k)} = A\overrightarrow{x}^{(k-1)} = A^k\overrightarrow{x}^{(0)} = PD^kP^{-1}\overrightarrow{x}^{(0)}$$

está formada por vectores de norma creciente hacia el infinito salvo que el radio espectral de A sea menor que 1, en cuyo caso tienen al vector 0. De modo más preciso

$$\frac{1}{\lambda_1^k}D^k = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \left(\frac{\lambda_2}{\lambda_1}\right)^k & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \left(\frac{\lambda_n}{\lambda_1}\right)^k \end{pmatrix}$$

que tiene por diagonal principal potencias cuya base tiene módulo menor o igual que 1 y consecuentemente el límite

$$\lim_{k\to\infty} \frac{1}{\lambda_1^k}D^k = D_0$$

es una matriz con ceros en la diagonal principal salvo en la posición $(1,1)$.

Es decir

$$\lim_{k\to\infty} \frac{\overrightarrow{x}^{(k)}}{\lambda_1^k} = \lim_{k\to\infty} P\frac{1}{\lambda_1^k}D^kP^{-1}\overrightarrow{x}^{(0)} = PD_0^{-1}\overrightarrow{x}^{(0)} = \overrightarrow{y}^{(0)}.$$

Este resultado tiene en principio poco interés ya que el límite $\overrightarrow{y}^{(0)}$ es un vector que depende linealmente de $\overrightarrow{x}^{(0)}$ a través de una matriz que implica la matriz P y que se supone que es uno de los objetivos de nuestro cálculo. Sin embargo, esta dependencia desaparece si se calcula el límite de los cocientes de primeras componentes del vector $\overrightarrow{x}^{(k)}$. En efecto, si $y_1^{(0)} \neq 0$ se tiene que

$$\lim_{k\to\infty} \frac{x_1^{(k)}}{x_1^{(k-1)}} = \lambda_1 \lim_{k\to\infty} \frac{\frac{x_1^{(k)}}{\lambda_1^k}}{\frac{x_1^{(k-1)}}{\lambda_1^{k-1}}} = \lambda_1.$$

Este límite no depende ni de $\overrightarrow{y}^{(0)}$ ni del vector de partida $\overrightarrow{x}^{(0)}$. El resultado sería aún válido si el cociente se efectúa entre las componentes i-ésima de los vectores de la sucesión en vez del cociente de las primeras componentes. Es importante tener en cuenta que $y_1^{(0)} = 0$ si la primera componente de $P^{-1}\overrightarrow{x}^{(0)}$ es igual a 0. Puesto que el primer vector columna de P es un vector propio asociado a λ_1, una condición necesaria y suficiente para que no se produzca una forma indeterminada en el límite anterior, es que $x^{(0)}$ tenga componente distinta de 0 en la dirección de un vector propio asociado a λ.

La principal crítica que merece este método es que cuando el número de iteraciones es elevado, nos ecnontramos

con el cociente de dos números o bien muy pequeños o bien muy grandes lo que produce el consiguiente deterioro del cálculo.

No obstante, puede salvarse esta dificultad si en cada iteración se normaliza el resultado con una norma

$$\overrightarrow{w}^{(k)} = \frac{A\overrightarrow{w}^{(k-1)}}{\|A\overrightarrow{w}^{(k-1)}\|}.$$

En este caso, la sucesión se mantendrá siempre en el conjunto de los vectores de norma 1. Además

$$\lim_{k\to\infty} w^k = \lim_{k\to\infty} \frac{A^k\overrightarrow{w}^{(0)}}{\|A\overrightarrow{w}^{(0)}\|} = \frac{\overrightarrow{y}^{(0)}}{\|\overrightarrow{y}^{(0)}\|}$$

$$\lim_{k\to\infty} A\overrightarrow{w}^{(k)} = \lim_{k\to\infty} \frac{A^{k+1}\overrightarrow{w}^{(0)}}{\|A\overrightarrow{w}^{(0)}\|} = \lambda_1 \frac{\overrightarrow{y}^{(0)}}{\|\overrightarrow{y}^{(0)}\|}$$

Consecuentemente se tiene que

$$\lim_{k\to\infty} \frac{(A\overrightarrow{w}^{(k)})_1}{w_1^k} = \lambda_1$$

Esta es la idea básica del método de la potencia iterada con normalización que permite calcular el autovalor dominante de una matriz.

La cuestión que ahora se plantea es cómo se pueden calcular los restantes autovalores. La siguiente idea da una orientación sobre el modo en que esto podría llevarse a cabo: Los autovalores de la matriz $A - \alpha I$ para cualquier escalar α son $\{\lambda - \alpha : \lambda$ es un autovalor de $A\}$. En otras palabras, se pueden desplazar los autovalores de una matriz perturbando la matriz con un múltiplo de la identidad. Sin embargo, con un desplazamiento un autovalor real intermedio no puede ser convertido en el de máximo valor absoluto. Para calcular el autovalor de menor valor absoluto, se puede utilizar un razonamiento, en cierto modo inverso que el utilizado en el de la potencia. En efecto, si los autovalores de una matriz están ordenados como

$$|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_{n-1}| > |\lambda_n| > 0$$

entonces los autovalores de A^{-1} están ordenados como

$$\frac{1}{|\lambda_n|} > \frac{1}{|\lambda_{n-1}|} \geqslant \cdots \geqslant \frac{1}{|\lambda_1|}.$$

En estas condiciones se puede utilizar el método de la potencia a la matriz de A^{-1}. De ese modo se podría aproximar el autovalor de menor valor absoluto mediante el cociente $\frac{w_1^{(k-1)}}{x_1^{(k)}}$ asociado a la iteración

$$A\overrightarrow{x}^{(k)} = \overrightarrow{w}^{(k-1)}, \quad \overrightarrow{w}^{(k)} = \frac{\overrightarrow{x}^{(k)}}{\|\overrightarrow{x}^{(k)}\|}$$

para $k > 0$. Este método se conoce como el de la potencia inversa.

El método de la potencia no tiene una validez general ya que requiere que exista un único autovalor dominante. Este no es el caso de matrices como las ortogonales cuyos autovalores tienen módulo 1. En este caso el límite qye aparece en la ecuación $\lim_{k\to\infty} \frac{(A\overrightarrow{w}^{(k)})_1}{w_1^k} = \lambda_1$ puede no existir. De hecho, si A es una matriz que tiene como autovalores $\lambda_1 = 1$ y $\lambda_2 = -1$ el límite de la sucesión de potencias

$$D^k = \begin{pmatrix} 1 & 0 \\ 0 & (-1)^k \end{pmatrix}$$

no existe.

3.4. Método QR

Existen algunos métodos que permiten construir una matriz triangular que *casi* sea semejante a una matriz dada. Uno de ellos, llamado método QR para el cálculo de los autovalores para una matriz A, está basado en la factorización QR como sugiere su nombre. Para ponerlo en práctica, se construye la siguiente sucesión: $A^{(1)} = A$. Con la factorización $A^{(1)} = Q^{(1)}R^{(1)}$ se construye $A^{(2)} = R^{(1)}Q^{(1)}$. Es fundamental constatar que $A^{(1)}$ y $A^{(2)}$ tienen los mismos autovalores ya que son semejantes

$$A^{(1)} = Q^{(1)}A^{(2)}(Q^{(1)})^t.$$

Reiterando el procedimiento se construye la sucesión de matrices

$$\{A^{(1)}, A^{(2),\cdots,A^{(k)}}, \cdots\}.$$

Teorema 12 Si los autovalores de una matriz A verifican que

$$|\lambda_1| > \cdots > |\lambda_n| > 0$$

entonces la sucesión de matrices equivalentes construidas por el algoritmo QR converge a una matriz triangular superior.

4. Aproximación de funciones

4.1. Introducción

Cuando se quiere evaluar una función de cierta complejidad, es conveniente pensar que el cálculo automático que realiza un computador solamente emplea operaciones aritméticas básicas junto con las más simples operaciones de comparación de números en el rango de la máquina. Esta limitación implica que en estas circunstancias, debe introducirse algún tipo de aproximación por funciones simples que puedan ser programadas directamente mediante un número finito de instrucciones.

4.2. Evaluación de polinomios

La evaluación por sustitución directa de la variable x en el polinomio

$$p(x) = \sum_{i=0}^{n} a_i x^i$$

requiere

- $(1 + 2 + \cdots + n - 1) = \frac{n(n-1)}{2}$ multiplicaciones para calcular las potencias.
- n multiplicaciones de las potencias calculadas por los coeficientes.

- n sumas.

En total, requiere $\frac{n^2+n}{2}$ multiplicaciones $+n$ sumas.

Sin embargo, los polinomios admiten la siguiente representación anidada

$$p(x) = a_0 + x(a_1 + x(a_2 + \cdots + x(a_{n-1} + a_n x)\cdots)),$$

que podría ser utilizada para una evaluación más eficiente.

Si se representa por q_i el polinomio definido por

$$q_i(x) = a_{i-1} + a_i x$$

para $i = 1, \cdots, n$, se puede evaluar el polinomio p como

$$p(x) = q_1(q_2(\cdots(q_n(x))))$$

para cualquier valor de x.

Este modo de evaluar iterativamente las potencias, que se conoce como algoritmo de Horner, permite reducir el número de operaciones a $n-1$ multiplicaciones y $n-1$ sumas. Se puede probar que para la evaluación directa de un polinomio, este procedimiento es óptimo en el sentido de que no es posible encontrar otro procedimiento que implique menos multiplicaciones.

Tradicionalmente se llama regla de Ruffini al uso de la representación anidada en forma de cuadro que se utiliza para ilustrar el uso de la evaluación de un polinomio mediante representación anidada.

4.3. Aproximación de funciones

Sea V un espacio vectorial de funciones de una variable real, dotado de una norma $\|\ \|$. Es decir, el espacio V tiene asociada una aplicación $\|\ \| : V \to \mathbb{R}^+$ que verifican las siguientes propiedades

1. $\|f\| = 0 \Leftrightarrow f = 0$,
2. $\|\lambda f\| = |\lambda|\,\|f\|$,
3. $\|f + g\| \leqslant \|f\| + \|g\|$

para todas la elecciones de $\lambda \in \mathbb{R}$ y $f, g \in V$. Si se fija un subespacio vectorial $U \subset V$ de dimensión finita, se puede considerar el siguiente problema de aproximación:

Dada una función $f \in V$ arbitraria, hallar $g \in U$ tal que

$$\|f - g\| \leqslant \|f - h\|$$

para toda función $h \in U$. Una función g que minimiza la distancia a f, se denomina mejor aproximación ó proyección de f en U con la norma $\|\ \|$.

En muchas ocasiones en esta sección, el espacio de funciones V considerado será el espacio de las funciones continuas en un intervalo $[a, b]$. En este espacio uno de los modos más simples de medir la distancia entre elementos es mediante la siguiente norma

$$\|f\|_2 = \left(\int_a^b f(x)^2\,\mathrm{dx}\right)^{1/2}.$$

Más general es la norma $\|\ \|_p$ definida por

$$\|f\|_p = \left(\int_a^b |f(x)|^p\right)^{1/p}.$$

Como caso límite, la norma $\|\ \|_\infty$ está definida por

$$\|f\|_\infty = \max_{x\in[a,b]} |f(x)|.$$

Típicamente, la aproximación con la norma $\|\ \|_2$ se llama mínimos cuadrados mientras que la aproximación con la norma $\|\ \|_\infty$ se llama aproximación uniforme o min-max.

La mejor aproximación a una función dada puede ser bastante diferente, dependiendo de la norma usada ya que normas distintas pueden expresar diferentes criterios para medir la proximidad entre dos funciones.

Teorema 13 Sea V un espacio normado de funciones. Para toda función $f \in V$ y todo subespacio vectorial $U \subset V$ de dimensión finita, existe al menos una función $g \in U$ que realiza la mejor aproximación a f en U. Además, si la norma es estrictamente convexa, es decir, si se verifica la desigualdad

$$\|tg + (1-t)h\| < t\|g\| + (1-t)\|h\| = \|g\|$$

para todas las funciones $g, h \in V$ tales que $\|f\| = \|g\|$ y los escalares $t \in (0, 1)$, entonces existe una única función que realiza la mejor aproximación.

En la mayoría de las situaciones particulares consideradas en este capítulo, V será el espacio de las funciones continuas en un intervalo cerrado y acotado y el subespacio vectorial U será el espacio de los polinomios de una variable de grado menor ó igual que n, para un entero n fijado previamente. Se usará la notación $\mathcal{P}_n$ para representar este subespacio de dimensión $n+1$.

4.4. Aproximación por mínimos cuadrados

El concepto de producto escalar generaliza el concepto de producto escalar ecuclídeo y es particularmente útil para estudiar la mejor aproximación a una función cuando se usan normas que provienen de él. Se define como una aplicación $\langle\ ,\ \rangle : V \times V \to \mathbb{R}$ que verifica las siguientes propiedades

1. $\langle f, f\rangle = 0 \Leftrightarrow f = 0$,
2. $\langle f, g\rangle = \langle g, f\rangle$,
3. $\langle \alpha f + \eta g, h\rangle = \alpha\langle f, h\rangle + \eta\langle g, h\rangle$

para toda elección de $\alpha, \eta \in \mathbb{R}$ y $f, g, h \in V$.

Todo producto escalar permite definir una norma mediante la expresión

$$\|f\| = \langle f, f\rangle^{1/2}$$

para todo $f \in V$. En particular, el producto escalar

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

define la norma $\| \|_2$.

Se destaca, sobre las propiedades de productos escalares y normas, la relevante desigualdad de Cauchy-Schwarz

$$|\langle f, g\rangle| \leqslant \|f\| \, \|g\|$$

para todas las funciones $f, g \in V$. La igualdad se alcanza si y sólo si existen escalares λ y η, tales que $\lambda f + \eta g = 0$ y $|\lambda| + |\eta| > 0$.

Toda norma definida por un producto escalar es estrictamente convexa y el problema de mejor aproximación en un espacio de dimensión finita siempre tiene una única solución.

El vector $\overrightarrow{\alpha}$ que realiza el mínimo de J, anula su gradiente. Es decir, se cumple que

$$\frac{\partial J}{\partial \alpha_j} = 2 \left\langle f - \sum_0^n \alpha_i \sigma_i, \sigma_j \right\rangle = 0$$

para $j = 0, 1, \cdots, n$.

Estas relaciones pueden organizarse en forma de lo que se conoce como las ecuaciones normales del problema de aproximación

$$G\overrightarrow{\alpha} = \overline{f}$$

donde $\overline{f}$ es el vector de componentes $\overline{f}_i = \langle f, \sigma_i \rangle$ y G la matriz de Gram definida por

$$G_{ij} = \langle \sigma_i, \sigma_j \rangle$$

para $i, j = 0, 1, \cdots, n$.

Fácilmente se comprueba que la matriz de Gram asociada a la base, es simétrica y definida positiva. En efecto, G es definida positiva ya que

$$\overrightarrow{\alpha}^t G \overrightarrow{\alpha} = \left\langle \sum_0^n \alpha_i \sigma_i, \sum_0^n \alpha_i \sigma_i \right\rangle = \| \sum_0^n \alpha_i \sigma_i \|^2 > 0$$

si $\overrightarrow{\alpha} \neq 0$

4.5. Aproximación discreta por mínimos cuadrados

En determinadas circunstancias, el interés por la proximidad de dos funciones puede reducirse a observarla en un conjunto relevante dee puntos y no en todo el intervalo. Esta proximidad recortada puede medirse utilizando seminormas. Este término se refiere a una aplicación que cumple la segunda y tercera condición de norma pero no necesariamente la primera.

Se considera un conjunto finito de puntos (nodos)

$$x_0 < x_1 < \cdots < x_{n-1} < x_n$$

en el intervalo de interés. Con ayuda de estos puntos y las normas usadas en las secciones anteriores, se pueden definir seminormas para medir la proximidad discreta entre dos funciones. Así, relacionada con la norma $\| \|_2$, se puede definir la seminorma

$$|f|_2 = \left(\sum_{i=0}^n f(x_i)^2 \right)^{1/2}.$$

Más general es la seminorma p definida por

$$|f|_p = \left(\sum_{i=0}^n |f(x_i)|^p \right)^{1/p}.$$

Como caso límite, la seminorma ∞ discreta está definida por

$$|f|_\infty = \max_{0 \leqslant i \leqslant n} |f(x_i)|.$$

Se comprende bien que estas seminormas no verifican la primera propiedad que se exigía a las normas. Una función que se anule en todos los puntos del conjunto $\{x_i : i = 0, 1, \cdots, n\}$, no tiene forzosamente que ser nula en los demás puntos y sin embargo, el valor de su seminorma es 0.

El hecho de que los conjuntos de nivel de una seminorma no son necesariamente compactos, impide argumentas de mismo modo que en la sección anterior cuando se pretende establecer resultados de existencia de mejor aproximación discreta. Es decir, aunque se pruebe la existencia de una sucesión minimizante, acotada con una seminorma, no se puede concluir la existencia de una subsección convergente a una función que realice el mínimo.

Si una seminorma definida en el espacio de funciones continuas es una verdadera norma en el subespacio U en el que se busca la aproximación, aunque no sea en todo el espacio, los razonamientos del teorema 13 siguen siendo válidos y la existencia de mejor aproximación está garantizada. Si la seminorma proviene de un producto escalar degenerado (que puede que no cumpla la primera condición de producto escalar), la mejor aproximación es única, las ecuaciones normales son válidas y la matriz de Gram no es singular si y sólo si la seminorma consideraa es una norma en U.

La resolución de las ecuaciones normales se vuelve particularmente simple si la base es ortonormal, es decir, si $\langle \sigma_i, \sigma_j \rangle = 0$ para $i \neq j$ y $\langle \sigma_i, \sigma_i \rangle = 1$ para $i, j = 0, 1, \cdots, n$. Esta condición equivale a que la matriz de Gram asociada a esta base sea la matriz identidad. En este caso, la mejor aproximación a f es

$$g = \sum_{i=0}^n \langle f, \sigma_i \rangle \sigma_i.$$

En una situación general, se espera que la base no sea ortonormal. Sin embargo, conviene tener en cuenta que una base de un espacio vectorial con producto escalar, siempre puede ser transformada en otra ortonormal, mediante un proceso de ortonormalización como el de Gram-Schmidt.

Teorema 14 Si $\{p_n\}$ es una sucesión de polinomios ortogonales respecto al producto escalar

$$\langle f, g \rangle = \int_a^b f(x) g(x) \omega(x) \, \mathrm{dx}$$

donde ω es una función positiva en (a, b). Además, se supone que p_n es de grado n y tiene coeficiente principal

igual a 1 (coeficiente del monomio de mayor grado). En estas condiciones se tiene que

$$\begin{array}{rcl} p_0(x) & = & 1, \\ p_1(x) & = & x - a_1, \\ \cdots & & \cdots \\ p_n(x) & = & (x - a_n)p_{n-1} - b_n p_{n-2}(x) \\ \cdots & & \cdots \end{array}$$

donde

$$a_n = \frac{\langle xp_{n-1}, p_{n-1}\rangle}{\langle p_{n-1}, p_{n-1}\rangle}, \quad b_n = \frac{\langle xp_{n-1}, p_{n-2}\rangle}{\langle p_{n-2}, p_{n-2}\rangle}$$

Además, la sucesión definida por esta recurrencia de tres términos es la única sucesión ortogonal de grados distintos y coeficiente principal igual a 1.

Si el intervalo es simétrico respecto al origen y la función peso es par (es decir, verifica que $\omega(-x) = \omega(x)$ para todo $x \in [a, b]$), todos los coeficientes a_n en el teorema anterior, son nulos. En este caso, todos los polinomios p_n de orden par son funciones pares y los polinomios p_n de orden impar son funciones impares (es decir, verifican que $p(-x) = -p(x)$ para todo $x \in [a, b]$).

En ocasiones resulta interesante ponderar en la definición de producto escalar, la relevancia de una parte del intervalo $[a, b]$ frente a otras.

4.6. Polinomios de Chebyshev

Las funciones definidas por

$$T_n(x) = \cos(n \arccos(x)),$$

para $n = 0, 1, \cdots$, se llaman polinomios de Chebyshev en el intervalo $[-1, 1]$.

No es evidente al observar esta definición, que se trate de una sucesión de polinomios, ya que, en principio, intervienen funciones trigonométricas. Para $n = 0, 1$ directamente se comprueba que

$$T_0 = 1, \quad T_1 = x.$$

Para comprobar que la función T_n es un polinomio para cualquier n, se considera la identidad

$$\cos(a + b) + \cos(a - b) = 2 \cos a \cos b$$

que aplicada a $a = n \arccos x$ y $b = \arccos x$ lleva a la siguiente relación

$$\cos((n+1)\arccos x) + \cos((n-1)\arccos x) = 2x \cos(n \arccos x).$$

Si se usa la definición de polinomio de Chebyshev en esta igualdad, se obtiene que

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x).$$

Mediante esta recurrencia de tres términos y los polinomios de bajo grado T_0 y T_1, se pueden construir las expresiones polinomiales de los elementos de la sucesión T_n. Así pues, T_n es un polinomio de grado n.

Directamente de la definición primitiva de los polinomios de Chebyshev se deduce que las raíces de T_n son

$$x_i = \cos\left(\frac{2i - 1}{2n}\pi\right), \quad i = 1, \cdots, n$$

y las de su derivada

$$\overline{x}_i = \cos\left(\frac{i}{n}\pi\right), \quad i = 1, \cdots, n - 1.$$

En las raíces del polinomio T_n' se cumple que

$$T_n(\overline{x}_i) = \cos i\pi = (-1)^i$$

y puesto que

$$|T_n(x)| \leqslant 1, \quad \text{para todo } x \in [0, 1],$$

se deduce que el polinomio de Chebyshev T_n alcanza sus máximos y mínimos de modo alternativo en los $n-1$ puntos $\{\overline{x}_i : i = 1, \cdots, n - 1\}$.

Sin duda, una propiedad relevante de estos polinomios es que para n fijo, el conjunto $\{T_0, T_1, \cdots, T_n\}$ es una base ortogonal del espacio vectorial de polinomios de grado menor o igual que n con respecto al producto escalar ponderado

$$\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)\omega(x)\,\mathrm{d}x, \quad \text{con } \omega(x) = \frac{1}{\sqrt{1 - x^2}}.$$

En efecto, para probar que se anula el producto

$$\langle T_n, T_m\rangle = \int_{-1}^{1} \frac{\cos((n \arccos\,(x))\cos(m \arccos\,(x)))}{\sqrt{1 - x^2}}\mathrm{dx}$$

para $n \neq m$, se introduce el cambio de variabe $\theta = \arccos x$. Puesto que

$$\mathrm{d}\theta = -\frac{1}{\sqrt{1 - x^2}}\mathrm{dx}$$

se obtiene

$$\begin{array}{rcl} \langle T_n, T_m\rangle & = & \int_0^\pi \cos(n\theta)\cos(m\theta)\,\mathrm{d}\theta \\ & = & \frac{1}{2}\int_0^\pi (\cos((n + m)\theta)) + \cos((n - m)\theta)\,\mathrm{d}\theta \\ & = & \begin{cases} 0, & \text{si } m \neq n \\ \frac{\pi}{2}, & \text{si } m = n \neq 0 \\ \pi, & \text{si } m = n = 0 \end{cases} \end{array}$$

Así pues, la sucesión $\{\sqrt{\frac{1}{\pi}}, \sqrt{\frac{2}{\pi}}T_n : n > 0\}$ es una sucesión de polinomios ortonormales respecto al producto escalar de Chebyshev.

Sea a_n el coeficiente de la potencia x^n del polinomio T_n (en adelante, coeficiente principal). De la fórmula de recurrencia se deduce directamente que

$$a_{n+1} = 2a_n = 2^n.$$

Es frecuente estandarizar la sucesión de polinomios de Chebyshev dividiendo cada polinomio por su coficiente

principal. En este caso, se obtiene la sucesión de polinomios determinada por la recurrencia

$$\begin{aligned}\hat{T}_0 &= 1\\ \hat{T}_1 &= x\\ \hat{T}_2(x) &= x\hat{T}_1(x) - \tfrac{1}{2}\hat{T}_0(x)\\ \cdots &\quad \cdots\\ \hat{T}_{n+1}(x) &= x\hat{T}_n(x) - \tfrac{1}{4}\hat{T}_{n-1}(x)\end{aligned}$$

para $n \geqslant 2$.

Teorema 15 El polinomio estandarizado de Chebyshev $\hat{T}_n = \frac{1}{2^{n-1}}T_n$ verifica la siguiente desigualdad

$$\max_{x\in[-1,1]} |\hat{T}_n(x)| \leqslant \max_{x\in[-1,1]} |p(x)|,$$

para todo polinomio $p \in \mathcal{P}_n$ de coeficiente principal igual a 1.

4.7. Aproximación trigonométrica

La mayoría de los subespacios U usados en la aproximación han sido espacios de polinomios. Pero debe quedar claro que esto es solamente una conveniencia y no una exigencia. Es decir, los teoremas de existencia y unicidad de mejor aproximación son válidos si se utilizan otros subespacios funcionales. Obviamente, el requisito de que sean funciones de fácil evaluación es indispensable para que tenga sentido práctico la aproximación.

De modo general, se considera el subespacio U el espacio de polinomios trigonométricos generados por las funciones trigonométricas

$$\{1, \cos x, \cdots, \cos nx, \operatorname{sen} x, \cdots, \operatorname{sen} nx\}$$

para algún $n > 0$, en el intervalo $[-\pi, \pi]$ y el producto escalar definido por

$$\langle f, g\rangle = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)g(x)\,\mathrm{dx}.$$

Si se tienen en cuenta las siguientes igualdades trigonométricas para $n, m \geqslant 0$

$$\begin{aligned}\textstyle\int_{-\pi}^{\pi} \cos ix \operatorname{sen} jx\,\mathrm{dx} &= 0,\\ \textstyle\int_{-\pi}^{\pi} \cos ix \cos jx\,\mathrm{dx} &= \textstyle\int_{-\pi}^{\pi} \operatorname{sen} ix \operatorname{sen} jx\,\mathrm{dx}\\ &= \begin{cases} 0, & \text{si } i \neq j\\ \pi, & \text{si } i = j > 1\\ 2\pi, & \text{si } i = j = 0\end{cases}\end{aligned}$$

se puede comprobar que la mejor aproximación a una función f está dada por

$$g_n(x) = \frac{a_0}{2} + \sum_{i=1}^{n}(a_i \cos ix + b_i \operatorname{sen} ix)$$

donde

$$\begin{aligned}a_i &= \tfrac{1}{\pi}\textstyle\int_{-\pi}^{\pi} f(x)\cos ix\,\mathrm{dx}\\ b_i &= \tfrac{1}{\pi}\textstyle\int_{-\pi}^{\pi} f(x)\operatorname{sen} ix\,\mathrm{dx}\end{aligned}$$

Esta funcion se conoce como serie de Fourier finita de la función f.

Se puede observar que la precisión de la aproximación trigonométrica se deteriora bruscamente en las proximidades de los extremos del intervalo. De hecho, este deterioro se incrementa cuando se aumenta el grado de la aproximación, lo que hace aproximar mejor la función lejos de los extremos pero empeora en sus proximidades. Esta inestabilidad se conoce como fenómeno de Gibbs.

De modo similar al caso de polinomios algebraicos, se puede desarrollar también una versión discreta de la aproximación de mínimos cuadrados trigonométrica. Para ello, se considera una partición del intervalo $[-\pi, \pi]$ en subintervalos igualmente espaciados, definida por los nodos

$$x_k = \left(\frac{k}{m} - 1\right)\pi$$

para $k = 0, 1, \cdots, 2m$. Asociado a esta partición se considera el producto discreto

$$\langle f, g\rangle = \sum_{k=0}^{2m-1} f(x_k)g(x_k).$$

Teorema 16 Para $n < m$, los polinomios trigonométricos

$$\{\cos kx, \operatorname{sen} kx : k = 0, 1, \cdots, n\}$$

son ortogonales respecto al producto escalar discreto. La aproximación trigonométrica discreta de la función f está dada por

$$g(x) = \frac{a_0}{2} + \sum_{j=1}^{n}(a_j \cos jx + b_j \operatorname{sen} jx)$$

donde

$$\begin{aligned}a_j &= \tfrac{1}{m}\sum_{k=0}^{2m-1} f(x_k)\cos jx_k\\ b_j &= \tfrac{1}{m}\sum_{k=0}^{2m-1} f(x_k)\operatorname{sen} jx_k\end{aligned}$$

4.8. Aproximación uniforme

La mejor aproximación uniforme a una función es en cierto sentido más severa que la aproximación por mínimos cuadrados ya que esta proximidad debe mantenerse en todos los puntos del intervalo y una alteración de la función en un entorno de un punto alejaría notablemente de su mejor aproximación. Antes, de entrar en el análisis de la mejor aproximación uniforme en un espacio de polinomios de grado limitado, es importante destacar que para toda función continua existe un polinomio tan próximo a la función como se desee.

Teorema 17 (de Weierstrass) Si f es una función continua en un intervalo cerrado y acotado I, para todo número real positivo ϵ existe un polinomio p tal que

$\|f - p\|_\infty < \epsilon$.

La sucesión de polinomios que se construye a continuación, conocida como sucesión de polinomios de Bernstein, tiene por límite uniforme la función f. Por razones de sencillez, se supone que $I = [0, 1]$ pero un sencillo cambio de variable podría extender el razonamiento a cualquier intervalo cerrado y acotado.

Para $n \geqslant 0$, se define el polinomio de Bernstein de grado n como

$$B_{n,f}(x) = \sum_{i=0}^{n} \binom{i}{n} \binom{n}{i} x^i (1-x)^{n-i}.$$

Puesto que

$$\sum_{i=0}^{n} \binom{n}{i} x^i (1-x)^{n-i} = (x + 1 - x)^n = 1$$

$B_{n,f}$ puede interpretarse como un valor promediado de los valores de f en los puntos $\frac{i}{n}$.

Se establecen las dos siguientes propiedades de los polinomios de Bernstein:

1. El operador B_n es lineal y positivo en el espacio de las funciones continuas:
   - $B_{n,\alpha f+\beta g} = \alpha B_{n,f} + \beta B_{n,g}$
   - Si $f(x) \geqslant 0$ para todo $x \in I$ se cumple que $B_{n,f}(x) \geqslant 0$ para todo $x \in I$
2. La sucesión $\{B_{n,f}\}$ converge uniformemente a f si f es un polinomio de grado menor o igual que 2. De la relación
   $$B_{n,x^2} = \frac{n-1}{n} x^2 + \frac{1}{n} x$$
   se deduce que el límite uniforme de B_{n,x^2} es x^2. De la linealidad del operador B_n se desprende que este resultado se conserva para todo polinomio de grado menor o igual que 2.

Si se examina en detalle la última sucesión de polinomios de Bernstein dada se observa que para $f = x^2$

$$\|B_{n,f} - f\|_\infty = \max_{x \in [0,1]} \frac{|x^2 - x|}{n} = \frac{1}{4n}.$$

Es decir, para obtener una aproximación con error 10^{-4} se necesitaría el polinomio de grado $n = 2500$. A pesar de que f ya es polinomio de grado 2, es necesario ir a un polinomio de Bernstein de grado muy alto para aproximarse de modo preciso. Esto implica que, pese a que los polinomios de Bernstein dan aproximaciones explícitas para aproximar funciones continuas, no suministran una técnica muy eficiente debido a su lenta convergencia y su interés más conceptual que práctico.

Teorema 18 (Criterio de Kolmogorov) Sea U un subespacio vectorial de dimensión finita del espacio de las funciones continuas en un intervalo $[a, b]$ y g un elemento de U. Entonces g es una mejor aproximación de f en U si y sólo si ningún elemento de U tiene el mismo signo que $f - g$ en todos los puntos del conjunto

$$A_{f-g} = \{x \in [a, b] : |f(x) - g(x)| = \|f - g\|_\infty\}$$

Teorema 19 (de la alternancia de Chebyshev) Un polinomio g es la mejor aproximación de una función f continua en $[a, b]$ en $\mathcal{P}_n$ si y sólo si $|f - g|$ alcanza su máximo en $n + 2$ puntos distintos con signo alternante en $f - g$.

Teorema 20 Existe una única mejor aproximación uniforme de una función continua en $I = [a, b]$ en $\mathcal{P}_n$.

Una segunda consecuencia del teorema de alternancia de Chebyshev es el procedimiento para construir la mejor aproximación uniforme que se conoce como método de intercambio de Remez y que se describe a continuación:

1. Seleccionar $n + 2$ puntos $x_0 < x_1 < \cdots < x_n < x_{n+1}$ en $I = [a, b]$ arbitrariamente.
2. Calcular el polinomio $p_n \in \mathcal{P}_n$ y el parámetro d tales que
   $$f(x_i) - p_n(x_i) = (-1)^{i+1} d$$
   para $i = 0, 1, \cdots, n, n + 1$. Si se escoge una base de $\mathcal{P}_n$, las $n + 1$ componentes del polinomio en esa base y el valor de d son las $n + 2$ incógnitas de un sistema lineal.
3. Determinar los puntos en los que la función $|f - p_n|$ alcanza un máximo local y reemplazar con ellos, todos o parte de los x_i utilizado.
4. Volver a la etapa 2.

Se referencia el punto 2 para justificar las siguientes afirmaciones:

- La sucesión formada por los polinomios p_n obtenidos por el algoritmo de Remez es convergente a la mejor aproximación uniforme. La convergencia es cuadrática si f es diferenciable.
- Una buena elección de los puntos de partida es la de las raíces del polinomio de Chebyshev de grado $n+2$.

5. Interpolación de funciones

5.1. Introducción

Se considera una mejor aproximación p_n de una función f en $U = \mathcal{P}_n$, en el sentido de mínimos cuadrados discretos en el siguiente conjunto de nodos

$$x_0 < x_1 < \cdots < x_{m-1} < x_m.$$

Es conveniente separar tres situaciones distintas:

1. El número de nodos menos 1 es menor que el máximo grado de los polinomios de U, es decir, $m < n$. En este caso, la seminorma $|\ |_2$ inducida por el producto discreto, no es una norma ya que el polinomio $q \in \mathcal{P}_n$ definido por
   $$q(x) = (x - x_0)(x - x_1) \cdots (x - x_m)$$
   verifica que $|q|_2 = 0$ y no es nulo. Por lo tanto, la mejor aproximación p_n de f en $\mathcal{P}_n$ no es única.

2. El número de nodos menos 1 es igual al máximo grado de los polinomios de U, es decir, $m = n$. La seminorma inducida por el producto es una norma y la mejor aproximación es única. En este caso, se cumple que la aproximación es perfecta

$$|f - p_n|_2 = 0.$$

3. El número de nodos menos 1 es mayor al máximo grado de los polinomios de U, es decir, $m > n$. La seminorma inducida por el producto es una norma y la mejor aproximación es única pero $|f - p_n|_2$ no es necesariamente 0.

Esta sección está dedicada al análisis de la situación $m = n$. En este caso, el polinomio p_n de grado menor o igual que n, que es la mejor aproximación en el sentido de mínimos cuadrados discreta, alcanza la distancia 0, ya que

$$p_n(x_i) = f(x_i)$$

para $i = 0, 1, \cdots, n$ y se conoce como polinomio de interpolación de la función f en los nodos $\{x_i : i = 0, 1, \cdots, n\}$. El hecho de que el número de nodos de interpolación y la dimensión del espacio de polinomios coincidan es esencial para que esto ocurra.

Se puede considerar el polinomio de interpolación como una combinación lineal de los valores de f en los nodos

$$p_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i)$$

en cada punto x. Los coeficientes de la combinación lineal dependerán del punto considerado x. Es frecuente referirse a ellos como las funciones de forma de la aproximación. La propiedad más deseable de las funciones de forma es que no dependan de la función f.

En la determinación de las funciones de forma, se puede utilizar el hecho de que la aproximación es exacta cuando la propia función a interpolar es un polinomio de grado menor o igual que n.

5.2. Interpolación de Lagrange

Teorema 21 Existe un único conjunto $L = \{l_0, l_1, \cdots, l_n\}$ de funciones de forma, correspondientes al conjunto de $n + 1$ nodos $\{x_i : i = 0, 1, \cdots, n\}$, definidas por

$$l_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}$$

Además, L constituye una base de $\mathcal{P}_n$ que se conoce como base de Lagrange.

El conocimiento explícito de la base de Lagrange permite, dada una función cualquiera f, construir un polinomio de grado menor o igual que n

$$p(x) = \sum_{i=0}^{n} l_i(x) f(x_i)$$

que interpole a f en $n+1$ puntos $x_0, x_1, \cdots, x_n$. Conviene recvordad que la solución del problema de interpolación es la solución del problema de mejor aproximación de una función en el espacio de los polinomios de grado n respecto a la seminorma definida por el producto discreto

$$\langle f, g \rangle = \sum_{i=0}^{n} f(x_i) g(x_i).$$

La base de Lagrange es ortonormal respecto a $\langle\ ,\ \rangle$. En efecto, se cumple que

$$\langle l_i, l_j \rangle = \sum_{k=0}^{n} \delta_{ik} \delta_{kj} = \delta_{ij}$$

donde δ_{ij} representa la delta de Kronecker (1 si $i = j$ y 0 si $i \neq j$).

5.3. Método de Newton

La base $\{1, x, x^2, \cdots, x^n\}$ de $\mathcal{P}_n$ no es adecuada para el análisis y cálculo de los polinomios de interpolación porque no implica cómo los nodos de la interpolación están distribuidos en el intervalo. En la base de Lagrange, todos los polinomios básicos son del mismo grado n y el coste computacional para evaluarla en un conjunto de puntos es más elevado que en el caso de los monomios. Otro modo alternativo de construir el polinomio de Lagrange es el que se basa en el uso de la base

$$\{1, x - x_0, (x - x_0)(x - x_1), \cdots, (x - x_0)(x - x_1) \cdots (x - x_{n-1})\}$$

de $\mathcal{P}_n$ formada por los polinomios

$$\omega_i(x) = (x - x_0)(x - x_1) \cdots (x - x_{i-1})$$

para $i = 1, \cdots, n$ y el polinomio constante $\omega_0(x) = 1$. Esta base incoprpora la información de cómo se distribuyen los nodos de la interpolación (salvo el último) y su grado ascendente. Se conoce en este contexto, como la base de Newton.

Es importante destacar que un polinomio

$$p(x) = \sum_{i=0}^{n} a_i \omega_i(x),$$

expresado en la base de Newton, admite la siguiente representación anidada

$$\begin{aligned} p(x) = \quad & a_0 + (x - x_0)(a_1 + (x - x_1)(a_2 + \cdots \\ & \cdots + (x - x_{n-2})(a_{n-1} + a_n(x - x_{n-1}) \cdots))). \end{aligned}$$

Así pues, para calcular el valor de p en un punto x, se podría utilizar la regla de Ruffini adaptada a esta base.

Para estudiar cómo se puede expresar e lpolinomio de interpolación en términos de esta base, se hace uso de la propiedad de exactitud

$$p(x) = \sum_{i=0}^{n} l_i(x) p(x_i)$$

del polinomio de interpolación sobre los polinomios p de grado menor o igual que n. En particular, si se utiliza esta condición sobre los elementos básicos ω_i para $i = 0, 1, \cdots, n$ y se tiene en cuenta que $\omega_i(x_j) = 0$ si $j < i$, se obtiene el sistema lineal

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & \omega_1(x_1) & \cdots & \omega_1(x_n) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \omega_n(x_n) \end{pmatrix} \begin{pmatrix} l_0 \\ l_1 \\ \vdots \\ l_n \end{pmatrix} = \begin{pmatrix} \omega_0 \\ \omega_1 \\ \vdots \\ \omega_n \end{pmatrix}$$

que relaciona la base de Lagrange con la de Newton. La matriz triangular superior D asociada a este sistema es la matriz de cambio de base que permite relacionar las componentes de cualquier polinomio $p \in \mathcal{P}_n$ en ambas bases. En particular, las componentes del polinomio de interpolación en ambas bases están relacionadas por el sistema lineal

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & \omega_1(x_1) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_1(x_n) & \cdots & \omega_n(x_n) \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

donde a_i representan las componentes del polinomio de interpolación en la base de Newton. Las componentes a_i se llaman diferencias divididas (ó cociente) de la función f en los nodos $\{x_i : i = 0, 1, \cdots, n\}$ y la expresión del polinomio de interpolación se conoce como polinomio de Newton. Es habitual representar la componente a_i del polinomio de interpolación de Newton p_n por $f[x_0, x_1, \cdots, x_i]$, de modo que

$$p_n(x) = f[x_0] + f[x_0, x_1]\omega_1(x) + \cdots + f[x_0, x_1, \cdots, x_n]\omega_n(x).$$

En el caso general, se puede probar que las diferencias divididas de orden j están dadas por

$$f[x_0, x_1, \cdots, x_j] = \frac{f[x_1, x_2, \cdots, x_j] - f[x_0, x_1, \cdots, x_{j-1}]}{x_j - x_0}$$

para $j = 1, 2, \cdots, n$.

Tradicionalmente, se calculan las diferencias en tablas organizadas como sigue:

$$\begin{array}{llll} x_0 & f(x_0) & & \\ & & \frac{f(x_1)-f(x_0)}{x_1-x_0} & \\ x_1 & f(x_1) & & \frac{\frac{f(x_2)-f(x_1))}{x_2-x_1} - \frac{f(x_1)-f(x_0)}{x_1-x_0}}{x_2-x_0} \\ & & \frac{f(x_2)-f(x_1)}{x_2-x_1} & \\ x_2 & f(x_2) & & \vdots \\ & & \vdots & \\ \vdots & & & \vdots \\ & & \vdots & \\ x_n & f(x_n) & & \end{array}$$

Se examina ahora el caso en el que los nodos están uniformemente distribuido. Si $h = x_{i+1} - x_i$ para $i = 0, 1, \cdots, n-1$ entonces

$$\omega_i(x_j) = \begin{cases} \frac{j!}{(j-i)!}h^i, & \text{si } i \leqslant j \\ 0, & \text{si } i > j \end{cases} = i!h^i \begin{cases} \binom{j}{i}, & \text{si } i \leqslant j \\ 0, & \text{si } i > j \end{cases}$$

para $i, j = 0, 1, \cdots, n$. En este caso el sistema que permite calcular las diferencias divididas es el siguiente

$$\begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 1 & 1 & 0 & \cdots & 0 \\ 1 & 2 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & n & \frac{n(n-1)}{2} & \cdots & 1 \end{pmatrix} \begin{pmatrix} \Delta^0 f(x_0) \\ \Delta^1 f(x_0) \\ \vdots \\ \Delta^n f(x_0) \end{pmatrix} = \begin{pmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}$$

donde $\Delta^i f(x_0) = f[x_0, x_1, \cdots, x_i]h^i i!$ se llama diferencia finita de orden i. Si se resuelve este sistema lineal se obtiene

$$\Delta^i f(x_0) = \sum_{k=0}^{i} \binom{i}{k} (-1)^{i+k} f(x_k)$$

lo que permite interpretar la diferencia finita como la aplicación i-veces reiterada del operador $\Delta f(x_j) = f(x_{j+1}) - f(x_j)$.

Por otra parte, el polinomio de interpolación se puede expresar en términos de las diferencias finitas como

$$p_n(x) = \sum_{i=0}^{n} \frac{\Delta^i f}{i!h^i} \omega_i(x).$$

5.4. Error en la interpolación de Lagrange

Teorema 22 El error que se comete al aproximar una función f de clase $n+1$ en el intervalo $[a, b]$ por su polinomio de interpolación de Lagrange de grado menos o igual que n en los nodos

$$a = x_0 < x_1 < \cdots < x_n = b$$

está dado para cualquier valor de $x \in [a, b]$, por la siguiente fórmula

$$E(x) = f(x) - p_n(x) = \frac{(x - x_0) \cdots (x - x_n)}{(n+1)!} f^{(n+1)}(\xi)$$

para algún número $\xi \in (a, b)$ (que depende de x).

Mediante un cambio de variable lineal se transforma el problema de interpolación en el intervalo $[a, b]$ en otro equivalente en el intervalo $[-1, 1]$. En este intervalo, de acuerdo con el teorema 15 la elección de nodo óptimos corresponde a las raíces del polinomio de Chebyshev de grado $n + 1$ en cuyo caso se tiene que

$$\max_{x \in [-1,1]} |(x - x_0)(x - x_1) \cdots (x - x_n)| = \frac{1}{2^n}$$

5.5. Algoritmos de Aitken y Neville

Para comprender la base teórica en la que estos procedimientos se sustentan, se analiza la siguiente situación: Se conoce el valor de una función en $n + 1$ puntos $X = \{x_0, x_1, \cdots, x_n\}$ distintos. Se supone que se ha construido el polinomio de interpolación p_{n-1} de grado $n - 1$ correspondiente en los puntos $\{x_0, x_1, \cdots, x_{n-1}\}$, directamente se comprueba que

$$p_n(x) = p_{n-1}(x) + (f_n - p_{n-1}(x_n))l_n(x)$$

es el polinomio de interpolación correspondiente a todos los datos (l_n es la función básica de Lagrange asociada al nodo n) ya que es único y $p_n(x_i) = f_i$ para todo $i = 0, 1, \cdots, n$.

Esta consideración es aún válida si el punto separado del conjunto X es arbitrario y no necesariamente el último. Por esta razón, se puede diseñar el siguiente procedimiento: Del conjunto X se extrae un punto x_j y se construye el polinomio de interpolación p^j_{n-1} con los puntos restantes. A continuación, se repone el punto x_j a X y se repite el procedimiento para otro punto distinto x_k, obteniéndose un segundo polinomio de interpolación p^k_{n-1} de grado menor o igual que $n-1$. De acuerdo con la relación anterior se tiene que

$$p_n(x) = p^j_{n-1}(x) + (f_j - p^j_{n-1}(x_j))l_j(x)$$
$$p_n(x) = p^k_{n-1}(x) + (f_k - p^k_{n-1}(x_k))l_k(x)$$

Si se multiplican, la primera igualdad por $x - x_j$, y la segunda por $x - x_k$, se deduce que

$$(x_k - x_j)p_n(x) = (x - x_j)p^j_{n-1}(x) - (x - x_k)p^k_{n-1}(x) + q(x)$$

donde q es el polinomio definido por

$$q(x) = (x - x_j)(f_j - p^j_{n-1}(x_j))l_j(x) - (x - x_k)(f_k - p^k_{n-1}(x_k))l_k(x).$$

El polinomio q es de grado n ya que es la diferencia de dos polinomios de grado n. Además se anula en todos los puntos de X. Consecuentemente, el polinomio q es el nulo y por ello se obtiene la fórmula

$$p_n(x) = \frac{(x - x_j)p^j_{n-1}(x) - (x - x_k)p^k_{n-1}(x)}{x_k - x_j}.$$

Esta fórmula permite justificar los algoritmos de Aitken y Neville que se usan para evaluar el polinomio de interpolación en un punto sin necesidad de calcular sus coeficientes o evaluar las funciones de forma de la interpolación. Esta fórmula permite calcular fácilmente el polinomio de interpolación $p(x; x_0, x_1, x_2, x_3)$ construido en los puntos $\{x_0, x_1, x_2, x_3\}$ usando dos polinomios construidos usando tres puntos. Se pueden organizar los cálculos de modo recurrente en algun de las disposiciones siguientes:

Algoritmo de Aitken

x_0	$x - x_0$	f_0			
x_1	$x - x_1$	f_1	$p(x; x_0, x_1)$		
x_2	$x - x_2$	f_2	$p(x; x_0, x_2)$	$p(x; x_0, x_1, x_2)$	
x_3	$x - x_3$	f_3	$p(x; x_0, x_3)$	$p(x; x_0, x_1, x_3)$	$p(x; x_0, x_1, x_2, x_3)$

En esta tabla, el polinomio $p(x; x_0, x_1, x_2, x_3)$ se calcula del modo siguiente

$$p(x; x_0, x_1, x_2, x_3) =$$

$$\frac{(x - x_2)p(x; x_0, x_1, x_3) - (x - x_3)p(x; x_0, x_1, x_2)}{x_3 - x_2}$$

Algoritmo de Neville

x_0	$x - x_0$	f_0			
x_1	$x - x_1$	f_1	$p(x; x_0, x_1)$		
x_2	$x - x_2$	f_2	$p(x; x_1, x_2)$	$p(x; x_0, x_1, x_2)$	
x_3	$x - x_3$	f_3	$p(x; x_2, x_3)$	$p(x; x_1, x_2, x_3)$	$p(x; x_0, x_1, x_2, x_3)$

5.6. Interpolación compuesta

El control del error de interpolación está en la $(n+1)$ derivada de la función a interpolar. No obstante, funciones con expresiones analíticas sencillas pueden representar fuertes oscilaciones en las derivadas sucesivas.

Hay situaciones que evidencian que muchas veces es preferible, a fin de agilizar el cálculo o conseguir una mayor estabilidad en la aproximación, agrupar los datos en pequeños grupos e interpolar independientemente en cada uno de ellos. En la situación extrema, con cada dato se construye un polinomio constante, que se considera como aproximación de la función hasta estar próximos al siguiente dato, en donde se toma como constante el nuevo dato. Se construye de este modo un polinomio constante a trozos de modo similar a como se construye la función por redondeo o la función parte entera (interpolación por el punto más próximo).

Un procedimiento más lento pero más preciso, consiste en considerar los datos a pares y construir el polinomio de primer grado que interpola a ambos datos. Se construye de este modo un polinomio lineal a trozos.

5.7. Interpolación de Hermite

De un modo general se puede plantear un problema de interpolación de Hermite del modo siguiente: Dados los valores de una función y de su derivada en los nodos $a = x_0 < \cdots < x_n = b$ hallar un polinomio p de grado $2n+1$ tal que

$$p(x_i = f_i), \quad p'(x_i) = f'_i$$

para $i = 0, 1, \cdots, n$. Se puede interpretar el polinomio de interpolación de Hermite como la mejor aproximación con respecto a la norma inducida por el producto escalar

$$\langle f, g \rangle = \int_a^b (f(x)g(x) + f'(x)g'(x))\mathrm{dx}$$

correspondiente al caso en el que la distancia es nula.

A fin de obtener un resultado de existencia y unidad de solución a este problema, así como para desarrollar métodos eficientes de construcción, se consideran los $2n+1$ polinomios de Newton

$$\omega_0 = 1$$
$$\omega_1 = x - x_0$$
$$\omega_2 = (x - x_0)^2$$
$$\omega_3 = (x - x_0)^2(x - x_1)$$
$$\omega_4 = (x - x_0)^2(x - x_1)^2$$
$$\vdots$$
$$\omega_{2n} = (x - x_0)^2 \cdots (x - x_{n-1})^2$$
$$\omega_{2n+1} = (x - x_0)^2 \cdots (x - x_{n-1})^2(x - x_n)$$

que generan el espacio de los polinomios de grado menor o igual que $2n+1$. Se busca un polinomio de interpolación que en cada punto dependa linealmente de los datos

$$p(x) = \sum_{i=0}^{n} (l_i(x)f_i + h_i(x)f'_i).$$

Se impone la condición de que esta fórmula de interpolación sea exacta en el espacio de los polinomios de grado

menor o igual que $2n+1$, es decir, que el polinomio de interpolación de Hermite de un polinomio sea él mismo. Si se impone la exactitud en la base de Newton se obtiene el siguiente sistema lineal

$$\sum_{i=0}^{n}(l_{2i}(x)\omega_j(x_i)+l_{2i+1}(x)\omega_j'(x_i))=\omega_j(x)$$

para cada $x\in[a,b]$ y para $j=0,\cdots,2n+1$. La matriz de coeficientes D de este sistema resulta ser

$$D=\begin{pmatrix} \omega_0(x_0) & \omega'_{x_0} & \cdots & \omega_0(x_n) & \omega_0'(x_n)\\ \omega_1(x_0) & \omega_1'(x_0) & \cdots & \omega_1(x_n) & \omega_1'(x_n)\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \omega_{2n}(x_0) & \omega_{2n}'(x_0) & \cdots & \omega_{2n}(x_n) & \omega_{2n}'(x_n)\\ \omega_{2n+1}(x_0) & \omega_{2n+1}'(x_0) & \cdots & \omega_{2n+1}(x_n) & \omega_{2n+1}'(x_n)\end{pmatrix}$$

el vector de incógnitas, $l=(l_0,l_1,\cdots,l_{2n+1})^t$ y el de términos independientes, $\omega=(\omega_0,\omega_1,\cdots,\omega_{2n+1})^t$.

Si se tiene en cuenta que

$$\begin{array}{lcl}\omega_{2j}(x_i) & = & \omega_j(x_i)^2=0\\ \omega_{2j}'(x_i) & = & 2\omega_j'(x_i)\omega_j(x_i)=0\\ \omega_{2j+1}'(x_i) & = & 2\omega_j'(x_i)\omega_j(x_i)(x_i-x_j)+\omega_j(x_i)^2=0\end{array}$$

para $j>i$

$$\omega_{2j+1}(x_i) \quad = \quad \omega_{2j}(x_i)(x_i-x_j)=0$$

para $j\geqslant i$, se comprueba que la matriz D es triangular superior. Además, puesto que

$$\omega_i(x_i)>0,\quad \omega_i'(x_i)>0$$

para todo $i=0,1,\cdots,n$ el determinante de D es positivo.

La resolución de este sistema proporciona las funciones de forma o funciones básicas de Hermite. De modo similar a como se procedió en el caso de la interpolación de Lagrange se pueden calcular las componentes del polinomio de interpolación en la base $\{\omega_i : i=0,\cdots,2n+1\}$ resolviendo el sistema

$$D^t\begin{pmatrix} f[x_0]\\ f[x_0,x_0]\\ f[x_0,x_0,x_1]\\ \vdots\\ f[x_0,x_0,\cdots,x_n,x_n]\end{pmatrix}=\begin{pmatrix} f(x_0)\\ f'(x_0)\\ f(x_1)\\ \vdots\\ f'(x_n)\end{pmatrix}$$

Los cálculos pueden organizarse en tablas como construidas como en el caso de la interpolación de Lagrange, excepto que en las dos primeras columnas se repiten datos y en la tercera se toman las derivadas cuando se anula el denominador.

$$\begin{array}{cccc} x_0 & f(x_0) & & \\ & & f[x_0,x_0]=f'(x_0) & \\ x_0 & f(x_0) & & f[x_0,x_0,x_1]=\frac{f[x_0,x_1]-f[x_0,x_0]}{x_1-x_0}\\ & & f[x_0,x_1]=\frac{f(x_1)-f(x_0)}{x_1-x_0} & \\ x_1 & f(x_1) & & f[x_0,x_1,x_1]=\frac{f[x_1,x_1]-f[x_0,x_1]}{x_1-x_0}\\ & & f[x_1,x_1]=f'(x_1) & \\ x_1 & f(x_1) & & \\ \vdots & \vdots & \vdots & \vdots\\ x_n & f(x_n) & & \end{array}$$

5.8. Interpolación por esplines cúbicos

Si no se dispone de información sobre las derivadas de la función pero se necesita regularidad en la interpolación compuesta, se pueden utilizar las funciones esplines que son polinómicas a trozos y conservan algunos ordenes de regularidad en los nodos de interpolación. La idea básica consiste en la unión de trozos de los polinomios construidos en subintervalos contiguos, de modo que las derivadas laterales coincidan en el nodo común.

Se considera el siguiente conjunto de nodos de interpolación

$$a=x_0<x_1<\cdots<x_n=b.$$

Una función s de clase $C^2([a,b])$ es un esplín cúbico en $[a,b]$ si s es un polinomio de grado menor o igual que 3 en cada intervalo $[x_i,x_{i+1}]$. Se dice que es un esplín cúbico de interpolación a los datos $\{(x_i,y_i) : i=0,1,\cdots,n\}$ si $s(x_i)=y_i$ para $i=0,1,\cdots,n$.

Para poder utilizar una expresión analítica, se consideran las restricciones s_i del esplín s a cada intervalo $[x_i,x_{i+1}]$ para $i=0,1,\cdots,n-1$. De este modo se puede representar el esplín mediante los n polinomios de tercer grado

$$s_i(x)=a_i(x-x_i)^3+b_i(x-x_i)^2+c_i(x-x_i)+d_i$$

para cada $i=0,\cdots,n-1$. La dificultad reside en que los coeficientes a_i,b_i,c_i y d_i no pueden ser determinados separadamente para cada $i=0,\cdots,n-1$, ya que en cada intervalo solamente se dispone de las dos ecuaciones

$$s_i(x_i)=y_i,\quad s_i(x_{i+1})=y_{i+1}$$

que proporciona la condición de interpolación. El hecho de que el esplín tenga derivada primera y segunda continua en los nodos de interpolación propociona dos nuevas ecuaciones

$$\begin{array}{rcll} s_i'(x_i) & = & s_{i-1}'(x_i) & \text{si } i>0\\ s_i''(x_i) & = & s_{i-1}''(x_i) & \text{si } i>0\\ s_i'(x_{i+1}) & = & s_{i+1}'(x_{i+1}) & \text{si } i<n-1\\ s_i''(x_{i+1}) & = & s''_{x_{i+1}} & \text{si } i<n-1\end{array}$$

en cada uno de los extremos de cada intervalo. Pero, queda claro que no pueden resolverse de modo independiente en cada intervalo ya que involucran a los polinomios s_{i-1} y s_{i+1}. Esto pone en claro la naturaleza global del esplín.

Para determinar esos coeficientes se introducen las siguientes variables auxiliares

$$z_i=s''(x_i)$$

para $i=0,\cdots,n$. Aunque el objetivo final es el cálculo de los coeficientes a_i,b_i,c_i y d_i para cada uno de los subintervalos, como cálculo intermedio se determinarán los valores de z_i.

De la propia definición de las variables z_i se desprende que si se utiliza la expresión del esplín en cada subintervalo para evaluar la derivada segunda en x_i se obtienen

n relaciones entre las variables z_i y los coeficientes de los polinomios. Puesto que

$$\begin{aligned} s_i'(x) &= 3a_i(x-x_i)^2 + 2b_i(x-x_i) + c_i \\ s_i''(x) &= 6a_i(x-x_i) + 2b_i \end{aligned}$$

de la definición de las variables z_i, se obtiene que

$$b_i = \frac{z_i}{2} \quad \text{para } i = 0, 1, \cdots, n-1.$$

Por otra parte, si se emplean para evaluar s'' en x_i, las expresiones que le corresponden en los intervalos $[x_{i-1}, x_i]$ y $[x_i, x_{i+1}]$ para $i = 1, \cdots, n-1$, el resultado debe ser el mismo. Esto conduce a las siguientes relaciones

$$6a_{i-1}(x_i - x_{i-1}) + 2b_{i-1} = 2b_i$$

para $i = 1, \cdots, n-1$. De estas relaciones se deduce que

$$a_i = \frac{z_{i+1} - z_i}{6h_i}$$

donde $h_i = x_{i+1} - x_i$ para $i = 0, \cdots, n-1$.

Si se imponen las condiciones de interpolación en los extremos del intervalo $[x_i, x_{i+1}]$, se obtienen las ecuaciones siguientes

$$\begin{aligned} d_i &= y_i \\ a_i h_i^3 + b_i h_i^2 + c_i h_i + d_i &= y_{i+1} \end{aligned}$$

para $i = 0, \cdots, n-1$. Si se usan las expresiones de a_i, b_i y d_i en la segunda de estas relaciones se obtiene que

$$c_i = \frac{y_{i+1} - y_i}{h_i} - \frac{2z_i + z_{i+1}}{6} h_i$$

para $i = 0, \cdots, n-2$.

De la continuidad de las derivadas en el punto x_{i+1} se deduce que

$$3a_i h_i^2 + 2b_i h_i + c_i = c_{i+1}$$

para $i = 0, 1, \cdots, n-2$. Si se insertan en esta relación las expresiones que dan a_i, b_i y c_i en términos de x_i, y_i y h_i se obtiene

$$\frac{h_i}{6} z_i + \frac{h_i + h_{i+1}}{3} z_{i+1} + \frac{h_{i+1}}{6} z_{i+2} = \frac{y_{i+2} - y_{i+1}}{h_{i+1}} - \frac{y_{i+1} - y_i}{h_i}$$

para $i = 0, 1, \cdots, n-2$. Si se define $r_i = \frac{h_i}{h_i + h_{i+1}}$, la anterior relación se convierte en

$$r_i z_i + 2z_{i+1} + (1 - r_i) z_{i+2} = 6f[x_i, x_{i+1}, x_{i+2}]$$

para $i = 0, 1, \cdots, n-2$. Los segundos miembros de estas relaciones son diferencias divididas que pueden calcularse sin dificultad. De hecho, estas relaciones podrían ser interpetadas como un sistema lineal de $n-1$ ecuaciones de incógnitas z_i para $i = 0, 1, \cdots, n$

$$\begin{pmatrix} r_0 & 2 & 1-r_0 & \cdots & 0 & 0 & 0 \\ 0 & r_1 & 2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 2 & 1-r_{n-3} & 0 \\ 0 & 0 & 0 & \cdots & r_{n-2} & 2 & 1-r_{n-2} \end{pmatrix} \begin{pmatrix} z_0 \\ z_1 \\ \vdots \\ z_n \end{pmatrix}$$

$$= 6 \begin{pmatrix} f[x_0, x_1, x_2] \\ f[x_1, x_2, x_3] \\ \vdots \\ f[x_{n-2}, x_{n-1}, x_n]. \end{pmatrix}$$

En cualquier caso, se trata de un sistema de ecuaciones indeterminado ya que el número de incógnitas excede en 2 al de ecuaciones. Consecuentemente, el problema, tal y como está planteado, admite una infinidad de soluciones.

Existen varias posibilidades para completar este sistema de ecuaciones. Las más habituales son:

- Esplín natural

$$s_0''(x_0) = s_{n-1}''(x_n) = 0$$

- Esplín con pendiente fijada en los extremos (End Slope Spline)

$$s_0'(x_0) = y_0', \quad s_{n-1}'(x_n) ? y_n'$$

- Esplín periódico

$$s_0'(x_0) = s_{n-1}'(x_n), \quad s_0''(x_0) = s_{n-1}''(x_n)$$

- Esplín sin nudo (Not-a-Knot Spline)

$$s_0'''(x_1) = s_1'''(x_1), \quad s_{n-2}'''(x_{n-1}) = s_{n-1}'''(x_{n-1})$$

6. Derivación e integración numérica

6.1. Introducción

Las ideas que soportan las técnicas de derivación o cuadratura numérica que se emplearán en esta sección no serán nuevas. Simplemente, los procedimientos de aproximación de funciones introducidos en la sección anterior serán aplicados directamente al derivando o al integrando que será sustituido por una función elemental, posiblemente polinomial. Una vez aproximada la función, se integra o deriva directamente el polinomio que la aproxima.

El énfasis se hará en el desarrollo de métodos de cuadratura. Además, se considerarán únicamente los métodos basados en la interpolación.

6.2. Fórmulas de derivación numérica

Sea f una función definida en un intervalo $[a, b]$ y derivable en un punto $c \in [a, b]$. Si la evaluación de la derivada de f en c es complicada parece razonable construir un polinomio de interpolación en este intervalo y después derivarlo en el punto $x = c$. Si solamente se utilizan los extremos del intervalo $x_0 = a < x_1 = b$ el polinomio de interpolación construido con las funciones básicas de Lagrange es

$$p(x) = \frac{x - x_1}{x_0 - x_1} f(x_0) + \frac{x - x_0}{x_1 - x_0} f(x_1)$$

y su derivada en el punto $x = c$ está dada por

$$p'(x) = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Esta expresión no es otra que la que corresponde a un cociente incremental cuyo límite cuando el tamaño del intervalo tiende a 0 es el valor exacto de la derivada de f

en x_0. Sin embargo la puesta en práctica de una idea tan simple, tiene algunas dificultades cuando los cálculos de los cocientes incrementales se realizan con una aritmética finita. Cuando el incremento de la variable independiente es muy pequeño, el numerador podría aunlarse y por lo tanto también el cociente incremental aun cuando el valor teórico del límite fuese distinto de 0. Cuando se pretende calcular una derivada de alto orden de este modo, las dificultades se incrementan considerablemente.

El cálculo efectivo de los coeficientes de una fórmula de derivación numérica pueden llevarse a cabo del modo siguiente. Se construye el polinomio de interpolación del integrando f

$$p_n(x) = \sum_{i=0}^{n} l_i(x) f(x_i)$$

donde l_i representan las funciones básicas de Lagrange. La fórmula de derivación se construye como

$$D(f)(\overline{x}_0) = \sum_{i=0}^{n} l_i'(\overline{x}_0) f(x_i).$$

Es importante observar que los coeficientes $a_i = l_i'(\overline{x}_0)$ de la fórmula de derivación pueden ser calculados por el sistema lineal

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ \vdots \\ n\overline{x}_0^{n-1} \end{pmatrix}$$

obtenido derivando ambos miembros del sistema de ecuaciones del apartado 5.2. En principio, el punto de derivación $\overline{x}_0$ no tiene porque coincidir con ninguno de los nodos de interpolación.

Si lo que se pretende es construir una fórmula de derivación de mayor orden, el procedimiento es el mismo. Así para la derivada r-ésima, la fórmula tendrá los siguientes coeficientes

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ n(n-1)\cdots \\ \cdots (n-r+1)\overline{x}_0^{n-r} \end{pmatrix}$$

6.3. Método de extrapolación de Richardson

Si el conjunto de puntos usados en el procedimiento de interpolación anterior depende de un parámetro h destinado a tender a 0, es posible que por debajo de un determinado umbral para h, la precisión de la aproximación buscada, se deteriore drásticamente. Un modo de mejorar la precisión, que limita este umbral del parámetro h, es utilizar métodos de extrapolación de Richardson.

La idea básica de estos métodos es la siguiente: Si se representa por $N(h)$ la aproximación correspondiente al parámetro h, y se efectúan varia aproximaciones con distintos valores de h en el rango estable, con estos valores se puede reconstruir aproximadamente la función $N(h)$ y a continuación extrapolar el valor $N(0)$.

La cuestión es ahora cómo distribuir los valores de h, que se usan como nodos de interpolación, para que la precisión conseguida con la extrapolación mejore la precisión obtenida con cada uno de ellos. Se puede encontrar una respuesta a esta cuestión cuando teóricamente se dispone de una fórmula que establezca la precisión de la aproximación, del siguiente tipo

$$N(h) = N(0) + \alpha h^m + O(h^{m+1})$$

En esta fórmula se utiliza una O grande de Landau para representr una función para la que existe un h_0 tal que la función $\frac{O(h^{m+1})}{h^{m+1}}$ está acotada en el intervalo $(0, h_0)$. En esta terminología, se dice que $N(h)$ es una aproximación a $N(0)$ de orden m.

Si en la fórmula anterior se sustituye h por rh para un número $0 < r < 1$ arbitrario, se obtiene que

$$N(rh) = N(0) + \alpha h^m r^m + O(h^{m+1}).$$

Si se combinan estas relaciones se obtiene

$$\frac{r^m N(h) - N(rh)}{r^m - 1} = N(0) + O(h^{m+1})$$

Es decir, el cociente que está a la izquierda de la anterior igualdad

$$N_1(h) = \frac{r^m N(h) - N(rh)}{r^m - 1}$$

es una aproximación a $N(0)$ de orden $m+1$ mientras que $N(h)$ solamente es una aproximación de orden m. Es decir, combinando el resultado de la fórmula para h con el de rh se puede construir una aproximación mejor al valor buscado.

Esta idea se puede utilizar de modo recurrente para aumentar el orden de aproximación en una unidad en cada etapa. Se puede organizar el cálculo del modo siguiente: Se calculan $N(h), N(rh), N(r^2h), \cdots$ de modo que $r^i h$ se mantenga en el rango estable. Con cada pareja $N(r^{i-1}h), N(r^i h)$ se construye la aproximación de orden $m+1$

$$N_1(r^{i-1}h) = \frac{r^m N(r^{i-1}h) - N(r^i h)}{r^m - 1}$$

para $i = 1, 2, \cdots$. Si se usa el mismo procedimiento para el conjunto $N_1(h), N_1(rh), N_1(r^2h), \cdots$ se obtiene un conjunto de aproximaciones de orden $m+2$

$$N_2(r^{i-1}h) = \frac{r^{m+1} N_1(r^{i-1}h) - N_1(r^i h)}{r^{m+1} - 1}$$

para $i = 1, 2, \cdots$ y así sucesivamente.

Se pueden organizar los cálculos en la siguiente tabla

$$\begin{array}{ccccc} h & N(h) & & & \\ rh & N(rh) & N_1(h) & & \\ r^2h & N(r^2h) & N_1(rh) & N_2(h) & \\ r^3h & N(r^3h) & N_1(r^2h) & N_2(rh) & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{array}$$

6.4. Cuadratura basada en la interpolación

Sea f una función definida en un intervalo $[a,b]$ y

$$a \leqslant x_0 < x_1 < \cdots < x_n \leqslant b$$

un conjunto de nodos que serán usados para construir un polinomio de interpolación de Lagrange p_n de la función f. ES razonable pensar en aproximar la integral de la función f en $[a,b]$ por la integral de p_n en el mismo intervalo. De este modo, si se utiliza la expresión de Lagrange para el polinomio de interpolación, se obtiene

$$Q(f) = \int_a^b p(d)\,\mathrm{dx} = \sum_{i=0}^{n}(\int_a^b l_i(x)\,\mathrm{dx})f(x_i) = \sum_{i=0}^{n}\alpha_i f(x_i)$$

donde los números $\alpha_i = \int_a^b l_i(x)\,\mathrm{dx}$ son conocidos como los pesos de la fórmula de cuadratura. En principio, no hay ninguna restricción que obligue a que los extremos x_0 y x_n, coincidan con los extremos del intervalo $[a,b]$. Es frecuente llamar fórmulas de cuadratura cerradas a aquellos en que esto ocurre y abiertas en otro caso.

La primera consideración que es preciso hacer sobre una fórmula construida así, es que es exacta menos en la misma clase de polinomios que lo era la fórmula de interpolación. En este caso, puesto que la interpolación de Lagrange era exacta para polinomios de grado menor o igual que n, la fórmula de cuadratura también lo será.

En general, el cálculo de los pesos α_i puede realizarse mediante el sistema lineal

$$\begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_0 & x_1 & \cdots & x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_0^n & x_1^n & \cdots & x_n^n \end{pmatrix}\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \int_a^b 1\,\mathrm{dx} \\ \int_a^b x\,\mathrm{dx} \\ \vdots \\ \int_a^b x^n\,\mathrm{dx} \end{pmatrix}$$

$$= \begin{pmatrix} b-a \\ \frac{b^2-a^2}{2} \\ \vdots \\ \frac{b^{n+1}-a^{n+1}}{n+1} \end{pmatrix}$$

Este sistema se obtiene al imponer la condición de exactitud a los polinomios $\{1, x, x^2, \cdots, x^n\}$ o lo que es lo mismo, al integrar ambos miembros en la ecuación del apartado 5.2. En definitiva, una fórmula de cuadratura

$$Q(f) = \sum_{i=0}^{n}\alpha_i f(x_i)$$

que utiliza evaluaciones del integrando en $n+1$ puntos está basada en la interpolación de Lagrange en esos puntos si y sólo si es exacta para todo polinomio de grado menos o igual que n.

Alternativamente, se puede utilizar el sistema triangular superior de Newton

$$D\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} \int_a^b \mathrm{dx} \\ \int_a^b (x-x_0)\,\mathrm{dx} \\ \vdots \\ \int_a^b \omega_n(x)\,\mathrm{dx} \end{pmatrix}$$

que se obtiene integrando el sistema de ecuaciones del apartado 5.3.

Teorema 23 El error de cuadratura

$$e = \int_a^b f(x)\,\mathrm{dx} - Q(f)$$

de una fórmula basada en la interpolación, puede estimarse del modo siguiente

$$|e| \leqslant \frac{\max_{\xi\in[a,b]}|f^{(n+1)}(\xi)|}{(n+1)!}\int_a^b |\omega_{n+1}(x)|\,\mathrm{dx}.$$

6.5. Fórmulas cerradas de Newton-Cotes

Se considera ahora el caso en el que los nodos de interpolación están igualmente espaciados. Sea $h = x_{i+1} - x_i$ para $i = 0, 1, \cdots, n-1$. Inicialmente, se considera el intervalo de integración $[0,1]$. El sistema lineal que termina los pesos de la fórmukla de cuadratura verifican

$$\begin{pmatrix} 1 & 1 & 1 & \cdots & 1 \\ 0 & 1 & 2 & \cdots & n \\ 0 & 1 & 2^2 & \cdots & n^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & 2 & \cdots & n^n \end{pmatrix}\begin{pmatrix} \alpha_0 \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_n \end{pmatrix} = \begin{pmatrix} 1 \\ \frac{n}{2} \\ \frac{n^2}{3} \\ \vdots \\ \frac{n^n}{n+1} \end{pmatrix}$$

La resolución de los sistemas de Vandermonde correspondientes a $n = 0, 1, 2, 3$ conduce a la siguiente tabla de pesos

n	nombre	pesos
0	Extremo izquierdo	$\alpha_0 = 1$
1	Trapecios	$\alpha_0 = \alpha_1 = \frac{1}{2}$
2	Simpson	$\alpha_0 = \alpha_2 = \frac{1}{6},\ \alpha_1 = \frac{2}{3}$
3		$\alpha_0 = \alpha_3 = \frac{1}{8},\ \alpha_1 = \alpha_2 = \frac{3}{8}$

En el caso general de un intervalo $[a,b]$, la fórmula del cambio de variable permite probar directamente que los pesos de una fórmula de cuadratura son el resultado de multiplicar los pesos de la fórmula normalizada al intervalo $[0,1]$ por la longitud del intervalo. De este modo, las fórmulas de cuadratura del trapecio y de Simpson resultan ser las siguientes:

$$\int_a^b f(x)\,\mathrm{dx} \approx Q_T(f) = \frac{b-a}{2}(f(a)+f(b)) \quad \text{Trapecios}$$

$$\int_a^b f(x)\,\mathrm{dx} = \frac{b-a}{6}(f(a)+4f(\frac{a+b}{2})+f(b)) \quad \text{Simspon}$$

Teorema 24

1. (Valor medio integral) Si f es una función continua en $[a,b]$ y ω es una función integrable en $[a,b]$ tal que su signo no cambia en ese intervalo entonces

$$\int_a^b f(x)\omega(x)\,\mathrm{dx} = f(\xi)\int_a^b \omega(x)\,\mathrm{dx}$$

para algún $\xi \in (a,b)$.

2. (Valor medio discreto) Sean
   - f una función continua en $[a,b]$,
   - $\{x_i : i = 0,\cdots,n\}, n+1$ puntos distintos en $[a,b]$,
   - $\{\phi_i : i = 0,\cdots,n\}, n+1$ valores teniendo el mismo signo,

   entonces, existen $\eta \in [a,b]$ tal que

$$\sum_{i=0}^{n} \phi_i f(x_i) = \left(\sum_{i=0}^{n} \phi\right) f(\eta).$$

Lema 2 Para cualquier conjunto de puntos igualmente espaciados

$$\{x_i : i = 0,1,\cdots,n,\ x_0 = a, x_n = b\}$$

se cumple que la función $\omega_{n+1}(x - \frac{a+b}{2})$ es una función par si $n+1$ es par o impar si $n+1$ es impar.

6.6. Cuadratura compuesta

En ocasiones la mejora de la precisión no pasa por aumentar el grado de exactitud polinomial. En estas situaciones, suele ser más eficiente para calcular la integral de una función f sobre un intervalo $[a,b]$, dividir primero $[a,b]$ en un número fijado de subintervalos, y aplicar en cada uno de ellos una fórmula de Newton-Cotes de bajo grado. Las fórmulas resultantes se conocen como fórmulas de Newton-Cotes compuestas.

- *Fórmula de cuadratura de los trapecios compuesta*: Si se utiliza la fórmula de los trapecios en cada uno de los m subintervalos en que se divide el intervalo $[a,b]$ se obtiene

$$\begin{aligned} Q_{1,m}(f) &= \sum_{i=0}^{m-1} \frac{h}{2}(f(x_i) + f(x_{i+1})) \\ &= \frac{h}{2}\left(f(x_0 + 2\sum_{i=1}^{m-1} f(x_i) + f(x_m))\right) \end{aligned}$$

  donde $h = \frac{b-a}{m}$. Además el error está dado por

$$E_{1,m}(f) = -\frac{f''(\eta)}{12} m h^3 = -\frac{f''(\eta)}{12}(b-a)h^2$$

- *Fórmula de cuadratura de Simpson compuesta*: Se divide el intervalo $[a,b]$ en $2m$ subintervalos $[x_i, x_{i+1}]$ y se utiliza la fórmula de Simpson en cada uno de los intervalos $[x_{2i}, x_{2(i+1)}]$ para $i = 0,1,\cdots,m-1$. De este modo se obtiene

$$\begin{aligned} Q_{2,m}(f) &= \sum_{i=0}^{m-1} \frac{h}{6}(f(x_{2i}) + 4f(x_{2i+1}) + f(x_{2(i+1)})) \\ &= \frac{4}{3}Q_{1,2m}(f) - \frac{1}{3}Q_{1,m}(f) \end{aligned}$$

  donde $h = \frac{b-a}{m}$. Quedando relacionadas las fórmulas de cuadratura compuesta de Simpson y de los trapecios.

Para analizar el error de la fórmula de cuadratura compuesta de Simpson se tendrá en cuenta que sumando los errores de cada subintervalo, el error total está dado por

$$E_{2,m}(f) = -\sum_{i=0}^{m-1} \frac{f^{(4)}(\eta_i)}{90}\left(\frac{h}{2}\right)^5$$

para $\eta_i \in [x_i, x_{i+1}]$. Del teorema del valor medio discreto se deduce que

$$E_{2,m}(f) = -\frac{f^{(4)}(\eta)}{90} m \left(\frac{h}{2}\right)^5 = -\frac{f^{(4)}(\eta)}{180}(b-a)\left(\frac{h}{2}\right)^4$$

6.7. Fórmulas de Gauss

Cuando se intenta mejorar la precisión de una fórmula de cuadratura, es interesante saber si una elección estratégica de los puntos de interpolación puede ayudar en este sentido.

Lema 3 Si $f(x) = x^p$ con $p > n$, entonces la diferencia dividida $f[x_0, x_1, \cdots, x_n, x]$ es un polinomio de grado menor o igual que $p-n-1$.

Se escoge ω_{n+1} como el término de grado $n+1$ en una sucesión de polinomios ortogonales de coeficiente principal 1. Si f es el polinomio x^p con $p > n$ entonces $f[x_0, x_1, \cdots, x_n, x]$ es un polinomio de grado menor o igual que $p-n-1$. Por otra parte, ω_{n+1} es ortogonal a todo polinomio de grado menor o igual que n. Consecuentemente, el error

$$\begin{aligned} e &= \int_a^b f[x_0, x_1, \cdots, x_n, x]\omega_{n+1}(x)\,\mathrm{dx} \\ &= \langle f[x_0, x_1, \cdots, x_n, x], \omega_{n+1}(x)\rangle \end{aligned}$$

se anula para toda función f que sea un polinomio de grado menor o igual que $2n+1$. Las fórmulas de cuadratura basadas en la interpolación que utilizan como nodos de interpolación las raíces de un polinomio ortogonal se llaman genéricamente fórmulas de Gauss.

En el cálculo de integrales existen situaciones en las que en el integrando es posible distinguir un factor ω que es una función integrable, toma valores positivos y que permanece fijo y un segundo factor que podría cambiar según los datos del problema. Es decir, en estas situaciones

$$\int_a^b f(x)\omega(x)\,\mathrm{dx}$$

se distingue entre el integrando f (que se aproximará por interpolación) y el peso ω. En esta situación, se pueden utilizar los argumentos anteriores sin modificaciones esenciales, y de este modo, la fórmula del error de la cuadratura se convertiría en

$$\begin{aligned} e &= \int_a^b f[x_0, x_1, \cdots, x_n, x]\omega_{n+1}(x)\omega(x)\,\mathrm{dx} \\ &= \langle f[x_0, x_1, \cdots, x_n, x], \omega_{n+1}(x)\rangle_\omega. \end{aligned}$$

Las fórmulas de cuadratura Gaussianas con peso, tendrían exactitud $2n+1$ si utilizan como nodos de interpolación las raíces del polinomio ortogonal de grado n respecto

al producto escalar

$$\langle f, g\rangle_\omega = \int_a^b f(x)g(x)\omega(x)\,\mathrm{dx}$$