Apuntes de Análisis Numérico

Margarita Manterola

4 de mayo de 2009

Capítulo 1

Definiciones

1.1. Clasificación de errores

Se consideran tres tipos de errores distintos.

Inherentes o de entrada Son aquellos introducidos por los datos del problema. Representan una incertidumbre en los datos de entrada.

De redondeo Son los que se producen por la aplicación del redondeo simétrico, o del truncamiento o corte.

De truncamiento Son los que se producen al discretizar un proceso continuo, o al acotar un proceso infinito.

1.2. Clasificación de problemas

En el análisis numérico se trabaja con dos tipos de problemas: los **matemáticos**, que se refieren a procesos infinitos, las ecuaciones, el planteo algebraico, etc; y los **numéricos**, que se refieren a procesos finitos.

Cuando a un problema matemático se lo aproxima a través de un método numérico se comete un error de truncamiento.

En los problemas numéricos, los errores inherentes se propagan a través de los cálculos.

1.3. Decimales significativos

1.3.1. Notación

- A representa el valor real y exacto de un dato particular.
- $\hat{A}$ representa el valor estimado (es decir, numérico) de ese dato.
- δA es el error absoluto exacto cometido por la estimación.
- Y r_A es el error relativo cometido por la estimación: $r_A = \frac{\delta A}{A}$.

En síntesis:

$$A = \hat{A} + \delta A \tag{1.3.1}$$

$$A = \frac{\hat{A}}{1 - r_A} \tag{1.3.2}$$

Es decir que, en general, no será posible conocer el valor A con certeza, sino que siempre se tendrá un valor $\hat{A}$ con algún rango de incerteza, δA.

Por otro lado, al no conocer el valor exacto de A, en general será imposible conocer los valores δA y r_A. En cada problema, sin embargo, será posible acotarlos, de manera que $|\delta A| \leq \Delta A$ y $r_A \leq R_A$, donde ΔA y R_A son las cotas del error absoluto y relativo respectivamente.

1.3.2. Determinación de los decimales significativos

El rango de incerteza determinará cuántos decimales significativos tendrá $\hat{A}$. Por ejemplo, si $\Delta A = 0,5 \times 10^{-T}$, y $T > 0$, $\hat{A}$ tendrá T decimales significativos.

En cambio, si $T \leq 0$, no habrá decimales significativos.

Por ejemplo:

$$\begin{array}{l} \hat{A} = 124,3257058 \\ \Delta A = 0,5 \times 10^{-3} \quad = \quad 0,0005 \quad \Rightarrow \quad A = 124,326 \pm 0,0005 \\ \Delta A = 0,2 \times 10^{-3} \quad = \quad 0,0002 \quad \Rightarrow \quad A = 124,326 \pm 0,0002 \\ \Delta A = 0,5 \times 10^{1} \quad = \quad 5 \quad \Rightarrow \quad A = 120 \pm 5 \end{array}$$

La cantidad de decimales significativos está relacionada con el error del dato. Si un dato tiene 2 decimales significativos, el error será del orden 10^{-2}.

Si $x = 2{,}00$ y está **correctamente redondeado**, x tiene dos decimales significativos: $T = 2$ y $\Delta x = 0{,}005 = 0{,}5 \times 10^{-2}$. La cota de error asociado a un valor es la que les dá significado a los dígitos.

1.3.3. Dígitos medianamente significativos

Si la cota del error no es $\Delta x = 0{,}5 \times 10^{-T}$, sino que se encuentra en el rango $0{,}5 \times 10^{-T-1} < \Delta x < 0{,}5 \times 10^{-T}$, con $T > 0$, el valor tendrá T decimales significativos y un decimal medianamente significativo.

Por ejemplo, si $\hat{x} = 5{,}687$ y $\Delta x = 0{,}03$, entonces $x = 5{,}69 \pm 0{,}03$, donde el segundo decimal es medianamente significativo.

Si $T \leq 0$, el último dígito significativo está en 10^{-T} y en 10^{-T-1} se encuentra un dígito medianamente significativo.

1.4. Representación interna de números reales

Detro de la computadora, los números reales se representan utilizando la técnica del punto flotante.

De forma que: $A = m \times 10^q$, donde A es el número real que se quiere representar, m es la mantisa (que debe cumplir $0,1 \leq |m| < 1$) y q es el exponente. En la computadora, el cero inicial no se almacena, para no guardar *un cero a la izquierda*.

La mantisa debe tener una cantidad T de dígitos, que indica la precisión con que se está representando el número. Si se utilizan 32 bits para representar cada número, se la denomina **simple precisión** y, si se utilizan 64, **doble precisión**.

Para analizar la precisión máxima que puede tener un número dentro de una computadora, se utiliza la siguiente ecuación.

$$\left|\frac{\hat{A} - A}{A}\right| = \left|\frac{\hat{m} \times 10^q - m \times 10^q}{m \times 10^q}\right| = \left|\frac{\hat{m} - m}{m}\right| \leq \frac{0,5 \times 10^{-T}}{0,1} = 0,5 \times 10^{1-T} \tag{1.4.1}$$

Mediante esta ecuación podemos definir el valor μ, también llamado *unidad de máquina*. Se trata del máximo error relativo que comete la computadora cuando se representa un valor con T dígitos significativos.

$$\mu = 0,5 \times 10^{1-T} \tag{1.4.2}$$

1.4.1. Conversión entre dígitos decimales y binarios

Para poder calcular cuántos dígitos binarios ocupará un determinado número decimal se utiliza la siguiente deducción.

$$10^d = 2^b \tag{1.4.3}$$

$$d \log 10 = b \log 2 \tag{1.4.4}$$

$$\frac{d}{\log 2} = b \tag{1.4.5}$$

Es decir que para representar una cantidad d de números decimales, se necesitan aproximadamente $b = 3{,}32 \times d$ dígitos binarios (bits).

Capítulo 2

Operaciones con Errores

2.1. Errores Inherentes

Cuando se opera con datos que tienen una determinada incerteza, es necesario propagar esa incerteza a medida que se realizan las operaciones, para poder obtener la incerteza correspondiente al resultado final.

2.1.1. Operaciones Elementales

Se analiza a continuación la forma de obtener la incerteza de un resultado para las operaciones elementales.

$$\begin{aligned} x_1 &= \hat{x}_1 + \delta x_1 \\ x_2 &= \hat{x}_2 + \delta x_2 \\ y &= \hat{y} + \delta y \\ y &= x_1 \, op \, x_2 \end{aligned}$$

Donde, *op* es la operación a realizar entre x_1 y x_2.

Suma y resta

$$\begin{aligned} y &= x_1 \pm x_2 \\ y &= \hat{x}_1 + \delta x_1 \pm \hat{x}_2 + \delta x_2 \\ y &= (\hat{x}_1 \pm \hat{x}_2) + (\delta x_1 \pm \delta x_2) \end{aligned}$$

Por lo tanto:

$$\hat{y} = \hat{x}_1 \pm \hat{x}_2 \tag{2.1.1}$$

$$\delta y = \delta x_1 \pm \delta x_2 \tag{2.1.2}$$

Si ahora buscamos el módulo de δy:

$$|\delta y| = |\delta x_1 \pm \delta x_2| \leq |\delta x_1| + |\delta x_2| \leq \Delta x_1 + \Delta x_2 = \Delta y \tag{2.1.3}$$

Por ejemplo:

$$\begin{aligned} x_1 &= 23,5 \pm 0,2 \\ x_2 &= 11,7 \pm 0,5 = 12 \pm 0,5 \\ x_1 + x_2 &= 35,2 \pm 0,5 = 35 \pm 0,5 \end{aligned}$$

Producto

Para el caso del producto se utiliza la expresión del valor relativo de x e y, en lugar del valor absoluto.

$$\frac{\hat{y}}{1 - r_y} = \frac{\hat{x_1}}{1 - r_{x_1}} . \frac{\hat{x_2}}{1 - r_{x_2}} \tag{2.1.4}$$

$$\hat{y} = \hat{x_1}\hat{x_2} \tag{2.1.5}$$

$$\frac{1}{1 - r_y} = \frac{1}{1 - r_{x_1}} . \frac{1}{1 - r_{x_2}} \tag{2.1.6}$$

2.1.2. Fórmula general de propagación

Si $y = F(x_1, x_2, \ldots, x_n)$, la cota del error absoluto Δy está dada por:

$$\Delta y = \sum_{i=1}^{n} \left| \frac{\partial F}{\partial x_i} \right| \Delta x_i \tag{2.1.7}$$

Por ejemplo, si $w = 3x + y - z$:

$$\Delta w = \left| \frac{\partial w}{\partial x} \right| \Delta x + \left| \frac{\partial w}{\partial y} \right| \Delta y + \left| \frac{\partial w}{\partial z} \right| \Delta z \tag{2.1.8}$$

$$= 3\Delta x + \Delta y + \Delta z \tag{2.1.9}$$

El error relativo de w es:

$$r_w = \frac{\Delta w}{w} = \left| \frac{\partial w}{\partial x} \right| \frac{\Delta x}{w} + \left| \frac{\partial w}{\partial y} \right| \frac{\Delta y}{w} + \left| \frac{\partial w}{\partial z} \right| \frac{\Delta z}{w} \tag{2.1.10}$$

$$r_w = \left| \frac{\partial w}{\partial x} \right| \frac{r_x x}{w} + \left| \frac{\partial w}{\partial y} \right| \frac{r_y y}{w} + \left| \frac{\partial w}{\partial z} \right| \frac{r_z z}{w} \tag{2.1.11}$$

Se puede obtener una cota al error relativo R_w tomando los valores estimados de las variables.

2.1.3. Gráfica de procesos - Factores de Amplificación

Cuando se analiza un proceso en el que se efectúan una o más operaciones con los datos de entrada, se puede utilizar la técnica gráfica para obtener la incerteza total del proceso.

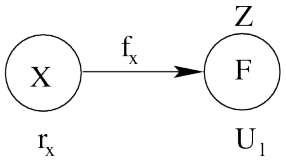

Figura 2.1.1: Gráfica general de un proceso unitario

En la figura 2.1.1, podemos ver la gráfica típica de un proceso unitario, donde $Z = F(X)$.

$$r_Z = f_X r_X + U_1 \tag{2.1.12}$$

En la figura 2.1.2, podemos ver la gráfica típica de un proceso binario, donde $Z = X\ op\ Y$ y $r_Z = f_X r_X + f_Y r_Y + U_1$.

En ambos casos U representa el error agregado por la unidad de máquina. Es un error que siempre debe tenerse en cuenta cuando se realizan operaciones con la computadora.

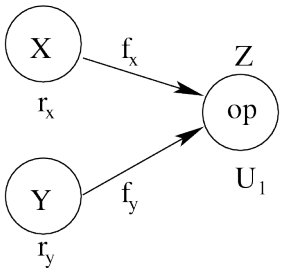

Figura 2.1.2: Gráfica general de un proceso binario

Por ejemplo, para la operación $Z = A^2 - B^2$, una gráfica posible para este proceso estará dada por la representada en la figura 2.1.3. En este caso, el error relativo del resultado estará dado por:

$$\begin{aligned} r_Z &= \frac{Z_1}{Z}(r_A + r_A + U_1) - \frac{Z_2}{Z}(r_B + r_B + U_2) + U_3 \\ r_Z &= \frac{2r_A Z_1}{Z} - \frac{2r_B Z_2}{Z} + \frac{Z_1 U_1}{Z} + \frac{Z_2 U_2}{Z} + U_3 \end{aligned}$$

El problema de este proceso en particular surge cuando utilizamos valores muy cercanos para A y B. Por ejemplo:

$$\begin{aligned} A &= 43{,}21 \pm 0{,}005 = 43{,}21(1 \pm 1{,}2 \times 10^{-4}) \\ B &= 43{,}11 \pm 0{,}005 = 43{,}11(1 \pm 1{,}2 \times 10^{-4}) \end{aligned}$$

Al acotar r_A, r_B y U (con $T = 6$):

$$\begin{aligned} |r_A| &\leq R_A = 1{,}2 \times 10^{-4} \\ |r_B| &\leq R_B = 1{,}2 \times 10^{-4} \\ |U_1| = |U_2| = |U_3| &\leq U = 0{,}5 \times 10^{1-T} = 0{,}5 \times 10^{-5} \end{aligned}$$

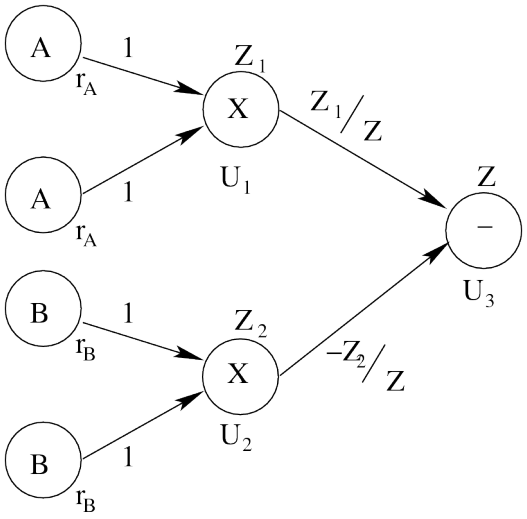

Figura 2.1.3: Gráfica de la operación $A^2 - B^2$

Ahora, acotamos r_Z:

$$|r_Z| \leq \left|\frac{Z_1}{Z}\right| 2R_A + \left|\frac{Z_2}{Z}\right| 2R_B + \left(\frac{Z_1}{Z} + \frac{Z_2}{Z} + 1\right) U = R_Z$$

El último término, $\left(\frac{Z_1}{Z} + \frac{Z_2}{Z} + 1\right) U$, es el error debido al redondeo, también llamado **factor de amplificación global de redondeo**.

Teniendo en cuenta que $Z_1 = 1867{,}1041$, $Z_2 = 1858{,}4721$, y $Z = 8{,}632$, tenemos que: $\frac{Z_1}{Z} = 216{,}3$ y $\frac{Z_2}{Z} = 215{,}3$. De modo que:

$$\begin{aligned} R_Z &\approx 420R_A + 420R_B + 421\mu \\ R_Z &\approx 0{,}05 + 0{,}05 + 0{,}002 = 0{,}102 \end{aligned}$$

Sin embargo, podemos recordar que los problemas numéricos pueden ser evaluados de varias maneras diferentes. De modo que efectuamos el mismo cálculo utilizando la ecuación: $Z = (A+B)(A-B)$, cuya gráfica se representa en la figura 2.1.4.

En este caso, el error relativo del resultado estará dado por:

$$\begin{aligned} r_Z &= \left(r_A\frac{A}{Z_1} + r_B\frac{B}{Z_1} + U_1\right)1 + \left(r_A\frac{A}{Z_2} - r_B\frac{B}{Z_2} + U_2\right)1 + U_3 \\ r_Z &= r_A\left(\frac{A}{Z_1} + \frac{A}{Z_2}\right) + r_B\left(\frac{B}{Z_1} - \frac{B}{Z_2}\right) + (U_1 + U_2 + U_3) \\ r_Z &= r_A\left(\frac{A(A-B) + A(A+B)}{Z}\right) + r_B\left(\frac{B(A-B) - B(A+B)}{Z}\right) + (U_1 + U_2 + U_3) \\ r_Z &= r_A\left(\frac{2A^2}{Z}\right) - r_B\left(\frac{2B^2}{Z}\right) + (U_1 + U_2 + U_3) \end{aligned}$$

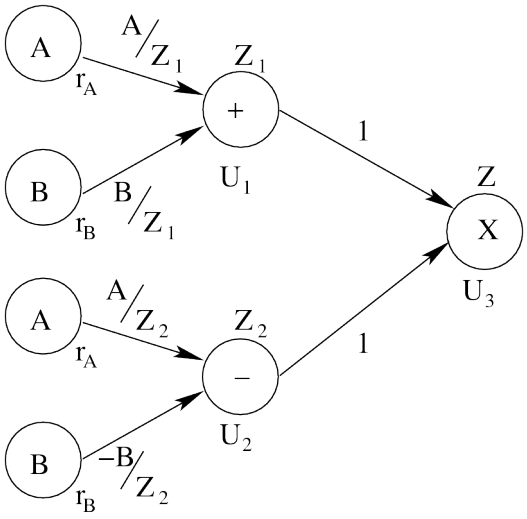

Figura 2.1.4: Gráfica de la operación $(A+B)(A-B)$

Y al acotar r_Z obtenemos:

$$|r_Z| \leq 2R_A \left|\frac{A^2}{Z}\right| + 2R_B \left|\frac{B^2}{Z}\right| + 3U$$

Es decir, la misma expresión que con el proceso anterior, excepto por el factor global de amplificación.

2.1.4. Condición de un problema

La **condición de un problema** es la suma de los factores de amplificación. En el caso en que R_A y R_B sean de un mismo orden, pueden sumarse. Si los órdenes de magnitud son similares, puede tomarse el mayor, y si son muy distintos, deben considerarse por separado.

En el ejemplo considerado anteriormente, cuando se utilizaba la fórmula $A^2 - B^2$, teníamos que $R_Z = 420R_A + 420R_B + 421\mu$. Mientras que cuando se utilizaba la fórmula, $(A+B)(A-B)$, obteníamos $R_Z = 420R_A + 420R_B + 3\mu$.

En ambos casos, la condición del problema es $C_P = 840$. Podemos observar que se trata de un problema muy *mal condicionado*, y esto se debe a que estamos operando con números muy cercanos.

> **Regla**: la cantidad T de dígitos significativos indica que el error va a estar en 10^{1-T}. Si la condición del problema es de orden 10^X, hace que se pierdan X dígitos significativos. Por lo tanto, si se quiere tener Z dígitos significativos en el resultado, se debe operar con un $T = Z + X$.

Un problema es **estable** cuando la condición del problema es mucho menor que la cantidad de términos del problema. Mientras que cuando la condición del problema es mayor, el problema es **inestable**.

De la misma manera, cuando el factor de amplificación que acompaña a la entrada es menor que 1, se dice que la entrada es **estable**, mientras que si el factor de amplificación es mayor que 1, la entrada será **inestable**.

2.1.5. Condición del Algoritmo

Si se tiene un problema como el explicado anteriormente, que está muy mal condicionado, no se puede cambiar la condición del problema, pero sí se puede trabajar sobre la **condición del algoritmo**, para obtener un menor factor global de amplificación.

La condición del algoritmo está dada por:

$$C_A = \frac{F_{GA}}{C_P} \tag{2.1.13}$$

De modo que, en el ejemplo anterior, $C_{A_1} = \frac{421}{840}$ mientras que $C_{A_2} = \frac{3}{840}$. Como podemos ver, el segundo algoritmo es mucho mejor que el primero.

La unidad de máquina y el algoritmo que se elijan deben lograr -al menos- que no se pierdan los dígitos significativos de los datos de entrada.

Si, por ejemplo, se utilizara $T = 4$ dígitos, se tendría una unidad de máquina $\mu = 0{,}5 \times 10^{-3}$. Con este valor, para el primer algoritmo, se tendría $R_Z \approx 1008 \times 10^{-4} + 2150 \times 10^{-4}$. Es decir que el error relativo debido al redondeo es mayor que el inherente. Esto es **malísimo**.

Si, en cambio, se elige el otro algoritmo: $R_Z \approx 1008 \times 10^{-4} + 15 \times 10^{-4}$.

2.1.6. Perturbaciones Experimentales

El método de las perturbaciones experimentales consiste en realizar un cambio en alguno de los valores de entrada, para analizar cómo cambia la salida.

En un proceso que tiene N datos de entrada y una salida:

$$\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_j, \ldots, \hat{x}_N \overset{proceso}{\longrightarrow} \hat{y}$$

Se puede realizar una perturbación en el dato $\hat{x}_j$:

$$\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_j + \xi_j, \ldots, \hat{x}_N \overset{proceso}{\longrightarrow} \hat{y}_j$$

Se define la **perturbación de entrada**:

$$\rho_j = \left| \frac{\xi_j}{\hat{x}_j} \right| \tag{2.1.14}$$

La **perturbación de salida**:

$$\omega_j = \left| \frac{\hat{y}_j - \hat{y}}{\hat{y}} \right| \tag{2.1.15}$$

Y el **factor de amplificación** del dato $\hat{x}_j$:

$$f_j = \frac{\omega_j}{\rho_j} \tag{2.1.16}$$

Si al graficar f_j en función de ξ_j se obtiene una gráfica lo suficientemente estable, esto implica que el algoritmo utilizado para el proceso funciona aceptablemente bien.

Si, en cambio, no se logra obtener una zona lo suficientemente estable, esto implica que el algoritmo es inestable.

Si, por ejemplo, se intenta mejorar el algoritmo utilizado en el proceso aumentando la precisión, se obtendrá una variación relativa del resultado debido a que la precisión es mejor. Para el doble de precisión, sería:

$$\omega_{2T} = \left| \frac{\hat{y} - \hat{y}_{2T}}{\hat{y}_{2T}} \right| \tag{2.1.17}$$

Si se quiere propagar los errores en un problema, sin conocer el algoritmo, se puede utilizar el método de las perturbaciones experimentales para aproximar la derivada.

$$\frac{\partial y}{\partial x_j} \approx \frac{\Delta y}{\Delta x_j} = \frac{\hat{y_j} - \hat{y}}{\hat{x_j} - \xi_j} \tag{2.1.18}$$

Para el caso de una función $z(x, y)$, la propagación de errores podría hacerse según:

$$\Delta z = \frac{\Delta z_x}{\hat{x} - \xi_x} \Delta x + \frac{\Delta z_y}{\hat{y} - \xi_y} \Delta y \tag{2.1.19}$$

Donde Δz_x y Δz_y representan las variaciones de la salida al variar x o y y mantener la otra variable constante.

2.2. Errores de truncamiento

Los errores de truncamiento son aquellos que se introducen al volver finito un proceso infinito. Por ejemplo:

$$S = \sum_{i=1}^{\infty} a_i \tag{2.2.1}$$

$$P = \sum_{i=1}^{n} a_i \tag{2.2.2}$$

P implica tomar una cantidad N de términos, de un problema con términos infinitos. El error cometido en este caso se puede obtener de:

$$S = P + \Delta P \tag{2.2.3}$$

$$\Delta P \approx |S[n+1] - S[n]| \tag{2.2.4}$$

$$\Delta P \approx a_{n+1} \tag{2.2.5}$$

Por ejemplo:

$$S = \sum_{i=1}^{\infty} \frac{1}{y^2} \tag{2.2.6}$$

$$P = \sum_{i=1}^{10} \frac{1}{y^2} = 1{,}549767731 \tag{2.2.7}$$

$$\Delta P \approx \frac{1}{11^2} = 0{,}008 \tag{2.2.8}$$

De manera que $S = 1{,}55 \pm 0{,}008$.

En un caso en el que haya cancelación de términos, es posible introducir un error de truncamiento, de forma que la fórmula aproximada sea más precisa que sin el truncamiento.

Tomamos como ejemplo un caso en el que queremos calcular la diferencia entre los senos de dos números muy cercanos:

$$\begin{array}{rclcrcl} x_1 &=& 0,57640 & \quad & \Delta x_1 &=& 0,000005 = 0,5 \times 10^{-5} \\ x_2 &=& 0,57630 & & \Delta x_2 &=& 0,000005 = 0,5 \times 10^{-5} \end{array}$$

Por el método directo:

$$\begin{aligned} \hat{y} &= \operatorname{sen}(\hat{x}_1) - \operatorname{sen}(\hat{x}_2) \\ \hat{y} &= 0{,}54501 - 0{,}54493 \\ y &= 0{,}00008 \pm 0{,}00001 \end{aligned}$$

Ahora, para utilizar el método del truncamiento, vamos a tomar el desarrollo en serie de Taylor de la función $\operatorname{sin} x_2$ alrededor del punto x_1:

$$\operatorname{sen} x_2 = \operatorname{sen}(x_1) + \cos(x_1)(x_2 - x_1) - sen(x_1)\frac{(x_2 - x_1)^2}{2} + \ldots$$

Ahora, al truncar la serie y quedarnos únicamente con los primeros dos términos, podemos utilizarlos para calcular la diferencia deseada, y luego utilizar el tercero para obtener una cota del error:

$$\begin{aligned} y_{[1]} &= \operatorname{sen}(x_1) - \operatorname{sen}(x_2) \\ y_{[1]} &= -\cos(x_1)(x_2 - x_1) \\ y_{[1]} &= -(0{,}83843 \pm 0{,}000005)(-0{,}00010) \\ y_{[1]} &= 0{,}00083843 \pm 0,5 \times 10^{-9} \end{aligned}$$

El error debido al truncamiento estará dado por:

$$\Delta y_{[1]} = \left| \operatorname{sen}(x_1)\frac{(x_2 - x_1)^2}{2} \right| = 3 \times 10^{-9}$$

De forma que el resultado final será: $y = 0{,}000083843 \pm 0{,}000000003$.

2.3. Errores de Redondeo

$$\begin{array}{rcllll} \hat{A} & = & 0{,}1234567 \times 10^0 & \rightarrow & 0{,}00001234567 & \times 10^4 \\ \hat{B} & = & 0{,}4711325 \times 10^4 & \rightarrow & 0{,}4711325 & \times 10^4 \\ \hline \xi & = & \hat{A} + \hat{B} & = & 0{,}47114484567 & \times 10^4 \end{array}$$

Ahora, redondeamos el valor de ξ a: $\xi = 0{,}4711448 \times 10^4$. Y operamos con el valor $\hat{C} = -0{,}4711325 \times 10^4$.

$$z = \xi + \hat{C} = 0{,}4711448 \times 10^4 - 0{,}4711325 \times 10^4 = 0{,}0000123 \times 10^4$$

Es decir que:

$$z = (\hat{A} + \hat{B}) + \hat{C} = 0{,}0000123 \times 10^4$$

Si, en cambio hubiéramos operado primero con $\hat{B}$ y $\hat{C}$:

$$\begin{array}{rcllll} \hat{B} & = & 0{,}4711325 \times 10^4 & \rightarrow & 0{,}4711325 & \times 10^4 \\ \hat{C} & = & -0{,}4711325 \times 10^4 & \rightarrow & -0{,}4711325 & \times 10^4 \\ \hline \xi & = & \hat{B} + \hat{C} & = & 0 & \times 10^4 \end{array}$$

Y al operar con A:

$$z = \xi + \hat{A} = 0 + 0{,}1234567 \times 10^0 = 0{,}1234567 \times 10^0$$

En conclusión, cuando se opera con números que no son todos de la misma magnitud, el orden en que se realizan las operaciones puede ser muy significativo a la hora de obtener los resultados.

Capítulo 3

Sistemas de ecuaciones lineales

3.1. Repaso básico

Sea A una matriz perteniciente al espacio $\Re^{N\times N}$. Si A es invertible, existe solución del sistema de ecuaciones $A\vec{x} = B$.

Se define la matriz **triangular inferior** L como aquella cuyos valores por encima de la diagonal son cero. Y la matriz **triangular superior** U como aquella cuyos valores por debajo de la diagonal son cero.

Una matriz es **diagonal dominante** si el valor absoluto de cada elemento de la diagonal es mayor que la suma de los valores absolutos de todos los otros elementos de su misma fila o columna.

3.1.1. Eliminación gaussiana

La eliminación gaussiana es un método para obtener la solución de problema. Se trata primero de realizar la **eliminación** de ciertos valores de A para llegar a la matriz diagonal superior U, luego realizar la **sustitución** de los valores y llegar al vector $\vec{x}$.

Para realizar la eliminación, se utiliza la técnica del **pivoteo**. Se elige un número que será el pivote, y a partir de allí se crea un multiplicador m.

Por ejemplo, para una matriz de 2×2, el multiplicador que se utilizaría sería: $m_{21} = \frac{A_{21}}{A_{11}}$. Donde el pivote es A_{11}.

Para obtener los nuevos valores de la matriz:

$$\begin{array}{rcccl} A_{21} & = & A_{21} - A_{11}m_{21} & = & A_{21} - A_{11}\frac{A_{21}}{A_{11}} = 0 \\ A_{22} & = & A_{22} - A_{12}m_{21} & = & A_{22} - A_{12}\frac{A_{21}}{A_{11}} \end{array}$$

En el caso en que uno de los pivotes a elegir se anule, será necesario intercambiar esa fila con alguna otra cuyo pivote no se haya anulado. Si no es posible, el sistema no tiene solución.

Además pueden intercambiarse las filas con la finalidad de obtener un pivote más grande y tener un menor error de pivoteo. Esta técnica es la que comunmente se llama *de pivoteo.*

Si la matriz es **simétrica** y **definida positiva**, se puede realizar la eliminación gaussiana sin preocuparse de que el pivote se anule.

3.1.2. Descomposición LU

Este método resulta útil cuando se tienen que resolver varios sistemas que tienen una misma matriz de coeficientes.

Se tiene el sistema $A\vec{x} = \vec{B}$. Y se descompone a la matriz A en dos matrices L y U, de forma que $LU\vec{x} = \vec{B}$. Se define $\vec{y}$ tal que $U\vec{x} = \vec{y}$ y $L\vec{y} = \vec{B}$.

Cuando los valores de la diagonal de L son todos 1, se la llama *descomposición de Doolittle*. En este caso, U es la matriz triangular obtenida al utilizar eliminación gaussiana, y L es la matriz de los pivotes utilizados, completada con unos en la diagonal.

Al utilizar la eliminación gaussiana, dejamos entre paréntesis el pivote utilizado, en lugar del cero.

Por ejemplo, para el caso de una matriz de 2×2:

$$\left[\begin{array}{cc|c} 3 & 4 & 5 \\ 2 & 3 & 4 \end{array}\right] = \left[\begin{array}{cc|c} 3 & 4 & 5 \\ (\frac{2}{3}) & \frac{1}{3} & \frac{2}{3} \end{array}\right]$$

A partir de estos datos podemos descomponer A en las matrices L y U:

$$A = \begin{bmatrix} 3 & 4 \\ 2 & 3 \end{bmatrix} = LU = \begin{bmatrix} 1 & 0 \\ \frac{2}{3} & 1 \end{bmatrix} \begin{bmatrix} 3 & 4 \\ 0 & \frac{1}{3} \end{bmatrix}$$

3.1.3. Norma Vectorial

Se llama **Norma** P, con $1 \leq P < \infty$, a la norma dada por:

$$\|X\|_P = \left(|X_1|^P + |X_1 2^P + \ldots + |X_N|^P\right)^{\frac{1}{P}} \tag{3.1.1}$$

Esta norma debe cumplir las siguientes propiedades:

$$\begin{aligned} \vec{x} \neq \vec{0} \Rightarrow \|\vec{x}\| &> 0 \\ \vec{x} = \vec{0} \Rightarrow \|\vec{x}\| &= 0 \\ \|\alpha\vec{x}\| &= |\alpha| \, \|\vec{x}\| \\ \|\vec{x} + \vec{y}\| &\leq \|\vec{x}\| + \|\vec{y}\| \end{aligned}$$

Según el valor que tome P, serán distintas las operaciones que debamos realizar sobre las componentes del vector. En el caso en que $P = 2$, obtendremos la **distancia euclídea**. En el caso en que $P = 1$, obtendremos la suma de las componentes, y si $P = \infty$, se tratará del máximo valor absoluto del vector.

3.1.4. Norma Matricial

La norma matricial se define con las mismas propiedades que la norma vectorial, y un par más:

$$\begin{aligned} A^{N\times M} \neq 0^{N\times M} \Rightarrow \|A^{N\times M}\| &> 0 \\ \|\alpha A^{N\times M}\| &= |\alpha|\,\|A^{N\times M}\| \\ \|A^{N\times M} + B^{N\times M}\| &\leq \|A^{N\times M}\| + \|B^{N\times M}\| \\ \|A^{N\times M}\ B^{M\times P}\| &\leq \|A^{N\times M}\|\,\|B^{M\times P}\| \\ \|A^{N\times M}\vec{x}\| &\leq \|A^{N\times M}\|\,\|\vec{x}\| \end{aligned}$$

En particular, la última propiedad es la que sirve para definir la norma matricial.

Al igual que en el caso de la norma vectorial, se a partir de estas propiedades se pueden definir varias normas matriciales:

$$\begin{aligned} \|A^{N\times M}\| &= \max_{x\neq 0} \frac{\|A^{N\times M}\ \vec{x}\|}{\|\vec{x}\|} \\ \|A^{N\times M}\|_\infty &= \max_{1\leq i\leq N} \sum_{j=1}^{M} |A_{ij}| \\ \|A^{N\times M}\|_1 &= \max_{1\leq j\leq M} \sum_{i=1}^{N} |A_{ij}| \\ \|A^{N\times M}\|_2 &= \frac{\lambda\,\text{máx}}{\lambda\,\text{mín}} \end{aligned}$$

Es decir que la norma infinita equivale a la máxima suma por filas, y la norma uno equivale a la máxima suma por columnas. La norma 2 se define solamente para matrices simétricas.

3.2. Errores

Los errores con los que se trabaja al analizar sistemas de ecuaciones lineales, son errores inherentes o de redondeo. No hay errores de truncamiento.

3.2.1. Errores Inherentes

Tomamos el sistema $A\vec{x} = \vec{B}$, teniendo en cuenta una matriz de incertezas δA y un vector de incertezas $\delta\vec{x}$ con los cuales se obtiene el vector resultado $\vec{B}$, para analizar la matriz.

$$
\begin{aligned}
A\vec{x} &= \vec{B} \\
(A+\delta A)(\vec{x}+\delta\vec{x}) &= \vec{B} \\
A\vec{x} + A\delta\vec{x} + \delta A(\vec{x}+\delta\vec{x}) &= \vec{B} \\
A\delta\vec{x} &= -\delta A(\vec{x}+\delta\vec{x}) \\
\delta\vec{x} &= A^{-1}(-\delta A(\vec{x}+\delta\vec{x})) \\
\|\delta\vec{x}\| &\leq \|A^{-1}\|\,\|\delta A\|\,\|\vec{x}+\delta\vec{x}\| \\
\frac{\|\delta\vec{x}\|}{\|\vec{x}+\delta\vec{x}\|} &\leq \|A^{-1}\|\,\frac{\|\delta A\|}{\|A\|}\,\|A\| \\
\frac{\|\delta\vec{x}\|}{\|\vec{x}+\delta\vec{x}\|} &\leq K(A)\,\frac{\|\delta A\|}{\|A\|}
\end{aligned}
$$

$K(A) = \|A^{-1}\|\,\|A\|$ es el número de condición de la matriz. Juega el mismo rol que el número de condición del problema.

A continuación, se analiza el caso en el que hay incertezas únicamente en los datos de entrada y salida, pero no en la matriz del sistema.

$$
\begin{aligned}
A\vec{x} &= \vec{B} \\
A(\vec{x}+\delta\vec{x}) &= \vec{B}+\delta\vec{B} \\
A\vec{x} + A\delta\vec{x} &= \vec{B}+\delta\vec{B} \\
\delta\vec{x} &= A^{-1}(\delta\vec{B}) \\
\|\delta\vec{x}\| &\leq \|A^{-1}\|\,\|\delta\vec{B}\| \\
\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} &\leq \frac{\|A^{-1}\|\,\|\delta\vec{B}\|}{\|\vec{x}\|} \\
\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} &\leq \frac{\|A^{-1}\|}{\|A\|}\,\frac{\|\delta\vec{B}\|}{\|\vec{B}\|} \\
\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} &\leq K(A)\,\frac{\|\delta\vec{B}\|}{\|\vec{B}\|}
\end{aligned}
$$

3.2.2. Residuos

Tomando un sistema $A\vec{x} = \vec{B}$, se propone un $\tilde{x}$ como solución del sistema. Se define al residuo R:

$$
\begin{aligned}
R &= b - A\tilde{x} && (3.2.1)\\
R &= A(x-\tilde{x}) && (3.2.2)\\
A^{-1}R &= x-\tilde{x} \\
A^{-1}R &= e && (3.2.3)\\
& && (3.2.4)
\end{aligned}
$$

Donde e es el error que se comete al estimar x con $\tilde{x}$. Es importante notar que no alcanza con saber si el residuo es pequeño, es necesario saber que el error sea pequeño.

Podemos proponer una cota (no muy buena) para el error:

$$\begin{aligned} \|e\| &\leq \|A^{-1}\| \, \|R\| \\ \frac{\|e\|}{\|\vec{x}\|} &\leq K(A) \frac{\|R\|}{\|\vec{B}\|} \end{aligned}$$

3.2.3. Propagación de los errores de redondeo

3.3. Refinamiento Iterativo

El método de refinamiento iterativo, es un método **directo**, que busca corregir el error de redondeo. No se ocupa de los errores inherentes, sino únicamente de los de redondeo.

3.3.1. Definición

Teniendo en cuenta la ecuación (3.2.2), se puede afirmar que si se conociera δx dada por $x = \tilde{x} + \delta x$, es posible conocer con exactitud el valor de x.

El método del refinamiento iterativo se basa en este concepto, para tratar de lograr una mejor aproximación de $\tilde{x}$ a x.

$$\tilde{x}_{N+1} = \tilde{x}_N + \delta x_N \tag{3.3.1}$$

Para obtener δx_N, se necesita obtener el residuo, evaluando la matriz A en $\tilde{x}_N$, y luego multiplicar la inversa de A por el residuo obtenido. Es decir:

$$\delta x_N = A^{-1}(B - A\tilde{x}_N) \tag{3.3.2}$$

La resta $(B - A\tilde{x}_N)$ debe efectuarse con el doble de los dígitos significativos deseados, y luego guardar el resultado con la cantidad de dígitos requerida.

Para que todas estas operaciones se hagan más sencillas, se utiliza la descomposición $A = LU$:

$$\begin{aligned} LU\delta\vec{x} &= R \\ L\vec{y} &= R \\ U\delta x &= \vec{y} \end{aligned}$$

Es decir que teniendo la matriz L, se averigua en primer lugar el valor de $\vec{y}$, y luego, teniendo la matriz U se averigua el vector $\delta\vec{x}$.

3.3.2. Número de condición de la matriz

Ahora se puede tomar otra cota del número de condición de la matriz:

$$K(A) \approx 10^T \frac{\|\delta x\|}{\|\hat{x}\|} \tag{3.3.3}$$

Que nos permite definir el número p:

$$p = \log K(A) \approx T + \log \frac{\|\delta x\|}{\|\hat{x}\|} \tag{3.3.4}$$

Si $p > T$, se pierden todos los dígitos significativos. Pero si $p < T$, se pierden algunos dígitos significativos, pero no todos. Se define la variable $q = T - p$, que identifica la cantidad de dígitos significativos que se quedan.

Con cada refinamiento iterativo, se gana un dígito significativo.

3.4. Métodos Iterativos

Se llama métodos iterativos a aquellos métodos que comienzan con una semilla cualquiera, y van iterando alrededor de la matriz hasta obtener valores que se acerquen al valor real de la solución.

Si tenemos un valor asignado para cada una de las las componentes de $\vec{x}$, el valor de $\vec{x_i}$ estará dado por:

$$x_i = (b_i - a_{i1}x_1 - a_{i2}x_2 - \ldots - a_{ii-1}x_{i-1} - a_{ii+1}x_{i+1} - \ldots - a_{iN}x_N)\frac{1}{a_{ii}} \tag{3.4.1}$$

3.4.1. Método de Jacobi

Requiere una matriz que cumpla con ciertas condiciones. Por ejemplo, este método puede utilizarse para matrices que sean *diagonal dominante*.

En primer lugar, es necesario tener una semilla con valores asignados para cada una de las componentes. Por ejemplo, puede tomarse el valor 0 para todas las componentes.

A continuación se realiza una iteración utilizando la ecuación (3.4.1) y se encuentran nuevos valores para las componenetes. Y así.

La ecuación que define a este método es:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(k)} \right) \tag{3.4.2}$$

3.4.2. Método de Gauss-Seidel

Es muy similar al método de Jacobi. La diferencia reside en que a medida que va calculando los nuevos valores de las componentes, los va utilizando en el cálculo de la siguiente.

La ecuación que lo define es:

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(k)} \right) \tag{3.4.3}$$

3.4.3. Método SOR

Este método está dado por la ecuación:

$$x_i^{(k+1)} = \omega \left\{ \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{N} a_{ij} x_j^{(k)} \right) - x_i^{(k)} \right\} + x_i^{(k)} \tag{3.4.4}$$

Donde ω es el factor de aceleración o frenado. Cuando $\omega = 1$, esta ecuación es equivalente a la (3.4.3). Cuando $\omega > 1$ el método está *sobre relajado* y cuando $\omega < 1$, está *sub relajado*.

3.4.4. Análisis de los métodos iterativos

Para analizar los métodos explicados, vamos a utilizar una descomposición de la matriz $A = -L + D - U$. Buscando obtener siempre una ecuación de la forma $\vec{x}^{(k+1)} = T\vec{x}^{(k)} + \vec{C}$.

La descomposición será tal que L es la matriz que tiene los valores recíprocos de A por debajo de la diagonal y cero en las otras posiciones.

U es la matriz que tiene los valores recíprocos de A por encima de la diagonal y cero en las otras posiciones.

Y D es la matriz que tiene los valores de la diagonal de A, y cero en las otras posiciones.

De esta manera, la fórmula para el método de Jacobi estará dada por:

$$\begin{aligned} D\vec{x}^{(k+1)} &= \vec{B} + L\vec{x}^{(k)} + U\vec{x}^{(k)} \\ \vec{x}^{(k+1)} &= \underbrace{D^{-1}(L+U)}_{T_J}\vec{x}^{(k)} + \underbrace{D^{-1}\vec{B}}_{C_J} \end{aligned}$$

La fórmula para el método de Gauss-Seidel:

$$\begin{aligned} D\vec{x}^{(k+1)} &= \vec{B} + L\vec{x}^{(k+1)} + U\vec{x}^{(k)} \\ (D-L)\vec{x}^{(k+1)} &= \vec{B} + U\vec{x}^{(k)} \\ \vec{x}^{(k+1)} &= \underbrace{(D-L)^{-1}U}_{T_{GS}}\vec{x}^{(k)} + \underbrace{(D-L)^{-1}\vec{B}}_{C_{GS}} \end{aligned}$$

En el caso del método SOR es más difícil obtener la fórmula deseada, de manera que se expresan los valores T_ω y C_ω, que dependen del factor de aceleración:

$$\begin{aligned} T_\omega &= (D - \omega L)^{-1}\left[(1-\omega)D + \omega U\right] \\ C_\omega &= \omega(D - \omega L)^{-1}\vec{B} \end{aligned}$$

Convergencia

Si el sistema es convergente, va a llegar un punto en que $\vec{x}^{(k)} = \vec{x}^{(k+1)}$:

$$\begin{aligned} \vec{x}^{(k+1)} - \vec{x}^{(k)} &= T(\vec{x}^{(k)} - \vec{x}^{(k-1)}) \\ \vec{e}^{(k+1)} &= T\vec{e}^{(k)} \\ \vec{e}^{(k+1)} &= T(T\vec{e}^{(k-1)}) \\ \vec{e}^{(k+1)} &= T^{k+1}\vec{e}^{(0)} \end{aligned}$$

Donde $\vec{e}^{(k+1)}$ es el error de truncamiento y $\vec{e}^{(0)}$ es el error de arranque. De esta forma, podemos acotar el error de truncamiento:

$$\begin{aligned} \|\vec{e}^{(k+1)}\| &\leq \|T^{k+1}\| \, \|\vec{e}^{(0)}\| \\ \|\vec{e}^{(k+1)}\| &\leq \|T\|^{k+1} \, \|\vec{e}^{(0)}\| \end{aligned}$$

Es decir que el sistema será convergente siempre y cuando $\|T\| < 1$.

Autovectores

Si ahora tenemos en cuenta los autovectores de la matriz, dados por $A\vec{x} = \lambda\vec{x}$.

$$\begin{aligned} \vec{e}^{(0)} &= \vec{x}^{(0)} - x &&= \textstyle\sum_{i=1}^{N} \alpha_i V_i \\ \vec{e}^{(k+1)} &= T^{k+1}\vec{e}^{(0)} &&= T^{k+1}\left(\textstyle\sum_{i=1}^{N} \alpha_i V_i\right) \end{aligned}$$

Utilizando los autovectores puedo demostrar que el método de Gauss-Seidel converge más rápidamente que el de Jacobi.

Capítulo 4

Ecuaciones No Lineales

Se trata de buscar las raíces que hacen que $F(x) = 0$, de ecuaciones que no son lineales.

$$\textbf{Teorema: } \sup F(x) \in C^0[a_0, b_0] F(a_0)F(b_0) < 0 \Rightarrow F(\alpha_k) = 0 \quad (4.0.1)$$
$$\text{para un número impar de valores } x_k \in [a_0, b_0]$$

En este caso, vamos a trabajar con la hipótesis de que en el intervalo en el cual estamos operando hay una sola raíz.

4.1. Método de la bisección

El método de la bisección consta de tres pasos:

1. Partir el intervalo en dos:

$$m_{k+1} = \frac{a_k + b_k}{2}$$

2. Localizar el intervalo en que la función cambia de signo:

$$F(a_k)F(m_{k+1}) < 0 \text{ o } F(m_{k+1})F(b_k) < 0$$

3. Seleccionar ese intervalo para la próxima iteración:

$$I_{k+1} = [a_{k+1}, b_{k+1}] = \begin{cases} [a_k, m_{k+1}] & \text{si} \quad F(a_k)F(m_{k+1}) < 0 \\ [m_{k+1}, b_k] & \text{si} \quad F(a_k)F(m_{k+1}) > 0 \end{cases}$$

Propiedades

1. Por construcción, se cumple que $F(a_k)F(b_k) < 0 \; \forall k$

2. Luego de N pasos, se cumple que: $\alpha \in [a_N, b_N]$ y además, la longitud del intervalo I_N es $\frac{b_0 - a_0}{2^N}$.

4.1.1. Ejemplo

Se sabe que la función $F(x) = x^3 + 4x^2 - 10$ tiene una sola raíz en el intervalo $I = [1, 2]$. En primer lugar, evaluamos la función en ambos extremos:

$$\begin{aligned} F(1) &= 1 + 4 - 10 = -5 \\ F(2) &= 8 + 16 - 10 = 14 \end{aligned}$$

A continuación, vamos operando siguiendo los pasos indicados:

a_k	b_k	$F(a_k)$	$F(b_k)$	m_{k+1}	$F(m_{k+1})$
1	2	-5	14	1.5	2.375
1	1.5	-5	2.375	1.25	-1.8
1.25	1.5	-1.8	2.375	1.375	0.16
1.25	1.375	-1.8	0.16	1.3125	-0.84
1.3125	1.375	-0.84	0.16	1.34375	-0.35
1.34375	1.375	-0.35	0.16	1.359375	-0.10
1.359375	1.375	-0.10	0.16	1.3671875	-0.03

4.1.2. Error de truncamiento

El error más común es el de truncamiento. Se puede dar el caso en el que se encuentre el valor exacto de α, pero muchas veces simplemente se continúa con la iteración hasta que una determinada cantidad de dígitos de la operación $F(m_{k+1})$ sea cero.

Si m_{k+1} es una estimación de α, tenemos que:

$$\begin{aligned} \alpha &= m_{k+1} \pm \Delta_{k+1} \\ \Delta_{k+1} &= \frac{b_k - a_k}{2} \\ \Delta_{k+1} &= \frac{b_0 - a_0}{2^{k+1}} \end{aligned} \tag{4.1.1}$$

Es importante tener en cuenta que la cota de error de truncamiento no depende de la función que se está analizando.

4.1.3. Error de redondeo

Puede suceder que al evaluar la función en alguno de los puntos, se cometa un error de redondeo, de modo que $\hat{F}(x_k) = F(x_k) \pm e_k$. Lo importante es poder garantizar que el signo de $\hat{F}(x_k)$ sea el mismo que el signo de $F(x_k)$. Para poder garantizarlo, e_k debe ser menor que mín$(|\hat{F}(a_k)|, |\hat{F}(a_k)|)$.

4.1.4. Convergencia

Analizando la ecuación (4.1.1), podemos observar que:

$$\Delta_{k+1} = \frac{b_0 - a_0}{2^{k+1}} \overset{k\to\infty}{\longrightarrow} 0 \tag{4.1.2}$$

Es decir, el método de la bisección converge incondicionalmente.

4.2. Método de Regula-Falsi

Este método es un poco más inteligente. En el caso del método de la bisección, puede suceder que se esté muy cerca de la raíz buscada, pero al siguiente paso se aleje mucho. El método de Regula-Falsi continúa acercándose a la raíz constantemente.

Para lograr esto, se utiliza una expresión bastante más compleja para el término m_{k+1}:

$$m_{k+1} = a_k - (b_k - a_k)\left(\frac{F(a_k)}{F(b_k) - F(a_k)}\right) \tag{4.2.1}$$

4.2.1. Ejemplo

Consideramos la misma función que se evaluó en el ejemplo del método de la bisección.

a_k	b_k	$F(a_k)$	$F(b_k)$	m_{k+1}	$F(m_{k+1})$
1	2	-5	14	1.263	-1.60
1.263	2	-1.60	14	1.34	-0.41
1.34	2	-0.41	14	1.358	-0.12
1.358	2	-0.12	14	1.363	-0.04
1.363	2	-0.04	14	1.365	-0.004
1.365	2	-0.004	14	1.3652	-0.0005
1.3652	2	-0.0005	14	1.36522	-0.0002

Podemos ver que en la misma cantidad de pasos, el método de Regula Falsi ha llegado mucho más cerca de la raíz que el método de la bisección.

4.2.2. Convergencia

Al igual que el método de la bisección, el método de Regula Falsi converge siempre, con $k \to \infty$.

4.2.3. Errores

De Truncamiento

Si $\partial f \neq 0 y \in C$, entonces podemos acotar el error de truncamiento con la siguiente inecuación:

$$|m_{k+1} - \alpha| \leq |m_{k+1} - m_k|$$

De Redondeo

Si llamamos E_{k+1} una cota del error absoluto producido a causa del redondeo, podemos introducir esta cota en la inecuación anterior:

$$|m_{k+1} - \alpha| \leq |m_{k+1} - m_k| + 2E_{k+1}$$

4.3. Método del punto fijo

Definimos una función $g(x) = x - qF(x)$. Y llamamos punto fijo de g, al punto β que satisface $g(\beta) = \beta$. Al analizar lo que significa que β se punto fijo de $g(x)$, llegamos a la conclusión de que β es raíz de $F(x)$:

$$\begin{aligned} g(\beta) &= \beta - qF(\beta) \\ \beta &= \beta - qF(\beta) \\ F(\beta) &= 0 \end{aligned}$$

Teorema:

$$g(x) \in C[a,b];\ g(x) \in [a,b]\ \forall x \in [a,b] \Rightarrow \exists \beta /\ g(\beta) = \beta \tag{4.3.1}$$

Además:

$$\exists g'(x) \in [a,b];\ \exists k/\ 0 < k < 1;\ |g'(x)| \leq k\ \forall x \in [a,b] \Rightarrow \beta \text{ es único} \tag{4.3.2}$$

La técnica que este método propone consiste en elegir una semilla, con alguno de los métodos de arranque, y luego iterar alrededor de esa semilla, tomándola como aproximación al punto fijo.

$$x_{k+1} = g(x_k) \tag{4.3.3}$$

4.3.1. Ejemplo

Se resuelve a continuación el mismo ejemplo elegido para los dos métodos anteriores, es decir, $F(x) = x^3 + 4x^2 - 10$, buscando una raíz en el intervalo $[1, 2]$.

Construcción de la función $g(x)$

La elección de la función $g(x)$ a utilizar es muy importante, ya que de esta función dependerá que el método converja rápidamente, lentamente, o no converja nunca.

Hay muchas formas de construir estas funciones. En primer lugar, podemos utilizar la fórmula: $g(x) = x - qF(x)$ y elegir algún valor determinado para q, como por ejemplo, $q = 1$. Por otro lado, podemos igualar $F(x) = 0$ y despejar x de alguna manera, hasta obtener una ecuación de la forma $x = g(x)$. Otra técnica que se puede aplicar cuando $F(x)$ se puede derivar, es hacer: $g(x) = x - \frac{F(x)}{F'(x)}$.

1. Caso más simple:

$$\begin{aligned} g_1(x) &= x - F(x) \\ g_1(x) &= -x^3 - 4x^2 + x + 10 \end{aligned}$$

2. Despejando:

$$\begin{aligned} x^3 &= 10 - 4x^2 \\ x^2 &= \frac{10}{x} - 4x^2 \\ g_2(x) &= \sqrt{\frac{10}{x} - 4x^2} \end{aligned}$$

3. Otro despeje:

$$\begin{aligned} 4x^2 &= 10 - x^3 \\ g_3(x) &= \sqrt{\frac{10 - x^3}{4}} \end{aligned}$$

4. Otro despeje posible:

$$\begin{aligned} x^2(x+4) &= 10 \\ x^2 &= \frac{10}{(x+4)} \\ g_4(x) &= \sqrt{\frac{10}{x+4}} \end{aligned}$$

5. Utilizando la derivada:

$$\begin{aligned} g_5(x) &= x - \frac{F(x)}{F'(x)} \\ g_5(x) &= x - \frac{x^3 + 4x^2 - 10}{3x^2 + 8x} \end{aligned}$$

Tablas de valores

Una vez que tenemos las funciones $g(x)$ que vamos a utilizar, tenemos que aplicarlas para ver cuál funciona mejor.

k	$g_1(x)$	$g_2(x)$	$g_3(x)$	$g_4(x)$	$g_5(x)$
0	1,5	1,5	1,5	1,5	1,5
1	−0,87	0,81	...	...	1,373
2	6,7	2,99	...	...	1,3634
3	−470	ERROR	...	...	1,3652
4	...		...	...	1,36523
5	-10^{25}		...	...	α
...	...		...	...	
11	...		...	α	
...	...		...		
31	...		α		

De los resultados podemos concluir que las dos primeras funciones elegidas no converjen, las siguientes dos converjen lentamente, y la última converje muy rápidamente.

4.3.2. Convergencia

Si se cumplen las condiciones indicadas en las ecuaciones (4.3.1) y (4.3.2), también se cumple que:

$$\forall x_0 \in [a,b]\ x_{k+1} = g(x_k) / \ \exists!\ \lim_{k\to\infty} g(x_k) = \beta \in [a,b]. \tag{4.3.4}$$

Al analizar la función $g_1(x)$, que en el ejercicio realizado no produjo una sucesión convergente, vemos que: $g_1'(x) = -3x^2 - 8x + 1$, de modo que $|g_1'(1)| = 10 \notin [1,2]$. Es decir que no la condición de que $g'(x) \in [a,b]\ \forall x \in [a,b]$, expresada por la ecuación (4.3.2).

Suele suceder cuando se utilizan los métodos no lineales, que haya intervalos para los que una función converge, y otros intervalos para los que no converge.

4.3.3. Corolarios para la elegir la función más veloz

Cota para el error absoluto

$$|x_n - \beta| \leq k^n \,\text{máx}(x_0 - a, b - x_0) \tag{4.3.5}$$

Con un k cercano a cero, converge muy rápidamente. Si, en cambio, k es cercano a uno, converge lentamente.

Orden de convergencia - Constante asintótica de error

Definimos P, orden de convergencia y C, constante asintótica de error, de forma que:

$$\lim_{k\to\infty} \frac{e_k + 1}{e_k^P} = C \tag{4.3.6}$$

Es decir que, en el límite, $\ln e_{k+1} = \ln C + p \ln e_k$.

Ahora, aplicamos estas definiciones al método del punto fijo.

$$\begin{aligned} x_{k+1} &= g(x_k) \\ x_{k+1} - \beta &= g(x_k) - \beta \\ x_{k+1} - \beta &= g(x_k) - g(\beta) \end{aligned}$$

Utilizando el desarrollo de Taylor alrededor de β, aplicado en x_k, $\xi \in (x_k, \beta)$, se obtiene:

$$\begin{aligned} g(x_k) &= g(\beta) + (x_k - \beta)g'(\xi) \\ x_{k+1} - \beta &= g(\beta) + (x_k - \beta)g'(\xi) - g(\beta) \\ x_{k+1} - \beta &= (x_k - \beta)g'(\xi) \\ e_{k+1} &= e_k g'(\xi) \\ \frac{e_{k+1}}{e_k} &= g'(\beta) \end{aligned}$$

De modo que el orden de convergencia de este método es $p = 1$, es decir que el error cometido se reduce linealmente en cada paso. Y la constante asintótica del error es $g'(\beta)$.

En el caso en que $g'(\beta) = 0$, será necesario tomar un término más de la serie de Taylor. Volviendo a aplicar los pasos anteriores:

$$\begin{aligned} g(x_k) &= g(\beta) + \frac{(x_k - \beta)^2}{2} g''(\xi) \\ x_{k+1} - \beta &= \frac{(x_k - \beta)^2}{2} g''(\xi) \\ \frac{e_{k+1}}{e_k^2} &= g''(\beta) \end{aligned}$$

De modo que, en este caso el orden de convergencia de este método es $p = 2$, y por lo tanto el error se reduce en forma cuadrática. Por otro lado, la la constante asintótica del error es $g''(\beta)$.

4.3.4. Error de truncamiento

Se analiza a continuación el error que se comete al truncar la sucesión en el término x_k.

$$\begin{aligned} x_{k+1} - \beta &= (x_k - \beta) g'(\xi) \\ x_{k+1} - \beta &= g'(\xi)(x_k - x_{k+1} + x_{k+1} - \beta) \\ x_{k+1} - \beta &= g'(\xi)(x_k - x_{k+1}) + g'(\xi)(x_{k+1} - \beta) \\ (1 - g'(\xi))(x_{k+1} - \beta) &= g'(\xi)(x_k - x_{k+1}) \\ x_{k+1} - \beta &= \frac{g'(\xi)}{1 - g'(\xi)} (x_k - x_{k+1}) \end{aligned}$$

Teniendo en cuenta las condiciones planteadas en la ecuación (4.3.2), su puede afirmar que $|g'(x)| \leq K$. Por lo tanto, se puede acotar el error de truncamiento con la siguiente expresión.

$$|x_{k+1} - \beta| \leq \frac{K}{1 - K} |x_{k+1} - x_k| \tag{4.3.7}$$

Es decir que, cuando más cercano a cero sea K, más rápida será la converengencia de la sucesión.

4.3.5. Error de redondeo

Llamamos δ_k al error de redondeo cometido en el paso k. Este valor es tal, que $|delta_k| \leq \delta$.

$$x_{k+1} - \beta = g(x_k) - g(\beta) + \delta_k \tag{4.3.8}$$

$$|x_{k+1} - \beta| \leq \frac{K}{1 - K} |x_{k+1} - x_k| + \frac{\delta}{1 - K} \tag{4.3.9}$$

$$\tag{4.3.10}$$

Es importante notar que el valor δ_k tiende a ser mucho menor que el error de truncamiento.

4.4. Método de Newton Raphson

Este método consiste en construir una buena función para utilizar con el método de punto fijo. De tal manera que se consiga orden de convergencia cuadrático.

Para llegar a obetener esa expresión, se analiza el desarrollo de Taylor de la función $F(x)$ alrededor del punto $\tilde{x}$, y luego alrededor del punto β.

$$\begin{aligned} F(\tilde{x}) &= F(x) + (\tilde{x} - x)F'(x) + \frac{1}{2}(\tilde{x} - x)^2 F''(\xi) \\ F(\beta) = 0 &= F(x) + (\beta - x)F'(x) + \frac{1}{2}(\beta - x)^2 F''(\xi) \end{aligned}$$

Tomando un x cercano a β, se puede eliminar el tercer término de la serie, dejándolo como cota del error de truncamiento cometido. De modo que podemos obtener una expresión para β.

$$\begin{aligned} -F(x) &= +(\beta - x)F'(x) + \\ x - \frac{F(x)}{F'(x)} &= \beta \end{aligned}$$

Así se llega a la expresión recursiva que en el ejercicio efectuado con el método de punto fijo logró una convergencia muy veloz.

$$x_{k+1} = x_k - \frac{F(x_k)}{F'(x_k)} \tag{4.4.1}$$

4.4.1. Condiciones de convergencia

Para que esta sucesión recursiva sea convergente, será necesario que se cumplan las siguientes condiciones.

1. $g(x) \in [a, b]\ \forall x \in [a, b]$
2. $F'(x) \neq 0 \in [a, b]$, ya que es el denominador de la función.
3. $\exists F''(x)$, pues es necesaria para que exista $g'(x)$.
4. $g'(x) = |\frac{F(x)F''(x)}{F'(x)^2}| < 1$

4.4.2. Orden de convergencia

Continuando con el análisis realizado para el método del punto fijo, se puede observar que, al aplicar β a $g'(x)$, se obtiene el valor $g'(\beta) = 0$, lo que nos indica

que el orden de convergencia es cuadrático.

$$\begin{aligned} g'(x) &= \frac{F(x)F''(x)}{F'(x)^2} \\ g'(\beta) &= \frac{F(\beta)F''(\beta)}{F'(\beta)^2} \\ g'(\beta) &= \frac{0 \times F''(\beta)}{F'(\beta)^2} = 0 \end{aligned}$$

4.5. Método de la secante

El método de la secante es similar al método de Newton Raphson, la diferencia reside en que en lugar de utilizar la derivada de la función, se la aproxima mediante diferencias. De esta forma, es posible realizar los cálculos mediante una computadora, sin tener que agregar programación especial para calcular la derivada.

La ecuación de la secante será:

$$F'(x_k) = \frac{F(x_k) - F(x_{k-1})}{x_k - x_{k-1}} \tag{4.5.1}$$

Y la ecuación de x_{k+1}, será:

$$x_{k+1} = x_k - \frac{F(x_k)(x_k - x_{k-1})}{F(x_k) - F(x_{k-1})} \tag{4.5.2}$$

La desventaja de este método es que se pierde el orden 2 de convergencia que se conseguía con el método de Newton Raphson. Es un método intermedio, que tiene un orden de convergencia $1 < p < 2$, llamado *super lineal*.

4.6. Resúmen de los distintos métodos

Método	Ventajas	Desventajas
Tablas / Gráficos	Inmediato	No tiene precisión
Bisección	Muy fácil de implementar. Siempre converge	Es lento
Regula-Falsi	Es más convergente	Es lento
Punto fijo	Fácil aplicación	Muchas condiciones
Newton-Raphson	Orden 2 ($p = 2$)	Se necesita mucha información y la derivada
Secante	No se calcula la derivada. Es muy rápido	Pierde el orden 2

Capítulo 5

Sistemas de ecuaciones no lineales

Los sistemas de ecuaciones no lineales que se analicen, serán de la forma:

$$F(\vec{x}) = 0 \Rightarrow \begin{cases} F_1(x_1, \ldots, x_N) &= 0 \\ F_2(x_1, \ldots, x_N) &= 0 \\ \ldots & \\ F_N(x_1, \ldots, x_N) &= 0 \end{cases}$$

5.1. Método del punto fijo

Para analizar estos sistemas, se utiliza un método análogo al de punto fijo, utilizado para analizar ecuaciones no lineales. En este caso, se define un sistema de ecuaciones $G(\vec{x}) = \vec{x} - F(\vec{x})$ y un vector fijo $\vec{\beta}$ tal que $G(\vec{\beta}) = \vec{\beta}$.

Teorema:

- Se define un dominio $D = \{x/\ a_i < x \leq b_i\ \forall 1 \leq i \leq N\}$.
- Si se tiene $G(\vec{x}) \in D \subset \Re^N$; tal que $G(\vec{x}) \in D\ \forall x \in D$
- Se puede afirmar
$$\Rightarrow \exists \beta/\ g(\beta) = \beta \tag{5.1.1}$$
- Si además, $\exists K < 1$ tal que $|\frac{\partial g_i(x)}{x_j}| < \frac{K}{N}\ \forall i, j$, la sucesión $a_{k+1} = G(\vec{x}_k)\ \forall x_0 \in D$ converge al único punto fijo $\vec{\beta} \in D$.
- Corolario.
$$\|\vec{x}_k - \vec{\beta}\|_\infty \leq \frac{K^k}{1-K}\|x_1 - x_0\|_\infty \tag{5.1.2}$$

5.1.1. Ejemplo

Para analizar el comportamiento de este método utilizaremos el siguiente sistema de ecuaciones no lineales.

$$\vec{F}(\vec{x}) = \begin{cases} x_1 - \frac{1}{3}\cos x_2 &= \frac{1}{3} \\ \frac{1}{20}e^{-x_1} + x_2 &= -\frac{1}{2} \end{cases}$$

Capítulo 6

Técnicas de Ajuste

Las técnicas de ajuste permiten tener una forma analiítica de una función, a partir de datos experimentales. Están basadas en la teoría lineal de aproximación.

$$f^*(x) = \sum_{j=0}^{N} C_j \phi_j(x) = C_0\phi_0(x) + C_1\phi_1(x) + \ldots + C_N\phi_N(x) \tag{6.0.1}$$

Donde C_j son $N+1$ incógnitas, y ϕ_j son $N+1$ funciones preestablecidas.

Se tienen $M+1$ datos, si se quiere que f^* pase por todos ellos, se formará un sistema de ecuaciones lineales, con $M+1$ ecuaciones y $N+1$ incógnitas. De modo que :

$$f^*(x_i) = \sum_{j=0}^{N} C_j \phi_j(x_i) = f^*(i); \tag{6.0.2}$$

En el caso en que $M < N$, el problema queda indeterminado. Si, en cambio, $M = N$, el problema suele tener una solución única (si ϕ_j son li), que se obtiene por interpolación o colocación.

Por último, si se diera el caso en que $M > N$, el problema está sobre determinado. Es necesario imponer algún criterio para poder resolverlo. Se denomina **ajuste** al criterio qeu se aplica a un problema sobre determinado.

6.1. Cuadrados mínimos

La idea del método de cuadrados mínimos consiste en tratar de minimizar los errores cometidos utilzando errores cuadráticos.

$$\begin{aligned} e_i &= f^*(x_i) - f_i \\ e &= \sum_{i=0}^{M} (f_i - f^*(x_i))^2 = \sum_{i=0}^{M} e_i^2 \end{aligned}$$

Para poder minimizar el error que se comete, es necesario que definir un juego de coeficientes C_k, que lleven al mínimo valor de e. Esto es posible lograrlo si $\frac{\partial e}{\partial C_k} = 0 \; \forall k = 0, \ldots, N$.

A continuación, se opera con estas ecuaciones en busca de los valores correspondientes a los coeficientes C_k.

$$\begin{aligned}
0 &= \frac{\partial}{\partial C_k}\left(\sum_{i=0}^{M}(f_i - f^*(x_i))^2\right) \\
&= \frac{\partial}{\partial C_k}\left(\sum_{i=0}^{M}\underbrace{\left[f_i - \left(\sum_{j=0}^{N} C_j\phi_j(x_i)\right)\right]^2}_{R_i}\right) \\
&= \frac{\partial}{\partial C_k}\left(\sum_{i=0}^{M} {R_i}^2\right) = \sum_{i=0}^{M}\frac{\partial}{\partial C_k}{R_i}^2 = \sum_{i=0}^{M} 2R_i\frac{\partial R_i}{\partial C_k} \\
&= \sum_{i=0}^{M} 2\left[f_i - \left(\sum_{j=0}^{N} C_j\phi_j(x_i)\right)\right]\frac{\partial}{\partial C_k}\left[f_i - \left(\sum_{j=0}^{N} C_j\phi_j(x_i)\right)\right] \\
&= \sum_{i=0}^{M} -2\left[f_i - \left(\sum_{j=0}^{N} C_j\phi_j(x_i)\right)\right]\phi_k(x_i) \\
&= -2\left(\sum_{i=0}^{M}[f_i\phi_k(x_i)] - \sum_{i=0}^{M}\sum_{j=0}^{N} C_j\phi_j(x_i)\phi_k(x_i)\right) \\
&= -2\left(\sum_{i=0}^{M}[f_i\phi_k(x_i)] - \sum_{j=0}^{N} C_j\left[\sum_{i=0}^{M}\phi_j(x_i)\phi_k(x_i)\right]\right) \\
&= -2\left(\vec{f}\cdot\vec{\phi_k} - \sum_{j=0}^{N} C_j\left(\vec{\phi_j}\cdot\vec{\phi_k}\right)\right)
\end{aligned}$$

Finalmente, llegamos a un sistema de ecuaciones comunes, que utiliza el producto interno de los vectores. Este sistema tendrá solución si las funciones ϕ_j son li.

$$\sum_{j=0}^{N} C_j\left(\vec{\phi_j}\cdot\vec{\phi_k}\right) = \left(\vec{f}\cdot\vec{\phi_k}\right);\ \forall k = 0,\ldots,N \tag{6.1.1}$$

6.1.1. Ejemplo

Se analizará el método de ajuste de cuadrados mínimos, teniendo la siguiente tabla de valores.

x	$-1{,}5$	-1	$-0{,}5$	0	$0{,}5$	1	$1{,}5$
$f(x)$	$-2{,}4$	$1{,}4$	$1{,}1$	$0{,}9$	$1{,}2$	$1{,}5$	$2{,}3$

La función f^* propuesta es: $f^*(x) = C_0 + C_1e^x + C_2e^{-x}$. De modo que $N = 2$ y $M = 6$. Para resolver el sistema, será necesario armar los vectores $\phi(x)$

$$\begin{array}{rcl|rcl} \phi_{0j} & = & 1 & \vec{\phi_0} & = & (1,1,1,1,1,1,1) \\ \phi_{1j} & = & e^{x_j} & \vec{\phi_0} & = & (e^{-1,5}, e^{-1}, e^{-0,5}, e^{0}, e^{0,5}, e^{1}, e^{1,5}) \\ \phi_{1j} & = & e^{-x_j} & \vec{\phi_0} & = & (e^{1,5}, e^{1}, e^{0,5}, e^{0}, e^{-0,5}, e^{-1}, e^{-1,5}) \end{array}$$

$$\begin{bmatrix} (\phi_0,\phi_0) & (\phi_1,\phi_0) & (\phi_2,\phi_0) \\ (\phi_0,\phi_1) & (\phi_1,\phi_1) & (\phi_2,\phi_1) \\ (\phi_0,\phi_2) & (\phi_1,\phi_2) & (\phi_2,\phi_2) \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} \begin{bmatrix} (f,\phi_0) \\ (f,\phi_1) \\ (f,\phi_2) \end{bmatrix}$$

Reemplazando los valores de los vectores en la matriz, obtenemos el sistema de ecuaciones que debemos resolver.

$$\begin{bmatrix} 7 & 11{,}047 & 11{,}047 \\ 11{,}047 & 31{,}749 & 7 \\ 11{,}047 & 7 & 31{,}749 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \end{bmatrix} \begin{bmatrix} 10{,}8 \\ 18{,}983 \\ 19{,}069 \end{bmatrix}$$

La solución a este sistema es: $C_0 = -0{,}0689$, $C_1 = 0{,}5089$, $C_2 = 0{,}5124$. De modo que la función final es: $f^*(x) = -0{,}0689 + 0{,}5089e^x + 0{,}5124e^{-x}$. Se puede corroborar que realmente esta es una buena función de ajuste para los datos dados, calculando la norma del error cometido.

$$\begin{array}{rcl} e_i & = & f_i - f^*(x_i) \\ \|e\| & = & 0{,}18 \\ \dfrac{\|e\|}{\|f\|} & \approx & 4\,\% \end{array}$$

6.1.2. Casos especiales

Cuando las funciones ϕ_j son ortogonales, la matriz obtenida es una matriz diagonal. Es decir que, los valores de C_k podrán obtenerse como $C_k = \frac{(f,\phi_k)}{(\phi_j,\phi_j)}$, si se cumple que:

$$(\phi_j,\phi_k) = \begin{cases} = & 0 \quad \forall j \neq k \\ \neq & 0 \quad \forall j = k \end{cases}$$

Si, además, el sistema es ortonormal ($(\phi_j,\phi_j) = 1$, se deduce que $C_k = (f,\phi_k)$.

6.1.3. Polinomios

Si se quiere tomar una $f^*(x) = Ax^P$, no es posible utilizar el método planteado, porque tanto A como P son incógnitas. Sin embargo, utilizando el logaritmo, es posible encontrar los valores.

$$\underbrace{\ln f^*(x_j)}_{y_j} = \underbrace{P}_{C_1}\underbrace{\ln x_j}_{z_j} + \underbrace{\ln A}_{C_0}$$

En este contexto, es posible resolver el problema con ajuste en escala lineal. Una vez realizado el ajuste, $A = e^{C_0}$ y $P = C_1$.

6.2. Ajuste de funciones a través de polinomios

Muchas veces se utiliza el ajuste a través de un polinomio para estudiar funciones más complejas, como el seno o el coseno, ya que el polinomio es mucho más sencillo de trabajar. En este caso, no se trata de ajustar 3 o más puntos, sino de ajustar la función completa.

Teorema de Weirstrass: el polinomio será de la forma $P_N(x) = \sum_{k=0}^{N} a_k x^k$, y la función que se quiera aproximar mediante el polinomio será $f(x) \in C[a,b]$. El ajuste se realizará de modo tal que el error $E(x) = f(x) - P(x)$, tienda a cero a medida que el grado N del polinomio tiende a ∞.

Es decir que cuánto mayor sea el grado del polinomio que se utiliza para aproximar la función, mejor será la aproximación que se realiza. Pero como se trabaja con problemas numéricos, N no puede ser ∞, de modo que se toman $N+1$ datos.

Para minimizar el error, tenemos en cuenta que la derivada en cada uno de los puntos debe anularse ($\frac{\partial E}{\partial a_j} = 0 \; \forall j$):

De modo que:

$$\begin{aligned}
E(a_0, \ldots, a_N) &= \int_a^b \left(f(x) - \sum_{k=0}^{N} a_K x^K \right)^2 dx \\
E &= \int_a^b f^2(x)dx - 2\int_a^b f(x) \sum_{k=0}^{N} a_K x^K dx + \int_a^b \left(\sum_{k=0}^{N} a_K x^K \right)^2 dx \\
\frac{\partial E}{\partial a_j} &= 0 - 2\int_a^b f(x)x^j dx + 2\int_a^b x^j \left(\sum_{k=0}^{N} a_K x^K \right) dx \\
0 &= -\int_a^b f(x)x^j dx + \sum_{k=0}^{N} a_K \left(\int_a^b x^{j+K} \right) dx
\end{aligned}$$

De modo que los coeficientes necesarios se pueden encontrar utilizando la siguiente ecuación.

$$\int_a^b f(x)x^j dx = \sum_{k=0}^{N} a_K \underbrace{\left(\int_a^b x^{j+K} \right)}_{H_{j+1,k+1}} dx \tag{6.2.1}$$

La matriz H es la matriz de Hilbert. No depende del problema, sino únicamente de los coeficientes a y b. Es una matriz mal condicionada.

$$H_{j+1,k+1} = \frac{b^{k+j+1} - a^{k+j+1}}{j+k+1} \tag{6.2.2}$$

6.2.1. Ejemplo

Se quiere determinar el polinomio $P_M(x)$, que coincida con la función $f(x_i) = f_i$, en $M+1$ puntos de una grilla. En primer lugar, se puede decir que el polinomio existe y es único. Es el polinomio que anula a la función error en ese intervalo.

$$\begin{aligned}
\exists!\, p(x) / \; p(x) = C_0 + C_1(x - x_0) + C_2(x - x_0)(x - x_1) \\
+ \ldots + C_M(x - x_0)(x - x_1)\ldots(x - x_{M-1})
\end{aligned}$$

Si $f(x) \in C^{N+1}[a,b]$, y $p(x) \in P_N(x)$, se puede afirmar que $p(x_i) = f(x_i)$. Y por lo tanto, el error estará dado por: $e(x) = f(x) - p(x)$, que puede ser acotado por la siguiente expresión.

$$e(x) = \frac{f^{(N+1)}(\xi)}{(M+1)!}(x - x_0)...(x - x_M) \tag{6.2.3}$$

Al término $\frac{f^{(N+1)}(\xi)}{(M+1)!}$ lo identificamos como $A(\xi)$. $\xi \in [a,b]$.

6.2.2. Polinomio de Taylor

Desarrolla la función alrededor de un punto $\tilde{x}$, no interpola $M+1$ puntos, sino solamente 1.

6.2.3. Interpolación de Lagrange

La Interpolación de Lagrange es una técnica alternativa de definir los coeficienes, teniendo en cuenta que $f(x_i) = f_i$.

$$p(x) = \underbrace{\left(\frac{x - x_1}{x_0 - x1}\right)}_{L_0} f_0 + \underbrace{\left(\frac{x - x_0}{x_1 - x0}\right)}_{L_1} f_1 = L_0 f_0 + L_1 f_1 \tag{6.2.4}$$

L_0 y L_1 son dos polinomios de Lagrange. Téngase en cuenta que no son coeficientes, sino polinomios en sí, que se combinan para formar el **polinomio interpolador de Lagrange**. Estos polinomios cumplen con las siguientes propiedades: $L_0(x_0) = L_1(x_1) = 1$ y $L_0(x_1) = L_1(x_0) = 0$.

Generalización para $N+1$ puntos

Es posible encontrar los polinomios de Lagrange para cualquier cantidad de puntos, utilizando la siguiente ecuación.

$$L_{Nk}(x) = \prod_{i=0; i\neq k}^{N} \frac{x - x_i}{x_k - x_i} \;\middle|\; L_{Nk}(x_j) = \begin{cases} 0 & j \neq k \\ 1 & j = k \end{cases} \tag{6.2.5}$$

Es decir que el polinomio se anula en todos los puntos excepto en el que $j = k$.

Definición

Sea f una función definida en $x_0, \ldots, x_N$, que conforman $N+1$ puntos distintos, $\Rightarrow \exists! p(x) \in P_M(x)$ con $M \leq N$, tal que $f(x_k) = p(x_k)$, para $k = 0, \ldots, N$. El polinomio interpolante con la técnica de Lagrange será:

$$p(x) = \sum_{k=0}^{N} f_k L_{Nk}(x) \tag{6.2.6}$$

Ejemplo

Se desea construir un polinomio interpolador para los siguientes datos.

k	0	1	2
x_k	0,25	0,5	0,75
f_k	$\frac{1}{\sqrt{2}}$	1	$\frac{1}{\sqrt{2}}$

Será necesario contar con tres polinomios de Lagrange para poder construir el polinomio interpolador.

$$\begin{aligned} L_0(x) &= \prod_{i=0;i\neq 0}^{2} \frac{x-x_i}{x_0-x_i} \\ L_0(x) &= \left(\frac{x-0{,}5}{0{,}25-0{,}5}\right)\left(\frac{x-0{,}75}{0{,}25-0{,}75}\right) \\ L_0(x) &= 8x^2-10x+3 \end{aligned}$$

$$\begin{aligned} L_1(x) &= \prod_{i=0;i\neq 1}^{2} \frac{x-x_i}{x_1-x_i} \\ L_1(x) &= \left(\frac{x-0{,}25}{0{,}5-0{,}25}\right)\left(\frac{x-0{,}75}{0{,}5-0{,}75}\right) \\ L_1(x) &= -16x^2+16x-3 \end{aligned}$$

$$\begin{aligned} L_2(x) &= \prod_{i=0;i\neq 2}^{2} \frac{x-x_i}{x_2-x_i} \\ L_2(x) &= \left(\frac{x-0{,}25}{0{,}75-0{,}25}\right)\left(\frac{x-0{,}5}{0{,}75-0{,}5}\right) \\ L_2(x) &= 8x^2-6x+1 \end{aligned}$$

Una vez obtenidos los tres polinomios se colocan en la ecuación (6.2.6), para obtener la ecuación general del polinomio interpolador de Lagrange.

$$\begin{aligned} p_L(x) &= L_0(x)f_0+L_1(x)f_1+L_2(x)f_2 \\ p_L(x) &= \frac{8x^2-10x+3}{\sqrt{2}}-16x^2+16x-3+\frac{8x^2-6x+1}{\sqrt{2}} \\ p_L(x) &= 16x^2(\frac{1}{\sqrt{2}}-1)+16x(1-\frac{1}{\sqrt{2}})+(\frac{4}{\sqrt{2}}-3) \end{aligned}$$

6.2.4. Interpolación de Newton

$$\begin{aligned} f(x) &= C_0 + C_1(x - x_0) + C_2(x - x_0)(x - x_1) \\ &\quad + \ldots + C_N(x - x_0)(x - x_1)\ldots(x - x_{M-1}) \\ &\quad + A(x)(x - x_0)(x - x_1)\ldots(x - x_{M-1}) \end{aligned}$$

$$\begin{aligned} A(x) &= \tfrac{d^{m+1}f(\xi)}{dx^{m+1}} \tfrac{1}{(M+1)!} \\ f(x_0) &= C_0 \end{aligned}$$

Notación de diferencia dividida

$$\begin{aligned} f[x_0, x] = \frac{f(x) - f(x_0)}{x - x_0} &= C_1 + C_2(x - x_1) + \ldots + \\ C_N(x - x_1)\ldots(x - x_{M-1}) &+ A(x)(x - x_1)\ldots(x - x_{M-1}) \end{aligned}$$

$$f[x_0, x_1] = C_1$$

$$f[x_0, x_1, x] = \frac{f[x_0, x_1] - C_1}{x - x_1} = C_2 + C_3(x - x_2) + \ldots$$

Generalización

$$f[x_0, x_1, \ldots, x_k, x] = \frac{f[x_0, x_1, \ldots, x_{k-1}] - f[x_0, x_1, \ldots, x_k]}{x - x_k}$$

$$p(x) = f_0 + \sum_{j=1}^{M} f[x_0, x_1, \ldots, x_j](x - x_0)(x - x_1)\ldots(x - x_{j-1})$$

$$R = f(x) - p(x) = f_0 + f[x_0, x_1, \ldots, x_M, x](x - x_0)(x - x_1)\ldots(x - x_M)$$

Ejemplo

Se desea realizar una interpolación de Newton con los siguientes datos.

k	0	1	2
x_k	0,25	0,5	0,75
f_k	$\frac{1}{\sqrt{2}}$	1	$\frac{1}{\sqrt{2}}$
$f[x_j, x_k]$		$\frac{1-\frac{1}{\sqrt{2}}}{0,5-0,25}$	$\frac{\frac{1}{\sqrt{2}}-1}{0,75-0,5}$

Es decir que , $f[x_0, x_1] = 4\left(1 - \frac{1}{\sqrt{2}}\right)$, y $f[x_1, x_2] = 4\left(\frac{1}{\sqrt{2}} - 1\right)$. La expresión para $f[x_0, x_1, x_2]$, en este caso, será:

$$f[x_0, x_1, x_2] = \frac{4\left(\frac{1}{\sqrt{2}} - 1\right) - 4\left(1 - \frac{1}{\sqrt{2}}\right)}{0,75 - 0,25} = -16\left(1 - \frac{1}{\sqrt{2}}\right)$$

Finalmente, el polinomio interpolante obtenido será:

$$p(x) = \frac{1}{\sqrt{2}} + 4\left(1 - \frac{1}{\sqrt{2}}\right)(x - 0{,}25) - 4\left(\frac{1}{\sqrt{2}} - 1\right)(x - 0{,}25)(x - 0{,}5)$$

6.2.5. Interpolación de Hermite

La interpolación de Hermite su utiliza cuando se posee información de f y de su derivada.

Sea $f \in C'[a,b]$, $x_0, \ldots, x_N \in [a,b]$, y $x_= \neq x_1 \neq \ldots \neq x_N$, $\Rightarrow \exists! P$ de mínimo grado, que coincide con f y f' en $x_0, \ldots, x_N$. Se trata de un polinomio cuyo grado es, a lo sumo, $2N+1$, y está dado por la siguiente ecuación.

$$H_{2M+1}(x) = \sum_{j=0}^{M} f_j H_{Mj}(x) + \sum_{j=0}^{M} f'_j \hat{H}_{Mj}(x) \tag{6.2.7}$$

Donde, $H_{Mj}(x)$ y $\hat{H}_{Mj}(x)$ están definidas por las siguientes ecuaciones.

$$\begin{aligned} H_{Mj}(x) &\equiv \left[1 - 2(x - x_j)L'_{Mj}(x_j)\right] L^2_{Mj}(x) \\ \hat{H}_{Mj}(x) &\equiv (x - x_j)L^2_{Mj}(x) \end{aligned}$$

Metodología

1. Construir el polinomio de Lagrange: $L_0, L_1, \ldots, L_N$.
2. Derivar $L'_0, L'_1, \ldots, L'_N$.
3. Construir H_{Mj}.
4. Construir $\hat{H}_{Mj}$.
5. Construir H_{2M+1}.

Ejemplo

En el siguiente ejemplo, $M = 1$, y el polinomio de Hermite que se busca es, por lo tanto, H_3.

x_i	f_i	$\frac{df}{dx}\big\vert_i$
0	0	1
1	3	6

1. $L_0 = \frac{x-1}{0-1} = 1 - x$ y $L_1 = \frac{x-0}{1-0} = x$
2. $L'_0 = -1$, $L'_1 = 1$.
3. $H_{10} = (1 - 2x(-1))(1-x)^2 = 2x^3 - 3x^2 + 4x - 1$.
   $H_{11} = (1 - 2(x-1))\,x^2 = -2x^3 + 3x^2$.

4. $\hat{H}_{10} = (x-0)(1-x)^2 = x^3 - 2x^2 + x$.
   $\hat{H}_{11} = (x-1)x^2 = x^3 - x^2$.

5. $H_3 = 0 \times H_{10}(x) + 3(-2x^3 + 3x^2) + 1(x^3 - 2x^2 + x) + 6(x^3 - x^2) = x^3 + x^2 + x$.

A continuación, se comprueba que el polinomio obtenido coincida verdaderamente con los valores tanto de la función como de la derivada. $H_3(0) = 0$ y $H_3(1) = 3$. $H_3'(0) = 1$ y $H_3(1) = 6$.

6.2.6. Fenómeno de Runge

El fenómeno de Runge ocurre en los casos en los que se calcula un polinomio de un grado alto (10, por ejemplo) de una función *buena*. Cerca del centro de la función se hace una buena aproximación, pero a medida que el polinomio se va alejando, la aproximación es cada vez peor, ya que aparecen ondas que se diferencian en gran medida de la función.

La solución a este problema es cambiar la escala de las abcisas, es decir, no utilizar puntos equidistantes, sino las abcisas de **Tcheby Cheff**. En un intervalo normalizado $I = [-1, 1]$ con una cantidad de M puntos, las abcisas equidistantes estarían dadas por $x_i = -1 + \frac{2i}{M}$, pero las abcisas de Tcheby Cheff serían:

$$x_i = \cos\left(\frac{2i+1}{M+1}\frac{\pi}{2}\right) \tag{6.2.8}$$

Una vez obtenidas las abcisas adecuadas, la interpolación se realiza utilizando el polinomio de Tcheby Cheff.

$$T_M(x) = \cos\left(M \arccos(x)\right) \tag{6.2.9}$$

Si $M = 0$, la ecuación (6.2.9) es equivalente a $T_0 = \cos(0) = 1$. Mientras que si $M = 1$, la misma ecuación será $T_1 = \cos(\arccos(x)) = x$. Finalmente, si $M = 2$, $T_2 = \cos(2 \arccos(x)) = 2x^2 - 1$

Interpolación de la función exponencial

Se desea realizar una interpolación de la función exponencial en el intervalo $[0, 1]$, con distintas metodologías. Se analiza el error que se comete en cada uno de los métodos.

- **Taylor**, es una buena aproximación alrededor del punto cero, pero es muy mala en los extremos.
- **Lagrange**, es una mejor aproximación, cuyo error vale cero en los puntos 0, 0,5 y 1, donde se tomaron las muestras.
- **Tcheby cheff**, el error no vale cero en los extremos, pero es el que presenta la banda de error más pequeña.

Capítulo 7

Integración

En este capítulo se estudian diversas técnicas de integración numérica. El problema a resolver será siempre de la forma $I = \int_a^b f(x)d(x)$, contará con $N+1$ puntos, que delimitarán N intervalos. Las operaciones que se realicen con esos puntos marcarán las diferencias entre una técnica y otra.

El primero de los puntos será $x_0 = a$, mientras que el punto final será $x_N = b$. Cada uno de los puntos comprendidos será $x_i = x_0 + iH$, donde $H = \frac{b-a}{N}$ es el ancho de cada uno de los intervalos.

7.1. Cuadratura Numérica

El método básico para calcular integrales definidas es el método de cuadratura. En este método, se multiplica la altura de la función en cada uno de los puntos por el ancho de cada uno de los intervalos.

$$\int_a^b f(x)dx \approx \sum_{i=0}^{N} Hf(x_i) \tag{7.1.1}$$

7.2. Rectángulos

En la técnica de los rectángulos, se toma el punto medio del intervalo como su altura y se lo multiplica por el intervalo para obtener el área de ese rectángulo.

$$\begin{aligned} I \approx R_H &= \sum_{i=1}^{N} H \times f(x_i - \frac{h}{2}) \\ R_H &= H \sum_{i=1}^{N} f_{i-\frac{h}{2}} \end{aligned}$$

7.2.1. Errores

El error absoluto que se comete al utilizar la aproximación de R_H, estará dado por las siguientes ecuaciones.

$$E_A = |R_H - I| = \sum_{i=1}^{N} \Delta I_i \leq n\Delta I_i$$

$$\Delta I_i = \int_{x_{i-1}}^{x_i} \left(f(x_i - \frac{H}{2}) - f(x) \right) dx$$

Para analizar el significado de estas expresiones, se desarrolla el polinomio de Taylor de la función $f(x)$ alrededor del punto $x_i - \frac{H}{2}$.

$$f(x) = f(x_i - \frac{H}{2}) + f'(x_i - \frac{H}{2})(x - (x_i - \frac{H}{2}) + \frac{f(\xi)}{2}(x - (x_i - \frac{H}{2}))^2$$

Luego se utiliza ese desarrollo en serie dentro de la integral anterior, retirando el término en ξ, para ser usado más adelante.

$$\begin{aligned}
\Delta I_i &= \int_{x_{i-1}}^{x_i} - \left(f'(x_i - \frac{H}{2})(x - (x_i - \frac{H}{2}) \right) dx \\
&= f'(x_i - \frac{H}{2})\frac{x^2}{2}\Big|_{x_{i-1}}^{x^i} - f'(x_i - \frac{H}{2})(x_i - \frac{H}{2})x\Big|_{x_{i-1}}^{x^i} \\
&= f'(x_i - \frac{H}{2})\frac{x_i^2}{2} - f'(x_i - \frac{H}{2})\frac{(x_{i-1})^2}{2} \\
&\quad -f'(x_i - \frac{H}{2})(x_i - \frac{H}{2})x_i + f'(x_i - \frac{H}{2})(x_i - \frac{H}{2})(x_{i-1}) \\
&= 0
\end{aligned}$$

Ahora, para acotar el error absoluto, se utiliza el término de la segunda derivada, que será la verdaderamente importante.

$$\begin{aligned}
\Delta I_i &= \int_{x_{i-1}}^{x_i} \frac{f''(\xi)}{2} \left(\overbrace{x - \underbrace{(x_i - \tfrac{H}{2})}_{\alpha}}^{\mu} \right)^2 dx \\
&\quad \alpha = x_i - \tfrac{H}{2};\ \mu = x - \alpha;\ x_i = \tfrac{H}{2};\ x_{i-1} = -\tfrac{H}{2} \\
&\quad |f''[a,b]| \leq M_2 \\
\Delta I_i &= \frac{f''(\xi)}{2} \int_{-\frac{H}{2}}^{\frac{H}{2}} u^2 du \\
\Delta I_i &= \frac{f''(\xi)}{2} \left(\frac{\frac{H}{2}^3}{3} - \frac{-\frac{H}{2}^3}{3} \right) \\
\Delta I_i &\leq \frac{M_2}{2}\frac{H^3}{12} \\
E_A = R_H - I &\leq n\Delta I_i \leq \frac{(b-a)}{H}\frac{M_2 H^3}{24} = \frac{(b-a)(M_2 H^2)}{24}
\end{aligned}$$

En conclusión, el error de truncamiento al utilizar el método de los rectángulos, es de orden 2. Mientras que el grado de precisión es 1, es decir que un polinomio de grado 1 puede ser acotado con precisión.

7.3. Trapecio

El método del trapecio utiliza los puntos de inicio y final de cada intervalo para obtener rectas interpolantes, de modo que el área que se calcula para cada intervalo no es el área de un rectángulo, sino el área de un trapecio.

Se deduce a continuación la fórmula para calcular la integral numérica utilizando este método. Partiendo del cálculo de cada una de las áreas como la suma de un rectángulo y un triángulo.

$$\begin{aligned} I \approx T_H &= \sum_{i=0}^{N} f(x_i)H + (f(x_{i+1}) - f(x_i))\tfrac{H}{2} \\ &= \sum_{i=0}^{N} H\left(f(x_i) + \frac{f(x_{i+1})}{2} - \frac{f(x_i)}{2}\right) \\ &= \sum_{i=0}^{N} \tfrac{H}{2}\left(f(x_i) + f(x_{i+1})\right) \\ I \approx T_H &= H\left[\frac{f_0 + f_N}{2} + \sum_{i=1}^{N-1} f_i\right] \end{aligned}$$

Es decir, en el último paso se han agrupado todas las apariciones duplicadas de $f(x_i)$, ya que solamente f_0 y f_N aparecen una sola vez en la sumatoria.

7.3.1. Error

El error en este método está acotado por la siguiente ecuación.

$$E_A \leq \frac{(b-a)}{12} M_2 H^2 \tag{7.3.1}$$

7.4. Simpson

Existen dos versiones del método del Simpson. La forma **simple** consiste en dos intervalos únicamente, la forma **compuesta** consiste en una cantidad par de intervalos.

La ecuación para el método simple será la siguiente.

$$I \approx S(h) = \frac{H}{3}\left[f_0 + 4f_1 + f_2\right] \tag{7.4.1}$$

La ecuación para el método complejo será la siguiente.

$$S(h) = \frac{H}{3}\left[f_0 + 4f_N + 2\sum_{i=1}^{\frac{N}{2}-1} f_{2i} + 4\sum_{i=1}^{\frac{N}{2}} f_{2i}\right] \tag{7.4.2}$$

7.4.1. Error

El error absoluto que se comete al utilizar este método está dado por la siguiente ecuación.

$$E_A \leq \frac{(b-a)}{180} M_4 H^4 \tag{7.4.3}$$

Es decir que con este método podemos integrar hasta una función cúbica sin error.

7.5. Extrapolación de Richardson

Llamamos $F(H)$ a una función que se utiliza para realizar la integración numérica, utilizando un paso de tamaño H.

Teniendo en cuenta que el paso que vamos a utilizar para los cáculos tiene un valor H, a medida que H se hace más pequeño, el esfuerzo de cálculo es mayor, y también son mayores los errores de redondeo.

El error que se comete en cada caso, es el que determina el orden del método. Es decir, $F = F_H + E_T$, donde $E_T = O(H^P)$. Siempre que se conozca el orden del método es posible utilizar la extrapolación de Richardson.

El efecto que produce la extrapolación de Richardson es correjir el error de truncamiento, teniendo en cuenta el orden.

En las ecuaciones siguientes, $H \to 0$, $R > P$, y se conoce el valor exacto de a_0.

$$\begin{aligned}
F(H_1) &= a_0 + a_1 H_1^P + O(H_1^R) \\
F(H_2) &= a_0 + a_1 H_2^P + O(H_2^R) \\
F(H_1) - F(H_2) &= a_1(H^P - H_2^P) + O(H_1^R) \\
a_1 &= \frac{F(H_1) - F(H_2)}{{H_1}^P - (H_2)^P} \\
F^* &= F(H_1) + \frac{F(H_1) - F(H_2)}{1 - \frac{H_2}{H_1}^P} + O(H_1^R)
\end{aligned}$$

Se requiere conocer el valor de P, el orden del método, la relación entre los dos valores H_1 y H_2 y las funciones $F(H_1)$ y $F(H_2)$.

La última ecuación es la que extrapola ambos valores y obtiene un valor corregido. Para poder hacer la extrapolación, se desprecia $O(H_1^R)$.

Una vez que se ha hecho la extrapolación, $F = F^* + O(H^R)$, es decir, se reduce el error de truncamiento.

La aplicación de la extrapolación de Richardson sucesivamente a valores obtenidos por trapecio se llama **Romberg**.

El error de truncamiento en el método del trapecio es de orden 2, al realizar una primera extrapolación, el orden es 4, y en una segunda extrapolación, el orden es 6. Los términos pares no aparecen utilizando el método de Romberg.

Para pasar de Trapecio (orden 2) a Simpson (orden 4), se utiliza la siguiente ecuación.

$$T^* = T_{H_1} - \frac{T_{H_1} - T_{H_2}}{1 - \frac{H_2}{H_1}^2} \tag{7.5.1}$$

Para avanzar de Simpson (orden 4) a Romberg 3 (orden 6), se utiliza la siguiente ecuación.

$$S^* = S_{H_1} - \frac{S_{H_1} - S_{H_2}}{1 - \frac{H_2}{H_1}^2} \tag{7.5.2}$$

Se puede generalizar esta fórmula, si los pasos siempre se van dividiendo por 2:

$$F^* = F_{H_1} - \frac{F_{H_1} - S_{H_2}}{1 - 2^P} \tag{7.5.3}$$

7.5.1. Ejemplo

Se obtienen en primer lugar los tres puntos por el método de trapecio.

$$T(h) = \frac{f_b + f_a}{2} H \tag{7.5.4}$$

$$T(\frac{h}{2}) = \frac{f_a + f_b + 2f_{a+\frac{h}{2}}}{2} H \tag{7.5.5}$$

$$T(\frac{h}{4}) = \frac{f_a + f_b + 2\left(f_{a+\frac{h}{4}} + f_{a+\frac{h}{2}} + f_{a+\frac{3h}{4}}\right)}{2} H \tag{7.5.6}$$

$$\tag{7.5.7}$$

Luego, a partir de esos tres puntos, se pueden obtener 2 puntos, mediante la extrapolación de Richardson, que son equivalentes a obtener esos puntos mediante Simpson.

A partir de esos dos valores, se calcula un nuevo valor, que pasa a ser de orden 6.

Este método permite obtener los valores deseados de una forma mucho más rápida que cualquier otro método.

7.5.2. Fórmula general

$$F(h) = a_0 + a_1 h^{p_1} + a_2 h^{p_2} + a_3 h^{p_3} + \ldots \tag{7.5.8}$$

Donde $p_1 < p_2 < p_3 < \ldots$.

$$F_{K+1}(H) = F_K(H) + \frac{F_K(H) - F_K(qH)}{q^{Pk} - 1};\ F_N(H) = a_0 + a_N h^{P_N} \tag{7.5.9}$$

7.5.3. Otro ejemplo

Se quiere evaluar la integral $I = \int_1^2 \frac{dx}{x}$. Con una cantidad T de dígitos significativos. Para poder evaluar la integral, se tomarán en total 5 puntos del intervalo $[1, 2]$.

x	1	1,25	1,5	1,75	2
$f(x)$	1	0,8	0,66	0,5714	0,5

Con estos valores, se encuentra en primer lugar los valores obtenidos mediante el método del trapecio, luego se extrapola a Simpson, y luego a Romberg 3.

h	$T(h)$	$S(h)$	$R_3(h)$
1	0,75		
0,5	0,7082	0,6945	
0,25	0,6970	0,6932	0,6931
0,125	0,6941	0,6931	0,6931

Efectivamente, $\ln 2 = 0{,}6931$.

7.6. Fórmulas de Newton Cotes

Las fórmulas de Newton Cotes pueden ser **cerradas** o **abiertas**.

7.6.1. Cerradas

La ecuación para las cerradas es la siguiente.

$$\int_a^b f(x)dx \approx \sum_{i=0}^{N} a_i f(x_i) \tag{7.6.1}$$

$$a_i = \int_{x_0}^{x_N} L_i(x)dx = \int_{x_0}^{x_N} \prod_{j=0}^{N} \frac{x - x_j}{x_i - x_j} dx \tag{7.6.2}$$

Al igual que en los casos anteriores, se trabaja con $N + 1$ puntos, donde $x_i = x_0 + iH$, $x_0 = a$, $x_N = b$, y $H = \frac{b-a}{N}$.

En el caso en que N sea par, la integral será exacta hasta polinomios de grado $N + 1$. Estará dada por la siguiente ecuación.

$$\int_a^b f(x)dx \approx \sum_{i=0}^{N} a_i f(x_i) + \frac{h^{(n+3)} f^{(n+2)}(\xi)}{(n+2)!} \int_0^N t^2(t-1)(t-2)\dots(t-N)dt \tag{7.6.3}$$

Mientras que si N es impar, la integral será exacta hasta polinomios de grado N, y estará dada por la siguiente ecuación.

$$\int_a^b f(x)dx \approx \sum_{i=0}^{N} a_i f(x_i) + \frac{h^{(n+2)} f^{(n+1)}(\xi)}{(n+1)!} \int_0^N t(t-1)(t-2)\dots(t-N)dt \tag{7.6.4}$$

En conclusión, no tiene sentido agregar un punto, es necesario agregar dos para que aumente la precisión.

7.6.2. Abiertas

Las premisas en este caso son distintas que las usadas anteriormente, si bien se trabaja con $N+1$ puntos, y $x_i = x_0+iH$, los valores de inicio y fin son: $x_0 = a+H$, $x_N = b-H$, y $H = \frac{b-a}{N+2}$.

Para el caso en que N es par, la integral estará dada por la siguiente ecuación.

$$\int_a^b f(x)dx \approx \sum_{i=0}^{N} a_i f(x_i) + \frac{h^{(n+3)} f^{(n+2)}(\xi)}{(n+2)!} \int_{-1}^{N+1} t^2(t-1)(t-2)\dots(t-N)dt \tag{7.6.5}$$

Los límites de la última integral indican que la precisión otorgada por esta ecuación es menor que la se consigue utilizando las fórmulas cerradas.

7.6.3. Relación con las otras fórmulas

Para las fórmulas cerradas, cuando $N = 1$, se obtiene la fórmula del trapecio, cuando $N = 2$ se obtiene la de Simpson, cuando $N = 3$, se obtiene la fórmula llamada **Simpson 3/8**.

$$I^* = \frac{3H}{8}[f_0 + 3f_i + 3f_2 + f_3] - \frac{3h^5}{80} f^{(4)}(\xi) \tag{7.6.6}$$

Para las fórmulas abiertas, cuando $N = 0$, se obtiene la fórmula llamada **del punto medio**.

$$I^* = 2Hf_0 + \frac{h^3}{3} f''(\xi) \tag{7.6.7}$$

Y en el caso $N = 1$, se obtiene la siguiente fórmula.

$$I^* = \frac{3H}{2}[f_0 f_1] + \frac{3h^3}{4} f''(\xi) \tag{7.6.8}$$

7.7. Métodos de cuadratura

Los métodos de cuadratura consisten en separar la función en dos partes, de forma que resulte más sencillo de calcular.

$$\int_a^b f(x)w(x)dx \tag{7.7.1}$$

7.7.1. Método de coeficientes indeterminados

En la siguiente ecuación, los términos f_0, f_1 y f_2 son valores conocidos mientras que los términos A_0, A_1 y A_2, son valores desconocidos. Se llama $H = x_1 - x_0$ y

$2H = x_2 - x_0$.

$$\begin{aligned}
\int_a^b f(x)dx &= \int_{x_0}^{x_2} f(x)dx \approx A_0 f_0 + A_1 f_1 + A_2 f_2 \\
\int_{x_0}^{x_2} f(x)dx &= \int_{x_0}^{x_2} f_0 dx + \int_{x_0}^{x_2} f_0'(x-x_0)dx \\
&\quad + \int_{x_0}^{x_2} \frac{f_0''(x-x_0)^2}{2!}dx + \int_{x_0}^{x_2} \frac{f_0'''(x-x_0)^3}{3!}dx + \ldots \\
A_0 f_0 + A_1 f_1 + A_2 f_2 &= A_0 f_0 + A_1 \left[f_0 + f_0' H + f_0'' \tfrac{H^2}{2} + \ldots \right] \\
&\quad + A_1 \left[f_0 + f_0' 2H + f_0'' \tfrac{(2H)^2}{2} + \ldots \right] \\
f_0 \int_{x_0}^{x_2} dx &= f_0(x_2 - x_0) = f_0 2H = f_0(A_0 + A_1 + A_2) \\
&\Rightarrow 2H = A_0 + A_1 + A_2 \\
f_0' \int_{x_0}^{x_2} (x-x_0)dx &= f_0' \left. \frac{(x_2-x_0)^2}{2} \right|_{x_0}^{x_2} = f_0' 2H^2 = f_0'(A_1 + 2A_2)H \\
&\Rightarrow 2H = A_1 + 2A_2 \\
f_0'' \int_{x_0}^{x_2} \frac{(x-x_0)^2}{2}dx &= f_0'' \left. \frac{(x_2-x_0)^3}{3} \right|_{x_0}^{x_2} = \frac{8H^3}{3} f_0'' = f_0''(\frac{H^2}{2}A_1 + \frac{2^2 H^2}{2}A_2) \\
&\Rightarrow \frac{4H}{3} = A_1 + 4A_2
\end{aligned}$$

Finalmente, se pueden obtener los valores de $A_0 = \frac{H}{3}$, $A_1 = \frac{4H}{3}$, $A_2 = \frac{H}{3}$. De modo que la integral obtenida es $\tilde{I} = \frac{H}{3}[f_0 + 4f_1 + f_2]$.

Se trata de un desarrollo **en adelanto**, porque se está centrando en el valor de x_0. Podría tambier desarrollarse **en atraso** o **centrado**.

Análisis del Error

Para f_0'''.

$$\begin{aligned}
D_1 &= \frac{f_0'''}{3!} \int_{x_0}^{x_2} (x-x_0)^3 dx \\
&= \frac{f_0'''}{3!} \left. \frac{(x-x_0)^4}{4} \right|_{x_0}^{x_2} \\
&= f_0''' \frac{2H^4}{3}
\end{aligned}$$

Por otro lado,

$$\begin{aligned}
D_2 &= \frac{f_0'''}{3!}(H^3 A_1 + 8H^3 A_2) \\
&= \frac{f_0'''}{3!}(\frac{4H^4}{3} + \frac{8H^4}{3}) \\
&= f_0''' \frac{2H^4}{3}
\end{aligned}$$

Es decir que no hay error, se avanza a la siguiente derivada, f_0''''.

$$\begin{aligned} D_3 &= \frac{f_0''''}{4!}\int_{x_0}^{x_2}(x-x_0)^4dx \\ &= \frac{f_0''''}{4!}\left.\frac{(x-x_0)^5}{5}\right|_{x_0}^{x_2} \\ &= f_0''''\frac{4H^5}{15} \end{aligned}$$

Por otro lado,

$$\begin{aligned} D_4 &= \frac{f_0''''}{4!}(H^4A_1+16H^4A_2) \\ &= \frac{f_0''''}{3!}(\frac{4H^5}{3}+\frac{16H^5}{3}) \\ &= f_0''''\frac{5H^5}{18} \end{aligned}$$

En este caso, son diferentes valores, y es posible proporcionar una cota para el error.

$$I-\tilde{I}=f''''\left(\frac{5}{18}-\frac{4}{15}\right)H^5=f''''\frac{1}{90}H^5 \tag{7.7.2}$$

Forma mecánica

Expresión de cuadratura.

$$\int_{x_0}^{x_2}f(x)=A_0f_0+A_1f_1+A_2f_2 \tag{7.7.3}$$

Si $f(x)=(x-x_0)^i$, con $i=0$:

$$\int_{x_0}^{x_2}f(x)dx=\int_{x_0}^{x_2}(x-x_0)^0dx=\int_{x_0}^{x_2}dx=A_0+A_1+A_2=2H \tag{7.7.4}$$

Con $i=1$:

$$\int_{x_0}^{x_2}(x-x_0)^1dx=\frac{4H^2}{2}=2H^2=A_1H+A_22H \tag{7.7.5}$$

Con $i=2$:

$$\int_{x_0}^{x_2}(x-x_0)^2dx=\frac{8H^3}{3}=A_1H^2+A_2(2H)^2 \tag{7.7.6}$$

7.7.2. Cuadratura de Gauss

El método de cuadratura gaussiana, utiliza una H que no es constante. Permite hacer una interpolación que utilizando una recta (grado1) puede integrar un polinomio de grado 2.

El método parte de la ecuación siguiente.

$$\int_a^b f(x)dx = \sum A_i \phi(\mu_i) \tag{7.7.7}$$

Donde A_i son los coeficientes, y μ_i son los puntos. Es necesario obtener valores para $2(N+1)$ incógnitas, ya que tanto los coeficientes como los puntos son incógnitas.

El método de elección de los puntos busca obtener una mayor precisión. La expresión hallada utilizando esos puntos será exacta hasta polinomios de grado $2(N+1)$.

La diferencia esencial con los otros métodos utilizados son los puntos que se van a usar para hacer la aproximación. Estos puntos son las raíces del polinomio de Legendre, del grado que corresponda a la cantidad de puntos.

En primer lugar, es necesario normalizar la función.

$$\int_a^b f(x)dx = \int_{-1}^1 \phi(\mu)d\mu = \sum A_i \phi(\mu_i) \tag{7.7.8}$$

En este caso, los coeficientes A_i pueden obtenerse mediante el método de coeficientes indeterminados, y μ_i son los ceros del polinomio de Legendre, que se encuentran tabulados.

Ejemplo para dos puntos

Se quiere aproximar la curva $\alpha_0 + \alpha_1\mu$, mediante la curva $\phi(\mu)$, de manera que el área encerrada por ambas curvas en el intervalo $[-1, 1]$ sea la misma.

$$\int_{-1}^1 \alpha_0 + \alpha_1\mu \approx= \int_{-1}^1 \phi(\mu)d\mu \tag{7.7.9}$$

Teniendo $I_G = A_0\phi(\mu_0) + A_1\phi(\mu_1)$. Tomo $\phi(\mu) = a_0 + a_1\mu + a_2\mu^2 + a_3\mu^3$, integrando cúbico. $\phi(\mu) = \alpha_0 + \alpha_1\mu + (\mu - \mu_0)(\mu - \mu_1)(\beta_0 + \beta_1\mu)$. De forma que $\phi(\mu_0) = y_0$ y $\phi(\mu_1) = y_1$.

Operando.

$$\begin{aligned}
\int_{-1}^1 \phi(\mu) &= \int_{-1}^1 \left[\alpha_0 + \alpha_1\mu + (\mu - \mu_0)(\mu - \mu_1)(\beta_0 + \beta_1\mu)\right] d\mu \\
&= \int_{-1}^1 (\alpha_0 + \alpha_1\mu)d\mu. \\
\text{y} \quad &\int_{-1}^1 (\mu - \mu_0)(\mu - \mu_1)(\beta_0 + \beta_1\mu)d\mu = 0 \\
\beta_0 &= \int_{-1}^1 (\mu - \mu_0)(\mu - \mu_1)d\mu) = \frac{2}{3} + 2\mu_0\mu_1 = 0 \\
\beta_1 &= \int_{-1}^1 (\mu - \mu_0)(\mu - \mu_1)\mu d\mu) = -\frac{2}{3} + (\mu_0 + \mu_1) = 0
\end{aligned}$$

Si $\phi(\mu) = 1$, se obtiene que $\int_{-1}^{1} d\mu = A_0 + A_1 \geq 2$, y por lo tanto $\mu_0 = -\frac{1}{sqrt3}$ y $\mu_1 = \frac{1}{sqrt3}$

Si, por otro lado, $\phi(\mu) = \mu$, se obtiene que $\int_{-1}^{1} \mu d\mu = A_0\mu_0 + A_1\mu_1 = 0$, y por lo tanto $A_0\left(-\frac{1}{sqrt3}\right) + A_1\left(\frac{1}{sqrt3}\right) = (A_1 - A_0)\frac{1}{sqrt3}$ y finalmente, $A_0 = A_1 = 1$.

En conclusión.

$$\tilde{I} = I_C = A_0\phi(\mu_0) + A_1\phi(\mu_1) = \phi(\frac{1}{sqrt3}) + \phi(-\frac{1}{sqrt3}) \tag{7.7.10}$$

Capítulo 8

Diferenciación Numérica

8.1. Derivadas de primer orden

8.1.1. Descentrada en adelanto

La fórmula más sencilla para el cálculo de una derivada de forma numérica está dada por:

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} \tag{8.1.1}$$

Cálculo del error

Para estimar el error cometido se utiliza el desarrollo de Taylor de $f(x)$ alrededor de x_0.

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(\xi)(x - x_0)^2}{2} \tag{8.1.2}$$

De manera que $f'(x_0)$, con $x = X_0 + h$ será:

$$f'(x_0) = \frac{f(x_0 + h) - f(x_0)}{h} + \frac{f''(\xi)h^2}{2h} \tag{8.1.3}$$

Es decir que al utilizar la estimación anterior, el error que se comete está dado por:

$$\frac{f''(\xi)h}{2} \tag{8.1.4}$$

Se trata de una fórmula con orden de precisión 1 y error de truncamiento $O(h)$.

8.1.2. Centrada

Otra fórmula para calcular la derivada en un punto puede obtenerse realizando el desarrollo de Taylor para dos puntos $x_0 + h$ y $x_0 - h$:

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{f''(x_0)h^2}{2} + \frac{f'''(\xi)h^3}{6} \quad (8.1.5)$$

$$f(x_0-h) = f(x_0) - f'(x_0)h + \frac{f''(x_0)h^2}{2} - \frac{f'''(\xi)h^3}{6} \quad (8.1.6)$$

Restando estas dos ecuaciones, se obtiene:

$$f'(x_0) = \frac{f(x_0+h) - f(x_0-h)}{2h} + 2\frac{f''(\xi)h^3}{12h} \quad (8.1.7)$$

De manera que en este caso, el orden de precisión es 2, mientras que el error de truncamiento es $O(h^2)$.

8.1.3. Descentrada, tomando 3 puntos

Se puede obtener una forma descentrada de la derivada de segundo orden, haciendo los desarrollos de Taylor en x_0-h y x_0-2h, y operando con las ecuaciones de tal manera que se anule el termino de la derivada segunda.

$$f(x_0-h) = f(x_0) - f'(x_0)h + \frac{f''(x_0)h^2}{2} - \frac{f'''(x_0)h^3}{6} \quad (8.1.8)$$

$$f(x_0-2h) = f(x_0) - f'(x_0)2h + 4\frac{f''(x_0)h^2}{2} - 8\frac{f'''(x_0)h^3}{6} \quad (8.1.9)$$

Si la primera ecuación se multiplica por 4 y se le resta la segunda, se obtiene:

$$4f(x_0-h) - f(x_0-2h) = 4f(x_0) - f(x_0) - 4f'(x_0)h + 2f'(x_0)h$$

$$f'(x_0) = \frac{3f(x_0) - 4f(x_0-h) + f(x_0-2h)}{2h} + 4\frac{f'''(x_0)h^3}{12h}$$

De manera que esta derivada tiene orden de precisión 2, y el error de truncamiento es $O(h^2)$.

8.2. Derivadas de segundo orden

8.2.1. Centrada

De la misma manera que antes, a partir del desarrollo de Taylor de la función evaluada en x_0+h y x_0-h, se puede obtener la fórmula para la derivada centrada.

$$f(x_0+h) = f(x_0) + f'(x_0)h + \frac{f''(x_0)h^2}{2} + \frac{f'''(x_0)h^3}{6} + \frac{f''''(\xi)h^4}{24} \quad (8.2.1)$$

$$f(x_0-h) = f(x_0) - f'(x_0)h + \frac{f''(x_0)h^2}{2} - \frac{f'''(x_0)h^3}{6} + \frac{f''''(\xi)h^4}{24} \quad (8.2.2)$$

En este caso, sumamos las ecuaciones, para obtener:

$$f(x_0+h)+f(x_0-h) = 2f(x_0)+2\frac{f''(x_0)h^2}{2}+\frac{f''''(\xi)h^4}{24} \quad (8.2.3)$$

$$f''(x_0)=\frac{f(x_0+h)-2f(x_0)+f(x_0-h)}{h^2}+\frac{f''''(\xi)h^4}{24h^2} \quad (8.2.4)$$

Con lo que el orden de precisión para esta estimación es 3, mientras que el error de truncamiento es $O(h^2)$.

Capítulo 9

Resolución numérica de ecuaciones diferenciales ordinarias

9.1. Problemas de Valores Iniciales

Se busca resolver problemas de la forma:

$$\frac{dy}{dt} = f(y,t) \quad y(0) = y_0 \tag{9.1.1}$$

Para resolver estos problemas, se discretizarán las funciones, de manera que la ecuación diferencial pasa a ser una ecuación en diferencias. La variable pasa a ser $t_n = hn$ y la función $y(t)$ se convierte en la sucesión $u_n = y(t_n)$.

9.1.1. Definiciones

Consistencia

Se dice que un método es consistente si el límite del máximo error local tiene a cero, cuando h tiende a cero.

La ecuacion en diferencias tiende a la ecuacion diferencial cuando el paso tiende a cero

Convergencia

Se dice que un método es convergente si el límite cuando h tiene a cero de $u_n - y(t_n)$ es cero.

La solucion de la ecuacion en diferencias tiende a la ecuacion diferencial.

Estabilidad

Esta situación aparece cuando los datos no son exactos. Un metodo es estable cuando las perturbaciones pequeñas en los valores iniciales, originan perturbaciones pequeñas en los resultados.

9.2. Método de Euler

El método de Euler es una discretización muy sencilla de una ecuación diferencial.

$$u_{n+1} = u_n + hf(t_n, x_n) \tag{9.2.1}$$

A partir de la condición inicial del problema, se pueden ir obteniendo los siguientes valores de u_n, de forma iterativa.

Cálculo del error

Al utilizar esta aproximación, se está descartando el siguiente término de la serie de Taylor, es decir:

$$\frac{y''(\xi)}{2}h^2 \tag{9.2.2}$$

Este término es el que corresponde al error **local**, debido a una sola iteración, y es $O(h^2)$. Sin embargo, después de m iteraciones iguales, el error acumulado será m veces mayor:

$$\sum_k = 1^m \frac{y''(\xi)}{2}h^2 = m\frac{y''(\xi)}{2}h^2 = \frac{b-a}{h}\frac{y''(\xi)}{2}h^2 = \frac{(b-a)y''(\xi)}{2}h \tag{9.2.3}$$

Es decir que el error **global** es $O(h)$.

9.2.1. Método de Taylor

El método de Taylor es una generalización del método de Euler, que permite resolver ecuaciones diferenciales de orden N (Euler sólo sirve para resolver ecuaciones diferenciales de orden 1).

9.2.2. Métodos implícitos

9.2.3. Métodos de Runge-Kutta

9.2.4. Métodos multi-paso

Los métodos multi-paso se refieren no solamente al valor obtenido para el paso anterior de la función (u_n), sino también a pasos previos (u_{n-1}, u_{n-2}, etc).

Es necesario contar con más de un valor inicial de la función. Para obtenerlos se suelen utilizar métodos de un sólo paso, como el método de Euler o el de Runge-Kutta.

Es importante tener en cuenta que si se utiliza un método multi-paso de orden N, se debe utilizar como método de arranque un método de un sólo paso, del mismo orden.

Los métodos multi-paso son los Adams-Bashforth, Adams-Moulton y el predictor-corrector de Milne.

9.2.5. Extrapolación de Richardson

9.2.6. Sistemas de ecuaciones

9.2.7. Problemas rígidos

Un problema rígido consiste en una ecuación diferencial cuya solución mediante métodos numéricos es inestable, a menos que se tome un paso extremadamente pequeño.

Por ejemplo, una ecuación diferencial de la forma:

$$\frac{dy}{dt} = -ky(t), \quad y(0) = y_0 \tag{9.2.4}$$

Cuya solución exacta es $y_0 e^{-kt}$.

Si esta ecuación se intenta resolver por el método de Euler, $u_{n+1} = u_n - hku_n$, exhibirá comportamientos oscilatorios, aún para valores del paso h pequeños.

Sin embargo, este mismo problema puede resolverse por otras técnicas, como por ejemplo la Adam-Bashforth, lográndose la solución esperada.

9.3. Problemas de Valores de Contorno

Los problemas de valores de contorno son aquellos en los cuales en lugar de tener las condiciones para el punto inicial de la ecuación diferencial, se tiene algún valor para el punto inicial y algún otro valor para el punto final.

9.3.1. Método directo centrado

9.3.2. Condiciones de contorno

9.3.3. Problemas de capa límite

Los problemas de capa límite son un caso especial de problemas de contorno que suelen aparecer en la física, especialmente en el estudio de la mecánica de los fluidos.

9.3.4. Refinamiento vs Upwinding

9.3.5. Método del tiro

El método del tiro se utiliza para transformar un problema de valores de contorno en un problema de valores iniciales. Consiste en proponer un valor ficticio como condición inicial del problema, luego resolverlo como un problema de valores iniciales y finalmente corregir el valor propuesto inicialmente para que las condiciones de contorno se cumplan.

La aproximación de S se puede ir mejorando de la siguiente manera:

$$S_j = S_{j-1} - \frac{[y(b, S_{j-1} - \beta](S_{j-1} - S_{j-2})}{y(b, S_{j-1}) - y(b, S_{j-2})} \tag{9.3.1}$$

Donde $y(b) = \beta$ e $y(a) = \alpha$. Utiliza el método de la secante, para obtener las raíces de la función $y(b, S)$.

9.4. Problemas de valores iniciales conservativos

9.4.1. Método de Taylor

9.4.2. Método de Newmark

9.4.3. Método de Nÿstrom