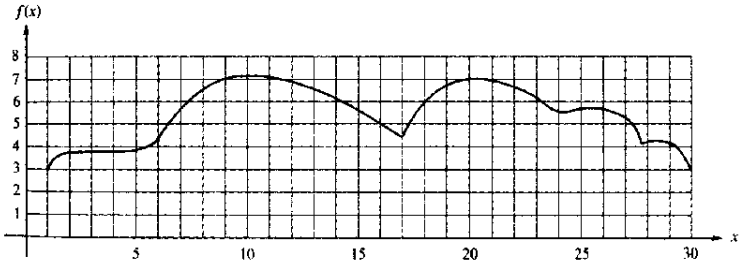
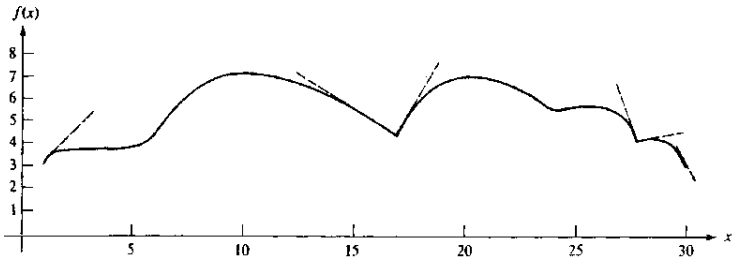
Departamento de Ingeniería Matemática
Facultad de Ciencias Físicas y Matemáticas
Universidad de Chile
Publicación:

INTRODUCCIÓN AL CÁLCULO NUMÉRICO

APUNTES PARA EL CURSO MA33A





Marzo 2000

María Leonor Varas S.

Presentación

Estos Apuntes han sido elaborados como apoyo al curso de Cálculo Numérico, que se dicta actualmente en la Escuela de Ingeniería y Ciencias de la Universidad de Chile. Se trata de un curso obligatorio, semestral, de Plan Común, que los alumnos siguen en segundo o tercer año.

La ubicación temprana de dicho curso en la malla curricular de la carrera de Ingeniería Civil tiene varias consecuencias. Por una parte, los estudiantes desconocen los problemas de ingeniería que requieren de métodos numéricos para su resolución y por lo tanto carecen de esta motivación. Por otra parte, cuentan con una sólida formación básica en matemáticas y conocimientos frescos que, en teoría, les permitirían seguir sin dificultad todas las demostraciones que aquí presentamos. La cantidad de materia incluida en el programa del curso y el breve lapso en que ésta debe ser cubierta, no permiten en la práctica un desarrollo tan detallado.

El Departamento de Ingeniería Matemática (DIM) ha decidido impulsar la creación de una Serie de Apuntes Docentes para los cursos de matemáticas de Plan Común, que faciliten el aprendizaje de estos tópicos y que recojan la riqueza diversa de una práctica extensa y seria. Este esfuerzo colectivo alberga y agrega valor al trabajo que aquí se presenta.

Especial mención merecen los aportes de Regina Mateluna y Eduardo Moreno. A Eduardo debemos el bello formato final, todos los gráficos, los ensayos computacionales y un entusiasmo contagioso que aligera la más pesada carga y convierte las dificultades en ocasión para el ingenio creativo. Regina puso su amable eficiencia, con el profesionalismo y compromiso que caracteriza a los funcionarios del DIM, en la transcripción de los manuscritos.

Otro de los motivos que originaron este Apunte, fue la incorporación de un Laboratorio Computacional asociado al curso. En él se desarrolla una actividad complementaria a la docencia en aula, es de carácter obligatorio, común a todas las secciones paralelas en que se imparte esta asignatura e incluye una evaluación. Se hizo así imprescindible aumentar la coordinación entre los profesores, homogeneizando el tratamiento de los contenidos previstos en el programa.

Esta primera versión del Apunte es apenas un esfuerzo de síntesis para iniciar un diálogo, sobre bases concretas, entre los profesores que habitualmente dictan el curso de Cálculo Numérico, lo que permite confiar en la aparición de mejores versiones posteriores a ésta.

María Leonor Varas S.
Santiago, Diciembre de 1999.

Índice General

Índice de Tablas

Índice de Figuras

CAPÍTULO 1

ERRORES

A lo largo de todo el curso de Cálculo Numérico hablaremos de errores. Todo el cálculo numérico se trata de que frente a la imposibilidad de conocer una cantidad o la solución a un problema tal cual ella es, condenados a errar, nos interesa saber cual es nuestro grado de imprecisión, o de que depende que éste aumente o disminuya.

Muchos de los métodos numéricos consisten en generar sucesiones que convergen a la solución del problema, cuando el número de términos de la sucesión tiende a *infinito*. En estos casos, estudiar el método obligatoriamente significa estudiar el error de no llegar jamás al límite, de *truncar* la sucesión en algún término. Este tipo de error, propio del método se llama **error de truncación**. Además de saber cuanto error se comete al truncar la sucesión en un término, tendremos que estudiar como evoluciona este error de truncación y por lo tanto deberemos definir de manera precisa la **velocidad de convergencia**, para saber con cuanto trabajo se mejora un resultado preliminar.

Definición 1.1. Sea $\{x_n\}$ una sucesión que converge a x, en un espacio vectorial normado. El error de truncación n-ésimo será

$$e_n = x - x_n.$$

Sea $\{\beta_n\}$ una sucesión de números reales positivos que converge a cero (típicamente $\beta_n = \frac{1}{n^p}$ para algún p entero positivo). Se dirá que la sucesión $\{x_n\}$ converge con una velocidad de convergencia caracterizada por un "o grande de β_n", $0(\beta_n)$, si existe una constante positiva K, tal que al menos a partir de cierto umbral, $\forall n \geq N_0$

$$\|e_n\| \leq K\beta_n.$$

Otros métodos son tan ambiciosos como para calcular la solución analítica directamente, como lo hace por ejemplo el conocido método de Gauss para resolver sistemas lineales (cuya solución aquí supondremos que existe y que es única). Estos métodos no están libres de error, pues como no podemos trabajar con números con infinitos decimales, habrá que redondearlos y al manipular (dividir, multiplicar, restar, etc...) estos números aproximados se producirá una **propagación de errores**. Hay situaciones (métodos y problemas) donde un pequeño error inicial se propaga mucho y se obtiene un resultado que difiere del resultado real mucho más que la diferencia inicial. Esto se llama **inestabilidad**. Un ejemplo de **problema inestable** es circular por las calles de Santiago en las horas de mayor tráfico. Si para llegar a un destino a las 8:30 AM. debemos salir a cierta hora de nuestra casa, sabemos que un atraso de 5 minutos en la hora de salida puede significar 30 minutos de retraso en la hora de llegada aún optimizando la ruta. En cambio si para

llegar desde San Bernardo a la Escuela de Ingeniería usamos la carretera Norte-Sur, tendremos un **método inestable**, pues de ocurrir un accidente en esta vía se producirá un atochamiento de difícil y lenta solución en comparación con otras rutas que tiene salidas alternativas.

Entre los problemas inestables uno de los más famosos son ciertos sistemas lineales llamados **mal condicionados**. La solución de un sistema de dos por dos, geométricamente corresponde a la intersección de dos rectas del plano. Si las dos rectas son paralelas o bien coinciden (habrá infinitas soluciones) o bien no se intersectan (no habrá solución del sistema). En ambos casos las dos filas de la matriz del sistema serán linealmente dependientes y por lo tanto la matriz no será invertible. Pero si estas dos rectas son casi paralelas (si la matriz es casi singular) entonces habrá solución única, pero moviendo muy poco una de las rectas, el punto de intersección (la solución) se desplazará mucho. Esto es lo que ocurre con el sistema del ejemplo que sigue.

Ejemplo 1. Consideremos el sistema

$$\begin{bmatrix} 1 & 2 \\ 0.499 & 1.001 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 1.5 \end{pmatrix}$$

cuya solución es $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$, si cambiamos ligeramente la matriz, el sistema

$$\begin{bmatrix} 1 & 2 \\ 0.5 & 1.001 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 1.5 \end{pmatrix}$$

tendrá por solución el vector $\begin{pmatrix} 3 \\ 0 \end{pmatrix}$, que difiere mucho más de la solución anterior que la perturbación sufrida por la matriz.

Para cuantificar estas perturbaciones y ser más precisos habría que tener normas matriciales relacionadas con las normas vectoriales con las cuales mediremos el desplazamiento de la solución. Esto será materia de otro capítulo posterior.

De los métodos inestables el más llamativo, por su inocente apariencia, es la resta de dos números parecidos. Para precisar el sentido de esta afirmación necesitaremos algunas definiciones acerca de como medimos el error cometido por números que aproximan a otros.

Definición 1.2. Si x^* es un número que aproxima al número x, se define el **error absoluto** cometido por x^*, como

$$E_A(x^*) = |x - x^*|$$

y el **error relativo** cometido por x^*, como

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

Si se conoce una aproximación x^* de un número desconocido x y se tiene una cota superior del error absoluto, es decir, un número ε tal que

$$E_A(x^*) \leq \varepsilon,$$

entonces se tiene un **intervalo de confianza** o intervalo de precisión para x,

$$x \in [x^* - \varepsilon, x^* + \varepsilon].$$

Recíprocamente, si se conoce un intervalo de confianza para x, $x \in [a,b]$, entonces $x^* = \frac{a+b}{2}$, es la mejor aproximación de x con que se cuenta y su error absoluto se acota por

$$E_A(x^*) \leq \frac{b-a}{2}.$$

En nuestros cálculos cotidianos acostumbrados a redondear cifras que nos parecen excesivas o irrelevantes. Así, por ejemplo un alumno que aparece en acta con nota 4.0 final pudo haber tenido un promedio desde 3.95 hasta 4.04, lo que redondeado a un decimal (o 2 cifras) se transforma en la nota del acta, cuyas dos cifras nos resultan significativas. Lo que entenderemos en este curso por **cifras significativas (o dígitos significativos)** se relaciona con el error relativo.

Supongamos que tenemos un número real x de n cifras escrito en "notación científica"como

$$x = s(0.a_1 a_2 \dots a_n) \times 10^{\alpha}$$

con signo s y dígitos $a_i \in \{0,1,2,\dots,9\}$ donde $a_1 \neq 0$. Si redondeamos este número a $p < n$ cifras obtendremos una aproximación x^* cuyo error se acota como

$$|x - x^*| \leq (0.0 \dots 05) \times 10^{\alpha} = \frac{1}{2} \cdot 10^{-p} \times 10^{\alpha}$$

Como $a_1 \geq 1$ entonces

$$|x| \geq 0.1 \times 10^{\alpha} = 10^{\alpha - 1}$$

Por consiguiente, el error relativo cometido por x^* quedará acotado por

$$E_R(x^*) = \frac{|x - x^*|}{|x|} \leq \frac{\frac{1}{2} \cdot 10\alpha - p}{10^{\alpha-1}} = \frac{1}{2} \cdot 10^{1-p}.$$

Definición 1.3. Un número x^* se dirá que tiene t dígitos significativos, si

$$E_R(x^*) \leq \frac{1}{2} \cdot 10^{1-t}.$$

Obviamente el número 10 que aparece en la cota se debe a que acostumbramos a usar una representación de los números en base 10 y que por lo tanto los dígitos que usamos son $0, 1, ..., 9$. Si representáramos los números en base 2, los únicos dígitos que requeriríamos serían 0 y 1.

Ejemplo 2. El número

$$1011.01 \text{ en base } 2,$$

corresponde a

$$1 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2^1 + 1 \cdot 2^0 + 0 \cdot 2^{-1} + 1 \cdot 2^{-2} = 11.25 \text{ en base } 10.$$

Una notación más clara, que identifique la base, sería

$$(1011.01)_2 = (11.25)_{10}.$$

Los errores en los números se originarán preferentemente en la forma en que los computadores los representan y esto no se refiere solo a una base.

Representación de punto flotante.

Cada computador usa una base β preestablecida y representa cada número con una cantidad finita de dígitos, p, de la forma

$$\pm 0.a_1 a_2 \ldots\ldots a_p \cdot \beta^{\alpha}$$

llamada representación de punto flotante.

Excepto que se trate del cero, los dígitos a_i pueden tomar valores entre 0 y $(\beta - 1)$, con $a_1 \neq 0$,cuando se quiere tener representación única la que se denomina "normalizada". El exponente α, también tendrá un recorrido limitado; será un entero, $m \leq \alpha \leq M$. De este modo, la precisión de un computador, quedará caracterizada por:

$$\text{(1.4)} \qquad \begin{cases} \text{La base, } \beta \\ \text{el máximo largo de la palabra, } p \\ \text{el rango del exponente, dado por } m \text{ y } M. \end{cases}$$

El conjunto de números representables mediante un computador con estas restricciones es un conjunto finito y cualquier número real que no pertenezca a este conjunto deberá ser aproximado por alguno del conjunto.

Ejemplo 3. Consideremos el conjunto de números representables por una máquina con

$$\begin{aligned} \beta &= 2 \\ p &= 3 \\ m &= -1 \\ M &= 1. \end{aligned}$$

Este conjunto será

$$A = \left\{ \begin{array}{cccccc} 0 & \pm 0.111 & \pm 0.110 & \pm 0.101 & \pm 0.100 & \pm 0.111 \cdot 2^{-1} \quad \pm 0.1110 \cdot 2^{-1} \\ & \pm 0.111 \cdot 2 & \pm 0.110 \cdot 2 & \pm 0.101 \cdot 2 & \pm 0.100 \cdot 2 & \pm 0.101 \cdot 2^{-1} \quad \pm 0.100 \cdot 2^{-1} \end{array} \right\}.$$

lo que corresponde a

$$A = \left\{ \begin{array}{cccccc} 0 & \pm\frac{7}{8} & \pm\frac{3}{4} & \pm\frac{5}{8} & \pm\frac{1}{2} & \pm\frac{7}{16} \quad \pm\frac{3}{8} \\ & \pm\frac{7}{4} & \pm\frac{3}{2} & \pm\frac{5}{4} & \pm 1 & \pm\frac{5}{16} \quad \pm\frac{1}{4} \end{array} \right\}$$

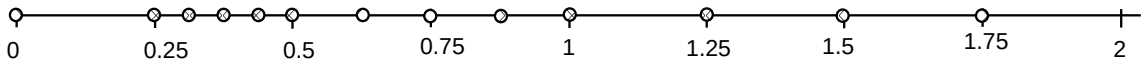


Figura 1.1: Representación de numeros positivos de A

Es bastante evidente que estos números no se distribuyen de manera uniforme en $\left[-\frac{7}{4}, \frac{7}{4}\right]$ y que si se quiere aproximar $\frac{13}{8}$ por un número de A, se cometerá un error mucho mayor que si se aproxima $\frac{9}{32}$ por un número representable por esta máquina.

Todo número cuya representación de punto flotante requiera de un exponente mayor que M, será considerado ∞ y el programa avisará que éste no es un número (Matlab avisa NaN). Todo número cuya representación de punto flotante requiera un exponente menor que m será considerado cero. Por ejemplo si se quiere invertir una matriz que tenga un valor propio con estas características, el programa dirá que esa matriz no es invertible. Para aproximar un número no representable (que no pertenece al conjunto finito A) debido a que su representación de punto flotante tiene más dígitos que p (el máximo largo de palabra admitido), por uno representable, el computador puede *truncar o redondear* a p dígitos. El error cometido será distinto, dependiendo si hace lo uno o lo otro.

Sea x el número no representable por largo excesivo

$$x = s \cdot 0.a_1 a_2 ... a_p a_{p+1} ... \cdot \beta^\alpha, \text{ con } m \leq \alpha \leq M \text{ y con } s \in \{-1, +1\}$$

y $fl(x)$ su representación de punto flotante de largo adecuado. Entonces el error cometido por $fl(x)$ será

(1.5)

- por redondeo

$$E_A(fl(x)) = |x - fl(x)| \leq \frac{1}{2}\beta^{\alpha-p}$$
$$E_R(fl(x)) = \frac{|x - fl(x)|}{|x|} \leq \frac{1}{2}\beta^{1-p},$$

- por truncación

$$E_A(fl(x)) \leq \beta^{\alpha-p}$$
$$E_R(fl(x)) \leq \beta^{1-p}.$$

Hacemos ver que para acotar el error relativo se debe recordar que $a_1 \geq 1$ y que por lo tanto $|x| \geq 0.1 \cdot \beta^\alpha = \beta^{\alpha-1}$.

A partir de estas relaciones también podríamos generalizar la definición de dígitos significativos de (1.3) a distintas bases y decir que un número redondeado a su representación de punto flotante de largo admisible, tiene todos sus p dígitos significativos.

Volvamos al problema de los métodos inestables. Supongamos que se quiere restar los números x e y de los cuales se tienen valores redondeados x^* e y^*.

Ejemplo 4. Sean

$$x = 1234566.5 \qquad x^* = 1234567$$
$$y = 1234568.4 \qquad x^* = 1234568.$$

El error relativo de cada una de estas aproximaciones es

$$E_R(x^*) = \frac{0.5}{1234566.5} = (0.4...) \cdot 10^{-6}$$
$$E_R(y^*) = \frac{0.4}{1234568.4} = (0.3...) \cdot 10^{-6}.$$

Es decir se trata de aproximaciones con 7 dígitos significativos, según la definición (1.3). Como

$$z = y - x = 1.9 \text{ y } z^* = y^* - x^* = 1,$$

entonces, el error relativo de la resta será $E_R(z^*) = \frac{|z-z^*|}{|z|} = \frac{0.9}{1.9} = 0.4...$, es decir z^* es una aproximación con apenas 1 dígito significativo. La pérdida de precisión que ha ocurrido es de 6 dígitos significativos o equivalentemente, corresponde a un aumento del error relativo de 10^6 veces.

Propagación de Errores.

Si se desea evaluar una expresión que depende de n números $x_1, x_2, ..., x_n$, que solo se conocen de manera aproximada como $x_1^*, x_2^*, ..., x_n^*$, entonces el error cometido por cada una de estas aproximaciones $\Delta_i = x_i - x_i^*$, contribuirá al error de la expresión final

$$z - z^* = \varphi(x_1, x_2, ..., x_n) - \varphi(x_1^*, x_2^*, ..., x_n^*).$$

Si las segundas derivadas parciales de φ son continuas entonces, desarrollando en serie de Taylor se tendrá que $\exists \xi$ en el segmento que une ambas n-tuplas, $x = (x_1, x_2, ..., x_n)^t$ y $x^* = (x_1^*, x_2^*, ..., x_n^*)^t$, tal que

$$z = z^* + \sum_{j=1}^{n} \frac{\partial \varphi}{\partial x_j}(x^*)\Delta_j + \sum_{i,j=1}^{n} \frac{\partial^2 \varphi}{\partial x_i \partial x_j}(\xi)\Delta_i \Delta_j.$$

Si los errores Δ_i son pequeños podemos despreciar los términos cuadráticos y estimar el error resultante como

$$z - z^* \approx \sum_{j=1}^{n} \frac{\partial \varphi}{\partial x_j}(x^*)\Delta_j. \tag{1.6}$$

En la evaluación de φ hay una segunda fuente de error que no se ha considerado en esta expresión. Por ejemplo si $\varphi(x_1, x_2) = x_1 + x_2$, entonces el computador aproximará la operación suma, pues aún si ambos argumentos son números representables (pertenecen al conjunto finito A) el resultado de la operación suma puede no serlo y deberá ser aproximada. Por lo tanto en lugar de obtenerse $z^* = \varphi(x_1^*, x_2^*, ..., x_n^*)$, se obtendrá $\tilde{z} = fl(\varphi(x^*))$.

Si consideramos el error producido por cada operación entonces deberemos preocuparnos del orden en que se hacen los cálculos, es decir del **algoritmo** usado para obtener el resultado de una expresión compleja (no elemental). Por ejemplo si se deben sumar tres números, la asociatividad de la suma **no** se cumplirá para la suma numérica o aproximada por el computador.

Ejemplo 5. Si $\varphi(x_1, x_2, x_3) = x_1 + x_2 + x_3$, con

$$\begin{aligned} x_1 &= 0.23371258 \cdot 10^{-4} \\ x_2 &= 0.33678429 \cdot 10^{2} \\ x_3 &= -0.33677811 \cdot 10^{2} \end{aligned}$$

entonces, con largo máximo de palabra $p = 8$ y redondeo se obtienen dos resultados distintos que se comparan con el valor exacto

$$\begin{aligned} fl(x_1 + fl(x_2 + x_3)) &= 0.64137126 \cdot 10^{-3} \\ fl(fl(x_1 + x_2) + x_3) &= 0.64100000 \cdot 10^{-3} \\ x_1 + x_2 + x_3 &= 0.641371258 \cdot 10^{-3}, \end{aligned}$$

mostrando la superioridad del primer algoritmo.

Para formalizar la influencia en el error del algoritmo utilizado, consideremos que la expresión φ se calcula a través de una sucesión de operaciones elementales

$$z = \varphi(x) = \varphi^{(r)}(\varphi^{(r-1)}(\ldots \varphi^{(0)}(x) \ldots)) = \varphi^{(r)} \circ \varphi^{(r-1)} \circ \ldots \circ \varphi^{(0)}(x).$$

Es evidente que la matriz Jacobiana de φ que participa en (1.6), $D\varphi(x^*)$, será el producto de las matrices jacobianas de $\varphi^{(k)}$, en el mismo orden en que se componen estas funciones.

Sean

$$x^{(k+1)} = \varphi^{(k)}(x^{(k)}), \forall k = 0, 1, ..., r, \text{ con } x^{(0)} = x.$$

El error aportado en cada etapa del cálculo será

$$\Delta x^{(k+1)} = fl(\varphi^{(k)}(\tilde{x}^{(k)})) - \varphi^{(k)}(x^{(k)}) = [fl(\varphi^{(k)}(\tilde{x}^{(k)})) - \varphi^{(k)}(\tilde{x}^{(k)})] + [\varphi^{(k)}(\tilde{x}^{(k)}) - \varphi^{(k)}(x^{(k)}]. \tag{1.7}$$

El segundo sumando se aproxima (despreciando los términos cuadráticos como hicimos para obtener (1.6)) por

$$D\varphi^{(k)}(x^{(k)})\Delta x^{(k)}.$$

Para el primer sumando recordaremos la expresión (1.5) para el error relativo de la representación de punto flotando de un número y, que permite escribir

$$fl(y) = y(1 + \varepsilon_y), \text{ con } |\varepsilon_y| = E_R(fl(y)).$$

Como $\varphi^{(k)}(\tilde{x}^{(k)})$, generalmente será un vector, su representación de punto flotante deberá satisfacer una ecuación matricial similar a la anterior,

$$fl(\varphi^{(k)}(\tilde{x}^{(k)})) = (I + E^{(k+1)})\varphi^{(k)}(\tilde{x}^{(k)}),$$

con lo cual en (1.7) se tendrá

$$\Delta x^{(k+1)} \approx E^{(k+1)}\varphi^{(k)}(\tilde{x}^{(k)}) + D\varphi^{(k)}(x^{(k)})\Delta x^{(k)}. \tag{1.8}$$

Despreciando términos cuadráticos del error podemos considerar a la matriz $E^{(k+1)}$ como diagonal y aproximar el primer sumando por

$$\alpha_{k+1} = E^{(k+1)}x^{(k+1)}.$$

De este modo el aporte de un paso del algoritmo a la propagación del error será

$$\Delta x^{(k+1)} = \alpha_{k+1} + D\varphi^{(k)}(x^{(k)})\Delta x^{(k)}, \tag{1.9}$$

y el error acumulado a lo largo de todo el algoritmo (despreciando nuevamente los términos cuadráticos y recordando que el Jacobiano de la composición es igual al producto de los Jacobianos) se aproxima como

$$\Delta z \approx \Delta x^{(r+1)} \approx D\varphi^{(r)} \cdot ... \cdot D\varphi^{(0)}\Delta x + D\varphi(r) \cdot ... \cdot D\varphi^{(1)}\alpha_1 + ... + \alpha_{r+1}. \tag{1.10}$$

Para ilustrar este esquema de análisis de la propagación del error revisaremos el ejemplo (5) de la suma, donde el algoritmo 1 se resume como

$$x = x^{(0)} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} \xrightarrow{\varphi^{(0)}} x^{(1)} = \begin{pmatrix} x_1 \\ x_2 + x_3 \end{pmatrix} = \begin{pmatrix} u \\ v \end{pmatrix} \xrightarrow{\varphi^{(1)}} x^{(2)} = u + v = z.$$

Por lo cual

$$D\varphi^{(0)} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \; D\varphi^{(1)} = [1, 1], \; E^{(1)} = \begin{bmatrix} 0 & 0 \\ 0 & \varepsilon_1 \end{bmatrix}, \; \alpha_1 = \begin{pmatrix} 0 \\ \varepsilon_1(x_2 + x_3) \end{pmatrix}, \; E^{(2)} = \varepsilon_2,$$
$$\alpha_2 = \varepsilon_2(x_1 + x_2 + x_3).$$

La propagación total del error será

$$\Delta z = [1,1,1]\begin{pmatrix}\Delta x_1\\ \Delta x_2\\ \Delta x_3\end{pmatrix} + [1,1]\begin{pmatrix}0\\ \varepsilon_1(x_2+x_3)\end{pmatrix} + \varepsilon_2(x_1+x_2+x_3).$$

El primer sumando corresponde a la propagación del error establecida en (1.6), es decir, aquella que no depende del algoritmo de cálculo, que es propia de la expresión o del problema. En cambio, el error relativo aportado por el algoritmo 1 será el módulo de

$$\frac{x_2+x_3}{x_1+x_2+x_3}\varepsilon_1 + \varepsilon_2.$$

Como todos los errores relativos usados satisfacen $|\varepsilon_i| \le \varepsilon$ (en caso de redondeo $\varepsilon = \frac{1}{2}\beta^{1-p}$, y en el caso del ejemplo $\beta = 10, p = 8$) y para el algoritmo 2 se deduce fácilmente una fórmula similar de propagación del error, cuyo aporte será el módulo de

$$\frac{x_1+x_2}{x_1+x_2+x_3}\varepsilon_3 + \varepsilon_4,$$

se concluye que la superioridad del algoritmo 1 sobre el algoritmo 2 se debe a que

$$\frac{|x_2+x_3|}{|x_1+x_2+x_3|} < \frac{|x_1+x_2|}{|x_1+x_2+x_3|}.$$

Problemas Resueltos.

1. Para calcular $z = a^2 - b^2$ se proponen dos algoritmos

$$\begin{array}{lll}\text{Algoritmo 1}: & \eta = a+b, & mu = a-b, \quad z = \eta\cdot\mu.\\ \text{Algoritmo 2}: & \eta = a\cdot a, & mu = b\cdot b, \quad z = \eta - \mu.\end{array}$$

Determinar para que valores de a y de b, el algoritmo 1 es superior al algoritmo 2.

Esquema del algoritmo 1

$$x^{(0)} = \begin{pmatrix}a\\ b\end{pmatrix} \xrightarrow{\varphi^{(0)}} x^{(1)} = \begin{pmatrix}a+b\\ a-b\end{pmatrix} = \begin{pmatrix}\eta\\ \mu\end{pmatrix} \xrightarrow{\varphi^{(1)}} x^{(2)} = \eta\cdot\mu = z.$$

Por lo tanto

$$D\varphi^{(0)} = \begin{bmatrix}1 & 1\\ 1 & -1\end{bmatrix},\ D\varphi^{(1)}(x^{(1)}) = [a-b, a+b],\ D\varphi = [2a, -2b],\ \alpha_1 = \begin{pmatrix}\varepsilon_1(a+b)\\ \varepsilon_2(a-b)\end{pmatrix}$$

$$\alpha_2 = \varepsilon_3 z$$

$$\Delta z = 2a\Delta a - 2b\Delta b + [a-b, a+b]\begin{pmatrix}\varepsilon_1(a+b)\\ \varepsilon_2(a-b)\end{pmatrix} + \varepsilon_3(a^2-b^2).$$

La contribución del algoritmo 1 a la propagación del error será

$$(a^2-b^2)(\varepsilon_1+\varepsilon_2+\varepsilon_3).$$

Esquema del algoritmo 2:

$$x^{(0)} = \begin{pmatrix}a\\ b\end{pmatrix} \xrightarrow{\varphi^{(0)}} x^{(1)} = \begin{pmatrix}a^2\\ b\end{pmatrix} = \xrightarrow{\varphi^{(1)}} x^{(2)} = \begin{pmatrix}a^2\\ b^2\end{pmatrix} \xrightarrow{\varphi^{(2)}} x^{(3)} = a^2 - b^2 = z.$$

Por lo tanto

$$D\varphi^{(0)}(x^{(0)})=\begin{bmatrix}2a & 0\\ 0 & 1\end{bmatrix},\ D\varphi^{(1)}(x^{(1)})=\begin{bmatrix}1 & 0\\ 0 & 2b\end{bmatrix},\ D\varphi^{(2)}(x^{(2)})=[1,-1],\ E^{(1)}=\begin{bmatrix}\varepsilon_1 & 0\\ 0 & 0\end{bmatrix},$$
$$E^{(2)}=\begin{bmatrix}0 & 0\\ 0 & \varepsilon_2\end{bmatrix},\quad E^{(3)}=\varepsilon_3,\quad \alpha_1=\begin{pmatrix}\varepsilon_1 a^2\\ 0\end{pmatrix},\quad \alpha_2=\begin{pmatrix}0\\ \varepsilon_2 b^2\end{pmatrix},\quad \alpha_3=\varepsilon_3(a^2-b^2),$$
$$\Delta z = 2a\Delta a - 2b\Delta b + [1,-1]\begin{bmatrix}1 & 0\\ 0 & 2b\end{bmatrix}\begin{pmatrix}\varepsilon_1 a^2\\ 0\end{pmatrix} + [1,-1]\begin{pmatrix}0\\ \varepsilon_2 b^2\end{pmatrix} + \varepsilon_3(a^2-b^2).$$

La contribución del algoritmo 2 a la propagación del error será

$$a^2\varepsilon_1 - b^2\varepsilon_2 + (a^2-b^2)\varepsilon_3.$$

Las cantidades a comparar serán entonces las cotas de los aportes en módulo

$$\begin{array}{ll}\text{del algoritmo 1} & 3\varepsilon|a^2-b^2|\\ \text{y del algoritmo 2} & \varepsilon(a^2+b^2+|a^2-b^2|).\end{array}$$

Si $\frac{1}{3} < \frac{a^2}{b^2} < 3$, lo que equivale a que $-a^2-b^2 < 2(a^2-b^2) < a^2+b^2$, entonces el algoritmo 1 será mejor que el algoritmo 2.

2. Un tipo especial de error de truncación se produce al reemplazar un dominio continuo por uno discreto. Por ejemplo, si en lugar del intervalo real $[a,b]$, dominio de una función, f, a valores reales, solo contamos con una malla

$$T=\{x_i\}_{i=0}^n \subset [a,b],\ \text{con } x_i = a+ih \text{ y } h=\frac{b-a}{n}.$$

Si se quiere calcular el valor de la derivada de f en un punto de la malla T, habrá que recurrir a una aproximación mediante alguna expresión que involucre solo a los nodos de la malla. proponemos dos alternativas

$$D_1 f(x_i) = \frac{f(x_{i+1})-f(x_i)}{h}$$
$$D_2 f(x_i) = \frac{f(x_{i+1})-f(x_{i-1})}{2h}.$$

Suponiendo que se trata de una discretización fina del dominio, es decir, que el paso h es pequeño, y que la función f es muy regular (con varias derivadas continuas), nos interesa saber cual de las dos alternativas comete un menor error de truncación.

Desarrollando en serie de Taylor en torno a x_i, se tendrá que

$$\exists \xi_1 \in [x_i, x_{i+1}],\ \text{tal que } f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2}f''(\xi_1)$$

y por lo tanto el error de truncación de la primera alternativa será

$$D_1 f(x_i) - f'(x_i) = \frac{h}{2}f''(\xi_1).$$

Por otra parte, si la regularidad de f lo permite, se tendrá que existen $\xi_2 \in [x_{i-1}, x_i]$ y $\xi_3 \in [x_i, x_{i+1}]$, tales que

$$f(x_{i-1}) = f(x_i) - hf'(x_i) + \frac{h^2}{2}f''(x_i) - \frac{h^3}{6}f'''(\xi_2)$$
$$f(x_{i+1}) = f(x_i) + hf'(x_i) + \frac{h^2}{2}f''(x_i) + \frac{h^3}{6}f'''(\xi_3).$$

Restando ambas expresiones y dividiendo por $2h$, se obtendrá

$$D_2 f(x_i) - f'(x_i) = \frac{h^2}{12}(f'''(\xi_3) + f'''(\xi_2)).$$

Estas expresiones son válidas si f''' es continua y en tal caso existirá un punto intermedio donde el valor de esta derivada tercera sea igual al promedio $f'''(\xi_4) = \frac{1}{2}(f'''(\xi_2) + f'''(\xi_3))$ y por lo tanto el error de truncación en este caso será

$$D_2 f(x_i) - f'(x_i) = \frac{h^2}{6} f'''(\xi_4),$$

que será menor que el error de truncación de la primera alternativa si $\|f'''\|_\infty \approx \|f''\|_\infty$ sobre el dominio $[a, b]$.

3. Sabemos que restar cantidades muy parecidas propaga mucho el error. Una manera de evitar este problema es transformar la expresión original en otra matemáticamente equivalente, pero numéricamente más estable. Si la expresión original es

$$F(x) = G(x) - H(x)$$

y la evaluación se realiza en un punto x donde las funciones G y H toman valores parecidos y del mismo signo, puede ser útil calcular

$$F(x) = \frac{G^2(x) - H^2(x)}{G(x) + H(x)}.$$

Consideremos el problema de evaluar $F(x) = 1 - cos(x)$ en x cerca de cero.

La técnica anterior nos llevará a evaluar $F(x) = \frac{sen^2(x)}{1+cos(x)}$, que es completamente estable.

4. Se quiere evaluar $F(x) = sen(x) - cos(x)$ en x cercano a $\frac{\pi}{4}$, la técnica del problema anterior no nos ayudará a encontrar una expresión estable. Se propone en cambio desarrollar ambas funciones $(sen(x), cos(x))$ en serie de Taylor, en torno a $\frac{\pi}{4}$, pagar un costo aceptable en errores de truncación y ganar en estabilidad.

$$sen(x) = \frac{\sqrt{2}}{2} + \frac{\sqrt{2}}{4}(x - \frac{\pi}{4}) - \frac{\sqrt{2}}{4}(x - \frac{\pi}{4})^2 - \frac{\sqrt{2}}{12}(x - \frac{\pi}{4})^3 + ...$$
$$cos(x) = \frac{\sqrt{2}}{2} - \frac{\sqrt{2}}{2}(x - \frac{\pi}{4}) - \frac{\sqrt{2}}{4}(x - \frac{\pi}{4})^2 + \frac{\sqrt{2}}{12}(x - \frac{\pi}{4})^3 + ...$$

Si se calcula $\widetilde{F}(x) = \sqrt{2}(x - \frac{\pi}{4}) - \frac{\sqrt{2}}{6}(x - \frac{\pi}{4})^3 + \frac{\sqrt{2}}{120}(x - \frac{\pi}{4})^5$ en lugar de $F(x)$, se cometerá un error de truncación de

$$\frac{\sqrt{2}}{5040}(\xi - \frac{\pi}{4})^7$$

para algún ξ entre x y $\frac{\pi}{4}$, que será menor en módulo que 10^{-10} si $|x - \frac{\pi}{4}| < 0.1199$, a cambio de la mayor estabilidad de la fórmula obtenida.

Ejercicios Propuestos.

1. De los números x, y, z se posee la siguiente información

$$x \in [100, 100.5],\ y^* = 0.0234 \quad E_A(y^*) \leq 0.0006, \quad z^* = 2.45 \quad E_R(z^*) \leq 0.03.$$

¿Cuál de estas cantidades se conoce con mayor precisión, es decir, con más dígitos significativos?

2. La pérdida de carga friccional en una tubería de diámetro D mt. y largo L mt. está dada por

$$h_f = 8f\frac{L}{D^3}\frac{Q^2}{\pi^2 g},$$

donde f denota el factor de fricción, Q es el caudal medido en m^3/seg. Considere los siguientes intervalos de precisión

$$I(g) = [9.8, 9.9]m/seg, \quad I(L) = [100.5, 110.5]mt, \quad I(D) = [25.5, 27.5]cm.,$$
$$I(Q) = [80, 90]lt/seg.$$

Considerando que los valores $\pi = 3.1415$, $f = 0.025$ son exactos, determine cuantos dígitos significativos deben tener las aproximaciones de todas las variables para que h_f se pueda conocer con al menos 5 dígitos significativos. Suponga igual distribución del error y atienda a las unidades. No considere la propagación del error debida al algoritmo.

3. Obtenga expresiones estables para evaluar las cantidades

$$\frac{1}{1+2x} - \frac{1-x}{1+x} \quad \text{para} \quad |x| << 1,$$

$$\sqrt{x+\frac{1}{x}} - \sqrt{x-\frac{1}{x}} \quad \text{para} \quad x >> 1,$$

$$\frac{1-cos(x)}{x} \quad \text{para} \quad |x| << 1.$$

4. Sean a, b, c tres números positivos menores que uno, con N decimales. Se define una operación *producto sustituto*

$$a^* b$$

como sigue:

sume $\frac{1}{2}10^{-N}$ *al producto exacto de a por b y luego borre las posiciones decimales desde la* $(N+1)$*-ésima en adelante.*

   (a) Obtenga una cota para $|(a^*b)^*c - abc|$.

   (b) ¿En cuántas unidades del N-ésimo decimal pueden diferir $(a^*b)^*c$ y $a^*(b^*c)$?

5. La función $tg\left(\frac{z}{2}\right)$ se puede calcular como $tg\left(\frac{z}{2}\right) = \pm\left(\frac{1-\cos(z)}{1+\cos(z)}\right)^{\frac{1}{2}}$.

Será estable esta fórmula para $z \approx 0, z \approx \frac{\pi}{2}$? De ser necesario obtenga una alternativa estable en cada caso.

6. Suponga que tiene un programa que entrega el valor de *arcsen*, en representación de punto flotante con largo de palabra t y para argumentos que sean en módulo menores que uno, comete un error ε, con $|\varepsilon| \leq 0.5 \cdot 10^{1-t}$. Dada la relación

$$arctg(x) = arcsen\left(\frac{x}{\sqrt{1+x^2}}\right)$$

que permite calcular *arctg* usando el mencionado programa, interesa saber para cuales valores de x esta fórmula será estable, estimando el error relativo.

7. Para calcular la varianza de un conjunto de n observaciones se proponen dos fórmulas:

$$S^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right), \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2, \text{ con } \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

¿Cuál de las dos fórmulas será más confiable?

CAPÍTULO 2

APROXIMACIÓN DE FUNCIONES

El problema que abordaremos en este capítulo es el de reconstruir una función f definida sobre un dominio real y a valores en $\mathbb{R}$, a partir de información incompleta o bien contaminada por errores. En tales circunstancias la reconstrucción no podrá ser perfecta y por lo tanto se tratará de una *aproximación* de la función f. En cursos de matemáticas previos a éste se ha resuelto este problema, en distintos contextos. Algunas de estas soluciones conocidas son:

- desarrollo en serie de Taylor en torno a un punto dado;
- desarrollo en serie de Fourier;
- aproximación polinomial de una función continua sobre un intervalo cerrado, según el teorema de Stone-Weierstrass.

Los polinomios serán nuestra principal herramienta y por lo tanto recordaremos este importante teorema y su demostración. Este dice que el conjunto de los polinomios es *denso* en el conjunto de las funciones continuas, lo que equivale a decir que toda función continua puede ser considerada como el límite de una sucesión de polinomios, en la norma de la convergencia uniforme.

Teorema 2.1 (Stone-Weierstrass). *Sea $f : [a,b] \to \mathbb{R}$ continua. $\forall \varepsilon > 0$ $\exists p$, polinomio, tal que*

$$\forall x \in [a,b] \qquad |f(x) - p(x)| \leq \varepsilon.$$

Demostración. Sin pérdida de generalidad se puede suponer que $[a,b] = [0,1]$ (si no es así, mediante el cambio de variables $u = \frac{x-a}{b-a}$, se obtiene el caso propuesto). Definamos el polinomio de Bernstein de grado n asociado a f como

$$B_n(f;x) = \sum_{k=0}^{n} \binom{n}{k} f(\frac{k}{n}) x^k (1-x)^{n-k}.$$

Notemos que para cada x, $B_n(f;x)$ corresponde a un promedio ponderado de los valores de f sobre la malla equiespaciada de $(n+1)$ puntos en $[0,1]$:

$$\{x_k\}_{k=0}^n = \{\frac{k}{n}\}_{k=0}^n = \{0, \frac{1}{n}, \frac{2}{n}, ..., \frac{n-1}{n}, 1\},$$

donde los ponderadores son $\lambda_k(x) = \binom{n}{k} x^k (1-x)^{n-k}$ y satisfacen

$$0 \leq \lambda_k(x) \leq 1 \qquad \forall k = 0, 1, ..., n \qquad \forall x \in [0,1] \tag{2.2a}$$

$$\sum_{k=0}^{n} \lambda_k(x) = 1 \qquad \forall x \in [0,1] \tag{2.2b}$$

y son mayores cuando $\frac{k}{n}$ está más cerca de x.

Separando los extremos de la sumatoria se obtiene

$$B_n(f;0) = f(0) \quad \text{y} \quad B_n(f;1) = f(1),$$

es decir, el polinomio de grado n, $B_n(f;x)$, coincide con f (o interpola a f) en los extremos del intervalo.

(Un ejercicio ilustrativo consiste en graficar estos polinomios para distintos n y con f una función de forma interesante como por ejemplo $f(x) = \frac{1}{x^2+1}$ en [-5,5] como se aprecia en la figura 2.1).

Como f es continua, se tendrá que es acotada y uniformemente continua en $[0,1]$ y por lo tanto para todo $\varepsilon > 0$ $\exists \delta > 0$ tal que

$$\forall x \in [0,1] \qquad \forall \frac{k}{n} \text{ tal que } |\frac{k}{n} - x| < \delta,\ |f(\frac{k}{n}) - f(x)| \leq \frac{\varepsilon}{2}$$

y por lo tanto, usando (2.2b)

$$\begin{aligned} |f(x) - B_n(f;x)| \leq & \sum_{\substack{k=0 \\ |\frac{k}{n}-x|<\delta}}^{n} |f(\frac{k}{n}) - f(x)| \binom{n}{k} x^k (1-x)^{n-k} \\ & + \sum_{\substack{k=0 \\ |\frac{k}{n}-x|\geq\delta}}^{n} |f(\frac{k}{n}) - f(x)| \binom{n}{k} x^k (1-x)^{n-k} \\ & \leq \frac{\varepsilon}{2} \sum_{\substack{k=0 \\ |\frac{k}{n}-x|<\delta}}^{n} \binom{n}{k} x^k (1-x)^{n-k} \\ & + 2M \sum_{\substack{k=0 \\ |\frac{k}{n}-x|\geq\delta}}^{n} \binom{n}{k} x^k (1-x)^{n-k} \end{aligned} \tag{2.3}$$

con $M = \max_{t \in [0,1]} |f(t)|$.

Para el primer sumando se tiene por (2.2b)

$$\sum_{\substack{k=0 \\ |\frac{k}{n}-x|<\delta}}^{n} \binom{n}{k} x^k (1-x)^{n-k} \leq \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} = 1. \tag{2.4}$$

Para acotar el segundo sumando debemos trabajar más. Notemos que $|\frac{k}{n} - x| \geq \delta \Rightarrow (k - nx)^2 \geq n^2\delta^2$ y que se puede demostrar sin mayor dificultad que

$$\sum_{k=0}^{n}(k-nx)^2\binom{n}{k}x^k(1-x)^{n-k} = xn(1-x).$$

Con lo cual podemos acotar el segundo sumando

$$\begin{aligned}\sum_{\substack{k=0\\ |\frac{k}{n}-x|\geq\delta}}^{n} 1\cdot\binom{n}{k}x^k(1-x)^{n-k} &\leq \sum_{\substack{k=0\\ |\frac{k}{n}-x|\geq\delta}}^{n}\frac{(k-nx)^2}{n^2\delta^2}\binom{n}{k}x^k(1-x)^{n-k}\\ &\leq \frac{1}{n^2\delta^2}\sum_{k=0}^{n}(k-nx)^2\binom{n}{k}x^k(1-x)^{n-k}\\ &\leq \frac{1}{n\delta^2}x(1-x)\\ &\leq \frac{1}{4n\delta^2}\quad \text{si } x\in[0,1].\end{aligned} \tag{2.5}$$

Usando (2.4) y (2.5) en (2.3) se obtiene que

$$|f(x)-B_n(f;x)| \leq \frac{\varepsilon}{2} + 2M\ \frac{1}{4n\delta^2}. \tag{2.6}$$

Por lo tanto $\forall n > \frac{M}{\varepsilon\delta^2}$ se tendrá que $\forall x\in[0,1]$

$$|f(x)-B_n(f;x)| \leq \frac{\varepsilon}{2}+\frac{\varepsilon}{2} = \varepsilon.$$

□

Este teorema no aborda directamente ninguno de los problemas específicos que resolveremos en este capítulo. De hecho para ajustar un polinomio a una función dada, con una precisión ε dada, habría que conocer los valores de f sobre una malla equiespaciada $\left\{\frac{k}{n}\right\}_{k=0}^{n}$ tan abundante (n tan grande) como resulte de la condición que sigue a (2.6) en la demostración. El polinomio resultante como aproximación de f solo coincidirá con f en los extremos del intervalo, pero a medida que aumenta n se puede observar como **reproduce** cada vez mejor **la forma de** f. Este tipo de objetivos es propio del Diseño Gráfico Asistido por Computadores, materia muy interesante que este curso no aborda.

El primer problema que trataremos será el de aproximar una función mediante un polinomio que coincida con ella en todos los puntos donde se conozca su valor, es decir, que **interpole** a la función f en todos los puntos $x_i, i = 0,1,...,n$, tales que se conozcan los valores $f(x_i)$. Denotaremos por $\mathcal{P}_n$ al conjunto de todos los polinomios a coeficientes reales de grado menor o igual que n.

INTERPOLACIÓN POLINOMIAL

Sean $f:[a,b]\to\mathbb{R}$, una malla $T=\{x_i\}_{i=0}^{n}\subset[a,b]$, y los valores de f $y_i = f(x_i)$ para $i=0,1,...,n$. Se desea encontrar $p\in\mathcal{P}_n$ un polinomio de grado menor o igual que n, que interpole a f sobre todos los puntos de la malla T, es decir,

$$p(x_i) = f(x_i) = y_i \quad \text{para todo } i=0,1,...,n.$$

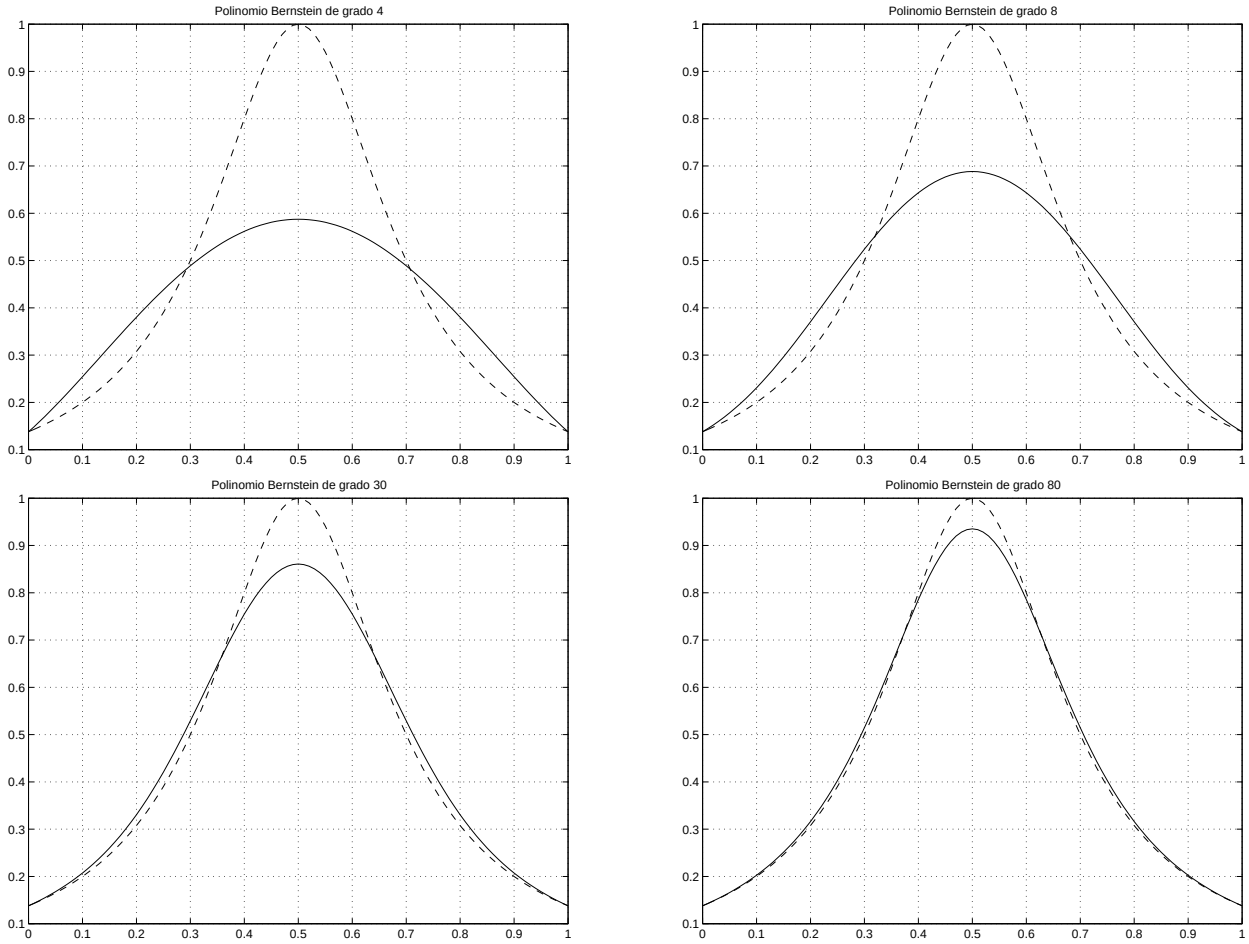


Figura 2.1: Aproximación por polinomios de Bernstein de la función $\frac{1}{1+x^2}$.

Este problema tiene solución única si los puntos de la malla T, también llamados nodos de interpolación, son todos distintos.

Teorema 2.7. *Si los $(n+1)$ nodos de interpolación de la malla T son todos distintos, entonces existe un único polinomio de interpolación de grado menor o igual que n, $p \in \mathcal{P}_n$, $p(x_i) = y_i$, $\forall i = 0, ..., n$.*

Demostración. Probaremos en primer lugar la unicidad. Supongamos que hay dos polinomios de interpolación de grado menor o igual que n, $p, q \in \mathcal{P}_n$,

$$p(x_i) = q(x_i) = y_i \quad \forall i = 0, 1, ..., n.$$

Entonces el polinomio $h = p - q$ será también un polinomio de grado menor o igual que n que tendrá $(n+1)$ raíces

$$h(x_i) = 0, \quad \forall i = 0, 1, ..., n.$$

Pero esto solo es posible si h es el polinomio nulo, es decir $p = q$.

La demostración de la existencia del polinomio de interpolación es constructiva. Sean

$$\ell_{n,i}(x) = \frac{\prod_{\substack{j=0 \\ j \neq i}}^{n} (x - x_j)}{\prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j)} \quad \text{para } i = 0, 1, ..., n. \tag{2.8}$$

Se verifica fácilmente que $\forall i = 0, 1, ..., n$

$$\ell_{n,i} \in \mathcal{P}_n$$

$$\ell_{n,i}(x_k) = \delta_{ik} = \begin{cases} 1 & \text{si } i = k \\ 0 & \text{si } i \neq k \end{cases}$$

y que por lo tanto

$$L_n(x) = \sum_{i=0}^{n} y_i \ell_{n,i}(x) \tag{2.9}$$

es de grado menor o igual a n e interpola, es decir, es el único polinomio de interpolación. □

La expresión (2.9) del polinomio de interpolación recibe el nombre de polinomio de **Lagrange** y los polinomios definidos en (2.8) se llaman polinomios de base de Lagrange.

La calidad del polinomio de interpolación como aproximación de f se establece en el teorema acerca del error que sigue. Denotaremos por $\overline{co}(x_0, x_1, ..., x_n, x)$ al menor intervalo cerrado que contenga a $x_0, x_1, ..., x_n, x$.

Teorema 2.10. *Si f es $(n+1)$ veces continuamente derivable sobre $[a, b]$, entonces $\forall x \in [a, b]$ $\exists \xi \in \overline{co}(x_0, x_1, ..., x_n, x)$ tal que*

$$e_n(x) = f(x) - L_n(x) = \frac{\prod_{j=0}^{n} (x - x_j) f^{(n+1)}(\xi)}{(n+1)!}.$$

Demostración. Consideremos un x arbitrario y fijo en $[a,b]$ ($x \neq x_i \quad \forall i = 0,1,...,n$, pues obviamente $e_n(x_i) = 0$) y la función definida sobre $[a,b]$

$$R(t) = e_n(t) - \frac{w(t)}{w(x)} e_n(x), \text{ donde } w(t) = \prod_{j=0}^{n} (t - x_j).$$

R es una función con $(n+1)$ derivadas continuas que se anula en $(n+2)$ lugares distintos

$$R(x_i) = 0 \quad \forall i = 0,1,...,n \quad \text{y} \quad R(x) = 0.$$

Utilizando el teorema de Rolle reiteradamente se concluirá que:
existen $(n+1)$ puntos $\xi_i^1 \in \overline{co}(x_0, x_1, ..., x_n, x)$ donde se anula R', $R'(\xi_i^1) = 0$,
existen n puntos $\xi_i^2 \in \overline{co}(x_0, x_1, ..., x_n, x)$ donde se anula R'', $R''(\xi_i^2) = 0$,
$\vdots$
existe 1 punto $\xi \in \overline{co}(x_0, x_1, ..., x_n, x)$ donde se anula $R^{(n+1)}$, $R^{(n+1)}(\xi) = 0$, es decir

$$e_n^{(n+1)}(\xi) - \frac{w^{(n+1)}(\xi)}{w(x)} e_n(x) = 0. \tag{2.11}$$

Pero $e_n(t) = f(t) - L_n(t)$ y L_n es un polinomio de grado n, de modo que su derivada $(n+1)$-ésima se anula y resulta

$$e_n^{(n+1)}(t) = f^{(n+1)}(t).$$

Por otra parte $w(t)$ es un polinomio mónico de grado $(n+1)$ y su derivada $(n+1)$-ésima será

$$w^{(n+1)}(\xi) = (n+1)!$$

En resumen, en (2.11) se tendrá

$$e_n(x) = \frac{w(x) f^{(n+1)}(\xi)}{(n+1)!}$$

con lo cual concluye la demostración del teorema. □

Como consecuencia de este teorema se obtiene el resultado de convergencia que sigue.

Corolario 2.12. Si f es $(n+1)$ veces continuamente derivable sobre $[a,b]$ y L_n es el polinomio de interpolación de f sobre la malla $T \subset [a,b]$, entonces

$$\forall x \in [a,b] \quad |f(x) - L_n(x)| \leq M h^{n+1},$$

donde $M = \sup_{t \in [a,b]} |f^{(n+1)}(t)|$ y h es el máximo paso, o distancia entre dos nodos consecutivos incluidos los extremos, es decir,

$$h = \max\{x_0 - a, b - x_n, \max_{0 \leq i \leq n-1} (x_{i+1} - x_i)\},$$

si los nodos estuvieran ordenados (lo que no es un requisito, excepto para facilitar la notación).

Demostración. Según el teorema (2.10) se tendrá que

$$\forall x \in [a,b] \quad |f(x) - L_n(x)| \leq \frac{M}{(n+1)!} \sup_{t\in[a,b]} \prod_{j=0}^{n} |t - x_j|.$$

El supremo que aparece en esta expresión se realiza en alguno de los extremos del intervalo. Sin perder generalidad, supongamos que esto ocurre en $t = a$, con lo cual

$$\sup_{t\in[a,b]} \prod_{j=0}^{n} |t - x_j| = |a - x_0||a - x_1|...|a - x_n| \leq h \cdot 2h... \cdot (n+1)h = (n+1)!h^{n+1}.$$

□

Notas:

2.13. *La malla T que produce la menor cota de error es la malla* **equiespaciada**, *es decir aquella en que los nodos se distribuyen uniformemente en el intervalo $[a,b]$ y por lo tanto el máximo paso es el menor posible, $h = \frac{b-a}{n+2}$, $x_i = a + (i+1)h$, para $i = 0, 1, ..., n$.*

2.14. *El error del polinomio de interpolación es menor al centro del intervalo y mayor en los extremos. Esta observación se basa en el comportamiento del término*

$$|w(x)| = \prod_{j=0}^{n} |x - x_j|.$$

2.15. *la convergencia que se establece en el corolario* (2.12) *sugiere que densificando la malla T sobre $[a,b]$, es decir, aumentando el número de nodos de interpolación de modo que el paso máximo, h, tienda a cero, se obtendrá que el error, e_n, tienda a cero. Pero esto solo será cierto si la derivada $(n+1)$-ésima de f no crece con el orden de derivación más rápido de lo que h^{n+1} tiende a cero.*

2.16. *Como corolario del teorema* (2.10) *se tiene un resultado de* **exactitud**. *Si $f \in \mathcal{P}_n$, entonces el error cometido por el polinomio de grado menor o igual a n que lo interpola sobre la malla T será cero (debido al factor $f^{(n+1)}(\xi)$), es decir, f y su polinomio de interpolación coinciden (el polinomio de interpolación es exacto). Este resultado es en extremo evidente debido a la unicidad del polinomio de grado n que pasa por $(n+1)$ puntos dados. Lo que nos interesa recalcar aquí, es que este hecho sea explícito en la expresión del error dada en dicho teorema.*

2.17. *La expresión del error dada en el teorema* (2.10) *tiene gran similitud con el error del polinomio de Taylor de grado n, T_n. De hecho, si colapsáramos todos los nodos de la malla T en un solo nodo x_0 (el teorema de existencia y unicidad del polinomio de interpolación prohíbe expresamente la repetición de nodos que aquí proponemos), entonces la expresión del error dada en el teorema* (2.10) *coincidiría completamente con el error del polinomio de Taylor. Esta observación sugiere que el polinomio de Taylor es un polinomio de interpolación con nodos repetidos $x_0 = x_1 = ... = x_n$, y que al repetir nodos de interpolación se obtiene interpolación de derivadas, puesto que $T_n^{(k)}(x_0) = f^{(k)}(x_0), \forall 0 \leq k \leq n$. Esta similitud es suficientemente interesante como para estudiar el comportamiento del polinomio de interpolación cuando los nodos tienden a colapsarse.*

2.18. *El conjunto de polinomios de grado n, $\{\ell_{n,i}\}_{n=0}^{n}$, definido en* (2.8), *forma una base de $\mathcal{P}_n$ muy distinta de la base canónica; todos los elementos de la base son polinomios del mismo grado (en cambio en la base canónica son todos de distinto grado) y no se pueden definir cuando los nodos se repiten. La ventaja que compensa estas deficiencias es la extrema simplicidad de los coeficientes de la combinación lineal que representa al polinomio de interpolación en esta base (los valores $y_i, \forall i = 0, 1, ..., n$). De estas dos últimas notas surge el interés por contar con una base de $\mathcal{P}_n$ formada por polinomios que estén definidos para nodos*

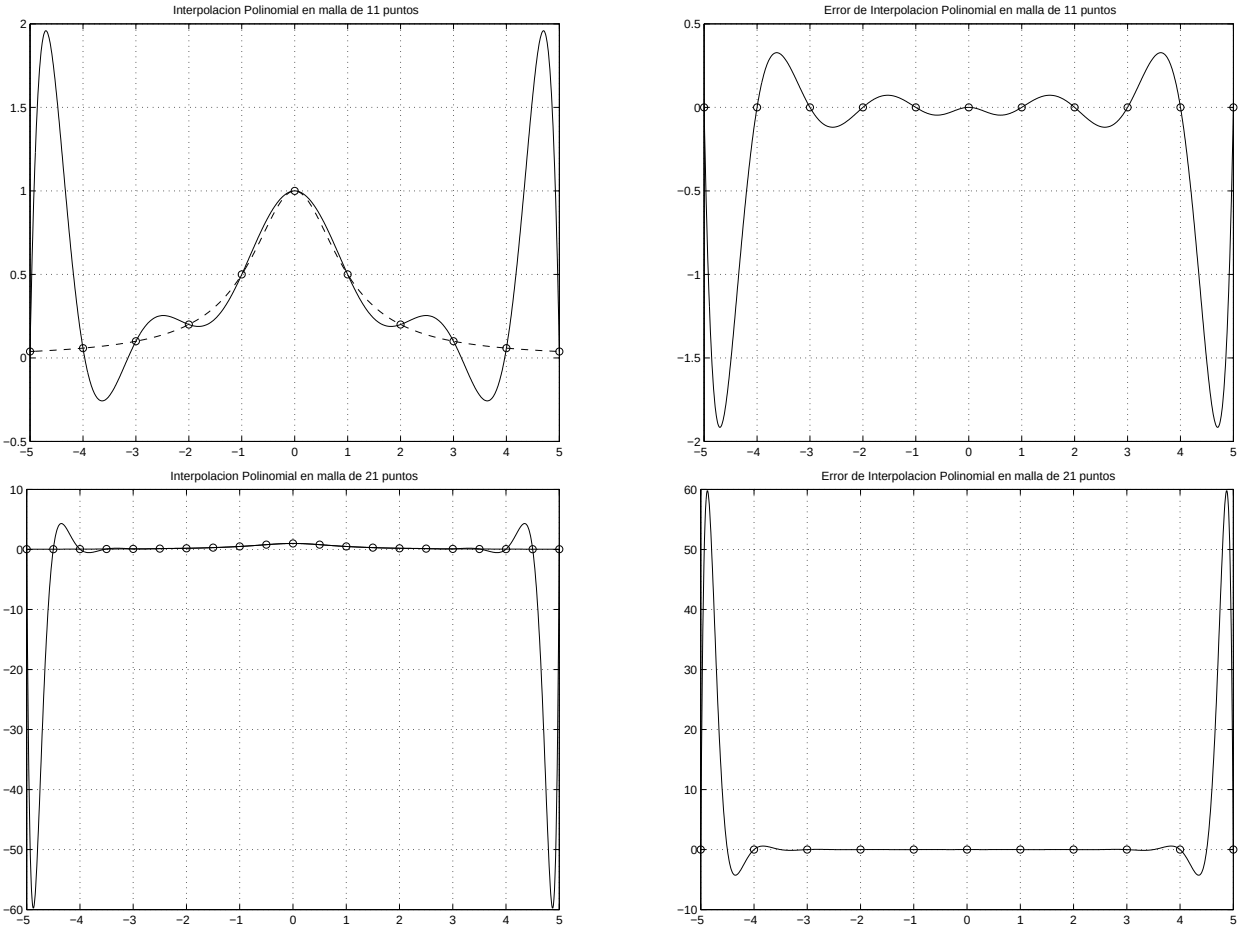


Figura 2.2: Interpolación de la función $\frac{1}{1+x^2}$ con una malla de 11 y 21 puntos equiespaciados.

repetidos, que sean de distinto grado y que permitan escribir la combinación lineal que representa al único polinomio de interpolación con coeficientes simples de calcular. La base de $\mathcal{P}_n$ que satisface todos estos requisitos es la base de **Newton** *asociada a la malla T, definida por*

$$\{1, (x-x_0), (x-x_0)(x-x_1), ..., \prod_{i=0}^{n-1}(x-x_i)\}. \tag{2.19}$$

Los coeficientes del polinomio de interpolación en esta base son conocidos y se llaman diferencias divididas.

Definición 2.20. Dada una colección cualquiera de puntos distintos $\{t_i\}_{i=0}^m$, se definen las **diferencias divididas** por recurrencia como

- $f[t_i] = f(t_i)$ será la diferencia dividida de orden 0 en $t_i \quad \forall i = 0, 1, ..., m$.
- $f[t_i, t_{i+1}, ..., t_{i+k}] = \frac{f[t_{i+1},...,t_{i+k}] - f[t_i,...,t_{i+k-1}]}{t_{i+k}-t_i}$ será la diferencia dividida de orden k en t_i,t_{i+1},. . .,t_{i+k}, $\forall k \geq 1, \forall i = 0, 1, ..., m-k$.

Los cálculos de estas diferencias divididas se ordenan en un arreglo triangular

$$\begin{array}{cccccc}
t_0 & y_0 & & & & \\
 & & > f[t_0,t_1] & & & \\
t_1 & y_1 & & > f[t_0,t_1,t_2] & & \\
 & & > f[t_1,t_2] & \vdots & \ddots & \\
t_2 & y_2 & \vdots & & & \\
\vdots & \vdots & & & & \\
\vdots & \vdots & & & & \\
t_{m-1} & y_{m-1} & & & & \\
 & & > f[t_{m-1},t_m] & & & \\
t_m & y_m & & & & \\
 & \uparrow & \uparrow & \uparrow & & \\
 & \text{orden 0} & \text{orden 1} & \text{orden 2.....} & &
\end{array}$$

Tabla 2.1: Tabla de Diferencias Divididas

Si bien la definición de las diferencias divididas dada en (2.20) facilita su cálculo ordenado y permite establecer varias propiedades, hay importantes propiedades que se deducen de una expresión que suele usarse como definición alternativa a la que hemos privilegiado. Por esta razón será ella quién encabece la lista de propiedades que sigue.

Propiedades de las Diferencias Divididas.

Propiedad 2.21.

$$f[x_i, x_{i+1}, ..., x_{i+k}] = \sum_{j=i}^{i+k} \frac{f(x_j)}{\prod\limits_{\substack{p=i \\ p\neq j}}^{i+k} (x_j - x_p)}.$$

Propiedad 2.22. las Diferencias Divididas son *invariantes bajo permutaciones* de los nodos en que se basan, es decir, el orden en que aparecen los argumentos $x_i, x_{i+1}, ..., x_{i+k}$, es completamente irrelevante.

Propiedad 2.23. Las $(n+1)$ Diferencias Divididas $f[x_0], f[x_0, x_1], ..., f[x_0, x_1, ..., x_n]$ son los coeficientes del polinomio de interpolación en la base de **Newton** de $\mathcal{P}_n$, dada por (2.19). Es decir, definiendo por recurrencia los polinomios de Newton

$$\begin{aligned} &N_0(x) = f[x_0] \\ &N_k(x) = N_{k-1}(x) + f[x_0, x_1, ..., x_k] \prod_{i=0}^{k-1} (x - x_i) \quad \forall 1 \leq k \leq n. \end{aligned} \tag{2.24}$$

El polinomio $N_n(x)$ será el polinomio de interpolación de f sobre la malla T.

Propiedad 2.25. Si f tiene k derivadas continuas en $[a, b]$ entonces existe

$$\xi \in \overline{co}(x_i, x_{i+1}, ..., x_{i+k})$$

tal que

$$f[x_i, x_{i+1}, ..., x_{i+k}] = \frac{f^{(k)}(\xi)}{k!}.$$

(Hacemos notar que si $k = 1$ se tiene el teorema del valor medio y por lo tanto este es un T.V.M. generalizado).

Propiedad 2.26. Si $f \in \mathcal{P}_n$, entonces todas las Diferencias Divididas de orden mayor que n, basadas en cualquier conjunto de nodos, serán nulas.

Propiedad 2.27. Definiendo las Diferencias Divididas con nodos repetidos como el límite cuando todos los demás nodos tiende al primero se cumple que, si f tiene k derivadas continuas en una vecindad de x_0, entonces

$$f[\underbrace{x_0, x_0, ..., x_0}_{(k+1)\text{veces}}] = \frac{f^{(k)}(x_0)}{k!}.$$

Demostraciones. Algunas de las propiedades anteriores son consecuencia directa de las previas, como por ejemplo (2.22) que se deduce de la expresión dada en (2.21), así como (2.26) y (2.27) se deducen de (2.25). Las propiedades restantes se demostrarán por inducción.

Comenzaremos por demostrar (2.21).

Si $k = 0$ la afirmación es evidentemente cierta.

Supongamos que la afirmación es cierta para $(k-1)$ y probemos para k. Por definición

$$f[x_i, x_{i+1}, ..., x_{i+k}] = \frac{1}{x_{i+k} - x_i} \{f[x_{i+1}, ..., x_{i+k}] - f[x_i, x_{i+1}, ..., x_{i+k-1}]\}.$$

Usando la hipótesis de inducción se tendrá que

$$f[x_i, x_{i+1}, ..., x_{i+k}] = \frac{1}{x_{i+k} - x_i} \left\{ \sum_{j=i+1}^{i+k} \frac{f(x_j)}{\prod\limits_{\substack{p=i+1 \\ p \neq j}}^{i+k} (x_j - x_p)} - \sum_{j=i}^{i+k-1} \frac{f(x_j)}{\prod\limits_{\substack{p=i \\ p \neq j}}^{i+k-1} (x_j - x_p)} \right\}$$

$$= \frac{1}{x_{i+k} - x_i} \left\{ \frac{f(x_{i+k})}{\prod\limits_{p=i+1}^{i+k-1} (x_{i+k} - x_p)} + \sum_{j=i+1}^{i+k-1} f(x_j) \left(\frac{1}{\prod\limits_{\substack{p=i+1 \\ p \neq j}}^{i+k} (x_j - x_p)} - \frac{1}{\prod\limits_{\substack{p=i \\ p \neq j}}^{i+k-1} (x_j - x_p)} \right) - \frac{f(x_i)}{\prod\limits_{p=i+1}^{i+k-1} (x_i - x_p)} \right\}$$

de lo cual se obtiene fácilmente la expresión propuesta.

Con el fin de demostrar (2.23) probaremos por inducción sobre k que $\forall k$ el polinomio de grado k, $N_k(x)$, interpola a f en $x_0, x_1, ..., x_k$ y por lo tanto es el único polinomio de interpolación asociado a estos datos.

Si $k = 0$ es claro que $N_0(x)$, definido en (2.24) interpola a f en x_0.

Supongamos que $N_{k-1}(x)$ interpola a f en $x_0, x_1, ..., x_{k-1}$ y probemos que $N_k(x)$ interpola a f en $x_0, x_1, ..., x_k$.

De la definición dada en (2.24), se ve fácilmente que

$$\begin{aligned} \forall j = 0, 1, ..., k-1, \quad N_k(x_j) &= N_{k-1}(x_j) + f[x_0, x_1, ..., x_k] \prod_{i=0}^{k-1} (x_j - x_i) \\ &= N_{k-1}(x_j) \\ &= f(x_j), \text{ por hipótesis de inducción.} \end{aligned}$$

De este modo todo lo que resta probar es que $N_k(x_k) = f(x_k)$.

La hipótesis de inducción dice que $N_{k-1}(x)$ es el único polinomio de grado $(k-1)$ que interpola a f en los k nodos $x_0, x_1, ..., x_{k-1}$, y por lo tanto se le puede expresar como el polinomio de Lagrange

$$N_{k-1}(x) = L_{k-1}(x) = \sum_{j=0}^{k-1} f(x_j) \frac{\prod\limits_{\substack{i=0 \\ i \neq j}}^{k-1} (x - x_i)}{\prod\limits_{\substack{i=0 \\ i \neq j}}^{k-1} (x_j - x_i)}.$$

Usando esta expresión y la propiedad (2.21) se tendrá que

$$N_k(x_k) = \sum_{j=0}^{k-1} f(x_j) \frac{\prod\limits_{\substack{i=0 \\ i\neq j}}^{k-1} (x_k - x_i)}{\prod\limits_{\substack{i=0 \\ i\neq j}}^{k-1} (x_j - x_i)} + \prod_{i=0}^{k-1} (x_k - x_i) \sum_{j=0}^{k} \frac{f(x_j)}{\prod\limits_{\substack{p=0 \\ p\neq j}}^{k} (x_j - x_p)}$$

$$= f(x_k) + \sum_{j=0}^{k-1} f(x_j) \left\{ \frac{\prod\limits_{\substack{i=0 \\ i\neq j}}^{k-1} (x_k - x_i)}{\prod\limits_{\substack{i=0 \\ i\neq j}}^{k-1} (x_j - x_i)} + \frac{\prod\limits_{i=0}^{k-1} (x_k - x_i)}{\prod\limits_{\substack{p=0 \\ p\neq j}}^{k} (x_j - x_p)} \right\}$$

$$= f(x_k).$$

Hemos probado así que $N_n(x)$ interpola a f sobre la malla T como se afirma en (2.23).

Para demostrar (2.25) basta estudiar el error cometido por el polinomio de interpolación usando la expresión de Newton. Probaremos que

$$f(x) - N_n(x) = \prod_{i=0}^{n} (x - x_i) f[x_0, x_1, ..., x_n, x]. \tag{2.28}$$

En efecto, dado x cualquiera en $[a, b]$ (obviamente $x \neq x_i, \forall i = 0, 1, ..., n$) consideraremos la malla de $(n+2)$ nodos $\widetilde{T} = T \cup \{x\}$ y el polinomio de interpolación de grado $(n+1)$ correspondiente

$$N_{n+1}(t) = N_n(t) + f[x_0, x_1, ..., x_n, x] \prod_{i=0}^{n} (t - x_i).$$

Por construcción, este polinomio interpola a f en $t = x$, es decir,

$$f(x) = N_{n+1}(x) = N_n(x) + f[x_0, x_1, ..., x_n, x] \prod_{i=0}^{n} (x - x_i),$$

lo que prueba (2.28). Hemos encontrado así una nueva expresión del error del polinomio de interpolación, para el cual contábamos con el teorema (2.10). Juntando ambos resultados obtenemos que si f tiene $(n+1)$ derivadas continuas en $[a, b]$ entonces $\exists \xi \in \overline{co}(x_0, x_1, ..., x_n, x)$ tal que

$$e_n(x) = f[x_0, x_1, ..., x_n, x] \prod_{i=0}^{n} (x - x_i) = \frac{\prod\limits_{i=0}^{n} (x - x_i) f^{(n+1)}(\xi)}{(n+1)!}$$

y por lo tanto

$$f[x_0, x_1, ..., x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Como n es arbitrario y podemos incluir el punto x, cualquiera, en la numeración de los nodos, se concluye la demostración de (2.25). □

Notas.

Los nodos $x_0, x_1, ..., x_n$ no están obligados a seguir ningún orden y además hemos probado que las diferencias divididas son invariantes bajo permutaciones de los argumentos. Sin embargo las diferencias

divididas que participan en el polinomio de interpolación corresponden exclusivamente a la diagonal superior de la tabla (2.1) y los polinomios de la base de Newton también privilegian un orden. Es decir, una vez que se ha decidido un orden debe ser respetado tanto en la confección de la tabla de diferencias divididas como en los polinomios de base.

La expresión del polinomio de interpolación de Newton dada por recurrencia en (2.24) corresponde a

(2.29)

$$N_n(x) = f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1) + \ldots + f[x_0, x_1, \ldots, x_n] \prod_{i=0}^{n-1} (x - x_i).$$

En los Problemas Resueltos 3. y 4. se proponen alternativas prácticas para usar del modo más provechoso esta libertad en la elección del orden de los nodos.

La propiedad (2.27) resuelve el problema de la repetición de nodos y permite relacionar el polinomio de Taylor con el polinomio de interpolación como se muestra en el problema 2. . En el problema 6. se muestra como utilizar esta relación para satisfacer condiciones de interpolación diversas.

Los seis problemas que se resuelven a continuación tienen por finalidad extraer consecuencias prácticas de resultados aparentemente teóricos, establecidos previamente.

Problemas Resueltos.

1. Considere los datos de la tabla que sigue y calcule el polinomio de Newton del mayor grado posible.

$$\begin{array}{rcccccccc} x_i: & -2 & -1 & 0 & 1 & 2 & 3 & 4 & 5 \\ f(x_i): & 0 & -4 & 0 & 0 & 8 & 60 & 216 & 560 \end{array}$$

   ¿Puede estimar el error cometido?

2. Compare el polinomio de Newton de grado n y el polinomio de Taylor del mismo grado en torno a x_0. Utilizando la propiedad (2.27) muestre que el polinomio de Taylor es un polinomio de Newton con nodos repetidos.

3. Considere la base de $\mathcal{P}_n$ siguiente

$$\left\{1, (x - x_n), (x - x_n)(x - x_{n-1}), ..., \prod_{i=1}^{n} (x - x_i)\right\}.$$

   Encuentre la expresión del polinomio de interpolación de grado n en esta base.

4. Considere una tabla ordenada ($x_0 < x_1 < ... < x_N$) y extensa (N grande).

$$\begin{array}{rcccc} x_i: & x_0 & x_1 & \cdots & x_N \\ f(x_i): & f(x_0) & f(x_1) & \cdots & f(x_N) \end{array}$$

   Se desea aproximar $f(x)$ por $N_n(x)$ con $n << N$ y x está ubicado aproximadamente al centro de la tabla

$$x_{i^*} < x < x_{i^*+1} \text{ con } i^* \approx N/2$$

   ¿Cuál será la mejor estrategia para calcular $N_n(x)$? Utilice esta estrategia para calcular $N_3(0.5)$ en el ejemplo del problema 1.

5. Demuestre que $\forall n \sum_{i=0}^{n} \ell_{n,i}(x) = 1$, donde $\ell_{n,i}(x)$ corresponde al i-ésimo polinomio de base de Lagrange de grado n definido en (2.8) .

6. Considere conocidos los siguientes datos:

$$x_0 < x_1 < x_2 < x_3$$
$$f(x_0), f'(x_0), f''(x_0), f(x_1), f(x_2), f(x_3), f'(x_3).$$

Calcule el polinomio de Newton de grado 6 que interpola todos estos datos, es decir,

$$\begin{aligned} N_6(x_i) &= f(x_i), \quad \forall i = 0,1,2,3 \\ N_6'(x_i) &= f'(x_i), \quad \text{para } i = 0,3 \\ N_6''(x_0) &= f''(x_0) \end{aligned}$$

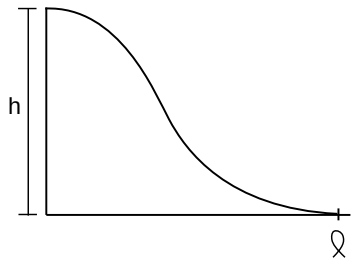
y pruebe que si f es 7 veces continuamente derivable, entonces el error cometido es

$$e_6 = (x-x_0)^3(x-x_1)(x-x_2)(x-x_3)^2 \frac{f^{(7)}(\xi)}{7!} \text{ para algún } \xi \in [x_0, x_3].$$

Compare el error en cada uno de los subintervalos

$$(x_0, x_1), \quad (x_1, x_2), \quad (x_2, x_3).$$

7. Considere el problema de construir un acceso a un paso sobre nivel de altura h sobre un camino plano, a partir de un punto que está a distancia ℓ como lo muestra la figura.



La subida no debe tener pendiente más pronunciada que el ángulo de 45° y los empalmes deben tener la suavidad correspondiente a primera derivada continua. Debemos encontrar la función que representa la subida y determinar la distancia mínima ℓ a la que ésta debe comenzar para que la pendiente satisfaga la restricción impuesta.

Desarrollo de los problemas.

1. En primer lugar construimos la tabla de diferencias divididas asociada a los datos dados:

x_i	$f(x_i)$											
-2	0											
		$>$	-4									
-1	-4			$>$	4							
		$>$	4			$>$	-2					
0	0			$>$	-2			$>$	1			
		$>$	0			$>$	2			$>$	0	
1	0			$>$	4			$>$	1			
		$>$	8			$>$	6			$>$	0	
2	8			$>$	22			$>$	1			
		$>$	52			$>$	10			$>$	0	
3	60			$>$	52			$>$	1			
		$>$	156			$>$	14					
4	216			$>$	94							
		$>$	344									
5	560											

$$N_4(x) = -4(x+2) + 4(x+2)(x+1) - 2(x+2)(x+1)x + 1(x+2)(x+1)x(x-1).$$

Como no sabemos de f nada más que los datos de la tabla, no **podremos acotar** el error cometido. Una manera de **estimar** el error, cuando se tiene una tabla más extensa que lo que se utiliza para construir el polinomio de grado n, es con el término siguiente es decir,

$$e_n(x) \approx \prod_{i=0}^{n}(x - x_i) f[x_0, x_1, ..., x_n, x_{n+1}]. \tag{2.30}$$

En nuestro caso

$$e_4(x) \approx 0.$$

Una diferencia dividida nula no significa nada en particular. En cambio, si la tabla es extensa y si a partir de cierto orden k todas las diferencias divididas se anulan, entonces se tendrá que f es un polinomio de grado $(k-1)$. En el caso de nuestro ejercicio, efectivamente, los datos corresponden a $f(x) = (x-1)^2 x(x+2) \in \mathcal{P}_4$ y $e_4(x) = 0 \quad \forall x$.

2. El polinomio de Taylor de grado n en torno a x_0 es

$$T_n(x) = f(x_0) + (x - x_0) f'(x_0) + (x - x_0)^2 \frac{f''(x_0)}{2!} + \ldots + (x - x_0)^n \frac{f^{(n)}(x_0)}{n!}.$$

De la propiedad (2.27) tenemos que

$$T_n(x) = f[x_0] + (x - x_0) f[x_0, x_0] + (x - x_0)^2 f[x_0, x_0, x_0] + ... + (x - x_0)^n f[\underbrace{x_0, x_0, ..., x_0}_{(n+1)\text{veces}}]$$

lo que coincide con el polinomio de Newton de grado n correspondiente a la malla $T = \{x_i\}_{i=0}^n$ con $x_i = x_0 \quad \forall i = 1, 2, ..., n$.

3. El polinomio de interpolación en esta base corresponde al polinomio de Newton que resulta de ordenar los nodos al revés

x_n	x_{n-1}	x_{n-2}		x_1	x_0
$\downarrow$	$\downarrow$	$\downarrow$	$\cdots$	$\downarrow$	$\downarrow$
x_0	x_1	x_2		x_{n-1}	x_n

y será

$$N_n(x) = f[x_n] + (x-x_n)f[x_{n-1},x_n] + (x-x_n)(x-x_{n-1})f[x_{n-2},x_{n-1},x_n] + \ldots + \prod_{j=1}^{n}(x-x_j)f[x_0,x_1,\ldots,x_n] \tag{2.31}$$

y equivale a utilizar la diagonal inferior de la tabla de diferencias divididas. Esta expresión recibe el nombre de polinomio de **Newton atrás**.

Para ejemplificar y como sabemos que los datos de el problema 1 proviene de $f(x) = (x+2)(x-1)^2x$, lo que coincide con $N_4(x)$ obtenido antes, calculemos el polinomio de grado 4 en la nueva base de Newton atrás:

$$\begin{aligned} P_4(x) &= 560 + 344(x-5) + 94(x-5)(x-4) + 14(x-5)(x-4)(x-3) \\ &\quad + (x-5)(x-4)(x-3)(x-2) \\ &= 560 + 344x - 1720 + 94(x^2-9x+20) + 14(x^3-12x^2+47x-60) \\ &\quad + x^4 - 14x^3 + 71x^2 - 154x + 120 \\ &= 2x - 3x^2 + x^4 \\ &= x(x^3-3x+2) \\ &= x(x+2)(x-1)^2. \end{aligned}$$

4. La misma idea de reenumerar los nodos puede utilizarse para seguir un camino zigzagueante en la tabla de diferencias divididas.

$$\begin{array}{cccccccc} x_{i^*-3} & y_{i^*-3} & & & & & & \\ & & > & f[x_{i^*-3},x_{i^*-2}] & & & & \\ x_{i^*-2} & y_{i^*-2} & & & > & f[x_{i^*-3},x_{i^*-2},x_{i^*-1}] & & \\ & & > & f[x_{i^*-2},x_{i^*-1}] & & & > & \\ x_{i^*-1} & y_{i^*-1} & & & > & f[x_{i^*-2},x_{i^*-1},x_i^*] & & \\ & & > & f[x_{i^*-1},x_{i^*}] & & & > & \\ x_{i^*} & y_{i^*} & & & > & f[x_{i^*-1},x_{i^*},x_{i^*+1}] & & \\ & & > & f[x_{i^*},x_{i^*+1}] & & & > & f[x_{i^*-1},x_{i^*},x_{i^*+1},x_{i^*+2}] \\ x_{i^*+1} & y_{i^*+1} & & & > & f[x_{i^*},x_{i^*+1},x_{i^*+2}] & & \\ & & > & f[x_{i^*+1},x_{i^*+2}] & & & > & \\ x_{i^*+2} & y_{i^*+2} & & & > & f[x_{i^*+1},x_{i^*+2},x_{i^*+3}] & & \\ & & > & f[x_{i^*+2},x_{i^*+3}] & & & & \\ x_{i^*+3} & y_{i^*+3} & & & & & & \end{array}$$

La reenumeración de nodos según su cercanía $a\ x$ será

$$\begin{array}{cccccccc} x_{i^*} & x_{i^*+1} & x_{i^*-1} & x_{i^*+2} & x_{i^*-2} & x_{i^*+3} & \ldots \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \\ x_0 & x_1 & x_2 & x_3 & x_4 & x_5 & \ldots \end{array}$$

y la expresión del polinomio de interpolación, que se llamará de Stirling, es

$$N_n(x) = y_{i^*} + (x-x_{i^*})f[x_{i^*},x_{i^*+1}] + (x-x_{i^*})(x-x_{i^*+1})f[x_{i^*},x_{i^*+1},x_{i^*-1}] + \ldots + (x-x_{i^*})(x-x_{i^*+1})\ldots(x-x_{i^*-q})f[x_{i^*},x_{i^*+1},x_{i^*-1},\ldots,x_{i^*-q},x_{i^*-p}] \tag{2.32}$$

con

$$\begin{cases} p = n/2 & q = -\frac{n}{2} & \text{si } n \text{ es par} \\ p = -\frac{(n+1)}{2} & q = \frac{(n-1)}{2} & \text{si } n \text{ es impar.} \end{cases}$$

En el caso del problema 1 con $x = 0.5$

$$\begin{aligned} N_4(x) &= -2x(x-1) + 2x(x-1)(x+1) + 1x(x-1)(x+1)(x-2) \\ &= x(x-1)^2(x+2) \\ N_4(0.5) &= 5/16. \end{aligned}$$

5. Se tiene:

$$\sum_{i=0}^{n} \ell_{n,i}(x) = \sum_{i=0}^{n} y_i \ell_{n,i}(x) \quad \text{si } y_i = 1 \quad \forall i = 0, 1, ..., n.$$

Es decir, $\sum_{i=0}^{n} \ell_{n,i}(x) = L_n(x)$ si $f(x_i) = y_i = 1 \quad \forall i = 0, 1, ..., n,$ corresponde al polinomio de interpolación de Lagrange de grado n que interpola a la función constante $f(x) = 1 \quad \forall x \in I\!R$.
Pero esta función $f \in \mathcal{P}_0$ y $\forall n \geq 0$ el polinomio de interpolación será exacto, es decir

$$L_n(x) = f(x) \quad \forall x \in I\!R.$$

Así

$$\sum_{i=0}^{n} \ell_{n,i}(x) = f(x) = 1 \quad \forall x \in I\!R.$$

6. De la analogía con el polinomio de Taylor y por la propiedad (2.27) se propone construir la tabla de diferencias divididas con nodos repetidos:

$$\begin{array}{llllllllllll}
x_0 & f(x_0) & & & & & & & & & & \\
 & & > & f'(x_0) & & & & & & & & \\
x_0 & f(x_0) & & & > & \frac{f''(x_0)}{2} & & & & & & \\
 & & > & f'(x_0) & & & > & f[x_0,x_0,x_0,x_1] & & & \\
x_0 & f(x_0) & & & > & f[x_0,x_0,x_1] & & & > & f[x_0,x_0,x_0,x_1,x_2] & & \\
 & & > & f[x_0,x_1] & & & > & f[x_0,x_0,x_1,x_2] & & & > & \\
x_1 & f(x_1) & & & > & f[x_0,x_1,x_2] & & & > & f[x_0,x_0,x_1,x_2,x_3] & & \\
 & & > & f[x_1,x_2] & & & > & f[x_0,x_1,x_2,x_3] & & & > & \\
x_2 & f(x_2) & & & > & f[x_1,x_2,x_3] & & & > & f[x_0,x_1,x_2,x_3,x_3] & & \\
 & & > & f[x_2,x_3] & & & > & f[x_1,x_2,x_3,x_3] & & & & \\
x_3 & f(x_3) & & & > & f[x_2,x_3,x_3] & & & & & & \\
 & & > & f'(x_3) & & & & & & & & \\
x_3 & f(x_3) & & & & & & & & & &
\end{array}$$

$$\begin{array}{lll}
> & f[x_0,x_0,x_0,x_1,x_2,x_3] & \\
 & & > \quad f[x_0,x_0,x_0,x_1,x_2,x_3,x_3] \\
> & f[x_0,x_0,x_1,x_2,x_3,x_3] &
\end{array}$$

$$\begin{aligned} N_6(x) = f(x_0) &+ (x-x_0)f'(x_0) + (x-x_0)^2\frac{f''(x_0)}{2!} + (x-x_0)^3 f[x_0,x_0,x_0,x_1] \\ &+ (x-x_0)^3(x-x_1)f[x_0,x_0,x_0,x_1,x_2] + (x-x_0)^3(x-x_1)(x-x_2)f[x_0,x_0,x_0,x_1,x_2,x_3] \\ &+ (x-x_0)^3(x-x_1)(x-x_2)(x-x_3)f[x_0,x_0,x_0,x_1,x_2,x_3,x_3] \end{aligned}$$

Se observa fácilmente que

$$\begin{aligned}
N_6(x_0) &= f(x_0)\\
N_6(x_1) &= f(x_0) + (x_1 - x_0)f'(x_0) + (x_1 - x_0)^2\frac{f''(x_0)}{2!} + (x_1 - x_0)^3 f[x_0, x_0, x_0, x_1]\\
&= f(x_0) + (x_1 - x_0)f'(x_0) + (x_1 - x_0)^2\frac{f''(x_0)}{2!} + (x_1 - x_0)^3\frac{\{f[x_0, x_0, x_1] - f''\frac{(x_0)}{2}\}}{(x_1 - x_0)}\\
&= f(x_0) + (x_1 - x_0)f'(x_0) + (x_1 - x_0)^2 f[x_0, x_0, x_1]\\
&= f(x_0) + (x_1 - x_0)f'(x_0) + (x_1 - x_0)^2\frac{\{f[x_0, x_1] - f'(x_0)\}}{(x_1 - x_0)}\\
&= f(x_1).
\end{aligned}$$

De manera similar se prueba que

$$\begin{aligned}
N_6(x_2) &= f(x_2),\\
N_6(x_3) &= f(x_3).
\end{aligned}$$

Para la interpolación de derivadas se debe calcular $N_6'(x)$.

$$\begin{aligned}
N_6'(x) = {} & f'(x_0) + (x - x_0)f''(x_0) + 3(x - x_0)^2 f[x_0, x_0, x_0, x_1]\\
& + \{3(x - x_0)^2(x - x_1) + (x - x_0)^3\} f[x_0, x_0, x_0, x_1, x_2]\\
& + \{3(x-x_0)^2(x-x_1)(x-x_2) + (x-x_0)^3(x-x_2) + (x-x_0)^3(x-x_1)\} f[x_0, x_0, x_0, x_1, x_2, x_3]\\
& + \big\{3(x-x_0)^2(x-x_1)(x-x_2)(x-x_3) + (x-x_0)^3(x-x_2)(x-x_3) + (x-x_0)^3(x-x_1)(x-x_3)\\
& + (x - x_0)^3(x - x_1)(x - x_2)\big\} f[x_0, x_0, x_0, x_1, x_2, x_3, x_3].
\end{aligned}$$

Derivando una vez más y evaluando ambas expresiones en x_0 se observa fácilmente que

$$N_6'(x_0) = f'(x_0) \text{ y } N_6''(x_0) = f''(x_0).$$

Desarrollando cada una de las diferencias divididas y evaluando la expresión de N_6' en $x = x_3$ se obtiene

$$N_6'(x_3) = f'(x_3).$$

Obviamente el error de este polinomio con nodos repetidos será

$$e_6(x) = (x - x_0)^3(x - x_3)^2(x - x_1)(x - x_2)\frac{f^{(7)}(\xi)}{7!}.$$

Si $h = Max\{(x_1 - x_0), (x_2 - x_1), (x_3 - x_2)\}$ con nodos ordenados, entonces

$$\begin{aligned}
x \in (x_0, x_1) &\Rightarrow |e_6(x)| \leq 18h^7 M,\\
x \in (x_1, x_2) &\Rightarrow |e_6(x)| \leq 32h^7 M,\\
x \in (x_2, x_3) &\Rightarrow |e_6(x)| \leq 54h^7 M,
\end{aligned}$$

con $M = \max_{x\in[x_0,x_3]} |f^{(7)}(x)|$.

Es decir, la cota al centro del intervalo aún es menor que al extremo derecho donde se agregó 1 condición de interpolación (en $f'(x_3)$), pero es mayor que al extremo izquierdo donde se agregaron 2 condiciones de interpolación (en $f'(x_0), f''(x_0)$).

7. Sea $S(x)$ la función que representa el acceso al paso sobre nivel. Las condiciones de interpolación de este problema son 4 : $S(0) = h, S(\ell) = 0,\ S'(0) = 0, S'(\ell) = 0$. Por lo tanto S puede determinarse como el único polinomio de grado menor o igual a 3 que satisface estas cuatro condiciones de interpolación de la tabla de Diferencias Divididas con nodos repetidas

$$\begin{array}{cccccccc} 0 & h & & & & & & \\ & & > & 0 & & & & \\ 0 & h & & & > & -\frac{h}{\ell^2} & & \\ & & > & -\frac{h}{\ell} & & & > & \frac{2h}{\ell^3} \\ \ell & 0 & & & > & \frac{h}{\ell^2} & & \\ & & > & 0 & & & & \\ \ell & 0 & & & & & & \end{array}$$

De la que se obtiene

$$S(x) = h - x^2\frac{h}{\ell^2} + x^2(x-\ell)\cdot\frac{2h}{\ell^3}.$$

La máxima pendiente se obtiene en x_0 tal que $S''(x_0) = 0$.

$$S'(x) = -2x\frac{h}{\ell^2} + [2x(x-\ell) + x^2]\frac{2h}{\ell^3},$$
$$S''(x) = -2\frac{h}{\ell^2} + [2(x-\ell) + 4x]\frac{2h}{\ell^3}$$

y por lo tanto

$$x_0 = \frac{\ell}{2}, \qquad S'(x_0) = -\frac{3h}{2\ell}.$$

Para que la subida no sea más pronunciada que lo permitido se debe tener que

$$\ell \geq \frac{3h}{2}.$$

Interpolación de Hermite.

Supongamos que se conocen los valores de f y de su derivada f' en los $(n+1)$ puntos distintos de la malla $T = \{x_i\}_{i=0}^n$, es decir, conocemos los $2(n+1)$ reales $y_0, y_1, ..., y_n, y_0', y_1', ..., y_n'$ tales que

$$\left.\begin{array}{l} y_i = f(x_i) \\ y_i' = f'(x_i) \end{array}\right\} i = 0, 1, ..., n.$$

Consideremos el problema de encontrar un polinomio (de un grado adecuado, que determinaremos luego) que interpole a f y su derivada en la malla T, es decir, buscamos p polinomio tal que

$$\left.\begin{array}{l} p(x_i) = y_i \\ p'(x_i) = y_i' \end{array}\right\} i = 0, 1, ..., n.$$

Estas son $2(n+1)$ condiciones "independientes" de interpolación y como $dim(\mathcal{P}_{2n+1}) = 2n+2$, buscamos un polinomio $p \in \mathcal{P}_{2n+1}$. Este polinomio efectivamente existe y es único, se llama polinomio de Hermite y se expresa como

$$H_n(x) = \sum_{i=0}^{n} y_i h_i(x) + \sum_{i=0}^{n} y_i' \widetilde{h}_i(x), \tag{2.33}$$

donde $h_i, \widetilde{h}_i \in \mathcal{P}_{2n+1},\ \forall i = 0, 1, ..., n$, son tales que

$$\left.\begin{array}{l} h_i(x_k) = \delta_{ik} \\ h_i'(x_k) = 0 \\ \widetilde{h}_i(x_k) = 0 \\ \widetilde{h}_i'(x_k) = \delta_{ik} \end{array}\right\} \quad \begin{array}{l} \forall i = 0, 1, ..., n, \\ \forall k = 0, 1, ..., n, \end{array}$$

y se expresan en términos de los polinomios de base de Lagrange, definidos en (2.8), como

$$\left.\begin{array}{rl} h_i(x) & = [1 - 2\ell_{n,i}'(x_i)(x - x_i)](\ell_{n,i}(x))^2 \\ \widetilde{h}_i(x) & = (x - x_i)(\ell_{n,i}(x))^2 \end{array}\right\} \quad i = 0, 1, ..., n. \tag{2.34}$$

Esta construcción prueba la existencia del polinomio de interpolación buscado. A continuación probaremos la unicidad.

Demostración de la unicidad del polinomio de Hermite. Supongamos que tenemos 2 polinomios $p, q \in \mathcal{P}_{2n+1}$ tales que

$$\left.\begin{array}{rl} p(x_i) & = y_i = q(x_i) \\ p'(x_i) & = y_i' = q'(x_i) \end{array}\right\} \quad i = 0, 1, ..., n,$$

$g = p - q \in \mathcal{P}_{2n+1}$ y $g(x_i) = 0 \quad i = 0, 1, ..., n$, es decir, los nodos $x_i\ i = 0, 1, ..., n$ son raíces de g, lo que implica que g se puede factorizar como

$$g(x) = \prod_{i=0}^{n} (x - x_i) h(x) \text{ con } h \in \mathcal{P}_n$$

y por lo tanto

$$g'(x) = w'(x)h(x) + w(x)h'(x),$$

con $w(x) = \prod_{i=0}^{n} (x - x_i)$.

Pero debido a las condiciones de interpolación de la derivada se tiene que

$$g'(x_i) = p'(x_i) - q'(x_i) = y_i' - y_i' = 0 \quad i = 0, 1, ..., n$$

y como $w(x_i) = 0\ \forall i = 0, 1, ..., n$, se cumple que

$$g'(x_i) = w'(x_i)h(x_i) = 0 \quad i = 0, 1, ..., n. \tag{2.35}$$

Pero

$$w'(x) = \sum_{k=0}^{n} \prod_{\substack{j=0 \\ j \neq k}}^{n} (x - x_j) \quad \text{y} \quad w'(x_i) = \prod_{\substack{j=0 \\ j \neq i}}^{n} (x_i - x_j) \neq 0 \quad \forall i = 0, 1, ..., n. \tag{2.36}$$

De modo que (2.35) implica $h(x_i) = 0 \quad \forall i = 0, 1, ..., n$, es decir, los $(n + 1)$ nodos $x_0, x_1, ..., x_n$ son raíces de $h \in \mathcal{P}_n$, lo que solamente es posible si h es el polinomio nulo, lo que a su vez implica que

$$g(x) = w(x)h(x) = 0 \quad \forall x,$$

que equivale a $p(x) = q(x) \quad \forall x$. □

La expresión del polinomio de Hermite dada en (2.33),(2.34), tendrá varias aplicaciones interesantes, como veremos en otros capítulos.

Por ahora hacemos ver que la función auxiliar $w(x)$ y (2.36) permiten escribir el i-ésimo polinomio de Lagrange, definido en (2.8), mediante la muy cómoda y popular expresión

$$\ell_{n,i}(x) = \frac{w(x)}{(x - x_i)w'(x_i)}. \tag{2.37}$$

Producto de la propiedad (2.27) y tal como se hizo en el problema 6.- tenemos otra expresión del polinomio que resuelve el problema de interpolación de Hermite, como un polinomio de Newton con nodos repetidos, es decir, basado en la tabla

$$\begin{array}{cccccccc} x_i & : & x_0 & x_0 & x_1 & x_1 & \dots & x_n & x_n \\ f(x_i) & : & y_0 & y_0 & y_1 & y_1 & \dots & y_n & y_n \end{array}$$

Con esta interpretación del polinomio de interpolación de Hermite podemos concluir la fórmula del error.

Teorema 2.38. *Si $f \in \mathcal{C}^{2n+2}[a,b]$, entonces $\forall x \in [a,b]$ $\exists \xi \in \overline{co}(x_0, x_1, ..., x_n, x)$ tal que*

$$e_n(x) = f(x) - H_n(x) = \prod_{i=0}^{n} (x - x_i)^2 \frac{f^{(2n+2)}(\xi)}{(2n+2!)}.$$

Demostración. Seguiremos las mismas técnicas usadas en la demostración del teorema (2.10) lo que nos evitará repetir los comentarios y explicaciones. Sea x fijo en $[a,b]$ cualquiera distinto de x_i, $i = 0, 1, ..., n$. Sea $\widetilde{w}(t) = \prod_{i=0}^{n} (x - x_i)^2$ y

$$R(t) = e_n(t) - \frac{\widetilde{w}(t)}{\widetilde{w}(x)} e_n(x).$$

Como $\widetilde{w}'(t) = \sum_{i=0}^{n} 2(t - x_i) \prod_{\substack{j=0 \\ j \neq i}}^{n} (t - x_j)^2$ se tiene que

$$\begin{aligned} R(x_i) &= 0 \quad i = 0, 1, ..., n, \\ R'(x_i) &= 0 \quad i = 0, 1, ..., n, \\ R(x) &= 0. \end{aligned}$$

Por el teorema de Rolle tendremos que R' se anula entre cada par de puntos adyacentes de $\{x_0, \dots, x_n, x\}$ $((n+1)$ veces) además de anularse en todos los nodos, es decir, R' se anula en $(n+1) + (n+1) = (2n+2)$ puntos distintos.

Como R tiene la misma regularidad de f podemos repetir la aplicación del teorema de Rolle hasta concluir que $\exists \xi \in \overline{co}(x_0, x_1, ..., x_n, x)$ donde se anula $R^{(2n+2)}$

$$0 = R^{(2n+2)}(\xi) = e_n^{(2n+2)}(\xi) - \frac{\widetilde{w}^{(2n+2)}(\xi)}{\widetilde{w}(x)} e_n(x).$$

Pero $H_n \in \mathcal{P}_{2n+1}$ implica $e^{(2n+2)}(\xi) = f^{(2n+2)}(\xi)$ y como $\widetilde{w} \in \mathcal{P}_{2n+2}$ es un polinomio mónico, entonces $\widetilde{w}^{(2n+2)}(\xi) = (2n+2)!$.

En consecuencia tendremos que

$$f^{(2n+2)}(\xi) - \frac{(2n+2)!}{\widetilde{w}(x)} e_n(x) = 0 \;\Leftrightarrow \qquad e_n(x) = \prod_{i=0}^{n} (x - x_i)^2 \frac{f^{(2n+2)}(\xi)}{(2n+2)!}$$

□

Como consecuencia de este resultado se obtiene el correspondiente teorema de convergencia.

Teorema 2.39. *Si $f \in \mathcal{C}^{2n+2}[a,b]$ y h es el máximo paso de la malla $T = \{x_i\}_{i=0}^{n}$, entonces, el error cometido por el polinomio de Hermite H_n se acota en norma de la convergencia uniforme como*

$$\|f - H_n\|_{\infty,[a,b]} \leq \frac{((n+1)!)^2}{(2n+2)!} h^{2n+2} M = \frac{h^{2n+2}}{\binom{2n+2}{n+1}} M$$

con $M = \|f^{(2n+2)}\|_{\infty,[a,b]}$.

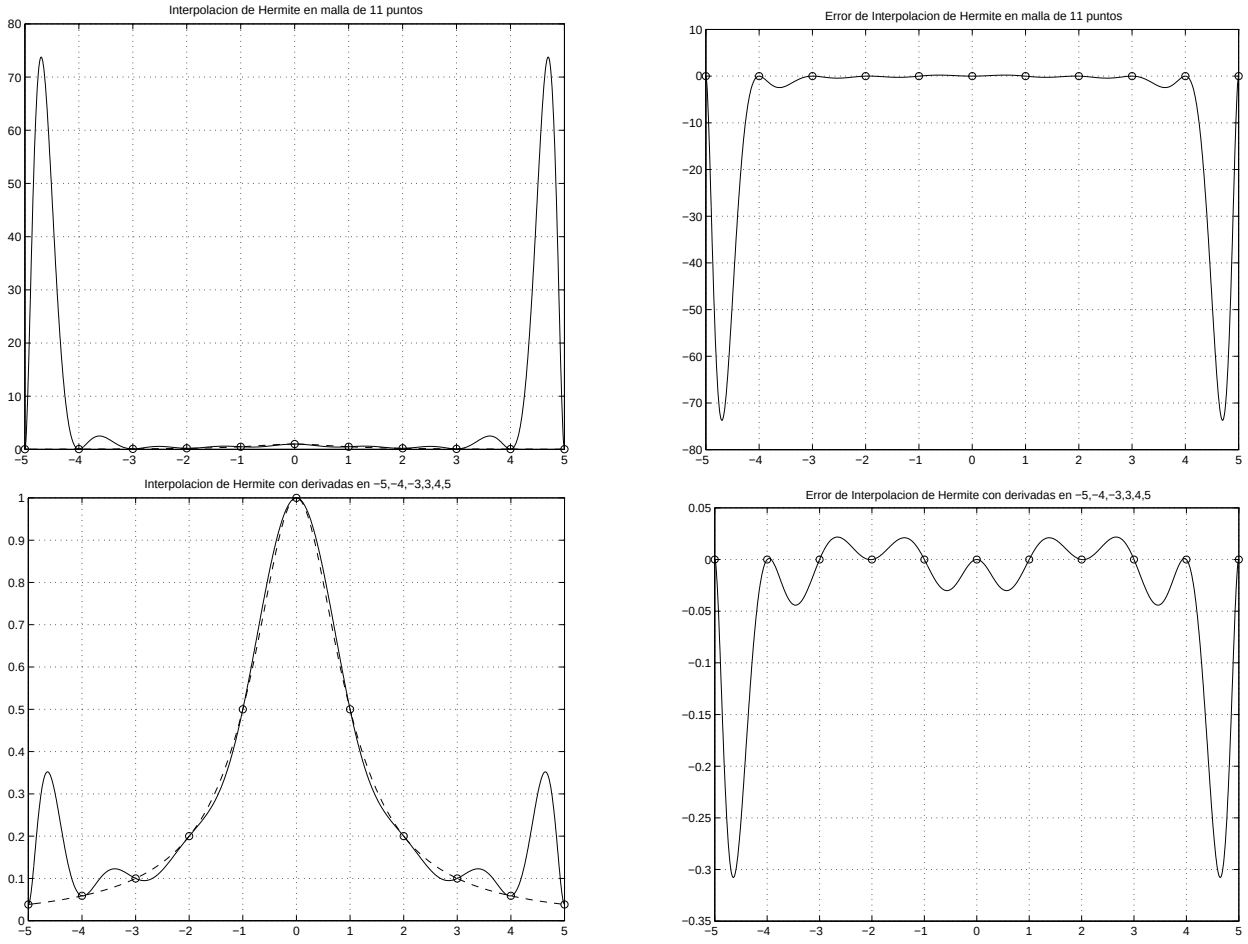


Figura 2.3: Interpolación de Hermite de la función $\frac{1}{1+x^2}$ con dos mallas distintas.

APROXIMACIÓN POLINOMIAL POR PEDAZOS; FUNCIONES SPLINE

Una idea muy intuitiva para diseñar una curva que interpole un conjunto de datos $\{(x_i, y_i)\}_{i=1}^n$ es la de unir los puntos de abscisas consecutivas mediante rectas:

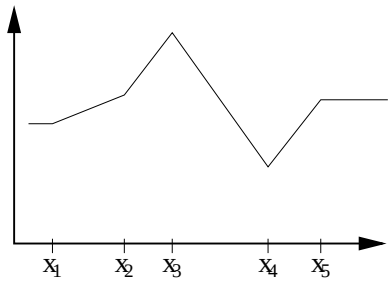


Figura 2.4: Ejemplo de interpolación lineal por pedazos

La función σ que resulta de este método es lineal por pedazos y continua. Claramente es única, pues existe una sola recta que una al punto (x_i, y_i) con el punto (x_{i+1}, y_{i+1}). Se intuye que si los datos corresponden a valores de una función continua entonces al aumentar el número de puntos a interpolar, en el mismo intervalo inicial, la poligonal debería converger a la función, tal como lo hacen los dibujos generados por ploteos computacionales. Pero si no se pretende mejorar el aspecto de la curva interpolante por la vía de aumentar los puntos de interpolación, la alternativa sería hacerla menos quebrada, es decir que tuviera una o más derivadas continuas. Esto obliga a aumentar el grado de los polinomios a considerar. ¿Cómo se relaciona el grado con la suavidad y la unicidad de la interpolante? Las funciones **Spline** de interpolación que definimos a continuación entregan la respuesta a esta pregunta.

Definición 2.40. Dados una malla ordenada $T = \{x_i\}_{i=1}^n$ en $[a, b]$, es decir, $a < x_1 < x_2 < ... < x_n < b$ y los valores $y_i, i = 1, 2, ..., n$, para un natural m, $1 \leq m \leq n$, se define la **función Spline de interpolación de orden** m asociada a estos datos como $\sigma_m : [a, b] \to I\!R$, tal que

1. $\sigma_m(x_i) = y_i, \qquad \forall i = 1, 2, ..., n,$
2. $\sigma_m|_{[x_i, x_{i+1}]} \in \mathcal{P}_{2m-1} \qquad \forall i = 1, 2, ..., n-1,$
   $\sigma_m|_{[a, x_1]} \in \mathcal{P}_{m-1},$
   $\sigma_m|_{[x_n, b]} \in \mathcal{P}_{m-1},$
3. $\sigma_m \in C^{2m-2}_{[a,b]}$.

Note que los polinomios interiores a la malla son siempre de grado impar.

La poligonal dibujada en la figura (2.4) satisface esta definición con $m = 1$ y por lo tanto es una función Spline. Para $m = 2$ se obtiene la más popular de las funciones Spline, la cúbica por pedazos con dos derivadas continuas. Esta interpolante se encuentra disponible en todos los software matemáticos. Nosotros mostraremos su construcción y la utilizaremos para facilitar la interpretación del modelo del cual surgen las funciones Spline.

La unicidad de la función σ_m definida en (2.40) se obtiene luego de establecer una propiedad de optimalidad de gran importancia teórica y práctica: **la función Spline es una proyección ortogonal**.

Siempre que una aproximante corresponda a una proyección ortogonal, se tendrá la posibilidad de develar el modelo subyacente, al reconocer el *criterio de optimalidad* representado por la *distancia* que ella minimiza y los *grados de libertad* o la riqueza del *subespacio* donde se busca esta solución optimal. El caso de las funciones Spline es extremadamente sorprendente e interesante. Como veremos a continuación, en el problema optimal resuelto por la función Spline la única decisión a priori que restringirá el modelo subyacente, consiste en la elección de la distancia a minimizar. El subespacio donde se busca la proyección es el más amplio posible, no tiene más restricciones que la de poder definir correctamente la distancia elegida y por supuesto debe contener solamente funciones que interpolen los datos dados. En ninguna parte se pide que la búsqueda se limite a los polinomios ni a las polinomiales por pedazos, como se podría suponer.

Definición 2.41. Se define el espacio de funciones $\mathcal{H}_m$ como

$$\mathcal{H}_m = \{u : [a,b] \to I\!R | \int_a^b (u^{(m)}(x))^2 dx < +\infty\}$$

y el producto de funciones de $\mathcal{H}_m$ como

$$\forall u, v \in \mathcal{H}_m \quad \langle u, v \rangle_m = \int_a^b u^{(m)}(x) v^{(m)}(x) dx.$$

Propiedad 2.42. El producto $\langle \cdot, \cdot \rangle_m$ definido en (2.41) es un **semi producto interno** en $\mathcal{H}_m$, es decir, satisface todas las propiedades de un producto interno excepto tener núcleo reducido al cero,

$$\langle u, u \rangle_m = 0 \not\Rightarrow u = 0.$$

Demostración. Esta propiedad se demuestra de manera directa, probando la conmutatividad, la bi-linealidad y que

$$\forall u \in \mathcal{H}_m, \langle u, u \rangle_m \geq 0. \tag{2.43}$$

$$\langle u, u \rangle_m = 0 \Rightarrow u \in \mathcal{P}_{m-1}. \tag{2.44}$$

□

Consideremos el **subespacio vectorial** I_0 de H_m definido por

$$I_0 = \{u \in \mathcal{H}_m | u(x_i) = 0, \forall i = 1, 2, ..., n\} \tag{2.45}$$

y el traslado de I_0, el **subespacio afín**

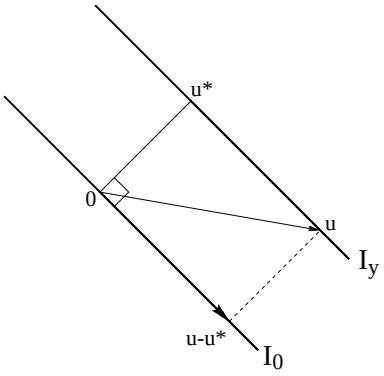
$$I_y = \{u \in \mathcal{H}_m | u(x_i) = y_i, \forall i = 1, 2, ..., n\}.$$

(RECUERDE I_0 es subespacio vectorial pues $u, v \in I_0 \Rightarrow (u - v) \in I_0$ e I_y es subespacio afín o trasladado de I_0 pues $u, v \in I_y \Rightarrow (u - v) \in I_0$)

La proyección ortogonal de la función nula o 0 de $\mathcal{H}_m$ en el subespacio afín I_y con respecto al semiproducto $\langle \cdot, \cdot \rangle_m$, que denotaremos u^*, se caracteriza (como es sabido) por

$$\begin{aligned} u^* &\in I_y \\ \langle u - u^*, u^* \rangle_m &= 0 \quad \forall u \in I_y, \end{aligned} \tag{2.46}$$

lo que se ilustra en el esquema que sigue



Como en toda proyección ortogonal, se tiene que ésta realiza una distancia mínima, es decir, si definimos la seminorma inducida por el semiproducto $|u|_m = \sqrt{\langle u, u\rangle_m}$, entonces la proyección ortogonal del cero en I_y se caracteriza por ser solución del problema de minimización

$$|u^*|_m = \min_{u \in I_y} |u|_m \tag{2.47}$$

A continuación veremos que $\sigma_m = u^*$ probando que σ_m satisface (2.46) y que es única.

Teorema 2.48. *La función Spline σ_m definida en* (2.40) *es la única solución del problema*

$$\min_{u \in I_y} |u|_m^2. \tag{2.49}$$

Demostración. Es evidente de la definición (2.40) que la polinomial por pedazos, σ_m tiene m-ésima derivada de cuadrado integrable e interpola los datos, es decir, pertenece a I_y y por lo tanto satisface la primera condición de (2.46). Para probar que satisface la condición de ortogonalidad que resta, consideremos una función $u \in I_y$, cualquiera, y calculemos el producto integrando por partes

$$\langle u - \sigma_m, \sigma_m\rangle_m = (u^{(m-1)} - \sigma_m^{(m-1)})(x)\sigma_m^{(m)}(x)|_a^b - \int_a^b (u^{(m-1)} - \sigma_m^{(m-1)})(x)\sigma_m^{(m+1)}(x)dx.$$

Pero en los extremos σ_m es un polinomio de grado $(m-1)$ y por consiguiente

$$\sigma_m^{(m)}(a) = \sigma_m^{(m)}(b) = 0,$$

lo que implica que el primer sumando se anula. Integrando por partes $(m-1)$ veces y repitiendo el argumento anterior, se obtiene que

$$\langle u - \sigma_m, \sigma_m\rangle_m = (-1)^{(m-1)} \int_a^b (u' - \sigma_m')(x)\sigma_m^{(2m-1)}(x)dx.$$

Como σ_m es un polinomio de grado $(2m-1)$ en cada intervalo $[x_i, x_{i+1}]$ su derivada $(2m-1)$-ésima será constante en cada uno de estos tramos y la llamaremos c_i, a su vez, en los extremos esta derivada es nula

pues σ_m es polinomial de grado $(m-1)$. De esto se concluye que

$$\begin{aligned}\langle u-\sigma_m,\sigma_m\rangle_m &= (-1)^{(m-1)}\sum_{i=1}^{n-1} c_i \int_{x_i}^{x_{i+1}} (u'-\sigma_m')(x)dx \\ &= (-1)^{(m-1)}\sum_{i=1}^{n-1} c_i\{u(x_{i+1})-\sigma_m(x_{i+1})-u(x_i)+\sigma_m(x_i)\} \\ &= 0,\end{aligned} \tag{2.50}$$

pues tanto u como σ_m interpolan los datos.

Hemos probado así que σ_m es la proyección ortogonal caracterizada por (2.46).

Probaremos ahora la unicidad.

Supongamos que hay 2 funciones que satisfacen (2.46): $\widetilde{u}, \hat{u}$. Como ambas pertenecen a I_y podemos intercambiar sus roles en la condición de ortogonalidad (segunda condición de (2.46)):

$$\begin{aligned}\langle \widetilde{u}-\hat{u},\hat{u}\rangle_m &= 0, \\ \langle \hat{u}-\widetilde{u},\widetilde{u}\rangle_m &= 0.\end{aligned}$$

De modo que

$$|\widetilde{u}-\hat{u}|_m^2 = \langle \widetilde{u}-\hat{u},\widetilde{u}-\hat{u}\rangle_m = \langle \widetilde{u},\widetilde{u}-\hat{u}\rangle_m - \langle \hat{u},\widetilde{u}-\hat{u}\rangle_m = 0,$$

lo que implica que

$$(\widetilde{u}-\hat{u}) \in \mathcal{P}_{m-1}.$$

Como ambas funciones $\widetilde{u}, \hat{u} \in I_y$, se tendrá que

$$(\widetilde{u}-\hat{u})(x_i) = 0,\ \forall i = 1,2,...,n$$

y considerando que $m \leq n$, se concluye que $(\widetilde{u}-\hat{u})$ es el polinomio nulo, lo que prueba la unicidad de $u^* = \sigma_m$.

De este modo se tendrá que σ_m es la única solución del problema (2.49). □

La interpretación de este resultado es particularmente clara en el caso $m = 2$, la popular Spline cúbica natural. El teorema (2.48),. en este caso, dice

$$\int_a^b (\sigma_2''(x))^2 dx \leq \int_a^b (u''(x))^2 dx \quad \forall u \in I_y,$$

lo que significa que de entre todas las funciones que interpolan y para las cuales el cuadrado de su segunda derivada sea integrable sobre $[a,b]$, σ_2 es la más plana posible, la de menores concavidades, la que menos se flecta, **la que minimiza la energía de flexión**. Si los datos provienen de una función f, es decir, si

$$f(x_i) = y_i,\ \forall i = 1,2,...,n,$$

entonces σ_2 será una buena aproximación de f si f tiene esta característica de tener energía de flexión pequeña. Por el contrario, si los datos muestran grandes concavidades de f, entonces la función Spline corresponderá a un modelo errado, que no se ajusta a un criterio acorde a la realidad. Esta condición, que hace depender la calidad de la aproximación de la adecuación del modelo a la realidad, debería reflejarse

en el estudio del error cometido por la función Spline. Efectivamente, el siguiente teorema explícita esta dependencia al presentar cotas del error que dependen de la energía de flexión de f.

Teorema 2.51. *1. Si f pertenece a $\mathcal{H}_m$, entonces el error cometido por la función Spline, medida en la seminorma de $\mathcal{H}_m$, satisface*

$$|f - \sigma_m|_m \leq |f|_m. \tag{2.52}$$

2. Si $f \in C^m_{[a,b]}$ y si h es el paso máximo de la malla T, es decir,

$$h = \max\{x_1 - a, b - x_n, \max_{1 \leq i \leq n-1}(x_{i+1} - x_i)\},$$

entonces $\forall k = 0, 1, ..., m-1$, existen constantes C_k tal que

$$\|f^{(k)} - \sigma_m^{(k)}\|_{\infty,[a,b]} \leq C_k h^{m-k} |f|_m. \tag{2.53}$$

Demostración. 1. Si $f \in \mathcal{H}_m$ entonces $f \in I_y$, y por (2.46) se tendrá que

$$\langle f - \sigma_m, \sigma_m \rangle_m = 0$$

y por lo tanto

$$\begin{aligned} |f - \sigma_m|_m^2 &= \langle f - \sigma_m, f - \sigma_m \rangle_m = \langle f, f - \sigma_m \rangle_m - \langle \sigma_m, f - \sigma_m \rangle_m \\ &= \langle f, f \rangle_m - \langle f, \sigma_m \rangle_m \\ &= |f|_m^2 - |\sigma_m|_m^2 \\ &\leq |f|_m^2, \end{aligned}$$

lo que prueba (2.52).

2. La continuidad de la función de error,

$$e = f - \sigma_m$$

y las condiciones de interpolación

$$e(x_i) = 0 \quad \forall i = 1, 2, ..., n$$

permitirán usar el teorema de Rolle reiteradamente para concluir que para $1 \leq k \leq m-1$ existen $(n-k)$ ceros distintos de la derivada k-ésima de la función de error e en $[a, b]$, que llamaremos $\xi_i^k, i = 1, ..., n-k$, tales que

$$e^{(k)}(\xi_i^k) = 0.$$

La distancia entre dos de estos ceros consecutivos del mismo nivel k es

$$\xi_{i+1}^k - \xi_i^k \leq (k+1)h.$$

Para x cualquiera en $[a, b]$, sea j el índice del cero de la derivada $(m-1)$-ésima de la función de error más cercano de x, es decir

$$|x - \xi_j^{m-1}| \leq mh.$$

Por el teorema fundamental del Cálculo se tendrá que

$$e^{(m-1)}(x) = \int_{\xi_j^{m-1}}^{x} e^{(m)}(t)dt$$

y por consiguiente

$$|e^{(m-1)}(x)| \leq mh \max_{t\in[\zeta_j^{m-1}x]} |e^{(m)}(t)|. \tag{2.54}$$

Sea $\hat{t}$ el punto donde se realiza el máximo anterior. Por la continuidad de la derivada m-ésima y su cuadrado, existirá una tolerancia $\delta > 0$ tal que

$$\forall t \in [\hat{t}-\delta, \hat{t}+\delta]$$
$$(e^{(m)}(t))^2 \geq \frac{(e^{(m)}(\hat{t}))^2}{2}.$$

Integrando entre $(\hat{t}-\delta)$ y $(\hat{t}+\delta)$ la desigualdad anterior se concluye que

$$2\delta(e^{(m)}(\hat{t}))^2 \leq 2\int_{\hat{t}-\delta}^{\hat{t}+\delta}(e^{(m)}(t))^2dt \leq 2\int_{a}^{b}(e^{(m)}(t))^2dt = 2|e|_m^2$$

y por lo tanto en (2.54) se tendrá

$$|e^{(m-1)}(x)| \leq \frac{mh}{\sqrt{\delta}}|e|_m.$$

Utilizando (2.52) se obtiene

$$|e^{(m-1)}(x)| \leq \frac{mh}{\sqrt{\delta}}|f|_m, \tag{2.55}$$

lo que prueba (2.53) si $m = 1$. Si $m \geq 2$, (2.55) prueba (2.53) con $k = m-1$. Para x cualquiera en $[a,b]$ se elige el cero de la derivada $(m-2)$-ésima más cercano de x. Supongamos que esto ocurre para el índice j de esta colección y utilizando nuevamente el teorema fundamental del Cálculo, se tendrá

$$|e^{(m-2)}(x)| \leq \int_{\xi_j^{m-2}}^{x} |e^{(m-1)}(t)|dt \leq \frac{m(m-1)h^2}{\sqrt{\delta}}|f|_m,$$

lo que prueba (2.53) con $k = m-2$.

Repitiendo este argumento m veces y utilizando la desigualdad (2.55) se obtiene finalmente que

$$\forall x \in [a,b] \qquad |e(x)| \leq \frac{m!h^m}{\sqrt{\delta}}|f|_m,$$

lo que concluye la demostración del teorema.

□

Observaciones. El teorema anterior entrega un resultado de **exactitud**:

si lo datos de interpolación provienen de una función $f \in \mathcal{P}_{m-1}$, entonces la interpolante Spline de orden m es exacta, es decir, no comete error, pues el error está acotado por $|f|_m$, que en este caso es cero.

La desigualdad (2.53) permite establecer la **convergencia** de la función Spline de orden m y de sus derivadas hasta el orden $(m-1)$ cuando se densifica la malla de modo tal que el paso máximo h tienda a cero. A diferencia de la convergencia de la interpolación polinomial en este caso no se requiere ni aumentar la hipótesis de regularidad de f ni aumentar el grado de los polinomios, pues m permanece fijo. Como ejemplo ilustrativo mencionamos la Spline cúbica que converge a la velocidad de h^2 a la función f cuando h tiende a cero. Su derivada en cambio, converge a la derivada de f a la misma velocidad con que h tiende a cero.

Construcción de la Spline Cúbica.

Para cada intervalo interior de la malla $[x_i, x_{i+1}], \forall i = 1, 2, ..., n-1$, se necesita construir un polinomio S_i,

$$\sigma_2|_{[x_i, x_{i+1}]} = S_i$$

de grado 3, para el cual se tienen 2 condiciones de interpolación en los nodos x_i, x_{i+1}. Las dos condiciones que faltan para tener su determinación única, son las condiciones de continuidad de ambas derivadas en los nodos, que serán los puntos donde se pegan dos polinomios distintos. La tabla de diferencias divididas con ambos nodos repetidos y donde los valores desconocidos de las derivadas en los nodos se reemplazan por parámetros, λ_i, permite escribir el polinomio de interpolación de Newton de grado 3:

(2.56) $\forall i = 1, 2, ..., n-1, \quad \forall x \in [x_i, x_{i+1}]$

$$S_i(x) = y_i + (x - x_i)\lambda_i + (x - x_i)^2 \frac{(m_i - \lambda_i)}{h_i} + (x - x_i)^2(x - x_{i+1})\frac{(\lambda_i - 2m_i + \lambda_{i+1})}{h_i^2},$$

donde se ha usado la notación siguiente para la diferencia dividida y el paso i-ésimo

$$m_i = f[x_i, x_{i+1}], \quad h_i = x_{i+1} - x_i.$$

Cualquiera sea el valor de los parámetros $\lambda_i, \forall i = 1, 2, ..., n$, al pegar todos estos polinomios se obtiene una cúbica por pedazos de derivada continua que interpola los datos. Esta propiedad de continuidad se extiende a los extremos, si en ellos ponemos las rectas:

(2.57)
$$\begin{aligned} \forall x \in [a, x_1] \quad & S_0(x) = y_1 + (x - x_1)\lambda_1, \\ \forall x \in [x_n, b] \quad & S_n(x) = y_n + (x - x_n)\lambda_n. \end{aligned}$$

Para que la polinomial por pedazos obtenida sea la Spline cúbica de interpolación solo falta escoger los parámetros de modo que la segunda derivada sea continua. Esto equivale a satisfacer las ecuaciones:

$$\forall i = 1, 2, ..., n \quad S''_{i-1}(x_i) = S''_i(x_i).$$

En resumen, la función Spline que se busca está dada por los polinomios $S_i, i = 0, 1, ..., n$, descritos en (2.56) y (2.57) con $\Lambda = (\lambda_i)_{i=1}^n$ solución del sistema tridiagonal y simétrico $A\Lambda = b$, donde A y b son como sigue:

$$A = \begin{bmatrix} \frac{2}{h_1} & \frac{1}{h_1} & & & & & \\ \frac{1}{h_1} & 2\left(\frac{1}{h_1} + \frac{1}{h_2}\right) & \frac{1}{h_2} & & & & \\ & \ddots & & & & & \\ & & \ddots & & & & \\ & & & \frac{1}{h_{n-2}} & 2\left(\frac{1}{h_{n-2}} + \frac{1}{h_{n-1}}\right) & \frac{1}{h_{n-1}} & \\ & & & & \frac{1}{h_{n-1}} & \frac{2}{h_{n-1}} \end{bmatrix}, b = \begin{pmatrix} 3\frac{m_1}{h_1} \\ 3\left(\frac{m_1}{h_1} + \frac{m_2}{h_2}\right) \\ \vdots \\ \vdots \\ 3\left(\frac{m_{n-2}}{h_{n-2}} + \frac{m_{n-1}}{h_{n-1}}\right) \\ 3\frac{m_{n-1}}{h_{n-1}} \end{pmatrix}.$$

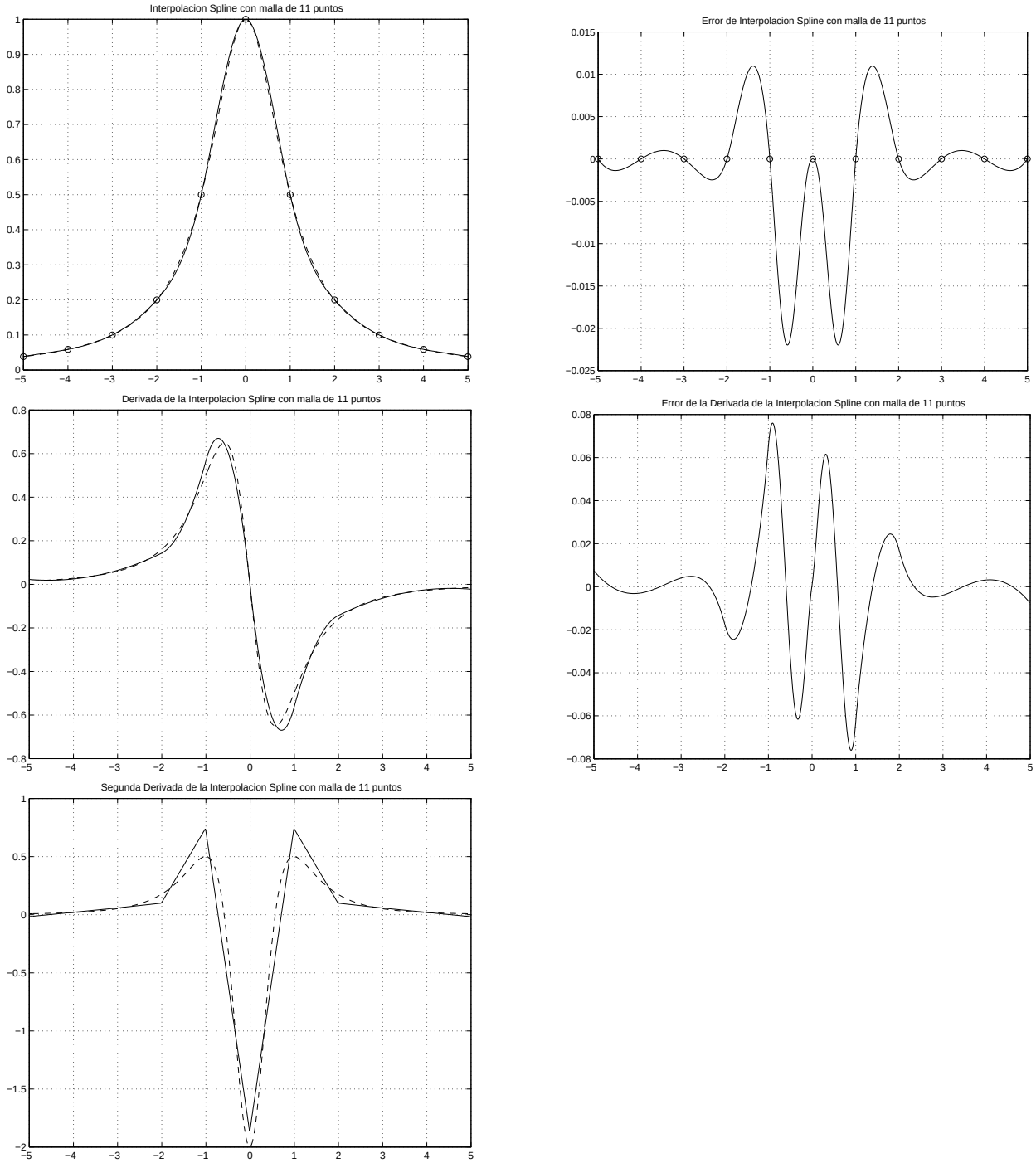


Figura 2.5: Interpolación Spline de la función $\frac{1}{1+x^2}$ con malla equiespaciada.

Esta técnica de construcción se generaliza fácilmente a otros órdenes m mayores que 2.

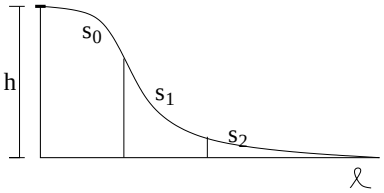
Problemas Resueltos.

1. Consideremos el problema de hacer un acceso a un puente de altura h sobre el camino plano a partir del punto que está a distancia ℓ del puente, que resolvimos en la sección anterior, pero ahora se pide que la suavidad sea la correspondiente a segunda derivada continua. Si llamamos S a la curva que representa la subida que se quiere construir, las condiciones de continuidad en los empalmes serán

$$\begin{aligned} S(0) &= h, & S(\ell) &= 0, \\ S'(0) &= 0, & S'(\ell) &= 0, \\ S''(0) &= 0, & S''(\ell) &= 0. \end{aligned}$$

Con estas 6 condiciones se puede construir un polinomio de grado 5, pero será inapropiado para establecer el largo mínimo ℓ para el cual la pendiente no excede la cota permitida. Estas consideraciones llevan a preferir una curva S cúbica por pedazos y para mayor comodidad pensamos en pedazos del mismo largo. Con 2 pedazos, es decir con 2 polinomios cúbicos que se pegan al centro del intervalo, no podremos satisfacer todas las condiciones de continuidad en el punto donde se juntan ambos, por lo que habrá que considerar 3 pedazos y 3 polinomios cúbicos:

$$\begin{aligned} \forall x \in [0, \frac{\ell}{3}] \quad & S(x) = s_0(x), \\ \forall x \in [\frac{\ell}{3}, \frac{2\ell}{3}] \quad & S(x) = s_1(x), \\ \forall x \in [\frac{2\ell}{3}, \ell] \quad & S(x) = s_2(x). \end{aligned}$$



Para construir s_0 de grado tres que satisfaga todas las condiciones en 0 (son 3) utilizamos una tabla de Diferencias Divididas con los datos:

$$\begin{array}{rcccc} x_i: & 0 & 0 & 0 & \ell/3 \\ f(x_i): & h & h & h & y_1 \end{array}$$

donde y_1 es el valor desconocido $y_1 = S\left(\frac{\ell}{3}\right)$.

El polinomio resultante es

$$s_0(x) = h + x^3(y_1 - h)\left(\frac{3}{\ell}\right)^3.$$

Al valor desconocido de S en $\frac{2\ell}{3}$ lo denotaremos por $y_2 = S\left(\frac{2\ell}{3}\right)$, con lo cual el polinomio de interpolación de Newton (atrás) de grado tres que se obtiene de la tabla de datos

$$\begin{array}{rcccc} x_i: & 2\ell/3 & \ell & \ell & \ell \\ f(x_i): & y_2 & 0 & 0 & 0 \end{array}$$

es

$$s_2(x) = -y_2(x-\ell)^3\left(\frac{3}{\ell}\right)^3.$$

Para garantizar la continuidad de la función S y de su primera derivada, construimos el polinomio cúbico s_1 con la tabla de datos

x_i :	$\frac{\ell}{3}$	$\frac{\ell}{3}$	$\frac{2\ell}{3}$	$\frac{2\ell}{3}$
$f(x_i)$:	y_1	y_1	y_2	y_2

Expresando las derivadas de los polinomios de los extremos en términos de los valores desconocidos y_1 e y_2, se obtiene

$$s_0'\left(\frac{\ell}{3}\right) = (y_1 - h)\left(\frac{9}{\ell}\right),$$
$$s_2'\left(\frac{2\ell}{3}\right) = -y_2\left(\frac{9}{\ell}\right),$$

con lo cual el polinomio cúbico que resta es

$$s_1(x) = y_1 + \left(x-\frac{\ell}{3}\right)(y_1-h)\left(\frac{9}{\ell}\right) + \left(x-\frac{\ell}{3}\right)^2(3y_2-12y_1+9h)\left(\frac{3}{\ell^2}\right) + \left(x-\frac{\ell}{3}\right)^2\left(x-\frac{2\ell}{3}\right)(15y_1-15y_2-9h)\left(\frac{9}{\ell^3}\right).$$

Las 2 incógnitas y_1 e y_2 se obtienen de las condiciones de continuidad de la segunda derivada:

$$s_0''\left(\frac{\ell}{3}\right) = s_1''\left(\frac{\ell}{3}\right) \text{ y } s_1''\left(\frac{2\ell}{3}\right) = s_2''\left(\frac{2\ell}{3}\right),$$

lo que se cumple con

$$y_1 = \frac{5}{6}h, \quad y_2 = \frac{1}{6}h.$$

Para obtener la máxima pendiente y su ubicación derivamos 2 veces todos los polinomios e igualamos a cero. Se encuentra que la pendiente más pronunciada se alcanza en $x = \frac{\ell}{2}$ y vale

$$s_1'(\frac{\ell}{2}) = -\frac{9h}{4\ell}.$$

De modo que si se quiere que estas pendientes no sean más pronunciadas que el ángulo de 45° se necesita que $\ell \geq \left(\frac{9}{4}\right)h$.

2. Dados los datos $\{(x_i, y_i)\}_{i=1}^n$ y el parámetro de ajuste ρ se define la Spline cúbica de ajuste asociada a estos datos como $\sigma_{2,\rho}$ tal que

   (a)
   - $\sigma_{2,\rho}|_{[x_i, x_{i+1}]} \in \mathcal{P}_3, \quad \forall i = 1, 2, ..., n-1,$
   - $\sigma_{2,\rho}|_{[a, x_1]} \in \mathcal{P}_1,$
   - $\sigma_{2,\rho}|_{[x_n, b]} \in \mathcal{P}_1,$

   (b) $\sigma_{2,\rho} \in \mathcal{C}^2_{[a,b]}$,

   (c) $\sigma^{(3)}_{2,\rho}(x_i^+) - \sigma^{(3)}_{2,\rho}(x_i^-) = 2\rho(y_i - \sigma_{2,\rho}(x_i))$.

i. Demuestre que $\sigma_{2,\rho}$ es solución del problema

$$P) \qquad \min_{u\in\mathcal{H}_m}\left\{\int_a^b (u^{(2)}(x))^2dx + \rho\sum_{i=1}^n (y_i - u(x_i))^2\right\}.$$

ii. Diseñe un método de construcción de esta aproximante.

Notemos que la solución del problema $P)$ será una función que minimiza un compromiso (medido por ρ) entre la distancia al cero de $\mathcal{H}_2$, en la seminorma 2, y la distancia a I_y, medida con la suma de los cuadrados. Para resolver la parte i) denotemos por u^* la solución de $P)$ y por β_i los valores, $\beta_i = u^*(x_i)\ \forall i = 1, 2, ..., n.$ resulta evidente de la definición de u^*, que $\forall u \in \mathcal{H}_2$, en particular, $\forall u \in \mathcal{H}_2$ tal que $u(x_i) = \beta_i, \forall i = 1, 2, ..., n.$

$$\int_a^b (u^{*(2)}(x))^2dx + \rho\sum_{i=1}^n (y_i - \beta_i)^2 \leq \int_a^b (u^{(2)}(x))^2dx + \rho\sum_{i=1}^n (y_i - u(x_i))^2,$$

lo que es equivalente a

$$|u^*|_2^2 \leq |u|_2^2 \quad \forall u \in I_\beta = \{u \in \mathcal{H}_2 | u(x_i) = \beta_i, \forall i = 1, 2, ..., n\},$$

debido a que los segundos sumandos se cancelan. Pero esta última desigualdad equivale a decir que u^* es la Spline cúbica de interpolación asociada a los valores $\beta_i, \forall i = 1, 2, ..., n$, y de ella ya sabemos que satisface a) y b). Por lo tanto podemos restringir la búsqueda de u^* minimizando solo entre aquellas funciones que satisfagan a) y b), para las cuales se tendrá, integrando por partes, que

$$\int_a^b (u^{(2)}(x))^2dx = \sum_{i=1}^n u(x_i)(u^{(3)}(x_i^+) - u^{(3)}(x_i^-)).$$

Reemplazando esto en el problema de minimización $P)$ se obtiene el problema equivalente en $\mathbb{R}^n$

$$\min_{\beta\in\mathbb{R}^n}\{\sum_{i=1}^n \beta_i(u^{(3)}(x_i^+) - u^{(3)}(x_i^-)) + \rho\sum_{i=1}^n (y_i - \beta_i)^2\}.$$

Derivando con respecto a β_i e igualando a cero se obtienen las ecuaciones de c).

Para la construcción de los polinomios cúbicos sobre los intervalos interiores usamos el mismo procedimiento que para la Spline de interpolación, es decir escribimos el polinomio de Newton de grado 3, con nodos repetidos, reemplazando ahora también los valores desconocidos por los parámetros β_i, con lo cual se obtiene

$$\begin{aligned}
\forall i = 1, 2, ..., n-1, \quad \forall x \in [x_i, x_{i+1}], \quad s_i(x) &= \beta_i + (x - x_i)\lambda_i + (x - x_i)^2\frac{m_i - \lambda_i}{h_i} \\
&\quad + (x - x_i)^2(x - x_{i+1})\frac{\lambda_i - 2m_i + \lambda_{i+1}}{h_i}, \\
\forall x \in [a, x_1] \quad s_0(x) &= \beta_1 + (x - x_1)\lambda_1, \\
\forall x \in [x_n, b] \quad s_n(x) &= \beta_n + (x - x_n)\lambda_n,
\end{aligned}$$

donde

$$m_i = \frac{\beta_{i+1} - \beta_i}{h_i} \quad \forall i = 1, 2, ..., n-1.$$

Esta polinomial por pedazos satisface a) y tiene hasta la primera derivada continua, cualquiera sean los valores de los parámetros $\{\beta_i\}_{i=1}^n, \{\lambda_i\}_{i=1}^n$. Para obtener la Spline de ajuste se deben escoger los parámetros de modo de satisfacer las condiciones c) y la continuidad de la segunda derivada. El sistema homogéneo que resulta de imponer estas $2n$ condiciones será pentadiagonal y simétrico (tridiagonal por bloques de 2 por 2), si se define el vector de incógnitas:

$$\Lambda = \begin{bmatrix} \beta_1 \\ \lambda_1 \\ \beta_2 \\ \lambda_2 \\ \vdots \\ \\ \vdots \\ \beta_{n-1} \\ \lambda_{n-1} \\ \beta_n \\ \lambda_n \end{bmatrix} \in I\!R^{2n}.$$

El vector de parámetros pedido es aquel que resuelve $A\Lambda = b$, con A y b dados por

$$A = \begin{bmatrix} \left(-\rho + \frac{6}{h_1^3}\right) & \frac{3}{h_1^2} & \frac{-6}{h_1^3} & \frac{3}{h_1^2} & & \\ \frac{3}{h_1^2} & \frac{2}{h_1} & \frac{-3}{h_1^2} & \frac{1}{h_1} & & \\ \frac{-6}{h_1^3} & \frac{-3}{h_1^2} & \left(-\rho + \frac{6}{h_1^3} + \frac{6}{h_1^3}\right) & 3\left(\frac{1}{h_2^2} - \frac{1}{h_1^2}\right) & \frac{-6}{h_2^3} & \frac{3}{h_2^2} \\ \frac{3}{h_1^2} & \frac{1}{h_1} & 3\left(\frac{1}{h_2^2} - \frac{1}{h_1^2}\right) & 2\left(\frac{1}{h_1} + \frac{1}{h_2}\right) & \frac{-3}{h_2^2} & \frac{1}{h_2} \\ & & & \ddots & & \\ & & & & \ddots & \\ & & & & & \frac{2}{h_{n-1}} \end{bmatrix}, b = \begin{bmatrix} -\rho \cdot y_1 \\ 0 \\ -\rho \cdot y_2 \\ 0 \\ \vdots \\ \\ \vdots \\ 0 \end{bmatrix}.$$

Ejercicios Propuestos.

1. Muestre con un ejemplo que la función Spline cúbica que definimos no es exacta para los polinomios de grado 3 ni siquiera sobre el intervalo $[x_1, x_n]$.

2. A continuación se define otra función Spline cúbica de interpolación que será exacta para los polinomios de grado 3, como se prueba en el desarrollo de este ejercicio. Sean $T = \{x_i\}_{i=0}^n$ una malla ordenada: $a = x_0 < x_1 < ... < x_n = b$, y los datos $y_i = f(x_i),\ \forall i = 0, 1, ..., n,\ z_0 = f'(x_0),\ \ z_n = f'(x_n)$.

   La función Spline que estudiaremos será S, definida por

   (a) $S(x_i) = y_i \quad \forall i = 0, 1, ..., n, \quad S'(x_0) = z_0, \quad S'(x_n) = z_n.$

   (b) $S|_{[x_i, x_{i+1}]} \in \mathcal{P}_3, \quad \forall i = 0, 1, ..., n-1.$

   (c) $S \in \mathcal{C}^2_{[a,b]}$.

   Respecto de esta función se pide:

   - Dé una construcción de S; exprese el polinomio cúbico $S|_{[x_i, x_{i+1}]}$, en términos de parámetros $\lambda_i = S'(x_i)$ que se obtenga resolviendo un sistema lineal.

- Suponga que $f \in \mathcal{H}_2$, considere el conjunto

$$\widetilde{I} = \{v \in \mathcal{H}_2 | v(x_i) = y_i, \forall i = 0, 1, ..., n,\ v'(x_0) = z_0,\ v'(x_n) = z_n\}$$

y demuestre que S es la proyección ortogonal del cero de $\mathcal{H}_2$ en $\widetilde{I}$, con respecto al semi-producto de $\mathcal{H}_2$. Identifique el problema optimal resuelto por S. Si $f \in \mathcal{H}_2$, acote el error $f - S$ en la seminorma de $\mathcal{H}_2$.

- Sean $h_i = x_{i+1} - x_i, \quad m_i = \frac{y_{i+1} - y_i}{h_i}, \quad \forall i = 0, 1, ..., n-1.$
$\forall x \in [x_i, x_{i+1}], \quad \forall i = 0, 1, ..., n-1$, pruebe que $S^{(3)}(x) = \frac{6}{h_i^2}(\lambda_i + \lambda_{i+1} - 2m_i)$, y que si $f \in C^4_{[a,b]}$, entonces

$$(*) \qquad |S^{(3)}(x) - f^{(3)}(x)| \leq \frac{6}{h_i^2}\ \{|\lambda_i - f'(x_i)| + |\lambda_{i+1} - f'(x_{i+1})|\} + hM_4$$

con $M_4 = \|f^{(4)}\|_{[a,b],\infty}$. (*Indicación: escriba de dos maneras distintas* $f[x_i, x_i, x_{i+1}, x_{i+1}]$ *y compare con* $S^{(3)}(x)$).
Si $f \in C^4_{[a,b]}$, pruebe que $\forall i = 1, 2, ..., n-1,\ \exists t_i^1, t_i^2, t_i^3, t_i^4 \in [x_{i-1}, x_{i+1}]$, tales que

$$\frac{e_{i-1}^1}{h_{i-1}} + 2e_i^1(\frac{1}{h_{i-1}} + \frac{1}{h_i}) + \frac{e_{i+1}^1}{h_i} = h_i^2(\frac{3}{4!}f^{(4)}(t_i^1) - \frac{1}{3!}f^{(4)}(t_i^4)) - h_{i-1}^2(\frac{3}{4!}f^{(4)}(t_i^2) - \frac{1}{3!}f^{(4)}(t_i^3))$$

con $e_i^1 = \lambda_i - f'(x_i)$.
(*Indicación: desarrolle en serie de Taylor* f *y su derivada* f'*; use las ecuaciones de continuidad de* S'').
Concluya que si la malla es uniforme de paso h, entonces

$$|e_i^1| \leq \frac{1}{2} \max_{1 \leq j \leq n-1} |e_j^1| + \frac{7}{48}h^3 M_4 \quad \forall i = 1, 2, ..., n-1.$$

Escribiendo esta última desigualdad para el índice i donde se alcanza el máximo y usándola en $(*)$ obtenga que

$$|S^{(3)}(x) - f^{(3)}(x)| \leq \frac{9}{2}hM_4 \quad \forall x \in [a, b].$$

Enuncie el teorema que acota el error en norma uniforme de

$$(S^{(k)} - f^{(k)}) \quad \forall k = 0, 1, 2, 3,$$

para el caso de malla equiespaciada.

APROXIMACIÓN DE MÍNIMOS CUADRADOS

Como vimos en el Problema Resuelto **2.** de la sección anterior, dados los datos $\{(x_i, y_i)\}_{i=1}^N$, un criterio de aproximación puede ser el de ajustar una curva que pase globalmente cerca de los datos sin que coincida necesariamente con alguno de ellos. Si V es un espacio vectorial de funciones, de dimensión finita n y $n \leq N$ (en la práctica $n << N$) la aproximación de mínimos cuadrados, a los datos dados, en V será la solución del problema optimal

$$\min_{v \in V} \sum_{i=1}^{N} (v(x_i) - y_i)^2. \tag{2.58}$$

Con el fin de caracterizar la solución de este problema, así como develar el modelo del cual surge, conviene mirarlo como un problema de proyección ortogonal en V. Consideremos el producto de funciones

$$\langle v, w\rangle = \sum_{i=1}^{N} v(x_i)w(x_i). \tag{2.59}$$

Resulta evidente que el producto así definido será conmutativo y bi-lineal, además de satisfacer $\langle v, v\rangle \geq 0$ cualquiera sea la función v (obviamente con un dominio que contenga los puntos $\{x_i\}_{i=1}^N$ y por lo tanto el producto esté bien definido). De modo que en el peor de los casos se tendrá al menos un semi-producto interno. Si se trabaja en un espacio de funciones donde además se tenga

$$\langle v, v\rangle = 0 \Rightarrow v = 0,$$

se tratará de un producto interno. Nótese que siempre se cumplirá

$$\langle v, v\rangle = 0 \Rightarrow v(x_i) = 0, \quad \forall i = 1, 2, ..., N.$$

Si g es una función tal que $g(x_i) = y_i, \quad \forall i = 1, 2, ..., N$, entonces, v^*, la solución del problema (2.58) será la proyección ortogonal de g en V con respecto al semi-producto dado en (2.59), y por lo tanto se podrá caracterizar en términos de una base de V, $\{v_j\}_{j=1}^n$ como sigue:

$$\begin{aligned} v^* &= \sum_{j=1}^{n} \alpha_j v_j \\ \langle g - v^*, v_i\rangle &= 0, \quad \forall i = 1, 2, ..., n. \end{aligned} \tag{2.60}$$

Las n ecuaciones de ortogonalidad se resumen en un sistema lineal cuyas incógnitas son los coeficientes $\alpha_j, \forall j = 1, 2, ..., n$, que caracterizan la proyección v^*.

De este modo, la aproximación de mínimos cuadrados se obtendrá calculando $\alpha = (\alpha_j)_{j=1}^n$ solución del sistema lineal simétrico $A\alpha = b$, con A y b dados por

$$A = \begin{bmatrix} \sum_{i=1}^{N} (v_1(x_i))^2 & \sum_{i=1}^{N} v_1(x_i)v_2(x_i) & \cdots & \cdots & \sum_{i=1}^{N} v_1(x_i)v_n(x_i) \\ \sum_{i=1}^{N} v_2(x_i)v_1(x_i) & \sum_{i=1}^{N} (v_2(x_i))^2 & \cdots & \cdots & \sum_{i=1}^{N} v_2(x_i)v_n(x_i) \\ \vdots & & & & \\ \vdots & & & & \\ \sum_{i=1}^{N} v_n(x_i)v_1(x_i) & \sum_{i=1}^{N} v_n(x_i)v_2(x_i) & \cdots & \cdots & \sum_{i=1}^{N} (v_n(x_i))^2 \end{bmatrix}, b = \begin{bmatrix} \sum_{i=1}^{N} v_1(x_i)y_i \\ \sum_{i=1}^{N} v_2(x_i)y_i \\ \vdots \\ \vdots \\ \sum_{i=1}^{N} v_n(x_i)y_i \end{bmatrix}.$$

Este sistema admite una factorización cuya utilidad se apreciará mejor en el capítulo de sistemas lineales. Sea B la matriz rectangular dada por

$$B_{i,j} = v_j(x_i) \quad \forall i = 1, 2, ..., N, \quad \forall j = 1, 2, ..., n.$$

Se comprueba fácilmente que

$$A = B^t B \text{ y } b = B^t y,$$

donde y es el vector de los valores $y = (y_i)_{i=1}^N$.

De esta factorización resulta directa la respuesta a la pregunta acerca de la solución única del problema planteado. La condición para que A sea definida positiva (y por lo tanto invertible) es que

$$\alpha^t A\alpha = 0 \Rightarrow \alpha = 0.$$

Pero

$$\alpha^t A\alpha = \alpha^t B^t B\alpha = \|B\alpha\|^2_{I\!R^N}$$

y para la norma euclidiana en $I\!R^N$ se tiene que

$$\|B\alpha\|^2_{I\!R^N} = 0 \Rightarrow B\alpha = 0.$$

Así se concluye que A será definida positiva si y solo si la matriz B es de rango completo n, es decir,

$$B\alpha = 0 \Rightarrow \alpha = 0,$$

lo que a su vez es equivalente a lo que se había observado antes, como condición para que el semi-producto fuera un auténtico producto interno en V,

$$v \in V, v(x_i) = 0, \quad \forall i = 1, 2, ..., N, \Rightarrow v = 0.$$

Esta condición se llama de *uni-solvencia*. Cuando esto ocurre se dice que la malla $T = \{x_i\}_{i=1}^N$ contiene un conjunto V-*unisolvente*.

En el caso particular en que V es el espacio de los polinomios de grado menor o igual que $(n-1)$, bastará con que los N puntos de la malla sean distintos y que $n \leq N$, para tener un conjunto $\mathcal{P}_{n-1}$-*unisolvente* y solución única del problema de mínimos cuadrados.

Cuando no importa lo mismo la distancia a cada dato, es decir si hay datos más confiables que otros a los que interesa ajustar mejor la curva de mínimos cuadrados, se consideran **pesos** que ponderen esta importancia relativa. Sean éstos los reales $p_i > 0$, $\forall i = 1, 2, ..., N$. La curva de **mínimos cuadrados** que ajusta los datos **con** los **pesos** dados, es $v^* \in V$, solución de

$$\min_{v \in V} \sum_{i=1}^{N} p_i (y_i - v(x_i))^2. \tag{2.61}$$

No representa mayor dificultad reconocer también aquí una proyección ortogonal en V y obtener el sistema lineal que permite calcular la solución en términos de una base $\{v_i\}_{i=1}^n$ de V:

$$v^* = \sum_{i=1}^{N} \alpha_i v_i$$

con $\alpha = (\alpha_i)_{i=1}^n$ solución del sistema lineal $A\alpha = b$, donde la matriz A y el vector de lado derecho b están dados por

$$\forall i = 1, 2, ..., n, \quad \forall j = 1, 2, ..., n$$
$$A_{i,j} = \sum_{k=1}^{N} p_k v_i(x_k) v_j(x_k), \quad b_i = \sum_{k=1}^{N} p_k v_i(x_k) y_k.$$

Este sistema admite también una factorización $A = B^t B$, $b = B^t \tilde{y}$, con la matriz B y el vector $\tilde{y}$ dados por

$$B_{i,j} = \sqrt{p_i} v_j(x_i), \quad \forall i = 1, 2, ..., N, \quad \forall j = 1, 2, ..., n.$$
$$\tilde{y}_i = \sqrt{p_i} y_i, \quad \forall i = 1, 2, ..., N.$$

Cuanto mayor sea un peso p_i, mayor será el beneficio aportado por la disminución de la distancia entre y_i y el valor de la curva en esa posición $v(x_i)$, a la minimización de (2.61), y por lo tanto más cerca del dato i-ésimo pasará la curva de mínimos cuadrados.

En un problema de mínimos cuadrados, el criterio de optimalidad (la distancia que se minimiza) está fijo, excepto por los pesos que se pueden asignar con información adicional. En cambio el subespacio de dimensión finita V puede ser elegido por del analista numérico, quién reconociendo patrones de comportamiento en los datos debe escoger las funciones de base de V, suficientes en número y tipo para dar cuenta de este comportamiento y permitir una buena aproximación. El número de oscilaciones de un polinomio queda limitado por su grado (o número máximo de raíces). Este también limita su velocidad de crecimiento. De manera que por ejemplo, datos rápidamente crecientes sugerirán incorporar exponenciales a la base de V y datos con oscilaciones periódicas sugerirán usar funciones como senos y cosenos.

Mínimos Cuadrados Continuos.

Cuando se conoce una **señal**, es decir una muestra muy abundante de la función f, se puede optar por un criterio continuo para construir una aproximación de ajuste polinomial. Esta será también una proyección ortogonal, esta vez en un espacio de polinomios y con los criterios de optimalidad inducidos por los productos internos que presentamos a continuación.

Sea D un dominio conexo en $\mathbb{R}$, sea w una función de peso sobre D, es decir, $w : D \to \mathbb{R}$, con $w(x) > 0, \forall x \in D$. Se define el espacio vectorial de funciones $L^2_{D,w}$ como

$$L^2_{D,w} = \left\{ v : D \to \mathbb{R} | \int_D w(x)(v(x))^2 dx < \infty \right\}.$$

Para las funciones en este espacio se puede definir el producto interno siguiente

$$\langle u, v \rangle_{L^2_{D,w}} = \int_D w(x)u(x)v(x)dx.$$

Si la señal considerada es una muestra de una función f que pertenece a este espacio, entonces la proyección ortogonal de f en el espacio de polinomios $\mathcal{P}_n$, relativa al producto interno dado, p^*, se caracterizará por

$$\begin{aligned} p^* &\in \mathcal{P}_n \\ \langle f - p^*, p \rangle_{L^2_{D,w}} &= 0, \quad \forall p \in \mathcal{P}_n. \end{aligned} \tag{2.62}$$

Como es sabido, este polinomio p^* será el que minimiza la distancia entre f y el espacio de los polinomios de grado menor o igual que n, medida en la norma inducida por el producto interno considerado, es decir, p^* es la solución del problema optimal

$$\min_{p \in \mathcal{P}_n} \|f - p\|_{L^2_{D,w}}, \tag{2.63}$$

donde la norma inducida es $\|g\|_{L^2_{D,w}} = (\langle g, g \rangle_{L^2_{D,w}})^{1/2}$.

Para varias combinaciones de funciones de peso w y de dominios D, se conocen bases ortogonales de $\mathcal{P}_n$. En estos casos la caracterización de la proyección p^* se simplifica pues se utilizan los coeficientes de Fourier.

Si $\{p_0, p_1, ..., p_n\}$ es base ortogonal de $\mathcal{P}_n$, entonces, la proyección, p^*, de f en $\mathcal{P}_n$, que satisface (2.62), está dada por

$$p^* = \sum_{i=0}^{n} \alpha_i p_i, \text{ con } \quad \alpha_i = \frac{\langle f, p_i \rangle_{L^2_{D,w}}}{\langle p_i, p_i \rangle_{L^2_{D,w}}}, \quad \forall i = 0, 1, ..., n.$$

A continuación se presentan los polinomios que forman bases ortogonales para algunos de estos productos internos.

Definición 2.64 (Polinomios de Legendre). Cuando

$$w(x) = 1, \quad \forall x \in D = [-1, +1],$$

la base ortogonal de $\mathcal{P}_n$, relativa a este producto interno, está dada por los polinomios que se definen recursivamente por

$$\begin{aligned} p_0(x) &= 1 \quad \forall x \\ p_1(x) &= x \quad \forall x \\ (n+1)p_{n+1}(x) &= (2n+1)x \cdot p_n(x) - np_{n-1}(x) \quad \forall x, \quad \forall n \geq 1. \end{aligned}$$

Definición 2.65 (Polinomios de Tchebycheff). Si $w(x) = \frac{1}{\sqrt{1-x^2}}, \quad \forall x \in D = [-1, +1]$, los polinomios ortogonales correspondientes son

$$p_n(x) = cos(n \cdot arcos(x)), \quad \forall x \in [-1, +1], \quad \forall n \geq 0,$$

los que a su vez satisfacen la recursión

$$\begin{aligned} p_0(x) &= 1 \\ p_1(x) &= x \\ p_{n+1}(x) &= 2x \cdot p_n(x) - p_{n-1}(x) \quad \forall n \geq 1. \end{aligned}$$

Definición 2.66 (Polinomios de Laguerre). Para el producto interno asociado a la función de peso

$$w(x) = e^{-x}, \quad \forall x \in D = [0, +\infty),$$

la base ortogonal de $\mathcal{P}_n$ está dada por los polinomios definidos por

$$\begin{aligned} p_0(x) &= 1 \quad \forall x \\ p_1(x) &= 1 - x \quad \forall x \\ p_{n+1}(x) &= (2n+1-x)p_n(x) - n^2 p_{n-1}(x), \quad \forall x, \quad \forall n \geq 1. \end{aligned}$$

Definición 2.67 (Polinomios de Hermite). Si

$$w(x) = exp(-x^2), \quad \forall x \in D = I\!R,$$

entonces la base ortogonal de $\mathcal{P}_n$, corresponde a los polinomios definidos por la recursión

$$\begin{aligned} p_0(x) &= 1, \quad \forall x \\ p_1(x) &= 2x, \quad \forall x \\ p_{n+1}(x) &= 2x \cdot p_n(x) - 2np_{n-1}(x) \quad \forall x \quad \forall n \geq 1. \end{aligned}$$

Si el intervalo en que se quiere aproximar la señal de f es cerrado y acotado del tipo $[a,b]$, entonces mediante un cambio de variables se lo puede llevar al intervalo $[-1,+1]$ y dependiendo si se quiere una mejor aproximación en los extremos o si se prefiere una aproximación que no asigne mayor importancia a ninguna zona del dominio se decidirá la función de peso conveniente, es decir, si se utilizan los polinomios de Tchebycheff o los de Legendre.

Si se quiere aproximar la señal de f sobre un dominio de alguno de los dos tipos $[a,+\infty)$ o $(-\infty,b]$ entonces se pueden usar los polinomios de Laguerre haciendo previamente el cambio de variables apropiado.

Los polinomios ortogonales presentados tienen muchas e interesantes propiedades, algunas de las cuales utilizaremos más adelante. Los polinomios de Tchebycheff resultan una herramienta muy útil en un problema clásico de aproximación que no abordamos en este curso: encontrar la **mejor aproximación en norma uniforme**, es decir en una norma que no proviene de un producto interno y por lo tanto en ausencia de una geometría que permita proyectar ortogonalmente. Esta falencia no puede inhibir la búsqueda de soluciones a este problema, ya que la distancia en norma uniforme es una medida fácilmente visualizable y comprensible del error de una aproximación. De hecho, fue ésta la razón de establecer teoremas que acotaran el error de las interpolantes presentadas (polinomios de interpolación y Splines de interpolación) en esa norma.

Para comprender esta propiedad de *uniformidad* de los polinomios de Tchebycheff, es útil graficarlos en $[-1,+1]$ y ver como, para cada grado ocupan el cuadrado $[-1,+1]\times[-1,+1]$ de manera óptima, con todos sus puntos críticos y todas sus raíces en él. Una propiedad general acerca de la distribución de las raíces de los polinomios que conforman bases ortogonales, que necesitaremos en el próximo capítulo, puede ser demostrada aquí, y sirve para ilustrar el comportamiento mencionado de los polinomios de Tchebycheff.

Teorema 2.68. *Si $\{p_i\}_{i=0}^n$ es base ortogonal de $\mathcal{P}_n$, y el grado de p_i es exactamente i, donde la ortogonalidad se relaciona con una función de peso w y un dominio D, es decir,*

$$\int_D w(x)p_i(x)p_j(x)dx = 0, \quad \forall i \neq j,$$

entonces p_n tiene exactamente n raíces simples en el interior de D.

Demostración. Sean $x_i \quad i = 1,2,...,m$, todos los puntos del interior de D donde p_n, cambia de signo. Obviamente $m \leq n$, debido al máximo número de raíces que puede tener un polinomio de grado n.

Sea $B(x) = \prod_{i=1}^{m}(x - x_i)$, un polinomio de grado m que cambiará de signo exactamente en los mismos lugares que p_n al interior de D, y por lo tanto la función producto $(B \cdot p_n)$ no cambiará de signo al interior de D.

Esto implica que

$$\int_D w(x)B(x)p_n(x)dx \neq 0. \tag{2.69}$$

Si $m < n$, entonces B se escribe como combinación lineal de polinomios $p_i, \forall i \leq m < n$, y por lo tanto B será ortogonal a p_n, lo que contradice (2.69). La única manera de evitar la contradicción, es con $m = n$, lo que concluye la demostración del teorema. □

Para calcular la proyección ortogonal contamos en general con bases **ortogonales** pero no **ortonormales** y por lo tanto en los coeficientes de Fourier aparecen denominadores (independientes de f) que se pueden calcular a priori. En el caso de la base de **Tchebycheff**, los denominadores son constantes (excepto para

$n = 0$, cuando vale π):

$$\langle p_n, p_n \rangle_{L^2_{D,w}} = \int_{-1}^{1} \frac{cos^2(n \cdot arcos(x))}{\sqrt{1-x^2}} dx = \int_0^{\pi} cos^2(n\theta) d\theta = \frac{\pi}{2}, \quad \forall n \geq 1,$$

con el cambio de variables $cos(\theta) = x$.

Utilizando el mismo cambio de variables se puede ver la expansión en serie de Tchebycheff de una función f de manera muy similar a la expansión de Fourier en serie de cosenos

$$p^*(x) = \sum_j \alpha_j cos(j \cdot arcosx),$$
$$p^*(cos\theta) = \sum_j \alpha_j cos(j \cdot \theta),$$

con $\alpha_j = \frac{2}{\pi} \int_0^{\pi} cos(j\theta) f(cos\theta) d\theta$.

Los polinomios de Legendre y los de Laguerre tienen interesantes expresiones en términos de derivadas que pueden resultar de utilidad al calcular integrales por partes. Los polinomios de **Legendre** definidos en (2.64) satisfacen

$$p_n(x) = \frac{(-1)^n}{2^n n!} \frac{d^n}{dx^n} \{(1-x^2)^n\} \quad \forall n \geq 1.$$

Además los denominadores de los coeficientes de Fourier serán en este caso

$$\langle p_n, p_n \rangle_{L^2_{D,w}} = \frac{2}{2n+1}.$$

Los polinomios de **Laguerre** definidos en (2.66) son **ortonormales** y satisfacen

$$p_n(x) = \frac{1}{n! exp(-x)} \frac{d^n}{dx^n} \{x^n exp(-x)\} \quad \forall n \geq 0.$$

Ejercicios Propuestos.

1. Calcule la aproximación de mínimos cuadrados continuos en el conjunto de los polinomios de grado menor o igual que 3, según el producto interno de Legendre y luego según el producto interno de Tchebycheff, de la función $f(x) = cos(4\pi \cdot x)$, para $x \in [-1, +1]$.

2. Suponga que tiene una nube de N puntos $\{(\theta_i, y_i)\}_{i=1}^N$, con $\theta_i \in [0, \pi], \forall i = 1, 2, ..., N$ distintos. Se quiere encontrar su aproximación de Mínimos Cuadrados en el conjunto $V = \{cosj\theta\}_{j=1}^3$. Bajo que condición ésta aproximación será única?

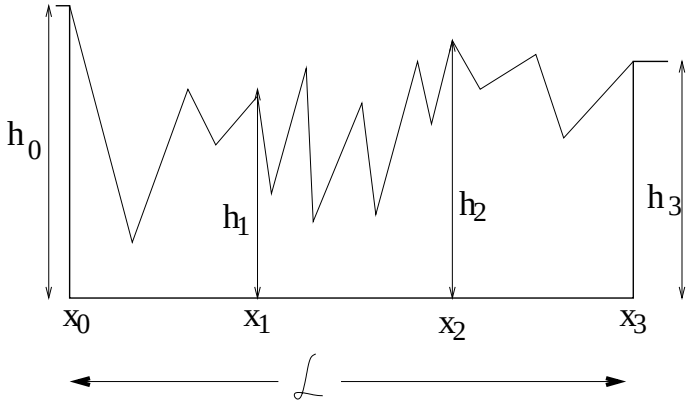
   Si $N = 4$, ¿puede caracterizar la solución, con funciones conocidas?

3. Considere el problema de calcular una Spline cúbica de ajuste con parámetro de ajuste ρ y con pesos p_i $\forall i = 1, 2, ..., n$. Cambie la definición dada antes (problema resuelto **2.-** de la sección anterior) de modo de obtener una función Spline que resuelva el problema

$$\min_{u \in \mathcal{H}_2} \left\{ \int_a^h (u^{(2)}(x)) dx + \rho \sum_{i=1}^n p_i (y_i - u(x_i))^2 \right\}.$$

   Como cambia el sistema lineal que permite su construcción en relación al que conocíamos?. Explicítelo para $n = 3$ y malla equiespaciada.

4. En la construcción de una autopista por un terreno montañoso se enfrenta el problema de unir 2 tramos separados de carretera elevada plana, uno a una altura h_0 y el otro a una altura h_3, separados por una distancia $\mathcal{L}$, medida horizontalmente, pero de valles profundos y cerros, que **no** se quiere interpolar. En dos puntos equidistantes entre si y de los tramos más próximos se conoce la altura exacta del terreno: h_1 y h_2. Para cada uno de estos puntos se conoce el costo (p_1 y p_2), en movimiento de tierras y/o de construcción de pilares de soporte, al no interpolar el terreno es decir, de que la autopista pase a distancia del terreno. Por otra parte la autopista debe tener una suavidad que corresponda al menos a una derivada continua (si es más suave que esto, tanto mejor). Modele este problema con las herramientas que maneja y encuentre un trazado de la autopista que minimice costos y satisfaga las condiciones impuestas.



CAPÍTULO 3

INTEGRACIÓN NUMÉRICA

El tema de este capítulo se vincula estrechamente con el del capítulo anterior y para muchos efectos pueden considerarse una sola unidad. El problema de aproximar $I(f) = \int_a^b f(x)dx$, se relaciona, de modo evidente, con el aproximar la función f sobre el dominio $[a, b]$. La estrategia que seguiremos será la de aproximar la cantidad $I(f)$ por la integral de una aproximante de la función f. Por consiguiente, el error de la fórmula de integración corresponderá a la integral del error de la aproximación de f, de la cual provenga y la regularidad que se supondrá de f, en cada caso, dependerá de las hipótesis que hayan permitido obtener la expresión del error de la mencionada aproximante. Las fórmulas de integración así obtenidas, serán todas del tipo

$$I_n(f) = \sum_{i=1}^{n} c_i f(x_i)$$

con coeficientes reales c_i y nodos $x_i \in [a, b]$.

Los métodos que estudiaremos se originan en interpolantes polinomiales o polinomiales por pedazos de la función f. Para el primer caso supondremos inicialmente mallas equiespacidas, obteniendo las fórmulas llamadas de **Newton Cotes**, y posteriormente elegiremos los nodos $\{x_i\}_{i=1}^n$ de una manera tal que la exactitud sea máxima, es decir, de modo que el espacio de polinomios para el cual el error de la fórmula se anula, sea lo más amplio posible. Estos métodos reciben el nombre de **Cuadratura Gaussiana**. Para la interpolación polinomial por pedazos solo usaremos mallas equiespaciadas y polinomios de grados bajos, obteniendo las **fórmulas compuestas** de **Trapecio** y **Simpson**. El estudio de la velocidad de convergencia de estas fórmulas permite introducir de manera muy natural el análisis del **comportamiento asintótico del error** y caracterizar las velocidades de convergencia mediante un indicador llamado **orden de convergencia**. Esto a su vez motiva la introducción a ciertos procedimientos clásicos de aceleración de convergencia, conocidos como **métodos de extrapolación**, cuyos beneficios se extenderán al problema de aproximación de raíces de funciones no lineales, abordado en el capítulo 6.

MÉTODOS DE NEWTON COTES

Del capítulo anterior sabemos que con n datos de interpolación, $y_i = f(x_i), i = 1, 2, ..., n$, sobre una malla equiespaciada $T = \{x_i\}_{i=1}^n$, con $x_i = a + (i-1)h$ y $h = \frac{b-a}{n-1}$, se construye un único polinomio de grado $(n-1)$ que interpola a la función f en los nodos x_i. La fórmula de Newton Cotes de n puntos, del tipo $I_n(f) = \sum_{i=1}^{n} c_i f(x_i)$, que aproxima la integral $I(f) = \int_a^b f(x)dx$, se obtiene integrando dicho polinomio de interpolación entre a y b.

Ejemplo 1. Para $n = 2$.

Si, $x_1 = a, x_2 = b$, la fórmula de Newton Cotes de 2 puntos correspondiente será

$$I_2(f) = \int_a^b \{f(a) + (x-a)\frac{f(b)-f(a)}{b-a}\}dx = \frac{b-a}{2}(f(a)+f(b)), \tag{3.1}$$

es decir, $c_1 = c_2 = \frac{b-a}{2}$.

Para acotar el error de la aproximación tendremos que remitirnos al error que comete el polinomio de interpolación de grado 1, del cual sabemos que, para poder expresarlo (y luego acotarlo) necesitamos que la función f, tenga al menos 2 derivadas continuas sobre $[a, b]$. En tal caso, $\exists \xi \in [a, b]$, tal que

$$I(f) - I_2(f) = \int_a^b (x-a)(x-b)f[a,b,x]dx = f[a,b,\xi]\int_a^b (x-a)(x-b)dx$$

y por lo tanto, $\exists \eta \in [a, b]$, tal que

$$I(f) - I_2(f) = -\frac{(b-a)^3}{12}f^{(2)}(\eta). \tag{3.2}$$

Ejemplo 2. Para $n = 3$ con malla equiespaciada, es decir, con paso $h = \frac{b-a}{2}$.

Integrando el polinomio de interpolación de Newton de grado 2 sobre la malla de tres puntos equiespaciados: $x_0 = a, x_1 = \frac{a+b}{2}, x_2 = b$, se obtiene la fórmula de Newton Cotes

$$\begin{aligned} I_3(f) &= \int_a^b \left\{ f(a) + (x-a)\frac{f(x_1)-f(a)}{h} + (x-a)(x-x_1)\frac{f(a)+f(b)-2f(x_1)}{2h^2} \right\} dx \\ &= \frac{h}{3}\{f(a) + 4f(x_1) + f(b)\}. \end{aligned} \tag{3.3}$$

El polinomio de interpolación de grado 2 del cual se deriva esta fórmula comete un error que depende de la tercera derivada de f, es decir, el error es nulo si f es un polinomio de grado 2. Si f es un polinomio de grado 3, el polinomio de interpolación de grado 2 antes mencionado cometerá error, pero su integral (la fórmula (3.3) será exactamente igual a $I(f)$.

Como ejercicio ilustrativo graficamos en la figura 3.1 la función $f(x) = 10x^3 + x^2$, sobre el intervalo $[-1, 1]$ y el polinomio de grado 2 que la interpola en $-1, 0, 1$. Comparando las áreas encerradas entre ambas curvas a ambos lados del origen, notamos que la integral del error de la interpolación se cancela. Esta mayor exactitud de la fórmula de integración se obtiene gracias al equiespaciamiento de la malla considerada.

Aparece aquí la primera cuestión interesante y propia de las fórmulas de integración de **Newton Cotes con un número** *impar* **de nodos** *equiespaciados*: **el error de interpolación que se cometerá al aproximar una función polinomial de grado** n **por un polinomio de interpolación de grado** $(n-1)$**, tendrá integral nula** y por lo tanto el error de la fórmula de integración dependerá de la derivada siguiente de f, es decir, $f^{(n+1)}$.

Demostraremos esta afirmación para la fórmula de Newton Cotes de 3 puntos equiespaciados.

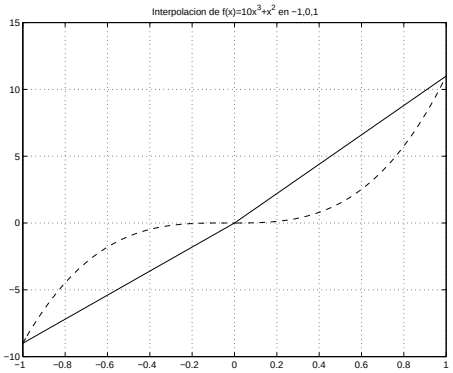


Figura 3.1: Interpolación de $f(x) = 10x^3 + x^2$ en $-1, 0, 1$, con el polinomio de grado 2, $p(x) = x^2 + 10x$

Supongamos que $f \in C^4_{[a,b]}$ y utilicemos los desarrollos de Taylor en torno a x_1

$$
\begin{aligned}
f(a) &= f(x_1) - hf'(x_1) + \frac{1}{2}h^2 f''(x_1) - \frac{1}{3!}h^3 f^{(3)}(x_1) + \frac{1}{4!}h^4 f^{(4)}(\eta_1) \quad \exists \eta_1 \in [a, x_1], \\
f(b) &= f(x_1) + hf'(x_1) + \frac{1}{2}h^2 f''(x_1) + \frac{1}{3!}h^3 f^{(3)}(x_1) + \frac{1}{4!}h^4 f^{(4)}(\eta_2) \quad \exists \eta_2 \in [x_1, b], \\
f(x) &= f(x_1) + (x - x_1)f'(x_1) + \frac{1}{2}(x - x_1)^2 f''(x_1) + \frac{1}{3!}(x - x_1)^3 f^{(3)}(x_1) + \\
&\quad \frac{1}{4!}(x - x_1)^4 f^{(4)}(\eta_3), \qquad \exists \eta_3 \in \overline{co}(x, x_1).
\end{aligned}
$$

Reemplazando estos desarrollos en (3.3) se obtiene

$$
\begin{aligned}
I(f) - I_3(f) &= \int_a^b \{\frac{1}{3!}f^{(3)}(x_1)[(x - x_1)^3 - (x - x_1)h^2] + \frac{1}{4!}v(x)\}dx \\
&= \frac{1}{4!}\int_a^b v(x)dx,
\end{aligned}
$$

con

$$
v(x) = (x - x_1)^4 f^{(4)}(\eta_3) - \frac{h^2}{2}f^{(4)}(\eta_2)[(x - x_1)^2 + h(x - x_1)] + \frac{h^2}{2}f^{(4)}(\eta_1)[h(x - x_1) - (x - x_1)^2],
$$

con lo cual se obtiene finalmente que $\exists \eta \in [a, b]$, tal que

$$
I(f) - I_3(f) = -\frac{h^5}{90}f^{(4)}(\eta). \tag{3.4}
$$

FÓRMULAS COMPUESTAS

La idea de aproximar una función mediante una curva polinomial por pedazos, se aplica también al problema de integración numérica. Los métodos más usados son **Trapecio** y **Simpson**. Ambas fórmulas se construyen a partir de mallas equiespaciadas: $T = \{x_i\}_{i=0}^n$ con $x_i = a + ih, \forall i = 0, 1, ..., n$, y $h = \frac{b-a}{n}$.

El método del **Trapecio** surge de aproximar f por una poligonal, es decir, una función lineal por pedazos que interpole a f sobre la malla T, y por lo tanto la fórmula de integración se obtiene repitiendo (y luego sumando) la fórmula de Newton Cotes con $n = 2$ calculada en (3.1) sobre cada tramo $[x_i, x_{i+1}]$, $\forall i = 0, 1, ..., n-1$, como se muestra en la figura 3.2.

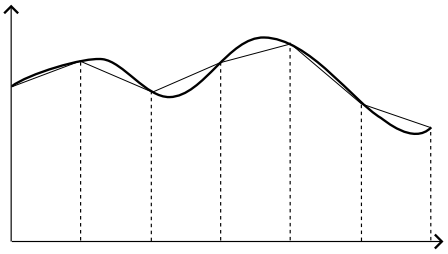


Figura 3.2: Ejemplo de interpolación para método del Trapecio

Si denotamos por $I_n^T(f)$ a la fórmula del Trapecio con $(n+1)$ puntos y por $I_{2,i}(f)$ a la fórmula de Newton Cotes con 2 puntos en el tramo $[x_i, x_{i+1}]$, entonces

$$I_n^T(f) = \sum_{i=0}^{n-1} I_{2,i}(f) = h\left\{\frac{f(x_0)+f(x_1)}{2} + \frac{f(x_1)+f(x_2)}{2} + ... + \frac{f(x_{n-1})+f(x_n)}{2}\right\} \tag{3.5}$$
$$= \frac{h}{2}\{f(x_0) + 2f(x_1) + 2f(x_2) + ... + 2f(x_{n-1}) + f(x_n)\}.$$

El error cometido por este método se obtiene también sumando los errores cometidos en cada tramo y dados por (3.2):

$$I(f) - I_n^T(f) = -\frac{h^3}{12}\sum_{i=0}^{n-1} f''(\eta_i) = -\frac{nh^3}{12}f''(\eta) = -\frac{(b-a)h^2}{12}f''(\eta), \tag{3.6}$$

considerando que si f'' es continua (lo que supusimos para obtener la fórmula (3.2), entonces debe existir un $\eta \in [a, b]$, donde se realice el promedio $f''(\eta) = \frac{1}{n}\sum_{i=0}^{n-1} f''(\eta_i)$, y recordando que $h = \frac{b-a}{n}$.

El método de **Simpson** resulta de aproximar f por una función cuadrática que la interpole, sobre cada pedazo de 3 nodos consecutivos tal como se aprecia en la figura 3.3.

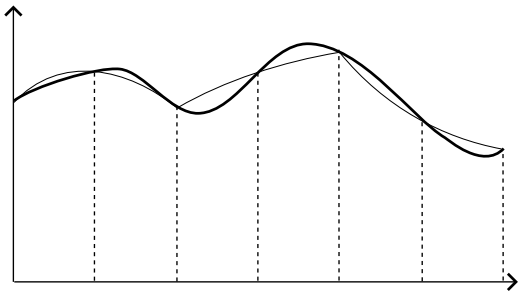


Figura 3.3: Ejemplo de interpolación para método de Simpson

(Cuidado; esta interpolante polinomial por pedazos no es una Spline y no se supone ninguna suavidad o continuidad de derivadas).

Para que esta estrategia de aproximación resulte adecuada se necesita que n sea par, es decir, un número impar de puntos de interpolación. Sobre cada tramo $[x_{i-1}, x_{i+1}]$ la integral de la polinomial corresponderá a la fórmula de Newton Cotes con 3 puntos equiespaciados y por lo tanto la fórmula de Simpson será

(3.7)

$$\begin{aligned} I_n^S(f) &= \frac{h}{3}\{f(x_0)+4f(x_1)+f(x_2)\}+\frac{h}{3}\{f(x_2)+4f(x_3)+f(x_4)\}+\ldots+\frac{h}{3}\{f(x_{n-2})+4f(x_{n-1})+f(x_n)\} \\ &= \frac{h}{3}\{f(x_0)+4f(x_1)+2f(x_2)+4f(x_3)+\ldots+2f(x_{n-2})+4f(x_{n-1})+f(x_n)\}. \end{aligned}$$

Sumando el error, dado por (3.4) para un tramo, sobre los $\frac{n}{2}$ tramos que cubren $[a,b]$, y razonando como hicimos con el error del Trapecio, se obtiene que si $f^{(4)}$ es continua sobre $[a,b]$, entonces el error de este método será

$$I(f) - I_n^S(f) = -\frac{(b-a)h^4}{180} f^{(4)}(\eta) \tag{3.8}$$

para algún $\eta \in [a,b]$.

Análisis Asintótico del Error.

Denotaremos por $E_n(f)$ al error de una fórmula de integración numérica que utilice los $(n+1)$ puntos equiespaciados $\{x_i\}_{i=0}^n \subset [a,b]$. Para Trapecio $E_n(f)$ está dado por (3.6) y para Simpson por (3.8).

Una expresión $\widetilde{E}_n(f)$ se dirá error asintótico del método numérico si

$$\lim_{n\to\infty} \frac{\widetilde{E}_n(f)}{E_n(f)} = 1.$$

El error de fórmulas compuestas recién presentadas puede ser mirado como una suma de Riemann cuyo comportamiento asintótico es conocido. Por ejemplo, el error cometido por el método del Trapecio, dado por (3.6), se puede expresar como

$$I(f) - I_n^T(f) = -\frac{h^2}{12}\sum_{i=0}^{n-1} h f''(y_i),$$

donde cada $\eta_i \in [x_i, x_{i+1}]$, y por lo tanto, la sumatoria corresponde a una suma de Riemann que aproxima a la integral $\int_a^b f''(x)dx$. De este modo, para n suficientemente grande se tendrá que el error del Trapecio será aproximadamente

$$I(f) - I_n^T(f) \approx -\frac{h^2}{12}[f'(b) - f'(a)]. \tag{3.9}$$

De manera similar se puede interpretar el error del método de Simpson dado por (3.8) como

$$I(f) - I_n^S(f) = -\frac{h^4}{180}\sum_{k=0}^{\frac{n}{2}-1} 2h f^{(4)}(\eta_{2k}),$$

donde $\eta_{2k} \in [x_{2k}, x_{2k+2}]$ y la sumatoria corresponde a la suma de áreas de rectángulos con base de largo $2h$ y altura igual al valor de $f^{(4)}$ en algún punto del intervalo que forma su base, es decir una suma de Riemann

que aproxima a $\int_a^b f^{(4)}(x)dx$. Por lo tanto para n suficientemente grande, el error del método de Simpson se puede aproximar como

$$I(f) - I_n^S(f) \approx -\frac{h^4}{180}[f^{(3)}(b) - f^{(3)}(a)]. \tag{3.10}$$

Se comprueba fácilmente que las expresiones del lado derecho de (3.9) y (3.10) corresponden a los errores asintóticos $\widetilde{E}_n(f)$ de los métodos de Trapecio y Simpson respectivamente.

La velocidad de convergencia de un método numérico se mide con un indicador que llamamos **orden de convergencia**. Se dirá que un método es de orden p si su error asintótico es

$$\widetilde{E}_n(f) = \frac{c}{n^p},$$

donde c es alguna constante. (Otra forma habitual de caracterizar esta velocidad de convergencia es con la expresión $0(h^p)$, que se lee como "o grande de h^p").

Las expresiones (3.9) y (3.10) permiten afirmar que si f es dos veces continuamente derivable en $[a, b]$ entonces la velocidad de convergencia de Trapecio es de orden $p = 2$ y si f es cuatro veces continuamente derivable, entonces el orden de convergencia del método de Simpson es cuatro.

En la práctica, cuando se dispone de varias aproximaciones numéricas del valor de una integral con un mismo método y mallas refinadas a la mitad del paso h (duplicando n) se puede estimar la velocidad de convergencia del método, observando la evolución de los cuocientes

$$\begin{aligned} R_n(f) = \frac{I_{2n}(f) - I_n(f)}{I_{4n}(f) - I_{2n}(f)} &= \frac{[I_{2n}(f) - I(f)] - [I_n(f) - I(f)]}{[I_{4n}(f) - I(f)] - [I_{2n}(f) - I(f)]} \approx \frac{-\widetilde{E}_{2n}(f) + \widetilde{E}_n(f)}{-\widetilde{E}_{4n}(f) + \widetilde{E}_{2n}(f)} \\ &= \frac{\frac{c}{n^p} - \frac{c}{(2n)^p}}{\frac{c}{(2n)^p} - \frac{c}{(4n)^p}} = 2^p. \end{aligned} \tag{3.11}$$

Es decir, calculando el logaritmo en base 2 de $R_n(f)$ para n suficientemente grande se tendría una aproximación del orden p.

En las tablas 3.1,3.2 y 3.3, mostramos el comportamiento de los métodos de Trapecio y de Simpson en tres ejemplos ilustrativos, para mallas que se refinan dividiendo el paso por 2 cada vez. El índice i que aparece en la primera columna de cada tabla, indica que en esa fila el paso corresponde a $h = \frac{b-a}{2^i}$, es decir que se trata de fórmulas con $n + 1 = 2^i + 1$, puntos.

Ejemplo 3. $\int_0^1 \exp(-x^2)dx = 0.74682413$. (ver tabla 3.1)

Ejemplo 4. $\int_0^1 x^{5/2}dx = 0.28571429$ (ver tabla 3.2)

Ejemplo 5. $\int_0^1 \sqrt{x}\ell n(x)dx = -0.44442296$ (ver tabla 3.3)

La función del ejemplo 3 satisface las hipótesis que permiten expresar el error tanto de Trapecio como de Simpson y por lo tanto la estimación del orden de convergencia aproxima bien en ambos casos la velocidad de convergencia esperada. La función del ejemplo 4 tiene solo dos derivadas continuas en el intervalo de integración y por lo tanto no se puede esperar que el método de Simpson converja con una velocidad caracterizada por un orden cuatro, pues esto requiere que la función sea 4 veces continuamente derivable. Se aprecia claramente en dicha tabla que Simpson, a pesar de que no se satisface la hipótesis **suficiente**

i	Trapecio	Estimación del orden de convergencia	Simpson	Estimación del orden de convergencia
0	0.68393972			
1	0.73137025		0.74718043	
2	0.74298410	2.02997022	0.74685538	
3	0.74586561	2.01094535	0.74682612	3.47369238
4	0.74658460	2.00280111	0.74682426	3.97312284
5	0.74676425	2.00070345	0.74682414	3.99523151
6	0.74680916	2.00017605	0.74682413	3.99891059
7	0.74682039	2.00004402	0.74682413	3.99973425
8	0.74682320	2.00001101	0.74682413	3.99994144
9	0.74682390	2.00000275	0.74682413	4.00005049
10	0.74682407	2.00000069	0.74682413	3.99641388

Tabla 3.1: Comportamiento de los métodos Trapecio y Simpson para $\int_0^1 \exp(-x^2)dx = 0.74682413$

i	Trapecio	Estimación del orden de convergencia	Simpson	Estimación del orden de convergencia
0	0.50000000			
1	0.33838835		0.28451780	
2	0.29879150	2.02907360	0.28559255	
3	0.28897474	2.01206737	0.28570249	3.28918871
4	0.28652857	2.00472072	0.28571318	3.36242933
5	0.28591778	2.00178163	0.28571418	3.40906483
6	0.28576515	2.00065775	0.28571428	3.43881057
7	0.28572700	2.00023946	0.28571428	3.45821566
8	0.28571746	2.00008638	0.28571429	3.47116250
9	0.28571508	2.00003097	0.28571429	3.47995105
10	0.28571448	2.00001106	0.28571429	3.48604404

Tabla 3.2: Comportamiento de los métodos Trapecio y Simpson para $\int_0^1 x^{5/2}dx = 0.28571429$

i	Trapecio	Estimación del orden de convergencia	Simpson	Estimación del orden de convergencia
0	-0.00046052			
1	-0.12299278		-0.16383687	
2	-0.32719181	- 0.73681433	-0.39525816	
3	-0.40039846	1.47992942	-0.42480068	2.96965666
4	-0.42757936	1.42938181	-0.43663966	1.31924885
5	-0.43793000	1.39287223	-0.44138022	1.32041417
6	-0.44192300	1.37417607	-0.44325400	1.33910376
7	-0.44347033	1.36769362	-0.44398611	1.35583307
8	-0.44406960	1.36849440	-0.44426936	1.36995139
9	-0.44430093	1.37329176	-0.44437804	1.38206843
10	-0.44438980	1.38012511	-0.44441942	1.39277058

Tabla 3.3: Comportamiento de los métodos Trapecio y Simpson para $\int_0^1 \sqrt{x}\ell n(x)dx = -0.44442296$

de convergencia, converge más rápido que Trapecio, quién converge a la velocidad esperada, pero no alcanza el orden 4. La función del ejemplo 5 no satisface las hipótesis suficientes de convergencia de ninguno de los dos métodos. Aún así, ambos muestran convergencia, pero no llegan a tener las velocidades de convergencia que se alcanzan cuando se satisfacen las hipótesis de regularidad correspondientes.

Dos criterios útiles para medir la calidad de un método numérico son la **velocidad de convergencia**, que acabamos de presentar, y la **exactitud**, es decir el espacio de funciones para el cual el método no comete error. Por ejemplo el método del Trapecio es exacto $\forall f \in \mathcal{P}_1$, el espacio de polinomios de grado menor o igual a uno, en cambio el método de Simpson es exacto $\forall f \in \mathcal{P}_3$.

A continuación se presentan dos procedimientos destinados a mejorar, en el primer caso, la velocidad de convergencia y en el segundo caso, la exactitud.

EXTRAPOLACIÓN; MÉTODO DE ROMBERG

Una estrategia fructífera para acelerar la convergencia consiste en combinar aproximaciones obtenidas con un mismo método numérico pero con mallas refinadas a la mitad del paso anterior. Ese procedimiento se conoce con el nombre de **extrapolación**.

Sea $I_i^0(f)$ la fórmula del Trapecio que aproxima $I(f) = \int_a^b f(x)dx$ con paso $h = \frac{b-a}{2^i}$, es decir, que utiliza $(2^i + 1)$ nodos equiespaciados: $x_j = a + jh \quad \forall j = 0, 1, ..., 2^i$. Consideremos una lista de estos valores para $i = 0, 1, ..., K$ y a partir de ella construiremos una segunda lista definida por

$$I_i^1(f) = \frac{4I_{i+1}^0 - I_i^0(f)}{3} \quad \text{para } i = 0, 1, ..., K-1.$$

Se comprueba fácilmente que la fórmula $I_i^1(f)$ corresponde al método de Simpson con paso $h = \frac{b-a}{2^{i+1}}$.

En efecto, para este h las fórmulas del Trapecio $I_i^0(f), I_{i+1}^0(f)$ con pasos $2h = \frac{b-a}{2^i}$ y h, respectivamente, serán

$$I_i^0(f) = h\{f(a) + 2f(a+2h) + 2f(a+4h) + ... + 2f(a+2^i h) + f(b)\},$$
$$I_{i+1}^0(f) = \frac{h}{2}\{f(a) + 2f(a+h) + 2f(a+2h) + ... + 2f(a+(2^{i+1}-1)h) + f(b)\}$$

y por lo tanto

$$I_i^1(f) = \frac{h}{3}\{f(a) + 4f(a+h) + 2f(a+2h) + ... + 4f(a+(2^{i+1}-1)h) + f(b)\}.$$

Se tiene así que la combinación propuesta de 2 fórmulas de Trapecio, que cometen errores que dependen de $(2h)^2$ y h^2 respectivamente, produce una fórmula de Simpson cuyo error depende de h^4, si $f \in C^4_{[a,b]}$ y por lo tanto convergerá más rápido si h decrece. La pregunta que surge de inmediato es acerca de la posibilidad de combinar apropiadamente 2 fórmulas de Simpson (de la lista $I_i^1(f)$) de modo de aumentar aún más la velocidad de convergencia. Esto es efectivamente posible si f tiene la regularidad suficiente. En general se puede definir el procedimiento recurviso

para $1 \le k \le K, \quad 0 \le i \le K - k$

$$I_i^k(f) = \frac{4^k I_{i+1}^{k-1}(f) - I_i^{k-1}(f)}{4^k - 1} \tag{3.12}$$

y confeccionar una tabla triangular de aproximaciones de $I(f)$ según el esquema de la tabla 3.4.

$$
\begin{array}{ccccccc}
I_0^0 & & & & & & \\
& I_0^1 & & & & & \\
I_1^0 & & I_0^2 & & & & \\
& I_1^1 & & \ddots & & & \\
I_2^0 & & I_1^2 & & \ddots & & \\
& I_2^1 & & & & \ddots & \\
I_3^0 & & \vdots & & & & I_0^K \\
& \vdots & & & & & \\
\vdots & & \vdots & & & & \\
& \vdots & & & & & \\
\vdots & & I_{K-2}^2 & & & & \\
& I_{K-1}^1 & & & & & \\
I_K^0 & & & & & &
\end{array}
$$

Tabla 3.4: Tabla de Romberg

En cada columna k se tiene un método que converge con $h^{2(k+1)}$ si $f \in C^{2(k+1)}_{[a,b]}$.

La aproximación de Romberg de $I(f)$, es I_0^K y comete un error que depende de h^{2K+2} con $h = \frac{b-a}{2^K}$.

Sin llegar a constituir una demostración, el desarrollo que sigue ilustra este comportamiento de ganar 2 órdenes de convergencia por cada columna que se agrega.

Toda la construcción se hace a partir de la fórmula del Trapecio, cuyo error hemos expresado antes. Cuando se analizó el comportamiento asintótico de éste (para $n+1$ nodos), observamos la presencia de una suma de Riemann:

$$E_n(f) = \frac{-h^2}{12} \sum_{j=0}^{n-1} h f''(\eta_j),$$

que puede ser aproximada por la expresión asintótica del error

$$\widetilde{E}_n(f) = \frac{-1}{12} h^2 (f'(b) - f'(a)).$$

Para comparar $E_n(f)$ con $\widetilde{E}_n(f)$ notamos que integrando 2 veces por partes se tiene

$$\begin{aligned}
\frac{1}{2} \int_{x_j}^{x_{j+1}} (x - x_j)(x - x_{j+1}) f''(x) dx &= \frac{-1}{2} \int_{x_j}^{x_{j+1}} (x - x_j + x - x_{j+1}) f'(x) dx \\
&= \int_{x_j}^{x_{j+1}} f(x) dx - \frac{1}{2} h (f(x_j) + f(x_{j+1}))
\end{aligned}$$

y por lo tanto, el error de Trapecio con $(n+1)$ nodos se puede expresar como

$$E_n(f) = \sum_{j=0}^{n-1} \frac{1}{2} \int_{x_j}^{x_{j+1}} (x - x_j)(x - x_{j+1}) f''(x) dx.$$

Como

$$\widetilde{E}_n(f) = \sum_{j=0}^{n-1} \frac{-1}{12} h^2 \int_{x_j}^{x_{j+1}} f''(x)dx,$$

conviene realizar la comparación para cada sumando.

$$\int_{x_j}^{x_{j+1}} f''(x) \left\{ \frac{(x-x_j)(x-x_{j+1})}{2} + \frac{h^2}{12} - \frac{h^2}{12} \right\} dx =$$

$$- \frac{h^2}{12}(f'(x_{j+1}) - f'(x_j)) + \int_0^h f''(u+x_j) \left\{ \frac{u(u-h)}{2} + \frac{h^2}{12} \right\} du.$$

Si la regularidad de f lo permite, integrando por partes esta última integral se tendrá,

$$= -\frac{h^2}{12}(f'(x_{j+1}) - f'(x_j)) + f''(u+x_j) \left[\frac{u^3}{6} - h\frac{u^2}{4} + h^2\frac{u}{12} \right] \Big|_{u=0}^{h} - \int_0^h f^{(3)}(u+x_j) \left[\frac{u^3}{6} - h\frac{u^2}{4} + h^2\frac{u}{12} \right] du.$$

Eliminando el segundo sumando, que se anula tanto en 0 como en h, e integrando nuevamente por partes se obtiene

$$= -\frac{h^2}{12}(f'(x_{j+1}) - f'(x_j)) - f^{(3)}(u+x_j) \left[\frac{u^4}{24} - h\frac{u^3}{12} + h^2\frac{u^2}{24} \right] \Big|_{u=0}^{h} + \int_0^h f^{(4)}(u+x_j) \left[\frac{u^4}{24} - h\frac{u^3}{12} + h^2\frac{u^2}{24} \right] du.$$

Nuevamente se anula el segundo sumando y debe ser eliminado. Para obtener una vez más un término que se anule en la integración por partes, sumamos y restamos la cantidad

$$\frac{h^4}{720} \int_0^h f^{(4)}(u+x_j)du = \frac{h^4}{720}(f^{(3)}(x_{j+1}) - f^{(3)}(x_j)),$$

con lo cual se tiene

$$\int_{x_j}^{x_{j+1}} \frac{1}{2} f''(x)(x-x_j)(x-x_{j+1})dx = -\frac{h^2}{12}(f'(x_{j+1}) - f'(x_j)) + \frac{h^4}{720}(f^{(3)}(x_{j+1}) - f^{(3)}(x_j))$$

$$+ \int_0^h f^{(4)}(u+x_j) \left\{ \frac{u^4}{24} - h\frac{u^3}{12} + h^2\frac{u^2}{24} - h^4\frac{1}{720} \right\} du.$$

Integrando nuevamente por partes (si la regularidad de f lo permite), se cancelará un término y se tendrá que la última integral es igual a

$$- \int_0^h f^{(5)}(u+x_j) \left\{ \frac{u^5}{120} - h\frac{u^4}{48} + h^2\frac{u^3}{72} - h^4\frac{u}{720} \right\} du,$$

lo que corresponde a un término en $f^{(5)}$ y h^6 que denotaremos por

$$-h^6 C_{nj}.$$

Sumando esta expresión obtenemos finalmente que si $f \in C^5_{[a,b]}$ el error del Trapecio se puede escribir como

$$E_n(f) = -\frac{h^2}{12}(f'(b) - f'(a)) + \frac{h^4}{720}(f^{(3)}(b) - f^{(3)}(a)) - h^6 C_n.$$

El procedimiento mediante el cual se llegó a esta expresión del error del método del Trapecio se puede continuar hasta donde la regularidad de f lo permita. En el teorema que sigue se explicitan las fórmulas del error así obtenidas.

Teorema 3.13. *Sean $m \geq 0$ y $n \geq 1, h = \frac{b-a}{n}, x_j = a + jh$, para $j = 0, 1, ..., n$. Si $f \in C^{2m+2}_{[a,b]}$, entonces, el error de la fórmula del Trapecio asociada a esta malla se puede expresar como*

$$E_n(f) = -\sum_{i=1}^{m} \frac{B_{2i}}{(2i)!} h^{2i}(f^{(2i-1)}(b) - f^{(2i-1)}(a)) + \frac{h^{2m+2}}{(2m+2)!}\int_a^b \bar{B}_{2m+2}(\frac{x-a}{h}) f^{(2m+2)}(x)dx, \tag{3.14}$$

donde $\bar{B}_j(x) = \begin{cases} B_j(x) & si\ 0 \leq x < 1 \\ B_j(x-1) & si\ x \geq 1 \end{cases}$ *es la extensión periódica del polinomio de Bernoulli de grado j definido implícitamente por*

$$\sum_{j=1}^{\infty} \frac{t^j}{j!} B_j(x) = \frac{t(e^{xt} - 1)}{e^t - 1}$$

y $B_j = -\int_0^1 B_j(x)dx$ son los números de Bernoulli.

Para una demostración formal de este Teorema se necesita usar las propiedades de estos polinomios y números, lo que escapa a nuestro interés. (Véase por ejemplo Ralston [6]). En cambio podemos utilizar la expresión (3.14) para analizar el error de las fórmulas que aparecen en la tabla de Romberg.

Sea

$$E^k_j(f) = I(f) - I^k_i(f),$$

el error de la fórmula de integración $I^k_i(f)$ de la tabla de Romberg. la expresión (3.14) con $n = 2^i$ corresponde así a $E^0_i(f)$ y por lo tanto

$$\begin{aligned} E^1_i(f) &= \frac{4E^0_{i+1} - E^0_i}{3} \\ &= \sum_{j=2}^{m} \frac{B_{2j}}{(2j)!}\left(\frac{b-a}{2^i}\right)^{2j}(f^{(2j-1)}(b) - f^{(2j-1)}(a))\frac{1}{3}\left(1 - \frac{4}{2^{2j}}\right) + \\ &\quad + \frac{1}{3}\cdot\frac{\left(\frac{b-a}{2^i}\right)^{2m+2}}{(2m+2)!}\int_a^b f^{(2m+2)}(x)\left(\frac{4}{2^{2m+2}}\bar{B}_{2m+2}\left(\frac{2^{i+1}(x-a)}{b-a}\right) - \bar{B}_{2m+2}\left(\frac{2^i(x-a)}{b-a)}\right)\right)dx. \end{aligned}$$

Como $B_4 = -\frac{1}{30}$, el término para $j = 2$ en la sumatoria resulta ser

$$\frac{-1}{180}\left(\frac{b-a}{2^{i+1}}\right)^4 (f^{(3)}(b) - f^{(3)}(a)),$$

que corresponde al error asintótico de Simpson con paso $\frac{b-a}{2^{i+1}}$ como se esperaba. Del mismo modo se observa que la expresión del error de la fórmula

$$I_i^2 = \frac{4^2 I_{i+1}^1 - I_i^1}{4^2 - 1}$$

comienza en $j = 3$ con el término

$$-\frac{2}{945}\left(\frac{b-a}{2^{i+2}}\right)^6 (f^{(5)}(b) - f^{(5)}(a)),$$

ya que el término de la sumatoria para $j = 2$ se cancela.

En resumen, si la regularidad de f lo permite, utilizando la expresión del error del método del Trapecio entregada por el teorema anterior, se puede probar que en la columna k-ésima de la tabla de Romberg se tiene una fórmula de integración que converge con h^{2k+2}. Es decir, **mediante el método de Romberg se acelera la convergencia**.

En las tablas que siguen mostraremos el comportamiento del método de Romberg en dos casos en los cuales la función tiene la regularidad suficiente. El índice i de la primera columna indica, al igual que en las tablas anteriores, que el paso correspondiente, de la fórmula del Trapecio de esa misma línea, es $h = \frac{b-a}{2^i}$. La diagonal superior de la tabla corresponde a las sucesivas aproximaciones de Romberg.

La tabla 3.5 corresponde al **Ejemplo 3** anterior

$$\int_0^1 \exp(-x^2)dx = 0.74682413.$$

$i \setminus k$	$k=0$ (Trapecio)	$k=1$ (Simpson)	$k=2$	$k=3$	$k=4$
0	0.68393972				
		0.74718043			
1	0.73137025		0.74683371		
		0.74685538		0.74682402	
2	0.74298410		0.74682417		<u>0.74682413</u>
		0.74682612		<u>0.74682413</u>	
3	0.74586561		<u>0.74682413</u>		0.74682413
		0.74682426		0.74682413	
4	0.74658460		0.74682413		0.74682413
		0.74682414		0.74682413	
5	0.74676425		0.74682413		
		<u>0.74682413</u>			
6	0.74680916				

Tabla 3.5: Comportamiento del método de Romberg para $\int_0^1 \exp(-x^2)dx = 0.74682413$

Ejemplo 6. $\int_{-4}^{4} \frac{1}{1+x^2}dx = 2.6516353$. (ver tabla 3.6)

En el capítulo anterior vimos (Figura 2.2) como la función con forma de campana, $f(x) = \frac{1}{1+x^2}$, si bien es infinitamente derivable, tiene derivadas cuya norma crece con el orden de derivación y por lo tanto sus polinomios de interpolación sobre mallas equiespaciadas empeoraron al duplicar el número de puntos

$i \setminus k$	0 (Trapecio)	1 (Simpson)	2	3	4	5	6
0	4.2352941						
		2.4784314					
1	2.9176471		2.5788235				
		2.5725490		2.6539203			
2	2.6588235		2.6527469		2.6518651		
		2.6477346		2.6518731		2.6516307	
3	2.6505068		2.6518868		2.6516309		2.6516353
		2.6516273		2.6516318		2.6516353	
4	2.6513472		2.6516358		2.6516353		2.6516353
		2.6516353		2.6516353		2.6516353	
5	2.6515633		2.6516353		2.6516353		2.6516353
		2.6516353		2.6516353		2.6516353	
6	2.6516173		2.6516353		2.6516353		2.6516353
		2.6516353		2.6516353		2.6516353	
7	2.6516308		2.6516353		2.6516353		2.6516353
		2.6516353		2.6516353		2.6516353	
8	2.6516342		2.6516353		2.6516353		
		2.6516353		2.6516353			
9	2.6516351		2.6516353				
		2.6516353					
10	2.6516353						

Tabla 3.6: Tabla de Romberg para $f(x) = \frac{1}{1+x^2}$

considerados. Como advertimos que el error de las fórmulas de integración que aparecen en la tabla de Romberg en la columna k-ésima depende de la derivada $2(k+1)$-ésima, entonces la esperada aceleración de la convergencia puede verse desfavorecida, en este caso.

CUADRATURA DE GAUSS

La clase de métodos que se presentan bajo este título tienen la virtud de **aumentar la exactitud**.

Sea

$$I(f) = \int_{-1}^{1} f(x)dx$$

y consideremos una fórmula de integración numérica

$$I_n(f) = \sum_{i=1}^{n} c_i f(x_i).$$

donde los coeficientes c_i **y los nodos** $x_i, i = 1, 2, ..., n$, son tales que la fórmula sea exacta para $f \in \mathcal{P}_{2n-1}$, es decir,

$$I(f) = I_n(f) \ \forall f \in \mathcal{P}_{2n-1}.$$

Si $n = 2$ se pueden encontrar los coeficientes y nodos que satisfacen este requisito, resolviendo el sistema **no lineal** que resulta de imponer la condición de exactitud sobre los polinomios de la base canónica de

$\mathcal{P}_3 : f(x) = 1; f(x) = x; f(x) = x^2; f(x) = x^3$:

$$\begin{aligned} c_1 + c_2 &= 2 \\ c_1 x_1 + c_2 x_2 &= 0 \\ c_1 x_1^2 + c_2 x_2^2 &= \frac{2}{3} \\ c_1 x_1^3 + c_2 x_2^3 &= 0. \end{aligned}$$

La solución de este sistema es

$$c_1 = c_2 = 1, \quad x_1 = \sqrt{\frac{1}{3}}, \quad x_2 = -\sqrt{\frac{1}{3}}$$

y en consecuencia la fórmula de cuadratura de Gauss (que para este problema recibe el nombre de Gauss-Legendre) para $n = 2$, es

$$I_2(f) = f(-\sqrt{\frac{1}{3}}) + f(\sqrt{\frac{1}{3}}). \tag{3.15}$$

Es poco práctico repetir este esquema para obtener las fórmulas de cuadratura para otros valores mayores de n, pues habría que resolver sistemas no lineales de complejidad creciente y de los cuales no tenemos garantías de existencia ni unicidad de soluciones. Probaremos en cambio que el polinomio de **interpolación de Hermite**, definido en (2.33) y (2.34) que utiliza como nodos de definición las n **raíces del polinomio de Legendre de grado** n, **definido en** (2.64), permite obtener la fórmula buscada para cualquier n.

Sabemos que el polinomio de Hermite $H_n(x)$, de grado $(2n-1)$, que interpola a f y f' en n nodos, $x_1, x_2, ..., x_n$, es exacto sobre $\mathcal{P}_{2n-1}$, es decir, si $f \in \mathcal{P}_{2n-1}$, entonces $\forall x \;\; H_n(x) = f(x)$, y en consecuencia

$$\int_a^b w(x) f(x) dx = \int_a^b w(x) H_n(x) dx,$$

cualquiera sean los límites de integración, a y b y la función de peso w.

Dicho de otro modo, integrando el polinomio de interpolación de Hermite basado en n nodos cualesquiera, se obtendría una fórmula de integración con la exactitud pedida. Pero como en la expresión del polinomio de Hermite intervienen los valores $f'(x_i)$, para $i = 1, 2, ..., n$, y por consiguiente estos valores aparecen en la expresión de la integral de dicho polinomio, la fórmula de integración obtenida no es del tipo exigido a $I_n(f)$, a menos que los nodos no sean cualesquiera, sino que sean tales que anulen aquella parte de la integral de H_n, en la que intervienen los valores indeseados de la derivada de f.

Teorema 3.16. *Sean $x_1, x_2, ..., x_n$, las n raíces (distintas, simples y todas en el interior del dominio D como probamos en el teorema (2.68)) de p_n, el polinomio de grado n de la base ortogonal de $\mathcal{P}_n$, donde la ortogonalidad consiste en*

$$\int_D w(x) p_i(x) p_j(x) dx = 0, \quad \forall i \neq j,$$

con w una función de peso.
Sea H_n el polinomio de Hermite basado en estos nodos que interpola a f y f'.
Si $I(f) = \int_D w(x) f(x) dx$ y si $I_n(f) = \int_D w(x) H_n(x) dx$, entonces

1. *$I_n(f)$ es una fórmula exacta sobre $\mathcal{P}_{2n-1}$,*
2. *$I_n(f) = \sum_{i=1}^{n} c_i f(x_i)$.*

Demostración. La exactitud de la fórmula $I_n(f)$ queda garantizada por el comentario previo. Para probar la segunda afirmación debemos recordar la expresión del polinomio de Hermite dada en (2.33)

$$H_n(x) = \sum_{i=1}^{n} f(x_i)h_i(x) + \sum_{i=1}^{n} f'(x_i)\tilde{h}_i(x),$$

donde los polinomios de grado $(2n-1)$, $h_i(x)$ y $\tilde{h}_i(x)$ satisfacen

$$\left.\begin{array}{l} h_i(x_k) = \delta_{ik} \\ h_i'(x_k) = 0 \\ \tilde{h}_i(x_k) = 0 \\ \tilde{h}_i'(x_k) = \delta_{ik} \end{array}\right\} \forall i,k = 1,2,...,n.$$

Para demostrar la parte 2) debemos probar que

$$\int_D w(x) \sum_{i=1}^{n} f'(x_i)\tilde{h}_i(x)dx = 0,$$

lo que se tendrá si

$$\forall i = 1,2,...,n, \quad \int_D w(x)\tilde{h}_i(x)dx = 0. \tag{3.17}$$

Sea $\Psi_n(x) = (x-x_1)(x-x_2)...(x-x_n)$.

Como los nodos $\{x_i\}_{i=1}^n$ son las raíces de $p_n(x)$, entonces necesariamente este polinomio admite esta misma factorización, es decir, existe una constante c tal que

$$\forall x \quad p_n(x) = c\Psi_n(x).$$

Por otra parte, el polinomio $\tilde{h}_i(x)$ se define en términos del i-ésimo polinomio de Lagrange de grado $(n-1)$ como

$$\tilde{h}_i(x) = (x-x_i)[l_i(x)]^2,$$

donde dicho polinomio de Lagrange a su vez se puede expresar según (2.37) como

$$l_i(x) = \frac{\Psi_n(x)}{(x-x_i)\Psi_n'(x_i)} = \frac{p_n(x)}{(x-x_i)p_n'(x_i)}.$$

De este modo se tiene que

$$\int_D w(x)\tilde{h}_i(x)dx = \frac{}{p_n'(x_i)} \int_D w(x)p_n(x)l_i(x)dx.$$

Pero esta última integral es nula cualquiera sea i, debido a que $p_n(x)$ es ortogonal a todos los polinomios de grado inferior a n y el i-ésimo polinomio de Lagrange, $l_i(x)$, es de grado $(n-1)$, lo que prueba (3.17) y permite concluir la demostración del teorema. □

Nota.

El teorema anterior identifica a **los nodos** $\{x_i\}_{i=1}^n$ **y los coeficientes** $\{c_i\}_{i=1}^n$ que permiten obtener la exactitud deseada de la fórmula de integración, es decir, tales que

$$\int_D w(x)f(x)dx = \sum_{i=1}^{n} c_i f(x_i) \quad \forall f \in \mathcal{P}_{2n-1}.$$

En efecto, de la demostración se concluye que

$$c_i = \int_D w(x)h_i(x)dx.$$

Del teorema anterior se deduce la fórmula del error de este método de integración numérica. Integrando el error cometido por el polinomio de interpolación de Hermite, dado en el teorema (2.38), se obtiene el corolario que sigue.

Corolario 3.18. Si f tiene $2n$ derivadas continuas sobre $[a,b]$, y si $I(f)$ e $I_n(f)$ denotan lo mismo que en el teorema (3.16), entonces existe $\eta \in (a,b)$, tal que

$$I(f) - I_n(f) = \frac{f^{(2n}(\eta)}{(2n)!}\int_D w(x)\prod_{i=1}^{i=n}(x-x_i)^2dx = \frac{f^{(2n)}(\eta)}{(2n)!}\frac{1}{c^2}\int_D w(x)(p_n(x))^2dx,$$

donde la integral corresponde al cuadrado de la norma del polinomio ortogonal $p_n(x)$.

Observaciones Prácticas.

A pesar de la nota anterior, el problema de calcular en cada problema los nodos y coeficientes de esta fórmula de cuadratura resultaría extraordinariamente tedioso. Como los casos para los cuales conocemos bases ortogonales de polinomios son solo cuatro (presentados en (2.64) ... (2.67)), se dispone de tablas que contienen tanto los nodos como los coeficientes de las fórmulas de cuadraturas respectivas para cada uno de estos casos en una variada gama de número total de nodos a utilizar. Por ejemplo, para la ortogonalidad de Legendre (es decir, $w(x) = 1,\ D = [-1,1]$), la tabla 3.7 entrega los nodos y coeficientes de las fórmulas con 2, 4 y 8 puntos.

n	x_i	c_i
2	± 0.5773502692	1
4	± 0.8611363116	0.3478548451
	± 0.3399810436	0.6521451549
8	± 0.9602898565	0.1012285363
	± 0.7966664774	0.2223810345
	± 0.5255324099	0.3137066459
	± 0.1834346425	0.3626837834

Tabla 3.7: Nodos y Coeficientes del método de Gauss Legendre

Cuando se tiene el problema de calcular numéricamente una integral donde los límites de integración no corresponden a los casos de ortogonalidad conocida se requerirá un cambio de variables previo a la utilización

de la fórmula de cuadratura correspondiente. Por ejemplo, si la integral que se quiere aproximar es

$$I(f) = \int_a^b f(x)dx,$$

con a y b finitos, entonces el cambio de variables correspondiente será

$$u = \frac{2x - a - b}{b - a},$$

de modo que se utilice la fórmula de cuadratura de Gauss-Legendre que aproxime a la integral que aparece en la expresión

$$I(f) = \frac{b-a}{2} \int_{-1}^{1} f\left(\frac{b-a}{2}u + \frac{a+b}{2}\right) du.$$

Ejemplo 7. Para ilustrar este procedimiento consideremos la integral entre 0 y $\pi/2$ del coseno, cuyo valor sabemos que es uno, es decir, sea

$$I(f) = \int_0^{\pi/2} cos(x)dx.$$

El cambio de variables propuesto nos lleva a la expresión

$$I(f) = \frac{\pi}{4} \int_{-1}^{1} cos(\frac{\pi}{4}u + \frac{\pi}{4})du$$

y la fórmula de Gauss-Legendre con solo 2 nodos (recuerde que para este caso obtuvimos la expresión (3.15)), entregará la aproximación

$$I_2(f) = \frac{\pi}{4}\{cos\left(-\frac{\pi}{4\sqrt{3}} + \frac{\pi}{4}\right) + cos\left(\frac{\pi}{4\sqrt{3}} + \frac{\pi}{4}\right)\}.$$

Utilizando la fórmula del coseno de la suma de los ángulos, y el valor

$$cos\left(\frac{\pi}{4\sqrt{3}}\right) = 0.8989,$$

se obtendrá la aproximación

$$I_2(f) = 0.9985,$$

que aproxima bien al valor exacto, 1, considerando las restricciones que nos impusimos a fin de ilustrar mejor el procedimiento. Para una mayor precisión usaremos los valores de la tabla anterior correspondientes a $n = 4$. Simplificando la expresión resultante mediante la fórmula del coseno de la suma, se obtiene

$$I_4(f) = \frac{\pi}{2\sqrt{2}}\left\{0.3478548451cos\left(0.861136116\frac{\pi}{4}\right) + 0.6521451549cos\left(0.3399810436\frac{\pi}{4}\right)\right\},$$

que comete un error de apenas $2.28^{*}10^{-8}$.

Los métodos de cuadratura de Gauss-Laguerre, Gauss-Hermite y Gauss Tchebycheff, se utilizarán para aproximar las integrales del tipo

$$\int_0^\infty e^{-x} f(x)dx, \quad \int_{-\infty}^{\infty} e^{-x^2} f(x)dx, \quad \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} f(x)dx,$$

respectivamente, donde los coeficientes y nodos correspondientes se encuentran en tablas contenidas en la mayoría de los textos en uso. Obviamente, para cualquier otra combinación de límites de integración, se deberá hacer el cambio de variables que permita hacer corresponder el/los límites finitos con el/los dados aquí.

Para comparar los métodos presentados en este capítulo, mostraremos el comportamiento del método de Cuadratura de Gauss Legendre en los cuatro ejemplos resueltos mediante el método de Romberg y las fórmulas compuestas de Trapecio y Simpson.

Número de puntos usados	$\int_0^1 \exp(-x^2)dx =$ 0.74682413	$\int_0^1 x^{5/2}dx =$ 0.28571429	$\int_0^1 \sqrt{x}\ell n(x)dx =$ - 0.44442296	$\int_4^4 \frac{1}{1+x^2}dx =$ 2.6516353
2	0.74659469	0.28645943	- 0.46268470	1.2631579
4	0.74682447	0.28571991	- 0.44887755	2.0472850
8	<u>0.74682413</u>	0.28571434	- 0.44531016	2.5600802
16	0.74682413	<u>0.28571429</u>	- 0.44459283	2.6498577
32	0.74682413	0.28571429	- 0.44446791	2.6516347

Tabla 3.8: Ejemplos anteriores usando el método de Gauss-Legendre

EJERCICIOS PROPUESTOS

1. Considere el método de integración numérica que resulta de integrar la función Spline cúbica sobre malla equiespaciada $\{x_i\}_{i=0}^{i=n}$, entre x_0 y x_n para aproximar

$$\int_{x_0}^{x_n} f(x)dx.$$

A partir de la expresión dada en el capítulo anterior en (2.56), obtenga la fórmula de integración, y compárela con las fórmulas compuestas conocidas.

(*Cuidado: ahora la malla parte en x_0 y no integramos sobre los extremos lineales de la Spline*).

Se prueba fácilmente que la función Splines Cúbica que estamos considerando, **no** es exacta para los polinomios de grado 3. En cambio si adicionalmente se imponen condiciones de interpolación de la primera derivada en los extremos, la función Spline resultante será exacta para los polinomios de grado 3 sobre el intervalo $[x_0, x_n]$. Recordando que en la expresión entregada en (2.56) el valor de la derivada de la función Spline en el nodo x_i, se denota por λ_i, proponemos considerar una función Spline como la anterior pero con las propiedades adicionales $\lambda_0 = f'(x_0)$ y $\lambda_n = f'(x_n)$. Obtenga la fórmula de integración resultante de integrar esta nueva función Spline. (Será exacta sobre $\mathcal{P}_3$).

2. Considere el método del Trapecio corregido que se obtiene si para cada intervalo $[x_i, x_{i+1}], \forall i = 0, 1, ..., n-1$, se aproxima la integral de f sobre ese intervalo por

$$\frac{h}{2}[f(x_i) + f(x_{i+1})] + \frac{h^2}{12}[f'(x_i) - f'(x_{i+1})].$$

Se supone malla equiespaciada ordenada con $a = x_0 < x_1 < ... < x_n = b$, y conocidos todos los valores $f(x_i), f'(x_i), \forall i = 0, 1, ..., n$. Encuentre la exactitud de esta fórmula, dando una cota del error. Compare con **todos** los métodos estudiados antes y los dos del problema anterior.

3. Obtenga una fórmula de Simpson mejorada que resulte de integrar sobre cada intervalo $[x_i, x_{i+2}]$, un polinomio de Hermite que interpole a f y a f', en los 3 nodos x_i, x_{i+1}, x_{i+2}. Se supone malla equiespaciada, n par y conocidos los valores de f y f' sobre la malla. Obtenga una expresión para el error de esta fórmula.

4. Exprese el error de Gauss-Legendre y Gauss-Tchebycheff con n puntos, usadas para aproximar integrales sobre un intervalo $[a, b]$ no necesariamente $[-1, 1]$.

5. Considere una malla equiespaciada de paso $h, a = x_0 < x_1 < ... < x_n = b$, sobre la cual se conocen los valores de una función f cuya integral se desea aproximar por una fórmula que resulte de aproximar f por una función constante por pedazos. Se proponen 2 alternativas:

$$P(x)|_{[x_i,x_{i+1}]} = \begin{cases} f(x_i), & \text{si } x_i \le x < x_i + \frac{h}{2} \\ f(x_{i+1}), & \text{si } x_i + \frac{h}{2} \le x < x_{i+1}, \end{cases}$$
$$S(x)|_{[x_i,x_{i+1}]} = \frac{f(x_i) + f(x_{i+1})}{2}.$$

Encuentre ambas fórmulas, calcule el error en cada caso, compare con métodos conocidos, ilustre en un gráfico la exactitud de estos métodos.

6. Con la ayuda de una calculadora confeccione tablas de Romberg para aproximar

(a) $I = \int_0^1 e^x dx$,

(b) $F = \int_0^{10} \frac{1}{1+x^2} dx$,

con pasos $h = \frac{b-a}{2}, \frac{h}{2}, \frac{h}{4}, \frac{h}{8}, \frac{h}{16}$. $(I = 1.7182818... \quad , F = 1.4711277...)$
Calcule el orden o velocidad de convergencia en ambos casos. Explique el comportamiento distinto del error en ambos casos.

7. Use un método adecuado de Cuadratura de Gauss para aproximar las integrales que siguen, con fórmulas exactas en $\mathcal{P}_3$ y luego en $\mathcal{P}_4$.

$$\int_{-2}^{3} \frac{1}{(x^2+1)^2} dx, \qquad \int_{-1}^{1} \frac{cos(\pi \cdot y)}{\sqrt{1-y^2}} dy, \qquad \int_0^1 \frac{x^2}{\sqrt{x(1-x)}} dx, \qquad \int_0^1 exp(x^2) dx.$$

8. ¿Cuántos puntos son necesarios para calcular con una precisión de 0.001 por Trapecio y cuantos para calcular con la misma precisión por Simpson la integral $\int_0^1 e^x sen3x dx$? Calcule con ambos métodos con 5 puntos equiespaciados y compare con el valor exacto.

CAPÍTULO 4

SISTEMAS LINEALES

Los sistemas de ecuaciones lineales han sido estudiados previamente en el curso de Álgebra Lineal, desde múltiples perspectivas. Nuestro interés aquí se reduce a los aspectos numéricos de su resolución. Es decir, siempre supondremos que su solución existe y es única, o equivalentemente, que la matriz del sistema es invertible. Además de los métodos orientados al cálculo efectivo de la solución del sistema lineal deberemos estudiar la propagación de errores, la estabilidad del problema y la convergencia de sucesiones de aproximaciones de la solución. Este objetivo nos lleva a la necesidad de introducir una sección inicial dedicada a la presentación de las normas de matrices adecuadas a nuestro problema, como lo son aquellas que se relacionan con (o *"subordinan"* a) normas vectoriales.

ESTABILIDAD Y NORMAS MATRICIALES SUBORDINADAS

Definición 4.1. Sea $\|\ \|_p$ una norma en $\mathbb{R}^n$. Se define la norma matricial subordinada a esta norma vectorial como

$$\forall A \text{ matriz real cuadrada de tamaño } n \qquad \|A\|_p = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Ax\|_p}{\|x\|_p} = \sup_{\substack{x \in \mathbb{R}^n \\ \|x\|_p = 1}} \|Ax\|_p.$$

Es bastante sencillo comprobar que la función del espacio de las matrices (reales cuadradas de tamaño n) en los reales no negativos, así definida, es efectivamente una norma y se propone como ejercicio.

Ejemplo 1. Consideremos la norma infinito en $\mathbb{R}^n$, es decir, $\forall x \in \mathbb{R}^n, \|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$. La norma matricial subordinada correspondiente será, por definición,

$$\|A\|_\infty = \sup_{\|x\|_\infty = 1} \max_{1 \leq i \leq n} \left| \sum_{j=1}^{n} A_{i,j} x_j \right|.$$

Utilizando la propiedad de la desigualdad triangular del módulo, se tendrá

$$\|A\|_\infty \leq \sup_{\|x\|_\infty=1} \max_{1\leq i\leq n} \sum_{j=1}^{n} |A_{i,j}\|x_j| \leq \sup_{\|x\|_\infty=1} \max_{1\leq i\leq n} \left(\max_{1\leq j\leq n} |x_j|\right) \sum_{j=1}^{n} |A_{i,j}| \tag{4.2}$$
$$= \max_{1\leq i\leq n} \sum_{j=1}^{n} |A_{i,j}|\,.$$

Por otra parte, dada una matriz A no es difícil construir un vector $\tilde{x} \in \mathbb{R}^n$ cuyas coordenadas sean ± 1 (y por lo tanto su norma infinito sea igual a 1) y tal que

$$\|A\tilde{x}\|_\infty = \max_{1\leq i\leq n} \left|\sum_{j=1}^{n} A_{i,j}\tilde{x}_j\right| = \max_{1\leq i\leq n} \sum_{j=1}^{n} |A_{i,j}|.$$

Con lo cual en (4.2) se tiene $\|A\|_\infty \leq \|A\tilde{x}\|_\infty$. Pero al estar definida la norma matricial como un supremo sobre todos los vectores unitarios de $\mathbb{R}^n$, se cumple la desigualdad inversa $\|A\tilde{x}\|_\infty \leq \|A\|_\infty$ y por consiguiente la igualdad

$$\|A\|_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |A_{i,j}|. \tag{4.3}$$

De modo similar se puede probar que

$$\|A\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{n} |A_{i,j}|, \tag{4.4}$$
$$|A\|_2 = \sqrt{\max_{\lambda\in\sigma(A^tA)} \lambda},$$

donde $\sigma(A^tA)$ es el espectro, o conjunto de todos los valores propios, de la matriz simétrica y semidefinida positiva A^tA. Si A es invertible, como lo será en todos los sistemas lineales que trataremos, entonces A^tA es definida positiva.

Definición 4.5. Sea A una matriz real y cuadrada de tamaño n. Se define el **radio espectral** de A, que denotaremos por $\rho(A)$, como el módulo del mayor valor propio en módulo de A. Es decir,

$$\rho(A) = \max_{\lambda\in\sigma(A)} |\lambda|.$$

El radio espectral **no** es una norma matricial y sin embargo juega un rol muy parecido a las normas matriciales subordinadas. Su relación con éstas quedará establecida en las propiedades que siguen.

Propiedades de las normas subordinadas.

4.6. *Para cualquier matriz A, para cualquier vector $x \in \mathbb{R}^n$, la norma vectorial y su correspondiente norma matricial subordinada satisfacen*

$$\|Ax\| \leq \|A\| \cdot \|x\|.$$

4.7. *Para cualquier par de matrices A y B, para cualquier norma matricial subordinada*

$$\|AB\| \leq \|A\| \cdot \|B\|.$$

4.8. *Para cualquier norma matricial subordinada, se tiene que la norma de la matriz identidad vale 1, es decir,*

$$\|I\| = 1.$$

4.9. *Para cualquier norma subordinada, para cualquier matriz A*

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

4.10. *Para cualquier tolerancia $\varepsilon > 0$, para cualquier matriz A existe una norma matricial subordinada (de sofisticada construcción) para la cual se cumple*

$$\rho(A) \geq \|A\| - \varepsilon.$$

Esta propiedad (se trata en realidad de un teorema) combinada con la anterior dicen que el radio espectral se puede aproximar tanto como se desee a una norma matricial subordinada.

4.11. *Si A es una matriz invertible entonces toda norma subordinada satisface*

$$\|A^{-1}\| \geq \frac{1}{\|A\|}.$$

Demostración. Las propiedades (4.6), (4.7) y (4.8) se prueban fácilmente por definición de norma subordinada. Demostraremos solo (4.7), dejando de ejercicio las demostraciones de (4.6) y (4.8).

$$\|AB\| = \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|ABx\|}{\|x\|}.$$

Por la propiedad (4.6) se tiene que $\|ABx\| \leq \|A\| \cdot \|Bx\|$ y por lo tanto

$$\|AB\| \leq \|A\| \sup_{\substack{x \in \mathbb{R}^n \\ x \neq 0}} \frac{\|Bx\|}{\|x\|} = \|A\| \cdot \|B\|.$$

Para probar la propiedad (4.9) usaremos (4.6). Sea $\lambda \in \sigma(A)$ tal que $|\lambda| = \rho(A)$ y sea $x \in \mathbb{R}^n$ vector propio de A asociado al valor propio λ, es decir,

$$Ax = \lambda x$$

con $x \neq 0_{\mathbb{R}^n}$. Esto implica que $|\lambda| \cdot \|x\| = \|\lambda x\| = \|Ax\| \leq \|A\| \cdot \|x\|$ y dividiendo por $\|x\| \neq 0$ se concluye el resultado.

Para una demostración de la propiedad (4.10) se sugiere consultar el libro de Ciarlet [3]. La propiedad (4.11) es consecuencia directa de las propiedades (4.7) y (4.8). En efecto,

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\| \cdot \|A^{-1}\|.$$

□

El primer aspecto numérico de los sistemas lineales que abordaremos será el problema de la estabilidad, es decir, nos preguntaremos acerca de cuanto influye en el resultado (la solución) una pequeña perturbación de los datos del problema. Consideremos el sistema

$$Ax = b \tag{4.12}$$

cuya solución denotaremos por $\tilde{x}$.

Nuestro objetivo es estudiar el desplazamiento de la solución cuando se perturba la matriz A y/o el lado derecho b. Comenzaremos por perturbar el lado derecho, es decir, consideremos el sistema

$$Ax = \hat{b} \tag{4.13}$$

cuya solución denotaremos por $\hat{x}$. Las cantidades que nos interesa comparar son

$$\frac{\|\tilde{x} - \hat{x}\|}{\|\tilde{x}\|} \text{ y } \frac{\|b - \hat{b}\|}{\|b\|}.$$

Como $\tilde{x} - \hat{x} = A^{-1}(b - \hat{b})$, se tendrá para cualquier norma subordinada que

$$\|\tilde{x} - \hat{x}\| \leq \|A^{-1}\| \cdot \|b - \hat{b}\|. \tag{4.14}$$

Por otra parte

$$A\tilde{x} = b \Rightarrow \|b\| \leq \|A\| \cdot \|\tilde{x}\| \Rightarrow \frac{1}{\|\tilde{x}\|} \leq \frac{\|A\|}{\|b\|},$$

lo que combinado con (4.14) produce el resultado buscado

$$\frac{\|\tilde{x} - \hat{x}\|}{\|\tilde{x}\|} \leq \|A\| \cdot \|A^{-1}\| \frac{\|b - \hat{b}\|}{\|b\|}. \tag{4.15}$$

Definición 4.16. Sea A una matriz invertible. Se llama número de condicionamiento de A a la cantidad

$$K(A) = \|A\| \cdot \|A^{-1}\|.$$

Obviamente este número depende de cual sea la norma matricial subordinada que se use y debería indicarse. Debido a la relación entre radio espectral y normas subordinadas, se define también el condicionamiento en radio espectral, o condicionamiento estrella, como

$$K_*(A) = \rho(A)\rho(A^{-1}) = \frac{\max\limits_{\lambda \in \sigma(A)} |\lambda|}{\min\limits_{\lambda \in \sigma(A)} |\lambda|}.$$

De la propiedad (4.11) se deduce que cualquiera sea la norma subordinada usada el condicionamiento de una matriz será siempre mayor o igual que uno. Por lo tanto la desigualdad (4.15) dice que frente a una perturbación del lado derecho se debe esperar un desplazamiento proporcionalmente mayor de la solución. Por otra parte, la propiedad (4.9) dice que el condicionamiento en radio espectral será menor o igual que el condicionamiento en cualquier norma matricial. En resumen, se tiene que

$$1 \leq K_*(A) \leq K(A).$$

Ejemplo 2. Consideremos el sencillo sistema de dos por dos y de inocente apariencia

$$\begin{aligned} 7x_1 + 10x_2 &= 1 \\ 5x_1 + 7x_2 &= 0.7 \end{aligned}$$

cuya solución es $\tilde{x} = \begin{pmatrix} 0 \\ 0.1 \end{pmatrix}$. Proponemos perturbar el lado derecho y considerar el nuevo sistema

$$\begin{aligned} 7x_1 + 10x_2 &= 1.01 \\ 5x_1 + 7x_2 &= 0.69 \end{aligned}$$

de solución $\hat{x} = \begin{pmatrix} -0.17 \\ 0.22 \end{pmatrix}$. Por lo tanto $\frac{\|\tilde{x}-\hat{x}\|_\infty}{\|\tilde{x}\|_\infty} = 1.7$ y $\frac{\|\tilde{b}-\hat{b}\|_\infty}{\|\tilde{b}\|_\infty} = 0.01$, es decir, el desplazamiento de la solución es **170** veces mayor que la perturbación.

Los números de condicionamiento de la matriz de ambos sistemas en las normas más usadas y en radio espectral son

$$K_\infty(A) = K_1(A) = 289, \quad K_2(A) \approx 223, \quad K_*(A) \approx 198.$$

Ejemplo 3. Se definen las matrices de Hilbert de tamaño n como $(H_n)_{i,j} = \frac{1}{i+j-1}, \forall i, j = 1, 2, ..., n$.

Por ejemplo

$$H_3 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}$$

Estas matrices son invertibles pero muy mal condicionadas, es decir, con números de condicionamiento muy grandes y que crecen en la medida que aumenta el tamaño, n. Algunos de estos números de condicionamiento, en radio espectral, son

$$\begin{aligned} K_*(H_3) &= 524, \\ K_*(H_5) &= 4.77 \cdot 10^5, \\ K_*(H_{10}) &= 1.60 \cdot 10^{13}. \end{aligned}$$

Un ejercicio interesante, que pone a prueba la eficiencia de cualquier software matemático, capaz de invertir matrices, consiste en pedirle calcular la inversa de este tipo de matrices aumentando el tamaño n. A partir de cierto tamaño umbral, $\forall n \geq n_0$, responderá que la matriz es numéricamente singular (no invertible) o mal condicionada. En los casos en que no avisa de este problema conviene comprobar la exactitud de la inversa entregada, calculando el producto $H_n \cdot H_n^{-1}$ y comparando este resultado con la matriz identidad I_n. Por ejemplo, Matlab para Windows, versión 4.2, calcula sin problemas la inversa de la matriz de Hilbert de tamaño $n = 10$. A partir de $n_0 = 12$, avisa que la matriz es mal condicionada y por lo tanto la inversa será poco confiable. Pero para $n = 11$ calcula H_{11}^{-1} sin advertir de ningún inconveniente y al pedirle calcular el producto $H_{11} \cdot H_{11}^{-1}$ entrega una matriz bien distinta de la identidad. Resulta de todos modos notable que una inestabilidad caracterizada por un número de condicionamiento, $K_*(H_{10}) = 1.6 \cdot 10^{13}$ no sea un obstáculo para que este software calcule con precisión la matriz inversa.

La frase "numéricamente singular" para indicar mal condicionamiento de una matriz, es rigurosamente exacta, como se establece en el teorema de Gastinel que presentamos sin demostración.

Teorema 4.17. *Si A es una matriz invertible, entonces su número de condicionamiento, en cualquier norma subordinada, satisface*

$$\frac{1}{K(A)} = \min \left\{ \frac{\|A - B\|}{\|A\|} | B \text{ es una matriz singular} \right\}.$$

Es decir, que A tenga un número de condicionamiento grande equivale a que se pueda aproximar bien por una matriz no invertible.

La prueba (en el Ejemplo (3)) a que propusimos someter a las matrices inversas calculadas por el software para decidir acerca de su precisión (calcular el producto $H_n \cdot H_n^{-1}$ y comparar con la identidad) es muy intuitiva pero peligrosa. Consideremos el caso de un sistema lineal

$$Ax = b.$$

Supongamos que hemos obtenido un vector $\hat{x} \in I\!R^n$, solución numérica de este sistema, cuya verdadera solución es $\tilde{x}$. Se llama **residuo** al vector

$$r = b - A\hat{x}.$$

Si la solución numérica fuera exacta, si coincidieran $\hat{x}$ y $\tilde{x}$, entonces el residuo sería nulo. Como el residuo es un vector calculable, cabe preguntarse si un residuo pequeño implica que la solución numérica es una buena aproximación de la solución verdadera. La respuesta es que esto no siempre es así. De hecho, la desigualdad (4.15) entrega la relación entre el residuo y el error cometido por la solución numérica, denotando por $\hat{b}$ a $A\hat{x}$, con lo cual se muestra la dependencia del condicionamiento de la matriz del sistema, como

$$\frac{\|\tilde{x} - \hat{x}\|}{\|\tilde{x}\|} \leq K(A)\frac{\|r\|}{\|b\|}.$$

Estudiaremos ahora la estabilidad de un sistema lineal sometido a perturbaciones de la matriz del sistema. Con éste y otros fines, que se apreciarán más adelante, presentamos un teorema de gran utilidad.

Teorema 4.18. *Sea E una matriz cuadrada de tamaño n tal que $\|E\| < 1$ para alguna norma subordinada cualquiera. Bajo esta hipótesis la matriz $(I - E)$ es invertible y en la misma norma subordinada se tiene*

$$\|(I - E)^{-1}\| \leq \frac{1}{1 - \|E\|}$$

donde I denota la matriz identidad de tamaño n.

Demostración. Razonando por el absurdo, supondremos que $(I-E)$ no es invertible y que por lo tanto existe un vector no nulo $x \in I\!R^n$ tal que $(I - E)x = 0$, equivalentemente $x = Ex$. Esto implica que con cualquier norma subordinada se cumplirá $\|x\| \leq \|E\| \cdot \|x\|$ y dividiendo por $\|x\| \neq 0$, se concluye que $1 \leq \|E\|$, lo que contradice la hipótesis.

Para acotar la norma de la inversa usaremos la misma propiedad de las normas subordinadas, considerando que

$$I = (I - E)^{-1}(I - E) = (I - E)^{-1} - (I - E)^{-1}E$$

y por consiguiente

$$(I - E)^{-1} = I + (I - E)^{-1}E.$$

Usando la desigualdad triangular y las propiedades (4.7) y (4.8) se obtiene

$$\|(I - E)^{-1}\| \leq 1 + \|(I - E)^{-1}\| \cdot \|E\|,$$

de lo cual se concluye la desigualdad propuesta. □

Consideremos el sistema lineal

$$Ax = b$$

de solución $\tilde{x}$ y el sistema perturbado

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

de solución $\hat{x}$ donde la matriz $\hat{A} = A(I + E)$, perturbación de la matriz A, es tal que

$$\|AE\| < \frac{1}{\|A^{-1}\|}. \tag{4.19}$$

Notemos que esto implica que $\|E\| = \|A^{-1}AE\| \leq \|A^{-1}\| \cdot \|AE\| < 1$ y por lo tanto, según el teorema anterior, la matriz $(I+E)$ es invertible. De la definición de $\tilde{x}$ y $\hat{x}$ se tiene que

$$\tilde{x} = (I+E)\hat{x} \tag{4.20}$$

y por lo tanto

$$\tilde{x} - \hat{x} = E\hat{x} = A^{-1}(\hat{A} - A)\hat{x}. \tag{4.21}$$

De esta ecuación se obtiene una desigualdad parecida a (4.15)

$$\frac{\|\tilde{x} - \hat{x}\|}{\|\hat{x}\|} \leq K(A)\frac{\|A - \hat{A}\|}{\|A\|}. \tag{4.22}$$

Pero el error relativo que deberíamos acotar es $\frac{\|\tilde{x}-\hat{x}\|}{\|\tilde{x}\|}$ en lugar de $\frac{\|\tilde{x}-\hat{x}\|}{\|\hat{x}\|}$. Lamentablemente en este caso la desigualdad que se obtiene pierde su parecido con (4.15) y el factor, si bien depende del número de condicionamiento no coincide con él. De (4.19) y el teorema anterior se tiene

$$\frac{\|\tilde{x} - \hat{x}\|}{\|\tilde{x}\|} \leq \frac{\|E\|}{1 - \|E\|} \leq \frac{K(A)}{1 - K(A)\frac{\|A-\hat{A}\|}{\|A\|}} \frac{\|A - \hat{A}\|}{\|A\|}. \tag{4.23}$$

Hacemos ver que la desigualdad (4.22) no requiere de la hipótesis (4.19) que limita el tamaño de las perturbaciones consideradas. Por el contrario, las dos desigualdades que aparecen en (4.23) solo son válidas bajo esta hipótesis. En particular, debido a (4.19) se tiene que

$$K(A)\frac{\|A - \hat{A}\|}{\|A\|} < 1$$

y por lo tanto el factor que multiplica a la perturbación relativa en esta desigualdad es mayor que el número de condicionamiento, es decir,

$$\frac{K(A)}{1 - K(A)\frac{\|A-\hat{A}\|}{\|A\|}} > K(A).$$

Para ilustrar estas relaciones consideremos el sistema de dos por dos del ejemplo (2) anterior.

Ejemplo 4. Para este sistema tenemos

$$A = \begin{bmatrix} 7 & 10 \\ 5 & 7 \end{bmatrix}, A^{-1} = \begin{bmatrix} -7 & 10 \\ 5 & -7 \end{bmatrix}$$

y por lo tanto

$$\|A\|_\infty = \|A^{-1}\|_\infty = 17.$$

Consideraremos una perturbación de la matriz que satisfaga la hipótesis (4.19), es decir, tal que

$$\|AE\|_\infty < \frac{1}{17} \approx 0.0588.$$

Este será el caso si la matriz perturbada es

$$\hat{A} = \begin{bmatrix} 7 & 9.95 \\ 4.97 & 7 \end{bmatrix}.$$

Teníamos que el número de condicionamiento de la matriz A en esta norma es $K_\infty(A) = 289$ y que la solución del sistema original es $\tilde{x} = \begin{pmatrix} 0 \\ 0.1 \end{pmatrix}$ si el lado derecho es $b = \begin{pmatrix} 1 \\ 0.7 \end{pmatrix}$. La solución del sistema perturbado $\hat{A}x = b$ es $\hat{x} = \begin{pmatrix} -\frac{10}{129} \\ \frac{20}{129} \end{pmatrix}$.

El error relativo es $\frac{\|\tilde{x}-\hat{x}\|_\infty}{\|\tilde{x}\|_\infty} = \frac{\frac{10}{129}}{0.1} = \frac{100}{129} \approx 0.775$.

La perturbación relativa es $\frac{\|A-\hat{A}\|_\infty}{\|A\|_\infty} = \frac{0.05}{17} \approx 0.0029$.

Es decir el error relativo es más de **267** veces mayor que la perturbación relativa. (*Recordemos que para la perturbación del lado derecho, presentada en el ejemplo* (1). *Este factor fue* **170** *y que el condicionamiento en radio espectral de esta matriz es* $K_*(A) = 198$). Por otra parte, la relación (4.23) dice en este caso que

$$\frac{\|\tilde{x} - \hat{x}\|_\infty}{\|\tilde{x}\|_\infty} \leq C \frac{\|A - \hat{A}\|_\infty}{\|A\|_\infty}, \text{ con } C \approx 1785.$$

Para terminar esta sección presentaremos un resultado que permite usar el residuo, en varias situaciones reales, para decidir si una solución numérica aproxima de manera aceptable a la solución verdadera. Insistimos en que en la práctica no tiene sentido pretender calcular un número de condicionamiento para decidir acerca de la calidad de una solución numérica. (*Conocer la inversa de la matriz del sistema, tiene al menos el mismo grado de dificultad que conocer la solución*). Esto no disminuye el interés de estas relaciones pues la gran mayoría de los sistemas lineales que aparecen en la resolución de problemas reales provienen de situaciones acerca de las cuales se tiene mucho más información y con frecuencia se sabe si son bien o mal condicionados.

Supongamos que se desea resolver el sistema $Ax = b$ y que tanto los coeficientes de la matriz como los del lado derecho solo se conocen de manera aproximada, con cierto error acotado. Es decir, en lugar de contar con A y b se tiene $\hat{A}$ y $\hat{b}$, además de una matriz E, con $E_{i,j} > 0$ tal que

$$|A_{i,j} - \hat{A}_{i,j}| \leq E_{i,j}, \quad \forall i, j = 1, 2, ..., n \tag{4.24}$$

y un vector δ con $\delta_i > 0$ tal que

$$|b_i - \hat{b}_i| \leq \delta_i, \forall i = 1, 2, ..., n. \tag{4.25}$$

La solución del sistema perturbado $\hat{A}x = \hat{b}$, que denotaremos por $\hat{x}$, solo se podrá conocer de manera aproximada, como solución numérica, resultante de algún procedimiento de cálculo. Sea $\bar{x}$ esta solución numérica del sistema perturbado y consideremos el residuo

$$r(\bar{x}) = \hat{b} - \hat{A}\bar{x}. \tag{4.26}$$

En todo el estudio hecho anteriormente nos interesaba analizar la relación entre este residuo y el error cometido por la solución numérica $\bar{x}$ con respecto a la solución exacta $\hat{x}$. Pero esta solución *exacta* lo es de un problema *aproximado* (el problema perturbado). Por lo tanto lo que realmente interesa es que la solución numérica $\bar{x}$ aproxime bien a la solución del sistema *desconocido* $Ax = b$. Diremos que la solución numérica obtenida $\bar{x}$ será aceptable como una buena aproximación de la verdadera solución del verdadero problema (desconocido) si existe un sistema lineal aceptablemente parecido al sistema resuelto, en el sentido de que satisfaga las tolerancias (4.24) y (4.25), del cual $\bar{x}$ sea solución exacta. Este criterio se formaliza en el teorema que sigue.

Teorema 4.27. *Sean $\hat{A}$ una matriz invertible de tamaño n, $\hat{b} \in \mathbb{R}^n, E$ una matriz de tamaño n y $\delta \in \mathbb{R}^n$, ambos con coeficientes positivos. Sea $\bar{x} \in \mathbb{R}^n$ una solución aproximada del sistema $\hat{A}x = \hat{b}$.*
Existe un sistema lineal $Ax = b$ cuya solución exacta es $\bar{x}$ y tal que A y b satisfacen (4.24) y (4.25), si y solo si el residuo, definido en (4.26) satisface

$$|r(\bar{x})| \leq E|\bar{x}| + \delta \tag{4.28}$$

donde se ha denotado por $|\bar{x}|$ al vector cuyas coordenadas son el valor absoluto de las coordenadas de $\bar{x}$, es decir, $(|\bar{x}|)_i = |\bar{x}_i|, \quad \forall i = 1, 2, ..., n$.

La demostración se puede ver en el libro de Stoer y Bulirsch [7].

Ejercicios propuestos.

1. Demuestre que toda norma subordinada es efectivamente una norma.

2. Demuestre las propiedades (4.6) y (4.8).

3. Considere el producto de matrices $\langle A, B\rangle = tr(B^t A)$, donde $tr(C) = \sum_{i=1}^{n} C_{i,i}$ y las matrices A y B son ambas del mismo tamaño (no se requiere en general que sean cuadradas pero éste será el único caso que nos interese aquí). Demuestre que éste es un producto interno en el correspondiente espacio de matrices y que por lo tanto $\|A\| = (\langle A, A\rangle)^{\frac{1}{2}}$ es una norma matricial. Demuestre que esta norma, llamada de Frobenius, **no** es norma subordinada. Explicítela.

4. Suponga que se conocen la matriz A y el lado derecho b de un sistema lineal con una precisión caracterizada por un error relativo ε en todos sus coeficientes. Es decir, se conocen $\hat{A}$ y $\hat{b}$ tales que $\frac{|A_{i,j} - \hat{A}_{i,j}|}{|A_{i,j}|} \leq \varepsilon, \forall i, j = 1, 2, ..., n, \frac{|b_i - \hat{b}_i|}{|b_i|} \leq \varepsilon, \forall i = 1, 2, ..., n$. Explicite la cota de (4.28) para cada componente del residuo en este caso.

5. Demuestre que si A es una matriz simétrica entonces $\|A\|_2 = \rho(A)$.

6. Demuestre que si A es una matriz ortogonal ($A^t = A^{-1}$) entonces $\|A\|_2 = 1$.

7. Repase la definición de normas equivalentes y recuerde que en un espacio vectorial de dimensión finita (como lo es el espacio de las matrices cuadradas de tamaño n) todas las normas son equivalentes. Encuentre todas las constantes para la equivalencia de las normas matriciales $\|\cdot\|_1, \|\cdot\|_2, \|\cdot\|_\infty$ entre sí y con la norma de Frobenius definida en **3.**

MÉTODOS DIRECTOS

A esta clase de métodos pertenecen los procedimientos estudiados en el curso de Álgebra Lineal, como son el métodos de Gauss y la factorización LU. Presentaremos aquí otras factorizaciones útiles y analizaremos aspectos numéricos tales como la propagación de errores y el costo en número de operaciones.

Método de Gauss.

Recordemos que el objetivo de este algoritmo de reducción es transformar un sistema lineal cualquiera en otro equivalente cuya matriz sea triangular superior. La razón que justifica este esfuerzo es que la resolución de un sistema de este tipo, mediante el procedimiento de **sustitución en reversa**, es simple y poco costosa, en cuanto al número de operaciones requeridas. En efecto, si el sistema $Ax = b$ es triangular superior, es decir, si $A_{i,j} = 0 \ \forall i > j$, entonces la solución $x \in \mathbb{R}^n$, se obtiene coordenada a coordenada como

$$x_n = \frac{1}{A_{n,n}} b_n \tag{4.29}$$

$$x_i = \frac{1}{A_{i,i}} \left(b_i - \sum_{j=i+1}^{n} A_{i,j} x_j \right) \quad \forall i = n-1, n-2, ..., 1.$$

En todo el capítulo suponemos que la matriz del sistema es invertible y por lo tanto aquí la matriz triangular A tiene diagonal no nula. Se observa directamente que el número de divisiones que se realizan es n y el número de productos es $\frac{n(n-1)}{2}$ y por lo tanto se dirá que el número de operaciones requeridas para realizar la sustitución en reversa es del orden de

$$\frac{n^2}{2}.$$

La reducción de Gauss se realiza en $(n-1)$ pasos destinados a anular una subcolumna (bajo la diagonal) cada vez. Sea $A^{(0)} = A$ y $A^{(k)}$ la matriz resultante del paso k-ésimo. La misma notación se usará para las transformaciones del lado derecho del sistema. El algoritmo de Gauss se resume como sigue.

4.30. *Para* $k = 1, 2, ..., n-1$

si el pivote $A_{k,k}^{(k-1)} = 0$ *realizar las permutaciones de filas o columnas que permitan modificar esta situación;*

para $i = k+1, ..., n$

$$A_{i,k}^{(k)} = 0$$

$$\theta_i = -\frac{A_{i,k}^{(k-1)}}{A_{k,k}^{(k-1)}}$$

$$b_i^{(k)} = b_i^{(k-1)} + \theta_i b_k^{(k-1)}$$

para $j = k+1, ..., n, \quad A_{i,j}^{(k)} = A_{i,j}^{(k-1)} + \theta_i A_{k,j}^{(k-1)}$.

Todos los coeficientes no mencionados permanecen idénticos.

Este procedimiento es considerablemente más costoso que la sustitución en reversa. El número de divisiones es $\frac{n(n-1)}{2}$ y el número de productos es $\frac{n(n-1)(2n-1)}{6}$ y por lo tanto se dirá que el número de operaciones requeridas para la reducción de Gauss es del orden de $\frac{n^3}{3}$.

Con gran frecuencia los sistemas lineales que aparecen en la práctica tiene matrices poco densas (con muchos ceros) y estructuradas. Una atenta revisión del algoritmo de reducción de Gauss permite saber a priori cuales elementos de una matriz de este tipo se reajustarán en el paso k-ésimo y cuales no. Si bien todos los software matemáticos permiten resolver sistemas lineales con una sola instrucción, el procedimiento convocado por ella no hará uso de esta observación y por lo tanto introducirá errores debido a la manipulación innecesaria que conllevan estos cálculos. En estos casos conviene programar un algoritmo de Gauss ad-hoc.

Consideremos por ejemplo el caso de un sistema tridiagonal $Ax = b$, con

$$A = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \gamma_2 & \alpha_2 & \beta_2 & & \\ & \gamma_3 & \ddots & \ddots & \\ & & \ddots & \ddots & \beta_{n-1} \\ & & & \gamma_n & \alpha_n \end{bmatrix}.$$

En cada paso k-ésimo la subcolumna tiene solo un elemento a anular y en el reajuste de una fila solo cambia el elemento diagonal. El algoritmo de Gauss para esta matriz se reduce a

$$\begin{aligned} & \bar{\alpha}_1 = \alpha_1 \\ \forall k = 2, ..., n \qquad & \bar{\alpha}_k = \alpha_k - \frac{\gamma_k}{\bar{\alpha}_{k-1}} \beta_{k-1}. \end{aligned}$$

La propagación del error de redondeo debido a los cálculos puede llegar a ser considerable en sistemas grandes no estructurados y por lo tanto conviene tomar las medidas de control a nuestra disposición. Las divisiones por pivotes pequeños en módulo son conocidas como principal fuente de este problema y por lo tanto el mejor resguardo contra este factor de riesgo consiste en elegir como pivote del paso k-ésimo al elemento de mayor módulo de toda la submatriz inferior derecha de tamaño $(n - k + 1)$ y llevarlo a la posición (k, k) mediante permutaciones de filas y/o columnas. Esta estrategia se llama de *búsqueda total de pivote*. Como todas estas comparaciones pueden resultar muy costosas frecuentemente se restringe la búsqueda del mejor pivote a la fila o a la subcolumna del pivote (la k-ésima). La denominación que recibe este procedimiento es de *búsqueda parcial*. La importancia de estas búsquedas y los riesgos de la búsqueda parcial se ilustran en el ejemplo sencillo, de dos por dos, que sigue.

Ejemplo 5. Consideremos el sistema

$$\begin{bmatrix} 0.005 & 1 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}$$

de solución $x = \frac{5000}{9950} = 0.503...$, $y = \frac{4950}{9950} = 0.497...$

Si aplicamos el método de reducción de Gauss sin permutar ni filas ni columnas usando aritmética de punto flotante de 2 dígitos se obtendrá como resultado $x = 0, \quad y = 0.5$, con un error evidente.

Supongamos que realizamos búsqueda de pivote (en este ejemplo dará igual si es total o parcial) y permutamos las ecuaciones (las filas). En este caso, con la misma aritmética, obtendremos la solución $x = 0.5, \quad y = 0.5$, con un error considerablemente menor que el anterior.

Multiplicando la primera ecuación por 200 se tendrá el sistema equivalente

$$\begin{bmatrix} 1 & 200 \\ 1 & 1 \end{bmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 100 \\ 1 \end{pmatrix}.$$

Si se nos permite realizar búsqueda total de pivote obtendremos una solución de la misma calidad de la anterior. En cambio si realizamos búsqueda parcial en la columna del pivote, es decir, si las permutaciones previstas se reducen a las filas, entonces no habrá ninguna razón para intercambiarlas en este caso y sin embargo la aplicación del método de Gauss a este sistema llevará a los mismos errores que la primera alternativa (sin ninguna búsqueda de pivote).

Una técnica también recomendada con frecuencia para controlar la propagación de errores, es la de **escalar** la matriz del sistema. Esto es lo que hemos hecho al multiplicar la primera ecuación por 200. En general, consiste en pre- y post-multiplicar la matriz por matrices diagonales, cuidando de obtener un sistema equivalente, es decir, resolver el sistema

$$D_1 A D_2 \tilde{x} = D_1 b$$

y luego calcular

$$x = D_2\tilde{x}.$$

Es muy difícil encontrar una regla general para decidir el escalamiento apropiado y el criterio de elección de pivote que se debe aplicar al sistema escalado. El método más recomendado, basados en evidencia empírica, consiste en limitarse a la premultiplicación por

$$D_1 = diag(s_1, ..., s_n), \text{ con } s_i = \frac{1}{\sum_{j=1}^{n} |A_{i,j}|}$$

y luego aplicar la reducción de Gauss con búsqueda parcial de pivote solo en la subcolumna correspondiente (permutar filas). En la práctica este escalamiento puede no explicitarse jamás, incorporando su efecto solo en la selección del pivote. En resumen, esta técnica consiste en aplicar la reducción de Gauss al sistema original (sin escalar) con la elección del pivote k-ésimo (en el paso k descrito en el algoritmo (4.30) que sigue

4.31. *Determinar $r \geq k$ tal que $|A_{r,k}^{(k-1)}|s_r = \max_{i \geq k} |A_{i,k}^{(k-1)}|s_i$,*

permutar filas r y k.

Factorizaciones LU y de Cholesky.

Dada una matriz A se dice que ella posee factorización LU si existen dos matrices L, triangular inferior con unos en la diagonal y U, triangular superior sin ceros en la diagonal tales que

$$A = LU.$$

El hecho de que A sea invertible no garantiza que esta factorización exista, pero si ella existe entonces es única. Que A sea invertible implica que existe una factorización (no única) $A = LUP$, donde la matriz P es producto de matrices de permutación y por lo tanto es invertible y solo tiene ceros y unos. De hecho, la inversa de esta matriz se construye como la síntesis de todas las permutaciones de columnas (reordenamiento de las coordenadas del vector solución) realizadas en la reducción de Gauss de la matriz A.

En el curso de Álgebra Lineal se demuestra un teorema acerca de formas cuadráticas definidas positivas que en una de sus partes dice

Teorema 4.32. *A simétrica definida positiva si y solo si existe una matriz S triangular superior con diagonal estrictamente positiva tal que*

$$A = S^t S,$$

llamada factorización de Cholesky.

La importancia de las factorizaciones triangulares presentadas es que permiten resolver un sistema como dos sistemas triangulares, es decir mediante dos procedimientos de sustitución: uno hacia adelante y otro en reversa. Por ejemplo supongamos que se conoce la factorización LU de la matriz A y por lo tanto resolver el sistema $Ax = b$ se reduce a resolver

$Ly = b$ por sustitución hacia adelante:

$y_1 = b_1$,

$$y_i = b_i - \sum_{j=1}^{i-1} L_{i,j} y_j, \quad \forall i = 2, ..., n.$$

$Ux = y$ por sustitución en reversa.

La obtención de la factorización LU o LUP es equivalente a la reducción de la matriz A mediante el algoritmo de Gauss (como se ve en el curso de Álgebra Lineal) y por lo tanto el beneficio de calcular estas factorizaciones se obtiene cuando se desea resolver varios sistemas con distintos lados derechos y la misma matriz.

Las factorizaciones LU y de Cholesky se pueden calcular también de manera directa por identificación de coeficientes de la matriz producto. Siguiendo esta esquema presentaremos el algoritmo que permite obtener la matriz de Cholesky.

4.33. $S_{1,1} = (A_{1,1})^{1/2}$.

Para $j = 1, 2, ..., n-1$,

para $k = j+1, ..., n$

$$S_{j,k} = \frac{1}{S_{j,j}} \left(A_{j,k} - \sum_{i=1}^{j-1} S_{i,j} S_{i,k} \right)$$

$$S_{j+1,j+1} = \left(A_{j+1,j+1} - \sum_{i=1}^{j} S_{i,j+1}^2 \right)^{1/2}.$$

Tal como ocurre con la resolución de sistemas lineales, las factorizaciones son realizadas con gran eficiencia por los software matemáticos pero estos no aprovechan la estructura de matrices poco densas para las cuales conviene programar algoritmos ad-hoc. Por ejemplo, se sabe que la matriz del sistema lineal que se debe resolver para calcular la función Spline cúbica de interpolación con n nodos equiespaciados es definida positiva. Esta matriz tridiagonal y simétrica de tamaño n es de la forma

$$A = \begin{bmatrix} 2 & 1 & & & & & \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & 4 & 1 & \\ & & & & 1 & 4 & 1 \\ & & & & & 1 & 2 \end{bmatrix}.$$

Para obtener su factorización de Cholesky bastará notar que la matriz S debe tener solo 2 diagonales y que por lo tanto el algoritmo (4.32) se reduce a

$S_{1,1} = \sqrt{2}$. Para $j = 1, 2, ..., n-1$,

$S_{j,j+1} = \frac{1}{S_{j,j}}$, $S_{j+1,j+1} = (4 - S_{j,j+1}^2)^{1/2}$, excepto para el último elemento diagonal,
$S_{n,n} = (2 - S_{n-1,n}^2)^{1/2}$.

Factorización QR.

Se dirá que una matriz A tiene factorización QR si existen una matriz ortogonal Q y una matriz triangular superior R tales que

$$A = QR.$$

Teorema 4.34. *Si A es invertible siempre existe la factorización QR de A y es única.*

Demostración. Si la matriz A de tamaño n es invertible entonces sus n columnas son linealmente independientes y forman una base de $\mathbb{R}^n$. Usando el procedimiento de Gramm Schmidt se puede ortonormalizar esta base. Denotemos por $u_j = A_{*j}$ a la columna j-ésima de la matriz A y por v_j, para $j = 1, 2, ..., n$, a los vectores ortonormales obtenidos por Gramm Schmidt. Consideremos una matriz Q cuyas columnas sean los vectores v_j, es decir, se define columna a columna como $Q_{*j} = v_j$. Por construcción se tendrá que Q es invertible y que $Q^tQ = I$, la matriz identidad, por lo tanto Q es una matriz ortogonal. Para relacionar las matrices A y Q debemos recordar el procedimiento de Gramm Schmidt:

$v_1 = \alpha_1 u_1$, donde $\alpha_1 = \frac{1}{\|u_1\|}$,

$\forall j = 2, ..., n \; v_j = \alpha_j \left(u_j - \sum_{i=1}^{j-1} \beta_{i,j} v_i \right)$, donde α_j denota el inverso de la norma del vector dentro del paréntesis y los coeficientes $\beta_{i,j}$ son los coeficientes de Fourier.

Despejando los vectores u_j de estas ecuaciones se obtiene

$u_1 = R_{1,1} v_1$,

$\forall j = 2, ..., n \; u_j = \sum_{i=1}^{j} R_{i,j} v_i$.

Si consideramos la matriz formada por los coeficientes $R_{i,j}$ que aparecen en las ecuaciones anteriores en las posiciones (i, j) que intervienen y ceros en las demás ubicaciones, tendremos que R es una matriz triangular superior con diagonal no nula (el coeficiente $R_{j,j}$ corresponden a una norma de un vector no nulo) y tal que $A = QR$. Probamos así la existencia de esta factorización.

La unicidad será demostrada por el absurdo. Supongamos que A tiene dos factorizaciones QR, es decir, que existen dos matrices ortogonales Q y P, y dos matrices invertibles triangulares superiores R y U tales que $A = QR = PU$. Esto equivale a

$$P^tQ = UR^{-1}.$$

Se comprueba fácilmente que la matriz del lado izquierdo, producto de matrices ortogonales, es a su vez ortogonal. Por otra parte la matriz del lado derecho, producto de matrices triangulares superiores, es triangular superior. Pero una matriz H que es triangular superior y ortogonal solo puede ser la matriz identidad. En efecto, $H^tH = I$, implica

$$\begin{aligned} \langle H_{*1}, H_{*1} \rangle = H_{1,1}^2 = 1 &\Rightarrow H_{1,1} = 1 \\ \forall j = 2, ..., n, \langle H_{*1}, H_{*j} \rangle = H_{1,1} H_{1,j} = 0 &\Rightarrow H_{1,j} = 0 \\ \langle H_{*2}, H_{*2} \rangle = H_{2,2}^2 = 1 &\Rightarrow H_{2,2} = 1 \\ \forall j = 3, ..., n, \langle H_{*2}, H_{*j} \rangle = H_{2,2} H_{2,j} = 0 &\Rightarrow H_{2,j} = 0, \text{ etc...} \end{aligned}$$

Por lo tanto $Q = P$ y $R = U$. □

Ejercicios propuestos.

1. Escriba un algoritmo de Gauss para matrices de banda, con ancho de banda p, es decir, con $(2p-1)$ diagonales, o equivalentemente, $A_{i,j} = 0$ si $|i - j| \geq p$.

2. Escriba un algoritmo para encontrar la factorización de Cholesky para las matrices del problema anterior, simétricas definidas positivas.

3. Escriba algoritmos de Gauss y para encontrar la factorización de Cholesky para matrices simétricas definidas positivas, 3-diagonales por bloques de tamaño p,

$$A = \begin{bmatrix} B_1 & C_1 & & & \\ C_1^t & B_2 & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & B_{n-1} & C_{n-1} \\ & & & C_{n-1}^t & B_n \end{bmatrix}$$

con B_i, C_i, matrices cuadradas de tamaño p y $B_i = B_i^t$.

4. Considere una matriz B de m filas y n columnas con $m > n$ (no cuadrada) de rango n, es decir, tal que sus n columnas son vectores linealmente independientes en $\mathbb{R}^m$. Pruebe que existe una factorización QR de B, donde Q es una matriz cuadrada de tamaño m y R es una matriz de m filas y n columnas que tiene en las primeras n filas una matriz r triangular superior, sin ceros en al diagonal, y todas las restantes $(m-n)$ filas nulas.

5. Calcule una factorización QR de la matriz $B = \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$.

6. Sea $Ax = b$ un sistema de mínimos cuadrados como el descrito en el capítulo 2, que admite una factorización $A = B^tB$, $b = B^ty$, donde B es una matriz de N filas y n columnas, con $N > n, B$ de rango $n, y \in \mathbb{R}^n$.

   (a) Pruebe que A admite factorización de Cholesky, $A = S^tS$.

   (b) Considere la factorización QR de la matriz rectangular $B = QR$, con $R = \begin{bmatrix} r \\ 0_{(N-n)\times n} \end{bmatrix}$ y r triangular superior invertible. Considere la partición de la matriz Q inducida por la partición de R, es decir $Q = [Q_1, Q_2]$, donde Q_1 tiene las primeras n columnas de Q. Pruebe que $Q_1^tB = r$ y $Q_2^tB = 0_{(N-n)\times n}$.

   (c) Pruebe que el factor de Cholesky de A, satisface $S = r$ y que si se conoce la factorización QR de la matriz B, entonces la resolución del sistema de mínimos cuadrados se reduce a una sustitución en reversa del sistema

$$rx = Q_1^ty.$$

7. Escriba un algoritmo que permita calcular la factorización LU de una matriz dada (suponga que existe) mediante identificación de los coeficientes del producto, como se hizo para obtener el algoritmo (4.33). Determine el número de operaciones que requiere este algoritmo.

8. Encuentre la factorización LU de la matriz de la Spline cúbica de interpolación con nodos equiespaciados (dada como ejemplo de cálculo del factor de Cholesky) usando el algoritmo diseñado en el problema anterior.

9. Adapte el algoritmo obtenido en el problema 7 al caso de matrices 3-diagonales que admiten factorización LU.

10. Diseñe una adaptación del algoritmo obtenido en el problema 7 para el caso de matrices de banda con ancho de banda p, que admiten factorización LU. Cuente el número de operaciones requeridas.

MÉTODOS ITERATIVOS

Una estrategia distinta para resolver un sistema lineal, especialmente si éste es de un gran tamaño, es la de generar una sucesión de vectores que converja a la solución del sistema. Los métodos que presentaremos se basan en la transformación del sistema lineal

$$Ax = b$$

en un problema de punto fijo *equivalente (con la misma solución)*,

$$x = Mx + v, \tag{4.35}$$

donde M es una matriz de tamaño n, al igual que A.

La solución de este último se obtiene como el límite de las aproximaciones sucesivas

$$x^{(k+1)} = Mx^{(k)} + v, \text{ partiendo de un vector } x^{(0)}, \text{ dado.} \tag{4.36}$$

Se sabe que si la función de iteración

$$\begin{aligned} F : \mathbb{R}^n &\to \mathbb{R}^n \\ x &\to F(x) = Mx + v \end{aligned}$$

es contractante, entonces las iteraciones propuestas convergen al punto fijo de F, es decir, a la solución del sistema lineal. La condición suficiente mencionada se expresa en alguna norma vectorial como

$$\exists L, 0 < L < 1, \text{ tal que } \forall x, y \in \mathbb{R}^n, \|F(x) - F(y)\| \leq L\|x - y\|.$$

Como $\|F(x) - F(y)\| = \|Mx - My\| \leq \|M\| \cdot \|x - y\|$, entonces bastará que en alguna norma subordinada la matriz de iteración M satisfaga

$$\|M\| \leq 1 \tag{4.37}$$

para asegurar la convergencia de las iteraciones (4.36), cualquiera sea el punto de partida $x^{(0)}$.

Como sabemos de la estrecha relación entre las normas subordinadas y el radio espectral, buscamos una condición de convergencia en este último, para lo cual estudiaremos la evolución del error.

El error de la iteración k-ésima es

$$e^{(k)} = x - x^{(k)} = F(x) - F(x^{(k-1)}) = M(x - x^{(k-1)}) = Me^{(k-1)}$$

y repitiendo el mismo argumento se obtendrá

$$e^{(k)} = M^k e^{(0)}. \tag{4.38}$$

La matriz M puede ser mirada como una matriz compleja. En este espacio vectorial será similar a una matriz compleja de Jordan, J, (triangular superior con los valores propios de M en su diagonal)

$$M = PJP^{-1}$$

y si todos los valores propios de M son de módulo menor que 1, entonces $J^k \to 0_{n\times n}$, si k tiende a infinito. Por lo tanto

$$\rho(M) < 1 \Rightarrow M^k = PJ^kP^{-1} \to 0_{n\times n}.$$

La expresión (4.38) del error k-ésimo permite así afirmar que el método iterativo converge si $\rho(M) < 1$.

Es más, mientras menor sea el radio espectral de la matriz de iteración, más rápida será la convergencia del método iterativo.

Un criterio habitual para detener un proceso iterativo es que las iteraciones sucesivas sean muy parecidas. Desde un punto de vista práctico resulta obviamente inútil continuar un proceso que parece no producir modificaciones. Veremos que en el caso de estos métodos este criterio práctico está validado por la teoría y que la poca distancia entre iteraciones sucesivas garantiza la cercanía a la solución. Nuestro objetivo aquí será acotar superiormente el error k-ésimo con la distancia entre las dos últimas iteraciones.

Consideremos la distancia entre las iteraciones $(k+1)$ y k

$$x^{(k+1)} - x^{(k)} = x^{(k+1)} - x + x - x^{(k)} = -e^{(k+1)} + e^{(k)} = (I - M)e^{(k)}.$$

Si $\|M\| < 1$, entonces según el teorema (4.18) la matriz $(I - M)$ será invertible y se tendrá que

$$\|e^{(k)}\| \leq \frac{1}{1 - \|M\|}\|x^{(k+1)} - x^{(k)}\|.$$

Pero

$$\begin{aligned} \|x^{(k+1)} - x^{(k)}\| &= \|F\left(x^{(k)}\right) - F\left(x^{(k-1)}\right)\| \\ &= \|Mx^{(k)} - Mx^{(k-1)}\| \leq \|M\| \cdot \|x^{(k)} - x^{(k-1)}\|, \end{aligned} \tag{4.39}$$

de lo cual se obtiene la relación buscada

$$\|e^{(k)}\| \leq \frac{\|M\|}{1 - \|M\|}\|x^{(k)} - x^{(k-1)}\|. \tag{4.40}$$

Esta desigualdad tiene otra gran importancia práctica. Es evidente que se puede repetir el argumento usado en (4.39), lo que usado en (4.40) produce la desigualdad

$$\|e^{(k)}\| \leq \frac{\|M\|^k}{1 - \|M\|}\|x^{(1)} - x^{(0)}\|. \tag{4.41}$$

De esta manera, habiendo calculado tan solo una iteración se puede asegurar con toda certeza en cuantas iteraciones, a lo más, se habrá alcanzado una precisión exigida.

Ejemplo 6. Consideremos un ejemplo de solución conocida para ilustrar lo anterior. El sistema de tamaño n, $Ax = b$, donde A es la matriz de la Spline cúbica de interpolación con nodos equiespaciados, de ejemplos anteriores,

$$A = \begin{bmatrix} 2 & 1 & & & & & & \\ 1 & 4 & 1 & & & & & \\ & 1 & 4 & 1 & & & & \\ & & 1 & \ddots & \ddots & & & \\ & & & \ddots & \ddots & 1 & & \\ & & & & 1 & 4 & 1 & \\ & & & & & 1 & 2 & \end{bmatrix}, \text{ el lado derecho es } b = \begin{bmatrix} 3 \\ 6 \\ 6 \\ \vdots \\ 6 \\ 6 \\ 3 \end{bmatrix} \text{ y la solución es } x = \begin{bmatrix} 1 \\ \vdots \\ \vdots \\ 1 \end{bmatrix}.$$

Este sistema es equivalente al problema de punto fijo $x = Mx + v$, con

$$M = \begin{bmatrix} 0 & -\frac{1}{2} & & & & & \\ -\frac{1}{4} & 0 & -\frac{1}{4} & & & & \\ & -\frac{1}{4} & 0 & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & -\frac{1}{4} & \\ & & & & -\frac{1}{4} & 0 & -\frac{1}{4} \\ & & & & & -\frac{1}{2} & 0 \end{bmatrix} \text{ y } v = \begin{bmatrix} \frac{3}{2} \\ \frac{3}{2} \\ \vdots \\ \vdots \\ \vdots \\ \frac{3}{2} \\ \frac{3}{2} \end{bmatrix}.$$

Se ve fácilmente que $\|M\|_\infty = \frac{1}{2} < 1$.

Si se parte de $x^{(0)} = 0 \in I\!R^n$, entonces $x^{(1)} = v$ y por lo tanto

$$\|x^{(1)} - x^{(0)}\|_\infty = \frac{3}{2}.$$

Supongamos que nos piden calcular la solución de este problema con una precisión $\varepsilon = 10^{-10}$. Gracias a la desigualdad (4.41) podremos asegurar que a lo sumo en 35 iteraciones se obtendrá la precisión pedida, es decir, la iteración 35-ésima cometerá un error menor o igual que 10^{-10}. En efecto, como $log_2(3 \cdot 10^{10}) = 34.80...$, entonces $\forall k \geq 35, \frac{(\frac{1}{2})^k \frac{3}{2}}{1-\frac{1}{2}} < 10^{-10}$, y por lo tanto

$$\|e^{(k)}\|_\infty \leq \frac{\|M\|_\infty^k}{1 - \|M\|_\infty} \|x^{(1)} - x^{(0)}\| < \varepsilon.$$

El ejemplo que hemos dado, de extraordinaria simpleza y solución conocida, resulta adecuado para ilustrar los resultados previos, pero es muy inapropiado para apreciar la ventaja de los métodos iterativos en cuanto al ahorro en el número total de operaciones, cuando se tienen sistemas grandes, no estructurados. Si tanto la matriz A como la matriz M son llenas, entonces bastará con que la precisión deseada se alcance en menos de $\frac{n}{3}$ iteraciones para que el costo del método iterativo sea menor que el costo del método de Gauss (el producto de la matriz de iteración M por el vector de la iteración previa se realiza con n^2 multiplicaciones).

Obviamente la transformación del sistema lineal en un problema de punto fijo equivalente, del tipo estudiado, no es única. De hecho, podríamos inventar según el sistema particular, el problema de punto fijo equivalente que nos parezca más apropiado, respetando la condición de convergencia. Existen ciertas elecciones clásicas para realizar de manera sencilla esta transformación. Estas son conocidas como los métodos de Jacobi y de Gauss-Seidel. Ambos métodos se basan en una descomposición de la matriz del sistema en una suma de su diagonal, su triángulo inferior estricto y su triángulo superior estricto.

Sean $D = diag(A_{1,1}, A_{2,2}, ..., A_{n,n})$, E y F definidas por

$$E_{i,j} = \begin{cases} 0 & \text{si } i \leq j \\ A_{i,j} & \text{si } i > j \end{cases}, \qquad F_{i,j} = \begin{cases} A_{i,j} & \text{si } i < j \\ 0 & \text{si } i \geq j \end{cases}.$$

Obviamente $A = D + E + F$.

Métodos de Jacobi, Gauss-Seidel y Relajación.

Si la matriz A no tiene elementos diagonales nulos (D es invertible), entonces el sistema $Ax = b$ será equivalente al problema de punto fijo

$$x = D^{-1}(-E - F)x + D^{-1}b. \tag{4.42}$$

Este problema tiene la forma del problema estudiado previamente. La matriz de iteración de este método, llamado de **Jacobi**, será

$$M_J = -D^{-1}(E+F)$$

y la condición de convergencia (4.37) en $\|\ \|_\infty$ impondrá a la diagonal de la matriz A una condición más severa que la de no tener elementos nulos. En efecto,

$$\|M_J\|_\infty = \max_{1\le i\le n} \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{|A_{i,j}|}{|A_{i,i}|}$$

implica que

$$\|M_J\|_\infty < 1 \Leftrightarrow \forall i = 1,2,...,n, \sum_{\substack{j=1\\ j\neq i}}^{n} |A_{i,j}| < |A_{i,i}|.$$

Esta última condición se conoce como la exigencia de que A sea *diagonal dominante por filas*. Se comprueba fácilmente que la condición $\|M_J\|_1 < 1$ equivale a pedir que A sea diagonal dominante por columnas, es decir,

$$\forall j = 1,2,...,n, \sum_{\substack{i=1\\ i\neq j}}^{n} |A_{i,j}| < |A_{jj}|.$$

Recordamos aquí que, debido a la equivalencia de las normas en $\mathbb{R}^n$, estableciendo la convergencia del método iterativo en alguna norma, se habrá garantizado la convergencia en cualquier otra norma. Las diferencias pueden ser interesantes al aplicar la desigualdad (4.41) como hicimos en el ejemplo. Hacemos ver que el método iterativo usado para resolver ese problema corresponde al método de Jacobi y que usando $\|\ \|_1$ habríamos obtenido otro resultado para conseguir la misma precisión.

La fórmula de cálculo de cada coordenada del vector de la iteración $(k+1)$-ésima de Jacobi, que se desprende de la ecuación (4.42) será

$$\forall i = 1,2,...,n, \quad x_i^{(k+1)} = \sum_{\substack{j=1\\ j\neq i}}^{n} \frac{-A_{i,j}}{A_{i,i}} x_j^{(k)} + \frac{b_i}{A_{i,i}}. \tag{4.43}$$

Así, para calcular la coordenada i-ésima del vector $(k+1)$-ésimo se usan *todas* las coordenadas, menos la i-ésima, del vector k-ésimo. Si cada nueva iteración representa una mejoría con respecto a la anterior y si las coordenadas se calculan en orden, entonces cuando se calcule la segunda coordenada de la iteración $(k+1)$-ésima debería ser beneficioso usar la nueva primera coordenada, ya calculada, en lugar de usar la de la iteración anterior. Esta modificación del método de Jacobi, que no espera a que se complete el cálculo de todo un vector para usar sus coordenadas, se llama método de **Gauss-Seidel**. La fórmula de cálculo de cada coordenada del vector de la iteración $(k+1)$-ésima será

$$\forall i = 1,2,...,n, \quad x_i^{(k+1)} = \sum_{j=1}^{i-1} \frac{-A_{i,j}}{A_{i,i}} x_j^{(k+1)} + \sum_{j=i+1}^{n} -\frac{A_{ij}}{A_{ii}} x_j^{(k)} + \frac{b_i}{A_{i,i}}. \tag{4.44}$$

Para escribir esta relación en términos de las matrices D, E y F, conviene ilustrar estas ecuaciones según

el esquema

$$\rightarrow \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \vdots \\ x_i^{(k+1)} \\ \vdots \\ x_{n-1}^{(k+1)} \\ x_n^{(k+1)} \end{pmatrix} = \begin{bmatrix} 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\ * & \ddots & & & & & \vdots \\ \vdots & & \ddots & & & & \vdots \\ * & \cdots & * & 0 & \cdots & \cdots & 0 \\ \vdots & & & & \ddots & & \vdots \\ \vdots & & & & & 0 & \vdots \\ * & \cdots & \cdots & \cdots & \cdots & * & 0 \end{bmatrix} \begin{pmatrix} x_1^{(k+1)} \\ \vdots \\ x_{i-1}^{(k+1)} \\ x_i^{(k+1)} \\ \vdots \\ \vdots \\ x_n^{(k+1)} \end{pmatrix} +$$

$$\begin{bmatrix} 0 & * & \cdots & \cdots & \cdots & \cdots & * \\ \vdots & \ddots & & & & & \vdots \\ \vdots & & \ddots & & & & \vdots \\ 0 & \cdots & \cdots & 0 & * & \cdots & * \\ \vdots & & & & \ddots & & \vdots \\ \vdots & & & & & \ddots & * \\ 0 & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \end{bmatrix} \begin{pmatrix} x_1^{(k)} \\ \vdots \\ \vdots \\ x_i^{(k)} \\ x_{i+1}^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix} + v.$$

De aquí se deduce que la iteración del método de Gauss-Seidel se resume como

$$x^{(k+1)} = -D^{-1}Ex^{(k+1)} - D^{-1}Fx^{(k)} + D^{-1}b \tag{4.45}$$

y que la matriz de iteración de este método es

$$M_{GS} = -(I + D^{-1}E)^{-1}D^{-1}F. \tag{4.46}$$

Esta matriz de iteración es bastante más compleja que la matriz de iteración del método de Jacobi y por lo tanto, el estudio de la convergencia de este método se complica. La diagonal dominancia de la matriz A es una condición suficiente de convergencia del método de Jacobi, muy fácil de revisar y veremos que también resulta suficiente para la convergencia del método de Gauss-Seidel. Más aún cuando ésta se cumple, el método de Gauss-Seidel converge al menos a la misma velocidad que el método de Jacobi, como se prueba en el teorema que sigue.

Teorema 4.47. *Si A es diagonal dominante por filas, entonces el método de Gauss-Seidel para resolver el sistema $Ax = b$ converge en norma $\|\cdot\|_\infty$ al menos a la misma velocidad que el método de Jacobi.*

Demostración. Sean $\alpha_i = \sum_{j=1}^{i-1} \frac{|A_{i,j}|}{|A_{i,i}|}$, $\forall i = 2, ..., n$, y $\beta_i = \sum_{j=i+1}^{n} \frac{|A_{i,j}|}{|A_{i,i}|}$, $\forall i = 1, 2, ..., n-1$, $\alpha_1 = \beta_n = 0$.

Como A es diagonal dominante, se tiene que

$$\forall i = 1, 2, ..., n, \quad \mu_i = \alpha_i + \beta_i < 1.$$

Sean $x^{(k)}$, la k-ésima iteración de Jacobi, $\tilde{x}^{(k)}$, la k-ésima iteración de Gauss-Seidel $e^{(k)} = x - x^{(k)}$, el error de la iteración k-ésima de Jacobi y $\tilde{e}^{(k)} = x - \tilde{x}^{(k)}$, el error de la iteración k-ésima de Gauss-Seidel. Sabemos que

$$\|e^{(k)}\|_\infty \leq \mu \|e^{(k-1)}\|_\infty, \quad \text{si } \mu = \max_{1 \leq i \leq n} \mu_i. \tag{4.48}$$

Sea i^* el índice de la coordenada del error k-ésimo de Gauss-Seidel donde se realiza el máximo de los módulos (donde se alcanza la norma considerada), es decir,

$$|\tilde{e}_{i^*}^{(k)}| = \max_{1\leq i\leq n} |\tilde{e}_i^{(k)}| = \|\tilde{e}^{(k)}\|_\infty.$$

De la fórmula (4.44) se obtiene que

$$\left\|\tilde{e}^{(k)}\right\|_\infty \leq \left\|\tilde{e}^{(k)}\right\|_\infty \quad \alpha_{i^*} + \left\|\tilde{e}^{(k-1)}\right\|_\infty \quad \beta_{i^*} \Leftrightarrow \left\|\tilde{e}^{(k)}\right\|_\infty \leq \frac{\beta_{i^*}}{1-\alpha_{i^*}} \left\|\tilde{e}^{(k-1)}\right\|_\infty.$$

Definiendo

$$\eta = \max_{1\leq i\leq n} \eta_i, \text{ con } \eta_i = \frac{\beta_i}{1-\alpha_i}, \ \forall i = 1, 2, ..., n,$$

se tendrá que el error del método de Gauss-Seidel evoluciona según

$$\|\tilde{e}^{(k)}\|_\infty \ \leq \eta \|\tilde{e}^{(k-1)}\|_\infty. \tag{4.49}$$

Se comprueba fácilmente que $\forall i = 1, 2, ..., n, \ \eta_i \leq \mu_i$ y que por lo tanto $\eta \leq \mu$, con lo que concluye la demostración del teorema. □

La descomposición matricial del método de Gauss-Seidel se puede sofisticar aún más con el fin de obtener una matriz de iteración con menor radio espectral y por consiguiente un método iterativo de convergencia más veloz.

Consideremos un parámetro $w, 0 < w < 2$, que será usado para acelerar la convergencia combinando, en esa proporción, el vector generado por la iteración de Gauss-Seidel con el vector de la iteración anterior.

Sea $\tilde{x}^{(k+1)}$ la iteración $(k+1)$-ésima del método de Gauss-Seidel, definida según (4.45) como

$$\tilde{x}^{(k+1)} = -D^{-1}Ex^{(k+1)} - D^{-1}Fx^{(k)} + D^{-1}b.$$

En el método que proponemos, se define la iteración $(k+1)$-ésima, como

$$x^{(k+1)} = w\tilde{x}^{(k+1)} + (1-w)x^{(k)},$$

es decir,

$$x^{(k+1)} = -wD^{-1}Ex^{(k+1)} + ((1-w)I - wD^{-1}F)x^{(k)} + wD^{-1}b. \tag{4.50}$$

- Si $w < 1$ este método se llama de **Relajación**,
- Si $w = 1$ se tiene el método de **Gauss-Seidel**,
- Si $w > 1$ el método se llama de **Sobrerrelajación**.

La matriz de iteración correspondiente a la fórmula (4.50) es

$$M_R(w) = (1 + wD^{-1}E)^{-1}((1-w)I - wD^{-1}F). \tag{4.51}$$

El mejor parámetro w, acelerador de convergencia, será aquel que minimice el radio espectral de la matriz de iteración, es decir, w^* solución del problema

$$\min_{w\in(0,2)} \rho(M_R(w)).$$

Este problema optimal es muy complejo y solo en algunos casos particulares se conoce su solución. Sin embargo, las diferencias en las velocidades de convergencia pueden llegar a ser dramáticas y todo esfuerzo por aproximar el parámetro óptimo w^* será bien recompensado por esta vía.

Ejercicios propuestos.

1. Considere una matriz tridiagonal por bloques cuadrados del mismo tamaño p, del tipo

$$A = \begin{bmatrix} A_1 & B_1 & & & \\ C_2 & A_2 & B_2 & & \\ & C_3 & \ddots & \ddots & \\ & & \ddots & \ddots & B_{n-1} \\ & & & C_n & A_n \end{bmatrix},$$

donde las matrices A_i son todas invertibles. Considere una descomposición por bloques $A = D+E+F$, donde $D = \begin{bmatrix} A_1 & & \\ & \ddots & \\ & & A_n \end{bmatrix}$, E es triangular inferior y F es superior. Describa los métodos de **Jacobi-bloques** y **Gauss-Seidel-bloques** que se obtienen con esta descomposición. Encuentre una condición suficiente de convergencia en $\| \ \|_\infty$.

2. Considere una matriz del tipo anterior con $n = 3$, $A_i = \begin{bmatrix} 5 & 5 \\ 0 & 5 \end{bmatrix}$, para $i = 1,2,3$, $B_1 = B_2 = C_2 = C_3 = I_2$, la identidad de tamaño 2. Pruebe que la matriz de iteración de Jacobi usual no satisface la condición de convergencia en $\| \ \|_\infty$ y que en cambio la matriz de iteración Jacobi-bloques, desarrollado en el ejercicio anterior, satisface esta condición.

   Considerando al sistema $Ax = b$, de lado derecho $b = \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 1 \\ 1 \end{pmatrix}$, realice 3 iteraciones del método de Jacobi-bloques y 3 iteraciones del método de Gauss-Seidel bloques.

3. Para calcular la inversa de una matriz invertible A, se propone el método iterativo que sigue.

   Dada la matriz C_0, calcular la sucesión de matrices C_i definidas por $C_{k+1} = C_k(I + R_k)$, donde $R_k = I - AC_k$, es el residuo de la iteración k-ésima.

   Suponga que la adivinanza C_0 satisface la condición

$$\|R_0\| = \delta < 1$$

   en alguna norma subordinada y demuestre que

   (a) todas las matrices C_i son invertibles,

   (b) si el método converge entonces converge a la inversa de A,

   (c) la relación de residuos consecutivos es $R_{k+1} = R_k^2$,

   (d) la evolución del error $E_k = A^{-1} - C_k$, es tal que $\|E_k\| \leq \delta^{2^k-1}\|E_0\|$.

4. Demuestre que toda matriz diagonal dominante es invertible.

5. Utilice el algoritmo descrito en el ejercicio 3 Para calcular la inversa de la matriz A dada en el ejercicio 2, partiendo de la adivinanza $C_0 = \begin{bmatrix} G & U & V \\ U & G & U \\ V & U & G \end{bmatrix}$ con $G = \begin{bmatrix} 0.2 & -0.2 \\ 0 & 0.2 \end{bmatrix}$ $U = \begin{bmatrix} -0.04 & 0.09 \\ 0 & -0.04 \end{bmatrix}$, $V = \begin{bmatrix} 0.008 & -0.02 \\ 0 & 0.008 \end{bmatrix}$.

6. Considere la matriz de tamaño 6 del problema 2 y calcule la matriz de iteración $M_R(w)$ definida en (4.51). Calcule su norma infinito en función de w y pruebe que tampoco satisface la condición suficiente de convergencia $\forall w \in (0,2)$. A pesar de que no está garantizada la convergencia de este método realice 3 iteraciones de Gauss-Seidel usual partiendo de la misma adivinanza usada en las iteraciones de bloques.

CAPÍTULO 5

Valores y vectores propios

El problema del cálculo efectivo de los valores y vectores propios de una matriz dada es de una complejidad mucho mayor que el otro problema clásico del Álgebra Lineal, Sistemas Lineales, abordado en el capítulo anterior. En el desarrollo de este capítulo, supondremos conocidos todos aquellos resultados contenidos en el curso de Álgebra Lineal de primer año, en particular, los teoremas acerca de diagonalización de matrices. En lo que sigue, consideraremos solo matrices cuadradas y reales, agregando como comentario, cuando corresponda, una referencia al caso complejo. Resolver el problema general de encontrar todos los valores propios y todos los vectores propios de una matriz cualquiera, además de ser costoso tiene dificultades numéricas originadas en la inestabilidad que suele tener este problema. Por esta razón se estudian métodos especializados en casos particulares como son las matrices simétricas, para las cuales el problema es estable, o bien destinados a obtener solo parte de la información, como serían solo los valores propios o solo una pareja de valor y vector propio.

Los temas que abordaremos son:

- Localización de Valores Propios
- Estabilidad del problema de Valores Propios
- Cálculo efectivo de Vectores Propios
- Cálculo efectivo de Valores Propios

LOCALIZACIÓN DE VALORES PROPIOS

En muchas situaciones el interés por los valores propios de una matriz se limita a tener una idea aproximada de su magnitud. Este es el caso, por ejemplo, cuando nos preguntamos por el condicionamiento de un sistema lineal y también cuando se quiere estimar la velocidad de convergencia de un método iterativo para resolver sistemas lineales.

El primer resultado de localización ya fue presentado, como una de las propiedades de las normas matriciales subordinadas. La propiedad (4.9) dice que el radio espectral de una matriz está acotado superiormente por su norma, cualquiera sea la norma subordinada que se considere y por lo tanto se verifica la propiedad que sigue.

Propiedad 5.1. Sea $\sigma(A)$ el espectro de la matriz A, es decir, el conjunto de todos sus valores propios. Sea $\| \ \|$ una norma matricial subordinada. Entonces se cumple que

$$\forall \lambda \in \sigma(A) \quad |\lambda| \leq \|A\|.$$

Ejemplo 1. La matriz tri-diagonal y simétrica que se obtiene en el proceso de cálculo de la Spline cúbica de interpolación con nodos esquiespaciados es

$$A = \begin{bmatrix} 2 & 1 & & & & & \\ 1 & 4 & 1 & & & & \\ & 1 & 4 & \ddots & & & \\ & & \ddots & \ddots & 1 & & \\ & & & 1 & 4 & 1 & \\ & & & & 1 & 2 & \end{bmatrix}.$$

Como $\|A\|_\infty = \|A\|_1 = 6$, se puede concluir que $\forall \lambda \in \sigma(A), -6 \leq \lambda \leq 6$, ya que por ser A simétrica, todos sus valores propios son reales.

Estas cotas son bastantes gruesas y nunca permitirán acotar inferiormente el módulo de los valores propios. Este resultado de localización no reconoce que la matriz del ejemplo anterior es definida positiva y por lo tanto invertible. El teorema que sigue entrega información mucho más precisa acerca de la ubicación de los valores propios con un trabajo equivalente a calcular la norma infinito de la matriz. Recordaremos que para las matrices complejas el problema de existencia de los valores propios no es tal y por lo tanto una matriz real de tamaño n, mirada como matriz compleja, siempre tendrá n valores propios complejos. (*Insistiendo en los recuerdos: esto se debe al Teorema Fundamental del Álgebra que garantiza la existencia de exactamente n raíces de un polinomio complejo de grado n, en este caso, el polinomio característico*). Aún cuando hemos centrado nuestro interés en las matrices reales, presentaremos el teorema que sigue en su versión general, es decir, para matrices complejas, debido al comentario previo.

Teorema 5.2 (Gerschgorin). *Dada una matriz $A = (a_{i,j})_{i,j=1}^n$ compleja, se definen los radios*

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

y los círculos en el plano complejo

$$Z_i = \{z \in \boldsymbol{C} \| z - a_{i,i}| \leq r_i\}.$$

$\forall \lambda \in \sigma(A)$ existe algún círculo Z_k que lo contiene. Si la unión de m círculos forma un conjunto conexo S, disjunto de los restantes $(n-m)$ círculos, entonces existen exactamente m valores propios de A en S.

Demostración. Sea $\lambda \in \sigma(A)$ (aún si A es real, este valor propio puede ser complejo) y x un vector propio asociado (que será también, en general, complejo). Se define la norma infinito para vectores complejos, del mismo modo que sobre $\mathbb{R}^n$, usando el módulo complejo en lugar del módulo real. (La norma matricial subordinada correspondiente, para matrices complejas, coincidirá con la norma infinito de matrices reales, cuando la matriz sea real, como lo es en nuestro caso). Sea k el índice de la coordenada donde se realiza el máximo de los módulos de las coordenadas de x, es decir $|x_k| = \|x\|_\infty \neq 0$, por ser x un vector propio.

Como $Ax = \lambda x$, tenemos que $\sum_{j=1}^{n} a_{k,j} x_j = \lambda x_k$, lo que equivale a $(\lambda - a_{k,k}) x_k = \sum_{\substack{j=1 \\ j \neq k}}^{n} a_{k,j} x_j$, y por lo tanto

$$|\lambda - a_{k,k}| \; |x_k| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{k,j}| \; |x_j| \leq r_k \|x\|_\infty,$$

lo que prueba la primera afirmación del teorema.

Para demostrar la segunda afirmación, debemos recordar que los valores propios (raíces del polinomio característico, cuyos coeficientes son funciones continuas de los coeficientes de la matriz) son funciones continuas de los coeficientes de la matriz. Consideremos una perturbación "extra-diagonal" de la matriz A, es decir, sean

$$D = diag(a_{1,1}, a_{2,2}, ..., a_{n,n}), \quad E = A - D,$$

para $0 \leq \varepsilon \leq 1$, se define la matriz perturbada

$$A(\varepsilon) = D + \varepsilon E,$$

cuyos valores propios denotaremos por $\lambda_i(\varepsilon)$, para $i = 1, 2, ..., n$, que serán funciones continuas de ε. Obviamente $A(1) = A$ y $\forall i, \lambda_i(1) = \lambda_i$, los valores propios de A. De la primera parte de la demostración sabemos que todo valor propio de la matriz perturbada debe pertenecer a algún círculo

$$Z_k(\varepsilon) = \{z \in C || z - a_{k,k}| \leq \varepsilon \cdot r_k\}.$$

Cuando ε tiende a cero, los círculos se hacen pequeños y, a menos que coincidan sus centros, para ε suficientemente pequeño serán disjuntos, resulta evidente que

$$\lambda_i(0) = a_{i,i}, \quad \forall i = 1, 2, ..., n,$$

que para todo ε la suma de las multiplicidades algebraicas de los valores propios (complejos) debe ser exactamente n y que en la medida que ε avanza desde 0 hasta 1, cada $\lambda_i(\varepsilon)$ recorre un camino continuo dentro del círculo cuyo radio se va expandiendo junto con ε. Con esto concluye la demostración del teorema. □

Ejemplo 2. Aplicando este teorema a la matriz del ejemplo anterior, se tendrá que
$Z_1 = Z_n = \{z \in \boldsymbol{C} || z - 2| \leq 1\}$ y $\forall i = 2, 3, ..., n-1,\ Z_i = \{z \in \boldsymbol{C} || z - 4| \leq 2\}$.

Pero como los valores propios de A son reales (por tratarse de una matriz simétrica), se concluye que $\forall \lambda \in \sigma(A),\ 1 \leq \lambda \leq 6$ y el condicionamiento de A en radio espectral será

$$K_*(A) \leq \frac{6}{1} = 6.$$

Ejemplo 3. Otro ejemplo que permite comprender el argumento de continuidad usado en la demostración es el siguiente.

Consideremos la matriz $A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 1 & -4 \end{bmatrix}$. El teorema previo asegura que los tres valores propios complejo, se ubican en los círculos con centro real

$$|\lambda - 4| \leq 1, \quad |\lambda| \leq 2, \quad |\lambda + 4| \leq 2.$$

Como el primer círculo es disjunto de los otros dos, debe haber un valor propio, aislado en el primero, y como las raíces complejas de un polinomio a coeficientes reales (como es el polinomio característico de una matriz real) se presentan de a pares conjugados, entonces esta raíz aislada tiene que ser real, es decir,

$$\lambda_1 \in [3, 5].$$

Para todo $\varepsilon < 1$, los otros dos círculos perturbados quedan disjuntos, cada uno con un valor propio de la matriz perturbada y por el mismo argumento de paridad de las raíces complejas se concluirá que ambos son reales. Por continuidad se tiene que

$$\lambda_2 \in [-6, -2], \lambda_3 \in [-2, 2]$$

y como se comprueba fácilmente que el punto de intersección, 2, no es valor propio, entonces deben ser dos valores propios simples y aislados

$$\lambda_2 \in [-6, -2), \lambda_3 \in (-2, 2].$$

ESTABILIDAD DEL PROBLEMA DE VALORES PROPIOS

Tal como vimos en el problema de sistemas lineales, quisiéramos saber cuanto cambia la solución del problema actual, es decir, cuanto cambian los valores propios, al perturbar ligeramente los datos de entrada, es decir, los coeficientes de la matriz A. En el caso de sistemas lineales pudimos definir un indicador de la estabilidad: el número de condicionamiento de la matriz. Veremos que para el problema de valores propios es mucho más difícil definir un indicador similar. Debido a esta complejidad, nos limitaremos al caso de matrices **diagonalizables**. Recordemos que una matriz A, se dirá diagonalizable sí y sólo sí, existen las matrices P invertible y D diagonal, tales que

$$A = PDP^{-1},$$

es decir, A es similar a una matriz diagonal. (Esta factorización no es única.) En tal caso, en la diagonal de D están los valores propios de A y en las columnas de P están los vectores propios correspondientes en el mismo orden.

Teorema 5.3 (Bauer-Fike). *Sea A una matriz diagonizable y consideremos la matriz perturbada $\widetilde{A} = A + E$. Sea $\|\ \|$ una norma matricial subordinada con la propiedad adicional, para matrices diagonales, que sigue*

(5.4) $$\text{Si } G = diag(g_1, g_2, ..., g_n), \text{ entonces } \|G\| = \max_{1 \le i \le n} |g_i|.$$

Si $A = PDP^{-1}, D = diag(\lambda_1, \lambda_2, ..., \lambda_n)$ y λ es un valor propio de la matriz perturbada $\widetilde{A}$, entonces

(5.5) $$\min_{1 \le i \le n} |\lambda - \lambda_i| \le \|P\| \cdot \|P^{-1}\| \cdot \|E\|.$$

Demostración. Si λ fuera un valor propio de A, entonces el lado izquierdo de (5.5) sería cero y la desigualdad trivialmente cierta. Supondremos entonces que λ no es valor propio de A, es decir, que para todo $i, \lambda \neq \lambda_i$. Sea x, un vector propio asociado a λ. Entonces

$$(A + E)x = \lambda x,$$

lo que equivale a

$$(\lambda I - A)x = Ex$$

y por lo tanto se tiene

(5.6) $$P(\lambda I - D)P^{-1}x = Ex,$$

que implica

$$P^{-1}x = (\lambda I - D)^{-1}(P^{-1}EP)P^{-1}x$$

y

$$\|P^{-1}x\| \le \|(\lambda I - D)^{-1}\| \cdot \|P^{-1}EP\| \cdot \|P^{-1}x\|.$$

Cancelando $\|P^{-1}x\|$ y usando la propiedad (5.4) se llega a que

$$1 \leq \max_{1\leq i\leq n} \left| \frac{1}{\lambda - \lambda_i} \right| \|P^{-1}\| \cdot \|P\| \cdot \|E\|,$$

lo que equivale a (5.5) y completa la demostración. □

Hacemos ver que las normas matriciales subordinadas más usadas, en particular, $\| \ \|_1 \cdot \| \ \|_2$ y $\| \ \|_\infty$, satisfacen la propiedad (5.4).

El aspecto de la desigualdad (5.5) sugiere que un posible número de condicionamiento (a pesar de que no aparecen errores ni perturbaciones *relativas*, como en el caso de sistemas lineales) del problema de valores propios sea

$$\|P\| \cdot \|P^{-1}\|. \tag{5.7}$$

Pero esto no quedaría bien definido, ya que la diagonalización de A no es única. Para una definición precisa del condicionamiento de este problema habría que tomar el ínfimo de las cantidades descritas en (5.7) sobre todas las matrices P que permiten diagonalizar la matriz A.

Corolario 5.8. Si A es simétrica y si se considera la norma euclidiana, entonces el teorema anterior produce la desigualdad

$$\min_{1\leq i\leq n} |\lambda - \lambda_i| \leq \|E\|_2.$$

Demostración. Si la matriz A es simétrica, entonces existe una diagonalización ortogonal, es decir, la matriz P será tal que $P^t = P^{-1}$. En tal caso

$$\|P^{-1}x\|_2 = \|Px\|_2 = \|x\|_2$$

y por esta misma razón

$$\|P\|_2 = \|P^{-1}\|_2 = 1.$$

De este modo, de la ecuación (5.6) se deduce que

$$\|x\|_2 = \|P^{-1}x\|_2 \leq \|(\lambda I - D)^{-1}\|_2 \|E\|_2 \|x\|_2.$$

Cancelando $\|x\|_2$, se concluye (5.8). □

Hacemos ver que, en cierto sentido, el corolario dice que el condicionamiento de las matrices simétricas, para el problema de valores propios, es óptimo. la vaguedad introducida en esta aseveración se debe a que en la desigualdad (5.8) no participan ni errores relativos ni perturbaciones relativas, limitando así la calidad de la comparación. De todos modos, son las matrices simétricas acerca de las que tenemos la mayor cantidad de resultados, como el teorema que entregamos a continuación sin demostración.

Teorema 5.9 (Wielandt-Hoffman). *Si A y E son matrices reales y simétricas de tamaño n, y si $\widetilde{A} = A + E$, es una perturbación de A, cuyos valores propios denotaremos por $\widetilde{\lambda}_i$, para $i = 1, 2, ..., n$, ordenados de menor a mayor, para facilitar la comparación con $\lambda_i, i = 1, 2, ..., n$, los valores propios de A, ordenados del mismo modo, entonces la norma de Frobenius de la matriz de perturbación E acota el desplazamiento de los valores propios según*

$$\left(\sum_{i=1}^{n}(\lambda_i - \widetilde{\lambda}_i)^2\right)^{1/2} \leq \left(\sum_{i,j=1}^{n} E_{i,j}^2\right)^{1/2} = \|E\|_F. \tag{5.10}$$

Del capítulo anterior sabemos que difícilmente un teorema de estabilidad permite su aplicación práctica directa. Este último resultado, sin embargo, tiene la virtud de la simpleza de la norma de Frobenius y que con frecuencia se conoce una cota de la perturbación de los coeficientes de la matriz, es decir,

$$|E_{i,j}| \leq \varepsilon, \forall i, j = 1, 2, ..., n,$$

lo que simplifica aún más la desigualdad (5.10).

Veremos a continuación como usar el residuo, que es un vector calculable, para acotar el error de una solución numérica, por lo tanto afectada por perturbaciones.

Sea A una matriz simétrica, de valores propios $\lambda_1, \lambda_2, ..., \lambda_n$ y sean λ y x, una pareja de valor y vector propio calculados numéricamente. Consideremos el vector propio normalizado en norma euclidiana, es decir, $\|x\|_2 = 1$, y sea

$$\eta = Ax - \lambda x,$$

el residuo.

Como A es simétrica, existen una matriz ortogonal P y una matriz diagonal $D = diag(\lambda_1, \lambda_2, .., \lambda_n)$, tales que

$$A = PDP^t.$$

Por lo tanto, el residuo satisface

$$P^t\eta = DP^tx - \lambda P^tx = (D - \lambda I)P^tx,$$

lo que equivale a

$$P^tx = (D - \lambda I)^{-1}P^t\eta,$$

si $\lambda \neq \lambda_i, \forall i = 1, 2, ..., n$.

Recordando que $\|P^tx\|_2 = \|x\|_2 = 1$, $\|P^t\eta\|_2 = \|\eta\|_2$ y que la norma matricial subordinada correspondiente satisface la propiedad (5.4), se tendrá

$$\min_{1 \leq i \leq n} |\lambda - \lambda_i| \leq \|\eta\|_2. \tag{5.11}$$

La utilidad práctica de las desigualdades (5.10) y (5.11) es evidente; la única restricción de estos resultados, es que solo son válidos para matrices simétricas.

MÉTODOS DE CALCULO DE VALORES Y VECTORES PROPIOS

Los métodos de cálculo podrían clasificarse según la estrategia que utilizan. En primer lugar veremos un método iterativo cuyo objetivo es aproximar un vector propio y como consecuencia de este proceso se obtiene

paralelamente una sucesión de aproximaciones del valor propio correspondiente. Otra clase de métodos resulta de un proceso iterativo de reducciones de la matriz, mediante transformaciones de similitud y por lo tanto preservando los valores propios. Uno de los métodos de este tipo es el método de Jacobi que, aplicado a matrices simétricas, genera una sucesión de matrices que convergen a una matriz diagonal y por lo tanto entregará sucesivas aproximaciones de todos los valores propios. El método QR es un procedimiento de esta misma clase, que genera una sucesión convergente a una matriz triangular superior, por lo tanto con los valores propios en la diagonal. Otros métodos de reducción, menos ambiciosos que los primeros, en un número finito de pasos transforman la matriz original en otra matriz similar, cuya estructura permite buscar aproximaciones iterativas de los valores propios con mayor eficiencia que en ausencia de dicha estructura. En síntesis, para resolver el problema de valores propios no contamos con métodos directos.

Método de la Potencia Iterada.

Si A es una matriz diagonizable, cuyo valor propio mayor en módulo está aislado, entonces, el método iterativo que se describe a continuación, converge a un vector propio asociado a dicho valor propio, como se establece en el teorema que sigue.

Teorema 5.12. *Si los valores propios de A satisfacen*

$$|\lambda_1| > |\lambda_2| \geq ... \geq |\lambda_n|$$

y si $v_1, v_2, ..., v_n$, denotan vectores propios linealmente independientes, asociados respectivamente a los valores propios anteriores, entonces la sucesión de vectores generada por

$$\begin{aligned} z^{(0)} &\text{ dado } \quad \forall k \geq 0, \\ w^{(k+1)} &= Az^{(k)}, \\ z^{(k+1)} &= \frac{1}{\left\| w^{(k+1)} \right\|_\infty} w^{(k+1)}, \end{aligned}$$

converge a $\pm \frac{1}{\|v_1\|_\infty} v_1$, cuando $k \to \infty$, siempre que la adivinanza inicial, $z^{(0)}$, tenga alguna componente en v_1. La velocidad de esta convergencia será mayor, mientras menor sea el cuociente $\left|\frac{\lambda_2}{\lambda_1}\right| < 1$.

Demostración. Como los vectores propios $v_1, v_2, ..., v_n$, forman una base de $\mathbb{R}^n$, entonces $z^{(0)} = \sum_{i=1}^{n} \alpha_i v_i$, con $\alpha_1 \neq 0$, para satisfacer la condición de tener alguna componente en el primer vector propio. De este modo se tendrá que

$$A^k z^{(0)} = \sum_{i=1}^{n} \alpha_i A^k v_i = \sum_{i=1}^{n} \alpha_i \lambda_i^k v_i = \alpha_1 \lambda_1^k \left(v_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k v_i \right). \tag{5.13}$$

Por otra parte

$$\begin{aligned} z^{(1)} &= \frac{1}{\left\| Az^{(0)} \right\|_\infty} Az^{(0)}, \\ z^{(2)} &= \frac{1}{\left\| w^{(2)} \right\|_\infty} w^{(2)} = \frac{1}{\left\| Az^{(1)} \right\|_\infty} Az^{(1)} = \frac{1}{\left\| \frac{1}{\left\| Az^{(0)} \right\|_\infty} A^2 z^{(0)} \right\|_\infty} \cdot \frac{1}{\left\| Az^{(0)} \right\|_\infty} A^2 z^{(0)} \\ &= \frac{1}{\left\| A^2 z^{(0)} \right\|_\infty} A^2 z^{(0)} \end{aligned}$$

y en general, se puede probar por inducción que $\forall k \geq 0$

$$z^{(k)} = \frac{1}{\left\| A^k z^{(0)} \right\|_\infty} A^k z^{(0)}.$$

Usando (5.13) en esta ecuación se obtiene

$$z^{(k)} = \frac{\sigma}{\left\| v_1 + \sum_{i=2}^{n} \left(\frac{\alpha_i}{\alpha_1}\right) \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i \right\|_\infty} \left(v_1 + \sum_{i=2}^{n} \left(\frac{\alpha_i}{\alpha_1}\right) \left(\frac{\lambda_i}{\lambda_1}\right)^k v_i \right), \tag{5.14}$$

donde σ denota un signo.

De la ecuación (5.14) se deducen todas las afirmaciones del teorema (5.12), que queda así demostrado. □

Hacemos ver que en cada iteración, se produce simultáneamente una aproximación del valor propio λ_1. En efecto, cualquiera sea el índice j de una componente no nula del vector $z^{(k-1)}$, el cuociente

$$\lambda_1^{(k)} = \frac{w_j^{(k)}}{z_j^{(k-1)}}$$

será una aproximación del valor propio que convergerá en la misma medida en que la sucesión de vectores producida por el método de la Potencia Iterada converja al vector propio. Para garantizar la división por una coordenada no nula se escoge el índice j donde se alcanza el máximo de los módulos de éstas, es decir,

$$|z_j^{(k-1)}| = \|z^{(k-1)}\|_\infty = 1.$$

Para un estudio más acabado de la velocidad de convergencia y alternativas de aceleración de ésta recomendamos consultar los libros de K.E. Atkinson [1] y de J. Stoer & R. Bulirsch [7].

Un método similar al presentado, destinado al cálculo de un vector propio asociado a un valor propio aislado λ_j, de una matriz A diagonizable, del cual se tenga una aproximación λ, se conoce bajo el nombre de **Método de la Potencia Inversa**. Consiste en iteraciones del mismo tipo que las del método de la Potencia Iterada, cambiando la matriz A por la matriz $(A - \lambda I)^{-1}$. En la práctica, en lugar de multiplicar por esta matriz inversa, se resuelve el sistema lineal correspondiente, es decir,

$$\begin{aligned} &\text{dado } z^{(0)} \in I\!R^n,\ \forall k \geq 0 \\ &\text{obtener } w^{(k+1)},\ \text{solución de } (A - \lambda I) w^{(k+1)} = z^{(k)}. \\ &z^{(k+1)} = \frac{1}{\|w^{(k+1)}\|_\infty} w^{(k+1)}. \end{aligned}$$

Para el estudio de la convergencia de este método sugerimos las mismas referencias bibliográficas.

Secuencia de Givens para el cálculo de valores propios de matrices tri-diagonales simétricas.

El problema abordado por este método es más general de lo que aparece, debido a que se cuenta con varios métodos que permiten transformar, en un número reducido de pasos, una matriz simétrica en otra tri-diagonal y simétrica similar a la anterior (y por lo tanto con los mismos valores propios).

Sea

$$A = \begin{bmatrix} \alpha_1 & \beta_1 & & & \\ \beta_1 & \alpha_2 & \beta_2 & & \\ & \beta_2 & \ddots & \ddots & \\ & & \ddots & \alpha_{n-1} & \beta_{n-1} \\ & & & \beta_{n-1} & \alpha_n \end{bmatrix}.$$

Supondremos que $\beta_i \neq 0,\ \forall i = 1, 2, ..., n-1$. Esto no resta generalidad al caso estudiado, pues de no cumplirse esta condición, el problema del cálculo de los valores propios de la matriz original se reducirá a dos problemas del tipo propuesto.

Desarrollando el Determinante de $(A - xI)$, que denotaremos por $p_n(x)$, por la última fila, se obtendrá que

$$\begin{aligned} p_n(x) &= -\beta_{n-1} Det \begin{bmatrix} \alpha_1 - x & \beta_1 & & & \\ \beta_1 & \alpha_2 - x & \ddots & & \\ & \ddots & \ddots & \beta_{n-3} & \\ & & \beta_{n-3} & \alpha_{n-2} - x & \\ & & & \beta_{n-2} & \beta_{n-1} \end{bmatrix} + (\alpha_n - x) p_{n-1}(x) \\ &= -\beta_{n-1}^2 p_{n-2}(x) + (\alpha_n - x) p_{n-1}(x). \end{aligned}$$

Definiendo $p_0(x) = 1$ y como $p_1(x) = \alpha_1 - x$, se puede definir de manera recursiva el polinomio característico de la matriz A como $p_n(x)$, donde

$$\forall k \geq 2 \qquad p_k(x) = (\alpha_k - x) p_{k-1}(x) - \beta_{k-1}^2 p_{k-2}(x). \tag{5.15}$$

La secuencia de polinomios de grados crecientes $p_0(x), p_1(x), ..., p_n(x)$, es muy particular y pertenece a una clase conocida por el nombre de Secuencias de Sturm, que recomendamos estudiar de los textos antes citados. Las principales propiedades de nuestra secuencia, llamada Secuencia de Givens, son:

- $\forall k \geq 1$ $p_k(x)$ tiene exactamente k raíces reales simples.
- $\forall k \geq 1$, si ξ es una raíz de $p_k(x)$, entonces $p_{k-1}(\xi) p_{k+1}(\xi) < 0$.
- Cuando t tiende a $-\infty$, $\quad p_k(t) > 0, \quad \forall k \geq 0$.
- El número de raíces de $p_n(x)$ ubicadas a la izquierda de t coincide con el número de cambios de signo de la secuencia $p_0(t), p_1(t), ..., p_n(t)$.

Esta última propiedad permite, mediante simples evaluaciones de la secuencia de polinomios, obtener intervalos disjuntos que contengan cada uno un valor propio de la matriz A.

Ejemplo 4. Para ilustrar este resultado, consideremos la matriz del **Ejemplo 1** de este capítulo con $n = 4$. Gracias al teorema de Gerschgorin (**Ejemplo 2**) sabemos que sus 4 valores propios pertenecen al intervalo $[1, 6]$, y por esta razón no tienen sentido las evaluaciones fuera de dicho intervalo. La secuencia de polinomios será

$$
\begin{aligned}
p_0(x) &= 1,\\
p_1(x) &= 2-x,\\
p_2(x) &= (4-x)p_1(x)-1,\\
p_3(x) &= (4-x)p_2(x)-p_1(x),\\
p_4(x) &= (2-x)p_3(x)-p_2(x).
\end{aligned}
$$

Evaluando en $t=2$, se obtiene la secuencia

$$
\begin{aligned}
p_0(2) &= 1,\\
p_1(2) &= 0,\\
p_3(2) &= -1,\\
p_4(2) &= 1,
\end{aligned}
$$

que presenta 2 cambios de signo, lo que indica que hay 2 valores propios menores que 2 y como no hay valores propios menores que 1 podemos concluir que existen 2 valores propios entre 1 y 2.

Evaluando la secuencia en $t=\frac{3}{2}$ se obtiene

$$
\begin{aligned}
p_0(\frac{3}{2}) &= 1,\\
p_1(\frac{3}{2}) &= \frac{1}{2},\\
p_2(\frac{3}{2}) &= \frac{1}{4},\\
p_3(\frac{3}{2}) &= \frac{1}{8},\\
p_4(\frac{3}{2}) &= -\frac{3}{16},
\end{aligned}
$$

con solo 1 cambio de signo, de lo que se concluye que hay un valor propio en $(1,\frac{3}{2})$ y un valor propio en $(\frac{3}{2},2)$. Ambos intervalos son abiertos pues como $p_4(1)\neq 0, \quad p_4(\frac{3}{2})\neq 0, \quad p_4(2)\neq 0$, ninguno de los extremos será valor propio.

Evaluando en $t=3, \quad t=4, \quad t=5$, se concluye que los cuatro valores propios satisfacen

$$\lambda_1 \in (1,\frac{3}{2}), \quad \lambda_2 \in (\frac{3}{2},2), \quad \lambda_3 \in (3,4), \quad \lambda_4 \in (5,6).$$

Una primera aproximación de los valores propios correspondería al centro de cada intervalo. Para obtener aproximaciones más precisas se puede seguir dividiendo en dos cada uno de estos intervalos y mediante el conteo de los cambios de signo decidir a cual de los dos subintervalos pertenece el valor propio localizado allí. Repitiendo este procedimiento se obtendrán aproximaciones con la precisión que se desee.

Método de Jacobi para el cálculo de valores propios de matrices simétricas.

Sea A una matriz real y simétrica de tamaño n. El objetivo de este método es generar una sucesión de matrices simétricas similares a la matriz A, que converja a una matriz diagonal. Cada iteración consiste en una transformación de similitud destinada a producir un cero (por simetría serán dos ceros) en una posición extra-diagonal, la sucesión de matrices comienza con

$$A^{(0)} = A.$$

Supongamos que tenemos una matriz $A^{(k)}$ simétrica, que en la posición (p,q) tiene un coeficiente no nulo, es decir,

$$a_{p,q}^{(k)} \neq 0, \text{ con } p \neq q.$$

Debemos construir una matriz ortogonal Q_k $(Q_k^t = Q_k^{-1})$, tal que

$$A^{(k+1)} = Q_k A^{(k)} Q_k^t \tag{5.16}$$

tenga nulo el coeficiente de la posición (p,q), es decir,

$$a_{p,q}^{(k+1)} = 0.$$

Esto se consigue con Q_k matriz de rotación en el plano (p,q) para una elección acertada del ángulo de rotación que llamaremos θ_k. Sea

$$Q_k = \begin{bmatrix} 1 & & & & & & & & & \\ & \ddots & & & & & & & & \\ & & 1 & & & & & & & \\ & & & c & & & & -s & & \leftarrow p \\ & & & & 1 & & & & & \\ & & & & & \ddots & & & & \\ & & & & & & 1 & & & \\ & & & s & & & & c & & \leftarrow q \\ & & & & & & & & 1 & \\ & & & & & & & & & \ddots \end{bmatrix} \tag{5.17}$$

$$\begin{matrix} & & \uparrow & & \uparrow \\ & & p & & q \end{matrix}$$

donde

$$c = cos(\theta_k), s = sen(\theta_k)$$

y todos los coeficientes no detallados son nulos. Es evidente que cualquiera sea el ángulo de rotación considerado, la matriz Q_k será ortogonal. Por otra parte, debido a que la matriz Q_k difiere de la identidad solo en dos filas y las mismas dos columnas, la matriz $A^{(k+1)}$ diferirá de la matriz $A^{(k)}$ solo en esas dos filas y dos columnas. Para obtener las fórmulas del reajuste y para determinar el ángulo apropiado a nuestro fin, simplificaremos la notación, evitando los super-índices, mediante la matriz auxiliar $B = A^{(k)} Q_k^t$, de modo que $A^{(k+1)} = Q_k B$.

$$\begin{aligned} \forall i: \quad & \forall j \neq p,q, \ B_{i,j} = a_{i,j}^{(k)}, \\ & B_{i,p} = c a_{i,p}^{(k)} - s a_{i,q}^{(k)}, \\ & B_{i,q} = s a_{i,p}^{(k)} + c a_{i,q}^{(k)}. \end{aligned} \tag{5.18}$$

$$\begin{aligned} \forall j: \quad & \forall i \neq p,q, \ a_{i,j}^{(k+1)} = B_{i,j}, \\ & a_{p,j}^{(k+1)} = c B_{p,j} - s B_{q,j}, \\ & a_{q,j}^{(k+j)} = s B_{p,j} + c B_{q,j}. \end{aligned} \tag{5.19}$$

El ángulo apropiado será tal que

$$\begin{aligned} 0 = a_{p,q}^{(k+1)} &= cB_{p,q} - sB_{q,q} = c(sa_{p,p}^{(k)} + ca_{p,q}^{(k)}) - s(sa_{qp}^{(k)} + ca_{q,q}^{(k)}) \\ &= cs(a_{p,p}^{(k)} - a_{q,q}^{(k)}) + a_{p,q}^{(k)}(c^2 - s^2) \\ &= \frac{1}{2} sen(2\theta_k)(a_{p,p}^{(k)} - a_{q,q}^{(k)}) + cos(2\theta_k) a_{p,q}^{(k)}, \end{aligned}$$

lo que equivale a que

$$tg(2\theta_k) = \frac{2a_{p,q}^{(k)}}{a_{q,q}^{(k)} - a_{p,p}^{(k)}}. \tag{5.20}$$

Esta expresión es poco apropiada para calcular el ángulo θ_k. De hecho, nada asegura que el denominador que allí aparece sea distinto de cero. Por otra parte no es el ángulo de rotación el que interesa conocer, sino los valores de $c = cos(\theta_k)$ y $s = sen(\theta_k)$. Un método estable para obtener estos valores consiste en calcular

$$c = \frac{1}{\sqrt{1+t^2}}, \quad s = tc,$$

donde $t = tg(\theta_k)$ se calcula como la raíz más pequeña en módulo del polinomio
$p(t) = t^2 + 2t\varphi - 1$, con $\varphi = \frac{a_{q,q}^{(k)} - a_{p,p}^{(k)}}{2a_{p,q}^{(k)}} = cot(2\theta_k)$, que estará bien definida puesto que $a_{p,q}^{(k)} \neq 0$.

La convergencia del método de Jacobi a una matriz diagonal, se concluye de la disminución sistemática (en cada iteración) de una medida global del tamaño de los coeficientes extra-diagonales, definida por

$$S_k = \sum_{i \neq j} \left(a_{i,j}^{(k)} \right)^2.$$

De acuerdo a las fórmulas del reajuste (5.18) se tiene

$$\forall i, \quad \forall j \neq p, q, \quad B_{i,j}^2 = \left(a_{i,j}^{(k)} \right)^2$$

y

$$\begin{aligned} B_{i,p}^2 + B_{i,q}^2 &= \left(ca_{i,p}^{(k)} - sa_{i,q}^{(k)} \right)^2 + \left(sa_{i,p}^{(k)} + ca_{i,q}^{(k)} \right)^2 \\ &= (c^2 + s^2) \left((a_{i,p}^{(k)})^2 + (a_{i,q}^{(k)})^2 \right) \\ &= \left(a_{i,p}^{(k)} \right)^2 + \left(a_{i,q}^{(k)} \right)^2. \end{aligned}$$

Del mismo modo se puede probar un resultado análogo, a partir de las fórmulas (5.19), de lo que se concluye que las sumas totales se conservan, es decir,

$$\sum_{i,j} \left(a_{i,j}^{(k+1)} \right)^2 = \sum_{i,j} B_{i,j}^2 = \sum_{i,j} (a_{i,j}^{(k)})^2.$$

Como debemos descartar los elementos diagonales de esta cuenta y los únicos de éstos que se modifican en una iteración son aquellos de las posiciones (p,p) y (q,q), revisamos su participación en las sumas previas, de las filas p y q de la matriz $A^{(k+1)}$

$$\left(a_{p,p}^{(k+1)} \right)^2 + \left(a_{q,p}^{(k+1)} \right)^2 = B_{p,p}^2 + B_{q,p}^2 \text{ y } \left(a_{p,q}^{(k+1)} \right)^2 + \left(a_{q,q}^{(k+1)} \right)^2 = B_{p,q}^2 + B_{q,q}^2$$

y de las columnas p y q de la matriz B

$$B_{p,p}^2 + B_{p,q}^2 = \left(a_{p,p}^{(k)}\right)^2 + \left(a_{p,q}^{(k)}\right)^2$$

y

$$B_{q,p}^2 + B_{q,q}^2 = \left(a_{q,p}^{(k)}\right)^2 + \left(a_{q,q}^{(k)}\right)^2.$$

Por consiguiente se tendrá que

$$\left(a_{p,p}^{(k+1)}\right)^2 + \left(a_{q,q}^{(k+1)}\right)^2 = \left(a_{p,p}^{(k)}\right)^2 + 2\left(a_{p,q}^{(k)}\right)^2 + \left(a_{q,q}^{(k)}\right)^2, \text{ ya que } a_{p,q}^{(k+1)} = a_{q,p}^{(k+1)} = 0.$$

De lo que se concluye que

$$0 \leq S_{k+1} = S_k - 2\left(a_{p,q}^{(k)}\right)^2 < S_k \tag{5.21}$$

y por lo tanto la convergencia del método de Jacobi.

Ejemplo 5. Consideremos la matriz $A = \begin{bmatrix} 1 & 3 & 1 \\ 3 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$. Sean $p = 1,\ q = 2$, con lo cual se tiene que $\varphi = 0$ y $t = 1,\ c = s = \frac{1}{\sqrt{2}}$. De las fórmulas (5.18) se concluye que

$$B = \begin{bmatrix} \frac{-2}{\sqrt{2}} & \frac{4}{\sqrt{2}} & 1 \\ \frac{2}{\sqrt{2}} & \frac{4}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 1 \end{bmatrix}.$$

Usando la simetría de $A^{(1)}$, la elección del ángulo y calculando $a_{1,1}^{(1)}$ y $a_{2,2}^{(1)}$ con las fórmulas (5.19) se obtiene

$$A^{(1)} = \begin{bmatrix} -2 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 4 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 1 \end{bmatrix}.$$

Si seleccionamos ahora el elemento (3,1) para su anulación en la iteración siguiente, se obtiene

$$\varphi = \frac{3\sqrt{2}}{2},\quad t = \frac{\sqrt{22} - 3\sqrt{2}}{2},\quad c = \frac{1}{\sqrt{11 - 3\sqrt{11}}},\quad s = \frac{\sqrt{22} - 3\sqrt{2}}{2\sqrt{11 - 3\sqrt{11}}},$$

$$B = \begin{bmatrix} -u(\sqrt{2} + \sqrt{22}) & 0 & 2u(7 - 2\sqrt{11}) \\ -u(3\sqrt{2} - \sqrt{22}) & 4 & 2u \\ u(8 - \sqrt{44}) & \frac{1}{\sqrt{2}} & u(\sqrt{22} - \sqrt{2}) \end{bmatrix}, \text{ con } u = \frac{1}{2\sqrt{2}\sqrt{11 - 3\sqrt{11}}}.$$

Aún antes de calcular los coeficientes (1,1) y (3,3) con las fórmulas (5.19) se puede observar que se des-anulan los ceros obtenidos en la iteración anterior. En efecto

$$a_{2,1}^{(2)} = a_{1,2}^{(2)} = u(3\sqrt{2} - \sqrt{22}) \neq 0.$$

Por otra parte, aún sin completar la iteración se puede observar la convergencia, pues $S_0 = 20$, $S_1 = 2$, $S_2 = \frac{1}{4}$, lo que concuerda con (5.21).

Método de reducción de Givens.

Las mismas transformaciones de similitud del método de Jacobi, realizadas ordenadamente según un algoritmo que presentaremos y eligiendo el ángulo de rotación de modo de anular el elemento $(q, p-1)$ en lugar del elemento (q,p) permiten reducir una matriz simétrica a una matriz similar, tri-diagonal y simétrica.

El orden en que se elige el plano de rotación (p,q) es

$$\begin{array}{ccccc} (2,3) & (2,4) & (2,5) & \cdots & (2,n) \\ & (3,4) & (3,5) & \cdots & (3,n) \\ & & (4,5) & \cdots & (4,n) \\ & & & \ddots & \vdots \\ & & & & (n-1,n). \end{array} \tag{5.22}$$

Una posible elección del ángulo de rotación que produce la anulación del coeficiente $(q, p-1)$ corresponde a

$$s = \frac{a^{(k)}_{q,p-1}}{\left(\left(a^{(k)}_{p,p-1}\right)^2 + \left(a^{(k)}_{q,p-1}\right)^2\right)^{1/2}}, \quad c = \frac{-a^{(k)}_{p,p-1}}{\left(\left(a^{(k)}_{p,p-1}\right)^2 + \left(a^{(k)}_{q,p-1}\right)^2\right)^{1/2}}. \tag{5.23}$$

Método de reducción de Housholder.

Sea $u \in \mathbb{R}^n$ normalizado tal que $\|u\|_2 = 1$. Se define la matriz de Housholder asociada a este vector como

$$H_n(u) = I_n - 2uu^t. \tag{5.24}$$

Se comprueba fácilmente que

$$(H_n(u))^t = (H_n(u))^{-1} = H_n(u). \tag{5.25}$$

Dado un vector $v \in \mathbb{R}^n$ siempre se puede construir un vector u tal que el producto de la matriz de Housholder correspondiente por el vector v resulte ser una ponderación del primer vector de la base canónica de $\mathbb{R}^n$, es decir, que tenga todas las coordenadas, salvo la primera, nulas. En efecto, si

$$\begin{array}{l} w = v + \|v\|_2 e_1, \text{ con } e_1 \in \mathbb{R}^n, \text{ el primer vector canónico y} \\ u = \dfrac{1}{\|w\|_2} w, \text{ entonces} \end{array} \tag{5.26}$$

$$H_n(u) \cdot v = \mu e_1, \text{ con } \mu = -\|v\|_2. \tag{5.27}$$

El método de Housholder usará esta propiedad en un algoritmo que con $(n-2)$ transformaciones de similitud reduce la matriz original a una de Hessenberg superior, es decir, que tenga nulos los coeficientes de índices (i,j) $\forall i \geq j+2$.

Sea Ω_p una matriz cuadrada de tamaño n, definida por bloques como

$$\Omega_p = \begin{bmatrix} I_p & 0_{p\times(n-p)} \\ 0_{(n-p)\times p} & H_{n-p} \end{bmatrix}, \tag{5.28}$$

donde $H_{n-p} = I_{n-p} - 2uu^t$ es una matriz de Housholder de tamaño $(n-p)$. Es evidente que

$$(\Omega_p)^t = \Omega_p = (\Omega_p)^{-1} \quad \forall p.$$

Sean $A^{(0)} = A, \quad A^{(k)} = \Omega_k A^{(k-1)} \Omega_k \quad \forall k = 1, 2, ..., n-2$, donde la matriz de Housholder que interviene en la definición (5.28), se construye según (5.26) con

$$v = \begin{pmatrix} a^{(k)}_{k+1,k} \\ a^{(k)}_{k+2,k} \\ \vdots \\ a^{(k)}_{n,k} \end{pmatrix} \in I\!R^{n-k}$$

y que por lo tanto, separando por bloques la matriz $A^{(k)}$ y usando la propiedad (5.27), se tendrá que no se reajustan los ceros construidos en las transformaciones de similitud previas y se anulan los coeficientes $(k+2,k), (k+3,k), ..., (n,k)$. Ilustraremos este comportamiento para $k = 1$ y $k = 2$.

Como $\Omega_1 = \begin{bmatrix} 1 & 0 \\ 0 & H_{n-1} \end{bmatrix}$, describimos la matriz $A^{(0)} = A$ por bloques como

$$A = \begin{bmatrix} 1 & w^t \\ v & B \end{bmatrix} \text{ y por lo tanto } A^{(1)} = \begin{bmatrix} a_{1,1} & w^t H_{n-1} \\ H_{n-1}v & H_{n-1}BH_{n-1} \end{bmatrix}.$$

Pero por construcción de la matriz de Housholder según (5.26), se tendrá que

$$H_{n-1}v = \mu_1 e_1 \in I\!R^{n-1} \quad \text{de acuerdo (5.27)}$$

y por lo tanto

$$A^{(1)} = \begin{bmatrix} a_{1,1} & a^{(1)}_{1,2} & \cdots & \cdots & a^{(1)}_{1,n} \\ \mu_1 & a^{(1)}_{2,2} & \cdots & \cdots & a^{(1)}_{2,n} \\ 0 & \vdots & & & \vdots \\ \vdots & \vdots & & & \vdots \\ 0 & a^{(1)}_{n,2} & \cdots & \cdots & a^{(1)}_{n,n} \end{bmatrix}.$$

Como $\Omega_2 = \begin{bmatrix} I_2 & 0 \\ 0 & H_{n-2} \end{bmatrix}$ separaremos la matriz $A^{(1)}$ por bloques correspondientes $A^{(1)} = \begin{bmatrix} C & W^t \\ V & B \end{bmatrix}$, con V de $(n-2)$ filas y 2 columnas, $V = \begin{bmatrix} 0 & a^{(1)}_{3,2} \\ 0 & a^{(1)}_{4,2} \\ \vdots & \vdots \\ 0 & a^{(1)}_{n,2} \end{bmatrix}$.

Por consiguiente

$$A^{(2)} = \begin{bmatrix} C & W^t H_{n-2} \\ H_{n-2}V & H_{n-2}BH_{n-2} \end{bmatrix}.$$

Construyendo la matriz de Housholder H_{n-2} según (5.26) con $v = \begin{pmatrix} a_{3,2}^{(k)} \\ a_{4,2}^{(1)} \\ \vdots \\ a_{n,2}^{(1)} \end{pmatrix}$, se tendrá que

$$H_{n-2}V = \begin{bmatrix} 0 & \mu_2 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \end{bmatrix}.$$

Se observa claramente que tras $(n-2)$ de estas transformaciones de similitud se obtendrá una matriz $A^{(n-2)}$ con forma de Hessenberg superior o tri-diagonal simétrica en el caso de A simétrica.

Ejemplo 6. Consideremos la matriz

$$A = \begin{bmatrix} 5 & 4 & 3 & 0 \\ 4 & 1 & 0 & 1 \\ 3 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}.$$

Para la primera transformación de similitud $v = \begin{pmatrix} 4 \\ 3 \\ 0 \end{pmatrix}$ con $\|v\|_2 = 5$ y por lo tanto

$$u = \begin{pmatrix} \left(\frac{3}{\sqrt{10}}\right) \\ \left(\frac{1}{\sqrt{10}}\right) \\ 0 \end{pmatrix}, \quad H_3 = \begin{bmatrix} \left(\frac{-4}{5}\right) & \left(\frac{-3}{5}\right) & 0 \\ \left(\frac{-3}{5}\right) & \left(\frac{4}{5}\right) & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \Omega_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \left(\frac{-4}{5}\right) & \left(\frac{-3}{5}\right) & 0 \\ 0 & \left(\frac{-3}{5}\right) & \left(\frac{4}{5}\right) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

con lo cual se obtiene

$$A^{(1)} = \begin{bmatrix} 5 & -5 & 0 & 0 \\ -5 & 1 & 0 & \left(\frac{-4}{5}\right) \\ 0 & 0 & 1 & \left(\frac{-3}{5}\right) \\ 0 & \left(\frac{-4}{5}\right) & \left(\frac{-3}{5}\right) & 1 \end{bmatrix}.$$

Para la segunda transformación de similitud $v = \begin{pmatrix} 0 \\ \frac{-4}{5} \end{pmatrix}$, de lo que se obtiene $u = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix}$ y $H_2 = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$, por lo que

$$A^{(2)} = \begin{bmatrix} 5 & -5 & 0 & 0 \\ -5 & 1 & \left(\frac{-4}{5}\right) & 0 \\ 0 & \left(\frac{-4}{5}\right) & 1 & \left(\frac{-3}{5}\right) \\ 0 & 0 & \left(\frac{-3}{5}\right) & 1 \end{bmatrix}.$$

Método QR para el cálculo de valores propios.

Este método genera una sucesión de matrices ortogonalmente similares que convergen a una matriz triangular superior.

Si se tiene la factorización QR de una matriz $A^{(k)}$, es decir, se conocen las matrices Q_k ortogonal y R_k triangular superior, tales que $A^{(k)} = Q_k R_k$, entonces la matriz

$$A^{(k+1)} = Q_k^t A^{(k)} Q_k = R_k Q_k$$

es ortogonalmente similar a $A^{(k)}$. El método QR realizará iteraciones que constan de dos pasos. En cada iteración se debe obtener la factorización QR de la matriz proveniente de la iteración anterior y luego realizar el producto en el orden inverso al de la factorización. La sucesión de matrices se genera como

$$\begin{aligned} &A^{(0)} = A. \\ &\text{a) Obtener factorización } QR \text{ de } A^{(k)}, \text{ es decir calcular } Q_k \text{ ortogonal} \\ &\qquad \text{y } R_k \text{ triangular superior, tales que } A^{(k)} = Q_k R_k. \\ &\text{b) Calcular } A^{(k+1)} = R_k Q_k. \end{aligned} \tag{5.29}$$

Este procedimiento es particularmente eficiente cuando A es de Hessenberg superior. Usando matrices de Housholder para obtener la factorización QR pedida en a) se puede probar que toda la secuencia de matrices similares permanece de la forma de Hessenberg superior. Para evitar el manejo de superíndices probaremos esta afirmación para la primera iteración.

Sea A una matriz de tamaño n de Hessenberg superior. Consideremos las matrices Ω_p definidas en (5.28), donde las matrices de Housholder $H_{n-p}(u)$ han sido construidas según el procedimiento descrito en (5.26) de modo de obtener una matriz triangular superior

$$R = \Omega_{n-2} \cdot \ldots \cdot \Omega_1 \Omega_0 A. \tag{5.30}$$

Por la construcción de la matriz de Housholder según (5.26) se tendrá que el vector u apropiado para que el producto por Ω_0 anule la primera subcolumna de A, se obtendrá de

$$v = \begin{pmatrix} a_{1,1} \\ a_{2,1} \\ \vdots \\ a_{n,1} \end{pmatrix}.$$

Pero como A es de Hessenberg superior, v tiene solo 2 coordenadas no nulas y por consiguiente u también tendrá solo 2 coordenadas no nulas. La matriz de Housholder que resulta, será de Hessenberg superior y de la forma

$$\Omega_0 = \begin{bmatrix} * & * & 0 & \cdots & 0 \\ * & * & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix} \quad \text{y } \Omega_0 A = \begin{bmatrix} * & * & * & \cdots & \cdots & * \\ 0 & * & * & \cdots & \cdots & * \\ 0 & a_{3,2} & a_{3,3} & \cdots & \cdots & a_{3,n} \\ 0 & 0 & a_{4,3} & \cdots & \cdots & a_{4,n} \\ \vdots & \vdots & \vdots & & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n,n-1} & a_{n,n} \end{bmatrix}.$$

Del mismo modo se puede observar que en general estas matrices serán todas de Hessenberg superior, donde las únicas posiciones no nulas son las marcadas

$$\Omega_p = \begin{bmatrix} 1 & & & & & & \\ & \ddots & & & & & \\ & & 1 & & & & \\ & & & * & * & & \leftarrow p+1 \\ & & & * & * & & \leftarrow p+2 \\ & & & & & 1 & \\ & & & & & & \ddots \\ & & & & & & 1 \end{bmatrix} \tag{5.31}$$

y por consiguiente la premultiplicación por esta matriz solo afectará a dos filas (y la pos-multiplicación por esta matriz solo afectará dos columnas).

Definiendo

$$Q = (\Omega_{n-2} \cdot ... \cdot \Omega_1 \Omega_0)^t = \Omega_0 \Omega_1 \cdot ... \cdot \Omega_{n-2}, \tag{5.32}$$

se tiene la factorización QR de A.

Es fácil probar que la matriz Q, así definida, es de Hessenberg superior y que en general el producto de una matriz de Hessenberg superior por una matriz triangular superior, es de Hessenberg superior. En efecto, si Q es de Hessenberg superior y R es triangular superior, entonces

$$(QR)_{i,j} = \sum_k Q_{i,k} R_{k,j} = \sum_{k=i-1}^{j} Q_{i,k} R_{k,j} = 0, \quad \text{si} \;\; j \leq i-2.$$

Existen procedimientos para acelerar la convergencia, conocidos bajo el nombre de QR **con desplazamiento** (QR con "shift"), para cuyo estudio recomendamos la bibliografía citada en este capítulo.

Ejemplo 7. Consideremos la matriz simétrica

$$A = \begin{bmatrix} 4 & 3 & 0 \\ 4 & \left(\frac{9}{4}\right) & 3 \\ 0 & 3 & 1 \end{bmatrix}.$$

La primera matriz de Housholder para obtener factorización QR será

$$\Omega_0 = \begin{bmatrix} \left(\frac{-4}{5}\right) & \left(\frac{-3}{5}\right) & 0 \\ \left(\frac{-3}{5}\right) & \left(\frac{4}{5}\right) & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Como

$$\Omega_0 A = \begin{bmatrix} -5 & \left(\frac{-15}{4}\right) & \left(\frac{-9}{5}\right) \\ 0 & 0 & \left(\frac{12}{5}\right) \\ 0 & 3 & 1 \end{bmatrix},$$

entonces

$$\Omega_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & -1 & 0 \end{bmatrix},$$

$$R = \Omega_1 \Omega_0 A = \begin{bmatrix} -5 & \left(\frac{-15}{4}\right) & \left(\frac{-9}{5}\right) \\ 0 & -3 & -1 \\ 0 & 0 & \left(\frac{-12}{5}\right) \end{bmatrix}$$

y

$$Q = \Omega_0 \Omega_1 = \begin{bmatrix} \left(\frac{-4}{5}\right) & 0 & \left(\frac{3}{5}\right) \\ \left(\frac{-3}{5}\right) & 0 & \left(\frac{-4}{5}\right) \\ 0 & -1 & 0 \end{bmatrix},$$

con lo cual

$$A^{(1)} = RQ = \begin{bmatrix} \left(\frac{25}{4}\right) & \left(\frac{9}{5}\right) & 0 \\ \left(\frac{9}{5}\right) & 1 & \left(\frac{12}{5}\right) \\ 0 & \left(\frac{12}{5}\right) & 0 \end{bmatrix}.$$

EJERCICIOS PROPUESTOS

1. Considere la matriz $A = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 4 & 5 \end{bmatrix}$. Mediante el método de la Potencia Iterada partiendo de $z^{(0)} = \begin{pmatrix} 1/2 \\ 3/4 \\ 1 \end{pmatrix}$ aproxime el vector propio asociado al valor propio de mayor módulo y este mismo valor propio. En cada iteración utilice el residuo para acotar el error con que aproxima al valor propio y detenga sus iteraciones cuando este error sea menor que 0.01.

2. Usando el teorema de Gerschgorin localice los valores propios de la matriz anterior. Mediante el método de reducción de Givens transforme esta matriz en una matriz tri-diagonal simétrica similar a la primera. Se sabe que la matriz dada no es invertible y por lo tanto tiene un valor propio nulo. Utilizando la secuencia de Givens aproxime el otro valor propio con la misma precisión con que calculó el del problema **1**).

3. Describa un algoritmo que permita usar el método de Jacobi para aproximar los vectores propios de una matriz simétrica. Demuestre que al límite de este método encontrará una base ortonormal de vectores propios.

4. Utilice el método de la secuencia de Givens para obtener todos los valores propios de la matriz

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

5. Suponga que debe calcular los valores propios de una matriz simétrica cuyos coeficientes conoce con una precisión muy baja, pues provienen de mediciones hechas en terreno. Es decir, en lugar de conocer la matriz A, conoce una matriz $\widetilde{A}$ y sabe que $|a_{i,j} - \widetilde{a}_{i,j}| \leq 0.01 \quad \forall i, j = 1, 2, ..., n$. Cuenta con un software confiable para realizar estos cálculos y por lo tanto los errores numéricos serán despreciables en comparación del error en los coeficientes de la matriz. En estas condiciones, ¿qué puede decir acerca de la precisión con que conocerá los valores propios?

6. Demuestre que el Teorema de Gerschgorin sigue siendo válido si se cambia la definición de los radios y de los círculos por

$$r_j = \sum_{i \neq j} |a_{i,j}|, \quad Z_j = \{z \in \boldsymbol{C} \| z - a_{j,j}| \leq r_j\}, \quad \forall j = 1, 2, ..., n.$$

7. Utilizando el resultado anterior demuestre que todos los valores propios de la matriz

$$A = \begin{bmatrix} 4 & 1 & 0 \\ 1 & 0 & -1 \\ 1 & 1 & -4 \end{bmatrix}$$

son reales y pertenecen a los intervalos disjuntos $[-5, -3], [-2, 2]$ y $[3, 5]$.

8. Para completar la demostración de que el método QR genera una sucesión de matrices de Hessenberg superior, si la matriz A tiene esta forma, pruebe que la matriz Q definida en (5.31) y (5.32) es de Hessenberg superior.

9. Note que la matriz del problema **7)** tiene forma de Hessenberg inferior y por lo tanto su traspuesta (que tiene los mismos valores propios que A) es de Hessenberg superior. Aplique el método QR para obtener aproximaciones de los valores propios de la matriz del problema **7)**.

10. Explique la razón por la cual el método QR no permite aproximar los vectores propios como lo hace el método de Jacobi según el algoritmo obtenido en el problema **3)**.

11. Usando el Teorema de Gerschgorin pruebe que si A es diagonal dominante entonces es invertible.

12. Aplique todos los métodos para reducir y calcular valores propios a la matriz

$$A = \begin{bmatrix} 1 & 0 & 3 & 1 \\ 0 & 1 & 4 & 0 \\ 3 & 4 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

Realice iteraciones hasta que la suma de los cuadrados de los elementos extradiagonales sea inferior o igual a 0.02, o bien hasta que los valores propios aproximados tengan una precisión de 0.01.

13. Sea A simétrica de tamaño n con valores propios $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$. Demuestre que

$$\max_{\substack{x \in I\!R^n \\ x \neq 0}} \frac{x^t A x}{x^t x} = \lambda_1$$

y

$$\min_{\substack{x \in I\!R^n \\ x \neq 0}} \frac{x^t A x}{x^t x} = \lambda_n.$$

14. Para las matrices $A(\varepsilon)$ que siguen, estudie el comportamiento de los valores y vectores propios para $\varepsilon \to 0^+$, y para $\varepsilon = 0$

$$\begin{bmatrix} 1 & 1 \\ \varepsilon & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & \varepsilon \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 1 & 1 \\ 0 & 1+\varepsilon \end{bmatrix}.$$

CAPÍTULO 6

Ecuaciones No Lineales

Consideraremos, inicialmente, el problema de encontrar una raíz de una función no lineal con dominio y valores reales. Es decir, resolveremos la ecuación

$$f(x) = 0,$$

con $f : [a, b] \to I\!R$.

En una primera etapa supondremos que f es continua, que tiene solo 1 raíz $\alpha \in [a, b]$ y que ésta es simple, lo que implica que f cambia de signo en ese lugar. Un procedimiento muy intuitivo para generar una sucesión que converja a la raíz es el **Método de Bisección**, que describimos a continuación.

Sean x_0 y x_1 dos adivinanzas en $[a, b]$ con la propiedad de que los valores de f en esos puntos tengan signos distintos, es decir, $f(x_0)f(x_1)$ sea negativo. Como el único cambio de signo de f en $[a, b]$ ocurre en α, necesariamente α debe estar entre medio de x_0 y x_1, dicho de otro modo, tenemos un intervalo de precisión para α y por lo tanto el punto medio de este intervalo $x_2 = \frac{x_0+x_1}{2}$, no puede estar a mayor distancia de α que $\frac{|x_1-x_0|}{2}$, lo que significa que el error absoluto que comete x_2 como aproximación de la raíz se acota como

$$|\alpha - x_2| < \frac{|x_1 - x_0|}{2}.$$

Ahora comparamos el signo de f en x_2 con el signo de f en los puntos anteriores y elegimos el intervalo donde f cambie de signo, en cuyo centro pondremos a x_3, es decir, si $f(x_2)f(x_1) < 0$, entonces $x_3 = \frac{x_1+x_2}{2}$, si por el contrario $f(x_2)f(x_0) < 0$, entonces $x_3 = \frac{x_0+x_2}{2}$.

En cualquiera de los dos casos el nuevo intervalo de precisión para α será la mitad de largo que el intervalo anterior y x_3 aproxima a α con un error absoluto acotado por la mitad de la cota anterior

$$|\alpha - x_3| < \frac{|x_1 - x_0|}{4}.$$

En la figura 6.1 se ilustra la construcción presentada.

La generalización de este procedimiento se resume en el algoritmo siguiente.

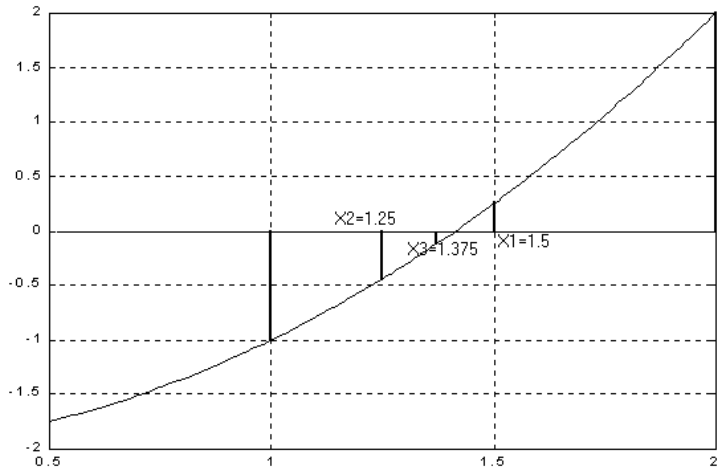


Figura 6.1: Iteraciones por Método de Bisección

Algoritmo 6.1. Dados x_0 y x_1 tales que $f(x_0)f(x_1) < 0$ para $n = 2, 3, ...$

$$x_n = \frac{x_{n-1} + x_{n-2}}{2}.$$

Si $f(x_n) = 0$ (en la práctica basta que sea suficientemente cercano a cero),
entonces termina el proceso con $\alpha = x_n$.
Si $f(x_n)f(x_{n-2}) < 0$,
entonces $x_{n-1} \leftarrow x_{n-2}$y continuar con $n \leftarrow n + 1$.
Si no,
entonces continuar con $n \leftarrow n + 1$.

También se puede generalizar el resultado de convergencia descrito anteriormente.

Lema 6.2. Si denotamos el error cometido por el término n-ésimo de la sucesión generada por el algoritmo anterior como e_n, es decir, $e_n \equiv \alpha - x_n$ y sea $M = |x_1 - x_0|$, entonces,

$$\forall n \geq 1 \quad |e_{n+1}| \leq \frac{M}{2^n}.$$

Una estrategia de la cual se derivan varios métodos más sofisticados que el anterior, consiste en aproximar localmente f por una función lineal y luego aproximar α, la raíz de f, por la raíz de esta función lineal. Los tres métodos que se presentan a continuación corresponden a este esquema.

Método de la Secante.

Sean x_n y x_{n-1} dos aproximaciones de α y consideremos la recta (secante) que intersecta al gráfico de f en los dos puntos del plano $(x_n, f(x_n))$ y $(x_{n-1}, f(x_{n-1}))$, es decir,

$$L_n(x) = f(x_n) + (x - x_n)\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}. \tag{6.3}$$

La nueva aproximación de α, calculada por este método, será x_{n+1} la raíz de L_n, que despejada de la ecuación que la define, $L_n(x_{n+1}) = 0$, resulta ser

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}. \tag{6.4}$$

Método de Regula Falsi.

Este método consiste en una ligera modificación del método anterior. La fórmula de cálculo de la iteración $(n+1)$-ésima será (6.4) al igual que en el método de la Secante, pero se exigirá que los dos puntos previos x_n y x_{n-1}, sean como en el primer método presentado, es decir, que f cambie de signo entre medio. Por lo tanto en cada iteración se deberá redefinir el punto considerado como iteración anterior. El algoritmo se puede resumir como sigue.

Algoritmo 6.5. Dados x_0 y x_1 tales que $f(x_0)f(x_1) < 0$ para $n = 1, 2, ...$

calcular x_{n+1} según la fórmula (6.4) de la Secante.
Si $f(x_{n+1}) = 0$ terminar con $\alpha = x_{n+1}$.
Si $f(x_{n+1})f(x_{n-1}) < 0$, entonces
$x_n \leftarrow x_{n-1}$ y continuar con $n \leftarrow n + 1$.
Si no, entonces
continuar con $n \leftarrow n + 1$.

Método de Newton.

Sea x_n una aproximación de α y consideremos la recta **tangente** al gráfico de f en el punto $(x_n, f(x_n))$,

$$\widetilde{L}_n(x) = f(x_n) + (x - x_n)f'(x_n). \tag{6.6}$$

La iteración $(n+1)$-ésima del método de Newton se define como la raíz de esta recta, es decir

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{6.7}$$

Los tres últimos métodos, son claramente más complejos que el primero y por consiguiente esperamos una mayor velocidad de convergencia. Para poder analizar este comportamiento necesitamos una definición precisa del indicador que usaremos para medir esta velocidad.

Definición 6.8. Sea $\{x_n\}$ una sucesión convergente a α, construída según algún método iterativo. Sea $e_n = \alpha - x_n$, el error cometido por la iteración n-ésima. Se dirá que el método es de **orden p** si existe una constante c tal que

$$|e_{n+1}| \leq c|e_n|^p, \tag{6.9}$$

al menos para todo n mayor o igual que cierto umbral.

El error del método de Bisección fue acotado en (6.2) y de la definición de la iteración se tendrá que

$$e_{n+1} = \alpha - x_{n+1} = \alpha - \frac{x_n + x_{n-1}}{2} = \frac{1}{2}(e_n + e_{n-1}).$$

Si bien no se deduce de aquí una desigualdad como la de (6.9), si se puede afirmar que en promedio el método es de orden uno.

Para estudiar el error de los otros tres métodos, necesitamos poder expresar el error de la aproximación lineal de f, lo que nos obligará a aumentar las hipótesis sobre f. Como es sabido, el error de una aproximación de primer orden se expresa en términos de la derivada segunda. La relación entre el error de la iteración $(n+1)$-ésima y el(los) error(es) anterior(es) de los tres últimos métodos se establece en el lema siguiente.

Lema 6.10. Si f tiene segunda derivada continua en una vecindad de α que contenga a los puntos de las iteraciones $(n-1)$ y n y si la derivada de f no se anula en esta vecindad, entonces, para Secante y Regula Falsi se tendrá

$$e_{n+1} = -e_n e_{n-1} \frac{f''(\xi_n)}{2f'(\eta_n)}, \tag{6.11}$$

para algún $\xi_n \in \overline{co}(\alpha, x_n, x_{n-1})$ y algún $\eta_n \in \overline{co}(x_n, x_{n-1})$.
En el caso del método de Newton, bajo las mismas hipótesis, la relación será

$$e_{n+1} = -e_n^2 \frac{f''(\xi_n)}{2f'(x_n)}, \tag{6.12}$$

para algún $\xi_n \in \overline{co}(\alpha, x_n)$.

Demostración. En ambos casos el error de la iteración se derivará del error de la aproximación lineal de la función no lineal f.

Comenzaremos por el error de la aproximación lineal $\widetilde{L}_n$, usada por el método de Newton. Como ésta corresponde al desarrollo de Taylor en torno a x_n, tenemos una expresión conocida del error cometido,

$$f(x) = \widetilde{L}_n(x) + (x - x_n)^2 \frac{f''(\xi_n)}{2}, \text{ para algún } \xi_n \in \overline{co}(x, x_n),$$

en particular en $x = \alpha$,

$$0 = f(\alpha) = \widetilde{L}_n(\alpha) + (\alpha - x_n)^2 \frac{f''(\xi_n)}{2}, \text{ para algún } \xi_n \in \overline{co}(\alpha, x_n).$$

Despejando $\widetilde{L}_n(\alpha)$ y recordando la definición del error de la iteración n-ésima, se tiene

$$\widetilde{L}_n(\alpha) = -e_n^2 \frac{f''(\xi_n)}{2}, \text{ para algún } \xi_n \in \overline{co}(\alpha, x_n). \tag{6.13}$$

Por otra parte, utilizando el teorema del valor medio (T.V.M.) para la función lineal, se llega a la relación

$$\widetilde{L}_n(\alpha) - \widetilde{L}_n(x_{n+1}) = (\alpha - x_{n+1})f'(x_n). \tag{6.14}$$

Recordando que por construcción $\widetilde{L}_n(x_{n+1}) = 0$, despejando el error de la iteración $(n+1)$-ésima de la última ecuación y reemplazando $\widetilde{L}_n(\alpha)$ por la expresión dada en (6.13), se obtiene la relación pedida, (6.12).

Para probar la relación entre errores de iteraciones sucesivas de los métodos de Secante y Regula Falsi, se seguirá el mismo esquema anterior, notando que toda la diferencia entre ambas aproximaciones lineales $\widetilde{L}_n$ y L_n es que el rol que juega $f'(x_n)$ en la primera, en la segunda lo asume la diferencia dividida

$$\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} = f'(\eta_n), \text{ para algún } \eta_n \in \overline{co}(x_n, x_{n-1}),$$

(por T.V.M. aplicado a f que satisface las hipótesis).

Por otra parte, en lugar del error del polinomio de Taylor, se tendrá el error de la interpolación lineal, que como sabemos del capítulo 2, es

$$f(x) - L_n(x) = (x - x_n)(x - x_{n-1})\frac{f''(\xi_n)}{2}, \text{ para algún } \xi_n \in \overline{co}(x, x_n, x_{n-1}),$$

en particular, para $x = \alpha$, recordando que $f(\alpha) = 0$ y usando la notación del error, se tiene

$$L_n(\alpha) = -e_n e_{n-1}\frac{f''(\xi_n)}{2}, \text{ para algún } \xi_n \in \overline{co}(\alpha, x_n, x_{n-1}). \tag{6.15}$$

Aplicando el T.V.M. a la función lineal L_n, recordando que por construcción se tiene que $L_n(x_{n+1}) = 0$ y la expresión de la diferencia dividida dada antes, se llega a

$$L_n(\alpha) = (\alpha - x_{n-1})f'(\eta_n), \text{ para algún } \eta_n \in \overline{co}(x_n, x_{n-1}). \tag{6.16}$$

Combinando, como antes, las ecuaciones (6.15) y (6.16), concluye la demostración de la relación (6.11). □

Con el fin de enunciar los teoremas que establecen el **orden** de estos métodos numéricos, tal como fue definido en (6.8), debemos preocuparnos de como se satisface la hipótesis de que toda la sucesión $\{x_n\}$ generada por el método correspondiente permanezca en una vecindad de la raíz α donde no se anule la primera derivada de f y por lo tanto las relaciones establecidas en el Lema precedente sean válidas. Esta garantía aparecerá en las demostraciones de un modo más preciso que en los enunciados de los teoremas que siguen.

Teorema 6.17. *Sea α una raíz simple de $f, (f(\alpha) = 0, f'(\alpha) \neq 0)$, que será una función de segunda derivada continua en una vecindad de α. Si x_0 está suficientemente cerca de α, entonces las iteraciones del método de Newton, $x_n, \forall n \geq 0$ permanecerán en una vecindad donde se satisfacen las hipótesis del Lema anterior, convergerán a la raíz α y se cumplirá*

$$\lim_{n\to\infty} \frac{e_{n+1}}{e_n^2} = -\frac{f''(\alpha)}{2f'(\alpha)}, \tag{6.18}$$

lo que implica que el método de Newton es de orden $p = 2$.

Demostración. Sea $I = [\alpha - \varepsilon, \alpha + \varepsilon]$ un intervalo donde no se anula la primera derivada de f y se satisface la condición de continuidad de su segunda derivada. Sea

$$M = \frac{\max_{x\in I} |f''(x)|}{2\min_{x\in I} |f'(x)|}.$$

Consideremos $x_0 \in I$, es decir $|\alpha - x_0| \leq \varepsilon$, donde la *cercanía suficiente* de x_0 a α se garantice con un ε tal que

$$M|\alpha - x_0| < 1.$$

Usando (6.12) se tendrá

$$|\alpha - x_1| \leq M|\alpha - x_0|^2$$

y por lo tanto

$$M|\alpha - x_1| \leq (M|\alpha - x_0|)^2 < M|\alpha - x_0|,$$

lo que implica que

$$|\alpha - x_1| < |\alpha - x_0| \leq \varepsilon,$$

es decir,

$$x_1 \in I.$$

Además, al igual que al comienzo, x_1 está *suficientemente cerca* de $\alpha, (M|\alpha - x_1| < 1)$, como para poder repetir el argumento y probar, por inducción, que toda la sucesión generada por el método de Newton permanece en el intervalo I.

Hemos garantizado así que se satisfacen las hipótesis del Lema (6.10) y por lo tanto podemos usar la ecuación (6.12) $\forall n \geq 0$, lo que implica la desigualdad

$$|e_{n+1}| \leq M|e_n|^2, \qquad \forall n \geq 0. \tag{6.19}$$

que establece la convergencia a una velocidad caracterizada por el orden $p = 2$ del método de Newton.

(*Note que* $|e_n| \leq \frac{1}{M}(M|e_0|)^{2^n}$ *y compare con el acotamiento del error en el método de Bisección dado en* (6.2)).

Volviendo a usar la expresión (6.12) se tendrá

$$\frac{e_{n+1}}{e_n^2} = -\frac{f''(\xi_n)}{2f'(x_n)},$$

para algún $\xi_n \in \overline{co}(\alpha, x_n)$.

Pero debido a la convergencia establecida en (6.19) se tiene que $x_n \to \alpha$ y por lo tanto $\xi_n \to \alpha$, cuando $n \to \infty$, con lo cual se concluye (6.18). □

El orden de convergencia del método de la Secante se establece en el teorema que sigue

Teorema 6.20. *Bajo las mismas hipótesis del teorema* (6.17)*, si* x_0 *y* x_1 *están suficientemente cerca de* α*, entonces el método de la Secante converge con un orden* $p = \frac{1+\sqrt{5}}{2}$.

Demostración. Sea $I = [\alpha - \varepsilon, \alpha + \varepsilon]$ un intervalo donde no se anule la primera derivada de f y se satisfaga la condición de continuidad de la segunda derivada.

Sea M la misma constante del teorema (6.17) y sean $x_0, x_1 \in I$.

Como se satisfacen las hipótesis del lema (6.10) para $n = 1$, podemos usar la expresión (6.11)

$$e_2 = -e_1 e_0 \frac{f''(\xi_1)}{2f'(\eta_1)}, \text{ para algún } \xi_1 \in \overline{co}(\alpha, x_0, x_1) \text{ y algún } \eta_1 \in \overline{co}(x_0, x_1)$$

y por lo tanto

$$|e_2| \leq |e_1||e_0|M,$$

lo que equivale a

$$M|e_2| \leq M|e_1|M|e_0|.$$

Si x_0 y x_1 están tan cerca de α (ε es tan chico) como para que $\delta = \max\{M|e_0|, M|e_1|\} < 1$, entonces

$$M|e_2| \leq \delta^2 < \delta$$

y como $|e_0| \leq \frac{\delta}{M} \leq \varepsilon$, se tendrá que $x_2 \in I$ y que $\max\{M|e_1|, M|e_2|\} < 1$.

Esto permite repetir el mismo argumento y demostrar por inducción que toda la sucesión generada por el método de la Secante pertenece al intervalo I y por lo tanto siempre se podrá utilizar la expresión recursiva del error dada en (6.11). Acotando esta expresión y multiplicando por M, se tendrá

$$M|e_3| \leq M|e_2|M|e_1| \leq \delta^3,$$
$$M|e_4| \leq M|e_3|M|e_2| \leq \delta^5$$

y en general

$$M|e_{n+1}| \leq \delta^{q_n}\delta^{q_{n-1}} = \delta^{q_{n+1}}.$$

La convergencia queda así establecida con

$$|e_n| \leq \frac{1}{M}\delta^{q_n}. \tag{6.21}$$

La sucesión de exponentes de la cota satisface la recurrencia

$$q_0 = q_1 = 1, \quad q_{n+1} = q_n + q_{n-1}, \quad \forall n \geq 1,$$

es decir, se trata de una sucesión de Fibonacci, de la cual se sabe que

$$q_n = \frac{1}{\sqrt{5}}(\lambda_1^{n+1} - \lambda_2^{n+1}),$$

donde los valores propios de la matriz que caracteriza la recurrencia son

$$\lambda_1 = \frac{1+\sqrt{5}}{2} \text{ y } \lambda_2 = \frac{1-\sqrt{5}}{2}.$$

Para establecer el orden de convergencia del método de la Secante utilizaremos la desigualdad (6.21) que se refiere a dos sucesiones de números positivos que convergen a 0:

$$\{|e_n|\} \text{ y } \{B_n\}, \text{ con } B_n = \frac{1}{M}\delta^{q_n}.$$

Si probamos que la sucesión $\{B_n\}$ converge con una cierta velocidad, entonces debido a la desigualdad (6.21) podremos afirmar que la sucesión de nuestro interés converge al menos a esa misma velocidad. Siguiendo esta estrategia calculemos

$$\frac{B_{n+1}}{B_n^{\lambda_1}} = M^{\lambda_1 - 1}\delta^{q_{n+1} - q_n\lambda_1}. \tag{6.22}$$

Pero para el exponente de δ se tiene

$$q_{n+1} - q_n\lambda_1 = \frac{1}{\sqrt{5}}(\lambda_1^{n+2} - \lambda_2^{n+2}) - \frac{\lambda_1}{\sqrt{5}}(\lambda_1^{n+1} - \lambda_2^{n+1}) = \lambda_2^{n+1} \approx (-0.618)^{n+1},$$

lo que permite acotar (6.22) con una constante, $\forall n \geq 1$

$$\frac{B_{n+1}}{B_n^{\lambda_1}} \leq M^{\lambda_1 - 1}\delta^{\frac{-1}{2}}$$

y probar así que el orden de convergencia de B_n es λ_1, y por consiguiente que el método de la Secante converge al menos con ese mismo orden, como se sostiene en el teorema. □

El orden de convergencia mide la velocidad de ésta en cuanto al número de iteraciones requeridas para alcanzar cierta precisión dada. Para comparar las velocidades de dos métodos en tiempo real, habría que considerar adicionalmente el costo de cada iteración. Si bien Newton es un método más eficiente que Secante en cuanto a número de iteraciones, puede llegar a ser peor que Secante debido a que en cada iteración debe realizar dos evaluaciones: $f(x_n)$ y $f'(x_n)$. En cambio Secante requiere solo de una evaluación, $f(x_n)$, pues se puede guardar de la iteración anterior el valor de f en x_{n-1}. Si denotamos por m el tiempo requerido para evaluar f y consideramos que el tiempo requerido para evaluar f' es $s \cdot m$. Se puede probar que si $s > 0.44$, entonces el método de la Secante es más eficiente que el método de Newton, en términos reales.

Ejemplo 1. Para ilustrar el comportamiento de los métodos estudiados veremos su aplicación al sencillo problema de calcular $\sqrt{2} = 1.41421356\ldots$, es decir, la solución positiva de

$$x^2 - 2 = 0.$$

Las primeras iteraciones se registran en las figuras 6.2, 6.3 y 6.4.

k	x_k	$f(x_k)$
0	1	-1
1	2	2
2	1.5	0.25
3	1.25	-0.4375
4	1.375	-0.109375
5	1.4375	0.06640625

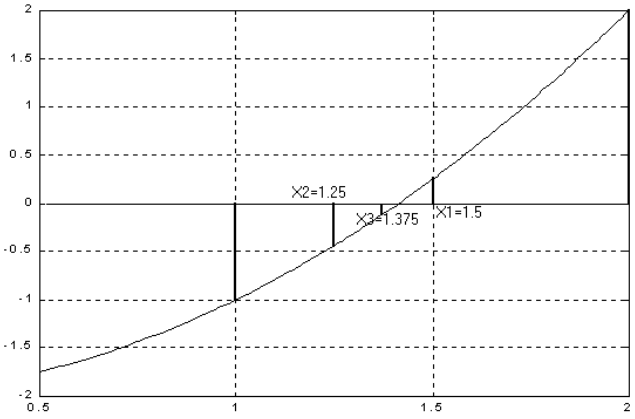


Figura 6.2: Iteraciones del Método de Bisección para $f(x) = x^2 - 2$

k	x_k	$f(x_k)$
0	1	-1
1	2	2
2	$\frac{4}{3}$	$-\frac{2}{9}$
3	1.4	-0.04
4	1.41463414...	-0.00118976...

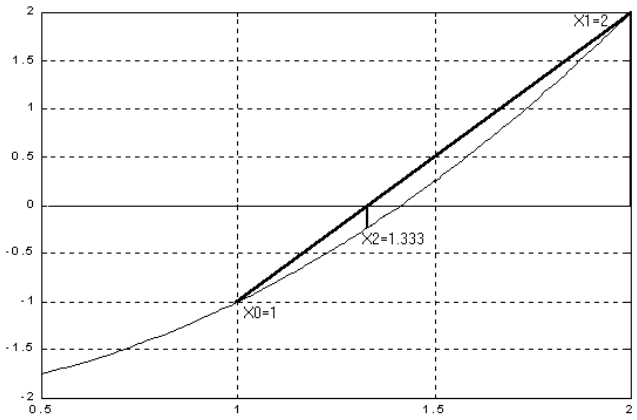


Figura 6.3: Iteraciones del Método de la Secante para $f(x) = x^2 - 2$

k	x_k	$f(x_k)$
0	1	-1
1	1.5	0.25
2	$\frac{17}{12}$	0.00694444...
3	1.41421568...	0.00000600...

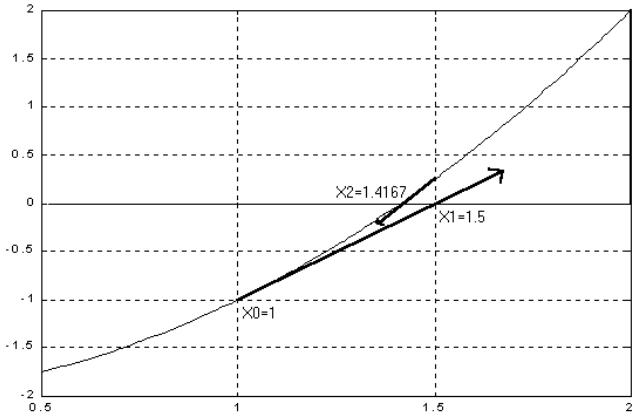


Figura 6.4: Iteraciones del Método de Newton para $f(x) = x^2 - 2$

MÉTODOS DE UN PUNTO

El método de Newton genera una sucesión de aproximaciones, de manera iterativa, donde el término $(n+1)$-ésimo depende solo de un término anterior, lo que permite escribir una iteración como

$$x_{n+1} = g(x_n). \tag{6.23}$$

En este caso la **función de iteración** g será $g(x) = x - \frac{f(x)}{f'(x)}$.

Pero las sucesiones definidas por recurrencia como en (6.23) $\forall n \geq 0$ y con x_0 dado, son bien conocidas y se sabe que cuando convergen, lo hacen a un punto fijo de la función de iteración, es decir a un α, tal que

$$g(\alpha) = \alpha.$$

Podemos pensar entonces en una clase general de métodos de este tipo, es decir, tales que los problemas

$$f(x) = 0 \text{ y } g(x) = x$$

sean equivalentes y que la sucesión definida por

$$\begin{aligned} x_0 &\text{ dado} \\ x_{n+1} &= g(x_n) \quad \forall n \geq 0 \end{aligned}$$

converja.

Normalmente hay muchas posibilidades para transformar un problema de obtención de raíz en un problema de punto fijo equivalente, pero sólo nos interesan aquellas donde la convergencia de la sucesión correspondiente esté garantizada. Para ello contamos con el conocido Teorema de Punto Fijo, que aquí recordamos.

Teorema 6.24. *Sea $g : [a,b] \to [a,b]$, una función contractante (es decir, existe una constante positiva $L < 1$ tal que $\forall x, y \in [a,b] \quad |g(x) - g(y)| \leq L|x - y|$), entonces existe solo un punto fijo de $g, \alpha \in [a,b]$ y las iteraciones $x_{n+1} = g(x_n)$ convergen a α, cualquiera sea el punto de partida $x_0 \in [a,b]$.*

Suponiendo mayor regularidad de la función g se obtiene el resultado que sigue y que permite analizar la evolución del error.

Teorema 6.25. *Si la función $g : [a,b] \to [a,b]$ es continuamente derivable y si $\max\limits_{x \in [a,b]} |g'(x)| = L < 1$, entonces g es contractante en $[a,b]$, con lo cual se tendrá lo establecido en el teorema anterior y además*

$$|\alpha - x_n| \leq L^n |\alpha - x_0| \quad \forall n \geq 0, \tag{6.26}$$

$$\lim_{n \to \infty} \frac{\alpha - x_{n+1}}{\alpha - x_n} = g'(\alpha), \tag{6.27}$$

Demostración. La primera afirmación es un resultado directo del T.V.M. En efecto, como g es continuamente derivable

$$\forall x, y \in [a,b] \quad g(x) - g(y) = g'(\xi)(x - y) \text{ para algún } \xi \in \overline{co}(x,y),$$

lo que implica que

$$\forall x, y \in [a,b] \quad |g(x) - g(y)| \leq L|x - y|, \text{ con } L < 1.$$

En particular, para α y x_{n-1}, se tiene

$$|\alpha - x_n| = |g(\alpha) - g(x_{n-1})| \leq L|\alpha - x_{n-1}|$$

y repitiendo el mismo argumento resulta que

$$|\alpha - x_n| \leq L^n|\alpha - x_0|.$$

Para demostrar la última afirmación utilizaremos nuevamente el T.V.M.

$$\alpha - x_{n+1} = g(\alpha) - g(x_n) = g'(\xi_n)(\alpha - x_n), \text{ para algún } \xi_n \in \overline{co}(\alpha, x_n)$$

y por lo tanto

$$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{\alpha - x_n} = \lim_{n\to\infty} g'(\xi_n) = g'(\alpha),$$

pues $x_n \to \alpha$ y por lo tanto $\xi_n \to \alpha$, cuando $n \to \infty$. □

NOTA.

Si α es un punto fijo de g, es decir, α es solución de la ecuación $x = g(x)$, si g es continuamente derivable en una vecindad de α y si $|g'(\alpha)| < 1$, entonces los resultados de convergencia (6.26) y (6.27) seguirán siendo válidos, siempre que x_0 esté suficientemente cerca de α.

La expresión (6.27) caracteriza el orden de convergencia $p = 1$. Para el método de Newton, del cual sabemos que su orden de convergencia es 2 se tiene que

$$g'(\alpha) = 1 - \frac{(f'(\alpha))^2 - f(\alpha)f''(\alpha)}{(f'(\alpha))^2} = 0.$$

Cabe preguntarse entonces ¿cómo se caracteriza en términos de las derivadas de la función de iteración los métodos de un punto de orden superior a uno?. El teorema que sigue responde a esta pregunta.

Teorema 6.28. *Sea α punto fijo de g, y sea g una función p veces continuamente derivable en una vecindad de α, para algún $p \geq 2$. Si*

(6.29) $$g'(\alpha) = ... = g^{(p-1)}(\alpha) = 0,$$

entonces, si x_0 está suficientemente cerca de α, las iteraciones

$$x_{n+1} = g(x_n) \quad \forall n \geq 0$$

convergen con un orden p y

(6.30) $$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{(\alpha - x_n)^p} = (-1)^{p-1} \frac{g^{(p)}(\alpha)}{p!}.$$

Demostración. Desarrollando $g(x_n)$ en serie de Taylor en torno a α

$$x_{n+1} = g(x_n) = g(\alpha) + (x_n - \alpha)g'(\alpha) + ... + (x_n - \alpha)^{p-1}\frac{g^{(p-1)}(\alpha)}{(p-1)!} + (x_n - \alpha)^p \frac{g^{(p)}(\xi_n)}{p!}$$

para algún $\xi_n \in \overline{co}(\alpha, x_n)$.

Utilizando la hipótesis (6.29) y $\alpha = g(\alpha)$, se tendrá

$$\alpha - x_{n+1} = -(x_n - \alpha)^p \frac{g^{(p)}(\xi_n)}{p!}.$$

Usando el teorema (6.25) y la nota anterior se prueba la convergencia y (6.30), lo que completa la demostración. □

Volviendo al método de Newton, recordaremos que una de las condiciones que nos permitió expresar el error de una iteración y luego estudiar su convergencia, fue $f'(\alpha) \neq 0$.

Los dos últimos teoremas nos permiten estudiar la convergencia del método de Newton en ausencia de esta condición, es decir, cuando α es una **raíz múltiple** de f.

Sea α una raíz de orden $m > 1$ de f. Es decir, existe una función h, que supondremos dos veces derivable en una vecindad de α y tal que

$$f(x) = (x - \alpha)^m h(x) \text{ con } h(\alpha) \neq 0.$$

La función de iteración del método de Newton en este caso será

$$\begin{aligned} g(x) &= x - \frac{(x-\alpha)^m h(x)}{m(x-\alpha)^{m-1}h(x) + (x-\alpha)^m h'(x)} \\ &= x - \frac{(x-\alpha)h(x)}{mh(x) + (x-\alpha)h'(x)}. \end{aligned}$$

Calcularemos $g'(\alpha)$ para revisar la condición de convergencia y el orden de ella, según los teoremas precedentes. Para simplificar los cálculos denotemos por

$$z(x) = mh(x) + (x - \alpha)h'(x).$$

Con esta notación se tendrá

$$g'(x) = 1 - \frac{[h(x) + (x-\alpha)h'(x)]z(x) - (x-\alpha)h(x)z'(x)}{[z(x)]^2}$$

y en $x = \alpha$

$$\begin{aligned} g'(\alpha) &= 1 - \frac{h(\alpha)z(\alpha)}{[z(\alpha)]^2} = 1 - \frac{m[h(\alpha)]^2}{[mh(\alpha)]^2} \\ &= 1 - \frac{1}{m} < 1, \end{aligned} \tag{6.31}$$

lo que prueba la convergencia del método de Newton, según el teorema (6.25) y la nota que le sigue.

Por consiguiente, si la multiplicidad de la raíz es mayor que uno, el método habrá perdido el orden $p = 2$ de su velocidad de convergencia, pues $g'(\alpha) \neq 0$. Pero en el mismo desarrollo anterior se puede observar que con una leve modificación de la función de iteración (y si la función h se puede derivar por tercera vez) se recupera la velocidad de convergencia característica de Newton, es decir, el orden 2.

Método de Newton modificado.

Si α es una raíz de orden $m > 1$ de f, entonces el método de Newton modificado, definido por las iteraciones

$$x_{n+1} = x_n - m\frac{f(x_n)}{f'(x_n)}, \quad \forall n \geq 0,$$

converge con un orden $p = 2$, si la función de iteración es dos veces continuamente derivable en una vecindad de α y si x_0 se escoge suficientemente cerca de α.

NOTA.

El orden de multiplicidad de una raíz se puede *estimar* estudiando la evolución de los cuocientes R_n, definidos por

$$R_{n+1} = \frac{x_{n+1} - x_n}{x_n - x_{n-1}} = g'(\xi_n),$$

para algún $\xi_n \in \overline{co}(x_n, x_{n-1})$, por T.V.M.

Pero si el método de un punto es convergente

$$\lim_{n\to\infty} R_n = g'(\alpha).$$

De modo que en el caso del método de Newton usual y α raíz de multiplicidad m, se tendrá por (6.31)

$$\lim_{n\to\infty} R_n = 1 - \frac{1}{m}.$$

Por ejemplo, si se aplica el método de Newton usual partiendo de $x_0 = 1$ a la función

$$f(x) = x^3 - 5.56x^2 + 9.1389x - 4.68999,$$

se encontrará que las iteraciones convergen a $\alpha = 1.23$ del modo que se ilustra en la tabla 6.1. Se aprecia

n	x_n	$f(x_n)$	$x_n - x_{n-1}$	R_n
1	1.10903	- 0.0291	0.109	
2	1.16773	- 0.00749	0.0587	0.539
3	1.19837	- 0.00190	0.0306	0.513
4	1.21406	- 0.000479	0.0157	0.506
5	1.22199	- 0.000120	0.00794	0.506
6	1.22599	- 0.0000302	0.00399	0.501
7	1.22799	- 0.00000755	0.00200	0.500
8	1.22900	- 0.00000189	0.00100	0.500

Tabla 6.1: Método de Newton para $f(x) = x^3 - 5.56x^2 + 9.1389x - 4.68999$, con $x_0 = 1$

claramente que $R_n \to 0.5 = 1 - \frac{1}{m}$ si $m = 2$, lo que corresponde a la multiplicidad de esta raíz. Se debe advertir que el cuociente R_n puede llegar a ser muy inestable para valores grandes de n, es decir, cuando las aproximaciones de la raíz, x_{n-1}, x_n y x_{n+1} sean muy parecidas entre sí.

Criterio de parada.

Los métodos de un punto que estamos analizando aquí se basan en el teorema de punto fijo, al igual que los métodos iterativos para sistemas lineales estudiados en el capítulo 4. Cabe esperar entonces que exista una relación entre el error de la iteración n-ésima y la distancia entre las dos últimas iteraciones, similar a la establecida en (4.40).

Sea g la función de iteración de un método de un punto y $L < 1$ su constante de Lipschitz, de modo que se cumple

$$|\alpha - x_{n+1}| \leq L|\alpha - x_n|.$$

Por lo tanto

$$|x_{n+1} - x_n| = |\alpha - x_n - (\alpha - x_{n+1})| \geq |\alpha - x_n| - |\alpha - x_{n+1}|$$
$$\geq |\alpha - x_n| - L|\alpha - x_n| = (1 - L)|\alpha - x_n|.$$

Con lo cual se tendrá

$$|\alpha - x_n| \leq \frac{1}{1 - L}|x_{n+1} - x_n|,$$

lo que implica

$$|\alpha - x_n| \leq \frac{L}{1 - L}|x_n - x_{n-1}|. \tag{6.32}$$

Una práctica frecuente para detener las iteraciones será la escasa diferencia entre una iteración y otra. Es decir, dada una tolerancia $\varepsilon > 0$, si $|x_n - x_{n-1}| \leq \varepsilon$, finalizar con $\alpha = x_n$.

La desigualdad (6.32) garantiza que este criterio sea razonable y lo será tanto más, mientras más pequeña sea la constante L que domina la convergencia según se establece en (6.26).

En el ejemplo de la tabla (6.1) compare la sucesión de valores de la cuarta columna con el error correspondiente: $e_n = 1.23 - x_n$ y note el parecido.

Por otra parte de la desigualdad (6.32) se deduce fácilmente la relación

$$|\alpha - x_n| \leq \frac{L^n}{1 - L}|x_1 - x_0|, \tag{6.33}$$

cuya similitud con (4.41) es evidente y que por tanto tendrá la misma utilidad práctica comentada en ese punto.

EXTRAPOLACIÓN DE AITKEN PARA SUCESIONES QUE CONVERGEN LINEALMENTE

Si $\{x_n\}$ es una sucesión que converge a α con un orden $p = 1$, entonces mediante una combinación adecuada de varias iteraciones consecutivas se puede acelerar la convergencia. Este tipo de procedimientos, (cuyo beneficio comprobamos en el capítulo 3 con el método de Romberg) se conocen con el nombre de **métodos de extrapolación**.

En el caso actual, la convergencia lineal significa que para n suficientemente grande

$$\frac{\alpha - x_{n+1}}{\alpha - x_n} \approx \text{ constante}$$

y por lo tanto

$$\frac{\alpha - x_{n+1}}{\alpha - x_n} \approx \frac{\alpha - x_{n+2}}{\alpha - x_{n+1}},$$

lo que equivale a

$$(\alpha - x_{n+1})^2 \approx (\alpha - x_{n+2})(\alpha - x_n). \tag{6.34}$$

Con tres iteraciones consecutivas x_n, x_{n+1} y x_{n+2}, el método de Aitken construye una nueva aproximación, $\tilde{x}_{n+2}$, como una mejora de la última iteración, x_{n+2}, definida por

$$\tilde{x}_{n+2} = x_{n+2} - \frac{(x_{n+2} - x_{n+1})^2}{(x_n - x_{n+1}) - (x_{n+1} - x_{n+2})}. \tag{6.35}$$

Para apreciar el beneficio de esta construcción debemos comparar $|\alpha - \tilde{x}_{n+2}|$ con $|\alpha - x_{n+2}|$. Con este fin obtendremos una expresión conveniente para $(x_{n+2} - x_{n+1})^2$.

$$\begin{aligned}(x_{n+2} - x_{n+1})^2 &= (x_{n+2} - \alpha + \alpha - x_{n+1})^2 \\ &= (\alpha - x_{n+2})^2 - 2(\alpha - x_{n+2})(\alpha - x_{n+1}) + (\alpha - x_{n+1})^2 \\ &= (\alpha - x_{n+2}) \left\{ (\alpha - x_{n+2}) - 2(\alpha - x_{n+1}) + \frac{(\alpha - x_{n+1})^2}{\alpha - x_{n+2}} \right\}.\end{aligned} \tag{6.36}$$

Reemplazando esta expresión en (6.35) se obtiene

$$\alpha - \tilde{x}_{n+2} = (\alpha - x_{n+2})\varepsilon_n, \tag{6.37}$$

con $\varepsilon_n = 1 + \frac{1}{(x_n - x_{n+1}) - (x_{n+1} - x_{n+2})} \left\{ \alpha - x_{n+2}) - 2(\alpha - x_{n+1}) + \frac{(\alpha - x_{n+1})^2}{\alpha - x_{n+2}} \right\}$.

De este modo bastará probar que $|\varepsilon_n| < 1$ para poder concluir de (6.37) que $\tilde{x}_{n+2}$ comete menos error que x_{n+2}. Pero de (6.34) tenemos que

$$\varepsilon_n \approx 1 + \frac{1}{(x_n - x_{n+1}) - (x_{n+1} - x_{n+2})} \{(\alpha - x_{n+2}) - 2(\alpha - x_{n+1}) + (\alpha - x_n)\} = 0.$$

Es decir, en la medida que la aproximación en (6.34) sea más cercana a la igualdad, mayor será el beneficio del método de extrapolación de Aitken.

Como ejemplo de la mejoría producida por el procedimiento de extrapolación propuesto veremos el comportamiento de la sucesión siguiente

$$x_{n+1} = 1.6 + 0.99cos(x_n), \quad \forall n \geq 0 \text{ y } x_0 = \pi/2$$

que converge al punto fijo

$$\alpha = 1.585471802.$$

En la tabla 6.2 se aprecia, por una parte, la lenta convergencia de estas iteraciones y por otra, el aumento de la velocidad de convergencia gracias a la extrapolación.

n	x_n	$x_n - x_{n-1}$	$\tilde{x}_n$	$\tilde{x}_n - \tilde{x}_{n-1}$
1	1.60000	0.0292	1.59999636	0.0292
2	1.57109	- 0.0289	1.57109607	- 0.0289
3	1.59971	0.0286	1.58547286	0.0144
4	1.57138	- 0.0283	1.58547075	- 0.00000211
5	1.59942	0.0280	1.58547284	0.00000209
6	1.57167	- 0.0278	1.58547180	- 0.00000104

Tabla 6.2: Comportamiento de $x_{n+1} = 1.6 + 0.99cos(x_n) \quad \forall n \geq 0$ y $x_0 = \pi/2$

SISTEMAS DE ECUACIONES NO LINEALES

El problema que abordaremos aquí es el de resolver la ecuación

$$f(x) = 0$$

con f una función no lineal de $\mathbb{R}^n$ en $\mathbb{R}^n$, es decir, se trata de calcular un vector $\alpha \in \mathbb{R}^n$ que satisfaga un sistema de ecuaciones no lineales

$$\begin{aligned} f_1(x_1, x_2, ..., x_n) &= 0, \\ f_2(x_1, x_2, ..., x_n) &= 0, \\ \vdots \qquad & \qquad \vdots \\ f_n(x_1, x_2, ..., x_n) &= 0, \end{aligned}$$

donde cada función f_i, para $i = 1, ..., n$ va de $\mathbb{R}^n$ en $\mathbb{R}$.

La generalización del método de Newton a este caso es evidente, al menos desde el punto de vista formal.

La aproximación lineal de f (que ocupará el lugar de la recta definida en (6.6)), en torno a un punto calculado previamente $x^{(k)} \in \mathbb{R}^n$, será

$$L_k(x) = f(x^{(k)}) + J(x^{(k)}) \cdot (x - x^{(k)}), \tag{6.38}$$

donde $J(x^{(k)}) = \left(\frac{\partial f_i}{\partial x_j}(x^{(k)}\right)_{i,j}$, representa a la matriz Jacobiana de f evaluada en $x^{(k)}$.

Como antes, la iteración de Newton se definirá como un vector $x^{(k+1)} \in \mathbb{R}^n$ que sea raíz de la aproximación lineal, es decir,

$$L_k(x^{(k+1)}) = 0, \tag{6.39}$$

lo que equivale a pedir que el vector diferencia $\Delta^{(k)} = x^{(k+1)} - x^{(k)}$ resuelva el sistema lineal

$$J(x^{(k)}) \cdot \Delta^{(k)} = -f(x^{(k)}). \tag{6.40}$$

Calcular la iteración de Newton de este modo es más eficiente que invertir la matriz Jacobiana y despejar $x^{(k+1)}$ de la ecuación (6.39).

Si f es continuamente diferenciable, si su matriz Jacobiana es invertible en una vecindad de la raíz $\alpha \in \mathbb{R}^n$ y si la adivinanza inicial $x^0 \in \mathbb{R}^n$ se escoge suficientemente cerca de α, entonces, al igual que en el caso uni-dimensional, el método de Newton convergerá con velocidad cuadrática a α.

El problema del costo de cada iteración de Newton, ya mencionado antes, puede volverse dramático en el caso actual. Por una parte habrá que resolver un sistema lineal diferente en cada iteración y por otra, se necesitarán las evaluaciones de muchas funciones:

$$f_i, i = 1, ..., n, \qquad \frac{\partial f_i}{\partial x_j}, \quad i, j = 1, ..., n.$$

Una alternativa usualmente aceptada para reducir estos costos es **no** recalcular la matriz Jacobiana en todas las iteraciones. De este modo todos los sistemas lineales a resolver en las iteraciones donde se ha mantenido constante la matriz son de bajo costo pues el reajuste de la matriz (equivalentemente, su factorización LU) se hace solo una vez y sirve para todos ellos. Además se tendrá un ahorro considerable en el número de evaluaciones a realizar al evitar las n^2 derivadas parciales.

Para clarificar esta estrategia, supongamos que se decide calcular la matriz Jacobiana solo cada tres iteraciones, es decir, solo cada tres iteraciones se realizará una iteración auténtica del método de Newton, tal como se define en (6.40).

Los cálculos a realizar en las primeras cuatro iteraciones serían:

- $f(x^{(0)})$; $J(x^{(0)})$; factorización LU de la matriz Jacobiana $L_0U_0 = J(x^{(0)})$; $\Delta^{(0)}$ solución del sistema lineal $J(x^{(0)})\Delta^{(0)} = -f(x^{(0)})$; $x^{(1)} = x^{(0)} + \Delta^{(0)}$.

- $f(x^{(1)})$; $\Delta^{(1)}$ solución del sistema lineal de matriz con factorización LU ya calculada $J(x^{(0)})\Delta^{(1)} = -f(x^{(1)})$; $x^{(2)} = x^{(1)} + \Delta^{(1)}$.
- $f(x^{(2)})$; $\Delta^{(2)}$ solución del sistema lineal de matriz con factorización LU ya calculada $J(x^{(0)})\Delta^{(2)} = -f(x^{(2)})$; $x^{(3)} = x^{(2)} + \Delta^{(2)}$.
- $f(x^{(3)})$; $J(x^{(3)})$; factorización LU de la matriz Jacobiana $L_3U_3 = J(x^{(3)})$; $\Delta^{(3)}$ solución del sistema lineal $J(x^{(3)})\Delta^{(3)} = -f(x^{(3)})$; $x^{(4)} = x^{(3)} + \Delta^{(3)}$.

RAÍCES DE POLINOMIOS

Los polinomios son funciones no lineales suficientemente particulares como para merecer un párrafo aparte. Supondremos aquí que la expresión conocida del polinomio es la canónica

$$p(x) = a_0 + a_1x + ... + a_{n-2}x^{n-2} + a_{n-1}x^{n-1} + a_nx^n. \tag{6.41}$$

(*Otras expresiones interesantes son aquellas que resultan al escribir el polinomio característico de una matriz tridiagonal simétrica, según vimos en el capítulo* 5.)

Como hemos insistido previamente, la velocidad de convergencia del método de Newton puede no ser tal, en términos reales, debido al problema del costo en evaluaciones de cada iteración. para evaluar eficientemente un polinomio y su derivada se cuenta con el **Método de Horner**, que pasamos a describir.

El polinomio dado en (6.41) se puede escribir también de la forma

$$p(x) = a_0 + x(a_1 + x(a_2 + ... + x(a_{n-2} + x(a_{n-1} + xa_n))...)). \tag{6.42}$$

Esta última expresión permite evaluar el polinomio en un z dado, de manera recursiva, aprovechando el *anidamiento*, según el procedimiento

$$\begin{aligned} b_n &= a_n \\ b_i &= a_i + zb_{i+1} \text{ para } i = n-1, n-2, ..., 0. \end{aligned} \tag{6.43}$$

De este modo el valor del polinomio en z será

$$p(z) = b_0.$$

El ahorro en el número de productos es considerable. Al calcular $p(z)$ reemplazando x por el valor z en la expresión (6.41) se realizan $(2n-1)$ productos, en cambio con el procedimiento (6.43) solo se realizan n de estas operaciones. Si el objetivo fuera conocer solo el valor de p en z, entonces los valores $b_n, b_{n-1}, ..., b_1$ podrían ser olvidados una vez utilizados. Pero si la razón de estos cálculos es realizar una iteración del método de Newton (z *será en ese caso la iteración anterior* x_k) entonces tendrán una segunda utilidad, pues se relacionan con la derivada del polinomio, como se muestra a continuación. Sea

$$q(x) = b_1 + b_2x + ... + b_nx^{n-1}.$$

Se comprueba fácilmente, identificando coeficientes y usando la definición de los coeficientes b_i, que $b_0 + (x-z)q(x) = p(x)$, lo que implica que $p'(x) = q(x) + (x-z)q'(x)$ y por lo tanto

$$p'(z) = q(z).$$

De modo que la evaluación de la derivada del polinomio p se realizará según el mismo método de Horner, es decir, siguiendo la recurrencia (6.43) para calcular ahora c_i a partir de los coeficientes $b_i, i = n, n-1, ..., 1$

$$c_n = b_n$$

$$c_i = b_i + zc_{i+1}, \text{ para } i = n-1, ..., 1.$$

El único valor que interesa es $c_1 = p'(z)$ y por lo tanto los demás términos de la recurrencia pueden ser olvidados una vez utilizados.

En síntesis una iteración de Newton se realizará en este caso como sigue.

$$\begin{aligned}
&\text{Sea } z = x_k \\
&\qquad b_n = a_n, \quad c = a_n, \\
&\qquad \text{para } i = n-1, ..., 1, \\
&\qquad\quad b_i = a_i + zb_{i+1}, \quad c = b_i + zc, \\
&\qquad b_0 = a_0 + zb_1, \\
&\qquad x_{k+1} = x_k - \frac{b_0}{c}.
\end{aligned} \tag{6.44}$$

Método de Bairstow.

Un polinomio a coeficientes reales puede tener raíces complejas y reales pero siempre será factorizable en factores polinomiales *reales* de grado menor o igual que dos. El método que presentamos aquí se ocupa de encontrar un factor cuadrático y por lo tanto, usándolo repetidas veces, puede entregarnos todas las raíces, reales y complejas, de un polinomio a coeficientes reales.

Abordaremos el problema de encontrar un factor cuadrático del tipo

$$x^2 - rx - q \tag{6.45}$$

del polinomio (de grado mayor o igual que 3) $p(x)$. Es decir, buscamos los coeficientes r y q, tales que la división de $p(x)$ por $(x^2 - rx - q)$ arroje resto nulo. Dados r y q, en general, la división producirá un polinomio cuociente $p_1(x)$, cuyos coeficientes dependerán de r y q, y un resto lineal, cuyos coeficientes también dependerán de r y q,

$$p(x) = p_1(x)(x^2 - rx - q) + Ax + B. \tag{6.46}$$

Así, los coeficientes r y q buscados, son aquellos que resuelvan el sistema no lineal de ecuaciones

$$\begin{aligned}
A(r, q) &= 0, \\
B(r, q) &= 0.
\end{aligned} \tag{6.47}$$

A pesar de no contar con una expresión analítica de estas funciones no lineales veremos que podemos establecer las relaciones que nos permitan calcular la matriz Jacobiana **evaluada en un vector de coeficientes** (r, q) **dado** y de este modo utilizar el método de Newton para sistemas no lineales. Es decir, el vector de coeficientes (r, q) buscado, se aproximará mediante una sucesión (r_k, q_k), calculada según el método de Newton

$$\begin{bmatrix} \frac{\partial A}{\partial r}(r_k, q_k) & \frac{\partial A}{\partial q}(r_k, q_k) \\ \frac{\partial B}{\partial r}(r_k, q_k) & \frac{\partial B}{\partial q}(r_k, q_k) \end{bmatrix} \begin{bmatrix} r_{k+1} - r_k \\ q_{k+1} - q_k \end{bmatrix} = - \begin{bmatrix} A(r_k, q_k) \\ B(r_k, q_k) \end{bmatrix}. \tag{6.48}$$

Con el fin de establecer las relaciones que permitan el cálculo de la matriz Jacobiana anterior de manera sencilla, simplificaremos la notación y evitaremos los subíndices relativos a la iteración. Sean

$$A_r = \frac{\partial A}{\partial r}, A_q = \frac{\partial A}{\partial q}, B_r = \frac{\partial B}{\partial r}, B_q = \frac{\partial B}{\partial q}.$$

Derivando parcialmente (6.46) con respecto a r y q se obtiene

$$\begin{aligned} 0 \equiv \frac{\partial p}{\partial r}(x) &= -xp_1(x) + (x^2 - rx - q)\frac{\partial p_1}{\partial r}(x) + xA_r + B_r \\ 0 \equiv \frac{\partial p}{\partial q}(x) &= -p_1(x) + (x^2 - rx - q)\frac{\partial p_1}{\partial q}(x) + xA_q + B_q. \end{aligned} \tag{6.49}$$

Consideremos una nueva división por el mismo factor cuadrático,

$$p_1(x) = p_2(x)(x^2 - rx - q) + A_1 x + B_1. \tag{6.50}$$

Si x_0 y x_1 son dos raíces distintas del factor cuadrático $(x^2 - rx - q)$, entonces $p_1(x_i) = A_1 x_i + B_1$, para $i = 0, 1$, y evaluando (6.49) en x_i para $i = 0, 1$, se llega a

$$-x_i(A_1 x_i + B_1) + A_r x_i + B_r = 0, \qquad \text{para } i = 0, 1, \tag{6.51}$$

$$-A_1 x_i - B_1 + A_q x_i + B_q = 0, \qquad \text{para } i = 0, 1. \tag{6.52}$$

De esta última expresión se obtiene directamente que

$$A_q = A_1 \text{ y } B_q = B_1.$$

Como $x_i^2 = rx_i + q$, en (6.51) se tendrá $x_i(A_r - rA_1 - B_1) - qA_1 + B_r = 0$, para $i = 0, 1$, lo que implica que

$$A_r = rA_1 + B_1 \quad \text{y} \quad B_r = qA_1.$$

En resumen, dado el vector de coeficientes (r_k, q_k), bastará dividir el polinomio $p(x)$ dos veces por el factor cuadrático $(x^2 - r_k x - q_k)$ para obtener

$$A(r_k, q_k), B(r_k, q_k), A_1(r_k, q_k), B_1(r_k, q_k)$$

y reemplazando las derivadas parciales como sigue

$$\begin{aligned} &\frac{\partial A}{\partial r}(r_k, q_k) = r_k A_1(r_k, q_k) + B_1(r_k, q_k), \frac{\partial A}{\partial q}(r_k, q_k) = A_1(r_k, q_k), \\ &\frac{\partial B}{\partial r}(r_k, q_k) = q_k A_1(r_k, q_k), \qquad \frac{\partial B}{\partial q}(r_k, q_k) = B_1(r_k, q_k), \end{aligned}$$

se obtiene una nueva aproximación de (r, q) mediante la iteración de Newton (6.48).

Para dividir polinomios eficientemente usamos un algoritmo de tipo Horner. Sean

$$p(x) = a_0 x^n + a_1 x^{n-1} + ... + a_{n-1} x + a_n = p_1(x)(x^2 - rx - q) + Ax + B,$$

$$p_1(x) = b_0 x^{n-2} + ... + b_{n-2}.$$

Identificando coeficientes se obtiene

$$\begin{aligned} b_0 &= a_0, \\ b_1 &= a_1 + rb_0, \\ b_i &= a_i + rb_{i-1} + qb_{i-2}, \text{ para } i = 2, ..., n-2, \\ A &= a_{n-1} + rb_{n-2} + qb_{n-3}, \\ B &= a_n + qb_{n-2}. \end{aligned}$$

Del mismo modo, a partir de los coeficientes del polinomio $p_1(x)$, se obtienen A_1 y B_1, según

$$\begin{aligned} c_0 &= b_0, \\ c_1 &= b_1 + rc_0, \\ c_i &= b_i + rc_{i-1} + qc_{i-2}, \text{ para } i = 2, ..., n-4, \\ A_1 &= b_{n-3} + rc_{n-4} + qc_{n-5}, \\ B_1 &= b_{n-2} + qc_{n-4}. \end{aligned}$$

Deflación y estabilidad de las raíces de polinomios.

Se sabe que las raíces de un polinomio son funciones continuas de sus coeficientes. Pero ocurre con frecuencia que bajo una pequeña perturbación de los coeficientes, las raíces se desplacen mucho, es decir, sean *inestables*. Este peligro debe ser tomado en cuenta cuando se pretende factorizar completamente un polinomio, o equivalentemente obtener todas las raíces, utilizando la técnica de *deflactar*. Para describir esta técnica supongamos que hemos obtenido $\tilde{\alpha}_1$, una aproximación de α_1, una raíz del polinomio de grado $n, p(x)$, mediante algún método numérico. Para encontrar una aproximación de otra raíz α_2 del mismo polinomio se puede considerar el polinomio de grado $(n-1)$,

$$q(x) = p(x) : (x - \alpha_1),$$

ya que α_2 será raíz de $p(x)$ si y solo si es raíz de $q(x)$. Pero en la práctica solo disponemos de una aproximación de la raíz α_1 y por lo tanto al *deflactar* y reducir el problema al cálculo de una raíz de $q(x)$, dividiendo $p(x)$ por el factor lineal $(x - \tilde{\alpha}_1)$, en realidad estaremos calculando una raíz del polinomio cuociente $\tilde{q}(x)$, definido por

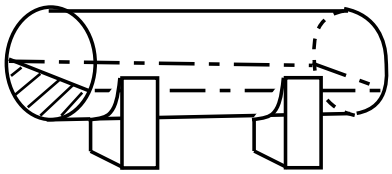
$$p(x) = \tilde{q}(x)(x - \tilde{\alpha}_1) + \delta(\tilde{\alpha}_1),$$

cuyos coeficientes corresponden a una perturbación de los coeficientes de $q(x)$. Como norma general, si se desean calcular una a una, todas las raíces de un polinomio mediante esta técnica de deflación, se cometerá menor error, debido a inestabilidades, si éstas se calculan en orden creciente en magnitud (calculando primero aquella de menor módulo).

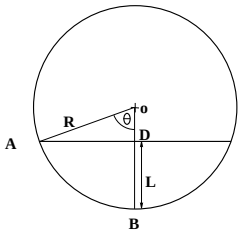
El método de Baisrtow se usa generalmente en un proceso de deflación para encontrar todos los factores lineales y cuadráticos, y de ese modo todas las raíces reales y complejas, de un polinomio. Tanto en este caso como en el anterior, se pueden compensar los errores debidos a inestabilidades producidas por la deflación, utilizando cada raíz aproximada resultante como adivinanza inicial en un proceso iterativo de cálculo de raíz que considere el polinomio original, sin deflactar.

EJERCICIOS PROPUESTOS

1. Un estanque cilíndrico, de radio R, contiene ácido sulfúrico para uso industrial. (Ver figura). Se requiere saber la altura h del nivel del líquido en su interior correspondiente a $1/4$ del estanque lleno. Modele este problema y proponga un método iterativo que le permita aproximar h con cualquier exigencia de precisión que le impongan.



Indicaciones: recuerde las fórmulas de $sen(2\alpha), cos(2\alpha)$.



$$\forall \alpha \in \left[0, \frac{\pi}{2}\right], \quad \cos\left(\frac{\pi - \alpha}{2}\right) = sen(\alpha).$$

$$\text{Área de la sección de círculo de arco } \overline{AB} = R^2\frac{\theta}{2}.$$

$$\text{Área del triángulo } OAD = \frac{R^2}{2} sen(\theta)cos(\theta).$$

2. Se propone calcular π mediante el método de Newton aplicado a la función $f(x) = sen(x)$, con adivinanza inicial $x_0 = \frac{5\pi}{6}$. Una manera de garantizar la convergencia de estas iteraciones es probar que x_0 pertenece a una vecindad de π donde la función de iteración es contractante.

   (a) Demuestre que la sucesión generada del modo propuesto permanece en una vecindad de π donde la función de iteración es contractante con constante de Lipschitz $L \leq \frac{1}{3}$.

   (b) Determine el número máximo de iteraciones en el que se obtendrá una aproximación de π con una precisión de $\varepsilon = 0.5 \cdot 3^{-20}$. *Indicación: recuerde que* $sen\left(\frac{5\pi}{6}\right) = \frac{1}{2}$ *y* $cos\left(\frac{5\pi}{6}\right) = -\frac{\sqrt{3}}{2}$.

3. Considere el método iterativo de un punto para aproximar una raíz α de la función f,

$$x_{n+1} = x_n - \frac{f(x_n)}{D_n}, \text{ con } D_n = \frac{f(x_n + f(x_n)) - f(x_n)}{f(x_n)}.$$

   Pruebe que si $f'(\alpha) \neq 0$, entonces este método converge con una velocidad caracterizada por el orden $p = 2$, partiendo de una adivinanza apropiada y bajo condiciones de regularidad de la función f que debe determinar. *Indicación: factorice* f *como* $f(x) = (x - \alpha)g(x)$, *con* $g(\alpha) \neq 0$.

4. Considere el algoritmo $x_{n+1} = x_n + sen(x_n)$ para calcular π. Encuentre un intervalo I que contenga a π y tal que $\forall x_0 \in I$, la sucesión converja a π. Pruebe que este método es de orden 2.

5. Considere las iteraciones

$$x_{n+1} = \frac{x_n(x_n^2 + 3a)}{3x_n^2 + a}, \quad \forall n \geq 0.$$

   Demuestre que este es un método de orden 3 para aproximar $\sqrt{a}$. Haga un estudio completo de la convergencia. Suponga que $|e_0| = |\sqrt{a} - x_0| \leq \varepsilon$ y encuentre condiciones para ε que garanticen que $|e_1| < |e_0|$. Calcule $\lim_{n\to\infty} \frac{e_{n+1}}{e_n^3}$. Programe este método y compare su comportamiento para $\alpha = 2$, con los métodos de Biseción, Secante, Regula Falsi y Newton del ejemplo 1.

6. Muestre que $x = \frac{1}{2}cosx$ tiene una solución α. Encuentre un intervalo $[a, b]$ que contenga a α y tal que $\forall x_0 \in [a, b]$ las iteraciones $x_{n+1} = \frac{1}{2}cos(x_n)$ converjan a α. Calcule una iteración y diga en cuantas iteraciones, a lo más, obtendrá una aproximación de α que satisfaga una tolerancia exigida ε.

7. Para encontrar una raíz α de la función f se propone el método iterativo

$$x_{n+1} = x_n + cf(x_n).$$

Si $f'(\alpha) \neq 0$, entonces con cual valor de c asegura la convergencia del método y con cual obtendría una convergencia de orden 2.

8. Considere la ecuación no lineal $x - tg(x) = 0$. Determine intervalos que contengan exactamente una solución (grafique la situación planteada). Analice el comportamiento del método de Newton en cada uno de estos intervalos. ¿En qué medida el comportamiento del método será distinto si se lo usa para aproximar la raíz más cercana a 100 a cuando se aproxima la raíz más cercana a cero?

9. Usando el método de Newton con la adivinanza inicial $(x_0, y_0) = (1.6, 1.2)$ realice iteraciones para resolver el sistema no lineal

$$\begin{aligned} x^2 + y^2 &= 4, \\ x^2 - y^2 &= 1, \end{aligned}$$

cuya solución es evidente.

10. Escriba un algoritmo que permita factorizar completamente un polinomio arbitrario mediante el método de Bairstow, incluyendo el algoritmo de tipo Horner para realizar las divisiones de polinomios que se requieran.

CAPÍTULO 7

Ecuaciones Diferenciales Ordinarias

Las Ecuaciones Diferenciales son algunas de las herramientas matemáticas más usadas en el modelamiento de fenómenos físicos. En este capítulo presentaremos, inicialmente, varios métodos para resolver numéricamente el problema de primer orden con condición inicial

$$\begin{aligned} y' &= f(x,y), \\ y(x_0) &= y_0. \end{aligned} \tag{7.1}$$

Los métodos que estudiaremos se generalizan fácilmente al caso de sistemas de primer orden, del tipo

$$\begin{array}{lll} y_1' = f_1(x,y_1,y_2,...,y_n), & \text{con condiciones iniciales} & y_1(x_0) = \eta_1, \\ y_2' = f_2(x,y_1,y_2,...,y_n), & & y_2(x_0) = \eta_2, \\ \quad\vdots & & \quad\vdots \\ y_n' = f_n(x,y_1,y_2,...,y_n), & & y_n(x_0) = \eta_n. \end{array} \tag{7.2}$$

y por lo tanto permiten resolver el problema de orden n, mayor que uno, con condiciones iniciales

$$\begin{aligned} y^{(n)} &= f(x,y,y',...,y^{(n-1)}), \\ y(x_0) &= \mu_0, \\ y'(x_0) &= \mu_1, \\ &\vdots \\ y^{(n-1)}(x_0) &= \mu_{n-1}. \end{aligned} \tag{7.3}$$

ya que éste se puede reducir al problema anterior mediante el cambio de variables

$$y_k = y_{k-1}' \quad \forall k = 2,...,n, \quad y_1 = y,$$

con lo que se obtiene el sistema

$$\begin{array}{lll} y_1' = y_2, & \text{con condiciones iniciales} & y_1(x_0) = \mu_0, \\ \quad\vdots & & y_2(x_0) = \mu_1, \\ y_{n-1}' = y_n, & & \quad\vdots \\ y_n' = f(x,y_1,...,y_{n-1}), & & y_n(x_0) = \mu_{n-1}. \end{array}$$

El problema de segundo orden con condiciones de borde será tratado más adelante pues los métodos utilizados en ese caso no corresponden a aplicaciones de los métodos anteriores.

Antes de abordar el problema numérico conviene establecer las condiciones bajo las cuales el problema (7.1) tiene solución única y un resultado de estabilidad de este problema.

Sea D un dominio del plano que contenga al punto (x_0, y_0) y sobre el cual la función f sea continua. Una función y será solución del problema (7.1) sobre $[a, b]$ si para todo $a \leq x \leq b, (x, y(x)) \in D$, existe $y'(x)$, la derivada, tal que $y'(x) = f(x, y(x))$, además de satisfacer la condición inicial $y(x_0) = y_0$.

A lo largo de todo el capítulo supondremos que el dominio D satisface la condición siguiente.

7.4. *Si los dos puntos $(x, y_1), (x, y_2)$ pertenecen a D entonces la recta vertical que los une también pertenece a D, es decir, $(x, \lambda y_1 + (1 - \lambda)y_2) \in D, \quad \forall 0 \leq \lambda \leq 1$.*

Teorema 7.5. *Sea f una función continua sobre el dominio D y Lipschitz en el segundo argumento, es decir, $\exists K \geq 0$ tal que*

$$|f(x, y_1) - f(x, y_2)| \leq K|y_1 - y_2| \quad \forall (x, y_1), (x, y_2) \in D.$$

Si (x_0, y_0) pertenece al interior del dominio D entonces existe un intervalo $I = (x_0 - \alpha, x_0 + \alpha)$, sobre el cual existirá una única solución $y(x)$ del problema (7.1).

Demostración. Ver libro de K. Atkinson [1]. □

Para estudiar la convergencia de los métodos numéricos deberemos imponer hipótesis más fuertes que ésta sobre la función f. En particular habrá que suponer que las derivadas parciales de f son continuas y bastará que $\frac{\partial f}{\partial y}$ sea continua para que, debido al teorema del valor medio, se tenga que f sea de Lipschitz en la segunda variable.

Para estudiar la estabilidad del problema (7.1) consideraremos un problema perturbado adecuado, en el sentido que también tenga solución única sobre un intervalo I sobre el cual exista solución única del problema (7.1). Bajo las mismas hipótesis para D y f del teorema anterior consideremos el problema perturbado

$$\begin{aligned} y' &= f(x, y) + \delta(x), \\ y(x_0) &= y_0 + \varepsilon, \end{aligned} \tag{7.6}$$

donde δ es una función continua para todo x tal que $(x, y) \in D$, dado y. Se puede probar que este problema tendrá una única solución, que denotaremos por $y(x; \delta, \varepsilon)$, sobre un intervalo fijo $I = [x_0 - \alpha, x_0 + \alpha]$, para toda perturbación que satisfaga $|\varepsilon| \leq \varepsilon_0, \ \|\delta\|_\infty \leq \varepsilon_0$ para algún ε_0 suficientemente pequeño.

En estas condiciones se tiene el siguiente resultado de estabilidad del problema diferencial.

Teorema 7.7. *Sean y la única solución del problema (7.1) e $y(\cdot; \delta, \varepsilon)$ la única solución del problema (7.6), sobre el intervalo $I = [x_0 - \alpha, x_0 + \alpha]$ del párrafo anterior. Entonces*

$$\|y(\cdot; \delta, \varepsilon) - y\|_{\infty, I} \leq \frac{1}{1 - K\alpha}(|\varepsilon| + \alpha\|\delta\|_{\infty, I}).$$

Demostración. Para todo $x \in I$

$$y(x) = \int_{x_0}^{x} f(t, y(t))dt + y_0,$$

$$y(x;\delta,\varepsilon) = \int_{x_0}^{x} (f(t, y(t;\delta,\varepsilon)) + \delta(t))dt + \varepsilon + y_0.$$

Restando estas dos expresiones se obtiene la desigualdad

$$\begin{aligned} |y(x;\delta,\varepsilon) - y(x)| &\leq \int_{x_0}^{x} |f(t, y(t;\delta,\varepsilon)) - f(t, y(t))|dt + \int_{x_0}^{x} |\delta(t)|dt + |\varepsilon| \\ &\leq K \int_{x_0}^{x} |y(t;\delta,\varepsilon) - y(t)|dt + \alpha\|\delta\|_{\infty,I} + |\varepsilon| \\ &\leq K\alpha\|y(\cdot;\delta,\varepsilon) - y\|_{\infty,I} + \alpha\|\delta\|_{\infty,I} + |\varepsilon| \end{aligned}$$

y con ello la demostración del resultado de estabilidad. □

El teorema (7.7) prueba que, bajo las condiciones descritas, el problema es estable, es decir, para perturbaciones pequeñas, el error será pequeño. Aún cuando la constante que amplifica la perturbación sea grande, el hecho que la relación entre el tamaño de la perturbación y la cota del error sea lineal, permite afirmar que se trata de un problema estable.

Para un estudio más profundo de las Ecuaciones Diferenciales Ordinarias existen numerosos textos y cursos dedicados a este tópico. Como nuestro interés se concentra en los aspectos numéricos dejaremos esta introducción hasta aquí.

Los métodos numéricos para Ecuaciones Diferenciales Ordinarias, **no** construyen una función $\widetilde{y}$ que aproxime a la solución y, sino que generan un conjunto de números

$$\{Y_k\}_{k=0}^{n}$$

que aproxima al conjunto

$$\{y(x_k)\}_{k=0}^{n}$$

de valores de la solución sobre una malla equiespaciada. En este sentido, los métodos numéricos construyen un objeto de naturaleza distinta a la solución analítica del problema, que es una función. Si se busca aproximar la solución del problema (7.1) sobre el intervalo $[a, b]$,

$$\text{con } x_k = a + kh \text{ y } h = \frac{b-a}{n},$$

para h suficientemente pequeño, la función lineal por pedazos que generará un ploteo computacional con los valores $\{Y_k\}_{k=0}^{n}$ se percibirá como una curva suave que aproxima a $y(x), \quad \forall x \in [a, b]$. Los métodos que veremos a continuación consisten en fórmulas para generar los valores Y_k. Estas fórmulas pueden ser vistas como sofisticaciones de un método muy intuitivo y simple, que incluiremos en esta introducción, por dos vías distintas, que dan origen a una clasificación en dos grupos; los métodos multipaso y los métodos de un paso o de Runge Kutta.

Método de Euler.

$$Y_{k+1} = Y_k + hf(x_k, Y_k), k \geq 0, \qquad \text{partiendo de } Y_0 = y_0. \tag{7.8}$$

Ejemplo 1. Consideremos la sencilla ecuación

$$\begin{aligned} y' &= 2y \\ y(0) &= 1 \end{aligned}$$

cuya solución es $y(x) = e^{2x}$. El método de Euler con paso $h = 0.1$, es decir, malla dada por $x_k = 0.1 \cdot k$, se producirá una sucesión de valores

$$Y_{k+1} = Y_k + 0.1 \cdot 2 \cdot Y_k = 1.2 \cdot Y_k$$

y por lo tanto $Y_k = (1.2)^k$ será el valor que aproxima a $y(x_k) = \exp(2x_k) = \exp(0.2 \cdot k)$. El comportamiento de estas aproximaciones se aprecia en la figura 7.1

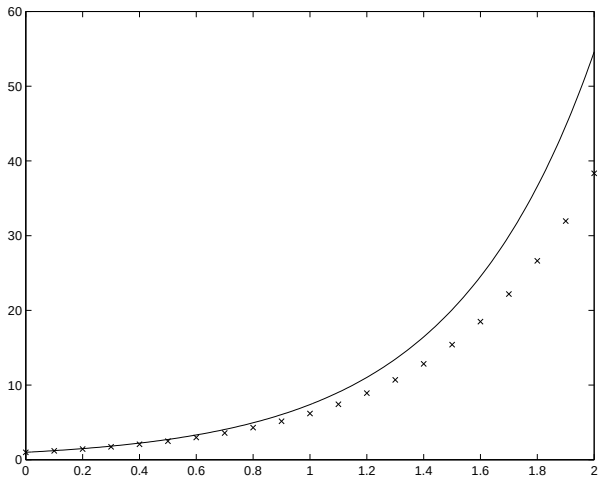


Figura 7.1: Método de Euler aplicado a $y' = 2y$ con condición inicial $y(0) = 1$

Una manera de mirar la formula (7.8) es a partir de la relación

$$\begin{aligned} y(x_{k+1}) &= y(x_k) + \int_{x_k}^{x_{k+1}} y'(t)dt \\ &= y(x_k) + \int_{x_k}^{x_{k+1}} f(t, y(t))dt. \end{aligned} \tag{7.9}$$

Si se aproxima la integral mediante una fórmula de Newton Cotes de un punto que consiste en integrar la constante correspondiente al valor de la función en el extremo izquierdo del intervalo de integración, es decir,

$$\int_{x_k}^{x_{k+1}} f(t, y(t))dt \approx hf(x_k, y(x_k)),$$

entonces bastará reemplazar $y(x_k)$ por Y_k para obtener la fórmula de Euler. Una estrategia, inspirada en esta observación será usar fórmulas de integración con más puntos para aproximar la integral y reemplazar los valores de la función y en los puntos anteriores de la malla por los valores Y_k calculados previamente. De este modo se originan los **métodos multipaso**.

Otra manera de mirar la fórmula de Euler es compararla con el desarrollo de Taylor de la función y en torno a x_k,

$$y(x_{k+1}) = y(x_k) + hy'(x_k) + \frac{h^2}{2}y''(\xi_k), \qquad \text{para algún} \quad \xi_k \in [x_k, x_{k+1}],$$

si la función y tiene dos derivadas continuas.

Como y es solución del problema (7.1) se puede reemplazar $y'(x_k)$ por $f(x_k, y(x_k))$, con lo cual se reconoce la fórmula de Euler y se puede concluir que el **error local** de este método, es decir el error que se produce en un paso, lo que equivale a suponer que $Y_k = y(x_k)$, será

$$y(x_{k+1}) - Y_{k+1} = \frac{h^2}{2}y''(\xi_k), \qquad \text{para algún } \xi_k \in [x_k, x_{k+1}].$$

Como el error local depende de h^2 se dice que el método de Euler es de **orden** 1. Según esta última perspectiva, para obtener fórmulas de mayor orden, es decir, con un error local que tienda a cero con una potencia más alta de h, habría que considerar desarrollos de Taylor de la función y con más términos. Así se originan las fórmulas de Runge-Kutta que estudiaremos a continuación.

FORMULAS DE RUNGE-KUTA

Estas son fórmulas de un paso, es decir, el cálculo de Y_{k+1} se realiza a partir de Y_k. Todas ellas se derivan de desarrollos de Taylor de la función y, de la que se requerirá la regularidad correspondiente, lo que a su vez implicará condiciones de regularidad de las derivadas parciales de la función f. El error local de estos métodos se obtendrá simultáneamente con la derivación de la fórmula. Sin embargo las expresiones que se obtienen no son tan simples como parece y se deben realizar desarrollos de Taylor al interior del segundo argumento de la función f. Para comprender el origen de esta particularidad de desarrollos *anidados*, que es típica de las fórmulas de Runge-Kutta, derivaremos la de orden 2.

Desarrollando la función y por Taylor en torno a x_k se obtiene

$$y(x_{k+1}) = y(x_k) + hy'(x_k) + \frac{h^2}{2}y''(x_k) + \frac{h^3}{3!}y'''(\xi_k), \text{ para algún } \xi_k \in [x_k, x_{k+1}], \tag{7.10}$$

si y''' es continua.

Para simplificar el desarrollo usaremos la notación $0(h^3)$ para el término del error $\frac{h^3}{3!}y'''(\xi_k)$. Como y es la solución de (7.1) se tiene que

$$y'(x_k) = f(x_k, y(x_k)) \text{ e } y''(x) = \frac{d}{dx}f(x, y(x)).$$

Desarrollando $f(x, y(x))$ en serie de Taylor en torno a x_k se obtiene

$$f(x_{k+1}, y(x_{k+1})) = f(x_k, y(x_k)) + h\frac{d}{dx}f(x_k, y(x_k)) + 0(h^2)$$

y por lo tanto

$$\frac{d}{dx}f(x_k, y(x_k)) = \frac{f(x_{k+1}, y(x_{k+1})) - f(x_k, y(x_k))}{h} + 0(h).$$

Reemplazando en (7.10) se obtiene

$$y(x_{k+1}) = y(x_k) + hf(x_k, y(x_k)) + \frac{h}{2}\{f(x_{k+1}, y(x_{k+1})) - f(x_k, y(x_k)) + 0(h^2)\} + 0(h^3)$$

y por lo tanto

$$y(x_{k+1}) = y(x_k) + \frac{h}{2}f(x_k, y(x_k)) + \frac{h}{2}f(x_{k+1}, y(x_{k+1})) + 0(h^3). \tag{7.11}$$

Usando nuevamente el desarrollo de Taylor de y en torno a x_k que usamos para analizar la fórmula de Euler, se obtendrá que

$$f(x_{k+1}, y(x_{k+1})) = f(x_{k+1}, y(x_k) + hf(x_k, y(x_k)) + 0(h^2))$$

y como f es de Lipschitz en el segundo argumento tenemos

$$f(x_{k+1}, y(x_{k+1})) = f(x_{k+1}, y(x_k) + hf(x_k, y(x_k))) + 0(h^2),$$

lo que reemplazado en (7.11) entrega

$$y(x_{k+1}) = y(x_k) + \frac{h}{2}\{f(x_k, y(x_k)) + f(x_{k+1}, y(x_k) + hf(x_k, y(x_k)))\} + 0(h^3). \tag{7.12}$$

Reemplazando $y(x_k)$ por Y_k se obtiene la fórmula de **Runge Kutta de orden 2**

$$Y_{k+1} = Y_k + \frac{h}{2}\{f(x_k, Y_k) + f(x_{k+1}, Y_k + f(x_{k+1}, Y_k))\}. \tag{7.13}$$

Ejemplo 2. El método de Runge-Kutta de orden 2 aplicado al mismo problema del ejemplo 1 con el mismo paso $h = 0.1$, produce los valores dados por la fórmula

$$Y_{k+1} = Y_k + \frac{0.1}{2} \cdot 2 \cdot Y_k + \frac{0.1}{2} \cdot 2 \cdot (Y_k + 0.1 \cdot 2 \cdot Y_k) = 1.22 \cdot Y_k.$$

Por lo tanto el valor que aproxima al valor de la solución sobre el k-ésimo punto de la malla es

$$Y_k = (1.22)^k$$

El la figura 7.2 se aprecia el comportamiento de estas aproximaciones y comparando con la figura 7.1 se observa la superioridad de este método sobre el método de Euler.

Los desarrollados de Taylor anidados se vuelven más complejos para órdenes mayores. Por ejemplo la fórmula de **Runge Kutta de orden 4** es

$$\begin{aligned} Y_{k+1} &= Y_k + \frac{h}{6}\{a_1 + 2a_2 + 2a_3 + a_4\}, \\ \text{con} \quad a_1 &= f(x_k, Y_k), \\ a_2 &= f(x_k + \frac{h}{2}, Y_k + \frac{1}{2}a_1), \\ a_3 &= f(x_k + \frac{h}{2}, Y_k + \frac{1}{2}a_2), \\ a_4 &= f(x_{k+1}, Y_k + a_3). \end{aligned} \tag{7.14}$$

Para establecer el error local de Runge-Kutta de orden 2, hay que explicitar el término del error $0(h^3)$ en (7.12), que dependerá de las primeras y segundas derivadas parciales de la función f, las que deberán ser continuas. Explicitar el error local de Runge-Kutta de orden 4, requerirá de cálculos aún más fatigosos debido a los múltiples desarrollos de Taylor que permiten derivar la expresión anterior.

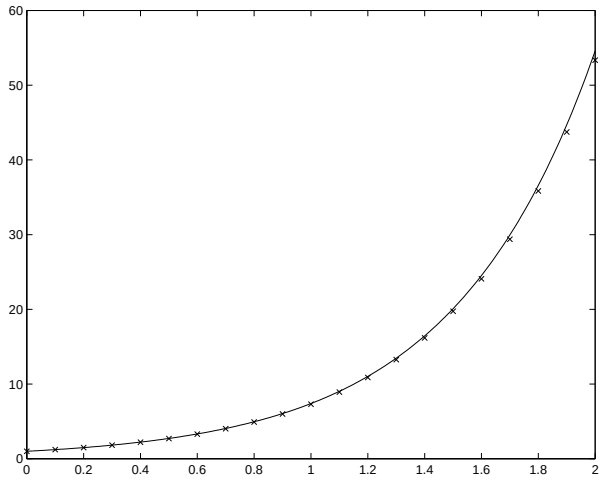


Figura 7.2: Método de Runge-Kutta de orden 2 aplicado a $y' = 2y$ con condición inicial $y(0) = 1$

MÉTODOS MULTIPASO

Vimos que el método de Euler podía ser visto como originado en una fórmula de integración numérica de Newton-Cotes de un punto. Veamos como se deriva un método de orden 2 usando una fórmula de integración numérica de 2 puntos.

$$\begin{aligned} y(x_{k+1}) &= y(x_k) + \int_{x_k}^{x_{k+1}} f(t, y(t))dt \\ &= y(x_k) + \frac{1}{2}\{f(x_k, y(x_k)) + f(x_{k+1}, y(x_{k+1}))\} - \frac{h^3}{12}\frac{d^2}{dx^2}f(\xi_k, y(\xi_k)), \end{aligned}$$

para algún $\xi_k \in [x_k, x_{k+1}]$, si y''' (o f'') es continua.

La fórmula que se obtiene reemplazando $y(x_k)$ por Y_k, llamada del **Trapecio** será

$$Y_{k+1} = Y_k + \frac{h}{2}\{f(x_k, Y_k) + f(x_{k+1}, Y_{k+1})\}. \tag{7.15}$$

Esta es una **fórmula implícita** pues el valor que se desea calcular, Y_{k+1}, aparece a ambos lados de la ecuación y, en general, no puede ser despejado de allí. Una manera de obtener una **fórmula explícita** con la misma estrategia de integración numérica consiste en integrar entre los mismos límites anteriores la recta que interpola a $f(x, y(x))$ en los dos puntos previos, x_{k-1} y x_k, con lo que se obtiene la fórmula de **Adams-Bashforth de orden 2**

$$Y_{k+1} = Y_k + \frac{h}{2}\{3f(x_k, Y_k) - f(x_{k-1}, Y_{k-1})\}, \tag{7.16}$$

cuyo error local será

$$y(x_{k+1}) - Y_{k+1} = \frac{5h^3}{12}\frac{d^2}{dx^2}f(\xi_k, y(\xi_k)).$$

Ejemplo 3. Aplicando el método de Adams-Bashforth de orden 2 al problema del ejemplo 1, sobre la misma malla de paso $h = 0.1$, se obtiene la fórmula

$$Y_{k+1} = 1.3 \cdot Y_k - 0.1 \cdot Y_{k-1}$$

En la figura 7.3 se muestra el comportamiento de esta solución numérica, usando el valor entregado por Runge-Kutta de orden 2 para $Y_1 = 1.22$.

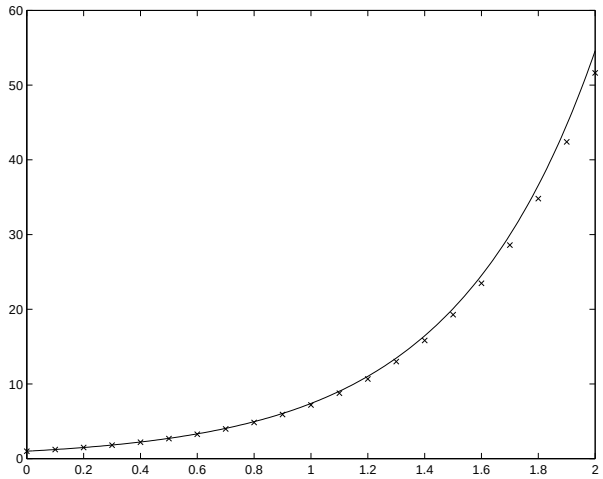


Figura 7.3: Método de Adams-Bashforth aplicado a $y' = 2y$ con condición inicial $y(0) = 1$

Sabemos que la fórmula de Newton-Cotes con tres puntos sobre una malla equiespaciada gana dos ordenes, es decir, comete un error que depende de h^5 y $\frac{d^4}{dx^4}f$ y por lo tanto la fórmula implícita de Simpson o de **Milne** (como se llama en este contexto), que sigue, será de orden 4.

$$Y_{k+1} = Y_{k-1} + \frac{h}{3}\{f(x_{k-1}, Y_{k-1}) + 4f(x_k, Y_k) + f(x_{k+1}, Y_{k+1})\} \tag{7.17}$$

comete un error local

$$y(x_{k+1}) - Y_{k+1} = -\frac{h^5}{90}\frac{d^4}{dx^4}f(\xi_k, y(\xi_k)),$$

para algún $\xi_k \in [x_{k-1}, x_{k+1}]$, si $\frac{d^4}{dx^4}f$ es continua.

Para obtener una fórmula explícita del mismo orden no se puede proceder como hicimos para pasar de la fórmula del Trapecio a la de Adams-Bashforth de orden 2, pues al perderse la simetría por el desplazamiento de los nodos de integración, se perderá la exactitud. La fórmula de **Adams-Bashforth de orden 4** se obtiene integrando entre x_k y x_{k+1} el polinomio de grado 3 que interpola a $f(x, y(x))$ en los nodos $x_{k-3}, x_{k-2}, x_{k-1}, x_k$,

$$Y_{k+1} = Y_k + \frac{h}{24}\{55f(x_k, Y_k) - 59f(x_{k-1}, Y_{k-1}) + 37f(x_{k-2}, Y_{k-2}) - 9f(x_{k-3}, Y_{k-3})\}, \tag{7.18}$$

cuyo error local es

$$y(x_{k+1}) - Y_{k+1} = \frac{251}{720}h^5\frac{d^4}{dx^4}f(\xi_k, y(\xi_k)),$$

para algún $\xi_k \in [x_{k-3}, x_{k+1}]$, si $\frac{d^4}{dx^4}f$ es continua.

En general los métodos multipaso deben ser complementados con los métodos de un paso para poder partir. Por ejemplo, si se pretende usar la fórmula de Adams-Bashforth de orden 4, anterior, ésta recién se podrá comenzar a emplear para el cálculo de Y_4. Para calcular Y_1, Y_2, Y_3 se recomienda usar una fórmula de un paso del mismo orden, por ejemplo Runge-Kutta de orden 4 dado por (7.4).

Las fórmulas implícitas que hemos obtenido, a pesar de no servir para cálculos directos, son de gran utilidad en una clase de métodos que utilizan iteraciones de punto fijo y que presentamos a continuación.

MÉTODO PREDICTOR-CORRECTOR

En las fórmulas implícitas como (7.15) y (7.17) el valor que se desea calcular Y_{k+1} aparece como un punto fijo de una función G_k. Por ejemplo, en la fórmula (7.15) la función mencionada será

$$G_k(t) = Y_k + \frac{h}{2}\{f(x_k, Y_k) + f(x_{k+1}, t)\}.$$

De este modo, si G_k es contractante, entonces partiendo de una adivinanza Y^0_{k+1} apropiada, se encontrará Y_{k+1} como el límite de la sucesión definida por recurrencia por

$$Y^{j+1}_{k+1} = G_k(Y^j_{k+1}). \tag{7.19}$$

Como hemos debido suponer continuidad de las derivadas parciales de f, para poder expresar el error local de estos métodos, y por lo tanto que f es Lipschitz en el segundo argumento, siempre se podrá encontrar un paso h suficientemente pequeño, de manera que la función G_k resulte contractante. Para la función del ejemplo, bastará tomar $h < \frac{2}{K}$ con K la constante de Lipschitz mencionada, para que G_k sea contractante y por lo tanto las iteraciones definidas por (7.19) converjan a Y_{k+1}. Para el método implícito de Milne se tendrá la función de iteración

$$G_k(t) = Y_{k-1} + \frac{h}{3}\{f(x_{k-1}, Y_{k-1}) + 4f(x_k, Y_k) + f(x_{k+1}, t)\},$$

que será contractante si $h < \frac{3}{K}$.

Del capítulo 6 sabemos que una adivinanza apropiada permitirá ahorra iteraciones. En este caso la adivinanza que se usará será aquella que resulte de usar una fórmula explícita del mismo orden que la fórmula implícita empleada como función de iteración. Por ejemplo, si se han de usar fórmulas de orden dos, el método de Predictor-Corrector calculará Y_{k+1} mediante aproximaciones sucesivas

- prediciendo $Y^0_{k+1} = Y_k + \frac{h}{2}\{3f(x_k, Y_k) - f(x_{k-1}, Y_{k-1})\}$,
- corrigiendo $Y^{j+1}_{k+1} = Y_k + \frac{h}{2}\{f(x_k, Y_k) + f(x_{k+1}, Y^j_{k+1})\}$ $j = 0, 1, ...$

En la práctica, las iteraciones se detendrán cuando dos valores consecutivos sean suficientemente parecidos, ya que como sabemos del capítulo anterior

$$|Y_{k+1} - Y^j_{k+1}| \leq \frac{L}{1-L}|Y^j_{k+1} - Y^{j-1}_{k+1}|,$$

donde L es la constante de Lipschitz de la función de iteración G_k, que en el caso de este ejemplo será

$$L = \frac{h}{2}\left\|\frac{\partial f}{\partial y}\right\|_\infty.$$

SISTEMAS DE ECUACIONES DE PRIMER ORDEN

Tanto los métodos de Runge-Kutta como los de Multipaso puede ser empleados para problemas del tipo (7.2), es decir, sistemas de Ecuaciones Diferenciales Ordinarias de primer orden con condiciones iniciales. para simplificar la notación, evitando el exceso de subíndices, consideraremos el problema de dos ecuaciones

$$\begin{aligned} y' &= f(x, y, z), \quad \text{con condiciones iniciales} \quad y(x_0) = y_0, \\ z' &= g(x, y, z), \qquad\qquad\qquad\qquad\quad z(x_0) = z_0. \end{aligned} \tag{7.20}$$

El método de **Euler** para este sistema es

$$\begin{aligned} Y_{k+1} &= Y_k + hf(x_k, Y_k, Z_k), \\ Z_{k+1} &= Z_k + hg(x_k, Y_k, Z_k). \end{aligned} \tag{7.21}$$

El método de **Runge-Kutta de orden 2** para este sistema es

$$\begin{aligned} Y_{k+1} &= Y_k + \frac{h}{2}\{f(x_k, Y_k, Z_k) + f(x_{k+1}, Y_k + hf(x_k, Y_k, Z_k), Z_k + hg(x_k, Y_k, Z_k))\}, \\ Z_{k+1} &= Z_k + \frac{h}{2}\{g(x_k, Y_k, Z_x) + g(x_{k+1}, Y_k + hf(x_k, Y_k, Z_k), Z_k + hg(x_k, Y_k, Z_k))\}. \end{aligned} \tag{7.22}$$

Los métodos multipaso se generalizan del mismo modo, por ejemplo, el método de **Adams-Bashforth de orden 4** para el sistema propuesto será

$$\begin{aligned} Y_{k+1} &= Y_k + \frac{h}{24}\left\{\begin{array}{ll} 55f(x_k, Y_k, Z_k) - 59f(x_{k-1}, Y_{k-1}, Z_{k-1}) & +37f(x_{k-2}, Y_{k-2}, Z_{k-2}) \\ & -9f(x_{k-3}, Y_{k-3}, Z_{k-3}) \end{array}\right\}, \\ Z_{k+1} &= Z_k + \frac{h}{24}\left\{\begin{array}{ll} 55g(x_k, Y_k, Z_k) - 59g(x_{k-1}, Y_{k-1}, Z_{k-1}) & +37g(x_{k-2}, Y_{k-2}, Z_{k-2}) \\ & -9g(x_{k-3}, Y_{k-3}, Z_{k-3}) \end{array}\right\}. \end{aligned} \tag{7.23}$$

Estos ejemplos ilustran la forma que adquieren los métodos antes vistos cuando se utilizan para sistemas de ecuaciones y permiten su generalización a sistemas más grandes así como desarrollar otros métodos de manera análoga.

ERROR GLOBAL Y ESTABILIDAD

Resulta bastante evidente que el error global de un método numérico será algo más complejo que la suma de los errores locales cometidos en cada uno de los pasos previos al k-ésimo. Recordemos que el error local se define como la diferencia de los valores

$$y(x_{k+1}) - Y_{k+1}, \tag{7.24}$$

suponiendo que

$$y(x_j) = Y_j, \quad \forall 0 \leq j \leq k. \tag{7.25}$$

Observando las fórmulas estudiadas y la manera en que estos valores previos participan en el cálculo de Y_{k+1}, queda claro que las diferencias $y(x_j) - Y_j \quad \forall 0 \leq j \leq k$, no solo se acumularán sino que se *propagarán*. Y esto ocurrirá de un modo que dependerá de la estabilidad del método considerado. Se define el error global como la diferencia (7.24) en ausencia del supuesto (7.25). Es en general bastante difícil estudiar este error pero, según el comentario previo, una información relevante para analizar su comportamiento se refiere a la estabilidad de la fórmula respectiva.

Todos los métodos numéricos de este capítulo en lugar de resolver Ecuaciones Diferenciales resuelven **Ecuaciones de Diferencias**, este es el nombre bajo el cual se conocen las expresiones del tipo

$$\begin{aligned} F_k(Y_k, Y_{k+1}, ..., Y_{k+p}) = 0, \quad \forall k = 0, 1, 2, \ldots \\ Y_i = y_i, \text{ dados}, \quad \forall i = 0, 1, ..., p-1. \end{aligned} \tag{7.26}$$

Estudiar la estabilidad de los métodos numéricos implicará estudiar la estabilidad de las Ecuaciones de Diferencias. Tal objetivo puede parecer muy ambicioso dado que en la función F_k que aparece en (7.26) sabemos que intervendrá la función f, del lado derecho de la ecuación diferencial, y por lo tanto la Ecuación

de Diferencias puede adquirir formas muy complejas, dependiendo, no sólo del método particular sino del problema diferencial al que se aplique. El único caso en que el estudio de la estabilidad resulta sencillo, es aquel en que la Ecuación de Diferencias es lineal y a coeficientes constantes, es decir, cuando

$$F_k(Y_k, Y_{k+1}, ..., Y_{k+p}) = \alpha_0 Y_k + \alpha_1 Y_{k+1} + ... + \alpha_p Y_{k+p}, \text{ con } \alpha_p \neq 0. \tag{7.27}$$

En este caso, la Ecuación de Diferencias (7.26) se puede reescribir usando una notación matricial que facilitará analizar su evolución, cuando $k \to \infty$.

Sea $u_k = \begin{pmatrix} Y_k \\ Y_{k+1} \\ \vdots \\ Y_{k+p-1} \end{pmatrix} \in I\!R^p$. La ecuación (7.26) será equivalente a

$$\begin{bmatrix} 0 & 1 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 0 & & & 0 \\ 0 & 0 & 0 & 1 & & & 0 \\ \vdots & & & \ddots & \ddots & & \vdots \\ \vdots & & & & \ddots & \ddots & 0 \\ \vdots & & & & & 0 & 1 \\ \frac{-\alpha_0}{\alpha_p} & \frac{-\alpha_1}{\alpha_p} & \frac{-\alpha_2}{\alpha_p} & \cdots & \cdots & \cdots & \frac{-\alpha_{p-1}}{\alpha_p} \end{bmatrix} u_k = u_{k+1}, \quad \forall k \geq 0, \quad u_0 \text{ dado.} \tag{7.28}$$

Llamando A a la matriz de la ecuación anterior, se tendrá que la ecuación de diferencias equivale a

$$u_k = A^k u_0, \forall k \geq 1, \tag{7.29}$$

con u_0 dado.

Considerando la forma de Jordan de la matriz A como matriz compleja, es decir, la factorización de A con las matrices J, triangular superior con los valores propios de A, $\lambda_1, \lambda_2, ..., \lambda_n$, en su diagonal, y P invertible (en general complejas), tales que

$$A = PJP^{-1},$$

se tendrá que la ecuación (7.29) será equivalente a

$$u_k = PJ^kP^{-1}u_0, \quad \forall k \geq 1, \tag{7.30}$$

con u_0 dado.

De este modo se evidencia que la evolución de u_k (y por lo tanto de los valores Y_k) cuando $k \to \infty$, dependerá exclusivamente de los valores propios de la matriz A. Desarrollando el determinante de $(A - \lambda I)$ por la última columna se obtiene

$$Det(A - \lambda I) = (-1)^p \left\{ \lambda^p + \frac{\alpha_{p-1}}{\alpha_p}\lambda^{p-1} + ... + \frac{\alpha_1}{\alpha_p}\lambda + \frac{\alpha_0}{\alpha_p} \right\} \tag{7.31}$$

y por lo tanto los valores propios de A, raíces del polinomio característico de A, (7.31), serán también raíces del polinomio.

$$Q(\lambda) = \alpha_p\lambda^p + \alpha_{p-1}\lambda^{p-1} + ... + \alpha_1\lambda + \alpha_0, \tag{7.32}$$

llamado polinomio característico de la Ecuación de Diferencias (7.26) en el caso que ésta sea lineal del tipo caracterizado por (7.27).

Hacemos ver que $\forall \lambda_i$, raíz de $Q(\lambda)$, valor propio de A, se tiene que $Y_k = \lambda_i^k$ satisface

$$F_k(Y_{k+1}, ..., Y_{k+p}) = \alpha_0 Y_k + \alpha_1 Y_{k+1} + ... + \alpha_p Y_{k+p} = 0. \tag{7.33}$$

En efecto, si $Y_k = \lambda_i^k, \quad \forall k, \quad$ con λ_i raíz de $Q(\lambda)$, entonces

$$\begin{aligned} \alpha_0 Y_k + \alpha_1 Y_{k+1} + ... + \alpha_p Y_{k+p} &= \lambda_i^k(\alpha_0 + \alpha_1 \lambda_i + ... + \alpha_p \lambda_i^p) \\ &= \lambda_i^k Q(\lambda_i) \\ &= 0. \end{aligned}$$

Es bastante simple verificar que si $\lambda_1, \lambda_2, ..., \lambda_p$, son las p raíces de $Q(\lambda)$, entonces

$$Y_k = a_i \lambda_1^k + a_2 \lambda_2^k + ... + a_p \lambda_p^k \tag{7.34}$$

satisface (7.33) para cualquier juego de coeficientes $a_1, a_2, ..., a_p$. De este modo bastará identificar los coeficientes $a_1, a_2, ..., a_p$ que permitan satisfacer las condiciones iniciales

$$Y_i = y_i, \quad \forall i = 0, 1, ..., p-1,$$

para obtener la solución de la Ecuación de Diferencias (7.26).

Para la estabilidad de una solución de la ecuación de diferencias del tipo analizado se tiene la definición siguiente:

Definición 7.35. Se dirá que una solución del tipo (7.34) es estable, si los valores propios λ_i a quienes corresponda un coeficiente a_i nulo, satisfacen $|\lambda_i| < 1$.

Ejemplo 4. Consideremos la ecuación de diferencias

$$\begin{aligned} 8Y_k - 17Y_{k+1} + 2Y_{k+2} = 0, \quad \forall k \geq 0, \\ Y_0 = 2, Y_1 = 1. \end{aligned}$$

El polinomio característico de esta ecuación será $Q(\lambda) = 8 - 17\lambda + 2\lambda^2$, cuyas raíces son $\lambda_1 = \frac{1}{2}, \lambda_2 = 8$, por lo tanto la solución de la ecuación de diferencias será

$$Y_k = a_1(\frac{1}{2})^k + a_2 8^k, \forall k \geq 0,$$

con a_1, a_2 tales que

$$\begin{aligned} a_1 + a_2 = 2, \\ \frac{a_1}{2} + 6a_2 = 1, \end{aligned}$$

es decir, $a_1 = 2, a_2 = 0$ y la solución será $Y_k = \frac{1}{2^{k-1}}$.

Según la definición previa, esta solución no será estable, ya que $|\lambda_2| = 8 > 1$. En efecto, si se perturban levemente las condiciones iniciales, por ejemplo, si

$$Y_0 = 2.0000001, \quad Y_1 = 1.000008,$$

entonces $a_1 = 2 \quad a_2 = 10^{-6}$ y por lo tanto $Y_k = \frac{1}{2^{k+1}} + 10^{-6} 8^k$, que será muy diferente a la solución anterior aún para valores pequeños de k y esta diferencia aumentará enormemente a medida que k crezca.

La estabilidad de los métodos numéricos para resolver ecuaciones diferenciales se estudia, con las técnicas anteriores, para una ecuación diferencial particular, llamada **ecuación test**,

$$\begin{aligned} y'(x) &= wy(x), \quad x \geq 0, \\ y(0) &= 1, \end{aligned} \tag{7.36}$$

donde w es un parámetro que se elige para analizar distintos tipos de soluciones. La solución de (7.36) es obviamente $y(x) = e^{wx}$ y por lo tanto dependiendo del parámetro w se tendrán los casos siguientes

- si $w > 0$, entonces $y(x) \to \infty$ cuando $x \to \infty$,
- si $w = 0$, entonces $y(x) = 1 \quad \forall x$,
- si $w < 0$, entonces $y(x) \to 0$ cuando $x \to \infty$,
- si w es complejo, entonces $y(x)$ oscila.

De este modo la ecuación test permite cubrir una amplia gama de comportamientos de soluciones y tiene la ventaja que la función $f(x, y) = wy$, del lado derecho, producirá en todos los métodos numéricos estudiados una Ecuación de Diferencias lineal a coeficientes constantes, cuya estabilidad sabremos analizar a través de las raíces de su polinomio característico.

Ejemplo 5. Consideremos, por ejemplo, el método de Adams-Bashforth de orden 2, dado en (7.16). La ecuación de diferencias asociada a este método aplicado a la ecuación test será

$$\frac{1}{2}hwY_{k-1} - \left(1 + \frac{3}{2}hw\right) Y_k + Y_{k+1} = 0, \quad \forall k \geq 1. \tag{7.37}$$

Su polinomio característico será

$$Q(\lambda) = \frac{1}{2}hw - \left(1 + \frac{3}{2}hw\right) \lambda + \lambda^2,$$

cuyas raíces dependerán de w. Como el paso h es un real positivo y multiplica a w en todas partes donde ésta aparece en la expresión anterior, resulta más cómodo analizar los valores propios como funciones de hw,

$$\lambda_i(hw) \text{ para } i = 1, 2,$$

en los casos $hw > 0$, $hw < 0$, $hw = 0, hw$ complejo.

En el ejemplo analizado

$$\lambda_1(hw) = \frac{1}{2}\left\{1 + \frac{3}{2}hw + \sqrt{1 + hw + \frac{9}{4}h^2w^2}\right\},$$
$$\lambda_2(hw) = \frac{1}{2}\left\{1 + \frac{3}{2}hw - \sqrt{1 + hw + \frac{9}{4}h^2w^2}\right\}.$$

Si $w = 0$, entonces $\lambda_1(0) = 1, \quad \lambda_2(0) = 0$ y la solución numérica será

$$Y_k = 1, \quad \forall k,$$

lo que corresponde exactamente a $y(x_k)$, $\forall k$, en este caso y por lo tanto se trata de una solución exacta.

Si $w \neq 0$, el método de Adams-Bashforth necesitará el valor de Y_1, además de la condición inicial $Y_0 = 1$, calculado con algún método de un paso para poder partir. (Si se usa Euler se tendrá que $Y_1 = 1 + hw$, si se

usa Runge Kutta de orden 2 se tendrá que $Y_1 = 1 + hw + \frac{h^2}{2}w^2$, lo que corresponde a desarrollos truncados de Taylor de $y(x_1) = e^{wh}$, en torno a 0).

Como $\lambda_i(hw)$, para $i = 1, 2$, son funciones continuas de hw, entonces para hw suficientemente pequeño $\lambda_2(hw)$ será suficientemente parecido a 0, es decir, $|\lambda_2(hw)| < 1$.

Si $w > 0$, entonces $\lambda_1(hw) > 1$ y $a_1 > 0$, como se puede apreciar al resolver el sistema

$$\begin{aligned} a_1 + a_2 &= 1, \\ a_1\lambda_1(hw) + a_2\lambda_2(hw) &= Y_1, \end{aligned}$$

con cualquiera de las dos posibles elecciones de Y_1. De modo que la solución de la ecuación de diferencias

$$Y_k = a_1(\lambda_1(hw))^k + a_2(\lambda_2(hw))^k \tag{7.38}$$

crecerá cuando $k \to \infty$, tal cual lo hace la solución de la ecuación diferencial cuando $x_k = kh \to \infty$.

Si $w < 0$, entonces $\lambda_1(hw) < 1$ y positivo para h suficientemente pequeño, con lo cual la solución numérica (7.38) tenderá a cero cuando $k \to \infty$ tal cual lo hace la solución de la ecuación diferencial cuando $x_k = kh \to \infty$, en este caso.

En resumen, para los distintos valores reales de w analizados, para un paso h suficientemente pequeño, la solución de la ecuación de diferencias correspondiente al método de Adams-Bashforth de orden 2 se comporta del mismo modo que la solución de la ecuación diferencial aproximada por ésta. Esto permite afirmar que se trata de un método estable.

Todos los métodos estudiados previamente, excepto el de Milne, sometidos al análisis anterior, resultan ser métodos estables, para un paso h suficientemente pequeño. Debido al comportamiento de las soluciones de la ecuación test, la estabilidad del método numérico se caracteriza por el hecho que todas las raíces del polinomio característico correspondiente satisfacen

$$|\lambda_i(hw)| < 1,$$

con la única posible excepción de $\lambda_1(hw)$, si se escoge este índice para aquella en que $\lambda_1(0) = 1$.

PROBLEMA CON CONDICIONES DE BORDE

Existen muchos tipos de problemas con distintas condiciones de borde que podrían tener cabida bajo un título como éste. El problema que abordaremos en esta sección será

$$\begin{aligned} y''(x) &= f(x, y, y'), \quad a \le x \le b, \\ y(a) &= \eta_1, \\ y(b) &= \eta_2. \end{aligned} \tag{7.39}$$

Bajo supuestos poco exigentes sobre la función f (comparados con las hipótesis que ésta debe satisfacer para la convergencia de los métodos numéricos), existirá una única solución de este problema. En lo que sigue supondremos que el problema (7.39) tiene solución única y nos abocaremos a su obtención numérica.

El primer método que presentaremos se basa en la observación siguiente. Si en lugar de la segunda condición de borde $y(b) = \eta_2$, tuviéramos una condición inicial $y'(a) = \alpha$, entonces tendríamos un problema de segundo orden con condiciones iniciales, del tipo (7.3), que sabemos resolver, llevándolo a un sistema de dos ecuaciones de primer orden con condiciones iniciales, del tipo (7.2). En cambio, los métodos numéricos de que disponemos no permiten resolver el problema planteado.

Método del Disparo.

Si el problema (7.39) tiene una única solución $y(x)$, entonces existirá $\alpha \in \mathbb{R}$, tal que

$$\begin{aligned} y''(x) &= f(x, y, y'), \quad a \le x \le b, \\ y(a) &= \eta_1, \\ y(b) &= \alpha, \end{aligned} \tag{7.40}$$

sea equivalente al problema (7.39).

El método del disparo resolverá el problema (7.40) con las técnicas conocidas, para sucesivas aproximaciones de α.

Sea $y_i(x)$, la solución del problema

$$\begin{aligned} y''(x) &= f(x, y, y'), \quad a \le x \le b, \\ y(a) &= \eta_1, \\ y'(a) &= \alpha_i, \end{aligned}$$

para α_i dado.

Comparando $y_i(b)$ con η_2 se sabrá si la pendiente inicial con que fue disparada la curva $y_i(x)$ es excesiva o deficitaria y se podrá ajustar el valor de ésta, α_i. Así planteado el método se trata de una estrategia de *ensayo y error* para adivinar el valor de α. El problema de encontrar el valor de α que hace equivalentes los problemas (7.39) y (7.40) se puede plantear como un problema conocido y por lo tanto sistematizar su resolución.

Sea $G(\alpha) = y(b; \alpha)$ donde $y(\cdot; \alpha)$ denota la solución de (7.40) para α dado. Encontrar α tal que

$$G(\alpha) = \eta_2,$$

equivale a encontrar una raíz de una función no lineal que, en general, no sabremos explicitar. Pero varios de los métodos numéricos estudiados en el capítulo anterior solo requieren evaluaciones de la función no lineal y no se necesita la expresión analítica de ella. Difícilmente se podrá establecer a priori que se verifican las condiciones de convergencia de estos métodos en el caso planteado, pero sabemos que en todos ellos es importante partir de una buena adivinanza, cercana al valor de α buscado.

En la práctica, en el lugar de conocer $G(\alpha) = y(b; \alpha)$, para un valor dado de α, solo conoceremos una aproximación, esto es del valor de Y_n^α, si $\{Y_k^\alpha\}_{k=0}^n$ es el resultado de aplicar algún método numérico al problema (7.40) con una malla equiespaciada de paso $h = \frac{b-a}{n}$. El error global acumulado en esta aproximación puede producir inestabilidades.

Un algoritmo sencillo que utiliza el método de la Secante para la ecuación no lineal y que por lo tanto requiere de dos adivinanzas α_0, α_1 para partir, sería:

1. $\alpha \leftarrow \alpha_0$,
2. resolver numéricamente (7.40), poner $\beta_0 = Y_n$,
3. si $|\beta_0 - \eta_2| \le \varepsilon$ (tolerancia dada) parar,
4. $\alpha \leftarrow \alpha_1$,
5. resolver numéricamente (7.40), poner $\beta_1 = Y_n$,
6. mientras $|\beta_1 - \eta_2| > \varepsilon$
   (a) $\alpha \leftarrow \frac{\alpha_1(\beta_0 - \eta_2) - \alpha_0(\beta_1 - \eta_2)}{\beta_0 - \beta_1}$,

(b) $\beta_0 \leftarrow \beta_1$,

(c) resolver numéricamente (7.40), poner $\beta_1 = Y_n$,

(d) $\alpha_0 \leftarrow \alpha_1, \alpha_1 \leftarrow \alpha$,

7. salida $\{Y_k\}_{k=0}^n$.

Método de Diferencias Finitas.

Cuando la función del lado derecho del problema (7.39) es lineal en el segundo y tercer argumento, es decir,

$$f(x, y, y') = r(x) + q(x)y(x) + p(x)y'(x), \tag{7.41}$$

se puede aplicar un método muy popular en Ecuaciones Diferenciales Parciales. Este consiste en aproximar los operadores diferenciales por operadores de diferencias, sobre un dominio discreto constituido por la malla equiespaciada $\{x_k\}_{k=0}^n \subset [a, b]$, con $x_k = a + kh, \quad h = \frac{b-a}{n}$.

Por ejemplo

$$\begin{aligned} &\text{a)} \quad y'(x_k) \approx \frac{y(x_{k+1}) - y(x_k)}{h}, \\ &\text{b)} \quad y'(x_k) \approx \frac{y(x_{k+1}) - y(x_{k-1})}{2h}, \\ &\text{c)} \quad y''(x_k) \approx \frac{y(x_{k+1}) - 2y(x_k) + y(x_{k-1})}{h^2}. \end{aligned} \tag{7.42}$$

La calidad de estas aproximaciones se puede establecer mediante desarrollos truncados de Taylor en torno a x_k.

En lugar de encontrar una función $y(x)$ que satisfaga la ecuación diferencial $\forall x \in [a, b]$, el método de Diferencias Finitas encontrará los valores $\{Y_k\}_{k=1}^{n-1}$ que satisfagan el sistema de ecuaciones que resulta de imponer la ecuación de diferencias en cada punto interior de la malla. Note que $Y_0 = y(x_0) = \eta_1, Y_n = y(x_n) = \eta_2$. Ilustraremos este procedimiento en un ejemplo.

Ejemplo 6. Consideremos el problema

$$\begin{aligned} y''(x) &= -2y(x) - 2y'(x), \quad \forall x \in [0, \frac{\pi}{2}], \\ y(0) &= 1, \\ y\left(\frac{\pi}{2}\right) &= 0. \end{aligned} \tag{7.43}$$

Usando los operadores de diferencias de (7.42) b) y c), se tendrá el problema aproximado

$$\begin{aligned} \frac{Y_{k-1} - 2Y_k + Y_{k+1}}{h^2} &= -2Y_k - 2\frac{Y_{k+1} - Y_{k-1}}{2h}, \quad \forall k = 1, 2, ..., n-1, \\ Y_0 &= 1, \\ Y_n &= 0, \end{aligned} \tag{7.44}$$

que equivale al sistema lineal de tamaño $(n-1)$ $AY = b$, con

$$A = \begin{bmatrix} (2h^2-2) & (h+1) & & & & & & & \\ & \ddots & \ddots & & & & & & \\ & & \ddots & \ddots & & & & & \\ & & & (1-h) & (2h^2-2) & (h+1) & & & \\ & & & & & \ddots & \ddots & & \\ & & & & & & \ddots & \ddots & \\ & & & & & & & (1-h) & (2h^2-2) \end{bmatrix},$$

(7.45)

$$Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{k-1} \\ Y_k \\ Y_{k+1} \\ \vdots \\ Y_{n-2} \\ Y_{n-1} \end{pmatrix}, \qquad b = \begin{pmatrix} h-1 \\ 0 \\ \vdots \\ \vdots \\ 0 \\ \vdots \\ \vdots \\ \vdots \\ 0 \end{pmatrix}.$$

La solución del problema (7.43) es $y(x) = e^{-x}cosx$. Considerando $n = 8$ y por lo tanto $h = \frac{\pi}{16}$, la solución numérica obtenida por el método de Diferencias Finitas, es decir, la solución del sistema anterior, se consigna en la tabla 7.1.

k	$y(x_k)$	Y_k
1	0.8059	0.8053
2	0.6238	0.6226
3	0.4613	0.4598
5	0.3224	0.3207
6	0.1178	0.1167
7	0.0494	0.0488

Tabla 7.1: Solución de (7.43) por método de diferencias finitas

Se aprecia claramente la ausencia de un error global que se propague y aumente. Por otra parte, con una malla bastante pobre se obtiene una solución numérica que comete un error

$$\max_{0 \le k \le 8} |y(x_k) - Y_k| = 0.0017\ldots$$

que no se alcanza a apreciar en un gráfico. Aumentando el número de puntos de la malla, con $n = 80$ se obtiene una solución numérica que comete un error

$$\max_{0 \le k \le 80} |y(x_k) - Y_k| = 1.7 \cdot 10^{-5}$$

EJERCICIOS PROPUESTOS

1. Considere el problema

$$\begin{aligned} y'(x) &= 10y(x), \quad x \geq 0, \\ y(0) &= 1. \end{aligned}$$

Demuestre que con el paso $h = 0.1$, el método Predictor Corrector que utiliza Adams-Bashforth para predecir y Trapecio para corregir, converge. Calcule con este método y usando solo fórmulas de orden 2 (cuidado al comienzo) $Y_k, \quad 1 \leq k \leq 5$, corrigiendo tantas veces como sea necesario para que $|Y_{k+1}^{j+1} - Y_{k+1}^{j}| < 0.1$.

2. Diseñe un algoritmo que permita utilizar el método del Disparo con Bisección para resolver la ecuación no lineal.

3. Pruebe que si $p(x) \geq 0, q(x) > 0, \quad \forall x \in [a,b]$, entonces el método de Diferencias Finitas, que usa los operadores de diferencia (7.42) **b)** y **c)**, para resolver el problema (7.39) cuando $f(x,y,y')$ es de la forma (7.41), producirá un sistema lineal que podrá ser resuelto mediante los métodos iterativos de Gauss-Seidel y Jacobi.

4. Pruebe que los operadores de diferencia usados en el problema anterior aproximan a los respectivos operadores diferenciales con un error $0(h^2)$.

5. Considere el problema con condiciones de borde

$$\begin{aligned} y''(x) &= r(x) + q(x)y(x) + p(x)y'(x), a \leq x \leq b, \\ y'(a) &= \alpha_1, \\ y'(b) &= \alpha_2. \end{aligned}$$

Usando los operadores de diferencia (7.42) **b)** y **c)**, y aproximando las condiciones de borde por $\frac{Y_1 - Y_0}{h} = \alpha_1$, $\frac{\mathbf{Y_n - Y_{n-1}}}{\mathbf{h}} = \alpha_2$, obtenga el sistema producido por el método de Diferencias Finitas.

6. Usando los métodos del Disparo y de Diferencias Finitas resuelva los problemas con condiciones de borde

$$\begin{aligned} & y''(x) = x^2 y'(x) + y(x), \quad 1 \leq x \leq 2, \\ \text{(a)} \quad & y(1) = 1, \\ & y(2) = 1. \\ & y''(x) = (cosx)y'(x) + (senx)y(x) + \pi, \quad 0 \leq x \leq \pi, \\ \text{(b)} \quad & y(0) = 1, \\ & y(\pi) = 0. \end{aligned}$$

7. Considere la fórmula $Y_{k+1} = 4Y_k - 3Y_{k-1} - 2hf(x_{k-2}, Y_{k-2})$.

Expandiendo $y(x_{k-1}), y'(x_{k-2})$ en serie de Taylor en torno a x_k, pruebe que el error local es

$$-24h^2 \frac{d^2}{dx^2} f(\xi_k, y(\xi_k)).$$

Analice la estabilidad de esta fórmula.

8. Explicite el error de Runge-Kutta de orden 2, en términos de las derivadas parciales de la función f.

9. Use los métodos de Euler, Runge-Kutta de orden 2 y Adams Bashforth de orden 2 para resolver el problema

$$\begin{aligned} y'(x) &= \left(\frac{2}{x} - 1\right) y(x), 1 \leq x \leq 2, \\ y(1) &= 1. \end{aligned}$$

10. Usando el método de Euler para sistemas, resuelva

$$\begin{aligned} y''(x) &= xy'(x) + x^2 y(x) + 1, \quad 0 \leq x \leq 1, \\ y(0) &= 1, \\ y'(0) &= 1. \end{aligned}$$

11. Usando desarrollos de Taylor pruebe que la fórmula de Runge Kutta de orden 4 dada en (7.14) comete un error local $0(h^5)$.

12. Obtenga la fórmula implícita que resulta de aproximar la integral $\int_{x_k}^{x_{k+1}} f(x, y(x))dx$ por la integral del polinomio de grado 3 que interpola a $f(x, y(x))$ en los nodos $x_{k-2}, x_{k-1}, x_k, x_{k+1}$. Calcule el orden de esta fórmula.

Bibliografía

[1] Kendall E. Atkinson. *An Introduction to Numerical Analysis.* John Wiley & Sons, 1978.

[2] Richard I. Burden and J.Douglas Faires. *Análisis Numérico.* International Thomson Editores, 1998.

[3] Philippe G. Ciarlet. *Introduction À L'analyse Numérique Matricielle et À L'optimisation.* G. Mason, 1990.

[4] Samuel Daniel Conte and Carl De Boor. *Elementary Numerical Analysis.* McGraw-Hill, 1965.

[5] Gene Howard Golub and Charles F. Van Loan. *Matrix Computations.* J. Hopkins University Press, Baltimore, 1996.

[6] Anthony Ralston. *A First Course in Numerical Analysis.* McGraw-Hill, 1978.

[7] Josef Stoer and Roland Bulirsch. *Introduction to Numerical Analysis.* Springer-Verlag, 1980.

[8] James S. Vandergraft. *Introduction to Numerical Computations.* Academic Press, Inc, 1983.

[9] Richard S. Varga. *Matrix Iterative Analysis.* Prentice Hall, 1962.