ELEMENTOS DE CÁLCULO NUMÉRICO

Ricardo G. Durán, Silvia B. Lassalle y Julio D. Rossi

Índice General

CAPíTULO 1

Punto flotante y redondeo

El objeto de este capítulo es analizar la representación de los números en una computadora y la propagación de los errores de redondeo al realizar cálculos.

Como la cantidad de información que puede guardarse en una computadora es finita, la máquina trabajará sólo con un conjunto finito de números. A éstos los llamaremos *números de máquina*. En consecuencia, toda vez que de nuestros datos o cálculos surja un número que no pertenece a este conjunto finito, éste deberá ser reemplazado por una aproximación (el número de máquina más cercano). Este reemplazo da lugar a lo que llamamos errores de redondeo.

Al realizar cálculos estos errores de redondeo se propagan y esto puede llevar a resultados totalmente incorrectos como veremos en algunos ejemplos simples.

En las aplicaciones del cálculo numérico es prácticamente imposible determinar exactamente la magnitud de los errores de redondeo. Lo que si puede hacerse, y es de fundamental importancia, es identificar las posibles causas de que los errores se propaguen más de lo admisible. Esto permite mejorar los algoritmos o determinar que método es más conveniente para resolver un problema. Un claro ejemplo de esto, que veremos más adelante, aparece cuando se utiliza el método de eliminación de Gauss para resolver un sistema de ecuaciones lineales. En este caso, el análisis de la propagación de errores permite determinar la forma más eficiente de aplicar el método.

Por otra parte, es fundamental distinguir cuando la propagación excesiva de errores se debe a que el algoritmo utilizado es "malo" o inestable o a que el problema en sí mismo está "mal condicionado". En el primer caso se puede (se debe!) tratar de mejorar el método de resolución mientras que en el segundo caso el problema es más esencial. Los ejemplos que presentaremos ilustrarán estos dos casos.

1. Punto flotante

En lo que sigue supondremos que los *números de máquina* son los que aparecen en la pantalla. Esto no es exacto pues en realidad la computadora opera internamente con los números desarrollados en base 2 y no en base 10. Este abuso de lenguaje es sólo para mayor claridad (el lector podrá observar que todo nuestro análisis puede repetirse trabajando en base 2).

Observemos primero que un número real cualquiera, $x \in \mathbb{R}$, $x > 0$, puede escribirse como

$$x = 0, a_1 a_2 \dots a_k \dots 10^l = r \times 10^l, \qquad \frac{1}{10} \leq r < 1 \qquad (\text{es decir}, a_1 \neq 0)$$

Pero en una computadora no se pueden poner infinitos dígitos. Por lo tanto, se trabaja sólo con números de desarrollo finito y de una longitud dada. De la misma forma, el exponente l (es decir el orden del número) estará limitado a cierto rango. En consecuencia los *números de máquina* serán de la forma

$$x = 0, a_1 a_2 ....a_m 10^l = q \times 10^l \qquad -M_1 \leq l \leq M_2, \ , a_1 \neq 0$$

más los correspondientes negativos y el cero. Los números m, M_1 y M_2 dependen de la máquina. Esta representación de los números es la que se llama de punto flotante.

Los *números de máquina* forman un conjunto finito de números racionales. La cantidad de números de máquina que hay entre 1/10 y 1 es,

$$\#\{x \ / \ 1/10 \leq x < 1\} = 9 \times 10^{m-1}.$$

En general, la cantidad que hay entre 10^l y 10^{l+1} es también $9 \times 10^{m-1}$. Esto nos dice que los *números de máquina* no están uniformemente distribuidos. Sí lo está el subconjunto que está entre 10^l y 10^{l+1} para cada l. En particular, los números más grandes están más separados. Resulta útil para la comprensión hacer un dibujo de los números de máquina en la recta numérica (Ejercicio). Al analizar la propagación de errores de redondeo se aclarará por qué esta distribución de los números es más razonable que una uniforme.

Sea $x \in \mathbb{R}$. Para simplificar notación supondremos $x > 0$ (claramente todas las consideraciones que haremos se aplican análogamente a los números negativos). Hay dos posibilidades: que x esté o no en el rango de los *números de máquina*. Si al hacer cálculos aparece un número en la segunda situación, por ser muy grande o muy chico, la computadora nos dará un mensaje indicándolo.

Supongamos entonces que x está en el rango de la máquina, o sea,

$$x = 0, a_1 a_2 ....a_m .....10^l \qquad M_1 \leq l \leq M_2, \ a_1 \neq 0$$

Este x está entre dos números de máquina,

$$x' \leq x \leq x''$$

Supongamos, para simplificar, que $a_m \neq 9$ (analice el lector el caso contrario). Entonces tenemos

$$x' = 0, a_1 a_2 ....a_m 10^l$$

y

$$x'' = 0, a_1 a_2 ....(a_m + 1) 10^l$$

2. Redondeo

Hay dos formas de aproximar a x. Una es por *truncamiento*: se elige siempre x' es decir el mayor de los números de máquina que son menores que x. La otra forma es tomar el más próximo a x entre x' y x''. A esta forma de aproximar se la conoce por *redondeo* y es la que usan habitualmente las computadoras.

Veamos que error cometemos al aproximar un número por redondeo. Usaremos la notación

$$x^* = x \;\; redondeado$$

El error absoluto será

$$|x - x^*| \leq \frac{1}{2}\frac{1}{10^m}10^l$$

Mientras que el error relativo se puede acotar de la forma siguiente

$$\frac{|x - x^*|}{|x|} \leq \frac{1}{2}\,\frac{10^{l-m}}{0, a_1a_2....a_m....10^l}$$

y como

$$0, a_1a_2....a_m.... \geq \frac{1}{10}$$

se tiene que

$$\frac{|x - x^*|}{|x|} \leq 5 \times 10^{-m}$$

Es decir que el error relativo es del orden de 10^{-m} si nuestra máquina trabaja con m dígitos.

Es importante observar que, si bien el error absoluto que introduce el redondeo depende de la magnitud del número, el relativo, que es el más significativo, es independiente de ésta, está controlado en términos de la cantidad de dígitos con la que trabaja nuestra computadora (Ejercicio: meditar que tiene que ver esto con la distribución no uniforme de los *números de máquina*).

Si tenemos

$$x^* = 0, a_1a_2....a_m10^l, \qquad a_1 \neq 0$$

decimos que conocemos a x con m dígitos significativos, lo que equivale, según vimos, a conocerlo con un error relativo del orden de 10^{-m}. Observar la importancia de la condición $a_1 \neq 0$: de lo contrario los dígitos dejan de ser significativos.

Observemos que

$$x^* = x(1+\delta)$$

con

$$|\delta| \leq \varepsilon = 5 \times 10^{-m}$$

Este ε es el *error de redondeo unitario*, es decir, que el error que se comete al aproximar x por x^* es $x\delta$ con $|\delta| \leq \varepsilon$, de esta forma el error $|x - x^*|$ será menor o igual que $\varepsilon|x|$.

Este valor, ε, depende de la máquina y se lo llama *el ε de la máquina*. Recordemos que, según lo comentado antes, el verdadero ε de máquina no es exactamente éste debido a que la computadora trabaja en base 2. Pero desde el punto de vista práctico sólo interesa el orden de ε (y éste sí es correcto!).

Ejercicio: calcular el ε exacto suponiendo que la máquina trabaja en base 2 con k dígitos.

Otra forma de interpretar el significado del ε de la máquina es la siguiente: ε nos dice cuál es el menor número tal que, en la máquina,

$$1 + \varepsilon \neq 1$$

O sea, si se le suma a 1 algo menor que ε, "desaparece" debido al redondeo. En efecto, según la máquina tendremos

$$1 + 4 \times 10^{-m} = 1,0....04 = 0,10....00 \times 10 = 1$$

en cambio,

$$1 + \varepsilon = 1,0....06 = 0,10....01 \times 10 \neq 1$$

Si sumamos exactamente ε el resultado dependerá de como redondee la máquina en el caso en que x equidista de dos *números de máquina*, es decir, cuando la cifra $m+1$ es justo 5.

Más en general, el orden del menor número que sumado a un x da, en la máquina, un resultado distinto de x es $\varepsilon|x|$.

Es fundamental observar que ε no es el menor número que se puede representar en la máquina (y está muy lejos de éste!). Este último depende de M_1 y no de m.

Veamos ahora algunos de los problemas que surgen debido al redondeo.

Empecemos por sumar dos números. Como vimos, si sumamos dos números de ordenes muy distintos, el más chico puede "desaparecer". Por ejemplo, si $m = 5$ y

$$x = 78473000; \qquad y = 24$$

tenemos

$$x^* = 0,78473 \times 10^8; \qquad y^* = 0,24 \times 10^2$$

es decir que x e y son *números de máquina*. Entonces,

$$x + y = 78473024$$

y por lo tanto,

$$(x + y)^* = 0.78473 \times 10^8 = x^* = x$$

En particular, esto nos dice que si tenemos que sumar varios números $x_1, x_2, ...x_N$ conviene hacerlo de menor a mayor (¿Por qué?).

En el ejemplo anterior el problema no es muy importante ya que no se pierden dígitos significativos, es decir, el orden del error relativo no se agranda (se está perdiendo la información de un número que es despreciable respecto del otro sumando).

El problema es mucho más serio cuando deben restarse dos números parecidos. En este caso, debido a lo que se conoce como "cancelación catastrófica", pueden perderse dígitos significativos o, en otras palabras, agrandarse mucho el error relativo.

Consideremos como antes $m = 5$ y tomemos

$$x = 0,372154876; \qquad y = 0,372023264$$

entonces,

$$x^* = 0,37215; \qquad y^* = 0,37202$$

y por lo tanto,

$$x - y = 0,000131612$$

mientras que

$$x^* - y^* = 0,00013 = 0,13 \times 10^{-3}$$

Observemos qué sucedió: x e y estaban representados con 5 dígitos significativos pero al restarlos quedaron sólo 2 del resultado. En consecuencia el error relativo creció de manera tal que si bien el error relativo en x e y era del orden de 10^{-5} el del resultado es del orden de 10^{-2}.

Como conclusión observamos que hay que tratar de evitar restar números "casi iguales". Por ejemplo, supongamos que queremos calcular

$$y = \sqrt{x^2 + 1} - 1$$

para valores de x pequeños. Si lo hacemos directamente, estamos en la situación anterior. En cambio podemos proceder de la siguiente manera:

$$y = \sqrt{x^2 + 1} - 1 = \frac{(\sqrt{x^2 + 1} - 1)(\sqrt{x^2 + 1} + 1)}{(\sqrt{x^2 + 1} + 1)} = \frac{x^2}{(\sqrt{x^2 + 1} + 1)}$$

y utilizar la última expresión para calcular y. Si los cálculos fueran exactos ambas fórmulas darían el mismo resultado pero, debido al redondeo, dan resultados distintos. Por ejemplo, trabajando con 5 dígitos, si $x = 0,0312$ obtenemos con la primera fórmula $y = 0,0004$ (un solo dígito significativo si bien conociamos x exactamente). Mientras que con la segunda, $y = 0,00048662$ (que tiene cuatro dígitos significativos correctos).

El mismo problema surge al calcular $y = x - \sin x$ para x pequeño. En este caso se puede usar el desarrollo de Taylor,

$$y = x - \sin x = \frac{x^3}{3!} - \frac{x^5}{5!} + \frac{x^7}{7!}.....$$

y calcular y sumando los primeros términos de la serie.

Otro caso en el que la cancelación de dígitos significativos puede presentarse es en la resolución de una ecuación cuadrática

$$ax^2 + bx + c = 0$$

si utilizamos la fórmula habitual

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \qquad \text{y} \qquad x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a}$$

Consideremos por ejemplo,

$$x^2 - 10^5 x + 1 = 0$$

Los primeros dígitos exactos de las raíces son

$$x_1 = 99999.99999$$

y

$$x_2 = 0.000010000000001$$

Usando la fórmula para x_2 tenemos

$$x_2 = \frac{10^5 - \sqrt{10^{10} - 4}}{2}$$

Si trabajamos con ocho dígitos el 4 desaparece y x_2 resulta igual a cero. Otra vez hemos restado dos números parecidos!

Esto se puede solucionar calculando primero x_1 y luego obtener x_2 usando que $x_2 = \frac{c}{ax_1}$.

En general, la forma correcta de encontrar las raíces en el caso en que ac sea despreciable respecto de b, es calculando primero

$$x_1 = \frac{-b - \ \mathrm{sign}(b)\sqrt{b^2 - 4ac}}{2a}$$

y luego la otra raiz como hicimos en el ejemplo. De esta forma se evita la pérdida de dígitos significativos.

Un problema fundamental del cálculo numérico es la resolución de sistemas de ecuaciones lineales. Veamos como los errores de redondeo pueden afectar la solución aún en problemas de dos ecuaciones con dos incógnitas. Tratemos de resolver el siguiente sistema utilizando el método de eliminación de Gauss,

$$\begin{pmatrix} \varepsilon & 1 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

Pongamos, a modo de ejemplo, $\varepsilon = 10^{-6}$ y supongamos que la máquina trabaja con cinco dígitos.

Multiplicando la primera fila por $\frac{1}{\varepsilon} = 10^6$ y restándosela a la segunda obtenemos

$$\begin{pmatrix} 10^{-6} & 1 \\ 0 & 1 - 10^6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 2 - 10^6 \end{pmatrix}$$

pero, con el redondeo a cinco cifras nos queda

$$\begin{pmatrix} 10^{-6} & 1 \\ 0 & -10^6 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ -10^6 \end{pmatrix}$$

(perdimos la información de la segunda ecuación!).

Mientras que el resultado exacto es

$$\begin{pmatrix} 10^{-6} & 1 \\ 0 & -999999 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ -999998 \end{pmatrix}$$

Hasta aqui el error no parece grave. Pero veamos: si utilizamos la matriz obtenida con la máquina y despejamos de la segunda ecuación obtenemos la solución $y' = 1$ y luego, reemplazando en la primera, $x' = 0$.

Pero la solución verdadera es

$$y = \frac{1 - 2 \times 10^{-6}}{1 - 10^{-6}} \sim 1 = y'$$
$$x = \frac{1}{1 - 10^{-6}} \neq 0 = x'$$

Observemos que $x - x' = x = \frac{1}{1-10^{-6}}$ es aproximadamente 1. Además el error relativo es 1 (catastrófico!).

Analicemos qué sucedió. Al hacer las restas $1 - \frac{1}{\varepsilon}$, $2 - \frac{1}{\varepsilon}$ se introduce un pequeño error en la matriz triangulada que se propaga a la solución. Este error, al perder la sexta cifra, no es significativo respecto de y pero al reemplazar en la primera ecuación queda,

$$\varepsilon x' = 1 - y', \qquad \text{y entonces} \qquad x = \frac{1}{\varepsilon}(1 - y')$$

Esto implica que el error $y^* - y$ se amplifica por un factor $\frac{1}{\varepsilon}$ dando lugar a un error grande en x.

Veamos ahora que pasa si hacemos el mismo proceso de eliminación de Gauss pero intercambiando las filas de lugar. Queda

$$\begin{pmatrix} 1 & 1 \\ 10^{-6} & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

Operando (fila 2 - ε fila 1), obtenemos

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 - 10^{-6} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 \\ 1 - 2 \times 10^{-6} \end{pmatrix}$$

y con el redondeo a cinco cifras nos queda

$$\begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

que tiene como solución $y' = 1$, $x' = 1$.

¿Qué pasó? El intercambio de filas permitió obtener un resultado correcto evitando la propagación catastrófica del error que se daba en el otro caso. Veremos más adelante que esto es algo general: conviene elegir como "pivote" (elemento por el que se divide) para la eliminación de Gauss el que tenga mayor valor absoluto.

En este ejemplo, la primera forma de resolver era un algoritmo "malo" o inestable en el sentido de que amplificaba los errores llevando a resultados absolutamente erróneos. Sin embargo, esto

se solucionó intercambiando el orden de las filas, o sea, modificando el algoritmo. Esto muestra que el error en el primer caso se debía a la forma de resolver y no a algo inherente al problema en sí mismo.

Hay casos de naturaleza esencialmente diferente en los cuales el problema que se quiere resolver está "mal condicionado". Esto significa que pequeños cambios en los datos implican grandes cambios en la solución. Esto hace que los errores de redondeo puedan amplificarse mucho independientemente del método que usemos para resolverlo.

Veamos un ejemplo de esto. Supongamos que nuestra máquina trabaja con 3 dígitos y trunca. Resolvamos el sistema

$$\begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.217 \\ 0.254 \end{pmatrix}$$

La solución exacta es $x = 1$, $y = -1$.

Teniendo en cuenta lo visto antes, intercambiamos filas antes de hacer la eliminación de Gauss. Obtenemos

$$\begin{pmatrix} 0.913 & 0.659 \\ 0 & 0.001 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 0.254 \\ 0.001 \end{pmatrix}$$

y en consecuencia $y' = 1$ y $x' = -0.443$ que no tiene nada que ver con la solución exacta. En particular, el error es mayor que la solución misma!

Lo que sucede en este ejemplo es que la matriz está "mal condicionada" (más adelante precisaremos lo que esto significa) y habrá problemas independientemente del algoritmo que utilicemos.

Otro ejemplo de problema "mal condicionado" es el siguiente. Las raíces de

$$(x-2)^2 = 10^{-6}$$

son

$$x_1 = 2 + 10^{-3} \qquad x_2 = 2 - 10^{-3}$$

en cambio, las raíces de

$$(x-2)^2 = 0$$

son $x_1 = x_2 = 2$.

Este ejemplo trivial muestra que un pequeño cambio en un coeficiente de la ecuación polinomial puede dar lugar a un cambio de otro orden en las raíces. En este caso, un cambio de 10^{-6} en el término independiente origina un cambio de 10^{-3} en las raíces.

Un ejemplo más interesante es el estudiado por Wilkinson en 1963. Se trata de calcular las raíces de

$$p(x) = (x-1)(x-2).....(x-19)(x-20) = x^{20} - 210x^{19} + ...$$

Wilkinson demostró que al cambiar el coeficiente -210 por $-210 - 2^{-23}$ las raíces 16 y 17 se transforman en el par complejo

$$16.73... + i2.812... \qquad 16.73... - i2.812...$$

Para finalizar, veamos otro ejemplo de algoritmo inestable. El problema consiste en calcular

$$E_n = \int_0^1 x^n e^{x-1}\, dx \qquad n = 1, 2, ...$$

Integrando por partes se obtiene

$$E_n = \int_0^1 x^n e^{x-1}\, dx = x^n e^{x-1} \,|_0^1 - \int_0^1 n x^{n-1} e^{x-1}\, dx$$

es decir

$$E_n = 1 - nE_{n-1}$$

y es fácil ver que

$$E_1 = 1/e$$

con lo que tenemos definida la sucesión E_n en forma recursiva.

Usando esta relación recursiva para calcular con una máquina que trabaje con seis dígitos se obtiene,

$$\begin{gathered} E_1 \sim 0.367879 \\ E_2 \sim 0.264242 \\ E_3 \sim 0.207274 \\ \vdots \\ E_9 \sim -0.0684800 \end{gathered}$$

cuando en realidad

$$E_9 \sim 0.0916$$

con lo que el resultado computacional es pésimo.

En este caso lo que sucede es que el error de E_{n-1} se amplifica multiplicándose por n. Entonces en nueve pasos se multiplica por 9!, obteniéndose un error de

$$9! \times \text{ error inicial } = 9! \times 4.412 \times 10^{-7} \sim 0.1601$$

que resulta mayor que el verdadero E_9.

Como conclusión el algoritmo es malo. Pero observemos que no lo es el problema en sí mismo. Como alternativas podemos calcular E_n por integración numérica o bien hacer el siguiente truco. Observemos que

$$E_{n-1} = \frac{1 - E_n}{n}$$

y como

$$E_n \leq \int_0^1 x^n \, dx = \frac{1}{n+1} \to 0$$

podemos empezar de $E_{20} \sim 0$ e ir hacia atrás usando $E_{n-1} = \frac{1-E_n}{n}$. Este algoritmo es estable (el error en cada paso se multiplica por algo menor que uno).

Como conclusión, los ejemplos analizados en esta sección muestran la diferencia entre el caso en el cual la amplificación de los errores de redondeo se debe a que el problema está "mal condicionado" o "mal planteado" y el caso en el que dicha amplificación se debe al uso de un "algoritmo inestable". Es fundamental distinguir entre ambos casos y, por otra parte, encontrar las causas de la propagación indebida de errores con el objeto de mejorar los algoritmos.

3. Ejercicios

(1) Utilizando el método de redondeo:
   (a) Hallar el número de máquina más próximo a 125.6 y a= 126 si trabaja con
      - Base 10 y mantisa de 2 dígitos.
      - Base 2 y mantisa de 8 dígitos.
   (b) Verificar para $x = 125.6$, la conocida cota para el error relativo

$$\left|\frac{x - fl(x)}{x}\right| \leq \epsilon$$

   si $\epsilon = 1/2\beta^{1-d}$ donde β es la base y d la longitud de la mantisa.
   (c) ¿Cuál es, en cada caso, el valor que da la máquina como resultado de las operaciones $126 + 125.6$ y $126 - 125.6$? ¿Cuál es el error relativo de estos resultados?

(2) Utilizando el método de truncamiento:
   (a) Rehacer el Ejercicio 1, con el ϵ correspondiente, es decir: $\epsilon = \beta^{-d+1}$, donde β y d son como antes.
   (b) Demostrar que, en este caso, ϵ es el menor número de máquina tal que $1 + \epsilon \neq 1$. ¿Cuánto da $\beta + \epsilon$?

(3) Mostrar que $fl(x)$ tiene (para ambos métodos) una escritura de la forma

$$fl(x) = x(1 + \delta_x)$$

   donde $|\delta_x| \leq \epsilon$. (Usar la cota para el error relativo).

(4) Pérdida de dígitos significativos:
   (a) Si $x, y \geq 0$ demostrar que

$$\left|\frac{x + y - fl(fl(x) + fl(y))}{x + y}\right| \leq 2\epsilon + \epsilon^2.$$

Observar que en la expresión $2\epsilon + \epsilon^2$ el valor de ϵ^2 es despreciable dado que ϵ es pequeño.

(b) Si x e y no poseen el mismo signo, ¿puede repetir la misma cuenta? (Sugerencia: recordar el error relativo de $126 - 125.6$ en el ejercicio 1, item (c), utilizando la computadora binaria con mantisa de 8 dígitos.)

(5) Un ejemplo que muestra que algunas de las reglas de la aritmética no son válidas para operaciones de punto flotante.
   (a) Intentar anticipar el resultado de los siguientes cálculos:
      (i) $(1 + \frac{\epsilon}{2}) + \frac{\epsilon}{2}$ (ii) $1 + (\frac{\epsilon}{2} + \frac{\epsilon}{2})$
      (iii) $\big((1 + \frac{\epsilon}{2}) + \frac{\epsilon}{2}\big) - 1$ (iv) $\big(1 + (\frac{\epsilon}{2} + \frac{\epsilon}{2})\big) - 1$
   (b) Efectuar estos cálculos usando `Matlab` y comprobar las predicciones hechas.

(6) Hallar la raíz menor en módulo de la ecuación

$$x^2 - 40x + 0.25 = 0,$$

utilizando aritmética de 4 dígitos y comparar con el resultado obtenido utilizando aritmética exacta. Calcular el error relativo y asegurarse de comprender de dónde viene la pérdida de dígitos significativos. ¿Se le ocurre cómo calcular con mayor precisión dicha raíz? ¿Cuál es el error relativo con el nuevo método?

(7) Hallar una forma de calcular sin pérdida de dígitos significativos las siguientes cantidades, para $x \sim 0$:
   (a) $(\alpha + x)^n - \alpha^n$
   (b) $\alpha - \sqrt{\alpha^2 - x}$
   (c) $\cos x - 1$
   (d) $\operatorname{sin}(\alpha + x) - \operatorname{sin}(\alpha)$

(8) Se pretende calcular las sumas $S_N = \sum_{k=1}^{N} a_k$ con $N \in \mathbb{N}$. Llamemos $\widehat{S_N}$ al valor calculado que se logra de hacer $fl(\widehat{S_{N-1}} + a_N)$.
   (a) $S_N = \displaystyle\sum_{k=1}^{N} \frac{1}{k}$. Mostrar que $\widehat{S_N}$ se estaciona a partir de algún N suficientemente grande. Deducir que a partir de entonces $S_N \neq \widehat{S_N}$.
   (b) Idem (a) para la suma $S_N = \displaystyle\sum_{k=1}^{N} \frac{2^{-k+100} + 1}{k}$. Encontrar, haciendo un programa en `Matlab`, el valor de N para el cual $\widehat{S_N}$ se estaciona.

(9) El desarrollo de Taylor de la función e^x proporciona una forma muy inestable de calcular este valor cuando x es negativo. Hacer un programa en `Matlab` que estime e^{-12} evaluando el desarrollo de Taylor hasta grado n de la función e^x en $x = -12$, para $n = 1, \ldots, 100$. Comparar con el valor exacto: $0.000006144212353328210\ldots$ ¿Cuáles son las principales fuentes de error? Realizar otra estimación= de e^{-12} con algún otro método que evite los problemas del método anterior (Sugerencia: Considerar $e^{-x} = 1/e^x$).

(10) Calcular en `Matlab` los valores: $\operatorname{sen}(\pi/2 + 2\pi 10^j)$ c= on $1 \leq j \leq 18$. ¿Cuánto debería dar? ¿Qué está pasando?

(11) Aproximación de la derivada de una función.

(a) Llamamos derivada discreta de f en $x = 1$ al valor

$$d_h f(1) = \frac{f(1+h) - f(1)}{h}.$$

Utilizando el desarrollo de Taylor, demostrar que

$$|f'(1) - d_h f(1)| \leq |f''(1)|\frac{h}{2} + o(h) \qquad (h \to 0)$$

siempre que f sea suficientemente derivable.

(b) Considerar la función $f(x) = x^2$. Hacer un programa en `Matlab` que calcule los valores de $d_h f(1)$ para aproximar $f'(1)$, dándole a h los valores 10^{-18}, $10^{-17.9}$, $10^{-17.8}, \ldots, 10^{-1}$ y grafique los resultados obtenidos. Decidir si éstos se contradicen con el resultado del ítem anterior. Hacer un análisis de los cálculos efectuados para calcular $d_h f(1)$, teniendo en cuenta que la máquina utiliza aritmética de punto flotante.

(c) Repetir el ítem anterior, dándole otros valores a h, de modo que el resultado resulte más confiable.

(12) Las funciones de Bessel J_n se pueden definir del siguiente modo:

$$J_n(x) = \frac{1}{\pi}\int_0^\pi \cos(x\sin\theta - n\theta)d\theta.$$

y verifican que $|J_n(x)| \leq 1$. Se sabe además que $J_{n+1}(x) = 2n/xJ_n(x) - J_{n-1}(x)$. Con los valores estimados $J_0(1) \sim 0.7651976865$, $J_1(1) \sim 0.4400505857$ y la recurrencia dada, hacer un programa en `Matlab` para calcular $J_2(1)$, $J_3(1)$, $\ldots$, $J_{10}(1)$. Decidir si la condición $|J_n(x)| \leq 1$ deja de satisfacerse. ¿Qué esta sucediendo?

(13) Dada la función $\Phi : \mathbb{R} \to \mathbb{R}$ definida por

$$\Phi(x) = \sum_{k=1}^{\infty} \frac{1}{k(k+x)},$$

consideramos las siguiente dos maneras de estimar numéricamente el valor de $\Phi(x)$ para un x fijo:

- sumar los primeros n términos de la serie $\Phi(x)$,
- teniendo en cuenta que $\Phi(1) = 1$, definir

$$\Psi(x) = \Phi(x) - \Phi(1) = \sum_{k=1}^{\infty}\left(\frac{1}{k(k+x)} - \frac{1}{k(k+1)}\right) = \sum_{k=1}^{\infty}\frac{1-x}{k(k+1)(k+x)},$$

luego expresar $\Phi(x) = 1 + \Psi(x)$ y, de este modo, estimar $\Phi(x)$ como 1 más la suma de los primeros n términos de la serie $\Psi(x)$.

Predecir cuál de las dos maneras converge más rápidamente. Luego, hacer un programa que calcule y grafique el resultado obtenido con los dos métodos propuestos para calcular $\Phi(0)$, con $n = 1, \ldots, 100$. Comparar con el resultado exacto, que es $\frac{\pi^2}{6}$.

(14) Algoritmo para calcular π.

Comenzar inicializando las variables a, b, c, d y e del siguiente modo: $a = 0$, $b = 1$, $c = 1/\sqrt{2}$, $d = 1/4$, $e = 1$. Luego, iterar n veces en el orden dado las siguientes fórmulas:

$$a = b, \quad b = \frac{b+c}{2}, \quad c = \sqrt{ca}, \quad d = d - e(b-a)^2, \quad e = 2e.$$

Finalmente, el valor de π puede estimarse como $f = b^2/d$, o como $g = (b+c)^2/(4d)$.

Hacer un programa que calcule los valores de π estimados por f y g cuando $n = 1, 2, \ldots, 10$. ¿Qué estimación converge más rápido? ¿Cuán precisos son sus resultados? El valor de π correcto hasta 36 dígitos es

$$\pi = 3.141592653589793238462643383279502 88$$

CAPíTULO 2

Normas y condicionamiento de una matriz.

Consideramos el sistema de n ecuaciones con n incógnitas

$$Ax = b$$

con $A \in \mathbb{R}^{n\times n}$, $x \in \mathbb{R}^n$ y $b \in \mathbb{R}^n$ y nos preguntamos cuánto afectará a la solución un error en el dato b. Para poder dar una respuesta debemos decidir primero cómo medir el error. Es decir, necesitamos dar alguna forma de medir vectores de $\mathbb{R}^n$. Una forma posible es utilizar la longitud o norma euclídea del vector, o sea,

$$\|x\|_2 = \sqrt{x_1^2 + ... + x_n^2}$$

Pero ésta no es la única medida razonable y en muchos casos es conveniente trabajar con otras. Por ejemplo, podemos decir que un vector es "chico" si lo son todas sus componentes y tomar entonces como medida de x la siguiente, llamada "norma infinito",

$$\|x\|_\infty = \max_{1\le i\le n} |x_i|$$

Otra elección natural es la "norma uno",

$$\|x\|_1 = |x_1| + ... + |x_n|$$

o más en general la "norma p",

$$\|x\|_p = (|x_1|^p + ... + |x_n|^p)^{\frac{1}{p}}$$

con $1 \le p < \infty$.

Todas estas formas de medir resultan equivalentes en el sentido de que, si x es "chico" en una de las normas entonces lo es en cualquier otra, puesto que una norma mayora a la otra salvo una constante que depende sólo de n. Por ejemplo, utilizando la desigualdad de Schwartz se obtiene

$$\|x\|_1 \le \sqrt{n}\|x\|_2$$

y por otra parte, es fácil ver que,

$$\|x\|_2 \le \|x\|_1$$

También se verifica fácilmente que

$$\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$$

Más en general, decimos que una norma en $\mathbb{R}^n$ es una manera de asignar a cada x un número $\|x\|$ de tal forma que se verifiquen las siguientes propiedades, análogas a las que cumple la longitud usual,

1) $\|x\| \geq 0 \qquad \forall x \in \mathbb{R}^n$

2) $\|x\| = 0$ si y solo si $x = 0$.

3) $\|\lambda x\| = |\lambda|\|x\| \qquad \forall \lambda \in \mathbb{R}, \forall x \in \mathbb{R}^n$

4) $\|x + y\| \leq \|x\| + \|y\| \qquad \forall x \in \mathbb{R}^n, \forall y \in \mathbb{R}^n$ (desigualdad triangular)

Una vez que sabemos como medir vectores podemos hablar también de la distancia entre dos vectores x e y la cual está dada por $\|x - y\|$. En particular, esto permite hablar de convergencia de sucesiones: $x_n \to x$ si $\|x - x_n\| \to 0$.

Tanto para medir el error como para analizar la convergencia de una sucesión elegiremos la norma que nos resulte más conveniente en cada caso. Esto está justificado por el hecho de que todas las normas son equivalentes: convergencia en una de ellas implica convergencia en cualquier otra. Más aún, se tiene el siguiente resultado.

Teorema 2.1. *Dadas dos normas en $\mathbb{R}^n$, $\|\ \|$ y $\|\ \|'$, existen constantes C_1 y C_2 que dependen sólo de n y de las normas consideradas (en particular, son independientes de x) tales que*

$$C_1\|x\| \leq \|x\|' \leq C_2\|x\| \qquad \forall x \in \mathbb{R}^n$$

Demostración. Basta ver que una norma cualquiera $\|\ \|$ es equivalente a la norma euclídea usual, $\|\ \|_2$. Sea $\{e_i\}$ la base canónica de $\mathbb{R}^n$ y definamos la constante $C = (\sum_{i=1}^n \|e_i\|^2)^{1/2}$, la cual depende sólo de n y de $\|\ \|$. Utilizando las propiedades de la norma y la desigualdad de Schwartz obtenemos

$$\|x\| = \|\sum_{i=1}^n x_i e_i\| \leq \sum_{i=1}^n |x_i|\|e_i\| \leq (\sum_{i=1}^n |x_i|^2)^{1/2}(\sum_{i=1}^n \|e_i\|^2)^{1/2} = C\|x\|_2 \tag{2.1}$$

Queremos ver ahora que existe una constante K tal que

$$\|x\|_2 \leq K\|x\| \qquad \forall x \in \mathbb{R}^n \tag{2.2}$$

Supongamos que una tal K no existe y veamos que se llega a una contradicción. En efecto, si (2.2) no se cumple para ningún K tenemos, en particular, que dado $m \in \mathbb{N}$, existe $y_m \in \mathbb{R}^n$ tal que

$$\|y_m\|_2 \geq m\|y_m\|$$

y llamando $x_m = y_m/\|y_m\|_2$ obtenemos $\|x_m\|_2 = 1$ y

$$\|x_m\| \leq \frac{1}{m} \tag{2.3}$$

pero, toda sucesión acotada en la norma euclídea tiene una subsucesión convergente. Entonces existe una subsucesión de (x_m), (x'_m) tal que

$$\|x'_m - x\|_2 \to 0$$

para cierto $x \in \mathbb{R}^n$. Pero entonces por (2.1), también vale que

$$\|x'_m - x\| \to 0$$

Por otra parte, por (2.3) tenemos que $\|x'_m\| \to 0$ y en consecuencia, por unicidad del límite, resulta $x = 0$. Pero observemos finalmente que, por la desigualdad triangular,

$$\Big|\|x'_m\|_2 - \|x\|_2\Big| \leq \|x'_m - x\|_2$$

y entonces se llega a la contradicción $1 = \|x'_m\|_2 \to \|x\|_2 = 0$, finalizando la demostración. □

Ahora sí, estamos en condiciones de abordar el problema de cómo afecta el error en los datos a la solución de un sistema lineal cuya matriz A es inversible. Si se reemplaza el dato b por $b+\Delta b$, la solución x del sistema será modificada de tal forma que tendremos

$$A(x + \Delta x) = (b + \Delta b)$$

o equivalentemente,

$$A\Delta x = \Delta b$$

y nos preguntamos que relación hay entre

$$\frac{\|\Delta x\|}{\|x\|} \qquad \text{y} \qquad \frac{\|\Delta b\|}{\|b\|}$$

Veamos primero el siguiente ejemplo simple,

$$\begin{pmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4.1 \\ 9.7 \end{pmatrix}$$

La solución es

$$x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Observemos que

$$\|b\|_1 = 13.8 \qquad \|x\|_1 = 1$$

Si modificamos b poniendo

$$b' = b + \Delta b = \begin{pmatrix} 4.11 \\ 9.70 \end{pmatrix}$$

entonces la solución es

$$x' = x + \Delta x = \begin{pmatrix} 0.34 \\ 0.97 \end{pmatrix}$$

y se obtiene en consecuencia

$$\|\Delta b\|_1 = 0.01 \qquad \|\Delta x\|_1 = 1.63$$

con lo que el error relativo se amplificó mucho, en efecto,

$$\frac{\|\Delta b\|_1}{\|b\|_1} = 0.7246 \times 10^{-3}$$

mientras que

$$\frac{\|\Delta x\|_1}{\|x\|_1} = 1.63$$

Nuestro objetivo es tratar de entender a qué se debe este comportamiento y poder predecir, dada una matriz A, cuál será el factor de amplificación del error relativo o, al menos, dar una cota de éste en términos de A.

Analicemos primero el caso de una matriz diagonal.

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

La solución es (si $\lambda_1, \lambda_2 \neq 0$)

$$x_1 = \frac{b_1}{\lambda_1} \qquad x_2 = \frac{b_2}{\lambda_2}$$

Por ejemplo, si

$$A = \begin{pmatrix} 1000 & 0 \\ 0 & \frac{1}{100} \end{pmatrix}$$

entonces,

$$x_1 = \frac{b_1}{1000} \qquad x_2 = 100 b_2$$

Si ponemos $b' = b + \Delta b$ con

$$\Delta b = \begin{pmatrix} 0 \\ \Delta b_2 \end{pmatrix} \qquad b = \begin{pmatrix} b_1 \\ 0 \end{pmatrix}$$

entonces,

$$x_1 = \frac{b_1}{1000} \qquad \Delta x_2 = 100 \Delta b_2$$

obteniéndose

$$10^5 \frac{\|\Delta b\|_2}{\|b\|_2} = \frac{\|\Delta x\|_2}{\|x\|_2}$$

es decir que el error relativo se multiplicó por 10^5.

Si en cambio elegimos

$$\Delta b = \begin{pmatrix} \Delta b_1 \\ 0 \end{pmatrix} \qquad b = \begin{pmatrix} 0 \\ b_2 \end{pmatrix}$$

tenemos entonces,

$$\Delta x_1 = \frac{1}{1000} \Delta b_1 \qquad x_2 = 100 b_2$$

y en consecuencia,

$$\frac{1}{10^5} \frac{\|\Delta b\|_2}{\|b\|_2} = \frac{\|\Delta x\|_2}{\|x\|_2}$$

o sea que en este caso el error relativo se redujo en un factor 10^5. En general, para una matriz diagonal

$$A = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$$

con $|\lambda_1| > |\lambda_2|$, el error relativo puede multiplicarse por

$$\frac{|\lambda_1|}{|\lambda_2|}$$

en el peor de los casos y por

$$\frac{|\lambda_2|}{|\lambda_1|}$$

en el mejor de los casos.

En general, el error tendrá componentes en cualquier dirección por lo que es de esperar que si $\frac{|\lambda_1|}{|\lambda_2|}$ es grande los errores relativos se amplifiquen.

El mismo análisis puede hacerse en $\mathbb{R}^n$. Si A es una matriz diagonal

$$A = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ & \cdots & \\ 0 & \cdots & \lambda_N \end{pmatrix}$$

el error relativo se puede amplificar, a lo sumo, por un factor

$$\frac{|\lambda_{\max}|}{|\lambda_{\min}|}$$

siendo $\lambda_{\max}$ y $\lambda_{\min}$ los de máximo y mínimo valor absoluto entre los λ_j (observemos que $\lambda_{\min} \neq 0$ pues estamos suponiendo que A es inversible). Este cociente se llama número de condición o de condicionamiento de A en la norma $\| \ \|_2$ y lo denotaremos $Cond_2(A)$.

Ahora veamos como definir el número de condición para una matriz A arbitraria. Comencemos por el caso en que A sea simétrica, es decir $a_{ij} = a_{ji}$. En este caso A se puede diagonalizar, es decir, existe una base de autovectores $\{v_1, ..., v_n\}$. Además, por ser A simétrica podemos considerar que la base es ortonormal. Entonces, si $Ax = b$, $A(x + \Delta x) = b + \Delta b$ y

$$x = \sum_{i=1}^{n} \alpha_i v_i \qquad \Delta x = \sum_{i=1}^{n} \beta_i v_i$$

tenemos,

$$\|x\|_2^2 = \sum_{i=1}^{n} \alpha_i^2 \qquad \|\Delta x\|_2^2 = \sum_{i=1}^{n} \beta_i^2$$

y además, si llamamos λ_i al autovalor correspondiente a v_i,

$$b = \sum_{i=1}^{n} \alpha_i \lambda_i v_i \qquad \Delta b = \sum_{i=1}^{n} \beta_i \lambda_i v_i$$

y en consecuencia,

$$\|b\|_2^2 = \sum_{i=1}^{n} \alpha_i^2 \lambda_i^2 \qquad \|\Delta b\|_2^2 = \sum_{i=1}^{n} \beta_i^2 \lambda_i^2$$

Entonces, si $\lambda_{\max}$ y $\lambda_{\min}$ son los autovalores de máximo y mínimo valor absoluto respectivamente, obtenemos

$$\frac{\|\Delta x\|_2^2}{\|x\|_2^2} = \frac{\sum_{i=1}^{n} \beta_i^2}{\sum_{i=1}^{n} \alpha_i^2} \leq \frac{1/|\lambda_{\min}|^2}{1/|\lambda_{\max}|^2} \frac{\|\Delta b\|_2^2}{\|b\|_2^2}$$

o sea

$$\frac{\|\Delta x\|_2}{\|x\|_2} \leq \frac{|\lambda_{\max}|}{|\lambda_{\min}|} \frac{\|\Delta b\|_2}{\|b\|_2}$$

es decir que el número $Cond_2(A) = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$ es una cota para el factor de amplificación del error relativo. Más aún, esta cota es la mejor posible pues la desigualdad se convierte en una igualdad para cierta elección de b y Δb (b en la dirección correspondiente al máximo autovalor y Δb en la correspondiente al mínimo).

Para generalizar el número de condición a cualquier matriz observemos que, en el caso de una simétrica, la dirección correspondiente al autovalor de máximo valor absoluto nos da la dirección de "máxima expansión", es decir, si miramos el cociente entre la longitud de Ax y la de x

$$\frac{\|Ax\|_2}{\|x\|_2}$$

éste será máximo entre todos los x cuando x está en la dirección correspondiente a $\lambda_{\max}$. En efecto, si escribimos x en la base de autovectores $\{v_1, ..., v_n\}$

$$x = \sum_{i=1}^{n} \alpha_i v_i$$

entonces,

$$Ax = \sum_{i=1}^{n} \alpha_i \lambda_i v_i$$

y de acá resulta que

$$\|Ax\|_2 \leq |\lambda_{\max}| \|x\|_2$$

y tomando $x = v_j$ con $\lambda_j = \lambda_{\max}$ se ve que se verifica la igualdad.

Análogamente, la dirección de "mínima expansión" corresponde a la asociada a $\lambda_{\min}$, la cual corresponde también a la de "máxima expansión" de la inversa de A.

El análisis realizado para matrices simétricas nos muestra que el factor de amplificación del error relativo está relacionado con los máximos factores de expansión de A y de su inversa. Teniendo en cuenta esto, definimos para una matriz arbitraria $A \in \mathbb{R}^{n \times n}$ y una norma de vectores $\| \ \|$ cualquiera, la norma matricial asociada como

$$\|A\| = \max_{0 \neq x \in \mathbb{R}^n} \frac{\|Ax\|}{\|x\|}$$

Es decir, la norma de A nos da lo máximo que "agranda" el multiplicar por A medido en la norma de vectores dada. Es fácil ver que

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

y en particular, esto muestra que el máximo existe, o sea que la norma esta bien definida ($\|Ax\|$ es una función continua de x y por lo tanto, alcanza su máximo en el conjunto de vectores de norma igual a uno pues éste es cerrado y acotado).

De la definición se desprende la siguiente desigualdad que usaremos frecuentemente,

$$\|Ax\| \leq \|A\|\|x\| \qquad \forall x \in \mathbb{R}^n$$

valiendo la igualdad para algun x. También es fácil ver que

$$\|AB\| \leq \|A\|\|B\| \qquad \forall A \in \mathbb{R}^{n\times n}, \forall B \in \mathbb{R}^{n\times n} \tag{2.4}$$

Por otra parte puede verificarse que $\|A\|$ es la menor entre todas las constantes C para las cuales vale la desigualdad

$$\|Ax\| \leq C\|x\| \qquad \forall x \in \mathbb{R}^n$$

siendo ésta otra forma usual de definir la norma matricial.

Como ejemplo tenemos que, por lo visto antes, si A es simétrica entonces

$$\|A\|_2 = |\lambda_{\max}|$$

donde el subíndice 2 nos indica cual es la norma de vectores correspondiente.

Análogamente tenemos que para A inversible y simétrica

$$\|A^{-1}\|_2 = \frac{1}{|\lambda_{\min}|}$$

y por lo tanto,

$$\|A\|_2\|A^{-1}\|_2 = \frac{|\lambda_{\max}|}{|\lambda_{\min}|}$$

En general introducimos entonces la siguiente

Definición 2.2. *Sea $A \in \mathbb{R}^{n\times n}$ una matriz inversible y sea $\|\ \|$ una norma en $\mathbb{R}^n$ definimos el número de condición de A como*

$$Cond(A) = \|A\|\|A^{-1}\|$$

Es claro que $Cond(A)$ depende de la norma de vectores elegida.

Es fácil ver que valen las siguientes propiedades,

$$Cond(A) = Cond(A^{-1})$$

y

$$Cond(A) \geq 1 \qquad \forall A \in \mathbb{R}^{n\times n}$$

En efecto, la primera es obvia mientras que, para ver la segunda, utilizamos la propiedad (2.4) y obtenemos

$$1 = \|I\| = \|AA^{-1}\| \leq \|A\|\|A^{-1}\| = Cond(A)$$

Podemos ahora probar el siguiente resultado fundamental

TEOREMA 2.3. *Si $A \in \mathbb{R}^{n\times n}$ es inversible, b, $\Delta b \in \mathbb{R}^n$, $Ax = b$ y $A(x+\Delta x) = b+\Delta b$ entonces,*

$$\frac{\|\Delta x\|}{\|x\|} \leq Cond(A)\frac{\|\Delta b\|}{\|b\|} \tag{2.5}$$

valiendo la igualdad para alguna elección de b y Δb.

Además,

$$\frac{1}{Cond(A)}\frac{\|\Delta b\|}{\|b\|} \leq \frac{\|\Delta x\|}{\|x\|} \tag{2.6}$$

y nuevamente, vale la igualdad para ciertos b y Δb.

Demostración. Se tiene que

$$A(\Delta x) = \Delta b$$

y entonces,

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\Delta b\|}{\|x\|} = \frac{\|A^{-1}\|\|\Delta b\|}{\|b\|}\frac{\|b\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\Delta b\|}{\|b\|}\|A\|$$

donde para la última desigualdad hemos usado que $\|b\| = \|Ax\| \leq \|A\|\|x\|$. Por lo tanto (2.5) vale.

Observemos además que si elegimos Δb tal que $\|\Delta x\| = \|A^{-1}\Delta b\| = \|A^{-1}\|\|\Delta b\|$, x tal que $\|Ax\| = \|A\|\|x\|$ (lo que siempre se puede por la definición de la norma matricial) y $b = Ax$ resulta la igualdad en (2.5).

Ahora, para ver la desigualdad (2.6) observemos que ésta puede escribirse como

$$\frac{\|\Delta b\|}{\|b\|} \leq Cond(A)\frac{\|\Delta x\|}{\|x\|}$$

la cual, teniendo en cuenta que $Cond(A) = Cond(A^{-1})$, es exactamente la desigualdad (2.5) aplicada a A^{-1} con lo que el teorema está demostrado. □

Veamos ahora que el número de condición también tiene que ver con la propagación del error que se cometa en los coeficientes del sistema. Como veremos más adelante, el teorema siguiente es también de suma importancia en el análisis del error de redondeo en la eliminación de Gauss.

THEOREMA 2.4. *Si $A \in \mathbb{R}^{n\times n}$ es inversible, $E \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$, $Ax = b$ y $(A+E)(x+\Delta x) = b$ entonces, llamando $\tilde{x} = x + \Delta x$ tenemos*

$$\frac{\|\Delta x\|}{\|\tilde{x}\|} \leq Cond(A)\frac{\|E\|}{\|A\|} \tag{2.7}$$

y si

$$Cond(A)\frac{\|E\|}{\|A\|} \leq \delta < 1$$

entonces,

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{1}{1-\delta}Cond(A)\frac{\|E\|}{\|A\|} \tag{2.8}$$

Demostración. Tenemos

$$Ax = b \qquad A\tilde{x} = b - E\tilde{x}$$

y entonces,

$$-E\tilde{x} = A\Delta x$$

por lo que concluímos que

$$\|\Delta x\| \leq \|A^{-1}\|\|E\|\|\tilde{x}\| \leq Cond(A)\frac{\|E\|}{\|A\|}\|\tilde{x}\|$$

lo que prueba (2.7).

Ahora observemos que

$$\frac{\|\tilde{x}\|}{\|x\|} = \frac{\|x+\Delta x\|}{\|x\|} \leq \frac{\|x\| + \|\Delta x\|}{\|x\|} = 1 + \frac{\|\Delta x\|}{\|x\|}$$

lo cual, junto con la desigualdad anterior implica que

$$\frac{\|\Delta x\|}{\|x\|} \leq Cond(A)\frac{\|E\|}{\|A\|}\left(1+\frac{\|\Delta x\|}{\|x\|}\right) \leq Cond(A)\frac{\|E\|}{\|A\|}+\delta\frac{\|\Delta x\|}{\|x\|}$$

lo que concluye la demostración de (2.8). □

Veamos ahora como calcular algunas normas matriciales. Dada $A \in \mathbb{R}^{n\times n}$ se llama radio espectral de A a

$$\rho(A) = |\lambda_{\max}|$$

siendo $\lambda_{\max}$ el de máximo valor absoluto entre todos los autovalores de A, incluyendo los complejos.

Ya vimos que si A es simétrica entonces,

$$\|A\|_2 = \rho(A)$$

En general, para $A \in \mathbb{R}^{n\times n}$ arbitraria se tiene

$$\|A\|_2 = \sqrt{\rho(A^TA)}$$

donde A^T es la matriz traspuesta de A. En efecto, como A^TA es simétrica, existe una base ortonormal de autovectores $\{v_j\}$. Llamando μ_j a los autovalores correspondientes tenemos

$$A^TAv_j = \mu_j v_j \qquad j = 1, ..., n$$

y si $x = \sum_{i=1}^n \alpha_i v_i$ entonces, por la ortonormalidad de los v_j resulta $\|x\|_2^2 = \sum_{i=1}^n \alpha_j^2$ y en consecuencia, teniendo en cuenta que $A^TAx = \sum_{i=1}^n \alpha_j\mu_j v_j$, se tiene que para todo $x \in \mathbb{R}^n$

$$\frac{\|Ax\|_2^2}{\|x\|_2^2} = \frac{x^TA^TAx}{\|x\|_2^2} = \frac{\sum_{i=1}^n \alpha_j^2\mu_j}{\sum_{i=1}^n \alpha_j^2} \leq \rho(A^TA)$$

es decir que

$$\|A\|_2 \leq \sqrt{\rho(A^TA)}$$

y tomando $x = v_j$ con $\mu_j = \mu_{\max}$ se ve que vale la igualdad.

El cálculo de la norma 2 de una matriz involucra el cálculo de autovalores, el cual es un problema complicado. Sin embargo, otras normas son mucho más fáciles de calcular. Por ejemplo, se tiene que

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

Para ver esto observemos primero que, para todo $x \in \mathbb{R}^n$ vale

$$\|Ax\|_\infty = \max_{1\leq i\leq n} | \sum_{j=1}^{N} a_{ij}x_j | \leq \|x\|_\infty \max_{1\leq i\leq n} \sum_{j=1}^{N} |a_{ij}|$$

y entonces,

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

Para ver que vale la otra desigualdad, sea k tal que $\sum_{j=1}^{n} |a_{kj}|$ es máxima y tomemos

$$x = \begin{pmatrix} sg(a_{k1}) \\ sg(a_{k2}) \\ \cdots \\ sg(a_{kn}) \end{pmatrix}$$

donde $sg(a) = 1$ si $a \geq 0$ y $sg(a) = -1$ si $a < 0$. Entonces,

$$(Ax)_k = \sum_{j=1}^{n} |a_{kj}|$$

y en particular, $\|Ax\|_\infty \geq \sum_{j=1}^{n} |a_{kj}|$ y como $\|x\|_\infty = 1$ obtenemos

$$\frac{\|Ax\|_\infty}{\|x\|_\infty} \geq \sum_{j=1}^{n} |a_{kj}|$$

y concluímos que

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

De manera similar puede verse, aunque no lo haremos aqui, que

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

Hemos visto que el número $Cond(A)$ nos da una medida de cuán mala es una matriz en cuanto a la propagación de los errores relativos. Si este número es grande se dice que la matriz está "mal condicionada".

Si A es una matriz singular y, para cierto b, el sistema $Ax = b$ tiene alguna solución, entonces tendrá infinitas y éstas formarán una variedad lineal de $\mathbb{R}^n$. Es decir que sin cambiar nada b se pueden obtener soluciones tan distantes como se quiera. En otras palabras, en este caso tendríamos $\Delta b = 0$ mientras que Δx sería arbitrariamente grande.

En consecuencia, es natural esperar que el número $Cond(A)$ nos de una medida de cuán cerca está A de ser singular. Esto es efectivamente así y lo formalizaremos en el próximo teorema. Pero antes veamos algunos ejemplos. Sea $\varepsilon \sim 0$ entonces la matriz

$$A = \begin{pmatrix} 1 & 1+\varepsilon \\ 1-\varepsilon & 1 \end{pmatrix}$$

está cerca de la matriz singular

$$B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

y en este caso

$$A^{-1} = \varepsilon^{-2} \begin{pmatrix} 1 & -1-\varepsilon \\ -1+\varepsilon & 1 \end{pmatrix}$$

entonces,

$$\|A\|_\infty = 2 + \varepsilon \qquad \|A^{-1}\|_\infty = (2+\varepsilon)\varepsilon^{-2}$$

y en consecuencia,

$$Cond_\infty(A) = \left(\frac{2+\varepsilon}{\varepsilon}\right)^2 > \frac{4}{\varepsilon^2}$$

Es importante recalcar que esta "distancia a las matrices singulares" debe entenderse en forma relativa al tamaño de A. En este ejemplo tenemos no sólo que $\|A - B\|_\infty$ es chica sino que lo es en relación con $\|A\|_\infty$. En efecto,

$$\frac{\|A-B\|_\infty}{\|A\|_\infty} = \frac{\varepsilon}{2+\varepsilon} < \frac{\varepsilon}{2}$$

En particular, estar "cerca" de ser singular no tiene nada que ver con el tamaño del determinante. Para aclarar esto veamos algunos casos simples. Por ejemplo, si $\varepsilon \sim 0$, la matriz

$$A = \begin{pmatrix} \varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}$$

tiene determinante muy pequeño pero es una matriz buenísima en cuanto a la propagación de errores relativos pues $Cond(A) = 1$.

En cambio,

$$A = \begin{pmatrix} 1 & 0 \\ 0 & \frac{1}{\varepsilon} \end{pmatrix}$$

tiene determinante grande pero, en las normas 2, 1 o ∞, $Cond(A) = 1/\varepsilon$

Damos ahora el resultado que relaciona el número de condición de una matriz con su distancia relativa a las matrices singulares.

TEOREMA 2.5. *Dadas $A \in \mathbb{R}^{n\times n}$ inversible y una norma de vectores cualquiera se tiene*

$$\frac{1}{Cond(A)} = \inf_{B \text{ singular}} \frac{\|A - B\|}{\|A\|}$$

Demostración. Sea $B \in \mathbb{R}^{n\times n}$ una matriz singular y tomemos $x \neq 0$ tal que $Bx = 0$. Entonces,

$$\|x\| = \|A^{-1}(A - B)x\| \leq \|A^{-1}\|\|A - B\|\|x\|$$

y en consecuencia,

$$1 \leq \|A^{-1}\|\|A - B\|$$

lo cual muestra que

$$\frac{1}{Cond(A)} \leq \frac{\|A - B\|}{\|A\|} \qquad \forall B \in \mathbb{R}^{n\times n} \quad \text{singular} \tag{2.9}$$

Entonces, para concluir el teorema, falta ver que hay una B singular para la cual vale la igualdad en (2.9). Para esto, sean y tal que $\|A^{-1}y\| = \|A^{-1}\|\|y\|$ y x tal que $Ax = y$. Como y puede tomarse con norma arbitraria, lo elegimos de tal forma que $\|y\| = 1/\|A^{-1}\|$ y en consecuencia $\|x\| = 1$.

Sea ahora z un vector tal que

$$z^T x = 1 \tag{2.10}$$

y

$$z^T u \leq 1 \qquad \forall u \in \mathbb{R}^n \quad \text{tal que} \quad \|u\| = 1 \tag{2.11}$$

La existencia de un tal z es la parte más técnica de la demostración y omitiremos la escritura formal. Sin embargo, observemos que es intuitivamente claro si analizamos el significado geométrico: los u que verifican la ecuación $z^T u = 1$ forman un hiperplano (es decir, una variedad lineal de dimensión $n-1$). Por lo tanto, que haya un z verificando (2.10) y (2.11) significa que hay un hiperplano que pasa por x y que deja a la bola unitaria $\mathcal{B}_1 = \{u \in \mathbb{R}^n : \|u\| \leq 1\}$ de un lado. La existencia de tal hiperplano es clara si se tiene en cuenta que, para toda norma, $\mathcal{B}_1$ es un conjunto convexo y que x está en el borde de éste. Observemos también que, en el caso de la norma $\| \|_2$, se tiene que $z = x$.

Definamos ahora $B = A - yz^T$ y veamos que esta matriz es singular y cumple con la igualdad que queríamos. En efecto,

$$Bx = Ax - yz^T x = y - y = 0$$

y por lo tanto, B es singular.

Por otra parte, $\|A - B\| = \|yz^T\|$, pero por (2.11) tenemos que $|z^T u| \leq 1$ para todo u tal que $\|u\| = 1$ puesto que, si $z^T u < 0$ entonces, $|z^T u| = -z^T u = z^T(-u) \leq 1$ ya que $\| - u\| = 1$. Entonces,

$$\|yz^T u\| = \|y\||z^T u| \leq \frac{1}{\|A^{-1}\|} \qquad \forall u \in \mathbb{R}^n \quad \text{tal que} \quad \|u\| = 1$$

y por lo tanto,

$$\|A - B\| \leq \frac{1}{\|A^{-1}\|}$$

lo que concluye la demostración. □

1. Ejercicios

(1) Calcular la norma 2 de la matriz $A = \begin{pmatrix} 3 & 0 \\ 4 & 5 \end{pmatrix}$.

(2) Se quiere estimar la norma 2 de una matriz $A \in \mathbb{R}^{3\times 3}$ como el máximo del valor $\|Ax\|_2/\|x\|_2$ entre varios vectores $x \in \mathbb{R}^3$ no nulos generados al azar. Hacer un programa que pida el ingreso de una matriz A y luego
   - genere los primeros 100 términos de la siguiente sucesión:

$$s_1 = 0, \quad s_{k+1} = \max\left\{s_k, \frac{\|Ax_k\|_2}{\|x_k\|_2}\right\}$$

     donde los $x_k \in \mathbb{R}^3$ son vectores no nulos generados al azar cuyas coordenadas estén el intervalo $[-1, 1]$.
   - grafique la sucesión calculada, junto con el valor exacto de la norma de la matriz.

Recordar que tanto la norma de un vector como de una matriz se calculan en `Matlab` con el comando **norm**. Tener en cuenta que los vectores generados al azar (comando **rand**) tienen coordenadas en el intervalo $[0, 1]$. Chequear, además, que estos vectores generados resulten no nulos.

(3) Sea $A = \begin{pmatrix} 3 & 0 & 0 \\ 0 & \frac{5}{4} & \frac{3}{4} \\ 0 & \frac{3}{4} & \frac{5}{4} \end{pmatrix}$. Calcular $cond_2(A)$ y $cond_\infty(A)$.

(4) Probar que si $A \in \mathbb{R}^{n\times n}$ es una matriz inversible y $\| \ \|$ es una norma matricial, la condición de A verifica la desigualdad:

$$\frac{1}{cond(A)} \leq \inf\left\{\frac{\|A-B\|}{\|A\|} : B \text{ es singular}\right\}$$

Nota: Más aún, vale la igualdad, pero la otra desigualdad es un poco más complicada de probar. De dicha igualdad se puede concluir que $cond(A)$ mide la distancia relativa de A a la matriz singular más próxima.

(5) (a) Mostrar que $cond_\infty(A) \to \infty$ cuando $\varepsilon \to 0$ para

$$\text{(i)} \quad A = \begin{pmatrix} 1 & 1 & 1 \\ 1 & \varepsilon & \varepsilon^2 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \text{(ii)} \quad B = \begin{pmatrix} 1 & 0 & 1+\varepsilon \\ 2 & 3 & 4 \\ 1-\varepsilon & 0 & 1 \end{pmatrix}.$$

(b) Concluir que la condición de las matrices A y B del ítem anterior tienden a infinito, cualquiera sea la norma considerada.

(6) Sea A la matriz del ejercicio 3. Se quiere resolver el sistema $Ax = b$ para un valor de $b \neq 0$ que se conoce con una precisión mayor que 10^{-3}; es decir, se conoce el valor conjunto de $b + \Delta b$ y se sabe que el error relativo $\dfrac{\|\Delta b\|_2}{\|b\|_2} < 10^{-3}$.

(a) Estimar el error relativo de la solución hallada $\tilde{x} = x + \Delta x$.

(b) Encuentre un ejemplo para b y $\Delta b \neq 0$ de modo que $\frac{\|\Delta x\|_2}{\|x\|_2}$ sea exactamente $cond_2(A)\frac{\|\Delta b\|_2}{\|b\|_2}$.

(7) Sea x la solución exacta al sistema $Ax = b$ y $\tilde{x}$ la solución obtenida numéricamente. Se llama "vector residual" a $r := b - A\tilde{x}$. Si $e = x - \tilde{x}$ se tiene $Ae = r$. Mostrar que:

$$\frac{1}{cond(A)}\frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq cond(A)\frac{\|r\|}{\|b\|}.$$

Concluir que para una matriz mal condicionada los métodos numéricos no aseguran buena aproximación.

(8) Para cada $n \in \mathbb{N}$, se definen $A_n = \begin{pmatrix} 1 & 2 \\ 2 & 4+\frac{1}{n^2} \end{pmatrix}$, $b_n = (1, 2-\frac{1}{n^2})$ y se quiere resolver el sistema $A_n x = b_n$. Utilizando cierto método numérico se obtiene como resultado el vector $(1, 0)$.

(a) Calcular el vector residual producido por esta solución tentativa. ¿Puede decirse que para n grande la solución es razonablemente confiable?

(b) Resolver $A_n x = b_n$ en forma exacta, calcular $cond_\infty(A_n)$ y verificar la cota de error del ejercicio 7.

(9) Sea D_n la matriz diagonal de $n \times n$ con elementos diagonales iguales a $1/10$. Calcular el determinante de D_n y ver que $det(D_n) \to 0$ si $n \to \infty$. ¿D_n está mal condicionada?

(10) (a) Escribir un programa en `Matlab` que resuelva un sistema $Ax = b$, $A \in \mathbb{R}^{n\times n}$ usando eliminación gaussiana sin pivoteo.
(b) Adaptar el programa del ítem anterior para que calcule la matriz A^{-1}.

(11) Para cada $n \in \mathbb{N}$, se quiere calcular la solución del sistema lineal:

$$\begin{aligned} 10^{-n}x + 2y &= 8 \\ x + y &= 2 \end{aligned}$$

utilizando eliminación gaussiana sin pivoteo, con aritmética de punto flotante de 3 dígitos y sistema de redondeo.
(a) Para $n = 2$ y $n = 3$, analizar si el resultado difiere significativamente de la solución real.
(b) Para $n = 3$, repetir el método de eliminación gaussiana eligiendo el pivote más conveniente.

(12) Obtener la descomposición LU de la matriz $\begin{pmatrix} 2 & 4 & -1 & 0 \\ 4 & 10 & -1 & -1 \\ 6 & 10 & -7 & 1 \\ 0 & 2 & 1 & -2 \end{pmatrix}$ de las siguientes dos maneras:
   (a) mediante el algoritmo de eliminación gaussiana,
   (b) despejando los coeficientes de L y U ordenadamente.

(13) Sea $A \in \mathbb{R}^{n\times n}$ una matriz que admite descomposición LU.
   (a) Estimar cuántas operaciones se necesitan para calcular esta descomposición de A, despejando los coeficientes de L y U.
   (b) Se quiere calcular el determinante de A. Para $n \geq 2$, mostrar que si esto se hace mediante el desarrollo sucesivo por alguna fila o columna, entonces se requieren más de $n!$ operaciones. Estimar cuántas operaciones se necesitan para calcularlo si se utiliza la descomposición LU.

(14) Demostrar que si todos los menores principales de una matriz $A \in \mathbb{R}^{n\times n}$ son no singulares, entonces ésta admite descomposición LU.

(15) Probar que la matriz no singular:

$$\begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

no tiene una descomposición LU, mientras que la matriz singular $A - I$ sí la tiene. Dar la matriz de permutaciones P tal que PA tenga una factorización LU.

(16) Considerar el algoritmo de eliminación gaussiana sin pivoteo aplicado a un sistema $Ax = b$ donde $A \in \mathbb{R}^{n\times n}$ es una matriz tridiagonal. Demostrar que si A es además estrictamente diagonal dominante, entonces durante la ejecución del algoritmo no se encuentra ningún pivote nulo. (Ayuda: demostrar que si A es estrictamente diagonal dominante, entonces luego de hacer cada etapa de la eliminación la matriz resultante también lo es.)

(17) Sea $A \in \mathbb{R}^{n\times n}$ una matriz tridiagonal tal que en el proceso de eliminación gaussiana no se encuentra ningún pivote nulo. Demostrar que A admite descomposición LU con L y U también tridiagonales.

(18) Adaptar el programa del ejercicio 10 para que resuelva un sistema de ecuaciones $Ax = b$, donde $A \in \mathbb{R}^{n\times n}$ es una matriz tridiagonal. Utilizar el comando **flops** de `Matlab` para conocer la cantidad de operaciones efectuadas y comparar con las que se requieren al resolver el mismo sistema utilizando los comandos **inv** y \, que no están especialmente pensados para matrices tridiagonales.

(19) La n-ésima matriz de Hilbert $H_n \in \mathbb{R}^{n\times n}$, se define de la siguiente manera

$$(H_n)_{i,j} = \frac{1}{i+j-1}.$$

Estas matrices son un ejemplo de matrices mal condicionadas y por tal motivo se las utiliza habitualmente para testear rutinas numéricas.

(a) Demostrar que $cond_\infty(H_n) \geq n^2$.

(b) Utilizar su programa del ejercicio 10 para calcular la inversa de la matriz de Hilbert H_9. Verificar su resultado calculando los productos $H_9 H_9^{-1}$ y $H_9^{-1} H_9$. Comparar con el resultado obtenido mediante el comando **inv**.

Nota: En realidad, $cond_\infty(H_n)$ es mucho mayor que n^2. Estas matrices pueden obtenerse en `Matlab` mediante el comando **hilb(n)** y su condición infinito puede calcularse con el comando **cond**.

(20) Considerar el sistema de ecuaciones lineales $Ax = b$, con

$$A = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 3 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 4 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 7 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 8 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 9 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 10 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Utilizando el comando **lu** de `Matlab`, verificar que la eliminación gaussiana puede crear elementos no nulos en lugares donde inicialmente había ceros (es decir, se produce una matriz densa a pesar de partir de una matriz rala). En muchas aplicaciones, uno debe resolver un sistema de ecuaciones lineales del orden de $10^4 \times 10^4$ donde hay a lo sumo 5 elementos no nulos por fila. Es decir, hay a lo sumo 5×10^4 elementos no nulos en la matriz, cifra bastante inferior a la cantidad total de elementos. Calcular qué cantidad de bytes (2 bytes por elemento) ocuparía una matriz densa de esas dimensiones. Este tipo de situación motiva el estudio de métodos de resolución de sistemas con matrices ralas que no involucren un llenado excesivo.

(21) Utilizar el Teorema de Cholesky para demostrar que las siguientes propiedades de una matriz son equivalentes:

- A es simétrica y definida positiva
- hay un conjunto de vectores linealmente independientes $x^1, x^2, \cdots, x^n$ de $\mathbb{R}^n$, tales que $a_{ij} = (x^i)^t x^j$.

(22) Considerar la matriz $\begin{pmatrix} 4 & 2 & -2 \\ 2 & 5 & 5 \\ -2 & 5 & 11 \end{pmatrix}$.

Mostrar que es definida positiva y calcular su descomposición de Cholesky.

(23) Estimar cuántas operaciones se requieren para hallar la descomposición de Cholesky de una matriz simétrica y definida positiva $A \in \mathbb{R}^{n\times n}$.

CAPíTULO 3

Resolución de sistemas lineales.

El objetivo de este capítulo es estudiar diferentes formas de resolver un sistema lineal de n ecuaciones con n incógnitas. Para dar soilución a este problema se pueden emplear dos grandes subclases de métodos; los directos y los iterados. Dentro de los métodos de cálculo directo se encuentran el de triangulación de Gauss, el de descomposición LU y el método de Cholesky. Entre los métodos iterativos más usuales encontramos el de Jacobi, Gauss-Seidel y el de relajación.

1. Métodos directos

1.1. Triangulación de Gauss y descomposición LU. El proceso de triangulación de Gauss puede verse como el resultado que se obtiene de multiplicar por matrices de la siguiente forma,

Primer paso: Multiplicar por

$$L_1 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ m_{21} & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ m_{N1} & 0 & \cdots & 1 \end{pmatrix}$$

con

$$m_{i1} = -\frac{a_{i1}}{a_{11}}$$

Entonces $L_1 A$ tendrá la forma

$$L_1 A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ 0 & a^1_{22} & \cdots & a^1_{2N} \\ \vdots & & \ddots & \vdots \\ 0 & a^1_{N2} & \cdots & a^1_{NN} \end{pmatrix}$$

Segundo paso: Multiplicar por

$$L_2 = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & m_{32} & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & m_{N2} & \cdots & 1 \end{pmatrix}$$

y nos queda

$$L_2 L_1 A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ 0 & a_{22}^2 & \cdots & a_{2N}^2 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & a_{NN}^2 \end{pmatrix}$$

Así sucesivamente hasta llegar a una matriz triangular suprior

$$L_{N-1} L_{N-2} \cdots L_2 L_1 A = U = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1N} \\ 0 & u_{22} & \cdots & u_{2N} \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & u_{NN} \end{pmatrix}.$$

Es fácil ver que la inversa de L_1 viene dada por

$$(L_1)^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -m_{21} & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ -m_{N1} & 0 & \cdots & 1 \end{pmatrix}$$

y, en general, L_j^{-1} es como L_j pero cambiando los signos de los m_{ji}.

Entonces podemos escribir A como sigue,

$$A = L_1^{-1} L_2^{-1} \cdots L_{N-1}^{-1} U,$$

además, observemos que la matriz $L = L_1^{-1} L_2^{-1} \cdots L_{N-1}^{-1}$ es de la forma

$$L = L_1^{-1} L_2^{-1} \cdots L_{N-1}^{-1} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -m_{21} & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ -m_{N1} & -m_{N2} & \cdots & 1 \end{pmatrix}$$

Así hemos demostrado el siguiente teorema,

TEOREMA 3.1. *Si no hace falta intercambiar filas en la eliminación de Gauss se obtiene*

$$A = LU$$

donde U es triangular superior y L es triangular inferior con 1 en la diagonal.

Además tenemos el siguiente corolario.

COROLARIO 3.2.

$$det(A) = det(U).$$

En el caso general, si hace falta cambiar filas, se tiene

$$PA = LU$$

con P una matriz de permutaciones.

1.2. Descomposición de Cholesky. En el caso en que $A \in \mathbb{R}^{N\times N}$ es definida positiva y simétrica una descomposición $L-U$ (con $U = L^T$) puede obtenerse más eficientemente mediante el método de Cholesky.

DEFINICIÓN 3.3. *$A \in \mathbb{R}^{N\times N}$ se dice definida positiva si*

$$\langle x, Ax\rangle > 0 \qquad \forall x \neq 0$$

Observemos que si $A = LL^T$ con L una matriz inversible, entonces

(1) A es simétrica.
(2) A es definida positiva pues $\langle x, Ax\rangle = \|L^T x\|_2^2 > 0, \qquad \forall x \neq 0.$

En consecuencia, para que A pueda escribirse como LL^T con L inversible es necesario que A sea simétrica y definida positiva.

Ahora, para lograr una descomposición de la forma LL^T, analicemos primero el caso simple, $A \in \mathbb{R}^{3\times 3}$. Planteamos $A = LL^T$ y nos queda

$$A = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{pmatrix}$$

Entonces, despejando los coeficientes, se obtiene

$$a_{11} = l_{11}^2 \qquad l_{11} = \sqrt{a_{11}}$$
$$a_{12} = l_{11}l_{21} \qquad l_{21} = \frac{a_{12}}{l_{11}}$$

etc.

Ahora veamos algunas propiedades que nos serán muy útiles. En el caso general en que $A \in \mathbb{R}^{N\times N}$ sea simétrica (i.e. $A = A^T$) y definida positiva se tienen las siguientes propiedades,

(1) $a_{ii} > 0$ para cualquier $i = 1, ..., N$, pues

$$0 < \langle e_i, Ae_i \rangle = a_{ii}$$

(2) Los menores principales s_j son positivos (esto fue demostrado en algebra lineal).

Ahora observamos que lo que hicimos en 3×3 se puede hacer en $N \times N$, es decir,

$$A = LL^T$$

si y solo si

$$a_{ik} = \sum_{j=1}^{k} l_{ij} l_{kj}.$$

Es decir,

$$a_{ik} = \sum_{j=1}^{k-1} l_{ij} l_{kj} + l_{ik} l_{kk}.$$

Entonces, despejando,

$$l_{ik} = \frac{\left(a_{ik} - \sum_{j=1}^{k-1} l_{ij} l_{kj}\right)}{l_{kk}} \qquad i > k.$$

Además

$$a_{kk} = \sum_{j=1}^{k} l_{kj}^2 = \sum_{j=1}^{k-1} l_{kj}^2 + l_{kk}^2$$

y entonces

$$l_{kk} = \sqrt{\left(a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2\right)}$$

Obtenemos, de esta manera, una forma recursiva para el cálculo de los elementos l_{ij}.

Para $k = 1, 2, ..., N$ hacemos

$$\left\{\begin{array}{ccccc} l_{11} & \to & l_{21} & \cdots & l_{N1} \\ l_{22} & \to & l_{32} & \cdots & l_{N2} \\ \vdots & & \vdots & & \\ l_{N-1N-1} & \to & l_{NN-1} & & \\ l_{NN} & & & & \end{array}\right.$$

Para que el algoritmo de Cholesky esté bien definido necesitamos ver que el argumento de la raíz cuadrada involucrada en el cálculo de l_{kk} sea positivo; es decir,

$$a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2 > 0.$$

Veamos que ésto es una consecuencia de ser A definida positiva. Argumentemos por inducción.

El a_{11} es positivo, entonces existe l_{11} positivo tal que $l_{11}^2 = a_{11}$. Supongamos que llegamos hasta el paso k, es decir

$$l_{11}, l_{22}, \ldots, l_{k-1k-1}$$

son todos números reales positivos y supongamos que

$$a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2 \leq 0.$$

Entonces

$$l_{kk} = \sqrt{\left(a_{kk} - \sum_{j=1}^{k-1} l_{kj}^2\right)} = 0 \text{ ó es un número en } \boldsymbol{C}$$

Pero si llamamos A_k al menor principal de A y L_k al menor principal de L, las matrices que se obtienen son de tamaño $k \times k$

$$A_k = \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix} \qquad \text{y} \qquad L_k = \begin{pmatrix} l_{11} & \cdots & l_{1k} \\ \vdots & \ddots & \vdots \\ l_{k1} & \cdots & l_{kk} \end{pmatrix}$$

y resulta fácil ver que

$$A_k = L_k L_k^T.$$

Entonces

$$0 < \det(A_k) = (\det(L_k))^2 = l_{11}^2 \cdots l_{k-1k-1}^2 l_{kk}^2;$$

como los primeros factores son positivos el último, l_{kk}^2 debe también ser positivo, absurdo.

Para terminar, hagamos las siguientes observaciones, el algoritmo de Cholesky es más conveniente que el de Gauss $(L-U)$ porque,

(1) El número de operaciones es $O(N^3/6)$ (en lugar de $O(N^3/3)$).
(2) Es estable, sin necesidad de "pivoteo". Los l_{ij} no crecen respecto de A pues

$$a_{kk} = \sum_{j=1}^{k} l_{kj}^2$$

implica que

$$|l_{kj}| \leq \sqrt{a_{kk}}$$

2. Métodos iterativos

Estos métodos convienen en general para matrices ralas (i.e. con muchos ceros). Este tipo de matrices aparecen, por ejemplo, cuando se discretizan ecuaciones diferenciales.

Como antes el objetivo es resolver

$$Ax = b$$

con A una matriz inversible.

Los métodos iterativos generan una sucesión

$$x_0 \to x_1 \to x_2 \to \cdots$$

donde x_{k+1} se calcula a partir de x_k.

2.1. Método de Jacobi. Empecemos con un ejemplo para ver como funciona el método de Jacobi,

$$\begin{cases} 4x_1 + x_2 = 5 \\ x_1 + 4x_2 = 5 \end{cases}$$

La solución es

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

y llamaremos

$$b = \begin{pmatrix} 5 \\ 5 \end{pmatrix}$$

El método de Jacobi calcula x^{k+1} a partir de x^k de la siguiente forma

$$\begin{cases} 4x_1^{k+1} = 5 - x_2^k \\ 4x_2^{k+1} = 5 - x_1^k \end{cases}$$

Es decir

$$x^{k+1} = \begin{pmatrix} 0 & -\frac{1}{4} \\ -\frac{1}{4} & 0 \end{pmatrix} x^k + \frac{b}{4}.$$

Entonces si empezamos con $x^0 = (0, 0)$, tenemos

$$x^0 = \begin{pmatrix} 0 \\ \\ 0 \end{pmatrix} \to x^1 = \begin{pmatrix} \frac{5}{4} \\ \\ \frac{5}{4} \end{pmatrix} \to x^2 = \begin{pmatrix} \frac{15}{16} \\ \\ \frac{15}{16} \end{pmatrix} \to \cdots$$

Convergencia y estimación del error

En forma matricial la iteración de Jacobi se escribe como

$$x^{k+1} = Bx^k + c.$$

Por otro lado la solución exacta, x, cumple que

$$x = Bx + c,$$

entonces el error $e^k = x^k - x$ verifica

$$e^{k+1} = Be^k$$

y entonces, iterando esta última igualdad,

$$e^k = B^k e^0.$$

En nuestro ejemplo observamos que

$$\|B\|_\infty = \frac{1}{4}$$

y entonces

$$\|B^k\|_\infty \leq \|B\|_\infty^k \leq (\frac{1}{4})^k \to 0 \qquad k \to \infty.$$

De ésto concluímos que

$$\|e^k\|_\infty \leq \left(\frac{1}{4}\right)^k \|e^0\|_\infty \to 0$$

es decir la iteración converge cualquiera sea el dato inicial x^0.

Por supuesto, esto no es cierto en general. La convergencia depende de como sea la matriz B. Si $\|B\| < 1$ para alguna norma asociada a una norma de vectores entonces el método convergerá cualquiera sea la condición inicial y si no no.

En el caso general, $A \in \mathbb{R}^{N\times N}$ supongamos $a_{ii} \neq 0$, $\forall i$ (si A es inversible esto se puede obtener reordenando). Despejamos x_i de la $i-$ésima ecuación, para $i = 1, ..., N$ tenemos

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

Resulta natural utilizar las componentes ya calculadas $x_1^{k+1}, \ldots, x_{i-1}^{k+1}$ para calcular la nueva aproximación x_i^{k+1}, resultando el método de Gauss-Seidel.

2.2. Método de Gauss-Seidel. Para $i = 1, \ldots, N$;

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

Escritura matricial de la iteración Escribimos

$$A = D + L + U$$

$$A = \begin{pmatrix} a_{11} & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & a_{NN} \end{pmatrix} + \begin{pmatrix} 0 & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix} + \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ a_{N1} & \cdots & 0 \end{pmatrix}$$

Entonces

$$Ax = b$$

si y solo si

$$Dx = -(L + U)x + b$$

Tanto el método de Jacobi como el de Gauss-Seidel pueden escribirse en la forma

$$x^{k+1} = Bx^k + c.$$

(1) Jacobi

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

(2) Gauss-Seidel

$$x^{k+1} = -(D + L)^{-1}Ux^k + (D + L)^{-1}b$$

Si escribimos $e^k = x^k - x$ y usamos que la solución exacta x cumple

$$x = -D^{-1}(L + U)x + D^{-1}b$$

y

$$x = -(D + L)^{-1}Ux + (D + L)^{-1}b$$

respectivamente, tenemos

$$e^{k+1} = -D^{-1}(L + U)e^k = B_J e^k$$
$$e^{k+1} = -(D + L)^{-1}Ue^k = B_{GS} e^k$$

En general, si la iteración está dada por una matriz B, o sea,

$$e^{k+1} = Be^k$$

tenemos

$$e^k = B^k e^0$$

Entonces si queremos que $e^k \to 0$ para todo dato inicial, es necesario que $B^k \to 0$. El siguiente objetivo es dar una caracterización de las matrices con esta propiedad.

En el ejemplo dedujimos que $B^k \to 0$ del hecho de que $\|B\| < 1$. Sin embargo $\|B\|_\infty$ podría ser grande y $B^k \to 0$. Por ejemplo

$$B = \begin{pmatrix} \frac{1}{2} & 1000 \\ 0 & \frac{1}{2} \end{pmatrix}$$

Observemos que $\|B\|_\infty = 1000.5$. Sin embargo $B^k \to 0$. En efecto $B = \frac{1}{2}\begin{pmatrix} 1 & 2000 \\ 0 & 1 \end{pmatrix}$.

En general las matrices de la forma $C = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$ verifican que $C^k = \begin{pmatrix} 1 & ka \\ 0 & 1 \end{pmatrix}$

Entonces

$$B^k = (\frac{1}{2})^k \begin{pmatrix} 1 & k2000 \\ 0 & 1 \end{pmatrix}$$

y se tiene que $(B^k)_{ij} \to 0$, $\forall i, j$ y esto implica que $B^k \to 0$.

Vale destacar que para que $B^k \to 0$ basta que exista alguna norma tal que $\|B\| < 1$.

El segundo ejemplo trata el caso en que A es simétrica. En este caso se puede diagonalizar, es decir, existe S tal que

$$SAS^{-1} = \begin{pmatrix} \lambda_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \lambda_N \end{pmatrix}$$

con λ_i los autovalores de A. En este caso

$$A^k = S^{-1} \begin{pmatrix} \lambda_1^k & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \lambda_N^k \end{pmatrix} S$$

y se tiene que

$$A^k \to 0$$

si y solo si

$$\max_i |\lambda_i| = \rho(A) < 1$$

Esto es cierto en general, pero si A no es diagonalizable es más difícil de probar.

En el caso en que A es simétrica se tiene

$$\rho(A) = \|A\|_2$$

entonces, si $\rho(A) < 1$ se tiene que $\|A\|_2 < 1$ y entonces $A^k \to 0$.

En general vale que,

$$\rho(A) \leq \|A\| \qquad \text{para cualquier norma}$$

y aunque $\rho(A)$ no es una norma, se tiene

TEOREMA 3.4.

$$\rho(A) = \inf_{\|\ \|} \|A\|$$

O sea $\forall \varepsilon > 0$ *existe una norma tal que*

$$\rho(A) \leq \|A\| \leq \rho(A) + \varepsilon.$$

Demostración. Primero veamos que

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

Observamos que $\|A\|$ en $\mathbb{R}$ es igual a $\|A\|$ en C (ejercicio).

Sea x tal que

$$Ax = \lambda_{\max} x \qquad x \neq 0$$

entonces

$$|\lambda_{\max}|\|x\| = \|Ax\| \leq \|A\|\|x\|$$

y así

$$|\lambda_{\max}| = \rho(A) \leq \|A\|$$

Ahora veamos que dado $\varepsilon > 0$ existe una norma con

$$\|A\| \leq \rho(A) + \varepsilon.$$

Queremos definir $\|x\|$, para $x \in \mathbb{R}^N$. Recordemos la forma de Jordan de una matriz. En alguna base $\{v_i\}$, una matriz B se transforma en

$$\begin{pmatrix} J_1 & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & J_r \end{pmatrix}$$

donde los J_i son los "bloques de Jordan",

$$J_i = \begin{pmatrix} \lambda_i & 1 & \cdots & 0 \\ 0 & \lambda_i & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \lambda_i \end{pmatrix}$$

Esta es la forma normal usual. Sin embargo, puede obtenerse una nueva base en la cual la transformación lineal toma la forma análoga pero con los

$$\tilde{J}_i = \begin{pmatrix} \lambda_i & \varepsilon & \cdots & 0 \\ 0 & \lambda_i & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \lambda_i \end{pmatrix}$$

donde ε es positivo y arbitrario. Esto se logra re-escalando la base. Miremos, por ejemplo, el bloque 1,

$$J_1 = \begin{pmatrix} \lambda_1 & 1 & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \lambda_1 \end{pmatrix}$$

en la base $v_1, ..., v_m$. Si T es la transformación lineal asociada a B tenemos

$$\begin{aligned} Tv_1 &= \lambda_1 v_1 \\ Tv_2 &= v_1 + \lambda_1 v_2 \\ Tv_3 &= v_2 + \lambda_1 v_3 \\ &\vdots \\ Tv_m &= v_{m-1} + \lambda_1 v_m \end{aligned}$$

Ahora definamos $\tilde{v}_1 = v_1$, $\tilde{v}_2 = \varepsilon v_2$, $\tilde{v}_3 = \varepsilon^2 v_3$,....., $\tilde{v}_m = \varepsilon^{m-1} v_m$. Tenemos entonces

$$\begin{aligned} T\tilde{v}_1 &= \lambda_1 \tilde{v}_1 \\ T\tilde{v}_2 &= \varepsilon \tilde{v}_1 + \lambda_1 \tilde{v}_2 \\ T\tilde{v}_3 &= \varepsilon \tilde{v}_2 + \lambda_1 \tilde{v}_3 \\ &\vdots \\ T\tilde{v}_m &= \varepsilon \tilde{v}_{m-1} + \lambda_1 \tilde{v}_m \end{aligned}$$

Por lo tanto en la base $\tilde{v}_1, ..., \tilde{v}_m$ queda el bloque

$$\tilde{J}_1 = \begin{pmatrix} \lambda_1 & \varepsilon & \cdots & 0 \\ 0 & \lambda_1 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \lambda_1 \end{pmatrix}$$

Hecho esto, definamos la norma de la siguiente manera, dado x lo escribimos en la base $\tilde{v}_i$,

$$x = \sum \alpha_i \tilde{v}_i$$

y definimos

$$\|x\| = \max |\alpha_i|$$

es decir la norma $\| \ \|_\infty$ en esa base.

Entonces, es fácil ver que,

$$\|A\| = \rho(A) + \varepsilon$$

pues $\|A\|$ es el máximo de $\sum_j |a_{ij}|$ si $\|A\|$ es la norma asociada a $\|x\|$. □

Corolario 3.5.

$$B^k \to 0 \text{ } si \text{ } y \text{ } solo \text{ } si \text{ } \rho(B) < 1.$$

Demostración. Veamos primero la vuelta. Si $\rho(B) < 1$ por el teorema anterior existe una norma tal que $\|B\| < 1$, entonces

$$\|B^k\| \leq \|B\|^k \to 0$$

. Ahora para la ida, supongamos que $\rho(B) \geq 1$, entonces existe $z \in \mathbb{C}^N$, $z \neq 0$, tal que

$$Bz = \lambda_{\max} z$$

y entonces

$$B^k z = \lambda_{\max}^k z$$

Esto implica que

$$\|B^k z\| = \rho(B)^k \|z\|$$

no tiende a cero.

Tomando parte real e imaginaria se ve que hay algún $x \in \mathbb{R}^N$ tal que $B^k x$ no tiende a cero. □

Ahora veamos otra forma de probar estos resultados.

Teorema 3.6.

$$B^k \to 0 \text{ } si \text{ } y \text{ } solo \text{ } si \text{ } \rho(B) < 1$$

Demostración. B es semejante a una matriz triangular (forma de Jordan), o sea, existe una matriz C tal que

$$CBC^{-1} = J$$

con J la forma de Jordan de B.

Ahora dado $\varepsilon > 0$ multiplicamos por la matriz diagonal

$$D = \begin{pmatrix} \varepsilon^{-1} & 0 & \cdots & 0 \\ 0 & \varepsilon^{-2} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \varepsilon^{-N} \end{pmatrix}$$

y su inversa para obtener

$$DJD^{-1} = \begin{pmatrix} \varepsilon^{-1} & 0 & \cdots & 0 \\ 0 & \varepsilon^{-2} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \varepsilon^{-N} \end{pmatrix} J \begin{pmatrix} \varepsilon & 0 & \cdots & 0 \\ 0 & \varepsilon^{2} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \varepsilon^{N} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \varepsilon & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \lambda_k \end{pmatrix}$$

En general la matriz $J = (\alpha_{ij})$ tiene coeficientes no nulos solo en la diagonal y en los lugares $(i, i+1)$. Y al multiplicar por D y D^{-1} quedan ε y 0 en lugar de 1 y 0 en la forma de Jordan.

En conclusión, queda

$$DCBC^{-1}D^{-1} = \begin{pmatrix} \lambda_1 & \varepsilon & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \lambda_k \end{pmatrix} = A$$

Para simplificar llamemos $S = DC$ y nos queda

$$SBS^{-1} = A$$

Ahora observamos que

$$\|A\|_\infty = \rho(B) + \varepsilon$$

pues

$$\|A\|_\infty = \max_j \sum |a_{ij}|$$

Pero,

$$B^k = S^{-1}A^kS$$

Entonces,

$$\begin{array}{rcl} \|B^k\|_\infty & \leq & \|S^{-1}\|_\infty \|A^k\|_\infty \|S\|_\infty \\ & = & Cond(S)\|A^k\|_\infty \leq \\ & \leq & Cond(S)\|A\|_\infty^k \\ & \leq & Cond(S)(\rho(B)+\varepsilon)^k \to 0 \end{array}$$

si ε es tal que $\rho(B) + \varepsilon < 1$.

Observemos que $Cond(S) \sim \frac{1}{\varepsilon^{N-1}}$. □

Corolario 3.7.

$$\|B^k\|^{1/k} \to \rho(B)$$

Demostración. Basta probarlo para la norma $\| \ \|_\infty$ pues todas las normas son equivalentes. Ya vimos que $\forall \varepsilon$,

$$\|B^k\|_\infty \leq Cond(S)(\rho(B)+\varepsilon)^k \leq \frac{C}{\varepsilon^{N-1}}(\rho(B)+\varepsilon)^k$$

Por otro lado se ve fácil que

$$\rho(B)^k \leq \rho(B^k) \leq \|B^k\|_\infty$$

Entonces

$$\rho(B)^k \leq \|B^k\|_\infty \leq \frac{C}{\varepsilon^{N-1}}(\rho(B)+\varepsilon)^k.$$

Luego,

$$\rho(B) \leq \|B^k\|_\infty^{1/k} \leq (\frac{C}{\varepsilon^{N-1}})^{1/k}(\rho(B)+\varepsilon) \to (\rho(B)+\varepsilon)$$

O sea, $\forall \varepsilon$ existe k a partir del cual

$$\rho(B) \leq \|B^k\|_\infty^{1/k} \leq (\rho(B)+2\varepsilon)$$

es decir

$$\|B^k\|_\infty^{1/k} \to \rho(B)$$

□

Ahora observemos lo siguiente, para B simétrica ya sabíamos que $\|B^k\| = \rho(B)^k$. En general esto no es cierto pero, para k *grande* vale que $\|B^k\| \sim \rho(B)^k$. Esto da una manera de comparar dos métodos iterativos (por ejemplo Jacobi y Gauss-Seidel). Supongamos que el método i ($i = 1, 2$) tiene la matriz de iteración B_i. Si

$$\rho(B_1) < \rho(B_2)$$

entonces

$$\|B_1^k\| < \|B_2^k\|$$

para k grande. O sea el método 1 es mejor asintóticamente (aunque para un número dado de iteraciones podría ser mejor el 2).

2.3. Análisis de los métodos de Jacobi y Gauss-Seidel.

DEFINICIÓN 3.8. *Una matriz $A \in \mathbb{R}^{N\times N}$ es estrictamente diagonal dominante si*

$$|a_{ii}| > \sum_{i\neq j} |a_{ij}| \qquad \forall i$$

Si A es estrictamente diagonal dominante entonces tanto Jacobi como Gauss-Seidel convergen.

TEOREMA 3.9. *Si A es estrictamente diagonal dominante el método de Jacobi converge.*

Demostración. Recordemos que

$$A = D + L + U \qquad B_J = -D^{-1}(L + U)$$

En este caso es fácil ver que

$$\|B_J\|_\infty < 1$$

En efecto, $B_J = (b_{ij})$ con

$$b_{ij} = \frac{a_{ij}}{a_{ii}} \; i \neq j \qquad b_{ii} = 0$$

entonces

$$\|B_J\|_\infty = \max_i \sum_{i\neq j} \frac{|a_{ij}|}{|a_{ii}|} < 1$$

pues A es estrictamente diagonal dominante. □

TEOREMA 3.10. *Si A es estrictamente diagonal dominante el método de Gauss-Seidel converge.*

Demostración. Como antes recordemos que

$$A = D + L + U \qquad B_{GS} = -(L + D)^{-1}U.$$

Hay que ver que $\rho(B) < 1$. Sea λ un autovalor de B y x un autovector con $\|x\|_\infty = 1$. Entonces tenemos,

$$-(L + D)^{-1}Ux = \lambda x$$

y esto es equivalente a

$$-Ux = \lambda(L+D)x$$

$$-\sum_{j=i+1}^{N} a_{ij}x_j = \lambda\sum_{j=1}^{i} a_{ij}x_j.$$

O bien,

$$\lambda a_{ii}x_i = -\lambda\sum_{j=1}^{i} a_{ij}x_j - \sum_{j=i+1}^{N} a_{ij}x_j$$

Sea i tal que $\|x\|_\infty = |x_i| \geq |x_j|$, entonces

$$|\lambda||a_{ii}| \leq |\lambda|\sum_{j=1}^{i-1}|a_{ij}| + \sum_{j=i+1}^{N}|a_{ij}|.$$

De esta forma obtuvimos

$$|\lambda| \leq \frac{\sum_{j=i+1}^{N}|a_{ij}|}{|a_{ii}| - \sum_{j=1}^{i-1}|a_{ij}|} < 1$$

pues A es estrictamente diagonal dominante. □

2.4. Matrices simétricas definidas positivas. Un caso importante es el de A simétrica y definida positiva. En este caso veremos,

(1) Jacobi no es necesariamente convergente.
(2) Gauss-Seidel es convergente.

Empecemos con un ejemplo. Sea a tal que $0 < a < 1$ y tomemos

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$

Esta matriz A es simétrica y definida positiva. Para ver que es definida positiva hay que ver que los menores principales son positivos.

$$A_1 = (1)$$

$$A_2 = \begin{pmatrix} 1 & a \\ a & 1 \end{pmatrix} \qquad \det(A) = 1 - a^2 > 0$$

y además

$$\det(A) = 1 + 2a^3 - 3a^2 = (a-1)^2(a+\frac{1}{2}) > 0 \qquad \text{si } a > -\frac{1}{2}$$

Analicemos el método de Jacobi en este caso,

$$B_J = -D^{-1}(L+U) = \begin{pmatrix} 0 & -a & -a \\ -a & 0 & -a \\ -a & -a & 0 \end{pmatrix}.$$

Calculemos los autovalores de B, el polinomio característico es

$$p(\lambda) = \det \begin{pmatrix} \lambda & a & a \\ a & \lambda & a \\ a & a & \lambda \end{pmatrix} = \lambda^3 + 2a^3 - 3a^2\lambda.$$

Observemos que $p(a) = 0$ entonces $\lambda_1 = a$ y como $p'(a) = 0$, $\lambda_1 = a$ es una raiz doble de p. Dividiendo p por $(\lambda - a)^2$ se obtiene que la tercer raiz es $\lambda_3 = -2a$. Entonces si

$$1 > a \geq \frac{1}{2}$$

se tiene que

$$\rho(B) = 2a \geq 1$$

y con esto no se puede asegurar que el método de Jacobi converja para cualquier dato inicial.

Conclusión: A simétrica definida positiva no implica que el método de Jacobi sea necesariamente convergente.

Observación 3.11. Tomando

$$A = \begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{pmatrix}$$

tenemos que $\rho(B_J) = 1$, con lo cual Jacobi no converge. Entonces la condición estricta es necesaria en el teorema que muestra la convergencia para A estrictamente diagonal dominante.

Observación 3.12.

$$A = D + L + L^T$$

Puede demostrarse que si $D - (L + L^T)$ es definida positiva, entonces

$$\rho(B_J) < 1$$

si y sólo si A es definida positiva . En particular, si

$$A = D + L + L^T \qquad \text{y} \qquad \tilde{A} = D - (L + L^T)$$

son definidas positivas, se tiene que

$$\rho(B_J) < 1$$

o sea Jacobi converge para A y para $\tilde{A}$ (la matriz $\tilde{B}_J = -B_J$, sólo cambia de signo). Para una demostración de este hecho ver Isaacson-Keller (pag 72).

Ejemplo 3.13. Para la matriz

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$

con $\frac{1}{2} \leq a < 1$ se tiene

$$\tilde{A} = \begin{pmatrix} 1 & -a & -a \\ -a & 1 & -a \\ -a & -a & 1 \end{pmatrix}.$$

Y resulta que $\tilde{A}$ no es definida positiva.

Ahora analicemos que pasa con el método de Gauss-Seidel en este mismo ejemplo.

$$A = \begin{pmatrix} 1 & a & a \\ a & 1 & a \\ a & a & 1 \end{pmatrix}$$

y entonces

$$L + D = \begin{pmatrix} 1 & 0 & 0 \\ a & 1 & 0 \\ a & a & 1 \end{pmatrix} \qquad \text{y} \qquad U = \begin{pmatrix} 0 & a & a \\ 0 & 0 & a \\ 0 & 0 & 0 \end{pmatrix}.$$

Calculando obtenemos

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

Entonces $B = B_{GS} = -(L + D)^{-1}U$ es

$$B = \begin{pmatrix} 0 & -a & -a \\ 0 & a^2 & a^2 - a \\ 0 & a^2 - a^3 & 2a^2 - a^3 \end{pmatrix}.$$

Veamos los autovalores de B,

$$\lambda I - B = \begin{pmatrix} \lambda & a & a \\ 0 & \lambda - a^2 & a - a^2 \\ 0 & -a^2 + a^3 & \lambda - 2a^2 + a^3 \end{pmatrix}.$$

Tenemos $\lambda_1 = 0$. Para simplificar, ahora consideremos el caso particular $a = \frac{1}{2}$, para este valor de a se obtiene

$$B = \begin{pmatrix} 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{4} & -\frac{1}{4} \\ 0 & \frac{1}{8} & \frac{3}{8} \end{pmatrix}.$$

Y entonces,

$$8B = \begin{pmatrix} 0 & -4 & -4 \\ 0 & 2 & -2 \\ 0 & 1 & 3 \end{pmatrix} \qquad \text{y} \qquad \lambda I - 8B = \begin{pmatrix} \lambda & 4 & 4 \\ 0 & \lambda - 2 & 2 \\ 0 & -1 & \lambda - 3 \end{pmatrix}.$$

Con lo cual,

$$\det(\lambda I - 8B) = \lambda((\lambda - 2)(\lambda - 3) + 2).$$

Las raíces de $(\lambda - 2)(\lambda - 3) + 2$ son

$$\lambda_2 = \frac{5 + \sqrt{-7}}{2} \qquad \lambda_3 = \frac{5 - \sqrt{-7}}{2}$$

Como éstos son los autovalores de $8B$ los autovalores de B son

$$\lambda_1 = 0 \qquad \lambda_2 = \frac{5 + \sqrt{-7}}{16} \qquad \lambda_2 = \frac{5 - \sqrt{-7}}{16}.$$

Observemos que

$$|\lambda_2| = |\lambda_3| = \frac{\sqrt{32}}{16} = \frac{\sqrt{2}}{4} < 1.$$

Entonces el método de Gauss-Seidel converge.

Más adelante veremos que si A es simétrica y definida positiva entonces Gauss-Seigel converge.

En particular este ejemplo nos da un caso donde el método de Gauss-Seigel converge pero Jacobi no, o sea $\rho(B_{GS}) < 1$ y $\rho(B_J) \geq 1$.

Ahora veremos un ejemplo "al revés´´, es decir donde Jacobi converge pero Gauss-Seigel no.

Ejemplo 3.14. (Collatz 1942, ver Varga, pag 74). Sea

$$A = \begin{pmatrix} 1 & 2 & -2 \\ 1 & 1 & 1 \\ 2 & 2 & 1 \end{pmatrix}.$$

Para el método de Jacobi nos queda

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

$$\lambda I - B_J = \begin{pmatrix} \lambda & 2 & -2 \\ 1 & \lambda & 1 \\ 2 & 2 & \lambda \end{pmatrix}$$

entonces

$$p(\lambda) = \lambda^3$$

y los autovalores resultan ser

$$\lambda_1 = \lambda_2 = \lambda_3 = 0$$

Entonces B_J es nilpotente, $\rho(B_J) = 0$ y el método converge en tres pasos.

$$e^3 = B_J^3 e^0 = 0$$

Ahora analicemos el método de Gauss-Seidel para este ejemplo.

$$L + D = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 2 & 2 & 1 \end{pmatrix}; \quad -U = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 0 & -1 \\ 0 & 0 & 0 \end{pmatrix} \quad \text{y} \quad (L + D)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -2 & 1 \end{pmatrix}$$

En consecuencia,

$$B_{GS} = -(L + D)^{-1}U = \begin{pmatrix} 0 & -2 & 2 \\ 0 & 2 & -3 \\ 0 & 0 & 2 \end{pmatrix}$$

y

$$\lambda I - B_{GS} = \begin{pmatrix} \lambda & 2 & -2 \\ 0 & \lambda - 2 & 3 \\ 0 & 0 & \lambda - 2 \end{pmatrix}$$

Los autovalores resultan ser

$$\lambda_1 = 0 \qquad \lambda_2 = 2 \qquad \lambda_3 = 2$$

entonces $\rho(B_{GS}) = 2$ y el método de Gauss-Seidel no converge.

Concluímos que en este ejemplo el método de Jacobi converge en tres pasos pero Gauss-Seidel no converge (existen datos iniciales para los cuales no converge).

Modificando trivialmente este ejemplo puede obtenerse un ejemplo en que ambos métodos convergen y se tiene $\rho(B_J) < \rho(B_{GS}) < 1$. Sea

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

que es estrictamente diagonal dominante y entonces Jacobi y Gauss-Seidel ambos convergen.

Veamos un ejemplo más.

Ejemplo 3.15. Este es un ejemplo para el cual ninguno de los dos métodos converge.

$$A = \begin{pmatrix} 2 & 1 & -1 \\ -2 & 2 & -2 \\ 1 & 1 & 2 \end{pmatrix}.$$

Aplicando Jacobi resulta

$$B_J = \begin{pmatrix} 0 & -\frac{1}{2} & \frac{1}{2} \\ 1 & 0 & 1 \\ -\frac{1}{2} & -\frac{1}{2} & 0 \end{pmatrix}$$

$$\lambda I - B_J = \begin{pmatrix} \lambda & \frac{1}{2} & -\frac{1}{2} \\ -1 & \lambda & -1 \\ \frac{1}{2} & \frac{1}{2} & \lambda \end{pmatrix}$$

con lo cual,

$$p(\lambda) = \lambda^3 + \frac{5}{4}\lambda$$

y los autovalores de B_J resultan

$$\lambda_1 = 0 \qquad \lambda_2 = i\frac{\sqrt{5}}{2} \qquad \lambda_3 = -i\frac{\sqrt{5}}{2}.$$

Entonces,

$$\rho(B_J) = \frac{\sqrt{5}}{2} > 1$$

y en consecuencia Jacobi no converge.

Para Gauss-Seidel se tiene

$$L + D = \begin{pmatrix} 2 & 0 & 0 \\ -2 & 2 & 0 \\ 1 & 1 & 2 \end{pmatrix}; \qquad (L+D)^{-1} = \begin{pmatrix} \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \\ -\frac{1}{2} & -\frac{1}{4} & \frac{1}{2} \end{pmatrix} \qquad \text{y}$$

$$-(L+D)^{-1}U = \frac{1}{4}\begin{pmatrix} 0 & -2 & 2 \\ 0 & -2 & 6 \\ 0 & 2 & -4 \end{pmatrix}$$

de donde $p(\lambda) = \lambda^3 + \frac{5}{4}\lambda^2 - \frac{1}{4}\lambda$ y calculando las raíces de $p(\lambda)$ obtenemos con aproximación los autovalores $\lambda_1 = 0, \lambda_2 = 0.1514$ y $\lambda_3 = -1.6514$ y por tanto

$$\rho(B_{GS}) = |\lambda_3| > 1$$

y entonces el método de Gauss-Seigel no converge.

EJEMPLO 3.16. CASO $\mathbb{R}^2$ En este caso es fácil analizar el método de Jacobi y el de Gauss-Seidel.

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$$

entonces

$$B_J = -D^{-1}(L+U) = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} \\ -\frac{a_{21}}{a_{22}} & 0 \end{pmatrix}$$

y

$$B_{GS} = -(D+L)^{-1}U = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} \\ 0 & \frac{a_{12}a_{21}}{a_{11}a_{22}} \end{pmatrix}.$$

Entonces,

$$\rho(B_J) = \sqrt{\frac{|a_{12}a_{21}|}{|a_{11}a_{22}|}}$$

$$\rho(B_{GS}) = \frac{|a_{12}a_{21}|}{|a_{11}a_{22}|}.$$

Es decir,

$$\rho(B_{GS}) = \rho(B_J)^2.$$

Como conclusión en $\mathbb{R}^2$ Jacobi converge si y solo si Gauss-Seidel converge. Y si convergen (o sea $\rho < 1$) es mejor asintóticamente Gauss-Seidel, pues en este caso $\rho(B_{GS}) < \rho(B_J)$.

Por ejemplo, si A es estrictamente diagonal dominante, entonces convergen los dos métodos. Y esto en $\mathbb{R}^2$ es decir $|a_{12}| < |a_{11}|$ y $|a_{21}| < |a_{22}|$.

Si A es simétrica y definida positiva entonces convergen ambos métodos en $\mathbb{R}^2$, pues

$$a_{11} > 0 \qquad a_{12} = a_{21} \qquad \det A = a_{11}a_{22} - a_{12}^2 > 0$$

entonces

$$a_{11}a_{22} > a_{12}^2$$

$$\frac{a_{12}^2}{a_{11}a_{22}} = \rho(B_{GS}) = \rho(B_J)^2 < 1.$$

El ejemplo anterior se generaliza para el método de Gauss-Seidel en $\mathbb{R}^N$ pero no para el método de Jacobi.

Veamos ahora el último ejemplo de esta serie.

EJEMPLO 3.17. Sea $A \in \mathbb{R}^3$ una matriz tridiagonal entonces,

$$\rho(B_{GS}) = \rho(B_J)^2$$

Si ponemos $A = \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{pmatrix}$ entonces $B_J = -D^{-1}(L+U) = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & 0 \\ -\frac{a_{21}}{a_{22}} & 0 & -\frac{a_{23}}{a_{22}} \\ 0 & -\frac{a_{32}}{a_{33}} & 0 \end{pmatrix}$ y

$$\det(\lambda I - B_J) = \lambda^3 - \lambda\left(\frac{a_{12}a_{21}}{a_{11}a_{22}} + \frac{a_{23}a_{32}}{a_{22}a_{33}}\right).$$

Y entonces

$$\rho(B_J) = \sqrt{\left|\frac{a_{12}a_{21}}{a_{11}a_{22}} + \frac{a_{23}a_{32}}{a_{22}a_{33}}\right|}.$$

Para el método de Gauss-Seidel se tiene

$$L + D = \begin{pmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & 0 \\ 0 & a_{32} & a_{33} \end{pmatrix} \qquad \text{y} \qquad U = \begin{pmatrix} 0 & a_{12} & 0 \\ 0 & 0 & a_{23} \\ 0 & 0 & 0 \end{pmatrix}.$$

Entonces

$$B_{GS} = -(L+D)^{-1}U = \begin{pmatrix} 0 & -\frac{a_{12}}{a_{11}} & 0 \\ 0 & \frac{a_{12}a_{21}}{a_{11}a_{22}} & -\frac{a_{23}}{a_{22}} \\ 0 & -\frac{a_{12}a_{21}a_{32}}{a_{11}a_{22}a_{33}} & \frac{a_{23}a_{32}}{a_{22}a_{33}} \end{pmatrix}$$

$$\det(\lambda I - B_{GS}) = \lambda^3 - \lambda^2\left(\frac{a_{12}a_{21}}{a_{11}a_{22}} + \frac{a_{23}a_{32}}{a_{22}a_{33}}\right).$$

Y los autovalores resultan ser

$$\lambda_1 = \lambda_2 = 0 \qquad \lambda_3 = \left(\frac{a_{12}a_{21}}{a_{11}a_{22}} + \frac{a_{23}a_{32}}{a_{22}a_{33}}\right).$$

Entonces,

$$\rho(B_{GS}) = \left|\frac{a_{12}a_{21}}{a_{11}a_{22}} + \frac{a_{23}a_{32}}{a_{22}a_{33}}\right| = \rho(B_J)^2$$

Los autovalores de Jacobi son los autovalores de $B_J = -D^{-1}(L+U)$ y resultan ser las raíces μ de

$$\begin{aligned} \det(\mu I + D^{-1}(L+U)) &= \det(D^{-1}(\mu D + L + U)) \\ &= \det(D^{-1})\det(\mu D + L + U) \end{aligned}$$

y como por hipótesis $\det(D^{-1}) \neq 0$ (asumimos $a_{ii} \neq 0$), μ son las raíces de

$$\det(\mu D + L + U).$$

Análogamente, los autovalores λ de $B_{GS} = -(L+D)^{-1}U$ son las raíces de

$$\begin{aligned}\det(\mu I + (L+D)^{-1}U) &= \det((L+D)^{-1}(\mu(L+D)+U)) \\ &= \det((L+D)^{-1})\det(\mu(L+D)+U)\end{aligned}$$

y como $\det((L+D)^{-1}) \neq 0$, λ son las raíces de

$$\det(\mu(L+D)+U).$$

LEMA 3.18. *Sea $A \in \mathbb{R}^{N\times N}$ tridiagonal entonces, para todo $\alpha \neq 0$ se tiene que*

$$\det(D+L+U) = \det(D+\alpha L+\alpha^{-1}U)$$

Demostración. Basta ver que las matrices $A = D+L+U$ y $D+\alpha L+\alpha^{-1}U$ son semejantes. Pongamos

$$A = \begin{pmatrix} d_1 & a_1 & 0 & \cdots & 0 \\ b_2 & d_2 & a_2 & \cdots & 0 \\ 0 & b_3 & d_3 & \ddots & \vdots \\ & & \ddots & \ddots & a_{N-1} \\ 0 & & \dots & b_{N-1} & d_N \end{pmatrix}$$

y consideremos

$$C = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \alpha & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \cdots & & \alpha^{N-1} \end{pmatrix}$$

Entonces,

$$CAC^{-1} = \begin{pmatrix} d_1 & \alpha^{-1}a_1 & 0 & \cdots & 0 \\ \alpha b_2 & d_2 & \alpha^{-1}a_2 & \cdots & 0 \\ 0 & \alpha b_3 & d_3 & \ddots & \vdots \\ & & \ddots & \ddots & \alpha^{-1}a_{N-1} \\ 0 & & \dots & \alpha b_{N-1} & d_N \end{pmatrix} = D+\alpha L+\alpha^{-1}U$$

□

TEOREMA 3.19. *Sea $A \in \mathbb{R}^{N\times N}$ una matriz tridiagonal y sean λ los autovalores no nulos de B_{GS}, μ los autovalores no nulos de B_J, entonces λ y μ se relacionan de la siguiente manera*

$$\lambda = \mu^2.$$

En particular,

$$\rho(B_{GS}) = \rho(B_J)^2.$$

Demostración. Los λ son autovalores de B_{GS} y por lo anterior son raíces de

$$\det(\lambda(L+D)+U)=0$$

pero $\lambda(L+D)+U$ es tridiagonal y por el lema anterior

$$\det(\lambda D+\alpha\lambda L+\alpha^{-1}U)=0$$

Si $\lambda\neq 0$ sea α tal que $\alpha^2=\frac{1}{\lambda}$. Entonces,

$$0=\alpha^{-N}\det(\lambda\alpha D+L+U)$$

y como los autovalores de B_J son las raíces μ de $\det(\mu D+L+U)=0$ resulta que

$$\mu=\lambda\alpha$$

Pero como $\alpha^2\frac{1}{\lambda}$ se tiene que

$$\mu^2=\lambda.$$

Ahora observemos que en lo anterior, dado $\lambda\neq 0$ autovalor de B_{GS} encontramos μ autovalor de B_J con $\mu^2=\lambda$, pero en realidad es un si y solo si. Es decir, dado μ autovalor de B_J, $\lambda=\mu^2$ resulta ser un autovalor de B_{GS},

$$\det(\mu D+L+U)=0$$

si y solo si

$$\det(\mu^2 D+\mu(L+U))=0$$

si y solo si (por el lema previo)

$$\det(\mu^2 D+\alpha\mu L+\alpha^{-1}\mu U)=0$$

y tomando $\alpha=\mu$ se tiene

$$\det(\mu^2(D+L)+U))=0.$$

Entonces μ^2 es autovalor de B_{GS}. □

Convergencia del método de Gauss-Seidel para $A\in\mathbb{R}^{N\times N}$ simétrica

Como los autovalores de $A\in\mathbb{R}^{N\times N}$ pueden ser complejos, trabajaremos directamente con $A\in\mathbb{C}^{N\times N}$.

Recordemos algunas definiciones

Definición 3.20. *Si $A\in\mathbb{C}^{N\times N}$ de define $A^*=\overline{A^T}$ o sea A^* es la matriz que en el lugar i,j tiene al elemento $a^*_{ij}=\overline{a_{ji}}$.*

Definición 3.21. *$A\in\mathbb{C}^{N\times N}$ es Hermitiana si*

$$A^*=A.$$

Observemos que si $A \in \mathbb{R}^{N\times N}$, $A^* = A^T$ y Hermitiana significa simétrica.

Si $z \in \mathbb{C}^N$ y A es Hermitiana entonces

$$z^*Az \in \mathbb{R}.$$

En general para A cualquiera $\overline{z^*Az} = z^*A^*z$. Veamos esto,

$$z^*Az = \sum_{ij} \overline{z_i} a_{ij} z_j$$

y entonces

$$\overline{z^*Az} = \sum_{ij} z_i \overline{a_{ij} z_j} = z^*A^*z$$

TEOREMA 3.22. *Sea $A \in \mathbb{C}^{N\times N}$ Hermitiana y definida positiva (o sea $z^*Az > 0\ \forall z \neq 0$), entonces el método de Gauss-Seidel es convergente.*

Demostración.

$$A = L + D + L^*$$

con

$$L = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ a_{21} & 0 & \cdots & 0 \\ \vdots & & \ddots & \\ a_{N1} & \cdots & & 0 \end{pmatrix}$$

Hay que ver que $\rho(B_{GS}) < 1$ donde $B_{GS} = -(L+D)^{-1}L^*$.

Observemos que B_{GS} puede escribirse como

$$B_{GS} = I - (L+D)^{-1}A$$

Sea $\lambda \in C$ un autovalor de B_{GS} y $z \in \mathbb{C}^N$, $z \neq 0$, un autovector correspondiente a λ, es decir

$$(I - (L+D)^{-1}A)z = \lambda z$$

o bien

$$(L+D)z - Az = \lambda(L+D)z$$

y esto es equivalente a

$$Az = (1-\lambda)(L+D)z$$

Como $Az \neq 0$ se deduce que $\lambda \neq 1$. Multiplicando por z^* se obtiene

$$\frac{1}{1-\lambda} = \frac{z^*(L+D)z}{z^*Az}$$

tomando conjugado y recordando que $z^*Az \in \mathbb{R}$ y que D es real se tiene

$$\frac{1}{1-\overline{\lambda}} = \frac{z^*(L+D)^*z}{z^*Az} = \frac{z^*(L^*+D)z}{z^*Az}.$$

Y sumando estas dos igualdades se obtiene

$$2Re(\frac{1}{1-\lambda}) = 1 + \frac{z^*Dz}{z^*Az} > 1$$

pues $z^*Az > 0$ y $z^*Dz > 0$ ($a_{ii} > 0$ si A es definida positiva).

Entonces si $\lambda = \alpha + i\beta$, tenemos

$$\frac{1}{1-\lambda} = \frac{1}{1-\alpha-i\beta}\frac{1-\alpha+i\beta}{1-\alpha+i\beta} = \frac{1-\alpha+i\beta}{(1-\alpha)^2+(\beta)^2}.$$

Y de esta forma

$$Re(\frac{1}{1-\lambda}) = \frac{1-\alpha}{(1-\alpha)^2+(\beta)^2}$$

y por lo anterior

$$\frac{1-\alpha}{(1-\alpha)^2+(\beta)^2} > \frac{1}{2}$$

es decir

$$2(1-\alpha) > 1 - 2\alpha + \alpha^2 + \beta^2$$

y esto es equivalente a

$$1 > \alpha^2 + \beta^2.$$

Hemos conseguido ver que

$$|\lambda| < 1.$$

Se puede probar que si $A \in \mathbb{C}^{N\times N}$ es Hermitiana y con $a_{ii} > 0$ entonces el método de Gauss-Seigel converge si y solo si A es definida positiva (ver Isaacson-Keller pag. 71). □

2.5. Método SOR. La idea del método SOR (succesive overrrelaxation / sobre relajación sucesiva) es tomar un "promedio" entre el x_i^k y el x_i^{k+1} de Gauss-Seidel (promedio entre comillas pues los pesos no son necesariamente menores o iguales que 1).

Dado ω un parámetro se define

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

En forma matricial escribimos como antes $A = L + D + U$ queda

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

entonces

$$x^{k+1} = B_\omega x^k + (D+\omega L)^{-1}\omega b$$

con

$$B_\omega = (D+\omega L)^{-1}((1-\omega)D - \omega U)$$

Observemos que $B_1 = B_{GS}$.

En principio, ω es arbitrario. Sin embargo el siguiente teorema nos dice que es necesario que $|\omega - 1| < 1$ para que haya convergencia (o sea para que $\rho(B_\omega) < 1$). Si $\omega \in \mathbb{R}$ entonces hace falta que $0 < \omega < 2$.

TEOREMA 3.23. *(Kahan) Sea $A \in \mathbb{C}^{N\times N}$, con $a_{ii} \neq 0$, entonces*

$$\rho(B_\omega) \geq |1-\omega|$$

Demostración. Si L es triangular inferior con ceros en la diagonal entonces $\det(D^{-1}) = \det((D+\omega L)^{-1})$. En consecuencia,

$$\begin{aligned}\det(B_\omega) &= \det((D+\omega L)^{-1})\det((1-\omega)D-\omega U)\\ &= \det(D^{-1})\det((1-\omega)D-\omega U)\\ &= \det((1-\omega)I-\omega D^{-1}U)\\ &= \det((1-\omega)I) = (1-\omega)^N\end{aligned}$$

Pero como $\det(B_\omega) = \prod_i \lambda_i$ se tiene que

$$\rho(B_\omega) \geq |1-\omega|$$

Si $\omega \in \mathbb{R}$, una condición necesaria para que el método converja es que $0 < \omega < 2$. Esta condición es también suficiente si A es simétrica definida positiva.

El problema consiste en encontrar el parámetro óptimo (o cercano al óptimo) para acelerar la convergencia. Para ciertas clases de matrices esto puede hacerse (ver libro Varga, Ortega o Smith). □

3. Ejercicios

(1) Escribir un programa que implemente el método de Jacobi y otro el de Gauss-Seidel con las siguientes condiciones:
- que incluya una restricción al número de iteraciones
- que finalice si el método se estaciona

(2) Decidir para cada uno de los siguientes sistemas, si los métodos de Jacobi y de Gauss-Seidel son convergentes (sugerencia: utilizar los comandos **tril**, **diag** y **eig** de `Matlab`). En caso afirmativo usarlos para resolver el sistema. Si ambos métodos convergen, determinar cuál converge más rápido. ¿Es la matriz del sistema diagonal dominante? ¿y simétrica y definida positiva?

(a) $\begin{pmatrix} 3 & 1 & 1 \\ 2 & 6 & 1 \\ 1 & 1 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 5 \\ 9 \\ 6 \end{pmatrix}$, (b) $\begin{pmatrix} 5 & 7 & 6 & 5 \\ 7 & 10 & 8 & 7 \\ 6 & 8 & 10 & 9 \\ 5 & 7 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 23 \\ 32 \\ 33 \\ 31 \end{pmatrix}$

(3) Dar ejemplos donde converja el método de Jacobi y no lo haga el de Gauss-Seidel y viceversa.

(4) Considerar el sistema $Ax = b$ para $A = \begin{pmatrix} 2 & 1 \\ 3 & 6 \end{pmatrix}$ y $b = (8, 21)$. Mostrar que el método de Jacobi converge; hacer un programa que lo modele y a la vez grafique en el plano la sucesión de aproximaciones obtenidas empezando en cada uno de lo siguientes valores iniciales

(a) $x_0 = (1, 4)$ (b) $x_0 = (1, 0)$ (c) $x_0 = (5, 2)$

(5) Considerar el sistema $\begin{cases} x - y = 0 \\ x + y = 0 \end{cases}$. Estudiar autovalores y autovectores de la matriz de iteración asociada al método de Gauss-Seidel, decidir si el método es convergente o

no y, sin hacer cálculos, predecir el comportamiento de las sucesiones que se obtienen con los siguientes valores iniciales.

(a) $x_0 = (2, 0)$ (b) $x_0 = (-0.03, 0.03)$ (c) $x_0 = (0, 1)$

Decidir si en este caso el método de Jacobi resulta convergente.

(6) (a) Mostrar que toda matriz $A \in \mathbb{R}^{n\times n}$ con $\det(A) > 1$ tiene un autovalor λ, real o complejo, con $|\lambda| > 1$.

(b) Decidir si el método de Jacobi converge o no para un sistema dado por la matriz

$$A = \begin{pmatrix} -1 & 1 & 2 \\ 4 & -1 & 3 \\ 5 & 6 & -1 \end{pmatrix}.$$

(7) Sean $A, B \in \mathbb{R}^{3\times 3}$ las matrices

$$A = \begin{pmatrix} a & c & 0 \\ c & a & c \\ 0 & c & a \end{pmatrix}; \qquad B = \begin{pmatrix} 0 & b & 0 \\ b & 0 & b \\ 0 & b & 0 \end{pmatrix}.$$

(a) Probar que $\lim_{n\to\infty} B^n = 0$ si y sólo si $|b| < \sqrt{2}/2$.

(b) Dar condiciones necesarias y suficientes sobre $a, c \in \mathbb{R}$ para la convergencia de los métodos de Jacobi y de Gauss-Seidel aplicados a la resolución de $Ax = v$.

(8) (a) Probar que si A tiene una base de autovectores v_i, con autovalores λ_i, la matriz

$$B = I + sA, \quad s \in \mathbb{R}$$

tiene los mismos autovectores, con autovalores $\nu_i = 1 + s\lambda_i$.

(b) Sabiendo que los autovalores de la matriz $A \in \mathbb{R}^{(n-1)\times(n-1)}$

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

son $\lambda_j = -4 \sin \frac{\pi j}{2n}$, $j = 1, \ldots, n-1$, decidir si el método de Jacobi aplicado a $Ax = b$ es convergente o no.

(c) Decidir si el método de Gauss-Seidel resulta convergente. En caso afirmativo, ¿qué método converge más rápido?

Comentario: Este problema es interesante por sus aplicaciones, pues corresponde a la discretización de la ecuación de Poisson en una dimensión espacial:

$$\begin{cases} \dfrac{d^2u}{dx^2} = f(x), & x \in [0, 1]; \\ \\ u(0) = u(1) = 0. \end{cases}$$

(9) Sea B_J la matriz asociada al método de Jacobi de un sistema dado. Estimar

(a) cuántas multiplicaciones y divisiones se requieren para calcular B_J.

(b) cuántas multiplicaciones y divisiones se requieren para para realizar una iteración con el método de Jacobi.

(c) si $\rho(B_J) < 1$, cuántas iteraciones se necesitan para reducir el error del método en más de 10^{-m} (en función de $\rho(B_J)$).
(d) cuántas multiplicaciones y divisiones se requieren para calcular la solución del sistema por el método de eliminación gaussiana.
(e) cuántas iteraciones del método de Jacobi podrían realizarse antes de igualar la cantidad de operaciones necesarias al usar el método de eliminación gaussiana.

(10) Sean B_J y B_{GS} las matrices asociadas al método de Jacobi y de Gauss-Seidel respectivamente del sistema $Ax = b$.
(a) Mostrar que si $A(i,k) = 0$ entonces, el elemento $B_J(i,k) = 0$. Notar que si A es una matriz rala (con muchos ceros) entonces B_J también lo es. Luego, en cada iteración se requieren pocas multiplicaciones.
(b) Mostrar que $\lambda = 0$ siempre es un autovalor de B_{GS}. ¿De qué autovector?

(11) Dada una matriz

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix}$$

y un vector $b \in \mathbb{R}^3$, se quiere resolver el sistema de ecuaciones $Ax = b$; para lo cual se considera el siguiente método iterativo, que es un caso particular de los métodos llamados *Jacobi por bloques*:

$$x_{k+1} = -\begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}^{-1} \cdot \begin{pmatrix} 0 & 0 & a_{13} \\ 0 & 0 & a_{23} \\ a_{31} & a_{32} & 0 \end{pmatrix} \cdot x_k + \begin{pmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & 0 & a_{33} \end{pmatrix}^{-1} \cdot b,$$

Este método resulta convergente para los siguientes datos:

$$A = \begin{pmatrix} 8 & 2 & -3 \\ -3 & 9 & 4 \\ 3 & -1 & 7 \end{pmatrix} \quad \text{y} \quad b = \begin{pmatrix} -20 \\ 62 \\ 0 \end{pmatrix}.$$

Hacer un programa que calcule la sucesión de aproximaciones generada con valor inicial el vector nulo y que se detenga cuando $\|x_{k+1} - x_k\|_\infty \leq 10^{-4}$ (es decir, cuando la iteración "se estabiliza").

(12) Sea $A \in \mathbb{R}^{n\times n}$. Probar que $\lambda = 1$ es autovalor de la matriz de Jacobi (o Gauss-Seidel) de A si y solo si A es no inversible.

(13) Para resolver el sistema $Ax = b$, se utiliza un método iterativo cuya matriz de iteración J es diagonalizable y satisface $\rho(J) < 1$. Sea e_k el vector error en el k-ésimo paso.
(a) Demostrar que $\|e_k\|_\infty = O(\rho(J)^k)$.
(b) Probar que si $e_k \neq 0$ para todo $k \in \mathbb{N}$ y $\rho(J) \neq 0$, la sucesión $(\|e_k\|_\infty)_{k\in\mathbb{N}}$ tiende a 0 linealmente.

(14) Utilizar la iteración de Gauss-Seidel para resolver el sistema $A_n x = b_n$ para

$$A_n = \begin{pmatrix} 1 & 2 \\ 2 & 4 + \frac{1}{n^2} \end{pmatrix} \qquad \text{y} \qquad b_n = (1, 2 - \frac{1}{n^2}).$$

¿Cómo es la convergencia? ¿Tiene ésto que ver con el mal condicionamiento de A? Dar un ejemplo de una matriz mal condicionada para la cual la convergencia sea rápida.

(15) Hacer un programa que pida el ingreso de una matriz A y un vector b y luego
- calcule las matrices de iteración de los métodos de Jacobi y Gauss-Seidel.

- calcule el menor de los radios espectrales de las dos matrices anteriores y, si este valor resulta menor a 1, entonces realice las primeras 10 iteraciones del método correspondiente (o de cualquiera de los dos métodos en caso de que los radios espectrales resulten coincidentes), con valor inicial el vector nulo.

(16) Considerar el sistema $Ax = b$ para $A = \begin{pmatrix} 64 & -6 \\ 6 & -1 \end{pmatrix}$ y $b = (1, 2)$.

(a) Demostrar que el método de Jacobi converge para todo dato inicial. Verificar, sin embargo, que la matriz no es diagonal dominante.

(b) Sea J la matriz de iteración. Hallar las normas 1, ∞ y 2 de J. Hallar una norma $\| \ \|$ en la cual $\|J\|$ sea < 1.

CAPíTULO 4

Resolución de ecuaciones no lineales.

En muchos problemas, aparece en forma natural, la necesidad de calcular el valor de x donde una función f se anula, es decir, una raíz de f. En general, con las herramientas analíticas que se usan para estudiar y graficar funciones suaves (derivables) sólo podemos analizar si hay un intervalo $[a, b]$ donde el gráfico de f cruza el eje x.

En este capítulo, veremos distintos métodos que nos permitirán aproximar el valor de una raíz, éste valor suele hallarse por aproximaciones sucesivas y por ende los métodos a utilizar son iterativos. En muchas ocasiones, sólo tiene sentido encontrar una solución aproximada. A veces, el cálculo exacto no es posible ya sea porque se trata de una raíz irracional ($f(x) = x^2 - 2$) o porque la función viene dada por coeficientes cuyos valores se conocen sólo en forma aproximada. Lo importante al utilizar métodos que estimen el valor deseado es, como venimos remarcando en estas notas, poder controlar el error que se comete al utilizar un valor aproximado en lugar del exacto.

El problema se plantea de la siguiente manera: Dada $f : \mathbb{R} \to \mathbb{R}$ (o bien $f : [a, b] \to \mathbb{R}$) se quiere encontrar r tal que

$$f(r) = 0.$$

El cálculo aproximado de raíces puede dividirse en dos etapas. En la primera, se separan las raíces. Es decir, se busca un subintervalo de $[a, b]$ que contenga una y sólo una raíz de f. Para asegurar la existencia de al menos una raíz en el intervalo propuesto se utiliza el teorema de Bolzano. Para asegurar que no hay más de una raíz se usa el teorema de Rolle, es decir, se verifica que la derivada primera no cambie de signo en dicho intervalo. En la segunda etapa, se aplica un método para aproximar la raíz aislada.

Antes de describir el primer método de estas notas, el de bisección, recordamos el teorema de Bolzano.

TEOREMA 4.1. **Bolzano** *Sea $f : [a, b] \to \mathbb{R}$ continua en $[a, b]$. Si $f(a)f(b) < 0$ (o sea $f(a)$ y $f(b)$ tienen distinto signo) entonces existe alguna raíz de f en el intervalo $[a, b]$.*

1. Método de bisección.

Este método, que se apoya en la idea geométrica del teorema de Bolzano, permite construir una sucesión $(x_n)_{n\in\mathbb{N}}$ que converge a la solución de $f(x) = 0$ de la siguiente manera.

Supongamos que $f(a)f(b) < 0$. Calculemos $c = \frac{a+b}{2}$. Supongamos, por ejemplo, que $f(a) > 0$ y $f(b) < 0$, entonces

(1) Si $f(c) = 0$ listo.
(2) Si $f(c) < 0$, habrá una raíz en $[a, c]$.
(3) Si $f(c) > 0$, habrá una raíz en $[c, b]$.

Ahora se elige el subintervalo, cuya longitud es la mitad de $[a, b]$ y que contiene a la raíz. Este proceso se sigue sucesivamente.

Así se genera una sucesión $x_1 = \frac{a+b}{2} \in [a_1, b_1]$, $x_2 \in [a_2, b_2]$, $x_3 \in [a_3, b_3]$ ... donde cada intervalo $[a_n, b_n]$ mide la mitad del anterior,

$$\begin{aligned} b_1 - a_1 &= \frac{b-a}{2} \\ b_2 - a_2 &= \frac{b_1 - a_1}{2} = \frac{b-a}{4} \\ &\vdots \\ b_n - a_n &= \cdots = \frac{b-a}{2^n} \end{aligned}$$

Además,

$$a \leq a_1 \leq a_2 \leq \quad \ldots \quad \leq b$$

$$b \geq b_1 \geq b_2 \geq \quad \ldots \quad \geq a$$

Entonces a_n y b_n son sucesiones monótonas y acotadas y en consecuencia convergen, es decir, existen los límites

$$\lim_{n\to\infty} a_n \qquad \text{y} \qquad \lim_{n\to\infty} b_n.$$

Y como

$$|b_n - a_n| \leq \frac{b-a}{2^n} \to 0$$

se tiene que

$$\lim_{n\to\infty} a_n = \lim_{n\to\infty} b_n = r.$$

En cada paso se verifica $f(a_n)f(b_n) \leq 0$ y tomando límite (usando que f es continua) resulta

$$f(r)^2 \leq 0.$$

Entonces r es la raíz buscada pues cumple, $f(r) = 0$.

Por otra parte el error puede acotarse de la siguiente forma. Tenemos que

$$x_n = \frac{a_{n-1} + b_{n-1}}{2}$$

entonces

$$|r - x_n| \leq \frac{1}{2}(b_{n-1} - a_{n-1}) \leq \frac{b-a}{2^n}.$$

Resumiendo, hemos demostrado,

TEOREMA 4.2. *Si $f : [a, b] \to \mathbb{R}$ es continua y $f(a)f(b) < 0$ entonces, el método de bisección genera una sucesión x_n tal que,*

(1) $x_n \to r$ *con* $f(r) = 0$,
(2) $|r - x_n| \leq \dfrac{b-a}{2^n}$.

Una de las ventajas que tiene el método de bisección es que converge para cualquier f continua, es decir no hace falta derivabilidad como en otros métodos que veremos mas adelante.

EJEMPLO 4.3. Calculemos $\sqrt{2}$.

Tomemos $f(x) = x^2 - 2$ y $[a, b] = [1, 3]$. Se tiene $f(1) = -1 < 0 < f(3) = 7$ y con un gráfico de f podemos asegurar que no hay otra raíz positiva. La suecsión que produce el método es:

$x_1 = 2$	$f(x_1) = 2$	$[a_1, b_1] = [1, 2]$
$x_2 = 1.5$	$f(x_2) = 0.25$	$[a_2, b_2] = [1, 1.5]$
$x_3 = 1.25$	$f(x_3) = -0.4375$	$[a_3, b_3] = [1.25, 1.5]$
$x_4 = 1.375$	$f(x_4) = -0.109375$	$[a_4, b_4] = [1.375, 1.5]$
$x_5 = 1.4375$	$f(x_5) = 0.06640625$	$[a_5, b_5] = [1.375, 1.4375]$
$x_6 = 1.40625$	$f(x_6) = -0.022\ldots$	$[a_6, b_6] = [1.40625, 1.4375]$
$x_7 = 1.421875$	$f(x_7) = 0.02\ldots$	$[a_7, b_7] = [1.40625, 1.421875]$
$x_8 = 1.4140625$	$\ldots$	

Para x_8, vemos que la aproximación lograda tiene 4 cifras exactas. Fue necesario hacer ocho pasos para obtener cuatro cifras exactas ($\sqrt{2} = 1.4142\ldots$).

Del análisis hecho en general sabíamos que,

$$|\sqrt{2} - x_8| \leq \frac{b-a}{2^8} = \frac{2}{2^8} = \frac{1}{128}.$$

Entonces el error relativo es

$$\frac{|\sqrt{2} - x_8|}{\sqrt{2}} \leq \frac{1}{128\sqrt{2}} \leq 0.005\ldots \sim \frac{5}{1000}.$$

La desventaja del método de bisección es que converge muy lentamente, por ejemplo en comparación con el método de Newton-Raphson que veremos más adelante.

En cada paso la cota del error, $(b-a)/2^n$, se reduce a la mitad,

$$|e_{n+1}| \leq \frac{b-a}{2^n}.$$

En consecuencia se reduce $\frac{1}{10}$ en tres o cuatro pasos (se gana una cifra en tres o cuatro pasos).

2. Método regula falsi

Este método llamado "regula falsi" o de falsa posición puede verse tanto como una variante del método de bisección como del método Newton-Raphson, que veremos en la próxima sección.

Supongamos, nuevamente, que tenemos una función $f : [a,b] \to \mathbb{R}$ continua que verifica $f(a)f(b) < 0$ (entonces existe una raíz, r, en $[a,b]$, por el teorema de Bolzano) y supongamos que la raíz es única en ese intervalo.

Definimos x_1 como la intersección de la recta secante L con el eje x (en lugar de tomar el promedio $\frac{b-a}{2}$, como se hace con el método de bisección).

La recta L, que une los puntos $(a, f(a))$ con $(b, f(b))$ tiene ecuación:

$$y - f(a) = \frac{f(b) - f(a)}{b-a}(x-a).$$

Como x_1 es el valor de x que cumple $y = 0$, se tiene,

$$x_1 = a - \frac{f(a)}{f(b) - f(a)}(b-a) = \frac{af(b) - bf(a)}{f(b) - f(a)}$$

Si $f(x_1) \neq 0$ entonces $f(a)f(x_1) < 0$ o bien $f(b)f(x_1) < 0$. Supongamos $f(b)f(x_1) < 0$, definimos x_2 con el mismo procedimiento anterior con el intervalo $[x_1, b] = I_1$, y así sucesivamente.

Observemos que puede suceder que $|I_n|$ no tienda a cero, pero sin embargo $x_n \to r$ para toda f continua.

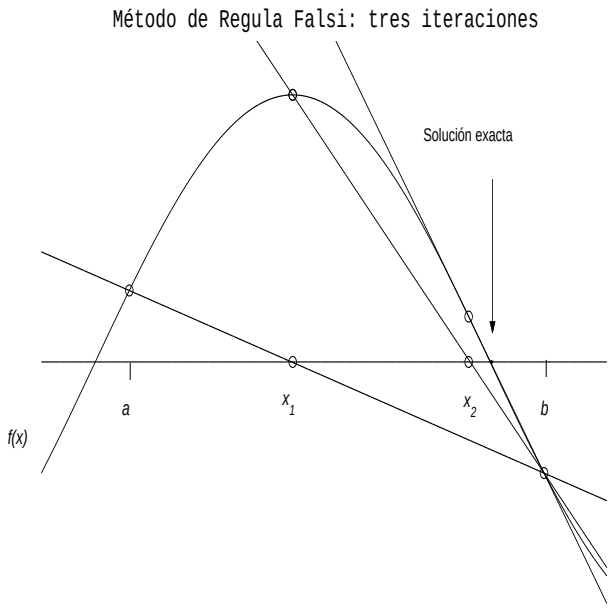

FIGURA 4.1.

3. Método de Newton-Raphson.

La idea del método es "ir por la tangente" como se describe a continuación.

Se empieza con x_0. Se traza la tangente en x_0 y se define x_1 como la intersección de la tangente con el eje x. Luego se traza la tangente por x_1 y se toma x_2 la intersección de la tangente con el eje x, y así sucesivamente. Esto genera una sucesión x_n como muestra la Figura 4.2.

Observemos que hace falta que f sea derivable. Además, puede ocurrir que la sucesión que produce este método no sea convergente. Esto último se puede ver gráficamente con el ejemplo que muestra la Figura 4.3.

Sin embargo veremos que el método converge muy rápidamente si x_0 está "suficientemente cerca" de una raíz, bajo condiciones bastante generales sobre la función f.

Descripción analítica de método de Newton-Raphson.

Sea $f : [a, b] \to \mathbb{R}$ derivable, $x_0 \in [a, b]$, se toma x_1 tal que

$$f(x_0) + (x_1 - x_0)f'(x_0) = 0$$

Y en general, se toma x_{n+1} tal que

$$f(x_n) + (x_{n+1} - x_n)f'(x_n) = 0$$

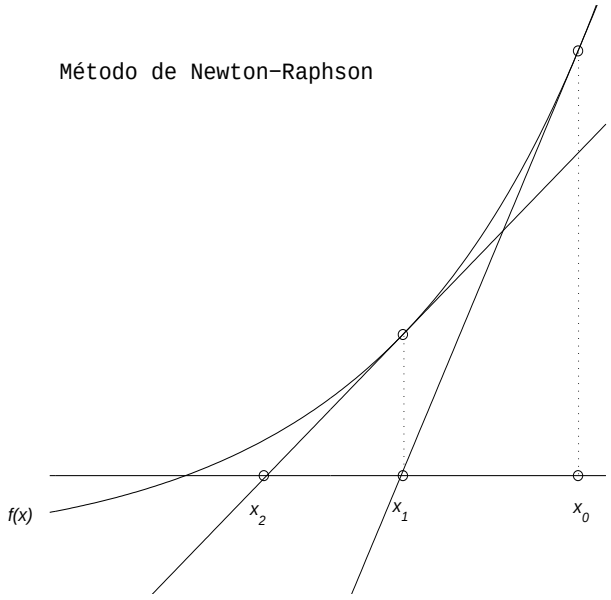

FIGURA 4.2.

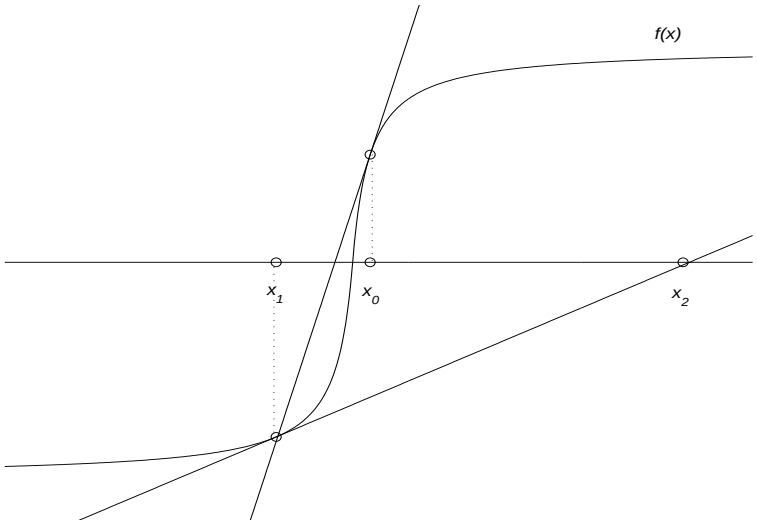

FIGURA 4.3.

o sea,

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

Observemos que para que esto tenga sentido, hay que suponer $f'(x_n) \neq 0$, esto es obvio gráficamente como muestra la figura 4.3.

Ahora analicemos la convergencia del método. Sea r una raíz simple de f, es decir, $f(r) = 0$, $f'(r) \neq 0$ y supongamos que f'' es acotada.

Debemos estimar el error que se comete al usar x_n en lugar de la solución exacta (y desconocida) r. Esto es, estudiamos la expresión $e_n = x_n - r$ y vemos si $e_n \to 0$.

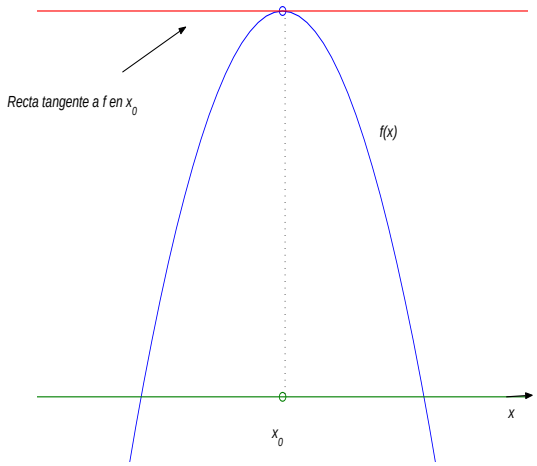

FIGURA 4.4.

Para analizar la convergencia del error miramos la sucesión recursiva

$$e_{n+1} = x_{n+1} - r = x_n - \frac{f(x_n)}{f'(x_n)} - r = e_n - \frac{f(x_n)}{f'(x_n)}$$

entonces

$$e_{n+1} = \frac{e_n f'(x_n) - f(x_n)}{f'(x_n)} \tag{4.1}$$

Observemos que si $f'(r) \neq 0$ entonces $f'(x_n) \neq 0$ para x_n cercano a r (esto lo justificaremos con más precisión después).

Usando el desarrollo de Taylor de orden 2 centrado en la raíz r se tiene,

$$0 = f(r) = f(x_n) - (x_n - r)f'(x_n) + \frac{1}{2}(x_n - r)^2 f''(\xi)$$

donde ξ es un valor intermedio entre x_n y r. Entonces

$$e_n f'(x_n) - f(x_n) = \frac{1}{2} f''(\xi) e_n^2$$

Reemplazando en la igualdad 4.1 queda

$$e_{n+1} = \frac{1}{2} \frac{f''(\xi)}{f'(x_n)} e_n^2 \tag{4.2}$$

Con todo esto podemos demostrar el siguiente teorema.

TEOREMA 4.4. **(de convergencia)** *Si r es un cero simple de f (i.e. $f'(r) \neq 0$) y sea $I = [r - \alpha, r + \alpha]$ un intervalo tal que $|f'(x)| \geq \delta > 0$ y $|f''(x)| \leq M$ en I. Entonces,*

Existe $\varepsilon > 0$ tal que $I_\varepsilon = [r - \varepsilon, r + \varepsilon] \subset I$ y se tiene que $|e_n| \to 0$ y

$$|e_{n+1}| \leq \frac{1}{2} \frac{M}{\delta} |e_n|^2, \tag{4.3}$$

siempre que $x_0 \in I_\varepsilon$.

Demostración. Como las cotas para f' y f'' siguen siendo ciertas para cualquier subintervalo de I, podemos elegir $\varepsilon > 0$ tal que

$$\frac{1}{2} \frac{M}{\delta} \varepsilon = \lambda < 1.$$

Entonces, si $x_0 \in I_\varepsilon$ tenemos que $|e_0| = |x_0 - r| < \varepsilon$ y usando (4.2) obtenemos

$$|e_1| = |x_1 - r| \leq \lambda |e_0|.$$

En particular, $x_1 \in I_\varepsilon$. Análogamente,

$$|e_2| = |x_2 - r| \leq \lambda |e_1| \leq \lambda^2 |e_0|$$

y $x_2 \in I_\varepsilon$. Continuando de esta manera, obtenemos una sucesión $(x_n)_{n \in \mathbb{N}} \subset I_\varepsilon$ tal que

$$|e_n| \leq \lambda^n |e_0|.$$

Como $0 < \lambda < 1$ se tiene que $|e_n| \to 0$ si $n \to \infty$. Finalmente, la desigualdad (4.3) se obtiene de (4.2). □

COROLARIO 4.5. *Si f' es continua y f'' es acotada en $[a, b]$ y $r \in [a, b]$ es una raíz simple de f, entonces existe un $\varepsilon > 0$ tal que si $x_0 \in I_\varepsilon = [r - \varepsilon, r + \varepsilon] \subset [a, b]$, el método de Newton empezando en x_0 converge a r.*

Demostración. Como $f'(r) \neq 0$ y f' es continua, existen $\alpha > 0$ y $\delta > 0$ tales que $I = [r-\alpha, r+\alpha] \subset [a,b]$ y $|f'(x)| > \delta$ para todo $x \in I$. Ahora estamos en las condiciones del teorema 4.4. □

OBSERVACIÓN 4.6. *Un caso particular del corolario 4.5 es una función $C^2([a,b])$ que tiene a $r \in [a,b]$ como raíz simple.*

Ahora, queremos estudiar la rápidez con la que una sucesión generada por un método, converge a la solución exacta. Para eso necesitamos la siguiente

DEFINICIÓN 4.7. *En general podemos definir que un método es de orden p si*

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = cte \qquad y \qquad \lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^{p-\varepsilon}} = 0$$

Observemos primero que cuanto más grande sea p mejor. Ahora, veamos que significa ésto geométricamente. Para valores grandes de n, es decir, asintóticamente, se puede considerar que el comportamiento de las sucesiones $|e_{n+1}|$ y $|e_n|^p$ son equivalentes, lo que se expresa como

$$|e_{n+1}| \sim C|e_n|^p.$$

Por otra parte, si se obtiene una desigualdad de la forma

$$|e_{n+1}| \leq C|e_n|^p$$

podemos asegurar que el orden de convergencia es por lo menos p.

La convergencia para el método de Newton-Raphson es cuadrática, es decir, $p = 2$. Si bien, con la desigualdad (4.3) podemos asegurar que existe $C > 0$ tal que

$$|e_{n+1}| \leq C|e_n|^2$$

de la igualdad (4.2) se deduce que el método, en general, converge cuadráticamente. Esto es, en cada paso el error se reduce cuadráticamente (o sea es menor o igual que el cuadrado del error del paso anterior).

Esta es la gran ventaja del método de Newton. El número de cifras correctas se duplica (esencialmente) en un paso.

Este resultado de convergencia es "local", o sea, el teorema garantiza la convergencia si se empieza "suficientemente cerca" de r. En la práctica es un tema difícil determinar lo que es "suficientemente cerca". Muchas veces, se combinan unos pasos del método de bisección para encontrar un intervalo en el que se aplique el Teorema 4.4. Sin embargo, el método de Newton funciona en forma excelente (incluso en N variables) y es de los más usados.

Ejemplo 4.8. Calculemos, aplicando el método de Newton una aproximación de $\sqrt{2}$. Comparemos el resultado con el que se obtuvo al aplicar el método de bisección. Como antes la función es $f(x) = x^2 - 2$ y elegimos $x_0 = 3$. Tenemos

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n - \frac{x_n^2 - 2}{2x_n} = \frac{x_n}{2} + \frac{1}{x_n} \tag{4.4}$$

Y aplicando esto obtenemos

$$x_0 = 3 \qquad\qquad x_3 = 1.41499843\ldots$$

$$x_1 = 1.833\ldots \qquad\qquad x_4 = 1.41421378\ldots$$

$$x_2 = 1.4621212\ldots \qquad\qquad x_5 = 1.414213562\ldots$$

Observemos que

$$\sqrt{2} = 1.414213562\ldots$$

Es decir, con cinco pasos del método tenemos más de diez cifras exactas, mientras que con bisección en ocho pasos teníamos cuatro cifras exactas.

Comentario. Hacia el año 2000 a.C. los Babilonios usaban el siguiente método para "calcular" el número $\sqrt{p}$ si $p \in \mathbb{N}$. Si $a > \sqrt{p}$ se tiene que $\frac{p}{a} < \sqrt{p}$. Luego $\sqrt{p}$ es un número entre $\frac{p}{a}$ y a. Entonces, consideraban el promedio $\frac{1}{2}(a + \frac{p}{a})$ como primera aproximación, así sucesivamente. Esto coincide con el método de Newton, de 1669 d.C., aplicado a la función $x^2 - p$. Comparar con (4.4).

Ejemplo 4.9. Como segundo ejemplo veamos que sucede con $f(x) = x^3$, $r = 0$. Es claro que la única raíz es $r = 0$. Lo que se pretende con este ejemplo es mostrar alguna de las dificultades a tener en cuenta cuando se aplica el método de Newton. La sucesión que produce este método es:

$$x_{n+1} = x_n - \frac{x_n^3}{3x_n^2} = \frac{2}{3}x_n$$

Entonces

$$|e_{n+1}| = \frac{2}{3}|e_n|$$

En este caso, observamos que la convergencia es lineal y no es cuadrática. Lo que sucede es que no se verifica la hipótesis de que r sea una raíz simple ($f'(r) = 0$ en este caso).

Ejemplo 4.10. Este es un ejemplo donde el método de Newton-Raphson no converge. En este caso, la hipótesis que no se cumple es la derivabilidad de f. Consideremos la función

$$f(x) = \begin{cases} \sqrt{x} & x \geq 0 \\ -\sqrt{-x} & x < 0, \end{cases}$$

con $r = 0$.

Un cálculo sencillo permite ver que f no es derivable en la raíz. En cualquier otro valor se tiene

$$f'(x) = \begin{cases} \frac{1}{2}x^{-\frac{1}{2}} & x > 0 \\ \frac{1}{2}(-x)^{-\frac{1}{2}} & x < 0. \end{cases}$$

Es decir,

$$f'(x) = \frac{1}{2}|x|^{-\frac{1}{2}}.$$

La sueción que produce el método se escribe como

$$x_{n+1} = x_n - \frac{x_n^{\frac{1}{2}}}{\frac{1}{2}x_n^{-\frac{1}{2}}} = x_n - 2x_n = -x_n.$$

Ahora, salvo que comencemos en la raíz (con lo cual no necesitaríamos de un método para hallarla) se tiene que x_n es positivo o negativo.

Supongamos que $x_n > 0$, entonces $x_{n+1} = -x_n < 0$ y $x_{n+2} = -x_{n+1} > 0$.

Si seguimos el proceso que genera x_n desde un x_0 inicial vemos que la sucesión es:

$$x_0 \quad \to \quad -x_0 \quad \to \quad x_0 \quad \to \quad -x_0 \quad \to \ldots$$

Concluímos que, en este ejemplo, el método de Newton no converge para ningún x_0 por más cerca de $r = 0$ que esté.

Ahora veamos un teorema de convergencia global para el método de Newton que se aplica a funciones convexas. Una función se dice convexa en (a, b) si la recta tangente al gráfico de f está por debajo de éste para todo los x en el intervalo. Si la función es dos veces derivable esto corresponde con la condición $f'' > 0$. En este caso, f puede tener un valor mínimo. Digamos, si existe, que lo alcanza en x^*.

TEOREMA 4.11. *Sea f dos veces derivable en $[a,b]$ tal que $f'' > 0$ (f es convexa), entonces el método de Newton-Raphson converge para todo $x_0 \neq x^*$. Es decir, en este caso no hace falta pedir que x_0 esté cerca de r.*

Demostración. Si pérdida de generalidad podemos suponer que f es "monótona" (si estamos a la derecha de x^*, la iteración de Newton nunca irá a la izquierda, ver figura 4.2).

Si $x_0 > r$ entonces $r < x_1 < x_0$ y en general

$$x_0 > x_1 > x_2 > .... > x_n > .... > r$$

y entonces la sucesión x_n converge pues es monótona. Veamos que converge a una raíz de f. Supongamos que $x_n \to \alpha$, luego tomando límite en la expresión $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$ y usando que f' es continua queda

$$\alpha = \alpha - \frac{f(\alpha)}{f'(\alpha)}$$

de dónde $f(\alpha) = 0$ y $\alpha = r$ pues supusimos f monótona. □

Si bien este teorema es bastante claro geométricamente para funciones definidas en $\mathbb{R}$, su interés radica en su extensión a $\mathbb{R}^N$.

4. Método de punto fijo

El método de Newton puede verse como un caso particular del método de punto fijo.

La idea es reemplazar la ecuación $f(x) = 0$ por otra de la forma $x = g(x)$ de manera que la solución de ésta sea la solución del problema original.

Esto puede hacerse de diversas maneras, por ejemplo, si

$$f(x) = x^3 - 13x + 18$$

podemos tomar $g(x)$ como cualquiera de las siguientes funciones

$$g_1(x) = \frac{x^3 + 18}{13}, \qquad g_2(x) = (13x - 18)^{\frac{1}{3}}, \qquad g_3(x) = \frac{13x - 18}{x^2}.$$

Una vez encontrada g una función continua, el problema se reduje a encontrar puntos fijos de g, es decir, r tales que

$$r = g(r).$$

Se define una sucesión por iteración, se elige un x_0 y después se toma

$$x_{n+1} = g(x_n). \tag{4.5}$$

Observemos que si la sucesión generada por (4.5) x_n converge, entonces lo hace a un punto fijo de g. En efecto, tomando límite y usando que g es continua se tiene que si $x_n \to r$ entonces $r = g(r)$.

TEOREMA 4.12. *Sea $I = [a, b]$ si $g(I) \subset I$ entonces g tiene al menos un punto fijo en I.*

Demostración. Como $g(I) \subset I$ se tiene que $a \leq g(a) \leq b$ y $a \leq g(b) \leq b$, si $a = g(a)$ o $b = g(b)$ listo. Si no, $g(a) - a > 0$ y $g(b) - b < 0$. Entonces la función $F(x) = g(x) - x$ cumple, $F(a) > 0$ y $F(b) < 0$ y como F es continua existe un r en I tal que $0 = F(r) = g(r) - r$. □

TEOREMA 4.13. *Si g es además derivable y $|g'(x)| \leq \lambda < 1$ $\forall x \in I$ y $g(I) \subset I$ entonces g tiene un único punto fijo.*

Demostración. Si hay dos puntos fijos, r_1, r_2 con $r_1 \neq r_2$, tenemos

$$|r_1 - r_2| = |g(r_1) - g(r_2)| = |g'(\xi)(r_1 - r_2)| \leq \lambda |r_1 - r_2| < |r_1 - r_2|$$

una contradicción. □

Bajo estas mismas hipótesis, la sucesión generada iterativamente converge y se puede dar una cota del error en términos de λ.

TEOREMA 4.14. *Sea g tal que $|g'(x)| \leq \lambda < 1$ $\forall x \in I$ y $g(I) \subset I$ entonces la sucesión x_n definida por*

$$x_{n+1} = g(x_n)$$

converge al único punto fijo de g y además,

1. $|x_n - r| \leq \lambda^n |x_0 - r|$
2. $|e_n| \leq \frac{\lambda^n}{1-\lambda} |x_1 - x_0|$. *O sea, se tiene una acotación en términos de $|x_1 - x_0|$ que es conocido.*

Demostración. Por el teorema anterior sabemos que existe un único punto fijo de g que llamamos r. La hipótesis sobre la derivada de g implica que $|g(x) - g(y)| \leq \lambda |x - y|$, o sea g es Lipschitz con constante λ. Entonces

$$|x_{n+1} - r| = |g(x_n) - g(r)| \leq \lambda |x_n - r|$$

y de aquí, como $\lambda < 1$ se tiene que

$$|x_n - r| \leq \lambda^n |x_0 - r| \to 0.$$

En particular demostramos que $x_n \to r$.

Por otra parte, intercalando x_1 y usando desigualdad triangular,

$$|x_0 - r| \leq |x_0 - x_1| + |x_1 - r| \leq |x_0 - x_1| + \lambda|x_0 - r|.$$

Entonces

$$(1 - \lambda)|x_0 - r| \leq |x_1 - x_0|$$

y como

$$|x_n - r| \leq \lambda^n |x_0 - r|$$

se obtine la estimación *2)*. □

La figura 4.5 muestra gráficamente como se genera una sucesión por el método de punto fijo. En dicho gráfico $0 < f'(x) < 1$.

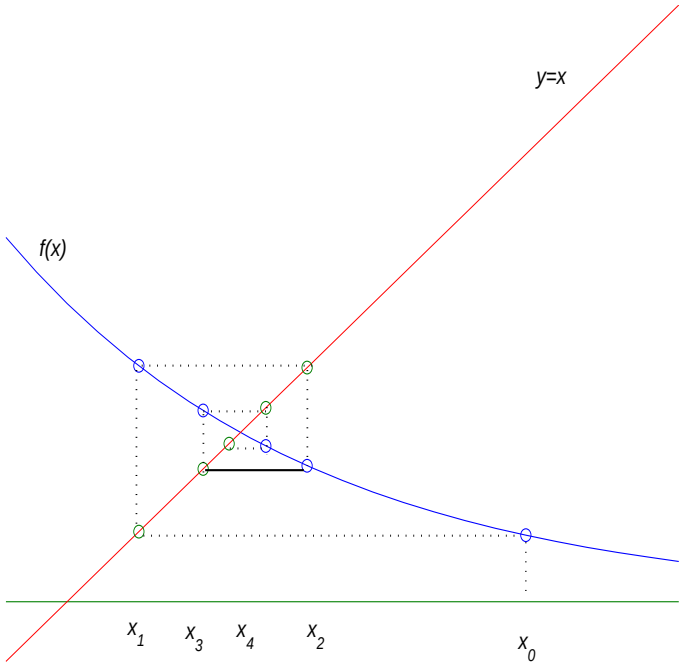

FIGURA 4.5.

Para aplicar el teorema 4.14 hay que garantizar que $g(I) \subset I$ (o sea primero hay que encontrar un tal I).

Si r es un punto fijo de g con $|g'(r)| < 1$ este intervalo I existe, resultado que probamos en el siguiente teorema.

TEOREMA 4.15. *g' continua en (a,b), $r \in (a,b)$ un punto fijo de g. Si $|g'(r)| < 1$, entonces existe $\varepsilon > 0$ tal que la iteración es convergente siempre que $x_0 \in I_\varepsilon = (r-\varepsilon, r+\varepsilon)$.*

Demostración. Como $|g'(r)| < 1$, existe una constante $K < 1$ y un $\varepsilon > 0$ tales que $|g'(x)| < K$, $\forall x \in I_\varepsilon = (r-\varepsilon, r+\varepsilon)$ (por la continuidad de g'). Entonces, $\forall x \in I_\varepsilon$,

$$|g(x) - r| = |g(x) - g(r)| \leq K|x - r| \leq K\varepsilon < \varepsilon$$

o sea, $g(I_\varepsilon) \subset I_\varepsilon$, y podemos aplicar el teorema anterior en I_ε. □

5. Método de la secante

En este método tenemos que x_{n+1} es función de x_n y de x_{n-1}. La idea es la misma que en el método "regula falsi", trazar la secante, pero este método es diferente pues se usan las dos últimas aproximaciones x_{n-1} y x_n en lugar de encerrar la raíz como en "regula falsi". Para empezar hay que dar dos valores x_0 y x_1.

La ecuación de la secante que une los puntos $(x_{n-1}, f(x_{n-1}))$ y $(x_n, f(x_n))$ es

$$y = f(x_n) + (x - x_n)\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

entonces se define x_{n+1} como la intersección de esta recta con el eje x, así, x_{n+1} verifica

$$0 = f(x_n) + (x_{n+1} - x_n)\frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

es decir,

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})}.$$

Observemos que esta fórmula es análoga a la de Newton reemplazando f' por un cociente incremental.

La ventaja es que no hay que calcular la derivada de f (esto es de gran ayuda en un caso en que f' sea difícil de calcular).

La desventaja es, según veremos, que la convergencia de la sucesión es más lenta que la que produce el étodo de Newton.

Observemos que la iteración del método de la secante también puede escribirse como

$$x_{n+1} = \frac{f(x_n)x_{n-1} - f(x_{n-1})x_n}{f(x_n) - f(x_{n-1})}.$$

Analicemos el orden de convergencia de este método, según la definición 4.7. Tenemos

$$f(r) = 0 \quad \text{y} \quad e_n = r - x_n.$$

Luego,

$$\begin{aligned} e_{n+1} = r - x_{n+1} &= r - \frac{f(x_n)x_{n-1} - f(x_{n-1})x_n}{f(x_n) - f(x_{n-1})} \\ &= \frac{e_{n-1}f(x_n) - e_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} \\ &= \frac{e_{n-1}(f(x_n) - f(r)) - e_n(f(r) - f(x_{n-1}))}{f(x_n) - f(x_{n-1})} \\ &= \frac{-e_n e_{n-1}\frac{f(x_n)-f(r)}{x_n - r} + e_n e_{n-1}\frac{f(r)-f(x_{n-1})}{x_{n-1}-r}}{f(x_n) - f(x_{n-1})}. \end{aligned}$$

Es decir,

$$e_{n+1} = e_n e_{n-1} \frac{\frac{f(x_{n-1})-f(r)}{x_{n-1}-r} - \frac{f(r)-f(x_n)}{x_n - r}}{f(x_n) - f(x_{n-1})}. \tag{4.6}$$

Definamos ahora las *diferencias.*

Primera diferencia :

$$f[a, b] = \frac{f(b) - f(a)}{b - a} = f'(\xi).$$

Segunda diferencia :

$$f[a, b, c] = \frac{\frac{f(c)-f(b)}{c-b} - \frac{f(b)-f(a)}{b-a}}{c - a}.$$

Entonces el error del método de la secante verifica,

$$e_{n+1} = -e_n e_{n-1} \frac{f[x_{n-1}, r, x_n]}{f[x_{n-1}, x_n]}.$$

LEMA 4.16.

$$f[a,b,c] = \frac{1}{2}f''(\eta).$$

Demostración.

$$f(x) = f(a) + f[a,b](x-a) + f[a,b,c](x-a)(x-b) + Resto.$$

Despreciamos el resto y nos quedamos con el polinomio de grado 2:

$$f(a) + f[a,b](x-a) + f[a,b,c](x-a)(x-b)$$

Se verá más adelante que este polinomio es el polinomio interpolador de grado dos.

Sea

$$g(x) = f(x) - f(a) + f[a,b](x-a) + f[a,b,c](x-a)(x-b) \tag{4.7}$$

g cumple que $g(a) = 0,\ g(b) = 0$ y $g(c) = 0$.

Entonces g' se anula en por lo menos dos puntos y de ahí que existe η con $g''(\eta) = 0$. Ahora, derivando dos veces la expresión (4.7) y evaluando en η se obtiene

$$0 = g''(\eta) = f''(\eta) - 2f[a,b,c],$$

es decir,

$$f[a,b,c] = \frac{1}{2}f''(\eta).$$

□

Aplicando el lema 4.16 a nuestra expresión de e_{n+1} dada en (4.6) queda

$$e_{n+1} = -\frac{f''(\eta_n)}{f'(\xi_n)}e_n e_{n-1}$$

y de acá se puede deducir la convergencia local.

TEOREMA 4.17. *Si $f'(r) \neq 0$, $|f''| \leq K$ en un entorno de r y x_0, x_1 están suficientemente cerca de r, es decir existe $\varepsilon > 0$ tal que si $x_0, x_1 \in I_\varepsilon = (r-\varepsilon, r+\varepsilon)$, entonces*

$$e_n \to 0.$$

Demostración. Existe $\varepsilon > 0$ tal que $|f'| > \delta$ en I_ε, entonces si $x_0, x_1 \in I_\varepsilon$ tenemos que

$$|e_2| \leq \frac{K}{2\delta}|e_1||e_0| \leq \frac{K}{2\delta}\varepsilon^2$$

y si ahora pedimos (quizás achicando el ε) que

$$\frac{K}{2\delta}\varepsilon = \lambda < 1$$

nos queda que

$$|e_2| \leq \lambda\varepsilon < \varepsilon$$

y entonces $x_2 \in I_\varepsilon$. Ahora bien

$$|e_3| \leq \frac{K}{2\delta}|e_2||e_1| \leq \frac{K}{2\delta}\lambda\varepsilon^2 \leq \lambda^2\varepsilon$$

$$|e_4| \leq \frac{K}{2\delta}|e_3||e_2| \leq \frac{K}{2\delta}\varepsilon\lambda^2\varepsilon\lambda \leq \lambda^3\varepsilon.$$

Y podemos concluir por inducción que

$$|e_n| \leq \lambda^{n-1}\varepsilon \to 0.$$

□

Veamos el orden de convergencia del método de la secante, teníamos

$$|e_{n+1}| = |-\frac{1}{2}\frac{f''(\eta_n)}{f'(\xi_n)}||e_n||e_{n-1}| = c_n|e_n||e_{n-1}|,$$

y además, si llamamos c_∞ al límite de c_n tenemos

$$c_n \to c_\infty = \left|\frac{1}{2}\frac{f''(r)}{f'(r)}\right|.$$

Supongamos que $f''(r) \neq 0$, de esta forma $c_\infty \neq 0$.

Buscamos p tal que

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} = C \neq 0.$$

Tenemos

$$\frac{|e_{n+1}|}{|e_n|^p} = c_n|e_n|^{1-p}|e_{n-1}| = c_n\left(\frac{|e_n|}{|e_{n-1}|^p}\right)^{\alpha}.$$

Si $\alpha = 1-p$ y $\alpha p = -1$, o sea

$$p - p^2 = \alpha p = -1,$$

entonces p es solución de

$$p^2 - p - 1 = 0,$$

y como $p > 0$,

$$p = \frac{1+\sqrt{5}}{2} = 1.618...$$

Con esta elección de p tenemos que

$$y_n = \frac{|e_{n+1}|}{|e_n|^p}$$

cumple la iteración de punto fijo (salvo que c_n es variable pero converge a c_∞),

$$y_{n+1} = c_n y_n^{-\frac{1}{p}}.$$

Entonces, y_n converge al punto fijo de la ecuación $x = c_\infty x^{-\frac{1}{p}}$ (esto es cierto porque $p > 1$ y se ve directamente escribiendo la iteración). El punto fijo es $\overline{x} = c_\infty^{\frac{1}{p}}$. Entonces nos queda que

$$\frac{|e_{n+1}|}{|e_n|^p} \sim \left|\frac{1}{2}\frac{f''(r)}{f'(r)}\right|^{\frac{1}{p}},$$

para n grande.

Ahora veamos una idea intuitiva de la demostración directamente.

Si suponemos $|e_{n+1}| \sim |e_n|^p$ tenemos $|e_{n+1}| \sim |e_n|^p \sim |e_{n-1}|^{p^2}$ y de la relación entre e_{n+1}, e_n y e_{n-1} se tiene $|e_{n-1}|^{p^2} \sim |e_{n-1}|^p|e_{n-1}|$. O sea, $|e_{n-1}|^{p^2-p-1} \sim cte$. Luego $p > 0$ tiene que ser solución de $p^2 - p - 1 = 0$ y entonces $p = \frac{1+\sqrt{5}}{2} = 1.618....$

6. Ejercicios

(1) Usar el método de bisección para hallar una raíz positiva de la ecuación trascendente:

$$2x = \tan(x)$$

¿Cuántos pasos hay que hacer para garantizar que el error sea menor que 10^{-5}?

(2) Hacer un programa en `Matlab` que ejecute los primeros 20 pasos de los métodos de bisección y Regula-Falsi para hallar una raíz de la ecuación $2x^3+x-2=0$ comenzando con el intervalo $[0,1]$.
(3) Hacer un programa en `Matlab` que ejecute los primeros 20 pasos de los métodos de bisección y N-R, para calcular $\sqrt[3]{2}$ comenzando con valores iniciales apropiados.
(4) Demostrar que la ecuación

$$f(x) = e^x + 5\sin x - 2 = 0$$

tiene una única raíz en el intervalo $(0, \frac{3}{2})$. Encontrar las cotas necesarias de $|f'|$ y $|f''|$ para determinar un valor inicial de modo que el método N-R converja a la raíz. Aplicar el método para hallar una aproximación de ésta. ¿Cuál es el orden de convergencia?
(5) Considerar la función $f(x) = \dfrac{x}{1+|x|}$. Determinar para qué valores de x_0 la iteración N-R es convergente, para cuáles es divergente, y cuándo se obtienen ciclos periódicos.
(6) Se quiere resolver la ecuación $f(x)=0$, donde $f(x)=e^x-2$. Calcular los 10 primeros términos de las sucesiones generadas por los métodos N-R y de la secante, comenzando con los valores iniciales $x_1=3$ para el primer método e $y_1=3, y_2=2.3$ para el segundo. Graficar simultáneamente las dos sucesiones obtenidas.
(7) Sea f una función C^1 y sea $(x_n)_{n\in\mathbb{N}}$ la sucesión que se obtiene de aplicar el método N-R a f. Supongamos que x_n converge a r y $f'(r)\neq 0$, mostrar que r es raíz de f.
(8) Sea f una función suave, y a tal que $f(a)=0$, y $f'(a)\neq 0$.
   (a) Suponiendo que en $(a,b]$, $f,\ f',\ f''$ son positivas, probar que la iteración de N-R generada a partir de $x_0\in(a,b)$ converge decrecientemente hacia a.
   (b) Con las mismas hipótesis, si $x_1\in(a,x_0)$, probar que la sucesión generada por el método de la secante a partir de $x_0,\ x_1$ converge decrecientemente hacia a.
(9) Sea $f(x)=x^\alpha$. Se desea utilizar el método N-R para resolver la ecuación $f(x)=0$, comenzando con $x_0>0$. Analizar el comportamiento del método en los casos
   (a) $\alpha\geq 1$ (b) $\alpha=\frac{1}{3}$ (c) $\alpha=\frac{1}{2}$
(10) (a) Sea $f(x)=(x-r_1)(x-r_2)\dots(x-r_d)$ donde $r_1<r_2<\dots<r_d$. Probar que si $x_0>r_d$ la sucesión de N-R converge a r_d.
   (b) Para un polinomio $P\in\mathbb{R}[x]$, $P(x)=a_dx^d+\dots+a_0, a_d\neq 0$, tal que sus d raíces son reales y distintas, se propone el siguiente método que aproxima los valores de todas sus raíces:
      (i) Se comienza con un valor x_0 mayor que $M=\max\{1,\sum_{i=0}^{d-1}\frac{|a_i|}{|a_d|}\}$ (Dato: M es una cota para el módulo de todas las raíces del polinomio).
      (ii) Se genera a partir de x_0 la sucesión de N-R, que, según el ítem anterior, converge a la raíz más grande de P, llamémosla r_d; obteniéndose de este modo un valor aproximado $\tilde{r}_d$.
      (iii) Se divide P por $x-\tilde{r}_d$ y se desprecia el resto, dado que $r_d\sim\tilde{r}_d$. Se redefine ahora P como el resultado de esta división y se comienza nuevamente desde el primer ítem, para hallar las otras raíces.
   Aplicar este método para aproximar todas las raíces del polinomio $P(x)=2x^3-4x+1$.
(11) Recordar que una raíz múltiple de un polinomio f es una raíz simple del polinomio $f/\gcd(f,f')$, donde gcd indica el máximo común divisor. Hacer un programa en `Matlab`

que aplique el método N-R a $f(x)$ y a $f(x)/\gcd(f, f')$ para hallar la raíz múltiple de

$$f(x) = (x-1)(x-2)^2.$$

Demostrar que, a pesar que la función f no está en las hipótesis del método N-R, éste converge (aunque no tan velozmente como cuando la raíz múltiple se halla como solución de $f/\gcd(f, f')$).

(12) Para f una función C^2 que tiene una raíz de orden 2 en x_0:
   (a) Demostrar que el método N-R converge sólo linealmente a x_0.
   (b) ¿Cuál es el orden de convergencia de la siguiente modificación?

$$x_{n+1} = x_n - 2\frac{f(x_n)}{f'(x_n)}$$

(13) Sea $f(x) = 4x^3 - 3x + 1 = 0$. La ecuación $f(x) = 0$ tiene una raíz doble. Aproximarla calculando las 10 primeras iteraciones de los métodos N-R y N-R con la modificación del ejercicio anterior, comenzando con los valores iniciales $x_1 = y_1 = 25$. Graficar simultáneamente las dos sucesiones obtenidas.

(14) Se quiere aplicar el método N-R para dar una tabla de valores de la función $y(x)$ definida implícitamente por la ecuación $G(x, y) = 0$ en un intervalo $[a, b]$.

El método consiste en comenzar la tabla en un par de valores x_0, y_0 que verifican $x_0 = a$ y $G(x_0, y_0) = 0$ y proceder por incrementos en x hasta llegar al valor $x_N = b$.

En cada paso se obtiene el valor de y_{n+1} aplicando el método N-R a la función $G(x_{n+1}, y)$ donde y es la variable y x_{n+1} permanece fijo; con valor inicial el valor de y_n obtenido en el paso anterior. Dado que la función $y(x)$ se supone continua, esta elección del valor inicial se supone apropiada.

   (a) Aplicar el método para la ecuación $G(x, y) = x^2 + y^2 - 1 = 0$, comenzando en $x_0 = 0, y_0 = 1$ para valores de x en $[0, 1]$. Graficar junto con la solución que se obtiene de despejar analíticamente y comparar. Utilizar distintos valores para el incremento y para la cantidad de iteraciones del método N-R en cada paso.
   (b) Aplicar el método para $G(x, y) = 3x^7 + 2y^5 - x^3 + y^3 - 3$. Comenzar la tabla en $x_0 = 0, y_0 = 1$ y proceder por incrementos en x de 0.2 hasta llegar a $x_{50} = 10$.

(15) Dada $F : \mathbb{R}^n \to \mathbb{R}^n$ el método N-R generalizado consiste en realizar la iteración vectorial

$$x^{k+1} = x^k - (DF|_{x^k})^{-1}.F(x^k),$$

donde $(DF|_{x^k})^{-1}$ es la inversa de la matriz diferencial de F evaluada en x^k.

Usar la versión generalizada a varias variables del método N-R para para resolver el sistema de ecuaciones

$$2x - 3y = 0, \quad x^2 - y^2 - 3 = 0$$

comenzando con valores iniciales $(x_0, y_0) = (2, 1)$.

(16) Resolver $\cos(x) = 2x$, $x > 0$ comenzando con $x_0 = 0.5$ y utilizando:
   (a) La iteración de punto fijo $x_{n+1} = \frac{1}{2}\cos(x_n)$
   (b) El método N-R.

Graficar, usando `Matlab`, las sucesiones obtenidas y comparar.

(17) Sea g una función tal que g' es continua en $[s, b]$, donde s es un punto fijo de g. Si además, se verifica que $0 \leq g'(x) \leq K < 1$ para todo $x \in [s, b]$, mostrar que la iteración, comenzando con $x_0 \in [s, b]$, converge decrecientemente a s.

(18) Sea $f : \mathbb{R}_{>0} \to \mathbb{R}$ definida como $f(x) = \dfrac{8x-1}{x} - e^x$.
   (a) Dibujar la gráfica de f y determinar el número de raíces de la ecuación $f(x) = 0$, localizando cada raíz entre dos enteros consecutivos.
   (b) Para cada una de las siguientes funciones:
$$f_1(x) = \frac{1}{8}(1 + xe^x), \quad f_2(x) = \ln\left(\frac{8x-1}{x}\right)$$
   consideramos el siguiente método iterativo: dado $x_0 = 1$ sea
$$x_{n+1} = f_i(x_n),\ n \in \mathbb{N},\ (i = 1, 2).$$
   Estudiar si estas sucesiones convergen hacia alguna de las raíces de $f = 0$.
   (c) Utilizando `Matlab`, estimar las raíces con estos dos métodos.

(19) Sea $f(x) = x^3 - x - 1$. Se consideran las dos siguientes iteraciones de método de punto fijo.
$$g(x) = x^3 - 1, \qquad h(x) = \sqrt[3]{x+1}.$$
   (a) Determinar cuáles de estas funciones son apropiadas para la iteración.
   (b) Para las que sí lo sean:
   - Determinar un intervalo inicial I en el cual el método converja.
   - Dar un valor inicial $x_0 \in I$ y la cantidad de iteraciones necesarias para aproximar la raíz de f con error menor que 10^{-5} comenzando con el x_0 dado.

(20) Dada la función $f(x) = x + 1/x - 2$, $f : \mathbb{R}_{>0} \to \mathbb{R}$, se construye el siguiente algoritmo para aproximar la raíz $r = 1$:
$$x_{n+1} = 2 - 1/x_n$$
   (a) Verificar que si $x_0 > 1$ entonces la sucesión $\{x_n\}$ es monótona decreciente y acotada inferiormente por 1. Concluir que $x_n \to 1$, aunque esta iteración no está en las hipótesis del teorema del punto fijo. ¿Qué hipótesis no se cumple?
   (b) Dar un algoritmo para aproximar la raíz de f que converja cuadráticamente.

(21) Sea f una función C^1 en las condiciones del método N-R. Sea $g(x) = x - \frac{f(x)}{f'(x)}$. Mostrar que el método N-R es un método de punto fijo.

CAPíTULO 5

Interpolación

El objetivo de este capítulo es estudiar cómo puede aproximarse una función por polinomios. Una forma de hacer ésto es construir los polinomios de manera que coincidan con la función dada en algunos puntos predeterminados, lo que recibe el nombre de *interpolación polinomial*. Analizaremos distintos métodos para resolver este problema y estudiaremos el error que se comete al reemplazar una función por un polinomio interpolante.

Hay diversos motivos para estudiar este problema. Por un lado, el polinomio interpolante puede utilizarse para reconstruir una función f a partir de una tabla de valores. Por otra parte, es una herramienta fundamental para integración y diferenciación numérica, como veremos más adelante.

1. Interpolación de Lagrange

En lo que sigue, si $n \in \mathbb{N}_0$, llamaremos $\mathcal{P}_n$ al conjunto de polinomios de grado menor o igual que n, incluyendo el polinomio nulo.

Supongamos que se sabe que la tabla de valores

(x_j):	x_0	x_1	x_2	$\dots$	x_n
(y_j):	y_0	y_1	y_2	$\dots$	y_n

corresponde con datos de una función continua que se desconoce. Queremos poder modelizar dicha función a por medio de un polinomio. Es decir, queremos encontrar un polinomio tal que

$$p(x_j) = y_j, \qquad \forall j = 0, 1, \dots, n. \tag{5.1}$$

Nuestro primer paso será dar un resultado básico que establece que ésto es posible. Mostraremos una forma concreta de hallar un polinomio p que verifique (5.1) y además veremos que si el polinomio es de grado menor o igual que n, éste es único. Vamos a basar nuestra demostración en la *Base de Lagrange* que es una base de polinomios que construimos a continuación.

Base de Lagrange: Para cada punto $x_j, j = 1, \dots, n$, buscamos un polinomio de grado n que se anule en todos los x_i salvo x_j donde queremos que valga 1. Por ejemplo, ℓ_0 será un polinomio en $\mathcal{P}_n$ tal que se anula en $x_1, \dots, x_n$ y $\ell_0(x_0) = 1$.

Como $x_1, \ldots, x_n$ son raíces de ℓ_0, $\ell_0(x) = \alpha \prod_{i=1}^{n}(x - x_i)$; donde α es una constante que se elige de modo que $\ell_0(x_0) = 1$. Imponiendo esta condición obtenemos

$$\ell_0(x) = \frac{\prod_{i=1}^{n}(x - x_i)}{\prod_{i=1}^{n}(x_0 - x_i)}.$$

De manera análoga, para cada $j = 1, \ldots, n,$, el polinomio $\ell_j \in \mathcal{P}_n$ tal que

$$\ell_j(x_i) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

estará dado por

$$\ell_j(x) = \frac{\prod_{i \neq j}(x - x_i)}{\prod_{i \neq j}(x_j - x_i)} \tag{5.2}$$

Los polinomios $\{\ell_0, \ell_1, \ldots, \ell_n\}$ se conocen como la *base de Lagrange*. Vale destacar que estos polinomios sólo dependen de los datos $\{x_0, x_1, \ldots, x_n\}$.

TEOREMA 5.1. *Dados x_0,...,x_n y valores $y_0, \ldots, y_n$ existe un único polinomio $p_n \in \mathcal{P}_n$ tal que*

$$p_n(x_j) = y_j; \quad \forall j = 0, \ldots, n.$$

Demostración. Usando la base de Lagrange definimos

$$p_n(x) = \sum_{j=0}^{n} y_j \ell_j(x). \tag{5.3}$$

obteniendo un polinomio $p_n \in \mathcal{P}_n$ que verifica (5.1). Veamos que es único. Supongamos que hay dos polinomios p_n, $q_n \in \mathcal{P}_n$ que interpolan la tabla de pares (x_i, y_i), esto es

$$(p_n - q_n)(x_j) = 0 \qquad \forall j = 0, \ldots, n.$$

Entonces $p_n - q_n$ es un polinomio de grado menor o igual que n con $n + 1$ raíces distintas; es decir, $p_n - q_n$ es el polinomio nulo. □

OBSERVACIÓN 5.2.

(1) La escritura (5.3) se llama *forma de Lagrange* del polinomio interpolador.
(2) El polinomio p_n puede tener grado estrictamente menor que n. Por ejemplo, si se considera la tabla de 5 valores

(x_j):	-4	-2	0	1	3
(y_j):	9	5	1	-1	-5

El polinomio de grado menor o igual que 4 que interpola la tabla es $p_4(x) = -2x + 1$. Gracias a la unicidad, como se trata de un polinomio de grado 1, es suficiente mostrar que en cada x_j, p_4 toma el valor y_j; esto es inmediato.
(3) Si los datos corresponden con una función f que es un polinomio de grado menor o igual que n, es decir, $f \in \mathcal{P}_n$ y los valores $y_j = f(x_j)$; entonces $f = p_n$ (la interpolación es exacta para polinomios).
(4) El polinomio que interpola en $n + 1$ puntos distintos es único en $\mathcal{P}_n$. Si se permite mayor grado hay infinitos. Por ejemplo, si q es un polinomio cualquiera, el polinomio

$$p(x) = (-2x + 1) + q(x)(x + 4)(x + 2)x(x - 1)(x - 3),$$

también interpola la tabla dada arriba.

Otra forma de demostrar la existencia (y de encontrar el polinomio) es por el método de los coeficientes indeterminados. El polinomio será de la forma

$$p_n(x) = a_0 + a_1 x + \cdots + a_n x^n$$

y se buscan $a_0, \ldots, a_n$ tales que

$$p_n(x_j) = y_j.$$

Al evaluar, queda formado un sistema $(n + 1) \times (n + 1)$

$$\begin{pmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & & & \ddots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{pmatrix} = \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{pmatrix}$$

La matriz de la izquierda se llama matriz de Van der Monde y como sólo depende de los datos $\{x_0, \ldots, x_n\}$ suele notarse por $V(x_0, \ldots, x_n)$.

Para ver que existe una solución $(a_0, \ldots, a_n)$ y que es única hay que ver que la matriz $V(x_0, \ldots, x_n)$ es inversible. Esto equivale a ver que el núcleo es nulo. Ahora, si $(a_0, \ldots, a_n) \in Nu(V(x_0, \ldots, x_n))$ tendríamos

$$a_0 + a_1 x_j + a_2 x_j^2 + \ldots + a_n x_j^n = 0 \qquad \forall j = 0, \ldots, n.$$

Entonces $a_0 = \ldots = a_n = 0$ (pues un polinomio de grado n no nulo no puede tener $n+1$ raíces distintas).

EJEMPLO 5.3. *Se quiere interpolar la función* $f(x) = x^{\frac{2}{3}}$ *en el intervalo* $[-1, 1]$ *por un polinomio.*

Si tenemos en cuenta la paridad de la función, podemos pensar que un polinomio de grado par será una buena elección. La Figura 5.1 muestra el gráfico de f junto con el polinomio interpolante p que se obtiene al considerar 11 puntos equiespaciados. Si consideramos la diferencia máxima entre f y el polinomio p evaluados en una malla suficientemente fina (puntos equiespaciados con distancia $h = 0.01$), el error que se obtiene es grande como puede observarse en el gráfico; el error numérico = 1.4886...

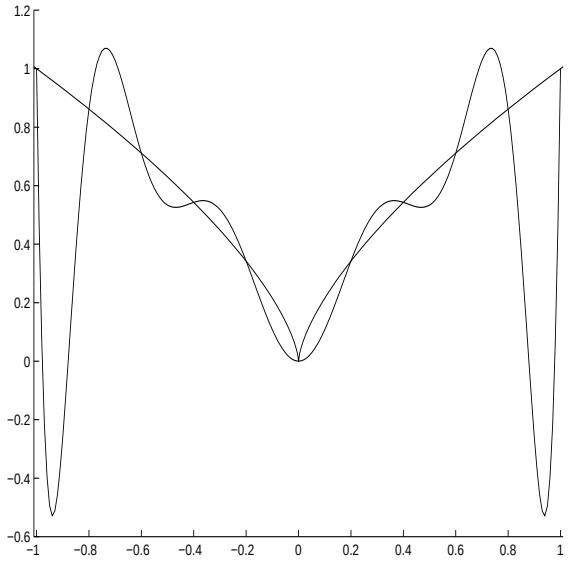

FIGURA 5.1. *Interpolación de* $f(x) = x^{\frac{2}{3}}$ *en* $[-1, 1]$*, (11 puntos equiespaciados)*

2. Error de interpolación

Cuando los datos obtenidos corresponden con datos de una función f definida en $[a, b]$ y x_0, $x_1, \ldots, x_n \in [a, b]$, son $n+1$ puntos distintos; el polinomio interpolador a encontrar será un polinomio $p_n \in \mathcal{P}_n$ que coincida con f en dichos puntos, es decir p_n verifica que

$$p_n(x_j) = f(x_j) \qquad \forall j = 0, \ldots, n.$$

La ventaja de obtener un polinomio que interpole a una función f de la cual sólo se conocen sus valores en los puntos $\{x_0, \ldots, x_n\}$ es que, el polinomio, arroja una fórmula que permite sustituir la función f y hacer evaluaciones en puntos diferentes a los conocidos. Para que este reemplazo tenga alguna validez numérica es importante conocer una estimación del error que

se comete. Para ésto será necesario suponer que la función f verifica algunas condiciones de suavidad. Llamemos a este error:

$$E_n(x) = f(x) - p_n(x), \quad x \in [a, b].$$

Con el siguiente teorema, damos el primer paso para poder estimar el error cometido; es decir, damos una expresión para $E_n(x)$.

Dados los puntos $x_0, \ldots, x_n$, utilizaremos la notación W_{n+1} para designar al polinomio mónico de grado $n+1$ que se anula en esos puntos. Es decir,

$$W_{n+1}(x) = (x - x_0) \cdots (x - x_n)$$

TEOREMA 5.4. *Sean $f \in C^{n+1}[a, b]$ y $p_n \in \mathcal{P}_n$ el polinomio interpolador de f en $x_0, \ldots, x_n$ puntos del intervalo $[a, b]$. Para cada $x \in [a, b]$, existe $\xi \in [a, b]$, $\xi = \xi(x)$, tal que*

$$E_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} W_{n+1}(x).$$

Demostración. Notar que $E_n(x_j) = 0$ y $W_{n+1}(x_j) = 0$ para todo j. Por lo tanto, podemos suponer $x \neq x_j$. Fijado x definimos la siguiente función de t,

$$F(t) = f(t) - p_n(t) - \alpha W_{n+1}(t)$$

donde α se elige de modo que $F(x) = 0$. O sea, $\alpha = \frac{f(x) - p_n(x)}{W_{n+1}(x)}$, que está bien definida pues $W_{n+1}(x) \neq 0$. Observemos que para todo j,

$$F(x_j) = f(x_j) - p_n(x_j) - \alpha W_{n+1}(x_j) = 0.$$

Entonces F se anula en los $n+2$ puntos $x_0, \ldots, x_n, x$. En consecuencia, por el teorema de Rolle, F' tiene al menos $n+1$ ceros, F'' al menos n ceros y así siguiendo se tiene que existe un punto $\xi \in (a, b)$ tal que $F^{(n+1)}(\xi) = 0$. Como

$$F^{(n+1)}(t) = f^{(n+1)}(t) - (n+1)!\alpha$$

Se obtiene,

$$\frac{f^{(n+1)}(\xi)}{(n+1)!} = \frac{f(x) - p_n(x)}{W_{n+1}(x)}$$

lo que concluye la demostración. □

EJEMPLO 5.5. *Se quiere interpolar la función $f(x) = cos(x)^3$ en el intervalo $[-3, 3]$ por un polinomio.*

Si se eligen 10 puntos equiespaciados se obtiene un polinomio como muestra la Figura 5.2. Si consideramos el error numérico, que es el que se obtiene como diferencia máxima entre f y el polinomio evaluados en una malla suficientemente fina (puntos equiespaciados con paso $h = 0.01$) se tiene un error de 0.4303... Tan sólo al considerar 25 puntos equiespaciados (tomados a intervalos de longitud 0.25) se obtiene un error numérico menor que 10^{-6}. En este caso, en una figura como la anterior, los gráficos del polinomio y la función se confunden.

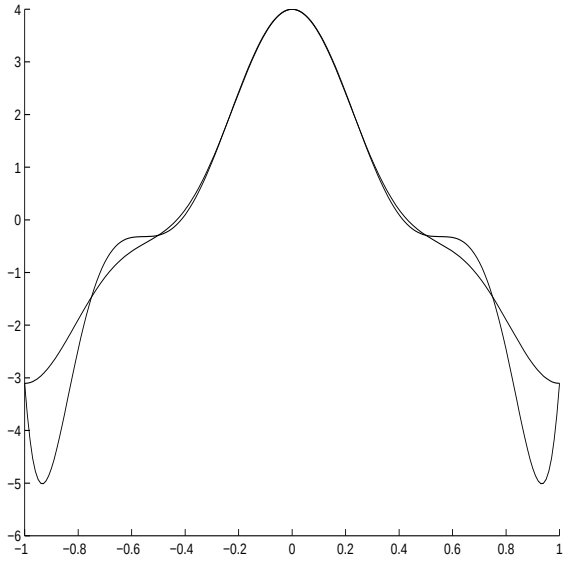

FIGURA 5.2. *Interpolación de* $f(x) = \cos(x)^3$ *en* $[-3, 3]$*, (10 puntos equiespaciados)*

3. Forma de Newton

La forma de Newton es conveniente para calcular el polinomio, en $\mathcal{P}_n$, que interpola a una función f en $x_0, \ldots, x_{n-1}, x_n$ una vez conocido el polinomio interpolador de f en $x_0, \ldots, x_{n-1}$.

La forma de Newton del polinomio interpolador puede verse como una generalización del polinomio de Taylor asociado a una función. En esta construcción aparecen las diferencias divididas que presentamos a continuación.

Primera diferencia dividida

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Segunda diferencia dividida

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}.$$

Así sucesivamente se define la diferencia de orden k asociada a los puntos $x_0, \ldots, x_k$,

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

La construcción de la forma de Newton se basa en la siguiente idea. Una vez obtenido $p_k \in \mathcal{P}_k$ que interpola a f en $x_0, \ldots, x_k$ escribimos $p_{k+1} \in \mathcal{P}_{k+1}$ como

$$p_{k+1}(x) = p_k(x) + a_{k+1}(x - x_0) \cdots (x - x_k).$$

Observemos que como el término agregado no modifica el valor de p_k en $x_0, \ldots, x_k$, p_{k+1} también interpola a f en esos puntos independientemente del valor de a_{k+1}. Por otra parte, podemos elegir

$$a_{k+1} = \frac{f(x_{k+1}) - p_k(x_{k+1})}{(x_{k+1} - x_0) \cdots (x_{k+1} - x_k)}$$

de modo que $p_{k+1}(x_{k+1}) = f(x_{k+1})$.

Iterando este procedimiento desde $k = 1$ hasta $k = n - 1$ se obtiene la forma de Newton

$$p_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \ldots + a_n(x - x_0) \cdots (x - x_{n-1})$$

En lo que sigue veremos que los a_j resultan ser las diferencias divididas y por lo tanto esta expresión es análoga al polinomio de Taylor.

Por ejemplo si $n = 1$,

$$p_1(x) = a_0 + a_1(x - x_0)$$

y como

$$p_1(x_0) = f(x_0) \qquad p_1(x_1) = f(x_1)$$

tenemos

$$a_0 = f(x_0) \qquad a_1 = f[x_0, x_1].$$

Si $n = 2$,

$$p_2(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1).$$

Como en el caso $n = 1$, de las igualdades $p_1(x_0) = f(x_0)$ y $p_1(x_1) = f(x_1)$ queda

$$a_0 = f(x_0) \qquad a_1 = f[x_0, x_1].$$

Veamos ahora que

$$a_2 = f[x_0, x_1, x_2].$$

Sabemos ya que el polinomio $p_1(x)$ que interpola a f en x_0, x_1 se escribe como

$$p_1(x) = f(x_0) + f[x_0, x_1](x - x_0).$$

Análogamente, si $q_1(x) \in \mathcal{P}_1$ interpola a f en x_1, x_2 tenemos,

$$q_1(x) = f(x_1) + f[x_1, x_2](x - x_1).$$

Entonces, el polinomio

$$r(x) = \frac{(x - x_0)q_1(x) - (x - x_2)p_1(x)}{x_2 - x_0}$$

tiene grado menor o igual que 2 y verifica $r(x_j) = f(x_j)$ para $j = 0, 1, 2$. Por lo tanto, coincide con p_2.

En consecuencia, igualando los coeficientes de x^2 de r y p_2 se obtiene

$$a_2 = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = f[x_0, x_1, x_2].$$

El mismo argumento puede aplicarse para demostrar el siguiente teorema.

TEOREMA 5.6. *El polinomio $p_n \in \mathcal{P}_n$, que interpola a f en los puntos $x_0, \ldots, x_n$ está dado por*

$$p_n(x) = f(x_0) + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n](x - x_0) \ldots (x - x_{n-1}) \tag{5.4}$$

No sólo los coeficientes del polinomio interpolador pueden expresarse en términos de las diferencias divididas, sino también el error como lo muestra el siguiente teorema.

TEOREMA 5.7. *Si $p_n \in \mathcal{P}_n$ interpola a f en los puntos $x_0, \ldots, x_n$, se tiene la siguiente expresión del error*

$$E_n(x) = f(x) - p_n(x) = f[x_0, \ldots, x_n, x]W_{n+1}(x).$$

Demostración. Agregamos x_{n+1} a la sucesión $\{x_0, \ldots, x_n\}$ y consideramos p_n y p_{n+1} como en (5.4), entonces se tiene

$$\begin{aligned} p_{n+1}(x) &= f(x_0) + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_{n+1}](x - x_0) \cdots (x - x_n) \\ &= p_n(x) + f[x_0, \ldots, x_{n+1}]W_{n+1}(x). \end{aligned}$$

Por lo tanto

$$f(x_{n+1}) = p_{n+1}(x_{n+1}) = p_n(x_{n+1}) + f[x_0, \ldots, x_{n+1}]W_{n+1}(x_{n+1}).$$

De aquí se deduce que el error satisface

$$E_n(x_{n+1}) = f(x_{n+1}) - p_n(x_{n+1}) = f[x_0, \ldots, x_{n+1}]W_{n+1}(x_{n+1}).$$

Como tomamos x_{n+1} cualquier punto distinto de $x_0, \ldots, x_n$ se tiene para todo x,

$$E_n(x) = f(x) - p_n(x) = f[x_0, \ldots, x_n, x]W_{n+1}(x).$$

□

COROLARIO 5.8. *Dados $x_0, \ldots, x_n$ puntos distintos, existe ξ intermedio, es decir ξ entre $x_0, \ldots, x_n$ tal que*

$$f[x_0, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$

Demostración. Evaluando en $x = x_n$ la expresión del error $E_{n-1} = f - p_{n-1}$, dada por el teorema anterior tenemos,

$$E_{n-1}(x_n) = f[x_0, \ldots, x_n](x_n - x_0) \cdots (x_n - x_{n-1})$$

lo que junto con la fórmula del error dada en el Teorema 5.4 concluye la demostración. □

4. Polinomios de Tchebychev - Minimización del Error

Una pregunta natural es cómo elegir los puntos de interpolación para optimizar la aproximación. El Teorema 5.4 nos dice que el error depende de f^{n+1} en algún punto del intervalo y de los puntos x_j a través del polinomio $W_{n+1}(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$. Como se pretende obtener una buena aproximación sin tener información sobre la función f, la idea es elegir los puntos de manera tal que $\|W_{n+1}(\cdot)\|_\infty$ sea mínima. Este problema, que en principio parece complicado, fue resuelto por Tchebychev en el siglo XIX introduciendo una sucesión de polinomios, que hoy llevan su nombre.

Para simplificar la presentación resolveremos el problema para funciones definidas en el intervalo $[-1, 1]$. Más adelante veremos que se puede trasladar la construcción a cualquier intervalo $[a, b]$ mediante un cambio de variables.

Los polinomios de Tchebychev se definen para $k = 0, 1, 2, \ldots$ por

$$T_k(x) = \cos(k \cos^{-1} x)$$

donde $\cos^{-1}$ es la inversa de $\cos : [0, \pi] \to [-1, 1]$.

En principio no es evidente que T_k sea un polinomio. Pero esto puede verse utilizando identidades trigonométricas. En efecto,

$$T_0(x) = 1, \qquad T_1(x) = x$$

y como $\cos(\alpha + \beta) + \cos(\alpha - \beta) = 2 \cos \alpha \cos \beta$, si ponemos $x = \cos \theta$ resulta

$$T_{k+1}(x) = \cos((k+1)\theta) = 2 \cos \theta \cos(k\theta) - \cos((k-1)\theta),$$

es decir,

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x). \tag{5.5}$$

Algunos ejemplos que siguen a T_0 y T_1 cuyos gráficos se muestran en la figura 5.1 son

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

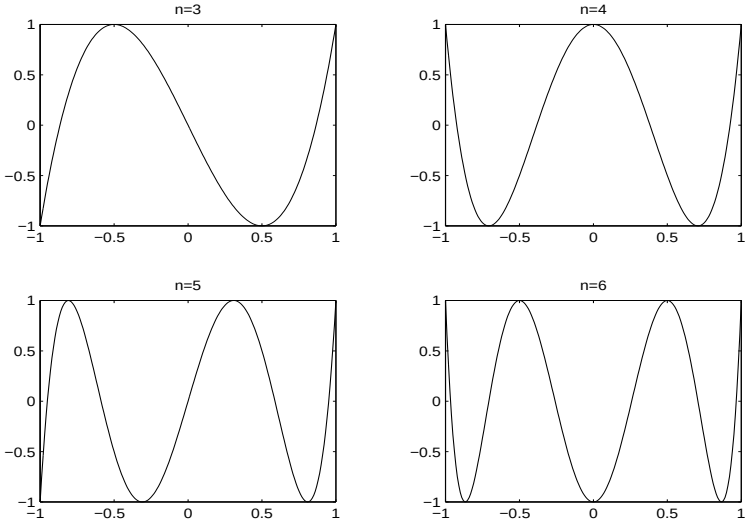

FIGURA 5.3. *Polinomios de Tchebychev*

Los polinomios de Tchebychev tienen las siguientes propiedades.

PROPOSICIÓN 5.9. *Sea T_k el polinomio de Tchebychev de grado k.*

(1) *El coeficiente principal de T_k es 2^{k-1}, para todo $k \in \mathbb{N}$.*
(2) *Las raíces del polinomio T_k se encuentran en el intervalo $[-1,1]$ y son de la forma*
$$x_i = \cos\left(\frac{(2i+1)\pi}{2k}\right)$$
para $i = 0, 1, \ldots, k-1$. En particular, son todas distintas.
(3) *$\|T_k\|_\infty = 1$. Además, T_k alcanza los valores 1 y -1 en $k+1$ puntos, es decir,*
$$\|T_k\|_\infty = |T_k(y_i)| = 1 \qquad \textit{para} \qquad y_i = \cos(\frac{i\pi}{k})$$
con $i = 0, \ldots, k$.

Demostración. La primer afirmación puede verse de la relación de recurrencia (5.5).

Como $T_k(x) = \cos(k \cos^{-1} x)$, $T_k(x) = 0$ si y sólo si el argumento es múltiplo impar de $\frac{\pi}{2}$. Es decir, para $i \in \mathbb{Z}$,
$$\begin{aligned} k\cos^{-1}(x) &= (2i+1)\tfrac{\pi}{2} \\ x &= \cos(\tfrac{(2i+1)}{k}\tfrac{\pi}{2}) \end{aligned}$$
ambas afirmaciones de *2* quedan probadas. Es decir, las raíces pertenecen al intervalo $[-1,1]$ y variando los valores de $i = 0, 1, \ldots, k-1$ se obtienen todas.

Para probar *3*, basta notar que $|T_k(x)| \leq 1$ por ser imagen de la función coseno. Además, sobre los puntos $y_i = \cos(\frac{i\pi}{k})$, T_k toma alternativamente los valores $1, -1$ y por lo tanto la norma es exactamente 1. □

Ahora sí, estamos en condiciones de enunciar y probar el resultado que anticipamos. Es decir, entre todas las posibles elecciones de $n+1$ puntos en $[-1,1]$, los ceros de T_{n+1} son los puntos de interpolación que hay que elegir para minimizar la expresión $\|(x-x_0)\ldots(x-x_n)\|_\infty$ que aparece en la fórmula del error.

TEOREMA 5.10. *Entre todos los polinomios mónicos de grado $n+1$,*
$$W_{n+1}(x) = \frac{1}{2^n} T_{n+1}(x)$$
minimiza la norma $\|\ \|_\infty$ en $[-1,1]$. O sea, si $P \in \mathcal{P}_{n+1}$ y es mónico entonces,
$$\|W_{n+1}\|_\infty \leq \|P\|_\infty.$$

Demostración. Como el coeficiente principal de T_{n+1} es 2^n se tiene que W_{n+1} es mónico.

Supongamos que existe un polinomio $P \in \mathcal{P}_{n+1}$, mónico tal que
$$\|P\|_\infty < \|W_{n+1}\|_\infty.$$

Por la proposición anterior, $|W_{n+1}(x)|$ alcanza su máximo (que es $\frac{1}{2^n}$) en los $n+2$ puntos $y_i = \cos(\frac{i\pi}{n+1})$, $i = 0, \ldots, n+1$. Esto es, si restringimos W_{n+1} a $[y_i, y_{i+1}]$, W_{n+1} alcanza la norma infinito en cada uno de estos subintervalos. Entonces, en cada subintervalo se mantiene la relación

$$\|P\|_{L^\infty[y_i,y_{i+1}]} < \frac{1}{2^n} = \|W_{n+1}\|_{L^\infty[y_i,y_{i+1}]}. \tag{5.6}$$

Por otra parte, $W_{n+1}(y_i) = -W_{n+1}(y_{i+1})$. Supongamos, por ejemplo, que $W_{n+1}(y_i) > 0$ (en el caso contrario se procede de manera análoga). Entonces, de la desigualdad (5.6) se sigue que $P(y_i) < W_{n+1}(y_i)$ y que $P(y_{i+1}) > W_{n+1}(y_{i+1})$.

Luego, el polinomio $Q(x) = P(x) - W_{n+1}(x)$ tiene al menos un cero en el intervalo (y_i, y_{i+1}) y como hay $n+2$ valores de y_i, resulta que Q tiene al menos $n+1$ ceros. Pero tanto P como W_{n+1} son polinomios de grado $n+1$ y ambos son mónicos de donde se deduce que Q tiene grado a lo sumo n. Esto es una contradicción pues acabamos de ver que Q tiene $n+1$ raíces distintas. Luego, un tal P no puede existir. □

Observación 5.11. Puede demostrarse, aunque no lo haremos aquí, que la desigualdad del teorema es estricta, o sea, $\|W_{n+1}\|_\infty < \|P\|_\infty$ si $P \neq W_{n+1}$ es un polinomio mónico $P \in \mathcal{P}_{n+1}$. Es decir el minimizante es único.

Ejemplo 5.12. *Se quiere aproximar la función $f(x) = x^{\frac{2}{3}}$ en el intervalo $[-1, 1]$ por un polinomio que la interpole en 11 puntos.*

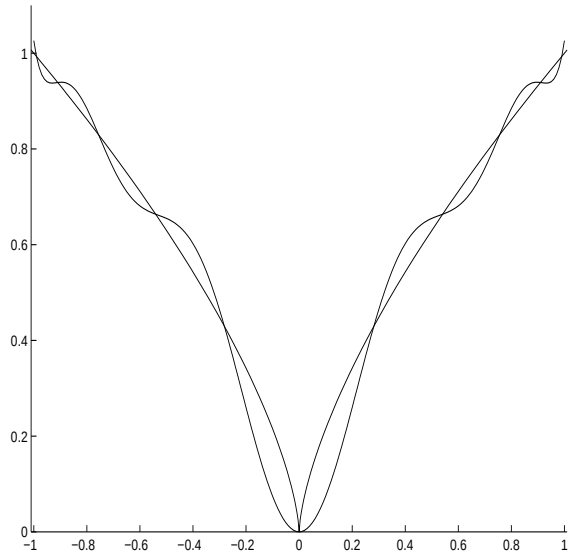

Figura 5.4. *Interpolación de $f(x) = x^{\frac{2}{3}}$ en los ceros de T_{11}*

Si se eligen los nodos como los ceros de T_{11} se obtiene un polinomio como muestra la Figura 5.4 (comparar con Figura 5.1). En este caso el error numérico cometido es menor que 0.1408, (comparar con Ejemplo 5.3).

Veamos ahora como se aplica el Teorema 5.7 para acotar el error cuando se usan las raíces de T_{n+1} como puntos de interpolación.

TEOREMA 5.13. *Sea $f \in C^{n+1}[-1,1]$. Si $p_n \in \mathcal{P}_n$ es el polinomio que interpola a f en las raíces de T_{n+1} entonces,*

$$\|f - p_n\|_\infty \leq \frac{\|f^{(n+1)}\|_\infty}{2^n(n+1)!}.$$

Demostración. Basta observar que $W_{n+1} = \frac{1}{2^n} T_{n+1}$ para obtener

$$f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} W_{n+1}(x) = \frac{f^{(n+1)}(\xi)}{2^n(n+1)!} T_{n+1}(x),$$

donde $\xi \in [-1,1]$. Entonces, el resultado se sigue del hecho de que $|T_{n+1}(x)| \leq 1$. □

EJEMPLO 5.14. *Sea $f : [-1,1] \to \mathbb{R}$ dada por $f(x) = e^{3x}$. Queremos comparar las cotas del error que se produce al estimar el valor de $f(0.8)$ al usar el polinomio interpolador de grado 4 construído con puntos equiespaciados y con los ceros del polinomio T_5.*

Comencemos observando que $f^{(5)}(x) = 243e^{3x}$ y por lo tanto

$$\frac{\|f^{(5)}\|_\infty}{5!} \leq \frac{243e^3}{5!} \leq \frac{4880.79}{5!}.$$

Si interpolamos f en cinco puntos equiespaciados tenemos que

$$W_5(x) = (x+1)(x+0.5)x(x-0.5)(x-1),$$

entonces $|W_5(0.8)| = 0.11232$ y usando la fórmula del error obtenemos

$$|(f - p_4)(0.8)| \leq \frac{4880.79}{5!} 0.11232 \sim 4.57.$$

Cuando en realidad

$$|E_4(0.8)| = 0.4591\ldots$$

Notar que en este caso, se sobre estima el error en un factor de 10.

Ahora, interpolamos usando los ceros de T_4. La cota que se obtiene de la fórmula de error es

$$|E_4(0.8)| \leq \frac{4880.79}{5!2^4} = 2.54,$$

mientras que $E_4(0.8) = f(0.8) - p_4(0.8) = 0.2544$.

Observemos que tanto el error como su estimación se reducen aproximadamente la mitad que en el caso de puntos equiespaciados.

OBSERVACIÓN 5.15. Una traslación lineal del intervalo $[a, b]$ al intervalo $[-1, 1]$ nos permite dar los polinomios de Tchebychev correspondientes al intervalo $[a, b]$.

En efecto, es fácil ver que el cambio de variables $t = \dfrac{2(x-a)}{b-a} - 1$ es la transformación mencionada. Por lo tanto

$$\widetilde{T}_k(x) = T_k(t) = T_k\left(\frac{2(x-a)}{b-a} - 1\right) = \cos\left(k\cos^{-1}\left(\frac{2(x-a)}{b-a} - 1\right)\right)$$

es un polinomio de grado k que tiene propiedades análogas a T_k pero ahora en el intervalo $[a, b]$. En particular se tiene:

(1) La relación de recurrencia:

$$\widetilde{T}_{k+1}(x) = 2\left(\frac{2(x-a)}{b-a} - 1\right)\widetilde{T}_k(x) - \widetilde{T}_{k-1}(x)$$

(2) El coeficiente principal de $\widetilde{T}_k(x)$ es $2^{k-1}(\frac{2}{b-a})^k$.

(3) Los ceros de $\widetilde{T}_k(x)$ son de la forma

$$x_j = \frac{b-a}{2}\cos\left(\frac{(2j+1)\pi}{2k}\right) + \frac{b+a}{2} \quad \forall j = 0, \ldots, k-1.$$

(4) Interpolando en los ceros de $\widetilde{T}_{n+1}$

$$W_{n+1}(x) = \frac{1}{2^n}\left(\frac{b-a}{2}\right)^{n+1}\widetilde{T}_{n+1}(x) \qquad \text{y} \qquad \|W_{n+1}\|_\infty = \frac{1}{2^n}\left(\frac{b-a}{2}\right)^{n+1}$$

obteniéndose, para $x \in [a, b]$, la cota del error

$$|f(x) - p_n(x)| \leq \frac{\|f^{(n+1)}\|_\infty}{(n+1)!2^n}\left(\frac{b-a}{2}\right)^{n+1}.$$

Antes de proceder con algunos comentarios finales estudiemos el análogo al Ejemplo 5.5 considerando como nodos los ceros del correspondiente polinomio de Tchebychev.

EJEMPLO 5.16. *Se quiere aproximar la función $f(x) = cos(x)^3$ en el intervalo $[-3, 3]$ por un polinomio que la interpole en los ceros de T_{10}.*

Al elegirse como nodos los ceros de T_{10} se obtiene un polinomio como muestra la Figura 5.5 (comparar con Figura 5.2). En este caso el error numérico cometido es menor que 4×10^{-3}. Comparar con Ejemplo 5.5 en el que se interpola la misma función en 10 puntos equiespaciados.

Comentarios:

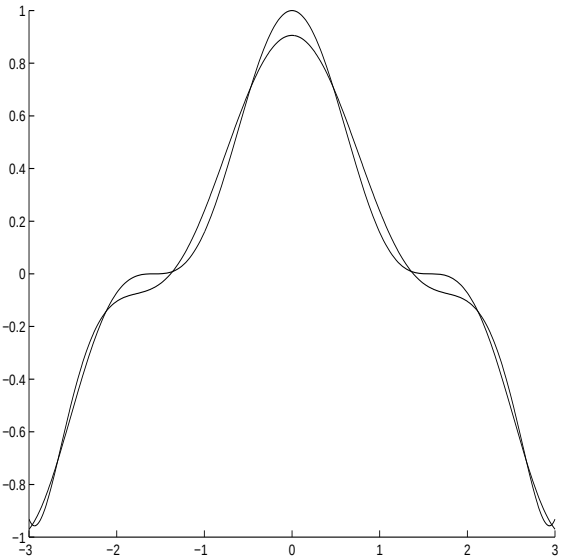

FIGURA 5.5. Interpolación de $f(x) = \cos(t)^3$ en $[-3, 3]$, (en los ceros de T_{10})

(1) A partir de la fórmula del error dada en el Teorema 5.4 puede demostrarse que si f es una función entera, es decir, admite desarrollo de Taylor convergente en todo $\mathbb{R}$, entonces

$$\|f - p_n\|_{L^\infty[a,b]} \to 0 \qquad (n \to \infty).$$

cualesquiera sean los puntos de interpolación.

(2) No podemos asegurar convergencia uniforme, es decir, en norma infinito, si se cambia la hipótesis f *entera* por $f \in C^\infty(\mathbb{R})$. Por ejemplo, si se eligen puntos equidistribuidos en el intervalo $[-1, 1]$ se sabe que el error no tiende a cero para la función de Runge

$$f(x) = \frac{1}{1 + 25x^2}$$

(3) El comportamiento de la interpolación en los puntos de Tchebychev es mucho mejor. Por ejemplo, puede demostrarse que si la función f es derivable

$$\|f - p_n\|_\infty \to 0 \qquad (n \to \infty).$$

(4) La interpolación en los puntos de Tchebychev no converge para cualquier función continua. O sea, puede verse que existe f continua tal que $\|f - p_n\|_\infty \not\to 0$. Más aún puede demostrarse el siguiente

Teorema. (Faber) Dados puntos

$$\begin{array}{cccc} x_0^0 & & & \\ x_0^1 & x_1^1 & & \\ x_0^2 & x_1^2 & x_2^2 & \\ x_0^3 & x_1^3 & x_2^3 & x_3^3 \\ \vdots & & & \end{array}$$

arbitrarios en $[a,b]$, existe f continua tal que $\|f-p_n\|_\infty \not\to 0$, donde p_n es el polinomio interpolador en $x_0^n,\dots,x_n^n$.

5. Interpolación de Hermite

En algunos casos interesa también considerar junto con los valores de una función f datos relacionados con sus derivadas. Por ejemplo, puede buscarse un polinomio p que interpole a f en determinados puntos y que además p' coincida con f' en algunos de esos puntos. Más en general, se tiene el siguiente teorema que fue probado por *Hermite*.

TEOREMA 5.17. *Dada una función f, puntos $x_0,\dots,x_k$ y $m_0,\dots,m_k \in \mathbb{N}_0$ tales que $m_0+\ldots+m_k = n+1$, existe un único polinomio $p\in\mathcal{P}_n$ que satisface*

$$\begin{cases} p(x_0)=f(x_0), & p'(x_0)=f'(x_0), & \dots & p^{(m_0-1)}(x_0)=f^{(m_0-1)}(x_0),\\ p(x_1)=f(x_1), & p'(x_1)=f'(x_1), & \dots & p^{(m_1-1)}(x_1)=f^{(m_1-1)}(x_1),\\ \quad\vdots & \quad\vdots & & \quad\vdots\\ p(x_k)=f(x_k), & p'(x_k)=f'(x_k), & \dots & p^{(m_k-1)}(x_k)=f^{(m_k-1)}(x_k). \end{cases}$$

No haremos una demostración de este teorema pero para dar una idea mostramos la construcción del polinomio interpolador en un caso particular; donde además puede verse cómo se generaliza la definición de diferencias divididas para valores de x_i no todos distintos.

Se busca un polinomio $p\in\mathcal{P}_3$ que cumpla

$$\begin{cases} (i)\ p(x_0)=f(x_0), & (iii)\ p(x_1)=f(x_1),\\ (ii)\ p'(x_0)=f'(x_0), & (iv)\ p'(x_1)=f'(x_1). \end{cases}$$

Como $\{1, x-x_0, (x-x_0)^2, (x-x_0)^2(x-x_1)\}$ forman una base de $\mathcal{P}_3$ por ser todos de distinto grado, cualquier polinomio en $\mathcal{P}_3$ se puede escribir de la forma

$$p(x) = a_0 + a_1(x-x_0) + a_2(x-x_0)^2 + a_3(x-x_0)^2(x-x_1).$$

Las condiciones $(i),(ii)$ se satisfacen si y solo si $a_0 = f(x_0)$ y $a_1 = f'(x_0)$. Ahora hay que determinar a_2 y a_3 para que se cumplan las dos condiciones restantes. Para simplificar la notación ponemos $h=(x_1-x_0)$, entonces se tiene

$$p(x_1) = a_0 + ha_1 + h^2a_2 = f(x_0) + f'(x_0)h + a_2h^2$$

Para que se satisfaga la condición (iii) debe ser

$$a_2 = \frac{f(x_1)-f(x_0)-f'(x_0)h}{h^2} = \left(\frac{f(x_1)-f(x_0)}{h} - f'(x_0)\right)\frac{1}{h}.$$

Observemos que $\lim_{x_1 \to x_0} f[x_0, x_1] = f'(x_0)$ por lo que resulta natural generalizar la primer diferencia dividida poniendo

$$f[x_0, x_0] = f'(x_0).$$

De esta manera se obtiene, de la definición de segunda diferencia dividida,

$$f[x_0, x_0, x_1] = \frac{f[x_0, x_1] - f[x_0, x_0]}{x_1 - x_0} = \left(\frac{f(x_1) - f(x_0)}{h} - f'(x_0) \right) \frac{1}{h}$$

y por lo tanto, $a_2 = f[x_0, x_0, x_1]$.

Por último, queremos que $p'(x_1) = f'(x_1)$. Entonces, debemos elegir a_3 para que se cumpla

$$f'(x_1) = a_1 + 2a_2 h + a_3 h^2 = f'(x_0) + 2f[x_0, x_0, x_1]h + a_3 h^2$$

de dónde,

$$\begin{aligned} a_3 &= \frac{1}{h^2} \left(f'(x_1) - f'(x_0) - 2f[x_0, x_0, x_1]h \right) \\ &= \frac{1}{h^2} \left(f[x_1, x_1] - f[x_0, x_0] - 2f[x_0, x_0, x_1]h \right) \\ &= \frac{1}{h^2} \left(f[x_1, x_1] - f[x_0, x_1] + f[x_0, x_1] - f[x_0, x_0] - 2f[x_0, x_0, x_1]h \right) \\ &= \frac{1}{h} \left(f[x_0, x_1, x_1] + f[x_0, x_0, x_1] - 2f[x_0, x_0, x_1] \right) \\ &= \frac{f[x_0, x_1, x_1] - f[x_0, x_0, x_1]}{x_1 - x_0}. \end{aligned}$$

O sea

$$a_3 = f[x_0, x_0, x_1, x_1].$$

En consecuencia, hemos demostrado que el único polinomio en $\mathcal{P}_3$ que satisface las condiciones pedidas es

$$p_3(x) = f[x_0] + f[x_0, x_0](x - x_0) + f[x_0, x_0, x_1](x - x_0)^2 + f[x_0, x_0, x_1, x_1](x - x_0)^2(x - x_1).$$

Esto generaliza la forma de Newton para x_i no todos distintos.

6. Interpolación por polinomios a trozos

En muchos casos para lograr una mejor aproximación es conveniente utilizar funciones polinomiales a trozos como interpolantes. De esta manera se parte el intervalo de manera tal que en cada subintervalo se elige un polinomio distinto que interpole los datos. Por ejemplo, al interpolar con polinomios de grado uno a trozos, quedan poligonales.

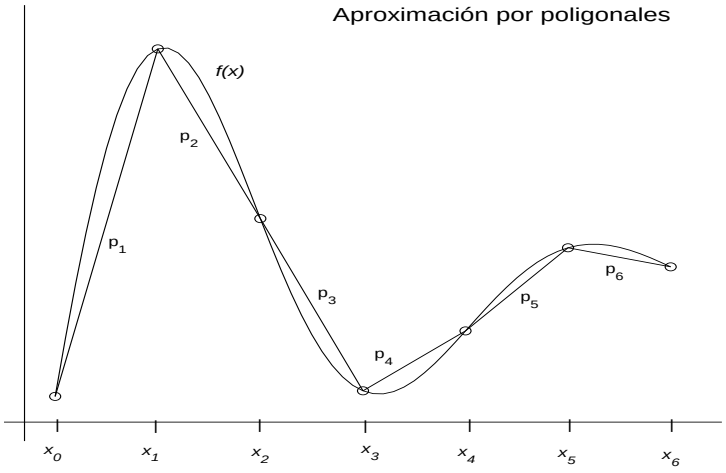

Figura 5.2

Partimos el intervalo $[a, b]$ en subintervalos $[x_j, x_{j+1}]$, $a = x_0 < x_1 < x_2 \ldots < x_n = b$. Dada $f : [a, b] \to \mathbb{R}$ definimos la función interpolante $q_n(x)$ tal que

$$q_n \mid_{[x_j, x_{j+1}]} \in \mathcal{P}_1.$$

Consideremos el caso de puntos equiespaciados o sea, $x_j = a + jh$ con $h = \frac{b-a}{n}$. Para cualquier $f \in C^2[a, b]$ y cualquier $x \in [x_j, x_{j+1}]$, usando el Teorema tenemos 5.4

$$f(x) - q_n(x) = \frac{f''(\xi_j)}{2}(x - x_j)(x - x_{j+1}).$$

Entonces

$$\|f - q_n\|_\infty \leq \frac{\|f''\|_\infty}{2}\frac{h^2}{4} = \frac{\|f''\|_\infty}{8}\frac{(b-a)^2}{n^2} \to 0$$

cuando $n \to \infty$.

Ejemplo 5.18. Sea $f : [-1, 1] \to \mathbb{R}$, la función de Runge $f(x) = \dfrac{1}{1 + 25x^2}$. Puede verse que $\|f''\|_\infty = 50$. En consecuencia, aproximando por poligonales obtenemos

$$\|f - q_n\|_\infty \leq \frac{\|f''\|_\infty}{8}\left(\frac{2}{n}\right)^2 = \frac{25}{n^2}.$$

Luego, en este caso, interpolando por poligonales obtenemos una aproximación mucho mejor que la que se obtiene interpolando con un polinoimio, si se utilizan los mismos puntos.

Splines cúbicos.

En muchos problemas interesa aproximar por funciones derivables. Esto no puede lograrse aproximando por poligonales y por lo tanto es necesario aumentar el grado de los aproximantes. Un método clásico es el corrospondiente a grado tres, *splines cúbicos*. Vamos a ver que, de esta manera, puede obtenerse un aproximante C^2.

Dada $f \in C[a,b]$ y $a = x_0 < x_1 < x_2 \ldots < x_n = b$ buscamos S tal que S, S' y S'' sean continuas en $[a,b]$ y además se verifique

$$S(x_j) = f(x_j) \quad \text{para} \quad 0 \leq j \leq n \quad \text{y} \quad S\,|_{[x_j,x_{j+1}]} \in \mathcal{P}_3.$$

Por ejemplo si $n = 3$, como $S_j \in \mathcal{P}_3$ tenemos $4 \times 3 = 12$ coeficientes a determinar. Veamos cuántas condiciones se tienen que satisfacer. Tenemos que verificar,

$$\begin{array}{lll} S_0(x_0) = f(x_0), & S_1(x_1) = f(x_1), & S_2(x_2) = f(x_2), \\ S_0(x_1) = S_1(x_1), & S_1(x_2) = S_1(x_2), & S_2(x_3) = f(x_3). \end{array}$$

es decir, seis condiciones. Si además queremos S' y S'' continuas en x_1, x_2 tenemos cuatro condiciones mas. O sea, en total diez condiciones para doce incógnitas. Una cuenta análoga puede hacerse en el caso general para ver que la cantidad de coeficientes a determinar supera en dos al número de condiciones.

Luego, si hay solución habrá infinitas pues tenemos dos coeficientes para fijar arbitrariamente. Lo natural entonces es fijar $S'(x_0)$ y $S'(x_n)$ o bien $S''(x_0)$ y $S''(x_n)$. Elegiremos esta última opción por ser más simple.

TEOREMA 5.19. *Dada $f \in C[a,b]$ y $a = x_0 < x_1 < x_2 \ldots < x_n = b$, existe un única $S \in C^2[a,b]$ tal que*

$$\begin{cases} S(x_j) = f(x_j) & 0 \leq j \leq n \\ S\,|_{[x_j,x_{j+1}]} \in \mathcal{P}_3 \end{cases}$$

con $S''(a) = S''(b) = 0$.

Demostración. Para $j = 0, \ldots, n-1$ usaremos la notación

$$S_j = S\,|_{[x_j, x_{j+1}]} \qquad \text{y} \qquad h_j = x_{j+1} - x_j.$$

La función S buscada debe cumplir que S'' es una poligonal. Por lo tanto, si $S''(x_j) = y_j$, S_j'' se escribe como

$$S_j''(x) = y_j \frac{x_{j+1} - x}{h_j} + y_{j+1} \frac{x - x_j}{h_j} \qquad 0 \le j \le n-1.$$

Veremos que es posible encontrar valores y_j de tal forma que se cumplan las condiciones requeridas para S. Integrando dos veces obtenemos, para $x \in [x_j, x_{j+1}]$, con $0 \le j \le n-1$

$$S_j(x) = \frac{y_j}{6h_j}(x_{j+1} - x)^3 + \frac{y_{j+1}}{6h_j}(x - x_j)^3 + c_j(x - x_j) + d_j(x_{j+1} - x) \tag{5.7}$$

donde c_j, d_j son constantes a determinar que provienen de la integración.

Observemos que para cualquier elección de y_j, c_j y d_j, S'' resulta continua por ser poligonal. Por lo tanto resta ver que esas constantes pueden elegirse de manera que se verifiquen las otras condiciones requeridas sobre S.

Para que S sea continua e interpole a f tenemos que elegir c_j, d_j tal que

$$S_j(x_j) = f(x_j) \quad \text{y} \quad S_j(x_{j+1}) = f(x_{j+1}), \qquad 0 \le j \le n-1$$

de lo que, reemplazando en (5.7), obtenemos

$$c_j = \frac{f(x_{j+1})}{h_j} - \frac{y_{j+1}h_j}{6} \qquad \text{y} \qquad d_j = \frac{f(x_j)}{h_j} - \frac{y_j h_j}{6}$$

y por lo tanto, para cada $0 \le j \le n-1$

$$\begin{aligned} S_j(x) = \; & \frac{y_j}{6h_j}(x_{j+1} - x)^3 + \frac{y_{j+1}}{6h_j}(x - x_j)^3 + \\ & + \left(\frac{f(x_{j+1})}{h_j} - \frac{y_{j+1}h_j}{6}\right)(x - x_j) + \left(\frac{f(x_j)}{h_j} - \frac{y_j h_j}{6}\right)(x_{j+1} - x). \end{aligned}$$

Derivando, y utilizando la notación $\Delta f_j = f(x_{j+1}) - f(x_j)$, obtenemos

$$S_j'(x) = -\frac{y_j}{2h_j}(x_{j+1} - x)^2 + \frac{y_{j+1}}{2h_j}(x - x_j)^2 + \frac{\Delta f_j}{h_j} - \frac{h_j}{6}(y_{j+1} - y_j)$$

y tenemos que elegir y_j para que se cumpla la condición que falta, es decir, que S' sea continua, o sea

$$S_j'(x_j) = S_{j-1}'(x_j) \qquad 1 \le j \le n-1$$

de lo que resulta que las $n+1$ incógnitas y_j deben ser solución del siguiente sistema de $n-1$ ecuaciones,

$$h_{j-1}y_{j-1} + 2(h_j + h_{j-1})y_j + h_j y_{j+1} = b_j$$

con

$$b_j = 6\left(\frac{\Delta f_j}{h_j} - \frac{\Delta f_{j-1}}{h_{j-1}}\right)$$

Como tenemos dos incógnitas más que ecuaciones, podemos dar valores arbitrarios a y_0, y_n y, pasando los términos correspondientes al lado derecho, obtenemos el sistema tridiagonal,

$$\begin{pmatrix} \gamma_1 & h_1 & 0 & \cdots & 0 \\ h_1 & \gamma_2 & h_2 & \cdots & 0 \\ \vdots & \vdots & & \ddots & \vdots \\ 0 & \cdots & & \cdots & \gamma_{n-1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \end{pmatrix} = \begin{pmatrix} b_1 - h_0 y_0 \\ b_2 \\ \vdots \\ b_{n-1} - h_{n-1} y_n \end{pmatrix}$$

donde $\gamma_i = 2(h_i + h_{i-1})$.

Ahora, como A es diagonal estrictamente dominante, entonces es inversible. Por lo tanto existe solución única una vez elegidos y_0, y_n.

Por ejemplo podemos elegir $y_0 = y_n = 0$ para que se satisfagan las condiciones $S''(x_0) = 0$ y $S''(x_n) = 0$, lo que concluye la demostración. □

Observemos que en general $S'(x_j) \neq f'(x_j)$ y $S''(x_j) \neq f''(x_j)$.

7. Ejercicios

(1) Para cada uno de los conjuntos de datos dados, calcular el polinomio $p(x)$ interpolador de grado menor o igual que 3, en la forma de Lagrange. Verificar utilizando el comando **polyfit** de `Matlab`. Graficar el polinomio interpolador, usando el comando **polyval**.

x	-1	0	2	3
y	-1	3	11	27

x	-1	0	1	2
y	-3	1	1	3

(2) Repetir el problema anterior, usando el método de coeficientes indeterminados.

(3) (a) Construir las tablas de diferencias divididas para los datos del Ejercicio 1, y emplearlas para construir los polinomios interpoladores.

(b) Agregar a las tablas de datos del Ejercicio 1 el punto $x = 4,\ y = 1$. Aumentar las tablas de diferencias divididas y calcular los polinomios interpoladores.

(4) Considerar la función $f(x) = \dfrac{1}{1+25x^2}$ en el intervalo [-1,1]. Graficar f junto con los polinomios que resultan de interpolar a f en los $n+1$ puntos equiespaciados $x_0 = -1, \ldots, x_i = x_0 + \dfrac{2i}{n}, \ldots, x_n = 1$; para $n = 5, 10, 15$.

(5) Repetir el Ejercicio **??** para la función $f_1 : [-1,1] \to \mathbb{R}$, $f_1(x) = |x|$ y para la función $f_2 : [-1,1] \to \mathbb{R}$, $f_2(x) = \sin(\pi x)$.

(6) Sea $f : [0,5] \to \mathbb{R}$, $f(x) = 2^x$. Sea P_n un polinomio de grado n que interpola a f en $n+1$ puntos distintos cualesquiera de dicho intervalo. Demostrar que para todo $x \in [0,5]$,

$$|P_n(x) - f(x)| \leq \frac{32.5^{n+1}}{(n+1)!}$$

(7) Sea f una función C^∞ tal que para todo $k \in \mathbb{N}$ y para todo $x \in [a,b]$ se tiene:

$$|f^k(x)| \leq C^k k!$$

Mostrar que, si $0 < C < \dfrac{1}{b-a}$ y P_n en un polinomio de grado n que interpola a f en $n+1$ puntos distintos, entonces P_n converge a f uniformemente, es decir, $\|f - P_n\|_\infty \to 0$ cuando n tiende a ∞.

(8) Sea $f : [-1,1] \to \mathbb{R}$, $f(x) = \dfrac{1}{a+x}$. Sean $(x_n)_{n\geq 0}$ una sucesión arbitraria de puntos en $[-1,1]$ y $P_n(x)$ el polinomio que interpola a $f(x)$ en $x_0, x_1, \ldots, x_n$. Demostrar que si $a > 3$ entonces P_n converge a f uniformemente.

(9) (a) Dado el intervalo $[a,b]$, sea m el punto medio entre a y b y sea $h < (b-a)/2$. Sea $p = m - h$ y $q = m + h$. Demostrar que para todo x en $[a,b]$,

$$|(x-p)(x-q)| \leq \frac{(b-a)^2}{4}.$$

(b) Sean $x_0 = a, \ldots, x_i = x_0 + \frac{b-a}{n}, \ldots, x_n = b$, $n+1$ puntos equiespaciados en el intervalo $[a,b]$. Demostrar que para todo x en $[a,b]$,

$$|(x-x_0)\ldots(x-x_n)| \leq \frac{(b-a)^{n+1}}{2^{n+1}}.$$

(10) Sea $f : [-\pi, \pi] \to \mathbb{R}$, $f(x) = \sin(x)$. Sea P_n un polinomio de grado n que interpola a f en $n+1$ puntos equiespaciados en dicho intervalo.

(a) Demostrar que para todo $x \in [-\pi, \pi]$

$$|P_n(x) - f(x)| \leq \frac{\pi^{n+1}}{(n+1)!}$$

(b) Concluir que P_n converge uniformemente a f.

(11) Sea $f : [0,1] \to \mathbb{R}$, $f(x) = \sin(\pi x) + e^x$. Sea P_n el polinomio de grado n que interpola a f en $n+1$ puntos equiespaciados.

(a) Usando el ejercicio 9, acotar el error $\|f - P_n\|_\infty$.

(b) Sea C_n la cota hallada en (a). Para $n = 1, 3, 5$ graficar simultáneamente f, $f + C_n$, $f - C_n$ y P_n.

(12) Dado un intervalo $[a,b]$, decidir como tienen que estar distribuidos $n+1$ nodos $x_0 < x_1 < \cdots < x_n$ en el intervalo de modo que exista $x \in [a,b]$ tal que

$$|(x-x_0)\ldots(x-x_n)| \sim (b-a)^{n+1}.$$

(13) Calcular el grado mínimo n que debe tener un polinomio P_n que interpola en los ceros de T_{n+1} a la función $f(x) = e^{2x}$, $x \in [-1, 1]$, para que el error $\|f - P_n\|_\infty \leq 10^{-2}$.

(14) Repetir el ejercicio anterior para $f(x) = e^x$, $x \in [0, 4]$.

(15) Para $n = 5, 10, 15$; graficar simultáneamente el polinomio $W_{n+1}(x) = \prod_{i=0}^{n}(x - x_i)$, donde $x_i = -1 + 2i/n$; $i = 0, \ldots, n$ y el polinomio de Tchebychev T_{n+1}.

(16) Repetir los Ejercicios 4 y 5 usando los polinomios que interpolan a la función f en los ceros del polinomio de Tchebychev de grado $n + 1$, para $n = 5, 10, 15$.

(17) Utilizar el método de coeficientes indeterminados para hallar un polinomio p de grado 2 que satisfaga:

$$p(1) = 0, \quad p'(1) = 7, \quad p(2) = 10$$

(18) (a) Sea $f(x) = \cos(\pi x)$, hallar un polinomio de grado menor o igual que 3 que verifique

$$p(-1) = f(-1),\ p(0) = f(0),\ p(1) = f(1),\ p'(1) = f'(1).$$

(b) Hallar un polinomio de grado menor o igual que 4 que verifique las condiciones del item anterior, más la condición

$$p''(1) = f''(1).$$

(19) Sea $f : [-1, 1] \to \mathbb{R}$ la función $f(x) = e^{2x-1}$ y sean $x_0 < \ x_1 < \ldots < x_n$ los ceros del polinomio de Tchebychev, T_{n+1}. Se interpola a f con un polinomio P de grado $\leq n + 1$ de modo que $P(x_0) = f(x_0),\ P(x_1) = f(x_1), \ldots, P(x_n) = f(x_n)$ y además $P'(x_n) = f'(x_n)$. Probar que si $n \geq 6$ entonces, el error cometido en la interpolación sobre el intervalo $[-1, 1]$ es menor que 10^{-3}.

(20) Para ilustrar qué pasa cuando se desea interpolar no sólo una función sino también sus derivadas, consideramos el problema de hallar p de grado a lo sumo 3 que verifique:

(a) $p(0) = 1, \quad p'(0) = 1, \quad p'(1) = 2, \quad p(2) = 1;$
(b) $p(-1) = 1, \quad p'(-1) = 1, \quad p'(1) = 2, \quad p(2) = 1;$
(c) $p(-1) = 1, \quad p'(-1) = -6, \quad p'(1) = 2, \quad p(2) = 1.$

Usando el método de coeficientes indeterminados, demostrar que el problema (a) tiene solución única, el problema (b) no tiene solución, y el problema (c) tiene infinitas soluciones.

(21) Analizar para que valores de $x_0,\ x_1,\ x_2$, y $\alpha_0,\ \alpha_1,\ \alpha_2$ existe un polinomio de grado 2 que satisface:

$$p(x_0) = \alpha_0,\ p(x_1) = \alpha_1,\ p'(x_2) = \alpha_2.$$

(22) Sea $f \in C^2[a, b]$, y sean $x_0 = a, x_1 = a+h, \ldots, x_n = b$, donde $h = (b-a)/n$. Considerar la poligonal $l(x)$que interpola a f en los puntos x_i, $i = 0 \ldots n$. Probar que

(a)

$$|f(x) - l(x)| \leq \frac{h^2}{2} \max_{x\in[a,b]} |f''(x)|$$

(b)

$$|f'(x) - l'(x)| \leq h \max_{x\in[a,b]} |f''(x)|$$

(23) (a) Determinar valores de α, β y γ en $\mathbb{R}$ para que S sea una función spline cúbica, siendo:

$$S(x) = \begin{cases} \alpha x^3 + \gamma x, & 0 \leq x \leq 1 \\ -\alpha x^3 + \beta x^2 - 5\alpha x + 1, & 1 \leq x \leq 2. \end{cases}$$

(b) Con los valores de α, β y γ obtenidos en el ítem anterior, decidir si S interpola a la función $f(x) = 2^x + 0.5x^2 - 0.5x - 1, 0 \leq x \leq 2$ respecto de la partición $\{0, 1, 2\}$.

(c) Graficar simultáneamente f y S en el intervalo $[0, 2]$.

(24) Sea f como en el Ejercicio 4. Utilizando `Matlab`, graficar la función f junto con una spline cúbica que la interpole en la red $\{-1, -0.75, \ldots, 0.75, 1\}$, tomando como condiciones de borde las derivadas de f.

(25) Encontrar una función del tipo $2^{ax^3+bx^2+cx+d}$ que interpole la siguiente tabla de datos:

x	-1	0	1	2
y	1	1	0.5	4

(26) Utilizando `Matlab`, encontrar y graficar una función del tipo $e^{a_4x^4+a_3x^3+\cdots+a_0}$ que interpole a la función $f(x) = 1/x$ en 5 nodos equiespaciados en el intervalo $[1, 10]$.

CAPíTULO 6

Polinomios ortogonales y aproximación por cuadrados mínimos.

En el capítulo anterior hemos discutido como aproximar una función por polinomios que interpolan a la función misma y/o a sus derivadas en algunos puntos. Hasta ahora, los métodos analizados nos permiten construir polinomios de grado n a partir de $n+1$ datos. Cierto es que, en un problema a modelizar, cuántos más datos se conocen es de esperar que se pueda lograr mayor precisión. Pero, como vimos, muchas veces polinomios de alto grado producen efectos no deseados como por ejemplo grandes oscilaciones. En este capítulo consideraremos otra forma de aproximar funciones conocida como el método de *cuadrados mínimos*. Este método nos permitirá, cuando se trate de aproximar por polinomios, contemplar una tabla de valores sin sujetar el grado del polinomio a la cantidad de datos. También será posible considerar funciones más generales que ajusten de manera natural los valores predeterminados.

En general, en esta clase de problemas uno sabe a priori a qué tipo de función corresponden los datos. Una situación frecuente es la de aproximar una tabla de más de dos valores por una recta (como muestra la Figura 6.1). Es decir, se tienen valores (x_i, y_i), $i = 0, \ldots, n$ y se quiere encontrar una recta que ajuste estos datos lo mejor posible. Si escribimos la ecuación de la recta como $y = mx + b$ nuestro problema consiste en encontrar valores de m y b que hagan que el error $|y_i - (mx_i + b)|$ sea lo mas chico posible para todo i. Por ejemplo, una manera de lograr esto sería pedir que m y b minimicen

$$\max_{0 \le i \le n} |y_i - (mx_i + b)|$$

o también podríamos pedir que minimicen

$$\sum_{i=0}^{n} |y_i - (mx_i + b)| \qquad \text{o} \qquad \sum_{i=0}^{n} |y_i - (mx_i + b)|^2.$$

De todas estas opciones es usual considerar la última, llamada "aproximación por cuadrados mínimos", debido a que es la más simple ya que el problema se reduce a resolver ecuaciones lineales.

En este capítulo estudiaremos distintos métodos para resolver éste y otros problemas. Como en general los valores de y_i corresponden a datos de una función f, podemos plantear estos problemas en el contexto de aproximación de funciones. Dada una función f consideramos:

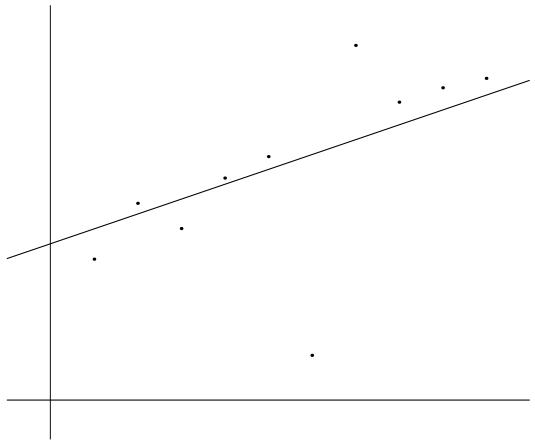

FIGURA 6.1. *Aproximación de 10 valores por una recta*

Problema A. Dados $w_0, \ldots, w_n$ constantes positivas (pesos), $m < n$ y valores $(x_i, f(x_i))$, con $i = 0, \ldots, n$ se trata de hallar $p \in \mathcal{P}_m$ que minimice

$$\sum_{i=0}^{n} w_i (p(x_i) - f(x_i))^2.$$

Problema B. Dada $w(x)$ una función positiva en $[a, b]$, dada f y $m \in \mathbb{N}$ se trata de hallar $p \in \mathcal{P}_m$ que minimice

$$\int_a^b w(x)(f(x) - p(x))^2 \, dx.$$

1. Preliminares

Nos dedicaremos especialmente al estudio de aproximaciones por polinomios. Comenzamos esta sección presentando un resultado clásico de Weierstrass que muestra que toda función continua puede aproximarse uniformemente por polinomios, en todo intervalo cerrado y acotado.

TEOREMA 6.1. *(***Weierstrass***) Sea $f \in C[a, b]$. Para todo $\varepsilon > 0$ existe un polinomio p tal que*

$$\|f - p\|_\infty < \varepsilon$$

Demostración. Damos la demostración para el intervalo $[0, 1]$, el caso general se obtiene fácilmente mediante un cambio de variables.

Definimos los polinomios de Bernstein,

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

y vamos a demostrar que $B_n f$ converge uniformemente a f en el intervalo $[0,1]$. Para eso necesitaremos calcular $B_n h_j$ para $h_j(x) = x^j$, $j = 0, 1, 2$.

Usando la fórmula del binomio de Newton, se tiene:

$$\begin{aligned}
B_n h_0(x) &= \sum_{k=0}^{n} \binom{n}{k} x^k (1-x)^{n-k} = (x+1-x)^n = 1. \\
B_n h_1(x) &= \sum_{k=0}^{n} \binom{n}{k} \frac{k}{n}\, x^k (1-x)^{n-k} = \sum_{k=1}^{n} \binom{n-1}{k-1} x^k (1-x)^{n-k} \\
&= x \sum_{k=1}^{n} \binom{n-1}{k-1} x^{k-1} (1-x)^{n-k} = x(x+1-x)^{n-1} = x. \\
B_n h_2(x) &= \sum_{k=0}^{n} \binom{n}{k} \left(\frac{k}{n}\right)^2 x^k (1-x)^{n-k} = \sum_{k=0}^{n} \binom{n-1}{k-1} \frac{k}{n}\, x^k (1-x)^{n-k} \\
&= \sum_{k=0}^{n} \binom{n-1}{k-1} \left(\frac{n-1}{n}\frac{k-1}{n-1} + \frac{1}{n}\right) x^k (1-x)^{n-k} \\
&= \frac{n-1}{n} x^2 \sum_{k=2}^{n} \binom{n-2}{k-2} x^{k-2} (1-x)^{n-k} + \frac{x}{n} \\
&= \frac{n-1}{n} x^2 (x+1-x)^{n-2} + \frac{x}{n} = x^2 + \frac{x(1-x)}{n}.
\end{aligned}$$

Dado $y \in \mathbb{R}$ consideremos la función $g_y(x) = (x-y)^2$. Desarrollando $(x-y)^2 = x^2 - 2xy + y^2$ y usando que B_n es lineal (o sea, $B_n(f_1 + f_2) = B_n f_1 + B_n f_2$ y $B_n(kf) = k B_n f$) se obtiene

$$B_n g_y(x) = g_y(x) + \frac{x(1-x)}{n}. \tag{6.1}$$

Por otra parte, como toda función continua en un intervalo cerrado es uniformemente continua, dado $\varepsilon > 0$, existe $\delta > 0$ tal que,

$$|f(x) - f(y)| \leq \varepsilon \quad \text{si} \quad |x-y| < \delta.$$

Además, para los x, y tales que $|x-y| \geq \delta$ se tiene

$$|f(x) - f(y)| \leq 2\|f\|_\infty \leq \frac{2\|f\|_\infty}{\delta^2}(x-y)^2.$$

Luego, para todo x, y, podemos asegurar que,

$$|f(x) - f(y)| \leq \varepsilon + \frac{2\|f\|_\infty}{\delta^2}(x-y)^2$$

es decir,

$$-\varepsilon - \frac{2\|f\|_\infty}{\delta^2}(x-y)^2 \leq f(x) - f(y) \leq \varepsilon + \frac{2\|f\|_\infty}{\delta^2}(x-y)^2.$$

Ahora, si $f_1 \leq f_2$, de la definición de B_n puede verse que $B_n f_1 \leq B_n f_2$; esto es B_n preserva el orden. En consecuencia, aplicando B_n en la desigualdad anterior, teniendo en cuenta (6.1), y recordando que B_n es lineal y que $B_n 1 = 1$ se obtiene (tener presente que hasta aquí estamos considerando y como una constante),

$$|B_n f(x) - f(y)| \leq \varepsilon + \frac{2\|f\|_\infty}{\delta^2}(x-y)^2 + \frac{2\|f\|_\infty}{\delta^2}\frac{x(1-x)}{n}$$

y por lo tanto, evaluando ambos lados de la desigualdad en y, resulta

$$|B_n f(y) - f(y)| \leq \varepsilon + \frac{2\|f\|_\infty}{\delta^2}\frac{y(1-y)}{n}$$

y por lo tanto

$$|B_n f(y) - f(y)| \leq 2\varepsilon$$

para n suficientemente grande independientemente de y, es decir que $B_n f$ converge uniformemente a f en el $[0, 1]$. □

Los Problemas A y B se enmarcan dentro de la teoría de espacios con producto interno. El producto interno no sólo permite definir distancia entre vectores, como vimos que se puede hacer mediante la noción de norma; sino que, además, permite introducir el concepto de ángulo entre vectores y por tanto tiene sentido hablar de ortogonalidad.

DEFINICIÓN 6.2. *Sea V un $\mathbb{R}$ espacio vectorial. Un producto interno (o producto escalar) sobre V es una función $\langle .,. \rangle : V \times V \to \mathbb{R}$ que asigna a cada par de vectores un número real de manera tal que, para todo $x, y, z \in V$ y todo $\alpha \in \mathbb{R}$ se satisfacen las propiedades:*

(i) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$;
(ii) $\langle \alpha x, y \rangle = \alpha \langle x, y \rangle$;
(iii) $\langle x, y \rangle = \langle y, x \rangle$;
(iv) $\langle x, x \rangle > 0$ *si* $x \neq 0$.

EJEMPLOS 6.3. (1) El producto interno usual en $\mathbb{R}^n$, para $x = (x_1, \ldots, x_n); y = (y_1, \ldots, y_n)$, está dado por

$$\langle x, y\rangle = \sum_{j=1}^{n} x_j y_j.$$

Es fácil ver (queda como ejercicio) que se satisfacen todas las condiciones de la definición.

(2) Otros productos internos para $\mathbb{R}^n$ similares al usual son los dados por pesos $w_j > 0$ para $j = 1, \ldots, n$:

$$\langle x, y\rangle_w = \sum_{j=1}^{n} w_j x_j y_j.$$

Ahora, si definimos la matriz $D_w \in \mathbb{R}^{n\times n}$ como:

$$D_w = \begin{pmatrix} w_1 & 0 & \ldots & 0 \\ 0 & w_2 & \ldots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \ldots & w_n \end{pmatrix}$$

el producto interno con pesos $(w_j)_{j=1}^n$ puede darse a través del producto interno usual $\langle .\ ,.\ \rangle$ y la matriz D_w,

$$\langle x, y\rangle_w = \sum_{j=1}^{n} w_j x_j y_j = \langle x, D_w y\rangle.$$

(3) Si $V = C[0, 1]$ es el espacio de funciones continuas y $f, g \in V$,

$$\langle f, g\rangle = \int_0^1 f(x)g(x)\ dx,$$

define un producto interno. Las condiciones (i)-(iii) se satisfacen gracias a la linealidad del producto y de la integral. Para asegurar que vale la condición (iv) basta ver que la integral de una función no negativa y continua, $g = f^2$, sólo puede ser nula si la función lo es. En efecto, supongamos que existe un $x_0 \in [0, 1]$ para el cual $g(x_0) = \delta > 0$. Ahora, por continuidad, existe un subintervalo $[a, b]$ tal que, para todo $x \in [a, b]$ es $g(x) > \frac{\delta}{2}$ y por ser g no negativa se tiene

$$\int_0^1 g(x)\ dx \geq \int_a^b g(x)\ dx > \frac{\delta}{2}(b - a) > 0,$$

lo que es una contradicción.

(4) Otros productos internos para espacios de funciones son los dados por una función de peso w, con $w(x) > 0$ para todo $x \in (a, b)$:

$$\langle f, g\rangle = \int_a^b f(x)g(x)w(x)\ dx.$$

En los espacios vectoriales con producto interno se tiene la norma inducida por dicho producto:

$$\|x\| = \langle x, x\rangle^{\frac{1}{2}}, \qquad \text{para todo } x \in V.$$

No es inmediato ver que con esta definición se obtiene efectivamente una norma. Esto es posible gracias a la siguiente desigualdad.

PROPOSICIÓN 6.4. **(Desigualdad de Cauchy - Schwarz)** *Si $\langle .,. \rangle$ es un producto interno sobre un espacio vectorial V, entonces*

$$|\langle x, y\rangle| \leq \langle x, x\rangle^{\frac{1}{2}} \langle y, y\rangle^{\frac{1}{2}}$$

para todo $x, y \in V$.

Demostración. Sean $x, y \in V$ dos vectores fijos. Si $\langle y, y\rangle = 0$, no hay nada que probar. Supongamos entonces que $\langle y, y\rangle \neq 0$.

Para cada $t \in \mathbb{R}$ consideramos $x - ty$, entonces

$$\begin{aligned} 0 &\leq \langle x - ty, x - ty\rangle \\ &= \langle x, x\rangle - t\langle x, y\rangle - t\langle y, x\rangle + t^2\langle y, y\rangle \\ &= \langle x, x\rangle - 2t\langle x, y\rangle + t^2\langle y, y\rangle \\ &= c - 2bt + at^2 = p(t). \end{aligned}$$

De esta manera se obtiene una función cuadrática donde $a = \langle y, y\rangle, b = \langle x, y\rangle$ y $c = \langle x, x\rangle$. Como $p(t) \geq 0$ para todo $t \in \mathbb{R}$, esta cuadrática tiene a lo sumo una raíz real y por lo tanto $4b^2 - 4ac \leq 0$. Luego,

$$0 \geq b^2 - ac = \langle x, y\rangle^2 - \langle x, x\rangle\langle y, y\rangle$$

de donde se sigue el resultado. □

COROLARIO 6.5. *Si $\langle .,. \rangle$ es un producto interno sobre un espacio vectorial V , entonces*

$$\|x\| = \langle x, x\rangle^{\frac{1}{2}}$$

define una norma sobre V.

Demostración. La única dificultad está en probar la desigualdad triangular, para eso notemos que dados $x, y \in V$ se tiene,

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y\rangle \\ &= \|x\|^2 + 2\langle x, y\rangle + \|y\|^2. \end{aligned}$$

Usando la desigualdad de Cauchy - Schwarz vale, $\langle x, y\rangle \leq |\langle x, y\rangle| \leq \|x\|\|y\|$. Luego,

$$\begin{aligned} \|x + y\|^2 &\leq \|x\|^2 + 2\|x\|\|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2 \end{aligned}$$

La desigualdad triangular se obtiene al tomar raíz cuadrada. □

La norma asociada al producto escalar usual en $\mathbb{R}^2$ o $\mathbb{R}^3$ definido en el Ejemplo 6.3 (1) corresponde a la norma $\|x\|_2$ y da la longitud del vector x. Recordemos además que este producto escalar puede escribirse, para x e y no nulos, en términos de las longitudes de ambos vectores y de θ, el ángulo entre éstos, a saber,

$$\langle x, y\rangle = \|x\|\|y\| \cos\theta.$$

En particular, x e y son ortogonales si y solo si $\langle x, y\rangle = 0$.

La gran ventaja de trabajar en espacios con producto interno es que se puede generalizar esta noción de ortogonalidad.

Notar que la desigualdad de Cauchy - Schwartz da, para todo $x, y \neq 0$

$$\frac{|\langle x, y\rangle|}{\|x\|\|y\|} \leq 1.$$

Esto permite definir el ángulo entre dos vectores x, y no nulos mediante la función coseno. Es decir $\theta \in [0, \pi]$ será el ángulo entre x e y si verifica

$$\cos(\theta) = \frac{\langle x, y\rangle}{\|x\|\|y\|}.$$

Luego resulta natural la siguiente definición.

Definición 6.6. *Si V es un espacio con producto interno $\langle ., .\rangle$, se dice que x e y son ortogonales si $\langle x, y\rangle = 0$. En este caso suele notarse $x \perp y$.*

Definición 6.7. *Dos conjuntos $A, B \subset V$ se dicen ortogonales ($A \perp B$) si $x \perp y$ para todo $x \in A$ e $y \in B$.*

El siguiente teorema relaciona los problemas de aproximación que queremos estudiar con la noción de ortogonalidad.

Teorema 6.8. *Dados S un subespacio de un espacio V con producto interno, $x \in V$ e $y \in S$, son equivalentes:*

(1) $\|x - y\| = \min_{s \in S}\{\|x - s\|\}$
(2) $\langle x - y, s\rangle = 0, \qquad \forall s \in S.$

Además, un elemento $y \in S$ que verifique alguna de las propiedades anteriores es único.

Demostración. Veamos primero que (1) implica (2). Sabemos que $y \in S$ minimiza la distancia de x a S. Como S es un subespacio, se tiene que $y + s \in S$ para todo $s \in S$, y por lo tanto,

$$\|x - y\|^2 \leq \|x - (y + s)\|^2 = \|(x - y) - s)\|^2 = \|x - y\|^2 - 2\langle x - y, s\rangle + \|s\|^2.$$

Así,

$$2\langle x-y, s\rangle \leq \|s\|^2$$

para todo $s \in S$. Si ahora consideramos $t \in \mathbb{R}$ y $s \in S$ se tiene que $ts \in S$ y de la desigualdad anterior obtenemos

$$\begin{aligned} 2\langle x-y, ts\rangle \leq & \quad \|ts\|^2 \\ 2t\langle x-y, s\rangle \leq & \quad t^2\|s\|^2 \end{aligned}$$

para todo $t \in \mathbb{R}$ y para todo $s \in S$. Para los $t > 0$ tenemos $2\langle x-y, s\rangle \leq t\|s\|^2$ y haciendo $t \to 0$ queda $2\langle x-y, s\rangle \leq 0$. Los $t < 0$ dan la otra desigualdad, $0 \leq 2\langle x-y, s\rangle$; de dónde

$$\langle x-y, s\rangle = 0 \qquad \text{para todo } s \in S.$$

Para ver que (2) implica (1), supongamos que $y \in S$ es tal que $x-y \perp s$ para todo $s \in S$. Como S es un subespacio $x-y \perp y-s$ para todo $s \in S$. Luego,

$$\begin{aligned} \|x-s\|^2 = & \quad \|(x-y)+(y-s)\|^2 \\ = & \quad \|x-y\|^2 + \|y-s\|^2 \\ \geq & \quad \|x-y\|^2. \end{aligned}$$

Tomando raíz cuadrada se obtiene que $\|x-y\| = \min_{s\in S}\{\|x-s\|\}$.

Nos queda mostrar que no puede haber más de un elemento que cumpla las condiciones (1) o (2). Para esto, veamos que si $y, \widetilde{y} \in S$ verifican (2) entonces, $y = \widetilde{y}$. En efecto, para cada $s \in S$ fijo se tiene

$$\langle x-y, s\rangle = 0, \qquad \text{y} \qquad \langle x-\widetilde{y}, s\rangle = 0,$$

luego, restando miembro a miembro, queda

$$\langle \widetilde{y}-y, s\rangle = 0,$$

en particular, tomado $s = \widetilde{y}-y \in S$ obtenemos $\|\widetilde{y}-y\| = 0$ de dónde $\widetilde{y} = y$. □

Veremos más adelante que cuando S es de dimensión finita siempre existe y en las condiciones del teorema anterior. Este y se llama proyección ortogonal de x sobre S.

2. Solución de los Problemas de Aproximación

Ahora sí, estamos en condiciones de describir los métodos para hallar las soluciones de los problemas A y B planteados. En lo que sigue de este capítulo trabajaremos sobre espacios con un producto interno.

El primer problema se puede reformular de la siguiente manera: Se considera en $\mathbb{R}^n$ el producto escalar dado por los pesos $w_0, \ldots, w_n$, es decir,

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i w_i.$$

Para los datos $(x_i, f(x_i))$, se quiere encontrar un polinomio $p \in \mathcal{P}_m$ con $n > m+1$ que minimice la distancia entre los vectores $(f(x_1), \ldots, f(x_n))$ y $(p(x_1), \ldots, p(x_n))$ en la norma asociada al producto escalar.

Si $p(x) = a_m x^m + \ldots + a_1 x + a_0$ entonces

$$\begin{pmatrix} p(x_1) \\ p(x_2) \\ \vdots \\ p(x_n) \end{pmatrix} = \begin{pmatrix} 1 & x_1 & \cdots & x_1^m \\ 1 & x_2 & \cdots & x_2^m \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^m \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_m \end{pmatrix} \tag{6.2}$$

Ahora, llamando $b = (f(x_1), \ldots, f(x_n))$ el problema se reduce a encontrar un vector $a = (a_0, \ldots, a_m) \in \mathbb{R}^{m+1}$ que minimice

$$\|Aa - b\|,$$

donde $A \in \mathbb{R}^{n \times (m+1)}$ es la matriz de 6.2.

En forma genérica el problema puede plantearse de la siguiente manera:

Dada $A \in \mathbb{R}^{n \times m}$ y $b \in \mathbb{R}^n$ se quiere hallar x tal que

$$\|Ax - b\|$$

sea lo menor posible.

Considerando el subespacio S:

$$S = \{y \in \mathbb{R}^n,\ y = Ax, \quad \text{para algún } x \in \mathbb{R}^m\}$$

el problema se transforma en hallar $y \in S$ tal que

$$\|y - b\| \leq \|s - b\| \text{ para todo } s \in S$$

y luego x tal que $Ax = y$.

En el caso del producto interno usual, es decir $\langle x, y \rangle = \sum_{j=1}^n x_j y_j$, la solución de este problema puede obtenerse resolviendo las llamadas ecuaciones normales que pueden obtenerse fácilmente a partir del Teorema 6.8 como veremos en el teorema que sigue. Recordemos que A^T denota la matriz traspuesta de A.

LEMA 6.9. *Sea $A \in \mathbb{R}^{n \times m}$, $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$. Si $\langle\ .\ ,\ .\ \rangle$ indica el producto interno usual (tanto en $\mathbb{R}^n$ como en $\mathbb{R}^m$) entonces,*

$$\langle A^T y, x \rangle = \langle y, Ax \rangle$$

Demostración.

$$\langle A^T y, x\rangle = \sum_{i=1}^{n}\left(\sum_{j=1}^{m} a_{ji}y_j\right)x_i = \sum_{j=1}^{m} y_j\left(\sum_{i=1}^{n} a_{ji}x_i\right) = \langle y, Ax\rangle$$

□

TEOREMA 6.10. *Sea $A \in \mathbb{R}^{n\times m}$ y $b \in \mathbb{R}^n$. Si $\langle\ .\ ,\ .\ \rangle$ indica el producto interno usual (tanto en $\mathbb{R}^n$ como en $\mathbb{R}^m$) entonces, son equivalentes*

(1) *$x_0 \in \mathbb{R}^m$ minimiza $\|Ax - b\|$*
(2) *x_0 es solución del sistema $A^TAx = A^Tb$.*

Además, si los vectores columnas de la matriz A son linealmente independientes, existe x_0 solución del sistema $A^TAx = A^Tb$ y es único.

Demostración. Considerando el subespacio $S = \{y \in \mathbb{R}^n,\ y = Ax,\ \text{ para } x \in \mathbb{R}^m\}$, por el Teorema 6.8 $y \in S$ es tal que $\|b - y\| = \min_{s\in S}\{\|b - s\|\}$ si y solo si $\langle b - y, s\rangle = 0$ para todo $s \in S$. Como $y \in S$ existe $x_0 \in \mathbb{R}^m$ tal que $y = Ax_0$ y $s = Ax$, con x variando en $\mathbb{R}^m$, luego la condición $\langle b - y, s\rangle = 0$ para todo $s \in S$ podemos reescribirla

$$\langle b - Ax_0, Ax\rangle = 0 \quad \forall\ x \in \mathbb{R}^m,$$

o equivalentemente, por el Lema 6.9,

$$0 = \langle A^T(b - Ax_0), x\rangle = \langle A^Tb - A^TAx_0), x\rangle \quad \forall\ x \in \mathbb{R}^m,$$

lo que ocurre si y solo si $A^TAx_0 = A^Tb$.

Para mostrar la existencia y unicidad de un elemento x_0 que cumpla con el enunciado, llamemos $A_j \in \mathbb{R}^n$ a los vectores columna de la matriz A, para $j = 1, \ldots, m$.

Si $x = (x_1, x_2, \ldots, x_m)$ el vector Ax puede escribirse en términos de las columnas de A por $Ax = \sum_{j=1}^{m} A_jx_j$. Luego, si las columnas de A son linealmente independientes resulta $Ax = 0$ si y solo si $x = 0$. Veamos que esto implica que A^TA es una matriz inversible. En efecto, si $A^TAx = 0$ entonces $\langle A^TAx, x\rangle = 0$, y por el Lema 6.9 tenemos que $\langle Ax, Ax\rangle = 0$. Es decir $\|Ax\|^2 = 0$, con lo cual $Ax = 0$ y por tanto $x = 0$.

Como la única solución del sistema $A^TAx = 0$ es la trivial, se deduce que A^TA es inversible y hay una única solución para el sistema $A^TAx = A^Tb$. □

OBSERVACIÓN 6.11. Si el producto interno no es el usual sino que viene dado por pesos w_j, o sea, $\langle x, y\rangle_w = \sum_{j=1}^{n} w_jx_jy_j$, entonces $x_0 \in \mathbb{R}^m$ minimiza $\|Ax - b\|_w$ si y solo si x_0 es solución del sistema $A^TD_wAx = A^TD_wb$.

La demostración es análoga a la del teorema anterior considerando la escritura $\langle x, y\rangle_w = \langle x, D_wy\rangle$ (ver Ejemplo 6.3 (b)).

Notemos que si el problema original

$$Ax = b$$

tiene una solución exacta x, este x también es solución de

$$A^T Ax = A^T b$$

Este sistema de $m \times m$ puede resolverse por el método de eliminación de Gauss o bien por métodos iterativos.

Ejemplo 6.12. Veamos un ejemplo sencillo, como presentamos a través de la Figura 6.1. Se quiere trazar una recta ($p(x) = a_0 + a_1 x$) que aproxime los puntos

(a_j):	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
(y_j):	0.35	0.5	0.45	0.55	0.6	0.1	0.9	0.75	0.8	0.8

Siguiendo (6.2) el sistema a resolver es:

$$\begin{pmatrix} 1 & 0.1 \\ 1 & 0.2 \\ 1 & 0.3 \\ 1 & 0.4 \\ 1 & 0.5 \\ 1 & 0.6 \\ 1 & 0.7 \\ 1 & 0.8 \\ 1 & 0.9 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 0.35 \\ 0.5 \\ 0.45 \\ 0.55 \\ 0.6 \\ 0.1 \\ 0.9 \\ 0.75 \\ 0.8 \\ 0.8 \end{pmatrix}$$

Que, después de multiplicar por la transpuesta de A, queda

$$\begin{pmatrix} 10 & 5.5 \\ 5.5 & 3.85 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \end{pmatrix} = \begin{pmatrix} 5.8 \\ 3.6 \end{pmatrix}$$

La solución es

$$a_1 = 0.497$$

$$a_0 = 0.3067$$

Es decir la recta que mejor aproxima, en el sentido de minimizar $\sum (f(x_i) - p(x_i))^2$, a f en los puntos dados es

$$p(x) = 0.497x + 0.3067$$

Esta forma de resolver no puede aplicarse, en general, para resolver el problema B. Para abordar esta clase de problemas necesitamos desarrollar más teoría relacionada con el producto interno y la idea de ortogonalización. Es decir, vamos a dar una descripción de la solución a través de la proyección ortogonal sobre un subespacio.

DEFINICIÓN 6.13. *Un subconjunto A de un espacio vectorial V se dice ortonormal si A es ortogonal y además $\langle f, f\rangle = 1$ para cualquier $f \in A$.*

Si se considera en el espacio V una base ortonormal $\mathcal{B} = \{v_1, \ldots, v_n\}$ cada elemento $x \in V$ admite una única escritura de la forma

$$x = \sum_{i=1}^{n} x_i v_i.$$

La ventaja de trabajar con una base ortonormal es que podemos describir fácilmente los escalares x_i en términos de x y de la base. En efecto, tenemos

$$\langle x, v_k\rangle = \langle \sum_{i=1}^{n} x_i v_i, v_k\rangle = \sum_{i=1}^{n} x_i \langle v_i, v_k\rangle = x_k.$$

Luego,

$$x = \sum_{i=1}^{n} \langle x, v_i\rangle v_i.$$

Ahora, dado un subespacio S de dimensión finita de V, una base ortonormal de S puede encontrarse a partir de una base dada de S mediante el proceso de ortonormalización de **Gram-Schmidt** que damos en el siguiente teorema.

TEOREMA 6.14. *Dada una base de S,*

$$\mathcal{B}_S = \{r_1, r_2, \ldots, r_m\},$$

se consideran

$$u_1 = r_1, \qquad v_1 = u_1/\|u_1\|.$$

y, para $k = 2, \ldots, m$,

$$u_k = r_k - \sum_{i=1}^{k-1} \langle r_k, v_i\rangle v_i, \quad y \quad v_k = u_k/\|u_k\|.$$

Entonces, el conjunto $\{u_1, \ldots, u_m\}$ es ortogonal y el conjunto $\{v_1, \ldots, v_m\}$ es una base ortonormal del subespacio S.

Demostración. Se hace por inducción. □

Con el siguiente teorema demostramos la existencia de la proyección ortogonal sobre S un subespacio de V, cuando S tienen dimensión finita.

TEOREMA 6.15. *Dado $x \in V$ y un subespacio $S \subset V$ de dimension finita, existe un único $y \in S$ que satisface*

$$\langle x - y, s\rangle = 0, \qquad \forall s \in S. \tag{6.3}$$

Demostración. Sea $\{v_1, \ldots, v_m\}$ una base ortonormal de S (que sabemos que existe gracias al Teorema 6.14). Veamos que el elemento $y \in S$ buscado es

$$y = \sum_{i=1}^{m} \langle x, v_i\rangle v_i. \tag{6.4}$$

En efecto, es claro que $y \in S$ ya que es una combinación lineal de elementos de la base y. Por otra parte, para verificar (6.3) es suficiente ver que se cumple para $s = v_j$, $j = 1, \ldots, m$. Pero

$$\langle x - y, v_j\rangle = \langle x, v_j\rangle - \langle \sum_{i=1}^{m} \langle x, v_i\rangle v_i, v_j\rangle = \langle x, v_j\rangle - \langle x, v_j\rangle = 0$$

donde en el último paso hemos usado la ortonormalidad de la base. La unicidad la probamos en el Teorema 6.8 □

El teorema anterior nos permite definir una aplicación

$$P : V \longrightarrow S$$

que a cada elemento $x \in V$ le asigna $Px \in S$ de tal forma que

$$\langle x - Px, s\rangle = 0, \qquad \forall s \in S$$

generalizando a espacios con producto interno la noción de proyección ortogonal conocida en $\mathbb{R}^n$. Teniendo en cuenta el Teorema 6.8, Px nos da la mejor aproximación a x por elementos del subespacio S en la norma asociada al producto interno.

Estos resultados nos permiten encontrar la mejor approximación a una función continua por polinomios de un grado dado, en la norma asociada a un producto interno. Para esto basta considerar el espacio $V = C[a, b]$ y el subespacio $S = \mathcal{P}_n$.

Aplicando el proceso de ortogonalización dado en el Teorema 6.14 a la base canónica de $\mathcal{P}_n$, es decir $\mathcal{B} = \{1, x, x^2, \ldots, x^n\}$, en un producto interno dado, obtenemos los polinomios ortogonales q_k asociados a dicho producto y los correspondientes polinomios ortonormales p_k. Estos polinomios están dados por,

$$q_0(x) = 1, \qquad p_0(x) = 1/\|q_0\|.$$

y, definiendo $h_k(x) = x^k$, para $k = 1, \ldots, n$,

$$q_k(x) = x^k - \sum_{i=1}^{k-1} \langle h_k, p_i \rangle p_i(x), \quad \text{y} \quad p_k(x) = q_k(x)/\|q_k\|.$$

Observemos que, como este procedimiento puede hacerse para cualquier $n \in \mathbb{N}$ lo que se obtiene es una sucesión de polinomios ortogonales $q_0, q_1, \ldots, q_n, \ldots$ cuyas propiedades básicas resumimos en el siguiente teorema.

TEOREMA 6.16. *Dado el espacio $V = C[a, b]$ con un producto interno, los polinomios ortogonales $q_0, q_1, \ldots, q_n, \ldots$ obtenidos mediante el proceso de Gram-Schmidt aplicado a la base canónica dada por las potencias satisfacen las siguientes propiedades. Para todo $k \in \mathbb{N}_0$,*

(1) *q_k es un polinomio mónico de grado k.*
(2) *$\{q_0, q_1, \ldots, q_k\}$ es una base ortogonal de $\mathcal{P}_k$.*
(3) *q_k es ortogonal a todo polinomio de grado menor que k.*

Las conclusiones del teorema son válidas si se considera la base canónica de $\mathcal{P}_k$: $\mathcal{B} = \{1, x, x^2, \ldots, x^k\}$. El orden en que se toman los elementos es importante. La demostración se sigue de todo lo anterior y por tanto la omitimos.

En lo que sigue consideramos fijado el producto interno y usamos la notación p_k para indicar la sucesión de polinomios ortonormales asociados a dicho producto, es decir $p_k = q_k/\|q_k\|$. Una vez obtenidos estos polinomios podemos encontrar la mejor aproximación a una función continua utilizando la teoría general que hemos visto. En efecto, tenemos

TEOREMA 6.17. *Si $f \in C[a, b]$ entonces el polinomio $p_n^* \in \mathcal{P}_n$ que satisface*

$$\|f - p_n^*\| \leq \|f - p\|, \qquad \forall p \in \mathcal{P}_n,$$

está dado por $p_n^* = Pf$, donde $P : C[a, b] \longrightarrow \mathcal{P}_n$ es la proyección ortogonal, o sea,

$$p_n^* = \sum_{i=0}^{n} \langle f, p_i \rangle p_i,$$

Demostración. Se sigue del Teorema 6.4. □

Observemos que esto resuelve simultáneamente los problemas A y B. Para resolver cualquiera de los dos hay que, primero generar los polinomios ortonormales p_j y luego calcular $\langle f, p_i \rangle$. En el caso continuo (problema B) aplicamos la teoría trabajando en el espacio de dimensión infinita $C[a, b]$ mientras que en el caso discreto (problema A) trabajamos en el espacio de dimensión finita $\mathbb{R}^{n+1}$ identificando a los valores de una función continua f con el vector $(f(x_0), \ldots, f(x_n))$. De esta forma se tiene un procedimiento alternativo al dado en el Teorema 6.10 para el problema discreto. En algunos casos el método basado en el uso de los polinomios ortogonales resulta mejor respecto de la propagación de errores de redondeo.

El teorema de Weierstrass nos permite demostrar que el error entre la f y su mejor aproximación en la norma asociada al producto interno tiende a cero cuando el grado del polinomio aproximante tiende a infinito. Este es el objetivo del siguiente teorema.

TEOREMA 6.18. *Si el producto interno en $C[a,b]$ está dado por*

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

donde w es una función positiva e integrable en (a,b) entonces,

$p_n^* \longrightarrow f$ *cuando* $n \longrightarrow \infty$.

Demostración. Por el teorema de Weierstrass, dado $\varepsilon > 0$ existe un polinomio $p \in \mathcal{P}_n$ (n depende de ε) tal que

$$\max_{a\leq x\leq b} |f(x) - p(x)| = \|f - p\|_\infty < \varepsilon.$$

Entonces

$$\begin{aligned} \|f - p_n^*\|^2 &\leq \|f - p\|^2 = \textstyle\int_a^b w(x)(f(x) - p(x))^2 \, dx \\ &\leq \|f - p\|_\infty^2 \textstyle\int_a^b w(x) \, dx \leq \varepsilon^2 \int_a^b w(x) \, dx. \end{aligned}$$

Por lo tanto,

$$\lim_{n\to\infty} \|f - p_n^*\| = 0.$$

□

COROLARIO 6.19. **(Igualdad de Parseval)** *Para un producto interno como el del teorema anterior se tiene,*

$$\|f\|^2 = \sum_{j=0}^{\infty} \langle f, p_j\rangle^2$$

Demostración. Recordemos que $p_n^* = \sum_{i=0}^n \langle f, p_i\rangle p_i$, y por lo tanto

$$\|p_n^*\|^2 = \sum_{j=0}^{n} \langle f, p_j\rangle^2.$$

Entonces, de la ortogonalidad entre $f - p_n^*$ y p_n^* se obtiene

$$\|f\|^2 = \|f - p_n^*\|^2 + \|p_n^*\|^2 = \|f - p_n^*\|^2 + \sum_{j=0}^{n} \langle f, p_j\rangle^2$$

pero por el teorema sabemos que el primer sumando del término de la derecha tiende a cero cuando n tiende a infinito con lo que concluye la demostración. □

Terminamos el capítulo dando una forma más eficiente de encontrar los polinomios ortogonales asociados a un producto interno. En efecto, el siguiente teorema muestra que cada polinomio q_n se escribe en función de los dos anteriores y por lo tanto la sucesión de polinomios ortogonales mónicos puede obtenerese por recurrencia.

TEOREMA 6.20. *Si un producto interno en $C[a,b]$ satisface $\langle xf, g\rangle = \langle f, xg\rangle$ entonces los polinomios ortogonales mónicos q_n satisfacen la relación de recurrencia*

$$q_n(x) = (x - a_n)q_{n-1}(x) - b_n q_{n-2}(x) \ , \qquad \forall n \geq 2 \tag{6.5}$$

donde a_n y b_n estan dados por

$$a_n = \frac{\langle xq_{n-1}, q_{n-1}\rangle}{\langle q_{n-1}, q_{n-1}\rangle} \qquad y \qquad b_n = \frac{\langle q_{n-1}, q_{n-1}\rangle}{\langle q_{n-2}, q_{n-2}\rangle}.$$

Demostración. Sea $n \geq 2$. Como cero es raíz del polinomio $q_n(x) - q_n(0)$ podemos escribir

$$q_n(x) - q_n(0) = xr_{n-1}$$

donde r_{n-1} es un polinomio de grado menor o igual que $n-1$. Además, como q_n es mónico, r_{n-1} también lo es. Tenemos entonces,

$$q_n(x) = xr_{n-1}(x) + q_n(0) = xq_{n-1}(x) + x(r_{n-1}(x) - q_{n-1}(x)) + q_n(0). \tag{6.6}$$

Pero como r_{n-1} y q_{n-1} son mónicos su diferencia resulta un polinomio de grado menor o igual que $n-2$ y por lo tanto, como $q_0, \ldots, q_{n-1}$ forman una base de $\mathcal{P}_{n-1}$, existen coeficientes β_j tales que

$$x(r_{n-1}(x) - q_{n-1}(x)) + q_n(0) = \sum_{j=0}^{n-1} \beta_j q_j(x)$$

y reemplazando en (6.6) obtenemos

$$q_n(x) = xq_{n-1}(x) + \sum_{j=0}^{n-1} \beta_j q_j(x). \tag{6.7}$$

Ahora, para $i < n-2$, tenemos

$$0 = \langle q_n, q_i\rangle = \langle (x + \beta_{n-1})q_{n-1} + \sum_{j=0}^{n-2} \beta_j q_j, q_i\rangle = \langle xq_{n-1}, q_i\rangle + \beta_i \langle q_i, q_i\rangle$$

donde en el último paso hemos usado la ortogonalidad de los q_j. Pero, como xq_i es un polinomio de grado menor que $n-1$, resulta

$$\langle xq_{n-1}, q_i\rangle = \langle q_{n-1}, xq_i\rangle = 0$$

y en consecuencia $\beta_i = 0$ para todo $i < n-2$. Por lo tanto, definiendo $a_n = -\beta_{n-1}$ y $b_n = -\beta_{n-2}$, (6.5) se obtiene de (6.7).

Finalmente, usando (6.5) y la ortogonalidad de los q_j tenemos,

$$0 = \langle q_n, q_{n-1}\rangle = \langle xq_{n-1}, q_{n-1}\rangle - a_n \langle q_{n-1}, q_{n-1}\rangle$$

de donde se obtiene la expresión para a_n. Análogamente,

$$0 = \langle q_n, q_{n-2}\rangle = \langle xq_{n-1}, q_{n-2}\rangle - b_n \langle q_{n-2}, q_{n-2}\rangle$$

y por lo tanto,

$$b_n = \frac{\langle xq_{n-1}, q_{n-2}\rangle}{\langle q_{n-2}, q_{n-2}\rangle}.$$

Para terminar la demostración falta ver que

$$\langle xq_{n-1}, q_{n-2}\rangle = \langle q_{n-1}, q_{n-1}\rangle,$$

pero como

$$\langle xq_{n-1}, q_{n-2}\rangle = \langle q_{n-1}, xq_{n-2}\rangle,$$

basta ver que

$$\langle q_{n-1}, xq_{n-2} - q_{n-1}\rangle = 0$$

lo que resulta del hecho de que $xq_{n-2} - q_{n-1}$ es un polinomio de grado menor que $n-1$ porque tanto xq_{n-2} como q_{n-1} son mónicos de grado $n-1$. □

Observación 6.21. Los productos internos asociados a los problemas A y B satisfacen trivialmente la hipótesis del teorema.

3. Ejercicios

(1) (a) Encontrar el polinomio de grado 1 que aproxima en el sentido de cuadrados mínimos la siguiente tabla de datos:

x	0	1	2	3	4	5	6	7	8	9
y	-.1	1.1	1.9	3.2	3.8	5	6	7.3	8.1	8.9

y el polinomio de grado 2 que aproxima en el mismo sentido la siguiente tabla de datos:

x	-1	0	1	3	6
y	6.1	2.8	2.2	6	26.9

(b) En cada caso, comparar gráficamente, usando Matlab, con el polinomio interpolador.

(2) Considerar la función $f(x) = \dfrac{1}{1+25x^2}$ en el intervalo [-1,1].
Para $n = 5, 10, 15$; graficar simultáneamente f junto con
- los polinomios que aproximan a f en el sentido de cuadrados mínimos en $n+1$ puntos equiespaciados y tienen grado $\frac{2}{5}n$ y $\frac{4}{5}n$,
- el polinomio que resulta de interpolar a f en los puntos anteriores.

(3) Probar que si se tienen $n+1$ puntos distintos, el polinomio de cuadrados mínimos de grado n coincide con el polinomio interpolador.
Concluir que para ciertas aplicaciones puede ser una mala idea aumentar el grado del polinomio de cuadrados mínimos, hasta hacerlo cercano al grado del polinomio interpolador.

(4) Sea A la matriz en $\mathbb{R}^{3\times 2}$ dada por $A = \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix}$. Mostrar que

(a) $\det(A^TA) = (ad-bc)^2 + (af-be)^2 + (cf-ed)^2$.
(b) Los rangos de las matrices A^TA y A coinciden.
(c) El polinomio de grado 1 que aproxima en el sentido de cuadrados mínimos una tabla de 3 datos es único.

(5) Aproximar la siguiente tabla de datos en el sentido de cuadrados mínimos

x	-1	0	2	3
y	0.3	-0.2	7.3	23.3

con funciones del tipo: (a) $y = a2^x + b3^x$, (b) $y = a2^x + b3^x + c$.

(6) Considerar $\text{erf} : \mathbb{R} \to \mathbb{R}$ la función dada por

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

(a) Graficar la función con el comando erf de Matlab en el intervalo $[-5, 5]$ y verificar numéricamente que $\lim_{x\to\pm\infty} \text{erf}(x) = \pm 1$.
(b) Ajustar la función erf en el sentido de cuadrados mínimos con polinomios de grado 1, 2, 5 y 10; considerando 15 puntos equiespaciados en el intervalo $[-1, 1]$. Graficar erf junto con estos polinomios en el intervalo $[-5, 5]$. Observar que la aproximación es mala fuera del intervalo $[-1, 1]$.
(c) Utilizando los mismos puntos, hallar la aproximación de cuadrados mínimos que utiliza el siguiente modelo:

$$\text{erf}(t) \sim c_1 + c_2\ e^{-t^2} + c_3\ \frac{e^{-t^2}}{1+t} + c_4\ \frac{e^{-t^2}}{(1+t)^2} + c_5\ \frac{e^{-t^2}}{(1+t)^3}.$$

Comparar el error obtenido al aproximar por la función hallada con el del item anterior.

(7) Aproximar los datos de la tabla siguiente

x	-1	0	1	2
y	8.1	3	1.1	0.5

con un modelo de la forma: $f(x) \sim a\ e^{bx}$; en el sentido de cuadrados mínimos para la función $\ln(f(x))$.

(8) Aproximar los datos de la tabla siguiente

x	-1	0	1	2
y	- 1.1	- 0.4	- 0.9	- 2.7

con un modelo de la forma: $f(x) \sim -e^{ax^2+bx+c}$, en el sentido de cuadrados mínimos para la función $\ln(f(x))$.

(9) Decidir cuáles de las siguientes aplicaciones $< , >: X \times X \to \mathbb{R}$, son productos internos, siendo $X = \{$polinomios de grado menor o igual a 1 definidos en$[0, 1]\}$.
   (a) $< f, g >= f(0) + 2g(0)$
   (b) $< f, g >= (f(0) + g(0))^2$
   (c) $< f, g >= f(0)g(0) + \int_0^1 f'(t)g'(t)dt$
   (d) $< f, g >= f(0)g(0) + f(1)g(1)$

(10) Sea $< f, g >$ cualquiera de los siguientes productos escalares:

$$\text{(a)} \quad < f, g >= \sum_0^n f(x_j)g(x_j)w_j, \qquad \text{(b)} \quad < f, g >= \int_a^b f(x)g(x)w(x)dx$$

Probar que $S = \{1, x, x^2, \ldots, x^n\}$ no puede ser un conjunto ortogonal para $n \geq 2$.

(11) **Polinomios de Laguerre.** Utilizando el método de Gram-Schmidt, calcular los primeros cuatro polinomios mónicos ortogonales con respecto al producto escalar:

$$< f, g >= \int_0^\infty e^{-x} f(x)g(x)dx.$$

(12) **Polinomios de Hermite.** Repetir el ejercicio anterior con el producto escalar

$$< f, g >= \int_{-\infty}^\infty e^{-x^2} f(x)g(x)dx.$$

(13) Considerar

$$< f, g >= \int_{-1}^1 f'(x)g'(x)\ dx$$

   (a) Probar que $< , >$ es un producto interno en S_m, el espacio generado por $\{x, x^2, x^3, \cdots, x^m\}$.
   (b) Hallar una base ortonormal para S_3.
   (c) Hallar la mejor aproximación en el sentido de cuadrados mínimos sobre S_3 para $f(x) = x^4$ y para $g(x) = 1$.

(14) Sea S el subespacio de las funciones derivables definidas en el intervalo $[-\pi, \pi]$ generado por $\{1, \cos(x), \sin(x)\}$ y considerar

$$< f, g >= f'(-\frac{\pi}{2})g'(-\frac{\pi}{2}) + f'(0)g'(0) + f(\frac{\pi}{2})g(\frac{\pi}{2}).$$

   (a) Probar que $< , >$ es un producto interno en S.
   (b) Hallar una base ortonormal para S.
   (c) Hallar la mejor aproximación en el sentido de cuadrados mínimos sobre S para $f(x) = \sin(2x)$, $g(x) = \cos(2x)$ y $h(x) = \frac{3}{2}\sin(2x) - 5\cos(2x)$.

(15) (a) Probar que el conjunto de funciones: $\{1,\ \sin(kx),\ \cos(mx),\ k, m \in \mathbb{N}\}$ es ortogonal con el producto escalar

$$< f, g >= \int_0^{2\pi} f(x)g(x)dx.$$

y calcular las normas de cada una de estas funciones.

(b) Verificar la ortogonalidad y calcular la norma de los polinomios de Tchebychev, con el producto escalar

$$< f, g >= \int_{-1}^{1} \frac{f(x)g(x)}{\sqrt{1-x^2}} dx.$$

(Sugerencia: usar el cambio de variables $u = \arcsin(x)$).

(16) Hallar los primeros 5 términos de la expansión en serie de Tchebychev para la función $f(x) = |x|$. Graficar en el intervalo $[-1, 1]$.

(17) Sea T_j el polinomio de Tchebychev de grado j; $(j \in \mathbb{N})$. Considerar las relaciones de ortogonalidad discretas para éstos polinomios:

$$\sum_{k=1}^{m} T_i(x_k)T_j(x_k) = \begin{cases} 0 & i \neq j \\ m/2 & i = j \neq 0 \\ m & i = j = 0 \end{cases}$$

donde $\{x_k;\ k = 1, \ldots, m\}$ es el conjunto de ceros de T_m.

Para una función $f : [-1, 1] \to \mathbb{R}$ se definen m coeficientes c_j, $j = 1, \ldots, m$ según

$$c_j = \frac{2}{m} \sum_{k=1}^{m} f(x_k)T_{j-1}(x_k).$$

Probar que el polinomio $\left[\sum_{k=1}^{m} c_k T_{k-1}(x)\right] - 0.5c_1$ interpola a f en las raíces de T_m.

(Sugerencia: usar Ejercicio 3).

Notar que esta fórmula proporciona una manera más directa de encontrar el polinomio interpolador en los ceros de T_m.

CAPíTULO 7

Integración numérica

En este capítulo estudiamos métodos para aproximar el valor de una integral definida en un intervalo $[a, b]$. En los cursos elementales de Cálculo se aprende que el valor $\int_a^b f(x)dx$ puede obtenerse a partir de una primitiva de f mediante la regla de Barrow. Sin embargo, en muchos casos no es posible encontrar una primitiva expresable en términos de funciones conocidas. Un ejemplo es el de la integral $\int_a^b e^{-x^2}dx$ que juega un papel muy importante en la teoría de probabilidades. Puede demostrarse que la función e^{-x^2} no tiene primitiva expresable mediante composiciones y operaciones algebraicas de las funciones conocidas (polinomios, trigonométricas, logaritmos y exponenciales). Si bien este es un ejemplo clásico, esta situación se da en una gran variedad de funciones.

En consecuencia será necesario recurrir a las llamadas reglas de integración numérica o de cuadratura. La idea básica para construir estas reglas es reemplazar la función por un polinomio puesto que:

(1) Es fácil integrar polinomios.
(2) Toda función continua puede aproximarse por polinomios.

Entonces, dada una función $f \in C[a, b]$ aproximamos el valor $\int_a^b f(x)dx$ por $\int_a^b p(x)dx$ donde p es algún polinomio que esta cerca de f.

A continuación describimos el procedimiento más usual para construir reglas de integración, el cual consiste en elegir el polinomio aproximante como uno que interpole a f. Para esto se eligen en primer lugar $n + 1$ puntos $x_0, \ldots, x_n \in [a, b]$. Sabemos que existe un único $p_n \in \mathcal{P}_n$ tal que $p_n(x_j) = f(x_j)$ para $j = 0, \ldots, n$ y definimos entonces la regla de integración numérica $Q(f)$ por

$$Q(f) = \int_a^b p_n(x)\, dx.$$

Si escribimos p_n en la forma de Lagrange (ver (5.3)), o sea,

$$p_n(x) = \sum_{i=0}^{n} f(x_j)\ell_j(x),$$

donde $\ell_j(x_j) = 1$ y $\ell_j(x_i) = 0$ para $i \neq j$, tenemos

$$\int_a^b p_n(x)\, dx = \int_a^b \sum_{j=0}^{n} f(x_j)\ell_j(x)\, dx = \sum_{j=0}^{n} f(x_j) \int_a^b \ell_j(x)\, dx = \sum_{j=0}^{n} A_j f(x_j).$$

Luego, obtenemos las fórmulas de cuadratura usuales Q para aproximar una integral buscada, de la forma:

$$\int_a^b f(x)dx \quad \sim \quad Q(f) = \sum_{j=0}^{n} A_j f(x_j) \tag{7.1}$$

donde los puntos x_j son llamados los nodos y los A_j los pesos de la integración numérica $(j = 0, \ldots, n)$.

Los pesos $A_j = \int_a^b \ell_j(x)\ dx$ dependen sólo de los nodos x_j, una vez calculados se usan para aproximar la integral de cualquier función f.

Notemos que si f es un polinomio de grado n entonces, como la interpolación en $n+1$ puntos, es exacta, la fórmula que obtuvimos para aproximar la integral será exacta sobre los polinomios de grado menor o igual que n. En otro caso, habrá que estudiar el error que se comete al utilizar este tipo de aproximaciones. Es decir, estudiaremos para cada fórmula de cuadratura el error que viene dado por:

$$R(f) = \int_a^b f(x)\,dx - \int_a^b p_n(x)\,dx = \int_a^b (f - p_n)(x)\,dx.$$

Hay, esencialmente, dos maneras de determinar una fórmula de cuadratura como en (7.1).

- Los nodos $\{x_0, x_1, \ldots, x_n\}$ están prefijados. En este caso, se trata de hallar los pesos $\{A_0, A_1, \ldots, A_n\}$. Cuando los nodos se toman equiespaciados, el problema se conoce como las *Fórmulas de Newton-Côtes*.
- Se buscan a la vez los nodos $\{x_0, x_1, \ldots, x_n\}$ y los pesos $\{A_0, A_1, \ldots, A_n\}$. Este método se conoce como *Fórmulas de cuadratura gaussiana*.

1. Fórmulas de Newton-Côtes

Si queremos aproximar la integral una función continua $f : [a, b] \to \mathbb{R}$ por la integral de un polinomio interpolador de grado n, probablemente la elección más natural para los nodos x_j es tomarlos equiespaciados en el intervalo $[a, b]$. para esto consideramos $h = (b-a)/n$ y $x_j = a + jh$ con $j = 0, \ldots, n$. Una fórmula de aproximación basada en estos puntos se conoce como "formula de Newton-Côtes cerrada" y si los puntos son tomados como $x_j = a + jh$ con $j = 1, \ldots, n-1$ se llama "fórmula de Newton-Côtes abierta" (no incluye a los extremos del intervalo). Estas fórmulas serán exactas cuando el polinomio interpolador coincida con la función f, esto es, para todo polinomio en $\mathcal{P}_n$.

1.1. Fórmulas simples de Newton-Côtes. Veamos primero las fórmulas de cuadratura si se considera el intervalo $[a, b]$ considerando nodos equiespaciados. Comencemos interpolando por una recta o por una función cuadrárica.

Regla de Trapecios: es la que se obtiene si se reemplaza en el intervalo $[a, b]$ la integral de la función f por la de la recta que une los puntos $(a, f(a))$ con $(b, f(b))$. De ahí el nombre de *trapecios* (ver Figura 7.1). Como los nodos de interpolación son los extremos del intervalo, esta fórmula también suele llamarse de *trapecios cerrada*.

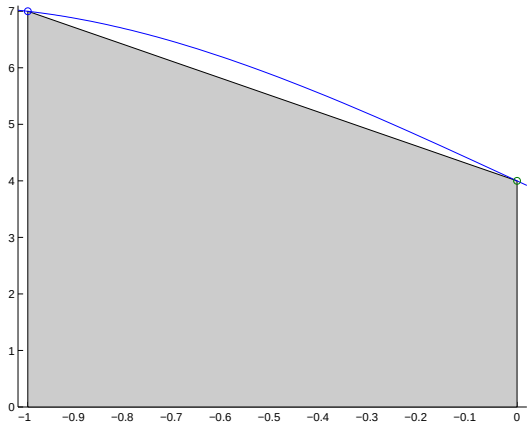

FIGURA 7.1. Regla de los Trapecios simple cerrada

La recta está dada por $p(x) = f(a) + \dfrac{f(b) - f(a)}{b - a}(x - a)$, integrado p obtenemos

$$\begin{aligned}\int_a^b p(x)\ dx &= f(a)x + \frac{f(b) - f(a)}{b - a}\frac{(x - a)^2}{2}\Big|_a^b \\ &= f(a)(b - a) + \frac{f(b) - f(a)}{2}(b - a),\end{aligned}$$

es decir:

$$\int_a^b f(x)dx \quad \sim \quad T(f) = \frac{(b - a)}{2}(f(a) + f(b)). \tag{7.2}$$

EJEMPLO 7.1. *Consideremos la función $f(x) = x^3 - 4x + 4$ en el intervalo $[-1, 0]$. ¿Cuál es el valor aproximado de la integral en este intervalo que da la regla de trapecios?*

Según vimos, se tiene

$$\int_{-1}^0 x^3 - 4x + 4\ dx \ \sim\ \frac{0 - (-1)}{2}(f(-1) + f(0)) = \frac{1}{2}(7 + 4) = \frac{11}{2}.$$

En este caso, es sencillo calcular el valor exacto de $\int_{-1}^{0} x^3 - 4x + 4 \; dx = \frac{23}{4}$ con lo cual se puede calcular exactamente el error que se comete, $R(f) = \frac{1}{4} = 0.25$.

Más adelante nos dedicaremos al estudio del error. Veamos ahora una pequeña modificación a la regla de trapecios.

Regla de Trapecios abierta: en este caso, en lugar de considerar como nodos los extremos del intervalo $[a, b]$ vamos a usar dos puntos interiores equiespaciados $\{x_1, x_2\}$. Luego, sustituímos la función f por la recta que la interpola en esos nodos (ver Figura 7.2). Para ésto partimos al intervalo $[a, b]$ en tercios, es decir en subintervalos de longitud $h = \frac{b-a}{3}$. De esta manera consideramos $\{x_1, x_2\}$ los extremos del intervalo medio, es decir $x_j = a + jh$ para $j = 1, 2$. El polinomio de grado 1 que interpola a f en esos nodos es

$$p(x) = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1).$$

Integrando p en $[a, b]$ y recordando que $h = \frac{b-a}{3}$ (ésto es: $b - a = 3h,\ x_2 - x_1 = h,\ b - x_1 = 2h$, y $a - x_1 = -h$) tenemos

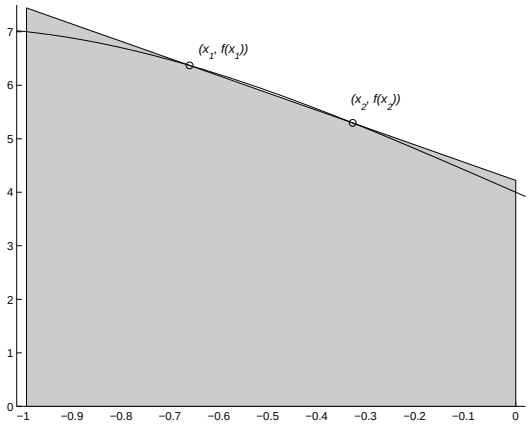

FIGURA 7.2. Regla de Trapecios simple abierta

$$\begin{aligned}
\int_a^b p(x)\; dx &= f(x_1)x + \frac{f(x_2) - f(x_1)}{x_2 - x_1}\frac{(x - x_1)^2}{2}\Big|_a^b \\
&= f(x_1)3h + \frac{f(x_1) - f(x_2)}{h}\Big[\frac{(2h)^2 - (-h)^2}{2}\Big] \\
&= 3hf(x_1) + \frac{f(x_1) - f(x_2)}{h}\frac{3h^2}{2} \\
&= 3h\Big(\frac{f(x_1) + f(x_2)}{2}\Big)
\end{aligned}$$

Luego, para $h = \dfrac{b-a}{3}$,

$$\int_a^b f(x)dx \quad \sim \quad \frac{3h}{2}(f(x_1)+f(x_2)). \tag{7.3}$$

EJEMPLO 7.2. *Consideremos nuevamente la función* $f(x)=x^3-4x+4$ *en el intervalo* $[-1,0]$. *Queremos calcular la aproximación que da la fórmula de Trapecios abierta.*

La regla de trapecios abierta tienen por nodos $\{x_1=-\frac{2}{3},\ x_2=-\frac{1}{3}\}$ con $h=\frac{1}{3}$. El valor aproximado de la integral de f en $[-1,0]$ es

$$\int_{-1}^{0} x^3-4x+4\ dx \ \sim\ \frac{1}{2}(f(-\frac{2}{3})+f(-\frac{1}{3}))=\frac{1}{2}\Big(-\frac{8}{27}+\frac{8}{3}+4-\frac{1}{27}+\frac{4}{3}+4\Big)=\frac{1}{2}\frac{53}{3}=5.8333...$$

Usando el valor exacto, ya calculado, de $\int_{-1}^{0} x^3-4x+4\ dx=\frac{23}{4}$ podemos asegurar que el error cometido es, $R(f)=-0.08333...$

Regla de Simpson: es la que se obtiene si se reemplaza en el intervalo $[a,b]$ la integral de la función f por la de una función cuadrática que interpola a f. Como para dar un único polinomio de grado 2 que interpole a f se necesitan tres nodos, se consideran los extremos del intervalo y su punto medio, es decir, $\{a,\frac{a-b}{2},b\}$ (ver Figura 7.3). Como a y b forman parte de los nodos, esta fórmula también suele llamarse de *Simpson cerrada*.

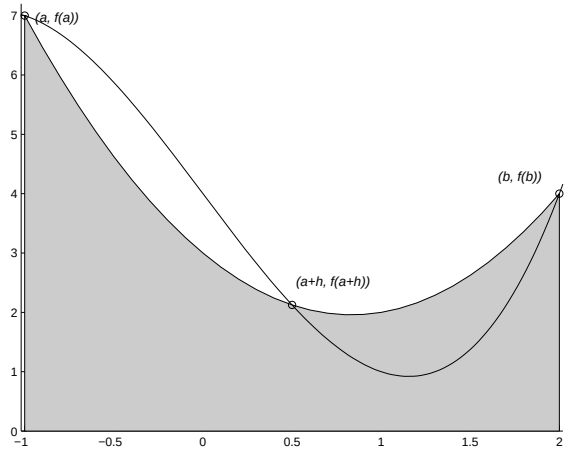

FIGURA 7.3. Regla de Simpson

Para simplificar los cálculos, queremos hallar la fórmula que le corresponde a una función continua cuando se considera el intervalo $[-1,1]$ y derivar de ésta la fórmula general. Para ésto necesitaremos el siguiente lema.

LEMA 7.3. *Si* $Q_0(f) = \sum_{j=0}^{n} A_j f(t_j)$ *es una fórmula de cuadratura para aproximar la integral* $\int_{-1}^{1} f(x)\ dx$ *entonces, para*

$$x_j = \frac{(b-a)}{2} t_j + \frac{(a+b)}{2}, \qquad \forall j = 0, \ldots, n;$$

se tiene una fórmula de cuadratura para el intervalo $[a, b]$*:*

$$\int_a^b f(x)\ dx \quad \sim \quad Q(f) = \sum_{j=0}^{n} \frac{(b-a)}{2} A_j f(x_j). \tag{7.4}$$

Demostración. Consideremos el cambio de variables $x = \alpha t + \beta$, con $\alpha = (b-a)/2$ y $\beta = (a+b)/2$ que transforma el intervalo $[-1, 1]$ en $[a, b]$. Así,

$$\int_a^b f(x)\ dx = \int_{-1}^{1} f(\alpha t + \beta)\ \alpha\ dt.$$

Aplicando la fórmula Q_0 a la función $g(t) = \alpha f(\alpha t + \beta)$ para el intervalo $[-1, 1]$ tenemos,

$$\int_a^b f(x)\ dx = \int_{-1}^{1} f(\alpha t + \beta)\ \alpha\ dt \quad \sim \quad Q_0(g) \tag{7.5}$$

con,

$$Q_0(g) = \sum_{j=0}^{n} A_j g(t_j) = \sum_{j=0}^{n} \alpha A_j f(\alpha t_j + \beta).$$

Si llamamos $x_j = \alpha t_j + \beta$, para $j = 0, \ldots, n$, tenemos que

$$x_j = \frac{(b-a)}{2} t_j + \frac{(a+b)}{2} \qquad \forall j = 0, \ldots, n.$$

Luego, podemos re-escribir la aproximación en $[a, b]$ dada en (7.5) como en la fórmula (7.4). □

Ahora sí, procedemos a dar la fórmula de Simpson, que aproxima a $\int_{-1}^{1} f(x)\ dx$, usando el polinomio interpolador p en los nodos equiespaciados $\{-1, 0, 1\}$. Si $p(x) = a_0 + a_1 x + a_2 x^2$,

$$\int_{-1}^{1} p(x)\ dx = 2\big[a_0 + \frac{a_2}{3}\big] = \frac{2}{6}\big[6a_0 + 2a_2\big].$$

Por otro lado, sabemos que $p(x) = a_0 + a_1x + a_2x^2$ verifica el sistema:

$$\begin{pmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} = \begin{pmatrix} f(-1) \\ f(0) \\ f(1) \end{pmatrix}$$

Por lo tanto, $a_0 = f(0), a_1 = \dfrac{f(1) - f(-1)}{2}$ y $a_2 = \dfrac{f(-1) - 2f(0) + f(1)}{2}$.

Luego,

$$\int_{-1}^{1} f(x)\ dx \quad \sim \quad \int_{-1}^{1} p(x)\ dx = \frac{2}{6}[f(-1) + 4f(0) + f(1)].$$

Ahora, por el Lema 7.4, se tiene para un intervalo $[a, b]$ la fórmula de Simpson simple cerrada:

$$S(f) = \frac{(b-a)}{6}\big[f(a) + 4f((a+b)/2) + f(b)\big].$$

Si escribimos la fórmula en términos de la distancia entre un nodo y otro, $h = \frac{b-a}{2}$, se tiene:

$$S(f) = \frac{h}{3}\big[f(a) + 4f(a+h) + f(b)\big]. \tag{7.6}$$

EJEMPLO 7.4. *Para la función $f(x) = x^3 - 4x + 4$ consideremos ahora el intervalo $[-1, 2]$. ¿Cuál es el valor que se obtiene al aproximar la integral de f en este intervalo si se emplea la fórmula de Simpson cerrada?*

Para este intervalo tenemos, $b - a = 3$, luego $h = \frac{3}{2}$ y $a + h = \frac{a+b}{2} = \frac{1}{2}$, entonces la fómula (7.6) nos da

$$S(f) = \frac{1}{2}\big[f(-1) + 4f(\frac{1}{2}) + f(2)\big] = \frac{1}{2}\big[7 + 4\frac{17}{8} + 4\big] = \frac{39}{4}.$$

En este caso el cálculo es exacto puesto que al calcular la integral de f en $[-1, 2]$ obtenemos por resultado $\frac{39}{4}$.

Regla de Simpson abierta: es la que se obtiene al reemplazar f por un polinomio de grado 2 que la interpole en nodos equiespaciados en el interior del intervalo $[a, b]$. Para ésto partimos al intervalo $[a, b]$ en cuartos, es decir en subintervalos de longitud $h = \frac{b-a}{4}$. De esta manera consideramos $\{x_1, x_2, x_3\}$ los extremos de los intervalos medios, es decir $x_j = a + jh$ para $j = 1, 2, 3$; (ver Figura 7.4). Como a y b no forman parte de los nodos, esta fórmula recibe el nombre de *Simpson abierta*.

Si procedemos como antes, podemos hallar el polinomios de grado 2 que interpola a una función en el intervalo $[-1, 1]$ y luego por el Lema 7.4 extendemos la fórmula a cualquier intervalo $[a, b]$. En este caso, el polinomio $p(x) = a_0 + a_1x + a_2x^2$ interpola a f en los nodos $\{-\frac{1}{2}, 0, \frac{1}{2}\}$. y resultan $a_0 = f(0),\ a_1 = f(\frac{1}{2}) - f(-\frac{1}{2})$, y $\ a_2 = 2f(-\frac{1}{2}) - 4f(0) + 2f(\frac{1}{2})$.

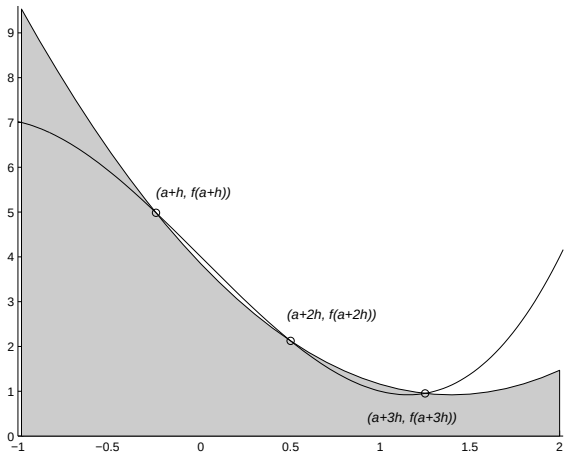

FIGURA 7.4. Regla de Simpson abierta

Luego, $\int_{-1}^{1} p(x)\ dx = \frac{2}{3}\big[3a_0 + a_2\big] = \frac{2}{3}\big[2f(-\frac{1}{2}) - f(0) + 2f(\frac{1}{2})\big]$. Al pasar a un intervalo $[a, b]$ por medio del Lema 7.4 y escribiendo la fórmula en términos del paso $h = \frac{b-a}{4}$ obtenemos,

$$\int_a^b f(x)\ dx \quad \sim \quad \frac{4h}{3}\big[2f(a+h) - f(a+2h) + 2f(a+3h)\big].$$

EJEMPLO 7.5. *Consideremos nuevamente la función $f(x) = x^3 - 4x + 4$ en el intervalo $[-1, 2]$. Queremos hallar una aproximación de la integral de f en dicho intervalo por medio de la regla de Simpson abierta.*

Como $h = \frac{b-a}{4} = \frac{3}{4}$, entonces $a + h = -\frac{1}{4},\ a + 2h = \frac{1}{2},\ a + 3h = \frac{5}{4}$, así tenemos

$$\int_{-1}^{2} x^3 - 4x + 4\ dx \quad \sim 1\big[2f(-\frac{1}{4}) - f(\frac{1}{2}) + 2f(\frac{5}{4})\big] = \big[2\frac{319}{64} - \frac{17}{8} + 2\frac{61}{64}\big] = \frac{624}{64} = \frac{39}{4},$$

que vuelve a ser un cálculo exacto.

Es claro que la fórmula de Simpson es exacta para polinomios de $\mathcal{P}_2$, al igual que la fórmula de Trapecios lo es para polinomios de $\mathcal{P}_1$. Esto motiva la siguiente definición.

DEFINICIÓN 7.6. *Decimos que una fórmula de cuadratura $Q(f) = \sum_{j=0}^{n} A_j f(x_j)$ tienen grado de exactitud k, si $\int_a^b p(x)\ dx = Q(p)$ para todo polinomio $p \in \mathcal{P}_k$ y no para $\mathcal{P}_{k+1}$.*

OBSERVACIÓN 7.7. Toda fórmula de cuadratura $\int_a^b f(x)\,dx \quad \sim \quad Q(f) = \sum_{j=0}^{n} A_j f(x_j)$ es lineal. Es decir, $Q(\alpha f + g) = \alpha Q(f) + Q(g)$ y este valor aproxima en $[a, b]$ la integral de $\alpha f + g$ para todo $\alpha \in \mathbb{R}$, f y g funciones.

En virtud de este resultado, podemos reducir el estudio del grado de exactitud al comportamiento de la fórmula sobre una base del espacio de polinomios. Esto queda expresado en la siguiente observación.

OBSERVACIÓN 7.8. Una fórmula de cuadratura Q tiene grado de exactitud k si y solo si es exacta para la base de $\mathcal{P}_k$, $\mathcal{B} = \{1, x, x^2, \ldots, x^k\}$ y no lo es para el polinomio x^{k+1}. Esto es, la igualdad

$$\int_a^b x^m\,dx = \sum_{j=0}^{n} A_j x_j^m$$

debe verificarse para todo $m = 0, \ldots, k$ y no para $m = k + 1$.

Además, gracias al Lema 7.4, una fórmula de cuadratura tiene grado de exactitud k independientemente del intervalo $[a, b]$ para el cual está calculada.

EJEMPLO 7.9. *Se quiere calcular el grado de exactitud de la fórmula de Simpson cerrada.*

Es claro que las fórmulas de Simpson, tanto abiertas como cerradas, son exactas para polinomios de $\mathcal{P}_2$. Luego, por Observación 7.8, basta ver qué sucede con $x^3, x^4, \ldots$, y esto puede hacerse, sin perder generalidad, en el intervalo $[-1, 1]$. Tenemos

$$\begin{aligned}\int_{-1}^{1} x^3\,dx &= 0 = \frac{2}{6}[(-1)^3 + 4(0)^3 + (1)^3],\\ \int_{-1}^{1} x^4\,dx &= \frac{2}{5} \neq \frac{2}{3} = \frac{1}{3}[(-1)^4 + 4(0)^4 + (1)^4].\end{aligned}$$

Luego, la fórmula de Simpson cerrada tiene grado de exactitud $k = 3$. Lo mismo sucederá para la fórmula abierta.

2. Estimación del error

Antes de estudiar cómo pueden mejorarse las aproximaciones obtenidas, veamos cuál es el error que se comete al utilizarlas. Las fórmulas de Trapecios y Simpson se obtienen de integrar el polinomio interpolador de grado 1, 2 respectivamente. También podríamos buscar otras fórmulas interpolando con polinomios de mayor grado.

En lo que sigue notaremos por $I(f) = \int_a^b f(x)\,dx$.

El Teorema 5.4 da, para cualquier función $f \in C^{n+1}[a,b]$ una fórmula que mide el error cuando se considera su polinomio interpolador $p_n \in \mathcal{P}_n$ en lugar de f:

$$E_n(x) = f(x) - p_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} W_{n+1}(x) \qquad (\text{con } \xi \in (a,b) \text{ que depende de) } x.$$

Luego, si Q es una fórmula de cuadratura como en (7.1) podemos expresar el error de integración por

$$R(f) = I(f) - Q(f) = \int_a^b (f - p_n)(x)\ dx.$$

Es decir

$$R(f) = \int_a^b \frac{f^{(n+1)}(\xi)}{(n+1)!} W_{n+1}(x)\ dx.$$

Error para la Regla de Trapecios: Para estimar el error citamos el siguiente teorema del que omitimos dar una demostración.

TEOREMA 7.10. **(Valor Medio Integral Generalizado).** *Si g, h son funciones continuas en $[a,b]$ tales que g no cambia de signo, entonces existe $\eta \in (a,b)$ tal que*

$$\int_a^b g(x)h(x)\ dx = h(\eta) \int_a^b g(x)\ dx.$$

Para calcular el error de la regla de los trapecios recordemos que los nodos son $\{a,b\}$, y por tanto $W_2(x) = (x-a)(x-b)$, que no cambia de signo en $[a,b]$, entonces usando el teorema anterior,

$$\begin{aligned} R(f) &= I(f) - Q(f) = \int_a^b \frac{f''(\xi)}{2!} W_2(x)\ dx \\ &= \frac{f''(\eta)}{2} \int_a^b (x-a)(x-b)\ dx \\ &= -\frac{f''(\eta)}{12}(b-a)^3. \end{aligned}$$

Notar que hay una forma directa de calcular $\int_a^b (x-a)(x-b)\ dx$ que es calcularla por partes. Si derivamos $(x-b)$ e integramos $(x-a)$, tenemos

$$\int_a^b (x-a)(x-b)\ dx = (x-b)\frac{(x-a)^2}{2}\Big|_a^b - \int_a^b \frac{(x-a)^2}{2}\ dx = -\frac{(x-a)^3}{6}\Big|_a^b = -\frac{(b-a)^3}{6}.$$

Finalmente, si $h = b - a$ es la distancia que hay entre los dos nodos, el error puede expresarse por

$$R(f) = -\frac{h^3}{12} f''(\eta), \qquad \text{para algún } \eta \in (a,b). \tag{7.7}$$

EJEMPLO 7.11. *Estimar el error cometido al aproximar* $\int_0^{\frac{1}{2}} x^4 - x^2 + 2x + 3 \, dx$ *por la fórmula de trapecios cerrada. ¿Cuál es el valor de dicha aproximación? ¿Qué análisis se puede hacer si se considera la misma función en el intervalo* $[-1,1]$*?*

Tenemos $h = \frac{1}{2}$, $f(x) = x^4 - x^2 + 2x + 3$ y $f''(x) = 12x^2 - 2$. Como $R(f) = -\frac{h^3}{12} f''(\eta)$ para algún $\eta \in (0, \frac{1}{2})$, acotamos $|f''(x)|$ para todo $x \in [0, \frac{1}{2}]$. Esto es, $|f''(x)| \leq 2$, puesto que alcanza su valor máximo en el extremo izquierdo del intervalo.

Luego, $|R(f)| = \frac{h^3}{12} |f''(\eta)| \leq \frac{1}{8}\frac{1}{12} 2 = 0.020833...$

El valor de la aproximación está dado por $T(f) = \frac{1}{4}(f(0) + f(\frac{1}{2})) = \frac{1}{4}(3 + \frac{61}{16}) = 3.81250$

A veces, al estimar el error perdemos información. Por ejemplo, si queremos estimar el error cometido al considerar $x \in [-1,1]$, no logramos un resultado muy fino. Por un lado $|f''(x)| = |12x^2 - 2| \leq 10$, pues alcanza su valor máximo en los extremos del intervalo y $h = b - a = 2$, así,

$$|R(f)| = \frac{h^3}{12} |f''(\eta)| \leq \frac{8}{12} 10 = \frac{20}{3} = 6.666...$$

Aunque $\int_{-1}^{1} x^4 - x^2 + 2x + 3 \, dx = \frac{86}{15} = 5.7333...$ y el valor que arroja la fórmula es $T(f) = \frac{2}{2}(f(-1) + f(1)) = 6$ y el error real es $0.2666...$

Error para la Regla de Simpson: es un poco más difícil de analizar pues el polinomio $W_3(x) = (x-a)(x - \frac{a+b}{2})(x-b)$ cambia de signo en $[a,b]$,

Sean $x_0 = a$, $x_1 = (a+b)/2$ y $x_2 = b$. Definamos el polinomio cúbico auxiliar, $p_3(x)$ como aquel polinomio que verifica

$$\begin{aligned} p_3(x_0) &= f(x_0), \\ p_3(x_1) &= f(x_1), \\ p_3(x_2) &= f(x_2), \\ p_3'(x_1) &= f'(x_1), \end{aligned}$$

(dejamos como ejercicio probar que un tal p_3 existe). Observemos que $S(f) = S(p_3)$ pues p_3 interpola a f en x_0, x_1, x_2. Además, como la regla de Simpson es exacta en polinomios de grado 3, tenemos que $S(p_3) = I(p_3)$, entonces

$$I(f) - S(f) = I(f) - S(p_3) = I(f) - I(p_3).$$

Para acotar esto necesitamos acotar $f(x) - p_3(x)$. Para x fijo distinto de x_0, x_1, x_2, definimos $\phi(t)$ para $t \in [a,b]$ por

$$\phi(t) = f(t) - p_3(t) - (f(x) - p_3(x)) \left(\frac{(t-x_0)(t-x_1)^2(t-x_2)}{(x-x_0)(x-x_1)^2(x-x_2)} \right).$$

Entonces $\phi(x)$ tiene al menos cuatro ceros en $[a,b]$, x_0, x_1, x_2 y x. Por el teorema de Rolle ϕ' tiene al menos tres ceros en (a,b) que están entre los cuatro ceros de ϕ. Por construcción $\phi'(x_1) = 0$, en consecuencia ϕ' tiene al menos cuatro ceros. Si f es C^4 encontramos que existe un punto $\xi \in (a,b)$ tal que

$$\phi^{(iv)}(\xi) = 0.$$

De la definición de ϕ esto es equivalente a

$$f(x) - p_3(x) = \frac{f^{(iv)}(\xi)}{4!}(x-x_0)(x-x_1)^2(x-x_2).$$

Como la función $(x-x_0)(x-x_1)^2(x-x_2)$ no cambia de signo en (a,b) podemos obtener,

$$I(f) - I(p_3) = \int_a^b \frac{f^{(iv)}(\xi)}{4!}(x-x_0)(x-x_1)^2(x-x_2)\ dx.$$

O bien, por el valor medio

$$I(f) - I(p_3) = \frac{f^{(iv)}(\eta)}{4!} \int_a^b (x-x_0)(x-x_1)^2(x-x_2)\ dx.$$

Ahora observamos que, llamando $h = (b-a)/2$,

$$\int_a^b (x-x_0)(x-x_1)^2(x-x_2)\ dx = \frac{-4h^5}{15}.$$

Entonces

$$R(f) = I(f) - S(f) = I(f) - I(p_3) = \frac{f^{(iv)}(\eta)}{4!}(\frac{-4h^5}{15}) = \frac{-h^5}{90} f^{(iv)}(\eta).$$

Ejemplo 7.12. *Aproximar* $\int_0^1 e^{-x^2}\ dx$ *mediante la regla de Simpson cerrada y estimar el error que se comete al efectuar dicho cálculo.*

Tenemos $h = \frac{1}{2}$, $f(x) = e^{-x^2}$, así

$$\int_0^1 e^{-x^2}\ dx \quad \sim \quad \frac{1}{6}(f(0) + 4f(\frac{1}{2}) + f(1)) = \frac{1}{6}(1 + e^{-\frac{1}{4}} + e^{-1}) = 0.74951...$$

Para estimar el error consideramos $f^{iv}(x) = 4(4x^4 - 12x^2 + 3)e^{-x^2}$. Como $R(f) = \frac{-h^5}{90} f^{(iv)}(\eta)$ para algún $\eta \in (0,1)$, acotamos $|f''(x)|$ para todo $x \in [0,1]$.

Por una parte tenemos que en $[0,1]$, $e^{-x^2} \leq 1$ y por otra parte puede verse que $|4x^4 - 12x^2 + 3|$ alcanza su valor máximo en el extremo superior del intervalo. Luego, $|4x^4 - 12x^2 + 3| \leq 5$ en $[0,1]$ y por lo tanto

$$|R(f)| \leq \frac{1}{90}\left(\frac{1}{2}\right)^5 20 = 0.006944...$$

3. Fórmulas de cuadratura compuestas

Si queremos aumentar la presición al aproximar $\int_a^b f(x)\ dx$, podemos aumentar el número de nodos. Esto es, considerar $n+1$ nodos y el polinomio de $\mathcal{P}_n$ que interpola a f en esos nodos, con $n \in \mathbb{N}$ grande. Veremos más adelante que ésto no siempre es conducente. Como vimos en el Capítulo 5, aumentar el grado del polinomio interpolador puede producir errores grandes en la aproximación, los que se trasladarían al cálculo de la integral. Otro método, que es que que vamos a desarrollar en esta sección, es el de partir el intervalo $[a,b]$ en pequeños subintervalos y en cada uno de estos aplicar una aproximación del tipo Trapecios o Simpson. Este último procedimiento se conoce como "cuadraturas compuestas".

La idea general es como sigue. Partimos el intervalo $[a,b]$ en subintervalos eligiendo puntos x_j con $a = x_0 < x_1 < \ldots < x_n = b$. Sabemos que

$$I(f) = \int_a^b f(x)\ dx = \sum_{j=0}^{n-1} \int_{x_j}^{x_{j+1}} f(x)\ dx.$$

Ahora si para cada

$$I_j(f) = \int_{x_j}^{x_{j+1}} f(x)\ dx$$

tenemos una fórmula de cuadratura $Q_j(f)$ y consideramos el error respectivo,

$$R_j(f) = I_j(f) - Q_j(f)$$

obtenemos

$$\begin{aligned} R(f) = \sum_{j=0}^{n-1} R_j(f) &= \sum_{j=0}^{n-1} (I_j(f) - Q_j(f)) \\ &= \int_a^b f(x)\ dx - \sum_{j=0}^{n-1} Q_j(f) \end{aligned}$$

Esto es, la fórmula de cuadratura será

$$\int_a^b f(x)\ dx \sim \sum_{j=0}^{n-1} Q_j(f) \quad \text{con error} \quad R(f) = \sum_{j=0}^{n-1} R_j(f). \tag{7.8}$$

Para dar una exprexión de las fórmulas y estudiar simultáneamente el error cometido en cada caso vamos a necesitar el siguiente lema.

LEMA 7.13. *Sea $g \in C([a,b])$ y sean $\{a_0, \ldots, a_k\}$ constantes con el mismo signo y $\{t_0, \ldots, t_k\} \in [a,b]$, entonces se tiene*

$$\sum_{j=0}^{k} a_j g(t_j) = g(\eta) \sum_{j=0}^{k} a_j$$

para algún $\eta \in [a, b]$.

Demostración. Sea $m = \min g(x)$ y $M = \max g(x)$ en $[a, b]$. Podemos suponer que $a_j \geq 0$ para todo $j = 1, \ldots, k$, luego

$$\begin{array}{lccccc} \text{para cada } j: & m & \leq & g(t_j) & \leq & M \\ (a_j \geq 0) & m\ a_j & \leq & a_j g(t_j) & \leq & M\ a_j \\ \text{(sumando)} & m \sum_{j=0}^{k} a_j & \leq & \sum_{j=0}^{k} g(t_j) a_j & \leq & M \sum_{j=0}^{k} a_j \end{array}$$

Ahora, definimos la función $G : [a, b] \to \mathbb{R}$,

$$G(x) = g(x) \sum_{j=0}^{k} a_j,$$

como G es un múltiplo de g, resulta continua en $[a, b]$. Además, el valor máximo de G en $[a, b]$ es $M \sum_{j=0}^{k} a_j$ y el valor mínimo es $m \sum_{j=0}^{k} a_j$. Entonces, por el teorema del valor medio, existe $\eta \in [a, b]$ tal que

$$G(\eta) = \sum_{j=0}^{k} a_j g(t_j),$$

es decir

$$g(\eta) \sum_{j=0}^{k} a_j = \sum_{j=0}^{k} a_j g(t_j),$$

como queríamos demostrar. □

Ahora estamos en condiciones de desarrollar las fórmulas de cuadratura compuestas. Consideraremos el caso de nodos equiespaciados. Esto nos permitirá aprovechar las fórmulas ya calculadas (Trapecios y Simpson) dado que la distancia entre dos nodos, que también llamaremos 'paso h' no varía.

Regla de Trapecios compuesta: para la fórmula cerrada se tiene que tanto a como b son nodos, luego tomamos los nodos $x_j = a + jh$ para $j = 0, \ldots, n-1$ con $h = (b-a)/n$.

La fórmula (7.2) nos da para cada integral

$$\int_{x_j}^{x_{j+1}} f(x)\ dx \quad \sim \quad T_j(f) = \frac{h}{2}(f(x_j) + f(x_{j+1})),$$

Luego,

$$\begin{aligned} T(f) &= \sum_{j=0}^{n-1} \frac{h}{2}(f(x_j)+f(x_{j+1})) \\ &= \frac{h}{2}\left[f(x_0)+f(x_1)+f(x_1)+f(x_2)+\ldots+f(x_{n-1})+f(x_n)\right] \\ &= \frac{h}{2}\left[f(x_0)+\sum_{j=1}^{n-1} 2f(x_j)+f(x_n)\right] \end{aligned}$$

Entonces la cuadratura compuesta usando la regla de Trapecios cerrada viene dada por

$$T(f) = \frac{h}{2}\left[f(x_0)+\sum_{j=1}^{n-1} 2f(x_j)+f(x_n)\right] \tag{7.9}$$

Como para cada subintervalo se comete un error (ver (7.7)) $R_j(f) = -\frac{f''(\eta_j)}{12}h^3$ se tiene

$$R(f) = \sum_{j=0}^{n-1} -\frac{f''(\eta_j)}{12}h^3 = -\frac{h^3}{12}\sum_{j=0}^{n-1} f''(\eta_j).$$

Ahora, gracias al Lema 7.13 (con $a_j = 1$ para todo j) y teniendo en cuenta que $h = (b-a)/n$ si y solo si $n = (b-a)/h$, tenemos que existe $\eta \in (a,b)$ tal que

$$R(f) = -\frac{h^3}{12}nf''(\eta) = -\frac{h^3}{12}\frac{b-a}{h}f''(\eta) = -\frac{h^2}{12}(b-a)f''(\eta). \tag{7.10}$$

Ejemplo 7.14. *Determinar el número n de subintervalos necesario para que el error cometido con la regla de Trapecios compuesta de una aproximación de la integral $\int_0^1 e^{-x^2}\,dx$ con error menor que 10^{-4}.*

Para hallar el número de subintervalos a considerar usamos la expresión del error (7.10). Debemos acotar $|f''(x)|$ para $x \in [0,1]$ siendo $f''(x) = (4x^2-2)e^{-x^2}$, Como $e^{-x^2} \leq 1$ y $|4x^2-2| \leq 2$ en este intervalo, se tiene:

$$|R(f)| = \frac{h^2}{12}(b-a)|f''(\eta)| \leq \frac{h^2}{12}2 = \frac{1}{6}\Big(\frac{1}{n}\Big)^2$$

Si tomamos $n > 40.8248...$ podemos asegurar que $|R(f)| < 10^{-4}$. Es decir, basta tomar $n = 41$.

Regla de Simpson compuesta: se trata de obtener una fórmula del tipo (7.8) cuando se usa la fórmula de Simpson en cada partición del intervalo $[a, b]$.

La fórmula (7.6) nos da para cada integral

$$\int_{x_j}^{x_{j+1}} f(x)\, dx \quad \sim \quad S_j(f) = \frac{h}{3}(f(x_j) + 4f(\frac{x_j + x_{j+1}}{2}) + f(x_{j+1})).$$

Como interpolamos f por un polinomio de grado 2, en puntos equiespaciados, en cada integral intervienen los nodos $\{x_j, \frac{x_j+x_{j+1}}{2}, x_{j+1}\}$. Así, el paso h entre dos nodos de cada integral es la longitud media del intervalo $[x_j, x_{j+1}]$. Es decir, $h = \frac{1}{2}\frac{b-a}{n} = \frac{b-a}{2n}$. Luego,

$$S(f) = \sum_{j=0}^{n-1} \frac{h}{3}(f(x_j) + 4f(\frac{x_j + x_{j+1}}{2}) + f(x_{j+1}))$$

fórmula que podemos expresar como

$$S(f) = \frac{h}{3}\left[f(a) \ + \ 2\sum_{j=0}^{n-1} f(x_j) \ + \ 4\sum_{j=0}^{n-1} f(\frac{x_j + x_{j+1}}{2}) \ + \ f(b)\right] \tag{7.11}$$

Para analizar el error cometido al usar esta fórmula, recordemos que en cada subintervalo el error está dado por

$$R_j(f) = \frac{-h^5}{90} f^{(iv)}(\eta_j).$$

$$R(f) = \sum_{j=0}^{n-1} \frac{-h^5}{90} f^{(iv)}(\eta_j) = \frac{-h^5}{90} \sum_{j=0}^{n-1} f^{(iv)}(\eta_j).$$

Por el Lema 7.13 (con $a_j = 1$ para todo j) y teniendo en cuenta que $h = \frac{b-a}{2n}$ si y solo si $n = \frac{b-a}{2h}$, tenemos que existe $\eta \in (a, b)$ tal que

$$R(f) = \frac{-h^5}{90} n f^{(iv)}(\eta) = \frac{-h^5}{90}\frac{b-a}{2h} f^{(iv)}(\eta) = \frac{-h^4}{180}(b-a) f^{(iv)}(\eta). \tag{7.12}$$

Ejemplo 7.15. *Determinar el número n de subintervalos necesario para que el error cometido con la regla de Simpson compuesta de una aproximación de la integral $\int_0^1 e^{-x^2}\, dx$ con error menor que 10^{-4}. Comparar con el ejemplo 7.14*

El error viene dado por la fórmula (7.12). Necesitamos acotar $f^{(iv)}(x)$ en el intervalo $[0,1]$. Usamos la cota hallada en el Ejemplo 7.12. Esto es, $|f^{(iv)}(x)| = |4(4x^4 - 12x^2 + 3)e^{-x^2}| \leq 20$. Entonces, con $h = \frac{b-a}{2n} = \frac{1}{2n}$ se tiene

$$|R(f)| = \frac{h^4}{180}|f^{(iv)}(\eta)| \leq \frac{1}{9}\Big(\frac{1}{2n}\Big)^4.$$

Si tomamos $n > 2.886...$ podemos asegurar que $|R(f)| < 10^{-4}$. Es decir, basta tomar $n = 3$, mientras que para la regla de Trapecios compuesta podíamos asegurar el mismo error partiendo en 41 subintervalos de igual longitud.

Finalizaremos esta sección con una aplicación de los métodos hasta ahora estudiados al cálculo aproximado de integrales múltiples.

Observación 7.16. Es posible aplicar en forma iterada las reglas de integración que acabamos de desarrollar para aproximar integrales múltiples de funciones continuas, para las cuales el teorema de Fubini puede aplicarse.

Por ejemplo, si $D \subset \mathbb{R}^2$ es una región que podemos describir por medio de funciones reales, podemos escribir (si la región es de Tipo 1)

$$\iint_D f(x,y)\ dxdy = \int_a^b \int_{\phi(x)}^{\psi(x)} f(x,y)\ dydx,$$

y calcular las integrales iteradas por los procedimientos anteriores.

Ejemplo 7.17. *Para $f : \mathbb{R}^2 \to \mathbb{R}$ una función continua y $D = [0,1]\times[0,1]$, se define la función* $F(x) = \displaystyle\int_0^1 f(x,y)\ dy$ *y luego*

- *Se aproximan los valores $F(0), F(\frac{1}{2}), F(1)$ con la regla de Simpson.*
- *Se aproxima $\displaystyle\int_0^1 F(x)\ \ dx$ usando otra vez la misma regla.*

El valor de $F(0)$ es el valor de la integral $\int_0^1 g(y)\ dy$ con $g(y) = f(0,y)$, luego aplicando la regla de Simpson simple cerrada tenemos

$$F(0) \ \sim\ \tfrac{1}{6}\Big[f(0,0) + 4f(0,\tfrac{1}{2}) + f(0,1)\Big]$$

Análogamente se obtiene los otros dos valores:

$$\begin{aligned} F(\tfrac{1}{2}) &\sim \tfrac{1}{6}\Big[f(\tfrac{1}{2},0) + 4f(\tfrac{1}{2},\tfrac{1}{2}) + f(\tfrac{1}{2},1)\Big] \\ F(1) &\sim \tfrac{1}{6}\Big[f(1,0) + 4f(1,\tfrac{1}{2}) + f(1,1)\Big] \end{aligned}$$

Ahora,

$$\int_0^1 F(x)\ \ dx \ \sim\ \tfrac{1}{6}\Big[F(0) + 4F(\tfrac{1}{2}) + F(1)\Big],$$

donde cada valor $F(0), F(\frac{1}{2})$ y $F(1)$ se reemplazan por los valores aproximados ya calculados.

En la forma explícita de esta regla aparecen los 9 nodos:

$$\Big\{(0,0),(\frac{1}{2},0),(1,0),(0,\frac{1}{2}),(\frac{1}{2},\frac{1}{2}),(1,\frac{1}{2}),(0,1),(\frac{1}{2},1),(1,1)\Big\}$$

para los cuales se debe calcular el valor de $f(x,y)$.

4. Convergencia de los métodos de cuadratura

Es natural preguntarse si el procedimiento estudiado converge a la integral cuando el núnero de nodos tiende a infinito. Esto es, si $Q_n(f)$ es la aproximación de $\int_a^b f(x)\ dx$ que se hace a través de un polinomio que interpola a f en $n+1$ nodos, ¿vale que $Q_n(f) \to \int_a^b f(x)\ dx$ cuando $n \to \infty$? En esta sección veremos una condición para que esto suceda.

Si bien hasta ahora aproximamos el valor de $\int_a^b f(x)\ dx$, podríamos haber tratado un caso más general, el de aproximar $\int_a^b f(x)w(x)\ dx$, con $w : [a,b] \to \mathbb{R}$ una función de peso positiva. Se recupera exactamente lo estudiado hasta ahora al considerar $w(x) = 1$ para todo $x \in [a,b]$

En tal caso una cuadratura

$$\int_a^b f(x)w(x)dx \quad \sim \quad Q(f) = \sum_{j=0}^{n} A_j f(x_j)$$

tendrá pesos A_j dependiendo de los nodos $\{x_0, x_1, \ldots, x_n\}$ y de la función w.

Cuando aproximamos f por el polinomio interpolador de Lagrange y usamos la base de Lagrange para tal fin, se tendrá que $A_j = \int_a^b \ell_j(x)w(x)\ dx$ para $j = 0, \ldots, n$.

Para estudiar la convergencia de estas cuadraturas cuando se incrementa el número de nodos usaremos notación más específica. Es decir, tanto la base de Lagrange como la cuadratura, los nodos y pesos de la misma se indicarán con el n correspondiente. El siguiente teorema da una condición que asegura la convergencia.

Teorema 7.18. *Dada $f \in C[a,b]$ y dado $n \in \mathbb{N}$ notamos por*

$$I(f) = \int_a^b f(x)w(x)\ dx, \quad \textit{y definimos} \quad Q_n(f) = \sum_{j=0}^{n} A_j^{(n)} f(x_j^{(n)})$$

donde los $A_j^{(n)}$ están dados por

$$A_j^{(n)} = \int_a^b \ell_j^{(n)}(x)w(x)\ dx.$$

Si existe una constante K tal que

$$\sum_{j=0}^{n} |A_j^{(n)}| \leq K \qquad \forall n$$

entonces

$$\lim_{n\to\infty} Q_n(f) = I(f)$$

Demostración. Por el teorema de Weierstrass, dado $\varepsilon > 0$ existe un polinomio $q_N \in P_N$ (N depende de ε) tal que

$$\max_{a\leq x\leq b} |f(x) - q_N(x)| = \|f - q_N\|_\infty \leq \varepsilon.$$

Observemos que como Q_n es exacta para polinomios de grado menos o igual que n, entonces se tiene que $Q_n(q_N) = I(q_N)$ para todo $n > N$. Tenemos que,

$$\begin{aligned} |I(f) - Q_n(f)| &= |I(f) - I(q_N) + Q_n(q_N) - Q_n(f)| \\ &\leq |I(f) - I(q_N)| + |Q_n(q_N) - Q_n(f)|. \end{aligned}$$

Ahora bien, si llamamos $c = \int_a^b w(x)\ dx$, se tiene

$$\begin{aligned} |I(f) - I(q_N)| &= \left| \int_a^b w(x)(f(x) - q_N(x))\ dx \right| \\ &\leq \|f - q_N\|_\infty \int_a^b w(x)\ dx \leq c\varepsilon, \end{aligned}$$

y además,

$$\begin{aligned} |Q_n(q_N) - Q_n(f)| &= \left| \textstyle\sum_{j=0}^n A_j^{(n)} (q_N(x_j^{(n)}) - f(x_j^{(n)})) \right| \\ &\leq \|f - q_N\|_\infty \textstyle\sum_{j=0}^n |A_j^{(n)}| \leq K\varepsilon. \end{aligned}$$

Con todo esto, hemos probado que dado $\varepsilon > 0$ existe N tal que si $n > N$,

$$|I(f) - Q_n(f)| \leq (c + K)\varepsilon.$$

□

También vale la implicación recíproca, que enunciamos en el siguiente teorema, de la cual omitimos una demostración.

TEOREMA 7.19. *Con las notaciones anteriores, si* $\lim_{n\to\infty} Q_n(f) = I(f)$ *entonces existe una constante* K *tal que*

$$\sum_{j=0}^{n} |A_j^{(n)}| \leq K \qquad \forall n.$$

COROLARIO 7.20. *Si los pesos* $A_j^{(n)}$ *son todos positivos tenemos entonces*

$$Q_n(f) \to \int_a^b f(x)w(x)\ dx,$$

para toda función continua f.

Demostración. Como la aproximación por cuadraturas Q_n es exacta para polinomios de $\mathcal{P}_n$, en particular se tiene que $Q_n(1) = \int_a^b w(x)\ dx$, para todo $n \in \mathbb{N}$.

Como $w(x) > 0$ y los pesos son todos positivos tenemos que,

$$0 < I(1) = \int_a^b w(x)\ dx = Q_n(1) = \sum_{j=0}^{n} A_j^{(n)} = \sum_{j=0}^{n} |A_j^{(n)}|.$$

La constante K que satisface la hipótesis del teorema anterior es

$$K = \int_a^b w(x)\ dx = \sum_{j=0}^{n} |A_j^{(n)}|,$$

y tenemos que estas aproximaciones de la integral convergen a la misma. □

Se sabe que si $n \geq 10$ los pesos de las fórmulas de Newton-Côtes cambian de signo. Peor aún, cuando n crece, los pesos crecen y no están acotados, por lo tanto existen funciones continuas para las cuales este procedimiento para aproximar la integral no converge. Así, el camino seguro para aumentar la precisión usando fórmulas de Newton-Côtes es por medio de fórmulas compuestas. Sin embargo es posible aumentar precisión aumentando el número de los nodos de interpolación. Estudiaremos este método en la sección siguiente.

5. Cuadratura Gaussiana

Queremos aumentar la presición al aproximar $\int_a^b f(x)\ dx$, o $\int_a^b f(x)w(x)\ dx$ como vimos en la sección anterior. Si consideramos nodos fijos $\{x_0, x_1, \ldots, x_n\}$ e interpolamos f con un polinomio $p \in \mathcal{P}_n$ pueden producirse errores grandes en la aproximación como vimos en el Capítulo 5. Una forma de mejorar el error que se comete es elegir los puntos de interpolación para optimizar la aproximación. Los polinomios de Tchebychev dan una solución en este sentido. Los nodos a considerar son los ceros del polinomio de grado $n + 1$. El método de *cuadratura gaussiana* generaliza este tipo de elección.

En síntesis, queremos encontrar una fórmula de cuadratura

$$\int_a^b f(x)w(x)dx \quad \sim \quad Q(f) = \sum_{j=0}^{n} A_j f(x_j)$$

donde podamos elegir tanto los pesos $\{A_0, A_1, \ldots, A_n\}$ como los nodos $\{x_0, x_1, \ldots, x_n\}$, es decir que tenemos $2n+2$ variables. Si pedimos que las formulas de inetgración sean exactas para polinomios del mayor grado posible nos quedan las siguientes ecuaciones

$$\int_a^b x^k w(x)\ dx = \sum_{j=0}^{n} A_j x_j^k \qquad 0 \leq k \leq 2n+1$$

Esto es un sistema no lineal de $2n+2$ ecuaciones con $2n+2$ incógnitas. Gauss demostró que este sistema tiene solución única cuando $w = 1$ y el intervalo es $[-1, 1]$. El Lema 7.4 nos permite independizarnos del intervalo mientras que la teoría de espacios con producto interno y polinomios ortogonales vistos en el Capítulo 6 nos permiten trabajar con un peso arbitrario w, $(w(x) > 0)$.

Así como los ceros del polinomio de Tchebychev de grado n son todos distinto, para cada n fijo, y éstos son tomados como nodos para interpolar una función f, el Teorema que sigue nos da un resultado análogo para cualquier familia de polinomios ortogonales.

Consideremos sobre $V = C[a, b]$ el producto interno

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

y llamemos $\{q_j\}$ a los polinomios ortogonales y mónicos con respecto a este producto interno. Respectivamente, notemos $\{p_j\}$ los polinomios ortonormales.

Teorema 7.21. *Las raíces de p_n son todas reales, distintas entre sí y pertenecen al intervalo (a, b).*

Demostración. Sea $n \geq 1$ fijo. Veamos primero p_n tiene al menos una raíz real. Si no fuera así, p_n tiene signo constante, en particular, tiene signo constante en el intervalo (a, b). Supongamos que $p_n(x) > 0$ en (a, b). Como p_n y q_0 $(q_0 = 1)$ son ortogonales tenemos

$$0 = \langle p_n, 1\rangle = \int_a^b p_n(x)w(x)\ dx > 0$$

esta contradicción muestra que p_n no sólamente tiene una raíz real sino que tiene al menos un cero en (a, b).

El segundo paso será ver que las raíces de p_n son simples. Supongamos que p_n tiene algún cero múltiple y veamos que esto no es posible. Si $x_0 \in \mathbb{R}$ es una raíz muúltiple, entonces p_n es divisible por $(x - x_0)^2$, y entonces $q(x) = \dfrac{p_n(x)}{(x - x_0)^2}$ es un polinomio de grado $n-2$. Luego, q es una combinación lineal de $\{p_0, p_1, \ldots, p_{n-2}\}$ y resulta ser ortogonal a p_n, por consiguiente,

$$\begin{aligned} 0 = \langle p_n, q\rangle &= \int_a^b p_n(x)\frac{p_n(x)}{(x-x_0)^2}w(x)\ dx \\ &= \int_a^b \frac{p_n(x)^2}{(x-x_0)^2}w(x)\ dx > 0, \end{aligned}$$

que resulta, nuevamente, es una contradicción. En consecuencia todos lo ceros de p_n son simples.

Finalmente, resta ver que todas las raíces de p_n pertenecen al intervalo (a, b).

Supongamos que $x_0, \ldots, x_k$ son los ceros de p_n que están en (a, b) y supongamos que $k < n-1$, es decir que p_n tiene ceros que no pertenecen a (a, b). Como las raíces $x_0, \ldots, x_k$ son simples el polinomio r dado por

$$r(x) = p_n(x)/(x-x_0)(x-x_1)\ldots(x-x_k).$$

tiene grado $n-(k+1)$, con lo cual es ortogonal a p_n y tiene signo constante en (a, b). Supongamos que $r(x) > 0$.

$$\begin{aligned} 0 = \langle p_n, (x-x_0)\ldots(x-x_k)\rangle &= \int_a^b p_n(x)(x-x_0)\ldots(x-x_k)w(x)\ dx \\ &= \int_a^b r(x)(x-x_0)^2\ldots(x-x_k)^2(x)w(x)\ dx > 0, \end{aligned}$$

Esta contradicción proviene de suponer que el grado de r es no nulo, luego $k = n-1$ y todos lo ceros de p_n están en (a, b). □

Ahora probemos el teorema básico de las cuadraturas de Gauss.

Teorema 7.22. *La fórmula*

$$\int_a^b p(x)w(x)\ dx = \sum_{j=0}^{n} A_j p(x_j),$$

vale para cualquier polinomio de grado menor o igual a $2n+1$ *si y solo si los puntos* $\{x_j\}$ *son los ceros de* $p_{n+1}(x)$.

Demostración. Sea $p(x) \in P_{2n+1}$ y supongamos que los puntos x_j están dados por

$$p_{n+1}(x_j) = 0 \qquad 0 \le j \le n.$$

Por el algoritmo de división para polinomios se puede escribir

$$p(x) = p_{n+1}(x)S(x) + R(x)$$

con $S(x)$ y $R(x)$ en $P_n(x)$.

Por la definición de los pesos A_j, como el grado de R es a lo sumo n, tenemos

$$I(R) = Q_n(R).$$

Entonces

$$\begin{aligned} I(p) = & \int_a^b p(x)w(x)\ dx \\ = & \int_a^b p_{n+1}(x)S(x)w(x)\ dx + \int_a^b R(x)w(x)\ dx \\ = & \langle p_{n+1}, S\rangle + I(R) \\ = & 0 + Q_n(R) \\ = & \sum_{j=0}^{n} A_j R(x_j) \\ = & \sum_{j=0}^{n} A_j p(x_j) = Q_n(p). \end{aligned}$$

Ahora supongamos que $\{x_j\}$ es un conjunto de puntos distintos y que se verifica

$$\int_a^b p(x)w(x)\ dx = \sum_{j=0}^{N} A_j p(x_j)$$

para todo $p \in P_{2n+1}$. Dado un entero k sea r_k un polinomio de grado menor o igual que k. Sea $W(x) = \prod_{j=0}^{n}(x - x_j)$ y definamos $p(x) = r_k(x)W(x)$. Es claro que $p \in P_{2n+1}$, por nuestra hipótesis $I(p) = Q_n(p)$, en consecuencia

$$\begin{aligned} \langle r_k, W\rangle = & \int_a^b r_k(x)W(x)w(x)\ dx \\ = & \int_a^b p(x)w(x)\ dx \\ = & \sum_{j=0}^{n} A_j p(x_j) \\ = & \sum_{j=0}^{n} A_j r_k(x_j)W(x_j) = 0, \end{aligned}$$

pues $W(x)$ se anula en los x_j. Entonces $W(x)$ es un polinomio mónico de grado $(n+1)$ que resulta ortogonal a cualquier polinomio de grado menor o igual que n, en consecuencia $W(x) = q_{n+1}(x)$

y entonces los x_j son los ceros de p_{n+1}. □

COROLARIO 7.23. *Sea* $Q_n(f) = \sum_{j=0}^{n} A_j f(x_j)$ *la cuadratura gaussiana, entonces*

$$\lim_{n\to\infty} Q_n(f) = I(f).$$

Demostración. Por la definición de los $l_k(x)$ cada $(l_k)^2 \in P_{2n}$ y entonces $I(l_k^2) = Q_n(l_k^2)$ y en consecuencia, como $l_k(x_j) = \delta_{kj}$,

$$0 < \int_a^b l_k^2(x)w(x)\ dx = \sum_{j=0}^{n} A_j(l_k(x_j))^2 = A_k.$$

Es decir, para la cuadratura gaussiana todos los pesos son positivos y por lo visto antes esto implica la convergencia. □

6. Ejercicios

(1) Usar las fórmulas cerradas de Newton-Cotes de dos y tres puntos (reglas de trapecios y de Simpson, respectivamente) para calcular las integrales:

$$\int_0^1 x^4\ dx \qquad \int_{0.1}^{0.2} \ln(x)\ dx \qquad \int_0^{.3} \frac{1}{1+x}\ dx$$

Calcular, además, en forma exacta cada una de las integrales anteriores y verificar la cota del error.

(2) Interpolando las funciones de base de Lagrange, hallar una fórmula de cuadratura por interpolación de la forma

$$\int_0^{2h} f(x)\ dx \sim A_0 f(0) + A_1 f(h).$$

(3) Usar el método de coeficientes indeterminados para dar una fórmula de cuadratura por interpolación:

$$\int_0^{3h} f(x)\ dx \sim A_0 f(0) + A_1 f(h) + A_2 f(3h).$$

(4) Construir la fórmula abierta de Newton-Cotes para calcular $\int_{-1}^{1} f(x)\ dx$ con nodos $-1/2,\ 0,\ 1/2$, y la fórmula cerrada de Newton-Cotes con nodos en los puntos $-1,\ -1/3,\ 1/3,\ 1$.

(5) Considerar la función definida en $[-h, h]$ $(h > 0)$:

$$f(x) = \begin{cases} 0, & \text{si } -h \le x \le 0 \\ x, & \text{si } 0 < x \le h. \end{cases}$$

Hallar el error de la regla de trapecios aplicada a $f(x)$. ¿El orden es igual al obtenido para una función suficientemente suave?

(6) La fórmula de cuadratura

$$\int_a^b f(x)\ dx \sim f(\frac{a+b}{2})\big(b-a\big)$$

es conocida como *Regla de los Rectángulos*. Para $f \in C^1[a,b]$ acotar el error que se comete al utilizarla.

(7) Para f una función C^2 probar que el error cometido al usar la fórmula de cuadratura del Ejercicio 2 no excede el valor $\dfrac{\|f''\|_\infty}{2}h^3$.

(8) (a) Hallar una fórmula de cuadratura del tipo:

$$\int_{-1}^1 f(x)\ dx \sim Af(-2) + Bf(0) + Cf(2).$$

(b) Para $f \in C^3[-2,2]$ probar que el error cometido no excede el valor $\frac{7}{12}\|f^{(3)}\|_\infty$.

(9) Escribir un programa que utilice las reglas de trapecios, de Simpson, de trapecios compuesta y de Simpson compuesta para calcular aproximaciones de la integral de una función $f(x)$ en un intervalo $[a,b]$.

(10) Se sabe que $\displaystyle\int_0^1 \frac{1}{1+x^2}\ dx = \frac{\pi}{4}$.

(a) Para $n = 1, \ldots, 100$, utilizar las reglas de trapecios y Simpson compuestas para aproximar numéricamente la integral y dar un valor cercano a π.

(b) Graficar las sucesiones obtenidas junto con el valor de π que arroja `Matlab` y el valor que se obtiene al aplicar la rutina **quad** de `Matlab`.

(11) (a) Calcular exactamente la integral

$$I = \int_0^{2\pi} [1 - \cos(32x)]\ dx.$$

(b) Aproximar el valor de I usando el programa del Ejercicio 9 con los métodos de los trapecios, Simpson, trapecios compuesta y Simpson compuesta para $n = 2, 4, 8$ y 16.

(c) Calcular el valor de I que produce la rutina **quad**.

(12) Se quiere calcular $\displaystyle\int_{-1}^1 e^{-x^2}\ dx$ utilizando la regla de trapecios compuesta, partiendo el intervalo $[-1,1]$ en n subintervalos. Hallar n de modo que el error sea menor que 10^{-3}.

(13) La expresión $Q_n(f) = \sum_{j=0}^n A_j f(x_j)$ define una fórmula de cuadratura.

(a) Probar que Q_n es lineal en f (el conjunto de funciones).

(b) Supongamos que $Q_n(f) \sim \int_a^b f(x)w(x)\ dx$ y que es exacta para las funciones $1,\ x, \ldots,\ x^k$. Mostrar que la fórmula tiene grado de precisión por lo menos k.

(14) Determinar el grado de precisión de las fórmulas para $\int_{-1}^1 f(x)\ dx$:

(a) $\frac{4}{3}f(-0.5) - \frac{2}{3}f(0) + \frac{4}{3}f(0.5)$.

(b) $\frac{1}{4}f(-1) + \frac{3}{4}f(-\frac{1}{3}) + \frac{3}{4}f(\frac{1}{3}) + \frac{1}{4}f(1)$.

(15) Hallar reglas de cuadratura de grado de precisión máximo para aproximar $\int_{-3}^3 f(x)\ dx$, de las siguientes formas:

(a) $A[f(x_0) + f(x_1)]$ (repitiendo el coeficiente).

(b) $Af(x_0) + Bf(x_0 + 4)$.

y determinar cuáles son dichos grados.

(16) Calcular $\int_{-1}^{1} f(x)x^2\ dx$ mediante una regla de cuadratura de la forma

$$\int_{-1}^{1} f(x)x^2\ dx \sim A_0 f(x_0) + A_1 f(x_1)$$

que sea exacta para polinomios de grado menor o igual que 3.

(17) (a) Hallar una regla de cuadratura del siguiente tipo

$$\int_{-1}^{1} f(x)\sqrt{|x|}dx \sim A_0 f(x_0) + A_1 f(x_1).$$

que tenga grado de precisión máximo. ¿Cuál es dicho grado?

(b) Hallar una regla de cuadratura del siguiente tipo

$$\int_{0}^{4} f(x)\sqrt{\Big|\frac{x-2}{2}\Big|}dx \sim A_0 f(x_0) + A_1 f(x_1).$$

que tenga grado de precisión máximo. ¿Cuál es dicho grado?

(18) Sea w una función de peso. Se considera la regla de cuadratura de 1 punto:

$$\int_{a}^{b} f(x)w(x)\ dx \sim A_0 f(s).$$

(a) Probar que, cualquiera sea w, la fórmula tiene grado de precisión máximo si $s = \dfrac{\int_a^b xw(x)\ dx}{\int_a^b w(x)\ dx}$.

(b) Probar que si $w(x) \equiv 1$, esta regla coincide con la regla de los rectángulos.

(c) Considerar el intervalo $[-1, 1]$ y $w(x) = (x-1)^2$. Acotar el error que produce el uso de esta regla para funciones C^1.

(19) Hallar los pesos y los nodos de las fórmulas de Gauss-Legendre de dos y tres puntos. (Los polinomios de Legendre mónicos de grado dos y tres son $x^2 - \frac{1}{3}$ y $x^3 - \frac{3}{5}x$).

(20) Usar las fórmulas de Gauss-Legendre de tres puntos para estimar:

$$\text{(a)}\quad \int_{-1}^{1} \sin(3x)\ dx, \qquad \text{(b)}\quad \int_{1}^{3} \ln(x)\ dx, \qquad \text{(c)}\quad \int_{1}^{2} e^{x^2}\ dx.$$

(21) Probar que una fórmula de cuadratura

$$\int_{a}^{b} f(x)w(x)\ dx \sim Q_n(f) = \sum_{j=0}^{n} A_j f(x_j)$$

no puede tener grado de precisión mayor que $2n+1$, independientemente de la elección de los coeficientes (A_j) y de los nodos (x_j).

Sugerencia: Hallar un polinomio $p \in \mathcal{P}_{2n+2}$ para el cual $Q_n(p) \neq \int_a^b p(x)w(x)\ dx$.

(22) Para $f : \mathbb{R}^2 \to \mathbb{R}$ una función continua, se quiere dar una fórmula de cuadratura que aproxime $\iint_D f(x,y)\ dx\ dy$ con $D \subset \mathbb{R}^2$ usando el Teorema de Fubini.

(a) Repetir el procedimiento hecho en el Ejemplo 7.17 y dar la fórmula correspondiente para D el triángulo de vértices $(0,0), (0,1), (1,0)$.
Sugerencia: considerar $F(x) = \int_0^x f(x,y)\ dy$.

(b) Probar que si D es el triángulo de vértices $(0,0), (0,1), (1,0)$ la fórmula anterior es exacta para $f(x,y) = x^2 + y^2$.

CAPíTULO 8

Resolución de ecuaciones diferenciales ordinarias.

En este capítulo abordaremos el problema de resolver ecuaciones diferenciales con valores iniciales. Es decir, desarrollaremos métodos numéricos para aproximar una función conociendo una ecuación que involucra sus derivadas.

Se llama orden de una ecuación al máximo orden de derivada que aparece en ella. En su forma más general una ecuación diferencial de orden n, puede escribirse como

$$F(t, x(t), x'(t), x''(t), \ldots, x^{(n)}(t)) = 0,$$

donde $t \in \mathbb{R}$, $F : \mathbb{R}^{n+2} \to \mathbb{R}$ es una función conocida y x es la función que se desea encontrar.

Vamos a suponer que la derivada de mayor orden puede despejarse de tal forma que la ecuación se escribe como

$$x^{(n)}(t) = f(t, x(t), x'(t), x''(t), \ldots, x^{(n-1)}(t)), \tag{8.1}$$

para $t \in \mathbb{R}$ y $f : \mathbb{R}^{n+1} \to \mathbb{R}$ una función dada.

Algunos ejemplos de ecuaciones diferenciales son:

Ejemplos 8.1.

(1) Para λ una constante dada, la ecuación

$$x'(t) = \lambda x(t).$$

es una ecuación lineal de primer orden. Su solución general es

$$x(t) = Ce^{\lambda t}$$

con C una constante arbitraria, es decir, hay infinitas soluciones. Esto es lo que pasa en general y por lo tanto, para poder determinar una solución es necesario tener más datos. En este ejemplo se ve fácilmente que si se conoce el valor inicial $x(0) = x_0$ entonces la solución es

$$x(t) = x_0 e^{\lambda t}.$$

Esto es algo general: dada una ecuación diferencial para determinar una solución es necesario conocer ciertos datos iniciales.

(2) Veamos un ejemplo elemental de ecuación que surge de un problema físico. Supongamos que se tiene una partícula de masa m que se mueve en una dirección debido a la acción de un resorte y se quiere conocer la posición $x(t)$ de la masa en el instante t. La ley de

Hooke dice que la fuerza $F(t)$ que ejerce el resorte en el instante t es proporcional a su estiramiento o compresión, es decir,

$$F(t) = -kx(t)$$

donde k es la constante de rigidez del resorte. Por otra parte, la ley de Newton nos dice que

$$F(t) = ma(t)$$

siendo $a(t)$ la aceleración en el instante t. En consecuencia, como $a(t) = x''(t)$, obtenemos la ecuación

$$mx''(t) + kx(t) = 0.$$

Esta es una ecuación lineal de segundo orden que, como tiene coeficientes constantes, puede resolverse analíticamente. Si llamamos $\omega = \sqrt{k/m}$, la solución general de esta ecuación es

$$x(t) = C_1 \cos(\omega t) + C_2 \operatorname{sen}(\omega t)$$

donde C_1 y C_2 son constantes arbitrarias. Introduciendo $A = \sqrt{C_1^2 + C_2^2}$ y $\varphi \in [0, 2\pi)$ tal que $\cos\varphi = C_1/A$ y $\operatorname{sen}\varphi = C_2/A$ (notar que tal φ existe porque $(C_1/A)^2 + (C_2/A)^2 = 1$), la solución general puede escribirse como

$$A\cos(\varphi - \omega t)$$

donde A representa la amplitud, ω la frecuencia y φ la fase.

Para poder determinar la posición en el instante t necesitamos conocer ciertas condiciones iniciales. Lo más natural es conocer la posición y la velocidad iniciales, es decir $x(0) = x_0$ y $x'(0) = v_0$. Veamos que con estos datos podemos encontrar A y φ de tal forma que la solución queda unívocamente determinada. En efecto, es fácil ver que de las condiciones iniciales se deduce que $A = \sqrt{x_0^2 + (v_0/\omega)^2}$ y $\varphi \in [0, 2\pi)$ es el único ángulo que satisface $\cos\varphi = x_0/A$ y $\operatorname{sen}\varphi = v_0/\omega A$.

(3) Veamos ahora un ejemplo de ecuación no lineal. Para esto consideremos otro ejemplo de los cursos básicos de física que es el problema del péndulo. En este caso se quiere determinar el ángulo $\theta(t)$ que un péndulo con masa m forma respecto de la vertical en el instante t. Despreciando el rozamiento con el aire podemos suponer que la única fuerza que actúa es la de la gravedad, es decir, una fuerza en dirección vertical hacia abajo y de magnitud mg. La proyección F de esta fuerza en la dirección del movimiento (o sea tangencial al arco de circunferencia que describe el péndulo) resulta entonces,

$$F(t) = mg \operatorname{sen}\theta(t).$$

Teniendo en cuenta que la longitud recorrida en un tiempo t es $L(\theta(t) - \theta(0))$, donde L es la longitud del péndulo, resulta que la aceleración en la dirección tangencial al movimiento es $L\theta''(t)$. Por lo tanto, aplicando nuevamente la ley de Newton obtenemos

$$L\theta''(t) = g \operatorname{sen}\theta(t)$$

o sea, una ecuación no lineal de segundo orden. También en este caso hay una única solución si se conocen la posición y la velocidad inicial, o sea, $\theta(0)$ y $L\theta'(0)$. Esto es consecuencia del teorema de existencia y unicidad que enunciaremos más adelante.

Como vimos en los ejemplos una ecuación diferencial puede tener muchas soluciones y para obtener una solución única hace falta conocer ciertos datos que pueden ser valores de la función y de algunas de sus derivadas en un valor inicial t_0. Más precisamente, puede demostrarse que para la ecuación de orden n (8.1) se tiene una solución única dados los datos iniciales $x(t_0), x'(t_0), \ldots, x^{(n-1)}(t_0)$, bajo hipótesis adecuadas sobre la función f.

En lo que sigue, vamos a estudiar métodos numéricos para ecuaciones de grado 1. La razón por la que hacemos esto es que las ecuaciones de grado n pueden reducirse a sistemas de ecuaciones de orden 1 y los métodos que presentaremos pueden extenderse a sistemas.

Para simplificar la notación usaremos $t_0 = 0$. La ecuación de orden 1 que corresponden con la escritura 8.1 tienen la forma

$$\begin{cases} x'(t) = f(t, x(t)), \\ x(0) = a. \end{cases} \tag{8.2}$$

Se trata de una ecuación diferencial de primer orden porque la derivada de mayor orden que aparece es la derivada primera.

Por ejemplo, podemos considerar ecuaciones como las siguientes:

$$\begin{array}{ll} \text{(i)} \quad x' = 1, & \text{(iii)} \quad x' = x, \\ \text{(ii)} \quad x' = t, & \text{(iv)} \quad x' = x^2. \end{array}$$

Primero enunciamos un teorema de existencia y unicidad de solución, cuya demostración no damos.

Teorema 8.2. *Si $f(t, x)$ es continua y Lipschitz en x, es decir*

$$|f(t, x) - f(t, y)| \leq L|x - y|$$

Entonces para cualquier valor $a \in \mathbb{R}$ existe una única función derivable $x(t)$ que verifica

$$\begin{cases} x'(t) = f(t, x(t)) \\ x(0) = a. \end{cases}$$

Nuestro estudio de aproximaciones numéricas para ecuaciones ordinarias empieza con los métodos conocidos como métodos de un solo paso.

Dado el problema (8.2) buscamos una aproximación de $x(t)$ en un cierto intervalo $[0, T]$. Para esto buscaremos una forma de generar N valores $x_1, \ldots, x_N$ que aproximen $x(t_1), \ldots, x(t_N)$ con $t_1 < t_2 < \ldots < t_N$. Después se puede interpolar en esos valores para obtener una aproximación de $x(t)$. A veces solo interesa conocer el valor de $x(T)$ y en este caso los pasos intermedios $x_1, \ldots, x_{N-1}$ pueden verse como pasos auxiliares para calcular $x_N \sim x(T)$.

El método general a un paso tiene la forma

$$x_{i+1} = x_i + h\Phi(x_i, t_i, h).$$

La función Φ se conoce como la función de incremento y nos dice como calcular la aproximación x_{i+1} de $x(t_i+h)$ a partir de x_i, t_i y de h. Una ventaja de estos métodos es que se puede cambiar fácilmente el paso h. Es decir calculamos una aproximación en un tiempo posterior (t_{i+1}) a partir de una aproximación en el tiempo t_i.

1. Métodos de Euler y Taylor de orden k

El método de Euler se basa en el desarrollo de Taylor de orden 1. En general, el desarrollo de orden 1 es: $x(t+h) = x(t) + x'(t)h + x''(\xi)\frac{h^2}{2}$. Supongamos que podemos despreciar el valor de $x''(\xi)\frac{h^2}{2}$; el primer valor aproximado de x en t_0+h se logra al calcular $x(t_0+h) \sim x(t_0)+x'(t_0)h$; esto es, partimos del valor $x(t_0)$ y nos movemos una longitud h con velocidad $x'(t_0)$. Notar que $x'(t)$ es una función conocida, es decir, para cada valor de t sabemos exactamente cuánto vale $x'(t) = f(t,x)$.

Repitiendo este procedimiento se obtiene:

$$x(t_i + h) \sim x(t_i) + hx'(t_i) = x(t_i) + hf(t_i, x(t_i)),$$

para valores pequeños de h. En consecuencia, si x_i es una aproximación para $x(t_i)$ se tiene que

$$x_{i+1} = x_i + hf(t_i, x_i)$$

es una aproximación razonable para $x(t_i + h) = x(t_{i+1})$.

Ejemplo 8.3. *Resolvamos la ecuación*

$$\begin{cases} x'(t) = x(t) \\ x(0) = 1. \end{cases}$$

La solución exacta es $x(t) = e^t$. Ahora, aplicando el método de Euler se obtiene la siguiente tabla,

t	e^t	$h = 0.25$	$h = 0.125$
0.125	1.1331		1.125
0.250	1.2840	1.25	1.2656
0.375	1.4549		1.4238
0.500	1.6487	1.5625	1.6018
0.625	1.8682		1.8020
0.750	2.1170	1.9531	2.0272

El método de Euler es de la forma

$$x_{i+1} = x_i + hf(t_i, x_i).$$

que como vimos responde a usar el polinomio de Taylor de grado uno en t_i para calcular x_{i+1}. En general se puede usar una expansión en más términos como sigue:

$$x(t_i + h) \sim x(t_i) + hx'(t_i) + \frac{1}{2}h^2x''(t_i) + ... + \frac{1}{k!}h^k x^{(k)}(t_i).$$

Si conociéramos las derivadas de x hasta orden k podríamos usar esto para implementar un método que aproxime $x(t_i+h)$ a partir de una aproximación de $x(t_i)$. La ecuación diferencial $x' = f(t,x)$ nos proporciona todas las derivadas de orden superior de la siguiente forma, derivando una vez se obtiene

$$x'' = f_x(t,x)x' + f_t(t,x),$$

es decir,

$$x'' = f_x(t,x)f(t,x) + f_t(t,x).$$

A partir de esto se puede poner

$$x(t_i+h) \sim x(t_i) + hf(t_i,x(t_i)) + \frac{1}{2}h^2 f_x(t_i,x(t_i))f(t_i,x(t_i)) + f_t(t_i,x(t_i))$$

e intentar la aproximación

$$x_{i+1} = x_i + hf(t_i,x_i) + \frac{1}{2}h^2 f_x(t_i,x_i)f(t_i,x_i) + f_t(t_i,x_i).$$

Podemos seguir derivando

$$x'' = f_x(t,x)f(t,x) + f_t(t,x)$$

para obtener x''' en términos de f y sus derivadas y continuando de esta manera calcular las derivadas de x hasta orden k en términos de f y sus derivadas. Esto nos da una manera de calcular métodos de aproximación a un paso (solo usamos x_i, t_i y h para calcular la siguiente aproximación x_{i+1}). Es decir, a partir de

$$x(t_i+h) \sim x(t_i) + hx'(t_i) + \frac{1}{2}h^2x''(t_i) + \ldots + \frac{1}{k!}h^k x^{(k)}(t_i),$$

proponemos el método,

$$x_{i+1} \sim x_i + hf(t_i,x_i) + \ldots + \frac{1}{k!}h^k x^{(k)}(t_i) = x_i + hT_k(t_i,x_i,h).$$

Estos métodos se conocen como métodos basados en Taylor de orden k. Su mayor problema es que requieren encontrar y evaluar derivadas sucesivas de $f(x,t)$.

EJEMPLO 8.4. *Calculemos el método de Taylor de orden 2 para el problema*

$$\begin{cases} x'(t) = 2tx \\ x(0) = 1. \end{cases}$$

En este caso, al derivar, se obtiene

$$x'' = 2x + 2tx' = 2x(1+2t^2).$$

Entonces,

$$x_{i+1} = x_i + h\left[2t_i x_i + x_i(1+2t_i^2)h\right].$$

EJEMPLO 8.5. **Ejercicio:** *Calcular el método de Taylor de orden 3 en este ejemplo.*

2. Métodos de Runge-Kutta

Los métodos generales a un paso son de la forma

$$x_{i+1} = x_i + h\Phi(t_i, x_i, h)$$

Podemos pensar que $h\Phi(t_i, x_i, h)$ es una aproximación de la variación de x cuando pasamos de t_i a t_{i+1}. Esta variación, para la ecuación original esta gobernada por $f(x,t)$, así que proponemos una forma particular de Φ dada por,

$$h\Phi(x_i, t_i, h) = A_1 f(\theta_1, \gamma_1) + \ldots + A_1 f(\theta_N, \gamma_N),$$

donde (θ_i, γ_i) son puntos próximos a (t_i, x_i). Para especificar el método que usamos es necesario especificar los A_i y los puntos (θ_i, γ_i).

Veamos un caso particular de lo anterior donde usamos solo dos puntos uno (t_i, x_i) y el otro $(t_i + \alpha h, \alpha h f(x_i, t_i))$. Todavía nos queda α libre (por ejemplo al considerar $\alpha = 1$ usamos $(t_{i+1}, \tilde{x}_{i+1})$ donde $\tilde{x}_{i+1}$ es la aproximación que da el método de Euler).

En general tenemos,

$$x_{i+1} = x_i + h\left[A_1 f(t_i, x_i) + A_2 f(t_i + \alpha h, x_i + \alpha h f(t_i, x_i))\right].$$

La estrategia del método de Runge-Kutta es elegir A_1, A_2 y α para que esto se aproxime todo lo posible a un método de Taylor.

Veamos como hacer esto. Primero expandimos

$$f(t_i + \alpha h, x_i) + \alpha h f(t_i, x_i) = f(t_i, x_i) + f_t(t_i, x_i)\alpha h + f_x(t_i, x_i)\alpha h f(t_i, x_i) + E,$$

donde E tiene la forma $E = Ch^2$. Agrupando términos se obtiene

$$\Phi(t_i, x_i, h) = (A_1 + A_2) f(t_i, x_i) + A_2 h\left[f_t(t_i, x_i)\alpha + f_x(t_i, x_i)\alpha f(t_i, x_i) + Ch\right].$$

Recordemos que en el método de Taylor se tiene

$$T_2(t_i, x_i, h) = f(t_i, x_i) + \frac{h}{2}\left[f_t(t_i, x_i) + f_x(t_i, x_i) f(t_i, x_i)\right]$$

Igualando los coeficientes, llegamos a

$$A_1 + A_2 = 1,$$
$$A_2\alpha = \frac{1}{2}.$$

Despejando, obtenemos

$$A_2 = \frac{1}{2\alpha},$$
$$A_1 = 1 - \frac{1}{2\alpha}.$$

Es decir, hay infinitas soluciones dependiendo de α (y por ende, infinitos métodos posibles).

Una elección posible es $\alpha = 1/2$ con lo que nos queda el método $(A_1 = 0, A_2 = 1)$

$$x_{i+1} = x_i + \frac{h}{2}\left[f(t_i + \frac{h}{2}, x_i + \frac{h}{2}f(t_i, x_i))\right] \tag{8.3}$$

que usualmente se conoce como *Método de Euler modificado.*

Otra elección posible es $\alpha = 1$ que nos da $(A_1 = A_2 = \frac{1}{2})$

$$x_{i+1} = x_i + \frac{h}{2}\left[f(t_i, x_i) + f(t_{i+1}, x_i + hf(t_i, x_i))\right], \tag{8.4}$$

que se conoce como *Método de Heun.*

Una observación importante es que estos métodos no requieren la evaluación de derivadas de f.

Considerando más términos en el desarrollo de Taylor se pueden deducir métodos de Runge-Kutta de mayor orden. Específicamente, un método de Runge-Kutta de orden k tiene la forma

$$x_{i+1} = x_i + h\Phi(t_i, x_i, h)$$

donde

$$\Phi(t_i, x_i, h) = T_k(t_i, x_i, h) + O(h^k).$$

También

$$\Phi(t_i, x_i, h) = \sum_{j=1}^{m} A_j K_j(t_i, x_i, h).$$

Los términos K_j están dados por

$$K_1(t_i, x_i, h) = f(t_i, x_i),$$

$$K_j(t_i, x_i, h) = f(t_i + \alpha_j h, y_i + h\sum_{r=1}^{j-1} \beta_{jr} K_r(t_i, x_i, h)),$$

donde $0 < \alpha_j \leq 1$ y $\alpha_j = \sum_{r=1}^{j-1} \beta_{jr}$.

Ejemplo 8.6. *Una forma de Runge-Kutta de orden cuatro es,*

$$x_{i+1} = x_i + \frac{h}{6}\left[K_1 + 2K_2 + 2K_3 + K_4\right],$$

$$\begin{array}{ll} K_1 = f(t_i, x_i), & K_3 = f(t_i + \frac{h}{2}, x_i + \frac{h}{2}K_2), \\ K_2 = f(t_i + \frac{h}{2}, x_i + \frac{h}{2}K_1), & K_4 = f(t_i + h, x_i + hK_3). \end{array}$$

3. Análisis de los Errores

Primero definimos el error de truncación local. Si $x(t)$ es la solución de la ecuación diferencial $x'(t) = f(x(t), t)$ y consideramos t^* y h fijos, se define el error de truncación local, τ por medio de la expresión que da:

$$x(t^* + h) = x(t^*) + h\Phi(t^*, x(t^*), h) + h\tau \tag{8.5}$$

3.1. Métodos de Euler y Taylor de orden k. En este caso el error de truncación local está dado por

$$x(t^* + h) = x(t^*) + hf(t^*, x(t^*)) + h\tau$$

Si la solución $x(t)$ es suave (por ejemplo x'' es acotada) se tiene que

$$x(t^* + h) = x(t^*) + hx'(t^*) + \frac{h^2}{2}x''(\gamma) \quad \text{para algún } \gamma.$$

De aquí se tiene que

$$\tau = \frac{h}{2}x''(\gamma). \tag{8.6}$$

El mismo argumento usado anteriormente nos da

$$\tau = \frac{h^k}{(k+1)!}x^{(k+1)}(\gamma) \tag{8.7}$$

Definición 8.7. *Diremos que el método $x_{i+1} = x_i + h\phi(t_i, x_i, h)$ es de orden k si el error de truncación local satisface*

$$\tau = O(h^k)$$

Definición 8.8. *Si $u(t)$ es una solución de la ecuación $u'(t) = f(t, u(t))$ con valor inicial x_i, es decir, u es solución de:*

$$\begin{cases} u'(t) & = f(t, u(t)) \\ u(t_i) & = x_i \qquad \forall i = 1, \ldots, n \end{cases}$$

(1) *El error local se define como: $u(t_{i+1}) - x_{i+1}$.*
(2) *El error global es: $x(t_{i+1}) - x_{i+1}$.*

donde x_{i+1} está dado por el método $x_{i+1} = x_i + h\phi(t_i, x_i, h)$.

OBSERVACIÓN 8.9. *El error global y el error local se relacionan por la fórmula*

$$x(t_{i+1}) - x_{i+1} = x(t_{i+1}) - u(t_{i+1}) + u(t_{i+1}) - x_{i+1}.$$

Esto muestra que el error global esta formado por dos "componentes" uno dado por la ecuación diferencial (el primero) y otro dado por el método (el segundo).

El error local puede ser expresado en términos del error de truncación local,

$$u(t_{i+1}) = u(t_i) + h\Phi(t_i, x(t_i), h) + h\tau.$$

Como $u(t_i) = y_i$ se sigue,

$$u(t_{i+1}) = y_i + h\Phi(t_i, x(t_i), h) + h\tau$$

y entonces

$$u(t_{i+1}) - y_{i+1} = h\tau.$$

Si el método es de orden p, entonces el error local satisface

$$u(t_{i+1}) - y_{i+1} = O(h^p).$$

3.2. Convergencia y análisis del error. Ahora nos restringiremos a problemas regulares, es decir

$$\begin{cases} x'(t) = f(t, x(t)), \\ x(0) = x_0. \end{cases}$$

donde la solución existe, es única y es regular en $[0, t_0]$.

Diremos que el método es convergente si dado t^* en $[0, t_0]$ se tiene

$$\lim_{n\to\infty,\ t^*=nh} x_n = x(t^*)$$

TEOREMA 8.10. *Para un método a un paso dado por*

$$x_{i+1} = x_i + h\Phi(t_i, x_i, h)$$

Si Φ es Lipschitz en la variable segunda variable, con constante K entonces, para e_j al error global, es decir, $e_j = x(t_j) - x_j$ se tiene:

$$|e_j| \leq \frac{\tau}{K}(e^{K(t_j)} - 1).$$

Demostración. Por un lado tenemos la propiedad de Lipschitz de ϕ en la segunda variable, esto es:

$$|\Phi(t, x, h) - \Phi(t, y, h)| \leq K|x - y|,$$

con K una constante independiente de t y h. Fijamos un paso $h = (b-a)/n$ y usamos el método y el desarrollo de Taylor,

$$\begin{aligned} x_{i+1} = &\quad x_i &+ &\quad h\Phi(t_i, x_i, h) \\ x(t_{i+1}) = &\quad x(t_i) &+ &\quad h\Phi(t_i, x(t_i), h) + h\tau_i \end{aligned}$$

Donde $h\tau_i$ es el error de truncación local en t_i dado en la fórmula (8.5). Restando miembro a miembro y usando la Definición 8.8 se obtiene

$$e_{i+1} = e_i + h(\Phi(t_i, x(t_i), h) - \Phi(t_i, x_i, h)) + h\tau_i.$$

Empleando la hipótesis de que ϕ es Lipschitz se obtiene

$$|e_{i+1}| \leq (1 + Kh)|e_i| + h\tau_i.$$

Si llamamos $\tau = \max\{\tau_i\}$ se tiene

$$|e_{i+1}| \leq (1 + Kh)|e_i| + h\tau.$$

Con la misma iteración calculada para $|e_i|$, se obtiene

$$|e_{i+1}| \leq (1 + Kh^2)|e_{i-1}| + h\tau(1 + (1 + hK)).$$

Continuando de esta manera queda

$$|e_{i+1}| \leq (1 + Kh)^{i+1}|e_0| + h\tau(\sum_{j=0}^{i}(1 + hK)^j$$

Como $e_0 = x(0) - x_0$ se sigue que $e_0 = 0$. Entonces, usando las sumas parciales de una serie geométrica,

$$|e_{i+1}| \leq h\tau \frac{(1+Kh)^{i+1}-1}{(1+Kh)-1} = \frac{\tau}{K}((1+hK)^{i+1}-1)$$

Ahora observamos que

$$(1+\alpha)^{i+1} \leq e^{(i+1)\alpha}$$

entonces

$$|e_{i+1}| \leq \frac{\tau}{K}(e^{K(t_{i+1})}-1).$$

□

Observación 8.11.
(1) El teorema anterior requiere de la existencia de K una constante de Lipschitz para Φ. De existir una constante así, cualquier otra mayor también sirve para acotar el error utilizando el teorema.
(2) En cuanto a la hipótesis sobre Φ (debe ser Lipschitz) esto puede extraerse de una condición Lipschitz sobre f.
(3) Del Teorema 8.10 se deduce que los métodos a un paso son convergentes si $\lim_{h\to 0} \tau = 0$. Esto se llama la condición de consistencia para el método. En métodos de un paso consistencia y convergencia son conceptos equivalentes.
(4) Si asumimos que ϕ es continua, se tiene que la consistencia equivale a la igualdad $x'(t_i) = \Phi(t_i, x(t_i), 0)$. Como x es solución de la ecuación diferencial se tiene $x'(t) = f(t, x(t))$, y por tanto un método de un paso es consistente si y sólo si

$$\Phi(t, x, 0) = f(t, x). \tag{8.8}$$

Observación 8.12. los métodos de Euler, los de Taylor y de Rungge-Kutta son convergentes.

Demostración. Para los métodos de Taylor se tiene

$$x(t_i + h) \sim x(t_i) + hx'(t_i) + \frac{1}{2}h^2x''(t_i) + \ldots + \frac{1}{k!}h^k x^{(k)}(t_i),$$

y por tanto

$$\Phi(t, x, h) = x'(t) + \frac{1}{2}hx''(t) + \ldots + \frac{1}{k!}h^{k-1}x^{(k)}(t).$$

Al evaluar $h = 0$ tenemos, $\Phi(t, x, 0) = x'(t) = f(t, x)$.

En cuanto a los métodos de Runge-Kutta, sólo verificaremos los dados por las fórmulas (8.3) y (8.4). El Ejemplo 8.6 queda como ejerecicio.

Para Runge-Kutta de orden 2 tenemos:

$$\begin{aligned} x_{i+1} &= x_i + h\big[f(t_i + \tfrac{h}{2}, x_i + \tfrac{h}{2}f(t_i, x_i))\big] \\ x_{i+1} &= x_i + \tfrac{h}{2}\left[f(t_i, x_i) + f(t_{i+1}, x_i + hf(t_i, x_i))\right]. \end{aligned}$$

Luego $\Phi(t,x,h) = f(t_+\frac{h}{2}, x + \frac{h}{2}f(t,x))$, y $\Phi(t,x,h) = \frac{1}{2}\big[f(t,x) + f(t,x+hf(t,x))\big]$ respectivamente. En ambos casos resulta

$$\Phi(t,x,0) = f(t,x).$$

□

Observemos que si en vez de una ecuación debemos lidiar con un sistema

$$U'(t) = F(U(t),t))$$

$$U = (u_1, \ldots, u_N)$$

podemos usar los mismos meťodos que antes, por ejemplo el método de Euler nos queda

$$U_{i+1} = U_i + hF(U_i, t_i)$$

En general, los métodos a un paso tienen la forma

$$U_{i+1} = U_i + h\phi(U_i, t_i, h)$$

4. Métodos multipaso lineales

Hasta ahora para aproximar las soluciones de $x' = f(x,t)$ nos basamos en el punto inmediato anterior para calcular el siguiente.

La filosofía de los métodos multipaso es usar la "historia", es decir los k puntos anteriores a t_i para calcular la aproximación en t_{i+1}.

Los métodos multipaso lineales (de k *pasos*) tienen la forma:

$$x_{n+k} = -\sum_{j=0}^{k-1} \alpha_j x_{n+j} + h\sum_{j=0}^{k} \beta_j f(t_{n+j}, x_{n+j})$$

Más precisamente a un método como el anterior se lo llama *método multipaso de k pasos.*

Por ejemplo, si aproximamos la derivada por

$$x'(t) \sim \frac{x(t+h) - x(t-h)}{2h}$$

considerando puntos equiespaciados nos queda

$$x'(t_i) \sim \frac{x_{i+1} - x_{i-1}}{2h}$$

y como

$$x'(t_i) = f(t_i, x(t_i)) \sim f(t_i, x_i)$$

podemos poner

$$\frac{x_{i+1} - x_{i-1}}{2h} = f(t_i, x_i)$$

es decir

$$x_{i+1} = x_{i-1} + 2hf(x_i, t_i)$$

que es un método multipaso de dos pasos (para calcular x_{i+1} se usan x_i y x_{i-1}).

Los métodos de integración también nos proporcionan ejemplos de métodos multipaso. Por ejemplo, si aproximamos la integral

$$x(t_{n+2}) - x(t_n) = \int_{t_n}^{t_{n+2}} x'(s)\ ds$$

por la regla de Simpson se obtiene

$$x(t_{n+2}) - x(t_n) \sim \frac{h}{3}(x'(t_n) + 4x'(t_{n+1}) + x'(t_{n+2}))$$

Recordando que $x' = f(x,t)$ podemos proponer el siguiente método

$$x_{n+2} - x_n = \frac{h}{3}(f(x_n, t_n) + 4f(x_{n+1}, t_{n+1}) + f(x_{n+2}, t_{n+2}))$$

que es un método multipaso a dos pasos.

Ahora usaremos la notación

$$\sum_{j=0}^{k} \alpha_j x_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}$$

para el método.

Si $\beta_k = 0$ el método es explícito, mientras que si $\beta_k \neq 0$ el método es implícito.

Si usamos una fórmula del tipo

$$\int_{t_{n+k}}^{t_{n+k-1}} x'(s)\ ds \sim h(A_0 x'(t_n) + \ldots + A_k x'(t_{n+k}))$$

para aproximar integrales nos encontramos con un método de la forma,

$$x_{n+k} - x_{n+k-1} = h(A_0 f_n + \ldots + A_k f_{n+k})$$

Si es explícito se conoce como método de Adams-Bashforth y si es implícito como método de Adams-Moulton (ver ejercicios).

A menudo se usan estos métodos de a pares y se usa la diferencia entre los dos métodos, $x_i - \tilde{x}_i$ para estimar el error local. Este procedimiento se conoce como "predictor-corrector".

Para comenzar a aplicar los métodos multipaso se necesitan los primeros k valores que usualmente se calculan con métodos a un paso.

4.1. Convergencia de los métodos multipaso. Empecemos por el error de truncación para un método de k pasos

$$\sum_{j=0}^{k} \alpha_j x_{n+j} = h \sum_{j=0}^{k} \beta_j f_{n+j}$$

Si $x(t)$ es la solución de $x' = f(x,t)$el error de truncación local está dado por

$$\sum_{j=0}^{k} \alpha_j x(t+jh) - h \sum_{j=0}^{k} \beta_j x'(t+jh) = h\tau$$

Si $x(t)$ es suficientemente regular, podemos expresar $h\tau$ en la forma

$$h\tau = C_0 x(t) + C_1 h x'(t) + C_2 h^2 x''(t) + \ldots + C_q h^q x^{(q)}(t) + \ldots$$

Para ver esto, escribimos

$$x(t+jh) = x(t) + x'(t)jh + \frac{x''(t)}{2}(jh)^2 + \ldots$$

$$x'(t+jh) = x'(t) + x''(t)jh + \frac{x'''(t)}{2}(jh)^2 + \dots$$

Metiendo esto en la expresión de τ e igualando potencias de h se obtiene

$$C_0 = \alpha_0 + \dots + \alpha_k$$

$$C_1 = \alpha_1 + 2\alpha_2 + 3\alpha_3 + \dots + k\alpha_k - \beta_0 - \beta_1 - \beta_2 - \dots - \beta_k$$

En general para cualquier $q \geq 1$

$$C_q = \frac{1}{q!}(\alpha_1 + 2^q\alpha_2 + 3^q\alpha_3 + \dots + k^q\alpha_k) - \frac{1}{(q-1)!}(\beta_1 + 2^{q-1}\beta_2 + \dots + k^{q-1}\beta_k)$$

Si $C_0 = C_1 = \dots = C_p = 0$ y $C_{p+1} \neq 0$ el método se dice de orden p. Para un método de orden p el error τ satisface

$$\tau h = C_{p+1}h^{p+1}x^{(p+1)}(t) + O(h^{p+2})$$

Para ver la convergencia, como antes, fijamos t^* y ponemos n y h tales que $t^* = (n+k)h$. Queremos

$$\lim_{h\to 0} x_{n+k} = x(t^*)$$

Queremos ver que condiciones debemos imponer al método para que esto ocurra.

Primero pongamos el problema $x'(t) = 0$, $x(0) = 1$ (solución $x \equiv 1$). El método aplicado a este problema se reduce a

$$\sum_{j=0}^{k} \alpha_j x_{n+j} = 0$$

Para cualquier método multipaso debemos tener (como k está fijo y $h \to 0$) que $x_{n+k} \to x(t^*)$, $\dots$, $x_n \to x(t^*)$. Entonces podemos escribir $x_{n+j} = x(t^*) + \varphi_j(h)$ con $\varphi_j(h) \to 0$ cuando $h \to 0$.

Usando esto se obtiene

$$\sum_{j=0}^{k} \alpha_j x(t^*) + \sum_{j=0}^{k} \alpha_j \varphi_j(h) = 0$$

Como la segunda suma tiende a cero con h nos queda

$$x(t^*)\sum_{j=0}^{k}\alpha_j = 0$$

Es decir

$$C_0 = 0.$$

Para ver que C_1 también debe ser cero para que el método converja, consideremos el problema $x'(t) = 1$, $x(0) = 0$ que tiene como solución $x(t) = t$. El método para este problema se reduce a

$$\sum_{j=0}^{k}\alpha_j x_{n+j} = h\sum_{j=0}^{k}\beta_j$$

Es fácil verificar que la sucesión dada por

$$x_l = lhM$$

es solución de este esquema donde

$$M = \frac{\sum_{j=0}^{k}\beta_j}{\sum_{j=0}^{k}j\alpha_j}$$

Si los valores iniciales se eligen de la forma $x_l = lhM$ el método va a producir la solución $x_l = lhM$ y en particular

$$x_{n+k} = (n+k)hM$$

Como suponemos que el método es convergente se tiene que lo valores iniciales satisfacen $x_i \to x(0) = 0$ cuando $h \to 0$, y además $x_{n+k} \to x(t^*)$, pero

$$x_{n+k} = (n+k)hM = t^*M \to t^*$$

Entonces concluímos que

$$M = 1$$

lo que nos da

$$C_1 = 0$$

Esto se denomina consistencia del método multipaso.

Para los métodos de un paso, la consistencia implicaba la convergencia, para los métodos multipaso se requiere una condición adicional la *condición de la raíz*.

Veamos esto. Ahora consideremos el problema $x'(t) = 0$, $x(t) = 0$, cuya solución es $x(t) \equiv 0$.

En este caso el método se reduce a

$$\sum_{j=0}^{k} \alpha_j x_{n+j} = 0$$

Esto describe una ecuación en diferencias que admite como solución a la sucesión

$$x_m = h(r_i)^m$$

donde r_i es cualquiera de las raíces del polinomio $p(r)$ dado por

$$p(r) = r^k + \alpha_{k-1} r^{k-1} + \ldots + \alpha_1 r + \alpha_0$$

Si asumimos que el método es convergente se tiene que

$$x_{n+k} \to x(t^*) = 0 \qquad h \to 0$$

Además

$$x_{n+k} = h(r_i)^{n+k}$$

Para verificar que $x_{n+k} \to 0$, como $n = t^*/h \to \infty$ cuando $h \to 0$ se debe tener que

$$|r_i| \leq 1$$

Entonces la convergencia implica que todo cero de $p(r)$ debe satisfacer

$$|r_i| \leq 1.$$

Además, si r_i es una raíz múltiple de $p(r)$ podemos poner

$$x_j = hj^q (r_i)^j$$

con $q \leq m-1$ (m es la multiplicidad de la raíz). Tenemos

$$x_{n+k} = h(n+k)^q (r_i)^{n+k}$$

y como $h(n+k) = t^*$ para que $x_{n+k} \to 0$ se debe tener

$$|r_i| < 1.$$

Es decir debemos pedir la

CONDICIÓN 8.13. *(de la raíz)*

(1) $|r_i| \leq 1$ *si* r_i *es un cero simple de* $p(r)$.
(2) $|r_i| < 1$ *si* r_i *es un cero múltiple de* $p(r)$.

Ahora podemos enunciar el teorema

TEOREMA 8.14. *Un método multipaso es convergente si y solo si el método es consistente y satisface la condición de la raíz.*

Terminaremos este capítulo con una sección de métodos de paso variable.

5. Métodos de paso variable

La idea es elegir los pasos h_i en forma variable. Por ejemplo,

5.1. Euler adaptivo para una ecuación que explota. Ahora analizaremos un método adaptivo para elegir los pasos τ_j en el método de Euler.

Sea $y(t)$ la solución de la ecuación diferencial

$$\begin{cases} y'(t) = f(y(t)) \\ y(0) = y_0 > 0 \end{cases} \tag{8.9}$$

donde f es positiva, creciente y regular.

Supongamos que $\int^{+\infty} 1/f < +\infty$, entonces $y(t)$ explota en tiempo finito T y podemos calcular exactamente el tiempo de explosión. De hecho vale

$$T = \int_{y_0}^{+\infty} \frac{1}{f(s)} ds$$

Si aproximamos $y(t)$ usando el método de Euler nos queda

$$\begin{cases} y^{j+1} = y^j + \tau_j f(y^j) \\ y^0 = y_0 \end{cases} \tag{8.10}$$

Ahora elegimos

$$\tau_j f(y^j) = \lambda \tag{8.11}$$

como los pasos τ_j. Usando (8.10) y (8.11) tenemos

$$y^{j+1} = y^j + \lambda = \ldots = y^0 + (j+1)\lambda$$

y entonces,

$$\tau_j = \frac{\lambda}{f(y^j)} = \frac{\lambda}{f(y_0 + j\lambda)}.$$

De esto podemos concluir que el esquema numérico también explota en el sentido siguiente: $y^j \to \infty$ mientras $\sum_j \tau_j < +\infty$.

Además nos provee de una estimación del tiempo de explosión. De hecho, vale

$$\sum_{j=0}^{\infty} \tau_j = \tau_0 + \sum_{j=1}^{\infty} \frac{\lambda}{f(y_0 + j\lambda)} \leq \frac{\lambda}{f(y_0)} + \int_0^{+\infty} \frac{\lambda}{f(y_0 + s\lambda)} ds =$$
$$\frac{\lambda}{f(y_0)} + \int_{y_0}^{+\infty} \frac{1}{f(s)} ds = \frac{\lambda}{f(y_(0)} + T$$

Entonces $T_\lambda = \sum_j \tau_j < +\infty$ y

$$T_\lambda - T \leq \frac{\lambda}{f(y_0)}.$$

Además, como y es convexa, es fácil ver que la solución numérica está por debajo de la continua y entonces

$$T \leq T_\lambda.$$

Concluimos que,

$$|T - T_\lambda| \leq \frac{\lambda}{f(y_0)}.$$

6. Ejercicios

(1) Utilizar el método de Euler para resolver $\begin{cases} x' = 2x & \text{en } [0,1] \\ x(0) = 1 \end{cases}$
empleando pasos $h = 0.1$, $h = 0.05$ y $h = 0.01$. Graficar las tres soluciones numéricas obtenidas junto con la solución exacta.

(2) Hacer el mapa de curvas integrales en la región $[0, 10] \times [0, 10]$ de la ecuación diferencial

$$x'(t) = (x(t) - 5).(\cos^2(t) - 0.5),$$

graficando simultáneamente, para $k = 0, 1, \ldots, 10$, la solución que se obtiene utilizando el método de Euler con paso $h = 0.01$ y con condición inicial

$$x(0) = k.$$

(3) Considerar el problema $\begin{cases} x' = \lambda x \\ x(0) = x_0 \end{cases}$.
   (a) Probar que el método de Euler con paso h genera la sucesión:
   $$x_i = (1 + \lambda h)^i x_0 \quad i = 0, 1, \ldots.$$
   (b) Mostrar que si $\lambda < 0$, la solución exacta tiende a cero a medida que x crece.
   (c) Para $\lambda < 0$, determinar para qué valores de h ocurre que $x_i \to 0$ cuando $i \to \infty$.

(4) Se considera el problema
$$\begin{cases} x'(t) = x(t) + t^2 + 3 \quad \text{en } [0, 2] \\ x(0) = -2 \end{cases}$$
   (a) Demostrar que la solución es una función convexa.
   (b) Utilizar los métodos de Euler explícito e implícito, con paso $h = 0.05$ para obtener dos aproximaciones de la solución y graficarlas. Decidir en que región del gráfico deberá situarse la solución analítica del problema.
   (c) Graficar la solución que se logra al utilizar el comando **ode45** de `Matlab`.

(5) Se considera la siguiente ecuación diferencial:
$$\begin{cases} x'(t) = & 2x(t) - 5\sin(t) \\ x(0) = & 1 \end{cases}$$
   cuya solución exacta es la función $x(t) = 2\sin(t) + \cos(t)$. Graficar simultáneamente en el intervalo $[0, 4]$ la solución exacta y las que se obtienen con los métodos de Euler y Taylor de orden 2, ambos con paso $h = 0.05$.

(6) Escriba un programa que resuelva la ecuación diferencial del Ejercicio 5 por algún método de Runge-Kutta de orden 2 y de orden 4. Agregar estas soluciones al gráfico realizado en dicho ejercicio.

(7) Verificar que la función error, **erf**, puede ser definida como la solución de la ecuación diferencial
$$\begin{cases} x'(t) = \frac{2}{\sqrt{\pi}} e^{-t^2} \\ x(0) = 0 \end{cases}.$$
   Utilizar un método de Runge-Kutta de orden 2 para hallar $\text{erf}(t_i)$ con $t_i = 0, 0.05, 0.1, 0.15, \ldots, 1$. Comparar con los valores obtenidos directamente con el comando **erf** de `Matlab`.

(8) Considerar la ecuación
$$x' = x^2, \quad x(0) = 0.5$$
   (a) Calcular el tiempo T en que la solución analítica explota.
   (b) Calcular y graficar en el intervalo $[0, T - 0.1]$ la aproximación a la solución de la ecuación utilizando el método de Euler adaptativo de parámetro λ con un λ tal que el tiempo en que explota la solución numérica T_λ diste de T en menos que 10^{-1}.
   (c) Agregar al gráfico anterior las aproximaciones obtenidas en el mismo intervalo con el método de Euler usual con paso $h = 0.05$ y con el comando **ode45**.

(9) Probar que el método de Runge-Kutta de oprden 4 dado en el Ejemplo 8.6 es consistente.

(10) Hallar el error local para los métodos de Euler explícito e implícito.

(11) Se quiere estimar, aplicando el método de Euler, el valor de e como $x(1)$ donde $x(t)$ es solución de $x' = x,\ x(0) = 1$. Hallar un paso h de modo que el error cometido resulte menor que 10^{-3}. Realizar el mismo trabajo para el método de Taylor de orden 2.

(12) Considerar el problema $x' = -2tx$, $x(0) = 1$, con $t \geq 0$.
   (a) Determinar una cota, en términos de h, para el error cometido si se usa el método de Euler para calcular $x(1)$.
   (b) ¿Cómo debería tomar h si se desea que el error cometido sea menor que 10^{-2}?
   (c) Calcular la solución en $t = 1$ usando el valor de h obtenido en el item previo, y verificar las estimaciones previstas comparando con la solución exacta.

(13) Repetir los items (a) y (b) del ejercicio anterior para el problema:

$$\begin{cases} x'(t) = & t\sin^2(x(t)) \\ x(0) = & 1 \end{cases}$$

(14) La trayectoria de una partícula que se se mueve en el plano está dada por las curva $(x_1(t), x_2(t))$, donde las funciones x_1, x_2 son la solución del siguiente sistema de ecuaciones diferenciales:

$$\begin{array}{rcl} x_1'(t) & = & -x_2(t) \\ x_2'(t) & = & x_1(t) - x_2(t) \end{array}.$$

Resolver este sistema en el intervalo $[0, 20]$ con el método de Euler utilizando paso $h = 0.05$ y graficar la trayectoria de la partícula, sabiendo que en tiempo $t = 0$ se encontraba en el punto $(1, -1)$. Realizar nuevamente el gráfico utilizando la solución obtenida con el comando **ode45**.

(15) Probar que una ecuación de orden n se puede escribir como un sistema de n ecuaciones de primer orden. Mostrar que un problema de valores iniciales para la primera se transforma en un problema de valores iniciales para el sistema.

(16) Considerar el siguiente problema:

$$x'' - 3x' + 2x = 0, \quad \text{con } x(0) = 1,\ x'(0) = 0.$$

Resolver la ecuación analíticamente y aproximar el valor $x(1)$ con un método de Runge-Kutta de orden 2 para distintos valores de h.

(17) Considerar la ecuación $x'(t) = f(t, x(t))$.
   (a) Deducir la fórmula de Milne:

$$x_n = x_{n-2} + h(\frac{1}{3}f_n + \frac{4}{3}f_{n-1} + \frac{1}{3}f_{n-2}),$$

aproximando la integral

$$\int_{t_{n-2}}^{t_n} f(t, x(t))dt = \int_{t_{n-2}}^{t_n} x'(t)dt = x(t_n) - x(t_{n-2}),$$

con la fórmula de Simpson.
   (b) Analizar la convergencia (estabilidad y consistencia) del método y calcular su órden.

(18) Analizar la convergencia de los siguientes métodos y calcular su órden.
   - **Adams-Bashforth.**

$$x_{n+3} - x_{n+2} = \frac{h}{12}(23f_{n+2} - 16f_{n+1} + 5f_n).$$

- **Adams-Moulton.**

$$x_{n+3} - x_{n+2} = \frac{h}{12}(5f_{n+3} + 8f_{n+2} - f_{n+1}).$$

(19) Considerar el método de 2 pasos

$$x_{n+2} + ax_{n+1} + ax_n = h(\beta_2 f_{n+2} + \beta_1 f_{n+1} + \beta_0 f_n).$$

Determinar $a, \beta_2, \beta_1, \beta_0$ de modo que el método resultante tenga orden 4.

(20) Decidir si existe algún valor de $a \in \mathbb{R}$ para el cual el siguiente método multipaso sea convergente:

$$x_{n+3} - 3x_{n+2} + (3 - a^2)x_{n+1} + (a^2 - 1)x_n = h[5f_{n+2} + (-a^2 - 5)f_n].$$

(21) **Miscelánea.** Considerar la ecuación $x' = \sqrt{|x|}$.

(a) Para el valor inicial $x(0) = 0$, seguir las iteraciones del método de Euler, con paso $h = 0.1$ hasta llegar al valor de $x(10)$.

(b) Graficar la solución que se obtiene al aplicar el método de Euler, si el valor de $x(0)$ es dado con un error de 10^{-6}, es decir $x(0) = 0.000001$.

Nota: La gran propagación del error en el dato inicial se debe a que esta ecuación tiene infinitas soluciones si $x(0) = 0$. En particular, cualquiera sea $\alpha > 0$

$$x(t) = \begin{cases} 0 & t \leq \alpha \\ \dfrac{(t-\alpha)^2}{4} & t > \alpha \end{cases}$$

es solución de la misma.