Universidad de Vigo

Notas de Clase

de la asignatura

Matemáticas de la Especialidad

(Regulación Automática y Electrónica)

(plan de 1981)

Curso 2001-02

Emilio Faro Rivas

Índice General

Capítulo 1

Resolución de ecuaciones no lineales

Uno de los primeros problemas de las matemáticas ha sido el de la resolución de ecuaciones. Pero, ¿qué es una ecuación?

El concepto general de ecuación

El concepto más general de ecuación se obtiene al igualar dos funciones $f : X \to A$, $g : Y \to A$ cuyos valores pertenecen a un mismo rango. Resolver la ecuación de f y g es, en tal situación, hallar todos los pares (x, y) para los cuales se verifica la igualdad

$$f(x) = g(y). \tag{1.1}$$

Pero nada se gana con tanta generalidad porque esta formulación tan general es, curiosamente, equivalente a una simplificación de uno de sus casos particulares: a saber, el caso más sencillo en que las dos funciones tienen el mismo dominio, es decir, son de la forma:

$$X \underset{g}{\overset{f}{\rightrightarrows}} A$$

y la solución de la ecuación de f y g está formada por todos los elementos $x \in X$ tales que

$$f(x) = g(x). \tag{1.2}$$

Para convencerse de que saber resolver ecuaciones del tipo (1.2) es suficiente para resolver todas las del tipo (1.1) no hay más que ver que el

problema de (1.1) se puede reducir al de (1.2) utilizando el producto cartesiano, $X \times Y$, de los dominios. Un razonamiento más sutil (que utiliza la diagonal $\Delta : X \to X \times X$) muestra que saber resolver todos los problemas de tipo (1.1) es suficiente para resolver todos los del tipo (1.2).

En el planteamiento general que acabamos de presentar no hace falta suponer que f y g sean funciones numéricas. Sin embargo, en el caso de que lo sean, (1.2) es equivalente a la ecuación $f(x) - g(x) = 0$. Este es el caso que nos interesará en lo que sigue y por tanto podemos decir que las soluciones de una ecuación como (1.2) son los *ceros* de la función diferencia $f - g$. Así, nuestro problema general será el de resolver ecuaciones de la forma

$$f(x) = 0\,, \tag{1.3}$$

es decir, el de hallar ceros de funciones.

¿Qué es un punto fijo?

También es posible plantear la resolución de una ecuación como el problema de hallar los *puntos fijos* de una función g, es decir, las soluciones de

$$g(x) = x\,.$$

Por ejemplo, dar una solución de (1.3) es equivalente a dar un punto fijo de cualquier función de la forma $g(x) = x + f(x)/\lambda$ con $\lambda \neq 0$ (entre otras muchas posibles). Planteado el problema de esta forma se puede someter a los métodos generales de resolución de problemas de punto fijo, que se llaman *métodos iterativos*.

Así pues, estudiaremos dos tipos de métodos de resolución de ecuaciones: *métodos generales* (o búsqueda de ceros) y *métodos iterativos* (o búsqueda de puntos fijos). Además de esto estudiaremos algunos métodos particulares para las ecuaciones polinómicas, es decir para el caso en que la función $f(x)$ en (1.3) sea un polinomio.

1.1 Métodos generales

1.1.1 Método de la bisección

El método de la bisección sirve para aproximarse tanto como se quiera a un cero de una función real de variable real continua y que toma valores de signos opuestos en los extremos de un intervalo $[a, b]$. Suponemos, pues, que f es continua en $[a, b]$ y que $f(a)\,f(b) < 0$. Esto implica que f

tiene (al menos) un cero en $[a, b]$. El punto medio de $[a, b]$, $x_0 = (a + b)/2$, aproxima al cero con un error menor que $|b - a|/2$. Existen dos posibilidades: (1) $f(x_0) = 0$, en cuyo caso ya hemos hallado el cero, y (2) $f(x_0) \neq 0$, en cuyo caso, según que el signo de $f(x_0)$ sea opuesto del de $f(a)$ o del de $f(b)$, el intervalo $[a, x_0]$ ó el $[x_0, b]$ tiene con certeza un cero. Si se da la posibilidad (2) hemos reducido la acotación del cero a la mitad, de forma que si al principio conocíamos el cero con un error menor que $(b - a)/2$, después de un paso de bisección el error en el nuevo punto medio es menor que $(b-a)/2^2$. Continuando el proceso vamos reduciendo el error a la mitad en cada paso. Así, llegamos al siguiente algoritmo para el método de la bisección. En él se calcula el número N de pasos que hay que dar para garantizar que el error sea menor que ϵ en base a la relación $(b - a)/2^{N+1} \leq \epsilon$, que es equivalente a $N \geq (\ln(b - a) - \ln\epsilon)/\ln 2 - 1$.

Algoritmo del Método de la Bisección

1 **Datos:** f (la función de la que se quiere un cero)**,** a**,** b (extremos del intervalo en que $f(a)\,f(b) < 0$)**,** ϵ (precisión deseada).
2 $N = \text{ent}((\ln(b - a) - \ln\epsilon)/\ln 2)$
3 **Si** $N = (\ln(b - a) - \ln\epsilon)/\ln 2$ **entonces** $N = N - 1$
4 $x = (a + b)/2$
5 **Para** $i = 1$ **hasta** N
  Si $f(x) = 0$ **entonces ir a 7**
  Si $f(x)\,f(a) < 0$ **entonces** $b = x$
  Si $f(x)\,f(a) > 0$ **entonces** $a = x$
  $x = (a + b)/2$
6 **siguiente** i
7 **Resultado:** x (el valor del cero con error menor que ϵ).

El principal inconveniente de este método es su lentitud. Esto es más de notar cuando es necesario aplicarlo repetidamente muchas veces y lo comparamos con otros métodos que veremos a continuación. Por otro lado, tiene la ventaja de ser un método sencillo, robusto y que ofrece un control excelente de la precisión de la aproximación y la posibilidad de calcular el número exacto de pasos que hay que dar para aproximar la solución con un error inferior a un valor dado. Por estas razones este método es

adecuado cuando necesitamos el cálculo sencillo y esporádico de un cero de una función continua con una cierta precisión dada.

Ejercicio 1.1
Hallar el número de pasos que se han de dar con el algoritmo de la bisección para hallar, con un error menor que una milésima, el cero de una función de la que se sabe que toma valores de signos opuestos en los extremos del intervalo $[-1, 1]$*. ¿Y si sólo quisiéramos que el error fuese menor que un octavo?*

1.1.2 Regula Falsi (falsa posición)

El siguiente método es una variante del anterior en el que se pretende acelerar la convergencia utilizando un punto de corte del intervalo que tenga mayor probabilidad de estar cerca del cero, a saber la posición que tendría el cero en caso de que la función fuese una línea recta (pasando por los puntos $(a,\ f(a))$ y $(b,\ f(b))$).

Ejercicio 1.2
Demostrar que el punto de corte de la recta que pasa por los puntos $(a,\ f(a))$ *y* $(b,\ f(b))$ *con el eje* x *es*

$$x_0 = a - f(a)\frac{b-a}{f(b)-f(a)} = \frac{af(b)-bf(a)}{f(b)-f(a)}.$$

El siguiente ejercicio muestra que otra forma de describir el punto de corte es diciendo que es el promedio ponderado de los puntos a y b con los pesos $|f(b)|$ y $|f(a)|$.

Ejercicio 1.3
Demostrar que si $f(a)f(b) < 0$ *entonces*

$$\frac{a|f(b)|+b|f(a)|}{|f(b)|+|f(a)|} = \frac{af(b)-bf(a)}{f(b)-f(a)}.$$

Algoritmo de la Regula Falsi

1 **Datos:** f (la función de la que se quiere un cero)**,** a**,** b (extremos del intervalo en que $f(a)f(b) < 0$)**,** ϵ (precisión deseada).

2 $x_0 = (af(b) - bf(a))/(f(b) - f(a))$
3 $x = x_0$
4 **Si** $f(x) = 0$ **entonces ir a 10**
5 **Si** $f(x)f(a) < 0$ **entonces** $b = x$
6 **Si** $f(x)f(a) > 0$ **entonces** $a = x$
7 $x_0 = (af(b) - bf(a))/(f(b) - f(a))$
8 **Si** $|x - x_0| < \epsilon$ **entonces ir a 10**
9 **Ir a 3**
10 **Resultado:** x_0.

Este algoritmo produce rápidamente un punto x donde $|f(x)| \simeq 0$, pero no produce un intervalo pequeño que contenga el cero. La razón es que en general va a acercarnos al cero por la derecha (manteniendo fijo el extremo inferior del intervalo) si la función es cóncava en el intervalo, o por la izquierda (manteniendo fijo el extremo superior del intervalo) si la función es convexa. Por ejemplo, en la siguiente figura (fig. 1.1) en que tenemos una función cóncava vemos que las aproximaciones se van acercando al cero por la derecha.

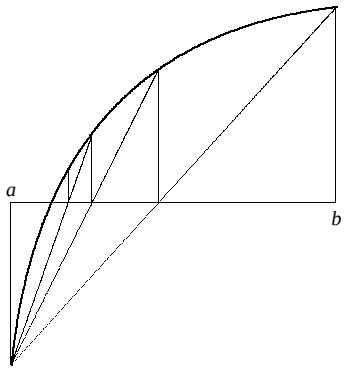


Figura 1.1: Regula Falsi.

Este problema se evita en el método de la *regula falsi modificada*.

1.1.3 Regula Falsi Modificada

Este método es una pequeña variante de la regula falsi. Con él se intenta evitar el que sucesivas aproximaciones se mantengan del mismo lado del cero, de forma que ningún extremo del intervalo quede fijo todo el tiempo. De esta forma la longitud del intervalo de acotación se va reduciendo hasta hacerse tan pequeño como se quiera.

Para conseguir esta reducción del intervalo de acotación, si en un paso usamos la cuerda de extremos $(a,\ f(a))$ y $(b,\ f(b))$ para obtener la aproximación $x = (af(b) - bf(a))/(f(b) - f(a))$, antes del siguiente paso miramos si el signo de $f(x)$ coincide con el de $f(a)$ o con el de $f(b)$. Supongamos que coincide con el de $f(b)$ (como en la figura 1.2. —Si coincidiese con $f(a)$ se realizaría el proceso análogo correspondiente—), entonces el siguiente punto se obtendrá a partir del segmento de extremos $(a,\ \frac{f(a)}{2})$ y $(x,\ f(x))$ continuando de esta forma, cada vez reduciendo a la mitad la ordenada del extremo del segmento en a hasta que se obtenga un punto en el que el valor de f tenga el mismo signo que $f(a)$. En ese momento se vuelve a trazar una cuerda y se repite el proceso. Un algoritmo sencillo para implementar este método se indica a continuación

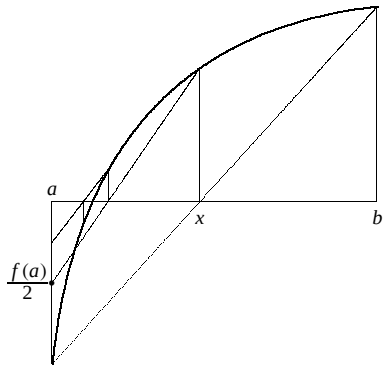


Figura 1.2: Regula Falsi Modificada.

Algoritmo de la Regula Falsi Modificada

1 **Datos:** f (la función de la que se quiere un cero)**,** a**,** b (extremos del intervalo en que $f(a)f(b) < 0$)**,** ϵ (precisión deseada).
2 $F = f(a)$**;** $G = f(b)$**;** $x_0 = a$
3 $x = (aG - bF)/(G - F)$
4 **Si** $f(a)f(x) < 0$ **entonces** $b = x$**;** $G = f(x)$ **en otro caso ir a 6**
5 **Si** $f(x)f(x_0) > 0$ **entonces** $F = F/2$
6 **Si** $f(x)f(a) > 0$ **entonces** $a = x$**;** $F = f(x)$ **en otro caso ir a 8**
7 **Si** $f(x)f(x_0) > 0$ **entonces** $G = G/2$
8 **Si** $f(x) = 0$ **entonces ir a 12**
9 **Si** $|b - a| < \epsilon$ **entonces ir a 12**
10 $x_0 = x$
11 **Ir a 3**
12 **Resultado:** x (el valor del cero con error menor que ϵ).

1.1.4 Método de la Secante e idea del Método de Müller

El método de la secante puede considerarse como otra modificación de la regula falsi porque al igual que en aquel método, en cada paso vamos a intersecar una secante a la gráfica de nuestra función con el eje x. Sin embargo ahora abandonamos el propósito de obtener intervalos que encierren a la solución en cada paso y simplemente construimos una sucesión que converge a dicha solución. Así, en este método cada secante es la que corresponde a los dos últimos puntos encontrados y por lo tanto x_{n+1} se obtiene al intersecar el eje x con la recta que pasa por los puntos $(x_{n-1},\ f(x_{n-1}))$ y $(x_n,\ f(x_n))$,

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})}. \tag{1.4}$$

La fórmula (1.4) es propensa a dar errores de pérdida de precisión cuando x_n es próximo a x_{n-1}. Como alternativa es preferible usar la fórmula equivalente

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} \tag{1.5}$$

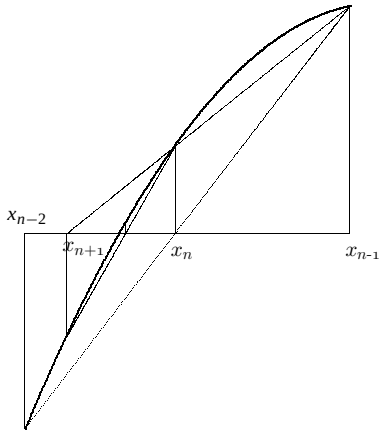


Figura 1.3: Método de la secante.

Cuando converge, el método de la secante nos acerca muy rápidamente a la solución. El problema principal de este método es que necesita partir de dos estimaciones iniciales, x_0 y x_1, *suficientemente cercanas* al punto buscado. Si las dos estimaciones iniciales no son suficientemente cercanas a la solución es fácil que el método no converja. En consecuencia este método debe ir acompañado de alguna técnica de acercamiento previo a la solución.

Un método parecido en su concepción al método de la secante es el método de Müller, que estudiamos más detenidamente en la sección 1.4. Aquí sólo queremos recalcar la relación que existe entre los dos métodos. La idea principal en el método de la secante es que, conocida la función en dos puntos cercanos a un cero, podemos aproximar ese cero por la raíz del polinomio de primer grado que coincide con la función en esos dos puntos. El método de Müller se obtiene al llevar esta idea un paso más lejos: aproximar el cero de nuestra función por una raíz del polinomio de segundo grado que coincide con la función en tres puntos cercanos al cero buscado.

1.2 Métodos iterativos

Iterar una función es evaluarla en un punto que es resultado de haberla evaluado previamente, es decir, hacer una evaluación de la forma $f(f(x))$. Un *endomorfismo* es una función $f : A \to A$ cuyo dominio es igual a su codominio o rango. Tales funciones pueden iterarse repetida e indefinidamente porque sus valores siempre pertenecerán a su dominio (es decir, al conjunto de los elementos en los que f se puede evaluar). Por tanto para una tal función todo elemento $x_0 \in A$ define una sucesión en A que consiste en las sucesivas iteraciones de f comenzando con x_0: $x_n = f(\cdots f(f(x_0))\cdots)$, ó $x_n = (f \circ \cdots \circ f)(x_0)$ ó

$$x_0 \in A, \qquad x_{n+1} = f(x_n).$$

Un importante concepto asociado con las funciones que son endomorfismos es el de *punto fijo*. Decir que un punto $x \in A$ es un punto fijo de la función $f : A \to A$ significa que

$$f(x) = x.$$

Por ejemplo, el número 1 es un punto fijo de la función $y = x^2$. Igualmente lo es el número 0. Éstos son lo únicos puntos fijos de la función x^2. Para la función identidad, $y = x$, todo número es un punto fijo.

Ejercicio 1.4
Si $f : A \to A$ es un endomorfismo idempotente *(es decir, tal que $f^2 = f$) entonces los puntos fijos de f son precisamente los valores de f, es decir los puntos de la forma $a = f(x)$ para algún $x \in A$.*

Ejercicio 1.5
Si g es una función continua y $\{x_n\}$ es una sucesión de iteraciones de g que converge, entonces su límite $\xi = \lim\{x_n\}$ es un punto fijo de g.

Basado en este resultado surge el método iterativo de resolución de problemas de punto fijo, que consiste en formar una sucesión de iteraciones que converja.

Recordemos que una condición sencilla que garantiza la convergencia de una sucesión de iteraciones de una función g es la existencia de una métrica en A para la que g sea una función *contractiva* (teorema de Banach del punto fijo).

Un método iterativo de resolución de una ecuación $f(x) = 0$ consiste, pues en los siguientes pasos:

1. Elegir una función g, continua, cuyos puntos fijos sean ceros de f.
2. Formar una sucesión de iteraciones de g que sea convergente.
3. Hallar el límite de dicha sucesión, que será necesariamente la solución de nuestra ecuación.

Tal función g se llama una *función de iteración* para la ecuación $f(x) = 0$. El ejemplo más famoso de función de iteración es la que corresponde al método iterativo llamado método de Newton, que describimos a continuación.

1.2.1 Método de Newton

Dada una ecuación $f(x) = 0$, donde f es una función diferenciable, el método de Newton para aproximarse a un cero de f es el método iterativo correspondiente a la función de iteración

$$g(x) = x - \frac{f(x)}{f'(x)}. \tag{1.6}$$

Ejercicio 1.6
Demostrar que todo punto fijo de la función definida en (1.6) es un cero de f.

Ejercicio 1.7
Demostrar que la intersección del eje x con la recta que pasa por el punto (x, y) con pendiente m tiene su abscisa igual a $x - y/m$. Como consecuencia la intersección de la tangente a la gráfica de f en el punto $(x, f(x))$ con el eje x tiene su abscisa igual a $x - f(x)/f'(x)$.

Según este último ejercicio, el método de Newton puede describirse geométricamente como una variante del método de la secante: en lugar de trazar secantes a la gráfica de nuestra función, trazamos rectas tangentes y por lo demás procedemos igual que en el método de la secante (figura 1.4).

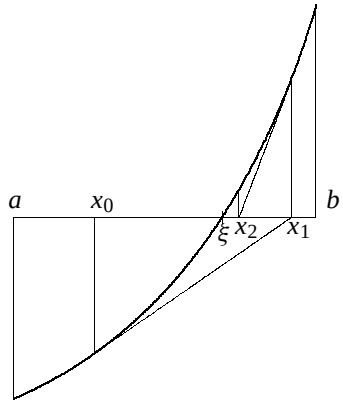


Figura 1.4: Método de Newton.

El método de Newton produce una sucesión en la que cada término, x_{n+1}, se obtiene del anterior por la fórmula

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}\,.$$

Nótese la semejanza entre esta fórmula y la del método de la secante en la forma (1.5). De hecho ésta es el caso límite de aquella para $x_{n-1} = x_n$, lo que corresponde al hecho geométrico de que la recta tangente es la posición límite de rectas secantes cuyos dos puntos de corte se aproximan uno al otro hasta coincidir en el punto de tangencia.

Para aplicar el método de Newton se puede utilizar el siguiente algoritmo:

Algoritmo del Método de Newton

1 **Datos:** f (la función de la que se quiere un cero)**,** x (estimación inicial de la solución)**,** M (número máximo de pasos que queremos dar)**,** ϵ (precisión deseada).
2 **Para** $i = 1$ **hasta** M
3 $x_0 = x$

4 $x = x - f(x)/f'(x)$
5 **Si** $|x - x_0| < \epsilon$ **entonces ir a 9**
6 **Siguiente** i
7 **Imprimir** "No hubo convergencia después de"**;** M**;** "pasos"
8 **PARAR**
9 **Resultado:** x (con error menor que ϵ)

Ejercicio 1.8
Aplicar el método de Newton a la ecuación $x^2 - N = 0$ para obtener un algoritmo para el cálculo de la raíz cuadrada de un número positivo N. Diseñar un algoritmo análogo para el cálculo de la raíz cúbica.

Como veremos más adelante, el método de Newton, cuando converge, converge incluso más rápido que el método de la secante. Sin embargo este método plantea varias dificultades. Una de ellas es la necesidad de evaluar la función derivada f' muchas veces, lo cual puede llegar a ser muy costoso. Para mitigar esta dificultad se han diseñado variantes del método de Newton en las que se evalúa la derivada un número menor de veces. La idea tras estos métodos es la siguiente: Supongamos que modificamos el método de Newton de forma que para calcular x_{n+1} con n par, en lugar de evaluar la derivada en x_n utilizamos el valor $f'(x_{n-1})$ que habíamos calculado en el paso anterior, de forma que *para n par* calculamos $x_{n+1} = x_n - f(x_n)/f'(x_{n-1})$ y para n impar calculamos $x_{n+1} = x_n - f(x_n)/f'(x_n)$. Entonces, dado que (al menos para n grande) las pendientes $f'(x_n)$ son todas próximas a su límite $f'(\xi)$, obtendremos una sucesión que diferirá muy poco de la del método de Newton y convergerá a ξ casi tan rápidamente como aquella, pero con la mitad de evaluaciones de la derivada. Llevando esto más lejos podemos elegir un entero positivo p y modificar el método de forma que se calculen solamente una derivada de cada p términos, o sea que definimos la sucesión de iteraciones de la siguiente forma:

$$x_{n+1} = \begin{cases} x_n - f(x_n)/f'(x_0) & \text{si } 0 < n \leq p-1 \\ x_n - f(x_n)/f'(x_p) & \text{si } p < n \leq 2p-1 \\ \vdots & \\ x_n - f(x_n)/f'(x_{rp}) & \text{si } rp < n \leq (r+1)p-1 \\ \vdots & \end{cases}$$

La otra dificultad básica del método de Newton consiste en la necesidad de conocer una estimación inicial suficientemente próxima al cero buscado. El siguiente teorema nos da una idea precisa de lo que es "suficientemente próximo" en este contexto.

Teorema 1 (Convergencia del método de Newton) *Supongamos que f es una función continua en un intervalo $[a, b]$, que tiene derivada segunda continua en (a, b) y que satisface las siguientes condiciones:*

1. *$f(a)\,f(b) < 0$, es decir, los valores de f en los extremos de $[a, b]$ tienen distinto signo.*
2. *La derivada primera de f no se anula en $[a, b]$, es decir, f es estrictamente monótona en $[a, b]$.*
3. *La derivada segunda de f no cambia de signo en $[a, b]$, es decir, la concavidad de f no cambia en $[a, b]$.*
4. $$\left|\frac{f(a)}{f'(a)}\right| < b - a \quad y \quad \left|\frac{f(b)}{f'(b)}\right| < b - a,$$
   (esto, supuestas las condiciones anteriores, equivale a que las tangentes a la gráfica de f en $(a,\ f(a))$ y $(b,\ f(b))$ intersecan al eje x dentro del intervalo $[a, b]$)

entonces f tiene un único cero, ξ, en $[a, b]$ y el método de Newton converge a ξ para cualquier estimación inicial $x_0 \in [a, b]$.

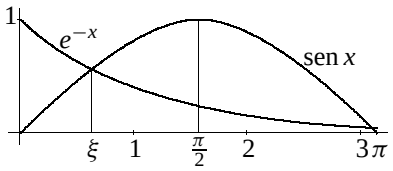
Demostración: Por la continuidad de f y las dos primeras condiciones, f tiene un único cero, ξ, en $[a, b]$. Sin perder generalidad se puede suponer que f es convexa y creciente, es decir, $f(a) < 0$ y $f''(x) \geq 0$ en (a, b) (como en la figura 1.4). Si $x_0 \neq \xi$ hay dos posibilidades:

1. $\xi < x_0$. En este caso, por la convexidad de f tendremos $\xi \leq x_1 \leq x_0$ y en general $\xi \leq x_{n+1} \leq x_n$, es decir, la sucesión producida por el método de Newton es monótona decreciente acotada inferiormente. En consecuencia converge siendo su límite necesariamente el cero, ξ, de f.
2. $x_0 < \xi$. En este caso tendremos $x_0 < \xi \leq x_1 \leq b$ con lo que a partir de x_1 estamos en la situación del caso anterior, obteniendo la misma conclusión.

Los demás casos (según la convexidad de f y si es creciente o decreciente) se analizan de forma análoga, llegándose a las mismas conclusiones. ■

Ejemplo de aplicación del método de Newton.–

Como ejemplo, supongamos que queremos hallar el menor cero positivo de $f(x) = e^{-x} - \operatorname{sen} x$. Vamos a buscar un intervalo que satisfaga las condiciones del teorema de convergencia del método de Newton. Comenzamos calculando las derivadas $f'(x) = -e^{-x} - \cos x$, $f''(x) = e^{-x} + \operatorname{sen} x$. Vemos que $f(0) = 1$, $f(1) \simeq -0.47$. Además $f'(x) < 0$ para $x \in [0, 1]$



y también $f''(x) > 0$ para $x \in [0, 1]$. Sólo nos falta estudiar la última condición en $[0, 1]$: Tenemos, $f(0) = 1$, $f'(0) = -2$, $|f(0)/f'(0)| = \frac{1}{2} < b - a = 1$; $f(1) = -0.47$, $f'(1) = -1.36$, $|f(1)/f'(1)| = 0.34 < b - a$. Luego las condiciones se satisfacen en el intervalo $[0, 1]$.

El método de Newton es uno de los métodos iterativos más importantes pero en ocasiones puede resultar más apropiado utilizar un método iterativo distinto (basado en otra función de iteración). A continuación estudiamos las propiedades generales de los métodos iterativos.

1.2.2 Propiedades generales de los métodos iterativos

Dada una ecuación de la forma $f(x) = 0$ existen diversas funciones $g(x)$ cuyos puntos fijos son precisamente los ceros de f. El ejemplo más sencillo, mencionado más arriba, es $g(x) = x + f(x)$, pero en general habrá muchas otras posibilidades como se ve en el siguiente ejercicio.

Ejercicio 1.9
Demostrar que los puntos fijos de las cuatro funciones dadas a continuación

son solución de la ecuación $x^2 - x - 2 = 0$*:*

$$g_1(x) = \sqrt{2 + x}\,, \qquad g_2(x) = 1 + \frac{2}{x}\,,$$

$$g_3(x) = x - \frac{x^2 - x - 2}{m} \ (m \neq 0)\,, \qquad g_4(x) = \frac{x^2 + 2}{2x - 1}.$$

Ahora bien, no toda función cuyos puntos fijos sean solución de nuestra ecuación sirve como función de iteración. Es necesario que las sucesivas iteraciones de la función den una sucesión convergente. En general vamos a pedir a toda función de iteración g que cumpla las siguientes condiciones:

Condiciones que ha de cumplir una función de iteración.–

(a) Que exista un intervalo $I = [a, b]$ en el que g esté definida y en el que g tome sus valores, es decir que g sea un endomorfismo de I, $I \xrightarrow{g} I$,

(b) Que g sea continua en I,

(c) Que g sea diferenciable en I y exista una constante $k < 1$ tal que para todo $x \in I$ se cumpla $|g'(x)| \leq k$.

Nótese que nos hemos permitido cierta redundancia en estas condiciones (por ejemplo (c) $\Rightarrow$ (b)). Lo más importante de estas condiciones es que garantizan la existencia y unicidad de punto fijo y la convergencia de la sucesión de iteraciones. La existencia de punto fijo es una sencilla consecuencia de (a) y (b), cuya demostración proponemos como ejercicio:

Ejercicio 1.10
Demostrar que toda función g que cumpla las propiedades (a) y (b) anteriores tiene un punto fijo en el intervalo I.

(Sugerencia: Estudiar los cambios de signo de la función $h(x) = g(x) - x$.)

Además de esto tenemos:

Teorema 2 *Si g es una función que satisface las condiciones (a), (b), (c) anteriores entonces g tiene un único punto fijo, $\xi \in I$ y para cualquier valor inicial $x_0 \in I$, la sucesión de iteraciones $x_{n+1} = g(x_n)$, converge a ξ.*

Demostración: Sabemos por el ejercicio 1.10 que existe en I un punto ξ que es punto fijo de g, es decir tal que, $g(\xi) = \xi$. Si demostramos que toda sucesión de iteraciones converge a ξ, la unicidad del punto fijo será consecuencia de la unicidad de límite de una sucesión. Ahora bien, los errores $e_n = \xi - x_n$ verifican

$$e_n = g(\xi) - g(x_{n-1}) = g'(\eta_n)(\xi - x_{n-1}) = g'(\eta_n)e_{n-1}$$

ya que según el teorema del valor medio existe un número η_n (entre ξ y x_{n-1}) tal que $g'(\eta_n) = (g(\xi) - g(x_{n-1}))/(\xi - x_{n-1})$. En consecuencia, por ser $|g'(x)| \leq k$ para todo $x \in I$, tenemos que para todo n $|e_n| \leq k|e_{n-1}|$, de donde

$$|e_n| \leq k|e_{n-1}| \leq k^2|e_{n-2}| \leq \cdots \leq k^n|e_0|$$

y por tanto, al ser $k < 1$,

$$\lim_{n\to\infty} |e_n| = |e_0| \lim_{n\to\infty} k^n = 0,$$

lo que significa que $\lim\{x_n\} = \xi$. ■

En aquellos casos en que sea difícil comprobar la condición (a), puede ser suficiente aplicar la siguiente versión débil del teorema:

Corolario 1 *Si g es una función continuamente diferenciable en algún intervalo abierto I que contenga un punto fijo ξ de g, y si $|g'(\xi)| < 1$, entonces existe un número $\epsilon > 0$ tal que para todo valor inicial x_0 que diste de ξ menos que ϵ la sucesión de iteraciones $x_{n+1} = g(x_n)$, converge a ξ.*

Demostración: Por ser $|g'(\xi)| < 1$ existe k tal que $|g'(\xi)| < k < 1$, y por ser g' continua en ξ existe un $\epsilon > 0$ tal que para todo x que diste de ξ menos que ϵ, $|g'(x)| < k$. Entonces se cumple la condición (c) en el intervalo $I = [\xi - \epsilon, \xi + \epsilon]$ (lo que implica la condición (b)). Además para todo $x \in I$

$$|g(x) - \xi| = |g(x) - g(\xi)| = |g'(\xi)||x - \xi| \leq k\epsilon \leq \epsilon\,,$$

luego $g(x) \in I$, de forma que también se cumple (a) en I y podemos aplicar el teorema al intervalo I. ■

Ejercicio 1.11
Para cada una de las cuatro funciones del ejercicio 1.9 hallar un intervalo I, si existe, en el que se cumplan las condiciones (a), (b) y (c).

1.2.3 Velocidad de convergencia

Definición: Sea $\{x_n\}$ una sucesión convergente cuyo límite es ξ. Sea, para cada entero positivo n, $e_n = \xi - x_n$ (el "error" del término n-ésimo). Decimos que $\{x_n\}$ tiene *orden de convergencia* igual al número real p si la sucesión $\{|e_{n+1}|/|e_n|^p\}$ es convergente y su límite es distinto de cero.

A la vista de esta definición quizás a uno le pueda quedar la duda de si el número real p tiene que ser único si existe. Por lo tanto conviene hacer el siguiente ejercicio para disipar esa duda.

Ejercicio 1.12
Demostrar que si p y q son dos números reales tales que

$$0 < \lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} < \infty$$

y

$$0 < \lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^q} < \infty$$

entonces $p = q$.

Si $\{x_n\}$ tiene orden de convergencia igual a p la constante

$$c = \lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p}$$

se llama *constante de error* o constante de error asintótico de $\{x_n\}$.

Si una sucesión tiene orden de convergencia igual a 1 decimos que tiene *convergencia lineal* y si tiene orden de convergencia igual a 2 decimos que tiene *convergencia cuadrática*.

Proposición 1 *Sea g una función de iteración continuamente diferenciable para la que la sucesión de iteraciones $x_{n+1} = g(x_n)$ converge al punto fijo ξ de g y para la que $g'(\xi) \neq 0$. Entonces la sucesión de iteraciones tiene convergencia lineal y su constante de error es igual a $|g'(\xi)|$.*

Demostración: Al igual que en la demostración del teorema 2, deducimos (aplicando el teorema del valor medio) que para cada entero positivo n existe un número η_n en el intervalo de extremos x_n y ξ tal que

$|e_{n+1}| = |g'(\eta_n)||e_n|$. Como evidentemente $\lim\{\eta_n\} = \xi$,

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|} = |g'(\xi)|\,.$$

■

Igualmente, si g es una función de iteración con derivada segunda continua y con punto fijo ξ, tal que $g'(\xi) = 0$ y $g''(\xi) \neq 0$ entonces $e_{n+1} = \frac{1}{2}g''(\eta_n)e_n^2$ de donde se deduce que la sucesión de iteraciones de g tiene convergencia cuadrática y la constante de error asintótico es $|\frac{1}{2}g''(\xi)|$. En general tenemos lo siguiente

Ejercicio 1.13
En las condiciones de la proposición 1, si las $n-1$ primeras derivadas de g se anulan en ξ y la derivada n-ésima no se anula entonces la sucesión de iteraciones tiene orden de convergencia igual a n. ¿Cuánto vale la constante de error?

Orden de convergencia del método de Newton.–

Podemos ahora justificar la ventaja del método de Newton respecto a su velocidad de convergencia demostrando que, si la derivada de f en ξ no se anula, tiene convergencia cuadrática. Para ello calculamos la derivada de la función de iteración $g(x) = x - f(x)/f'(x)$ obteniendo

$$g'(x) = 1 - (f'(x)^2 - f(x)f''(x))/f'(x)^2 = f(x)f''(x)/f'(x)^2.$$

Ahora hemos de evaluar esta derivada en el punto ξ, en el que $f(\xi) = 0$. La forma de hacerlo es calculando el límite

$$g'(\xi) = \lim_{x\to\xi} \frac{f(x)f''(x)}{f'(x)^2} = 0.$$

Esto prueba que la convergencia es al menos cuadrática.

Para hallar la constante de error asintótico hemos de hallar la derivada segunda de g y evaluarla en $x = \xi$. El resultado es el siguiente:

Ejercicio 1.14
Si $f'(\xi)$, $f''(\xi) \neq 0$ entonces la constante de error asintótico del método de Newton aplicado a la función f para hallar el cero ξ es igual a

$$\frac{1}{2}\left|\frac{f''(\xi)}{f'(\xi)}\right|$$

Ejercicio 1.15
Si ξ es un cero de f de orden dos, (es decir, $f(\xi) = f'(\xi) = 0$ y $f''(\xi) \neq 0$) entonces el método de Newton no converge cuadráticamente (probar que $g'(\xi) = \frac{1}{2}$). En este caso se puede modificar el algoritmo de Newton usando la función de iteración

$$g(x) = x - \frac{2f(x)}{f'(x)}$$

la cual, si f''' es continua, da lugar a una convergencia cuadrática.

(Sugerencia: Usar el hecho de que

$$\lim_{x\to\xi} \frac{f(x)f''(x)}{f'(x)^2} = \lim_{x\to\xi} \frac{f(x)}{f'(x)^2} \lim_{x\to\xi} f''(x)$$

junto con la regla de L'Hôpital.)

El resultado de este ejercicio se puede generalizar:

Ejercicio 1.16
Si ξ es un cero de f de orden m entonces la función de iteración

$$g(x) = x - m\frac{f(x)}{f'(x)}$$

da lugar a una convergencia cuadrática.

Orden de convergencia del método de la secante.–

El método de la secante no es un método iterativo en el sentido que aquí entendemos. Sin embargo es un método que da lugar, igual que los métodos iterativos a una sucesión de la que podemos estudiar la velocidad de convergencia siguiendo los conceptos generales vistos más arriba. De este estudio se deduce que la convergencia es casi tan buena como la del método de Newton, ya que su orden de convergencia es cercano a 1.62.

Para hallar el orden de convergencia del método de la secante necesitamos establecer primeramente la siguiente representación de una función (lema 1):

$$f(x) = f(\alpha) + f[\alpha,\beta](x-\alpha) + f[\alpha,\beta,x](x-\alpha)(x-\beta)\,. \qquad (1.7)$$

La notación usada en esta fórmula significa lo siguiente: en primer lugar tenemos la *primera diferencia dividida*,

$$f[a,b] = \begin{cases} \frac{f(a)-f(b)}{a-b} & \text{si } a \neq b, \\ f'(a) & \text{si } a = b. \end{cases}$$

En segundo lugar tenemos la *segunda diferencia dividida*:

$$f[a,b,c] = \begin{cases} \frac{f[a,b]-f[a,c]}{b-c} & \text{si } b \neq c, \\ \frac{f[a,c]-f[b,c]}{a-b} & \text{si } a \neq b, \\ \frac{1}{2}f''(a) & \text{si } a = b = c. \end{cases}$$

Ejercicio 1.17

Demostrar que con la notación que acabamos de introducir, la sucesión obtenida por el método de la secante se puede expresar de la siguiente forma:

$$x_{n+1} = x_n - \frac{f(x_n)}{f[x_n, x_{n-1}]}.$$

Ejercicio 1.18

Demostrar que, suponiendo $a \neq b$ y $b \neq c$,

$$\frac{f[a,b]-f[a,c]}{b-c} = \frac{f[a,c]-f[b,c]}{a-b}$$

y

$$\lim_{b\to a} \frac{f[a,b]-f[a,a]}{b-a} = \frac{1}{2}f''(a).$$

Estas diferencias divididas se estudiarán con más detalle en el próximo capítulo. De momento sólo las necesitamos para establecer la fórmula (1.7):

Lema 1 *Sea f una función con derivada segunda continua en un intervalo I y sean $\alpha, \beta \in I$. Entonces, para todo $x \in I$ se cumple*

$$f(x) = f(\alpha) + f[\alpha,\beta](x-\alpha) + f[\alpha,\beta,x](x-\alpha)(x-\beta). \qquad (1.7)$$

Demostración: Si $x = \alpha$ la fórmula no es más que la identidad $f(x) = f(x)$. Podemos suponer, pues, $x \neq \alpha$. Si $x = \beta$ esta fórmula no es más que una reescritura de la definición del símbolo $f[\alpha,\beta]$, mientras que si $x \neq \beta$ la fórmula es una reescritura de la definición del símbolo $f[\alpha,\beta,x]$. ■

Ahora usaremos la fórmula (1.7) para hallar el orden de convergencia del método de la secante. Sea ξ un cero de la función f. Poniendo $x = \xi$ en (1.7) tenemos:

$$0 = f(\alpha) + f[\alpha,\beta](\xi-\alpha) + f[\alpha,\beta,\xi](\xi-\alpha)(\xi-\beta)$$

de donde

$$\xi = \alpha - \frac{f(\alpha)}{f[\alpha,\beta]} - \frac{f[\alpha,\beta,\xi]}{f[\alpha,\beta]}(\xi-\alpha)(\xi-\beta)$$

Supongamos ahora que β y α son sucesivas aproximaciones $\beta = x_{n-1}$, $\alpha = x_n$ obtenidas al aplicar el método de la secante a la función f para hallar el cero ξ. Entonces $\xi - \beta = \xi - x_{n-1}$ y $\xi - \alpha = \xi - x_n$ son los sucesivos errores e_{n-1} y e_n, mientras que, según el ejercicio 1.17, $\alpha - \frac{f(\alpha)}{f[\alpha,\beta]} = x_n - \frac{f(x_n)}{f[x_n,x_{n-1}]} = x_{n+1}$, de donde deducimos la siguiente relación entre tres errores consecutivos:

$$e_{n+1} = -\frac{f[x_n, x_{n-1}, \xi]}{f[x_n, x_{n-1}]} e_n e_{n-1}. \qquad (1.8)$$

De esta relación junto con la hipótesis de que tanto la derivada primera como la derivada segunda de f en ξ son distintas de cero deducimos inmediatamente que el orden de convergencia del método de la secante es $p = (1+\sqrt{5})/2 = 1.618\ldots$, debido al siguiente teorema:

Teorema 3 *Sea $\{x_n\}$ una sucesión convergente al número ξ para la que los errores $e_n = \xi - x_n$ verifican $e_{n+1} = c_n e_n e_{n-1}$ donde las constantes c_n convergen a $c_\infty = \lim\{c_n\}$. Entonces $\{x_n\}$ tiene orden de convergencia $p = (1+\sqrt{5})/2 \simeq 1.618$ y tiene constante de error asintótico igual a $|c_\infty|^{\frac{1}{p}}$.*

Demostración: Tenemos que hallar el valor de p tal que los cocientes $y_n = |e_{n+1}|/|e_n|^p$ tengan límite distinto de cero, es decir, tal que $\lim\{y_n\} \neq 0$. Para conseguir esto buscamos el valor de p que hace que y_n pueda expresarse en términos de y_{n-1}. Supongamos que el valor de p es tal que para algún α se cumple

$$y_n = |c_n||e_n|^{1-p}|e_{n-1}| = |c_n|(y_{n-1})^\alpha. \qquad (1.9)$$

Entonces la sucesión de cocientes y_n satisfará la relación

$$y_{n+1} = |c_n| y_n^\alpha,$$

y el límite $c = \lim\{y_n\}$ verificará $c = (\lim\{|c_n|\})c^\alpha = |c_\infty|c^\alpha$, es decir, $c = |c_\infty|^{\frac{1}{1-\alpha}}$. Para calcular p y α sólo hace falta escribir (1.9) en la forma

$$y_n = |c_n||e_n|^{1-p}|e_{n+1}| = |c_n|\left(\frac{|e_n|}{|e_{n+1}|^p}\right)^\alpha$$

para deducir que p y α han de verificar $\alpha p = -1$ y $\alpha = 1 - p$. De aquí se deduce $c = |c_\infty|^{\frac{1}{p}}$ y $p^2 - p - 1 = 0$, de donde $p = (1 + \sqrt{5})/2$. ■

Corolario 2 *Supongamos que ξ es un cero de primer orden de la función f y que la derivada segunda de f en ξ es distinta de cero. Entonces suponiendo que el método de la secante aplicado a f converge a ξ, su orden de convergencia es $p = (1 + \sqrt{5})/2$ y su constante de error asintótico es*

$$\left|\frac{1}{2}\frac{f''(\xi)}{f'(\xi)}\right|^{\frac{1}{p}}$$

Demostración: Según el teorema anterior y la fórmula (1.8) sólo necesitamos demostrar

$$\lim_{n\to\infty} \frac{f[x_n, x_{n-1}, \xi]}{f[x_n, x_{n-1}]} \neq 0.$$

Pero usando el resultado del ejercicio 1.18, es inmediato comprobar que

$$\lim_{n\to\infty} \frac{f[x_n, x_{n-1}, \xi]}{f[x_n, x_{n-1}]} = \frac{f[\xi, \xi, \xi]}{f[\xi, \xi]} = \frac{1}{2}\frac{f''(\xi)}{f'(\xi)} \neq 0,$$

como queríamos demostrar. ■

1.2.4 Aceleración de la convergencia

Como hemos visto, si g es una función de iteración cuya sucesión de iteraciones tiene convergencia lineal al límite ξ entonces $g'(\xi) \neq 0$. Además para cada n existe η_n en el intervalo de extremos ξ y x_n tal que

$$\xi - x_{n+1} = g'(\eta_n)(\xi - x_n)$$

de donde

$$\begin{aligned}\xi\big(1 - g'(\eta_n)\big) &= x_{n+1} - g'(\eta_n)x_n \\ &= x_{n+1} - g'(\eta_n)x_{n+1} + g'(\eta_n)x_{n+1} - g'(\eta_n)x_n \\ &= \big(1 - g'(\eta_n)\big)x_{n+1} + g'(\eta_n)(x_{n+1} - x_n).\end{aligned}$$

De aquí obtenemos la siguiente expresión para el límite ξ:

$$\xi = x_{n+1} + \frac{g'(\eta_n)(x_{n+1} - x_n)}{1 - g'(\eta_n)} = x_{n+1} + \frac{x_{n+1} - x_n}{g'(\eta_n)^{-1} - 1}.$$

En consecuencia, si podemos aproximar $g'(\eta_n)$, podríamos utilizar esa aproximación en esta fórmula para obtener una aproximación de ξ de la que se puede esperar que sea más exacta que x_{n+1}.

Aceleración Δ^2 de Aitken.–

Una forma de aproximar $g'(\eta_n)$ nos viene sugerida por el hecho de que para todo n el cociente $(g(x_n) - g(x_{n-1}))/(x_n - x_{n-1})$ es igual al valor de la derivada de g en algún punto ζ_n entre x_n y x_{n-1}. Este punto ζ_n estará necesariamente próximo a η_n (al menos para valores grandes de n) ya que η_n se encuentra entre x_n y ξ y la sucesión $\{x_n\}$ converge a ξ (véase la figura 1.5). Así pues, teniendo en cuenta que $g(x_n) = x_{n+1}$ podemos usar la aproximación

$$g'(\eta_n) \approx g'(\zeta_n) = \frac{g(x_n) - g(x_{n-1})}{x_n - x_{n-1}} = \frac{x_{n+1} - x_n}{x_n - x_{n-1}}$$

para aproximar ξ mediante

$$\hat{x}_n = x_{n+1} + \frac{x_{n+1} - x_n}{r_n - 1} \tag{1.10}$$

con

$$r_n = \frac{1}{g'(\zeta_n)} = \frac{x_n - x_{n-1}}{x_{n+1} - x_n}. \tag{1.11}$$

Como se puede ver en la figura 1.5 esta corrección corresponde a una interpolación lineal de g en los puntos x_n y x_{n-1}. Este proceso de aceleración se conoce como aceleración Δ^2 de Aitken debido a que la fórmula (1.10) puede expresarse en términos de los operadores de incremento Δ y Δ^2 definidos de la siguiente forma:

$$\begin{aligned}\Delta x_n &= x_{n+1} - x_n \\ \Delta^2 x_n &= \Delta(\Delta x_n) = \Delta x_{n+1} - \Delta x_n = x_{n+2} - 2x_{n+1} + x_n.\end{aligned}$$

Ejercicio 1.19
Demostrar que la fórmula (1.10) es equivalente a

$$\hat{x}_n = x_{n+1} - \frac{(\Delta x_n)^2}{\Delta^2 x_{n-1}}.$$

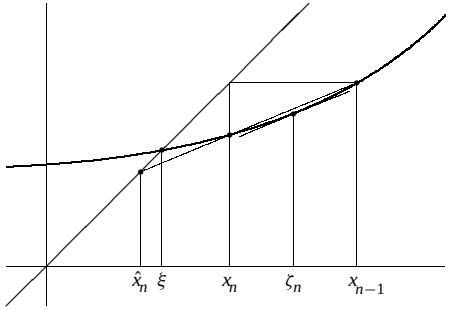


Figura 1.5: Aceleración Δ^2.

Se puede demostrar que al aplicar este proceso de aceleración a una sucesión cualquiera que tenga convergencia lineal se obtiene otra que tiene un orden de convergencia mayor.

Iteración de Steffensen.–

Cuando aplicamos la aceleración Δ^2 a la sucesión que resulta de un método iterativo con orden de convergencia lineal, una vez calculada la corrección $\hat{x}_n$, es mejor usar este valor en el siguiente paso de iteración, en lugar de usar x_{n+1}. De esta forma se llega al siguiente algoritmo que acelera un proceso iterativo lineal intercalando un paso de aceleración entre cada dos pasos de iteración:

Algoritmo de Iteración de Steffensen

1 **Datos:** g (la función de iteración)**,** x_0 (estimación inicial)**,** M (número máximo de pasos que queremos dar)**,** ϵ (precisión deseada).

2 $x = x_0$

3 **Para** $i = 1$ **hasta** M

$x_0 = x$

$x_1 = g(x_0)$**;** $x_2 = g(x_1)$**;**

$d = \Delta x_1 = x_2 - x_1$**;** $r = \Delta x_0/d = (x_1 - x_0)/d$

$x = x_2 + d/(r-1)$

Si $|d/(r-1)| < \epsilon$ **entonces ir a 6**

4 **Siguiente** i

5 **Imprimir** "No hubo convergencia después de"**;** M**;** "pasos" **y PARAR**

6 **Solución:** x**.**

1.3 Ecuaciones polinómicas

1.3.1 Introducción

Las ecuaciones polinómicas son aquellas ecuaciones $f(x) = g(x)$ en las que ambas funciones, f y g, son polinomios. Como la diferencia $f - g$ es de nuevo un polinomio, la resolución de las ecuaciones polinómicas se reduce al problema del cálculo de las raíces de polinomios.

Las raíces de los polinomios de segundo grado se obtendrán directamente mediante las conocidas fórmulas

$$\text{solución de menor valor absoluto} = \frac{-2c}{b + \text{sign}(b)\sqrt{b^2 - 4ac}},$$

$$\text{solución de mayor valor absoluto} = \frac{b + \text{sign}(b)\sqrt{b^2 - 4ac}}{-2a}.$$

Existen también fórmulas para las raíces de los polinomios de tercer y cuarto grado. En el caso de la ecuación cúbica existe siempre una raíz real, la cual puede hallarse de la siguiente forma:

$$\boxed{\text{Ecuación cúbica:} \quad x^3 + a_1x^2 + a_2x + a_3 = 0}$$

Se calculan primeramente las cantidades:

$$Q = \frac{3a_2 - a_1^2}{9}, \quad R = \frac{9a_1a_2 - 27a_3 - 2a_1^3}{54}, \quad D = Q^3 + R^2.$$

y entonces:

1. Si $D \geq 0$ una solución real es $x_1 = \sqrt[3]{R + \sqrt{D}} + \sqrt[3]{R - \sqrt{D}} - \frac{1}{3}a_1$.

2. Si $D < 0$ el cálculo se puede hacer mediante trigonometría:

$$x_1 = 2\sqrt{-Q}\cos\left[\tfrac{1}{3}\arccos\left(R/\sqrt{-Q^3}\right)\right] - \tfrac{1}{3}a_1.$$

Una vez hallada la raíz real x_1 es fácil reducir la cúbica a una ecuación cuadrática para hallar las restantes raíces: no hay más que dividir el polinomio de tercer grado por el binomio $x - x_1$.

La fórmula de las raíces de una ecuación cuártica nos las da en función de las raíces de una de tercer grado:

$$\boxed{\text{Ecuación cuártica:} \quad x^4 + a_1x^3 + a_2x^2 + a_3x + a_4 = 0}$$

Sea Y una solución real de la ecuación cúbica asociada

$$y^3 - a_2y^2 + (a_1a_3 - 4a_4)y + (4a_2a_4 - a_3^2 - a_1^2a_4) = 0$$

Entonces las soluciones de la cuártica son las raíces de las dos ecuaciones cuadráticas en z

$$z^2 + \tfrac{1}{2}\left(a_1 \pm \sqrt{a_1^2 - 4a_2 + 4Y}\right)z + \tfrac{1}{2}\left(Y \mp \sqrt{Y^2 - 4a_4}\right) = 0\,.$$

Un famoso trabajo de Galois demuestra que para los polinomios de grado superior al cuarto no existe ninguna fórmula algebraica que nos de sus raíces y por lo tanto es imprescindible recurrir en esos casos a los métodos numéricos. Es más, los métodos numéricos son incluso preferibles para polinomios de grado cuatro e incluso tres.

Por supuesto, las ecuaciones polinómicas pueden siempre resolverse por los métodos generales de resolución de ecuaciones no lineales que hemos visto hasta ahora. Sin embargo, para los polinomios existen métodos especiales que nos permiten mayor flexibilidad y eficacia. Estos métodos se basan en las propiedades especiales de estas funciones.

1.3.2 Propiedades generales de los polinomios

Cuando nos enfrentamos con la tarea de hallar las raíces de un polinomio es con frecuencia posible deducir información útil acerca de las raíces mediante un breve estudio preliminar del polinomio dado. En este estudio conviene tener en cuenta resultados generales acerca de los polinomios como son los siguientes:

Evaluación de un polinomio y de su derivada.–

El conocido algoritmo de la división de números puede efectuarse también con polinomios y aplicarlo al caso particular de la división de un polinomio cualquiera $p(x) = a_nx^n + \cdots + a_1x + a_0$ por un polinomio lineal de la forma $x - a$.

La demostración de que el resto de la división $p(x)/(x-a)$ es $p(a)$ se reduce fácilmente al caso de binomios mónicos:

Ejercicio 1.20
Demostrar que $x^n - a^n$ es divisible por $x - a$ siendo el cociente el polinomio

$$x^{n-1} + ax^{n-2} + \cdots + a^{n-2}x + a^{n-1}\,.$$

Deducir de ello que para todo polinomio $p(x)$ la diferencia $p(x) - p(a)$ es divisible por $x - a$, lo que significa que $p(a)$ es el resto al dividir $p(x)$ entre $x - a$.

Una demostración alternativa de este resultado es la siguiente. Sea $q(x) = b_{n-1}x^{n-1} + \cdots + b_1x + b_0$ el cociente y r el resto de la división de $p(x)$ entre $x - a$, de forma que

$$p(x) = (x - a)q(x) + r\,. \tag{1.12}$$

Queremos hallar los coeficientes b_k y r. Realizando la multiplicación obtenemos

$$\begin{aligned}(x-a)q(x) &= b_{n-1}x^n + \cdots + b_1x^2 + b_0x \\ &\quad - (ab_{n-1}x^{n-1} + \cdots + ab_1x + ab_0) \\ &= b_{n-1}x^n + (b_{n-2} - ab_{n-1})x^{n-1} \\ &\quad + \cdots + (b_0 - ab_1)x - ab_0\end{aligned}$$

Sumando r e igualando coeficientes con los de $p(x)$

$$\begin{aligned} b_{n-1} &= a_n \\ b_{n-2} - ab_{n-1} &= a_{n-1} \\ &\vdots \\ b_0 - ab_1 &= a_1 \\ r - ab_0 &= a_0\end{aligned}$$

de lo que deducimos que los coeficientes b_k satisfacen la relación recursiva

$$b_k = a_{k+1} + ab_{k+1}$$

(con la condición inicial $b_{n-1} = a_n$) y el resto es $r = a_0 + ab_0$, que podría ser llamado b_{-1}. En consecuencia el resto puede considerarse como el resultado final de un proceso iterativo que va produciendo los coeficientes b_k uno en cada paso. Por ejemplo, suponiendo que el grado es $n = 4$ el resto es

$$r = a_0 + a\Big(a_1 + a\big(a_2 + a(a_3 + aa_4)\big)\Big).$$

¡Pero esto no es más que la evaluación de $p(x)$ en $x = a$ por el método de Horner o de las multiplicaciones encajadas! De esto se deduce que el resto r es igual al valor $p(a)$ que el polinomio toma en a.

Vemos pues que

Los resultados intermedios en el proceso de evaluación de un polinomio $p(x)$ en a por el método de las multiplicaciones encajadas son los coeficientes del polinomio obtenido al dividir $p(x)$ entre $x - a$ y el resultado final, el valor del polinomio en a, $p(a)$, es el resto en dicha división.

Dichos resultados intermedios tienen una importancia adicional en la evaluación de la derivada de $p(x)$ en el punto a. De la expresión (1.12) deducimos

$$p'(x) = q(x) + (x - a)q'(x)$$

de donde

$$\boxed{p'(a) = q(a)}.$$

Es decir,

El valor $q(a)$ que toma en a el cociente $q(x)$ de dividir $p(x)$ entre $x - a$ es igual al valor que toma la derivada de p en el mismo punto.

Esto era de esperar ya que siendo

$$\frac{p(x) - p(a)}{x - a} = q(x),$$

por la definición de derivada

$$p'(a) = \lim_{x \to a} \frac{p(x) - p(a)}{x - a} = \lim_{x \to a} q(x) = q(a).$$

La consecuencia más importante de este resultado es que podemos hacer una sencilla modificación del algoritmo de Horner para que nos evalúe tanto el polinomio como su derivada en un punto dado:

Datos: $a_0, \ldots, a_n$ (coeficientes del polinomio p)**,** a
$c = a_n\,;\ b = a \cdot a_n + a_{n-1}$
Para $k = n - 2$ **hasta** 0 **paso** -1
  $c = a \cdot c + b$
  $b = a \cdot b + a_k$
siguiente k
Resultados: b (el valor de $p(a)$), c (el valor de $p'(a)$).

Raíces o ceros de polinomios.–

Puesto que el resto de la división de un polinomio $p(x)$ entre un polinomio lineal de la forma $x - a$ es la constante $p(a)$ que resulta al evaluar p en a, se puede concluir lo siguiente:

Ejercicio 1.21
Un número a es un cero o raíz de un polinomio $p(x)$ si y sólo si el polinomio $x - a$ es un factor de p.

Un polinomio *irreducible* es aquél que es distinto de cero y no es igual a un producto de polinomios de menor grado. Con este concepto podemos enunciar de forma precisa el teorema que tiene como corolario que, contando cada raíz tantas veces como indica su multiplicidad, "todo polinomio de grado n tiene n raíces":

Teorema 4 (Teorema fundamental del álgebra) *En el campo de los números complejos los polinomios irreducibles son precisamente los no nulos de grado menor o igual que uno, o sea: los polinomios lineales y las constantes no nulas.*

Nótese que parte del contenido de este teorema es trivial, pues *En cualquier campo de números los polinomios lineales y las constantes no nulas son irreducibles*.

Si una potencia $(x-a)^k$, $k > 0$, de un polinomio mónico lineal aparece como factor de un polinomio $p(x)$ entonces el número a es una raíz de p.

Si $(x-a)^k$ es pero $(x-a)^{k+1}$ no es un factor de p entonces se dice que la multiplicidad del factor $x-a$ en p es k y que la raíz a tiene *multiplicidad* k. En consecuencia el teorema anterior nos dice que todo polinomio de grado n con coeficientes complejos tiene n raíces complejas contando su multiplicidad.

Para polinomios con coeficientes reales conviene recordar que:

Lema 2 *Los polinomios irreducibles en el campo de los números reales son los de grado uno y aquellos de grado dos que tienen discriminante negativo.*

Ejercicio 1.22
Si $p(x)$ es un polinomio con coeficientes reales entonces para todo número complejo z, $p(\bar{z}) = \overline{p(z)}$ (p evaluado en el conjugado de z es el conjugado de $p(z)$).

Corolario 3 *Las raíces complejas no reales de un polinomio con coeficientes reales aparecen en pares conjugados. En otras palabras: si $z = a + bi$ es raíz de un polinomio $p(x)$ con coeficientes reales entonces también su conjugado, $\bar{z} = a - bi$, es raíz de p.*

De esto se deduce que todo polinomio de grado impar y con coeficientes reales tiene al menos una raíz real. Es interesante comparar la demostración de este hecho basada en los resultados anteriores (una demostración puramente "algebraica") con una demostración analítica del mismo hecho basada en la continuidad de los polinomios y el cambio de signo que ocurre en los polinomios de grado impar (cuando se considera un intervalo suficientemente grande) debido a que para n impar, $\lim_{x\to\pm\infty} x^n = \pm\infty$.

Regla de los signos de Descartes.–

La regla de los signos de Descartes nos permite limitar y a veces determinar el número de ceros reales positivos y el número de ceros reales negativos de un polinomio de coeficientes reales. Dice lo siguiente:

El número de raíces positivas de un polinomio de coeficientes reales restado del número de cambios de signo que ocurren en sus coeficientes da un número par no negativo.

Una consecuencia evidente de esto es:

Corolario 4 *Para todo polinomio en el que ocurra exactamente un cambio de signo entre sus coeficientes el número de raíces reales positivas es igual a uno.*

Para estudiar las raíces reales negativas de $p(x)$ basta estudiar el polinomio $q(x) = p(-x)$ ya que es evidente que si a es un cero positivo de $q(x)$ entonces $-a$ es un cero negativo de $p(x)$ de manera que el número de raíces negativas de $p(x)$ es igual al número de raíces positivas de $q(x)$.

Por ejemplo, consideremos el polinomio

$$p(x) = x^4 - x^3 - x^2 + x - 1$$

Puesto que tiene tres cambios de signo, el número de raíces positivas es un impar menor o igual que tres, esto es, uno o tres. Por otro lado, el polinomio

$$q(x) = p(-x) = x^4 + x^3 - x^2 - x - 1$$

sólo tiene un cambio de signo, por lo que $p(x)$ sólo tiene una raíz real negativa. Por tanto este polinomio tiene o bien tres raíces reales positivas y una negativa o bien una raíz real positiva, dos complejas conjugadas y una real negativa.

Acotaciones de las raíces.–

Existen varios teoremas que nos dan información acerca de la situación de los ceros de polinomios en el plano complejo.

Teorema 5 *Todos los ceros del polinomio $p(x) = a_n x^n + \cdots + a_1 x + a_0$ tienen módulo menor o igual que*

$$r = 1 + \max_{0\le k\le n-1} \left|\frac{a_k}{a_n}\right|$$

es decir, todos se encuentran en el disco de centro cero y radio r.

Por ejemplo, para el polinomio

$$p(x) = x^3 - x - 1$$

tenemos $\frac{a_2}{a_3} = 0$, $\frac{a_1}{a_3} = -1$, $\frac{a_0}{a_3} = -1$, luego

$$r = 1 + \max_{0 \le k \le n-1} \left| \frac{a_k}{a_n} \right| = 2$$

por lo que toda raíz x de p satisface $|x| \le 2$.

Teorema 6 *Todo polinomio $p(x) = a_n x^n + \cdots + a_1 x + a_0$ tiene al menos un cero cuyo módulo es menor que cualquiera de los números*

$$\rho_1 = n \left| \frac{a_0}{a_1} \right| , \quad \rho_n = \left| \frac{a_0}{a_n} \right|^{\frac{1}{n}}$$

es decir, p tiene al menos un cero dentro del disco del plano complejo de centro cero y radio $\min\{\rho_1, \rho_2\}$.

Para el polinomio del ejemplo anterior, $p(x) = x^3 - x - 1$, tenemos

$$\rho_1 = 3\left|\frac{-1}{-1}\right| = 3 \quad \text{y} \quad \rho_3 = \left|\frac{-1}{1}\right|^{\frac{1}{3}} = 1$$

por lo que p tiene al menos una raíz x tal que $|x| \le 1$.

Finalmente tenemos un teorema debido a Cauchy:

Teorema 7 *Dado un polinomio $p(x) = a_n x^n + \cdots + a_1 x + a_0$, si formamos los polinomios*

$$P(x) = |a_n| x^n - \cdots - |a_1| x - |a_0| ,$$
$$Q(x) = |a_n| x^n + \cdots + |a_1| x - |a_0| ,$$

dado que ambos tienen exactamente un cambio de signo, cada uno de ellos tiene exactamente un cero positivo y toda raíz de p tiene su módulo comprendido entre dichos ceros.

Reducción de grado.–

Cuando se quieren hallar todos los ceros de un polinomio $p(x)$ se puede emplear la técnica de *reducción de grado* que consiste en, hallada una raíz a, obtener el polinomio $q(x) = p(x)/(x - a)$. El grado de q es menor que el de p y sus raíces son las restantes raíces de p. A continuación se busca una raíz de q, lo que permite una nueva reducción de grado y así sucesivamente hasta llegar a un polinomio de segundo grado.

1.3.3 Algoritmo de Newton para ecuaciones polinómicas

Veremos a continuación un sencillo algoritmo en el que se adapta el método de Newton para cálculo de ceros de funciones al cálculo de una raíz de un polinomio. El método de Newton consiste en iterar la función $g(x) = x - f(x)/f'(x)$ a partir de una estimación del cero, en consecuencia nos resultará útil el algoritmo que hemos visto para la evaluación de un polinomio y de su derivada. Así llegamos al siguiente:

Algoritmo de Newton para Polinomios

1 **Datos:** n (grado del polinomio $p(x)$); $a_0, \ldots, a_n$ (coeficientes del polinomio p); a (estimación inicial del cero); ϵ (tolerancia de error); M (máximo número de iteraciones)

2 **Para** $i = 1$ **hasta** M

$c = a_n$; $b = a \cdot a_n + a_{n-1}$

Para $k = n - 2$ **hasta** 0 **paso** -1

$c = a \cdot c + a_k$

$b = a \cdot b + a_k$

Siguiente k

$d = b/c$

Si $|d| \le \epsilon$ **entonces ir a 5**

$a = a - d$

3 **Siguiente** i

4 **Imprimir "**No hubo convergencia después de **";** M**; "** pasos.**"** **y PARAR**

5 **Imprimir "**Solución es: **";** $a - d$**; "."** **y PARAR**

1.3.4 Ecuaciones polinómicas mal condicionadas

En esta sección vamos a estudiar cómo cambian las raíces de un polinomio por el hecho de cometer pequeños errores en sus coeficientes. Supongamos que $p(x)$ es un polinomio de grado n con coeficientes reales y con raíces simples $r_1, \ldots, r_n$. Vamos a suponer también que producimos una alteración en los coeficientes de p de la forma

$$P(x, \epsilon) = p(x) + \epsilon q(x)$$

donde $q(x)$ es otro polinomio de grado n.

Sean $z_1(\epsilon), \ldots, z_n(\epsilon)$ las funciones que nos dan las raíces de P (como polinomio en x) en función de ϵ. Evidentemente para cada $i = 1, \ldots, n$ es $z_i(0) = r_i$. En estas condiciones puede demostrarse que estas funciones z_i son funciones diferenciables (incluso analíticas) en algún entorno del origen $\epsilon = 0$ y en consecuencia para ϵ pequeño podemos aproximar $z_i(\epsilon) \simeq z_i(0) + \epsilon z_i'(0) = r_i + \epsilon z_i'(0)$ de donde podemos estimar el cambio sufrido por las raíces de p mediante:

$$|z_(\epsilon) - r_i| \simeq |\epsilon||z_i'(0) \tag{1.13}$$

Así, sólo nos falta determinar los valores de las derivadas $z_i'(0)$. Para ello hacemos lo siguiente: Sabemos que

$$0 = P(z_i(\epsilon), \epsilon) = p(z_i(\epsilon)) + \epsilon q(z_i(\epsilon))$$

de donde, derivando respecto a ϵ,

$$0 = p'(z_i(\epsilon))z_i'(\epsilon) + q(z_i(\epsilon)) + \epsilon q'(z_i(\epsilon))z_i'(\epsilon)$$

y despejando $z_i'(\epsilon)$,

$$z_i'(\epsilon) = -\frac{q(z_i(\epsilon))}{p'(z_i(\epsilon)) + \epsilon q'(z_i(\epsilon))}.$$

Por la hipótesis de que todas las raíces son simples, podemos evaluar estas derivadas en $\epsilon = 0$ con lo que obtenemos

$$z_i'(0) = -\frac{q(r_i)}{p'(r_i)}.$$

En consecuencia, la estimación (1.13) del cambio sufrido por las raíces de $p(x)$ nos da

$$|z_i(\epsilon) - r_i| \simeq \frac{|\epsilon q(r_i)|}{|p'(r_i)|}.$$

De esta estimación aprendemos dos cosas: Primeramente, como cabe esperar para funciones en general, cuando la derivada es pequeña en valor absoluto cerca o en un cero la búsqueda de ese cero es un problema mal condicionado. La segunda cosa que aprendemos es que pequeños errores en un coeficiente de una potencia alta de x en $p(x)$ puede afectar seriamente el cálculo de las raíces de valor absoluto grande (tómese $q(x) = x^n$ y supongamos que r_i tiene $|r_i|$ grande y se verá que $|\epsilon q(r_i)|/|p'(r_i)|$ puede ser muy grande).

1.3.5 El método de Bernoulli

Vamos a terminar este tema de raíces de polinomios con una breve referencia a un antiguo método atribuido a Daniel Bernoulli (de la famosa familia Bernoulli de matemáticos) que (en su versión más sencilla) sirve para aproximar la raíz real de mayor valor absoluto de un polinomio.

El estudiante conoce seguramente el método para obtener una fórmula para el término general de una sucesión $\{x_n\}$ que satisface una ecuación de recurrencia de la forma $a_2x_n + a_1x_{n-1} + a_0x_{n-2} = 0$ (donde $a_2 \neq 0$). Claramente, para a_0, a_1, a_2 dados, el conjunto de tales sucesiones forma un espacio vectorial de dos dimensiones (pues cada sucesión está completamente determinada por sus dos términos iniciales x_0, x_1). En consecuencia la forma general de una tal sucesión está dada por una combinación lineal de dos de ellas que sean linealmente independientes.

Por otro lado, supongamos que $\{x_n\}$ es una sucesión de ese espacio vectorial, es decir, que satisface la relación de recurrencia dada. Entonces es muy fácil de ver que en caso de que la sucesión de cocientes sucesivos $y_n = x_{n+1}/x_n$ converja, su límite necesariamente será una raíz del polinomio $a_2x^2 + a_1x + a_0$. Este hecho es la base del método de Bernoulli, el cual consiste en construir de forma más o menos arbitraria una sucesión que satisfaga la recurrencia e intentar calcular su límite (con la esperanza de que exista). A continuación veremos una situación en que la existencia de dicho límite está garantizada.

Ejercicio 1.23
Si z es una raíz real del polinomio $a_2x^2 + a_1x + a_0$ entonces la sucesión $x_n = z^n$ satisface la relación de recurrencia $a_2x_n + a_1x_{n-1} + a_0x_{n-2} = 0$. Si z_1, z_2 son dos raíces reales no nulas distintas del mismo polinomio entonces las dos sucesiones z_1^n, z_2^n (satisfacen aquella relación de recurrencia y) son linealmente independientes

Ejercicio 1.24
Si el polinomio $a_2x^2 + a_1x + a_0$ tiene dos raíces reales distintas z_1, z_2 entonces la forma general de las sucesiones que satisfacen la relación de recurrencia $a_2x_n + a_1x_{n-1} + a_0x_{n-2} = 0$ es

$$x_n = c_1z_1^n + c_2z_2^n$$

donde c_1 y c_2 son dos constantes arbitrarias.

La importancia del ejercicio 1.24 consiste en que de él se deduce lo siguiente. Supongamos que $|z_1| > |z_2|$. Entonces podemos poner

$$\frac{x_{n+1}}{x_n} = \frac{c_1 z_1^{n+1} + c_2 z_2^{n+1}}{c_1 z_1^n + c_2 z_2^n} = \frac{z_1^{n+1}}{z_1^n} \frac{c_1 + c_2 (z_2/z_1)^{n+1}}{c_1 + c_2 (z_2/z_1)^n} = z_1 \frac{c_1 + c_2 (z_2/z_1)^{n+1}}{c_1 + c_2 (z_2/z_1)^n}.$$

Esto implica (por ser $\lim_{n\to\infty} (z_2/z_1)^n = 0$) que, si $c_1 \neq 0$, la sucesión $y_n = x_{n+1}/x_n$ tiene como límite la raíz de $a_2x^2 + a_1x + a_0$ de mayor valor absoluto. Y esto se cumple cualquiera que sea la sucesión $\{x_n\}$ que satisface la relación de recurrencia asociada con el polinomio cuadrático $a_2x^2 + a_1x + a_0$, con tal que x_1/x_0 no sea ya una raíz. Pero es muy sencillo construir sucesiones que satisfacen dicha recurrencia; no hay más que elegir "arbitrariamente" (pero juiciosamente también) los dos términos iniciales y los demás quedan determinados por la propia relación de recurrencia. Así pues, llegamos al siguiente método de estimar la raíz de mayor valor absoluto de un polinomio cuadrático:

Proposición 2 *Si $p(x) = a_2x^2 + a_1x + a_0$ es un polinomio de coeficientes reales con dos raíces reales distintas z_1 y z_2 y $|z_1| > |z_2|$ entonces para cualesquiera valores iniciales x_0, x_1 tales que x_1/x_0 no es raíz de $p(x)$, la sucesión $\{x_n\}$ definida recurrentemente por*

$$x_n = -(a_1x_{n-1} + a_0x_{n-2})/a_2$$

tiene la propiedad

$$\lim_{n\to\infty} \frac{x_{n+1}}{x_n} = z_1$$

Todo lo que acabamos de decir para ecuaciones cuadráticas vale también para ecuaciones polinómicas de cualquier grado. Además no es nada más difícil demostrar el caso general:

Teorema 8 (Método de Bernoulli) *Si $z_1, \ldots, z_m$ (ordenados por sus valores absolutos, es decir $|z_1| > \cdots > |z_m|$) son m raíces reales de un polinomio de grado m con coeficientes reales $a_0, \ldots, a_m$ (es decir, si $\sum_{k=0}^m a_k z_i^k = 0$) entonces la forma general de una sucesión $\{x_n\}$ que satisfaga la relación de recurrencia $\sum_{k=0}^m a_k x_{n+k-m} = 0$ es $x_n = \sum_{i=1}^m c_i z_i^n$ y para cualquier tal sucesión con $c_1 \neq 0$ se verifica*

$$\lim_{n\to\infty} \frac{x_{n+1}}{x_n} = z_1.$$

Una tal sucesión se puede obtener mediante la fórmula

$$x_n = \sum_{k=0}^{m-1} a_k x_{n+k-m}$$

y una elección juiciosa de los m términos iniciales $x_0, \ldots, x_{m-1}$.

El método de Bernoulli, como método general para hallar raíces de polinomios tiene muchas dificultades, especialmente en su forma simple explicada aquí. Incluso en los casos en los que es aplicable para hallar la raíz de mayor valor absoluto, la sucesión de aproximaciones suele converger muy despacio. Evidentemente es un método que tiene que ir acompañado de la técnica de reducción de grado explicada más arriba. Con todo, cuando aplicable, es un método muy sencillo que se puede utilizar como paso previo a otro más eficaz con la intención de proporcionar una estimación inicial de la raíz buscada.

Hay que decir también que este método ha sido objeto de exhaustivas investigaciones y con modificaciones apropiadas puede utilizarse incluso para estimar pares complejo-conjugados de raíces en algunos casos. Una variante moderna de este método (el algoritmo QD o de *cociente-diferencia*) resuelve muchas de las dificultades del algoritmo de Bernoulli y proporciona estimaciones de todas las raíces simultáneamente (incluyendo raíces complejas).

1.4 El Método de Müller

Ya hemos indicado en la sección 1.1.4 la idea que hay tras el método de Müller y su semejanza con el método de la secante. Como aquél, es un método que no requiere la evaluación de la derivada y que converge casi cuadráticamente. Además tiene la ventaja de que no necesita una estimación inicial precisa, aunque sí es necesario elegir no dos sino tres valores iniciales.

Aunque en lo que sigue se supone tácitamente que buscamos un cero real y que las variables son todas reales, todo lo que se dice vale sin cambios para el caso complejo. Esta es una de las ventajas que hacen del método de Müller una herramienta excelente a la hora de buscar ceros de funciones, pues no tenemos que preocuparnos a priori de si el cero buscado es real o complejo.

Dados tres puntos a, b, c en la recta real, comenzamos por calcular el polinomio cuadrático, $q(x)$, que coincide con la función dada, f, en esos tres puntos, es decir, tal que $q(a) = f(a)$, $q(b) = f(b)$, $q(c) = f(c)$.

Una vez hallado dicho polinomio cuadrático $q(x)$ calculamos, de sus dos raíces, aquella que más se acerque al cero buscado. Esa raíz será el siguiente término de la sucesión que converge al cero buscado. Así pues el método de Müller progresa de la siguiente forma: Conocidos tres términos consecutivos de la sucesión, x_{n-2}, x_{n-1}, x_n, el siguiente término se obtiene haciendo lo siguiente:

1. Primeramente hacemos $a = x_{n-2}$, $b = x_{n-1}$, $c = x_n$.
2. Calculamos el polinomio $q(x)$ determinado por $q(a) = f(a)$, $q(b) = f(b)$, $q(c) = f(c)$.
3. Hallamos el cero de $q(x)$ más cercano a la solución buscada, sea este cero x, y ponemos $x_{n+1} = x$.
4. Por último preparamos los datos para el siguiente paso haciendo $a = b$, $b = c$, $c = x$ y vamos al paso 2.

Para elegir cuál de las dos posibles soluciones de $q(x) = 0$ nos interesa, razonamos así: a medida que el método progresa los tres puntos a, b, c estarán más y más cerca del cero y por tanto más cerca entre sí, lo que implica que la parábola representada por $y = q(x)$ se aproximará más y más a una línea recta que pasa cerca del cero buscado. Por lo tanto, mientras una raíz de $q(x)$ se acerca a la solución buscada la otra tenderá a $+\infty$ o a $-\infty$. Esto nos dice que la raíz de $q(x)$ que debemos desechar es la de mayor valor absoluto. Otra forma de plantear esto es buscar la ecuación cuadrática satisfecha por la diferencia $x - x_n$, es decir, expresando la ecuación $q(x) = 0$ en la forma $p(x - x_n) = 0$ (ecuación (1.14)). En este caso de nuevo buscaríamos la solución de menor valor absoluto ya que a medida que progresa el método esta diferencia irá acercándose a cero.

Este proceso continúa hasta que alcancemos un error (medido en términos de las correcciones) menor que el máximo tolerado.

La forma más sencilla de hallar el polinomio $q(x)$ es expresándolo en la forma

$$q(x) = y_3 + A(x - x_3) + B(x - x_2)(x - x_3)$$

así se garantiza que pasa por (x_3, y_3) y sólo tenemos que calcular dos coeficientes, los cuales se obtienen muy fácilmente. De $q(x_2) = y_2$ deducimos que A vale

$$\boxed{A = \frac{y_2 - y_3}{x_2 - x_3}}$$

y de $q(x_1) = y_1$ deducimos

$$y_1 = y_3 + A(x_1 - x_3) + B(x_1 - x_2)(x_1 - x_3)$$

de donde

$$\boxed{B = \frac{\frac{y_1 - y_3}{x_1 - x_3} - A}{x_1 - x_2}}.$$

Para hallar el cero que nos interesa expresaremos $q(x)$ en la forma

$$q(x) = y_3 + C(x - x_3) + D(x - x_3)^2, \tag{1.14}$$

cuyos coeficientes se determinan también fácilmente. En primer lugar, por ser igual al coeficiente principal, ha de ser $D = B$. Por otro lado el coeficiente C ha de verificar $C + (x - x_3)B = A + (x - x_2)B$, de donde

$$\boxed{C = A + B(x_3 - x_2)}$$

y la corrección que tenemos que hacer sobre x_3 es

$$\boxed{x - x_3 = \frac{-2y_3}{C + \text{sign}(C)\sqrt{C^2 - 4y_3 B}}}.$$

Como dijimos más arriba, este método sirve para obtener tanto los ceros reales como los ceros complejos no reales de una función.

En el caso de que el cero buscado sea real, y a pesar de que todas las variables involucradas sean reales, los cálculos intermedios pueden producir

ocasionalmente valores complejos no reales por obtenerse un discriminante negativo en la resolución de la ecuación cuadrática.

Esas salidas del eje real no influyen en la aproximación al cero, pero puede ser conveniente evitarlas (en el caso de que el cero buscado se sepa con seguridad que es real) especialmente si se realizan los cálculos en un lenguaje de programación que no admite aritmética compleja.

En tal situación puede evitarse que aparezcan en los cálculos cantidades complejas no reales sin más que evitar el cálculo de la raíz cuadrada $\sqrt{C^2 - 4y_3B}$ cuando $C^2 - 4y_3B < 0$. Ésta es la técnica empleada en el siguiente algoritmo que implementa el método de Müller para el cálculo de ceros reales usando sólo aritmética real.

Algoritmo del Método de Müller

(ceros reales)

1 **Datos:** x_1, x_2, x_3 (valores iniciales)**,** f (función de la que se desea un cero)**,** ϵ **(tolerancia de error),** M (máximo número de pasos del algoritmo)**.**

2 $h = x_3 - x_2$**;** $y_1 = f(x_1)$**;** $y_2 = f(x_2)$**:**

3 **Para** $k = 1, M$

$y_3 = f(x_3)$**;** $A = (y_3 - y_2)/h$**;**

$B = \frac{y_1 - y_3}{x_1 - x_3}$**;** $B = B - A$**;** $B = B/(x_1 - x_2)$**;**

$C = A + Bh$**;** $Q = C^2 - 4y_3B$

Si $Q < 0$ **entonces** $Q = 0$

$h = -2y_3/(C + \text{sign}(C)\sqrt{Q})$**;**

Si $|h| < \epsilon$ **entonces ir a 6**

$x_1 = x_2$**;** $x_2 = x_3$**;** $x_3 = x_3 + h$**;** $y_1 = y_2$**;** $y_2 = y_3$**;**

Imprimir x_3

4 **Siguiente** k

5 **Imprimir "**No hubo convergencia después de **";** M**; "** pasos.**" y PARAR**

6 **Imprimir "**Solución es: **";** x_3**; "." y PARAR**

Naturalmente sería necesario modificar este algoritmo para que encontrase ceros complejos no reales. Para ello sería suficiente eliminar la línea

Si $Q < 0$ **entonces** $Q = 0$

siempre y cuando se pudiesen realizar los cálculos en aritmética de números complejos. Si se quieren buscar ceros complejos usando solamente aritmética real sería necesario hacer más modificaciones; principalmente en el cálculo de $h = -2y_3/(C + \text{sign}(C)\sqrt{Q})$. Dejamos este trabajo al estudiante trabajador.

En el siguiente ejercicio se busca un cero complejo de la función Coseno Integral de Fresnel, cuyo único cero real es el obvio $x = 0$.

Ejercicio 1.25
Utilizar el método de Müller para hallar el cero no nulo de menor módulo de la función Coseno Integral de Fresnel:

$$C(x) = \sum_{n=0}^{\infty} \frac{(-1)^n (\frac{\pi}{2})^{2n}}{(2n)!(4n+1)} x^{4n+1}.$$

(Sugerencias: Evaluar C truncando su serie de potencias en un valor de n par (por ejemplo, en $n = 2$ con lo cual $C(x) \simeq a_0 x + a_1 x^5 + a_2 x^9$). Con ello se evitará la aparición de un cero real espurio. Dividir por x para obviar la solución $x = 0$. Evaluar cada coeficiente a_k en términos de a_{k-1} para mejorar la precisión, y evaluar el polinomio resultante por el método de las multiplicaciones encajadas. Comenzar las iteraciones con $x_1 = 1.6$, $x_2 = 1.7$, y $x_3 = 1.8$, lo cual permitirá obtener dos dígitos de precisión en cinco o seis iteraciones. Ir aumentando progresivamente el número de términos de la serie (siempre truncando en n par) para ir aumentando la precisión de la solución.)

Respuestas a Algunos Ejercicios del Capítulo 1

Ejercicio 1.1 (Esto está hecho para otros datos.) Con N pasos a partir de un intervalo de longitud l se obtiene un intervalo de longitud $l/2^N$. El punto medio de este intervalo tendrá un error menor que el radio, o sea, menor que $\frac{1}{2}l/2^N$ (el error no puede ser igual al radio porque eso significaría que el cero es uno de los extremos, lo que habría sido detectado por el algoritmo y se hubiera hallado el cero exacto). Para que dicho error sea menor que ϵ necesitamos que el número de pasos, N, sea tal que $2^N \geq \frac{1}{2}l/\epsilon$, es decir N mayor o igual que $\log(l/2\epsilon)/\log 2 = \log(l/\epsilon)/\log 2 - 1$, que es lo mismo que decir

$$N = \begin{cases} \text{ent}(\log(l/\epsilon)/\log 2) & \text{si frc}(\log(l/\epsilon)/\log 2) \neq 0 \\ \log(l/\epsilon)/\log 2 - 1 & \text{si frc}(\log(l/\epsilon)/\log 2) = 0. \end{cases}$$

En nuestro caso ($l = 20$ y $\epsilon = 10^{-6}$) $\log(l/\epsilon)/\log 2 = \log(2{\cdot}10^7)/\log 2 = 1 + 7/\log 2 = 24.25\ldots$ luego $N = 24$.

Ejercicio 1.2 La pendiente de la recta es $\frac{f(a)}{a-x_0} = \frac{f(b)-f(a)}{b-a}$. despejando x_0 se obtiene

$$x_0 = a - f(a)\frac{b-a}{f(b)-f(a)} = \frac{af(b)-bf(a)}{f(b)-f(a)}\,.$$

Ejercicio 1.3 En realidad dicho promedio ponderado es también igual a $\frac{-af(b)+bf(a)}{-f(b)+f(a)}$. Ambas expresiones se obtienen al tener en cuenta que $f(a)f(b) < 0$ lo cual implica que o bien $|f(a)| = f(a)$ y $|f(b)| = -f(b)$ o bien $|f(a)| = -f(a)$ y $|f(b)| = f(b)$.

Ejercicio 1.4 Sea $a = f(x)$. Calculemos $f(a)$. $f(a) = f(f(x)) = f^2(x) = f(x) = a$. Luego a es un punto fijo. Recíprocamente, sea a un punto fijo de f. Entonces tomando $x = a$ se verifica que $a = f(x)$

Ejercicio 1.5 $\xi = \lim\{x_n\} = \lim\{g(x_{n-1})\} = g(\lim\{x_{n-1}\}) = g(\xi)$.

Ejercicio 1.6 Sea ξ el punto fijo, entonces $\xi = g(\xi) = \xi - \frac{f(\xi)}{f'(\xi)}$. Esto implica $f(\xi) = 0$.

Ejercicio 1.7 Sea a dicha abscisa. Entonces $f'(x) = f(x)/(x-a)$, de donde se deduce inmediatamente la fórmula dada.

Ejercicio 1.8 Como $f(x) = x^2 - N$ y $f'(x) = 2x$ la función de iteración del método de Newton es $g(x) = x - \frac{x^2-N}{2x} = x - \frac{x}{2} + \frac{N}{2x} = \frac{1}{2}(x + \frac{N}{x})$. Para la ecuación $x^3 - N = 0$, $f'(x) = 3x^2$ y la función de iteración es $g(x) = x - \frac{x^3-N}{3x^2} = x - \frac{2x}{3} + \frac{N}{3x^2} = \frac{1}{3}(2x + \frac{N}{x^2})$. Luego el algoritmo es: $x_{n+1} = \frac{1}{3}(2x_n + N/x_n^2)$.

Ejercicio 1.9 (a) Para g_1: si $x = \sqrt{2+x}$ entonces $x^2 = 2 + x$, que es la ecuación dada. (b) Para g_2: si $x = 1 + \frac{2}{x}$ entonces $x^2 = x + 2$, que es la ecuación dada. (c) Para g_3: si $x = x - \frac{x^2-x-2}{m}$ entonces $0 = -\frac{x^2-x-2}{m}$, de donde $0 + x^2 - x - 2$ que es la ecuación dada. (d) Para g_4: si $x = \frac{x^2+2}{2x-1}$ entonces $x(2x-1) = x^2+2$, de donde $2x^2 - x = x^+2$, que es equivalente a la ecuación dada.

Ejercicio 1.10 Si $g(a) = a$ o $g(b) = b$ ya tenemos un punto fijo. En caso contrario, puesto que $I \xrightarrow{g} I$, tenemos $a < g(a) < b$, y $a < g(b) < b$, lo que implica que la función $h(x) = g(x) - x$ es positiva en a y negativa en b, además de ser continua en I. En consecuencia h tiene un cero en I el cual es un punto fijo de g.

Ejercicio 1.11 Como $g'(x) = 1/(2\sqrt{2+x})$, para $x > 0$ se cumple $|g'(x)| < 1/\sqrt{8} < 1$, luego la condición (c) se cumple para todo $x > 0$. Como g es creciente, para que la condición (a) se cumpla en $[a, b]$ basta que sea $g(a) \geq a$ y $g(b) \leq b$. Lo primero se cumple para $a = 0$ y lo segundo es $\sqrt{2+b} \leq b$ o $2 + b \leq b^2$ o $b \geq 2$. Luego podemos tomar $I = [0, 3]$.

Ejercicio 1.12

Ejercicio 1.13 La constante de error es igual a

$$\frac{1}{n!}|g^{(n)}(\xi)|\,.$$

Ejercicio 1.14 La derivada segunda de g es $\big(f'(x)^2 f''(x) + f(x)f'''(x) - 2f(x)f''(x)^2\big)/f'(x)^3$. De aquí, teniendo en cuenta que $f(\xi) = 0$, se obtiene $g''(\xi) = f''(\xi)/f'(\xi)$.

Ejercicio 1.15 (a)

$$g'(\xi) = \lim_{x\to\xi} \frac{f(x)\,f''(x)}{f'(x)^2} = \lim_{x\to\xi} \frac{f(x)}{f'(x)^2} \lim_{x\to\xi} f''(x)$$
$$= f''(\xi) \lim_{x\to\xi} \frac{f'(x)}{2\,f'(x)\,f''(x)} = f''(\xi)\frac{1}{2\,f''(\xi)} = \frac{1}{2}.$$

(b) $g'(x) = 1 - 2\frac{[f'(x)]^2 - f(x)\,f''(x)}{[f'(x)]^2} = 1 - 2\frac{[f'(x)]^2}{[f'(x)]^2} + 2\frac{f(x)\,f''(x)}{[f'(x)]^2} = -1 + 2\frac{f(x)\,f''(x)}{[f'(x)]^2}$. Para hallar $g'(\xi)$ calculamos el límite

$$\lim_{x\to\xi} g'(x) = -1 + 2\lim_{x\to\xi} \frac{f(x)\,f''(x)}{[f'(x)]^2} = -1 + 2\lim_{x\to\xi} \frac{f(x)}{[f'(x)]^2} \lim_{x\to\xi} f''(x)$$
$$= -1 + 2\left(\lim_{x\to\xi} \frac{f(x)}{[f'(x)]^2}\right) f''(\xi) = -1 + 2\left(\lim_{x\to\xi} \frac{f'(x)}{2\,f'(x)\,f''(x)}\right) f''(\xi)$$
$$= -1 + 2\left(\lim_{x\to\xi} \frac{1}{2\,f''(x)}\right) f''(\xi) = -1 + 2\frac{1}{2\,f''(\xi)}\, f''(\xi) = 0$$

de donde efectivamente $g'(\xi) = 0$. Ahora tenemos que demostrar que g'' es continua en $x = \xi$. Para ello hemos de ver que $\lim_{x\to\xi} g''(x)$ es finito. Primero hallamos

$$g'' = \left(-1 + 2\frac{ff''}{[f']^2}\right)' = 2\frac{(f'f'' + ff''')f' - 2ff''f''}{[f']^3}$$
$$= 2\frac{[f']^2 f'' + ff'''f' - 2f[f'']^2}{[f']^3} = 2\frac{([f']^2 - 2ff'')f'' + ff'''f'}{[f']^3}$$
$$= 2\frac{([f']^2 - 2ff'')}{[f']^3} f'' + 2\frac{f}{[f']^2} f'''$$

y ahora, aplicando la regla de l'Hôpital y el hecho de que

$$\lim_{x\to\xi} \frac{f(x)}{[f'(x)]^2} = \lim_{x\to\xi} \frac{f'(x)}{2\,f'(x)\,f''(x)} = \frac{1}{2\,f''(\xi)},$$

obtenemos,

$$\lim_{x\to\xi} g''(x) = 2\lim_{x\to\xi} \frac{([f'(x)]^2 - 2f(x)\,f''(x))'}{3f'(x)^2 f''(x)} \lim_{x\to\xi} f''(x) + 2\frac{f'''(\xi)}{2\,f''(\xi)}$$
$$= 2\lim_{x\to\xi} \frac{2f'(x)\,f''(x) - 2f'(x)\,f''(x) - 2f(x)\,f'''(x)}{3f'(x)^2} + \frac{f'''(\xi)}{f''(\xi)}$$
$$= \frac{f'''(\xi)}{f''(\xi)} - \tfrac{4}{3} f'''(\xi) \lim_{x\to\xi} \frac{f(x)}{[f'(x)]^2} = \frac{f'''(\xi)}{f''(\xi)} - \tfrac{2}{3}\frac{f'''(\xi)}{f''(\xi)} = \tfrac{1}{3}\frac{f'''(\xi)}{f''(\xi)}$$

Ejercicio 1.16

Ejercicio 1.17

Ejercicio 1.18 Supongamos primeramente que $a = c$. Entonces:

$$\frac{f[a,b] - f[a,c]}{b-c} = \frac{f[a,c] - f[a,b]}{c-b} = \frac{f[a,c] - f[c,b]}{a-b} = \frac{f[a,c] - f[b,c]}{a-b}$$

donde hemos usado la propiedad obvia $f[c,b] = f[b,c]$. Supongamos ahora que $a \neq c$. Tenemos que demostrar

$$(f[a,b] - f[a,c])(a-b) = (f[a,c] - f[b,c])(b-c),$$

que es equivalente a

$$f(a) - f(b) - f[a,c]a + f[a,c]b = f[a,c]b - f[a,c]c - f(b) + f(c)$$

y cancelando el término $-f(b) + f[a,c]b$ y operando, el problema queda reducido a probar

$$f(a) - f[a,c]a = -f[a,c]c + f(c),$$

lo cual no es más que una reescritura de la definición del símbolo $f[a,c]$.

Ejercicio 1.19 Esto es un simple cálculo:

$$\hat{x}_n = x_{n+1} + \frac{x_{n+1} - x_n}{\frac{x_n - x_{n-1}}{x_{n+1} - x_n} - 1} = x_{n+1} + \frac{(x_{n+1} - x_n)^2}{x_n - x_{n-1} - (x_{n+1} - x_n)}$$
$$= x_{n+1} + \frac{(\Delta x_n)^2}{x_n - x_{n-1} - x_{n+1} + x_n} = x_{n+1} - \frac{(\Delta x_n)^2}{x_{n-1} + x_{n+1} - 2x_n}$$
$$= x_{n+1} - \frac{(\Delta x_n)^2}{\Delta^2 x_{n-1}}.$$

Ejercicio 1.20

Ejercicio 1.21 Si a es una raíz de p entonces al dividir $p(x)$ entre $x-a$ el resto es $p(a) = 0$ y por lo tanto $x-a$ es un factor de $p(x)$. Por otro lado, es evidente que si $x-a$ es un factor de p, $p(x) = (x-a)q(x)$ para algún otro polinomio $q(x)$ y entonces a es un cero de p: $p(a) = (a-a)q(a) = 0$.

Ejercicio 1.22 $p(\bar{z}) = a_n\bar{z}^n+\cdots+a_1\bar{z}+a_0 = a_n\overline{z^n}+\cdots+a_2\overline{z^2}+a_1\bar{z}+a_0$. Teniendo en cuenta que los coeficientes son reales y por tanto verifican $\bar{a_k} = a_k$ deducimos

$$p(\bar{z}) = \overline{a_n z^n + \cdots + a_1 z + a_0} = \overline{p(z)}\,.$$

Ejercicio 1.23 La satisfacción de la relación de recurrencia es una comprobación inmediata. La independencia lineal de $\{z_1^n\}$ y $\{z_2^n\}$ se reduce a la no singularidad de la matriz

$$\begin{pmatrix} 1 & z_1 \\ 1 & z_2 \end{pmatrix},$$

la cual es equivalente a $z_1 \neq z_2$.

Ejercicio 1.24

Ejercicio 1.25 $z = 1.7437 + 0.3057i$

Capítulo 2

Resolución de Sistemas de Ecuaciones Lineales

2.1 Métodos directos

La *regla de Cramer* es un método teórico de resolución directa de sistemas de ecuaciones lineales, sin embargo no lo consideramos como un método numérico por la complejidad de cálculo que conlleva. Para un sistema de orden n la regla de Cramer requiere el cálculo de $n+1$ determinantes de orden n, lo que significa realizar $(n+1)!(n-1)$ multiplicaciones: un número de operaciones excesivamente grande comparado con el requerido por los métodos que vamos a estudiar.

2.1.1 Eliminación de Gauss

El método de eliminación de Gauss es el método más sencillo de resolución de sistemas de ecuaciones lineales. La única diferencia entre este método y lo que suele hacer un estudiante principiante está en que el método de Gauss es una forma *ordenada y sistemática* de realizar la eliminación de las incógnitas, lo que da lugar a un algoritmo sencillo y eficaz, que se puede programar. Este algoritmo puede aplicarse tanto a sistemas sobredeterminados como a sistemas incompletos, pero en lo que sigue supondremos que

el sistema dado es no-singular y tiene tantas ecuaciones como incógnitas. El algoritmo consta de dos partes: (1) la *eliminación* propiamente dicha, que consiste en reducir el sistema a uno triangular superior (cuya matriz de coeficientes tiene todo ceros debajo de la diagonal); y (2) *sustitución* u obtención de la solución por el método llamado de *sustitución regresiva*. Veamos a continuación cada una de estas dos partes ilustradas con un ejemplo.

Eliminación.–

Supongamos que queremos resolver el siguiente sistema de ecuaciones lineales:

$$\begin{aligned} 10x - 7y &= 7 \\ -3x + 2y + 6z &= 4 \\ 5x - y + 5z &= 6 \end{aligned} \tag{2.1}$$

El primer paso del método de eliminación de Gauss es eliminar la primera incógnita de todas las ecuaciones que siguen a la primera, para lo cual se les resta sucesivamente la primera ecuación multiplicada por el factor adecuado a cada una (llamado *multiplicador*). Para ello, en este caso, restamos de la segunda ecuación la primera multiplicada por el *multiplicador* $-\frac{3}{10}$ y restamos de la tercera ecuación la primera multiplicada por el multiplicador $\frac{5}{10}$ con lo que el sistema se convierte en uno equivalente de la forma

$$\begin{aligned} 10x - 7y &= 7 \\ -0.1y + 6z &= 6.1 \\ 2.5y + 5z &= 2.5 \end{aligned}$$

El siguiente paso y sucesivos consisten en repetir el primer paso pero aplicado al subsistema formado por las ecuaciones que hayan sido modificadas en el paso anterior. Así pues, en nuestro ejemplo el paso siguiente es eliminar el término en y de la tercera ecuación para lo que le restamos la segunda multiplicada por $2.5/(-0.1) = -25$, de forma que nos queda

$$\begin{aligned} 10x - 7y &= 7 \\ -0.1y + 6z &= 6.1 \\ 155z &= 155\,. \end{aligned}$$

El resultado final (al cabo de $n-1$ pasos en un sistema de orden n) es haber transformado el sistema original en otro equivalente pero triangular superior.

Ejercicio 2.1
Describir un algoritmo que efectúe el proceso de eliminación que acabamos de describir para un sistema de n ecuaciones lineales con n incógnitas (de forma que lo transforme en uno equivalente pero triangular superior).

Sustitución regresiva.–

Después de haberlo reducido a forma triangular es inmediato encontrar la solución del sistema dado empezando por la última ecuación y procediendo hacia atrás. La última ecuación es ahora una ecuación lineal en una incógnita que se resuelve de forma inmediata. Llevando esta solución a la ecuación anterior podremos hallar la incógnita anterior, y así sucesivamente hasta hallar todas las incógnitas. En nuestro ejemplo,

$$\begin{aligned} z &= 1\,; \\ y &= \frac{6.1 - 6z}{-0.1} = \frac{6.1 - 6}{-0.1} = -1\,; \\ x &= \frac{7 + 7y + 0z}{10} = \frac{7 + 7(-1)}{10} = 0\,. \end{aligned}$$

Esta resolución en escalada hacia atrás se conoce como *sustitución regresiva*.

Ejercicio 2.2
Describir un algoritmo que efectúe el proceso de sustitución regresiva en un sistema de n ecuaciones lineales con n incógnitas supuesto dado en forma triangular superior.

Ejercicio 2.3
Hallar el número de operaciones (contando sólo multiplicaciones y divisiones) que es necesario realizar para resolver un sistema de n ecuaciones lineales con n incógnitas por el método de eliminación de Gauss.

Dificultades que pueden surgir en la eliminación.–

Veamos ahora dos tipos de dificultades que pueden surgir en la aplicación del método de eliminación.

En primer lugar puede ocurrir que al comienzo de alguno de los pasos el coeficiente por el que tenemos que dividir los demás de su columna para hallar los multiplicadores de ese paso sea cero, con lo cual no podremos continuar la eliminación. Por ejemplo, esto ocurre ya en el primer paso en el sistema

$$\begin{aligned} -7y + 10z &= 7 \\ 6x + 2y - 3z &= 4 \\ 5x - y + 5z &= 6 \end{aligned}$$

sin embargo es evidente que, a menos que el sistema sea singular, no pueden ser cero todos los coeficientes de la primera columna, con lo cual siempre podremos intercambiar la ecuación cuyo primer coeficiente es cero con una ecuación *posterior* cuyo primer coeficiente no sea cero (intercambio que evidentemente no afecta en nada a la solución). Este mecanismo de intercambio de ecuaciones en la búsqueda de un coeficiente adecuado para ser usado como divisor en el cálculo de los multiplicadores de un cierto paso de eliminación se llama *pivotación* y el coeficiente encontrado (que pasará a ocupar la posición $(1, 1)$ del subsistema sobre el que vamos a trabajar en ese paso) se llama el *pivote* de ese paso. Existen diversas estrategias para realizar la pivotación. La que acabamos de describir consiste en buscar el primer coeficiente no nulo en la primera columna del subsistema, estrategia conocida como *pivotación simple*. Según lo dicho,

Teorema 9 *Todo sistema no singular de n ecuaciones lineales con n incógnitas puede resolverse por el método de eliminación de Gauss con pivotación simple.* ■

Asimismo se conoce una caracterización de los sistemas que se pueden resolver por el método de eliminación de Gauss sin necesidad de realizar pivotación alguna. Esta propiedad, evidentemente, es simplemente una propiedad del orden en que decidimos tomar las ecuaciones del sistema durante el proceso de eliminación. Recordemos que los *menores principales* de una matriz son las submatrices cuadradas que se obtienen al eliminar todas las filas y columnas *a partir* de la columna k y de la fila k para un mismo k fijo.

Así, el menor principal de orden k de una matriz $A = (a_{ij})$ puede definirse como la matriz cuadrada cuyos elementos son todos los a_{ij} con $1 \leq i \leq k$ y $1 \leq j \leq k$. Dicho esto podemos enunciar:

Teorema 10 *Un sistema de n ecuaciones lineales con n incógnitas puede resolverse por el método de eliminación de Gauss sin pivotación alguna si y sólo si en la matriz de coeficientes todos los menores principales son no singulares.*

Demostración: Primeramente hay que observar que las operaciones elementales realizadas sobre la matriz de coeficientes durante el proceso de eliminación no cambian los valores de los determinantes de los menores principales. Denotemos $a_{ij}^{(k)}$ los elementos obtenidos al realizar el paso $k-1$ de eliminación. Si se puede realizar la eliminación sin pivotación entonces los elementos

$$a_{11},\ a_{22}^{(2)},\ a_{33}^{(3)},\ \ldots,\ a_{nn}^{(n)}$$

son todos distintos de cero. Además, para todo $r \in \{1, \ldots, n\}$ el menor principal de orden r, A_{rr}, tiene determinante dado por

$$\det(A_{rr}) = a_{11} \cdot a_{22}^{(2)} \cdot a_{33}^{(3)} \cdots a_{rr}^{(r)},$$

luego todos los menores principales son no singulares.

Supongamos ahora que algún menor principal sea singular. Sea r el menor de los órdenes de los menores principales que sean singulares. Entonces podemos realizar $r-1$ pasos de eliminación sin pivotación y los elementos

$$a_{11},\ a_{22}^{(2)},\ a_{33}^{(3)},\ \ldots,\ a_{r-1r-1}^{(r-1)}$$

son todos distintos de cero, pero el elemento $a_{rr}^{(r)}$ obtenido en el paso $r-1$ será necesariamente igual a cero y en consecuencia la eliminación no puede proseguir sin pivotación. ■

Evidentemente la propiedad de un sistema de ecuaciones lineales mencionada en este teorema (de poder ser resuelto por eliminación sin pivotación) no es una propiedad del sistema como tal, sino que es una propiedad del *orden en que están dadas las ecuaciones*. En todo sistema no singular pueden reordenarse las ecuaciones de tal forma que el sistema tenga dicha propiedad.

La segunda dificultad que puede surjir durante la eliminación es un poco más sutil y es debida al hecho de que los cálculos necesarios para obtener la solución se realicen en aritmética de coma flotante con una precisión prefijada. Como consecuencia de esto puede ocurrir que al comienzo de alguno de los pasos el coeficiente por el que tenemos que dividir los demás de su columna para hallar los multiplicadores de ese paso, sin ser cero, sea, en valor absoluto, tan pequeño en relación a los demás coeficientes que produzca uno o varios multiplicadores con valor absoluto excesivamente grande, resultando en graves errores de redondeo. Esto puede dar lugar a una solución completamente inaceptable como en el siguiente ejemplo en que los coeficientes y términos independientes están dados con una precisión de cuatro dígitos decimales y se usa aritmética de coma flotante con cuatro dígitos:

$$\begin{aligned} 0.0003\,x + 1.566\,y &= 1.569 \\ 0.3454\,x - 2.436\,y &= 1.018\,. \end{aligned}$$

El primer (y único) multiplicador es $\frac{0.3454}{0.0003} = 1151$, luego en el primer paso de eliminación obtenemos

$$\begin{aligned} 0.0003\,x + 1.566\,y &= 1.569 \\ -\,1804\,y &= -1805\,, \end{aligned}$$

de donde

$$\begin{aligned} y &= -1805/1804 = 1.001 \\ x &= \frac{1.569 - 1.566 \times 1.001}{0.0003} = 3.333\,. \end{aligned}$$

mientras que la solución exacta es $x = 10\,,\ y = 1$.

Una sencilla solución de esta dificultad, que simultáneamente resuelve la primera mencionada, consiste en incluir en la fase de eliminación la llamada *pivotación parcial*.

Pivotación parcial.–

El método de pivotación parcial consiste en elegir como *pivote* en cada paso de la eliminación el coeficiente de mayor valor absoluto de la primera

columna del subsistema correspondiente a ese paso. Por ejemplo, en el caso del sistema

$$\begin{aligned} -\,7y + 10z &= 7 \\ 6x + 2y - \ 3z &= 4 \\ 5x - \ y \ + \ 5z &= 6 \end{aligned}$$

en el primer paso queremos eliminar la x. La ecuación que tiene el coeficiente de x con máximo valor absoluto es la segunda, luego el pivote en este paso es 6 y comenzamos el primer paso intercambiando las ecuaciones primera y segunda quedando el sistema en la forma:

$$\begin{aligned} 6x + 2y - \ 3z &= 4 \\ -\,7y + 10z &= 7 \\ 5x - \ y \ + \ 5z &= 6 \end{aligned}$$

Con este método el cálculo de los multiplicadores a utilizar en cada paso se realiza mediante una división por el divisor de mayor valor absoluto que es posible tener sin permutar las variables, lo que resuelve en gran medida las dificultades mencionadas. Nótese que el pivote necesariamente será no nulo en cada paso si el sistema es no-singular. Este método nos ofrece, pues, un test de singularidad.

Algoritmo del método de eliminación de Gauss con pivotación parcial.–

Lo que hemos dicho hasta ahora queda reflejado en el siguiente pseudo código para resolver un sistema de ecuaciones lineales:

```
Suponemos dados: n, el orden del sistema,
  A(i, j), la matriz de coeficientes y
  B(i), la matriz de términos independientes.
La solución sera X(i).
********* BUCLE DE ELIMINACION *********
 para k = 1 hasta n-1
********* BUSQUEDA DEL pivote *********
 c = abs(A(k, k)); p = k
 para i = k+1 hasta n
  si abs(A(i, k)) > c entonces c = abs(A(i, k)) y p = i
 siguiente i
```

```
 si c = 0 entonces imprimir 'SISTEMA SINGULAR' y parar
********* INTERCAMBIO DE LA ECUACION k CON LA p *********
 para j = k hasta n
  t = A(p, j); A(p, j) = A(k, j); A(k, j) = t
 siguiente j
 t = B(p); B(p) = B(k); B(k) = t
********* PASO k DE LA ELIMINACION *********
 para i = k+1 hasta n
  m = A(i, k)/A(k, k); A(i, k) = 0
  para j = k+1 hasta n
 A(i, j) = A(i, j) - m*A(k, j)
  siguiente j
  B(i) = B(i) - m*B(k)
 siguiente i
 siguiente k
********* SUSTITUCION REGRESIVA *********
 si A(n, n) = 0 entonces imprimir 'SISTEMA SINGULAR' y parar
 X(n) = B(n)/A(n, n)
 para k = n-1 hasta 1 incremento -1
 s = 0
 para j = k+1 hasta n
 s = s + A(k, j)*X(j)
 siguiente j
 X(k) = (B(k) - s)/A(k, k)
 siguiente k
```

2.1.2 La factorización triangular realizada por la eliminación de Gauss

El proceso de transformación a que se somete un sistema de ecuaciones al resolverlo por el método de eliminación consiste en una serie de *operaciones elementales* realizadas sobre las filas de la matriz de coeficientes ampliada por la columna de términos independientes.

Recordemos que el resultado de realizar una serie de operaciones elementales sobre las filas de una matriz cualquiera es el mismo que si (1) realizamos dichas operaciones elementales sobre (las filas de) la matriz identidad que tiene el mismo número de filas que la dada y (2) multiplicamos la matriz resultante (por la izquierda) por la matriz dada.

Por ejemplo, la eliminación que hemos realizado en el sistema (2.1)

consiste en las operaciones elementales:

1. Restar de la segunda fila la primera multiplicada por $-\frac{3}{10}$,

2. Restar de la tercera fila la primera multiplicada por $\frac{5}{10}$,

3. Restar de la tercera fila la segunda multiplicada por -25.

Efectuando estas operaciones sobre la matriz identidad de orden 3 ésta se convierte sucesivamente en

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ \frac{-5}{10} & 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ \frac{25\times 3-5}{10} & 25 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix}.$$

Multiplicando ahora la matriz resultante por la matriz de coeficientes obtenemos

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix} \begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix} = \begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{pmatrix}$$

que es la matriz de coeficientes del sistema reducido, es decir, del obtenido tras la eliminación. Análogamente, si multiplicamos por la matriz de términos independientes obtendremos los términos independientes del sistema reducido:

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix} \begin{pmatrix} 7 \\ 4 \\ 6 \end{pmatrix} = \begin{pmatrix} 7 \\ 6.1 \\ 155 \end{pmatrix}$$

La consecuencia más importante de todo esto es que el proceso de eliminación puede considerarse como la construcción de la matriz M que multiplicada por la matriz de coeficientes, A, nos da una matriz triangular superior U

$$M \cdot A = U.$$

La matriz U resultante es la matriz de los coeficientes del nuevo sistema de ecuaciones y M es la matriz que se obtiene al aplicar las operaciones elementales de eliminación a la matriz identidad. Así pues una forma sencilla de obtener la matriz M es realizando el proceso de eliminación sobre la matriz de coeficientes ampliada con la matriz identidad (igual que hacemos

en el cálculo de la matriz inversa). Cuando la matriz de coeficientes se ha puesto en forma triangular ($A \mapsto U$), la matriz identidad se ha convertido en la matriz M ($I \mapsto M$).

Una característica importante de la matriz M de las transformaciones elementales realizadas durante la eliminación es que si, como ocurre en el ejemplo, durante la eliminación no ha habido ninguna pivotación entonces M es una matriz triangular inferior con unos en la diagonal. En este caso su inversa, $L = M^{-1}$ nos proporciona la factorización triangular de la matriz de coeficientes del sistema: $A = LU$. Además esta inversa, L, tiene la propiedad de ser también una matriz triangular inferior con unos en la diagonal pero que además sus elementos bajo la diagonal son los multiplicadores utilizados en la eliminación.

El proceso de eliminación de Gauss nos proporciona, en caso de poderlo realizar sin pivotación alguna como en nuestro ejemplo, la factorización triangular (o factorización LU) de una matriz sin más que hallar $L = M^{-1}$. En nuestro ejemplo,

$$A = \begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix}^{-1} \begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{pmatrix}$$

es decir,

$$L = \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix}^{-1}, \quad U = \begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{pmatrix}.$$

Nótese que L resultará ser (suponiendo que no ha habido pivotaciones) la inversa de una matriz triangular inferior con unos en la diagonal, lo que justifica que ella misma es triangular inferior con unos en la diagonal.

Pero esto no es todo. El proceso de eliminación de Gauss nos proporciona directamente la matriz L sin necesidad de calcular ninguna inversa. La razón de esto es que L es la inversa de un producto de matrices elementales y por tanto se puede obtener ella misma como resultado de operaciones elementales sobre las filas de la matriz identidad. Lo único que necesitamos es realizar las inversas de las operaciones elementales usadas en la eliminación y en el orden inverso. Es decir, en nuestro ejemplo,

1. Sumar a la tercera fila la segunda multiplicada por -25.

2. Sumar a la tercera fila la primera multiplicada por $\frac{5}{10}$,

3. Sumar a la segunda fila la primera multiplicada por $-\frac{3}{10}$,

que efectuadas sucesivamente sobre la matriz identidad de orden 3 dan lugar a

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -25 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{5}{10} & -25 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 \\ \frac{-3}{10} & 1 & 0 \\ \frac{5}{10} & -25 & 1 \end{pmatrix} = L.$$

Ejercicio 2.4
Comprobar que

$$\begin{pmatrix} 1 & 0 & 0 \\ \frac{-3}{10} & 1 & 0 \\ \frac{5}{10} & -25 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{3}{10} & 1 & 0 \\ 7 & 25 & 1 \end{pmatrix}^{-1}.$$

Así pues, la factorización LU de nuestra matriz de coeficientes es:

$$\begin{pmatrix} 10 & -7 & 0 \\ -3 & 2 & 6 \\ 5 & -1 & 5 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{-3}{10} & 1 & 0 \\ \frac{5}{10} & -25 & 1 \end{pmatrix} \begin{pmatrix} 10 & -7 & 0 \\ 0 & -0.1 & 6 \\ 0 & 0 & 155 \end{pmatrix}.$$

Vemos que los elementos de la matriz L bajo la diagonal son precisamente los multiplicadores utilizados en la eliminación. Esto es un resultado general:

El proceso de eliminación de Gauss (sin pivotación) lleva a cabo una factorización de la matriz de coeficientes en una matriz (L) triangular inferior con unos en la diagonal y una matriz (U) triangular superior. Los elementos de la matriz L (bajo la diagonal) son los multiplicadores utilizados durante la eliminación mientras que los de U son los coeficientes del sistema reducido.

Efecto de la pivotación.–

En lo que precede hemos supuesto explícitamente que en el proceso de eliminación no se realizaba ninguna reordenación de filas, es decir, se realizaba sin pivotación. ¿Cómo se altera lo dicho en el caso de realizar la eliminación con algún tipo de pivotación?

Para contestar a esta pregunta supongamos que llevamos a cabo un proceso de eliminación con pivotación sobre un sistema de ecuaciones lineales. Esto significa que algunas de las operaciones elementales realizadas durante la eliminación pueden ser intercambios de filas. Por ejemplo, en nuestro ejemplo pivotación parcial daría lugar a las siguientes operaciones elementales:

1. Restar de la segunda fila la primera multiplicada por $-\frac{3}{10}$,

2. Restar de la tercera fila la primera multiplicada por $\frac{5}{10}$,

3. Intercambiar la segunda fila con la tercera,

4. Restar de la tercera fila la segunda multiplicada por $(-1/25)$.

Ahora bien, el intercambio de filas (o pivotación) realizado antes del paso k de la eliminación puede realizarse inmediatamente antes de realizar el paso $k-1$ sin afectar en nada al proceso (ya que es un intercambio de dos ecuaciones *posteriores* a la $k-1$, que es la ecuación pivote en el paso $k-1$). Podemos luego, para cada ecuación, calcular su multiplicador, que es independiente del orden que ocupa (aunque esto implica una "reordenación" de los multiplicadores, pero que se realiza automáticamente). Como consecuencia se tiene que todos los intercambios de ecuaciones necesarios en un proceso de eliminación con pivotación pueden realizarse juntos (pero sin alterar su orden) antes de comenzar la eliminación. En nuestro ejemplo, la eliminación puede realizarse mediante los pasos (nótese el intercambio de los dos primeros multiplicadores):

1. Intercambiar la segunda fila con la tercera,

2. Restar de la segunda fila la primera multiplicada por $\frac{5}{10}$,

3. Restar de la tercera fila la primera multiplicada por $-\frac{3}{10}$,

4. Restar de la tercera fila la segunda multiplicada por $(-1/25)$.

La consecuencia de lo que acabamos de decir es que la eliminación de Gauss con pivotación es equivalente a la eliminación sin pivotación pero realizada después de una reordenación de las ecuaciones del sistema. Esta

reordenación puede realizarse sobre la matriz de coeficientes, A, (así como sobre la matriz de términos independientes, B), mediante la multiplicación (por la izquierda) por una matriz de permutación P que se ha obtenido sometiendo la matriz identidad a los sucesivos intercambios de filas que se han de realizar sobre A. Así pues, el proceso de eliminación con pivotación es equivalente a la transformación de la matriz A en la forma

$$MPA = U$$

donde P es una matriz de permutación.

Como consecuencia de todo lo dicho tenemos:

Teorema 11 (Factorización triangular) *Para toda matriz cuadrada no singular A se puede hallar una permutación P tal que PA admite una factorización triangular, es decir, de la forma*

$$PA = LU$$

donde L una matriz triangular inferior con unos en la diagonal y U una matriz triangular superior. Las matrices L y U están completamente determinadas por A y P. Además, se podrá elegir $P = 1$ (matriz identidad) si y sólo si todos los menores principales de A son no singulares. ■

Además sabemos cómo encontrar P así como los elementos de L y U mediante el proceso de eliminación de Gauss con pivotación parcial,

Ejercicio 2.5
Adaptar el algoritmo de eliminación de Gauss con pivotación parcial para convertirlo en un algoritmo que efectúe la factorización triangular de una matriz no singular dada, es decir, que a partir de una matriz A encuentre las matrices P, L, y U del teorema anterior.

Uso de la factorización triangular en la resolución de sistemas de ecuaciones lineales.–

El mayor interés de obtener la factorización triangular mencionada más arriba está en el hecho de que permite resolver cualquier sistema de ecuaciones lineales con la misma matriz de coeficientes A mediante una sustitución progresiva seguida de una sustitución regresiva. Esto es: del sistema

$Ax = b$ pasamos a $PAx = Pb$, o sea, $LUx = b'$. Esto significa que si hemos realizado la eliminación en A obteniendo P, L y U, la resolución de un sistema $Ax = b$ se reduce a estos pasos:

1. Obtener el vector $b' = Pb$.
2. Resolver $Ly = b'$ mediante una sustitución progresiva.
3. Resolver $Ux = y$ mediante una sustitución regresiva.

Esto es especialmente útil cuando necesitamos resolver varios sistemas de ecuaciones que tienen la misma matriz de coeficientes.

Según se ha visto, al resolver un sistema de ecuaciones lineales mediante el proceso de eliminación de Gauss con pivotación parcial se lleva a cabo en realidad una factorización triangular $PA = LU$ de una permutación de la matriz de coeficientes así como una sustitución progresiva y una sustitución regresiva. En la parte de factorización no interviene la columna de términos independientes, sino solamente la matriz de coeficientes. Esto abre la posibilidad de separar las dos partes del proceso, de forma que si necesitamos resolver varios sistemas de ecuaciones lineales en los que sólo difieren los términos independientes, sólo tengamos que realizar la parte de factorización una vez. Puesto que esa parte es la más costosa (en términos del número de operaciones a realizar), esta estrategia redundará en una mayor eficacia del método.

Una de las aplicaciones de este método es el cálculo de la matriz inversa. Este cálculo es claramente equivalente a la resolución de n sistemas de ecuaciones lineales todos ellos con la matriz dada como matriz de coeficientes y teniendo como columnas de términos independientes las columnas de la matriz identidad. Es interesante observar que el cálculo de la matriz inversa de una matriz A de orden n por este método requiere realizar el mismo número de operaciones que el cálculo de A^2.

Otra aplicación de este método es el algoritmo de la mejora iterativa de las soluciones aproximadas que explicamos más adelante (página 72).

2.1.3 Método de Gauss-Jordan

Si en cada paso del proceso de eliminación

$$(\text{ecuación } i) = (\text{ecuación } i) - m_i \times (\text{ecuación } k) \quad i = k+1, \ldots n$$

aplicamos la eliminación no sólo a las ecuaciones siguientes a la ecuación k, sino también a las ecuaciones anteriores entonces estaremos convirtiendo nuestro sistema en uno diagonal, no sólo triangular. Es decir si en el paso k realizamos

$$(\text{ecuación } i) = (\text{ecuación } i) - m_i \times (\text{ecuación } k)$$

para $i = 1, \ldots, k-1, k+1, \ldots n$, nuestro sistema queda convertido en uno de n ecuaciones con una incógnita, para cuya solución no es necesario el proceso de sustitución regresiva. Solamente necesitaremos n divisiones adicionales. Este método se conoce con el nombre de *método de Gauss-Jordan*

Ejercicio 2.6
Hallar el número de operaciones (contando sólo multiplicaciones y divisiones) que es necesario realizar para resolver un sistema de n ecuaciones lineales con n incógnitas por el método de eliminación de Gauss-Jordan.

2.1.4 Otras técnicas de pivotación

Una técnica de pivotación es un criterio para la elección de la fila o ecuación *pivote* en cada paso del proceso de eliminación. Hasta ahora hemos encontrado dos técnicas de pivotación: pivotación simple y pivotación parcial, pero existen muchas otras posibilidades. Lo ideal sería descubrir una técnica sencilla de pivotación que nos asegurase que hacemos las operaciones en el orden óptimo en el sentido de que se minimizan tanto la pérdida de precisión como los errores de redondeo. En ausencia de esa técnica ideal hemos de guiarnos por la experiencia y por consideraciones generales tales como la citada más arriba. Veremos a continuación dos técnicas adicionales que suelen dar resultados adecuados.

Pivotación parcial escalada.–

Para algunos sistemas podemos obtener algún aumento en la precisión si al elegir la fila pivote tenemos en cuenta de alguna manera no tanto el tamaño del elemento pivote como su tamaño *relativo* a los demás elementos de su fila. De esta forma el proceso no quedaría afectado si se re-escala alguna de las ecuaciones del sistema, es decir, si se multiplica por un número toda una ecuación (lo cual, en principio no debería afectar a la solución).

Así pues, realizaremos una *pivotación parcial escalada* cuando en cada paso de la eliminación elijamos como pivote aquel candidato cuyo cociente por el tamaño de su fila sea máximo. Por "tamaño" de una fila se puede entender alguna norma vectorial conveniente tal como la norma del máximo, la cual se calcula con poco esfuerzo.

Para realizar la pivotación parcial escalada de forma eficaz calcularemos los tamaños de todas las filas antes de comenzar el primer paso de la eliminación,

$$d_i = \max_{1 \le j \le n} |a_{ij}|$$

Después, al comienzo del paso k para $k = 1, \ldots, n-1$, si los posibles pivotes son $a_{kk}, \ldots, a_{nk}$, elegimos como pivote aquel para el que el cociente $|a_{ik}|/d_i$ sea máximo.

Ejercicio 2.7
Modificar el algoritmo de eliminación de Gauss para que realice pivotación parcial escalada en lugar de mera pivotación parcial.

Pivotación total.–

La técnica de pivotación total es considerada como la que da mejores resultados en general, aunque su uso es bastante limitado debido al coste adicional de programación que supone. Esta técnica es una variante de la pivotación parcial en la que no sólo se barajan las ecuaciones restantes en la búsqueda del pivote sino que en cada paso se busca la incógnita más adecuada a eliminar en ese paso. Así, no sólo nos permitimos permutar las ecuaciones, sino que también nos permitimos permutar las incógnitas.

Al comienzo del paso k se elegirá como pivote aquel elemento de mayor valor absoluto en la submatriz de coeficientes con que se trabajará en ese paso. En consecuencia un algoritmo que lleve a cabo pivotación total realizará en cada paso tanto una permutación de filas (o ecuaciones) como una permutación de columnas (o variables).

2.2 Análisis de errores y Condición de un sistema de ecuaciones lineales

2.2.1 Introducción: Distintas medidas del error

En la práctica del cálculo científico raramente se conocen con exactitud los coeficientes y términos independientes del sistema de ecuaciones lineales que se quiere resolver. En general, esos coeficientes se calculan a partir de observaciones experimentales sujetas como mínimo a errores que vienen de la limitada precisión de los aparatos de medida. En consecuencia se plantea la cuestión de la *confianza* que podemos tener en los resultados de nuestros cálculos sabiendo que los propios datos están sujetos a error. En otras palabras, nos preguntamos en que medida nuestros cálculos pueden haber amplificado el error inherente a los datos y si éste se habrá mantenido dentro de límites razonables.

Normalmente la confianza que podemos tener en la solución dependerá del sistema de que se trate: distintas matrices de coeficientes darán lugar a distintos comportamientos respecto a la propagación y amplificación de errores. Nuestro objetivo es poder determinar a priori el tipo de comportamiento que una matriz dada va a tener. Para ello comenzaremos repasando varias formas de medir la exactitud o precisión de una solución aproximada, $\hat{x}$, de un sistema de ecuaciones lineales $Ax = b$.

La primera forma es mediante el **error absoluto** o diferencia $e = x - \hat{x}$ entre la solución exacta y la solución aproximada. Obviamente esto es normalmente desconocido. Otra forma de medir la exactitud de $\hat{x}$ como solución de $Ax = b$ es mediante el **error residual**, que mide cuán lejos está $\hat{x}$ de satisfacer el sistema de ecuaciones lineales. El error residual de $\hat{x}$ es

$$r = Ax - A\hat{x} = Ae\,,$$

lo cual puede calcularse sin dificultad, pues es igual a la diferencia $b - A\hat{x}$ entre el vector de términos independientes dado y el calculado a partir de la solución aproximada.

En general el error residual no será el vector cero, aunque podemos esperar que su tamaño sea pequeño. Al hablar del "tamaño" de vectores en $\mathbf{R}^n$ nos referimos a una norma particular fija que suponemos elegida de antemano. Qué norma se use no tiene demasiada importancia mientras uno se atenga a usar siempre esa norma. Una norma muy en uso es la

"norma infinito" o del máximo ya que se calcula con muy poco esfuerzo (no requiere operaciones aritméticas). En general se puede suponer que la norma usada es una "norma p", para algún real $p \geq 1$, definida por

$$\|x\|_p = \left(|x_1|^p + \cdots + |x_n|^p\right)^{1/p}.$$

De éstas la norma del máximo es el caso límite para $p \to \infty$. Los casos más utilizados son $p = 1, 2, \infty$.

El tamaño del error residual está relacionado con el tamaño del error absoluto. Al multiplicar la matriz A por un vector columna u la norma de éste puede variar mucho, pero hay que observar, sin embargo, que el cambio *relativo* o factor por el que ha cambiado la norma (el cociente $\|Au\|/\|u\|$) no puede ser arbitrario. Necesariamente permanecerá dentro de ciertos límites que dependen de la matriz A. Concretamente $\|Au\|/\|u\|$ será necesariamente menor o igual que el *máximo factor* por el que A multiplica las normas de los vectores. Veremos a continuación que debido a esto el **error residual relativo**

$$\frac{\|b - Ax\|}{\|b\|} = \frac{\|r\|}{\|b\|} = \frac{\|Ae\|}{\|Ax\|}$$

es un buen indicador, no del error e, sino del **error relativo**

$$\frac{\|x - \hat{x}\|}{\|x\|} = \frac{\|e\|}{\|x\|}.$$

(ver fórmula (2.3)).

2.2.2 Mejora iterativa

Algunos ordenadores y lenguajes de programación admiten realizar aritmética de coma flotante en dos modos distintos de precisión: sencilla y doble, de los que el segundo modo cuesta aproximadamente el doble de tiempo-máquina que el primero pero permite reducir los errores (especialmente los de pérdida de precisión) al realizar las operaciones con aritmética de coma flotante con el doble de dígitos de precisión que en el modo de precisión sencilla. Cuando se dispone de esta posibilidad se puede reducir notablemente el error en la solución de sistemas de ecuaciones lineales sin un coste excesivo (como el que implicaría el realizar todos los cálculos con doble precisión). La técnica que permite realizar esta reducción de error se conoce como *mejora iterativa* y consiste en lo siguiente:

Supongamos que hemos utilizado el método de la factorización triangular (con precisión sencilla) para hallar la solución de un sistema de ecuaciones lineales $Ax = b$ tal como

$$\begin{pmatrix} 0.20000 & 0.16667 & 0.14286 \\ 0.16667 & 0.14286 & 0.12500 \\ 0.14286 & 0.12500 & 0.11111 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 0.50953 \\ 0.43453 \\ 0.37897 \end{pmatrix}.$$

O sea, suponemos que ya hemos obtenido una factorización $A = PLU$ de la matriz de coeficientes, que en nuestro ejemplo es

$$L = \begin{pmatrix} 1 & 0 & 0 \\ 0.83335 & 1 & 0 \\ 0.71430 & 0.89673 & 1 \end{pmatrix}, \quad U = \begin{pmatrix} 0.20000 & 0.16667 & 0.14286 \\ 0 & 0.00397 & 0.00595 \\ 0 & 0 & 0.00015 \end{pmatrix},$$

sin realizar permutaciones, o sea que P es la identidad, y hemos obtenido la solución

$$x^{(1)} = \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} 1.0384 \\ 0.89673 \\ 1.0667 \end{pmatrix},$$

realizando las operaciones con aritmética de coma flotante con precisión de cinco dígitos. Este resultado puede mejorarse de la siguiente forma:

Primeramente calculamos el error residual $r^{(1)} = b - Ax^{(1)}$ utilizando aritmética de doble precisión. Esto significa realizar para cada $i \in \{1, \ldots, n\}$ las siguientes $(n-1)$ multiplicaciones y n sumas con doble precisión:

$$r_i^{(1)} = b - \sum_{k=1}^{n} a_{ik} x_k^{(1)},$$

lo que en nuestro ejemplo da

$$Ax^{(1)} = \begin{pmatrix} 0.5095324653 \\ 0.4345190593 \\ 0.3789619207 \end{pmatrix} \quad \text{y} \quad r^{(1)} = \begin{pmatrix} -0.24653 \\ 1.0941 \\ 0.80793 \end{pmatrix} 10^{-5}$$

Obtenida esta aproximación del vector de error residual la utilizamos para calcular una estimación $e^{(1)}$ del error absoluto e resolviendo la ecuación $Ae = r$. Esta ecuación representa un sistema de ecuaciones lineales que sólo se diferencia del dado en los términos independientes. En consecuencia, como ya disponemos de la factorización de la matriz de coeficientes la nueva resolución tendrá un coste mucho menor ya que se reduce

a una sustitución progresiva y una sustitución regresiva como las que nos llevaron de b a $x^{(1)}$, pero esta vez utilizando $r^{(1)}$ en lugar de b como vector de términos independientes. De esta forma obtenemos el vector $e^{(1)}$, que sumado a $x^{(1)}$ resultará en la mejora

$$x^{(2)} = x^{(1)} + e^{(1)} .$$

En nuestro ejemplo obtenemos

$$e^{(1)} = \begin{pmatrix} -0.03709 \\ 0.09955 \\ -0.06424 \end{pmatrix} \quad \text{y} \quad x^{(2)} = x^{(1)} + e^{(1)} = \begin{pmatrix} 1.0014 \\ 0.99628 \\ 1.0024 \end{pmatrix}$$

Repitiendo este proceso hallaríamos el resto $r^{(2)}$ asociado con $x^{(2)}$ utilizando doble precisión, después hallaríamos el 'error' $e^{(2)}$ mediante la resolución de $Ae^{(2)} = r^{(2)}$ y obtendríamos una nueva mejora $x^{(3)} = x^{(2)} + e^{(2)}$, que en nuestro ejemplo es

$$x^{(3)} = \begin{pmatrix} 1.0001 \\ 0.99986 \\ 1.0001 \end{pmatrix} .$$

Continuando de este modo se obtiene una sucesión de vectores

$$\{x^{(1)}, x^{(2)}, \dots\}$$

que son sucesivas mejoras que bajo hipótesis bastante razonables convergerán a la solución exacta x (en el ejemplo la solución exacta es

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

como se comprueba fácilmente).

Resumiendo, tenemos el siguiente algoritmo para la mejora iterativa de la solución de un sistema de ecuaciones lineales:

Algoritmo de la Mejora Iterativa

1 Factorización triangular $A = PLU$.

2 Utilizar dos sustituciones a partir de P, L, U y b para obtener x.

3 Con doble precisión calcular el resto de x, $r = b - Ax$.

4 Utilizar dos sustituciones a partir de P, L, U y r para obtener e.

5 Sumar e al viejo x para obtener el nuevo x.

6 Ir al paso **3**.

La única condición que se ha de cumplir para poder garantizar que este algoritmo converge a la solución exacta es que el proceso de resolución (una sustitución progresiva y una regresiva basadas en la factorización PLU de la matriz de coeficientes) sea equivalente a aplicar una matriz C "suficientemente cercana" a la inversa de la matriz de coeficientes, como se demuestra en el siguiente teorema:

Teorema 12 (Convergencia del método de mejora iterativa) *Sea $x^{(1)}$ un vector cualquiera y sea C una matriz cercana a la inversa de A en el sentido de que*

$$\|I - CA\| < 1 .$$

Entonces la sucesión de vectores $\{x^{(n)}\}$ definida inductivamente por

$$x^{(n+1)} = x^{(n)} + C(b - Ax^{(n)})$$

converge a la solución x del sistema de ecuaciones lineales $Ax = b$.

Demostración: Basta demostrar que $\lim_{n\to\infty} \|x^{(n)} - x\| = 0$. Ahora bien,

$$\begin{aligned} \|x^{(n)} - x\| &= \|x^{(n-1)} + C(b - Ax^{(n-1)}) - x\| \\ &= \|x^{(n-1)} - x + C(Ax - Ax^{(n-1)})\| \\ &= \|x^{(n-1)} - x - CA(x^{(n-1)} - x)\| \\ &= \|(I - CA)(x^{(n-1)} - x)\| \le \|I - CA\| \cdot \|x^{(n-1)} - x\| \end{aligned}$$

y continuando así se llega a

$$\|x^{(n)} - x\| \le \|I - CA\|^{n-1} \cdot \|x^{(1)} - x\|$$

de donde

$$0 \le \lim_{n\to\infty} \|x^{(n)} - x\| \le \|x^{(1)} - x\| \lim_{n\to\infty} \|I - CA\|^{n-1} = \|x^{(1)} - x\| \cdot 0 = 0$$

ya que por ser $\|I - CA\| < 1$, obviamente $\lim_{n\to\infty} \|I - CA\|^{n-1} = 0$. ■

Detención del algoritmo.–

Para implementar cualquier algoritmo iterativo es necesario establecer un criterio de detención, es decir, establecer una norma a seguir para determinar en qué paso conviene detener los cálculos. Un posible criterio (en ocasiones el único) es realizar un número fijo de pasos determinado a priori. Por supuesto el criterio mejor sería poder decidir cuál es el error absoluto que estamos dispuestos a aceptar y detener el algoritmo justo cuando se alcance un error inferior al que vamos a aceptar. Esto se complica por el hecho de que el error absoluto es desconocido, sin embargo en el algoritmo de la mejora iterativa que acabamos de estudiar el error absoluto puede ser acotado en términos de las sucesivas *correcciones* como indica el siguiente resultado,

Ejercicio 2.8
En el proceso de mejora iterativa definido por

$$x^{(n+1)} = x^{(n)} + C(b - Ax^{(n)})$$

el tamaño del error $x^{(n)} - x$ se puede acotar en cada paso por el siguiente múltiplo de la corrección $x^{(n+1)} - x^{(n)}$ a realizar en ese paso:

$$\|x^{(n)} - x\| \leq \|(CA)^{-1}\| \|x^{(n+1)} - x^{(n)}\| .$$

Como la matriz CA (y por tanto también $(CA)^{-1}$) es cercana a la identidad, el factor $\alpha = \|(CA)^{-1}\|$ será cercano a la unidad, por lo que podemos utilizar como criterio de detención el tamaño de las sucesivas correcciones $\|x^{(n+1)} - x^{(n)}\|$, es decir, detener el algoritmo cuando la corrección realizada tenga un tamaño menor que el error máximo admisible.

2.2.3 Normas de matrices

El máximo factor por el que A multiplica las normas de los vectores se llama la *norma de la matriz* A (relativa a la norma vectorial empleada). Es decir, definimos la norma de la matriz A como

$$\|A\| = \max_{x \neq 0} \frac{\|Ax\|}{\|x\|} \tag{2.2}$$

y en consecuencia se verifica evidentemente para todo $u \neq 0$,

$$\frac{\|Au\|}{\|u\|} \leq \|A\|,$$

y por lo tanto, suponiendo que A es no-singular, también

$$\frac{\|v\|}{\|Av\|} \leq \|A^{-1}\|,$$

para todo $v \neq 0$.

Ejercicio 2.9
Multiplicando esas dos desigualdades deducir que para cualquier matriz cuadrada no-singular A y cualesquiera dos vectores no nulos u, v se verifica

$$\frac{\|Au\|}{\|u\|} \leq \|A\|\|A^{-1}\| \cdot \frac{\|Av\|}{\|v\|} .$$

Concluir que el número $\|A\| \cdot \|A^{-1}\|$ es siempre mayor o igual que 1.

La primera consecuencia de nuestra definición de norma de una matriz es la siguiente acotación del tamaño del error absoluto en términos del tamaño del error residual:

$$\frac{\|r\|}{\|A\|} \leq \|e\| \leq \|A^{-1}\|\|r\| .$$

Ejercicio 2.10
Del resultado del ejercicio A.11 se deduce que si A es una matriz cuadrada no-singular cualquiera se verifica

$$\frac{1}{\|A\|\|A^{-1}\|} \cdot \frac{\|Av\|}{\|Au\|} \leq \frac{\|v\|}{\|u\|} \leq \|A\|\|A^{-1}\| \cdot \frac{\|Av\|}{\|Au\|}$$

para cualesquiera vectores no nulos u, v.

Ejercicio 2.11
Demostrar que la definición (2.2) de norma de una matriz es equivalente a

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

y también a la conjunción de

$$\big(\forall x \|Ax\| \leq \|A\|\|x\|\big) \quad y \quad \exists x \neq 0 \mid \|Ax\| \geq \|A\|\|x\|.$$

Ejercicio 2.12
Demostrar que el concepto de norma de matrices definido antes satisface los axiomas usuales de normas y dos axiomas adicionales: (1) Para toda matriz identidad, I, $\|I\| = 1$, y (2) Si A y B son matrices cuadradas del mismo orden, $\|AB\| \leq \|A\| \cdot \|B\|$. Usar estas dos propiedades para probar que si A es inversible, $\|A\|\|A^{-1}\| \geq 1$.

Cálculo de algunas normas de matrices.–

El cálculo de la norma de una matriz asociada a una norma vectorial dada puede llegar a ser bastante engorroso. Un caso especial por su sencillez nos lo da la norma del máximo o "norma infinito". Obsérvese que para todo x,

$$\|Ax\|_\infty = \max_{1\leq i\leq n} \left| \sum_{j=1}^n a_{ij}x_j \right| \leq \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}||x_j|$$
$$\leq \max_{1\leq i\leq n} \Big(\big(\max_{1\leq j\leq n} |x_j| \big) \sum_{j=1}^n |a_{ij}| \Big) = \|x\|_\infty \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$$

lo cual demuestra que $\|A\|_\infty \leq \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$. Por otro lado es fácil dar un vector x tal que $\|Ax\|_\infty \geq \|x\|_\infty \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$. Para ello sea i_0 la fila de A para la que la suma $\sum_{j=1}^n |a_{ij}|$ es máxima, es decir, tal que

$$\sum_{j=1}^n |a_{i_0 j}| = \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$$

Eligiendo x de forma que $x_j = \text{signo}(a_{i_0 j})$, tendremos (suponiendo $A \neq 0$) $\|x\|_\infty = 1$ y

$$\|Ax\|_\infty = \max_{1\leq i\leq n} \left| \sum_{j=1}^n a_{ij}x_j \right| \geq \left| \sum_{j=1}^n a_{i_0 j}x_j \right|$$
$$= \sum_{j=1}^n |a_{i_0 j}| = \|x\|_\infty \max_{1\leq i\leq n} \sum_{j=1}^n |a_{ij}|$$

En conclusión tenemos que para toda matriz cuadrada $A = (a_{ij})$,

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

Por ejemplo, en el sistema de ecuaciones

$$0.0003\,x + 1.566\,y = 1.569$$
$$0.3454\,x - 2.436\,y = 1.018\,.$$

la matriz de coeficientes

$$A = \begin{pmatrix} 0.0003 & 1.566 \\ 0.3454 & -2.436 \end{pmatrix}$$

tiene norma-infinito

$$\|A\|_\infty = \max\{|0.0003| + |1.566|\,,\ |0.3454| + |-2.436|\} = 2.781\,.$$

Al extremo opuesto de la norma infinito está la norma uno que, curiosamente, su cálculo cuesta lo mismo que el de la norma infinito ya que

Ejercicio 2.13
Para toda matriz cuadrada A, su norma-uno es igual a la norma-infinito de su traspuesta

$$\|A\|_1 = \|A^t\|_\infty\,,$$

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

2.2.4 Condición de una matriz y acotación de errores

Acotación del error relativo.–

Volviendo a nuestro sistema de ecuaciones lineales $Ax = b$ y a los errores absoluto, $e = x - \hat{x}$, y residual, $r = b - A\hat{x} = Ae$, podemos interpretar el resultado del ejercicio 2.10 como estableciendo la siguiente acotación del error relativo $\|e\|/\|x\|$ en términos del error residual relativo $\|r\|/\|b\|$ y del número $c(A) = \|A\| \cdot \|A^{-1}\|$,

$$\frac{1}{c(A)} \cdot \frac{\|r\|}{\|b\|} \leq \frac{\|e\|}{\|x\|} \leq c(A) \cdot \frac{\|r\|}{\|b\|}\,. \tag{2.3}$$

Este número

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

que podemos asociar con cada matriz no singular A una vez elegida una norma, se llama *condición de la matriz A* (relativo a la norma elegida). Según la última parte del ejercicio A.11 la condición de una matriz verifica

$$c(A) \geq 1.$$

Por ejemplo, para la matriz del ejemplo anterior, con precisión de cuatro dígitos,

$$A^{-1} = \frac{1}{\det A}\begin{pmatrix} -2.436 & -1.566 \\ -0.3454 & 0.0003 \end{pmatrix} = \begin{pmatrix} 4.498 & 2.891 \\ 0.6377 & -5.539 \times 10^{-4} \end{pmatrix}$$

y por tanto

$$\begin{aligned} \|A^{-1}\|_\infty &= \max\{|4.498| + |2.891|\,,\ |0.6377| + |-0.0006|\} \\ &= 7.389\,, \end{aligned}$$

de donde obtenemos

$$c_\infty(A) = \|A\|_\infty \|A^{-1}\|_\infty = 2.781 \times 7.389 = 20.55$$

indicando un pobre condicionamiento de la matriz en cuestión.

Condición en términos de operadores lineales.–

Hay que acentuar el hecho de que la condición de una matriz está definida solamente si se ha especificado una norma vectorial, aunque algunas matrices (como por ejemplo las que son un múltiplo de la identidad) tienen la misma condición relativa a cualquier norma. En realidad el concepto de condición que acabamos de definir pertenece, rigurosamente hablando, a los operadores lineales en espacios normados. Cuando se habla de norma de una matriz se está identificando esa matriz con el operador lineal en $\mathbf{R}^n$ que la tiene por matriz en la base canónica. El siguiente ejercicio nos indica el significado geométrico de la condición de un operador lineal en un espacio normado: es una medida de la distorsión que el operador produce en las bolas.

Ejercicio 2.14
La condición de un operador lineal en un espacio normado es igual al cociente del radio máximo al radio mínimo de la imagen de la bola unitaria.

Ejercicio 2.15
La condición de una matriz A es igual a 1 *si y sólo si (el operador lineal representado por) A conserva las bolas, lo cual es decir que multiplica las normas de todos los vectores por el mismo factor.*

Dado que la condición de una matriz depende de la norma vectorial con que se calcula, se plantea la cuestión de si un sistema de ecuaciones lineales puede parecer bien condicionado en términos de una norma y mal condicionado en términos de otra distinta. Como respuesta parcial a esta cuestión tenemos:

Ejercicio 2.16
Si A es una matriz cuadrada real no singular que está bien condicionada respecto a una p-norma entonces también estará bien condicionada respecto a cualquier otra p-norma. Además, si la condición de A es igual a 1 *en relación a una p-norma para* $p \neq 2$ *entonces la condición de A también es igual a* 1 *en relación a cualquier otra p-norma. (Demuéstrese en el caso de matrices de orden* 2*.)*

Errores en los términos independientes.–

La condición de una matriz A nos sirve para estimar, mediante las desigualdades (2.3) el error relativo de una solución aproximada de cualquier sistema de ecuaciones lineales cuya matriz de coeficientes sea A. Pero además la condición de A es un indicador de la sensibilidad de la solución de dicho sistema de ecuaciones lineales a errores en el vector de términos independientes ya que las desigualdades (2.3) pueden interpretarse también de la siguiente forma: Supongamos que existe un error Δb en el vector b de términos independientes lo cual produce un error Δx en la solución, es decir, tenemos $A(x + \Delta x) = b + \Delta b$ donde x es la solución exacta de $Ax = b$. Entonces $A\Delta x = \Delta b$, $e = \Delta x$ y resulta $r = Ae = \Delta b$ en (2.3), y tenemos la siguiente acotación del error relativo resultante de los errores en los términos independientes:

$$\frac{1}{c(A)} \cdot \frac{\|\Delta b\|}{\|b\|} \leq \frac{\|\Delta x\|}{\|x\|} \leq c(A) \cdot \frac{\|\Delta b\|}{\|b\|}\,.$$

Así pues, el condicionamiento de un sistema de ecuaciones lineales está indicado por la condición de la matriz de coeficientes. Si este número es cercano a la unidad el sistema está bien condicionado porque pequeños

errores (relativos) en los datos (términos independientes) no pueden dar lugar a grandes errores (relativos) en la solución.

Errores en los coeficientes del sistema.–

Nos interesa también estudiar el efecto que tienen los errores que pueda haber en los coeficientes del sistema.

Supongamos primeramente que A y B son dos matrices de las que A es inversible y que u, v son dos vectores tales que $Au = Bv$. Entonces $u = A^{-1}Bv$ y por tanto

$$\|u\| = \|A^{-1}Bv\| \leq \|A^{-1}\|\|Bv\| \leq \|A^{-1}\|\|B\|\|v\|$$

de donde deducimos, suponiendo $v \neq 0$,

$$\|B\|\frac{\|v\|}{\|u\|} \geq \frac{1}{\|A^{-1}\|} \quad \text{o bien} \quad \frac{\|u\|}{\|v\|} \leq c(A)\frac{\|B\|}{\|A\|}\,. \tag{2.4}$$

Esta acotación puede aplicarse al caso de un sistema de ecuaciones lineales en el que los coeficientes se conozcan sólo aproximadamente como, por ejemplo, cuando se han obtenido como resultado de medidas experimentales. Tal situación es la de un sistema $Ax = b$ del que en lugar de la matriz A sólo disponemos de la matriz $\hat{A} = A + E$ donde E es una matriz de error. Si resolvemos exactamente el sistema $\hat{A}\hat{x} = b$, ¿Cuál será el error relativo en que hemos incurrido?

Nuestra situación es la de tener $\hat{A}\hat{x} = b = Ax = (\hat{A}-E)x = \hat{A}x - Ex$, de donde $\hat{A}(x - \hat{x}) = Ex$ y por tanto la acotación que acabamos de ver nos proporciona una acotación del error relativo en términos de algo que es aproximadamente el error relativo de los coeficientes,

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq c(\hat{A})\frac{\|E\|}{\|\hat{A}\|}\,.$$

Alternativamente podemos poner $(A + E)\hat{x} = Ax$ para escribir $A(x - \hat{x}) = E\hat{x}$ de donde el error relativo en los coeficientes nos da una acotación de una cantidad que es aproximadamente igual al error relativo en la solución,

$$\frac{\|x - \hat{x}\|}{\|\hat{x}\|} \leq c(A)\frac{\|E\|}{\|A\|}\,.$$

2.2.5 Estimación de la condición de una matriz

El cálculo de la condición de una matriz A tiene la dificultad de requerir calcular la matriz inversa, A^{-1}, lo cual es, en general, demasiado costoso. Resulta, pues, importante encontrar métodos de estimación o acotación de la condición. Uno de tales métodos se basa en el siguiente teorema que se deduce fácilmente de (2.4),

Teorema 13 *Si A es una matriz inversible entonces para cualquier matriz singular E se verifica*

$$\|A + E\| \geq \frac{1}{\|A^{-1}\|} \quad \textit{y por tanto} \quad c(A) \geq \frac{\|A\|}{\|A + E\|}\,.$$

Demostración: Por ser E singular existe un vector no nulo x tal que $Ex = 0$. Entonces para este vector, $(A + E)x = Ax$ y la conclusión se deduce de (2.4) con $B = A + E$ y $u = v = x$. ■

Ejercicio 2.17
Usar el resultado de este teorema para demostrar que toda matriz con diagonal estrictamente dominante *es inversible, es decir, no singular.*

Corolario 5 (Condición de matrices triangulares) *Si $A = (a_{ij})$ es una matriz triangular inversible entonces para cualquier elemento a_{ii} de su diagonal se verifica*

$$c_\infty(A) \geq \frac{\|A\|_\infty}{|a_{ii}|}\,,$$

en particular,

$$c_\infty(A) \geq \frac{\|A\|_\infty}{\min\limits_{1\leq i\leq n}|a_{ii}|}\,.$$

Demostración: Sea E la matriz que resulta al sustituir el elemento a_{jj} de A por cero. Entonces E es obviamente singular y por tanto

$$c_\infty(A) \geq \frac{\|A\|_\infty}{\|A - E\|_\infty} = \frac{\|A\|_\infty}{|a_{jj}|}\,.$$ ■

Ejercicio 2.18
Demostrar que las conclusiones de este corolario siguen siendo ciertas si se sustituye la norma infinito por la norma uno, obteniéndose una fórmula

de acotación de la condición, relativa a la norma uno, de cualquier matriz triangular.

Corolario 6 (Estimación de la condición de cualquier matriz) *Si $A = (a_{ij})$ es una matriz inversible de orden n entonces para cualquier $j \in \{1, \ldots, n\}$ se verifica*

$$c_\infty(A) \geq \frac{\|A\|_\infty}{\max\limits_{1\leq i\leq n} |a_{ij}|}$$

y en particular,

$$c_\infty(A) \geq \frac{\|A\|_\infty}{\min\limits_{1\leq j\leq n} \left(\max\limits_{1\leq i\leq n} |a_{ij}| \right)}.$$

Demostración: Sea E la matriz obtenida al sustituir todos los elementos de la columna j de la matriz A por cero. Obviamente E es una matriz singular y la matriz $A - E$ es una matriz todos cuyos elementos son cero excepto que la columna j es igual a la columna j de A. Por tanto tenemos $\|A - E\|_\infty = \max\limits_{1\leq i\leq n} |a_{ij}|$. En consecuencia, según el teorema 13,

$$c_\infty(A) \geq \frac{\|A\|_\infty}{\max\limits_{1\leq i\leq n} |a_{ij}|}. \quad \blacksquare$$

Ejercicio 2.19
Establecer, por analogía con el resultado anterior, una fórmula de acotación de la condición de cualquier matriz relativa a la norma uno.

2.2.6 Condición de las matrices ortogonales y sus múltiplos

Las matrices ortogonales tienen condición óptima para la norma euclídea ya que conservan esta norma y por tanto conservan sus bolas, lo cual, según el ejercicio 2.15, implica que el número de condición es 1. Pero asimismo, cualquier múltiplo de una matriz ortogonal conservará las bolas de la norma euclídea por lo que también tendrá condición 1 en esa norma.

2.3 Métodos iterativos

Los sistemas de ecuaciones lineales que surgen de ciertos problemas (principalmente en la resolución de ecuaciones diferenciales por métodos de diferencias finitas) tienen la característica de ser de órdenes excesivamente elevados para su tratamiento por métodos directos de resolución, los cuales exigirían la utilización de gran cantidad de memoria para almacenar todos los elementos de la matriz de los coeficientes del sistema. Por otro lado los coeficientes de tales sistemas suelen venir dados por una fórmula sencilla, de forma que pueden ser generados cuando se necesiten. Esto último ocurre también con las matrices *escasas* en las que una gran mayoría de elementos son cero. Tales sistemas suelen cumplir las condiciones de convergencia de métodos iterativos como los utilizados en la determinación del punto fijo de un operador (véase (19)), métodos que se pueden aplicar sin necesidad de tener todos los coeficientes del sistema almacenados. Así pues, para la resolución de algunos sistemas de ecuaciones lineales resultan más convenientes los métodos iterativos como los que estudiamos a continuación.

2.3.1 Ejemplo introductorio del método iterativo

Supongamos que nos dan el sistema de ecuaciones lineales

$$\begin{aligned} 10x + 2y - 3z &= 5 \\ 2x + 8y + z &= 6 \\ 3x - y + 15z &= 12 \end{aligned} \tag{2.5}$$

Ahora "despejamos" x de la primera ecuación, y de la segunda y z de la tercera obteniendo

$$\begin{aligned} x &= \frac{5 - 2y + 3z}{10}, \\ y &= \frac{6 - 2x - z}{8}, \\ z &= \frac{12 - 3x + y}{15}. \end{aligned}$$

Esto obviamente no nos proporciona una solución; es simplemente una reformulación del sistema de ecuaciones lineales dado, que ha tomado la forma general

$$\mathbf{x} = f(\mathbf{x}) \tag{2.6}$$

para f dada por $f(\mathbf{x}) = G\mathbf{x} + d$ donde

$$G = \begin{pmatrix} 0 & -\frac{2}{10} & \frac{3}{10} \\ -\frac{2}{8} & 0 & -\frac{1}{8} \\ -\frac{3}{15} & \frac{1}{15} & 0 \end{pmatrix} \quad \text{y} \quad d = \begin{pmatrix} \frac{5}{10} \\ \frac{6}{8} \\ \frac{12}{15} \end{pmatrix} .$$

La reformulación (2.6) indica que la solución del sistema (2.5) es un punto fijo de la función f. ¿Qué ocurrirá si aplicamos f a un vector $\mathbf{x}$ arbitrario? Supongamos que elegimos una estimación inicial arbitraria como por ejemplo

$$x^{(0)} = 0.5\,, \quad y^{(0)} = 1\,, \quad z^{(0)} = 1\,.$$

Aplicando f a este vector obtenemos otro vector $\mathbf{x}^{(1)} = f(\mathbf{x}^{(0)}) = C\mathbf{x}^{(0)} + d$ cuyas coordenadas son

$$x^{(1)} = 0.6\,, \quad y^{(1)} = 0.5\,, \quad z^{(1)} = 0.76\,.$$

Si ahora aplicamos f al resultado que acabamos de obtener llegamos a un vector $\mathbf{x}^{(2)}$ dado por

$$x^{(2)} = 0.63\,, \quad y^{(2)} = 0.504\,, \quad z^{(2)} = 0.7133\,,$$

y continuando de esta forma en la octava iteración se obtiene

$$x^{(8)} = 0.611831\,, \quad y^{(8)} = 0.508104\,, \quad z^{(8)} = 0.711507\,,$$

lo cual está muy cerca de la solución del sistema.

Este es el procedimiento básico de los métodos iterativos: construir una función contractiva cuyo punto fijo es la solución buscada e *iterar* repetidamente dicha función hasta acercarnos suficientemente al punto fijo. Decimos que se *itera* una función cuando se evalúa ésta en el resultado de haberla evaluado previamente, es decir, la evaluación de $f(f(x))$. Si se empieza con un valor $x^{(0)}$ las sucesivas iteraciones de f se obtienen siempre evaluando f en distintos valores, pero producen los mismos resultados que si aplicamos sucesivamente las funciones f, f^2, f^3, etc. a $x^{(0)}$.

Reformulado un sistema de ecuaciones lineales como un problema de punto fijo de una función contractiva (llamada *función de iteración*), por la teoría general de las funciones contractivas sabemos que sucesivas iteraciones de la función empezando con cualquier vector dará lugar a una sucesión que converge a la solución del sistema. Veamos a continuación un método general de obtener una función de iteración para un sistema de ecuaciones lineales dado para la cual es relativamente sencillo determinar si es contractiva o no.

2.3.2 Esquema general del Método Iterativo

No es difícil inventar formas de asociar a un sistema de ecuaciones lineales una función para la que la solución del sistema sea un punto fijo. Menos trivial es hacerlo de forma que la función obtenida sea contractiva. El siguiente método general tiene la ventaja de que permite confirmar de forma sencilla si la función asociada al sistema es contractiva y por tanto si el método iterativo converge.

Sea $Ax = b$ nuestro sistema de ecuaciones. Sea $A = A_1 + A_2$ una descomposición de la matriz de coeficientes A como suma de dos matrices de forma tal que A_1 es inversible.

Ejercicio 2.20
Si calculamos

$$G = -A_1^{-1}A_2 \quad y \quad d = A_1^{-1}b$$

entonces el sistema $Ax = b$ es equivalente a

$$x = Gx + d\,.$$

En consecuencia tenemos,

Si $A = A_1 + A_2$ y $G = -A_1^{-1}A_2$, $d = A_1^{-1}b$ entonces cada solución del sistema $Ax = b$ es un punto fijo de la función

$$f(x) = Gx + d$$

y viceversa.

El método general descrito en el ejercicio 2.20 da lugar a muchos métodos particulares cuando se eligen diversas formas de descomponer la matriz de coeficientes A como suma de dos matrices una de las cuales es inversible. Una de las formas más sencillas de hacer esto es llamada el método de *Jacobi*, que estudiaremos en breve.

Ejercicio 2.21
Indicar tres formas diferentes de descomponer una matriz cuadrada A sin ceros en la diagonal como suma de dos matrices una de las cuales es inversible. Describir el método iterativo de resolución de sistemas de ecuaciones lineales a que cada una de esas formas de descomposición da lugar.

2.3.3 Convergencia del método iterativo general

Sea $Ax = b$ un sistema de ecuaciones lineales y supongamos que tenemos una descomposición $A = A_1 + A_2$ de la matriz de coeficientes tal que A_1 es inversible, de modo que el sistema original es equivalente al problema de punto fijo $x = Gx + d$ donde $G = -A_1^{-1}A_2$ y $d = A_1^{-1}b$. Nos interesa saber bajo qué condiciones se puede resolver el problema de punto fijo mediante iteraciones de la función $f(x) = Gx + d$, es decir, nos interesa conocer las condiciones que garantizan que dado un vector $x^{(0)}$, la sucesión

$$\{x^{(k)}\} = \{f(x^{(0)}),\ f(f(x^{(0)})), \ldots,\ f^k(x^{(0)}), \ldots\} \tag{2.7}$$

converge al punto fijo de f. Pero nótese que si dicha sucesión converge a un vector x, la continuidad de f implica

$$x = \lim_{k\to\infty} x^{(k)} = \lim_{k\to\infty} f(x^{(k-1)}) = f(\lim_{k\to\infty} x^{(k-1)}) = f(x),$$

o sea, x es un punto fijo de f. O sea, que basta saber que la sucesión de iteraciones converge, pues si lo hace el límite es necesariamente la solución de nuestro sistema.

Supongamos, que $\{x^{(k)}\}$ converge y que su límite es x. Entonces la sucesión de "errores" $\{x^{(k)} - x\}$ converge a cero, pero teniendo en cuenta que

$$\begin{aligned} x^{(k)} - x &= G(x^{(k-1)}) + d - (G(x) + d) = G(x^{(k-1)} - x) = \cdots \\ &= G^k(x^{(0)} - x) \end{aligned}$$

vemos que

$$\lim_{k\to\infty} G^k(x^{(0)} - x) = 0.$$

De esto se deduce que si la sucesión (2.7) converge para cualquier elección del vector inicial $x^{(0)}$ entonces la sucesión de potencias de la matriz G tiende a la matriz cero. Trivialmente se cumple también el recíproco (hay que recordar que f *tiene* un punto fijo por hipótesis), de forma que tenemos:

Teorema 14 *Un método iterativo $x = Gx + d$ de resolución de un sistema de ecuaciones lineales converge para toda elección del vector inicial $x^{(0)}$ si y sólo si $G^n \to 0$, o sea, si y sólo si la sucesión de potencias de la matriz G tiende a la matriz cero.* ■

Una forma trivial de comprobar que $G^n \to 0$ sería el comprobar que para alguna norma se verifica $\|G\| < 1$ ya que en tal caso $\|G^n\| \leq \|G\|^n \to 0$. Por tanto es evidente que una sencilla condición suficiente de convergencia es:

Condición suficiente de convergencia de un método iterativo. *Si $\|G\|_\infty < 1$ entonces el método iterativo $x = Gx + d$ converge para toda elección del vector inicial $x^{(0)}$.*

En realidad no cuesta demasiado esfuerzo adicional el dar una visión completa de la situación demostrando el siguiente teorema que da las dos condiciones necesarias y suficientes fundamentales para la convergencia del método iterativo general:

Teorema 15 (Condiciones necesarias y suficientes) *Cada una de las condiciones enunciadas a continuación es necesaria y suficiente para que un método iterativo, $x = Gx + d$, de resolución de un sistema de ecuaciones lineales converja para toda elección del vector inicial $x^{(0)}$:*

1. *Que exista una norma para la que se verifique $\|G\| < 1$,*
2. *Que todo autovalor de G tenga valor absoluto menor que* 1.

Demostración: Como ya hemos observado más arriba, la condición 1 es suficiente. Por otro lado es inmediato que la condición 2 es necesaria ya que si G tiene un autovalor, λ, cuyo valor absoluto es $|\lambda| \geq 1$, este autovalor tendrá un autovector y (para el cual $G^k y = \lambda^k y$) que hace imposible que $G^k \to 0$ (ya que para todo k, $\|G^k\| \geq \frac{\|G^k y\|}{\|y\|} = |\lambda|^k \geq 1$). Así pues, la demostración del teorema queda completa con el siguiente lema:

Lema 3 *Si todo autovalor de G tiene valor absoluto menor que* 1 *entonces existe una norma para la que se verifica $\|G\| < 1$.*

Demostración: Si todo autovalor de G tiene valor absoluto menor que 1 entonces el radio espectral, $\rho(G)$, de G (el máximo de los valores absolutos de los autovalores de G) es menor que 1, por lo tanto existe un número real, ϵ, tal que $0 < \epsilon < 1 - \rho(G)$. Sea D la matriz diagonal $D = \mathrm{diag}(1, \epsilon^{-1}, \epsilon^{-2}, ..., \epsilon^{1-m})$. Sea J la forma canónica de Jordan de G y S la matriz de semejanza tal que $J = SGS^{-1}$. Para toda matriz X definimos

$$\|X\|_T = \|TXT^{-1}\|_\infty$$

donde $T = DS$. Basta demostrar dos cosas: (1) que $\|\cdot\|_T$ es una norma, y (2) que $\|G\|_T < \rho(G) + \epsilon$. La primera se deduce de que $\|\cdot\|_T$ es la norma matricial asociada a la norma vectorial $\|x\| = \|Tx\|_\infty$, y la segunda se comprueba con un sencillo cálculo teniendo en cuenta que en DJD^{-1} en cada bloque de Jordan quedan sustituidos los unos por ϵ. ■

Veamos ahora un criterio para la detención de las iteraciones. ¿Cuándo podemos considerar que ya hemos realizado suficientes iteraciones? Para poder detener las iteraciones de una forma eficaz es necesario poder estimar el error. El resultado siguiente nos permite hallar una cota superior de la magnitud del error en cada paso en términos de la corrección realizada en ese paso:

Proposición 3 *Sea $x^{(k+1)} = Gx^{(k)} + d$ un método iterativo dado. Para cualquier norma se verifica que en cada paso de iteración*

$$\|x^{(k)} - x\| \leq \frac{\|G\|}{1 - \|G\|}\|x^{(k)} - x^{(k-1)}\| .$$

Por lo tanto, si existe alguna norma para la que $\|G\| < \frac{1}{2}$ entonces el error en cada paso (medido con esa norma) está acotado superiormente por la corrección realizada en ese paso.

Demostración: Dado que $x^{(k)} - x = -G(x^{(k)} - x^{(k-1)}) + G(x^{(k)} - x)$, tomando normas obtenemos,

$$\|x^{(k)} - x\| \leq \|G\|\|x^{(k)} - x^{(k-1)}\| + \|G\|\|x^{(k)} - x\|$$

y despejando $\|x^{(k)} - x\|$ obtenemos

$$\|x^{(k)} - x\| \leq \frac{\|G\|}{1 - \|G\|}\|x^{(k)} - x^{(k-1)}\| ,$$

como queríamos demostrar. ■

2.3.4 Método de Jacobi

Estamos ahora en situación de volver al ejemplo ilustrativo de los métodos iterativos explicado al principio y analizarlo de forma un poco más teórica.

Descrito en forma general, el método en que se basaba aquel ejemplo es el siguiente:

Dado un sistema de ecuaciones lineales $Ax = b$, la ecuación i puede escribirse

$$\sum_{j=1}^{n} a_{ij}x_j = b_i .$$

Supongamos que los elementos diagonales de A sean todos distintos de cero. Entonces podemos despejar la primera incógnita de la primera ecuación, la segunda incógnita de la segunda ecuación, y así hasta despejar la última incógnita de la última ecuación de forma que el sistema queda re-escrito como

$$x_i = \Big(b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j \Big)\Big/ a_{ii} . \tag{2.8}$$

Esto lo podemos expresar fácilmente en forma matricial si observamos que

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j = \begin{pmatrix} 0 & a_{12} & \cdots & \cdots & a_{1n} \\ a_{21} & 0 & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 0 & a_{n-1\,n} \\ a_{n1} & \cdots & \cdots & a_{n\,n-1} & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = A_2 x ,$$

donde A_2 es la matriz obtenida a partir de A al hacer cero todos sus elementos diagonales. Además, si A_1 es "la diagonal de A", es decir, tal que $A = A_1 + A_2$, entonces el vector de coordenadas b_i/a_{ii} es simplemente $A_1^{-1}b$ y el vector de coordenadas $\big(\sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j\big)/a_{ii}$ (para $i = 1, \ldots, n$) es simplemente $A_1^{-1}A_2x$, de forma que la expresión (2.8) en forma matricial es

$$x = A_1^{-1}b - A_1^{-1}A_2x ,$$

es decir, la expresión (2.8) tiene la forma

$$\boxed{x = Gx + d} \qquad \text{con} \quad G = -A_1^{-1}A_2 , \text{ y } d = A_1^{-1}b .$$

Esta elección de descomposición de la matriz de coeficientes de un sistema de ecuaciones lineales en "diagonal" mas "el resto" para resolverlo

por el método iterativo, se conoce con el nombre de *Método de Jacobi* o de los *desplazamientos simultáneos*. Resumiendo,

Supongamos que la matriz de coeficientes A de nuestro sistema de ecuaciones lineales no tiene ningún cero en la diagonal. En tal situación, la matriz A_1 cuya diagonal es igual a la de A y cuyos otros elementos son cero es inversible y el método iterativo basado en la descomposición $A = A_1 + A_2$ se conoce como método de Jacobi.

Convergencia del método de Jacobi.–

Según la condición suficiente de convergencia establecida justo antes del teorema 15, el método de Jacobi será aplicable siempre que la matriz A de los coeficientes del sistema de ecuaciones lineales tenga la propiedad de que $\|D^{-1}(A-D)\|_\infty = \|D^{-1}A - I\|_\infty < 1$ donde D es la matriz "diagonal de A". Ahora bien,

$$\|D^{-1}A - I\|_\infty = \max_{1\le i\le n}\left\{\Big(\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\Big)\Big/|a_{ii}|\right\}$$

luego una condición suficiente para la convergencia del método de Jacobi es:

$$\max_{1\le i\le n}\left\{\Big(\sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\Big)\Big/|a_{ii}|\right\} < 1 \qquad \text{ó}\,, \qquad (\forall i)_{1\le i\le n}\ |a_{ii}| > \sum_{\substack{j=1\\ j\neq i}}^{n}|a_{ij}|\,.$$

Expresando esta condición con palabras:

Condición suficiente para la convergencia del método de Jacobi es que cada elemento de la diagonal de la matriz de coeficientes tenga un valor absoluto mayor que la suma de los valores absolutos de los demás elementos de su fila.

Las matrices descritas en la condición suficiente anterior se llaman *matrices de diagonal estrictamente dominante (por filas)*. Concluimos, pues, que el método de Jacobi es aplicable a cualquier sistema cuya matriz de coeficientes tiene la diagonal estrictamente dominante (por filas).

Obsérvese que toda matriz de diagonal estrictamente dominante (lo que implica que es no singular según el ejercicio 2.17) necesariamente tiene

todos los elementos de la diagonal distintos de cero y por tanto tiene sentido el plantearse usar el método de Jacobi.

2.3.5 Aceleración de la convergencia; método de Gauss-Seidel

La principal dificultad que plantean los métodos iterativos tales como el de desplazamientos simultáneos (o de Jacobi) es su lentitud en la convergencia. Por ello es necesario diseñar métodos de aceleración de la convergencia, El primero que estudiaremos es el de *desplazamientos sucesivos* (o de Gauss-Seidel). Más tarde estudiaremos los llamados métodos de *relajación* por su origen en el estudio de problemas mecánicos. Libros aconsejados: (17), (38) y (21).

El método de Jacobi se puede describir como basado en las ecuaciones (2.8), que nos dan las siguientes ecuaciones de iteración:

$$x_i^{(k+1)} = \Big(b_i - \sum_{\substack{j=1\\ j\neq i}}^{n} a_{ij}x_j^{(k)}\Big)\Big/ a_{ii}\,.$$

Según estas ecuaciones cada coordenada del paso $k+1$ se halla mediante las coordenadas del paso k. Pero obsérvese que habiendo calculado $x_1^{(k+1)}$, y dado que en principio $x_1^{(k+1)}$ está más cerca de la solución exacta que $x_1^{(k)}$, parece lógico pensar que sería más provechoso usar $x_1^{(k+1)}$ que $x_1^{(k)}$ en el cálculo de $x_2^{(k+1)}$. Asimismo sería más ventajoso usar $x_1^{(k+1)}$ y $x_2^{(k+1)}$ en lugar de $x_1^{(k)}$ y $x_2^{(k)}$ en el cálculo de $x_3^{(k+1)}$ y así sucesivamente. Esto sugiere el sustituir las ecuaciones de iteración anteriores por estas otras

$$x_i^{(k+1)} = \Big(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\Big)\Big/ a_{ii}\,. \qquad (2.9)$$

en las que en cada paso se usan para cada coordenada su estimación más actual.

Este método se conoce con el nombre de *Método de Gauss-Seidel* y puede llegar a aumentar la velocidad de convergencia en algunos casos a más del doble (respecto al método de Jacobi).

Las ecuaciones (2.9) pueden expresarse en forma matricial para descubrir a qué corresponden en el esquema general del método iterativo y poder así estudiar su convergencia basados en el teorema 15. Pasando todos los términos que hay en (2.9) con coordenadas del paso $k+1$ al miembro de la izquierda después de haber multiplicado por a_{ii} las ecuaciones (2.9) quedan

$$\sum_{j=1}^{i} a_{ij} x_j^{(k+1)} = b_i - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} ,$$

pero teniendo en cuenta que

$$\sum_{j=1}^{i} a_{ij} x_j^{(k+1)} = \begin{pmatrix} a_{11} & 0 & \cdots & \cdots & 0 \\ a_{21} & a_{22} & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & a_{n-1\,n-1} & 0 \\ a_{n1} & \cdots & \cdots & a_{n\,n-1} & a_{n\,n} \end{pmatrix} \begin{pmatrix} x_1^{(k+1)} \\ x_2^{(k+1)} \\ \vdots \\ x_n^{(k+1)} \end{pmatrix}$$

y

$$\sum_{j=i+1}^{n} a_{ij} x_j^{(k)} = \begin{pmatrix} 0 & a_{12} & \cdots & \cdots & a_{1n} \\ 0 & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & 0 & a_{n-1\,n} \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix} \begin{pmatrix} x_1^{(k)} \\ x_2^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix}$$

podemos escribir las ecuaciones anteriores como

$$A_1 x^{(k+1)} = b - A_2 x^{(k)} \qquad \text{ó} \qquad x^{(k+1)} = G x^{(k)} + d$$

donde A_1 es la matriz obtenida al hacer cero todos los elementos de A sobre la diagonal, y $A_2 = A - A_1$. En otras palabras, el método de Gauss-Seidel encaja en el esquema general de los métodos iterativos como el método obtenido con la descomposición

$$A = \begin{pmatrix} a_{11} & 0 & \cdots & \cdots & 0 \\ a_{21} & a_{22} & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & a_{n-1\,n-1} & 0 \\ a_{n1} & \cdots & \cdots & a_{n\,n-1} & a_{n\,n} \end{pmatrix} + \begin{pmatrix} 0 & a_{12} & \cdots & \cdots & a_{1n} \\ 0 & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ \vdots & & & 0 & a_{n-1\,n} \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix} .$$

Convergencia del método de Gauss-Seidel.–

Del teorema general de convergencia del método iterativo se deduce la siguiente condición suficiente de convergencia:

Teorema 16 (Convergencia del método de Gauss-Seidel) *El método iterativo de Gauss-Seidel para la resolución de un sistema de ecuaciones lineales converge para cualquier vector inicial siempre que la matriz de coeficientes del sistema sea de diagonal estrictamente dominante.*

Demostración: Sea $A = L + D + U$ la descomposición de la matriz de coeficientes en elementos bajo la diagonal (matriz L), elementos diagonales (matriz D) y elementos sobre la diagonal, de forma que la matriz de la función de iteración del método de Gauss-Seidel es $G = -(L+D)^{-1}U$ y la matriz de la función de iteración del método de Jacobi es $G_J = -D^{-1}(L+U)$, cumpliéndose $\|G_J\|_\infty < 1$. El teorema quedará demostrado si demostramos que $\|G\|_\infty \leq \|G_J\|_\infty$. Para ello basta demostrar que para cualquier vector x se cumple $\|Gx\|_\infty \leq \|G_J\|_\infty \|x\|_\infty$ o equivalentemente que cada componente y_k del vector $y = Gx = -(L+D)^{-1}Ux$ verifica

$$|y_k| \leq \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{|a_{ij}|}{|a_{ii}|} \|x\|_\infty . \tag{2.10}$$

Esto lo demostramos por inducción completa en las componentes. Primeramente, y teniendo en cuenta que $(L+D)y = -Ux$, expresamos el vector y en la forma $y = D^{-1}(-Ly - Ux)$ de donde se obtienen para las coordenadas de y fórmulas análogas a (2.9), a saber:

$$y_i = \Big(-\sum_{j=1}^{i-1} a_{ij} y_j - \sum_{j=i+1}^{n} a_{ij} x_j \Big) \Big/ a_{ii} .$$

De aquí se deduce (2.10) para $i = 1$ (¡ejercicio!). Supongamos ahora que $i > 1$ y que se cumple la hipótesis de inducción (2.10) para todo y_j con $j < i$. Entonces cada una de estas y_j también verifica $|y_j| \leq \|G\|_\infty \|x\|_\infty <$

$\|x\|_\infty$ y por lo tanto tenemos:

$$\begin{aligned}
|y_i| &\le \Big(\sum_{j=1}^{i-1}|a_{ij}||y_j| + \sum_{j=i+1}^{n}|a_{ij}||x_j|\Big)\Big/|a_{ii}| \\
&\le \Big(\sum_{j=1}^{i-1}|a_{ij}|\|x\|_\infty + \sum_{j=i+1}^{n}|a_{ij}|\|x\|_\infty\Big)\Big/|a_{ii}| \\
&\le \Big(\sum_{j=1}^{i-1}|a_{ij}| + \sum_{j=i+1}^{n}|a_{ij}|\Big)\|x\|_\infty\Big/|a_{ii}| \\
&\le \sum_{\substack{j=1\\ j\neq i}}^{n}\frac{|a_{ij}|}{|a_{ii}|}\|x\|_\infty
\end{aligned}$$

con lo que se completa la demostración. ■

Es interesante observar el hecho de que hay matrices (que necesariamente no son de diagonal estrictamente dominante ni por filas ni por columnas) para las que el método de Jacobi converge pero el de Gauss-Seidel no; y también hay matrices para las que el método de Gauss-Seidel converge pero el de Jacobi no.

2.3.6 Método de las relajaciones sucesivas

En los métodos iterativos que estamos estudiando, y en particular en el de Gauss-Seidel, cada paso de iteración puede considerarse como la realización de una corrección (también llamada *relajación*) a una aproximación de la solución. Visto desde este punto de vista podemos preguntarnos cuáles son las correcciones empleadas en cada paso del método de Gauss-Seidel y si hay alguna forma de mejorarlas. Evidentemente las correcciones son las diferencias $x_i^{(k+1)} - x_i^{(k)}$, que se encuentran fácilmente a partir de las fórmulas de iteración (2.9). De allí con poco esfuerzo se obtiene:

$$x_i^{(k+1)} - x_i^{(k)} = \Big(b_i - \sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n}a_{ij}x_j^{(k)}\Big)\Big/a_{ii}$$

(nótese que el índice inferior del segundo sumatorio ha cambiado de $i+1$ a i). Ahora bien, el hecho de que estas correcciones no proporcionen de inmediato la solución exacta indica que son (en valor absoluto) insuficientes

(caso de una convergencia monótona) o excesivas (caso de una convergencia alternada), y que la corrección exacta sería de la forma

$$c_i = \omega\Big(b_i - \sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n}a_{ij}x_j^{(k)}\Big)\Big/a_{ii}$$

para algún factor ω cercano a la unidad, que sería mayor o menor que la unidad según multiplicase a una corrección insuficiente o excesiva. Este factor se llama *coeficiente de relajación*. Si acertamos a elegir un valor apropiado para ω entonces el uso de las nuevas correcciones nos da las ecuaciones de iteración

$$x_i^{(k+1)} = x_i^{(k)} + \omega\Big(b_i - \sum_{j=1}^{i-1}a_{ij}x_j^{(k+1)} - \sum_{j=i}^{n}a_{ij}x_j^{(k)}\Big)\Big/a_{ii}$$

que pueden converger más rápidamente que (2.9). Este método de aceleración se conoce como *Método de las relajaciones sucesivas*.

El uso del término *relajación* surge del hecho de que gran parte de los trabajos iniciales sobre métodos para la solución iterativa de sistema de ecuaciones lineales estaban enfocados a la determinación de las fuerzas y momentos en estructuras mecánicas. Si se usan valores incorrectos de esas cantidades desconocidas, entonces es necesario aplicar en cada nodo de la estructura fuerzas restrictivas artificiales. A medida que uno se acerca a la solución correcta, esas fuerzas pueden ser "relajadas". Véase (17) p.92.

Ejercicio 2.22
Indicar la descomposición $A = A_1 + A_2$ con A_1 inversible que es necesario utilizar para que el esquema general de los métodos iterativos, $G = -A_1^{-1}A_2$, $d = A_1^{-1}b$, de lugar, con esta descomposición, al método de las relajaciones sucesivas, es decir, que se obtenga

$$G = (D+\omega L)^{-1}[(1-\omega)D - \omega U] \qquad y \qquad d = \omega(D+\omega L)^{-1}b$$

donde $A = L + D + U$ es la obvia descomposición en parte inferior, parte diagonal y parte superior.

La dificultad principal para poner en práctica el método de las relajaciones sucesivas consiste en realizar la elección adecuada del coeficiente ω. El valor óptimo de este coeficiente puede hallarse analíticamente sólo

en casos muy especiales y en general ha de hallarse a base de ensayos. Sin embargo es fácil demostrar, apoyados en el teorema 15 y en la expresión para G dada en el ejercicio 2.22, que los únicos valores de ω que pueden dar lugar a un método convergente son aquellos valores positivos menores que 2.

Proposición 4 *Para la matriz $G_\omega = (D+\omega L)^{-1}[(1-\omega)D-\omega U]$, donde D es diagonal y L, U son respectivamente triangulares inferior y superior con diagonal cero, se verifica que su radio espectral es mayor o igual que $|1-\omega|$ y en consecuencia para que un método iterativo de la forma*

$$x^{(k+1)} = G_\omega x^{(k)} + d$$

converja para cualquier valor inicial es condición necesaria que $|1-\omega| < 1$, es decir, $0 < \omega < 2$.

Demostración: Sean $\lambda_1, \lambda_2, \ldots, \lambda_n$ los autovalores de G_ω, entonces

$$\begin{aligned}\lambda_1 \cdots \lambda_n &= \det G_\omega = \det D^{-1} \det[(1-\omega)D] \\ &= (1-\omega)^n \det D^{-1} \det D = (1-\omega)^n\end{aligned}$$

de esto se deduce que $|1-\omega| \leq \max_i |\lambda_i|$. Como se quería demostrar. ■

Respuestas a Algunos Ejercicios del Capítulo 2

Ejercicio 2.1 Sin preocuparnos de posibles divisiones por cero, la base del algoritmo de eliminación es la siguiente

```
Datos: n, el orden del sistema,
A(i, j), la matriz de coeficientes y
B(i), la matriz de terminos independientes.
para k = 1 hasta n-1
para i = k+1 hasta n
m = A(i, k)/A(k, k)
para j = k hasta n
 A(i, j) = A(i, j) - m*A(k, j)
siguiente j
B(i) = B(i) - m*B(k)
siguiente i
siguiente k
```

Nótese que podemos ahorrarnos algunas operaciones en el bucle interno si evitamos calcular los ceros que se obtienen para $j = k$. Éstos se pueden introducir "manualmente" como en

```
para k = 1 hasta n-1
para i = k+1 hasta n
m = A(i, k)/A(k, k)
A(i, k) = 0
para j = k+1 hasta n
 A(i, j) = A(i, j) - m*A(k, j)
siguiente j
B(i) = B(i) - m*B(k)
siguiente i
siguiente k
```

Ejercicio 2.2 La sustitución regresiva se puede realizar mediante el siguiente algoritmo que se puede aplicar a sistemas singulares ya que detecta dicha condición.

```
si A(n, n) = 0 entonces
imprimir 'SISTEMA SINGULAR' y parar
X(n) = B(n)/A(n, n)
para k = n-1 hasta 1 incremento -1
si A(k, k) = 0 entonces
 imprimir 'SISTEMA SINGULAR' y parar
s = 0
para j = k+1 hasta n
 s = s + A(k, j)*X(j)
siguiente j
X(k) = (B(k) - s)/A(k, k)
siguiente k
```

Ejercicio 2.3 En el método de eliminación de Gauss hay dos partes: (1) Eliminación de los elementos bajo la diagonal y (2) sustitución regresiva.

En la primera parte, para eliminar el elemento en posición (k, l) es necesaria una división para calcular el multiplicador y, si el sistema es de orden n, $n-l+1$ multiplicaciones para multiplicar cada uno de los elementos $a_{k\,l+1}, a_{k\,l+2}, \ldots, a_{k\,n}, b_k$ por el multiplicador. Total $n-l+2$ operaciones para el elemento (k, l). Así pues, el número total de operaciones en esta parte es

$$\begin{aligned}
\sum_{l=1}^{n-1}\sum_{k=l+1}^{n}(n-l+2) &= \sum_{l=1}^{n-1}(n-l)(n-l+2)\\
&= (n-1)(n+1)+(n-2)n+(n-3)(n-1)+\cdots+1\cdot 3\\
&= 1\cdot(2+1)+2\cdot(2+2)+\cdots+(n-1)(2+n-1)\\
&= 2(1+2+\cdots+(n-1))+(1^2+2^2+\cdots+(n-1)^2)\\
&= 2\frac{n(n-1)}{2}+\tfrac{1}{6}(n-1)n(2(n-1)+1)\\
&= n^2-n+\tfrac{1}{6}(n^2-n)(2n-1)\\
&= n^2-n+\tfrac{1}{6}(2n^3-3n^2-n)\\
&= \tfrac{1}{3}n^3+\tfrac{1}{2}n^2-\tfrac{5}{6}n.
\end{aligned}$$

a esto hay que sumarle las operaciones realizadas durante la sustitución regresiva, que son: una división para la última ecuación, y en general una división y k multiplicaciones para la ecuación $n-k$ ($k=0,\ldots,n-1$), o sea $\sum_{k=0}^{n-1}(k+1)=\frac{1}{2}n(n+1)=\frac{n^2}{2}+\frac{n}{2}$. Así pues:

Operaciones en Eliminación de Gauss: $\boxed{\frac{1}{3}n^3+n^2-\frac{1}{3}n}$.

Este mismo resultado puede obtenerse de forma mucho menos trabajosa si se sabe que la función que buscamos es un polinomio en n de tercer grado, es decir de la forma $f(n)=an^3+bn^2+cn+d$. Entonces lo único que tenemos que hacer para averiguar los coeficientes a, b, c, d es resolver un sistema de cuatro ecuaciones y cuatro incógnitas que resulta al evaluar f respectivamente en 0, 1, 2, y 3: $f(0)=0$, $f(1)=1$, $f(2)=6$, $f(3)=17$, (esto se halla contando directamente el número de multiplicaciones y divisiones necesarios para resolver los sistemas de órdenes cero, uno, dos y tres) lo cual nos da el sistema de ecuaciones

$$\begin{pmatrix}0&0&0&1\\1&1&1&1\\8&4&2&1\\27&9&3&1\end{pmatrix}\begin{pmatrix}a\\b\\c\\d\end{pmatrix}=\begin{pmatrix}0\\1\\6\\17\end{pmatrix}$$

cuya solución es $a=\frac{1}{3}$, $b=1$, $c=-\frac{1}{3}$, $d=0$. Otra observación útil para este problema es que la función $f(n)$ puede definirse recursivamente a partir del valor inicial $f(0)=0$ por $f(n)=f(n-1)+(n^2+n-1)$. Con este método el cómputo se simplifica al reducirse a contar las operaciones necesarias para eliminar la primera columna y para el último paso de la sustitución regresiva (lo que nos da el sumando (n^2+n-1) de la fórmula de $f(n)$).

Ejercicio 2.4

Ejercicio 2.5 El siguiente algoritmo efectúa la descomposición triangular de una matriz. Suponemos dados: n, el orden del sistema, y $A(i, j)$, la matriz de coeficientes. La solución se almacenará en las variables $p(i)$, $L(i, j)$, y $U(i, j)$. Empezamos haciendo U igual a A y L igual a la matriz identidad.

```
para i = 1 hasta n
para j = 1 hasta n
U(i, j) = A(i, j); L(i, j) = 0
si i = j entonces L(i, j) = 1
siguiente j
siguiente i
para k = 1 hasta n-1
c = abs(U(k, k)); p(k) = k
para i = k+1 hasta n
```

```
si abs(U(i, k)) > c entonces c = abs(U(i, k)) y p(k) = i
siguiente i
si c = 0 entonces imprimir 'MATRIZ SINGULAR' y parar
para j = k hasta n
t = U(p(k), j); U(p(k), j) = U(k, j); U(k, j) = t
siguiente j
para i = k+1 hasta n
L(i, k) = U(i, k)/U(k, k); U(i, k) = 0
para j = k+1 hasta n
U(i, j) = U(i, j) - L(i, k)*U(k, j)
siguiente j
siguiente i
siguiente k
si U(n, n) = 0 entonces imprimir 'MATRIZ SINGULAR'
```

Ejercicio 2.6 En el método de Gauss-Jordan hay una parte de eliminación y una parte de sustitución. La segunda consta solamente de n divisiones mientras que en la primera se usan $n-l+2$ operaciones para eliminar el elemento (k,l) donde k toma $n-1$ valores y l varía desde la columna 1 hasta la n. Así pues, el número total de operaciones en esta parte es

$$\begin{aligned}\sum_{l=1}^{n}(n-1)(n-l+2) &= (n-1)\sum_{l=1}^{n}(n-l+2)\\ &= (n-1)[(n+1)+n+(n-1)+\cdots+2]\\ &= (n-1)(\tfrac{1}{2}(n+1)(n+2)-1)\\ &= (n-1)(\tfrac{1}{2}n^2+\tfrac{3}{2}n+1-1) = \tfrac{1}{2}n^3+n^2-\tfrac{3}{2}n.\end{aligned}$$

y sumando a éstas las n operaciones de la segunda parte, tenemos:

Operaciones en Eliminación de Gauss-Jordan: $\boxed{\frac{1}{2}n^3+n^2-\frac{1}{2}n}$.

De nuevo, al mismo resultado se llega resolviendo el sistema de ecuaciones

$$\begin{pmatrix}0&0&0&1\\1&1&1&1\\8&4&2&1\\27&9&3&1\end{pmatrix}\begin{pmatrix}a\\b\\c\\d\end{pmatrix}=\begin{pmatrix}0\\1\\7\\21\end{pmatrix}$$

satisfecho por los coeficientes de $g(n)=an^3+bn^2+cn+d$.

Ejercicio 2.7

Ejercicio 2.8 A partir de $x^{(n+1)}=x^{(n)}+C(b-Ax^{(n)})=x^{(n)}+CA(x-x^{(n)})$ se deduce $(CA)^{-1}(x^{(n+1)}-x^{(n)})=x-x^{(n)}$, de donde

$$\|x^{(n)}-x\|\le\|(CA)^{-1}\|\|x^{(n+1)}-x^{(n)}\|\,.$$

Ejercicio 2.9 Evidente. La última conclusión se obtiene tomando $u=v\neq 0$.

Ejercicio 2.10 La desigualdad de la derecha es consecuencia inmediata de la del ejercicio anterior. La desigualdad de la izquierda se obtiene de la de la derecha intercambiando u y v y tomando inversos.

Ejercicio 2.11 Si A es un operador lineal en un espacio normado, la obvia inclusión de conjuntos de números

$$E=\{\|Ax\| \mid \|x\|=1\}\subset\left\{\frac{\|Ax\|}{\|x\|}\mid x\neq 0\right\}=F$$

es en realidad una igualdad ya que si $x\neq 0$ entonces tomando $y=kx$ con $k=1/\|x\|$, su norma es $\|y\|=1$ y por lo tanto

$$\frac{\|Ax\|}{\|x\|}=k\|Ax\|=\|k\cdot Ax\|=\|A(kx)\|=\|Ay\|\in E,$$

o sea $F\subset E$. Así pues $E=F$, de donde

$$\max_{x\neq 0}\frac{\|Ax\|}{\|x\|}=\max F=\max E=\max_{\|x\|=1}\|Ax\|.$$

De las dos fórmulas dadas a continuación, la fórmula de la izquierda dice que $\|A\|$ es mayor o igual que el máximo de E, mientras que la de la derecha dice que es menor o igual que ese máximo.

Ejercicio 2.12 $\|I\|=\max_{x\neq 0}\frac{\|Ix\|}{\|x\|}=\max_{x\neq 0}\frac{\|x\|}{\|x\|}=1$. Por ejemplo el otro axioma adicional es consecuencia de que para todo x,

$$\|ABx\|\le\|A\|\|Bx\|\le\|A\|\|B\|\|x\|\,.$$

Usando estas dos propiedades, $1=\|I\|=\|AA^{-1}\|\le\|A\|\|A^{-1}\|$.

Ejercicio 2.13 Primeramente tenemos para todo vector x,

$$\begin{aligned}\|Ax\|_1&=\sum_{i=1}^{n}\Big|\sum_{j=1}^{n}a_{ij}x_j\Big|\le\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}||x_j|=\sum_{j=1}^{n}\Big(\sum_{i=1}^{n}|a_{ij}|\Big)|x_j|\\ &\le\Big(\max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}|\Big)\|x\|_1\,.\end{aligned}$$

Por otro lado, sea k una columna de A para la que la suma de los valores absolutos de sus elementos es máxima, es decir que

$$\max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}| = \sum_{i=1}^{n}|a_{ik}|.$$

Entonces para cualquier vector $x = (0,\dots,0,x_k,0,\dots,0)$ todas cuyas componentes sean cero excepto la del lugar k, se verifica (ya que $\|x\|_1 = |x_k|$)

$$\|Ax\|_1 = \sum_{i=1}^{n}|a_{ik}x_k| = \Big(\sum_{i=1}^{n}|a_{ik}|\Big)|x_k| = \Big(\max_{1\le j\le n}\sum_{i=1}^{n}|a_{ij}|\Big)\|x\|_1 .$$

En consecuencia

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

Ejercicio 2.14 Sabemos que $c(A) = \|A\|\|A^{-1}\|$. Por otro lado, según el ejercicio 2.11, $\|A\| = \max_{\|x\|=1}\|Ax\|$, y análogamente

$$\|A^{-1}\| = \max_{y\neq 0}\frac{\|A^{-1}y\|}{\|y\|} = \max_{x\neq 0}\frac{\|x\|}{\|Ax\|} = \frac{1}{\min_{x\neq 0}\frac{\|Ax\|}{\|x\|}} = \frac{1}{\min_{\|x\|=1}\|Ax\|}$$

por lo tanto

$$c(A) = \frac{\max_{\|x\|=1}\|Ax\|}{\min_{\|x\|=1}\|Ax\|}$$

Ejercicio 2.15 $c(A) = 1$ significa que

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

y por lo tanto $\|Ax\| = \text{const.}$ para $\|x\| = 1$. En otras palabras, A conserva las bolas de la norma en cuestión, lo cual es decir que conserva la norma salvo un factor constante.

Ejercicio 2.16 Si nos situamos en el plano es geométricamente evidente que las matrices A que conservan una p-norma para $p \neq 2$ son precisamente las matrices del grupo de simetrías del cuadrado y si $p = 2$ las del grupo de simetrías del círculo.

Por otro lado, si A tiene condicionamiento 1 respecto a la norma euclídea entonces puede no conservar las p-normas para $p \neq 2$, pero la imagen de la bola unitaria de una p-norma siempre estará contenida dentro del círculo de radio $\sqrt{2}$ y contendrá al círculo de radio $\frac{1}{2}\sqrt{2}$. En consecuencia, para cualquier x tal que $\|x\|_p = 1$ tendremos $\frac{1}{2}\sqrt{2} \le \|Ax\|_p \le \sqrt{2}$ y por lo tanto la condición de A relativa a la norma p verifica $c_p(A) \le \dfrac{\sqrt{2}}{\frac{1}{2}\sqrt{2}} = 2$.

Ejercicio 2.17 Sea A una matriz cuadrada y sea D la matriz diagonal cuya diagonal es igual a la diagonal de A. Suponemos que los elementos de esa diagonal son todos no nulos (en caso contrario A no puede ser de diagonal estrictamente dominante). Dividiendo cada elemento de A por el valor absoluto del elemento diagonal de su fila obtenemos una matriz B que es $B = D^{-1}A$, es decir, $A = DB$. Basta demostrar que si A es de diagonal estrictamente dominante entonces B es inversible. Pero decir que A es de diagonal estrictamente dominante es equivalente a decir que $\|B-I\|_\infty < 1$, lo cual, por el teorema 13 implica que $-B$ no es singular.

Ejercicio 2.18

Ejercicio 2.19

Ejercicio 2.20 $Gx + d = -A_1^{-1}A_2x + A_1^{-1}b = -A_1^{-1}(A_2x - b)$ luego $x = Gx + d$ es equivalente a $-A_1x = A_2x - b$, que a su vez es equivalente a $b = A_1x + A_2x = Ax$.

Ejercicio 2.21 Una descomposición que siempre es posible se obtiene al elegir $A_1 = I$ (identidad). En el caso de una matriz sin ceros en la diagonal también $A_1 = D = \text{diag}(A)$ es inversible, así como la matriz triangular obtenida al hacer cero en A los elementos que están encima (o debajo) de la diagonal.

Cada una de las descomposiciones anteriores da lugar a un método iterativo:

1. En este caso $G = -I^{-1}(A - I) = I - A$ y $d = I^{-1}b = b$, luego el método iterativo para resolver $Ax = b$ asociado con esta descomposición es hallar un punto fijo de la función $(I-A)x+b = x-Ax+b$. Efectivamente tal punto fijo verifica $x = x - Ax + b$ que es equivalente a $Ax = b$.

2. En el segundo caso tenemos el método de Jacobi.

3. En el tercer caso tenemos el método de Gauss-Seidel.

Ejercicio 2.22 Si tomásemos $A_1 = -(D+\omega L)$ y $A_2 = (1-\omega)D - \omega U$ tendríamos

$$A_1+A_2 = -(D+\omega L)+(1-\omega)D-\omega U = -D-\omega L+D-\omega D-\omega U = -\omega A$$

mientras que lo que queremos es $A_1 + A_2 = A$. Por tanto tenemos que dividir los valores que habíamos dado a A_1 y a A_2 por $-\omega$ y entonces tomaremos

$$A_1 = (D+\omega L)/\omega = L + \tfrac{1}{\omega}D$$

y

$$A_2 = -[(1-\omega)D - \omega U]/\omega = U - \left(\tfrac{1-\omega}{\omega}\right)D\,.$$

De esta forma tenemos $-A_1^{-1}A_2 = -\omega(D+\omega L)^{-1}(U - \left(\frac{1-\omega}{\omega}\right)D) = -(D+\omega L)^{-1}(\omega U - (1-\omega)D) = G$, y $A_1 + A_2 = A$. por lo tanto

$$d = A_1^{-1}b = \boxed{\omega(D+\omega L)^{-1}b}\,.$$

Capítulo 3

Cálculo de vectores y valores propios

El problema general de encontrar todos los valores y vectores propios de una matriz cuadrada es bastante amplio y su estudio podría llenar fácilmente un trimestre completo. En este capítulo haremos solamente una mera introducción al tema. Veremos el *método de la potencia* para el cálculo del autovalor de mayor valor absoluto de una matriz que tenga un autovalor dominante y, tras explicar la transformación, mediante matrices de Householder, de una matriz en otra semejante a ella pero de tipo Hessenberg, se explicará el llamado método de la *factorización ortogonal* (o método QR), el cual es el método general más eficaz y de mayor uso para el cálculo de todos los valores propios de una matriz arbitraria. Este método, descubierto por J. G. F. Francis y publicado en 1961, ofrece ciertas dificultades computacionales que se pueden reducir si se prepara previamente la matriz dada A reduciéndola a una forma de tipo *Hessenberg*, proceso que, por conservar la simetría, da lugar a una matriz tridiagonal semejante a A si ésta es simétrica.

3.1 Método de la potencia

El método de la potencia es un método iterativo de cálculo del autovalor de máximo valor absoluto de una matriz diagonalizable. Está basado en la siguiente observación:

Supongamos que A es una matriz diagonalizable de orden n, $x^{(0)}$ un vector arbitrario y $\{x^{(k)}\}$ la sucesión de vectores definida recurrentemente por:

$$x^{(k)} = Ax^{(k-1)} = A^k x^{(0)}.$$

Por la hipótesis de que A es diagonalizable existe una base, que denotaremos $\{u_1, u_2, \ldots, u_n\}$, formada por vectores propios de A, que podemos suponer dados en orden decreciente de magnitud de autovalores, es decir, si λ_i es el autovalor correspondiente al vector propio u_i,

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

Entonces, si $x^{(0)} = \alpha_1 u_1 + \cdots + \alpha_n u_n$, el término general de la sucesión anterior tiene la siguiente expresión en la base $\{u_i\}$ de vectores propios de A:

$$\begin{aligned} x^{(k)} &= A^k(\alpha_1 u_1 + \alpha_2 u_2 + \cdots + \alpha_n u_n) \\ &= \alpha_1 \lambda_1^k u_1 + \alpha_2 \lambda_2^k u_2 + \cdots + \alpha_n \lambda_n^k u_n \\ &= \lambda_1^k \left(\alpha_1 u_1 + \alpha_2 \left(\tfrac{\lambda_2}{\lambda_1}\right)^k u_2 + \cdots + \alpha_n \left(\tfrac{\lambda_n}{\lambda_1}\right)^k u_n \right). \end{aligned}$$

Supongamos ahora por un momento que el vector $x^{(0)}$ no es arbitrario como hemos dicho sino que tiene la propiedad de que $\alpha_1 \neq 0$ [1]. Si, además, la matriz A es tal que tiene un *autovalor dominante* (es decir, $\lambda_1 \neq \lambda_2$, o sea, la desigualdad $|\lambda_1| \geq |\lambda_2|$ es estricta de modo que $|\lambda_1| > |\lambda_2|$) entonces tenemos

$$\lim_{k\to\infty} \frac{1}{\lambda_1^k} x^{(k)} = \alpha_1 u_1,$$

[1] Esta hipótesis no se puede comprobar de antemano porque se desconocen los vectores u_i. Sin embargo no es una hipótesis demasiado restrictiva porque, por un lado, la probabilidad de que un vector elegido al azar tenga $\alpha_1 = 0$ es cero y, por otro lado, aunque fuese $\alpha_1 = 0$, a lo largo de los cálculos los errores de redondeo acabarían por introducir una componente no nula en la dirección de u_1.

y por lo tanto para cualquier índice $i \in \{1, \ldots, n\}$,

$$1 = \lim_{k\to\infty} \left(\frac{1}{\lambda_1^k} x_i^{(k)}\right) \Big/ \left(\frac{1}{\lambda_1^{k-1}} x_i^{(k-1)}\right) = \frac{1}{\lambda_1} \lim_{k\to\infty} \frac{x_i^{(k)}}{x_i^{(k-1)}},$$

de donde,

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

con lo que tenemos una sucesión que converge al autovalor dominante de la matriz A.

Si ya conocemos el signo del autovalor buscado (o su argumento, en caso de que sea complejo) sólo necesitamos hallar su valor absoluto. Para esto podemos obtener una sucesión que tiene mejor velocidad de convergencia que la anterior ya que

$$\lim_{k\to\infty} \frac{\|A^k x^{(0)}\|}{\|A^{k-1} x^{(0)}\|} = \lim_{k\to\infty} \frac{\|x^{(k)}\|}{\|x^{(k-1)}\|} = |\lambda_1|.$$

Lo que acabamos de decir significa que si A es una matriz cuadrada de orden n con un autovalor dominante, éste se puede hallar mediante el siguiente proceso iterativo:

Algoritmo del Método de la Potencia

1 Elegir un vector inicial $v \neq 0$ y una tolerancia de error T. Además, inicializar la variable $\lambda_0 = 0$.
2 Calcular el producto $y = Av$.
3 Hallar j tal que $|y_j| = \max_{1\leq i\leq n} |y_i|$.
4 Asignar $S = |y_j|$; $\lambda = y_j/v_j$; $v = y/S$.
5 Si $|\lambda_0 - \lambda| \leq T$ entonces PARAR.
6 Asignar $\lambda_0 = \lambda$ e ir al paso **2**.

3.2 Método QR o de la factorización ortogonal

Cuando se desean conocer todos los autovalores de una matriz de orden grande, el cálculo de su polinomio característico y subsiguiente estimación

de raíces no es siempre el método más sencillo o eficaz. El método recomendado en general es el método de la factorización ortogonal que se basa en los siguientes teoremas:

Teorema 17 (Exist. y unicid. de la factorización ortogonal) *Para toda matriz real A existen matrices Q, ortogonal, y R, triangular superior, tales que $A = QR$. Además, si A es no singular entonces estas matrices están unívocamente determinadas por A excepto en el signo de los elementos.*

Demostración: *Existencia*: Esto se demostrará en la sección siguiente en la que se da un método de construcción de las matrices Q y R.
Unicidad: En primer lugar, la matriz identidad no admite más factorización ortogonal $I = QR$ que la trivial, o sea, aquella con $Q = I = R$ ya que $Q^{-1} = R$ nos dice que R es una matriz ortogonal triangular, lo que implica que es la identidad salvo, quizás por los signos de los elementos diagonales. En segundo lugar, si $A = QR = Q'R'$ entonces $I = Q^{-1}Q'R'R^{-1}$, de donde $R'R^{-1} = S$, una matriz de signos, que solo difiere de la identidad en los signos de sus elementos y por tanto $S = S^{-1} = Q^{-1}Q'$ lo que implica $R' = SR$ y $Q' = QS$. ■

3.2.1 La Sucesión General del Método de la Factorización Ortogonal

Una vez que determinemos un método de construir una factorización ortogonal de una matriz cuadrada cualquiera, podemos utilizarlo para construir la siguiente sucesión definida recursivamente a partir de una matriz cuadrada cualquiera A:

$$A^{(0)} = A, \quad A^{(k+1)} = R^{(k)}Q^{(k)}$$

donde $R^{(k)}$ y $Q^{(k)}$ se obtienen mediante la factorización ortogonal

$$A^{(k)} = Q^{(k)}R^{(k)}$$

de la matriz $A^{(k)}$

Llamaremos a esta sucesión *la sucesión general del método de la factorización ortogonal determinada por A*. Esta sucesión goza de algunas propiedades evidentes pero de gran importancia:

1. Dado que dos términos consecutivos están relacionados por:

$$A^{(k+1)} = \left(Q^{(k)}\right)^{-1}A^{(k)}Q^{(k)},$$

Todas las matrices de esta sucesión son semejantes entre sí y por tanto todas tienen los mismos autovalores y éstos son también los autovalores de su límite

2. Dado que la inversa de una matriz ortogonal es su traspuesta, tenemos

$$\left(A^{(k+1)}\right)^t = \left(Q^{(k)}\right)^t\left(A^{(k)}\right)^tQ^{(k)},$$

y por tanto *si una matriz de esta sucesión es simétrica entonces todas lo son*.

Visto esto podemos enunciar el teorema en el que se basa el método de la factorización ortogonal:

Teorema 18 (Método QR para cálculo de autovalores) *Sea A una matriz real cuadrada no singular entre cuyos autovalores no hay dos con el mismo valor absoluto, es decir, que verifican*

$$|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0.$$

Entonces la sucesión general del método de la factorización ortogonal determinada por A converge a una matriz triangular. Si la matriz A es simétrica, el límite es una matriz diagonal. ■

Así pues, para una matriz A cuyos autovalores tengan todos distinto valor absoluto, estos son los elementos diagonales límite de la *sucesión general del método de la factorización ortogonal* determinada por A.

Ejercicio 3.1
Demostrar que la sucesión general del método de la factorización ortogonal *determinada por $A = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ no converge a una matriz triangular. ¿Cuál de las hipótesis del teorema anterior no cumple esta matriz?*

3.2.2 La factorización ortogonal de una matriz: Método de las rotaciones planas

La factorización ortogonal de una matriz puede realizarse por un método que es formalmente parecido al método de eliminación de Gauss utilizado

para realizar la factorización triangular. Al igual que en aquél la idea es aplicar sucesivas transformaciones a la matriz de partida para ir haciendo cero los elementos bajo la diagonal de forma que al final quede transformada en una matriz triangular superior. La diferencia es que si allí se utilizaban transformaciones elementales (que combinadas resultaban en una matriz triangular inferior con unos en la diagonal), aquí se utilizarán transformaciones ortogonales (concretamente rotaciones planas) que combinadas darán lugar a una matriz ortogonal.

Una rotación plana en el espacio n-dimensional queda determinada por el plano de rotación y el ángulo de rotación. Si el plano de rotación es el determinado por los ejes de coordenadas 1 y 2 y el ángulo de rotación es α entonces la matriz de rotación tiene la forma

$$P_\alpha(1,2) = \begin{pmatrix} \cos\alpha & -\operatorname{sen}\alpha & & & \\ \operatorname{sen}\alpha & \cos\alpha & & 0 & \\ & & 1 & & \\ & 0 & & \ddots & \\ & & & & 1 \end{pmatrix}$$

y en general una rotación sobre el plano de los ejes i y j (con $i > j$) tiene una matriz de la forma (escribiendo sólo los elementos no nulos)

$$\begin{array}{l} \\ \text{fila } j \to \\ \\ \\ \text{fila } i \to \\ \\ \\ \\ \end{array} \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & \cdots & -s & & & \\ & & & & 1 & & & & \\ & & & \vdots & \ddots & \vdots & & & \\ & & & & & 1 & & & \\ & & & s & \cdots & c & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix} = P_\alpha(i,j)$$

donde c y s son los cosenos directores de la rotación, es decir, el coseno y seno respectivamente del ángulo α de rotación.

Si queremos hallar los cosenos directores c, s de la matriz $P(i,j)$ para hacer cero el elemento (i,j) de la matriz $A = (a_{i,j})$ no tenemos más que

considerar la ecuación obtenida al multiplicar $P(i,j)$ por la columna j de A, es decir,

$$\begin{array}{l} \\ j \to \\ \\ \\ i \to \\ \\ \\ \\ \end{array} \begin{pmatrix} 1 & & & & & & & & \\ & \ddots & & & & & & & \\ & & 1 & & & & & & \\ & & & c & \cdots & -s & & & \\ & & & & 1 & & & & \\ & & & \vdots & \ddots & \vdots & & & \\ & & & & & 1 & & & \\ & & & s & \cdots & c & & & \\ & & & & & & 1 & & \\ & & & & & & & \ddots & \\ & & & & & & & & 1 \end{pmatrix} \begin{pmatrix} a_{1j} \\ \vdots \\ a_{jj} \\ \vdots \\ a_{ij} \\ \vdots \\ a_{nj} \end{pmatrix} = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{j-1\,j} \\ ca_{jj} - sa_{ij} \\ a_{j+1\,j} \\ \vdots \\ a_{i-1\,j} \\ sa_{jj} + ca_{ij} \\ a_{i+1\,j} \\ \vdots \\ a_{nj} \end{pmatrix},$$

de donde se deduce que los cosenos directores c, s han de verificar

$$c^2 + s^2 = 1 \quad \text{y} \quad sa_{jj} + ca_{ij} = 0,$$

y una posible solución de esto es

$$c = \frac{a_{jj}}{\sqrt{a_{jj}^2 + a_{ij}^2}}, \qquad s = \frac{-a_{ij}}{\sqrt{a_{jj}^2 + a_{ij}^2}} \tag{3.1}$$

Teniendo en cuenta lo que acabamos de decir el proceso de eliminación para la factorización ortogonal de una matriz A se lleva a cabo de la siguiente forma: En un primer paso se hacen cero los elementos de la primera columna bajo la diagonal mediante multiplicación sucesiva de la matriz A por matrices de rotación plana sobre los ejes $(1,2), (1,3), \ldots$. Es decir obtenemos $Q_1 A$ donde $Q_1 = P(1,n)P(1,n-1)\cdots P(1,2)$. Después se hacen cero los elementos de la segunda columna bajo la diagonal multiplicando la matriz $Q_1 A$ por matrices de rotación plana sobre los ejes $(2,3), (2,4), \ldots$. Así obtenemos una matriz $Q_2 Q_1 A$ donde $Q_2 = P(2,n)P(2,n-1)\cdots P(2,3)$. Continuando de esta forma se obtiene una matriz triangular superior

$$R = Q_{n-1}\cdots Q_1 A$$

donde todas las matrices Q_i son matrices ortogonales por ser producto de matrices ortogonales. Con esto queda demostrada la existencia de la factorización ortogonal.

Ejercicio 3.2
Hallar el número de matrices de rotación plana que es necesario multiplicar para hallar la factorización de una matriz arbitraria de orden n.

3.2.3 Simplificación que ocurre para las matrices tipo Hessenberg

Supongamos que la matriz A es de tipo Hessenberg, es decir tiene ceros bajo la subdiagonal. Entonces su factorización ortogonal puede llevarse a cabo con solamente $n-1$ rotaciones planas. Las Q_i constan cada una de una sola rotación. Además:

Lema 4 *Si R es una matriz triangular superior y para cada $i \in \{1, \ldots, n-1\}$ $Q_i = P(i, i+1)$ (una rotación sobre el plano de los ejes i e $i+1$) entonces la matriz $RQ_1 \cdots Q_{n-1}$ es una matriz tipo Hessenberg.*

Demostración: Por inducción. ■

La consecuencia de este lema es que si el algoritmo QR es aplicado a una matriz Hessenberg entonces todas las matrices de la sucesión obtenida son de tipo Hessenberg. Veremos ahora cómo podemos transformar por semejanza una matriz arbitraria en otra de tipo Hessenberg.

El resultado del ejercicio 3.2 muestra el enorme coste que tendría la aplicación del método QR a una matriz arbitraria. Afortunadamente es posible reducir drásticamente ese coste gracias a dos cosas: (1) Según acabamos de ver, el algoritmo QR conserva la forma Hessenberg de una matriz (es decir, el tener ceros bajo la subdiagonal), y (2) Toda matriz puede transformarse por semejanza en otra tipo Hessenberg. Veremos cada uno de estos dos hechos a continuación.

3.2.4 Método de las reflexiones de Householder para el cálculo de una matriz Hessenberg semejante a una dada

En esta sección haremos uso del producto interior natural en $\mathbf{R}^n$. Si representamos los vectores en una base ortonormal mediante matrices columna, el producto interior de $\mathbf{x}$ e $\mathbf{y}$, normalmente denotado $\mathbf{x} \cdot \mathbf{y}$ ó $\langle \mathbf{x}, \mathbf{y} \rangle$, es igual al producto de matrices $\mathbf{x}^t\mathbf{y}$ y también igual a $\mathbf{y}^t\mathbf{x}$.

Un vector unitario $\mathbf{w}$ determina un subespacio ortogonal a él y, para cada vector $\mathbf{x}$ podemos calcular su componente en la dirección de $\mathbf{w}$ como $\mathbf{x}_\mathbf{w} = (\mathbf{x} \cdot \mathbf{w})\mathbf{w}$. Este vector (como matriz columna) es igual al producto del escalar $\mathbf{x}^t\mathbf{w}$ por el vector $\mathbf{w}$. Esto se puede expresar como el producto de matrices $\mathbf{w}(\mathbf{x}^t\mathbf{w})$ que se puede reescribir como $\mathbf{w}(\mathbf{w}^t\mathbf{x})$, y usando la propiedad asociativa,

$$\mathbf{x}_\mathbf{w} = (\mathbf{w}\mathbf{w}^t)\mathbf{x}\,.$$

Si un vector unitario $\mathbf{w}$ lo representamos como un "vector columna", la matriz $\mathbf{w}\mathbf{w}^t$ representa la aplicación lineal 'proyección ortogonal sobre la recta de dirección $\mathbf{w}$, o "*componente longitudinal* en la dirección $\mathbf{w}$". La matriz $I - \mathbf{w}\mathbf{w}^t$ representa entonces la "*componente transversal* a la dirección $\mathbf{w}$".

Si a un vector $\mathbf{x}$ le restamos su componente en la dirección de un vector $\mathbf{w}$, obtenemos la proyección ortogonal de $\mathbf{x}$ sobre el hiperplano ortogonal a $\mathbf{w}$. Si a esta proyección le restamos una vez más la componente de $\mathbf{x}$ en la dirección de $\mathbf{w}$, obtenemos la reflexión, $P_\mathbf{w}\mathbf{x}$, de $\mathbf{x}$ sobre el hiperplano ortogonal a $\mathbf{w}$. En consecuencia tenemos

$$P_\mathbf{w}\mathbf{x} = \mathbf{x} - 2\mathbf{x}_\mathbf{w} = \mathbf{x} - 2(\mathbf{w}\mathbf{w}^t)\mathbf{x} = (I - 2\mathbf{w}\mathbf{w}^t)\mathbf{x}$$

de esta forma llegamos a la expresión general de la matriz de la reflexión sobre el hiperplano ortogonal a un vector unitario $\mathbf{w}$:

$$P_\mathbf{w} = I - 2\mathbf{w}\mathbf{w}^t\,.$$

A partir de esta fórmula es sencillo demostrar que las matrices de reflexión son ortogonales:

Ejercicio 3.3
Toda matriz de la forma $P_\mathbf{w} = I - 2\mathbf{w}\mathbf{w}^t$ con $\mathbf{w}$ unitario (matriz de una reflexión referida a una base ortonormal) es simétrica ($P_\mathbf{w}^t = P_\mathbf{w}$) e involutiva ($P_\mathbf{w}^2 = I$), por lo cual es una matriz ortogonal.

Ahora observamos el hecho fundamental que nos permitirá llegar al algoritmo de Householder, esto es: que toda reflexión sobre un hiperplano queda completamente determinada por la imagen de un vector (que no esté contenido en el hiperplano de reflexión), en otras palabras:

Proposición 5 *Para cualesquiera vectores $\mathbf{x}, \mathbf{y}$ no nulos y de igual norma existe una única reflexión P tal que $P\mathbf{x} = \mathbf{y}$.*

Demostración: Primero veamos la unicidad. Esto es consecuencia de que, para toda reflexión P, la dirección $\mathbf{w}$ de (el hiperplano de) reflexión está determinada por P porque es la misma que la del vector $\mathbf{x} - P\mathbf{x}$ para cualquier $\mathbf{x}$ tal que $\mathbf{x} \neq P\mathbf{x}$. Por tanto si dos reflexiones coinciden en un tal

vector, son reflexiones sobre el mismo hiperplano y por tanto son la misma reflexión. Ahora la existencia: Dados dos vectores distintos $\mathbf{x}$ e $\mathbf{y}$ de igual norma, sea P la reflexión sobre el hiperplano ortogonal a la dirección del vector diferencia $\mathbf{x}-\mathbf{y}$ ($\neq 0$) y sea $\mathbf{w}=(\mathbf{x}-\mathbf{y})/\|\mathbf{x}-\mathbf{y}\|$ la dirección de ese hiperplano. Como $(\mathbf{x}-\mathbf{y})/2=\frac{1}{2}\|\mathbf{x}-\mathbf{y}\|\mathbf{w}$ es el centro del rombo de vértices $\mathbf{x}$, $-\mathbf{y}$, $\mathbf{x}-\mathbf{y}$ y el origen, el producto escalar $\mathbf{w}^t\mathbf{x}$ o componente de $\mathbf{x}$ en la dirección $\mathbf{w}$ es igual a la norma de $(\mathbf{x}-\mathbf{y})/2$, por tanto

$$\begin{aligned} P_{\mathbf{w}}\mathbf{x} &= (I-2\mathbf{w}\mathbf{w}^t)\mathbf{x}=\mathbf{x}-2\mathbf{w}\mathbf{w}^t\mathbf{x}=\mathbf{x}-2\mathbf{w}\|(\mathbf{x}-\mathbf{y})/2\| \\ &= \mathbf{x}-\frac{\mathbf{x}-\mathbf{y}}{\|\mathbf{x}-\mathbf{y}\|}\|\mathbf{x}-\mathbf{y}\|=\mathbf{x}-(\mathbf{x}-\mathbf{y})=\mathbf{y}\,. \end{aligned}$$

como queríamos demostrar. ■

La consecuencia de esto es que si x es un vector no nulo de coordenadas $x_1,\ldots,x_n$, y si $k\in\{1,\ldots,n\}$, eligiendo el número S de forma tal que el vector y de coordenadas $x_1,\ldots,x_k,-S,0,\ldots,0$ tenga la misma norma que x (es decir, eligiendo S tal que $S^2=x_{k+1}^2+,\ldots,+x_n^2$), existirá una (única) reflexión P tal que $Px=y$. Ésta es la reflexión de dirección

$$\mathbf{w}=(x-y)/\|x-y\|=\frac{1}{R}\begin{pmatrix}0\\ \vdots\\ 0\\ x_{k+1}+S\\ x_{k+2}\\ \vdots\\ x_n\end{pmatrix}$$

donde

$$\begin{aligned} R=\|x-y\| &= \sqrt{(x_{k+1}+S)^2+x_{k+2}^2+\cdots+x_n^2} \\ &= \sqrt{(x_{k+1}+S)^2+S^2-x_{k+1}^2} \\ &= \sqrt{2(x_{k+1}+S)S} \end{aligned}$$

Para evitar una pérdida de precisión en estos cálculos conviene elegir el signo de S de tal forma que al sumarle x_{k+1} se sumen sus valores absolutos. Esto significa elegir para S el mismo signo que tenga x_{k+1}, de forma que la fórmula de S es:

$$S=\operatorname{sgn}(x_{k+1})\sqrt{x_{k+1}^2+\cdots+x_n^2}\,.$$

Con lo dicho hasta aquí estamos preparados para demostrar el siguiente

Teorema 19 *Para toda matriz cuadrada A existe una matriz ortogonal P tal que el producto PAP^t es una matriz tipo Hessenberg.*

Demostración: Empezamos construyendo una matriz de reflexión, $P_{\mathbf{w}_1}$ que haga cero todos los elementos bajo la subdiagonal de A en la primera columna. Para ello aplicamos lo dicho más arriba al caso de que el vector x es la primera columna de A. Entonces tendremos

$$\mathbf{w}_1=\frac{1}{R_1}\begin{pmatrix}0\\ a_{21}+S_1\\ a_{31}\\ \vdots\\ a_{n1}\end{pmatrix}$$

El producto $P_{\mathbf{w}_1}A$ tendrá todo ceros bajo la subdiagonal en la primera columna. A continuación multiplicamos por la derecha de esta matriz la inversa de $P_{\mathbf{w}_1}$ (que es ella misma) para obtener la matriz $P_{\mathbf{w}_1}AP_{\mathbf{w}_1}$, semejante a A. Ahora bien, como $\mathbf{w}_1$ tiene un cero como primer elemento, la matriz $\mathbf{w}_1{\mathbf{w}_1}^t$ tiene todo ceros tanto en la primera fila como en la primera columna, lo que implica que la matriz $P_{\mathbf{w}_1}$ es de la forma

$$P_{\mathbf{w}_1}=I-2\mathbf{w}_1{\mathbf{w}_1}^t=\begin{pmatrix}1&0&\cdots&0\\ 0&\times&\cdots&\times\\ \vdots&\vdots&&\vdots\\ 0&\times&\cdots&\times\end{pmatrix}$$

Esto implica que al multiplicar $(P_{\mathbf{w}_1}A)$ por $P_{\mathbf{w}_1}$ no se altera la primera columna de $P_{\mathbf{w}_1}A$ con lo cual la matriz $P_{\mathbf{w}_1}AP_{\mathbf{w}_1}$ sigue teniendo todo ceros bajo la subdiagonal en la primera columna. Similarmente construimos una matriz de reflexión, $P_{\mathbf{w}_2}$ que haga cero todos los elementos bajo la subdiagonal de $P_{\mathbf{w}_1}AP_{\mathbf{w}_1}$ en la segunda columna y hallamos el producto $P_{\mathbf{w}_2}P_{\mathbf{w}_1}AP_{\mathbf{w}_1}P_{\mathbf{w}_2}$. Continuando de esta forma, en el paso k, puesto que las

primeras k componentes de $\mathbf{w}_k$ son cero, usamos una reflexión de la forma

$$P_{\mathbf{w}_k} = I - 2\mathbf{w}_k{\mathbf{w}_k}^t = \begin{pmatrix} 1 & & & & \\ & \ddots & & 0 & \\ & & 1 & & \\ & & & \times & \cdots & \times \\ & 0 & & \vdots & & \vdots \\ & & & \times & \cdots & \times \end{pmatrix}$$

por lo cual al multiplicar

$$P_{\mathbf{w}_k}(P_{\mathbf{w}_{k-1}} \cdots P_{\mathbf{w}_1} A P_{\mathbf{w}_1} \cdots P_{\mathbf{w}_{k-1}}) P_{\mathbf{w}_k}$$

no se alteran las k primeras columnas de

$$P_{\mathbf{w}_k}(P_{\mathbf{w}_{k-1}} \cdots P_{\mathbf{w}_1} A P_{\mathbf{w}_1} \cdots P_{\mathbf{w}_{k-1}}),$$

las cuales tienen todo ceros bajo la subdiagonal.

En consecuencia este proceso produce una matriz ortogonal $P = P_{\mathbf{w}_{n-2}} \cdots P_{\mathbf{w}_1}$ tal que PAP^t es de tipo Hessenberg. ■

Para implementar de forma eficaz el algoritmo de Householder, en cada paso del cual se efectúa un producto de la forma $P_{\mathbf{w}} A P_{\mathbf{w}}$, conviene tener en cuenta que dicho producto se puede realizar de la siguiente forma:

Proposición 6 *Si A es una matriz cuadrada y $\mathbf{w}$ es un vector unitario entonces si $\mathbf{v} = A\mathbf{w}$, $\mathbf{z}^t = \mathbf{w}^t A$, y $\lambda = \mathbf{w}^t\mathbf{v}$*

$$P_{\mathbf{w}} A P_{\mathbf{w}} = A + 2(\mathbf{q}\mathbf{w}^t + \mathbf{w}\mathbf{r}^t)$$

donde $\mathbf{q} = \lambda\mathbf{w} - \mathbf{v}$ y $\mathbf{r}^t = \lambda\mathbf{w}^t - \mathbf{z}^t$.

Demostración: Sólo hay que calcular el producto $P_{\mathbf{w}} A P_{\mathbf{w}} = (I - 2\mathbf{w}\mathbf{w}^t)A(I - 2\mathbf{w}\mathbf{w}^t)$:

$$\begin{aligned} P_{\mathbf{w}} A P_{\mathbf{w}} &= (I - 2\mathbf{w}\mathbf{w}^t)(A - 2A\mathbf{w}\mathbf{w}^t) \\ &= (I - 2\mathbf{w}\mathbf{w}^t)(A - 2\mathbf{v}\mathbf{w}^t) \\ &= A - 2\mathbf{v}\mathbf{w}^t - 2\mathbf{w}\mathbf{w}^t A + 4\mathbf{w}(\mathbf{w}^t\mathbf{v})\mathbf{w}^t \\ &= A - 2\mathbf{v}\mathbf{w}^t - 2\mathbf{w}\mathbf{z}^t + 4\lambda\mathbf{w}\mathbf{w}^t \\ &= A + 2\lambda\mathbf{w}\mathbf{w}^t - 2\mathbf{v}\mathbf{w}^t + 2\lambda\mathbf{w}\mathbf{w}^t - 2\mathbf{w}\mathbf{z}^t \\ &= A + 2(\lambda\mathbf{w} - \mathbf{v})\mathbf{w}^t + 2\mathbf{w}(\lambda\mathbf{w}^t - \mathbf{z}^t) \\ &= A + 2\mathbf{q}\mathbf{w}^t + 2\mathbf{w}\mathbf{r}^t, \end{aligned}$$

que da el resultado requerido. ■

Ejercicio 3.4
Usar el resultado anterior para demostrar que si $\mathbf{w}$ es un vector unitario, el producto $P_{\mathbf{w}} A P_{\mathbf{w}}^t$ puede calcularse mediante el siguiente algoritmo:

1. *Para cada $i \in \{1, \ldots, n\}$ hallar $v_i = \sum_j a_{ij} w_j$ y $z_i = \sum_j a_{ji} w_j$.*
2. *Hallar $\lambda = \sum_i w_i v_i$.*
3. *Para cada $i \in \{1, \ldots, n\}$ y cada $j \in \{1, \ldots, n\}$ hallar*

$$a'_{ij} = a_{ij} + 2(2\lambda w_i w_j - w_j v_i - w_i z_j)\,.$$

Ejercicio 3.5
Usando los resultados anteriores escribir un programa que transforme una matriz dada en otra semejante a ella pero de tipo Hessenberg.

3.2.5 Simplificación aplicable a matrices simétricas

En el caso de que queramos transformar una matriz simétrica a forma Hessenberg, es útil observar dos cosas. En primer lugar cada paso del algoritmo de Householder produce una matriz simétrica ya que si A es una matriz simétrica también lo es $P_{\mathbf{w}} A P_{\mathbf{w}}$ por serlo $P_{\mathbf{w}}$. Esto implica que en este caso el resultado es una matriz *tridiagonal*. En segundo lugar, los cálculos de cada paso del algoritmo pueden simplificarse ya que

Proposición 7 *Si A es una matriz simétrica y $\mathbf{w}$ es un vector unitario entonces si $\mathbf{v} = A\mathbf{w}$ y $\lambda = \mathbf{w}^t\mathbf{v}$,*

$$P_{\mathbf{w}} A P_{\mathbf{w}} = A + 2(\mathbf{q}\mathbf{w}^t + \mathbf{w}\mathbf{q}^t)$$

donde $\mathbf{q} = \lambda\mathbf{w} - \mathbf{v}$.

Demostración: Esto es consecuencia directa del resultado de la proposición 6 teniendo en cuenta que si A es simétrica entonces $\mathbf{z}^t = \mathbf{w}^t A = \mathbf{v}^t$ con lo que $\mathbf{r}^t = \lambda\mathbf{w}^t - \mathbf{z}^t = \lambda\mathbf{w}^t - \mathbf{v}^t = \mathbf{q}^t$. ■

3.2.6 Algoritmo rápido para la factorización ortogonal de matrices Hessenberg

Sea A una matriz de tipo Hessenberg de orden n. Para hallar su factorización ortogonal por medio de rotaciones planas son necesarias n rotaciones,

$Q_1, \dots Q_n$, cada una de las cuales anulará un elemento de la subdiagonal de A. Para eliminar el primer elemento usamos la rotación sobre el plano de los ejes $(1, 2)$ dada por

$$Q_1^t = \begin{pmatrix} c_1 & -s_1 & & & \\ s_1 & c_1 & & 0 & \\ & & 1 & & \\ & 0 & & \ddots & \\ & & & & 1 \end{pmatrix}$$

donde los cosenos directores de la rotación, c_1 y s_1, están determinados, como ya hemos visto en las fórmulas (3.1) por

$$c_1 = \frac{a_{11}}{\sqrt{a_{11}^2 + a_{21}^2}}, \qquad s_1 = \frac{-a_{21}}{\sqrt{a_{11}^2 + a_{21}^2}}.$$

Al multiplicar Q_1^t por A obtenemos una matriz que difiere de A a lo sumo en las dos primeras filas y cuyo elemento en posición $(2, 2)$ es $s_1 a_{12} + c_1 a_{22}$. De esta matriz eliminaremos el segundo elemento de la subdiagonal multiplicándola por

$$Q_2^t = \begin{pmatrix} 1 & & & & & \\ & c_2 & -s_2 & & & \\ & s_2 & c_2 & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}$$

lo cual claramente no altera la primera columna. Continuando de esta manera llegaremos a obtener la matriz triangular superior R como el producto

$$R = Q_{n-1}^t \cdots Q_2^t Q_1^t A$$

lo que implica

$$A = QR$$

donde (teniendo en cuenta que $Q_i^t = Q_i^{-1}$ por ser ortogonal)

$$Q = Q_1 Q_2 \cdots Q_{n-1}.$$

Nos proponemos ahora describir un algoritmo que nos permite calcular la matriz Q en la forma más compacta posible. Para ello estudiamos la estructura de las matrices obtenidas al calcular los sucesivos productos $Q_0 = I, Q_1, Q_1 \cdot Q_2, \dots$ de los cuales el n-ésimo es la propia matriz Q. Las primeras dos matrices (aparte de Q_0) son:

$$Q_1 = \begin{pmatrix} c_1 & s_1 & & & \\ -s_1 & c_1 & & 0 & \\ & & 1 & & \\ & 0 & & \ddots & \\ & & & & 1 \end{pmatrix}, \quad Q_1 Q_2 = \begin{pmatrix} c_1 & s_1 c_2 & s_1 s_2 & & \\ -s_1 & c_1 c_2 & c_1 s_2 & & 0 \\ 0 & -s_2 & c_2 & & \\ & & & 1 & \\ & 0 & & & \ddots \\ & & & & 1 \end{pmatrix}.$$

Ejercicio 3.6
Comprobar que

$$\begin{pmatrix} c_1 & s_1 & 0 \\ -s_1 & c_1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & c_2 & s_2 \\ 0 & -s_2 & c_2 \end{pmatrix} = \begin{pmatrix} c_1 & s_1 c_2 & s_1 s_2 \\ -s_1 & c_1 c_2 & c_1 s_2 \\ 0 & -s_2 & c_2 \end{pmatrix}$$

En esta sucesión la matriz i-ésima (es decir el producto $Q_1 \cdots Q_i$) para $i = 1, \dots n$ se puede obtener de la anterior (el producto $Q_1 \cdots Q_{i-1}$) y de los coeficientes c_i, s_i aplicando los siguientes pasos: (Para $i = 1$ partimos de Q_0 que tomamos siempre igual a la matriz identidad.)

1. Hacer la columna $i + 1$ igual a la columna i multiplicada por s_i.
2. Multiplicar la columna i por c_i.
3. Hacer los elementos de la fila $i+1$ en posición $(i+1, i)$ e $(i+1, i+1)$ (elemento a la derecha de la diagonal y en la diagonal) iguales a $-s_i$ y c_i respectivamente.

Ejercicio 3.7
Comprobar que mediante aplicación de los pasos anteriores se obtiene la transformación (con $i = 3$):

$$\begin{pmatrix} c_1 & s_1 c_2 & s_1 s_2 & \\ -s_1 & c_1 c_2 & c_1 s_2 & \\ 0 & -s_2 & c_2 & \\ & & & 1 \end{pmatrix} \mapsto \begin{pmatrix} c_1 & s_1 c_2 & s_1 s_2 c_3 & s_1 s_2 s_3 \\ -s_1 & c_1 c_2 & c_1 s_2 c_3 & c_1 s_2 s_3 \\ 0 & -s_2 & c_2 c_3 & c_2 s_3 \\ 0 & 0 & -s_3 & c_3 \end{pmatrix}$$

y que el resultado es precisamente $Q_1 Q_2 Q_3$.

Ejercicio 3.8
Hallar el número de multiplicaciones necesarias para obtener la matriz Q (de orden n) mediante iteración de los pasos indicados más arriba a partir de los valores (supuestos conocidos) $(c_1, s_1), \ldots, (c_{n-1}, s_{n-1})$.

Estas observaciones nos llevan al siguiente algoritmo compacto para realizar la factorización ortogonal de una matriz tipo Hessenberg:

Algoritmo de factorización ortogonal de matrices Hessenberg

1 Sea A la matriz dada. Declaramos las variables Q y R como matrices del mismo orden que A.

2 Inicializamos las matrices $Q = I$ y $R = 0$.

3 Inicializamos el contador $k = 0$.

4 Incrementamos el contador $k = k + 1$.

5 Hallamos el valor provisional de la columna k de R:
Para $i = 1, \ldots, k$ **hacemos** $r_{ik} = \sum_{j=1}^{k} q_{ji} a_{jk}$.
Hallamos $r_{k+1\,k} = a_{k+1\,k}$.

6 Calculamos los cosenos directores de este paso:

$$c = r_{kk}/\sqrt{r_{kk}^2 + r_{k+1\,k}^2} \qquad s = -a_{k+1\,k}/\sqrt{r_{kk}^2 + r_{k+1\,k}^2}\,.$$

7 Corregimos la columna k de R:

$$r_{kk} = cr_{kk} - sr_{k+1\,k} \text{ y } r_{k+1\,k} = 0.$$

8 Calculamos la nueva Q haciendo
Para $j = 1, \ldots, k$ **hacemos** $q_{j\,k+1} = sq_{jk}$, $q_{jk} = cq_{jk}$.
Hecho eso $q_{k+1\,k+1} = c$, $q_{k+1\,k} = -s$.

9 Si $k < n - 1$ ir al paso **4**.

10 Si $k = n - 1$ calculamos la última columna de R:
Para $i = 1, \ldots, n$ **hacemos** $r_{in} = \sum_{j=1}^{n} q_{ji} a_{jn}$.

Nótese que este algoritmo admite una pequeña simplificación ya que el valor de $r_{k+1\,k}$ es inicialmente cero y finalmente también cero (paso 7). Ese valor sólo se usa en los pasos 6 y 7 y podemos evitar su introducción usando directamente $a_{k+1\,k}$ en esos pasos. Además debe notarse que el valor de r_{kk} calculado en el paso 7 es igual al denominador usado en el paso 6. Así pues los pasos 5 a 7 pueden sustituirse por los siguientes más eficaces:

5′. Hallamos el valor provisional de la columna k de R:
Para $i = 1, \ldots, k$ hacemos $r_{ik} = \sum_{j=1}^{k} q_{ji} a_{jk}$.

6′. Calculamos los cosenos directores de este paso:

$$t = \sqrt{r_{kk}^2 + a_{k+1\,k}^2} \qquad c = r_{kk}/t \qquad s = -a_{k+1\,k}/t\,.$$

7′. Corregimos la columna k de R: $r_{kk} = t$.

Respuestas a Algunos Ejercicios del Capítulo 3

Ejercicio 3.1 Su sucesión asociada es (por ejemplo) la sucesión constante igual a A porque A es ortogonal. No cumple la hipótesis de los autovalores porque sus autovalores son 1 y -1 (igual valor absoluto). Evidentemente ninguna matriz ortogonal cumple las hipótesis del teorema.

Ejercicio 3.2 Tantas como elementos hay bajo la diagonal, es decir la mitad del número total, n^2, menos los n de la diagonal: $\boxed{n(n-1)/2}$.

Ejercicio 3.3 Sea $P_{\mathbf{w}} = I - 2\mathbf{w}\mathbf{w}^t$. Entonces $P_{\mathbf{w}}^t = I^t - 2(\mathbf{w}\mathbf{w}^t)^t = I - 2(\mathbf{w}^t)^t\mathbf{w}^t = I - 2\mathbf{w}\mathbf{w}^t = P_{\mathbf{w}}$, luego $P_{\mathbf{w}}$ es simétrica. Además $P_{\mathbf{w}}^2 = (I - 2\mathbf{w}\mathbf{w}^t)^2 = I - 2I2\mathbf{w}\mathbf{w}^t + (2\mathbf{w}\mathbf{w}^t)^2 = I - 4\mathbf{w}\mathbf{w}^t + 4\mathbf{w}(\mathbf{w}^t\mathbf{w})\mathbf{w}^t = I - 4\mathbf{w}\mathbf{w}^t + 4\mathbf{w}\mathbf{w}^t = I$ donde se usa que $(\mathbf{w}^t\mathbf{w}) = 1$ por ser $\mathbf{w}$ un vector unitario.

Ejercicio 3.4 La respuesta es un mero e inmediato cálculo.

Ejercicio 3.5 El siguiente algoritmo transforma una matriz cualquiera en otra Hessenberg. Suponemos dados: n, el orden de la matriz, y $A(i, j)$, la matriz dada. La solución se almacenará en la misma variable A.

```
 para k = 1 hasta n-2
********* CALCULAMOS LA DIRECCION DE REFLEXION *********
 S = 0
 para i = k+1 hasta n
 S = S + A(i, k)^2
 siguiente i
 S = SGN(A(k+1, k))*SQRT(S)
 R = SQRT(2*(A(k+1, k)+S)*S)
 para j = 1 hasta k
 w(j) = 0
 siguiente j
 w(k+1) = (A(k+1, k)+S)/R
 para j = k+2 hasta n
 w(j) = A(j, k)/R
 siguiente j
********* CALCULAMOS LOS VECTORES v Y z *********
 para i = 1 hasta n
```



```
 v(i) = 0; z(i) = 0
 para j = 1 hasta n
  v(i) = v(i) + A(i, j)*w(j)
  z(i) = z(i) + A(j, i)*w(j)
 siguiente j
 siguiente i
********* CALCULAMOS EL ESCALAR lambda *********
 lambda = 0
 para j = 1 hasta n
 lambda = lambda + v(j)*w(j)
 siguiente j
********* CALCULAMOS LA NUEVA MATRIZ A *********
 para i = 1 hasta n
 para j = 1 hasta n
  t = 2*lambda*w(i)*w(j) - v(i)*w(j) - w(i)*z(j)
  A(i, j) = A(i, j) + 2*t
 siguiente j
 siguiente i
 siguiente k
```

Ejercicio 3.6

Ejercicio 3.7

Ejercicio 3.8 Las dos comprobaciones son rutinarias. El número de operaciones es la suma de las operaciones realizadas en cada paso. En el paso k, en el que se pasa de una matriz $(k-1)\times(k-1)$ a una $k\times k$ se realizan $2(k-1)$ multiplicaciones (supuesto $k \geq 3$) y por tanto el número total es

$$\sum_{k=3}^{n} 2(k-1) = 2(2+\cdots+(n-1)) = n(n-1)-2 = n^2-n-2$$

que da el valor (correcto) 0 para el menor valor posible de n (a saber: $n = 2$).

Capítulo 4

Teoría de la Interpolación y Aproximación

4.1 El problema general de interpolación

4.1.1 Introducción

Antes de la gran difusión que los ordenadores personales y las calculadoras digitales tuvieron en el último tercio del siglo XX, los cálculos necesarios para resolver los problemas de ingeniería y otras ciencias aplicadas se realizaban principalmente con la ayuda de tablas de funciones. No hace mucho aún se enseñaba en el bachillerato el uso de las *tablas trigonométricas y logarítmicas* en que se listan los valores (cuidadosamente calculados) que toma en ciertos puntos una función particular, como puede ser el seno o el logaritmo. Naturalmente las tablas no pueden contener todos los puntos que se puedan necesitar. Para los puntos no tabulados es necesario realizar una *interpolación* basada en los puntos de la tabla que sean más próximos al dado. Por ejemplo, como nos enseñaban en el bachillerato, si necesitamos el valor de la función f en el punto x y en nuestra tabla de la función f aparecen x_1 y x_2 como puntos más cercanos a x, tendremos

$$f(x) - f(x_2) \approx (x - x_2)\frac{f(x_1) - f(x_2)}{x_1 - x_2}$$

de donde se obtiene el valor interpolado de f en x.

Lo que acabamos de ver es un ejemplo de interpolación lineal. Consiste en hallar el polinomio lineal o de primer grado que "pasa" por los puntos $(x_1, f(x_1))$ y $(x_2, f(x_2))$ y tomar para $f(x)$ el valor de ese polinomio en x. Si este polinomio lineal es una aproximación suficientemente buena de la función f en el intervalo $[x_1, x_2]$ entonces sus valores ahí podrán ser tomados por los de f. Así pues, el problema general de interpolación está muy relacionado con la teoría de aproximación.

La esencia del problema de interpolación ha de ser entendida como la búsqueda de un criterio de cómo debe comportarse una función razonable entre los puntos dados. Después de todo, se puede usar un número infinito de curvas distintas para interpolar unos datos y debemos tener, en cada caso, un criterio para elegir entre ellas. Normalmente ese criterio incluye sencillez y minimizar la curvatura.

4.1.2 Interpolación, aproximación y ajuste de datos

La diferencia entre el problema de interpolación y el problema de aproximación de funciones es principalmente de enfoque. En el problema de aproximación se busca una función (de entre las de una cierta clase) que sea lo más parecida posible (en un sentido a definir en cada problema) a una función dada. En el problema de interpolación lo que nos interesa es la estimación del *valor* en un punto o unos puntos concretos de una cierta función que sólo es conocida de forma más o menos exacta en ciertos otros puntos. Por supuesto, este problema puede resolverse a veces si se encuentra una función que aproxime a la dada suficientemente bien en el intervalo de interés, pero también puede resolverse por otros métodos especiales adaptados al problema particular de interpolación.

Otro problema parecido y relacionado con el problema de interpolación es el de ajuste de una función a unos datos. En ambos casos se trata de poder hallar valores de una función en "puntos intermedios" a los de una tabla dada $\{(x_i, y_i)\}_{i=1,\dots,n}$. Pero en el problema de interpolación los datos son considerados exactos, y se requerirá que la función de interpolación pase exactamente por cada uno de los puntos (x_i, y_i), mientras que en el problema de ajuste lo que se considera correcto es la clase de funciones entre las que buscamos la nuestra, mientras que los datos se consideran como resultados aproximados de la evaluación de la función (ya sea mediante el

cálculo o como resultados de medidas experimentales), por lo que no se pide que la función pase exactamente por ellos (lo que, por otra parte no tendría normalmente solución).

4.2 Interpolación polinómica

Entre las distintas clases de funciones que se usan en la práctica para la interpolación y aproximación ocupa un lugar destacado por su sencillez, facilidad de evaluación y por su gran número de aplicaciones la clase de las *funciones polinómicas*. Debe recordarse que toda función diferenciable de clase k en un punto x_0 puede aproximarse en un entorno de x_0 por su polinomio de Taylor en torno a ese punto y que toda función continua puede aproximarse tanto como se quiera mediante polinomios en el sentido del famoso teorema de Weierstrass:

Teorema 20 (Teorema de Weierstrass) *Dada una función f continua en un intervalo $[a, b]$ sea $E_n(f)$, para cada entero positivo n, el ínfimo de los valores que toma la cantidad $\|f - p\|_\infty = \sup_{a\le x\le b} |f(x) - p(x)|$ cuando p recorre el conjunto $\mathcal{P}_n$ de los polinomios de grado menor o igual que n, es decir,*

$$E_n(f) = \inf_{p\in\mathcal{P}_n} \left(\sup_{a\le x\le b} |f(x) - p(x)| \right).$$

Entonces

$$\lim_{n\to\infty} E_n(f) = 0.$$

Así pues, los polinomios son las primeras funciones que se usarán en las técnicas de interpolación, lo cual nos lleva a la teoría de la *interpolación polinómica*. Para hacer una interpolación polinómica de una función f en unos puntos dados $x_0, \dots, x_n$, llamados los *nodos de interpolación*, es necesario conocer los valores exactos $y_k = f(x_k), k = 0, \dots, n$ que f toma en esos puntos. Entonces interpolar polinómicamente f en un punto x del intervalo que contiene a los $x_0, \dots, x_n$ es evaluar en x el polinomio del menor grado posible que coincida con f en los nodos $x_0, \dots, x_n$. Tal polinomio se llama *polinomio de interpolación* de f en los nodos $x_0, \dots, x_n$ o también polinomio de interpolación en los puntos $(x_0, y_0), \dots, (x_n, y_n)$. Veremos que el valor que el polinomio de interpolación toma en el punto x

se podrá considerar, bajo condiciones adecuadas, como una buena aproximación del valor de f en ese punto.

4.2.1 Existencia y unicidad del polinomio de interpolación

Dado que un polinomio de grado menor o igual que n está determinado por sus $n+1$ coeficientes, quedará completamente determinado por $n+1$ condiciones como son, por ejemplo, que tome valores prescritos en $n+1$ puntos distintos. Sean $a_0, \ldots, a_n$ los $n+1$ coeficientes desconocidos del polinomio $p(x) = a_n x^n + \cdots + a_1 x + a_0$. Para que $p(x)$ pase por los $n+1$ puntos (x_0, y_0), ..., (x_n, y_n) dichos coeficientes deben satisfacer el sistema de $n+1$ ecuaciones lineales en las $n+1$ incógnitas $a_0, \ldots, a_n$,

$$\begin{aligned} a_n x_0^n + \cdots + a_1 x_0 + a_0 &= y_0 \\ &\vdots \\ a_n x_n^n + \cdots + a_1 x_n + a_0 &= y_n \,. \end{aligned} \tag{4.1}$$

La matriz de coeficientes de este sistema es la matriz de Vandermonde

$$M = \begin{pmatrix} x_0^n & x_0^{n-1} & \cdots & x_0 & 1 \\ x_1^n & x_1^{n-1} & \cdots & x_1 & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_n^n & x_n^{n-1} & \cdots & x_n & 1 \end{pmatrix}$$

cuyo determinante es

Ejercicio 4.1

$$\det M = \prod_{0 \le i < j \le n} (x_i - x_j)$$

(Sugerencia: tómese la abscisa x_n como una variable, t. Entonces la función $p(t) = \det\big(M(t)\big)$ es un polinomio de grado n cuyas raices son las otras abscisas.)

Este determinante es ciertamente distinto de cero si y sólo si $x_i \neq x_j$ siempre que $i \neq j$. En consecuencia el sistema (4.1) tiene solución única para cualesquiera $n+1$ puntos con *distintas* abscisas $x_0, \ldots, x_n$. Hemos pues demostrado

Teorema 21 *Dados $n+1$ puntos del plano (x_0, y_0), ..., (x_n, y_n) con distintas abscisas $x_0, \ldots, x_n$, existe un único polinomio de grado menor o igual que n, $p(x)$ que pase por todos ellos, es decir tal que*

$$p(x_k) = y_k. \qquad (k = 0, 1, \ldots, n)$$

En base a este teorema podemos dar la siguiente:

Definición: Dados $n+1$ puntos del plano (x_0, y_0), ..., (x_n, y_n) con distintas abscisas $x_0, \ldots, x_n$, el único polinomio de grado menor o igual que n, $p(x)$ que pase por todos ellos se llama el *polinomio de interpolación* de dichos puntos. Dada una función real de variable real f, y abscisas distintas $x_0, \ldots, x_n$ el único polinomio de grado $\le n$ que toma el mismo valor que f en los nodos $x_0, \ldots, x_n$ se llama el *polinomio de interpolación de f en $x_0, \ldots, x_n$*.

Ejercicio 4.2
¿Qué condición deben cumplir los $n+1$ puntos $\{(x_k, y_k)\}_{0 \le k \le n}$) para que en la solución de (4.1) *resulte $a_n = 0$?*

Ejercicio 4.3
¿Cuál es el grado del polinomio de interpolación para los puntos (x_0, y_0), ..., (x_n, y_n) si éstos verifican $y_k = x_k$ para cada k?

La demostración de unicidad del polinomio de interpolación también puede hacerse con el siguiente razonamiento: Supongamos que $P(x)$ y $Q(x)$ son dos polinomios de grado $\le n$ que coinciden con f en los $n+1$ puntos $x_0, \ldots, x_n$. Entonces $P(x) - Q(x)$ es un polinomio de grado $\le n$ que se anula en los $n+1$ puntos $x_0, \ldots, x_n$, es decir, que tiene $n+1$ ceros. Esto sólo es posible si es el polinomio cero, es decir, si $P(x) = Q(x)$.

De lo dicho se deduce la siguiente importante consecuencia:

Si la función f es un polinomio de grado menor o igual que n, el polinomio de interpolación de f en $n+1$ puntos es la propia función f. En consecuencia, el polinomio de interpolación para $n+1$ nodos $(x_0, y_0), \ldots, (x_n, y_n)$ que tengan las mismas ordenadas, $y_0 = y_1 = \cdots = y_n = y$, es el polinomio constante (de grado cero) igual a y.

4.2.2 Fórmula de Lagrange para la interpolación polinómica. Polinomios de Lagrange

El cálculo del polinomio de interpolación de una función en unos puntos dados puede hacerse de varias maneras. Después de lo dicho más arriba quizás lo primero que se le ocurriría a uno sería resolver el sistema (4.1), sin embargo con ello estaríamos calculando más de lo necesario porque en general sólo necesitamos evaluar el polinomio de interpolación en un punto o dos, para lo cual no es necesario evaluar sus coeficientes explícitamente. A continuación veremos un método de cálculo que lleva el nombre de *fórmula de Lagrange* el cual tiene importantes aplicaciones en varias áreas de las matemáticas.

Suponemos dados $n+1$ nodos de interpolación, $x_0, \ldots, x_n$, todos ellos *distintos entre sí* y en cada uno de los cuales se conoce el valor de la función a interpolar. Para cada $k \in \{0, \ldots, n\}$ formamos el polinomio $q_k(x)$ definido como el polinomio *mónico* (coeficiente principal 1) de grado n que se anula en todos los nodos excepto en x_k, es decir, que $q_k(x)$ viene dado por la fórmula

$$q_k(x) = (x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n) = \prod_{\substack{j=0 \\ j \neq k}}^{n} (x - x_j).$$

Podemos hacer que este $q_k(x)$ tome el mismo valor que f en x_k si lo multiplicamos por $f(x_k)/q_k(x_k)$. Repitiendo lo mismo para cada nodo y sumando los resultados obtenemos un polinomio

$$p(x) = \frac{f(x_0)}{q_0(x_0)} q_0(x) + \cdots + \frac{f(x_n)}{q_n(x_n)} q_n(x)$$

que tiene grado $\leq n$ por ser suma de polinomios de grado n y que claramente coincide con f en todos los nodos. Éste es, pues, el polinomio de interpolación de f en los nodos $x_0, \ldots, x_n$.

Ejercicio 4.4
Usar el hecho de que los polinomios $q_k(x)$ son todos mónicos de grado n para deducir que el coeficiente del término de grado n del polinomio de interpolación de f en los nodos $x_0, \ldots, x_n$ es

$$\frac{f(x_0)}{q_0(x_0)} + \cdots + \frac{f(x_n)}{q_n(x_n)} \tag{4.2}$$

Si definimos los polinomios

$$L_k(x) = \frac{q_k(x)}{q_k(x_k)}$$

el polinomio de interpolación puede expresarse de la siguiente forma llamada *fórmula de Lagrange*:

$$\boxed{p(x) = f(x_0)L_0(x) + \cdots + f(x_n)L_n(x) = \sum_{k=0}^{n} f(x_k)L_k(x)}. \tag{4.3}$$

Los polinomios $L_k(x)$ se llaman *polinomios de Lagrange* relativos a los nodos $x_0, \ldots, x_n$.

Los polinomios de Lagrange tienen (entre otras propiedades algebraicas de interés) la ventaja de ser independientes de la función a interpolar; dependen únicamente de los nodos, no de los valores $y_k = f(x_k)$ de la función. Por tanto una vez calculados para unos nodos determinados pueden ser utilizados para interpolar cualquier función en esos nodos.

Ejemplo de interpolación usando los polinomios de Lagrange.–

Supongamos que queremos estimar el valor de ciertas funciones f, g en el punto $x = 2.12$ para lo cual disponemos de los valores dados en la siguiente tabla:

x	$f(x)$	$g(x)$
2.0	0.69315	0.56931
2.1	0.74194	0.57884
2.2	0.78846	0.67419

entonces procedemos de la siguiente forma: primero evaluamos los polinomios de Lagrange en los nodos, lo cual puede hacerse directamente escribiendo solamente:

$$L_0(2.12) = \frac{(2.12-2.1)(2.12-2.2)}{(2.0-2.1)(2.0-2.2)} = \frac{(0.02)(-0.08)}{(-0.1)(-0.2)} = -0.08$$

$$L_1(2.12) = \frac{(2.12-2.0)(2.12-2.2)}{(2.1-2.0)(2.1-2.2)} = \frac{(0.12)(-0.08)}{(0.1)(-0.1)} = 0.96$$

$$L_2(2.12) = \frac{(2.12-2.0)(2.12-2.1)}{(2.2-2.0)(2.2-2.1)} = \frac{(0.12)(0.02)}{(0.2)(0.1)} = 0.12$$

(Observación: $L_0(2.12) + L_1(2.12) + L_2(2.12) = 1$. ¿Casualidad?)

y ahora es inmediato calcular el valor estimado de $f(x)$:

$$\begin{aligned} f(2.12) &\simeq p_2(2.12) \\ &= 0.69315 \times (-0.08) + 0.74194 \times 0.96 + 0.78846 \times 0.12 \\ &= 0.75142, \end{aligned}$$

y el valor estimado de $g(x)$:

$$\begin{aligned} g(2.12) &\simeq q_2(2.12) \\ &= 0.56931 \times (-0.08) + 0.57884 \times 0.96 + 0.67419 \times 0.12 \\ &= 0.59104. \end{aligned}$$

4.2.3 Fórmula baricéntrica para el polinomio de interpolación

El cálculo de un punto de interpolación mediante la fórmula de Lagrange es excesivamente costoso si se compara con el número de operaciones necesarias para evaluar un polinomio de grado n (a saber, n sumas y n multiplicaciones).

Ejercicio 4.5
La evaluación de un punto de interpolación mediante la fórmula de Lagrange del polinomio de interpolación para $n+1$ puntos requiere

$$\begin{aligned} n(2n+3) &\ \textit{sumas (o } n(n+2) \textit{ si se guardan} \\ &\qquad \textit{los resultados intermedios)} \\ (n+1)(2n-1) &\ \textit{multiplicaciones} \\ n+1 &\ \textit{divisiones} \end{aligned}$$

lo cual representa un número de operaciones del orden de $4n^2$ (o $3n^2$ si se guardan resultados intermedios).

Por supuesto, al aplicar la fórmula de Lagrange no sólo estamos evaluando el polinomio de interpolación sino que también estamos, de alguna manera y hasta cierto punto, hallando ese polinomio. Esto nos indica que si necesitamos hacer varias interpolaciones con los mismos nodos estaremos de alguna forma repitiendo innecesariamente algunos cálculos. Así

pues podemos preguntarnos si existe alguna forma de disponer los cálculos para reducir al máximo el número de operaciones que hemos de realizar o para ahorrar trabajo en el caso de que necesitemos interpolar en varios puntos. La fórmula baricéntrica es una forma de escribir el polinomio de interpolación que da una solución parcial a esta cuestión.

En primer lugar notemos que los $n+1$ polinomios de Lagrange $L_k(x)$ están relacionados por el hecho de que su suma vale 1 en cualquier punto. Esto es consecuencia del hecho observado más arriba de que el polinomio de interpolación de la función constante $f(x) = 1$ es ella misma y por lo tanto

$$1 = \sum_{k=0}^{n} L_k(x).$$

Además de esto si definimos el polinomio de grado $n+1$

$$q(x) = (x - x_0) \cdots (x - x_n),$$

entonces para todo $k \in \{0, \ldots, n\}$ tenemos $q_k(x) = q(x)/(x - x_k)$ lo cual permite escribir el polinomio de interpolación como

$$\begin{aligned} p(x) &= \frac{q(x)\, f(x_0)}{(x - x_0) q_0(x_0)} + \cdots + \frac{q(x)\, f(x_n)}{(x - x_n) q_n(x_n)} \\ &= q(x) \sum_{k=0}^{n} \frac{f(x_k)}{(x - x_k) q_k(x_k)}. \end{aligned} \tag{4.4}$$

Ejercicio 4.6
Demostrar que el polinomio $q(x)$ definido más arriba tiene la propiedad de que para todo $k \in \{0, \ldots, n\}$ su derivada verifica $q'(x_k) = q_k(x_k)$.

Sugerencia: *Derivar $q(x) = q_k(x)(x - x_k)$ respecto a x.*

Ejercicio 4.7
De la relación $1 = \sum_{k=0}^{n} L_k(x)$ se deduce que el polinomio $q(x)$ puede expresarse como

$$q(x) = \frac{1}{\displaystyle\sum_{k=0}^{n} \frac{1}{(x - x_k) q_k(x_k)}}. \tag{4.5}$$

Como consecuencia de estos resultados, se llega a la fórmula baricéntrica del polinomio de interpolación:

Teorema 22 (Fórmula baricéntrica) *Dados $n+1$ puntos (x_0, y_0), ..., (x_n, y_n) y definidas las cantidades*

$$a_k = \frac{1}{q_k(x_k)},$$

el valor en x del polinomio de interpolación para los puntos (x_0, y_0), ..., (x_n, y_n) es el promedio ponderado o baricentro de las ordenadas y_k con pesos $\frac{a_k}{(x-x_k)}$, es decir,

$$p(x) = \frac{\frac{a_0}{(x-x_0)}y_0 + \cdots + \frac{a_n}{(x-x_n)}y_n}{\frac{a_0}{(x-x_0)} + \cdots + \frac{a_n}{(x-x_n)}}$$

Demostración: Esto no es más que la fórmula (4.4) en la que se ha expresado $q(x)$ según (4.5). ■

En consecuencia si se evalúa el polinomio de interpolación mediante esta fórmula y se guardan los valores de las cantidades a_k, así como los de las cantidades $b_k = a_k y_k$, posteriores evaluaciones para distintos valores de x sólo cuestan $3n+1$ sumas y $2(n+1)+1 = 2n+3$ divisiones, es decir, del orden de $5n$ operaciones.

4.2.4 Análisis de errores

Nos proponemos ahora estimar el error que se comete al realizar una interpolación polinómica. Sea $P_n(x)$ el polinomio de interpolación de $f(x)$ en los puntos $x_0, \ldots, x_n$ y sea $R_n(x)$ el *resto* o error cometido al tomar $P_n(x)$ por $f(x)$, $R_n(x) = f(x) - P_n(x)$. Evidentemente R_n es una función que se anula en los $n+1$ puntos $x_0, \ldots, x_n$ por lo que podemos poner

$$R_n(x) = C(x)(x-x_0)\cdots(x-x_n).$$

Fijamos ahora el punto $x \in [x_0, x_n]$, distinto de cualquiera de los nodos $x_0, \ldots, x_n$, y definimos, para este x fijo, la función

$$F(t) = f(t) - P_n(t) - C(x)(t-x_0)\cdots(t-x_n).$$

Esta función se anula para $t = x_0, \ldots, x_n$ y también para $t = x$, luego tiene $n+2$ ceros distintos en el intervalo $[x_0, x_n]$. En consecuencia, según el teorema de Rolle, su derivada tiene $n+1$ ceros distintos en $[x_0, x_n]$. Por la misma razón la derivada segunda $F''(t)$ tiene n ceros distintos en $[x_0, x_n]$ y siguiendo de esta forma llegamos a que la derivada $n+1$, $F^{(n+1)}(t)$, tiene un cero en $[x_0, x_n]$ que denotaremos η. Pero, teniendo en cuenta que la derivada $n+1$ de un polinomio de grado n es cero y que $\frac{d^{n+1}}{dt^{n+1}}\{(t-x_0)\cdots(t-x_n)\} = \frac{d^{n+1}}{dt^{n+1}}t^n = (n+1)!$, tenemos

$$\begin{aligned} F^{(n+1)}(t) &= f^{(n+1)}(t) - P_n^{(n+1)}(t) - C(x)\frac{d^{n+1}}{dt^{n+1}}\big\{(t-x_0)\cdots(t-x_n)\big\} \\ &= f^{(n+1)}(t) - C(x)(n+1)!, \end{aligned}$$

en consecuencia

$$f^{(n+1)}(\eta) - C(x)(n+1)! = 0$$

de donde deducimos

$$C(x) = \frac{f^{(n+1)}(\eta)}{(n+1)!}.$$

Con esto la fórmula del resto queda

$$R_n(x) = \frac{f^{(n+1)}(\eta)}{(n+1)!}(x-x_0)\cdots(x-x_n) \tag{4.6}$$

para algún $\eta \in [x_0, x_n]$. Nótese la semejanza de esta fórmula con la fórmula del resto del polinomio de Taylor (véase, por ejemplo (4.16)).

El valor de $f^{(n+1)}(\eta)$ que aparece en la fórmula (4.6) es desconocido (entre otras razones porque el punto η es desconocido). Sin embargo, con frecuencia es posible dar una acotación de $f^{(n+1)}$ en el intervalo $[x_0, x_n]$ donde sabemos se encuentra η. En tal caso podemos utilizar la fórmula del resto (4.6) para estimar el error de interpolación como ilustramos con el siguiente ejemplo:

Supongamos que estimamos el valor de $\log_{10} 7$ mediante interpolación

en la siguiente tabla

x	$\log_{10}(x)$
5	0.69897
6	0.77815
8	0.90309
9	0.95424
10	1

El error es

$$R_4(7) = C(7)(7-5)(7-6)(7-8)(7-9)(7-10) = -12\,C(7)$$

donde $C(7) = \frac{1}{5!}\frac{d^5 \log_{10}(x)}{dx^5}\big|_{x=\eta} = \frac{1}{5!}\frac{1}{\ln 10}\frac{4!}{\eta^5} = \frac{1}{5\ln 10\eta^5}$. Ahora bien, como $\eta > 5$, $\frac{1}{\eta^5} < \frac{1}{5^5}$ y por tanto

$$|R_4(7)| = |-12\,C(7)| = \frac{12}{5\ln 10\,\eta^5} < \frac{12}{5^6 \ln 10} \simeq 0.000334.$$

Este resultado nos indica que podemos esperar tres decimales exactos en el valor de $\log_{10} 7$ hallado por interpolación.

4.3 Método de Newton para simplificar el añadir nuevos puntos

Hemos visto que la fórmula baricéntrica representa una economía de operaciones respecto a la fórmula de Lagrange, especialmente si necesitamos interpolar la misma función en varios puntos con los mismos nodos. Sin embargo en ocasiones uno desearía hacer una interpolación de una misma función en un mismo punto pero con unos nodos diferentes. Por ejemplo, podemos encontrarnos en la situación de poder elegir entre unos nodos u otros, o, en caso de trabajar con una tabla en la que los nodos son fijos, podemos haber calculado una interpolación basándonos en ciertos puntos de la tabla y desear repetir la interpolación usando algún valor adicional con la intención de aumentar la exactitud. En este segundo caso ¿sería necesario repetir los cálculos desde el principio otra vez? ¿Podremos aprovechar parte de los cálculos ya realizados? Newton observó que se puede expresar el polinomio de interpolación, P_n, de $n+1$ puntos, como una suma de términos tal que el último término sea igual a $P_n - P_{n-1}$. De esta forma, al añadir más puntos de interpolación no es necesario recalcular todos los términos, sino simplemente añadir unos nuevos.

4.3.1 La fórmula de Newton

Sea $P_n(x)$ el polinomio de interpolación de los $n+1$ puntos

$$(x_0, y_0), \ldots, (x_n, y_n).$$

El cálculo de este polinomio puede hacerse con un mínimo de operaciones adicionales si se conoce ya el polinomio $P_{n-1}(x)$ de interpolación para los n puntos $(x_0, y_0), \ldots, (x_{n-1}, y_{n-1})$. Para ver lo que hemos de añadir a $P_{n-1}(x)$ para obtener $P_n(x)$ estudiemos la diferencia $P_n(x) - P_{n-1}(x)$. Claramente esta diferencia es un polinomio de grado menor o igual que n que se anula en los n puntos $x_0, \ldots, x_{n-1}$ por lo tanto es de la forma

$$P_n(x) - P_{n-1}(x) = A_n(x - x_0)\cdots(x - x_{n-1})$$

para alguna constante A_n que hemos de determinar a partir de los datos $(x_0, y_0), \ldots, (x_n, y_n)$.

Ejercicio 4.8
Demostrar que la constante A_n es igual al coeficiente del término de grado n del polinomio $P_n(x)$. Usar el resultado del ejercicio 4.4 para deducir la fórmula

$$A_n = \frac{y_0}{q_0(x_0)} + \cdots + \frac{y_n}{q_n(x_n)} \tag{4.7}$$

Así pues, dados los puntos $(x_0, y_0), \ldots, (x_n, y_n)$, para cada $k \in \{0, \ldots, n\}$ podemos calcular la constante A_k la cual es, por definición, el coeficiente principal del polinomio de interpolación de grado k en los puntos (x_0, y_0), $\ldots$, (x_k, y_k). Estas constantes tienen una importancia especial y reciben un nombre que indica una forma alternativa de calcularlas:

Definición: Las constantes $A_1, A_2, \ldots, A_n$ se denominan *diferencias divididas* de los puntos (x_0, y_0), $\ldots$, (x_n, y_n). Cuando los y_k se obtienen como los valores de una función f en los x_k, estas diferencias divididas se representan con la notación

$$\boxed{A_k = f[x_0, \ldots, x_k]}.$$

Evidentemente esta definición implica la siguientes expresiones para los sucesivos polinomios de interpolación:

$$
\begin{aligned}
P_0(x) &= f(x_0)\\
P_1(x) &= f(x_0) + f[x_0, x_1](x - x_0)\\
P_2(x) &= f(x_0) + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)\\
&\vdots\\
P_n(x) &= f(x_0) + f[x_0, x_1](x - x_0) + \cdots + f[x_0, \ldots, x_n](x - x_0)\cdots(x - x_{n-1})
\end{aligned}
$$

Esta forma de expresar los polinomios de interpolación se conoce como *fórmula de Newton*. Podemos expresarla en forma más compacta así:

$$\boxed{P_n(x) = \sum_{i=0}^{n} f[x_0, \ldots, x_i] \prod_{j=0}^{i-1} (x - x_j)} \tag{4.8}$$

Ejercicio 4.9
Utilizar el resultado del ejercicio 4.8 para establecer la fórmula

$$f[x_0, \ldots, x_n] = \frac{1}{n!} \frac{d^n P_n(x)}{dx^n} \tag{4.9}$$

Vamos ahora a estudiar la forma más eficaz de calcular las diferencias divididas $f[x_0, \ldots, x_i]$ que aparecen en la fórmula de Newton.

4.3.2 Las diferencias divididas

Dando a x los valores $x_0, \ldots, x_n$ en la fórmula de Newton obtenemos

$$
\begin{aligned}
y_0 &= f(x_0)\\
y_1 &= y_0 + f[x_0, x_1](x_1 - x_0)\\
y_2 &= y_0 + f[x_0, x_1](x_2 - x_0) + f[x_0, x_1, x_2](x_2 - x_0)(x_2 - x_1)\\
&\vdots\\
y_n &= y_0 + f[x_0, x_1](x_n - x_0) + \cdots + f[x_0, \ldots, x_n](x_n - x_0)\cdots(x_n - x_{n-1})
\end{aligned}
$$

De esto, suponiendo que los nodos son distintos dos a dos, deducimos:

$$
\begin{aligned}
f[x_0, x_1] &= \frac{y_1 - y_0}{x_1 - x_0}\\
f[x_0, x_1, x_2] &= \frac{\frac{y_2 - y_0}{x_2 - x_0} - f[x_0, x_1]}{x_2 - x_1} = \frac{f[x_0, x_2] - f[x_0, x_1]}{x_2 - x_1},
\end{aligned}
$$

y continuando de esta manera se puede llegar a establecer en general:

$$f[x_0, \ldots, x_{k+1}] = \frac{f[x_0, \ldots, x_{k-1}, x_{k+1}] - f[x_0, \ldots, x_{k-1}, x_k]}{x_{k+1} - x_k} \tag{4.10}$$

que es la fórmula que expresa $f[x_0, \ldots, x_{k+1}]$ como diferencia dividida. Sin embargo esta fórmula puede demostrarse mucho más sencillamente de la siguiente forma:

Ejercicio 4.10
Sean $x_0, \ldots, x_{k+1}$ nodos de interpolación distintos entre si. Si $p(x)$ es el polinomio de interpolación de f en $x_0, \ldots, x_{k-1}, x_{k+1}$, y $q(x)$ es el polinomio de interpolación de f en $x_0, \ldots, x_k$, entonces el polinomio

$$\frac{x - x_k}{x_{k+1} - x_k} p(x) + \frac{x_{k+1} - x}{x_{k+1} - x_k} q(x)$$

tiene grado $\leq k + 1$ y coincide con f en $x_0, \ldots, x_{k+1}$ (y por lo tanto es el polinomio de interpolación de f en los puntos $x_0, \ldots, x_{k+1}$).

Sugerencia: *Ver la demostración del Lema de Aitken.*

Utilizando el resultado de este ejercicio se llega fácilmente al siguiente resultado:

Ejercicio 4.11
Sean $f[x_0, \ldots, x_{k-1}, x_{k+1}]$ y $f[x_0, \ldots, x_k]$ los coeficientes del término de grado k de los polinomios p y q del ejercicio 4.10. Entonces, suponiendo que $x_{k+1} \neq x_k$, el coeficiente del término de grado $k + 1$ del polinomio de interpolación de f en $x_0, \ldots, x_{k+1}$ es igual a

$$\frac{f[x_0, \ldots, x_{k-1}, x_{k+1}] - f[x_0, \ldots, x_{k-1}, x_k]}{x_{k+1} - x_k}.$$

En realidad, la fórmula (4.10) es sólo una de las posibles fórmulas de las diferencias divididas. Todas esas fórmulas pueden deducirse a la vez al demostrar, con un razonamiento análogo al de los ejercicios 4.10 y 4.11, el siguiente teorema,

Teorema 23 *Si los nodos de interpolación son todos distintos entre si, entonces la n-ésima diferencia dividida* $f[x_0, \ldots, x_n]$ *puede calcularse, para cualesquiera* $i, j \in \{0, \ldots, n\}$ *con* $i \neq j$*, mediante*

$$f[x_0, \ldots, x_n] = \frac{f[x_0, \ldots, \widehat{x_i}, \ldots, x_n] - f[x_0, \ldots, \widehat{x_j}, \ldots, x_n]}{x_j - x_i} \quad (4.11)$$

donde el circunflejo indica una variable que no se considera incluida en la lista, por ejemplo,

$$f[x_0, \ldots, \widehat{x_i}, \ldots, x_n] = f[x_0, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n].$$

El caso particular de la fórmula (4.11) más utilizado es el que resulta al tomar $i = 0$ y $j = n$, es decir

$$\boxed{f[x_0, \ldots, x_n] = \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0}.} \quad (4.12)$$

Este teorema se deduce (con un razonamiento análogo al que nos lleva del ejercicio 4.10 al ejercicio 4.11) de un teorema conocido con el nombre de *Lema de Aitken*.

Teorema 24 (Lema de Aitken) *Sea* f *una función continua en un intervalo* I *y sean* $\{x_0, \ldots, x_n\} \in I$*. Para cada subconjunto* $S \subset \{x_0, \ldots, x_n\}$ *denotemos* $P_S(x)$ *el polinomio de interpolación de* f *en los nodos que son elementos de* S*. Si* $x_i, x_j \in S$ *y* $x_i \neq x_j$ *entonces*

$$P_S(x) = \frac{(x - x_j)P_{S-\{x_j\}}(x) - (x - x_i)P_{S-\{x_i\}}(x)}{x_i - x_j}.$$

Demostración: Sea $m + 1$ el número de elementos de S. Lo que hay que probar es que P_S tiene grado $\leq m$ y que coincide con f en los elementos de S. Lo último es evidente para x_i y para x_j porque lo cumplen $P_{S-\{x_i\}}$ y $P_{S-\{x_j\}}$. Para un $x_k \in S$ distinto de x_i y de x_j, dado que $P_{S-\{x_i\}}(x_k) = f(x_k)$ y $P_{S-\{x_j\}}(x_k) = f(x_k)$, tenemos

$$\begin{aligned} P_S(x_k) &= \frac{(x_k - x_j)f(x_k) + (x_i - x_k)f(x_k)}{x_i - x_j} \\ &= \frac{-x_j f(x_k) + x_i f(x_k)}{x_i - x_j} = f(x_k). \end{aligned}$$

Sólo falta demostrar que P_S tiene grado $\leq m$. Por ser polinomios de interpolación en m nodos tanto $P_{S-\{x_i\}}$ como $P_{S-\{x_j\}}$ tienen grado $\leq m - 1$ y al multiplicarlos por polinomios de primer grado se obtienen polinomios de grado $\leq m$, por lo que P_S tiene grado $\leq m$. En consecuencia P_S es el polinomio de interpolación en los nodos que son elementos de S. ■

4.3.3 Propiedades de simetría

Una propiedad importante de las diferencias divididas de un conjunto de puntos es el ser independientes del orden en que los puntos estén dados.

Teorema 25 (Propiedad de simetría de las diferencias divididas) *Dados* $n + 1$ *números* $x_0, \ldots, x_n$ *en el dominio de una función* f*, para cualquier permutación* $\sigma : \{0, \ldots, n\} \to \{0, \ldots, n\}$ *de los índices* $\{0, \ldots, n\}$ *se verifica*

$$f[x_0, \ldots, x_n] = f[x_{\sigma(0)}, \ldots, x_{\sigma(n)}]$$

Demostración: Una demostración inmediata se basa en la fórmula (4.7) en la que una reordenación de los puntos $x_0, \ldots, x_n$ sólo cambia el orden de los sumandos. Otra demostración puede hacerse por inducción en el número de puntos. La simetría es trivial para el caso de un punto ($n = 0$) y es evidente en el caso de dos puntos:

$$f[x_0, x_1] = \frac{f(x_1) - f(x_0)}{x_1 - x_0} = \frac{f(x_0) - f(x_1)}{x_0 - x_1} = f[x_1, x_0].$$

Supongamos que se cumple para el caso de n puntos y consideremos los $n+1$ puntos $x_0, \ldots, x_n$ y la permutación $\sigma : \{0, \ldots, n\} \to \{0, \ldots, n\}$. Sean $i = \sigma(0)$, $j = \sigma(n)$, entonces (por la hipótesis de inducción) $f[x_0, \ldots, \widehat{x_i}, \ldots, x_n] = f[x_{\sigma(1)}, \ldots, x_{\sigma(n)}]$ e igualmente $f[x_0, \ldots, \widehat{x_j}, \ldots, x_n] = f[x_{\sigma(0)}, \ldots, x_{\sigma(n-1)}]$

donde el circunflejo sobre una variable indica que esa variable no se considera incluida en la lista. Entonces, según las fórmulas (4.11) y (4.12),

$$\begin{aligned} f[x_0,\ldots,x_n] &= \frac{f[x_0,\ldots,\widehat{x_i},\ldots,x_n]-f[x_0,\ldots,\widehat{x_j},\ldots,x_n]}{x_j-x_i} \\ &= \frac{f[x_{\sigma(1)},\ldots,x_{\sigma(n)}]-f[x_{\sigma(1)},\ldots,x_{\sigma(n-1)}]}{x_{\sigma(n)}-x_{\sigma(0)}} \\ &= f[x_{\sigma(0)},\ldots,x_{\sigma(n)}] \end{aligned}$$

con lo que queda demostrado el teorema. ■

4.3.4 La tabla de diferencias divididas y adición de nuevos nodos

Si calculamos las diferencias divididas de una función f en los puntos $x_0,\ldots,x_n$ mediante (4.12) podemos disponerlas en una tabla de la siguiente forma:

x_0	$f[x_0]$				
		$f[x_0,x_1]$			
x_1	$f[x_1]$		$f[x_0,x_1,x_2]$		
		$f[x_1,x_2]$		$f[x_0,x_1,x_2,x_3]$	
x_2	$f[x_2]$		$f[x_1,x_2,x_3]$		$f[x_0,x_1,x_2,x_3,x_4]$
		$f[x_2,x_3]$		$f[x_1,x_2,x_3,x_4]$	
x_3	$f[x_3]$		$f[x_2,x_3,x_4]$		
		$f[x_3,x_4]$			
x_4	$f[x_4]$				

Esta tabla nos indica la forma de calcular cada diferencia dividida: la diferencia de las dos que están a su izquierda dividida entre la correspondiente diferencia de las x. Por ejemplo, la tabla de diferencias divididas para los datos

x	$f(x)$
1	1.5709
4	1.5727
6	1.5751

es

1	1.5709		
		0.0006	
4	1.5727		0.00012
		0.0012	
6	1.5751		

Supongamos ahora que nos dan el punto (0, 1.5708) como información adicional. Entonces sólo tenemos que hacer un poco más de trabajo para completar la tabla

1	1.5709			
		0.0006		
4	1.5727		0.00012	
		0.0012		−0.000001
6	1.5751		0.000121	
		0.000717		
0	1.5708			

4.4 El algoritmo de interpolación de Aitken

Gran parte del error cometido al hacer una interpolación proviene del redondeo en las operaciones aritméticas realizadas. Veremos a continuación un algoritmo para la evaluación del polinomio de interpolación, que está especialmente adaptado para reducir los errores de cálculo.

Antes de introducir dicho algoritmo, llamado el *algoritmo de Aitken*, vamos a introducir la siguiente notación para los polinomios de interpolación. Denotaremos por $P_{0,\ldots,n}$ el polinomio de interpolación de una función f en los nodos $x_0,\ldots,x_n$. Es decir, usamos como subíndices en el polinomio de interpolación el conjunto de índices correspondientes a los nodos de interpolación. De acuerdo con esto tenemos:

$$P_{0,1} = f(x_0)\frac{x-x_1}{x_0-x_1} + f(x_1)\frac{x-x_0}{x_1-x_0}, \qquad \text{(Lagrange)}$$

$$P_{0,1} = f(x_0) + \frac{f(x_1)-f(x_0)}{x_1-x_0}(x-x_0), \qquad \text{(Newton)}$$

$$\begin{aligned} P_{0,1} &= \frac{(x-x_0)f(x_1)-(x-x_1)f(x_0)}{x_1-x_0} \qquad \text{(Aitken)} \\ &= \frac{(x-x_0)P_1(x)-(x-x_1)P_0(x)}{x_1-x_0}. \end{aligned}$$

Para tres puntos la fórmula de Aitken es

$$P_{0,1,2}(x) = \frac{(x-x_1)P_{0,2}(x)-(x-x_2)P_{0,1}(x)}{x_2-x_1}$$

y en general, de acuerdo con la fórmula del lema de Aitken, tenemos

$$P_{0,\ldots,k+1}(x) = \frac{(x - x_k)P_{0,\ldots,k-1,k+1}(x) - (x - x_{k+1})P_{0,\ldots,k}(x)}{x_{k+1} - x_k}.$$

Esta fórmula es la base del algoritmo de Aitken:

Algoritmo de interpolación de Aitken

1 **Datos:** n (número de intervalos o segmentos), $x_0, \ldots, x_n$ (nodos de interpolación), $y_0, \ldots, y_n$ (ordenadas en los nodos), x (punto en el que se desea interpolar la función).
2 **Para** $k = 0, n$
  $A(k, 0) = y_k$;
  $D(k) = x - x_k$;
3 **Siguiente** k
4 **Para** $i = 1, n$
  Para $j = 1, i$
    $A(i, j) = [D(j-1)A(i, j-1) - D(i)A(j-1, j-1)]/(x_i - x_{j-1})$;
  Siguiente j
5 **Imprimir** $A(i, i)$
6 **Siguiente** i
7 **Imprimir** "Solución es: "; $A(n, n)$; "." **y PARAR**

En este algoritmo se imprimen todos los valores intermedios, $A(i, i)$, con el objetivo de tener una idea de la convergencia. Dichos valores de la diagonal de A forman una sucesión de aproximaciones al valor deseado. Si esta sucesión se acerca a un valor fijo podemos tomar ese valor como el valor de la interpolación deseado. Normalmente dichos valores parecen converger al cabo de unos pocos pasos y llegados a un punto empiezan a diverger con un empeoramiento de la estimación. Esto nos indica que pasado cierto punto no se gana precisión al utilizar más nodos de interpolación.

Cuando se hacen cálculos con una calculadora de bolsillo, se pueden sistematizar los cálculos según un esquema sencillo fácil de memorizar. Por ejemplo, el cálculo de $P_{0,1,2} = [(x - x_1)P_{0,2} - (x - x_2)P_{0,1}]/(x_2 - x_1)$ es como el de un determinante 2×2 dividido por $x_2 - x_1$. En esta situación conviene disponer los cálculos en una tabla de la siguiente forma:

x_i	$x - x_i$	$f(x_i)$			
x_0	$x - x_0$	P_0			
x_1	$x - x_1$	P_1	$P_{0,1}$		
x_2	$x - x_2$	P_2	$P_{0,2}$	$P_{0,1,2}$	
x_3	$x - x_3$	P_3	$P_{0,3}$	$P_{0,1,3}$	$P_{0,1,2,3}$

Ejercicio 4.12
Comprobar la siguiente tabla de interpolaciones de la función f en el punto $x = 4.5$,

x_i	$x - x_i$	$f(x_i)$				
4.0	0.5	0.60206				
4.2	0.3	0.62325	0.65504			
4.4	0.1	0.64345	0.65380	0.65318		
4.6	−0.1	0.66276	0.65264	0.65324	0.65321	
4.8	−0.3	0.68124	0.65155	0.65330	0.65321	0.65321

4.5 Interpolación con nodos igualmente espaciados

Vamos a suponer ahora que los nodos de interpolación están igualmente espaciados, es decir (suponiendo los nodos dados en orden creciente, $x_0 < \cdots < x_n$), las diferencias $x_{i+1} - x_i$ son todas iguales a un valor fijo que denotamos h. Entonces, si $a = x_0$ y $b = x_n$, tenemos $h = (b - a)/n$.

Dado un punto x definimos la variable $s = (x - x_0)/h$, de forma que $x = x_0 + sh$. Con este cambio de variable los nodos corresponden a los valores enteros no negativos de s, $s = 0, \ldots, n$ y podemos aplicar este cambio de variable a la función f para obtener la función de variable entera F tal que para valores correspondientes de x y s, $F(s) = f(x) = f(x_0 + sh) = f_s$.

Ahora definimos los incrementos de órdenes sucesivos,

$$\Delta^1 f_s = f_{s+1} - f_s = F(s+1) - F(s)$$
$$\Delta^2 f_s = \Delta^1 f_{s+1} - \Delta^1 f_s$$

y, en general

$$\Delta^{i+1} f_s = \Delta^i f_{s+1} - \Delta^i f_s.$$

Estos incrementos se llaman *diferencias progresivas* y se pueden calcular para cualquier conjunto de puntos (x_i, y_i) en los que las abscisas estén igualmente espaciadas. En base a estas diferencias progresivas se pueden calcular fácilmente las diferencias divididas correspondientes a los nodos dados mediante la fórmula que se establece en el siguiente teorema:

Teorema 26 *Si las abscisas de los puntos $(x_i,\ f(x_i))$ están igualmente espaciadas siendo el espaciamiento $h = x_{i+1} - x_i$, entonces*

$$f[x_k, \ldots, x_{k+n}] = \frac{\Delta^n f_k}{n! h^n}$$

Demostración: Por inducción. Si $n = 0$, $f[x_k] = f(x_k) = \Delta^0 f_k / h^0 = f_k$. También se puede comprobar inmediatamente que la fórmula es válida para $n = 1$. Supongamos el resultado cierto para cierto n. Entonces

$$\begin{aligned} f[x_k, \ldots, x_{k+n+1}] &= \frac{f[x_{k+1}, \ldots, x_{k+n+1}] - f[x_k, \ldots, x_{k+n}]}{x_{k+n+1} - x_k} \\ &= \frac{\frac{1}{n!h^n}\Delta^n f_{k+1} - \frac{1}{n!h^n}\Delta^n f_k}{(n+1)h} \\ &= \frac{\Delta^n f_{k+1} - \Delta^n f_k}{(n+1)!h^{n+1}} = \frac{\Delta^{n+1} f_k}{(n+1)!h^{n+1}} \end{aligned}$$

con lo que se completa la demostración. ■

Corolario 7 *Si P_n es el polinomio de interpolación de f en nodos igualmente espaciados $x_0, \ldots, x_n$ con $x_k = x_0 + kh$ entonces*

$$\frac{d^n P_n(x)}{dx^n} = \frac{\Delta^n f_0}{h^n}$$

Demostración:

Según la fórmula (4.9) que aparece en el ejercicio 4.9, tenemos $f[x_0, \ldots, x_n] = \frac{1}{n!} d^n P_n(x)/dx^n$. Ahora bien, según el teorema anterior

$$f[x_0, \ldots, x_n] = \frac{1}{n!}\Delta^n f_0 / h^n,$$

por tanto, igualando términos y cancelando el factorial se obtiene la fórmula del corolario. ■

Vamos a ver ahora la forma que adopta la fórmula de Newton del polinomio de interpolación para nodos igualmente espaciados con espaciamiento igual a $h = (x_n - x_0)/n$. Utilizando la expresión de las diferencias divididas dada en el teorema anterior, la fórmula de Newton queda

$$P_n(x) = f(x_0) + \frac{\Delta^1 f_0}{h}(x - x_0) + \cdots + \frac{\Delta^n f_0}{n! h^n}(x - x_0)\cdots(x - x_{n-1})$$

y teniendo en cuenta que para cada k, $x - x_k = (s-k)h$ (usando la variable s definida por $s = (x - x_0)/h$), podemos poner

$$P_n(x) = f_0 + s\Delta^1 f_0 + \frac{s(s-1)}{2!}\Delta^2 f_0 + \cdots + \frac{s(s-1)\cdots(s-n+1)}{n!}\Delta^n f_0.$$

Ahora bien, podemos definir los números combinatorios $\binom{s}{k}$ para s no necesariamente entero mediante

$$\binom{s}{k} = \frac{s(s-1)\cdots(s-k+1)}{k!},$$

con lo cual el polinomio de interpolación queda

$$P_n(x) = f_0 + \binom{s}{1}\Delta^1 f_0 + \binom{s}{2}\Delta^2 f_0 + \cdots + \binom{s}{n}\Delta^n f_0,$$

o en forma más compacta

$$P_n(x) = P_n(x_0 + sh) = \sum_{k=0}^{n} \binom{s}{k}\Delta^k f_0.$$

4.5.1 Problemas de la interpolación con nodos igualmente espaciados

Si nos fijamos en la función $q(x) = (x - x_0)\cdots(x - x_n)$ que determina el comportamiento del error de interpolación, nos encontramos con una función que, para nodos igualmente espaciados, tiene grandes oscilaciones hacia los extremos del intervalo de interpolación. Como ejemplo obsérvese el comportamiento de $q(x)$ para las abscisas obtenidas al dividir el intervalo $[-1, 1]$ en 9 subintervalos iguales:

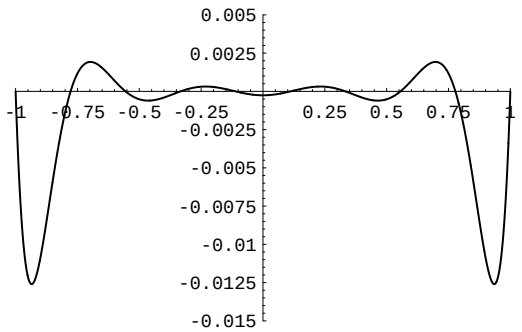


Figura 4.1: Gráfica del poinomio $q(x) = (x+1)(x+\frac{7}{9})(x+\frac{5}{9})(x+\frac{3}{9})(x+\frac{1}{9})(x-\frac{1}{9})(x-\frac{3}{9})(x-\frac{5}{9})(x-\frac{7}{9})(x-1)$

Por lo dicho, la interpolación mediante un solo polinomio que pase por todos los nodos suele ser de poca utilidad, especialmente cuando el número de nodos es elevado y están igualmente distribuidos. Vemos en las siguientes gráficas varios ejemplos del fenómeno conocido como fenómeno Runge, que consiste en las excesivas oscilaciones que aparecen en el polinomio de interpolación cerca de los extremos del intervalo de interpolación.

4.6 Interpolación mediante varillas flexibles (*splines*)

4.6.1 Introducción

Para evitar los problemas de la interpolación polinómica con un número elevado de nodos (sobre todo si están igualmente espaciados), surge la idea de interpolar por distintas funciones en intervalos adyacentes imponiendo condiciones de contorno entre cada dos de ellos para asegurar la continuidad, diferenciabilidad, etc. Es éste un desarrollo reciente, pero que está relacionado con un artilugio mecánico muy antiguo.

Los artesanos han usado desde hace mucho varillas flexibles (inglés: *splines*) hechas de algún material elástico como madera o (más recientemente) plástico, para trazar curvas que tomen una forma predeterminada al hacerlas pasar por unos puntos

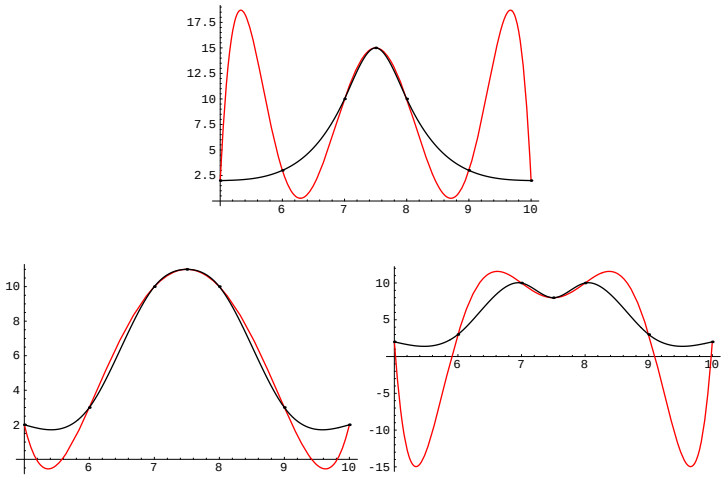


Figura 4.2: Ejemplos de oscilaciones del polinomio de interpolación. En gris el correspondiente polinomio de interpolación para las abscisas $x_0 = 5$, $x_1 = 6$, $x_2 = 7$, $x_3 = 7.5$, $x_4 = 8$, $x_5 = 9$, $x_6 = 10$.

dados (puntos de interpolación). Estas varillas se sujetan fijamente a los puntos de interpolación y automáticamente adoptan la forma que minimiza su energía potencial elástica, obteniéndose una curva suave y estéticamente agradable que pasa por todos los puntos prefijados.

Como la densidad lineal de energía elástica de una varilla es en cada uno de sus puntos proporcional al cuadrado de la curvatura $\kappa(x)$ en ese punto, la forma de la varilla será la solución del problema variacional $\delta \int \alpha\kappa(x)^2\, dl = 0$ (α = constante elástica), la cual, suponiendo que las fuerzas que los nodos ejercen sobre la varilla sean perpendiculares a ésta (ausencia de rozamiento), resulta ser una función que, entre cada dos nodos consecutivos, está dada por un polinomio de tercer grado cuyos coeficientes pueden calcularse fácilmente resolviendo un sistema de ecuaciones lineales.

Estas funciones polinómicas a trozos son también la solución de un problema similar: El de la trayectoria de un móvil que ha de pasar por unos puntos dados en el menor tiempo posible y de la forma más suave (esto es, minimizando en cada punto la curvatura).

La teoría de interpolación mediante "varillas flexibles" se ha desarrollado enormemente alcanzando un gran nivel de generalidad. Hoy día las varillas cúbicas son sólo un caso particular de unos métodos mucho más generales, que se salen del alcance de este curso. A pesar de ello el caso de las varillas cúbicas es de gran interés y utilidad práctica. Por ejemplo,

en las artes gráficas se siguen utilizando las varillas flexibles cúbicas, en su versión de software de diseño gráfico, bajo el nombre de *Curvas de Bezier*, que son el caso más sencillo ya que son curvas que pasan por dos puntos determinados y con tangentes en esos puntos de pendiente determinada. Así, las curvas de Bezier son polinomios de tercer grado que quedan completamente determinados por 3 puntos del plano; los *dos extremos* y un *punto de control*, que es el punto de intersección de las dos rectas tangentes en los extremos, como se indica en la figura 4.3.

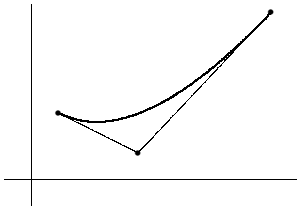


Figura 4.3: Curva de Bezier

4.6.2 Varillas flexibles cúbicas

Condiciones para su determinación.–

Sean $x_0 < \cdots < x_n$ nodos de interpolación en los que conocemos los valores $y_0 = f(x_0), \ldots, y_n = f(x_n)$ de cierta función f. Nuestro problema es hallar n polinomios de tercer grado, $p_1(x), \ldots, p_n(x)$, tales que para cada $k = 1, \ldots, n$, p_k pase por los puntos (x_k, y_k) y (x_{k-1}, y_{k-1}) y además de tal forma que en cada nodo los dos polinomios que coinciden en él tengan la misma pendiente y la misma curvatura. De esta forma obtenemos una "varilla flexible" como la que se muestra en la figura 4.4, es decir una curva polinómica de grado tres a trozos y con derivada segunda continua en todos los puntos del intervalo de x_0 a x_n.

Así pues, las condiciones que deben cumplir los polinomios $p_1(x), \ldots, p_n(x)$ para que constituyan una varilla flexible que pase por los puntos (x_0, y_0), $\ldots$, (x_n, y_n) son:

1. Que pasen por los puntos dados:

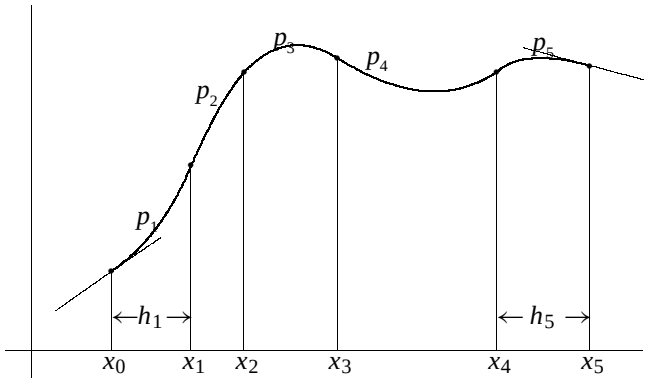


Figura 4.4: Varilla flexible

$p_1(x_0) = y_0$ y para cada $j = 1, \ldots, n$, $p_j(x_j) = y_j$.

2. Continuidad de la varilla:
Para cada $j = 1, \ldots, n-1$, $p_j(x_j) = p_{j+1}(x_j)$.

3. Continuidad de la pendiente o derivada:
Para cada $j = 1, \ldots, n-1$, $p'_j(x_j) = p'_{j+1}(x_j)$ y finalmente

4. Continuidad de la curvatura:
Para cada $j = 1, \ldots, n-1$, $p''_j(x_j) = p''_{j+1}(x_j)$.

Aquí hemos hecho un esfuerzo por expresar las tres condiciones de continuidad según fórmulas semejantes. En lugar de ello podríamos haber expresado las dos primeras condiciones conjuntamente como:

Para cada $j = 1, \ldots, n$,

1. $p_j(x_j) = y_j$, y
2. $p_j(x_{j-1}) = y_{j-1}$.

En cualquier caso lo que tenemos son un total de $4n-2$ ecuaciones para determinar los $4n$ coeficientes desconocidos de los n polinomios cúbicos

$$p_1(x), \ldots, p_n(x).$$

Necesitamos dos condiciones adicionales las cuales suelen imponerse en

los extremos x_0, x_n de la varilla. Estas condiciones adicionales pueden imponerse de varias maneras, dependiendo del contexto de nuestro problema.

Una de las condiciones más naturales, satisfecha por una varilla física que sólo está restringida a pasar sin rozamiento por los nodos (lo que da lugar al nombre de *condición de extremos libres*), consiste en tener curvatura cero en los extremos. Puesto que la curvatura es un múltiplo de la derivada segunda esta condición es equivalente a

$$p_1''(x_0) = 0 \quad \text{y} \quad p_n''(x_n) = 0.$$

Una segunda condición utilizada con frecuencia consiste en imponer en cada uno de los extremos la misma derivada segunda que en su nodo adyacente, es decir:

$$p_1''(x_0) = p_1''(x_1) \quad \text{y} \quad p_n''(x_n) = p_n''(x_{n-1}).$$

Por último veremos una condición en los extremos consistente en asignar en ellos una pendiente predeterminada a la curva, es decir, imponer un valor predeterminado a las derivadas

$$p_1'(x_0) = y_0' \quad \text{y} \quad p_n'(x_n) = y_n'.$$

Aunque estas tres condiciones son las más comunes, en modo alguno son las únicas posibles. Según las necesidades se pueden emplear distintas combinaciones de estas u otras condiciones.

Cálculo de las varillas flexibles cúbicas.–

Para el cálculo eficaz de los polinomios $p_1, \ldots, p_n$ que componen la varilla flexible los expresaremos en la forma $p_k(x) = a_k(x-x_k)^3 + b_k(x-x_k)^2 + c_k(x-x_k) + d_k$ de manera que las condiciones $p_k(x) = y_k$ nos proporcionan automáticamente el valor de los coeficientes de grado cero: $d_k = y_k$, con lo cual nos queda

$$p_k(x) = a_k(x-x_k)^3 + b_k(x-x_k)^2 + c_k(x-x_k) + y_k$$

y sólo necesitamos determinar tres coeficientes en cada polinomio.

En lo que sigue utilizaremos frecuentemente los valores de las distancias entre nodos consecutivos, las cuales denotaremos mediante $h_1, \ldots, h_n$, definidas por

$$h_k = x_k - x_{k-1}.$$

En términos de estas cantidades las condiciones de continuidad de la varilla (a saber: $p_k(x_{k-1}) = y_{k-1}$ para $k = 2, \ldots, n$) dan lugar a las ecuaciones

$$a_k(x_{k-1}-x_k)^3 + b_k(x_{k-1}-x_k)^2 + c_k(x_{k-1}-x_k) + y_k = y_{k-1},$$

es decir:

$$a_k h_k^3 - b_k h_k^2 + c_k h_k = y_k - y_{k-1} \tag{4.13}$$

para $k = 1, \ldots, n$.

Nuestro objetivo ahora es expresar cada uno de los coeficientes incógnita a_k, b_k, c_k en términos de una misma variable o parámetro, s_k, para reducir nuestro problema a un sistema de $n+1$ ecuaciones en las $n+1$ incógnitas $s_0, s_1, \ldots, s_n$. Estas nuevas incógnitas serán las derivadas segundas de los polinomios p_k en los nodos, es decir, definimos

$$s_k = p_k''(x_k),$$

para $k = 1, \ldots, n$, y además, el parámetro s_0 que es igual a la derivada segunda de p_1 en x_0, esto es,

$$s_0 = p_1''(x_0). \tag{4.14}$$

Teniendo en cuenta que las dos primeras derivadas de los p_k son

$$p_k'(x) = 3a_k(x-x_k)^2 + 2b_k(x-x_k) + c_k$$
$$p_k''(x) = 6a_k(x-x_k) + 2b_k$$

la definición anterior implica

$$\boxed{b_k = \tfrac{1}{2}s_k}$$

para $k = 1, \ldots, n$. Además,

Ejercicio 4.13
Las condiciones de continuidad de la curvatura (expresadas en términos de las s_k), junto con la definición de s_0 dada por la fórmula (4.14) *implican*

$$\boxed{a_k = \frac{s_k - s_{k-1}}{6h_k}}$$

para $k = 1, \ldots, n$.

Ejercicio 4.14
Sustituyendo en las ecuaciones (4.13) *la expresión de los coeficientes a_k dada en el ejercicio 4.13, así como las expresiones de los coeficientes b_k, obtenemos*

$$\boxed{c_k = \frac{y_k - y_{k-1}}{h_k} + \tfrac{1}{6}(s_{k-1} + 2s_k)h_k}$$

para $k = 1, \ldots, n$.

Con las expresiones dadas en estos dos ejercicios podemos obtener los coeficientes de los polinomios $p_1, \ldots, p_n$ una vez conocidos los parámetros s_k. Para hallar éstos usamos la condición de continuidad de la derivada, según la cual para $k = 1, \ldots, n-1$, $p'_k(x_k) = p'_{k+1}(x_k)$, es decir, $3a_k(x_k - x_k)^2 + 2b_k(x_k - x_k) + c_k = 3a_{k+1}(x_k - x_{k+1})^2 + 2b_{k+1}(x_k - x_{k+1}) + c_{k+1}$, lo que nos da

$$c_k = 3a_{k+1}h_{k+1}^2 - 2b_{k+1}h_{k+1} + c_{k+1}.$$

Introduciendo en esta ecuación las expresiones de los coeficientes a_k, b_k, c_k en términos de los parámetros s_k obtenemos las siguientes $n-1$ ecuaciones lineales en las $n+1$ incógnitas $s_0, \ldots, s_n$

$$\begin{aligned}&\frac{y_k - y_{k-1}}{h_k} + \tfrac{1}{6}(s_{k-1} + 2s_k)h_k \\ &= \tfrac{1}{2}(s_{k+1} - s_k)h_{k+1} - s_{k+1}h_{k+1} + \frac{y_{k+1} - y_k}{h_{k+1}} + \tfrac{1}{6}(s_k + 2s_{k+1})h_{k+1}\end{aligned}$$

y ordenando según los s_k,

$$h_k s_{k-1} + 2(h_k + h_{k+1})s_k + h_{k+1}s_{k+1} = \hat{b}_k, \tag{4.15}$$

donde hemos introducido los términos independientes

$$\hat{b}_k = 6\Big(\frac{y_{k+1} - y_k}{h_{k+1}} - \frac{y_k - y_{k-1}}{h_k}\Big).$$

En forma matricial las ecuaciones anteriores son:

$$\begin{pmatrix} h_1 & 2(h_1+h_2) & h_2 & & & \\ & h_2 & 2(h_2+h_3) & h_3 & & \\ & & & \ddots & & \\ & & & h_{n-1} & 2(h_{n-1}+h_n) & h_n \end{pmatrix} \begin{pmatrix} s_0 \\ s_1 \\ \vdots \\ s_{n-1} \\ s_n \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \vdots \\ \hat{b}_{n-1} \end{pmatrix}.$$

Necesitamos dos ecuaciones adicionales, las cuales, como hemos dicho, se obtienen generalmente de las condiciones en los extremos. Nosotros nos vamos a limitar a tres posibles conjuntos de condiciones:

Varillas de extremos libres.–

Este es el caso en el que se imponen las dos condiciones adicionales $s_0 = 0$ y $s_n = 0$. Entonces en nuestro sistema de ecuaciones se eliminan la primera (s_0) y última (s_n) incógnitas (lo que conlleva eliminar la primera y última columnas de la matriz de coeficientes) y queda

$$\begin{pmatrix} 2(h_1+h_2) & h_2 & & & \\ h_2 & 2(h_2+h_3) & h_3 & & \\ & \ddots & & & \\ & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} & \\ & & h_{n-1} & 2(h_{n-1}+h_n) & \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{n-2} \\ s_{n-1} \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \vdots \\ \hat{b}_{n-1} \end{pmatrix}$$

Nótese que el sistema que resulta tiene una matriz de coeficientes simétrica, de tipo Hessenberg y con diagonal dominante. Esto último por sí sólo garantiza la existencia de solución única.

Varillas de extremos con pendiente prefijada.–

Supongamos que se quiere imponer un valor determinado a la pendiente de la varilla flexible en los extremos x_0 y x_n: $p'_1(x_0) = y'_0$ y $p'_n(x_n) = y'_n$. Esto nos proporciona dos ecuaciones adicionales que podemos utilizar en lugar de las condiciones de extremos libres y el problema tendrá una solución que en general será distinta de la de extremos libres. Sólo se obtendrá la misma

solución si los valores de y'_0 e y'_n son precisamente los $p'_1(x_0)$ y $p'_n(x_n)$ de la varilla de extremos libres. En todo caso está claro que las varillas de extremos libres pueden considerarse un caso particular de la situación más general de varillas con pendiente prefijada en los extremos.

Por otro lado, se da la circunstancia de que este caso aparentemente más general de varillas flexibles puede obtenerse como un caso particular de las varillas de extremos libres. Una justificación geométrica de este hecho se basa en lo siguiente: Consideremos dos abscisas accesorias adicionales que denotamos x_{-1} y x_{n+1} y que haremos variar según los valores de un parámetro positivo ϵ del siguiente modo: $x_{-1} = x_0 - \epsilon$ y $x_{n+1} = x_n + \epsilon$. Asociada a cada una de estas abscisas consideramos la ordenada que nos da el punto sobre la recta que pasa por (x_0, y_0) con pendiente y'_0 (o, en el extremo opuesto, que pasa por (x_n, y_n) con pendiente y'_n), es decir, consideramos las ordenadas $y_{-1} = y_0 - y'_0\epsilon$ e $y_{n+1} = y_n + y'_n\epsilon$.

Ejercicio 4.15
La varilla de extremos libres que pasa por los puntos descritos más arriba

$$(x_{-1}, y_{-1}), (x_0, y_0), \ldots, (x_n, y_n), (x_{n+1}, y_{n+1})$$

tiene como posición límite cuando ϵ tiende a cero la de la varilla que pasa por $(x_0, y_0), \ldots, (x_n, y_n)$ con pendientes y'_0 e y'_n respectivamente en los extremos x_0 y x_n.

Teniendo en cuenta que los parámetros asociados con la varilla de extremos libres correspondiente a un cierto ϵ son:

$$h_0 = \epsilon, \quad h_{n+1} = \epsilon,$$

$$\hat{b}_0 = 6\Big(\frac{y_1 - y_0}{h_1} - y'_0\Big), \quad \hat{b}_n = 6\Big(y'_n - \frac{y_n - y_{n-1}}{h_n}\Big)$$

podemos concluir que el sistema de ecuaciones que determina la varilla flexible con pendiente prefijada en los extremos es

$$\begin{pmatrix} 2h_1 & h_1 & & & & \\ h_1 & 2(h_1+h_2) & h_2 & & & \\ & h_2 & 2(h_2+h_3) & h_3 & & \\ & & \ddots & & & \\ & & h_{n-1} & 2(h_{n-1}+h_n) & h_n & \\ & & & h_n & 2h_n \end{pmatrix} \begin{pmatrix} s_0 \\ s_1 \\ \vdots \\ s_{n-1} \\ s_n \end{pmatrix} = \begin{pmatrix} \hat{b}_0 \\ \vdots \\ \hat{b}_n \end{pmatrix}.$$

Naturalmente el mismo resultado se puede justificar algebraicamente a partir de las ecuaciones adicionales impuestas por las pendientes asignadas a los extremos, las cuales son,

$$3a_1h_1^2 - 2b_1h_1 + c_1 = y'_0 \qquad \text{y} \qquad c_n = y'_n$$

en las que hemos de expresar los coeficientes en términos de los s_k. Haciendo esto se obtienen las ecuaciones

$$\tfrac{1}{2}(s_1 - s_0)h_1 - s_1h_1 + \frac{y_1 - y_0}{h_1} + \tfrac{1}{6}(s_0 + 2s_1)h_1 = y'_0$$

y

$$\frac{y_n - y_{n-1}}{h_n} + \tfrac{1}{6}(s_{n-1} + 2s_n)h_n = y'_n,$$

que después de reordenar sus términos se pueden escribir como

$$2h_1s_0 + h_1s_1 = 6\Big(\frac{y_1 - y_0}{h_1} - y'_0\Big) \equiv \hat{b}_0$$

y

$$h_ns_{n-1} + 2h_ns_n = 6\Big(y'_n - \frac{y_n - y_{n-1}}{h_n}\Big) \equiv \hat{b}_n.$$

Unidas éstas a las (4.15) se llega al sistema de $n+1$ ecuaciones lineales en las $n+1$ incógnitas $s_0, \ldots, s_n$ que obtuvimos por el razonamiento geométrico.

Varillas con $s_0 = s_1$ y $s_n = s_{n-1}$.–

Al utilizar esta condición el sistema de ecuaciones (4.15) toma la forma

$$\begin{pmatrix} 3h_1+2h_2 & h_2 & & & \\ h_2 & 2(h_2+h_3) & h_3 & & \\ & \ddots & & & \\ & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} & \\ & & h_{n-1} & 2h_{n-1}+3h_n & \end{pmatrix} \begin{pmatrix} s_1 \\ s_2 \\ \vdots \\ s_{n-2} \\ s_{n-1} \end{pmatrix} = \begin{pmatrix} \hat{b}_1 \\ \vdots \\ \hat{b}_{n-1} \end{pmatrix}.$$

La matriz de coeficientes sigue siendo simétrica, de tipo Hessenberg y con diagonal dominante.

4.6.3 Un algoritmo para obtener varillas flexibles cúbicas de extremos libres

Para terminar esta sección sobre aproximación mediante varillas flexibles vamos a dar un algoritmo que evalúa varillas flexibles de extremos libres.

Algoritmo para evaluar varillas flexibles de extremos libres

1 **Datos:** n (número de intervalos o segmentos), $x_0, \ldots, x_n$ (abscisas de los nodos), $y_0, \ldots, y_n$ (ordenadas de los nodos), x (punto en el que se desea evaluar la varilla flexible).
2 $s_0 = 0$; $s_n = 0$;
3 **Para** $k = 1, n$
  $h_k = x_k - x_{k-1}$;
4 **Siguiente** k
5 **Para** $k = 1, n-1$
  $A(k,k) = 2(h_k + h_{k+1})$;
  $A(k,k+1) = h_{k+1}$;
  $A(k+1,k) = h_{k+1}$;
  $\hat{b}_k = 6[(y_{k+1} - y_k)/h_{k+1} - (y_k - y_{k-1})/h_k]$;
6 **Siguiente** k
7 **Para** $i = 2, n-1$
  $m = A(i,i-1)/A(i-1,i-1)$;
  $A(i,i) = A(i,i) - mA(i-1,i)$;
  $\hat{b}_i = \hat{b}_i - m\hat{b}_{i-1}$;
8 **Siguiente** i
9 **Para** $k = n-1, 1$ **incremento** -1
  $s_k = (\hat{b}_k - A(k,k+1)s_{k+1})/A(k,k)$;
10 **Siguiente** k

11 **Para** $k = 1, n$
  $a_k = (s_k - s_{k-1})/(6h_k)$;
  $b_k = s_k/2$;
  $c_k = (y_k - y_{k-1})/h_k + (s_{k-1} + 2s_k)h_k/6$;
12 **Siguiente** k
13 **Si** $x < x_0$ **entonces ir a 16**
14 **Para** $k = 1, n$
  Si $x \le x_k$ **entonces ir a 17**
15 **Siguiente** k
16 **Imprimir "**x no está contenido en el intervalo de interpolación.**" y PARAR**
17 $h = x - x_k$
18 $y = y_k + h(c_k + h(b_k + ha_k))$
19 **Imprimir "**Solución es: **";** y**; ".**" **y PARAR**

4.7 Interpolación Óptima y Polinomios de Chebyshev

4.7.1 Introducción y motivación

Estudiamos ahora un problema relacionado con el problema de interpolación polinómica pero planteado desde el punto ligeramente distinto de la teoría de la aproximación. Queremos poder evaluar una función dada, f, en cualquier punto de un intervalo $[a, b]$ contenido en su dominio de definición. En general no disponemos de una fórmula algebraica sencilla para evaluar una función dada a menos que ésta sea de un tipo muy especial. Por "fórmula sencilla" nos referimos a una fórmula que involucre solamente las operaciones aritméticas (como en los polinomios), lo cual descarta todas las funciones trascendentes.

Una posible idea es la de representar la función dada (supuesta diferenciable de orden suficientemente alto) mediante su polinomio de Taylor de cierto grado en torno a algún punto del intervalo de interés:

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \cdots + \frac{1}{n!} f^{(n)}(x_0)(x - x_0)^n + R_n$$

donde el resto R_n está dado por

$$R_n = \frac{f^{(n+1)}(\eta)}{(n+1)!}(x - x_0)^{n+1} \tag{4.16}$$

para algún η que dista de x_0 menos que x. Ya que los polinomios son funciones que se pueden evaluar fácilmente, el polinomio de Taylor nos ofrece una forma sencilla de evaluar aproximadamente f en cualquier punto cerca de x_0 siempre que podamos obtener con suficiente precisión el valor de f y el de sus derivadas en x_0.

El problema de este método es que, en general, el error aumenta rápidamente a medida que nos alejamos de x_0. Esto se deduce de la fórmula del error (4.16) que nos proporciona la fórmula de acotación

$$|R_n| \leq \frac{M}{(n+1)!}|x - x_0|^{n+1}$$

donde $M = \max_{a\leq\eta\leq b} |f^{(n+1)}(\eta)|$.

Intuitivamente este aumento del error se comprende fácilmente al pensar que el polinomio de Taylor de grado 1 representa la recta tangente a la gráfica de f en el punto de abscisa x_0 y cómo ésta aproxima en general peor y peor a f a medida que nos alejamos de x_0.

Como ilustración consideremos la evaluación de la función *seno* mediante su polinomio de Taylor de grado tres en torno al origen

$$\operatorname{sen} x \simeq x - \frac{x^3}{3!} = x\Big(1 - \frac{x^2}{6}\Big).$$

Los resultados se muestran en la tabla 4.1, donde se ve que el error aumenta muy rápidamente a medida que nos alejamos de $x = 0$.

Se plantea, pues, la cuestión de hallar un polinomio que aproxime una función dada en un intervalo $[a, b]$ con más uniformidad que el polinomio de Taylor. Una posible idea es utilizar un polinomio de interpolación. Para ello necesitamos elegir los nodos de interpolación y evaluar la función dada en esos nodos.

4.7.2 El concepto de interpolación óptima

Surge, pues, la cuestión de cómo elegir los nodos de interpolación y de si habrá algún criterio según el cual una elección es mejor que otra. Si aproximamos la función f mediante un polinomio de interpolación de grado prefijado, digamos n, el error cometido en un punto x está dado por una

x	$p_3(x)$	$\operatorname{sen} x$	Error
0.1	0.0998333	0.0998334	0.0000001
0.2	0.1986667	0.1986693	0.0000027
0.3	0.2955000	0.2955202	0.0000202
0.4	0.3893333	0.3894183	0.0000850
0.5	0.4791667	0.4794255	0.0002589
0.6	0.5640000	0.5646425	0.0006425
0.7	0.6428333	0.6442177	0.0013844
0.8	0.7146667	0.7173561	0.0026894
0.9	0.7785000	0.7833269	0.0048269

Tabla 4.1: Aproximación de Taylor de grado 3, $p_3(x) = x - x^3/6$.

expresión parecida a (4.16):

$$|f(x) - P_n(x)| = \left|\frac{f^{(n+1)}(\eta)}{(n+1)!}(x - x_0)\cdots(x - x_n)\right|$$

donde $x_0, \ldots, x_n$ son los nodos de interpolación y η es algún punto del intervalo de interpolación. Supongamos que la derivada $n+1$ de la función f está acotada en el intervalo de interés, o sea, que existe un número M tal que para todo $x \in [a, b]$, se cumple $|f^{n+1}(x)| \leq M$. Entonces el error cometido al interpolar f en los nodos $x_0, \ldots, x_n$ está acotado por

$$|f(x) - P_n(x)| \leq \frac{M}{(n+1)!}\max_{a\leq x\leq b}|(x - x_0)\cdots(x - x_n)|$$

y nuestra mejor estrategia es elegir los nodos $x_k \in [a, b]$ de tal forma que la cantidad

$$\max_{a\leq x\leq b}|(x - x_0)\cdots(x - x_n)| \tag{4.17}$$

sea lo más pequeña posible. Los $n+1$ puntos del intervalo $[a, b]$, $x_0, \ldots, x_n$, que minimizan la cantidad (4.17) serán llamados los *nodos de interpolación óptima de grado n en* $[a, b]$. Dado que la condición que define los nodos de interpolación óptima no depende de la función a aproximar, una vez hallados, éstos servirán para interpolar de forma óptima cualquier función en el intervalo $[a, b]$.

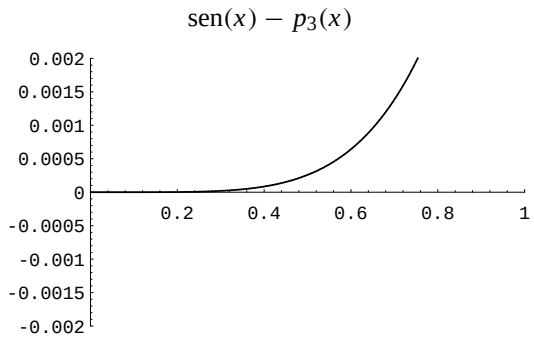


Figura 4.5: Error en la aproximación de la función seno mediante el polinomio de Taylor de grado 3, $p_3(x) = x - x^3/6$.

Podemos replantear nuestro problema como el problema de hallar un polinomio mónico de grado $n+1$, cuyos ceros estén en un intervalo dado $[a,b]$ y cuyo máximo valor absoluto en $[a,b]$ sea lo más pequeño posible.

Se puede reducir el problema de hallar los nodos de interpolación óptima en un intervalo arbitrario $[a,b]$ al mismo problema en el intervalo $[-1,1]$. Los resultados se pueden trasladar después al intervalo $[a,b]$ mediante la transformación compuesta de la homotecia que convierte la longitud del intervalo $[-1,1]$ en la longitud del intervalo $[a,b]$ (es decir, la razón $r = (b-a)/2$), seguida de la traslación que lleva el origen al centro del intervalo $[a,b]$ (o sea, "sumar $h = (a+b)/2$"), en resumen, la transformación:

$$t(x) = rx + h = \frac{b-a}{2}x + \frac{b+a}{2}.$$

Así pues, el problema toma ahora la siguiente forma: encontrar $n+1$ puntos $x_0, , \ldots, x_n \in [-1,1]$ tales que la cantidad

$$\max_{-1\le x\le 1} |(x-x_0)\cdots(x-x_n)|$$

tenga el menor valor posible. Equivalentemente: encontrar el polinomio mónico de grado $n+1$ (que denotaremos $\hat{T}_{n+1}(x)$) cuya norma ∞ en el intervalo $[-1,1]$ sea lo más pequeña posible. No es difícil descubrir cuáles son dichos polinomios para valores pequeños de n como $n=0$, $n=1$, e incluso $n=2$. En el caso general los polinomios de interpolación óptima

están directamente relacionados con los llamados *polinomios de Chebyshev* que estudiamos a continuación.

4.7.3 Los polinomios de Chebyshev

¿Qué son los polinomios de Chebyshev? Pensemos por un momento en el siguiente hecho conocido:

Ejercicio 4.16
Para $n = 0, 1, \ldots$ el coseno de $n\alpha$ puede expresarse como combinación lineal de las potencias

$$(\cos\alpha)^0, \ldots, (\cos\alpha)^n.$$

Definición: El *polinomio de Chebyshev* de grado n, $T_n(x)$, es el polinomio cuyos coeficientes son los de las potencias $(\cos t)^k$ en la expresión de $\cos nt$ en términos de $\cos t$. Es decir, $T_n(x)$ está definido por la propiedad

$$T_n(\cos\theta) = \cos n\theta.$$

Ejemplos:

$$T_0(x) = 1; \quad T_1(x) = x; \quad T_2(x) = 2x^2 - 1$$

ya que $\cos 0\theta = 1$; $\cos\theta = \cos\theta$; y $\cos 2\theta = 2\cos^2\theta - 1$.

La importancia de los polinomios de Chebyshev para nuestro problema radica en el siguiente resultado:

Teorema 27 (Nodos de Chebyshev) *Los nodos de interpolación óptima de grado n en el intervalo $[-1,1]$ son los $n+1$ ceros del polinomio de Chebyshev de grado $n+1$, $T_{n+1}(x)$.*

Es decir, los nodos de Chebyshev para el grado n son los puntos $x_k = \cos\theta_k$ donde $\theta_0, \ldots, \theta_n$ son los ángulos tales que

$$\cos[(n+1)\theta_k] = 0$$

o sea,

$$\theta_0 = \tfrac{1}{n+1}\cdot\tfrac{\pi}{2}\,,\ \theta_1 = \tfrac{3}{n+1}\cdot\tfrac{\pi}{2}\,,\ \ldots\,,\ \theta_k = \tfrac{2k+1}{n+1}\cdot\tfrac{\pi}{2}\,,\ \ldots\,,\ \theta_n = \tfrac{2n+1}{n+1}\cdot\tfrac{\pi}{2}$$

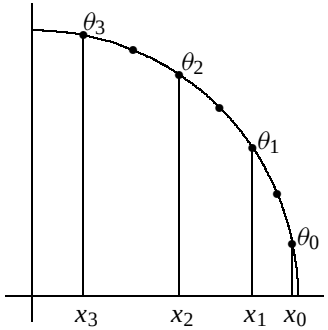


Figura 4.6: Nodos de Chebyshev no negativos para interpolar con polinomio de grado $n = 7$.

con lo que los nodos de interpolación óptima son

$$\boxed{x_k = \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right)}.$$

Nótese que estos puntos son las abscisas de los puntos del círculo unitario determinados por los ángulos θ_k. Dicho de otra forma, los x_k son las partes reales de los números complejos unitarios $e^{i\theta_k}$. En consecuencia pueden dibujarse fácilmente sin más que dividir la semicircunferencia cuyo diámetro es el intervalo $[-1, 1]$ en $n + 1$ arcos iguales y hallando las abscisas de los puntos medios de dichos arcos. En realidad, habida cuenta de la simetría, es suficiente hallar los nodos correspondientes a la parte positiva como se muestra en la figura 4.6.

4.7.4 Demostración del teorema de los nodos de Chebyshev

Para demostrar el teorema 27 observemos primero lo siguiente:

Proposición 8 *Supongamos que $T(x)$ es un polinomio mónico de grado n que tiene n raíces reales en el intervalo $[a, b]$ y tal que todos sus extremos en $[a, b]$ tienen el mismo valor absoluto. Entonces sus raíces son los nodos de interpolación óptima en $[a, b]$.*

Demostración: Tenemos que demostrar que todo otro polinomio mónico de grado n alcanza en $[a, b]$ valores (absolutos) mayores que $m = \max_{a\le x\le b} T(x)$. Sea, pues, $P(x)$ un polinomio mónico de grado n distinto de $T(x)$. Evidentemente la diferencia

$$Q(x) = T(x) - P(x)$$

es un polinomio de grado estrictamente menor que n. Supongamos que en todo punto $x \in [a, b]$ fuese $|P(x)| \le m$, entonces, tanto en los puntos críticos de T como en los extremos a, b del intervalo el signo de $Q(x)$ es igual al signo de $T(x)$. En consecuencia $Q(x)$ tiene al menos tantos cambios de signo como $T(x)$ y por tanto tiene también al menos tantas raíces como T, esto es, n. Pero como $Q(x)$ tiene grado menor que n, es necesariamente el polinomio nulo; esto es, $P(x)$ es igual a $T(x)$. ■

El siguiente es un simple ejercicio de máximos y mínimos. Con él se establece el hecho de que en el intervalo $[-1, 1]$ todo polinomio de Chebyshev toma sus valores entre -1 y 1. Éste es un importante resultado que se utilizará en la demostración del teorema de los nodos de Chebyshev.

Ejercicio 4.17
Evaluar $T_n(x)$ para $x = \pm 1$ y usar el resultado para demostrar que

$$\max_{x\in[-1,1]} |T_n(x)| = 1.$$

Ejercicio 4.18
Demostrar que los extremos relativos del polinomio de Chebyshev de grado n son los números de la forma $\bar{x}_k = \cos(k\pi/n)$ para $k \in \{1, \ldots, n-1\}$. Además también son extremos los puntos $a = -1$ y $b = 1$, que se obtienen de la fórmula de los $\bar{x}_k$ para $k = 0$ y $k = n$. Por lo tanto podemos decir que los extremos son los $\bar{x}_k$ para $k \in \{0, \ldots, n\}$, pero cuidado: para $k = 0$ y $k = n$ $T_n'(\bar{x}_k) \neq 0$ (demuéstrese).

Estamos ahora en condiciones de demostrar nuestro teorema. La demostración es muy sencilla; sólo necesitamos demostrar lo siguiente:

Proposición 9 *El máximo valor de $|\frac{1}{2^{n-1}} T_n(x)|$ en $[-1, 1]$ es $\frac{1}{2^{n-1}}$ y para todo otro polinomio mónico de grado n, $P_n(x)$, su valor absoluto alcanza, en alguno de los puntos $\bar{x}_k = \cos(k\pi/n)$, valores $\ge \frac{1}{2^{n-1}}$.*

Demostración: Por reducción al absurdo. Supongamos que para todo $k \in \{0, \ldots, n\}$ se verifica $|P_n(\bar{x}_k)| < \frac{1}{2^{n-1}}$. Esto es lo mismo que decir que el polinomio

$$Q(x) = \frac{1}{2^{n-1}} T_n(x) - P_n(x)$$

tiene en cada $\bar{x}_k$ el mismo signo que T_n (esto es, $(-1)^k$ —ejercicio (4.18)). Vamos a ver que bajo la suposición hecha el polinomio $Q(x)$ tendría grado menor que n y, al mismo tiempo, n cambios de signo en $[-1, 1]$, lo cual es absurdo ya que *"el número total de cambios de signo de un polinomio de grado n es menor o igual que n"*. El polinomio $Q(x)$ no es cero y, por ser diferencia de dos polinomios mónicos de grado n, su grado es menor que n. Por otra parte

$$Q(\bar{x}_k) = \frac{1}{2^{n-1}} T_n(\bar{x}_k) - P_n(\bar{x}_k) = \frac{(-1)^k}{2^{n-1}} - P_n(\bar{x}_k)$$

y el signo de la diferencia es el mismo que el del minuendo porque, por hipótesis, el sustraendo tiene menor valor absoluto. En consecuencia $Q(x)$ tiene n cambios de signo en $[-1, 1]$ y por ser una función continua ha de tener n ceros en $[-1, 1]$. Pero es un polinomio no nulo de grado menor que n. Contradicción. Luego para algún $\bar{x}_k \in [-1, 1]$ se verifica $|P_n(\bar{x}_k)| \geq \frac{1}{2^{n-1}}$. ■

Con esto queda establecida la fórmula de los nodos de Chebyshev o nodos de interpolación óptima en el intervalo $[-1, 1]$. Nuestra tarea ahora es obtener las fórmulas correspondientes para un intervalo arbitrario.

4.7.5 Interpolación óptima en un intervalo arbitrario

Si necesitamos aproximar una función $f(x)$ en un intervalo $[a, b]$ distinto de $[-1, 1]$ hemos de trasladar los nodos de Chebyshev

$$x_k = \cos \frac{(2k+1)\pi}{2(n+1)}$$

del intervalo $[-1, 1]$ al intervalo $[a, b]$. El resultado es:

Ejercicio 4.19
Demostrar que los nodos de interpolación óptima de grado n en un intervalo $[a, b]$ *son*

$$\boxed{y_k = \frac{b-a}{2} \cos\left(\frac{(2k+1)\pi}{2(n+1)}\right) + \frac{b+a}{2}}$$

x	$P_3(x)$	sen x	Error
0.1	0.0999441	0.0998334	−0.0001107
0.2	0.1987851	0.1986693	−0.0001158
0.3	0.2955310	0.2955202	−0.0000108
0.4	0.3893151	0.3894183	0.0001033
0.5	0.4792708	0.4794255	0.0001548
0.6	0.5645314	0.5646425	0.0001111
0.7	0.6442302	0.6442177	−0.0000125
0.8	0.7175005	0.7173561	−0.0001444
0.9	0.7834758	0.7833269	−0.0001489

Tabla 4.2: Interpolación de Chebyshev de grado 3

Nótese que, como dijimos más arriba, el resultado no es más que el de aplicar al intervalo $[-1, 1]$ la homotecia que cambia su longitud de ser 2 a ser la del intervalo $[a, b]$ y después aplicar una traslación al resultado para que sus extremos coincidan con los del intervalo $[a, b]$, es decir, después de la homotecia, trasladar el origen al punto medio del intervalo $[a, b]$. En consecuencia, dado que las homotecias y las traslaciones convierten circunferencias en circunferencias y rectas en rectas, la construcción geométrica dada para los nodos de Chebyshev puede aplicarse a cualquier intervalo.

Ejercicio 4.20
Demostrar que para una función $f(x)$ *cuya derivada* $n+1$ *está acotada en el intervalo* $[a, b]$ *por la constante* M*, el error cometido en la aproximación de Chebyshev de grado* n *en el intervalo* $[a, b]$*,* $P_n(x)$*, está acotado por*

$$|f(x) - P_n(x)| \leq \frac{2M}{(n+1)!} \left(\frac{b-a}{4}\right)^{n+1}.$$

4.7.6 Ejemplo de la interpolación de Chebyshev

Vamos a comparar el resultado de aproximar la función *seno* en $[0, 1]$ por el método de Chebyshev con un polinomio de grado tres que interpole en los nodos óptimos con la aproximación de la misma función en el mismo intervalo por el polinomio de Taylor de grado cinco hecha al principio de esta lección.

Los cuatro nodos en el intervalo $[0, 1]$ están dados por $x_k = \frac{1}{2}\cos\frac{(2k+1)\pi}{8} + \frac{1}{2}$, o sea que son:

$$x_0 = 0.5\Big(1+\cos\frac{\pi}{8}\Big) = 0.9619398\,, \quad x_1 = 0.5\Big(1+\cos\frac{3\pi}{8}\Big) = 0.6913417$$
$$x_2 = 0.5\Big(1+\cos\frac{5\pi}{8}\Big) = 0.3086583\,, \quad x_3 = 0.5\Big(1+\cos\frac{7\pi}{8}\Big) = 0.0380602$$

Para hallar el polinomio de interpolación calculamos las diferencias divididas

x	sen x			
0.9619398	0.8203025			
		0.6752862		
0.6913417	0.6375714		−0.3014799	
		0.8722374		−0.1444449
0.3086583	0.3037806		−0.1680302	
		0.9820084		
0.0380602	0.0380510			

con lo que el polinomio de interpolación es

$$\begin{aligned} P_3(x) &= 0.8203025 + (x - 0.9619398)[0.6752862 \\ &\quad + (x - 0.6913417)(-0.3014799 - 0.1444449(x - 0.3086583))] \\ &= -0.1444449x^3 - 0.01808759x^2 + 1.003947x - 0.0001252498 \\ &= [-(0.1444449x + 0.01808759)x + 1.003947]x - 0.0001252498. \end{aligned}$$

Podemos ahora evaluar este polinomio en varios puntos del intervalo $[0, 1]$ para comparar los resultados con los valores exactos de la función *seno* en esos puntos. El resultado se muestra en la tabla 4.2. Compárese con la tabla 4.1.

Ejercicio 4.21
Utilizar la fórmula del error dada en el ejercicio 4.20 para estimar la cota del error correspondiente a la aproximación de Chebyshev de grado 3 de la función seno *en el intervalo* $[0, 1]$*. Comprobar que los errores obtenidos en la tabla 4.2 concuerdan con la cota de error estimada. ¿Cuál sería la cota del error si se usase el polinomio de grado 5?*

$\text{sen}(x) - P_3(x)$

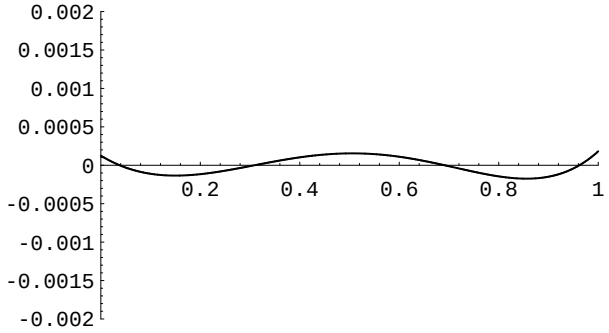


Figura 4.7: Error en la aproximación de Chebyshev de grado 3 de la función seno.

Respuestas a Algunos Ejercicios del Capítulo 4

Ejercicio 4.1 Esto se puede hacer por inducción en n con ayuda de la sugerencia dada. Para $n = 1$ se cumple $\det\begin{pmatrix} x_0 & 1 \\ x_1 & 1 \end{pmatrix} = x_0 - x_1$. Supuesto cierto para n puntos (distintos) $x_0, ..., x_{n-1}$, y añadamos a éstos un nuevo punto variable t, de forma que obtenemos una matriz variable

$$M(t) = \begin{pmatrix} x_0^n & x_0^{n-1} & \cdots & x_0 & 1 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ x_{n-1}^n & x_{n-1}^{n-1} & \cdots & x_{n-1} & 1 \\ t^n & t^{n-1} & \cdots & t & 1 \end{pmatrix}$$

y nos preguntamos qué tipo de función es la definida por $p(t) = \det M(t)$. Evidentemente es un polinomio en t. Su grado es n, tiene a cada una de las $x_0, \ldots, x_{n-1}$ como raiz (por lo que, dado su grado, no tiene más raices que

esas) y su coeficiente principal es

$$(-1)^n \det \begin{pmatrix} x_0^{n-1} & \cdots & x_0 & 1 \\ \vdots & \cdots & \vdots & \vdots \\ x_{n-1}^{n-1} & \cdots & x_{n-1} & 1 \end{pmatrix} = (-1)^n \prod_{0 \le i < j \le n-1} (x_i - x_j)$$

Todo esto determina completamente el polinomio $p(t)$ como

$$\begin{aligned} p(t) &= (\text{coef. ppal}) \times (t - x_0) \cdots (t - x_{n-1}) \\ &= (-1)^n \prod_{0 \le i < j \le n-1} (x_i - x_j) \times (t - x_0) \cdots (t - x_{n-1}), \end{aligned}$$

y por tanto

$$\det M(x_n) = p(x_n) = \prod_{0 \le i < j \le n} (x_i - x_j).$$

Ejercicio 4.2 Que estén sobre la gráfica de un polinomio de grado menor que n.

Ejercicio 4.3 Uno.

Ejercicio 4.4 Según la fórmula

$$p(x) = \frac{f(x_0)}{q_0(x_0)} q_0(x) + \cdots + \frac{f(x_n)}{q_n(x_n)} q_n(x)$$

el coeficiente del término de grado n de p es la suma de los coeficientes de los términos de grado n de los sumandos, es decir, suma de los $\frac{f(x_k)}{q_k(x_k)}$ ya que los $q_k(x)$ son polinomios mónicos de grado n.

Ejercicio 4.5

Ejercicio 4.6 Usando la fórmula de Leibnitz para la derivada de un producto se obtiene $q'(x) = \sum_{k=0}^{n} q_k(x)$ de donde $q'(x_k) = q_k(x_k)$. Alternativamente el resultado se obtiene de forma inmediata derivando $q(x) = q_k(x)(x - x_k)$ y evaluando el resultado en x_k.

Ejercicio 4.7 Es evidente que los polinomios de Lagrange son

$$L_k(x) = \frac{q(x)}{(x - x_k) q_k(x_k)},$$

de lo cual se deduce inmediatamente la fórmula dada.

Ejercicio 4.8 Evidentemente A_n es el coeficiente del término de grado n del polinomio $A_n(x - x_0) \cdots (x - x_{n-1})$, que es el polinomio diferencia $P_n(x) - P_{n-1}(x)$. Pero esta diferencia tiene el mismo coeficiente del término de grado n que $P_n(x)$ ya que $P_{n-1}(x)$ tiene grado estrictamente menor que n. La fórmula dada no es más que una reescritura de (4.2).

Ejercicio 4.9 Para todo polinomio $p(x) = a_n x_n + \cdots$ de grado menor o igual que n la derivada n-ésima de p es la constante $d^n p/dx^n = n! a_n$. Si p tuviese grado menor que n la derivada n-ésima de p sería cero. En cualquier caso $\frac{1}{n!} d^n p/dx^n$ es igual al coeficiente del término de grado n de p.

Ejercicio 4.10 Evidentemente es un polinomio de grado $\le k+1$ ya que p y q tienen grado $\le k$ y van multiplicados por polinomios de grado 1. Que coincide con f en x_k y en x_{k+1} es inmediato porque p coincide con f en x_{k+1} y q en x_k. En los demás puntos $x_i \in \{x_0, \ldots, x_{k-1}\}$ tenemos

$$\begin{aligned} \frac{x_i - x_k}{x_{k+1} - x_k} p(x_i) + \frac{x_{k+1} - x_i}{x_{k+1} - x_k} q(x_i) &= \frac{x_i - x_k}{x_{k+1} - x_k} f(x_i) + \frac{x_{k+1} - x_i}{x_{k+1} - x_k} f(x_i) \\ &= \left(\frac{x_i - x_k}{x_{k+1} - x_k} + \frac{x_{k+1} - x_i}{x_{k+1} - x_k} \right) f(x_i) = f(x_i). \end{aligned}$$

Ejercicio 4.11 Usando la forma

$$\frac{x - x_k}{x_{k+1} - x_k} p(x) + \frac{x_{k+1} - x}{x_{k+1} - x_k} q(x)$$

del polinomio de interpolación de f en $x_0, \ldots, x_{k+1}$ (ver ejercicio 4.10) y sustituyendo los coeficientes del término de grado k en p y en q, se obtiene inmediatamente la expresión dada para el término de grado $k+1$ del polinomio de interpolación en $x_0, \ldots, x_{k+1}$.

Ejercicio 4.12

$$0.65504 = \frac{0.5 \times 0.62325 - 0.3 \times 0.60206}{0.5 - 0.3}, \quad 0.65380 = \frac{0.5 \times 0.64345 - 0.1 \times 0.60206}{0.5 - 0.1}$$

$$0.65264 = \frac{0.5 \times 0.66276 + 0.1 \times 0.60206}{0.5 + 0.1}, \quad 0.65155 = \frac{0.5 \times 0.68124 + 0.3 \times 0.60206}{0.5 + 0.3}$$

$$0.65318 = \frac{0.3 \times 0.65380 - 0.1 \times 0.65504}{0.5 - 0.3}, \quad 0.65324 = \frac{0.3 \times 0.65264 + 0.3 \times 0.65504}{0.5 - 0.3}$$

$$0.65330 = \frac{0.3\times 0.65155 - 0.3\times 0.65504}{0.5-0.3}\ , \quad 0.65321 = \frac{0.1\times 0.65324 + 0.1\times 0.65318}{0.1+0.1}$$

$$0.65321 = \frac{0.1\times 0.65330 + 0.3\times 0.65318}{0.1+0.3},$$

y dado que los dos últimos coinciden, el siguiente debe ser también igual a ellos.

Ejercicio 4.13 Según la definición de los s_k, $p_k''(x) = 6a_k(x - x_k) + s_k$. Por tanto, la condición de continuidad de la curvatura, $p_{k+1}''(x_k) = p_k''(x_k)$, $(k = 1, \ldots, n-1)$ implica que $6a_{k+1}(x_k - x_{k+1}) + s_{k+1} = s_k$, de donde $s_{k+1} - s_k = 6a_{k+1}h_{k+1}$ para $k = 1, \ldots, n-1$, lo que es equivalente a la fórmula dada.

Ejercicio 4.14 Las ecuaciones (4.13) son $a_k h_k^3 - b_k h_k^2 + c_k h_k = y_k - y_{k-1}$ para $k = 2, \ldots, n$. Poniendo en ellas $a_k = (s_k - s_{k-1})/6h_k$ y $b_k = \frac{1}{2}s_k$ obtenemos $((s_k - s_{k-1})/6h_k)h_k^3 - \frac{1}{2}s_k h_k^2 + c_k h_k = y_k - y_{k-1}$, de donde $\frac{1}{6}(s_k - s_{k-1})h_k - \frac{1}{2}s_k h_k + c_k = (y_k - y_{k-1})/h_k$ y despejando c_k, $c_k = (y_k - y_{k-1})/h_k - \frac{1}{6}(s_k - s_{k-1})h_k + \frac{3}{6}s_k h_k$.

Ejercicio 4.15

Ejercicio 4.16 Esto puede hacerse fácilmente por inducción. Es claramente cierto para $n = 0, 1$. Si lo es para $n-1$ y para n entonces

$$\begin{aligned}\cos\big((n+1)\alpha\big) &= \cos\big((n\alpha + \alpha\big)\\ &= \cos n\alpha\cos\alpha + \operatorname{sen} n\alpha \operatorname{sen}\alpha\\ &= 2\cos n\alpha\cos\alpha - \big(\cos n\alpha\cos\alpha - \operatorname{sen} n\alpha\operatorname{sen}\alpha\big)\\ &= 2\cos n\alpha\cos\alpha - \cos\big((n-1)\alpha\big).\end{aligned}$$

Ejercicio 4.17 Sabemos que $\max_{x\in[-1,1]}|T_n(x)| \le 1$ porque en ese intervalo x es el coseno de un ángulo y por lo tanto T_n también. Evaluando, $T_n(1) = T_n(\cos 0) = \cos n0 = 1$, $T_n(-1) = T_n(\cos\pi) = \cos n\pi = (-1)^n$. En consecuencia $\max_{x\in[-1,1]}|T_n(x)| = 1$.

Ejercicio 4.18 No hay más que derivar $T_n(x)$ e igualar a cero para obtener que la condición de extremo es $\operatorname{sen}(n\theta)/\operatorname{sen}\theta = 0$, de donde $x = \cos\theta$ con $\theta \neq m\pi$ y $\operatorname{sen} n\theta = 0$ y por tanto $\bar{x}_k = \cos(k\pi/n)$ para $k \in \{1, \ldots, n-1\}$. En estos puntos tenemos $T_n(\cos\frac{k\pi}{n}) = \cos(n\frac{k\pi}{n}) = \cos k\pi = (-1)^k$. Además, en los extremos del intervalo $[-1, 1]$ tenemos

$T_n(1) = T_n(\cos 0) = \cos 0 = 1$ y $T_n(-1) = T_n(\cos\pi) = \cos n\pi = (-1)^n$ y por tanto el máximo valor absoluto de $T_n(x)$ para x en el intervalo $[-1, 1]$ es 1.

Ejercicio 4.19 La transformación lineal $y = px + q$ que lleva -1 en a y 1 en b es $y = \frac{b-a}{2}x + \frac{b+a}{2}$, con lo que se obtiene la fórmula dada.

Ejercicio 4.20 Sabemos que el error del polinomio de interpolación en los nodos $y_0, \ldots, y_n$ está acotado por:

$$|f(x) - P_n(x)| \le \frac{M}{(n+1)!}|(x - y_0)\cdots(x - y_n)|$$

en nuestro caso los nodos son, según el ejercicio 4.19, $y_k = px_k + q$ donde x_k son los correspondientes nodos de Chebyshev, $p = (b-a)/2$ y $q = (b+a)/2$. Entonces

$$x - y_k = x - (px_k + q) = x - q - px_k = p\Big(\frac{x-q}{p} - x_k\Big)$$

luego

$$\begin{aligned}|(x - y_0)\cdots(x - y_n)| &= \Big|p^{n+1}\Big(\frac{x-q}{p} - x_0\Big)\cdots\Big(\frac{x-q}{p} - x_n\Big)\Big|\\ &= \Big|p^{n+1}\frac{1}{2^n}T_{n+1}\Big(\frac{x-q}{p}\Big)\Big| = p^{n+1}\frac{1}{2^n}\Big|T_{n+1}\Big(\frac{x-q}{p}\Big)\Big|\\ &\le p^{n+1}\frac{1}{2^n} = \Big(\frac{b-a}{2}\Big)^{n+1}\frac{1}{2^n} = 2\Big(\frac{b-a}{4}\Big)^{n+1}\end{aligned}$$

donde hemos usado que para $x \in [a, b]$ se cumple $\frac{x-q}{p} \in [-1, 1]$ y por lo tanto $|T_{n+1}(\frac{x-q}{p})| \le 1$. Así pues,

$$|f(x) - P_n(x)| \le \frac{2M}{(n+1)!}\Big(\frac{b-a}{4}\Big)^{n+1}.$$

Ejercicio 4.21 El error estará acotado en valor absoluto por

$$\frac{2M}{4!}\left(\frac{1}{4}\right)^4 = \frac{M}{3\times 4^5} = \frac{\operatorname{sen} 1}{3072} = \frac{0.84147}{3072} = \boxed{0.000274},$$

donde hemos acotado la función $d^4\operatorname{sen}(x)/dx^4 = \operatorname{sen} x$ por su valor en $x = 1$. Si interpolamos con grado 5, tenemos que acotar la función $d^6\operatorname{sen}(x)/dx^6 =$

$-\operatorname{sen} x$ para la cual nos sirve la misma cota que en el caso anterior. La cota del error es

$$\frac{2M}{6!}\left(\frac{1}{4}\right)^6 = \frac{M}{90\times 4^7} = \frac{0.84147}{1474560} = \boxed{0.000000571}.$$

Capítulo 5

Resolución de ecuaciones diferenciales ordinarias

Las ecuaciones iferenciales ordinarias (esto es, las que no contienen derivadas parciales) de orden superior siempre se pueden reducir a sistemas de ecuaciones diferenciales de primer orden. Para hacer esto no hay más que dar nombre a las sucesivas derivadas de la incógnita (por ejemplo, $u(x) = f'(x)$, $v(x) = f''(x)$, $w(x) = f'''(x)$, etc.), de forma que la derivada más alta pasa a ser una derivada primera. Después sólo queda expresar estas igualdades en términos de derivadas primeras solamente (es decir: $u(x) = f'(x)$, $v(x) = u'(x)$, $w(x) = v'(x)$, etc.), y utilizar esos nombres en lugar de las correspondientes derivadas de la incógnita en la ecuación dada. Así, una ecuación tal como

$$2(y'')^2 - 5\operatorname{sen}(y') + xyy''' = 0$$

se sustituiria por el sistema

$$\begin{aligned} 2v^2 - 5\operatorname{sen} u + xyv' &= 0\\ u &= y'\\ v &= u'. \end{aligned}$$

Esto justifica la suficiencia de estudiar los métodos de resolución de ecuaciones de primer orden. En este curso nos limitaremos a los métodos

de resolución de una sola ecuación diferenciable de primer orden, pero se debe comprender que éstos se pueden adaptar fácilmente a la resolución de sistemas.

Una importante cuestión asociada con los problemas de ecuaciones diferenciales es la relativa a las condiciones de contorno. Los métodos de resolución dependen fuertemente de si estas condiciones se imponen en un único punto (problemas de valores iniciales) o en varios puntos (problemas de valores en la frontera). En este tema estudiaremos principalmente las nociones fundamentales de los métodos numéricos para problemas de valores iniciales, dedicando al final una breve sección a las ideas generales de tres tipos de métodos elementales para problemas de valores en la frontera (el de las *diferencias finitas*, el de *disparo* y los de *colocación*).

5.1 Problemas de valores iniciales

El problema de valores iniciales para las ecuaciones diferenciales ordinarias de primer orden consiste en evaluar en un intervalo dado $[a, b]$ funciones $y(x)$ que verifican

$$y'(x) = f(x, y(x)), \qquad y(x_0) = y_0 \tag{5.1}$$

donde f es una función real de dos variables reales. Obsérvese que la solución de (5.1) tiene la propiedad

$$y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\, dt, \tag{5.2}$$

expresión que debe considerarse como una formulación equivalente del problema de valores iniciales (5.1). En lo que sigue supondremos $x_0 = a$.

5.1.1 El algoritmo de Euler

El método numérico más sencillo para resolver un problema de valores iniciales como (5.1) es el algoritmo de Euler. Éste se basa en la fórmula

$$y(x+h) = y(x) + hy'(x) + \tfrac{1}{2}h^2 y''(\xi)$$

en base a la cual, elegido un número h suficientemente pequeño como para que h^2 sea despreciable, formamos las sucesiones

$$x_{i+1} = x_i + h, \qquad y_{i+1} = y_i + hf(x_i, y_i). \tag{5.3}$$

Esto se puede conseguir mediante el siguiente algoritmo

Algoritmo de Euler

Datos: x_0, b (extremos del intervalo)**,** h (longitud del paso)**,** y_0 (valor inicial)**.**

1 $y_0 = y_0 + hf(x_0, y_0)$**;** $x_0 = x_0 + h$**;**
Imprimir x_0**,** y_0**;**
Si $x_0 < b$ **entonces ir al paso 1;**
Stop.

Para poder hacer un uso eficaz de un método numérico como éste, es necesario calcular también la cota del error que se comete en cada paso. Esto puede hacerse de varias maneras, dependiendo de la cantidad que queramos tomar como indicativa del error. El error en que quizás se piense primero es la diferencia $y(x_i) - y_i$ entre el valor exacto de la función incógnita en cada nodo y el valor estimado en ese nodo. Esta cantidad se llama *error global*. Es sencillo demostrar que para el método de Euler,

Ejercicio 5.1
Supongamos que aplicamos el método de Euler al problema (5.1) *y que* $f(x, y)$ *y* $f_y = \partial f/\partial y$ *son funciones continuas en el rectángulo* $x_0 < x < b$, $c < y < d$ *y la solución de* (5.1) *satisface para todo* $x_0 < x < b$ *que* $c < y(x) < d$. *Si* M *es una cota superior de* $|y''|$ *y* L *una cota superior de* $f_y(x, y(x))$ *en* $x_0 < x < b$ *entonces la sucesión de errores globales,* $E_k = |y(x_k) - y_k|$, *satisface:*

$$E_{k+1} \leq (1 + hL)E_k + \tfrac{1}{2}Mh^2$$

Para el método de Euler el error global tiene una cota superior dada por el siguiente teorema:

Teorema 28 *Si* $f(x, y)$ *y* $f_y = \partial f/\partial y$ *son funciones continuas en el rectángulo* $x_0 < x < b$, $c < y < d$ *y la solución de* (5.1) *satisface para todo*

$x_0 < x < b$ que $c < y(x) < d$ entonces el error global en la sucesión y_n de (5.3) *tiene la siguiente cota superior en valor absoluto,*

$$E_n = |y(x_n) - y_n| < \frac{M}{2L}h\big((1+hL)^n - 1\big) \approx \frac{M}{2L}h(e^{(x_n - x_0)L} - 1)$$

donde M es una cota superior de $|y''|$ y L una cota superior de $f_y(x, y(x))$ en $x_0 < x < x_n$.

5.1.2 Métodos de Taylor

Los problemas asociados con el algoritmo de Euler sugieren llevar más lejos la aproximación mediante serie de Taylor

$$y(x+h) = y(x) + hy'(x) + \frac{1}{2}h^2y''(x) + \frac{1}{3!}h^3y'''(x) + \cdots \qquad (5.4)$$

Si esta aproximación tiene una aplicación práctica a la resolución de (5.1) esto es debido a que (suponiendo que f es suficientemente diferenciable) *todas las derivadas de $y(x)$ pueden calcularse en términos de f y sus derivadas parciales.* Por ejemplo, diferenciando $y'(x) = f(x, y(x))$ se obtiene $y''(x) = f_x + ff_y$. Así obtenemos una nueva función de dos variables

$$f^{(1)}(x, y) = f_x(x, y) + f(x, y)\, f_y(x, y)$$

que, al igual que ocurre con $f(x, y)$ respecto de $y'(x)$, para cada x, $y''(x) = f^{(1)}(x, y(x))$ Análogamente,

Ejercicio 5.2
Demostrar (suponiendo que f pueda diferenciarse tanto como sea necesario) que si $y'(x) = f(x, y(x))$ y definimos

$$f^{(2)} = f_{xx} + 2ff_{xy} + f^2 f_{yy} + f_x f_y + f(f_y)^2$$

entonces

$$\begin{aligned} y'''(x) &= f^{(2)}(x, y(x)) \\ &= f_{xx}(x, y(x)) + 2f(x, y(x))\, f_{xy}(x, y(x)) + f^2(x, y(x))\, f_{yy}(x, y(x)) \\ &\quad + f_x(x, y(x))\, f_y(x, y(x)) + f(x, y(x))(f_y(x, y(x)))^2. \end{aligned}$$

Si definimos el "polinomio de Taylor de orden p" asociado a f mediante

$$T_p(x, y, h) = f(x, y) + \frac{h}{2!}f^{(1)}(x, y) + \cdots + \frac{h^{p-1}}{p!}f^{(p-1)}(x, y);$$

entonces (5.4) se reduce a

$$y(x_1) = y_0 + hT_p(x_0, y_0, h) + \frac{h^{p+1}}{(p+1)!}y^{(p+1)}(\xi)$$

y en general, para $x_k = x_0 + kh$, definimos el algoritmo de Taylor de orden p mediante

$$y(x_{k+1}) = y_k + hT_p(x_k, y_k, h)\,, \qquad \text{con} \qquad y_0 = y(x_0).$$

Ejercicio 5.3
Demostrar que el algoritmo de Taylor de orden 1 es precisamente el algoritmo de Euler.

En la práctica, por su complicación, no se usan los algoritmos de Taylor de orden superior al 2. Pero es útil conocer en detalle el caso de orden dos. En este caso tenemos

$$\begin{aligned} y(x_{k+1}) &= y_k + hT_2(x_k, y_k, h) \\ &= y_k + h\Big(f(x_k, y_k) + \frac{h}{2!}\big(f_x(x_k, y_k) + f(x_k, y_k)\, f_y(x_k, y_k)\big)\Big) \end{aligned}$$

lo cual da lugar al siguiente algoritmo:

Algoritmo de Taylor de orden 2

Datos: x_0, b (extremos del intervalo)**,** h (longitud del paso)**,** y_0 (valor inicial)**.** Funciones predefinidas: $f(x, y)$**,** $f_x(x, y)$**,** $f_y(x, y)$**.**

1 $f_0 = f(x_0, y_0)$**;** $f_1 = f_x(x_0, y_0) + f_0 f_y(x_0, y_0)$**;**
$T_2 + f_0 + 0.5\, h f_1$**;** $y_0 = y_0 + hT_2$**;** $x_0 = x_0 + h$**;**
Imprimir x_0**,** y_0**;**
Si $x_0 < b$ **entonces ir al paso 1;**
Stop.

La principal ventaja de los métodos de Taylor es su potencial para obtener una precisión muy alta en los resultados sin necesidad de utilizar un número demasiado grande de pasos (es decir, sin necesidad de elegir un paso de longitud excesivamente pequeña). Pero, por otro lado, para alcanzar esa precisión es necesario disponer de las derivadas parciales de f, el cálculo de las cuales pueden requerir un gran número de operaciones. Éste es uno de los inconvenientes de estos métodos. Otro inconveniente más serio es la necesidad de llevar a cabo el trabajo preliminar de análisis (cálculo de las derivadas) y programación (para la evaluación de dichas derivadas). Este inconveniente puede resolverse hoy día (al menos para cierta clase de funciones) mediante el uso de programas de cálculo simbólico que automatizan el cálculo de derivadas de expresiones complejas. Algunos de estos programas se encuentran disponibles en calculadoras simbólicas modernas tales como la HP48 de *Hewlett-Packard*.

A pesar de lo dicho, los métodos de Taylor rara vez se usan en la práctica. Su principal interés es de tipo teórico porque la mayoría de los métodos prácticos intentan alcanzar la misma precisión que un método de Taylor del mismo orden sin la desventaja de tener que calcular las derivadas superiores. Esto nos lleva a los métodos de Runge-Kutta.

5.1.3 Métodos de Runge-Kutta

Según lo dicho más arriba, la motivación principal de los métodos de Runge-Kutta es el intento de imitar la precisión de los métodos de Taylor evitando realizar las derivadas de orden superior de la función f. La forma más sencilla e intuitiva de introducir los métodos de Runge-Kutta es comenzar con el caso particular conocido como *método de Heun* ((28), p. 515, (9), p. 541) o *método de Euler mejorado* ((38), p. 421), el cual puede plantearse como un método de aceleración del método de Euler.

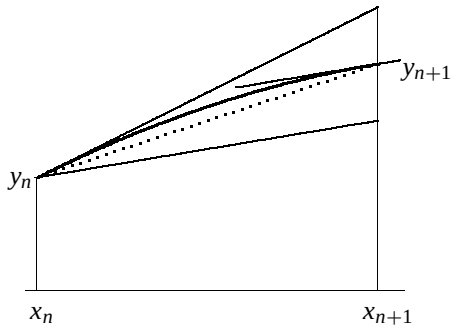
Consideremos la siguiente idea: el valor de $f(x_n, y_n)$ aproxima la pendiente de $y(x)$ en x_n y esto nos permite estimar y_{n+1} como $y_n + hf(x_n, y_n)$. Pero al disponer de este valor estimado podemos usarlo a su vez para estimar la pendiente de $y(x)$ en x_{n+1} como

$$y'(x_{n+1}) \approx f(x_{n+1}, y_{n+1})$$

y a partir de esta estimación realizar una corrección en y_{n+1}, recalculándolo en base a los valores estimados de las pendientes en x_n y en x_{n+1} y aproximando $y(x)$ en $[x_n, x_n + h]$ por una parábola. De esta forma se obtiene el

algoritmo del método de Heun,

$$y_{n+1} = y_n + \frac{h}{2}\Big(f(x_n, y_n) + f\big(x_n + h, y_n + hf(x_n, y_n)\big)\Big).$$



Es interesante también observar que el valor final de la pendiente en x_{n+1} es la media aritmética de las pendientes $f(x_n, y_n)$ y $f\big(x_n + h, y_n + hf(x_n, y_n)\big)$ inicialmente estimadas en x_n y x_{n+1}. Esto no debe sorprender porque la pendiente de la parábola varía linealmente.

La idea general de los métodos de Runge-Kutta de orden 2, es utilizar una fórmula de iteración de la forma:

$$y_{n+1} = y_n + h\Big(a_1 f(x_n, y_n) + a_2 f(x_n + b_1 h, y_n + b_2 hf(x_n, y_n))\Big)$$

donde las constantes a_1, a_2, b_1, b_2 han de determinarse de tal forma que

$$y_{n+1} = y_n + hT_2(x_n, y_n, h) + O(3)$$

Ejercicio 5.4
Demostrar que la restricción impuesta a las constantes a_1, a_2, b_1, b_2 implica que éstas deben verificar $a_1 + a_2 = 1$, $a_2 b_1 = \frac{1}{2}$, $a_2 b_2 = \frac{1}{2}$ con $a_2 > 0$ y que por tanto un método de Runge-Kutta de orden 2 está determinado por la constante $a = a_2$ y tiene la forma

$$y_{n+1} = y_n + (1-a)hf(x_n, y_n) + ahf(x_n + \tfrac{h}{2a}, y_n + \tfrac{h}{2a} f(x_n, y_n))$$

De esta forma general es inmediato ver que

Ejercicio 5.5
El método de Heun es el caso particular de los métodos de Runge-Kutta de orden 2 que se obtiene al tomar $a = \frac{1}{2}$ *en la fórmula que aparece en el ejercicio* (5.4).

Por otro lado, resulta interesante que el método original desarrollado por RUNGE hace más de un siglo (1895) era el caso particular para $a = 1$. Este caso también se conoce con el nombre de *método de Euler modificado* y se reduce a la ecuación de iteración siguiente:

$$y_{n+1} = y_n + hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}hf(x_n, y_n)).$$

En la práctica de programación de los algoritmos de los métodos de Runge-Kutta es común y útil ordenar y efectuar los cálculos mediante las cantidades

$$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf(x_n + \tfrac{1}{2a}h, y_n + \tfrac{1}{2a}k_1) \\ y_{n+1} &= y_n + (1-a)k_1 + ak_2 \end{aligned}$$

Estos cálculos se pueden implementar con el siguiente algoritmo:

Algoritmo de Runge-Kutta de orden 2

Datos: x_0, b (extremos del intervalo)**,** h (longitud del paso)**,** y_0 (valor inicial)**,** c (constante del método, Runge-Kutta original es $c = \frac{1}{2}a = \frac{1}{2}$)**.** Función predefinida: $f(x, y)$**.**

1 $k_1 = hf(x_n, y_n)$**;**
$k_2 = hf(x_n + ch, y_n + ck_1) - k_1$**;**
$y_0 = y_0 + k_1 + 2ck_2$**;** $x_0 = x_0 + h$**;**
Imprimir x_0**,** y_0**;**
Si $x_0 < b$ **entonces ir al paso 1;**
Stop.

A continuación explicamos la forma de obtener métodos de Runge-Kutta de órdenes superiores por un método análogo al utilizado para obtener la fórmula general de los métodos de orden 2. Sin entrar en los detalles

de cálculo, el método usual de orden 4 (utilizado por KUTTA en 1905), es el siguiente,

$$y_{n+1} = y_n + \tfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

donde

$$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf\left(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_1\right) \\ k_3 &= hf\left(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_2\right) \\ k_4 &= hf(x_n + h, y_n + k_3). \end{aligned}$$

Este método requiere cuatro evaluaciones de la derivada $y'(x)$ en cada paso de iteración, frente a las dos evaluaciones necesarias con los métodos de orden 2. Puesto que el orden de exactitud de este método es mayor que el de aquél, lo dicho significa que el método será superior si podemos usar un paso de longitud *al menos doble* que el necesario en el método de orden 2 para obtener la misma precisión. Este es importante para evitar caer en la creencia ciega en la superioridad del método de Runge-Kutta de cuarto orden frente al de segundo orden. Ciertamente el método de cuarto orden es superior en la mayoría de los casos, pero no es una propiedad del método sino de los problemas a los que se suele aplicar. Igual comentario puede hacerse sobre la superioridad del método de cuarto orden frente a los de orden superior. ((39), p.604.)

5.1.4 Error local y paso variable

Una consideración que debe quedar clara respecto al uso de los métodos de Runge-Kutta es la referente al tamaño del paso empleado en cada iteración, es decir, el hecho de que no hay razón por la cual se deba mantener fija la longitud h del paso. Hecha esta consideración se abre la posibilidad de ajustar el paso de cada iteración al error máximo que se quiere tolerar. Esto nos llevará a la cuestión de estimar el error local (ver CONTE y DE BOOR (14), p. 366).

Nos limitaremos a uno de los métodos conocidos de estimación del error local, a saber, el que se basa en dividir los intervalos en dos partes iguales. Explicaremos la idea de este método suponiendo que utilizamos un método de Runge-Kutta de orden p e indicando que el caso más corriente es $p = 4$. La idea es muy sencilla: integrar una vez de x_n a $x_{n+1} = x_n + h$

y después hacer dos integraciones de paso $h/2$ a partir de x_n. De esta forma se obtienen dos estimaciones $y_h(x_{n+1})$ e $y_{h/2}(x_{n+1})$ mediante cuya comparación podemos estimar el error local como

$$D_n = \frac{|y_{h/2}(x_{n+1}) - y_h(x_{n+1})|}{2^p - 1}.$$

Ahora podemos describir la forma de controlar el paso. Para ello es necesario fijar una tolerancia de error local, ϵ, con la cual limitaremos el error local *por unidad de paso*, es decir: $D_n \leq h\epsilon$. A partir de la tolerancia de error local calcularemos la cota inferior del error local $\epsilon' = \epsilon/2^{p+1}$, por debajo de la cual consideraremos que la precisión es excesiva. Con esto establecemos el siguiente criterio de actuación:

(a) Si $\epsilon' < \dfrac{D_n}{h} < \epsilon$ nos quedamos con el valor obtenido de $y_{h/2}(x_{n+1})$ y continuamos integrando a partir de x_{n+1} con el mismo paso h.

(b) Si $\epsilon < \dfrac{D_n}{h}$ reducimos h a la mitad y repetimos los cálculos desde x_n.

(c) Si $\dfrac{D_n}{h} < \epsilon'$ nos quedamos con $y_{h/2}(x_{n+1})$ pero cambiamos h al doble antes de seguir integrando a partir de x_{n+1}.

5.1.5 Métodos de paso múltiple

La idea de los métodos de paso múltiple es muy simple y natural. Para introducir un caso sencillo que ilustre la idea general utilizaremos la fórmula siguiente

Ejercicio 5.6
Si y es una función con derivada tercera continua, entonces para algún $\xi \in (a-h, a+h)$

$$y'(a) = \frac{y(a+h) - y(a-h)}{2h} + \tfrac{1}{6}h^2 y'''(\xi)$$

a partir de la cual es sencillo deducir el método

$$y_{n+1} = y_{n-1} + 2h f(x_n, y_n).$$

El error local de truncación en este método es de la forma $\tau_{n+1}(h) = \frac{1}{3}y'''(\xi_n)h^2$ y por tanto es un método de segundo orden. Compararemos este método con los métodos de Runge-Kutta de orden 2, destacando la ventaja que aquí tenemos por la gran simplicidad de los cálculos.

A continuación introduciremos la idea de utilizar la información dada por los puntos ya calculados, $(x_0, y_0), \ldots, (x_n, y_n)$, para aproximar $f(x, y(x))$ mediante interpolación en algunos de dichos puntos (los m últimos) y usar esta aproximación en la versión integral,

$$y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} f(x, y(x))\, dx$$

de nuestra ecuación diferencial. Así, llegamos a las fórmulas de Adams-Bashforth. Éstas se pueden escribir bien en términos de diferencias o en términos de ordenadas, de las cuales preferiremos la segunda forma. Así, en el caso $m = 3$ obtendremos la fórmula (usando la abreviación $f_n = f(x_n, y_n)$)

$$y_{n+1} = y_n + \tfrac{1}{24}h\big(55\, f_n - 59\, f_{n-1} + 37\, f_{n-2} - 9\, f_{n-3}\big),$$

cuya fórmula del error local se deduce integrando la fórmula del error de interpolación y da

$$E = h^4 y^{(5)}(\xi)\tfrac{251}{720}.$$

Otras fórmulas de paso múltiple se deducen de la misma idea aplicada a la integración de $f(x, y(x))$ en intervalos distintos del $[x_n, x_{n+1}]$, por ejemplo, integrando desde x_{n-1} hasta x_{n+1}. En general, estas fórmulas se obtienen integrando desde x_{n-p} hasta x_{n+1} para algún valor de $p \in \{0, \ldots, n\}$. Así se obtienen, por ejemplo (casos $m = 1$, $p = 1$ y $m = 3$, $p = 3$) las fórmulas

$$y_{n+1} = y_{n-1} + 2h f(x_n, y_n), \qquad E = \tfrac{1}{3}y'''(\xi)h^2$$
$$y_{n+1} = y_{n-3} + \tfrac{4}{3}h(f_n - f_{n-1} + 2 f_{n-2}), \qquad E = \tfrac{14}{45}y^{(5)}(\xi)h^4$$

entre las cuales reconocemos las del método presentado al principio de este apartado.

Por último mencionaremos una dificultad de que adolecen las fórmulas de paso múltiple: requieren conocer varios valores iniciales de la función $f\big(x, y(x)\big)$ para poder comenzar los cálculos. Estos valores iniciales deben ser provistos por otro método, y es necesario asegurarse de que son suficientemente precisos para la precisión total requerida.

5.1.6 Sistemas de ecuaciones

En este apartado indicaremos brevemente la forma de adaptar los métodos que hemos estudiado al caso de sistemas de ecuaciones. Para ello estudiaremos en detalle la reformulación de los métodos de Euler y de Heun para los sistemas que provienen de problemas de valores iniciales para una ecuación de segundo orden, es decir, para problemas de la forma

$$\begin{aligned} y_1' &= y_2\,, & y_1(a) &= A \\ y_2' &= f(x, y_1(x), y_2(x))\,, & y_2(a) &= B \end{aligned}$$

que serán utilizados en la siguiente sección.

5.2 Problemas de valores en la frontera

5.2.1 Métodos de tiro o disparo

Comenzamos esta sección recordando la forma general de los problemas unidimensionales de valores en la frontera, y cómo se reducen a un sistema de ecuaciones diferenciales de primer orden. Explicaremos las dificultades inherentes a los problemas de valores en la frontera dando una idea general del procedimiento a seguir en los métodos de disparo. En especial indicaremos la necesidad de utilizar una versión multidimensional de algún método de resolución numérica de ecuaciones no-lineales tal como (por ejemplo) el método de la secante o el de Newton-Raphson.

A continuación indicaremos el tipo particular de problema con que ilustraremos los métodos de resolución de disparo, asegurándonos de que el alumno comprenda la posibilidad de modificar los métodos explicados para aplicarlos a problemas más generales.

El problema particular que estudiaremos es el problema de segundo orden de dos puntos frontera, es decir el problema de evaluar una función $y(x)$ que satisface

$$y''(x) = f(x, y(x), y'(x))\,, \qquad y(a) = A\,, y(b) = B$$

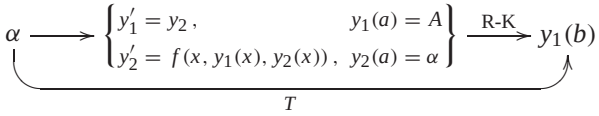
que reformularemos como el sistema

$$\begin{aligned} y_1' &= y_2\,, & y_1(a) &= A \\ y_2' &= f(x, y_1(x), y_2(x))\,, & y_1(b) &= B. \end{aligned}$$

Para explicar conceptualmente el proceso de resolución del método del disparo plantearemos considerar una función $T : \mathbf{R} \to \mathbf{R}$ que a cada número real α (quizás restringido a cierto dominio) le asocia el valor en b de la función y_1 solución del sistema

$$\begin{aligned} y_1' &= y_2\,, & y_1(a) &= A \\ y_2' &= f(x, y_1(x), y_2(x))\,, & y_2(a) &= \alpha, \end{aligned}$$

es decir, T es la función que se puede representar de la forma



Nuestro problema es equivalente al de hallar el número α tal que

$$T(\alpha) = B.$$

Este problema lo resolvemos por alguno de los métodos de resolución numérica de ecuaciones no lineales, por ejemplo por el método de la secante. Con ese método tenemos que realizar las iteraciones

$$\alpha_{n+1} = \alpha_n - \frac{\big(T(\alpha_n) - B\big)(\alpha_n - \alpha_{n-1})}{T(\alpha_n) - T(\alpha_{n-1})}$$

en cada una de las cuales hay que realizar una nueva evaluación de T, lo que significa una aplicación de un método numérico para problemas de valores iniciales tal como, por ejemplo, el usual de Runge-Kutta de cuarto orden.

Para terminar este apartado sobre métodos de disparo comentaremos brevemente la variante llamada del *disparo a un punto intermedio*, que resulta necesaria cuando el problema es de tal naturaleza o la estimación inicial α_0 está tan apartada de su valor correcto que los "disparos" iniciales no son capaces de atravesar todo el dominio de integración. En estos casos, y en aquellos en que los extremos del intervalo de integración son puntos singulares de la ecuación diferencial de partida, suele ser ventajoso, en lugar de integrar desde un extremo del intervalo $[a, b]$ hasta el otro, elegir un punto intermedio x_0 y realizar dos integraciones, una desde a hasta x_0 y otro desde b hasta x_0. En lugar de la ecuación no lineal a resolver que teníamos antes ($T(\alpha) = B$), ahora tendremos (en el caso unidimensional

de segundo orden que hemos tomado como ejemplo) un sistema de dos ecuaciones no lineales de la forma

$$T_a(\alpha) = T_b(\beta)$$
$$S_a(\alpha) = S_b(\beta)$$

de las que la primera expresa la igualdad de los dos valores obtenidos para $y(x)$ en x_0 y la segunda expresa la igualdad de los dos valores obtenidos para la derivada $y'(x)$ en x_0.

5.2.2 Problemas lineales: Método de las diferencias finitas

En este apartado explicaremos el método de las diferencias finitas para problemas lineales de valores en la frontera. De nuevo, nos restringiremos a problemas unidimensionales de segundo orden. Este caso lo usaremos para ilustrar el procedimiento general: elegir una *malla* de puntos ($x_0 = a$, con $x_n = x_0 + nh$, $h = (b-a)/N$) en el intervalo del problema y sustituir la ecuación diferencial por su aproximación mediante una ecuación en diferencias finitas resultante de sustituir las derivadas que aparecen en la ecuación diferencial por sus aproximaciones en términos de *diferencias centrales*, es decir, realizando las aproximaciones

$$y'(x) \approx \frac{y(x+h) - y(x-h)}{2h}$$
$$y''(x) \approx \frac{y(x+h) - 2y(x) + y(x-h)}{h^2}$$
$$y'''(x) \approx \frac{y(x+2h) - 3y(x+h) + 3y(x-h) - y(x-2h)}{h^3}$$
$$\vdots$$

De esta forma pasamos del problema

$$y''(x) + f(x)y'(x) + g(x)y(x) = q(x)\,, \qquad y(a) = \alpha\,,\ y(b) = \beta$$

a su aproximación en diferencias finitas, es decir, al sistema de $N-1$ ecuaciones lineales en $N-1$ incógnitas:

$$\big(1 - \tfrac{1}{2}hf_n\big)y_{n-1} + \big(-2 + h^2 g_n\big)y_n + \big(1 + \tfrac{1}{2}hf_n\big)y_{n+1} = h^2 q_n$$

donde $n = 1, \ldots, N-1$.

Observaremos que la matriz de coeficientes de este sistema siempre resulta ser tridiagonal, con lo cual el sistema es susceptible de ser resuelto por los métodos especiales descritos en el tema de métodos numéricos para el álgebra lineal.

5.2.3 Métodos de colocación

Haremos una breve mención a los métodos de *colocación*, cuya idea general se aplica a cualquier ecuación funcional lineal

$$Ly = f$$

y por tanto a las ecuaciones diferenciales lineales.

La idea es muy sencilla. Se comienza eligiendo una malla de puntos que denotamos $a = x_0, x_1, \ldots, x_N, x_{N+1} = b$ y, además, se eligen N *funciones básicas* $\varphi_1, \ldots, \varphi_N$ que supuestamente sirven para aproximar la solución mediante una combinación lineal de la forma

$$y(x) \approx c_1\varphi_1(x) + \cdots + \varphi_N(x).$$

Debido a la linealidad del operador L, la condición de que dicha aproximación sea *exacta* en los N puntos *interiores* de la malla (puntos de *colocación*) nos lleva a un sistema de N ecuaciones lineales en las N incógnitas $c_1, \ldots, c_N$. Además, si las condiciones de contorno son homogéneas, bastará elegir las funciones φ_i de forma que satisfagan dichas condiciones de contorno para que cualquier combinación lineal de ellas también las satisfaga. Esto hace que este método sea muy apropiado para los problemas de Sturm-Liouville con condiciones de contorno homogéneas.

Cuando las condiciones de contorno no son homogéneas tenemos que imponerlas sobre la combinación lineal $c_1\varphi_1(x) + \cdots + \varphi_N(x)$ y reducir correspondientemente el número de puntos de colocación para seguir teniendo un sistema determinado.

No insistiremos más en este método salvo para indicar que un conjunto de funciones básicas de gran utilidad nos las ofrecen las varillas flexibles (*splines*) adaptadas a las condiciones de contorno de nuestro problema. En particular son de interés las varillas flexibles cúbicas que hemos estudiado en un tema anterior.

Muy probablemente los problemas de contorno han aparecido ya en varias asignaturas de los tres primeros años de ingeniería. Los métodos para resolver estos problemas son en general muy distintos de los métodos para resolver problemas de valores iniciales, sin embargo estudiaremos aquí un método sencillo (llamado *método del disparo*) que reduce el problema de contorno a una serie de problemas de valores iniciales. Consiste este método en la búsqueda iterativa de condiciones iniciales tales que la solución determinada por ellas satisfaga las condiciones de contorno dadas.

5.2.4 Tratamiento Variacional

El método variacional o de la energía reduce la resolución de problemas elípticos de contorno a la de un problema variacional. Una clase de problemas unidimensionales que pueden tratarse por este método es la de aquellos que sean equivalentes a las ecuaciones del movimiento de un sistema mecánico en que las fuerzas que actúan provengan de una función de energía potencial V. Entonces se puede demostrar que existe una *función lagrangiana* L (la diferencia entre la energía cinética y la energía potencial) tal que el movimiento del sistema, para cualesquiera condiciones iniciales, minimiza la *acción* $\int L\,dt$, es decir, es la solución del problema variacional $\delta \int L\,dt = 0$. El teorema fundamental del cálculo de variaciones establece que una función $x(t)$ da un valor extremo a una integral de la forma

$$\int_{t_1}^{t_2} L(x, \dot{x}, t)\,dt,$$

si y sólo si es solución de la ecuación diferencial de Euler asociada al problema variacional:

$$\frac{\partial L}{\partial x} - \frac{d}{dt}\frac{\partial L}{\partial \dot{x}} = 0.$$

Si x es un vector con n dimensiones se obtendrán n ecuaciones de Euler, pero los razonamientos son enteramente iguales al caso unidimensional en que concentraremos nuestra atención. Una clase bastante general de problemas a los que se pueden aplicar estos métodos son los que se expresan mediante una ecuación diferencial de la forma

$$\ddot{x} + f(t)x = g(t) \tag{5.5}$$

la cual puede obtenerse como ecuación de Euler de un problema variacional tomando como lagrangiana (entre otras posibles) la función $L(x, \dot{x}, t) = \dot{x}^2 - f(t)x^2 + 2g(t)x$. Una función lagrangiana que da lugar a problemas no lineales para $a \neq 0$ es $L(x, \dot{x}, t) = (ax + b)\dot{x}^2 - f(t)x^2 + 2g(t)x$.

Un método eficaz para la resolución numérica de problemas variacionales es el *método de los elementos finitos* que en realidad se aplica a casos de varias dimensiones (siendo t en la ecuación (5.5) un vector) pero cuya aplicación ilustraremos solamente en el caso unidimensional descrito.

5.2.5 Método de los elementos finitos

Para hallar numéricamente una función $x(t)$ definida en el intervalo (t_1, t_2) que minimize la integral

$$I = \int_{t_1}^{t_2} \left(\dot{x}^2 - f(t)x^2 + 2g(t)x\right) dt$$

se procederá como sigue: Comenzamos dividiendo el intervalo de integración en subintervalos (o *elementos*, de donde el nombre del método). En cada elemento se aproxima la solución mediante una función polinómica de grado pre-especificado y los coeficientes de todos esos polinomios se someten a las condiciones de continuidad, diferenciabilidad y contorno apropiadas y se evalua en términos de ellos la integral I. Ésta resultará ser necesariamente una función lineal de los coeficientes y cuadrática en los valores que x toma en los nodos de los elementos. En consecuencia la condición de extremo de la integral conduce a un sistema de ecuaciones lineales con matriz de coeficientes simétrica y definida positiva. La solución de ese sistema será la solución aproximada de nuestra ecuación diferencial.

Respuestas a Algunos Ejercicios del Capítulo 5

Ejercicio 5.1 Ver Plybon, pp. 409–410.

Ejercicio 5.2

Ejercicio 5.3

Ejercicio 5.4 Ver Plybon, p. 420.

Ejercicio 5.5

Ejercicio 5.6 Si aplicamos la fórmula de Taylor de orden 2 con resto a $y(a+h)$ y a $y(a-h)$, y restamos, nos queda

$$y(a+h) - y(a-h) = 2hy'(a) + \tfrac{1}{3}h^3 \frac{y'''(\eta_1) + y'''(\eta_2)}{2}.$$

Pero la semisuma que aparece aquí es claramente (suponiendo y''' continua) igual a $y'''(\xi)$ para algún $\xi \in (a-h, a+h)$.

Bibliografía

[1] Abramowitz and Stegun. *Handbook of Mathematical Functions, Graphs, and Mathematical Tables*. Dover, 1965.

[2] Ireneo Peral Alonso. *Primer Curso de Ecuaciones en Derivadas Parciales*. Addison-Wesley Iberoamericana, S.A., 1995.

[3] Larry C. Andrews. *Special Functions of Mathematics for Engineers*. McGraw-Hill, second edition, 1992.

[4] Kendall E. Atkinson. *An introduction to Numerical Analysis*. John Wiley & Sons, 1978.

[5] N. Bakhvalov. *Métodos numéricos*. Paraninfo, 1980.

[6] Richard Barret and Others. *Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods*. SIAM, 1994.

[7] Ronald N. Bracewell. *The Fourier Transform and its Applications*. McGraw-Hill, 2 edition, 1986.

[8] Brice Carnahan, H. A. Luther, and James O. Wilkes. *Cálculo Numérico Métodos, Aplicaciones*. Editorial Rueda, Madrid, 1979.

[9] Steven C. Chapra and Raymond P. Canale. *Métodos Numéricos para Ingenieros*. McGraw-Hill, 1985.

[10] Ciarlet. *Introduction a l'analyse numérique matricielle et à l'optimisation*. Masson, 1982.

[11] P. G. Ciarlet and J. L. Lions, editors. *Handbook of Numerical Analysis*, volume I, II, III, IV. North-Holland, 1994.

[12] A. M. Cohen, J. F. Cutts, R. Fielder, D. E. Jones, J. Ribbans, and E. Stuart. *Análisis Numérico*. Editorial Reverté, 1977.

[13] S. D. Conte and C. De Boor. *Análisis Numérico Elemental*. Libros McGraw-Hill de Mexico, segunda edition, 1974.

[14] Samuel Daniel Conte and Carl De Boor. *Elementary Numerical Analysis*. McGraw-Hill, third edition, 1981.

[15] Crouzeix and Mignot. *Analyse numérique des equations différentielles*. Masson, 1983.

[16] Dahlquist and Bjorck. *Numerical Methods*. Prentice-Hall, 1974.

[17] Graham De Vahl Davis. *Numerical Methods in Engineering and Science*. Chapman and Hall, second edition, 1986.

[18] Deuflhard and Hohmann. *Numerical Analysis*. Walter de Gruyter, 1995.

[19] Forsythe, Malcom, and Moler. *Computer Methods for Mathematical Computations*. Prentice-Hall, 1977.

[20] Gene Golub and James M. Ortega. *Scientific Computing. An introduction to Parallel Computing*. Academic Press, Inc., 1993.

[21] Hämmerlin and Hoffmann. *Numerical Mathematics*. Springer UTM, 1991.

[22] Peter Henrici. *Elements of Numerical Analysis*. John Wiley & Sons, 5 edition, 1964.

[23] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*. Siam, 1996.

[24] Joe D. Hoffman. *Numerical Methods for Engineers and Scientists*. McGraw-Hill, 1992.

[25] Isaacson and Keller. *Analysis of Numerical Methods*. Wiley, 1966.

[26] Eugene Jahnke and Fritz Emde. *Tables of Functions with formulae and curves*. Dover, 5 edition, 1945.

[27] Jain, Iyengar, and Jain. *Numerical Methods for Scientific and Engineering Computation*. Wiley Eastern, 1993.

[28] David Kinkaid and Ward Cheney. *Análisis Numérico*. Addison-Wesley Iberoamericana, 1994.

[29] P. Lascaux and R. Theodor. *Analyse numérique matricielle appliquée a l'art de l'ingénieur*, volume 1. Masson, 1986.

[30] Charles Van Loan. *Computational Frameworks for the Fast Fourier Transform*. SIAM, 1992.

[31] Luenberger. *Optimization by vector space methods*. Wiley, 1969.

[32] F. Marcellán, L. Casasús, and A. Zarzo. *Ecuaciones Diferenciales*. McGraw-Hill, Madrid, 1990.

[33] John H. Mathews. *Métodos Numéricos para Matemáticas, Ciencia e Ingeniería*. Dover, 2 edition, 1945.

[34] John H. Mathews. *Numerical Methods for Mathematics, Science, and Engineering*. Prentice-Hall, second edition, 1992.

[35] A. V. Oppenheim and A. S. Willski. *Señales y Sistemas*. Prentice-Hall Hispanoamericana, 1994. Traducido de la segunda edición inglesa.

[36] Alan V. Oppenheim and Alan S. Willski. *Signals and Systems*. Prentice-Hall, second edition, 1983.

[37] James M. Ortega. *Numerical Analysis: a second course*. SIAM, 1990.

[38] Benjamin F. Plybon. *An Introduction to Applied Numerical Analysis*. PWS-KENT, 1992.

[39] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipies in Fortran*. Cambridge University Press, second edition, 1994.

[40] Robert W. Ramirez. *The FFT Fundamentals and Concepts*. Prentice-Hall, 1985.

[41] H. R. Schwarz. *Numerical Analysis, A Comprehensive Introduction*. Wiley, 1989.

[42] Murray R. Spiegel. *Mathematical Handbook of Formulas and Tables*. Schaum's Outline Series in Mathematics. McGraw-Hill, 1968.

[43] J. Stoer and R. Bulirsch. *Introduction to Numerical Analysis*. Texts in Applied Mathematics, 12. Springer TAM, 2 edition, 1993.

[44] R. Theodor. *Initiation a l'Analyse numérique*. Masson, 2eme edition, 1986.

[45] John Todd. *Basic Numerical Mathematics*, volume 2 (Numerical Algebra) of *International Series in Numerical Mathematics, 22*. Academic Press, 1977.

[46] John Todd. *Basic Numerical Mathematics*, volume 1 (Numerical Analysis) of *International Series in Numerical Mathematics, 14*. Academic Press, 1979.

[47] William T. Vetterling, Saul A. Teukolsky, William H. Press, and Brian P. Flannery. *Numerical Recipies Example Book [Fortran]*. Cambridge University Press, second edition, 1995.

[48] Norbert Wiener. *The Fourier Integral and certain of its Applications*. Cambridge University Press, 1933.

[49] Stephen Wolfram. *Mathematica: A System for Doing Mathematics by Computer*. Addison-Wesley, 3 edition, 1996.

Apéndice A

Ejercicios Complementarios

Ejercicio A.1
Para cada una de las funciones dadas abajo, comprobar que toma valores de signos opuestos en los extremos del intervalo $[0,1]$*. ¿Qué valor límite se obtiene al aplicar el algoritmo de la bisección? ¿Es ese valor un cero de* $f(x)$*?*

$$f(x) = \frac{1}{3x-1}, \qquad f(x) = \cos 10x.$$

Ejercicio A.2
Prueba que la ecuación $e^x - 4x^2 = 0$ *tiene una solución en el intervalo* $[4,5]$ *y otra en* $[0,1]$*. ¿Qué relación hay entre las soluciones de dicha ecuación y las de* $e^{x/2} - 2x = 0$*? Supón que se quiere usar* $g(x) = \frac{1}{2}e^{x/2}$ *como función de iteración para hallar las soluciones de la primera ecuación.*

(a) Demuestra que, a menos que dé la casualidad de que x_0 *sea ya el cero en* $[4,5]$*, las iteraciones de* g *no pueden converger a este cero.*

(b) Demuestra que si $x_0 \in [0,1]$*, entonces las iteraciones de* g *convergen al cero en* $[0,1]$*.*

(c) Halla una función de iteración que sirva para calcular el cero en $[4,5]$*.*

Ejercicio A.3 (Feb 2001)
Demuestra que el orden de convergencia de una sucesión es un número real bien definido*, es decir, que si $p > 0$ es un número real tal que*

$$0 < \lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^p} < \infty$$

entonces para $0 < q < p$

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^q} = 0,$$

y para $q > p$

$$\lim_{n\to\infty} \frac{|e_{n+1}|}{|e_n|^q} = \infty.$$

Ejercicio A.4
Sabemos que si $g(x)$ es una función de iteración tal que en su punto fijo g' se anula y g'' es continua entonces las iteraciones $x_{n+1} = g(x_n)$ convergen cuadráticamente. Enuncia condiciones que garanticen la convergencia cúbica de las iteraciones.

Ejercicio A.5
Esta es una pregunta sobre la regla de los signos de Descartes. Para entrar en materia, recuerda que la regla de los signos de Descartes es una regla sencilla que nos da alguna información sobre el número de ceros positivos de un polinomio de coeficientes... [ ] *(escribe la palabra adecuada), sin más que mirar los signos de sus coeficientes. Las preguntas son: (1) ¿Qué dice la regla de los signos de Descartes? (2) ¿Qué consecuencia importante tiene esta regla para los polinomios entre cuyos coeficientes sólo hay un cambio de signo? (3) ¿Cómo puede usarse la regla de los signos para averiguar algo sobre el número de raices negativas de un polinomio? (4) Fíjate en el polinomio $p(x) = 2x^4 - 7x^3 - 3x^2 + 5x - 1$ y di si el número de raices positivas que tiene es par o impar. (5) Escribe un polinomio tal que aplicándole la regla de los signos obtendremos información sobre el número de raices negativas de $p(x)$. (6) Sabiendo que todas las raices del polinomio $p(x)$ dado antes son reales, ¿cuántas de sus raices son positivas y cuántas son negativas?*

Ejercicio A.6
Aplica la idea de las multiplicaciones encajadas para encontrar una forma eficaz de evaluar las siguientes series (supuesto que se truncan para un valor $n = N$ dado).

$$e^x = \sum_{n=0}^{\infty} x^n/n!$$

$$\ln x = 2\sum_{n=0}^{\infty} \frac{1}{2n+1}\left(\frac{x-1}{x+1}\right)^{2n+1}$$

$$\operatorname{arcsen} x = x + \frac{1}{2}\frac{x^3}{3} + \cdots = \sum_{n=0}^{\infty} \frac{1\cdot 3\cdots(2n-1)}{2\cdot 4\cdots 2n}\left(\frac{x^{2n+1}}{2n+1}\right).$$

Ejercicio A.7 (Feb 2001)
Aplica la idea de las multiplicaciones encajadas para encontrar una forma eficaz de evaluar la siguiente serie (supuesto que se trunca para un valor $n = N$ dado).

$$\ln x = 2\sum_{n=0}^{\infty} \frac{1}{2n+1}\left(\frac{x-1}{x+1}\right)^{2n+1}.$$

Escribe, en tu propio pseudocódigo, un algoritmo que tome como datos N y x y evalúe $\ln x$ mediante dicha serie truncada en $n = N$, aplicando la idea que hayas descrito.

Ejercicio A.8
Demostrar que en una factorización triangular $A = LU$ de una matriz simétrica A, la matriz triangular superior U se puede expresar como $U = DL^t$ donde D es una matriz diagonal y L^t es la traspuesta de L. Justificar que D será necesariamente la diagonal de U.

Sugerencia: *Considérese la factorización triangular de U^t.*

Ejercicio A.9
Explica cómo se realiza la mejora iterativa *de la solución de un sistema de ecuaciones lineales. ¿En qué parte de los cálculos se saca ventaja de operar en modo de doble precisión?*

Ejercicio A.10

(a) ¿Cómo se calcula la condición, en la norma infinito, de una matriz 2×2 *como ésta:*

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}?$$

(b) Explica cómo encontrar cotas superior e inferior del error relativo de una solución aproximada de un sistema de ecuaciones lineales en términos de su resto y de la condición de su matriz de coeficientes.

Ejercicio A.11

(a) ¿Qué utilidad tiene conocer la condición de la matriz de coeficientes de un sistema de ecuaciones lineales?

(b) ¿Cuál es la condición, en la norma infinito, de la matriz de coeficientes del siguiente sistema de ecuaciones lineales:

$$0,0003\,x + 1,566\,y = 1,569$$

$$0,3454\,x - 2,436\,y = 1,018$$

(c) Halla el resto de la siguiente solución aproximada del sistema anterior: $x = 8$, $y = 1$. *Acota superior e inferiormente el error relativo en esta solución utilizando la norma infinito.*

Ejercicio A.12

Calcular la norma euclídea (es decir, la norma matricial correspondiente a la norma vectorial $\|\mathbf{x}\|_2 = \sqrt{x^2 + y^2}$*) de la matriz*

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

Ejercicio A.13

Sabemos que si $g(x)$ *es una función de iteración tal que en su punto fijo* g' *se anula y* g'' *es continua entonces las iteraciones* $x_{n+1} = g(x_n)$ *convergen cuadráticamente. Enuncia condiciones que garanticen la convergencia cúbica de las iteraciones.*

Ejercicio A.14

Supón que la función $f(x)$ *es un polinomio de grado* n. *Demuestra que el polinomio de interpolación de* $f(x)$ *en* $n+1$ *puntos distintos cualesquiera es la propia función* $f(x)$.

Ejercicio A.15 (Feb 2001)

Una forma Teórica de hallar un polinomio de interpolación para unos datos (x_0, y_0), ..., (x_n, y_n) *es resolver un sistema de ecuaciones lineales cuya matriz de coeficientes es de Vandermonde. Explica cómo el hecho de que conozcamos otras formas de hallar un polinomio de interpolación nos proporciona también una forma de hallar la inversa de una matriz de Vandermonde cualquiera.*

Pista: *Polinomios de Lagrange.*

Ejercicio A.16

Usa los polinomios de Lagrange para hallar la inversa de la matriz de Vandermonde

$$\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 5 & 25 & 125 \end{pmatrix}.$$

Pista: *Hacer primero el ejercicio A.15.*

Ejercicio A.17

Se desea expresar el polinomio $19 - 34x + 19x^2 - 3x^3$ *en la forma*

$$A_0 + A_1(x+1) + A_2(x+1)(x+2) + A_3(x+1)(x+2)(x+3).$$

Teniendo en cuenta que los coeficientes A_k *pueden considerarse como diferencias divididas, calcúlense sus valores evaluando el polinomio dado en los puntos* $x = 1$, $x = 2$, $x = 3$, *y* $x = 4$, *y construyendo la correspondiente tabla de diferencias divididas.*

Ejercicio A.18

Teniendo en cuenta que la diferencia dividida $f[x_0, \ldots, x_n]$ *es el coeficiente del término de grado* n *en el polinomio de interpolación de* f *en los puntos* $x_0, \ldots, x_n$, *demuestra que la forma de Lagrange del polinomio de interpolación nos proporciona la siguiente fórmula para las diferencias divididas, donde aparece la derivada del polinomio* $q(x) = (x - x_0) \cdots (x - x_n)$*:*

$$f[x_0, \ldots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{q'(x_i)}.$$

Ejercicio A.19
Usando la siguiente fórmula para las diferencias divididas:

$$f[x_0, \ldots, x_n] = \sum_{i=0}^{n} \frac{f(x_i)}{q'(x_i)},$$

(donde $q(x) = (x - x_0) \cdots (x - x_n)$) calcular el límite de $f[x_0, x_1, x_2]$ cuando $x_2 \to x_1$ mientras que x_0 y x_1 permanecen fijos.

Ejercicio A.20
Se desea construir una tabla de la función seno en el intervalo $[0, \pi/2]$ a intervalos igualmente espaciados con la que se pueda evaluar $\mathrm{sen}(x)$ *con 5 dígitos de precisión realizando a lo sumo interpolaciones lineales. ¿Cuántos valores debe contener la tabla?*

Ayuda: *La precisión deseada equivale a un error menor que* 0.5×10^{-5}.

Ejercicio A.21
Usa la fórmula del error de interpolación,

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

para dar una cota superior *del error al estimar el valor de* $\ln(\frac{3}{2})$ *mediante interpolación cúbica de* $\ln x$ *en los puntos $x_0 = 1$, $x_1 = \frac{4}{3}$, $x_2 = \frac{5}{3}$, $x_3 = 2$.*

Ejercicio A.22
Demostrar que la derivada $p'(x)$ de la parábola $p(x)$ que interpola a $f(x)$ en los puntos $x_0 < x_1 < x_2$ es la recta que toma valores $f[x_{i-1}, x_i]$ en los puntos $\frac{x_{i-1}+x_i}{2}$ para $i = 1, 2$. Generalizar este resultado al caso del polinomio de interpolación $p_n(x)$ de $f(x)$ en los puntos $x_0 < \cdots < x_n$ para describir $p_n'(x)$ como la interpolante de los datos $(\xi_i, f[x_i, x_{i+1}])$, para valores apropiados de $\xi_i \in [x_i, x_{i+1}]$.

Ejercicio A.23
Discutir las condiciones en que se puede o no se puede esperar que se pueda realizar una interpolación inversa, es decir, el obtener de una tabla de valores de una función $f(x)$ el valor $x = x_0$ en que la función f toma un determinado valor y_0.

Ejercicio A.24
Se desea construir una tabla de la función coseno en el intervalo $[0, \pi/2]$ a intervalos igualmente espaciados con la que se pueda evaluar $\cos(x)$ *con seis dígitos de precisión realizando a lo sumo interpolaciones lineales. ¿Cuántos valores debe contener la tabla?*

Ayuda: *Seis dígitos de precisión es un error menor que* 0.5×10^{-6}.

Ejercicio A.25
Usa la definición trigonométrica de los polinomios de Chebyshev $T_n(x)$ para demostrar que para todo n, éstos verifican

$$max_{-1 \le x \le 1} |T_n(x)| = 1$$

Ejercicio A.26
Describe el problema de interpolación óptima en un intervalo dado y explica la propiedad que tienen los polinomios de Chebyshev en relación con ese problema.

Ejercicio A.27
Supón que se te presenta un programa FORTRAN, "`FUNCTON F(X)`*", se te dice que es una rutina que evalua el polinomio conocido $p(x)$ (de grado $= r$) y te propones verificar esa afirmación. Al examinar el código de la rutina encuentras que, efectivamente, no contiene más operaciones que sumas, restas y multiplicaciones, y que la variable* `X` *no aparece nunca como factor más de r veces, lo cual prueba que la rutina evalua algún polinomio de grado $\le r$. ¿En cuántos puntos será necesario y suficiente evaluar la rutina para confirmar que representa a $p(x)$?*

Ejercicio A.28

(a) *¿Qué condiciones deben cumplir tres polinomios de grado ≤ 3, $p_1(x)$, $p_2(x)$, $p_3(x)$ para que constituyan la varilla flexible (spline) cúbica de extremos libres que pasa por los puntos (x_0, y_0), (x_1, y_1), (x_2, y_2), (x_3, y_3).*

(b) *Verificar si los tres polinomios*

$$p_1(x) = 1 + (12/23)\,(x-1) - (33/46)\,(x-1)^2 - (11/46)\,(x-1)^3,$$

$$p_2(x) = 1 - (15/46)\,(x-2) - (3/23)\,(x-2)^2 + (9/46)\,(x-2)^3,$$

$$p_3(x) = -(27/46)\,(x-4) + (1/46)\,(x-4)^3$$

constituyen o no la varilla flexible cúbica de extremos libres que pasa por los puntos

$$(0,0),\ (1,1),\ (2,1)\ y\ (4,0)$$

Ejercicio A.29
El método de Euler es un método sencillo de resolución numérica de problemas de valores iniciales en las ecuaciones diferenciales ordinarias de primer orden. ¿En qué idea se basa? Escribe la ecuación de iteración de dicho método aplicada al siguiente problema de valores iniciales:

$$y' = x + y\,, \qquad y(0) = 1,$$

y efectua los cálculos del primer paso de integración con una longitud de paso $h = 0.01$.

Ejercicio A.30
El método de Heun o método de Euler mejorado es un método sencillo de resolución numérica de problemas de valores iniciales en las ecuaciones diferenciales ordinarias de primer orden. ¿En qué idea se basa? Escribe la ecuación de iteración de dicho método y muestra que se trata de un método de segundo orden.

Ejercicio A.31 (Feb 2001)
La ecuación

$$y_{n+1} = y_{n-1} + 2hf(x_n, y_n)$$

constituye un sencillo método multipaso de grado 1 para la resolución numérica de problemas de valores iniciales en las ecuaciones diferenciales ordinarias de primer orden. Deduce dicha ecuación, junto con su fórmula del error, a partir del polinomio de Taylor de orden 2 con resto para $y(a+h)$ *e* $y(a-h)$.

Aplica dicho método al siguiente problema de valores iniciales:

$$y' = x + y\,, \qquad y(0) = 1,$$

de la siguiente forma: efectua los cálculos del primer paso de integración del método de Euler con una longitud de paso $h = 0.01$. *Con el* y_1 *así obtenido y el valor inicial dado,* y_0*, efectua los cálculos del primer paso de integración del método multipaso dado arriba para obtener* y_2 *con la misma longitud de paso* h.

Ejercicio A.32
La ecuación

$$y_{n+1} = y_{n-1} + 2hf(x_n, y_n)$$

constituye un sencillo método multipaso de grado 1 para la resolución numérica de problemas de valores iniciales en las ecuaciones diferenciales ordinarias de primer orden. Deduce dicha ecuación, junto con su fórmula del error, a partir del polinomio de Taylor de orden 2 con resto para $y(a+h)$ *e* $y(a-h)$.

Aplica dicho método al siguiente problema de valores iniciales:

$$y' = (x + y)^2\,, \qquad y(1) = 0,$$

de la siguiente forma: efectua los cálculos del primer paso de integración del método de Euler con una longitud de paso $h = 0.1$. *Con el* y_1 *así obtenido y el valor inicial dado,* y_0*, efectua los cálculos del primer paso de integración del método multipaso dado arriba para obtener* y_2 *con la misma longitud de paso* h.