Métodos Numéricos

Mercedes Román Ruiz

15 de enero de 2025

Índice general

Capítulo 1

Introducción

1. Espacios Vectoriales

Definición 1.1. Un espacio vectorial sobre el campo numérico K ($K = \mathbb{R}$ o $K = \mathbb{C}$) es un conjunto no vacío V, cuyos elementos se llaman vectores y en el cual se definen dos operaciones, denominadas suma y multiplicación por escalares, que cumplen las siguientes propiedades:

1. la suma es conmutativa y asociativa;
2. existe un elemento $0 \in V$ (el vector cero o vector nulo) tal que $v + 0 = v$ para cada $v \in V$;
3. $0 \cdot v = 0$, $1 \cdot v = v$, para cada $v \in V$, donde 0 y 1 son respectivamente el cero y la unidad de K;
4. para cada elemento $v \in V$ existe su opuesto, $-v$, en V tal que $v + (-v) = 0$;
5. se cumplen las siguientes propiedades distributivas:

$$\forall \alpha \in K, \forall v, w \in V, \alpha(v + w) = \alpha v + \alpha w$$

$$\forall \alpha, \beta \in K, \forall v \in V, (\alpha + \beta)v = \alpha v + \beta v$$

6. se cumple la siguiente propiedad asociativa:

$$\forall \alpha, \beta \in K, \forall v \in V, (\alpha\beta)v = \alpha(\beta v)$$

Definición 1.2. Decimos que una parte no vacía W de V es un subespacio vectorial de V si W es un espacio vectorial sobre K.

Definición 1.3. Un sistema de vectores $v_1, ..., v_n$ de un espacio vectorial V se llama linealmente independiente si la relación

$$\alpha_1 v_1 + \alpha_2 v_2 + ... + \alpha_m v_m = 0$$

con $\alpha_1, \alpha_2, ..., \alpha_m \in K$ implica que $\alpha_1 = \alpha_2 = ... = \alpha_m = 0$. En caso contrario, el sistema se llamará linealmente dependiente.

Llamamos base de V a cualquier sistema generador linealmente independiente de V. Si $u_1, ..., u_n$ es una base de V, la expresión $v = v_1 u_1 + ... + v_n u_n$ se llama la descomposición de v con respecto a la base y los escalares $v_1, ..., v_n \in K$ son los componentes de v con respecto a la base dada. Además, se cumple la siguiente propiedad.

Propiedad 1.1. Sea V un espacio vectorial que admite una base de n vectores. Entonces, todo sistema de vectores linealmente independientes de V tiene como máximo n elementos y cualquier otra base de V tiene n elementos. El número n se llama la dimensión de V y escribimos $\dim(V) = n$. Si, en cambio, para cualquier n siempre existen n vectores linealmente independientes en V, el espacio vectorial se llama de dimensión infinita.

2. Matrices

Sean m y n dos enteros positivos. Llamamos matriz de m filas y n columnas, o una matriz $m \times n$, o una matriz (m, n), con elementos en K, a un conjunto de mn escalares $a_{ij} \in K$, con $i = 1, ..., m$ y $j = 1, ..., n$, representada en el siguiente arreglo rectangular

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \tag{1.1}$$

Cuando $K = \mathbb{R}$ o $K = \mathbb{C}$, escribiremos respectivamente $A \in \mathbb{R}^{m \times n}$ o $A \in \mathbb{C}^{m \times n}$, para detallar explícitamente

los campos numéricos a los que pertenecen los elementos de A. Usaremos letras mayúsculas para denotar las matrices, mientras que las letras minúsculas correspondientes a esas letras mayúsculas denotarán las entradas de la matriz.

Abreviaremos (1.1) como $A = (a_{ij})$ con $i = 1, ..., m$ y $j = 1, ..., n$. El índice i se llama índice de fila, mientras que j es el índice de columna. El conjunto $(a_{i1}, a_{i2}, ..., a_{in})$ se llama la i-ésima fila de A; igualmente, $(a_{1j}, a_{2j}, ..., a_{mj})$ es la j-ésima columna de A.

Si $n = m$, la matriz se llama cuadrada o de orden n, y el conjunto de las entradas $(a_{11}, a_{22}, ..., a_{nn})$ se llama su diagonal principal.

Una matriz que tenga una fila o una columna se llama vector fila o vector columna, respectivamente. A menos que se especifique lo contrario, siempre asumiremos que un vector es un vector columna. En el caso $n = m = 1$, la matriz simplemente denotará un escalar de K.

Definición 2.1. Sea A una matriz $m \times n$. Sean $1 \leq i_1 < i_2 < ... < i_k \leq m$ y $1 \leq j_1 < j_2 < ... < j_l \leq n$ dos conjuntos de índices contiguos. La matriz $S(k \times l)$ de entradas $s_{pq} = a_{i_p j_q}$ con $p = 1, ..., k, q = 1, ..., l$ se llama submatriz de A. Si $k = l$ y $i_r = j_r$ para $r = 1, ..., k$, S se llama submatriz principal de A.

Definición 2.2. Una matriz $A(m \times n)$ se llama particionada en bloques o se dice que está particionada en submatrices si

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1l} \\ A_{21} & A_{22} & \cdots & A_{2l} \\ \vdots & \vdots & \ddots & \vdots \\ A_{k1} & A_{k2} & \cdots & A_{kl} \end{bmatrix}$$

donde A_{ij} son submatrices de A.

3. Operaciones con matrices

3.1. Matrices y transformaciones lineales

Definición 3.1. Una transformación lineal para $\mathbb{C}^n$ en $\mathbb{C}^m$ es una función $f : \mathbb{C}^n \to \mathbb{C}^m$ tal que $f(\alpha x + \beta y) = \alpha f(x) + \beta f(y)$, $\forall \alpha, \beta \in K$ y $\forall x, y \in \mathbb{C}^n$.

El siguiente resultado vincula matrices y transformaciones lineales.

Propiedad 3.1. Sea $f : \mathbb{C}^n \to \mathbb{C}^m$ una transformación lineal. Entonces, existe una matriz única $A_f \in \mathbb{C}^{m \times n}$ tal que

$$f(x) = A_f x \quad \forall x \in \mathbb{C}^n \tag{1.2}$$

Inversamente, si $A_f \in \mathbb{C}^{m \times n}$, entonces la función definida en (1.2) es una transformación lineal de $\mathbb{C}^n$ en $\mathbb{C}^m$.

4. Bien planteamiento y número de condición de un problema

Consideremos el siguiente problema: encontrar x tal que

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

donde d es el conjunto de datos del cual depende la solución y F es la relación funcional entre x y d. Según el tipo de problema representado en (1.3), las variables x y d pueden ser números reales, vectores o funciones. Típicamente, (1.3) se llama un problema directo si F y d están dados y x es desconocido, un problema inverso si F y x son conocidos y d es el desconocido, y un problema de identificación cuando x y d están dados mientras que la relación funcional F es desconocida.

El problema (1.3) está bien planteado si admite una solución única x que depende continuamente de los datos. Usaremos los términos bien planteado y estable de manera intercambiable y a partir de ahora solo trataremos problemas bien planteados.

Un problema que no goza de la propiedad anterior se llama mal planteado o inestable y antes de abordar su solución numérica debe ser regularizado, es decir, debe transformarse adecuadamente en un problema bien planteado. De hecho, no es adecuado pretender que el método numérico pueda curar las patologías de un problema intrínsecamente mal planteado.

Sea D el conjunto de datos admisibles, es decir, el conjunto de los valores de d en correspondencia con los cuales el problema (1.3) admite una solución única. La dependencia continua de los datos significa que pequeñas perturbaciones en los datos d de D producen "pequeñosçambios en la solución x. Precisamente, sea $d \in D$ y denotemos por δd una perturbación admisible en el sentido de que $d + \delta d \in D$ y por δx el cambio correspondiente en la solución, de manera que

$$F(x + \delta x, d + \delta d) = 0 \tag{1.4}$$

Entonces, requerimos que

$$\exists \eta_0 = \eta_0(d) > 0, \quad \exists K_0 = K_0(d) \text{ tal que si } \|\delta d\| \leq \eta_0 \text{ entonces } \|\delta x\| \leq K_0 \|\delta d\| \tag{1.5}$$

Las normas usadas para los datos y para la solución pueden no coincidir, siempre que d y x representen variables de diferentes tipos.

Observación 4.1. La propiedad de dependencia continua de los datos podría haberse expresado de la siguiente manera alternativa, que es más parecida a la forma clásica del Análisis $\forall \epsilon > 0\ \exists \delta = \delta(\epsilon)$ tal que si $\|\delta d\| \leq \delta$ entonces $\|\delta x\| \leq \epsilon$.

La forma (1.5) es, sin embargo, más adecuada para expresar en lo siguiente el concepto de estabilidad numérica, es decir, la propiedad de que pequeñas perturbaciones en los datos producen perturbaciones del mismo orden en la solución.

Con el objetivo de hacer que el análisis de estabilidad sea más cuantitativo, introducimos la siguiente definición.

Definición 4.1. Para el problema (1.3) definimos el número de condición relativo como

$$K(d) = \sup\{\frac{\|\delta x\|/\|x\|}{\|\delta d\|/\|d\|}, \delta d \neq 0, d + \delta d \in D\} \quad (1.6)$$

Siempre que $d = 0$ o $x = 0$, es necesario introducir el número de condición absoluto, dado por

$$K_{abs}(d) = \sup\{\frac{\|\delta x\|}{\|\delta d\|}, \delta d \neq 0, d + \delta d \in D\} \quad (1.7)$$

El problema (1.3) se llama mal condicionado si $K(d)$ es "grande"para cualquier dato admisible d.

La propiedad de que un problema esté bien condicionado es independiente del método numérico que se use para resolverlo. De hecho, es posible generar esquemas numéricos estables e inestables para resolver problemas bien condicionados. El concepto de estabilidad para un algoritmo o para un método numérico es análogo al utilizado para el problema (1.3) y se hará preciso en la siguiente sección.

Observación 4.2. (Problemas mal planteados) Incluso en el caso en que el número de condición no exista (formalmente, sea infinito), no necesariamente es cierto que el problema esté mal planteado. De hecho, existen problemas bien planteados para los cuales el número de condición es infinito, pero que pueden ser reformulados en problemas equivalentes con un número de condición finito.

Si el problema (1.3) admite una solución única, entonces necesariamente existe una aplicación G, que llamamos resolvente, entre los conjuntos de los datos y las soluciones, tal que

$$x = G(d), \text{ es decir } F(G(d), d) = 0 \quad (1.8)$$

Según esta definición, (1.4) da lugar a $x + \delta x = G(d + \delta d)$. Suponiendo que G sea diferenciable en d y denotando formalmente por $G'(d)$ su derivada con respecto a d (si $G : \mathbb{R}^n \to \mathbb{R}^m$, $G'(d)$ será la matriz jacobiana de G evaluada en el vector d), una expansión de Taylor de G truncada en primer orden asegura que

$$G(d + \delta d) - G(d) = G'(d)\delta d + o(\|\delta d\|) \quad \text{para } \delta d \to 0$$

donde $\|\cdot\|$ es una norma vectorial adecuada y $o(\cdot)$ es el símbolo clásico de infinitesimal que denota un término infinitesimal de orden superior con respecto a su argumento. Despreciando el infinitesimal de orden superior con respecto a $\|\delta d\|$, de (1.6) y (1.7) deducimos respectivamente que

$$K(d) \approx \|G'(d)\|\frac{\|d\|}{\|G(d)\|}, \quad K_{abs}(d) \approx \|G'(d)\| \quad (1.9)$$

donde el símbolo $\|\cdot\|$, cuando se aplica a una matriz, denota la norma de matriz inducida (**??**) asociada con la norma vectorial introducida anteriormente. Las estimaciones en (1.9) son de gran utilidad para realizar análisis cualitativos sobre la condición de un problema.

4.1. Relaciones entre Estabilidad y Convergencia

Los conceptos de estabilidad y convergencia están fuertemente conectados.

En primer lugar, si el problema (1.3) está bien planteado, una condición necesaria para que el problema numérico (**??**) sea convergente es que sea estable.

Supongamos entonces que el método es convergente, es decir, que (**??**) se cumple para un $\epsilon > 0$. Tenemos

$$\begin{aligned}\|\delta x_n\| = \|x_n(d + \delta d_n) - x_n(d)\| \leq \|x_n(d) - x(d)\| \\ + \|x(d) - x(d + \delta d_n)\| + \|x(d + \delta d_n) - x_n(d + \delta d_n)\| \\ \leq K(\delta(n_0, \epsilon), d)\|\delta d_n\| + \epsilon \quad (1.10)\end{aligned}$$

habiendo utilizado (1.5) y (**??**) dos veces. Ahora, eligiendo δd_n tal que $\|\delta d_n\| \leq \eta_0$, deducimos que $\|\delta d_n\| \leq \eta_0$, lo que nos lleva a que $\|\delta x_n\|/\|\delta d_n\|$ puede ser acotado por $K_0 = K(\delta(n_0, \epsilon), d) + 1$, siempre que $\epsilon \leq \|\delta d_n\|$, de modo que el método es estable. Así, nos interesan los métodos numéricos estables, ya que solo estos pueden ser convergentes.

La estabilidad de un método numérico se convierte en una condición suficiente para que el

problema numérico (**??**) converja si este último también es consistente con el problema (1.3). De hecho, bajo estas suposiciones, tenemos

$$\|x(d+\delta d_n) - x_n(d+\delta d_n)\| \leq \|x(d+\delta d_n) - x(d)\|$$

$$+\|x(d) - x_n(d)\| + \|x_n(d) - x_n(d+\delta d_n)\|$$

Gracias a (1.5), el primer término en el lado derecho puede ser acotado por $\|\delta d_n\|$. Un límite similar se aplica al tercer término, debido a la propiedad de estabilidad (**??**). Finalmente, respecto al término restante, si F_n es diferenciable con respecto a la variable x, una expansión en serie de Taylor da

$$F_n(x(d),d) - F_n(x_n(d),d) = \frac{\partial F_n}{\partial x}|_{(x,d)}(x(d) - x_n(d))$$

para un adecuado x .entre"$x(d)$ y $x_n(d)$. Suponiendo además que $\partial f_n/\partial x$ es invertible, obtenemos

$$x(d)-x_n(d) = \left(\frac{\partial F_n}{\partial x}\right)^{-1}_{|(x,d)} [F_n(x(d),d)-F_n(x_n(d),d)] \tag{1.11}$$

Por otro lado, al reemplazar $F_n(x_n(d),d)$ por $F_n(x(d),d)$ y pasar a las normas, encontramos

$$\|x(d)-x_n(d)\| \leq \| \left(\frac{\partial F_n}{\partial x}\right)^{-1}_{|(x,d)} \| \|F_n(x(d),d)-F(x(d),d)\|$$

Gracias a (**??**) podemos concluir que $\|x(d) - x_n(d)\| \to 0$ cuando $n \to \infty$. El resultado que acabamos de demostrar, aunque expresado en términos cualitativos, es un hito en el análisis numérico, conocido como el teorema de equivalencia (o teorema de Lax-Richtmyer): "para un método numérico consistente, la estabilidad es equivalente a la convergencia".

5. Fuentes de Error en Modelos Computacionales

Siempre que el problema numérico (**??**) sea una aproximación al problema matemático (1.3) y este último sea a su vez un modelo de un problema físico, diremos que (**??**) es un modelo computacional para el PP.

En este proceso, el error global, denotado por e, se expresa como la diferencia entre la solución realmente computada, $\hat{x}_n$, y la solución física, x_{ph}, de la cual x proporciona un modelo. El error global e del modelo matemático, dado por $x - x_{ph}$, y el error e_c del modelo computacional, $\hat{x}_n - x$, es decir, $e = e_m + e_c$.

El error e_m tendrá en cuenta el error del modelo matemático en sentido estricto y el error en los datos. De manera similar, e_c resulta ser la combinación del error de discretización numérica $e_n = x_n - x$, el error e_a introducido por el algoritmo numérico y el error de redondeo introducido por la computadora durante la solución del problema (**??**).

En general, podemos esbozar las siguientes fuentes de error:

1. Error debido al modelo, que puede ser controlado mediante una adecuada elección del modelo matemático;
2. Errores en los datos, que pueden reducirse mejorando la precisión en la medición de los propios datos;
3. Error de truncamiento, que surge al haber reemplazado en el modelo numérico límites por operaciones que involucran un número finito de pasos;
4. Errores de redondeo.

El error de los ítems 3 y 4 da lugar al error computacional. Un método numérico será convergente si este error puede hacerse arbitrariamente pequeño al aumentar el esfuerzo computacional. Por supuesto, la convergencia es el objetivo principal, aunque no único, de un método numérico, siendo los otros la precisión, la fiabilidad y la eficiencia.

La precisión significa que los errores son pequeños respecto a una tolerancia fija. Usualmente, se cuantifica por el orden de infinitesimal del error e_n con respecto al parámetro característico de discretización. Por cierto, cabe señalar que la precisión de la máquina no limita, en términos teóricos, la precisión.

La fiabilidad significa que es probable que el error global se garantice por debajo de una cierta tolerancia. Por supuesto, un modelo numérico solo puede considerarse confiable si ha sido adecuadamente probado, es decir, aplicado con éxito a varios casos de prueba.

La eficiencia significa que la complejidad computacional necesaria para controlar el error sea lo más pequeña posible.

Por algoritmo entendemos una directiva que indica, a través de operaciones elementales, todos los pasos necesarios para resolver un problema específico. Un algoritmo puede, a su vez, contener sub-algoritmos y debe tener la característica de terminar después de un número finito de operaciones elementales. Como consecuencia, el ejecutor del algoritmo debe encontrar dentro del algoritmo mismo todas las instrucciones para resolver completamente el problema planteado.

Finalmente, la complejidad de un algoritmo es una medida de su tiempo de ejecución. Calcular la complejidad de un algoritmo es, por lo tanto, parte del análisis de la eficiencia de un método numérico. Dado que se pueden emplear varios algoritmos con diferentes complejidades para resolver el mismo problema P,

es útil introducir el concepto de complejidad de un problema, que significa la complejidad del algoritmo que tiene la mínima complejidad entre los que resuelven P. La complejidad de un problema se mide típicamente mediante un parámetro directamente asociado con P.

6. Representación en Computadora de Números

Cualquier operación en una máquina está afectada por errores de redondeo o redondeo. Esto se debe al hecho de que en una computadora solo se puede representar un subconjunto finito del conjunto de números reales.

6.1. El Sistema Posicional

Sea una base $\beta \in \mathbb{N}$ fijada con $\beta \geq 2$, y sea x un número real con un número finito de dígitos x_k con $0 \leq x_k < \beta$ para $k = -m, ..., n$. La notación (convencionalmente adoptada)

$$x_\beta = (-1)^s [x_n x_{n-1} ... x_1 x_0 x_{-1} x_{-2} ... x_{-m}, x_n \neq 0] \tag{1.12}$$

se llama la representación posicional de x con respecto a la base β. El punto entre x_0 y x_{-1} se llama punto decimal si la base es 10, punto binario si la base es 2, mientras que s depende del signo de x ($s = 0$ si x es positivo, 1 si es negativo). La relación (1.12) significa en realidad

$$x_\beta = (-1)^s \left(\sum_{k=-m}^{n} x_k \beta^k \right)$$

Cualquier número real puede ser aproximado por números que tengan una representación finita. De hecho, al haber fijado la base β, se cumple la siguiente propiedad

$$\forall \epsilon > 0, \forall x_\beta \in \mathbb{R}, \exists y_\beta \in \mathbb{R} \text{ tal que } |y_\beta - x_\beta| < \epsilon$$

donde y_β tiene una representación posicional finita.

De hecho, dado el número positivo $x_\beta = x_n x_{n-1} ... x_1 x_0 x_{-1} x_{-2} ... x_{-m} ...$ con el número de dígitos, finitos o infinitos, para cualquier $r \geq 1$ se pueden construir dos números

$$x_\beta^{(l)} = \sum_{k=0}^{r-1} x_{x-k} \beta^{n-k}, x_\beta^{(u)} = x_\beta^{(l)} = \beta^{n-r+1}$$

que tienen r dígitos, tales que $x_\beta^{(l)} < x_\beta < x_\beta^{(u)}$ y $x_\beta^{(u)} - x_\beta^{(l)} = \beta^{n-r+1}$. Si se elige r de manera que $\beta^{n-r+1} < \epsilon$, entonces tomando y_β igual a $x_\beta^{(l)}$ o $x_\beta^{(u)}$ se obtiene la desigualdad deseada. Este resultado legitima la representación computacional de los números reales (y por lo tanto de un número finito de dígitos).

Aunque teóricamente hablando todas las bases son equivalentes, en la práctica computacional se emplean generalmente tres bases: la base 2 en binario, la base 10 o decimal y la base 16 o hexadecimal. En lo que sigue, supondremos que β es un número entero par.

Para simplificar las notaciones, escribiremos x en lugar de x_β, dejando la base β entendida.

Ejercicios

(1) Se considera la siguiente sucesión definida por recursión

$$x_0 = 1 \quad x_1 = \frac{1}{5} \quad x_{n+1} = \frac{36}{5} x_n - \frac{7}{5} x_{n-1}$$

Esta sucesión tiene como solución $x_n = \frac{1}{5^n}$. Utilizar Matlab para calcular $\frac{1}{5^n}$ utilizando la sucesión recusiva del principio, para $n \leq 32$. Hacer el estudio del error absoluto y relativo.

(2) Repetir el ejercicio anterior tomando $x_0 = 2$ y $x_1 = \frac{36}{5}$, teniendo en cuenta que la solución es ahora $x_n = \frac{1}{5^n} + 7^n$.

(3) Compara los resultados de los ejercicios 1 y 2 y dar una explicación formal de lo que está sucediendo.

Capítulo 2

Diferenciación Numérica

1. Diferenciación Numérica

Ya hemos visto una forma de aproximar la derivada de una función f:

$$f'(x) = \frac{f(x+h) - f(x)}{h} \tag{2.1}$$

para algún número pequeño h. Para determinar la precisión de esta aproximación, utilizamos el teorema de Taylor, asumiendo que $f \in C^2$:

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(\xi) \quad \xi \in [x, x+h] \rightarrow$$

$$f'(x) = \frac{f(x+h) - f(x)}{h} - \frac{h}{2}f''(\xi)$$

El término $\frac{h}{2}f''(\xi)$ se llama error de truncamiento, o error de discretización, y se dice que la aproximación es de primer orden ya que el error de truncamiento es $O(h)$.

Sin embargo, el redondeo también juega un papel en la evaluación del cociente de diferencias finitas (2.1). Por ejemplo, si h es tan pequeño que $x + h$ se redondea a x, entonces el cociente de diferencia calculado será 0. Más generalmente, incluso si el único error que se comete es en el redondeo de los valores $f(x+h)$ y $f(x)$, el cociente de diferencia calculado será

$$\frac{f(x+h)(1+\delta_1) - f(x)(1+\delta_2)}{h} =$$

$$= \frac{f(x+mh) - f(x)}{h} + \frac{\delta_1 f(x+h) - \delta_2 f(x)}{h}$$

Dado que cada $|\delta_i|$ es menor que la precisión de la máquina ϵ, esto implica que el error de redondeo es menor o igual a

$$\frac{\epsilon(|f(x)| + |f(x+h)|)}{h}$$

Dado que el error de truncamiento es proporcional a h y el error de redondeo es proporcional a $1/h$, la mejor precisión se alcanza cuando estas dos cantidades son aproximadamente iguales. Ignorando las constantes $|f''(\xi)/2|$ y $(|f(x)| + |f(x+h)|)$, esto significa que

$$h \approx \frac{\epsilon}{h} \rightarrow h \approx \sqrt{\epsilon}$$

y en este caso el error (error de truncamiento o error de redondeo) es aproximadamente $\sqrt{\epsilon}$. Así, con la fórmula (2.1), podemos aproximar una derivada solo a aproximadamente la raíz cuadrada de la precisión de la máquina.

Otra forma de aproximar la derivada es utilizar una fórmula de diferencia centrada:

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} \tag{2.2}$$

Podemos nuevamente determinar el error de truncamiento utilizando el teorema de Taylor. Expandiendo $f(x+h)$ y $f(x-h)$ alrededor del punto x, encontramos

$$f(x+h) = f(x) + hf'(x) + \frac{h^2}{2}f''(x) + \frac{h^3}{6}f'''(\xi)$$

$$\xi \in [x, x+h]$$

$$f(x-h) = f(x) - hf'(x) + \frac{h^2}{2}f''(x) - \frac{h^3}{6}f'''(\eta)$$

$$\eta \in [x-h, x]$$

Restando las dos ecuaciones y resolviendo para $f'(x)$ se obtiene

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{12}(f'''(\xi) + f'''(\eta))$$

Por lo tanto, el error de truncamiento es $O(h^2)$, y esta fórmula de diferencia es de segundo orden.

Para estudiar los efectos del redondeo, nuevamente hacemos la suposición simplificadora de que el único redondeo que ocurre es en el redondeo de los valores $f(x+h)$ y $f(x-h)$. Entonces, el cociente de diferencia calculado es

$$\frac{f(x+h)(1+\delta_1) - f(x-h)(1+\delta_2)}{2h} =$$

$$= \frac{f(x+h) - f(x-h)}{2h} + \frac{\delta_1 f(x+h) - \delta_2 f(x-h)}{2h}$$

y el término de redondeo $(\delta_1 f(x+h) - \delta_2 f(x-h))/2h$ está acotado en valor absoluto por $\epsilon(\delta_1 f(x+h) - \delta_2 f(x-h))/(2h)$. Nuevamente, ignorando los términos

constantes que involucran a f y sus derivadas, la mayor precisión se logra cuando

$$h^2 \approx \frac{\epsilon}{h} \rightarrow h \approx \epsilon^{1/3}$$

y el error (error de truncamiento o error de redondeo) es $\epsilon^{2/3}$. Con esta fórmula, podemos obtener mayor precisión, hasta aproximadamente la potencia 2/3 de la precisión de la máquina.

Se puede aproximar derivadas de orden superior de manera similar. Para derivar una aproximación de segundo orden a la segunda derivada, nuevamente utilizamos el teorema de Taylor:

$$f(x+h) = f(x)+hf'(x)+\frac{h^2}{2!}f''(x)+\frac{h^3}{3!}f'''(x)+\frac{h^4}{4!}f''''(\xi)$$
$$\xi \in [x, x+h]$$
$$f(x-h) = f(x)-hf'(x)+\frac{h^2}{2!}f''(x)-\frac{h^3}{3!}f'''(x)+\frac{h^4}{4!}f''''(\eta)$$
$$\eta \in [x-h, x]$$

Sumando estas dos ecuaciones se obtiene

$$f(x+h)+f(x-h) = 2f(x)+h^2f''(x)+\frac{h^4}{12}f''''(v)$$
$$v \in [\eta, \xi]$$

Resolviendo para $f''(x)$, obtenemos la fórmula

$$f''(x) = \frac{f(x+h)-2f(x)+f(x-h)}{h^2} - \frac{h^2}{12}f''''(v)$$

Usando la aproximación

$$f''(x) = \frac{f(x+h)-2f(x)+f(x-h)}{h^2} \tag{2.3}$$

el error de truncamiento es $O(h^2)$. Sin embargo, note que un análisis similar del error de redondeo predice un error de redondeo de tamaño ϵ/h^2, por lo que el menor error total ocurre cuando h es aproximadamente $\epsilon^{1/4}$ y luego el error de truncamiento y el error de redondeo son cada uno aproximadamente $\sqrt{\epsilon}$. Con una precisión de máquina $\epsilon \approx 10^{-16}$, esto significa que h no debe ser tomado como menor que aproximadamente 10^{-4}. La evaluación de cocientes estándar de diferencias finitas para derivadas de orden superior es aún más sensible a los efectos del redondeo.

1.1. Formulas de tres puntos

Fórmula de tres puntos hacia delante

$$f'(x) = \frac{-3f(x)+4f(x+h)-f(x+2h)}{2h} + O(h^2) \tag{2.4}$$

Fórmula de tres puntos hacia atrás

$$f'(x) = \frac{f(x-2h)-4f(x-h)+3f(x)}{2h} + O(h^2) \tag{2.5}$$

Fórmula de tres puntos centrada

$$f'(x) = \frac{-f(x-h)+f(x+h)}{2h} - O(h^2) \tag{2.6}$$

1.2. Fórmulas de Cinco Puntos

Una fórmula común de cinco puntos se utiliza para determinar aproximaciones de la derivada en el punto medio.

Fórmula de Cinco Puntos en el Punto Medio

$$f'(x) = \frac{1}{12h}[f(x-2h)-8f(x-h)+8f(x+h) - f(x+2h)] + O(h^4) \tag{2.7}$$

Fórmula de Cinco Puntos en el Extremo

$$f'(x) = \frac{1}{12h}[-25f(x)+48f(x+h)-36f(x+2h) + 16f(x+3h)-3f(x+4h)] + O(h^4) \tag{2.8}$$

1.3. Inestabilidad del Error de Redondeo

Es particularmente importante prestar atención al error de redondeo al aproximar derivadas. Para ilustrar esta situación, examinemos más de cerca la fórmula de tres puntos en el punto medio,

$$f'(x_0) = \frac{1}{2h}[f(x_0+h)-f(x_0-h)] - \frac{h^2}{6}f^{(3)}(\xi_1)$$

Supongamos que al evaluar $f(x_0+h)$ y $f(x_0-h)$ encontramos errores de redondeo $e(x_0+h)$ y $e(x_0-h)$. Entonces, nuestros cálculos realmente utilizan los valores $\tilde{f}(x_0+h)$ y $\tilde{f}(x_0-h)$, que están relacionados con los valores verdaderos $f(x_0+h)$ y $f(x_0-h)$ por $f(x_0+h) = \tilde{f}(x_0+h)+e(x_0+h)$ y $f(x_0-h) = \tilde{f}(x_0-h)+e(x_0-h)$.

El error total en la aproximación,

$$f'(x_0) - \frac{\tilde{f}(x_0+h)-\tilde{f}(x_0-h)}{2h} =$$

$$= \frac{e(x_0+h) - e(x_0-h)}{2h} - \frac{h^2}{6} f^{(3)}(\xi_1)$$

es debido tanto al error de redondeo, la primera parte, como al error de truncamiento. Si asumimos que los errores de redondeo $e(x_0 \pm h)$ están acotados por un número $\epsilon > 0$ y que la tercera derivada de f está acotada por un número $M > 0$, entonces

$$\left| f'(x_0) - \frac{\tilde{f}(x_0+h) - \tilde{f}(x_0-h)}{2h} \right| \leq \frac{\epsilon}{h} + \frac{h^2}{6} M$$

Para reducir el error de truncamiento, $h^2M/6$, necesitamos reducir h. Pero al reducir h, el error de redondeo ϵ/h crece. En la práctica, por lo tanto, rara vez es ventajoso hacer que h sea demasiado pequeño, porque en ese caso el error de redondeo dominará los cálculos.

Ejercicios

(1) Reproducir todos los ejemplos del Greenbaum sección 9.1

(2) Utilziar el Teorema de Taylor para hallar la siguiente fórmula de diferenciación numérica, dejando indicado el orden del error de truncamiento del problema.

$$f'(x) \approx \frac{1}{2h}(-3f(x) + 4f(x+h) - f(x+2h))$$

(3) Para la fórmula de aproximación de la derivada segunda

$$f''(x) \approx \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}$$

determinar el tamño de paso h, en función de ϵ_M a partir del cual el error cometido al aproximar f'' vuelve a crecer. Determinar el orden de ese error en función de ϵ_M.

(4) Utilizar el Teorema de Taylor para hallar la siguiente fórmula de diferenciación numérica, dejando indicado el orden del error de truncamiento del problema.

$$f'''(x) \approx \frac{f(x+2h) - f(x+h) + 2f(x-h) - f(x-2h)}{2h^3}$$

(5) Utilizar el Teorema de Taylor para determinar los coeficientes A, B, C y D de la fórmula

$$f''(x) \approx \frac{Af(x+3h) + Bf(x+2h) + Cf(x+h) + Df(x)}{h^2}$$

e indicar el orden del error de truncamiento.

Capítulo 3

Número de Condición de una Matriz

1. Normas de Matrices

Definición 1.1. Una norma de matriz es una función $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ tal que:

1. $\|A\| \geq 0 \forall A \in \mathbb{R}^{m\times n}$ y $\|A\| = 0$ si y solo si $A = 0$;
2. $\|\alpha A\| = |\alpha|\|A\| \forall \alpha \in \mathbb{R}, \forall A \in \mathbb{R}^{m\times n}$ (homogeneidad);
3. $\|A+B\| \leq \|A\| + \|B\| \forall A, B \in \mathbb{R}^{m\times n}$ (desigualdad triangular).

A menos que se indique lo contrario, utilizaremos el mismo símbolo $\|\cdot\|$ para denotar normas de matrices y normas de vectores.

Podemos caracterizar mejor las normas de matrices introduciendo los conceptos de norma compatible y norma inducida por una norma de vector.

Definición 1.2. Decimos que una norma de matriz $\|\cdot\|$ es compatible o consistente con una norma de vector $\|\cdot\|$ si

$$\|Ax\| \leq \|A\|\|x\|, \quad \forall x \in \mathbb{R}^n \tag{3.1}$$

Más generalmente, dado tres normas, todas denotadas por $\|\cdot\|$, aunque definidas en $\mathbb{R}^m$, $\mathbb{R}^n$ y $\mathbb{R}^{m\times n}$, respectivamente, decimos que son consistentes si $\forall x \in \mathbb{R}^n$, $Ax = y \in \mathbb{R}^m$, $A \in \mathbb{R}^{m\times n}$, tenemos que $\|y\| \leq \|A\|\|x\|$

Para identificar las normas de matrices de interés práctico, generalmente se requiere la siguiente propiedad:

Definición 1.3. Decimos que una norma de matriz $\|\cdot\|$ es sub-multiplicativa si $\forall A \in \mathbb{R}^{m\times n}$, $\forall B \in \mathbb{R}^{m\times q}$

$$\|AB\| \leq \|A\| \cdot \|B\| \tag{3.2}$$

Esta propiedad no es satisfecha por todas las normas de matrices. Por ejemplo, la norma $\|A\|_\Delta = \text{máx}|a_i j|$ para $i = 1, ..., n$, $j = 1, ..., m$ no satisface (3.2) si se aplica a las matrices

$$A = B = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

ya que $2 = \|AB\|_\Delta < \|A\|_\Delta \|B\|_\Delta = 1$.

Observa que, dada una norma de matriz sub-multiplicativa $\|\cdot\|_\alpha$, siempre existe una norma de vector consistente. Por ejemplo, dado cualquier vector fijo $y \neq 0$ en $\mathbb{C}^n$, basta definir la norma de vector consistente como

$$\|x\| = \|xy^H\|_\alpha \quad x \in \mathbb{C}^n$$

Como consecuencia, en el caso de normas de matrices sub-multiplicativas ya no es necesario especificar explícitamente la norma de vector con respecto a la cual la norma de matriz es consistente.

En vista de la definición de una norma natural, recordamos el siguiente teorema.

Teorema 1.1. Sea $\|\cdot\|$ una norma de vector. La función

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} \tag{3.3}$$

es una norma de matriz llamada norma de matriz inducida o norma natural de matriz.

Demostración. Comenzamos observando que (3.3) es equivalente a

$$\|A\| = \sup_{\|x\|=1} \|Ax\| \tag{3.4}$$

De hecho, se puede definir para cualquier $x \neq 0$ el vector unitario $u = x/\|x\|$, de modo que (3.3) se convierte en

$$\|A\| = \sup_{\|u\|=1} \|Au\| = \|Aw\| \quad \|w\| = 1$$

Dicho esto, comprobemos que (3.3) (o, de manera equivalente, (3.4)) es realmente una norma, haciendo uso directo de la Definición 1.1

1. Si $\|Ax\| \geq 0$, entonces se sigue que $\|A\| = \sup_{\|x\|=1} \|Ax\| \geq 0$. Además

$$\|A\| = \sup_{\|x\neq 0} \frac{\|Ax\|}{\|x\|} = 0 \Leftrightarrow \|Ax\| = 0 \forall x \neq 0$$

y $Ax = 0 \forall x \neq 0$ si y solo si $A = 0$; por lo tanto, $\|A\| = 0 \Leftrightarrow A = 0$.

2. Dado un escalar α

$$\|\alpha A\| = \sup_{\|x\|=1} \|\alpha Ax\| = |\alpha| \sup_{\|x\|=1} \|Ax\| = |\alpha| \|A\|$$

3. Finalmente, se cumple la desigualdad triangular. De hecho, por definición de supremo, si $x \neq 0$ entonces

$$\frac{\|Ax\|}{\|x\|} \leq \|A\| \Rightarrow \|Ax\| \leq \|A\| \|x\|$$

de modo que, al tomar x con norma unidad, obtenemos

$$\|(A+B)x\| \leq \|Ax\| + \|Bx\| \leq \|A\| + \|B\|$$

de donde se sigue que $\|A + B\| = \sup_{\|x\|=1} \|(A+B)x\| \leq \|A\| + \|B\|$

□

Instancias relevantes de normas de matrices inducidas son las llamadas normas p definidas como

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}$$

La norma 1 y la norma infinito son fácilmente computables, ya que

$$\|A\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{m} |a_{ij}|, \quad \|A\|_\infty = \max_{i=1,\ldots,m} \sum_{j=1}^{n} |a_{ij}|$$

y se les llama la norma de la suma de columnas y la norma de la suma de filas, respectivamente.

Además, tenemos $\|A\|_1 = \|A^T\|_\infty$ y, si A es autoadjunto o simétrico real, $\|A\|_1 = \|A\|_\infty$.

Una discusión especial merece la norma 2 o norma espectral, para la cual se cumple el siguiente teorema.

Teorema 1.2. Sea $\sigma_1(A)$ el mayor valor singular de A. Entonces

$$\|A\|_2 = \sqrt{\rho(A^H A)} = \sqrt{\rho(AA^H)} = \sigma_1(A) \tag{3.5}$$

En particular, si A es hermítica (o simétrica real), entonces

$$\|A\|_2 = \rho(A) \tag{3.6}$$

mientras que, si A es unitaria, $\|A\|_2 = 1$.

Demostración. Dado que $A^H A$ es hermítica, existe una matriz unitaria U tal que

$$U^H A^H A U = \text{diag}(\mu_1, ..., \mu_n)$$

donde μ_i son los valores propios (positivos) de $A^H A$. Sea $y = U^H x$, entonces

$$\|A\|_2 = \sup_{x \neq 0} \sqrt{\frac{(A^H Ax, x)}{(x,x)}} = \sup_{y \neq 0} \sqrt{\frac{(U^H A^H AUy, y)}{(y,y)}}$$

$$= \sup_{y \neq 0} \sqrt{\sum_{i=1}^{n} \mu_i |y_i|^2 / \sum_{i=1}^{n} |y_i|^2} = \sqrt{\max_{i=1,\ldots,n} |\mu_i|}$$

de donde sigue (3.5), gracias a

Si A es hermítica, las mismas consideraciones anteriores se aplican directamente a A. Finalmente, si A es unitaria, tenemos

$$\|Ax\|_2^2 = (Ax, Ax) = (x, A^H Ax) = \|x\|_2^2$$

de modo que $\|A\|_2 = 1$. □

Como consecuencia, el cálculo de $\|A\|_2$ es mucho más extenso que el de $\|A\|_\infty$ o $\|A\|_1$. Sin embargo, si solo se requiere una estimación de $\|A\|_2$, se pueden emplear con provecho las siguientes relaciones en el caso de matrices cuadradas

$$\max_{i,j} |a_{ij}| \leq \|A\|_2 \leq n \max_{i,j} a_{ij}$$

$$\frac{1}{\sqrt{n}} \|A\|_\infty \leq \|A\|_2 \leq \sqrt{n} \|A\|_\infty$$

$$\frac{1}{\sqrt{n}} \|A\|_1 \leq \|A\|_2 \leq \sqrt{n} \|A\|_1$$

$$\|A\|_2 \leq \sqrt{\|A\|_1 \|A\|_\infty}$$

Además, si A es normal, entonces $\|A\|_2 \leq \|A\|_p$ para cualquier n y todo $p \geq 2$.

Teorema 1.3. Sea $|||\cdot|||$ una norma de matriz inducida por una norma de vector $\|\cdot\|$. Entonces, se cumplen las siguientes relaciones:

1. $\|Ax\| \leq |||A|||x\|$, es decir, $|||\cdot|||$ es una norma compatible con $\|\cdot\|$;
2. $|||I||| = 1$
3. $|||AB||| \leq |||A|||||B|||$, es decir, $|||\cdot|||$ es submultiplicativa.

Demostración. La parte 1 del teorema ya está contenida en la prueba del Teorema 1.1, mientras que la parte 2 sigue del hecho de que $|||I||| = \sup_{x\neq 0} \|Ix\|/\|x\| = 1$. La parte 3 es fácil de verificar. □

Notemos que las normas p son submultiplicativas. Además, hacemos notar que la propiedad de submultiplicatividad por sí sola solo nos permitiría concluir que $|||I||| \geq 1$. De hecho, $|||I||| = |||I \cdot I||| \leq |||I|||^2$.

2. Análisis de Estabilidad de Sistemas Lineales

2.1. El Número de Condición de una Matriz

El número de condición de una matriz $A \in \mathbb{C}^{m\times n}$ se define como

$$K(A) = \|A\|\|A^{-1}\| \tag{3.7}$$

donde $\|\cdot\|$ es una norma de matriz inducida. En general, $K(A)$ depende de la elección de la norma; esto quedará claro al introducir un subíndice en la notación, por ejemplo, $K_\infty(A) = \|A\|_\infty\|A^{-1}\|_\infty$. Más generalmente, $K_p(A)$ denotará el número de condiciónd e A en la norma p.

Comencemos notando que $K(A) \geq 1$, ya que

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

Además, $K(A^{-1}) = K(A)$ y $\forall\alpha \in \mathbb{C}$ con $\alpha \neq 0$, $K(\alpha A) = K(A)$. Finalmente, si A es ortogonal, $K_2(A) = 1$ ya que $\|A\|_2 = \sqrt{\rho(A^TA)} = \sqrt{\rho I} = 1$ y $A^{-1} = A^T$. El número de condición de una matriz singular se establece como infinito.

Observación 2.1. Define la distancia relativa de $A \in \mathbb{C}^{m\times n}$ con respecto al conjunto de matrices singulares en la norma p como

$$dist_p(A) = min\left\{\frac{\|\delta A\|_p}{\|A\|_p} : A + \delta A \text{ es singular}\right\}$$

Entonces, se puede demostrar que

$$dist_p(A) = \frac{1}{K_p(A)} \tag{3.8}$$

La ecuación 3.8 sugiere que una matriz A con un número de condición alto puede comportarse como una matriz singular de la forma $A + \delta A$. En otras palabras, las perturbaciones nulas en el lado derecho no necesariamente producen cambios no nulos en la solución, ya que, si $A + \delta A$ es singular, el sistema homogéneo $(A+\delta A)z = 0$ ya no admite solo la solución nula. Observe que si se cumple la siguiente condición:

$$\|A^{-1}\|_p\|\delta A\|_p < 1 \tag{3.9}$$

entonces la matriz $A + \delta A$ es no singular.

3. Condicioneamiento de Sistemas Lineales

El problema de resolver $Ax = b$ involucra, como entrada, una matriz A y un vector b en el término derecho, mientras que la salida es un vector x. Para medir el cambio en la salida debido a un pequeño cambio en la entrada, debemos discutir las normas de vecotres y matrices.

3.1. Normas

Definición 3.1. Una norma para vectores es una función $\|\cdot\|$ que satisface, para todos los vecotres v, w de dimensión n:

1. $\|v\| > 0$, con igualdad si y solo si $v = 0$
2. $\|\alpha v\| = |\alpha|\|v\|$ para cualquier escalar α
3. $\|v + w\| \leq \|v\| + \|w\|$ (desigualdad triangular)

La norma de vectores más utilizada en $\mathbb{R}^n$ es la norma-2 o norma euclidiana:

$$\|v\|_2 = \sqrt{\sum_{i=1}^{n} |v_i|^2}$$

Otras normas frecuentemente empleadas son:

- Norma-∞:

$$\|v\|_\infty = \max_{1\le i\le n} |v_i|$$

- Norma-1:

$$\|v\|_1 = \sum_{i=1}^{n} |v_i|$$

Más generalmente, se puede demostrar que para cualquier $p \geq 1$ (sin necesidad de que p sea un entero), la norma-p, definida por:

$$\|v\|_p = \left(\sum_{i=1}^{n} |v_i|^p\right)^{1/p}$$

es una norma. Usualmente trabajaremso con la norma-1, la norma-2 o la norma-∞. De estas, solo la norma-2 proviene de un producto interno; es decir,

$$\|v\|_2 = \langle v, v\rangle^2, \text{ donde } \langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$$

3.2. Sensibilidad de las Soluciones de Sistemas Lineales

Considere el siguiente sistema lineal:

$$\begin{pmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{n} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+1} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n+1} & \cdots & \frac{1}{2n+-1} \end{pmatrix} x = b$$

Esta matriz es conocida como una matriz de Hilbert, y se sabe que es notoriamente mal condicionada.

Supongamos, entonces, que $\hat{b}$ es un vector tal que, en cierta norma, $\|b - \hat{b}\|$ es pequeño. Sea x la solución del sistema lineal $Ax = b$, y sea $\hat{x}$ la solución del sistema lineal $A\hat{x} = \hat{b}$. Restando estas dos ecuaciones, entontramos que $A(x - \hat{x}) = b - \hat{b}$, o, $x - \hat{x} = A^{-1}(b - \hat{b})$. Tomando normas en cada lado se obtiene la desigualdad

$$\|x - \hat{x}\| \leq \|A^{-1}\| \cdot \|b - \hat{b}\| \tag{3.10}$$

El factor $\|A^{-1}\|$ puede considerarse como un número de condición absoluto para este problema, sin embargo, suele ser el error relativo, en lugar del error absoluto, el que resulta de interés; en este caso, nos gustaría relacionar $\frac{\|x-\hat{x}\|}{\|x\|}$ con $\frac{\|b-\hat{b}\|}{\|b\|}$. Dividiendo cada lado de 3.10, encontramos

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

Dado que $\frac{\|b\|}{\|x\|} = \frac{\|Ax\|}{\|x\|}$, se sigue que

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \|A^{-1}\| \cdot \|A\| \cdot \frac{\|b - \hat{b}\|}{\|b\|} \tag{3.11}$$

El número $\|A\| \cdot \|A^{-1}\|$ sirve como una especie de número de condición relativo para el problema de resolver $Ax = b$.

Definición 3.2. El número $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ se llama el número de condición de la matriz no singular A.

Teorema 3.1. Sea A una matriz no singular $n \times n$, b un vector dado de n-dimensiones, y x que satisface $Ax = b$. Sea $A + E$ otra matriz no singular de $n \times n$ y $\hat{x}$ otro vector de n-dimensiones, y sea $\hat{x}$ que satisface $(A + E)\hat{x} = b$. Entonces,

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \left(\|(A + E)^{-1}\| \cdot \|A\|\right) \left(\frac{\|b - \hat{b}\|}{\|b\|} + \frac{\|E\|}{\|A\|}\right) \tag{3.12}$$

Si $\|E\|$ es suficientemente pequeño tal que $\|A^{-1}\| \cdot \|E\| < 1$, entonces

$$\frac{\|x - \hat{x}\|}{\|x\|} \leq \frac{\kappa(A)}{1 - \kappa(A)\|E\|/\|A\|} \cdot \left(\frac{\|b - \hat{b}\|}{\|b\|} + \frac{\|E\|}{\|A\|}\right) \tag{3.13}$$

Ejercicios

(1) Sea

$$\|\cdot\| : M_{n\times m} \to (\mathbb{R})$$

la aplicación definida por

$$\|A\| = max\{|a_{ij}| : i = 1, ..., n, j = 1, ..., m\}$$

Se piede probar que $\|\cdot\|$ es una norma pero que no está inducida por una norma vectorial (Quarteroni pág. 22).

(2) Dado $\gamma \geq 0$, se considera la matriz

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

Comprobar que

$$K_\infty(A) = K_1(A) = (a + \gamma)^2$$

Ahora, se considera el sistema $Ax = b$ donde $x = (1 - \gamma, 1)^T$ es la solución del sistema. Para $\gamma = 101$ y $\gamma = 0{,}01$ calcular $K_\infty(A)$ y $K_{rel}(b)$ (con norma infinito) y decidir si el problema está bien o mal condicionado.

(3) Utilizar interpolación de Vandermonde para aproximar la función $f(x) = cos(kx)$, con $k = 10, 20, 30, ..., 100$. Variar el grado del polinomio interpolador para apreciar la inestabilidad. Calcular el número de condición de las matrices de Vandermonde obtenidas, evaluar en los nodos y hacer la gráfica del error.

Capítulo 4

Interpolación

1. Interpolación Polinómica

En esta sección, resolvemos el siguiente problema: Se nos da una tabla de $n+1$ puntos de datos (x_i, y_i) y buscamos un polinomio p de menor grado posible tal que $p(x_i) = y_i \quad (0 \leq i \leq n)$. Dicho polinomio se dice que interpola los datos. Aquí está el teorema que rige este problema.

Teorema 1.1. Si $x_0, x_1, ..., x_n$ son números reales distintos, entonces para valores arbitrarios $y_0, y_1, ..., y_n$ hay un único polinomio p_n de grado como mucho n tal que

$$p_n(x_i) = y_i \quad (0 \leq i \leq n)$$

Demostración. Demostraremos primero la unicidad. Supongamos que existieran dos tales polinomios, p_n y q_n. Entonces, el polinomio $p_n - q_n$ tendría la propiedad de que $(p_n - q_n)(x_i) = 0$ para $0 \leq i \leq n$. Dado que el grado de $p_n - q_n$ puede ser como máximo n, este polinomio puede tener como máximo n ceros si no es el polinomio nulo. Como los x_i son distintos, $p_n - q_n$ tiene $n+1$ ceros; por lo tanto, debe ser cero. Así, $p_n \equiv q_n$.

Para la parte de existencia del teorema, procedemos inductivamente. Para $n = 0$, la existencia es obvia ya que se puede elegir una función constante p_0 (polinomio de grado ≤ 0) tal que $p_0(x_0) = y_0$. Ahora supongamos que hemos obtenido un polinomio p_{k-1} de grado $\leq k-1$ con $p_{k-1}(x_i) = y_i$ para $0 \leq i \leq k-1$. Intentamos construir p_k de la forma

$$p_k(x) = p_{k-1}(x) + c(x - x_0)(x - x_1)...(x - x_{k-1}) \quad (4.1)$$

Observa que esto es, sin lugar a dudas, un polinomio de grado como máximo k. Además, p_k interpola los datos que interpola p_{k-1}, porque

$$p_k(x_i) = p_{k-1}(x_i) = y_i \quad (0 \leq i \leq k-1)$$

Ahora determinamos el coeficiente desconocido c a partir de la condición $p_k(x_k) = y_k$. Esto conduce a la ecuación

$$p_{k-1}(x_k) + c(x_k - x_0)(x_k - x_1)...(x_k - x_{k-1} = y_k) \quad (4.2)$$

La ecuación 4.2 ciertamente puede resolverse para c porque los factores que multiplican a c no son cero.

□

1.1. Forma de Newton del Polinomio de Interpolación

Antes de intentar escribir un algoritmo que lleve a cabo el proceso recursivo en esta demostración, necesitamos hacer algunas observaciones. En primer lugar, los polinomios $p_0, p_1, \ldots, p_n$ construidos en la demostración tienen la propiedad de que cada p_k se obtiene simplemente añadiendo un término a p_{k-1}. Por lo tanto, al final del proceso, p_n será una suma de términos y cada $p_0, p_1, \ldots, p_{n-1}$ será claramente visible en la expresión para p_n. Cada p_k tiene la forma

$$p_k(x) = c_0 + c_1(x - x_0) + c_2(x - x_0)(x - x_1) + ... + c_k(x - x_0)...(x - x_{k-1}) \quad (4.3)$$

Su forma compacta es

$$p_k(x) = \sum_{i=0}^{k} c_i \prod_{j=0}^{i-1} (x - x_j) \quad (4.4)$$

Estos polinomios se denominan polinomios de interpolación en la forma de Newton.

1.2. Forma de Lagrange del Polinomio de Interpolación

Ahora presentamos una forma alternativa para el polinomio de interpolación p asociado con una tabla de puntos de datos (x_i, y_i) para $0 \leq i \leq n$. Es importante entender que existe un único polinomio de interpolación de grado $\leq n$ asociado con los datos. Sin embargo, ciertamente existe la posibilidad de expresar este polinomio en diferentes formas y de llegar a él mediante distintos algoritmos.

El método alternativo expresará p en la forma

$$p(x) = y_0 l_o(x) + y_1 l_1(x) + ... + y_n l_n(x) = \sum_{k=0}^{n} y_k l_k(x) \tag{4.5}$$

Aquí $l_0, l_1, \ldots, l_n$ son polinomios que dependen de los nodos $x_0, x_1, \ldots, x_n$ pero no de las ordenadas $y_0, y_1, \ldots, y_n$. Dado que las ordenadas podrían ser todas 0 excepto por un 1 ocupando la posición i-ésima, vemos que

$$\delta_{ij} = p_n(x_j) = \sum_{k=0}^{n} y_k l_k(x_j) = \sum_{k=0}^{n} \delta_{ki} l_k(x_j) = l_i(x_j)$$

La fórmula general de l_i es

$$l_i(x) = \prod_{\substack{j=0 \\ j\neq 0}}^{n} \frac{x - x_j}{x_0 - x_j} \quad (0 \leq i \leq n) \tag{4.6}$$

Para el conjunto de nodos $x_0, x_1, \ldots, x_n$, estos polinomios son conocidos como las funciones cardinales. Con los polinomios cardinales en mano, la Ecuación 4.5 proporciona la forma de Lagrange de los polinomios interpolantes.

Existen aún otros algoritmos para la interpolación polinómica, y estos tienen diversas ventajas y desventajas. Dado que existe un único polinomio de grado $\leq n$ que toma valores prescritos en $n+1$ puntos dados (y distintos), estos algoritmos producen el mismo polinomio en diferentes formas. Por ejemplo, podemos requerir que nuestro polinomio se exprese en potencias de x:

$$p(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n \tag{4.7}$$

Las condiciones de interpolación, $p(x_i) = y_i$ para $0 \leq i \leq n$, conducen a un sistema de $n+1$ ecuaciones lineales para determinar $a_0, a_1, \ldots, a_n$. Este sistema tiene la forma:

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ 1 & x_2 & x_2^2 & \ldots & x_2^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{4.8}$$

La matriz de coeficientes aquí se llama matriz de Vandermonde. Es no singular porque el sistema tiene una solución única para cualquier elección de $y_0, y_1, \ldots, y_n$. El determinante de la matriz de Vandermonde, por lo tanto, es no nulo para nodos distintos $x_0, x_1, \ldots, x_n$. Sin embargo, la matriz de Vandermonde a menudo está mal condicionada, por lo que los coeficientes a_i pueden ser determinados de manera imprecisa al resolver el Sistema 4.8. Además, la cantidad de trabajo necesario para obtener el polinomio en 4.7 es excesiva. Por lo tanto, este enfoque no se recomienda.

1.3. El Error en la Interpolación Polinómica

Ahora presentamos algunos teoremas que se refieren a la discrepancia entre una función y un polinomio interpolante de esta.

Teorema 1.2. Sea f una función en $C^{n+1}[a, b]$, y sea p el polinomio de grado $\leq n$ que interpola la función f en $n+1$ puntos distintos $x_0, x_1, \ldots, x_n$ en el intervalo $[a, b]$. Para cada x en $[a, b]$, existe un punto ξ_x en (a, b) tal que

$$f(x) - p(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \prod_{i=0}^{n} (x - x_i) \tag{4.9}$$

Demostración. Si x es uno de los nodos de interpolación x_i, la afirmación es obviamente verdadera, ya que ambos lados de la Ecuación 4.9 se reducen a cero. Entonces, sea x cualquier punto distinto de un nodo. Definamos

$$w(t) = \prod_{i=0}^{n} (t - x_i) \quad \phi \equiv f - p - \lambda w$$

donde λ es el número real que hace que $\phi(x) = 0$. Por lo tanto,

$$\lambda = \frac{f(x) - p(x)}{w(x)}$$

Ahora $\phi \in C^{n+1}[a, b]$ y ϕ se anula en los puntos $n+2$ $x, x_0, x_1, ..., x_n$. Por el Teorema de Rolle, ϕ' tiene al menos $n+1$ ceros distintos en (a, b). De manera similar, ϕ'' tiene al menos n ceros distintos en (a, b). Si este argumento se repite, concluimos eventualmente que $\phi^{(n+1)}$ tiene al menos un cero, digamos ξ_x, en (a, b). Ahora

$$\phi^{(n+1)} = f^{(n+1)} - p^{(n+1)} - \lambda w^{(n+1)} = f^{(n+1)} - (n+1)!\lambda$$

Entonces

$$0 = \phi^{(n+1)}(\xi_x) = f^{(n+1)}(\xi_x) - (n+1)!\lambda = f^{(n+1)}(\xi_x) - (n+1)!\frac{f(x)-p(x)}{w(x)}$$

□

1.4. Polinomios de Chebychev

En el Teorema 1.2, hay un término que puede optimizarse eligiendo los nodos de una manera especial. El proceso de optimización conduce naturalmente a un sistema de polinomios llamados polinomios de Chebyshev, y comenzamos con su definición y propiedades básicas.

Los polinomios de Chebyshev se definen recursivamente de la siguiente manera:

$$\begin{cases} T_0(x) = 1 \quad T_1(x) = x \\ T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x) \quad (n \geq 1) \end{cases} \tag{4.10}$$

Teorema 1.3. Para x en el intervalo $[-1, 1]$, los polinomios de Chebyshev tienen la siguiente expresión en forma cerrada:

$$T_n(x) = cos(ncos^{-1}x) \quad (n \geq 0)$$

Teorema 1.4. Si p es un polinomio mónico de grado n, entonces

$$\|p\|_\infty = \max_{-1 \leq x \leq 1} |p(x)| \geq 2^{1-n}$$

Demostración. Procedemos por contradicción. Supongamos que

$$|p(x)| < 2^{1-n} \quad (|x| \leq 1)$$

Sea $q = 2^{1-n}T_n$ y $x_i = cos(i\pi/n)$. Como se señaló anteriormente, q es un polinomio mónico de grado n. Entonces

$$(-1)^i p(x_i) \leq |p(x_i)| < 2^{1-n} = (-1)^i q(x_i)$$

En consecuencia

$$(-1)^i[q(x_i) - p(x_i)] > 0 \quad (0 \leq i \leq n)$$

Esto muestra que el polinomio $q-p$ oscila en signo $n+1$ veces en el intervalo $[-1, 1]$. Por lo tanto, debe tener al menos n raíces en $(-1, 1)$. Pero esto no es posible porque $q - p$ tiene grado como máximo $n - 1$. □

1.5. Eligiendo los Nodos

En el Teorema 1.2, supongamos que los nodos de interpolación están en el intervalo $[-1, 1]$. Si x está en este mismo intervalo, ξ_x también estará en ese intervalo. Por lo tanto, podemos deducir que $\max_{|x|\leq 1} |f(x) - p(x)| \leq \frac{1}{(n+1)!} \max_{|x|\leq 1} |f^{(n+1)}(x)| \max_{|x|\leq 1} |\prod_{i=0}^{n}(x - x_i)|$ Por el Teorema 1.4, tenemos

$$\max_{|x|\leq 1} \left| \prod_{i=0}^{n}(x - x_i) \right| \geq 2^{-n}$$

Este valor mínimo se alcanzará si $\prod_{i=0}^{n}(x - x_i)$ es el múltiplo monico de T_{n+1}; que es, $2^{-n}T_{n+1}$. Los modos serán entonces las raíces de T_{n+1}. Estas son

$$x_i = cos\left(\frac{2i+1}{2n+2}\pi\right) \quad (0 \leq i \leq n)$$

Estas consideraciones establecen el siguiente resultado.

Teorema 1.5. Si los nodos x_i son las raíces del polinomio de Chebyshev T_{n+1}, entonces la fórmula de error en el Teorema 2 da (para $|x| \leq 1$):

$$|f(x) - p(x)| \leq \frac{1}{2^n(n+1)!} \max_{|t|\leq 1} |f^{(n+1)}(t)|$$

1.6. El Teorema de Aproximación de Weiertrass

Teorema 1.6. Si f es continua en $[a, b]$ y si $\epsilon > 0$, existe un polinomio p tal que $|f(x) - p(x)| \leq \epsilon$ en el intervalo $[a, b]$.

2. Interpolación Unidimensional

El alcance de esta sección es la teoría de la interpolación de funciones definidas en un intervalo $[a, b]$. Para un entero $k \geq 0$, $\mathbb{P}_k$ denota el espacio de los polinomios en una variable, con coeficientes reales y de grado a lo sumo k.

2.1. La malla

Una malla de $\Omega = [a, b]$ es una colección indexada de intervalos de medida no nula $\{I_i = [x_{1,i}, x_{2,i}]\}_{0\leq i\leq N}$ que forma una partición de Ω, es decir,

$$\bar{\Omega} = \cup_{i=0}^{N} I_i \quad \text{y} \quad \mathring{I}_i \cap \mathring{I}_j = \emptyset \quad \text{para} \quad i \neq j \tag{4.11}$$

La forma más sencilla de construir una malla es tomar $N+2$ puntos en Ω tal que

$$a = x_0 < x_1 < ... < x_N < x_{N+1} = b \tag{4.12}$$

y definir $x_{1,i} = x_i$ y $x_{2,i} = x_{i+1}$ para $0 \leq i \leq N$. Los puntos en el conjunto $\{x_0, ..., x_{N+1}\}$ se llaman los vértices de la malla. El mallado puede tener tamaño de paso variable

$$h_i = x_{i+1} - x_i \quad 0 \leq i \leq N$$

y establecemos

$$h = \max_{0 \leq i \leq N} h_i$$

En lo sucesivo, los intervalor I_i también se llaman elementos (o celdas) y la malla se denota por $T_h = \{I_i\}_{0 \leq i \leq N}$. El subíndice h hace referencia al nivel de refinamiento.

2.2. El elemento finito $\mathbb{P}_1$ de Lagrange

Considera el espacio vectorial de funciones continuas, lineales a trozos

$$P_h^1 = \{v_h \in C^0(\bar{\Omega}); \forall i \in \{0, ..., N\}, v_{h|I_i} \in \mathbb{P}_1\} \tag{4.13}$$

Este espacio puede ser utilizado en conjunto con los métodos de Garlekin para aproximar ecuaciones en derivadas parciales unidimensionales. Por esta razón, P_1^h se llama un espacio de aproximación. Introducimos las funciones $\{\phi_0, ..., \phi_{N+1}\}$ definidas elemento por elemento como sigue: Para $i \in \{0, ..., N+1\}$

$$\phi_i(x) = \begin{cases} \frac{1}{h_{i-1}}(x - x_{i-1}) & \text{si } x \in I_{i-1} \\ \frac{1}{h_i}(x_{i+1} - x) & \text{si } x \in I_i \\ 0 \quad \text{en otro caso} \end{cases} \tag{4.14}$$

con modificaciones obvias si $i = 0$ o $i = N+1$. Claramente, $\phi_i \in P_i^h$. Estas funciones son comúnmente llamadas "funciones de sombrero" debido a la forma de su gráfico; véase la figura 4.1.

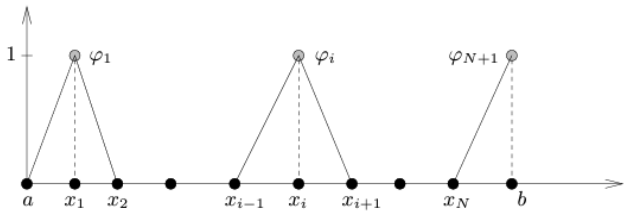

Figura 4.1: Funciones de Sombrero Unidimensionales

2.3. El elemento finito $\mathbb{P}_k$ de Lagrange

La técnica de interpolación presentada en la sección 2.2 se generaliza a polinomios de mayor grado. Consideremos la malla $\tau_h = \{I_i\}_{0 \leq i \leq N}$ introducida en la sección 2.1. Sea

$$P_h^k = \{v_h \in C^0(\bar{\Omega}); \forall i \in \{0, ..., N\}, v_{h|I_i} \in \mathbb{P}_k\} \tag{4.15}$$

Para investigar un operador de interpolación con codominio en P_h^k, es conveniente considerar los polinomios de Lagrange. Recuerda lo siguiente:

Definición 2.1. Polinomios de Lagrange Sea $k \geq 1$ y $\{s_0, ..., s_k\}$ un conjunto de $k+1$ números distintos. Los polinomios de Lagrange $\{\mathcal{L}_0^k, ..., \mathcal{L}_k^k\}$ asociados con los nodos $\{s_0, ..., s_k\}$ se definen como

$$\mathcal{L}_m^k(t) = \frac{\prod_{l \neq m}(t - s_l)}{\prod_{l \neq m}(s_m - s_l)}, \quad 0 \leq m \leq k \tag{4.16}$$

Los polinomios de Lagrange cumplen la propiedad importante

$$\mathcal{L}_m^k(s_l) = \delta_{ml}, \quad 0 \leq m, l \leq k$$

La Figura 4.2 presenta familias de polinomios de Lagrange con nodos equidistribuidos en el intervalo de referencia $[0, 1]$ para $k = 1, 2$ y 3.

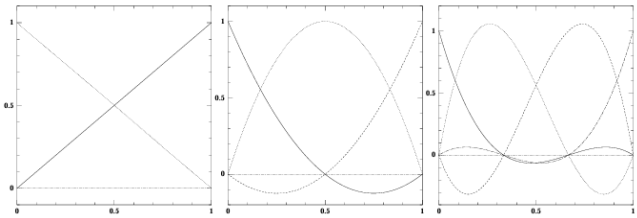

Figura 4.2: Familias de polinomios de Lagrange con nodos equidistribuidos en el intervalo de referencia $[0, 1]$ y de grado $k = 1$ (izquierda), $k = 2$ (centro) y $k = 3$ (derecha).

Para $i \in \{0, ..., N\}$, introducimos los nodos $\xi_{i,m} = x_i + m\frac{h_i}{k}$, $0 \leq m \leq k$, en el intervalo de la malla I_i. Sea $\{\mathcal{L}_{i,0}^k, ..., \mathcal{L}_{i,k}^k\}$ el conjunto de polinomios de Lagrange asociados con estos nodos. Para $j \in \{0, ..., k(N+1)\}$ con $j = ki + m$ y $0 \leq m \leq k-1$, definimos la función ϕ_j elemento a elemento de la siguiente manera: Para $\leq m \leq k-1$

$$\phi_{ki+m}(x) = \begin{cases} \mathcal{L}_{i,m}^k(x) & \text{si } x \in I_i \\ 0 \quad \text{en caso contrario} \end{cases}$$

y para $m = 0$,

$$\phi_{ki}(x) = \begin{cases} \mathcal{L}_{i,0}^k(x) & \text{si } x \in I_i \\ \mathcal{L}_{i-1,k}^k(x) & \text{si } x \in I_{i-1} \\ 0 \quad \text{en caso contrario} \end{cases}$$

con modificaciones evidentes si $i = 0$ o $i = N+1$. Las funciones ϕ_j se ilustran en la Figura 4.3 para $k = 2$.

Obsérvese la diferencia entre el soporte de las funciones asociadas a los vértices de la malla (dos intervalos adyacentes) y el de las funciones asociadas a los puntos medios de las celdas (un intervalo).

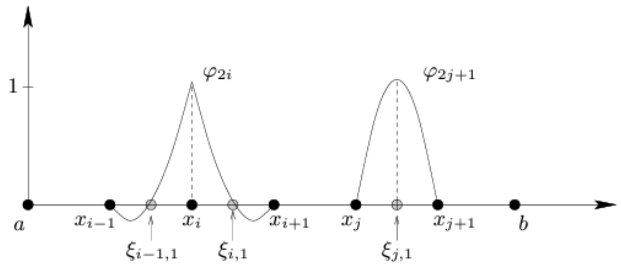

Figura 4.3: Funciones de forma globales en el espacio de aproximación P_h^2.

Ejercicios

(1) Sea $f(x) = (3+x)cos^2(\pi x/4)$, $x \in [0,3]$. Utilizar el polinomio interpolador de Lagrange cuadrático con nodos $x_1 = 0, x_2 = 1, x_3 = 3$ para aproximar $f(2)$, $f(12/5)$, $f(7/2)$ y $f(4)$.

(2) Para las siguientes funciones $f(x)$, escribir el término del error $E_2(x)$ del polinomio interpolador de Lagrange cuadrático con nodos $x_0 = -1, x_1 = 1, x_2 = 3$.

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

(3) Aproximar el valor $0{,}15^{1/7}$ mediante el polinomio interpolador cuadrático de la función $f(x) = 2^x$ con los nodos $x_0 = -1, x_1 = 0, x_2 = 1$. Acotar el error cometido.

(4) ¿Cuál es el polinomio interpolaor de Lagrange de grado 21 con nodos equidistantes, de la función $f(x) = x^5 + 3x^{12}$, $x \in [0,20]$.

(5) Se considera la función $f(x) = \frac{1}{1+x^2}$. Calcular y representar gráficamente el polinomio de interpolación de grado 14 con puntos de interpolación equiespaciados en el intervalo $[-5,5]$.

(6) Haciendo uso de la transformación lineal $F : [-1,1] \to [-5,5]$, con $F(x) = 5x$, repetir el ejercicio anterior utilizando como puntos de interpolación $x_i = F(\hat{x}_i)$, donde los puntos

$$\hat{x}_i = cos\left(\frac{2i+1}{2n+2}\pi\right) 0 \leq i \leq n$$

(7) Dado el intervalo $[0,2]$, se genera un mallado D_h formado por elementos de anchura $h = 0{,}5$, y sea V_h el esapcio de elementos finitos generado con polinomios de grado 1 o 2. Se considera la función $f : [0,2] \to \mathbb{R}$ dad por $f(x) = sin(\pi x)$ y sea $f_h \in V_h$ una aproximación de f en el espacio V_h verificando que $f(x_i) = f_h(x_i)$ para todo x_i nodo del mallado. Se piede:

- Calcular $|f(0{,}8) - f_h(0{,}8)|$
- Repetir el ejercicio utilizando $h = 1/2^j$, $j = 2,3,4$.
- Para qué valor de h podemos afirmar que el error está por debajo de una tolerancia de $\delta = 10^{-6}$.

(8) Determinar el valor de h para que la aproximación $f_h \in V_h$ de la función f verifique que

$$\|f - f_h\|_{L^\infty([a,b])} < \delta$$

suponiendo que el mallado asociado a V_h es equiespaciado, que el grado de los polinomios para construir V_h es $m = 1$ o 2, que $\delta = 10^{-8}$ y para las siguientes funciones f:

- $f : [0,10] \to \mathbb{R}$, $f(x) = xsinx$
- $f : [-1,1] \to \mathbb{R}$, $f(x) = sin(\pi x)$
- $f : [-1,1] \to \mathbb{R}$, $f(x) = xarctgx$
- $f : [-5,5] \to \mathbb{R}$, $f(x) = e^{-x^2}$

Capítulo 5

Integración

1. Elementos de integración numérica

A menudo surge la necesidad de evaluar la integral definida de una fucnión que no tiene una antiderivada o cuta antiderivada no es fácil de obtemer. El método básico asociado con la aproximación de $\int_a^b f(x)dx$ recibe el nombre de cuadratura numérica. Éste utiliza una suma $\sum_{i=0}^n a_i f(x_i)$ para aproximar $\int_a^b f(x)dx$.

Los métodos de cuadratura se basan en los polinomios de interpolación. La idea básica es seleccionar un conjunto de nodos distintos $\{x_0, ..., x_n\}$ del intervalo $[a, b]$. Entonces integramos el polinomio interpolante de Lagrange

$$P_n(x) = \sum_{i=0}^{n} f(x_i)L_i(x)$$

y su término de error de truncamiento sobre $[a, b]$ para obtener

$$\int_a^b f(x)dx = \int_a^b \sum_{i=0}^n f(x_i)L_i(x)dx + \int_a^b \prod_{i=0}^n (x - x_i)\frac{f^{(n+1)}(\xi(x))}{(n+1)!}dx$$
$$= \sum_{i=0}^n a_i f(x_i) + \frac{1}{(n+1)!}\int_a^b \prod_{i=0}^n (x - x_i)f^{(n+1)}(\xi(x))dx$$

donde $\xi(x)$ se encuentra en $[a, b]$ para cada x y

$$a_i = \int_a^b L_i(x)dx, \text{ para cada } i = 0, 1, ..., n$$

La fórmula de cuadratura es, por lo tanto,

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

con un error dado por

$$E(f) = \frac{1}{(n+1)!}\int_a^b \prod_{i=0}^{n}(x - x_i)f^{(n+1)}(\xi(x))dx$$

Antes de analizar la situación general de las fórmulas de cuadratura, consideremos las fórmulas producidas mediante el uso del primer y del segundo polinomios de Lagrange con nodos igualmente espaciados. Esto da la regla trapezoidal y la regla de Simpson.

1.1. La regla trapezoidal

Para derivar la regla trapezoidal (o regla del trapecio) para aproximar $\int_a^b f(x)dx$, sean $x_0 = a$, $x_1 = b$, $h = b - a$ y utilice el polinomio de Lagrange

$$P_1(x) = \frac{(x - x_1)}{(x_0 - x_1)}f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)}f(x_1)$$

Entonces

$$\int_a^b f(x)dx = \int_{x_0}^{x_1}\left[\frac{(x - x_1)}{x_0 - x_1}f(x_0) + \frac{(x - x_0)}{(x_1 - x_0)}f(x_1)\right]dx + \frac{1}{2}\int_{x_0}^{x_1} f''(\xi(x))(x - x_0)(x - x_1)dx \quad (5.1)$$

El producto $(x - x_0)(x - x_1)$ no cambia de signo en $[x_0, x_1]$, por lo que el teorema de valor promedio ponderado para integrales se puede aplicar al término de error para obtener, para algunos ξ en (x_0, x_1),

$$\int_{x_0}^{x_1} f''(\xi(x))(x - x_0)(x - x_1)dx$$
$$= f''(\xi)\int_{x_0}^{x_1}(x - x_0)(x - x_1)dx$$
$$= f''(\xi)\left[\frac{x^3}{3} - \frac{(x_1 + x_0)}{2}x^2 + x_0x_1x\right]_{x_0}^{x_1}$$
$$= -\frac{h^3}{6}f''(\xi)$$

Por consiguiente, la ecuación 5.1 implica que

$$\int_a^b f(x)dx =$$
$$= \left[\frac{(x - x_1)^2}{2(x_0 - x_1)}f(x_0) + \frac{(x - x_0)^2}{2(x_1 - x_0)}f(x_1)\right]_{x_0}^{x_1} - \frac{h^3}{12}f''(\xi) =$$
$$= \frac{(x_1 - x_0)}{2}[f(x_0) + f(x_1)] - \frac{h^3}{12}f''(\xi)$$

Por medio de la notación $h = x_1 - x_0$ obtenemos la siguiente regla:

Regla trapezoidal:

$$\int_a^b f(x)dx = \frac{h}{2}[f(x_0) + f(x_1)] - \frac{h^3}{12}f''(\xi)$$

Esto recibe el nombre de regla trapezoidal porque cuando f es una función con valores positivos, $\int_a^b f(x)dx$ se aproxima mediante el área de un trapecio, como se muestra en la figura 5.1

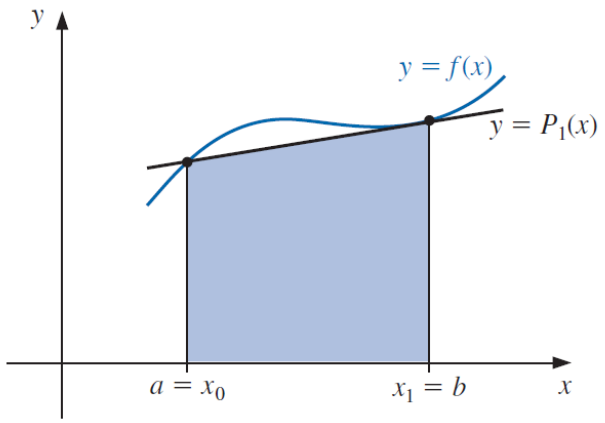

Figura 5.1: Regla del Trapecio.

El término de error para la regla trapezoidal implica f'', por lo que la regla da el resultado exacto cuando se aplica a cualquier función cuya segunda derivada es idénticamente cero, es decir, cualquier polinomio de grado uno o menos.

1.2. Regla de Simpson

La regla de Simpson resulta de la integración sobre $[a, b]$ del segundo polinomio de Lagrange con nodos igualmente espaciados $x_0 = a$, $x_2 = b$ y $x_1 = a + h$, en donde $h = (b - a)/2$ (véase la figura 5.2).

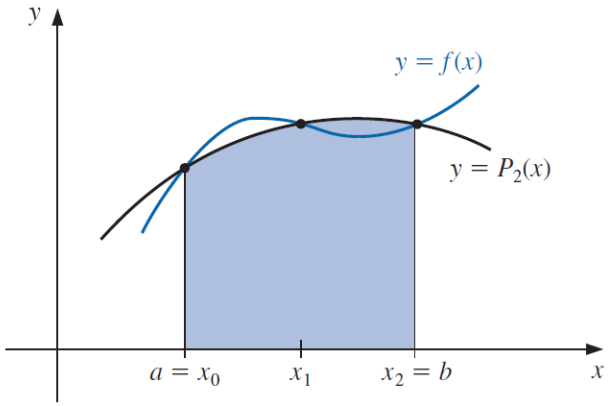

Figura 5.2: Regla de Simpson.

Por lo tanto,

$$\int_a^b f(x)dx = \int_{x_0}^{x_2} \left[\frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)} f(x_0) + \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)} f(x_1) + \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)} f(x_2) \right] dx + \int_{x_0}^{x_2} \frac{(x-x_0)(x-x_1)(x-x_2)}{6} f^{(3)}(\xi(x))dx$$

Al deducir la regla de Simpson de esta forma, sin embargo, da un solo término de error $O(h^4)$ relacionado con $f^{(3)}$. Al aproximar el problema de otra forma, se puede derivar otro término de orden superior relacionado con $f^{(4)}$.

Para ilustrar este método alternativo, suponga que f se expande en el tercer polinomio de Taylor alrededor de x_1. Entonces, para cada x en $[x_0, x_2]$, existe un número $\xi(x)$ en (x_0, x_2) con

$$f(x) = f(x_1) + f'(x_1)(x - x_1) + \frac{f''(x_1)}{2}(x - x_1)^2 + \frac{f'''(x_1)}{6}(x - x_1)^3 + \frac{f^{(4)}(\xi(x))}{24}(x - x_1)^4$$

y

$$\int_{x_0}^{x_2} f(x)dx = \left[f(x_1)(x - x_1) + \frac{f'(x_1)}{2}(x - x_1)^2 \frac{f''(x_1)}{6}(x - x_1)^3 + \frac{f'''(x_1)}{24}(x - x_1)^4 \right]_{x_0}^{x_2} + \frac{1}{24} \int_{x_0}^{x_2} f^{(4)}(\xi(x))(x - x_1)^4 dx \tag{5.2}$$

Puesto que $(x - x_1)^4$ nunca es negativo en $[x_0, x_2]$, el teorema de valor promedio ponderado para las integrales implica que

$$\frac{1}{24} \int_{x_0}^{x_2} f^{(4)}(\xi(x))(x - x_1)^4 dx = \frac{f^{(4)}(\xi_1)}{24} \int_{x_0}^{x_2} (x - x_1)^4 dx = \frac{f^{(4)}(\xi_1)}{120}(x - x_1)^5 \Big]_{x_0}^{x_2}$$

para algún número ξ_1 en (x_0, x_2).

Sin embargo, $h = x_2 - x_1 = x_1 - x_0$, por lo que

$$(x_2 - x_1)^2 - (x_0 - x_1)^2 = (x_2 - x_1)^4 - (x_0 - x_1)^4 = 0$$

mientras

$$(x_2 - x_1)^3 - (x_0 - x_1)^3 = 2h^3 \text{ y } (x_2 - x_1)^5 - (x_0 - x_1)^5 = 2h^5$$

Por consiguiente, la ecuacion 5.2 se puede escribir como

$$\int_{x_0}^{x_2} f(x)dx = 2hf(x_1) + \frac{h^3}{3} f''(x_1) + \frac{f^{(4)}(\xi_1)}{60} h^5$$

Ahora, si reemplazamos $f''(x_1)$ por medio de la aproximación determinada en la ecuación 2.3 de la sección 1 del capítulo 2, tenemos

$$\int_{x_0}^{x_2} f(x)dx = 2hf(x_1) + \frac{h^3}{3} \left\{ \frac{1}{h^2}[f(x_0) - 2f(x_1) + f(x_2)] - \frac{h^2}{12} f^{(4)}(\xi_2) \right\} + \frac{f^{(4)}(\xi_1)}{60} h^5 = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{12} \left[\frac{1}{3} f^{(4)}(\xi_2) - \frac{1}{5} f^{(4)}(\xi_1) \right]$$

Con métodos alternos se puede mostar que los valores ξ_1 y ξ_2 en esta expresión se pueden reemplazar mediante un valor común ξ en (x_0, x_2). Esto da la regla de Simpson.

Regla de Simpson:

$$\int_{x_0}^{x_2} f(x)dx = \frac{h}{3}[f(x_0) + 4f(x_1) + f(x_2)] - \frac{h^5}{90}f^{(4)}(\xi)$$

El término de error en la regla de Simpson implica la cuarta derivada de f, por lo que da resultados exactos cuando se aplica a cualquier polinomio de grado tres o menos.

1.3. Precisión de medición

Las derivadas estándar de las fórmulas de error de cuadratura están basadas al determinar la clase de polinomios para los que estas fórmulas producen resultados exactos. La siguiente definición se utiliza para facilitar el análisis de esta derivada.

Definición 1.1. El grado de precisión, o precisión, de una fórmula de cuadratura es el mayor entero positivo n, de tal forma que la fórmula es exacta para x^k, para cada $k = 0, 1..., n$.

Esta definición implica que las reglas trapezoidal y de Simpson tienen grados de precisión uno y tres, respectivamente.

2. Integración numérica compuesta

Teorema 2.1. Si $f \in C^4[a, b]$, n es par, $h = (b-a)/n$, y $x_j = a + jh$, para cada $j = 0, 1, ..., n$. Existe $\mu \in (a, b)$ para los que la regla compuesta de Simpson para n subintervalos se puede reescribir con su término de error como

$$\int_a^b f(x)dx = \frac{h}{3}\left[f(a) + 2\sum_{j=1}^{(n/2)-1} f(x_{2j}) + 4\sum_{j=1}^{n/2} f(x_{2j}-1) + f(b)\right] - \frac{b-a}{180}h^4 f^{(4)}(\mu)$$

Observe que el término de error para la regla compuesta de Simpson es $O(h^4)$, mientras que era $O(h^5)$ para la regla estándar de Simpson. Sin embargo, estos índices no son comparables ya que para la regla estándar de Simpson, tenemos h fija en $h = (b-a)/2$, pero para la regla compuesta de Simpson, tenemos $h = (b-a)/n$, para n un entero par. Esto nos permite reducir considerablemente el valor de h.

El enfoque de subdivisión se puede aplicar a cualquiera de las fórmulas de Newton-Cotes. Las extensiones de las reglas trapezoidal y de punto medio se dan sin prueba. La regla trapezoidal sólo requiere un intervalo para cada aplicación, por lo que el entero n puede ser tanto par como impar.

Teorema 2.2. Sean $f \in C^2[a, b]$, $h = (b-a)/n$, y $x_j = a+jh$, para cada $j = 0, 1, ..., n$. Existe $\mu \in (a, b)$ para el que la regla compuesta trapezoidal para n subintervalos se puede reescribir con este término de error como

$$\int_a^b f(x)dx = \frac{h}{2}\left[f(a) + 2\sum_{j=1}^{n-1} f(x_j) + f(b)\right] - \frac{b-a}{12}h^2 f''(\mu)$$

Ejercicios

(1) Calcular $\int_0^1 \frac{2}{x^2+4}$ utilizando:

- La regla del trapecio de Simpson simples
- La regla compuesta del trapecio con $h = 1/2^k$, $k = 2, ..., 6$
- La regla compuesta de Simpson con $h = 1/2^k$, $k = 2, ..., 6$

(2) Aproxime $\int_0^1 x^3 dx$ usando la regla compuesta del trapecio eligiendo el valor de h adecuado para que el error sea menor que $\delta = 10^-6$.

(3) Sea $f : [0, 1] \to \mathbb{R}$, se considera la fórmula de cuadratura

$$Q[f] = af(1/4) + bf(1/2) + cf(3/4)$$

para aproximar $\int_0^1 f(x)dx$, se pide:

- determinar los valores constantes a, b, y c de modo que la fórmula de cuadratura $Q[f]$ tenga grado de exactitud máximo.
- Aplicar la fórmula obtenida para aproximar

$$\int_0^1 \frac{x^2}{\sqrt{x^2+12}}dx$$

(4) De una fuerza $F(x)$ que depende de la posición x, contamos con las siguientes mediciones discretas a intervalor de 5 m.

$$x(m) : 0 \quad 5 \quad 10 \quad 15 \quad 20$$

$$F(N) : 0 \quad 1{,}53 \quad 9{,}51 \quad 8{,}70 \quad 2{,}81$$

- estimar el valor de F para un valor de $x = 6m$ utilizando un polinomio interpolador cuadrátivo

- estimar el trabajo realizado por la fuerza

$$W = \int_0^{20} F(x)dx$$

utilizando la regla de los trapecios compuestas y la regla de Simpson compuesta.

Capítulo 6

Raíces de Ecuaciones y Sistemas No Lineales

1. Soluciones de las ecuaciones en una variable

1.1. El método de bisección

Este proceso implica encontrar una raíz, o solución, para una ecuación de la forma $f(x) = 0$, para una función f dada. Una raíz de esta ecuación también recibe el nombre de cero de la función f.

La primera técnica, basada en el teorema del valor intermedio, recibe el nombre de bisección, o método de búsqueda binaria.

Suponga que f es una función continua definida dentro del intervalo $[a, b]$ con $f(a)$ y $f(b)$ de signos opuestos. El teorema del valor intermedio implica que existe un número p en (a, b) con $f(p) = 0$. A pesar de que el procedimiento operará cuando haya más de una raíz en el intervalo (a, b), para simplicidad, nosotros asumimos que la raíz en este intervalo es única. El método realiza repetidamente una reducción a la mitad (o bisección) de los subintervalos de $[a, b]$ y, en cada paso, localizar la mitad que contiene p.

Para comenzar, sea $a_1 = a$ y $b_1 = b$ y sea p_1 es el punto medio de $[a, b]$, es decir,

$$p_1 = a_1 + \frac{b_1 - a_1}{2} = \frac{a_1 + b_1}{2}$$

- Si $f(p_1) = 0$, entonces $p = p_1$ y terminamos.
- Si $f(p_1) \neq 0$, entonces $f(p_1)$ tiene el mismo signo que ya sea $f(a_1)$ o $f(b_1)$.
  - Si $f(p_1)$ y $f(a_1)$ tienen el mismo signo, $p \in (p_1, b_1)$. Sea $a_2 = p_1$ y $b_2 = b_1$
  - Si $f(p_1)$ y $f(a_1)$ tienen signos opuestos, $p \in (a_1, p_1)$. Sea $a_2 = a_1$ y $b_2 = p_1$.

Entonces, volvemos a aplicar el proceso al intervalo $[a_2, b_2]$.

Se pueden aplicar otros procedimientos de parada, por ejemplo, podemos seleccionar una tolerancia $\epsilon > 0$ y generar $p_1, ..., p_N$ hasta que se cumpla una de las siguientes condiciones:

$$|f(p_N)| < \epsilon \tag{6.1}$$

$$\frac{|p_N - p_{N-1}|}{|p_N|} < \epsilon, \quad p_N \neq 0 \tag{6.2}$$

$$|p_N - p_{N-1}| < \epsilon \tag{6.3}$$

El método de bisección, a pesar de que está conceptualmente claro, tiene desventajas significativas. Su velocidad de convergencia es más lenta y se podría descartar inadvertidamente una buena aproximación intermedia. Sin embargo, el método tiene la importante propiedad de que siempre converge a una solución y por esta razón con frecuencia se utiliza como iniciador para los métodos más eficientes que veremos más adelante en este capítulo.

Teorema 1.1. Suponga que $f \in C[a, b]$ y $f(a) \cdot f(b) < 0$. El método de bisección genera una sucesión $p_n{}_{n=1}^{\infty}$ que se aproxima a cero p de f con

$$|p_n - p| \leq \frac{b - a}{2^n}, \text{ cuando } n \geq 1$$

1.2. Iteración de punto fijo

Un punto fijo para una función es un número en el que el valor de la función no cambia cuando se aplica la función.

Definición 1.1. El número p es un punto fijo para la función dada g si $g(p) = p$.

Teorema 1.2.
- Si $g \in C[a, b]$ y $g(x) \in [a, b]$ para todas $x \in [a, b]$, entonces g tiene por lo menos un punto fijo en $[a, b]$.

- Si, además, $g'(x)$ existe en (a, b) y hay una constante positiva $k < 1$ con

$$|g'(x)| \leq k, \forall x \in (a, b)$$

entonces, existe exactamente un punto fijo en $[a, b]$. (Figura 6.1)

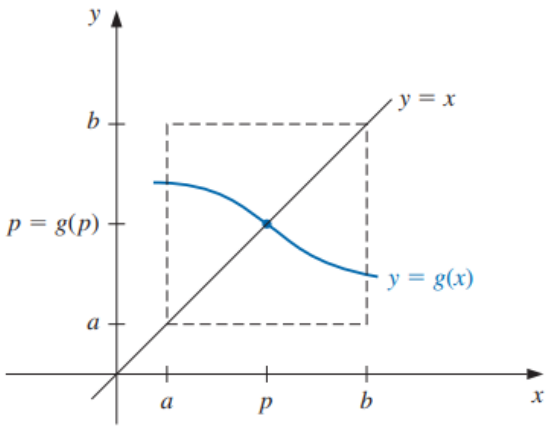

Figura 6.1: Teorema del Punto Fijo

Demostración. Si $g(a) = a$ o $g(b) = b$, entonces g tiene un punto fijo en un extremo. De lo contrario, entonces $g(a) > a$ y $g(b) < b$. La función $h(x) = g(x) - x$ es continua en $[a, b]$, con

$$h(a) = g(a) - a > 0 \quad \text{y} \quad h(b) = g(b) - b > 0$$

El teorema de valor intermedio implica que existe $p \in (a, b)$ para la cual $h(p) = 0$. Este número p es un punto fijo para g porque

$$0 = h(p) = g(p) - p \quad \text{implica que} \quad g(p) = p$$

Suponga, además, que $|g'(x)| \leq k < 1$ y que p y q son puntos fijos en $[a, b]$. Si $p \neq q$, entonces el teorema de valor medio implica que existe un número ξ entre p y q y por lo tanto en $[a, b]$ con

$$\frac{g(p) - g(p)}{p - q} = g'(\xi)$$

Por lo tanto

$$|p-q| = |g(p)-g(q)| = |g'(\xi)||p-q| \leq k|p-q| < |p-q$$

lo cual es una contradicción. Esta contradicción debe provenir de la única suposición $p \neq q$. Por lo tanto, $p = q$ y el punto fijo en $[a, b]$ es único. □

Iteración de punto fijo

Para aproximar el punto fijo de una función g, elegimos una aproximación inicial p_0 y generamos la sucesión $p_n{}_{n=0}^{\infty}$ al permitir $p_n = g(p_{n-1})$, para cada $n \geq 1$. Si la sucesión converge a p y g es continua, entonces

$$p = \lim_{n\to\infty} p_n = \lim_{n\to\infty} g(p_{n-1}) = g\left(\lim_{n\to\infty} p_{n-1}\right) = g(p)$$

y se obtiene una solución para $x = g(x)$. Esta técnica recibe el nombre de punto fijo o iteración funcional (Figura 6.2).

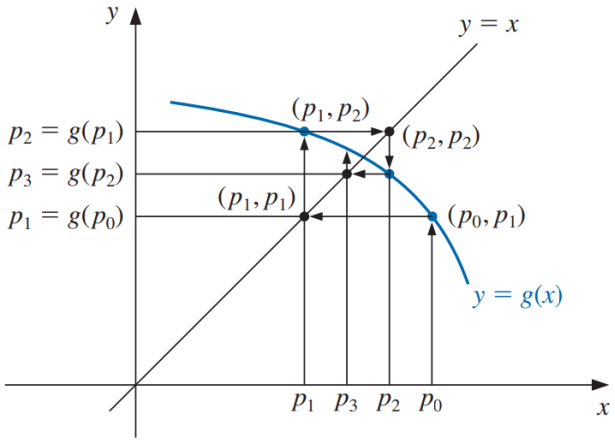

Figura 6.2: Iteración del Punto Fijo

Teorema 1.3. (Teorema de punto fijo) Sea $g \in C[a, b]$ tal que $g(x) \in [a, b]$ para todas las $x \in [a, b]$. Suponga, además, que existe g' en (a, b) y que existe una constante $0 < k < 1$ con

$$|g'(x)| \leq k, \text{ para todas } x \in (a, b)$$

Entonces, para cualquier número p_0 en $[a, b]$, la sucesión definida por

$$p_n = g(p_{n-1}), \quad n \geq 1$$

converge al único punto fijo p en $[a, b]$.

Demostración. El teorema 1.3 implica que existe un único punto p en $[a, b]$ con $g(p) = p$. Ya que g mapea $[a, b]$ en sí mismo, la sucesión $p_n{}_{n=0}^{\infty}$ se define para todas las $n \geq 0$ y $p_n \in [a, b]$ para todas las n. Al utilizar el hecho de que $|g'(x)| \leq k|$ y el teorema de valor medio **??**, tenemos, para cada n,

$$|p_n-p| = |g(p_{n-1})-g(p)| = |g'(\xi_n)||p_{n-1}-p| \leq k|p_{n-1}-p|$$

donde $\xi_n \in (a,b)$. Al aplicar esta desigualdad de manera inductiva obtenemos

$$|p_n - p| \leq k|p_{n-1} - p| \leq k^2|p_{n-2} - p| \leq ... \leq k^n|p_0 - p| \tag{6.4}$$

Ya que $0 < k < 1$, tenemos que $\lim_{n\to\infty} k^n = 0$ y

$$\lim_{n\to\infty} |p_n - p| \leq \lim_{n\to\infty} k^n|p_0 - p| = 0$$

Por lo tanto, $p_n{}_{n=0}^{\infty}$ converge a p. □

Observación 1.1. Si g satisface las hipótesis del teorema 1.3, entonces las octas del error relacionado con el uso de p_n para aproximar p, están dadas por

$$|p_n - p| \leq k^n \max\{p_0 - a, b - p_0\} \tag{6.5}$$

y

$$|p_n - p| \leq \frac{k^n}{1-k}|p_1 - p_0|, \text{ para toda } n \geq 1 \tag{6.6}$$

1.3. Método de Newton

El método de Newton (o de Newton-Raphson) es uno de los métodos numéricos más poderosos y reconocidos para resolver un problema de encontrar la raíz. Existen muchas formas de presentar el método de Newton.

Supona que $f \in C^2[a,b]$. Si $p_0 \in [a,b]$ es una aproximación para p, de tal forma que $f'(p_0 \neq 0)$ y $|p - p_0|$ es "pequeño". Considere que el primer polinomio de Taylor para $f(x)$ expandido alrededor de p_0 y evaluado en $x = p$:

$$f(p) = f(p_0) + (p - p_0)f'(p_0) + \frac{(p-p_0)^2}{2}f''(\xi(p))$$

donde $\xi(p)$ se encuentra entre p y p_0. Puesto que $f(p) = 0$, esta ecuación nos da

$$0 = f(p_0) + (p - p_0)f'(p_0) + \frac{(p-p_0)^2}{2}f''(\xi(p))$$

El método de Newton se deriva al suponer que como $|p - p_0|$ es pequeño, el término relacionado con $(p - p_0)^2$ es mucho más pequeño, entonces

$$0 \approx f(p_0) + (p - p_0)f'(p_0)$$

Al resolver para p obtenemos

$$p \approx p_0 - \frac{f(p_0)}{f'(p_0)} \equiv p_1$$

Esto constituye la base para el método de Newton, que empieza con una aproximación inicial p_0 y genera la sucesión $p_n{}_{n=0}^{\infty}$, mediante

$$p_n = p_{n-1} - \frac{f(p_{n-1})}{f'(p_{n-1})}, \text{ para } n \geq 1 \tag{6.7}$$

La figura 6.3 ilustra cómo se obtienen las aproximaciones usando tangentes sucesivas.

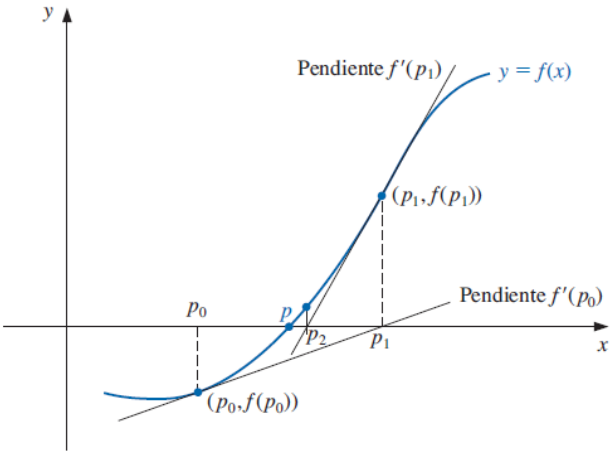

Figura 6.3: Método de Newton

Convergencia con el método de Newton

La derivación del método de Newton por medio de la serie de Taylor al inicio de la sección señala la importancia de una aproximación inicial precisa. La suposición crucial es que el término relacionado con $(p - p_0)^2$ es, en comparación con $|p - p_0|$, tan pequeño que se puede eliminar. Claramente esto será falso a menos que p_0 sea una buena aproximación para p. Si p_0 no está suficientemente cerca de la raíz real, existen pocas razones para sospechar que el método de Newton convergerá a la raíz. Sin embargo, en algunos casos, incluso las malas aproximaciones iniciales producirán convergencia.

Teorema 1.4. Teorema de Ostrowski. Sea $f \in C^2[a,b]$. Si $p \in (a,b)$ es tal que $f(p) = 0$ y $f'(p) \neq 0$, entonces existe una $\delta > 0$ tal que el método de Newton genera una sucesión $p_n{}_{n=0}^{\infty}$ que converge a p para cualquier aproximación inicial $p_0 \in [p - \delta, p + \delta]$.

El método de la secante

El método de Newton es una técnica en extremo poderosa, pero tiene una debilidad importante: la necesidad de conocer el valor de la derivada de f en

cada aproximación. Con frecuencia, $f'(x)$ es mucho más dificil y necesita más operaciones aritméticas para calcular $f(x)$.

Para evitar el problema de la evaluación de la derivada en el método de Newton, presentamos una ligera variación. Por definición,

$$f'(p_{n-1}) = \lim_{x \to p_{n-1}} \frac{f(x) - f(p_{n-1})}{x - p_{n-1}}$$

Si p_{n-2} está cerca de p_{n-1}, Entonces

$$f'(p_{n-1}) \approx \frac{f(p_{n-2}) - f(p_{n-1})}{p_{n-2} - p_{n-1}} = \frac{f(p_{n-1}) - f(p_{n-2})}{p_{n-1} - p_{n-2}}$$

Usando esta aproximación para $f'(p_{n-1})$ en la fórmula de Newton obtenemos

$$p_n = p_{n-1} - \frac{f(p_{n-1})(p_{n-1} - p_{n-2})}{f(p_{n-1}) - f(p_{n-2})} \tag{6.8}$$

Esta técnica recibe el nombre de método de la secante y se ilustra en la figura 6.4.

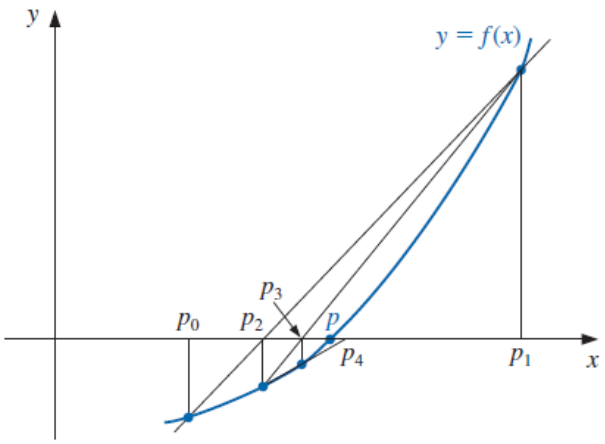

Figura 6.4: Método de la Secanta

El método de posición falsa

El método de posición falsa (también llamado Regula Falsi) genera aproximaciones de la misma manera que el método de la secante, pero incluye una prueba para garantizar que la raíz simepre se agrupa entre iteraciones sucesivas.

En primer lugar, seleccionamos las aproximaciones iniciales p_0 y p_1 con $f(p_0) \cdot f(p_1) < 0$. La aproximación p_2 se selecciona de la misma forma que en el método de la secante como la intersección en x de la recta que une $(p_0, f(p_0))$ y $(p_1, f(p_1))$. Para decidir cuál línea secante se usa para calcular p_3, considere $f(p_2) \cdot f(p_1)$ o, más concretamente, $sgnf(p_2) \cdot sgnf(p_1)$.

- Si $sgnf(p_2) \cdot sgnf(p_1) < 0$, entonces p_1 y p_2 agrupan una raíz. Seleccione p_3 como la intersección en x de la recta que une $(p_1, f(p_1))$ y $(p_2, f(p_2))$.
- Si no, seleccionamos p_3 como la intersección en x de la recta que une $(p_0, f(p_0))$ y $(p_2, f(p_2))$ y, a continuación intercambia los índices en p_0 y p_1.

De manera similar, una vez se encuentra p_3, el signo de $f(p_3) \cdot f(p_2)$ determina si usamos p_2 y p_3 o p_3 y p_1 para calcular p_4. En el último caso, se vuelve a etiquetar p_2 y p_1. Reetiquetar garantiza que la raíz se agrupa entre iteraciones sucesivas.

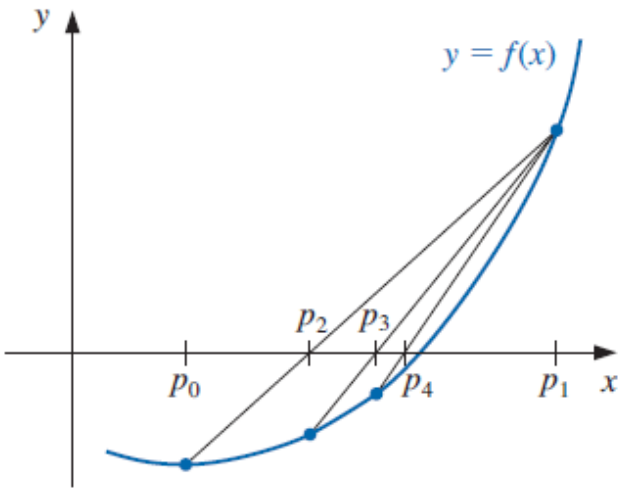

Figura 6.5: Método de la Posición Falsa

1.4. Análisis de error para métodos iterativos

En esta sección inverstigamos el orden de convergencia de esquemas de iteración funcional y, con el propósito de obtener convergencia rápida, redescubrimo el método de Newton. También consideramos consideramos formas para acelerar la convergencia del método de Newton en circunstancias especiales. Primero, sin embargo, necesitamos un nuevo procedimiento para medir qué tan rápido converge una sucesión.

Orden de convergencia

Definición 1.2. Suponga $p_n{}_{n=0}^{\infty}$ es una sucesión que converge a p, con $p_n \neq p$ para todas las n. Si existen constantes positivas λ y α con

$$\lim_{n \to \infty} \frac{|p_{n+1} - p|}{|p_n - p|^{\alpha}} = \lambda$$

Entonces $p_n{}_{n=0}^{\infty}$ converge a p de orden α, con constante de error asintótica λ

Se dice que una técnica iterativa de la forma $p_n = g(p_{n-1})$ es de orden α si la sucesión $p_n{}_{n=0}^{\infty}$ converge a la solución $p = g(p)$ de orden α.

En general, una sucesión con un alto orden converge más rápidamente uqe una sucesión con un orden más bajo. La constante asintótica afecta la velocidad de concergencia pero no el grado del orden. Se presta atención especial a dos casos:

1. Si $\alpha = 1$ (y $\lambda < 1$), la sucesión es linealmente convergente.
2. Si $\alpha = 2$, la sucesión es cuadráticamente convergente.

Teorema 1.5. Sea $g \in [a, b]$ tal que $g(x) \in [a, b]$ para todas las $x \in [a, b]$. Suponga además que g' es continua en (a, b) y que existe una constante positiva $k < 1$ con

$$|g'(x)| \leq k, \text{ para toda } x \in (a, b)$$

Si $g'(p) \neq 0$, entonces para cualquier número $np_0 \neq p$ en $[a, b]$, la sucesión

$$p_n = g(p_{n-1}), \text{ para } n \geq 1,$$

Converge sólo linealmente para el único punto fijo p en $[a, b]$.

Demostración. Sabemos que, a partir del teorema del punto fijo, la sucesión converge a p. Puesto que existe g' en (a, b), podemos aplicar el teorema del valor medio para g para demostrar que para cualquier n,

$$p_{n+1} - p = g(p_n) - g(p) = g'(\xi_n)(p_n - p)$$

donde ξ_n está entre p_n y p. Ya que $p_n{}_{n=0}^{\infty}$ converge a p, también tenemos que $\xi_n{}_{n=0}^{\infty}$ converge a p. Puesto que g_0 es continua en (a, b), tenemos

$$\lim_{n\to\infty} g'(\xi_n) = g'(p)$$

Por lo tanto

$$\lim_{n\to\infty} \frac{p_{n+1} - p}{p_n - p} = \lim_{n\to\infty} g'(\xi_n) = g'(p)$$

y

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|} = |g'(p)|$$

De este modo, si $|g'(p)| \neq 0$, la iteración de punto fijo muestra convergencia lineal con error asíntótico constante $|g'(p)|$. □

Teorema 1.6. Sea p una solución de la ecuación $x = g(x)$. Suponga que $g'(p) = 0$ y que g'' es continua con $|g''(x)| < M$ en un intervalo abierto I que contiene a p. Entonces existe $\delta > 0$ talo que para $p_0 \in [p - \delta, p + \delta]$, la sucesión definida por $p_n = g(p_{n-1})$, cuando $n \geq 1$, converge, por lo menos cuadráticamente a p. Además, con valores suficientemente grandes de n,

$$|p_{n+1} - p| < \frac{M}{2}|p_n - p|^2$$

Demostración. Seleccione k en $(0, 1)$ y $\delta > 0$ tal que el intervalo $[p - \delta, p + \delta]$, contenido en I, tenemos $|g'(x) \leq k$ y g'' continua.l Puesto que $|g'(x) \leq k < 1|$. Al expandir $g(x)$ en un polinomio lineal de Taylor, para $x \in [p - \delta, p + \delta]$ obtenemos

$$g(x) = g(p) + g'(p)(x - p) + \frac{g''(\xi)}{2}(x - p)^2$$

donde ξ se encuentra entre x y p. Las hipótesis $g(p) = p$ y $g'(p) = 0$ implican que

$$g(x) = p + \frac{g''(\xi)}{2}(x - p)^2$$

En especial, cuando $x = p_n$,

$$p_{n+1} = g(p_n) = p + \frac{g''(\xi_n)}{2}(p_n - p)^2$$

con ξ_n entre p_n y p. Por lo tanto,

$$p_{n+1} - p = \frac{g''(\xi_n)}{2}(p_n - p)^2$$

Puesto que $|g'(x)| \leq k < 1$ en $[p - \delta, p + \delta]$ en sí mismo, por el teorema de putno fijo se sigue que $p_n{}_{n=0}^{\infty}$ converge a p. Pero como ξ_n se encuentra entre p y p_n para cada n, entonces $\xi_n{}_{n=0}^{\infty}$ también converge a p y

$$\lim_{n\to\infty} \frac{|p_{n+1} - p|}{|p_n - p|^2} = \frac{|g''(p)|}{2}$$

Este reseultado implica que la sucesión $p_n{}_{n=0}^{\infty}$ es cuadráticamente convergente si $g''(p) \neq 0$ y de convergencia de orden superior si $g''(p) = 0$.

Puesto que g'' es continua y está estrictamente acotada por M en el intervalo $[p-\delta, p+\delta]$, esto también implica que, para los valores suficientemente grandes de n,

$$|p_{n+1} - p| < \frac{M}{2}|p_n - p|^2$$

□

Raíces múltiples

Definición 1.3. Una solución p de $f(x) = 0$ es un cero de multiplicidad m de f si para $x \neq p$, podemos escribir $f(x) = (x-p)^m g(x)$, donde $\lim_{x\to p} g(x) \neq 0$.

Teorema 1.7. La función $f \in C^1[a,b]$ tiene un cero simple en p en (a,b) si y sólo si $f(p) = 0$, pero $f'(p) \neq 0$.

Teorema 1.8. La función $f \in C^m[a,b]$ tiene un cero de multiplicidad m en p en (a,b) si y sólo si

$$0 = f(p) = f'(p) = f''(p) = ... = f^{(m-1)(p)}$$

pero

$$f^{(m)}(p) \neq 0$$

2. Soluciones numéricas de sistemas de ecuaciones no lineales

2.1. Puntos fijos para funciones de varias variables

Un sistema de ecuaciones no lineales tiene la forma

$$\begin{cases} f_1 = (x_1, x_2, ..., x_n) = 0 \\ f_2 = (x_1, x_2, ..., x_n) = 0 \\ \quad \vdots \\ f_2 = (x_1, x_2, ..., x_n) = 0 \end{cases} \tag{6.9}$$

donde cada función f_i se puede pensar como un mapeo de un vector $x = (x_1, x_2, ..., x_n)^t$ del espacio n dimensional $\mathbb{R}^n$ en la recta real $\mathbb{R}$

Este sistema de n ecuaciones no lineales en n variables también se puede representar al definir una función F de mapeo $\mathbb{R}^n$ en $\mathbb{R}^n$.

Si se utiliza notacion vectorial para representar las variables $x_a, x_2, ...x_n$, entonces el sistema asume la forma

$$F(x) = 0 \tag{6.10}$$

Las funciones $f_1, f_2, ..., f_n$ reciben el nombre de funciones coordenadas de F.

Definición 2.1. Sea f una función definida en un conjunto $D \subset \mathbb{R}^n$ en $\mathbb{R}$ y rango en $\mathbb{R}$. Se dice que la función f tiene límite L en x_0, escrito

$$\lim_{x\to x_0} f(x) = L$$

si, dado cualquier número $\epsilon > 0$, existe un número $\delta > 0$ con

$$|f(x) - L| < \epsilon$$

siempre que $x \in D$, y

$$0 < \|x - x_0\| < \delta$$

Definición 2.2. Sea f una función del conjunto $D \subset \mathbb{R}^n$. La función f es continua en $x_0 \in D$ siempre que exista $\lim_{x\to x_0} f(x_0)$ y

$$\lim_{x\to x_0} f(x) = f(x_0)$$

Además, f es continua en un conjunto D si f es continua en cada punto de D. Este concepto se expresa al escribir $f \in C(D)$

Definición 2.3. Sea F una función desde $D \subset \mathbb{R}^n$ a $\mathbb{R}^n$ de la forma

$$F(x) = (f_1(x), f_2(x), ..., f_n(x))^t$$

donde fi es un mapeo de $\mathbb{R}^n$ hasta $\mathbb{R}$ para cada i. Definimos

$$\lim_{x-x_0} F(x) = L = (L_1, L_2, ...L_n)^t$$

si y sólo si $\lim_{x\to x_0} f_i(x) = L_i$ para cada $i = 1, 2, ..., n$.

Teorema 2.1. Sea f una función de $D \subset \mathbb{R}^n$ a $\mathbb{R}$ y $x_0 \in D$. Suponga que existen todas las derivadas parciales de f y las constantes $\delta > 0$ y $K > 0$, de tal forma que siempre que $\|x - x_0\| < \delta$ y $x \in D$, tenemos

$$\left|\frac{\partial f(x)}{\partial x_j}\right| \leq K, \text{ para cada } j = 1, 2, ..., n$$

Entonces f es continua en x_0.

Puntos fijos en $\mathbb{R}^n$

Definición 2.4. Una función G desde $D \subset \mathbb{R}^n$ hasta $\mathbb{R}^n$ tiene un punto fijo en $p \in D$ si $G(p) = p$.

Teorema 2.2. Sea $D = (x_1, x_2, ..., x_n)^t | a_i \leq x_i \leq b_i, \text{ para cada } i = 1, 2, ..., n$ para algún conjunto de constantes $a_1, a_2, ..., a_n$ y $b_1, b_2, ..., b_n$. Suponga que G es una función continua

en $D \subset \mathbb{R}^n$ a $\mathbb{R}$ con la propiedad de que $G(x) \in D$, siempre que $x \in D$. Entonces G tiene un punto fijo en D.

Además, suponga que todas las funciones componentes de G tienen derivadas parciales continuas y que existe una constante $K < 1$ con

$$\left|\frac{\partial g_i(x)}{\partial x_j}\right| \leq G(x^{(k-1)}), \quad \text{siempre que } x \in D$$

para cada $j = 1, 2, ..., n$ y cada funcikón componente g_i. Entonces, la sucesión de punto fijo $x^{(k)}{}_{k=0}^{\infty}$ definida por $x^{(0)}$ seleccionada arbitrariamente en D y generada por medio de

$$x^{(k)} = G(x^{(k-1)}) \quad \text{para cada } k \geq 1$$

converge al único punto fijo $p \in D$ y

$$\|x^{(k)} - p\|_\infty \leq \frac{K^k}{1-K}\|x^{(1)} - x^{(0)}\|_\infty \tag{6.11}$$

2.2. Método de Newton

Para construir el algoritmo que conduce a un método de punto fijo adecuado en el caso unidemnsional, encontramos una función ϕ con la proopiedad de que

$$g(x) = x - \phi(x)f(x)$$

da convergencia cuadrática para el punto fijo p de la función g. A partir de esta condición el método de Newton evolucionó al seleccionar $\phi(x) = 1/f'(x)$ suponiendo que $f'(x) \neq 0$.

Un enfoque similar en el caso n-dimensional implica una matriz

$$A(x) = \begin{bmatrix} a_{11}(x) & a_{12}(x) & \dots & a_{1n}(x) \\ a_{21}(x) & a_{22}(x) & \dots & a_{2n}(x) \\ \vdots & \vdots & & \vdots \\ a_{n1}(x) & a_{n2}(x) & \dots & a_{nn}(x) \end{bmatrix} \tag{6.12}$$

donde cada una de las entradas $a_{ij}(x)$ es una función de $\mathbb{R}^n$ a $\mathbb{R}$. Esto requiere encontrar $A(x)$ de tal forma que

$$G(x) = x - A(x)^{-1}F(x)$$

da convergencia cuadrática para la solución de $F(x) = 0$, suponiendo que $A(x)$ es no singular en el punto fijo p de G.

Teorema 2.3. Si p es la solución de $G(x) = x$. Suponga que existe un número $\delta > 0$ con las propiedades:

- $\partial g_I/\partial x_j$ es continua en $N_\delta = x|\|x - p\| < \delta$, para cada $i = 1, 2, ..., n$ y $j = 1, 2, ..., n$
- $\partial^2 g_i(x)/(\partial x_j \partial x_k)$ es continua y $|\partial^2 g_i(x)/(\partial x_j \partial x_k)| \leq M$ para algunas constantes M, siempre que $x \in N_\delta$, para cada $i = 1, 2, ..., n$, $j = 1, 2, ..., n$ y $k = 1, 2, ..., n$.
- $\partial g_i(p)/\partial x_k = 0$, para cada $i = 1, 2, ..., n$ y $k = 1, 2, ..., n$.

Entonces, un número $\hat{\delta} \leq \delta$ existe de tal forma que la sucesión generada por $x^{(k)} = G(x^{(k-1)})$ converge de forma cuadrática en p para cualquier selección de $x^{(0)}$, siempre y cuando $\|x^{(0)} - p\| < \hat{\delta}$. Además,

$$\|x^{(k)} - p\|_\infty \leq \frac{n^2 M}{2}\|x^{(k-1)} - p\|_\infty^2, \text{ para cada } k \geq 1$$

La matriz jacobiana

Defina la matriz $J(x)$ mediante

$$J(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(x) & \frac{\partial f_1}{\partial x_2}(x) & \dots & \frac{\partial f_1}{\partial x_n}(x) \\ \frac{\partial f_2}{\partial x_1}(x) & \frac{\partial f_2}{\partial x_2}(x) & \dots & \frac{\partial f_2}{\partial x_n}(x) \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}(x) & \frac{\partial f_n}{\partial x_2}(x) & \dots & \frac{\partial f_n}{\partial x_n}(x) \end{bmatrix} \tag{6.13}$$

Una seleccion adecuada para $A(x)$ es $A(x) = J(x)$. La función G se define mediante

$$G(x) = x - J(x)^{-1}F(x)$$

y el procedimiento de iteración de punto fijo evoluciona al seleccionar $x^{(0)}$ y generar, para $k \geq 1$,

$$x^{(k)} = G(x^{(k-1)}) = x^{(k-1)} - J(x^{(k-1)})^{-1}F(x^{(k-1)}) \tag{6.14}$$

Esto recibe el nombre de método de Newton para sistemas no lineales y en general se espera que proporcione convergencia cuadrática, siempre y cuando se conozca un valor inicial suficientemente preciso y que $J(p)^{-1}$ exista.

Una debilidad en el método de Newton surge de la necesidad de calcular e invertir la matriz $J(x)$ en cada paso. En la práctica, el cálculo explícito de $J(x)^{-1}$ se evita al realizar la operación en una forma de dos pasos. Primero se encuentra un vector y que satisface $J(x^{(k-1)})y = -F(x^{(k-1)})$. Entonces, la nueva aproximación, $x^{(k)}$, se obtiene sumando y a $x^{(k-1)}$.

Ejercicios

(1) Calcular el número de iteraciones necesarias para aproximar la raíz del polinomio $p(x) = x^3 + x - 4$ en el intervalo $[1, 4]$ con el método de la Bisección.

(2) Calcular el número de iteraciones necesarias para aproximar la raíz del polinomio $p(x) = x^3 - x - 1$ en el intervalo $[1, 2]$ con el método de la Bisección.

(3) Probar que la función $g(x) = \pi + 0{,}5sin(x/2)$ tiene un único punto fijo en el intervalo $[0, 2\pi]$.

(4) Probar que la función $q(x) = 2^{-x}$ tiene un único punto fijo en el intervalo $[1/3, 1]$.

(5) Encontrar una función de punto fijo para cada una de las siguientes funciones que converja a una raíz positiva:

a) $f(x) = 3x^2 - e^x$

b) $f(x) = x - cos(x)$

(6) Analiza la convergencia del método del punto fijo $x_{k+1} = g(x_k)$ para calcular los ceros $\alpha_1 = -1$ y $\alpha_2 = 2$ de la función $f(x) = x^2 - x - 2$ cuando se usan las funciones $g_1(x) = x^2 - 2$, $g_2(x) = \sqrt{2 + x}$, $g_3(x) = -\sqrt{2 + x}$, $g_4(x) = 1 + \frac{2}{x}$, $x \neq 0$.

Capítulo 7

Optimización Sin Restricciones y Mínimos Cuadrados

1. Optimización Sin Restricciones

El punto $x^* \in \mathbb{R}^n$ es un minimizador global de f, mientras que x^* es un minimizador local de f si existe $R > 0$ tal que

$$f(x^*) \leq f(x), \quad \forall x \in B(x^*; R)$$

A lo largo de esta sección, siempre asumiremos que $f \in C^1(\mathbb{R}^n)$. Denotaremos por

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), ..., \frac{\partial f}{\partial x_n}(x) \right)^T$$

el gradiente de f en un punto x. Si d es un vector distinto de cero en $\mathbb{R}^n$ entonces la derivada direccional de f respecto a d es

$$\frac{\partial f}{\partial d}(x) = \lim_{\alpha \to \infty} \frac{f(x + \alpha f) - f(x)}{\alpha}$$

y satisface

$$\frac{\partial f(x)}{\partial d} = \nabla f(x)^T d$$

Además, denotando por $(x, x + \alpha d)$ el segmento en $\mathbb{R}^n$ que une los puntos x y $x + \alpha d$, con $\alpha \in \mathbb{R}$, la expansión de Taylor asegura que existe un $\xi \in (x, x + \alpha d)$ tal que

$$f(x + \alpha d) - f(x) = \alpha \nabla f(\xi)^T d \tag{7.1}$$

Si $f \in C^2(\mathbb{R}^n)$, denotaremos por $H(x)$ (o $\nabla^2 f(x)$) la matriz Hessiana de f evaluada en un punto x, cuyos elementos son

$$h_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j}, \quad i, j = 1, ..., n$$

En tal caso, se puede demostrar que, si $d \neq 0$, existe la derivada direccional de segundo orden y tenemos

$$\frac{\partial^2 f}{\partial d^2}(x) = d^T H(x) d$$

Para un ξ adecuado en $(x, x + d)$, también tenemos

$$f(x + d) - f(x) = \nabla f(x)^T d + \frac{1}{2} d^T H(\xi) d$$

Propiedad 1.1. Sea $x^* \in \mathbb{R}^n$ un minimizador local de f y supongamos que $f \in C^1(B(x^*; R))$ para un $R > 0$ adecuado. Entonces $\nabla f(x^*) = 0$. Además, si $f \in C^2(B(x^*; R))$, entonces $h(x^*)$ es semidefinida positiva. Por el contrario, si $\nabla f(x^*) = 0$ y $H(x^*)$ es definida positiva, entonces x^* es un minimizador local de f en $B(x^*; R)$.

Un punto x^* tal que $\nabla f(x^*) = 0$ se dice que es un punto crítico de f. Esta condición es necesaria para que se cumpla la optimalidad. Sin embargo, esta condición también se vuelve suficiente si f es una función convexa en $\mathbb{R}^n$, es decir, tal que para todo $x, y \in \mathbb{R}^n$ y para cualquier $\alpha \in [0, 1]$,

$$f[\alpha x + (1 - \alpha) y] \leq \alpha f(x) + (1 - \alpha) f(y) \tag{7.2}$$

1.1. Métodos de descenso

Dado un vector inicial $x^{(0)} \in \mathbb{R}^n$, calcular para $k \geq 0$ hasta la convergencia

$$x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)} \tag{7.3}$$

donde $d(k)$ es una dirección elegida de manera apropiada y α_k es un parámetro positivo (llamado tamaño de paso) que mide el paso a lo largo de la dirección $d(k)$. Esta dirección $d(k)$ es una dirección de descenso si

$$\begin{aligned} d^{(k)} \nabla f(x^{(k)}) < 0 \quad \text{si} \quad \nabla f(x^{(k)}) \neq 0, \\ d^{(k)} = 0 \quad \text{si} \quad \nabla f(x^{(k)}) = 0. \end{aligned} \tag{7.4}$$

Un método de descenso es un método como 7.3, en el que los vectores $d(k)$ son direcciones de descenso.

Si $d_k \in \mathbb{R}^n$ es una dirección de descenso, entonces existe $\alpha_k > 0$, suficientemente pequeña, tal que

$$f(x^{(k)} + \alpha_k d^{(k)}) < f(x^{(k)}) \tag{7.5}$$

siempre que f sea diferenciable de manera continua. De hecho, tomando en 7.1 $\xi = x^{(k)} + \vartheta\alpha_k d^{(k)}$ con $\vartheta \in (0,1)$, y empleando la continuidad de $\bigtriangledown f$ obtenemos

$$f(x^{(k)}+\alpha_k d^{(k)}) - f(x^{(k)}) = \alpha_k \bigtriangledown f(x^{(k)})^T d^{(k)} + \epsilon \tag{7.6}$$

donde ϵ tiende a cero cuando α_k tiende a cero. Como consecuencia, si $\alpha_k > 0$ es suficientemente pequeño, el signo del lado izquierdo de 7.6 coincide con el signo de $\bigtriangledown f(x^{(k)})^T d^{(k)}$, de modo que 7.5 se satisface si $d^{(k)}$ es una dirección de descenso.

Diferentes elecciones de $d^{(k)}$ corresponden a diferentes métodos. En particular, recordamos los siguientes:

- El método de Newton, en el que
$$d^{(k)} = -H^{-1}(x^{(k)}) \bigtriangledown f(x^{(k)})$$
siempre que H sea definida positiva dentro de un vecindario suficientemente grande del punto x^*.
- Métodos de Newton inexactos, en lo que
$$d^{(k)} = -B_k^{-1} \bigtriangledown f(x^{(k)})$$
donde B_k es una aproximación adecuada de $H(x^{(k)})$.
- El método del gradiente o método del máximo descenso, que corresponde a establecer $d^{(k)} = -\bigtriangledown f(x^{(k)})$. Este método es, por lo tanto, un método de Newton inexacto, en el que $B_k = I$. También puede considerearse un método tipo gradiente, ya que ${d^{(k)}}^T \bigtriangledown f(x^{(k)}) = -\| \bigtriangledown f(x^{(k)})\|_2^2$
- El método del gradiente conjugado, para el cual
$$d^{(k)} = - \bigtriangledown f(x^{(k)}) + \beta_k d^{(k-1)}$$
donde β_k es un escalar que se selecciona adecuadamente de manera que las direcciones $d^{(k)}$ resulten ser mutuamente ortogonales con respecto a un producto escalar adecuado.

Un método para calcular α_k consiste en resolver el siguiente problema de minimización en una dimensión:

$$\text{enc. } \alpha \text{ tal que } \phi(\alpha) = f(x^{(k)} + \alpha d^{(k)}) \text{ sea minimizado} \tag{7.7}$$

En tal caso, tenemos el siguiente resultado.

Teorema 1.1. Considerando el método de descenso 7.3, si en el paso genérico k, el parámetro α_k se establece igual a la solución exacta de 7.7, entonces se cumple la siguiente propiedad de ortogonalidad:

$$\nabla f(x^{(k+1)})^T d^{(k)} = 0$$

1.2. Técnicas de búsqueda en línea

Los métodos con los que vamos a tratar en esta sección son técnicas iterativas que terminan tan pronto como se satisface algún criterio de detención basado en la precisión de α_k.

La experiencia práctica revela que no es necesario resolver con precisión el problema (7.29) para desarrollar métodos eficientes. En cambio, es crucial imponer alguna limitación sobre las longitudes de paso (y, por lo tanto, sobre los valores admisibles para α_k). De hecho, sin introducir ninguna limitación, una solicitud razonable sobre α_k sería que el nuevo iterado $x^{(k+1)}$ cumpla la desigualdad

$$f(x^{(k+1)}) < f(x^{(k)}) \tag{7.8}$$

donde $x^{(k)}$ y $d^{(k)}$ se han fijado. Con este propósito, el procedimiento basado en comenzar con un valor (suficientemente grande) de la longitud de paso α_k y reducir este valor a la mitad hasta que se cumpla 7.8, puede dar lugar a resultados completamente erróneos.

Criterios más estrictos que 7.8 deberían adoptarse en la elección de los posibles valores para α_k. Con este fin, observamos que surgen dos tipos de dificultades con los ejemplos anteriores: una tasa de descenso lenta de la secuencia y el uso de tamaños de paso pequeños.

La primera dificultad puede superarse exigiendo que:

$$\begin{aligned} 0 \geq v_M(x^{(k+1)}) = \frac{1}{\alpha_k}\left[f(x^{(k)}) - f(x^{(k)} + \alpha_k d^{(k)})\right] \\ \geq -\sigma \bigtriangledown f(x^{(k)})^T d^{(k)} \end{aligned} \tag{7.9}$$

con $\sigma \in (0, \frac{1}{2})$. Esto equivale a exigir que la tasa de descenso promedio v_M de f a lo largo de $d^{(k)}$, evaluada en $x^{(k+1)}$, sea al menos igual a una fracción dada de la tasa de descenso inicial en $x^{(k)}$. Para evitar la generación de tamaños de paso demasiado pequeños, requerimos que la tasa de descenso en la dirección $d^{(k)}$ en $x^{(k+1)}$ no sea menor que una fracción dada de la tasa de descenso en $x^{(k)}$.

$$\left|\bigtriangledown f(x^{(k)+\alpha_k d^{(k)}})^T d^{(k)}\right| \leq \beta \left|\bigtriangledown f(x^{(k)})^T d^{(k)}\right| \tag{7.10}$$

con $\beta \in (\sigma, 1)$ de manera que también satisfaga 7.9. En la práctica computacional, $\sigma \in [10^{-5}, 10^{-1}]$ y $\beta \in [10^{-1}, \frac{1}{2}]$ son elecciones habituales. A veces, 7.10 se reemplaza por la condición más suave

$$\bigtriangledown f(x^{(k)} + \alpha_k d^{(k)})^T d^{(k)} \geq \beta \bigtriangledown f(c^{(k)})^T d((k)) \tag{7.11}$$

La siguiente propiedad garantiza que, bajo suposiciones adecuadas, es posible encontrar valores de α_k que satisfagan 7.9 - 7.10 o 7.9 - 7.11.

Propiedad 1.2. SUponga que $f(x) \geq M$ para cualquier $x \in \mathbb{R}^n$. Entonces existe un intervalo $I =$

$[c, C]$ para el método de descenso, con $0 < c < C$, tal que $\forall \alpha_k \in I$, 7.9, 7.10 (o 7.9 - 7.11) se satisfacen, con $\sigma \in \left(0, \frac{1}{2}\right)$ y $\beta \in (\sigma, 1)$.

Bajo la restricción de cumplir las condiciones 7.9 y 7.10, existen varias opciones para α_k. Entre las estrategias más actuales, recordamos aquí las técnicas de retroceso (backtraking): fijando $\sigma \in \left(0, \frac{1}{2}\right)$, se comienza con $\alpha_k = 1$ y luego se sigue reduciendo su valor mediante un factor de escala adecuado $p \in (0, 1)$ (paso de retroceso) hasta que satisfaga 7.9.

Otras estrategias comúnmente utilizadas son las desarrolladas por Armijo y Goldstein. Ambas utilizan $\sigma \in (0, \frac{1}{2})$. En la fórmula de Armijo, se toma $\alpha_k = \beta^{m_k}\bar{\alpha}$, donde $\beta \in (0, 1)$, $\bar{\alpha} > 0$ y m_k es el primer entero no negativo tal que se satisface 7.9. En la fórmula de Goldstein, el parámetro α_k se determina de manera que

$$\sigma \leq \frac{f(x^{(k)} + \alpha_k d^{(k)}) - f(x^{(k)})}{\alpha_k \bigtriangledown f(x^{(k)})^T d^{(k)}} \leq 1 - \sigma \qquad (7.12)$$

1.3. Métodos tipo Newton para la minimización de funciones

Otro ejemplo de método de descenso lo proporciona el método de Newotn, que se diferencia de su versión para sistemas no lineales en que ahora no se aplica f, sino su gradiente.

Usando la notación de la sección 7.1.1, el método de Newton para la minimización de funciones consiste en calcular, dado $x^{(0)} \in \mathbb{R}^n$, para $k = 0, 1, ...$, hasta la convergencia:

$$\begin{aligned} d^{(k)} &= -H_k^{-1} \bigtriangledown f(x^{(k)}) \\ x^{(k+1)} &= x^{(k)} + d^{(k)} \end{aligned} \qquad (7.13)$$

donde se ha establecido $H_k = H(x^{(k)})$. El método se puede derivar truncando la expansión de Taylor de $f(x^{(k)})$ hasta el segundo orden:

$$f(x^{(k)} + p) \approx f(x^{(k)}) + \bigtriangledown f(x^{(k)})^T p + \frac{1}{2} p^T H_k p \quad (7.14)$$

Seleccionando p en 7.14 de tal manera que el nuevo vector $x^{(k+1)} = x^{(k)} + p$ satisfaga $\bigtriangledown f(x^{(k+1)}) = 0$, obtenemos el método 7.13, que así converge en un paso si f es cuadrática.

1.4. Método de Quasi-Newton

En la iteración genérica k-ésima, un método cuasi-Newton para la minimización de funciones realiza los siguientes pasos:

1. Calcular la matriz Hessiana H_k, o una aproximación adecuada B_k;
2. Encontrar una dirección de descenso $d^{(k)}$ (no necesariamente coincidente con la dirección proporcionada por el método de Newton), utilizando H_k o B_k;
3. Calcular el parámetro de aceleración α_k;
4. Actualizar la solución, estableciendo $x^{(k+1)} = x^{(k)} + \alpha_k d^{(k)}$, según un criterio de convergencia global.

En el caso particular en que $d^{(k)} = -H_k^{-1}\nabla f(x^{(k)})$, el esquema resultante se denomina método de Newton amortiguado.

2. Problemas de Mínimos Cuadrados

Hemos discutido la resolución de un sistema no singular $Ax = b$ de n ecuaciones lineales con n incógnitas. Tal sistema tiene una solución única. Supongamos, por otro lado, que hay más ecuaciones que incógnitas. Entonces, el sistema probablemente no tenga solución. En su lugar, podríamos desear encontrar una solución aproximada x que satisfaga $Ax \approx b$.

Esto se llama un problema de mínimos cuadrados, y corresponde a minimizar el cuadrado de la norma 2 del residuo, $\|b - Ax\|_2^2$, en un sistema sobredeterminado.

Sea A una matriz de $m \times n$ con $m > n$ y b un vector dado de dimensión m. Buscamos un vector x de dimensión n tal que $Ax \approx b$. Más precisamente, elegimos x para minimizar el cuadrado de la norma del residuo,

$$\|b - Ax\|_2^2 = \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j \right) \qquad (7.15)$$

Este es el problema de mínimos cuadrados que se abordará en esta sección y discutiremos dos enfoques diferentes.

2.1. Las Ecuaciones Normales

Una forma de determinal el valor de x que logra el mínimo en 7.15 es derivar con respectoa a cada componente x_k y establecer estas derivadas igual a 0. Dado que esta es una función cuadrática, el punto donde estas derivadas parciales sean 0 será, de hecho, el mínimo. Al derivar, encontramos:

$$\frac{\partial}{\partial x_k}(\|b - Ax\|_2^2) = \sum_{i=1}^{m} 2 \left(b_j - \sum_{j=1}^{n} a_{ij} x_j \right) (-a_{ik})$$

y estableciendo estas derivadas igual a 0 para $k = 1, ..., n$ da:

$$\sum_{i=1}^{m} a_{ik} \left(\sum_{j=1}^{n} a_{ij} x_j \right) \equiv \sum_{i=1}^{m} a_{ik} (Ax)_i = \sum_{i=1}^{m} a_{ik} b_i$$

De manera equivalente, podemos escribir $\sum_{i=1}^{m} (A^T)_{ki} (Ax)_i = \sum_{i=1}^{m} (A^T)_{ki} b_i$, $k = 1, ..., n$ o,

$$A^T Ax = A^T b \tag{7.16}$$

El sistema 7.16 se llama las ecuaciones normales.

2.2. Descomposición QR

El problema de mínimos cuadrados puede abordarse de una manera diferente. Deseamos encontrar un vector x tal que $Ax = b_*$, donde b_* es el vector más cercano b (en la norma 2) en el rango de A. Esto es equivalente a minimizar $\|b - Ax\|_2$ ya que, por la definición de b_*, se cumple que $\|b - b_*\|_2 \leq \|b - Ay\|_2$ para todo y.

Tal vez recuerdes de álgebra lineal que el vector más cercano a un vector dado, desde un subespacio, es la proyección ortogonal de ese vector sobre el subespacio.

Si $q_1, ..., q_k$ forman una base ortonormal para el subespacio, entonces la proyección ortogonal de b sobre el subespacio es $\sum_{j=1}^{k} \langle b, q_j \rangle q_j$. Supongamos que las columnas de A son linealmente independientes, de modo que el rango de A (que es el espacio generado por sus columnas) tiene dimensión n. Entonces el vector más cercano a b en $range(A)$ es

$$b_* = \sum_{j=1}^{n} \langle b, q_j \rangle q_j$$

donde $q_1, ..., q_k$ forman una base ortonormal para $range(A)$. Sea Q la matriz $m \times n$ cuyas columnas son los vectores ortonormales $q_1, ..., q_n$. Entonces, $Q^T Q = I_{m \times n}$ y la fórmula para b_* puede escribirse de manera compacta como

$$b_* = Q(Q^T b)$$

ya que $Q(Q^T b) = \sum_{j=1}^{n} q_j (Q^T b)_j = \sum_{j=1}^{n} q_j (q_j^T b) = \sum_{j=1}^{n} q_j \langle b, q_j \rangle$.

También recordarás de álgebra lineal que, dado un conjuunto de vectores linealmente independientes, como las columnas de A, se puede construir un conjunto ortonormal que abarque el mismo espacio utilizando el algoritmo de Gram-Schmidt:

Dado un conjunto linealmente independiente $v_1, v_2, ...,$ e define $q_1 = \frac{v_1}{\|v_1\|}$, y para $j = 2, 3, ...,$

- $\tilde{q}_j = v_j - \sum_{i=1}^{j-1} \langle v_j, q_i \rangle q_i$
- $q_j = \frac{\tilde{q}_j}{\|\tilde{q}_j\|}$

Cabe señalar que si $v_1, v_2, ..., v_n$ son las columnas de A, entonces el algoritmo de Gram-Schmidt puede interpretarse como una factorización de la matriz A de tamaño $m \times n$ en la forma $A = QR$, donde $Q = (q_1, ..., q_n)$ es una matriz $m \times n$ con columnas ortonormales y R es una matriz triangular superior de tamaño $n \times n$. Para ver esto, se pueden escribir las ecuaciones del algoritmo de Gram-Schmidt en la forma

$$v_1 = r_{11} q_1, \quad r_{11} = \|v_1\|$$

$$v_j = r_{jj} q_j + \sum_{i=1}^{j-1} r_{ij} q_i, \quad r_{jj} = \|\tilde{q}_j\|, \quad r_{ij} = \langle v_j, q_i \rangle$$

$$j = 2, ..., n$$

Estas ecuaciones pueden escribirse utilizando matrices como

$$(v_1, ..., v_n) = (q_1, ..., q_n) \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ & r_{22} & \cdots & r_{2n} \\ & & \ddots & \vdots \\ & & & r_{nm} \end{pmatrix}$$

Así, si $v_1, ..., v_n$ son las columnas de A, entonces hemos escrito A en la forma $A = QR$. Esto se llama la descomposición QR reducida de A.

Habiendo factorizado A en la forma QR, el problema de los mínimos cuadrados se conbierte en $QRx = b_* = QQ^T b$. Sabemos que este conjunto de ecuaciones tiene una solución, ya que $b_* = QQ^T b$ está en el rango de A. Por lo tanto, si multiplicamos ambos lador por Q^T, la ecuación resultante tendrá la misma solución. Dado que $Q^T Q = I$, nos queda el sistema triangular superior $Rx = Q^T b$.

Ejercicios

(1) Utilizar el método de Newton y del máximo descenso para encontrar los extremos de las siguientes funciones:

a) $f(x, y) = -ln(1 - x - y) - lnx - lny$

b) $f(x) = 7x - lnx$

c) $f(x, y) = (a - x)^2 + b(y - x^2)^2$, dando distintos valores a y b. Dibujar juntas las curvas de nivel y la solución aproximada de cada iteración.

Encontrar el polinomio de grado 10 que mejor ajusta la función $b(t) = cos(4t)$ con 50 nodos equispaciados en $[0, 1]$. Determinar la matriz A de Vandermonde y el vector b del problema y determinar los coeficientes del polinomio de las siguientes maneras:

a) resolviendo las ecuaciones normales del problema;

b) utilizando la descomposición QR.

Dibujar los datos y la solución en una misma gráfica. Calcular el error mínimo cometido en el ajuste.

(3) Encontrar la elipse $x^3+By^2+Cxy+Dx+Ey+F = 0$ que mejor ajusta los puntos del plano (0, 2), (2, 1), (1, -1), (-1, -2), (-3, 1), (-1, -1). Después, dibujar los datos y la solución en una misma gráfica. Calcular el error mínimo cometido en el ajuste.

(4) Se considera dos vectores de datos, u = [0.132, 0.322, 0.511, 0.701, 0.891, 1.081, 1.27, 1.46, 1.65, 1.839, 2.029, 2.219] y v = [0.1, 0.258, 0.543, 0.506, 0.606, 0.622, 0.569, 0.453, 0.438, 0.316, 0.29, 0.195]. Se pide ajustar la función de Weibull $W(u, \alpha, \beta) = \alpha \cdot \beta \cdot u^{\beta-1} \cdot e^{\alpha \cdot u^{\beta}}$ con el Método de Gauss-Newton tomando como puntos de inicio de las iteraciones $\alpha = 0{,}8$ y $\beta = 1$. Después, dibujar los datos y la solución en una misma gráfica. Calcular el error mínimo cometido en el ajuste.

(5) Se consideran los datos (x, y): (-2, 0.5), (-1, 1), (0, 2), (1, 4). Ajustar la función $f(t, x, y) = e^{x+ty}$ a estos datos con el Método de Gauss-Newton. Después, dibujar los datos y la solución en una misma gráfica. Calcular el error mínimo cometido en el ajuste.

Capítulo 8

Optimización Con Restricciones

1. Introducción

El caso más simple de optimización con restricciones se puede formular de la siguiente manera. Dada una función $f : \mathbb{R}^n \to R$,

$$minimize \quad f(x), \quad \text{con } x \in \Omega \subset \mathbb{R}^n \tag{8.1}$$

Más precisamente, se dice que un punto x^* es un minimizador global en Ω si satisface 8.1, mientras que es un minimizador local si $\exists R > 0$ tal que

$$f(x^*) \leq f(x), \quad \forall x \in B(x^*; R) \subset \Omega$$

La existencia de soluciones para el problema 8.1 está, por ejemplo, asegurada por el Teorema de Weiertrass, en el caso en que f sea continua y Ω sea un conjunto cerrado y acotado. Bajo la suposición de que Ω es un conjunto convexo, se cumplen las siguientes condiciones de optimalidad.

Propiedad 1.1. Sea $\Omega \subset \mathbb{R}^n$ y $f \in C^1(B(x^*; R))$, para una $R > 0$ adecuado. Entonces:

1. Si x^* es un minimizador local de f, entonces:

$$\bigtriangledown f(x^*)^T(x - x^*) \geq 0, \quad \forall x \in \Omega \tag{8.2}$$

2. Además, si f es convexa en Ω, entonces x^* es un minimizador global de f.

Propiedad 1.2. Sea $\Omega \subset \mathbb{R}^n$ un conjunto cerrado y convexo, y sea f una función fuertemente convexa en Ω. Entonces, existe un único minimizador local $x^* \in \Omega$

Un caso notable de 8.1 es el siguiente problema: dada $f : \mathbb{R}^n \to \mathbb{R}$

$$minimize \quad f(x), \text{ bajo la restricción de que } h(x) = 0 \tag{8.3}$$

donde $h : \mathbb{R}^n \to \mathbb{R}$, con $m \leq n$, es una función dada con componentes $h_1, ..., h_m$. Los analogos de los puntos críticos en el problema 8.3 se denominan puntos regulares.

Definición 1.1. Un punto $x^* \in \mathbb{R}^n$, tal que $h(x^*) = 0$, se dice que es regular si los vectores de la matriz Jacobiana $J_h(x^*)$ son linealmente independientes, asumiendo que $h_i \in C^2(B(x^*; R))$, para un $R > 0$ adecuado y $i = 1, ..., m$

Nuestro objetivo ahora es convertir el problema 8.3 en un problema de minimización sin restricciones.

Por ese motivo, introducimos la función del Lagrangiano $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^{\geqslant} \to \mathbb{R}$

$$\mathcal{L}(x, \lambda) = f(x) + \lambda^T h(x)$$

donde el vector λ se conoce como multiplicador de Lagrange. Además, denotemos por $J_{\mathcal{L}}$ la matriz Jacobiana asociada a $\mathcal{L}$, pero donde las derivadas parciales se toman únicamente con respecto a las variables $x_1, ..., x_n$.

Propiedad 1.3. Sea x^* un minimizador local para 8.3 y suponiendo que, para un $R > 0$ adecuado, $f, h_i \in C^1(B(x^*; R))$, para $i = 1, ..., m$. Entonces, existe un vector único $\lambda^* \in \mathbb{R}^m$ tal que $J_{\mathcal{L}}(c^*, \lambda^*) = 0$.

Por el contrario, supongamos que $x^* \in \mathbb{R}^n$ satisface $h(x^*) = 0$ y que, para un $R > 0$ adecuado y $i = 1, ..., m$, $f, h_i \in X^2(B(x^*; R))$. Sea $H_{\mathcal{L}}$ la matriz de entradas $\frac{\partial^2 \mathcal{L}}{\partial x_i \partial x_j}$ para $i, j = 1, ..., n$. Si existe un vector $\lambda^* \in \mathbb{R}^m$ tal que $J_{\mathcal{L}}(x^*, \lambda^*) = 0$ y

$$z^T H_{\mathcal{L}}(x^*, \lambda^*) z > 0 \quad \forall z \neq 0, \quad \text{con } \bigtriangledown h(x^*)^T z = 0$$

entonces x^* es un minimizador estricto local de 8.3.

La última clase de problemas que vamos a tratar incluye el caso en el que también están presentes restricciones de desigualdad, es decir: dado $f : \mathbb{R}^n \to \mathbb{R}$,

$$minimize f(x), \text{ bajo la restricción } h(x) = 0 \text{ y } g(x) \leq 0 \tag{8.4}$$

donde $h : \mathbb{R}^n \to \mathbb{R}^m$, con $m \leq n$, y $g : \mathbb{R}^n \to \mathbb{R}^r$ son dos funciones dadas. Se entiende que $g(x) \leq 0$ significa que $g_i(x) \leq 0$ para $i = 1, ..., r$. Las restricciones de desigualdad generan una complicación formal adicional con respecto al caso examinado previamente, pero no impiden convertir la solución de 8.4 en la minimización de una función Lagangiana adecuada.

Definición 1.2. Supongamos que $h_i, g_j \in C^1(B(x^*; R))$ para un $R > 0$ adecuado, con $i = 1, ..., m$ y $j = 1, ..., r$, y denotemos por $J(x^*)$ el conjunto de índices j tales que $g_j(x^*) = 0$. Un punto $x^* \in \mathbb{R}^n$ tal que $h(x^*) = 0$ y $g(x^*) \leq 0$ se dice que es regular si los vectores columna de la matriz Jacobiana $J_h(x^*)$ junto con los vectores $\bigtriangledown g_j(x^*), j \in \mathcal{J}(x^*)$, forman un conjunto de vectores linealmente independientes.

2. Condiciones necesarias de Kuhn-Tucker para la programación no lineal

En esta sección recordamos algunos resultados, conocidos como las condiciones de Kuhn-Tucker, que aseguran en general la existencia de una solución local para el problema de programación no lineal. Bajo suposiciones adecuadad, también garantizan la existencia de una solución global. A lo largo de esta sección suponemos que un problema de minimización siempre se puede reformular como un problema de maximización.

Consideremos el problema general de programación no lineal:

$$\begin{aligned} &\text{dada } f : \mathbb{R}^n \to \mathbb{R} \\ &maximize \quad f(x), \text{ sujeto a} \\ &g_i(x) \leq b_i, \quad i = 1, ..., l \\ &g_i(x) \geq b_i, \quad i = l+1, ..., k \\ &g_i(x) = b_i, i = k+1, ..., m \\ &x \geq 0 \end{aligned} \tag{8.5}$$

Un vector x que satisface las restricciones anteriores se llama una solución factible de 8.5 y el conjunto de las soluciones factibles se llama la región factible. Suponemos en adelante que $f, g_i \in C^1(\mathbb{R}^n)$, con $i = 1, ..., m$ y definimos conjuntos $I_= = i : g_i(x^*) = b_i$, $I_{/==} i : g_i(x^*) \neq b_i$, $J_= = i : x_i^* = 0$, $J_> = i : x_i^* > 0$, denotando por x^* un maximizador local de f. Asociamos con 8.5 el siguiente Lagrangiano

$$\mathcal{L}(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i [b_i - g_i(x)] - \sum_{i=m+1}^{m+n} \lambda_i x_{i-m}$$

Propiedad 2.1. Si f tiene un máximo local restringido en el punto $x = x^*$, es necesario que el vector $\lambda^* \in \mathbb{R}^{m+1}$ exista tal que (primera condición de Kuhn-Tucker):

$$\bigtriangledown_x \mathcal{L}(x^*, \lambda^*) \leq 0$$

donde se cumple igualdad estricta para componente $i \in J_>$. Además (segunda condición de Kuhn-Tucker):

$$\bigtriangledown_x \mathcal{L}(x^*, \lambda^*)^T x^* = 0$$

La tercera condición de Kuhn-Tucker requiere que:

$$\begin{aligned} \bigtriangledown_\lambda \mathcal{L}(x^*, \lambda^*) \geq 0 \quad i = 1, ...l \\ \bigtriangledown_\lambda \mathcal{L}(x^*, \lambda^*) \leq 0 \quad i = l+1, ...k \\ \bigtriangledown_\lambda \mathcal{L}(x^*, \lambda^*) = 0 \quad i = k+1, ...m \end{aligned}$$

Además (cuarta condición de Kuhn-Tucker):

$$\bigtriangledown_x \mathcal{L}(x^*, \lambda^*)^T x^* = 0$$

Propiedad 2.2. Supongamos que la función f en 8.5 es una función cóncava (es decir, $-f$ es convexa) en la región factible. Supongamos también que el punto (x^*, λ^*) satisface todas las condiciones necesarias de Kuhn-Tucker y que las funciones g_i para las cuales $\lambda_i^* > 0$ son convexas, mientras que aquellas para las cuales $\lambda_i^* < 0$ son cóncavas. Entonces, $f(x^*)$ es el máximo global restringido de f para el problema 8.5.

3. Método del Penalti

La idea básica de este método es eliminar, parcial o completamente, las restricciones para transformar el problema restringido en uno no restringido. Este nuevo problema se caracteriza por la presencia de un parámetro que proporciona una medida de la precisión con la que se impone efectivamente la restricción.

Consideremos el problema restringido 8.3, suponiendo que buscamos la solución x^* únicamente en $\Omega \subset \mathbb{R}^n$. Supongamos que dicho problema admite al menos una solución en Ω y lo escribimos en la siguiente forma penalizada:

$$minimize \quad \mathcal{L}_\alpha(x) \quad \text{para } x \in \Omega \tag{8.6}$$

donde

$$\mathcal{L}_\alpha(x) = f(x) + \frac{1}{2}\alpha \|h(x)\|_2^2$$

La función $\mathcal{L}_\alpha : \mathbb{R}^n \to \mathbb{R}$ se llama el Lagrangiano penalizado, y α se llama el parámetro de penalización. Es claro que si la restricción se satisficiera exactamente, entonces minimizar f sería equivalente a minimizar $\mathcal{L}_\alpha$.

El método de penalización es una técnica iterativa para resolver 8.6.

Para $k = 0, 1, ...$, hasta la convergencia, se debe resolver la secuancia de problemas:

$$minimize \quad \mathcal{L}_{\alpha_k}(x) \quad \text{con } x \in \Omega \tag{8.7}$$

donde α_k es una secuencia monótamente creciente de parámetros de penalización positivos, tal que $\alpha_k \to \infty$ cuando $k \to \infty$.

Como consecuencia, después de elegir α_k, en cada paso del proceso de penalizaciónd ebemos resolver un problema de minimización con respecto a la variable x, lo que da lugar a una seguencia de valores x_k^*, soluciones de 8.7. Al hacer esto, la función objetivo $\mathcal{L}_{\alpha_k}(x)$ tiende a infinito, a menos que $h(x)$ sea igual a cero.

Los problemas de minimización pueden ser resueltos luego mediante uno de los métodos introducidos en el capítulo anterior. La siguiente propiedad garantiza la convergencia del método de penalización en la forma 8.6.

Propiedad 3.1. Supongamos que $f : \mathbb{R}^\kappa \to \mathbb{R}$ y $h : \mathbb{R}^n \to \mathbb{R}^m$, con $m \leq n$, son funciones continuas en un conjunto cerrado $\Omega \subset \mathbb{R}^n$ y supongamos que la secuencia de parámetros de penalización $\alpha_k > 0$ es monótamente divergente. Finalmente, sea x_k^* el minimizador global del problema 8.7 en el paso k. Entonces, al tomar el límite cuando $k \to \infty$, la secuencia x_k^* converge a x^*, que es un minimizador global de f en Ω y satisface la restricción $h(x^*) = 0$.

En cuanto a la selección de parámetros α_k, se puede demostrar que valores grandes de α_k hacen que el problema de minimización 8.7 esté mal condicionado, lo que hace que su solución sea bastante costosa, a menos que la suposición inicial esté particularmente cerca de x^*. Por otro lado, la secuencia α_k no debe crecer demasiado lentamente, ya que esto afectaría negativamente la convergencia global del método.

Una elección que se hace comúnmente en la práctica es seleccionar un valor no demasiado grande α_0 y luego establecer $\alpha_k = \beta\alpha_{k-1}$ para $k > 0$ donde β es un número entero entre 4 y 10.

4. Método de los multiplicadores de Lagrange

Una variante del método de penalización hace uso de la función Lagrangiana aumentada $\mathcal{G}_\alpha : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ dada por

$$\mathcal{G}_\alpha(x, \lambda) = f(x) + \lambda^T h(x) + \frac{1}{2}\alpha\|h(x)\|_2^2 \tag{8.8}$$

donde $\lambda \in \mathbb{R}^m$ es un multiplicador de Lagrange. Claramente, si x^* es una solución del problema 8.3, entonces también será solución de 8.8, pero con la ventaja, respecto a 8.6, de disponeer de un grado de libertad adicional λ. El método de penalización aplicado a 8.8 se formula de la siguiente manera: para $k = 0, 1...$, hasta la convergencia, resolver la secuencia de problemas

$$minimize \quad \mathcal{G}_{\alpha_k}(x, \lambda_k) \text{ para } x \in \Omega \tag{8.9}$$

donde α_k es una secuencia acotada de vectores desconocidos en $\mathbb{R}^m$, y los parámetros α_k se definen como antes.

En cuanto a la elección de los multiplicadores, la secuencia de vectores λ_k se asigna típicamente mediante la siguiente fórmula:

$$\lambda_{k+1} = \lambda_k + \alpha_k h(x^{(k)})$$

donde λ_0 es un valor dado, mientras que la secuencia de α_k puede establecerse a priori o modificarse durante la ejecución.

En lo que respecta a las propiedades de convergencia del método de los multiplicadores de Lagrange, se cumple el siguiente resultado local.

Propiedad 4.1. Supongamos que x^* es un minimizador local estricto regular del problema 8.3 y que:

1. $f, h_i \in C^2(B(x^*; R))$ con $i = 1, ..., m$ y para un $R > 0$ adecuado.
2. El par (x^*, λ^*) satisface $z^T H_{\mathcal{G}_0}(x^*, \lambda^*)z > 0, \forall z \neq 0$ tal que $J_h(x^*)^T z = 0$
3. Existe un $\overline{\alpha} > 0$ tal que $H_{\mathcal{G}_{\overline{\alpha}}}(x^*, \lambda^*) > 0$

Entonces, existen tres escalares positivos δ, γ y M tal que, para cualquier par $(\lambda, \alpha) \in V = \{(\lambda, \alpha) \in \mathbb{R}^{m+1} : \|\lambda - \lambda^*\|_2 < \delta\alpha, \alpha \geq \overline{\alpha}\}$, el problema

$$minimize \quad \mathcal{G}_\alpha(x, \lambda), \text{ para } x \in B(x^*; \gamma)$$

admite una solución única $x(\lambda, \alpha)$, diferenciable con respecto a sus argumentos. Además, para todo $(\lambda, \alpha) \in V$

$$\|x(\lambda, \alpha) - x^*\|_2 \leq M\|\lambda - \lambda^*\|_2$$

Bajo suposiciones adicionales se puede demostrar que el método de los multiplicadores de Lagrange converge. Además, si $\alpha_k \to \infty$ cuando $k \to \infty$, entonces

$$\lim_{k\to\infty} \frac{\|\lambda_{k+1} - \lambda^*\|_2}{\|\lambda_k - \lambda^*\|_2} = 0$$

y la convergencia del método es más que lineal.

En el caso en que la secuencia α_k tenga una cota superior, el método converge linealmente.

Finalmente, cabe señalar que, a diferencia del método de penalización, ya no es necesario que la secuencia α_k tienda a infinito. Esto, a su vez, limita el mal condicionamiento del problema 8.9 a medida que α_k crece. Otra ventaja se refiere a la tasa de convergencia del método, que resulta ser independiente de la tasa de crecimiento del parámetro de penalización en el caso de la técnica de los multiplicadores de Lagrange. Esto, por supuesto, implica una reducción considerable del coste computacional.

Ejercicios

(1) Un dado de 6 caras tiene la probabilidad p_k de sacar el número k, con $k \in \{1, 2, 3, 4, 5, 6\}$ donde $p_k \geq 0$ para todo k y $p_1+p_2+p_3+p_4+p_5+p_6 = 1$. La entropía de la distribución de probabilidad $p = (p_1, p_2, ...p_6)$ se define como $f(p) = -\sum_{i=1}^{6} p_k ln p_k$. Encontrar la distribución p que maximiza la entropía. Observación: entropía máxima implica menos contenido de información.

(2) La probabilidad de que un sistema físico o químico esté en un estado k es p_k y la energía de ese estado es E_k. Si las energías E_i están fijas, la naturaleza tiende a minimizar la energía libre $f(p_1, ...p_n) = -\sum_{i=1}^{n}(p_1 ln(p_i) - E_i p_i)$. Encontrar la distribución de probabilidad $p_1, \sum_{i=1}^{n} p_1 = 1$, que minimiza la energía libre. Observación: está distribución se llama distribución de Gibbs.

(3) Utilizar el método del Penalty y del Lagrangiano aumentado para aproximar el mínimo de la función

$$f(x, y) = (1 - x)^2 + 100(y - x^2)^2$$

restringida a:

a) $x + y - 2 = 0 \quad (x - 1)^3 - y + 1 = 0$

b) $x^2 + y^2 = 2$

(4) Utilizar el método del Penalty y del Lagrangiano aumentado para aproximar el mínimo de la función

$$f(x, y) = 4x^2 - 2{,}1x^4 + \frac{1}{3}x^6 + xy - 4y^2 + 4y^4$$

sujeto a: $-sin(4\pi x) + 2sin^2(2\pi y) = 1{,}5$

Capítulo 9

Métodos para Ecuaciones Diferenciales Ordinarias

En este capítulo estudiamos problemas de la forma

$$y'(t) = f(t, y(t)), \quad t \geq t_0$$
$$y(t_0) \equiv y_0 \tag{9.1}$$

donde la variable independiente t normalmente representa tiempo, $y \equiv y(t)$ es la función desconocida que buscamos, y y_0 es un valor inicial conocido. Esto se llama problema del valor inicial (PVI) para la ecuación diferencial ordinaria (EDO) $y' = f(t, y)$.

1. Existencia y unicidad de las soluciones

Al estudiar cuestiones sobre la existencia y unicidad de soluciones para la ecuación 9.1, consideramos que la función del lado derecho es una función de dos varibales independientes t y y, y hacemos suposiciones sobre su comportamiento como función de cada una de estas variables. El siguiente teorema proporciona condiciones suficientes para que el problema de valor inicial tenga una solución localmente.

Teorema 1.1. Si f es continua en un rectángulo R centrado en (t_0, y_0),

$$R = (t, y) : |t - t_0| \leq \alpha, |y - y_0| \leq \beta$$

entonces el PVI 9.1 tiene una solución $y(t)$ para $|t - t_0| \leq min(\alpha, \beta/M)$, donde $M = max_R|f(t, y)|$.

Aunque exista una solución, esta puede no ser única. El siguiente teorema proporciona condiciones suficientes para la existencia y unicidad locales.

Teorema 1.2. Si f y $\frac{\partial f}{\partial y}$ son continuas en el rectángulo R, entonces el problema de valor inicial tiene una única solución $y(t)$ para $|t - t_0| \leq min(\alpha, \beta/M)$ donde $M = max_R|f(t, y)|$.

1.1. Teoría elemental de problemas de valor inicial

Definición 1.1. Se dice que una función $f(t, y)$ satisface la condición de Lipschitz en la variable y en un conjunto $D \subset \mathbb{R}^n$ si existe una constante $L > 0$ con

$$|f(t, y_1) - f(t, y_2)| \leq L|y_1 - y_2|$$

siempre que (t, y_1) y (t, y_2) estén en D. La constante L recibe el nombre de constante de Lipschitz para f.

Definición 1.2. Se dice que un conjunto $D \subset \mathbb{R}^n$ es convexo siempre que (t_1, y_1) y (t_2, y_2) pertenezcan a D, entonces $((1 - \lambda)t_1 + \lambda t_2, (1 - \lambda)t_1 + \lambda t_2)$ también pertenece a D para cada λ en $[0, 1]$.

Teorema 1.3. Suponga que $f(t, y)$ se define sobre un conjunto convexo $D \subset \mathbb{R}^n$. Si existe una constante $L > 0$ con

$$\left|\frac{\partial f}{\partial y}(t, y)\right| \leq L, \text{ para todo } (t, y) \in D \tag{9.2}$$

entonces f satisface la condición de Lipschitz en D en la variable y con constante L de Lipschitz.

Definición 1.3. El problema de valor inicial

$$\frac{dy}{dt} = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha \tag{9.3}$$

se dice que es un problema bien planteado si:

- Existe una única solución $y(t)$
- Existen constantes $\epsilon > 0$ y $k > 0$, tales que para cualquier ϵ en $(0, \epsilon_0)$, siempre que $\delta(t)$ es continua con $|\delta(t)| < \epsilon$ para toda t en $[a, b]$, y cuando $|\delta_0| < \epsilon$, el problema de valor inicial

$$\frac{dz}{dt} = f(t, z) + \delta(t), \quad a \leq t \leq b, \quad z(a) = \alpha + \delta_0 \tag{9.4}$$

tiene una única solución $z(t)$ que satisface

$$|z(t) - y(t)| < k\epsilon \text{ para toda } t \in [a, b]$$

El problema especificado por la ecuación 9.4 recibe el nombre de problema perturbado relacionado con el problema original en la ecuación 9.3. Suponge la posibilidad de un error introducido en la declaración de la ecuación diferencial, así como un error ∂_0 presente en la condición inicial.

Teorema 1.4. Suponga que $D = (t, y =)|a \leq t \leq b$ y $-\infty < y < \infty$. Si f es continua y satisface la condición de Lipschitz en la variable y sobre el conjunto D, entonces el problema de valor inicial

$$\frac{dy}{dt} = f(t, y), a \leq t \leq b, y(a) = \alpha$$

está bien planteado.

2. Métodos de un paso

Suponiendo que el problema de valor inicial 9.1 estébien planteado, aproximaremos su solución en el tiempo T dividiendo el intervalo $[t_0, T]$ en pequeños subintervalos y reemplazando la derivada temporal sobre cada subintervalo por un cociente de diferencias finitas.

Sea $t_0, t_1, ..., T_N$ los puntos finales de los subintervalos (llamados nodos o puntos de malla), donde $t_N = T$, y sea la solución aproximada en el tiempo t_j denotada como y_j. Un método de un paso es aquel en el que la solución aproximada en el tiempo t_{k+1} se determina a partir de la solución en el tiempo t_k

2.1. Método de Euler

El método de Euler es la técnica de aproximación más básica para resolver problemas de valor inicial.

El objetivo del método de Euler es obtener aproximaciones para el problema de valor inicial bien planteado

$$\frac{dy}{dt} = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha \tag{9.5}$$

No se obtendrá una aproximación continua a la solución $y(t)$; en su lugarm las aproximaciones para y se generarán en varios valores, llamados puntos de malla, en el intervalo $[a, b]$. Una vez que se obtiene la solución aproximada en los puntos, la solución aproximada en otros puntos en el intervalo se puede encontrar a través de interpolación.

Primero estipulamos que los puntos de malla estén igualmente expaciados a lo largo del intervalo $[a, b]$. Esta condición se garantiza al seleccionar un entero positivo N, al establecer $h = (b - a)/N$, y seleccionar los puntos de malla

$$t_i = a + ih, \quad \forall i = 0, 1, 2, ..., N$$

La distancia común entre los puntos $h = t_{i+1} - t_i$ recibe el nombre de tamaño de paso.

Usaremos el teorema de Taylor para deducir el método de Euler. Suponga que $y(t)$, la única solución para 9.5, tiene dos derivadas continuas en $[a, b]$, de tal forma que cada $i = 0, 1, 2, ..., N - 1$

$$y(t_{i+1}) = y(t_i) + (t_{i+1} - t_i)y'(t_i) + \frac{(t_{i+1} - t_i)^2}{2}y''(\xi_i)$$

para algún número ξ_i en (t_i, t_{i+1}). Puesto que $h = t_{i+1} - t_i$, tenemos

$$y(t_{i+1}) = y(t_i) + hy'(t_i) + \frac{h^2}{2}y''(\xi_i)$$

y ya que $y(t)$ satisface la ecuació diferencial 9.5,

$$y(t_{i+1}) = y(t_i) + hf(t_i, y(t_i)) + \frac{h^2}{2}y''(\xi_i) \tag{9.6}$$

El método de Euler construye $w_i \approx y(t_i)$, para cada $i = 1, 2, ..., N$, al borrar el término restante. Por lo tanto, el método de Euler es

$$w_0 = \alpha$$

$$w_{i+1} = w_i + hf(t_i, w_i), \quad \forall i = 0, 1, ..., N - 1 \tag{9.7}$$

La ecuación 9.7 recibe el nombre de ecuación de diferencia relacionada con el método de Euler.

Cotas del error para el método de Euler

A pesar de que el método de Euler no es por completo apropiado para garantizar su uso en la práctica, es suficientemente básico para analizar el error producido a partir de esta aplicación.

Para derivar una cota del error para el método de Euler, necesitamos dos lemas de cálculo.

Lema 2.1. Para toda $x \geq -1$ y cualquier m positiva, tenemos $0 \leq (1+x)^m \leq e^{mx}$

Lema 2.2. Si s y t son números reales positivos, $a_{i\,i=0}^{k}$ es una sucesión que satisface $a_0 \geq -t/s$, y

$$a_{i+1} \leq (1+s)a_i + t, \quad \forall i = 0, 1, 2, ..., k-1 \tag{9.8}$$

entonces

$$a_{i+1} \leq e^{(i+1)s}\left(a_0 + \frac{t}{s}\right) - \frac{t}{s}$$

Teorema 2.1. Suponga que f es continua y satisface la condición de Lipschitz con constante L en

$$D = (t, y)|a \leq t \leq b \text{ y } -\infty < y < \infty$$

y que existe una constante M con

$$|y''(t)| \leq M, \quad \forall t \in [a, b]$$

donde $y(t)$ denota la única solución para el problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha$$

Sean $w_0, w_1, ..., w_N$ las aproximaciones generadas por el método de Euler para un entero positivo N. Entonces, para cada $i = 0, 1, 2, ..., N$

$$|y(t_i) - w_i| \leq \frac{hM}{2L}\left[e^{L(t_i - a)} - 1\right] \tag{9.9}$$

La principal importancia de la fórmula de la cota de error determinada en el teorema 2.1 es que la cota depende linealmente del tamaño de paso h. Por consiguiente, disminuir el tamaño de paso debería proporcionar mayor precisión para las aproximaciones en la misma medida.

Olvidado en el resultado del teorema 2.1 está el efecto que el error de redondeo representa en la selección del tamaño de paso. Conforme h se vuelve más pequeño, se necesitan más cálculos y se espera más error de redondeo. Entonces, en la actualidad, la forma de la ecuación de diferencia

$$w_0 = \alpha$$

$$w_{i+1} = w_i + hf(t_i, w_i), \quad \forall i = 0, 1, ..., N-1$$

no se utiliza para calcular la aproximación a la solución y_i en un punto de malla t_i. En su lugar, usamos una ecuación de la forma

$$u_0 = \alpha + \delta_0$$

$$u_{i+1} = u_i + hf(t_i, y_i) + \delta_{i+1}, \quad \forall i = 0, 1, ..., N-1 \tag{9.10}$$

donde δ_i denota el error de redondeo asociado con u_i.

Teorema 2.2. Si $y(t)$ denota la única solución para el problema de valor inicial

$$y' = f(t, y), \quad a \leq t \leq b, \quad y(a) = \alpha \tag{9.11}$$

y $u_0, u_1, ..., u_N$ son las aproximaciones obtenidas de la ecuación 9.10. Si $|\delta_i| < \delta$ para cada $i = 0, 1, ..., N$ y la hipótesis del teorema 2.1 son aplicables a la ecuación 9.11, entonces

$$|y(t_i - u_i)| \leq \frac{1}{L}\left(\frac{hM}{2} + \frac{\delta}{h}\right)\left[e^{L(t_i - a)} - 1\right] + |\delta_0|e^{L(t_i - a)} \tag{9.12}$$

para cada $i = 0, 1, ..., N$.

La cota de error 9.12 ya no es lineal en h. De hecho, puesto que

$$\lim_{h \to 0}\left(\frac{hM}{2} + \frac{\delta}{h}\right) = \infty$$

se esperaría que el error se vuelva más grande para los valores suficientemente pequeños de h. El cálculo se puede usar para determinar una cota inferior para el tamaño de paso h. Si $E(h) = (hM/2) + (\delta/h)$ implica que $E'(h) = (M/2) - (\delta/h^2)$:

- Si $h < \sqrt{2\delta/M}$, entonces $E'(h) < 0$ y $E(h)$ disminuye.
- Si $h > \sqrt{2\delta/M}$, entonces $E'(h) > 0$ y $E(h)$ aumenta.

El valor mínimo de $E(h)$ se presenta cuando

$$h = \sqrt{\frac{2\delta}{M}} \tag{9.13}$$

La disminución de h más allá de este valor tiende a incrementar el error total en la aproximación. Por lo general, sin embargo, el valor de δ es suficientemente pequeño para que esta cota inferior para h no afecte la operación del método de Euler.

2.2. Método del medio punto

El método del punto medio se define tomando un medio paso con el método de Euler para aproximar la solución en el tiempo $t_{k+1/2} \equiv (t_k + t_{k+1})/2$, y luego tomando un paso completo utilizando el valor de la solución en $t_{k+1/2}$ y la solución aproximada $y_{k+1/2}$:

$$y_{k+1/2} = y_k + \frac{h}{2}f(t_k, y_k) \tag{9.14}$$

$$y_{k+1} = y_k + hf(t_{k+1/2}, y_{k+1/2}) \tag{9.15}$$

Para determinar el error de truncamiento local de este método, expandimos la solución exacta en

una serie de Taylor alrededor de $t_{k+1/2} = t_k + h/2$:
$y(t_{k+1}) = y(t_{k+1/2}) + \frac{h}{2} f(t_{k+1/2}, y(t_{k+1/2})) + \frac{(h/2)^2}{2} y''(t_{k+1/2}) + O(h^3)$
$y(t_k = y(t_{k+1/2})) - \frac{h}{2} f(t_{k+1/2}, y(t_{k+1/2})) + \frac{(h/2)^2}{2} y''(t_{k+1/2}) + O(h^3)$

Restando estas dos ecuaciones se obtiene

$$y(t_{k+1}) - y(t_k) = hf(t_{k+1/2}, y(t_{k+1/2})) + O(h^3)$$

Ahora, expandiendo $y_(t_{k+1/2})$ alrededor de t_k se obtiene:

$$y(t_{k+1/2}) = y(t_k) + \frac{h}{2} f(t_k, y(t_k)) + O(h^2)$$

Y al hacer esta sustitución, tenemos:

$$\begin{aligned} &y(t_{k+1}) - y(t_k) = \\ &= hf(t_{k+1/2}, y(t_k) + \frac{h}{2} f(t_k, y(t_k)) + O(h^2)) + O(h^3) = \\ &hf(t_{k+1/2}, y(t_k) + \frac{h}{2} f(t_k, y(t_k))) + O(h^3) \end{aligned} \quad (9.16)$$

donde la segunda igualdad se sigue porque f es Lipschitz en su segundo argumento, de modo que $hf(t, y + O(h^2)) = hf(t, y) + O(h^3)$. Dado que a partir de 9.14 y 9.15, la solución aproximada satiface:

$$y_{k+1} = y_k + hf(t_{k+1/2}, y_k + \frac{h}{2} f(t_k, y_k))$$

o, de manera equivalente

$$\frac{y_{k+1} - y_k}{h} = f(t_{k+1/2}, y_k + \frac{h}{2} f(t_k, y_k))$$

y, a partir de 9.16, la solución exacta satisface: $\frac{y(t_{k+1}) - y(t_k)}{h} = f(t_{k+1/2}, y(t_k) + \frac{h}{2} f(t_k, y(t_k))) + O(h^2)$ el error de truncamiento local en el método del punto medio es $O(h^2)$; es decir, el método es de precisión de segundo orden.

2.3. Métodos basados en fórmulas de cuadratura

Integrando la ecuación diferencial 9.1 de t a $t+h$ obtenemos

$$y(t+h) = y(t) + \int_t^{t+h} f(s, y(s)) ds \quad (9.17)$$

La integral al lado derecho de la ecuación puede aproximarse con cualquier fórmula de cuadratura vista anteriormente. Por ejemplo, usando la regla del trapecio para aproximar la integral,

$$\int_t^{t+h} f(s, y(s)) ds = \frac{h}{2} [f(t, y(t)) + f(t+h, y(t+h))] + O(h^3)$$

conduce al método del trapecio para resolver el PVI 9.1,

$$y_{k+1} = y_k + \frac{h}{2} [f(t_k, y_k) + f(t_{k+1}, y_{k+1})] \quad (9.18)$$

Esto se llama un método implícito, ya que el nuevo valor y_{k+1} aparece en ambos lados de la ecuación. Para determinar y_{k+1}, se debe resolver una ecuación no lineal. Dado que el error en la aproximación de la regla del trapecio para la integral es $O(h^3)$, el error de truncamiento local para este método es $O(h^2)$.

Para evitar resolver la ecuación no lineal en el método del trapecio, se puede utilizar el método de Heun, que primero estima y_{k+1} utilizando el método de Euler y luego usa esta estimación en el lado derecho de 9.18:

$$\tilde{y}_{k+1} = y_k + hf(t_k, y_k)$$
$$y_{k+1} = y_k + \frac{h}{2} [f(t_k, y_k) + f(t_{k+1}, \tilde{y}_{k+1})]$$

El método de Heun sigue una línea cuya pendiente es el promedio de la pendiente de la curva solución en (t_k, y_k) y la pendiente de la curva solución en $(t_{k+1}, \tilde{y}_{k+1})$, donde $\tilde{y}_{k+1}$ es el resultado de un paso con el método de Euler.

2.4. Método de Runge - Kutta

Los métodos Runge-Kutta tienen el error de truncamiento local de orden superior a los métodos de Taylor, pero eliminan la necesidad de calcular y evaluar las derivadas de $f(t, y)$.

Teorema 2.3. Suponga que $f(t, y)$ y todas sus derivadas parciales de orden menor o igual a $n+1$ son continuas en $D = (t, y) | a \leq t \leq b, c \leq y \leq d$ y si $(t_0, y_0) \in D$. Para cada $(t, y) \in D$, existe ξ entre t y t_0 y μ entre y y y_0 con

$$f(t, y) = P_n(t, y) + R_n(t, y)$$

donde

$$\begin{aligned} P_n(t,y) = f(t_0,y_0) &+ \left[(t-t_0) \frac{\partial f}{\partial t}(t_0,y_0) + (y-y_0) \frac{\partial f}{\partial y}(t_0,y_0) \right] \\ &+ \left[\frac{(t-t_0)^2}{2} \frac{\partial^2 f}{\partial t^2}(t_0,y_0) + (t-t_0)(y-y_0) \frac{\partial^2 f}{\partial t \partial y}(t_0,y_0) \right. \\ &\left. + \frac{(y-y_0)^2}{2} \frac{\partial^2 f}{\partial y^2}(t_0,y_0) \right] + \dots \\ &+ \frac{1}{n!} \sum_{j=0}^{n} \binom{n}{j} (t-t_0)^{n-j} (y-y_0)^j \frac{\partial^n f}{\partial t^{n-j} \partial y^j}(t_0,y_0). \end{aligned}$$

y

$$R_n(t,y) = \frac{1}{(n+1)!} \sum_{j=0}^{n+1} \binom{n+1}{j} (t-t_0)^{n+1-j} (y-y_0)^j \frac{\partial^{n+1} f}{\partial t^{n+1-j} \partial y^j}(\xi, \mu)$$

La función $P_n(t, y)$ recibe el nombre del emésimo polinomio de Taylor en dos variables para la función f cerca de (t_0, y_0), y $R_n(t, y)$ es el término restante asociado con $P_n(t, y)$.

Métodos de Runge-Kutta de orden 2

El primer paso para deducir un método Runge-Kutta es determinar los valores para a_i, α_1, β_1 con la propiedad de que $a_1 f(t`\alpha_1, y + \beta_1)$ se aproxima a

$$T^{(2)}(t,y) = f(t,y) + \frac{h}{2} f'(t,y)$$

con error no mayor a $O(h^2)$, que es igual al orden del error de truncamiento local para el método de Taylor de orden 2. Ya que

$$f'(t,y) = \frac{df}{dt}(t,y) = \frac{\partial f}{\partial t}(t,y) + \frac{\partial f}{\partial y}(t,y) \cdot y'(t)$$

$$y'(t) = f(t,y)$$

tenemos

$$T^{(2)}(t,y) = f(t,y) + \frac{h}{2}\frac{\partial f}{\partial t}(t,y) + \frac{h}{2}\frac{\partial f}{\partial y}(t,y) \cdot f(t,y) \tag{9.19}$$

Al expandir $f(t+\alpha_1, y+\beta_1)$ en su polinomio de Taylor de grado 1, cerca de (t,y) obtenemos

$$\alpha_1 f(t+\alpha_1, y+\beta_1) = \alpha_1 f(t,y) + a_1\alpha_1 \frac{\partial f}{\partial t}(t,y) + a_1\beta_1 \frac{\partial f}{\partial y}(t,y) + a_1 \cdot R_1(t+\alpha_1, y+\beta_1) \tag{9.20}$$

donde

$$R_1(t+\alpha_1, y+\beta_1) = \frac{\alpha_1^2}{2}\frac{\partial^2 f}{\partial t^2}(\xi,\mu) + \alpha_1\beta_1\frac{\partial^2 f}{\partial t \partial y}(\xi,\mu) + \frac{\beta_1^2}{2}\frac{\partial^2 f}{\partial y^2}(\xi,\mu) \tag{9.21}$$

para algunas ξ entre t y $t+\alpha_1$ y μ entre y y $y+\beta_1$.

Al ajustar los coeficientes de ff y sus derivadas en las ecuaciones 9.19 y 9.20 obtenemos las tres ecuaciones

$$f(t,y) : a_1 = 1$$

$$\frac{\partial f}{\partial t}(t,y) : a_1\alpha_1 = \frac{h}{2}$$

$$\frac{\partial f}{\partial y}(t,y) : a_1\beta_1 = \frac{h}{2} f(t,y)$$

Los parámetos son por tanto: $a_1 = 1$, $\alpha_1 = \frac{h}{2}$ y $\beta_1 = \frac{h}{2} f(t,y)$; por lo que $T^{(2)}(t,y) = f\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right) - R_1\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right)$

y, a partir de la ecuación 9.21

$$R_1\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right) = \frac{h^2}{8}\frac{\partial^2 f}{\partial t^2}(\xi,\mu) + \frac{h^2}{4}f(t,y)\frac{\partial^2 f}{\partial t \partial y}(\xi,\mu) + \frac{h^2}{8}(f(t,y))^2\frac{\partial^2 f}{\partial y^2}(\xi,\mu)$$

Si todas las derivadas parciales de segundo orden de f están acotadas, entonces

$$R_1\left(t+\frac{h}{2}, y+\frac{h}{2}f(t,y)\right)$$

es $O(h^2)$. En consecuencia, el orden de error para este nuevo método es igual al del método de Taylor de orden 2.

El método de ecuación de diferencia que resulta de reemplazar $T^{(2)}(t,y)$ en el método de Taylor de orden 2 por $f(t+(h/2), y+(h/2)f(t,y))$ es un método Runge-Kutta específico, conocido como método de punto medio.

Método de punto medio

$$w_0 = \alpha$$

$$w_{i+1} = w_i + hf(t_i, w_i), \quad \forall i = 0, 1, ..., N-1$$

Solamente se encuentran tres parámetros en $a_1 f(t+\alpha_1, y+\beta_1)$, y todos son necesarios para ajustar $T^{(2)}$. Por lo que se requiere una forma más complicada para satisfaces las condiciones para cualquier de los métodos de Taylor de orden superior.

La forma de cuatro parámetros más adecuada para aproximar

$$T^{(3)}(t,y) = f(t,y) + \frac{h}{2}f'(t,y) + \frac{h}{6}f''(t,y)$$

es

$$a_1 f(t,y) + a_2 f(t+\alpha_2, y+\delta_2 f(t,y)) \tag{9.22}$$

e incluso con esto, no hay suficiente flexibilidad para ajustar el término

$$\frac{h^2}{6}\left[\frac{\partial f}{\partial y}(t,y)\right]^2 f(t,y)$$

lo cual resulta en la expansión de $(h^2/6)f''(t,y)$. Por consiguiente, lo mejor que se puede obtener al usar 9.22 son métodos con error de truncamiento local $O(h^2)$.

Sin embargo, el hecho de que 9.22 tenga cuatro parámetos proporciona una flexibilidad en su elección, por lo que se puede derivar una serie de métodos $O(h^2)$. Uno de los más importantes es el método modificado de Euler, que corresponde a seleccionar $a_1 = a_2 = \frac{1}{2}$ y $\alpha_2 = \delta_2 = h$. Éste tiene la siguiente forma de ecuación de diferencia.

Méotodo modificado de Euler

$$w_0 = \alpha$$

$$w_{i+1} = w_i + \frac{h}{2}\left[f(t_i, w_i) + f(t_{i+1}, w_i + hf(t_i, w_i))\right]$$

$$\forall i = 0, 1, ..., N-1$$

Métodos de Runge-Kutta de orden superior

El término $T^{(3)}(t,y)$ se puede aproximar con error $O(h^3)$ mediante una expresión de la forma

$$f(t+\alpha_1, y+\delta_1 f(t+\alpha 2, y+\delta_2 f(t,y)))$$

relacionada con cuatro parámetros y el álgebra implicada en la determinaciónd de α_1, δ_1, α_2 y δ_2 es bastante tediosa. El método $O(h^3)$ más común es el de Heun, dado por

$$w_0 = \alpha$$

$$w_{i+1} = w_i + \tfrac{h}{4}\left[f(t_i,w_i) + 3\left(f\left(t_i+\tfrac{2h}{3}, w_i+\tfrac{2h}{3}f\left(t_i+\tfrac{h}{3}, w_i+\tfrac{h}{3}f(t_i,w_i)\right)\right)\right)\right]$$

$$\forall i = 0,1,...,N-1$$

En general, los métodos de Runge-Kutta de orden 3 no se usan. El método Runge-Kutta que se usa de manera común es de orden 4 en forma de ecuación de diferencia, dado como sigue.

Runge-Kutta de orden 4

$$w_0 = \alpha$$
$$k_1 = hf(t_i, y_i)$$
$$k_2 = hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_1\right)$$
$$k_3 = hf\left(t_i + \frac{h}{2}, w_i + \frac{1}{2}k_2\right)$$
$$k_4 = hf(t_{i+1}, w_i + k_3)$$
$$w_{i+1} = w_i + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$$

para cada $i = 0,1,...,N-1$. Este método tiene error de truncamiento local $O(h^4)$, siempre y cuando la solución $y(t)$ tenga cinco derivadas continuas. Introducimos en el método la notación k_1, k_2, k_3 y k_4 para eliminar la necesidad de anidado sucesivo en la segunda variable de $f(t,y)$.

2.5. Analisis de métodos de un paso

Un método general explícito de un solo paso se puede escribir de la forma:

$$y_{k+1} = y_k + h\psi(t_k, y_k, h) \tag{9.23}$$

- Método de Euler

$$y_{k+1} = y_k + hf(t_k, y_k) \to \psi(t,y,h) = f(t,y)$$

- Método de Heun

$$y_{k+1} = y_k + \frac{h}{2}[f(t_k,y_k) + f(t_{k+1}, y_k + hf(t_k,y_k))] \to$$

$$\psi(t,y,h) = \frac{1}{2}[f(t,y) + f(t+h, y+hf(t,y))]$$

- Método del Medio Punto

$$y_{k+1} = y_k + hf(t_{k+1/2}, y_k + \frac{h}{2}f(t_k,y_k)) \to$$

$$\psi(t,y,h) = f(t+\frac{h}{2}, y+\frac{h}{2}f(t,y))$$

Definición 2.1. El método de un paso 9.23 es consistente si $\lim_{h\to 0}\psi(t,y,h) = f(t,y)$

Definición 2.2. El método de un paso 9.23 es estable si existe una constante K y un tamaño de paso $h_0 > 0$ tal que la diferencia entre las dos soluciones y_n y $\tilde{y}_n$ con valores iniciales y_0 y $\tilde{y}_0$, respectivamente, satiface

$$|y_n - \tilde{y}_n| \le K|y_0 - \tilde{y}_0|$$

para $h \le h_0$ y $nh \le T - t_0$

Definición 2.3. El error de truncamiento es

$$\tau(t,h) = \frac{y(t+h) - y(t)}{h} - \psi(t, y(t), h)$$

Hemos visto que para el método de Euler, el error local de truncamiento es $O(h)$ y para el método de Heun y el método del medio punto el error local de truncamiento es $O(h^2)$

Teorema 2.4. Si un método de un paso de la forma 9.23 es estable y consistente y $|\tau(t,h)| \le Ch^p$, entonces el error global está acotado por

$$\max_{k:t_0+kh\le T}|y_k - y(t_k)| \le Ch^p\frac{e^{L(T-t_0)}-1}{L} + e^{L(T-t_0)}|y_0 - y(t_0)|$$

donde L es la constante de Lipschitz de ψ.

Definición 2.4. El método de un paso 9.23 es convergente si, para todo PVI bien planteado, $\max_{k:t_k\in[t_o,T]}|y(t_k) - y_k| \to 0$ con $y_0 \to y(t_0)$ y $h \to 0$

3. Ecuaciones Stiff

Hasta ahora, nos hemos enfocado en lo que ocurre en el límite cuando $h \to 0$. Para que un método sea útil, ciertamente debe converger a la solución exacta conforme $h \to 0$. Sin embargo, en la práctica, utilizamos un tamaño de paso h fijo y no nulo, o al menos existe un límite inferior para h basado en el tiempo permitido para el cálculo y, posiblemente, en la precisión de la máquina, ya que, por debajo de cierto punto, los errores de redondeo comenzarán a causar inexactitudes. Por lo tanto, nos gustaría

entender cómo se comportan los diferentes métodos con un tamaño de paso h fijo. Esto, por supuesto, dependerá del problema, pero deseamos usar métodos que proporcionen resultados razonables con un tamaño de paso h fijo para la mayor cantidad posible de clases de problemas.

3.1. Estabilidad Absoluta

Para analizar el comportamiento de un método con un tamaño de paso h particular, se podría considerar una ecuación de prueba muy simple:

$$y' = \lambda y \tag{9.24}$$

donde λ es una constante compleja. La solución es $y(t) = e^{\lambda t}y(0)$. Ejemplos de soluciones se muestran en la figura 11.9 para los casos en que la parte real de λ es mayor que 0, igual a 0 y menor que 0. Nótese que $y(t) \to 0$ cuando $t \to \infty$ si y solo si $\mathrm{Re}(\lambda) < 0$, donde $\mathrm{Re}(\cdot)$ denota la parte real.

Definición 3.1. La región de estabilidad absoluta de un método es el conjunto de todos los números $h\lambda \in \mathbb{C}$ tales que $y_k \to 0$ cuando $k \to \infty$, al aplicar el método a la ecuación de prueba 9.24 usando un tamaño de paso h.

Definición 3.2. Un método es A-estable si su región de estabilidad absoluta contiene todo el semiplano izquierdo.

3.2. Métodos de Runge–Kutta Implícitos (IRK)

Otra clase de métodos útiles para ecuaciones rígidas son los métodos de Runge–Kutta implícitos (IRK). Estos tienen la forma:

$$\xi_j = y_k + h\sum_{i=1}^{v} a_{ij} f(t_k + c_i h, \xi_i) \quad j = 1, \ldots, v \tag{9.25}$$

$$y_{k+1} = y_k + h\sum_{j=1}^{v} b_j f(t_k + c_j h, \xi_j), \tag{9.26}$$

donde los valores a_{ij}, b_j y c_j pueden elegirse arbitrariamente, pero para garantizar la consistencia se requiere que:

$$\sum_{i=1}^{v} a_{ji} = c_j, \quad j = 1, ..., v$$

Se puede demostrar que para cada $\nu \geq 1$, existe un único método IRK de orden 2ν, y este es A-estable. El método IRK de orden 2ν se obtiene tomando los valores $c_1, \ldots, c_\nu$ como las raíces del ν-ésimo polinomio ortogonal en $[0, 1]$.

El método IRK de orden 2 ($\nu = 1$) es el método del trapecio 9.18. El método IRK de orden 4 ($\nu = 2$) tiene $c_1 = 1/2 - \sqrt{3}/6$ y $c_2 = 1/2 + \sqrt{3}/6$ como las raíces del segundo polinomio ortogonal en $[0, 1]$, y los demás parámetros se derivan para garantizar la precisión de cuarto orden.

Ejercicios

(1) Probar que los siguientes problemas de valor inicial están bien planteados:

a) $y' = -ty + 4t/y \quad 0 \leq t \leq 1 \quad y(0) = 1$

b) $y' = t^2 y + 1 \quad 0 \leq t \leq 1 \quad y(0) = 1$

c) $y' = t - y \quad 0 \leq t \leq 1 \quad y(0) = 1$

d) $y' = ty \quad 0 \leq t \leq 1 \quad y(0) = 1$

(2) Consultar la entrada List of Runge-Kutta methods en Wikipedia y escribir el método de Runge-Kutta para varias tablas de Butcher.

(3) Utilizar el método de Euler y los métodos de Runge-Kutta de orden 2 y orden 4 vistos en clase para aproximar la solución de los siguientes PVI:

a) $y' = (y/t)^2 + y/t \quad 1 \leq t \leq 1{,}2 \quad y(1) = 1$

b) $y' = sint + e^{-t} \quad 0 \leq t \leq 1 \quad y(0) = 0$

c) $y' = \frac{1}{t}(y^2 + y) \quad 0 \leq t \leq 3 \quad y(1) = -2$

d) $y' = t^2 \quad 0 \leq t \leq 2 \quad y(0) = 0$

Además, aproximar las soluciones, en cada caso, con el comando ode45 de Matlab y decidir que método es más preciso.

(4) Utilizar los métodos de Runge-Kutta de orden 2 y 4 vistos para resolver los siguiente problemas de valor inicial:

a) $t^2y'' - 2ty' + 2y = t^3 lnt \quad 1 \leq t \leq 2 \quad y(1) = 1 \quad y'(1) = 0$

b) $y''' + 2y'' - y' - 2y = e^t \quad 0 \leq t \leq 3 \quad y(0) = 1 \quad y'(0) = 2 \quad y''(0) = 0$

(5) Probar que el método de Runge-Kutta de orden 4 es estable mostando que cuando se escribe de la forma

$$y_{k+1} = y_k + \psi(t_k, y_k, h)$$

entonces ψ satisface la condición de Lipschitz para y.

Bibliografía

[1] David Kincaid; Ward Cheney. *Numerical Analysis. Mathematics of Scientific Computing.*

[2] Richard L. Burden; J. Douglas Faires. *Numerical Analysis.*

[3] Richard L. Burden; J. Douglas Faires. *Numerical Analysis.*

[4] Robert M. Corless; Nicolas Fillion. *A Graduate Introduction to Numerical Methods, From the Viewpoint of Backward Error Analysis.*

[5] Jean-Luc Guermond. *Finite Element Interpolation.*

[6] Anne Greenbaum; Timothy P.Chartier. *Numerical Methods: Design, Analysis, and Computer Implementarion of Algorithms.* 212.

[7] A. Quarteroni; R. Sacco; F. Sale. *Numerical Mathematics.* Springer, 2007.

[8] Jorge Nocedal; Stephen J. Wright. *Numerical Optimization.*