Capítulo 1

Varios

1.1 Representación de números

Respresentación interna de números doble precisión, norma IEEE: 64 bits:

1 bit: s: signo: $(-1)^s$

11 bits: c: exponente o característica; $2^{11} = 2048$; $-1023 \leq c \leq 1024$

52 bits: f: mantisa: $1/2 \leq f < 1$, $2^{52} = 4.5 \times 10^{15}$: mas o menos 15 o 16 cifras significativas,

el primer dígito de la mantisa no es cero.

Supongamos que, en lugar de punto flotante binario, tenemos punto flotante decimal con k cifras significativas. El **truncamiento** se obtiene al suprimir de la mantisa las cifras $k+1$, $k+2$, ..., dejando únicamente las primeras k cifras significativas. El **redondeo** se obtiene sumando a la mantisa 0.5×10^{-k} y en seguida se trunca a k cifras significativas.

Por ejemplo, consideremos $e = 2.718281828459... = 0.2718281828459... \times 10^1$. Al truncar a 5 cifras significativas se obtiene $\bar{E} = 0.27182 \times 10^1$. Para redondear, 0.2718281828459... + .000005 = 0.2718331828459... y al truncar se obtiene el valor redondeado $\tilde{E} = .27183 \times 10^1$.

1.2 Épsilon de la máquina

Hay dos maneras de definir el épsilon de la máquina: un épsilon absoluto y un épsilon relativo. Este último es el más usado. Como el conjunto de números

usados en el computador es finito, la siguiente definición tiene sentido:

$$\varepsilon_{\text{maq}} = \varepsilon = \min\{t > 0 : 1 + t \neq 1\}$$

El épsilon absoluto se define comparando con cero:

$$\varepsilon^{\text{abs}} = \min\{t > 0 : t \neq 0\}.$$

En realidad el épsilon depende de la máquina pero también del sistema operativo, del compilador y del tipo de números utilizados. El siguiente ejemplo da dos aproximaciones del épsilon de la máquina y una aproximación del épsilon absoluto

```
double eps, uno, t, t1;

uno = 1.0;

t = 1.0;
while( 1.0+t != 1.0 ){
    eps = t;
    t /= 2.0;
}
cout<<" eps1 = "<<eps<<endl;

t = 1.0;
t1 = uno + t;
while( uno != t1 ){
    eps = t;
    t /= 2.0;
    t1 = uno + t;
}
cout<<" eps2 = "<<eps<<endl;

t = 1.0;
while( t != 0.0 ){
    eps = t;
    t /= 2.0;
}
cout<<" eps3 = "<<eps<<endl;
```

Con el compilador Borland `bcc32 5.2` para Windows se obtienen los siguientes resultados:

```
eps1 = 1.0842e-19
eps2 = 2.22045e-16
eps3 = 4.94066e-324
```

Los dos primeros valores "teóricamente" deberían ser iguales, pero el uso de las variables `uno` y `t1`, más parecido a la mayoría de los casos reales, hace la diferencia. Si se trabaja con números de precisión sencilla (`float`) se obtiene:

```
eps1 = 1.0842e-19
eps2 = 1.19209e-07
eps3 = 1.4013e-45
```

Se puede tener una aproximación mejor del épsilon si en lugar de dividir por 2.0 se divide por un número mayor que 1.0 pero cercano a 1.0, por ejemplo 1.01. Obviamente el proceso es mucho más demorado. Con números doble precisión se obtiene para los dos primeros valores:

```
eps1 = 5.43645ee-20
eps2 = 1.12153e-16
```

El tercer valor puede requerir mucho tiempo de cómputo o bloquear el computador. En resumen, los resultados anteriores muestran que

$$\varepsilon \approx 10^{-16}.$$

Esto quiere decir, que en estas condiciones hay aproximadamente 16 cifras significativas. Usualmente en un método iterativo no se exige una precisión mejor que $\sqrt{\varepsilon}$, es decir, 10^{-8}.

Capítulo 2

Solución de sistema un sistema lineal de ecuaciones

2.1 Sistemas tridiagonales

Un sistema $Ax = b$ se llama tridiagonal si la matriz A es tridiagonal, o sea, si

$$a_{ij} = 0 \text{ si } |i - j| > 1,$$

o sea, A es de la forma

$$A = \begin{bmatrix} a_{11} & a_{12} & 0 & 0 & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & 0 & & 0 \\ 0 & a_{32} & a_{33} & a_{34} & & 0 \\ 0 & 0 & a_{43} & a_{44} & & 0 \\ \vdots & & & & & \\ 0 & 0 & 0 & 0 & \cdots & a_{nn} \end{bmatrix}.$$

Estos sistemas se presentan en algunos problemas particulares, por ejemplo, al resolver, mediante diferencias finitas, una ecuación diferencial lineal de segundo orden con condiciones de frontera.

Obviamente este sistema se puede resolver mediante el método de Gauss. Pero dadas las características especiales es mucho más eficiente sacar provecho de ellas. Se puede mostrar que si A admite descomposición LU, entonces estas dos matrices también guardan la estructura de A, es decir, L, además de ser triangular inferior, tiene ceros por debajo de la "subdiagonal" y U,

además de ser triangular superior, tiene ceros por encima de la "superdiagonal".

Para simplificar, denotemos con f_i los elementos de la suddiagonal de L, d_i los elementos de la diagonal de U y u_i los elementos de la superdiagonal de U. Se conoce A y se desea conocer L y U a partir de la siguiente igualdad:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ f_1 & 1 & 0 & 0 & & 0 & 0 \\ 0 & f_2 & 1 & 0 & & 0 & 0 \\ 0 & 0 & f_3 & 1 & & 0 & 0 \\ & & & & \ddots & & \\ 0 & 0 & 0 & 0 & & 1 & 0 \\ 0 & 0 & 0 & 0 & & f_{n-1} & 1 \end{bmatrix} \begin{bmatrix} d_1 & u_1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & u_2 & 0 & & 0 & 0 \\ 0 & 0 & d_3 & u_3 & & 0 & 0 \\ 0 & 0 & 0 & d_4 & & 0 & 0 \\ & & & & \ddots & & \\ 0 & 0 & 0 & 0 & & d_{n-1} & u_{n-1} \\ 0 & 0 & 0 & 0 & & 0 & d_n \end{bmatrix} = A \ .$$

Sean F_i la fila i de L y C_j la columna j de U. Entonces los productos de las filas de L por las columnas de U producen las siguientes igualdades:

$$\begin{aligned} F_1C_1 : &\quad d_1 = a_{11} \\ F_1C_2 : &\quad u_1 = a_{12} \\ F_2C_1 : &\quad f_1d_1 = a_{21} \\ F_2C_2 : &\quad f_1u_1 + d_2 = a_{22} \\ F_2C_3 : &\quad u_2 = a_{23} \\ F_3C_2 : &\quad f_2d_2 = a_{32} \\ F_3C_3 : &\quad f_2u_2 + d_3 = a_{33} \\ F_3C_4 : &\quad u_3 = a_{34} \\ &\ \vdots \\ F_iC_{i-1} : &\quad f_{i-1}d_{i-1} = a_{i,i-1} \\ F_iC_i : &\quad f_{i-1}u_{i-1} + d_i = a_{ii} \\ F_iC_{i+1} : &\quad u_i = a_{i,i+1} \end{aligned}$$

A partir de las igualdades anteriores se obtienen los valores u_i, f_i y d_i:

$$\begin{aligned} d_1 &= a_{11}, \\ \\ u_i &= a_{i,i+1}\,, \qquad i = 1, ..., n-1, \\ f_i &= \frac{a_{i+1,i}}{d_i}, \\ d_{i+1} &= a_{i+1,i+1} - f_i u_i \end{aligned} \tag{2.1}$$

Resolver $Ax = b$ es equivalente a resolver $LUx = b$. Entonces, si $Ux = y$, se resuelve $Ly = b$ y después $Ux = y$. Al explicitar las anteriores igualdades se tiene:

$$\begin{aligned} y_1 &= b_1, \\ f_{i-1}y_{i-1} + y_i &= b_i, \\ d_n x_n &= y_n, \\ d_i x_i + u_i x_{i+1} &= y_i\,. \end{aligned}$$

Las fórmulas explícitas son:

$$\begin{aligned} y_1 &= b_1, \\ y_i &= b_i - f_{i-1}y_{i-1}, \quad i = 2, ..., n, \\ x_n &= \frac{y_n}{d_n}, \\ x_i &= \frac{y_i - u_i x_{i+1}}{d_i}\,, \qquad i = n-1, n-2, ..., 1. \end{aligned} \tag{2.2}$$

Ejemplo 2.1. Resolver el sistema $Ax = b$, con

$$A = \begin{bmatrix} 2 & 4 & 0 & 0 \\ 3 & 5 & 6 & 0 \\ 0 & -4 & -5 & 1 \\ 0 & 0 & -1 & -2 \end{bmatrix}, \qquad b = \begin{bmatrix} -8 \\ 1 \\ -2 \\ -10 \end{bmatrix}.$$

Entonces

$$\begin{aligned} d_1 &= 2\,, \\ u_1 &= 4\,, \end{aligned}$$

$$
\begin{aligned}
f_1 &= \frac{3}{2} = 1.5\,,\\
d_2 &= 5 - 1.5 \times 4 = -1\,,\\
u_2 &= 6\,,\\
f_2 &= \frac{-4}{-1} = 4\,,\\
d_3 &= -5 - 4 \times 6 = -29\,,\\
u_3 &= 1,\\
f_3 &= \frac{-1}{-29} = 0.034483\,,\\
d_4 &= -2 - 0.034483 \times 1 = -2.034483\,,
\end{aligned}
$$

Ahora la solución de los sistemas $Ly = b,\ Ux = y$:

$$
\begin{aligned}
y_1 &= -8,\\
y_2 &= 1 - 1.5 \times (-8) = 13\,,\\
y_3 &= -2 - 4 \times 13 = -54\,,\\
y_4 &= -10 - 0.034483 \times -54 = -8.137931\,,\\
x_4 &= \frac{-8.137931}{-2.034483} = 4\,,\\
x_3 &= \frac{-54 - 1 \times 4}{-29} = 2\,,\\
x_2 &= \frac{13 - 6 \times 2}{-1} = -1\,,\\
x_1 &= \frac{-8 - 4 \times (-1)}{2} = -2\,.\ \diamond
\end{aligned}
$$

Las fórmulas (2.1) y (2.2) se pueden utilizar sin ningún problema si todos los d_i son no nulos. Algún elemento diagonal de U resulta nulo si la matriz A no es invertible o si simplemente A no tiene factorización LU.

Ejemplo 2.2. Consideremos las dos matrices siguientes:

$$
A = \begin{bmatrix} 2 & -3 \\ -8 & 12 \end{bmatrix}, \qquad A' = \begin{bmatrix} 0 & 2 \\ 3 & 4 \end{bmatrix}.
$$

La matriz A no es invertible y d_2 resulta nulo. La matriz A' es invertible pero no tiene factorización LU. En este último caso, se obtiene $d_1 = 0$. $\diamond$

Si la matriz A es grande no se justifica almacenar todos los n^2 elementos. Basta con almacenar la diagonal, la subdiagonal y la superdiagonal, es decir $3n-2$ números. Mejor aún, en el mismo sitio donde inicialmente se almacenan los elementos diagonales de A se pueden almacenar los elementos diagonales de U a medida que se van calculando, donde se almacenan los elementos subdiagonales de A se pueden almacenar los elementos subdiagoanles de L, los elementos superdiagonales de A son los mismos elementos superdiagonales de U, donde se almacena b se puede almacenar y y posteriormente x.

En resumen, una implementación eficiciente utiliza 4 vectores d, f, u y b. El primero y el cuarto están en $\mathbb{R}^n$, los otros dos están en $\mathbb{R}^{n-1}$. Al comienzo contienen datos de A y b, al final contienen datos de L, U y la solución final x.

SOLUCIÓN DE SISTEMA TRIDIAGONAL

datos: d, f, u, b, ε

si $|d_1| \leq \varepsilon$ **ent parar**
para $i = 1, ..., n-1$
    $f_i = \dfrac{f_i}{d_i}$
    $d_{i+1} = d_{i+1} - f_i * u_i$
    si $|d_{i+1}| \leq \varepsilon$ **ent parar**
fin-para
para $i = 2, ..., n$
    $b_i = b_i - f_{i-1} b_{i-1}$
fin-para
$b_n = \dfrac{b_n}{d_n}$
para $i = n-1, n-2, ..., 1$
    $b_i = \dfrac{b_i - u_i b_{i+1}}{d_i}$
fin-para

2.2 Factorización de Cholesky

Teorema 2.1. *Sea $A \in S_n$ (conjunto de matrices simétricas $n \times n$). A es definida positiva sssi A tiene factorización de Cholesky.*

Demostración.
$\Leftarrow$)
Si A tiene factorización de Cholesky, entonces existe U triangular superior invertible tal que $A = U^{\mathrm{T}}U$.

$$\begin{aligned} x^{\mathrm{T}}Ax &= x^{\mathrm{T}}U^{\mathrm{T}}Ux \\ &= (Ux)^{\mathrm{T}}(Ux) \\ &= ||Ux||_2^2 \\ &\geq 0, \ \ \forall x. \end{aligned}$$

Además $x^{\mathrm{T}}Ax = 0$ sssi $Ux = 0$. Como U es invertible, entonces $x^{\mathrm{T}}Ax = 0$ sssi $x = 0$. Luego $x^{\mathrm{T}}Ax > 0$ para todo $x \neq 0$, es decir, A es definida positiva.

$\Rightarrow$)
Por inducción sobre n, tamaño de la matriz. Para $n = 1$, $A = [a_{11}]$. Entonces $x^{\mathrm{T}}Ax = x_1 a_{11} x_1 = a_{11}x_1^2$, luego A es definida positiva sssi $a_{11} > 0$. Entonces $U = [\,\sqrt{a_{11}}\,]$. Ahora supongamos cierto para cualquier matriz simétrica de tamaño menor o igual a $n-1$. Ahora sea A simétrica $n \times n$. A se puede expresar por bloques

$$A = \begin{bmatrix} \tilde{A} & a \\ a^{\mathrm{T}} & \alpha \end{bmatrix},$$

donde $\tilde{A} \in S_{n-1}$, $a \in \mathbb{R}^{(n-1)\times 1}$, $\alpha \in \mathbb{R}$. Veamos que $\tilde{A}$ es definida positiva. Supongamos que no lo es. Entonces existe $\tilde{x} \in \mathbb{R}^{n-1}$, $\tilde{x} \neq 0$, tal que $\tilde{x}^{\mathrm{T}}\tilde{A}\tilde{x} \leq 0$. Sea $\xi = (\tilde{x}, 0) \in \mathbb{R}^n$, $\xi \neq 0$,

$$\xi^{\mathrm{T}}A\xi = \begin{bmatrix} \tilde{x}^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} \tilde{A} & a \\ a^{\mathrm{T}} & \alpha \end{bmatrix} \begin{bmatrix} \tilde{x} \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{x}^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} \tilde{A}\tilde{x} \\ a^{\mathrm{T}}\tilde{x} \end{bmatrix} = \tilde{x}^{\mathrm{T}}\tilde{A}\tilde{x} \leq 0$$

Pero A es definida positiva y $\xi \neq 0$. Luego lo supuesto no puede ser verdadero.

Como $\tilde{A}$ es definida postiva, entonces por hipótesis de inducción $\tilde{A} = \tilde{U}^{\mathrm{T}}\tilde{U}$. Sea

$$U = \begin{bmatrix} \tilde{U} & u \\ 0 & w \end{bmatrix},$$

Veamos que u y w se pueden obtener para que $A = U^{\mathrm{T}}U$.

$$\begin{bmatrix} \tilde{A} & a \\ a^{\mathrm{T}} & \alpha \end{bmatrix} = \begin{bmatrix} \tilde{U}^{\mathrm{T}} & 0 \\ u^{\mathrm{T}} & w \end{bmatrix} \begin{bmatrix} \tilde{U} & u \\ 0 & w \end{bmatrix}.$$

De la igualdad anterior se tiene:

$$\begin{aligned} \tilde{A} &= \tilde{U}^{\mathrm{T}}\tilde{U} \quad \surd \\ a &= \tilde{U}^{\mathrm{T}}u \\ \alpha &= u^{\mathrm{T}}u + w^2. \end{aligned}$$

Como $\tilde{U}$ es invertible, se puede obtener u y después w:

$$\begin{aligned} u &= (\tilde{U})^{-1}a \\ w &= \sqrt{\alpha - u^{\mathrm{T}}u}. \end{aligned}$$

Solamente falta por mostrar que $\alpha - u^{\mathrm{T}}u > 0$. Sean

$$v = -\tilde{U}^{-1}u, \qquad \zeta = \begin{bmatrix} \tilde{v} \\ 1 \end{bmatrix} \neq 0.$$

Entonces

$$\begin{aligned} \zeta^{\mathrm{T}}A\zeta &= \begin{bmatrix} v^{\mathrm{T}} & 1 \end{bmatrix} \begin{bmatrix} \tilde{A} & a \\ a^{\mathrm{T}} & \alpha \end{bmatrix} \begin{bmatrix} v \\ 1 \end{bmatrix} \\ &= v^{\mathrm{T}}\tilde{A}v + 2a^{\mathrm{T}}v + \alpha \\ &= (-\tilde{U}^{-1}u)^{\mathrm{T}}\tilde{U}^{\mathrm{T}}\tilde{U}(-\tilde{U}^{-1}u) + 2a^{\mathrm{T}}(-\tilde{U}^{-1}u) + \alpha \\ &= u^{\mathrm{T}}u - 2a^{\mathrm{T}}\tilde{U}^{-1}u + \alpha \\ &= u^{\mathrm{T}}u - 2(\tilde{U}^{\mathrm{T}}u)^{\mathrm{T}}\tilde{U}^{-1}u + \alpha \\ &= u^{\mathrm{T}}u - 2u^{\mathrm{T}}u + \alpha \\ &= \alpha - u^{\mathrm{T}}u > 0 \end{aligned}$$

por ser A definida positiva.

2.2.1 Relación entre factorizaciones de Cholesky y LU

Sean

$$A = U^{\mathrm{T}}U \quad \text{la factorización de Cholesky}$$

$$A = LU_G \quad \text{la factorización } LU \text{(en el método de Gauss)},$$

entonces

$$L = U^{\mathrm{T}}\Delta^{-1}$$
$$U_G = \Delta U,$$

donde

$$\Delta = \texttt{diag}(\texttt{diag}(U)).$$

2.3 Normas matriciales

En el conjunto de matrices cuadradas de orden n se puede utilizar cualquier norma definida sobre $\mathbb{R}^{n^2}$. Dado que en el conjunto de matrices cuadradas está definido el producto, es interesante contar con normas con características especiales relativas al producto.

Una norma $||\ \ ||$ definida sobre el $\mathcal{M}(n,n)$ (conjunto de matrices $n \times n$) se llama **matricial** si (además de las propiedades usuales de norma) para cualquier par de matrices A y B

$$||AB|| \leq ||A||\,||B||.$$

Sean $||\ \ ||_m$ una norma matricial sobre $\mathcal{M}(n,n)$ y $||\ \ ||_v$ una norma sobre $\mathbb{R}^n$. Estas dos normas se llaman compatibles si, para toda matriz $A \in \mathcal{M}(n,n)$ y para todo $x \in \mathbb{R}^n$

$$||Ax||_v \leq ||A||_m ||x||_v\,.$$

Una manera común de construir normas que sean matriciales y compatibles es generando una norma a partir de un norma sobre $\mathbb{R}^n$. Sea $||\ \ ||$ una norma sobre $\mathbb{R}^n$. La **norma generado o inducida** por esta norma se define de varias maneras, todas ellas equivalentes:

$$|||A||| = \sup_{x \neq 0} \frac{||Ax||}{||x||} \tag{2.3}$$

$$|||A||| = \max_{x \neq 0} \frac{||Ax||}{||x||} \tag{2.4}$$

$$|||A||| = \max_{||x||=1} ||Ax||. \tag{2.5}$$

Se puede mostrar que la definición está bien hecha, es decir, que $|||\ \ ||||$ es una norma. Además es matricial. También se tiene que $|||\ \ |||$ y $||\ \ ||$ son compatibles.

Para las 3 normas vectoriales más usadas, las normas matriciales generadas son:

$$|||A|||_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|, \tag{2.6}$$

$$|||A|||_2 = \sqrt{\rho(A^{\mathrm{T}}A)}\ \text{(norma espectral)}, \tag{2.7}$$

$$|||A|||_\infty = \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|. \tag{2.8}$$

Si la matriz A se considera como un vector, entonces se puede aplicar la norma euclidiana. Esta norma resulta ser matricial. Esta norma se conoce con el nombre de norma de Frobenius o también de Schur.

$$||A||_F = \left(\sum_{i,j} (a_{ij})^2 \right)^{1/2}. \tag{2.9}$$

Para cualquier norma generada $|||I||| = 1$. Como $||I||_F = \sqrt{n}$, entonces esta norma no puede ser generada por ninguna norma vectorial

Ejemplo 2.3. Sea

$$A = \begin{bmatrix} 1 & 2 \\ 3 & -4 \end{bmatrix}$$

Entonces

$$A^{\mathrm{T}}A = \begin{bmatrix} 10 & -10 \\ -10 & 20 \end{bmatrix}$$

Sus valores propios son 3.8196601 y 26.18034. Luego

$$|||A|||_1 = 6, \tag{2.10}$$
$$|||A|||_2 = 5.1166727, \tag{2.11}$$
$$|||A|||_\infty = 7. \tag{2.12}$$

2.4 Método de sobrerrelajación

Este método, conocido como SOR (Successive Over Relaxation), se puede considerar como una generalización del método GS. La fórmulas que definen el método GS son:

$$\begin{aligned} r_i &= b_i - A_i.x\,, \\ \delta_i &= \frac{r_i}{a_{ii}}\,, \\ x_i &= x_i + \delta_i\,. \end{aligned}$$

El el método SOR únicamente cambia la última asignación, introduciendo un parámetro ω,

$$\begin{aligned} r_i &= b_i - A_i.x\,, \\ \delta_i &= \frac{r_i}{a_{ii}}\,, \\ x_i &= x_i + \omega\delta_i\,. \end{aligned} \tag{2.13}$$

Si $0 < \omega < 1$ se tiene una subrrelajación, si $1 < \omega$ se tiene la sobrerrelajación propiamente dicha. Si $\omega = 1$, se tiene el método GS. Una escogencia adecuada de ω mejora la convergencia del método GS. Este método se usa en algunas técnicas de solución de ecuaciones diferenciales parciales.

Una condición necesaria para que el método SOR converja, ver [Dem97], es que

$$0 < \omega < 2\,.$$

Para matrices definidas positivas el método SOR converge para cualquier ω en el intervalo $]0, 2[$.

Ejemplo 2.4. Resolver el sistema $Ax = b$ por el método SOR con $\omega = 1.4$ partiendo de $x^0 = (1, 1, 1, 1)$.

$$A = \begin{bmatrix} 5 & -1 & 2 & -2 \\ 0 & 4 & 2 & 3 \\ 3 & 3 & 8 & -2 \\ -1 & 4 & -1 & 6 \end{bmatrix}, \qquad b = \begin{bmatrix} 25 \\ -10 \\ 35 \\ -33 \end{bmatrix}.$$

Entonces

$$r_1 = b_1 - A_1.x = 25 - 4 = 21$$

$$\delta_1 = \frac{21}{5} = 4.2$$
$$\omega\delta_1 = 5.88$$
$$x_1 = 1 + 5.88 = 6.88$$

$$r_2 = -10 - 9 = -19$$
$$\delta_2 = \frac{-19}{4} = -4.75$$
$$\omega\delta_2 = -6.65$$
$$x_2 = 1 - 6.65 = -5.65$$

$$r_3 = 35 - 9.69 = 25.31$$
$$\delta_3 = \frac{25.31}{8} = 3.163750$$
$$\omega\delta_3 = 4.429250$$
$$x_3 = 1 + 4.429250 = 5.429250$$

$$r_4 = -33 - -28.909250 = -4.090750$$
$$\delta_4 = \frac{-4.090750}{6} = -0.681792$$
$$\omega\delta_4 = -0.954508$$
$$x_4 = 1 - 0.954508 = 0.045492$$

$$r_1 = 25 - 50.817517 = -25.817517$$
$$\delta_1 = \frac{-25.817517}{5} = -5.163503$$
$$\omega\delta_1 = -7.228905$$
$$x_1 = 6.880000 + -7.228905 = -0.348905$$

La siguiente tabla muestra las primeras 15 iteraciones completas

Sobrerrelajación, $\omega = 1.4$.

k	x_1	x_2	x_3	x_4
0	1.000000	1.000000	1.000000	1.000000
1	6.880000	-5.650000	5.429250	0.045492
2	-0.348905	-5.088241	6.823724	-1.458380
3	1.076876	-4.710011	4.792473	-1.351123
4	1.810033	-3.552048	4.649676	-2.337041
5	1.368852	-2.880061	4.240550	-2.768266
6	1.721105	-2.409681	3.821389	-3.050409
7	1.788640	-2.008170	3.644054	-3.337915
8	1.812353	-1.742759	3.462571	-3.507443
9	1.883878	-1.543881	3.333868	-3.638593
10	1.909584	-1.395632	3.248121	-3.738508
11	1.932877	-1.289998	3.179762	-3.807650
12	1.952699	-1.211802	3.131447	-3.859624
13	1.964616	-1.154687	3.096340	-3.897553
14	1.974261	-1.113133	3.070228	-3.925007
15	1.981287	-1.082649	3.051371	-3.945238

La tabla siguiente muestra los resultados de la solución del mismo sistema por el método GS. La solución exacta es $x = (2, -1, 3, -4)$. Se aprecia que en la iteración 15 se tiene una mejor aproximación de la solución con el método de sobrerrelajación.

Gauss-Seidel

k	x_1	x_2	x_3	x_4
0	1.000000	1.000000	1.000000	1.000000
1	5.200000	-3.750000	4.081250	-1.453125
2	2.036250	-3.450781	4.542168	-2.103076
3	1.651746	-3.193777	4.427492	-2.357609
4	1.647204	-2.945539	4.272474	-2.549694
5	1.682025	-2.723966	4.128304	-2.715634
6	1.717631	-2.527427	3.999765	-2.862150
7	1.749749	-2.353270	3.885783	-2.991898
8	1.778274	-2.198968	3.784786	-3.106845
9	1.803554	-2.062259	3.695303	-3.208684
10	1.825953	-1.941139	3.616023	-3.298912
11	1.845798	-1.833828	3.545783	-3.378851
12	1.863381	-1.738753	3.483552	-3.449676
13	1.878958	-1.654519	3.428416	-3.512425
14	1.892760	-1.579890	3.379568	-3.568019
15	1.904987	-1.513770	3.336289	-3.617274

$\diamond$

El método SOR depende de la escogencia de ω y queda entonces la pregunta ¿Cómo escoger ω? La respuesta no es sencilla. Algunas veces se hace simplemente por ensayo y error. Si se desea resolver muchos sistemas de ecuaciones parecidos, por ejemplo provenientes del mismo tipo de problema pero con datos ligeramente diferentes, se puede pensar que un valor adecuado de ω para un problema puede servir para un problema parecido. Entonces se puede pensar en hacer ensayos con varios valores de ω para "ver" y escoger el ω que se supone sirva para este tipo de problemas.

En algunos caso muy particulares se puede hacer un estudio teórico. Tal es el caso de la solución, por diferencias finitas, de la ecuación de Poisson en un rectángulo. Allí se demuestra que

$$\omega_{\text{opt}} = \frac{2}{1 + \sin \dfrac{\pi}{m+1}}$$

Este resultado y otros teóricos se basan en el radio espectral de la matriz de la iteración de punto fijo.

El **radio espectral** de una matriz cuadrada M, denotado generalmente $\rho(M)$, es la máxima norma de los valores propios de M (reales o comple-

jos),

$$\rho(M) = \max_{1\leq i\leq n}\{|\lambda_i| : \lambda_i \in \text{esp}(M)\},$$

donde esp(M) es el conjunto de valores propios de M.

Los métodos de Jacobi, GS, SOR se pueden expresar de la forma

$$x^{k+1} = Mx^k + p. \tag{2.14}$$

Al aplicar varias veces la fórmula anterior, se está buscando un punto fijo de la función $f(x) = Mx + p$. Se puede mostrar que la iteración de punto fijo converge si existe una norma matricial $||\ \ ||$ tal que

$$||M|| < 1.$$

En algunos casos el criterio anterior se puede aplicar fácilmente al encontrar una norma adecuada. Pero por otro lado, si después de ensayar con varias normas, no se ha encontrado una norma que sirva, no se puede concluir que no habrá convergencia. El siguiente criterio es más preciso pero puede ser numéricamente más difícil de calcular.

La iteración de punto fijo (2.14) *converge si y solamente si*

$$\rho(M) < 1.$$

La convergencia es lenta cuando $\rho(M)$ es cercano a 1, es rápida cuando $\rho(M)$ es pequeño.

Cualquier matriz cuadrada A se puede expresar de la forma

$$A = L + D + U,$$

donde L es matriz triangular inferior correspondiente a la parte triangular estrictamente inferior de A, D es la matriz diagonal correspondiente a los elementos diagonales de A y U es matriz triangular superior correspondiente a la parte triangular estrictamente superior de A. Se puede mostrar que el método SOR se puede expresar como una iteracion de punto fijo con

$$\begin{aligned} M_{\text{SOR}} &= (D + \omega L)^{-1}\big((1-\omega)D - \omega U\big), \\ p_{\text{SOR}} &= \omega(D + \omega L)^{-1}b. \end{aligned}$$

La deducción anterior proviene de descomponer

$$\begin{aligned} A &= \frac{1}{\omega}D + L \quad + \quad (1-\frac{1}{\omega})D + U \\ &= \frac{1}{\omega}\big(D+\omega L\big) + \frac{1}{\omega}\big((\omega-1)D+\omega U\big) \\ &= \frac{D+\omega L}{\omega} + \frac{(\omega-1)D+\omega U}{\omega} \end{aligned}$$

Entonces

$$\begin{aligned} Ax &= b \\ \big(\frac{D+\omega L}{\omega} + \frac{(\omega-1)D+\omega U}{\omega}\big)x &= b \\ \big(D+\omega L \quad + \quad (\omega-1)D+\omega U\big)x &= \omega b \\ (D+\omega L)x &= -\big((\omega-1)D+\omega U\big)x + \omega b \\ (D+\omega L)x &= \big((1-\omega)D-\omega U\big)x + \omega b \\ x &= (D+\omega L)^{-1}\big((1-\omega)D-\omega U\big)x + \omega(D+\omega L)^{-1}b \end{aligned}$$

Para el caso particular del método GS

$$\begin{aligned} M_{\mathrm{GS}} &= -(D+L)^{-1}U, \\ p_{\mathrm{GS}} &= (D+L)^{-1}b. \end{aligned}$$

Para el ejemplo 2.4, con $\omega = 1.4$,

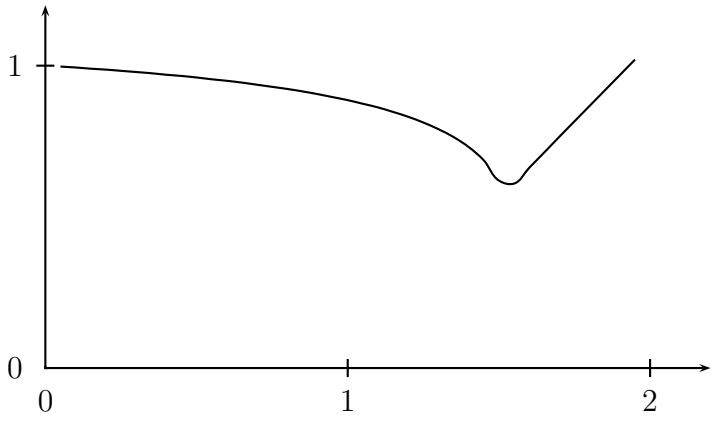
$$x^{k+1} = \begin{bmatrix} -0.400000 & 0.280000 & -0.560000 & 0.560000 \\ 0.000000 & -0.400000 & -0.700000 & -1.050000 \\ 0.210000 & 0.063000 & 0.261500 & 0.607250 \\ -0.044333 & 0.453367 & 0.583683 & 0.852358 \end{bmatrix} x^k + \begin{bmatrix} 7.000000 \\ -3.500000 \\ 4.287500 \\ -1.799583 \end{bmatrix}.$$

En este caso $\rho(M) = 0.730810$, lo que garantiza la convergencia.

La siguiente tabla nos muestra los valores del número de iteraciones y del radio espectral para diferentes valores de ω. El criterio de parada utilizado fue $\max\{|\delta_i| : i = 1, ..., n\} \leq 0.000001$.

ω	k	$\rho(M)$
0.10	999	0.994
0.20	641	0.987
0.30	415	0.979
0.40	301	0.970
0.50	232	0.961
0.60	185	0.950
0.70	151	0.937
0.80	125	0.923
0.90	105	0.906
1.00	88	0.886
1.10	74	0.862
1.20	61	0.831
1.30	50	0.790
1.40	40	0.731
1.50	29	0.620
1.60	33	0.662
1.70	50	0.765
1.80	92	0.867
1.90	408	0.969

La siguiente gráfica muestra la variación del radio espectral $\rho(M)$ al variar ω. Proviene de un conjunto de datos más amplio que el de la tabla anterior.



El mejor valor de ω es aproximadamente $\omega \approx 1.55$. Esto coincide, en la tabla, con el menor número de iteraciones.

El siguiente es el esquema del algoritmo de sobrerrelajación, muy parecido al de GS. Se supone que no hay elementos diagonales nulos.

SOR: SOBRERRELAJACIÓN

datos: $A,\ b,\ \omega,\ x^0,\ \varepsilon,$ maxit

$x = x^0$
para $k = 1, ...,$maxit
    difX $= 0$
    para $i = 1, ..., n$
        $r_i = b_i - A_{i.}\, x$
        $\delta_i = \dfrac{r_i}{a_{ii}}$
        $x_i = x_i + \omega\delta_i$
        difX $= \max\{\text{difX},\ |\omega\delta_i|\}$
    fin-para i
    si difX $\leq \varepsilon$ **ent** $x^* \approx x$, **salir**
fin-para k

El método de sobrerrelajación, como el de GS, es útil para sistemas dispersos en los que la matriz se ha almacenado de manera dispersa. Si la matriz es dispersa pero se almacena como si fuera densa, el método de Gauss, en la mayoría de los casos, debe resultar mejor.

2.5 Método del gradiente conjugado

Si A es una matriz simétrica y definida positiva, la solución del sistema

$$Ax = b \tag{2.15}$$

es exactamente el mismo punto x^* que resuelve el siguiente problema de optimización:

$$\min \ f(x) = \frac{1}{2} x^{\mathrm{T}} A x - b^{\mathrm{T}} c. \tag{2.16}$$

Como A es definida positiva, entonces f es convexa (más aún, es estrictamente convexa). Para funciones convexas diferenciables, un punto de gradiente nulo es necesariamente un minimizador global:

$$\nabla f(x) = f'(x) = Ax - b = 0.$$

Para minimizar funciones sin restricciones hay muchos métodos. Dentro del grupo de métodos de direcciones conjugadas, está el método del gradiente conjugado. Este método se adapta muy bien cuando la matriz es "dispersa". Tiene la ventaja adicional que, aunque es un método iterativo, a lo más en n iteraciones se obtiene la solución exacta, si no hay rerrores de redondeo.

La mayoría de los método de minimización son iterativos. En cada iteración, dado un punto x^k, hay dos pasos importantes: en el primero se calcula una dirección d^k. Normalmente esta dirección cumple con la propiedad

$$f'(x^k)^{\mathrm{T}} d^k < 0.$$

Esto garantiza que la dirección sea de descenso, es decir, que para t suficientemente pequeño

$$f(x^k + t d^k) < f(x^k).$$

El segundo paso consiste en encontrar el mejor t posible, o sea, encontrar

$$t_k = \operatorname{argmin} \ f(x^k + t\, d^k), \quad t \geq 0.$$

Con d^k y t_k se construye el siguiente punto

$$x^{k+1} = x^k + t_k\, d^k.$$

Un método muy popular, pero no necesariamente muy eficiente, es el método de Cauchy, también llamado método del gradiente o método del descenso

más pendiente. En este método la dirección es simplemente el opuesto del gradiente,

$$d^k = -f'(x^k).$$

En el método GC la dirección se construye agregando a $-f'(x^k)$ un múltiplo de la dirección anterior,

$$d^k = -f'(x^k) + \alpha_k \, d^{k-1}. \tag{2.17}$$

Dos direcciones diferentes, d^i y d^j, se llaman conjugadas con respecto a A si

$$d^{i^{\mathrm{T}}} A \, d^j = 0.$$

Para el caso de la solución de un sistema lineal por medio del método GC, es corriente denominar el vector residuo

$$r^k = Ax^k - b. \tag{2.18}$$

Obviamente $x^k = x^*$ si y solamente si $r^k = 0$. El vector residuo es exactamente el mismo gradiente de f en el punto x^k.

Las fórmulas que completan la definición del método GC son:

$$\alpha_1 = 0, \tag{2.19}$$

$$\alpha_k = \frac{||r^k||_2^2}{||r^{k-1}||_2^2}, \quad k = 2, ..., n, \tag{2.20}$$

$$t_k = \frac{||r^k||_2^2}{d^{k^{\mathrm{T}}} A d^k}, \quad k = 1, ..., n. \tag{2.21}$$

Suponiendo que A es definida positiva, el método GC tiene las siguientes propiedades:

- d^k es dirección de descenso.
- $f(x^k) < f(x^{k-1})$.
- las direcciones son conjugadas con respecto a A.
- Si no hay errores de redondeo, entonces $x^* = x^k$ para algún $k \leq n + 1$.

Cuando se llega a x^{n+1} y no se obtiene la solución con la precisión deseada, entonces se vuelve a empezar el proceso utilizando como nuevo x^1 el x^{n+1} obtenido.

MÉTODO DEL GRADIENTE CONJUGADO

datos: A, b, x^1, MAXIT, ε
para $K = 1, ..., \texttt{MAXIT}$
  para $k = 1, ..., n$
    $r^k = Ax^k - b$
    si $||r^k|| < \varepsilon$ **ent parar**
    si $k = 1$ **ent** $d^k = -r^k$
    sino
      $\alpha_k = \dfrac{||r^k||_2^2}{||r^{k-1}||_2^2}$
      $d^k = -r^k + \alpha_k d^{k-1}$
    fin-sino
    $t_k = \dfrac{||r^k||_2^2}{d^{k\mathrm{T}} A d^k}$
    $x^{k+1} = x^k + t_k d^k$
  fin-para k
  $x^1 = x^{n+1}$
fin-para K

Ejemplo 2.5. Resolver el sistema $Ax = b$ por el método GC, partiendo de $x^1 = (1, 1, 1)$, donde

$$A = \begin{bmatrix} 19 & 6 & 8 \\ 6 & 5 & 2 \\ 8 & 2 & 4 \end{bmatrix}, \qquad b = \begin{bmatrix} 55 \\ 22 \\ 24 \end{bmatrix}.$$

$$\begin{aligned} r^1 &= Ax^1 - b = (-22, -9, -10), \\ ||r^1||_2^2 &= 665, \\ d^1 &= -r^1 = (22, 9, 10), \\ d^{1\mathrm{T}} A d^1 &= 16257, \\ t_1 &= \frac{665}{16257} = 0.040905, \\ x^2 &= x^1 + t_1 d^1 = (1.899920, 1.368149, 1.409055), \end{aligned}$$

$$\begin{aligned} r^2 &= (0.579812, -0.941625, -0.428123), \\ ||r^2||_2^2 &= 1.406129, \end{aligned}$$

$$
\begin{aligned}
\alpha_2 &= \frac{1.406129}{665} = 0.002114, \\
d^2 &= (-0.533293, 0.960655, 0.449268), \\
{d^2}^{\mathrm{T}} A d^2 &= 2.570462, \\
t_2 &= 0.547034, \\
x^3 &= (1.608191, 1.893660, 1.654819),
\end{aligned}
$$

$$
\begin{aligned}
r^3 &= (0.156138, 0.427083, -0.727877), \\
||r^3||_2^2 &= 0.736584, \\
\alpha_3 &= 0.523838, \\
d^3 &= (-0.435497, 0.076145, 0.963221), \\
{d^3}^{\mathrm{T}} A d^3 &= 0.527433, \\
t_3 &= 1.396545, \\
x^4 &= (1, 2, 3),
\end{aligned}
$$

$$
\begin{aligned}
x^1 &= x^4 = (1, 2, 3), \\
r^1 &= (0, 0, 0).
\end{aligned}
$$

Si la matriz A es dispersa y se utiliza una estructura de datos donde solamente se almacenen los elementos no nulos, para poder implementar con éxito el método GC, se requiere simplemente poder efectuar el producto de la matriz A por un vector. Hay dos casos, Ax^k para calcular r^k y Ad^k para calcular t_k. Las otras operaciones necesarias son producto escalar entre vectores, sumas o restas de vectores y multiplicación de un escalar por un vector. Todo esto hace que sea un método muy útil para matrices muy grandes pero muy poco densas.

2.6 Condicionamiento de una matriz

Cuando se resuelve un sistema de ecuaciones $Ax = b$ se desea conocer cómo son los cambios en la solución cuando se cambia ligeramente el vector de términos independientes b.

De manera más precisa, sea $\bar{x}$ la solución de $Ax = b$ y $\bar{x}'$ la solución de $Ax = b'$. Se puede suponer que

$$\begin{aligned} b' &= b + \Delta b, \\ \bar{x}' &= \bar{x} + \Delta x. \end{aligned}$$

Se espera que si $||\Delta b||$ es pequeña, entonces también $||\Delta x||$ es pequeña. En realidad es mejor considerar cambios relativos. Se espera que si el valor $||\Delta b||/||b||$ es pequeño, entonces también $||\Delta x||/||\bar{x}||$ sea pequeño. Las deducciones que siguen relacionan los dos cambios relativos.

$$\begin{aligned} \Delta x &= \bar{x}' - \bar{x} \\ &= A^{-1}b' - A^{-1}b \\ &= A^{-1}(b + \Delta b) - A^{-1}b \\ &= A^{-1}\Delta b. \end{aligned}$$

Al utilizar una norma y la norma matricial generada se obtiene

$$||\Delta x|| \leq |||A^{-1}||| \; ||\Delta b||.$$

Por otro lado

$$\begin{aligned} b &= Ax \\ ||b|| &\leq |||A||| \, ||\bar{x}|| \\ \frac{1}{||\bar{x}||} &\leq \frac{|||A|||}{||b||} \end{aligned}$$

Multiplicando la primera y la última desigualdad

$$\frac{||\Delta x||}{||\bar{x}||} \leq |||A||| \; |||A^{-1}||| \frac{||\Delta b||}{||b||}. \tag{2.22}$$

El valor $|||A||| \; |||A^{-1}|||$ se llama condicionamiento o número de condición de la matriz A (invertible) y se denota

$$\kappa(A) = |||A||| \; |||A^{-1}|||.$$

Entonces

$$\frac{||\Delta x||}{||\bar{x}||} \leq \kappa(A) \frac{||\Delta b||}{||b||}. \tag{2.23}$$

Ejemplo 2.6. Calcular $\kappa_1(A)$, $\kappa_2(A)$ y $\kappa_\infty(A)$ para la matriz

$$A = \begin{bmatrix} -10 & -7 \\ 6 & 4 \end{bmatrix}.$$

Entonces

$$\begin{aligned}
A^{-1} &= \begin{bmatrix} 2 & 7/2 \\ -3 & -5 \end{bmatrix} \\
A^{\mathrm{T}}A &= \begin{bmatrix} 136 & 94 \\ 94 & 65 \end{bmatrix} \\
{A^{-1}}^{\mathrm{T}}A^{-1} &= \begin{bmatrix} 13 & 22 \\ 22 & 149/4 \end{bmatrix} \\
\mathrm{esp}(A^{\mathrm{T}}A) &= \{0.0199025, 200.9801\} \\
\mathrm{esp}({A^{-1}}^{\mathrm{T}}A^{-1}) &= \{0.0049756, 50.245024\} \\
|||A|||_2 &= 14.176745 \\
|||A^{-1}|||_2 &= 7.0883725 \\
\kappa_2(A) &= 100.49005 \\
|||A|||_1 &= 16 \\
|||A^{-1}|||_1 &= 17/2 \\
\kappa_1(A) &= 136 \\
|||A|||_\infty &= 17 \\
|||A^{-1}|||_\infty &= 8 \\
\kappa_\infty(A) &= 136. \; \diamond
\end{aligned}$$

El condicionamiento, definido para normas matriciales inducidas de normas vectoriales, tiene la siguientes propiedades:

- $\kappa(A) \geq 1$.
- $\kappa(\alpha A) = \kappa(A)$ si $\alpha \neq 0$.
- $\kappa_2(A) = 1$ si y solamente si A es un múltiplo de una matriz ortogonal (o unitaria).

La desigualdad (2.23) indica que si $\kappa(A)$ es pequeño, entonces un cambio relativo en b pequeño produce un cambio relativo en x pequeño.

Una matriz A es bien condicionada si $\kappa(A)$ es cercano a 1 y es mal condicionada si $\kappa(A)$ es grande. Para el condicionamiento κ_2 (definido con la norma espectral) las matrices mejor condicionadas son las matrices ortogonales.

Ejemplo 2.7. Resolver los sistemas $Ax = b$ y $Ax' = b'$, donde

$$A = \begin{bmatrix} 10 & 10 \\ 10 & -9 \end{bmatrix}, \qquad b = \begin{bmatrix} 20.01 \\ 19.99 \end{bmatrix}, \qquad b' = \begin{bmatrix} 20.02 \\ 19.98 \end{bmatrix}.$$

Entonces

$$\begin{aligned} \Delta b &= [0.01 \quad -0.01]^{\mathrm{T}}, \\ \frac{||\Delta b||}{||b||} &= 0.0005, \\ \kappa(A) &= 1.0752269. \end{aligned}$$

Al resolver los dos sistemas se obtiene:

$$\begin{aligned} x &= [1.9999474 \quad 0.0010526]^{\mathrm{T}}, \\ x' &= [1.9998947 \quad 0.0021053]^{\mathrm{T}}, \\ \Delta x &= [-0.0000526 \quad .0010526]^{\mathrm{T}}, \\ \frac{||\Delta x||}{||x||} &= 0.0005270, \\ \kappa(A)\frac{||\Delta b||}{||b||} &= 0.0005376. \end{aligned}$$

La matriz A es muy bien condicionada y entonces cambios pequeños en b producen cambios pequeños en x. $\diamond$

Ejemplo 2.8. Resolver los sistemas $Ax = b$ y $Ax' = b'$, donde

$$A = \begin{bmatrix} 10.01 & 10.00 \\ 10.00 & 9.99 \end{bmatrix}, \qquad b = \begin{bmatrix} 20.01 \\ 19.99 \end{bmatrix}, \qquad b' = \begin{bmatrix} 20.02 \\ 19.98 \end{bmatrix}.$$

Entonces

$$\begin{aligned} \Delta b &= [0.01 \quad -0.01]^{\mathrm{T}}, \\ \frac{||\Delta b||}{||b||} &= 0.0005, \\ A^{-1} &= \begin{bmatrix} -99900 & 100000 \\ 100000 & -100100 \end{bmatrix}, \\ \kappa(A) &= 4000002. \end{aligned}$$

Al resolver los dos sistemas se obtiene:

$$\begin{aligned} x &= [1 \quad 1]^{\mathrm{T}}, \\ x' &= [-1998 \quad 2002]^{\mathrm{T}}, \\ \Delta x &= [-1999 \quad 2001]^{\mathrm{T}}, \\ \frac{||\Delta x||}{||x||} &= 2000.0002, \\ \kappa(A)\frac{||\Delta b||}{||b||} &= 2000.0008. \end{aligned}$$

La matriz A es muy mal condicionada y entonces cambios pequeños en b pueden producir cambios muy grandes en la solución. $\diamond$

Ejemplo 2.9. Resolver los sistemas $Ax = b$ y $Ax'' = b''$, donde

$$A = \begin{bmatrix} 10.01 & 10.00 \\ 10.00 & 9.99 \end{bmatrix}, \qquad b = \begin{bmatrix} 20.01 \\ 19.99 \end{bmatrix}, \qquad b'' = \begin{bmatrix} 20.02 \\ 20.00 \end{bmatrix}.$$

Entonces

$$\begin{aligned} \Delta b &= [0.01 \quad 0.01]^{\mathrm{T}}, \\ \frac{||\Delta b||}{||b||} &= 0.0005, \end{aligned}$$

$$A^{-1} = \begin{bmatrix} -99900 & 100000 \\ 100000 & -100100 \end{bmatrix},$$
$$\kappa(A) = 4000002.$$

Al resolver los dos sistemas se obtiene:

$$x = [1 \quad 1]^{\mathrm{T}},$$
$$x'' = [2 \quad 0]^{\mathrm{T}},$$
$$\Delta x = [1 \quad -1]^{\mathrm{T}},$$
$$\frac{||\Delta x||}{||x||} = 1,$$
$$\kappa(A)\frac{||\Delta b||}{||b||} = 2000.0008.$$

La matriz A, la misma del ejemplo anterior, es muy mal condicionada y entonces cambios pequeños en b pueden producir cambios muy grandes en la solución. Sin embargo los cambios en la solución, aunque no despreciables, no fueron tan grandes como en el ejemplo anterior, o sea, $||\Delta x||/||x||$ está lejos de la cota superior. $\diamond$

Capítulo 3

Solución de ecuaciones no lineales

3.1 Método de Muller

Este método sirve para hallar raíces reales o complejas de polinomios. Sea $p(x)$ un polinomio real (con coeficientes reales), de grado n, es decir,

$$p(x) = a_0 + a_1 x + a_2 x^2 + ... + a_n x^n, \quad a_i \in \mathbb{R},\ i = 0, 1, ..., n, \quad a_n \neq 0.$$

En general no se puede garantizar que $p(x)$ tenga raíces reales. Sin embargo (teorema fundamental del Álgebra) se puede garantizar que tiene n raíces complejas (algunas de ellas pueden ser reales). De manera más precisa, existen $r_1, r_2, ..., r_n \in \mathbb{C}$ tales que

$$p(r_i) = 0, \quad i = 1, 2, ..., n.$$

El polinomio p se puede expresar en función de sus raíces:

$$p(x) = a_n(x - r_1)(x - r_2) \cdots (x - r_n).$$

Las raíces complejas, no reales, siempre vienen por parejas, es decir si $r = a + ib$, $b \neq 0$, es una raíz entonces $\bar{r} = a - ib$, el conjugado de r, también es raíz. Esto garantiza que los polinomios de grado impar tienen por lo menos una raíz real. Para los polinomios de grado par, el número de raíces reales es par y el número de raíces estrictamente complejas también es par. Así un polinomio de grado par puede tener cero raíces reales.

Para las raíces complejas $(x-r)(x-\bar{r})$ divide a $p(x)$.

$(x-r)(x-\bar{r}) = (x-a-ib)(x-a+ib) = (x-a)^2+b^2 = x^2-2ax+(a^2+b^2)$.

O sea, se tiene un polinomio real de grado 2 que divide a $p(x)$.

Si $q(x)$ divide a $p(x)$, entonces existe un polinomio $s(x)$ tal que

$$\begin{aligned} p(x) &= q(x)s(x), \\ \text{grado}(p) &= \text{grado}(q) + \text{grado}(s). \end{aligned}$$

Entonces para sequir obteniendo las raíces de $p(x)$ basta con obtener las raíces de $s(x)$, polinomio más sencillo.

Si se hallar una raíz real r entonces $q(x) = (x-r)$ divide a $p(x)$. Si se obtiene una raíz compleja $r = a+ib$, entonces $q(x) = x^2-2ax+(a^2+b^2)$ divide a $p(x)$. Este proceso de obtener un polinomio de grado menor cuyas raíces sean raíces del polinomio inicial se llama **deflación**.

En el método de la secante, dados dos valores x_0 y x_1 se busca la recta que pasa por los puntos $(x_0, f(x_0))$, $(x_1, f(x_1))$; el siguiente valor x_2 está dado por el punto donde la recta corta el eje x.

En el método de Muller, en lugar de una recta, se utiliza una parábola. Dados tres valores x_0, x_1 y x_2, se construye la parábola $P(x)$ que pasa por los puntos $(x_0, f(x_0))$, $(x_1, f(x_1))$ y $(x_2, f(x_2))$; el siguiente valor x_3 está dado por el (un) punto tal que $P(x_3) = 0$.

La parábola se puede escribir de la forma $P(x) = a(x-x_2)^2 + b(x-x_2) + c$. Entonces hay tres condiciones que permiten calcular a, b y c:

$$\begin{aligned} f(x_0) &= a(x_0-x_2)^2 + b(x_0-x_2) + c, \\ f(x_1) &= a(x_1-x_2)^2 + b(x_1-x_2) + c, \\ f(x_2) &= c. \end{aligned}$$

Después de algunos cálculos se obtiene

$$\begin{aligned} d &= (x_0-x_1)(x_0-x_2)(x_1-x_2), \\ a &= \frac{-(x_0-x_2)(f(x_1)-f(x_2)) + (x_1-x_2)(f(x_0)-f(x_2)}{d}, \\ b &= \frac{(x_0-x_2)^2(f(x_1)-f(x_2)) - (x_1-x_2)^2(f(x_0)-f(x_2)}{d}, \\ c &= f(x_2). \end{aligned} \tag{3.1}$$

Entonces

$$x_3 - x_2 = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Para reducir los errores de redondeo se "racionaliza" el numerador y se escoge el signo buscando que el denominador resultante sea grande (en valor absoluto)

$$\begin{aligned}
D &= b^2 - 4ac, \\
R &= \sqrt{D} \\
x_3 - x_2 &= \frac{-b \pm R}{2a} \ \frac{-b \mp R}{-b \mp R} \\
x_3 - x_2 &= \frac{b^2 - R^2}{2a\big(-b \mp R\big)} = \frac{b^2 - b^2 + 4ac}{2a\big(-b \mp R\big)} = \frac{2c}{-b \mp R} \\
x_3 - x_2 &= -\frac{2c}{b \pm R} \\
x_3 &= x_2 - \frac{2c}{b + \text{signo}(b)R}
\end{aligned} \tag{3.2}$$

En la siguiente iteración se obtiene la parábola utilizando x_1, x_2 y x_3 para obtener x_4.

Si en una iteración

$$D = b^2 - 4ac < 0$$

es necesario utilizar, a partir de ahí, aritmética compleja. Eso hace que los siguientes valores a, b y c no sean necesariamente reales. Muy posiblemente $b^2 - 4ac$ tampoco es real. Para utilizar (3.2) es necesario obtener la raíz cuadradada de un complejo.

Sean z un complejo, θ el ángulo (en radianes) formado con el eje real ("eje x"), llamado con frecuencia argumento de z, y ρ la norma o valor absoluto de z. La dos raíces cuadradas de z son:

$$\begin{aligned}
\sqrt{z} = \zeta_1 &= \sqrt{\rho}\,\big(\cos(\theta/2) + i\,\text{sen}(\theta/2)\big), \\
\zeta_2 &= -\zeta_1.
\end{aligned}$$

Ejemplo 3.1. Sea $z = 12 + 16i$. Entonces

$$\rho = 20,$$

$$\theta = \tan^{-1}(16/12) = 0.927295,$$
$$\zeta_1 = \sqrt{20}\left(\cos(0.927295/2) + i\ \mathrm{sen}(0.927295/2)\right) = 4 + 2i,$$
$$\zeta_2 = -4 - 2i. \quad \diamond$$

Cuando b no es real, es necesario modificar ligeramente (3.2). Se escoge el signo para que el denominador tenga máxima norma:

$$\begin{aligned} D &= b^2 - 4ac \\ R &= \sqrt{D} \\ \delta &= \begin{cases} b + R & \text{si} \quad |b+R| \geq |b-R| \\ b - R & \text{si} \quad |b+R| < |b-R| \end{cases} \\ x_3 &= x_2 - \frac{2c}{\delta}\ . \end{aligned} \tag{3.3}$$

Ejemplo 3.2. Hallar las raíces de $p(x) = 2x^5 + x^4 + 4x^3 + 19x^2 - 18x + 40$ partiendo de $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1$.

$$\begin{aligned} f(x_0) &= 40 \\ f(x_1) &= 36.375 \\ f(x_2) &= 48 \\ d &= -0.25 \\ a &= 30.5 \\ b &= 38.5 \\ c &= 48 \\ D &= -4373.75 \end{aligned}$$

Hay que utilizar aritmética compleja

$$\begin{aligned} R &= 66.134333i \\ \delta &= 38.5 + 66.134333i \\ x_3 &= 0.368852 + 1.084169i \\ f(x_3) &= 12.981325 - 9.579946i \end{aligned}$$

Ahora utilizamos x_1, x_2 y x_3

$$d = 0.546325 + 0.413228i$$

$$
\begin{aligned}
a &= 27.161207 + 11.293018i \\
b &= -21.941945 + 50.286087i \\
c &= 12.981325 - 9.579946i \\
D &= -3890.341507 - 1752.330850i \\
R &= 13.719321 - 63.863615i \\
\delta &= -35.661266 + 114.149702i \\
x_4 &= 0.586513 + 1.243614i \\
f(x_4) &= 3.760763 - 6.548104i \\
&\vdots \\
x_5 &= 0.758640 + 1.246582i \\
f(x_5) &= -2.013839 - 1.490220i \\
x_6 &= 0.748694 + 1.196892i \\
f(x_6) &= 0.123017 + 0.025843i \\
x_7 &= 0.750002 + 1.198942i \\
f(x_7) &= 0.000535 + 0.000636i \\
x_8 &= 0.750000 + 1.198958i \\
f(x_8) &= 0
\end{aligned}
$$

Ahora se construye el polinomio $q(x) = (x - r)(x - \bar{r})$. Para $r = 0.75 + 1.198958i$ se tiene $q(x) = x^2 - 1.5x + 2$.

$$
\frac{2x^5 + x^4 + 4x^3 + 19x^2 - 18x + 40}{x^2 - 1.5x + 2} = 2x^3 + 4x^2 + 6x^2 + 20.
$$

Ahora se trabaja con $p(x) = 2x^3 + 4x^2 + 6x^2 + 20$. Sean $x_0 = -3$, $x_1 = -2.5$ y $x_2 = -2$. También se hubiera podido volver a utilizar $x_0 = 0$, $x_1 = 0.5$ y $x_2 = 1$.

$$
\begin{aligned}
f(x_0) &= -16 \\
f(x_1) &= -1.25 \\
f(x_2) &= 8 \\
d &= -0.25 \\
a &= -11 \\
b &= 13
\end{aligned}
$$

$$
\begin{aligned}
c &= 8 \\
D &= 521 \\
R &= 22.825424 \\
\delta &= 35.825424 \\
x_3 &= -2.446610 \\
f(x_3) &= -0.026391
\end{aligned}
$$

Ahora utilizamos x_1, x_2 y x_3

$$
\begin{aligned}
d &= 0.011922 \\
a &= -9.893220 \\
b &= 22.390216 \\
c &= -0.026391 \\
D &= 500.277428 \\
R &= 22.366882 \\
\delta &= 44.757098 \\
x_4 &= -2.445431 \\
f(x_4) &= -0.000057 \\
&\vdots \\
x_5 &= -2.445428 \\
f(x_5) &= 0
\end{aligned}
$$

Para $r = -2.445428$ se tiene $q(x) = x + 2.445428$.

$$\frac{2x^3 + 4x^2 + 6x^2 + 20}{x + 2.445428} = 2x^2 - 0.890857x + 8.178526.$$

Ahora se trabaja con $p(x) = 2x^2 - 0.890857x + 8.178526$. Sus raíces son $0.2227142 + 2.009891i$ y $0.2227142 - 2.009891i$. En resumen, las 5 raíces de $p(x)$ son:

$$
\begin{aligned}
&0.75 + 1.1989579i \\
&0.75 - 1.1989579i \\
&-2.445428
\end{aligned}
$$

$$0.222714 + 2.009891i$$
$$0.222714 - 2.009891i. \ \diamond$$

El método de Muller tiene orden de convergencia no inferior a 1.84... Este valor proviene de la raíz mas grande de $\mu^3 - \mu^2 - \mu - 1 = 0$. Esto hace que sea un poco menos rápido que el método de Newton (orden 2) pero más rápido que el método de la secante (orden 1.68).

El método no tiene sentido si hay valores iguales (o muy parecidos) entre x_0, x_1 y x_2. Además esto haría que no se pueda calcular a ni b. Tampoco funciona si los valores $f(x_0)$, $f(x_1)$ y $f(x_2)$ son iguales o muy parecidos. En este caso $P(x)$ es una línea recta horizontal y no se puede calcular x_3 ya que $a = 0$, $b = 0$ y, principalmente, $\delta = b \pm R = 0$.

MÉTODO DE MULLER

datos: p, x_0, x_1, x_2, ε_f, ε_0, maxit
mientras grado$(p) \geq 3$
    tomar o redefinir x_0, x_1, x_2
    aritmética = real
    $f_0 = p(x_0)$, $f_1 = p(x_1)$, $f_2 = p(x_2)$
    fink= 0
    para $k = 1, ..., \text{maxit}$
        si $|f_2| \leq \varepsilon_f$ **ent** $r = x_2$, fink= 1, salir del bucle **para**
        $d = (x_0 - x_1)(x_0 - x_2)(x_1 - x_2)$
        si $|d| \leq \varepsilon_0$ **ent parar**
        calcular a, b y c según (3.1)
        $D = b^2 - 4ac$
        si aritmética=real y $D < 0$ **ent** aritmética=compleja
        $R = \sqrt{D}$
        si $|b + R| \geq |b - R|$ **ent** $\delta = b + R$
        sino $\delta = b - R$
        si $|\delta| \leq \varepsilon_0$ **ent parar**
        $x_3 = x_2 - 2c/\delta$
        $x_0 = x_1$, $x_1 = x_2$, $x_2 = x_3$, $f_0 = f_1$, $f_1 = f_2$
        $f_2 = p(x_2)$
    fin-para k
    si fink=0 **ent parar**
    si imag$(r) = 0$ **ent** $q(x) = (x - r)$
    sino $q(x) = (x - r)(x - \bar{r})$
    $p(x) = p(x)/q(x)$
fin-mientras
calcular raíces de p (de grado no superior a 2)

Si se espera que el número de raíces reales sea pequeño, comparado con el de raíces complejas, se puede trabajar todo el tiempo con aritmética compleja.

3.2 Método de Bairstow

Sirve para hallar las raíces reales o complejas de un polinomio de grado mayor o igual a 4, mediante la obtención de los factores cuadráticos "mónicos" del polinomio. Cuando es de grado 3, se halla una raíz real por el método de

Newton, y después de la deflación se calculan las 2 raíces del polinomio cuadrático resultante.
Sea

$$p(x) = \alpha_n x^n + \alpha_{n-1} x^{n-1} + \alpha_{n-2} x^{n-2} + ... + \alpha_1 x + \alpha_0$$

reescrito como

$$p(x) = u_1 x^n + u_2 x^{n-1} + u_3 x^{n-2} + ... + u_n x + u_{n+1}$$

Se desea encontrar $x^2 - dx - e$ divisor de p. Cuando se hace la división entre p y un polinomio cuadrático cualquiera, se obtiene un residuo $r(x) = Rx + S$. Entonces se buscan valores de d y e tales que $r(x) = 0$, es decir, $R = 0$ y $S = 0$. Los valores R y S dependen de d y e, o sea, $R = R(d, e)$ y $S = S(d, e)$ Tenemos dos ecuaciones con dos incógnitas,

$$R(d, e) = 0$$
$$S(d, e) = 0$$

Sea

$$q(x) = \beta_{n-2} x^{n-2} + \beta_{n-3} x^{n-3} + ... + \beta_1 x + \beta_0$$

reescrito como

$$q(x) = v_1 x^{n-2} + v_2 x^{n-3} + ... + v_{n-2} x + v_{n-1}$$

el cociente. Entonces

$$p(x) = q(x)(x^2 - dx - e) + Rx + S.$$

Es decir,

$$u_1 x^n + u_2 x^{n-1} + ... + u_n x + u_{n+1} = (v_1 x^{n-2} + v_2 x^{n-3} + ... + v_{n-2} x + v_{n-1})(x^2 - dx - e) + Rx + S.$$

$$u_1 = v_1$$
$$u_2 = v_2 - dv_1$$
$$u_3 = v_3 - dv_2 - ev_1$$
$$u_4 = v_4 - dv_3 - ev_2$$

$$u_i = v_i - dv_{i-1} - ev_{i-2}$$

$$\begin{aligned} u_{n-1} &= v_{n-1} - dv_{n-2} - ev_{n-3} \\ u_n &= -dv_{n-1} - ev_{n-2} + R \\ u_{n+1} &= -ev_{n-1} + S \end{aligned}$$

Para facilitar las fórmulas es útil introducir dos coeficientes adicionales, v_n y v_{n+1}, que no influyen sobre q, definidos por

$$\begin{aligned} v_n &= R \\ v_{n+1} &= S + dv_n \end{aligned}$$

Entonces:

$$\begin{aligned} u_n &= v_n - dv_{n-1} - ev_{n-2} \\ u_{n+1} &= dv_n - dv_n - ev_{n-1} + S \\ \text{o sea} \quad u_{n+1} &= v_{n+1} - dv_n - ev_{n-1} \end{aligned}$$

Las igualdades quedan:

$$\begin{aligned} u_1 &= v_1 \\ u_2 &= v_2 - dv_1 \\ u_i &= v_i - dv_{i-1} - ev_{i-2}, \qquad i = 3, ..., n+1. \end{aligned}$$

Las fórmulas para calcular los v_i son

$$\begin{aligned} v_1 &= u_1 \\ v_2 &= u_2 + dv_1 \\ v_i &= u_i + dv_{i-1} + ev_{i-2}, \qquad i = 3, ..., n+1. \end{aligned}$$

Una vez obtenidos los v_i, entonces

$$R = v_n$$
$$S = v_{n+1} - dv_n$$

El objetivo inicial era buscar $R = 0$ y $S = 0$. Esto se obtiene si $v_n = 0$ y $v_{n+1} = 0$. O sea

$$v_n(d, e) = 0$$
$$v_{n+1}(d, e) = 0$$

Al aplicar el método de Newton se tiene:

$$\begin{bmatrix} \frac{\partial v_n}{\partial d}(d^k, e^k) & \frac{\partial v_n}{\partial e}(d^k, e^k) \\ \frac{\partial v_{n+1}}{\partial d}(d^k, e^k) & \frac{\partial v_{n+1}}{\partial e}(d^k, e^k) \end{bmatrix} \begin{bmatrix} \Delta d^k \\ \Delta e^k \end{bmatrix} = - \begin{bmatrix} v_n(d^k, e^k) \\ v_{n+1}(d^k, e^k) \end{bmatrix}$$

$$\begin{bmatrix} d^{k+1} \\ e^{k+1} \end{bmatrix} = \begin{bmatrix} d^k \\ e^k \end{bmatrix} + \begin{bmatrix} \Delta d^k \\ \Delta e^k \end{bmatrix}$$

Cálculo de las derivadas parciales:

$$\frac{\partial v_1}{\partial d} = 0$$
$$\frac{\partial v_2}{\partial d} = v_1$$
$$\frac{\partial v_i}{\partial d} = v_{i-1} + d\frac{\partial v_{i-1}}{\partial d} + e\frac{\partial v_{i-2}}{\partial d}$$

$$\frac{\partial v_1}{\partial e} = 0$$
$$\frac{\partial v_2}{\partial e} = 0$$
$$\frac{\partial v_i}{\partial e} = d\frac{\partial v_{i-1}}{\partial e} + v_{i-2} + e\frac{\partial v_{i-2}}{\partial e}$$

$$\frac{\partial v_i}{\partial e} = v_{i-2} + d\frac{\partial v_{i-1}}{\partial e} + e\frac{\partial v_{i-2}}{\partial e}$$

Explicitando las derivadas parciales con respecto a d se tiene

$$\begin{aligned}
\frac{\partial v_1}{\partial d} &= 0 \\
\frac{\partial v_2}{\partial d} &= v_1 \\
\frac{\partial v_3}{\partial d} &= v_2 + d\frac{\partial v_2}{\partial d} + e\frac{\partial v_1}{\partial d} \\
\frac{\partial v_3}{\partial d} &= v_2 + d\frac{\partial v_2}{\partial d} \\
\frac{\partial v_4}{\partial d} &= v_3 + d\frac{\partial v_3}{\partial d} + e\frac{\partial v_2}{\partial d} \\
\frac{\partial v_i}{\partial d} &= v_{i-1} + d\frac{\partial v_{i-1}}{\partial d} + e\frac{\partial v_{i-2}}{\partial d}
\end{aligned}$$

Sea

$$\begin{aligned}
w_1 &= v_1 \\
w_2 &= v_2 + dw_1 \\
w_i &= v_i + dw_{i-1} + ew_{i-2}, \qquad i = 3, ..., n+1.
\end{aligned}$$

Entonces

$$\begin{aligned}
\frac{\partial v_1}{\partial d} &= 0 \\
\frac{\partial v_2}{\partial d} &= w_1 \\
\frac{\partial v_3}{\partial d} &= w_2 \\
\frac{\partial v_i}{\partial d} &= w_{i-1}
\end{aligned}$$

Explicitando las derivadas parciales con respecto a e se tiene

$$\frac{\partial v_1}{\partial e} = 0$$

$$\frac{\partial v_2}{\partial e} = 0$$
$$\frac{\partial v_3}{\partial e} = v_1$$
$$\frac{\partial v_4}{\partial e} = v_2 + dv_1$$
$$\frac{\partial v_5}{\partial e} = v_3 + d\frac{\partial v_4}{\partial e} + e\frac{\partial v_3}{\partial e}$$

Utilizando de nuevo los w_i

$$\frac{\partial v_1}{\partial e} = 0$$
$$\frac{\partial v_2}{\partial e} = 0$$
$$\frac{\partial v_3}{\partial e} = w_1$$
$$\frac{\partial v_4}{\partial e} = w_2$$
$$\frac{\partial v_5}{\partial e} = w_3$$
$$\frac{\partial v_i}{\partial e} = w_{i-2}$$

Entonces

$$\frac{\partial v_n}{\partial d} = w_{n-1}$$
$$\frac{\partial v_n}{\partial e} = w_{n-2}$$
$$\frac{\partial v_{n+1}}{\partial d} = w_n$$
$$\frac{\partial v_{n+1}}{\partial e} = w_{n-1}$$

Es decir la matriz jacobiana es simplemente

$$\begin{bmatrix} w_{n-1} & w_{n-2} \\ w_n & w_{n-1} \end{bmatrix}$$

```
 alfa  =  - 4.  - 1.    1.    0.    3.  - 2.    2.

d0  =  - 1.  e0  =  - 1
.
u     :    2.0000   -2.0000    3.0000    0.0000    1.0000   -1.0000   -4.0000

v     :    2.0000   -4.0000    5.0000   -1.0000   -3.0000    3.0000   -4.0000
R, S  :    3.0000   -1.0000
w     :    2.0000   -6.0000    9.0000   -4.0000   -8.0000   15.0000
J
   -8.0000   -4.0000
   15.0000   -8.0000
Delta :    0.3226    0.1048
d, e  :   -0.6774   -0.8952

v     :    2.0000   -3.3548    3.4823    0.6441   -2.5536    0.1532   -1.8179
R, S  :    0.1532   -1.7141
w     :    2.0000   -4.7097    4.8824    1.5526   -7.9759    4.1664
J
   -7.9759    1.5526
    4.1664   -7.9759
Delta :   -0.0280   -0.2426
d, e  :   -0.7054   -1.1377

v     :    2.0000   -3.4108    3.1307    1.6721   -3.7414   -0.2632    0.4423
R, S  :   -0.2632    0.2566
w     :    2.0000   -4.8217    4.2566    4.1552  -11.5153    3.1326
J
  -11.5153    4.1552
    3.1326  -11.5153
Delta :   -0.0100    0.0357
d, e  :   -0.7154   -1.1020

v     :    2.0000   -3.4308    3.2503    1.4555   -3.6232   -0.0120    0.0015
R, S  :   -0.0120   -0.0071
w     :    2.0000   -4.8616    4.5243    3.5765  -11.1677    4.0360
J
  -11.1677    3.5765
```

```
    4.0360  -11.1677
Delta :   -0.0012   -0.0003
d, e  :   -0.7166   -1.1023

v     :    2.0000   -3.4331    3.2554    1.4517   -3.6287    0.0000   -0.0000
R, S  :    0.0000   -0.0000
w     :    2.0000   -4.8663    4.5378    3.5642  -11.1847    4.0857
J
  -11.1847    3.5642
    4.0857  -11.1847
Delta :    0.0000   -0.0000
d, e  :   -0.7166   -1.1023

v     :    2.0000   -3.4331    3.2554    1.4517   -3.6287   -0.0000    0.0000
R, S  :   -0.0000   -0.0000
w     :    2.0000   -4.8663    4.5378    3.5642  -11.1848    4.0857
J
  -11.1848    3.5642
    4.0857  -11.1848
Delta :   -0.0000    0.0000
d, e  :   -0.7166   -1.1023

 raices  =

! - 0.3582816 + 0.9868895i !
! - 0.3582816 - 0.9868895i !
-------------------------------------------------------
u     :    2.0000   -3.4331    3.2554    1.4517   -3.6287

v     :    2.0000   -5.4331    6.6885    0.1963  -10.5135
R, S  :    0.1963  -10.3173
w     :    2.0000   -7.4331   12.1217   -4.4923
J
   12.1217   -7.4331
   -4.4923   12.1217
Delta :    0.6673    1.1146
d, e  :   -0.3327    0.1146
```

```
v     :    2.0000   -4.0985    4.8482   -0.6311   -2.8630
R, S  :   -0.6311   -3.0729
w     :    2.0000   -4.7639    6.6623   -3.3937
J
    6.6623   -4.7639
   -3.3937    6.6623
Delta :    0.6323    0.7518
d, e  :    0.2996    0.8664

v     :    2.0000   -2.8339    4.1392    0.2364    0.0285
R, S  :    0.2364   -0.0423
w     :    2.0000   -2.2347    5.2026   -0.1410
J
    5.2026   -2.2347
   -0.1410    5.2026
Delta :   -0.0484   -0.0068
d, e  :    0.2512    0.8596

v     :    2.0000   -2.9306    4.2384   -0.0027    0.0141
R, S  :   -0.0027    0.0148
w     :    2.0000   -2.4281    5.3476   -0.7465
J
    5.3476   -2.4281
   -0.7465    5.3476
Delta :   -0.0007   -0.0027
d, e  :    0.2505    0.8569
v     :    2.0000   -2.9321    4.2347    0.0000    0.0000
R, S  :    0.0000    0.0000
w     :    2.0000   -2.4311    5.3395   -0.7456
J
    5.3395   -2.4311
   -0.7456    5.3395
Delta :   -0.0000   -0.0000

d, e  :    0.2505    0.8569
v     :    2.0000   -2.9321    4.2347    0.0000    0.0000
R, S  :    0.0000    0.0000
w     :    2.0000   -2.4311    5.3395   -0.7456
```

```
J
    5.3395   -2.4311
   -0.7456    5.3395
Delta :   -0.0000   -0.0000
d, e  :    0.2505    0.8569

raices   = - 0.8088697  1.0593828

raices restantes =

   0.7330251 + 1.2569892 i
   0.7330251 - 1.2569892 i
```

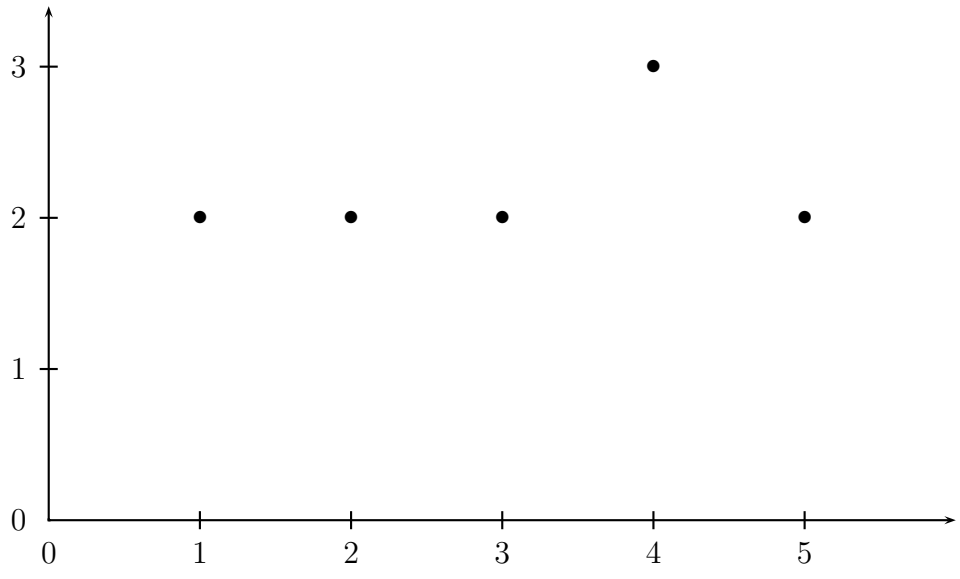
Capítulo 4

Interpolación y aproximación

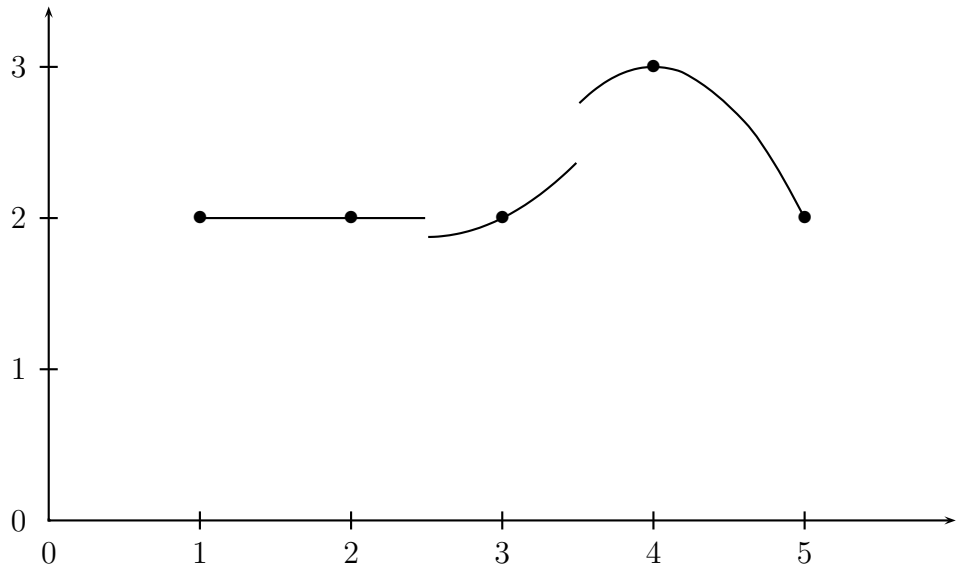
4.1 Interpolación polinomial por trozos

Dados $n+1$ puntos, al utilizar diferencias divididas o diferencias finitas, cuando se desea interpolar por un polinomio de grado m en un valor t, se escoge el mejor conjunto de puntos (x_k, y_k), (x_{k+1}, y_{k+1}), ..., (x_{k+m}, y_{k+m}), para obtener el valor $p_m(t)$. Sin embargo este método presenta un gran inconveniente cuando hay que interpolar en muchos valores t. Consideremos los siguientes puntos:

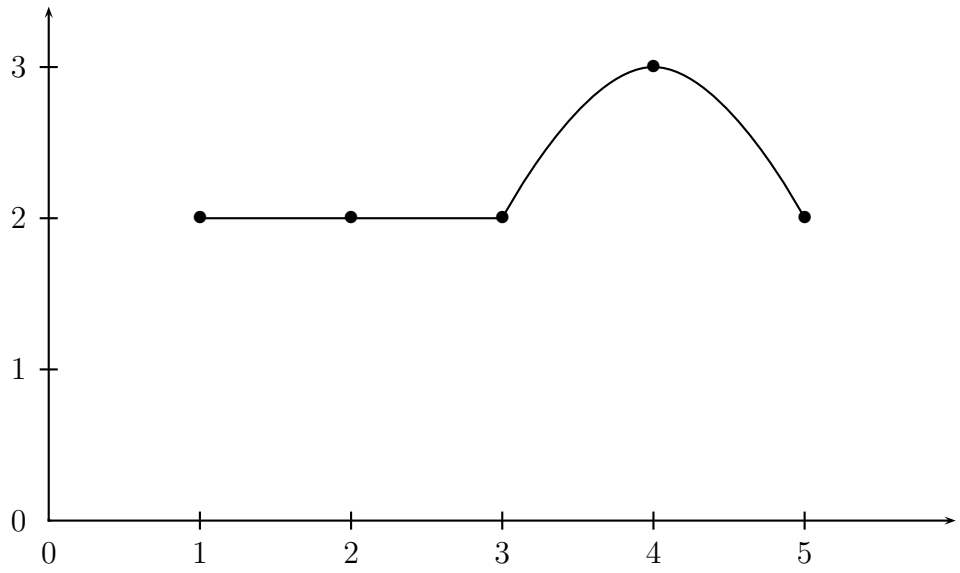
$$(1,2),\ (2,2),\ (3,2),\ (4,3),\ (5,2).$$



Para interpolar por polinomios de orden 2, si $t < 2.5$ se utilizan los puntos $(1, 2)$, $(2, 2)$ y $(3, 2)$. Entonces, por ejemplo, $p_2(2.49) = 2$. Si $2.5 < t < 3.5$, se utilizan los puntos $(2, 2)$, $(3, 2)$ y $(4, 3)$. Después de algunos cálculos se obtiene $p_2(2.51) = 1.87505$. Para $t = 2.501$ se obtiene $p_2(2.501) = 1.8750005$. El límite de $p_2(t)$, cuando $t \to 2.5^+$, es 1.875. Esto nos muestra una discontinuidad. En $t = 3.5$ también se presenta una discontinuidad.



Estas discontinuidades se pueden evitar utilizando en el intervalo $[1, 3]$ un polinomio $p_2(t)$ y en el intervalo $[3, 5]$ otro polinomio $p_2(t)$.



Obviamente ya no hay discontinuidades pero la gráfica no es suave, es decir, la función interpolante no es diferenciable.
Los trazadores cúbicos ("splines" cúbicos) remedian este inconveniente. En cada intervalo $[x_i, x_{i+1}]$ se utiliza un polinomio cúbico y los coeficientes de cada polinomio se escogen para que en los puntos x_i haya continuidad, diferenciabilidad y doble diferenciabilidad.

Dados $n+1$ puntos (x_0, y_0), (x_1, y_1), (x_2, y_2), ... (x_n, y_n), con

$$x_o < x_1 < x_2 < \cdots < x_n,$$

el trazador cúbico se define así:

$$S(x) = \begin{cases} S_0(x) & \text{si} \quad x \in [x_0, x_1] \\ S_1(x) & \text{si} \quad x \in [x_1, x_2] \\ \vdots & \\ S_{n-1}(x) & \text{si} \quad x \in [x_{n-1}, x_n] \end{cases} \tag{4.1}$$

En cada uno de los n intervalos, $S_i(x)$ es un polinomio cúbico.

$$S_i(x) = a_i(x - x_i)^3 + b_i(x - x_i)^2 + c_i(x - x_i) + d_i, \quad i = 0, 1, ..., n-1. \tag{4.2}$$

Conocer $S(x)$ quiere decir conocer $4n$ coeficientes: a_i, b_i, c_i, d_i, para $i = 0, 1, ..., n-1$.

Se requiere que $S(x)$ pase por los puntos, y que sea doblemente diferenciable. Los problemas se pueden presentar en los extremos de los intervalos. Entonces,

$$\begin{aligned} S(x_i) &= y_i, \quad i = 0, ..., n \\ S_i(x_{i+1}) &= S_{i+1}(x_{i+1}), \quad i = 0, ..., n-2 \\ S_i'(x_{i+1}) &= S_{i+1}'(x_{i+1}), \quad i = 0, ..., n-2 \\ S_i''(x_{i+1}) &= S_{i+1}''(x_{i+1}), \quad i = 0, ..., n-2 \end{aligned}$$

Sea $h_j = x_{j-1} - x_j$, el tamaño del intervalo $[x_j, x_{j+1}]$. Las condiciones anteriores se convierten en:

$$\begin{aligned} S_i(x_i) = d_i = y_i \qquad & i = 0, ..., n-1, \\ S_{n-1}(x_n) = a_{n-1}h_{n-1}^3 + b_{n-1}h_{n-1}^2 + c_{n-1}h_{n-1} + d_{n-1} = y_n & \\ a_i h_i^3 + b_i h_i^2 + c_i h_i + d_i = d_{i+1} \quad & i = 0, ..., n-2, \\ 3a_i h_i^2 + 2b_i h_i + c_i = c_{i+1} \quad & i = 0, ..., n-2, \\ 6a_i h_i + 2b_i = 2b_{i+1} \quad & i = 0, ..., n-2. \end{aligned}$$

Sea $d_n := y_n$ una variable adicional. Esta variable se utilizará únicamente en las fórmulas intermedias, pero no aparece en las fórmulas finales.

$$\begin{aligned} d_i = y_i \qquad & i = 0, ..., n, & (4.3) \\ a_i h_i^3 + b_i h_i^2 + c_i h_i + d_i = d_{i+1} \qquad & i = 0, ..., n-1, & (4.4) \\ 3a_i h_i^2 + 2b_i h_i + c_i = c_{i+1} \qquad & i = 0, ..., n-2, & (4.5) \\ 3a_i h_i + b_i = b_{i+1} \qquad & i = 0, ..., n-2. & (4.6) \end{aligned}$$

De (4.6):

$$a_i = \frac{b_{i+1} - b_i}{3h_i} \tag{4.7}$$

Reemplazando (4.7) en (4.4):

$$\begin{aligned} \frac{h_i^2}{3}(b_{i+1} - b_i) + b_i h_i^2 + c_i h_i + d_i &= d_{i+1} \\ \frac{h_i^2}{3}(b_{i+1} + 2b_i) + c_i h_i + d_i &= d_{i+1} \end{aligned} \tag{4.8}$$

Reemplazando (4.7) en (4.5):

$$\begin{aligned}(b_{i+1} - b_i)h_i + 2b_i h_i + c_i &= c_{i+1} \\ (b_{i+1} + b_i)h_i + c_i &= c_{i+1}\end{aligned} \tag{4.9}$$

Despejando c_i de (4.8):

$$c_i = \frac{1}{h_i}(d_{i+1} - d_i) - \frac{h_i}{3}(2b_i + b_{i+1}) \tag{4.10}$$

Cambiando i por $i - 1$:

$$c_{i-1} = \frac{1}{h_{i-1}}(d_i - d_{i-1}) - \frac{h_{i-1}}{3}(2b_{i-1} + b_i) \tag{4.11}$$

Cambiando i por $i - 1$ en (4.9):

$$(b_i + b_{i-1})h_{i-1} + c_{i-1} = c_i \tag{4.12}$$

Reemplazando (4.10) y (4.11) en (4.12):

$$(b_i + b_{i-1})h_{i-1} + \frac{1}{h_{i-1}}(d_i - d_{i-1}) - \frac{h_{i-1}}{3}(2b_{i-1} + b_i) = \frac{1}{h_i}(d_{i+1} - d_i) - \frac{h_i}{3}(2b_i + b_{i+1})$$

Las variables d_i son en realidad constantes ($d_i = y_i$). Dejando al lado izquierdo las variables b_j y al lado derecho los términos independientes, se tiene:

$$\frac{h_{i-1}}{3}\, b_{i-1} + \left(\frac{2h_{i-1}}{3} + \frac{2h_i}{3}\right) b_i + \frac{h_i}{3}\, b_{i+1} = \frac{1}{h_{i-1}}(d_{i-1} - d_i) + \frac{1}{h_i}(d_{i+1} - d_i).$$

Multiplicando por 3:

$$h_{i-1}b_{i-1} + 2(h_{i-1} + h_i)b_i + h_i b_{i+1} = \frac{3}{h_{i-1}}(d_{i-1} - d_i) + \frac{3}{h_i}(-d_i + d_{i+1}). \tag{4.13}$$

La igualdad anterior es válida para $i = 1, ..., n - 2$. Es decir, hay $n - 2$ ecuaciones con n incógnitas. El sistema se completa según las condiciones de frontera. Hay dos clases de condiciones sobre $S(x)$. La primera clase se

conoce con el nombre de condiciones de **frontera libre o natural**: en los extremos la curvatura es nula, o sea, $S''(x_0) = 0$ y $S''(x_n) = 0$,

$$\begin{aligned} S_0''(x_0) &= 0, \\ S_{n-1}''(x_n) &= 0. \end{aligned} \tag{4.14}$$

En la segunda clase de condiciones de frontera, **frontera sujeta**, se supone conocida la pendiente de $S(x)$ en los extremos:

$$\begin{aligned} S_0'(x_0) &= f'(x_0), \\ S_{n-1}'(x_n) &= f'(x_n). \end{aligned} \tag{4.15}$$

Al explicitar las condiciones de frontera libre se tiene:

$$S_0''(x) = 6a_0(x - x_0) + 2b_0$$
$$S_{n-1}''(x) = 6a_{n-1}(x - x_{n-1}) + 2b_{n-1}$$
$$S_0''(x_0) = 2b_0 = 0 \tag{4.16}$$
$$S_{n-1}''(x_n) = 3a_{n-1}h_{n-1} + b_{n-1} = 0. \tag{4.17}$$

Además del resultado anterior, $b_0 = 0$, se puede introducir una variable adicional $b_n = 0$. Esto permite que la ecuación (4.13) se pueda aplicar para $i = n - 1$. Recuérdese que ya se introdujo $d_n = y_n$ y que para todo i se tiene $d_i = y_i$. Entonces se tiene un sistema de $n + 1$ ecuaciones con $n + 1$ incógnitas, escrito de la forma

$$Ab = \zeta, \tag{4.18}$$

donde

$$A = \begin{bmatrix} 1 & 0 & 0 & & & & & 0 \\ h_0 & 2(h_0 + h_1) & h_1 & 0 & & & & \\ 0 & h_1 & 2(h_1 + h_2) & h_2 & & & & \\ 0 & 0 & h_2 & 2(h_2 + h_3) & h_3 & & & \\ & & & & & & & \\ & & & & & & & \\ 0 & 0 & & & & h_{n-2} & 2(h_{n-2} + h_{n-1}) & h_{n-1} \\ 0 & 0 & & & & & 0 & 1 \end{bmatrix}$$

$$b = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix}, \qquad \zeta = \begin{bmatrix} 0 \\ \frac{3}{h_0}(y_0 - y_1) + \frac{3}{h_1}(-y_1 + y_2) \\ \frac{3}{h_1}(y_1 - y_2) + \frac{3}{h_2}(-y_2 + y_3) \\ \vdots \\ \frac{3}{h_{n-2}}(y_{n-2} - y_{n-1}) + \frac{3}{h_{n-1}}(-y_{n-1} + y_n) \\ 0 \end{bmatrix}.$$

El sistema (4.18) tiene dos características importantes: es tridiagonal, lo cual facilita su solución; la matriz A es de diagonal estrictamente dominante, lo cual garantiza que A es invertible y que la solución existe y es única.

Una vez conocidos los valores b_0, b_1, ..., b_{n-1}, b_n, se puede aplicar (4.10) para calcular los c_i:

$$c_i = \frac{1}{h_i}(y_{i+1} - y_i) - \frac{h_i}{3}(2b_i + b_{i+1}), \quad i = 0, ..., n-1. \tag{4.19}$$

Como b_n existe y vale 0, la ecuación (4.7) se puede aplicar aún para $i = n-1$.

$$a_i = \frac{b_{i+1} - b_i}{3h_i}, \quad i = 0, ..., n-1. \tag{4.20}$$

Obsérvese que para $i = n-1$, la igualdad $a_{n-1} = (0 - b_{n-1})/(3h_{n-1})$ coincide con la segunda condición de frontera (4.17). El orden de aplicación de las fórmulas es el siguiente:

- $d_i = y_i, \quad i = 0, ..., n-1.$
- Obtener b_0, b_1, ..., b_n resolviendo (4.18).
  En particular $b_0 = 0$ y $b_n = 0$.
- Para $i = 0, ..., n-1$ calcular c_i según (4.19).
- Para $i = 0, ..., n-1$ calcular a_i según (4.20).

Ejemplo 4.1. Construir el trazador cúbico para los puntos $(1,2)$, $(2,2)$, $(3,2)$, $(4,3)$ y $(5,2)$.

De manera inmediata $d_0 = 2$, $d_1 = 2$, $d_2 = 2$ y $d_3 = 3$. Adicionalmente $d_4 = 2$. En este ejemplo $h_0 = h_1 = h_2 = h_3 = 1$. El sistema que permite obtener los b_i es:

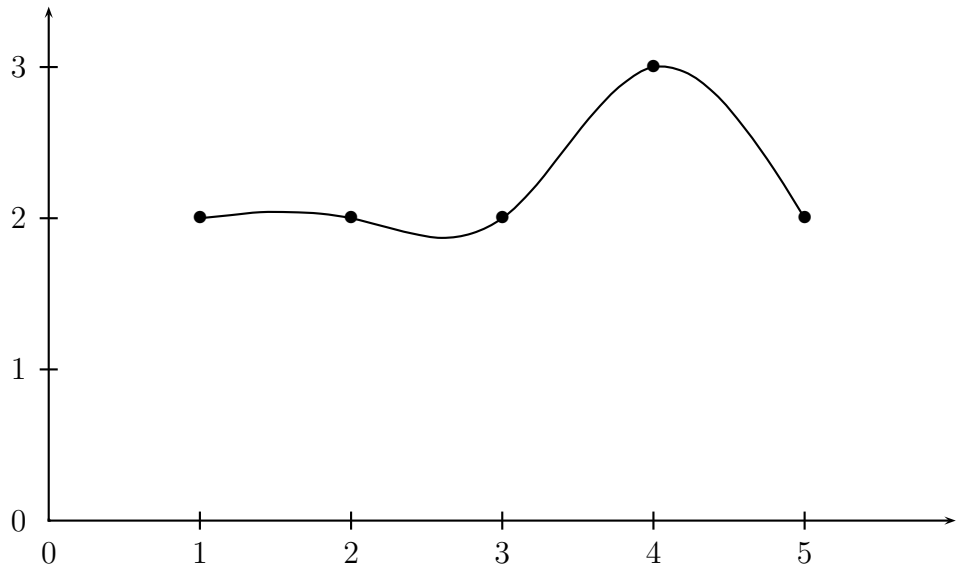
$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 \\ 0 & 1 & 4 & 1 & 0 \\ 0 & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 3 \\ -6 \\ 0 \end{bmatrix}.$$

Al resolver el sistema se obtiene $b_0 = 0$ (obvio), $b_1 = -0.321429$, $b_2 = 1.285714$, $b_3 = -1.821429$ y $b_4 = 0$ (también obvio). El cálculo de los otros coeficientes da:

$$\begin{aligned} c_0 &= \ \ 0.107143 \\ c_1 &= -0.214286 \\ c_2 &= \ \ 0.75 \\ c_3 &= \ \ 0.214286 \\ a_0 &= -0.107143 \\ a_1 &= \ \ 0.535714 \\ a_2 &= -1.035714 \\ a_3 &= \ \ 0.607143. \end{aligned}$$

Entonces

$$\begin{aligned} S_0(x) &= -0.107143(x-1)^3 + 0(x-1)^2 + 0.107143(x-1) + 2 \\ S_1(x) &= 0.535714(x-2)^3 - 0.321429(x-2)^2 - 0.214286(x-2) + 2 \\ S_2(x) &= -1.035714(x-3)^3 + 1.285714(x-3)^2 + 0.75(x-3) + 2 \\ S_3(x) &= 0.607143(x-4)^3 - 1.821429(x-4)^2 + 0.214286(x-4) + 3\,. \end{aligned}$$


3
2
1
0
0
1
2
3
4
5

Capítulo 5

Integración numérica

5.1 Cuadratura adaptativa

Sea $I = \int_a^b f(x)dx$, I_n la aproximación de I por un método fijo de Newton-Cotes (trapecio, Simpson,...) utilizando n subintervalos. La fórmula que relaciona I, I_n y el error global se puede expresar así:

$$I = I_n + F(b-a)h^p f^{(q)}(\xi), \text{ para algún } \xi \in [a,b],$$

donde F, p y q dependen del método escogido; ξ depende del método, de la función f, de n y del intervalo. Entonces

$$\begin{aligned} I &= I_n + F(b-a)(\frac{b-a}{n})^p f^{(q)}(\xi), \\ &= I_n + F\frac{(b-a)^{p+1}}{n^p} f^{(q)}(\xi). \end{aligned}$$

Sea $m = 2n$,

$$I = I_m + F\frac{(b-a)^{p+1}}{n^p 2^p} f^{(q)}(\zeta),$$

Supongamos que

$$f^{(q)}(\xi) \approx f^{(q)}(\zeta).$$

Entonces

$$I \approx I_n + 2^p G \approx I_n + e_n,$$
$$I \approx I_m + G \approx I_n + e_m,$$

donde $G = F\frac{(b-a)^{p+1}}{n^p 2^p} f^{(q)}(\zeta)$, e_n y e_m son los errores. Se puede despejar G:

$$\begin{aligned} e_m \approx G &= \frac{I_m - I_n}{2^p - 1} \\ &= \frac{I_m - I_n}{3} \qquad \text{trapecio} \\ &= \frac{I_m - I_n}{15} \qquad \text{Simpson} \end{aligned} \tag{5.1}$$

Con G se obtiene, supuestamente, una mejor aproximación de I:

$$I \approx I_m + G. \tag{5.2}$$

Los datos para el proceso iterativo para cuadratura adaptativa son: el método (la fórmula de Newton-Cotes), f, a, b, n_0, ε, $n_{\max}$.
Se empieza con un $n = n_0$ (debe ser adecuado) y se obtiene I_n. A partir de ahí se empieza a duplicar el número de subintervalos. El cálculo de la nueva aproximación I_m se hace **sin repetir evaluaciones de la función** f, ya que al duplicar el número de subintervalos los valores $f(x_i)$ de la etapa anterior hacen parte de los valores $f(x_j)$ de la etapa actual. Se calcula G aproximación de e_m, usando (5.1). Si $|G| \leq \varepsilon$, entonces se supone que el error es suficientemente pequeño y se toma como valor final $I_m + G$. En caso contrario, se continua duplicando el número de subintervalos. De todas está previsto un múmero máximo de subintervalos $n_{\max}$, ya que es posible que no se obtenga una aproximación del error suficientemente pequeña.

Ejemplo

$$I = \int_0^\pi \text{sen}(x)dx,$$

utilizando el método del trapecio ($n_0 = 1$) y el de Simpson, ($n_0 = 2$), $\varepsilon = 10^{-8}$

Método del trapecio:

n	I_n	G
1	0.0000000000000002	
2	1.5707963267948966	0.5235987755982988
4	1.8961188979370398	0.1084408570473811
8	1.9742316019455508	0.0260375680028370
16	1.9935703437723395	0.0064462472755962
32	1.9983933609701441	0.0016076723992682
64	1.9995983886400375	0.0004016758899645
128	1.9998996001842038	0.0001004038480554
256	1.9999749002350531	0.0000251000169498
512	1.9999937250705768	0.0000062749451746
1024	1.9999984312683834	0.0000015687326022
2048	1.9999996078171378	0.0000003921829181
4096	1.9999999019542845	0.0000000980457155
8192	1.9999999754885744	0.0000000245114300
16384	1.9999999938721373	0.0000000061278543

$I \approx$ 1.9999999938721373 + 0.0000000061278543= 1.9999999999999916.

Método de Simpson:

n	I_n	G
2	2.0943951023931953	
4	2.0045597549844207	-0.0059890231605850
8	2.0002691699483881	-0.0002860390024022
16	2.0000165910479355	-0.0000168385933635
32	2.0000010333694127	-0.0000010371785682
64	2.0000000645300013	-0.0000000645892941
128	2.0000000040322572	-0.0000000040331829

$I \approx$ 2.0000000040322572 - 0.0000000040331829 = 1.9999999999990743.

Capítulo 6

ECUACIONES DIFERENCIALES PARCIALES

Sea $u = u(x, y)$ una función de dos variables con derivadas parciales de orden dos. Una ecuación diferencial se llama cuasi-lineal si es de la forma

$$Au_{xx} + Bu_{xy} + Cu_{yy} = \varphi(x, y, u, u_x, u_y),$$

dosnde A, B y C son constantes. Hay tres tipos de ecuaciones cuasi-lineales.

elíptica si $B^2 - 4AC < 0$,
parabólica si $B^2 - 4AC = 0$,
hiperbólica si $B^2 - 4AC > 0$.

Un ejemplo típico de una ecuación elíptica es la ecuación de Poisson

$$\nabla^2 u = u_{xx} + u_{yy} = f(x, y).$$

Un caso particular es la ecuación de Laplace

$$u_{xx} + u_{yy} = 0.$$

Un ejemplo típico de una ecuación parabólica es la ecuación unidimensional del calor

$$u_t = c^2 u_{xx}.$$

Un ejemplo típico de una ecuación hiperbólica es la ecuación de onda

$$u_{tt} = c^2 u_{xx}.$$

6.1 Ecuación de Poisson en un rectángulo

Consideraremos un caso particular cuando el dominio es un rectangulo,

$$\begin{aligned} \Omega &= \{(x,y) : a < x < b,\ c < y < d\}, \\ \partial\Omega &= \text{frontera de } \Omega. \end{aligned}$$

La ecuación de Poisson con condiciones de frontera de Dirichlet es la siguiente:

$$\begin{aligned} \triangle u(x,y) &= f(x,y) \text{ en } \Omega, \\ u(x,y) &= g(x,y) \text{ en } \partial\Omega. \end{aligned} \tag{6.1}$$

Hay condiciones de frontera que utilizan derivadas con respecto al vector normal en la frontera. Estas condiciones se llaman condiciones de Neumann.

Resolver numéricamente la ecuación diferencial consiste en obtener aproximaciones de $u(x_i, y_j)$, donde los puntos (x_i, y_j) están en Ω. De manera más precisa,

$$\begin{aligned} &n_x \in \mathbb{Z},\ n_x \geq 1, \\ &n_y \in \mathbb{Z},\ n_y \geq 1, \\ &h_x = \frac{b-a}{n_x+1}, \\ &h_y = \frac{d-c}{n_y+1}, \\ &x_i = a + ih_x, \quad i = 1, ..., n_x, \\ &y_j = c + jh_y, \quad j = 1, ..., n_y, \\ &u_{ij} \approx u(x_i, y_j), \quad i = 1, ...n_x, \quad j = 1, ..., n_y. \end{aligned}$$

Usando la aproximación

$$\varphi''(t) \approx \frac{\varphi(t+h) - 2\varphi(t) + \varphi(t-h)}{h^2}$$

se obtiene

$$\triangle u(x_i, y_j) \approx \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h_x^2} + \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h_y^2}. \tag{6.2}$$

Sea $\eta = h_x/h_y$.

$$\triangle u(x_i, y_j) \approx \frac{u_{i+1,j} - 2u_{ij} + u_{i-1,j}}{h_x^2} + \eta^2 \frac{u_{i,j+1} - 2u_{ij} + u_{i,j-1}}{h_x^2}$$

$$\triangle u(x_i, y_j) \approx \frac{u_{i+1,j} + u_{i-1,j} + \eta^2 u_{i,j+1} + \eta^2 u_{i,j-1} - (2 + 2\eta^2)u_{ij}}{h_x^2}. \tag{6.3}$$

En el caso particular cuando $h = h_x = h_y$

$$\triangle u(x_i, y_j) \approx \frac{u_{i+1,j} + u_{i-1,j} + u_{i,j+1}u_{i,j-1} - 4u_{ij}}{h^2}. \tag{6.4}$$

Al aplicar la aproximación (6.3) en (6.1), y cambiando el signo aproximación por el signo de igualdad, se obtiene

$$-u_{i+1,j} - u_{i-1,j} - \eta^2 u_{i,j+1} - \eta^2 u_{i,j-1} + (2 + 2\eta^2)u_{ij} = -h_x^2 f_{ij}, \tag{6.5}$$

donde $f_{ij} = f(x_i, y_j)$ son valores conocidos. Al considerar los $n_x n_y$ puntos de la malla se obtiene un sistema de $n_x n_y$ ecuaciones con $n_x n_y$ incógnitas. Para simplificar la notación, sean

$$\begin{aligned}
n &= n_x \\
m &= n_y \\
N &= nm \\
h &= h_x \\
\eta &= \frac{h}{h_y} \\
\rho &= \eta^2 \\
\sigma &= 2 + 2\eta^2 \\
\alpha_j &= g(a, y_j) \\
\beta_j &= g(b, y_j) \\
\gamma_i &= g(x_i, c) \\
\delta_i &= g(x_i, d)
\end{aligned}$$

Entonces

$$-u_{i+1,j} - u_{i-1,j} - \rho u_{i,j+1} - \rho u_{i,j-1} + \sigma u_{ij} = -h^2 f_{ij} \tag{6.6}$$

Utilizaremos el siguiente orden para los puntos: primero los puntos de la primera fila (la fila horizontal inferior), en seguida los puntos de la segunda fila, ..., y finalmente los puntos de la fila superior. En cada fila el orden es el usual, de izquierda a derecha.
En este orden se plantean la ecuaciones: la ecuación en (x_1, y_1), en (x_2, y_1), ..., en (x_n, y_1), en (x_1, y_2), ... Para las variables utilizaremos el mismo orden

$$\begin{aligned}
\xi_1 &= u_{11} \\
\xi_2 &= u_{21} \\
&\vdots \\
\xi_n &= u_{n1} \\
\xi_{n+1} &= u_{12} \\
\xi_{n+2} &= u_{22} \\
&\vdots \\
\xi_{2n} &= u_{n2} \\
&\vdots \\
\xi_N &= u_{nm}
\end{aligned}$$

Con el anterior orden para las variables la igualdad (6.6) se reescribe así:

$$-\rho u_{i,j-1} - u_{i-1,j} + \sigma u_{ij} - u_{i+1,j} - \rho u_{i,j+1} = -h^2 f_{ij}$$

El sistema de N ecuaciones con N incógnitas se escribe simplemente:

$$A\xi = v. \tag{6.7}$$

En alguno de los siguientes cuatro casos: $i = 1$, $i = n$, $j = 1$ y $j = m$, alguno(s) de los valores u_{kl} corresponde al valor de u en la frontera. En este caso se utilizan las condiciones de frontera, es decir, los valores de g en el punto de frontera específico. Como son valores conocidos, entonces pasan al lado derecho de la igualdad. A continuación están algunas de las igualdades.

Al plantear la ecuación en el punto (x_1, y_1) se obtiene:

$$-\rho u_{10} - u_{01} + \sigma u_{11} - u_{21} - \rho u_{12} = -h^2 f_{11}.$$

Es necesario cambiar u_{10} por el valor conocido γ_1 y cambiar u_{01} por el valor conocido α_1. Utilizando la notación ξ_k se obtiene:

$$\sigma\xi_1 - \xi_2 - \rho\xi_{n+1} = -h^2 f_{11} + \rho\gamma_1 + \alpha_1.$$

En el punto (x_2, y_1) se obtiene:

$$\begin{aligned} -\rho u_{20} - u_{11} + \sigma u_{21} - u_{31} - \rho u_{22} &= h^2 - f_{21} \\ -\xi_1 + \sigma \xi_2 - \xi_3 - \rho \xi_{n+2} &= -h^2 f_{21} + \rho \gamma_2. \end{aligned}$$

En el punto (x_3, y_1) se obtiene:

$$\begin{aligned} -\rho u_{30} - u_{21} + \sigma u_{31} - u_{41} - \rho u_{32} &= -h^2 f_{31} \\ -\xi_2 + \sigma \xi_3 - \xi_4 - \rho \xi_{n+3} &= -h^2 f_{31} + \rho \gamma_3. \end{aligned}$$

En el punto (x_n, y_1) se obtiene:

$$\begin{aligned} -\rho u_{n0} - u_{n-1,1} + \sigma u_{n1} - u_{n+1,1} - \rho u_{n2} &= -h^2 f_{n1} \\ -\xi_{n-1} + \sigma \xi_n - \rho \xi_{2n} &= -h^2 f_{n1} + \rho \gamma_n + \beta_1. \end{aligned}$$

En el punto (x_1, y_2) se obtiene:

$$\begin{aligned} -\rho u_{11} - u_{02} + \sigma u_{12} - u_{22} - \rho u_{13} &= -h^2 f_{12} \\ -\rho \xi_1 + \sigma \xi_{n+1} - \xi_{n+2} - \rho \xi_{2n+1} &= -h^2 f_{12} + \alpha_2. \end{aligned}$$

En el punto (x_3, y_2) se obtiene:

$$\begin{aligned} -\rho u_{31} - u_{22} + \sigma u_{32} - u_{42} - \rho u_{33} &= -h^2 f_{32} \\ -\rho \xi_3 - \xi_{n+2} + \sigma \xi_{n+3} - \xi_{n+4} - \rho \xi_{2n+3} &= -h^2 f_{32}. \end{aligned}$$

Si $n = n_x = 3$ y $m = n_y = 4$, la matriz A tiene la siguiente forma:

$$A = \begin{bmatrix} \sigma & -1 & 0 & -\rho & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & \sigma & -1 & 0 & -\rho & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & \sigma & 0 & 0 & -\rho & 0 & 0 & 0 & 0 & 0 & 0 \\ -\rho & 0 & 0 & \sigma & -1 & 0 & -\rho & 0 & 0 & 0 & 0 & 0 \\ 0 & -\rho & 0 & -1 & \sigma & -1 & 0 & -\rho & 0 & 0 & 0 & 0 \\ 0 & 0 & -\rho & 0 & -1 & \sigma & 0 & 0 & -\rho & 0 & 0 & 0 \\ 0 & 0 & 0 & -\rho & 0 & 0 & \sigma & -1 & 0 & -\rho & 0 & 0 \\ 0 & 0 & 0 & 0 & -\rho & 0 & -1 & \sigma & -1 & 0 & -\rho & 0 \\ 0 & 0 & 0 & 0 & 0 & -\rho & 0 & -1 & \sigma & 0 & 0 & -\rho \\ 0 & 0 & 0 & 0 & 0 & 0 & -\rho & 0 & 0 & \sigma & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\rho & 0 & -1 & \sigma & -1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\rho & 0 & -1 & \sigma \end{bmatrix}$$

Se puede observar que A es una matriz simétrica, tridiagonal por bloques, de tamaño $m \times m$ bloques, donde cada bloque es de tamaño $n \times n$.

$$A = \begin{bmatrix} D & -\rho I_n & 0 & & \\ -\rho I_n & D & -\rho I_n & & \\ 0 & -\rho I_n & D & & \\ & & & D & -\rho I_n \\ & & & -\rho I_n & D \end{bmatrix}.$$

D es una matriz simétrica tridiagonal de tamaño $n \times n$.

$$D = \begin{bmatrix} \sigma & -1 & 0 & & \\ -1 & \sigma & -1 & & \\ 0 & -1 & \sigma & & \\ & & & \sigma & -1 \\ & & & -1 & \sigma \end{bmatrix}.$$

A es de diagonal positiva dominante. En la mayoría de las filas

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| = 2 + 2\rho = a_{ii}.$$

En algunas filas

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{ij}| < a_{ii}.$$

Para resolver $A\xi = v$ se puede utilizar el método de Gauss si m y n son pequeños. Si N es muy grande no se puede almacenar completamente A como matriz densa y, además, el tiempo de cálculo se vuelve muy grande. Hay varios métodos que pueden ser más eficientes para N grande, algunos son específicos para la ecuación de Poisson en un rectángulo. Por ejemplo se puede utilizar el método de Gauss Seidel o el de sobrerrelajación. Estos dos métodos se pueden implementar sin almacenar los elementos no nulos de A. Conociendo m, n, σ y ρ se tiene toda la información sobre A. También se pueden utilizar métodos basados en la FFT (Fast Fourier Transform). Otros métodos son: el de reducción cíclica, el método FACR (Fourier Analysis Cyclic Reduction) y el método de doble barrido de Cholesky.

Ejemplo 6.1. Resolver la ecuación diferencial

$$\begin{aligned} \triangle u &= 6x + 12y, \quad 1 < x < 13, \quad 2 < y < 7 \\ u(a, y) &= 1 + 2y^3 \\ u(b, y) &= 2197 + 2y^3 \\ u(x, c) &= 16 + x^3 \\ u(x, d) &= 686 + x^3 \end{aligned}$$

con $n_x = 3$ y $n_y = 4$.

Entonces $h_x = 3$, $h_y = 1$, $\rho = 9$, $\sigma = 20$,

$v = [235 \;\; 2529 \;\; 10531 \;\; -519 \;\; -810 \;\; 1353 \;\; -505 \;\; -918 \;\; 1367 \;\; 6319 \;\; 8235 \;\; 16615]^{\mathrm{T}}$.

Al resolver el sistema 12×12 se obtiene

$$u = [118 \;\; 397 \;\; 1054 \;\; 192 \;\; 471 \;\; 1128 \;\; 314 \;\; 593 \;\; 1250 \;\; 496 \;\; 775 \;\; 1432]^{\mathrm{T}}.$$

La ecuación diferencial es muy sencilla, su solución es $u(x, y) = x^3 + 2y^3$. En este caso, la solución numérica obtenida es exacta. $\diamond$

Capítulo 7

VALORES PROPIOS

7.1 Método de la potencia

Este método se puede aplicar para hallar λ_1, el valor propio dominante de una matriz diagonalizable A, cuando éste existe, o sea, si

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|.$$

Sea $\{v^1, v^2, ..., v^n\}$ una base formada por vectores propios asociados a los valores propios $\lambda_1, \lambda_2, ..., \lambda_n$ respectivamente. Sea $x^0 \neq 0$ un vector inicial. Éste se puede expresar como combinación de los vectores propios

$$\begin{aligned} x^0 &= \alpha_1 v^1 + \alpha_2 v^2 + ... + \alpha_n v^n \\ Ax^1 &= \alpha_1 \lambda_1 v^1 + \alpha_2 \lambda_2 v^2 + ... + \alpha_n \lambda_n v^n \\ A^k x^0 &= \alpha_1 \lambda_1^k v^1 + \alpha_2 \lambda_2^k v^2 + ... + \alpha_n \lambda_n^k v^n \\ A^k x^0 &= \alpha_1 \lambda_1^k \left(v^1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k v^i \right) \end{aligned}$$

Esta última factorización está bien definida si $\alpha_1 \neq 0$, o sea, si x^0 no es ortogonal a v^1. Como $|\lambda_i/\lambda_1| < 1$, entonces para valores grandes de k

$$A^k x^0 \approx \alpha_1 \lambda_1^k v^1.$$

De manera análoga

$$A^{k+1} x^0 \approx \alpha_1 \lambda_1^{k+1} v^1.$$

Entonces

$$A^{k+1}x^0 \approx \lambda_1 A^k x^0.$$

Si definimos

$$\xi^k = A^k x^0,$$

entonces

$$\xi^{k+1} \approx \lambda_1 \xi^k.$$

Al tomar ξ_j^k una componente no nula de ξ^k

$$\frac{\xi_j^{k+1}}{\xi_j^k} \approx \lambda_1.$$

El mecanismo anterior puede conducir hasta una buena aproximación de λ_1, pero tiene un inconveniente: $||\xi^k|| \to \infty$. La solución es normalizar. El siguiente esquema, además de incluir la normalización, tiene una manera más eficiente de aproximar λ.

$z^1 = Ax^0$

para $k = 1, ...,$`maxit`

$$x^k = \frac{z^k}{||z^k||_2}$$

$$z^{k+1} = Ax^k$$

$$\lambda_1^k = x^{k^{\mathrm{T}}} z^{k+1}$$

fin-para

El proceso se detiene satisfactoriamente cuando dos aproximaciones, λ_1^k y λ_1^{k-1}, son muy parecidas. La salida no deseada se tiene cuando se llega al número máximo de iteraciones.

La rapidez de la convergencia está ligada al valor $|\lambda_1/\lambda_2|$. Si este valor es cercano a 1 la convergencia es lenta. Si es mucho mayor que 1 la convergencia es rápida.

```
 A  =
  -1.   -2.   -3.
  -4.   -5.   -6.
  -7.   -8.   -8.
x0 :   1.0000000   1.0000000   1.0000000
z1 :  -6.0000000 -15.0000000 -23.0000000
x1 :  -0.2134704  -0.5336761  -0.8183033
z2 :   3.7357324   8.4320816  12.3101276
la1 = -15.370886
xk :   0.2428695   0.5481912   0.8003129
zk+1 :  -3.7401907  -8.5143118 -12.4881199
lak = -15.570253
xk :  -0.2402124  -0.5468287  -0.8020451
zk+1 :   3.7400052   8.5072639  12.4724775
lak = -15.553901
lak = -15.555409
lak = -15.555271
lak = -15.555284
lak = -15.555283
lak = -15.555283
v1 =   -0.2404342   -0.5469432   -0.8019006
```

7.2 Método de la potencia inversa

Este método se puede aplicar para hallar λ_n, el valor propio menos dominante de una matriz diagonalizable e invertible A, cuando éste existe, o sea, si

$$|\lambda_1| \geq |\lambda_2| \geq |\lambda_3| \geq \cdots > |\lambda_n| > 0.$$

Si A es invertible y tiene valores propios λ_1, λ_2,..., λ_n, entonces los valores propios de A^{-1} son

$$\frac{1}{\lambda_1}, \ \frac{1}{\lambda_2}, ..., \frac{1}{\lambda_n}.$$

El valor propio dominante de A^{-1} es justamente $1/\lambda_n$. Entonces se puede aplicar el método de la potencia a A^{-1}. En lugar de escribir explícitamente $z^{k+1} = A^{-1}x^k$ es preferible presentarlo como la solución del sistema $Az^{k+1} = x^k$.

resolver $Az^1 = x^0$

para $k = 1, ...,$`maxit`

$x^k = \dfrac{z^k}{||z^k||_2}$

resolver $Az^{k+1} = x^k$

$\sigma_1^k = x^{k^{\mathrm{T}}} z^{k+1}$

fin-para

Cuando se obtenga la convergencia, $\lambda_n \approx 1/\sigma_1^k$.

```
 A  =
 - 1.  - 2.  - 3.
 - 4.  - 5.  - 6.
 - 7.  - 8.  - 8.
x0 :   1.0000000   1.0000000   1.0000000
z1 :   1.0000000  -1.0000000   0.0000000
x1 :   0.7071068  -0.7071068   0.0000000
z2 :   3.7712362  -5.4211520   2.1213203
sigma1 = 6.500000
xk :   0.5437021  -0.7815718   0.3058324
zk+1 :   3.8398963  -5.8108165   2.4126782
sigmak = 7.367206
xk :   0.5209476  -0.7883366   0.3273210
zk+1 :   3.8187454  -5.8072592   2.4249418
sigmak = 7.361175
sigmak = 7.359910
sigmak = 7.359782
sigmak = 7.359769
sigmak = 7.359768
sigmak = 7.359768

la n = 1/7.359768 =  .1358738
 v1  =    0.5185346  -0.7889596     0.3296431
```