Análisis Numérico Matricial

Bibliografía:

- Antonio José Pallarés Ruiz (2019), Universidad de Murcia. *Introducción y complementos de análisis matricial.*
- P. G. Ciarlet (1990). *Introduction à l'analyse numérique matricielle et à l'optimisation.*
- G. Hammerlin, K. H. Hoffmann (1991). *Numerical Mathematic*s.
- J. Stoer y R. Bulirsch (1993). *Introduction to Numerical Analysis, Second Edition.*
- Wen Li, Ludwig Elsner, Linzhang Lu (2000). *Comparisons of spectral radii and the theorem of Stein–Rosenberg* (https://core.ac.uk/download/pdf/82203897.pdf).

Capítulo 1

Introducción

1.1. Matrices

Dada $M \in \mathcal{M}_{m\times n}(\mathbb{C})$, llamamos **matriz adjunta** de M a $M^* \coloneqq (\overline{M_{ji}})_{ij} \in \mathcal{M}_{n\times m}(\mathbb{C})$ y **matriz traspuesta** de M a $M^t \coloneqq (M_{ji})_{ij} \in \mathcal{M}_{n\times m}(\mathbb{C})$, que coincide con la adjunta cuando los coeficientes son reales, y se tiene $(AB)^* = B^*A^*$ y $A^{**} = A$.

Llamamos **traza** de una matriz $A \in \mathcal{M}_n$ a

$$\mathrm{tr}A \coloneqq \sum_{k=1}^{n} A_{kk}.$$

El **determinante** es la única aplicación $\det : \mathcal{M}_n(\mathbb{K}) \to \mathbb{K}$ multilineal (lineal en cada fila o columna) y alternada (que cambia de signo al permutar dos filas o columnas) que le asocia 1 a la identidad, y cumple $\det(AB) = \det(A)\det(B)$.

Sean $A \in \mathcal{M}_n(\mathbb{K})$, $\lambda \in \mathbb{K}$ y $p \in \mathbb{K}^n \setminus 0$, si $Ap = \lambda p$, λ es un **valor propio** y p es un **vector propio** de A. Los valores propios de A son los ceros de su **polinomio característico**, $p_A(\lambda) \coloneqq \det(A - \lambda I)$. Si estos son $\lambda_1, \ldots, \lambda_n$, el **espectro** de A es $\sigma(A) \coloneqq \{\lambda_1, \ldots, \lambda_n\}$ y su **radio espectral** es $\rho(A) \coloneqq \max\{|\lambda_1|, \ldots, |\lambda_n|\}$.

1.2. Sistemas de ecuaciones

Un **sistema** de m ecuaciones lineales con n incógnitas es uno de la forma

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n}x_n = b_1, \\ \qquad\qquad\qquad\quad \vdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n = b_m. \end{cases}$$

Llamamos **coeficientes** a los escalares a_{ij}, **términos independientes** a los escalares b_i y **soluciones** del sistema a los vectores $(x_1, \ldots, x_n)$ que cumplen todas las igualdades. Un sistema es **compatible** si tiene soluciones, en cuyo caso es **determinado** si solo tiene una o **indeterminado** si tiene más, y en otro caso es **incompatible**. Es **homogéneo** si todos los

términos independientes son 0, en cuyo caso tiene al menos la **solución trivial** 0. Dos sistemas de ecuaciones son **equivalentes** si tienen el mismo conjunto de soluciones.

Llamamos **matriz de coeficientes** a la matriz $m \times n$ $A \coloneqq (a_{ij})_{ij}$, **columna de términos independientes** a la matriz columna $b \coloneqq (b_i)_{ij}$ y **matriz ampliada** a $\left(\begin{array}{c|c} A & b \end{array} \right)$. Entonces podemos representar el sistema como $Ax = b$. Si A es invertible, el sistema es compatible determinado, pues $x = A^{-1}Ax = A^{-1}b$.

1.3. Aplicaciones lineales

Una **base** de un $\mathbb{K}$-espacio vectorial E de dimensión finita es una tupla $(v_1, \ldots, v_n)$ de vectores linealmente independientes de E tal que todo $x \in E$ se puede escribir como

$$\sum_{k=1}^{n} x_k v_k$$

con $x_1, \ldots, x_n \in \mathbb{K}$. Esto nos permite identificar los vectores $x \in E$ con sus coordenadas $(x_1, \ldots, x_n) \in \mathbb{K}^n$, y con la correspondiente matriz columna.

Un **producto escalar** en un $\mathbb{R}$-espacio vectorial E es una función $\langle \cdot, \cdot \rangle : E \times E \to \mathbb{R}$ bilineal simétrica tal que $\forall f \in E \setminus 0, \langle f, f \rangle > 0$.

Llamamos **producto escalar euclídeo** en $\mathbb{R}^n$ a

$$\langle x, y \rangle := \sum_{k=1}^{n} x_k y_k = x^t y = y^t x,$$

producto escalar hermitiano en $\mathbb{C}^n$ a

$$\langle x, y \rangle := \sum_{k=1}^{n} x_k \overline{y_k} = y^* x = \overline{x^* y}.$$

Sean $M \in \mathcal{M}_{m \times n}(\mathbb{C})$, $u \in \mathbb{C}^n$ y $v \in \mathbb{C}^m$, $\langle Mu, v \rangle = \langle u, M^* v \rangle$, y en particular, para $M \in \mathcal{M}_{m \times n}(\mathbb{R})$, $u \in \mathbb{R}^n$ y $v \in \mathbb{R}^m$, $\langle Mu, v \rangle = \langle u, M^t v \rangle$. En efecto, $\langle Me_i, e_j \rangle = \langle (M_{ki})_k, e_j \rangle = M_{ji}$, y $\langle e_i, M^* e_j \rangle = \langle e_i, ((M^*)_{kj})_k \rangle = \langle e_i, (\overline{M_{jk}})_k \rangle = \overline{\overline{M_{ji}}} = M_{ji}$.

Dos vectores $x, y \in \mathbb{C}^n$ son **ortogonales** si $\langle x, y \rangle = 0$.

1.4. Matrices especiales

Una matriz A es **cuadrada** si tiene el mismo número de filas que de columnas, en cuyo caso es **diagonal** si $\forall i \neq j, A_{ij} = 0$, **triangular superior** si $\forall i > j, A_{ij} = 0$ y **triangular inferior** si $\forall i < j, A_{ij} = 0$. Es **simétrica** si $A = A^t$, **hermitiana** si $A = A^*$, **ortogonal** si $A^{-1} = A^t$, **unitaria** si $A^{-1} = A^*$ y **normal** si $AA^* = A^*A$. Si $A \in \mathcal{M}_n(\mathbb{C})$:

1. Existe $U \in \mathcal{M}_n$ unitaria tal que $U^{-1}AU$ es triangular superior.

   Lo probamos primero para U cualquiera. Para $n = 1$ esto es claro. Sea ahora $n > 1$ y supongamos esto probado para $n - 1$. Si $f : \mathbb{C}^n \to \mathbb{C}^n$ es la aplicación lineal asociada a f, por el teorema fundamental del álgebra, el polinomio característico de A tendrá raíz y

por tanto f tendrá un valor propio $\lambda \in \mathbb{C}$ con vector propio asociado v_1. Sean $p_2, \dots, p_n$ tales que $(v_1, p_2, \dots, p_n)$ es base de $\mathbb{C}^n$ y $W \coloneqq \operatorname{span}(p_2, \dots, p_n)$, existen $g : W \to W$ y $\alpha_2, \dots, \alpha_n \in \mathbb{C}$ tales que, para $2 \leq k \leq n$, $f(p_k) = \alpha_k v_1 + g(p_k)$.

Por la hipótesis de inducción, existe una base $(v_2, \dots, v_n)$ de W en la que la matriz de g es triangular superior. Si, para $2 \leq i \leq n$, $v_i =: \sum_{j=2}^{n} \gamma_{ij} p_j$, como $f(v_1) = \lambda v_1$ y, para $2 \leq i \leq n$, $f(v_i) = \left(\sum_{k=2}^{n} \alpha_k \gamma_{ik}\right) v_1 + g(v_i)$, tenemos que la matriz (b_{ij}) de f con la base $(v_1, \dots, v_n)$ es triangular superior.

Por el método de Gram-Schmidt, existe una base ortonormal $(u_1, \dots, u_n)$ tal que, para $1 \leq k \leq n$, $\operatorname{span}(u_1, \dots, u_k) = \operatorname{span}(v_1, \dots, v_k)$, y como $f(v_k) = \sum_{i=1}^{j} b_{ik} v_i$ es combinación lineal de $(v_1, \dots, v_k)$, también lo es de $(u_1, \dots, u_k)$ y f se expresa en la base $(u_1, \dots, u_n)$ como matriz triangular. Como esta base es ortonormal respecto al producto escalar hermitiano, la matriz de paso es unitaria.

2. Si A es normal, existe $U \in \mathcal{M}_n$ unitaria tal que $U^{-1}AU$ es diagonal.

   Existe U unitaria tal que $T \coloneqq U^{-1}AU = U^*AU$ es triangular superior, pero $T^*T = (U^*AU)^*U^*AU = U^*A^*U^{**}U^*AU = U^*A^*UU^*AU = U^*A^*AU$ y $TT^* = U^*AU(U^*AU)^* U^*AUU^*A^*U = U^*AA^*U = U^*A^*AU = T^*T$, luego T es normal. Entonces, para $i < j$, $A_{ij} = (A^*)_{ij} = \overline{A_{ji}} = \overline{0} = 0$, pues $j > i$ y A es triangular superior. Por tanto T es diagonal.

AAlG

Toda matriz simétrica real $A \in \mathcal{M}_m(\mathbb{R})$ admite una matriz ortogonal P tal que $P^{-1}AP = P^tAP$ es diagonal.

Dada $A \in \mathcal{M}_n$, los valores propios de A^*A son no negativos, pues si $p \neq 0$, $A^*Ap = \lambda p \implies \lambda\|p\|^2 = \lambda\langle p, p\rangle = \lambda p^*p = p^*\lambda p = p^*A^*Ap = (Ap)^*(Ap) = \|Ap\|^2 \geq 0 \implies \lambda \geq 0$. Llamamos **valores singulares** de A a las raíces cuadradas de estos valores propios. Entonces:

1. Para $A \in \mathbb{C}^n$ con valores singulares $\mu_1, \dots, \mu_n$, existen U y V unitarias tales que $U^*AV = \operatorname{diag}(\mu_1, \dots, \mu_n)$.

   A^*A es normal, pues $(A^*A)^* = A^*A$ y por tanto $(A^*A)(A^*A)^* = (A^*A)^*(A^*A)$. Por el teorema anterior, existe V unitaria tal que $V^*A^*AV = \operatorname{diag}(\mu_1^2, \dots, \mu_n^2)$, donde los μ_i son los valores singulares de A. Si $f_1, \dots, f_n$ son las columnas de AV, entonces $f_i^* f_j = \mu_i^2 \delta_{ij}$ para $i, j \in \{1, \dots, n\}$. Podemos suponer que los valores singulares nulos son $\mu_1, \dots, \mu_r$, luego $f_1, \dots, f_r = 0$, y haciendo $u_j \coloneqq \frac{f_j}{\mu_j}$ para $j \in \{r+1, \dots, n\}$, queda $u_i^* u_j = \delta_{ij}$ para $i, j \in \{r+1, \dots, n\}$, es decir, $\{u_{r+1}, \dots, u_n\}$ son ortogonales. Completando con vectores ortogonales $u_1, \dots, u_r$ se obtiene una base ortonormal de $\mathbb{C}^n$, y llamando U a la matriz con columnas $(u_1, \dots, u_n)$, queda que

   $$(U^*AV)_{ij} = u_i^* f_j = \begin{cases} 0 & \text{si } 1 \leq j \leq r \\ \mu_j u_i^* u_j & \text{si } r+1 \leq j \leq n \end{cases} = \mu_i \delta_{ij}.$$

2. Para $A \in \mathbb{R}^n$ con valores singulares $\mu_1, \dots, \mu_n$, existen U y V ortogonales tales que $U^tAV = \operatorname{diag}(\mu_1, \dots, \mu_n)$.

   Análogo, viendo que A^tA es simétrica y cambiando adjuntas por traspuestas y unitarias por ortogonales.

1.5. Cocientes de Rayleigh

El **cociente de Rayleigh** de una matriz $A \in \mathcal{M}_n(\mathbb{C})$ es una aplicación $R_A : \mathbb{C}^n \setminus 0 \to \mathbb{C}$ dada por

$$R_A(v) := \frac{\langle Av, v\rangle}{\langle v, v\rangle} = \frac{v^* Av}{v^* v}.$$

Si A es hermitiana, este cociente toma valores reales, pues entonces $\overline{R_A(v)} = \frac{\overline{\langle Av,v\rangle}}{\langle v,v\rangle} = \frac{\langle v,Av\rangle}{\langle v,v\rangle} = \frac{\langle A^*v,v\rangle}{\langle v,v\rangle} = \frac{\langle Av,v\rangle}{\langle v,v\rangle} = R_A(v)$.

Sean $A \in \mathcal{M}_n$ hermitiana con valores propios $\lambda_1 \leq \cdots \leq \lambda_n$, $(p_1, \ldots, p_n)$ una base ortonormal $((p_i, p_j) = \delta_{ij})$ de vectores propios correspondientes $(Ap_k = \lambda_k p_k)$, $E_k \coloneqq \operatorname{span}\{p_1, \ldots, p_k\}$ para cada k $(E_0 = \{0\})$ y $\mathcal{S}_k$ la familia de todos los subespacios de $\mathbb{C}^n$ con dimensión k. Entonces, para $1 \leq k \leq n$:

1. $\lambda_k = R_A(p_k)$.

   Sean U unitaria tal que $D \coloneqq U^*AU = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $v \in \mathbb{C}^n \setminus 0$ y $v = Uw$,

   $$R_A(v) = \frac{v^*Av}{v^*v} = \frac{w^*U^*AUw}{w^*U^*Uw} = \frac{D}{w^*w} = \frac{\sum_k \lambda_k |w_k|^2}{\sum_k |w_k|^2}.$$

   Como w es v expresado respecto de la base $(p_1, \ldots, p_n)$, $Up_k = e_k$, luego $R_A(p_k) = \frac{\lambda_k}{1} = \lambda_k$.

2. $\lambda_k = \max_{v \in E_k \setminus 0} R_A(v)$.

   Si v es de la forma $\sum_{i=1}^k \alpha_i p_i$, $w = (\alpha_1, \ldots, \alpha_k, 0, \ldots, 0)$, y como los valores propios están ordenados,

   $$R_A(v) = \frac{\sum_{i=1}^k \lambda_i |\alpha_i|^2}{\sum_{i=1}^k |\alpha_i|^2} \leq \frac{\sum_{i=1}^k \lambda_k |\alpha_k|^2}{\sum_{i=1}^k |\alpha_k|^2} = \lambda_k = R_A(p_k).$$

3. $\lambda_k = \min_{0 \neq v \perp E_{k-1}} R_A(v)$.

   En este caso, v es de la forma $\sum_{i=k}^n \alpha_i p_i$, y el razonamiento es análogo al del punto anterior.

4. $\lambda_k = \min_{W \in \mathcal{S}_k} \max_{v \in W \setminus 0} R_A(v)$.

   $\geq$] $\lambda_k = \max_{v \in E_k \setminus \{0\}} R_A(v) \overset{E_k \in \mathcal{S}_k}{\leq} \inf_{W \in \mathcal{S}_k} \max_{v \in W \setminus \{0\}} R_A(v)$.

   $\leq$] Queremos ver que $\forall W \in \mathcal{S}_k, \lambda_k \leq \max_{v \in W \setminus \{0\}} R_A(v)$. Si $E_{k-1}^\perp \coloneqq \{v \in V \mid v \perp E_{k-1}\}$, basta ver que para todo subespacio W de dimensión k, $W \cap E_{k-1}^\perp \neq 0$, pues entonces, para $v \in (W \cap E_{k-1}^\perp) \setminus 0$, como $0 \neq v \perp E_{k-1}$, $\lambda_k \leq \min_{0 \neq v \perp E_{k-1}} R_A(v)$. Pero como $E_{k-1}^\perp$ tiene dimensión $n-k+1$, por Grassmann, $\dim(W \cap E_{k-1}^\perp) = \dim W + \dim E_{k-1}^\perp - \dim(W \oplus E_{k-1}^\perp) \leq \dim W + \dim E_{k-1}^\perp - \dim \mathbb{C}^n = k + n - k + 1 - n = 1$.

5. $\lambda_k = \max_{E \in \mathcal{S}_{k-1}} \min_{0 \neq v \perp E} R_A(v)$.

   Análogo.

1.6. Normas matriciales

Sea E un espacio vectorial sobre $\mathbb{R}$ o $\mathbb{C}$, una **norma** sobre E es una aplicación $\|\cdot\| : E \to [0, +\infty)$ tal que:

1. $\|x\| = 0 \iff x = 0$.
2. $\|x + y\| \leq \|x\| + \|y\|$.
3. $\|ax\| = |a|\|x\|$.

Llamamos **espacio vectorial normado** al par $(E, \|\cdot\|)$. La función $d(x, y) = \|x - y\|$ es una distancia en E. Todas las normas en un espacio de dimensión finita son equivalentes, es decir, definen la misma topología.

Si E es un $\mathbb{R}$-espacio vectorial con un producto escalar $\langle\cdot,\cdot\rangle$, $\|\cdot\| : E \to \mathbb{R}$ dada por $\|f\| \coloneqq \sqrt{\langle f, f\rangle}$ define una norma en E. Llamamos **norma** p de $x \in \mathbb{C}^n$ a

$$\|x\|_p := \sqrt[p]{\sum_{k=1}^{n} |x_k|^p},$$

norma euclídea a $\|x\|_2 = \sqrt{\langle x, x\rangle}$ y **norma infinito** a

$$\|x\|_\infty := \max{}_{k=1}^{n} |x_k|.$$

Una **norma matricial** en $\mathcal{M}_n(\mathbb{K})$, donde $\mathbb{K}$ es $\mathbb{R}$ o $\mathbb{C}$, es una que cumple $\forall A, B \in \mathcal{M}_n(\mathbb{K}), \|AB\| \leq \|A\|\|B\|$. Dada una norma $\|\cdot\|$ en $\mathbb{K}^n$, llamamos **norma matricial subordinada** a la norma $\|\cdot\|$ a la norma matricial en $\mathcal{M}_n(\mathbb{K})$ dada por

$$\|A\| := \sup_{x \in \mathbb{K}^n \setminus \{0\}} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\| \leq 1} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|.$$

Entonces, para $A \in \mathcal{M}_n(\mathbb{K})$ y $x \in \mathbb{K}^n$, $\|Ax\| \leq \|A\|\|x\|$.

Sea $A \coloneqq (a_{ij})_{ij} \in \mathcal{M}_n(\mathbb{C})$:

1. $\|A\|_1 = \max_j \sum_i |a_{ij}|$.
   $\sup\{\|Ax\| \mid \|x\| = 1\} = \sup\{\sum_k |Ax|_k \mid \sum_k |x_k| = 1\} = \sup\{\sum_{k,i} |a_{ki}||x_i| \mid \sum_i |x_i| = 1\}$.
   Sea j tal que $\max_i \sum_k |a_{ki}| = \sum_k |a_{kj}|$, para x con $\sum_i |x_i|$,

   $$\sum_{k,i} |a_{ki}||x_i| = \sum_i \left(|x_i| \sum_k |a_{ki}| \right) \leq \left(\sum_i |x_i| \right) \left(\sum_k |a_{kj}| \right) = \sum_k |a_{kj}|,$$

   luego $\sup\{\sum_{k,i} |a_{ki}||x_i| \mid \sum_i |x_i| = 1\} = \max_i \sum_k |a_{ki}|$.

2. $\|A\|_2 = \sqrt{\rho(A^*A)} = \|A^*\|_2$. En particular, $\|A\|_2$ es el mayor valor singular de A, y si A es unitaria o real ortogonal, $\|A\|_2 = 1$.

$$\|A\|_2^2 = \sup \left\{ \frac{\|Ax\|_2^2}{\|x\|_2^2} \;\middle|\; \|x\|_2 = 1 \right\} = \sup \left\{ \frac{\langle Ax, Ax\rangle}{\langle x, x\rangle} = \frac{\langle A^*Ax, x\rangle}{\langle x, x\rangle} = R_{A^*A}(x) \;\middle|\; \|x\|_2 = 1 \right\}$$

pero si $\lambda_1, \ldots, \lambda_m \geq 0$ son los valores propios de A^*A y $E_1, \ldots, E_m$ son los subespacios propios asociados, $\rho(A^*A) = \max\{\lambda_1, \ldots, \lambda_m\} = \max_{k=1}^m \max\{R_{A^*A}(v) \mid v \in E_k \setminus \{0\}\} = \max\{R_{A^*A}(v) \mid v \neq 0\}$, y como

$$R_{A^*A}(v) = \frac{\langle Av, v\rangle}{\langle v, v\rangle} = \left\langle A\frac{v}{\sqrt{\langle v, v\rangle}}, \frac{v}{\sqrt{\langle v, v\rangle}} \right\rangle = \left\langle A\frac{v}{\|v\|_2}, \frac{v}{\|v\|_2} \right\rangle,$$

queda $\rho(A^*A) = \max\{R_{A^*A}(v) \mid v \neq 0\} = \max\{R_{A^*A}(v) \mid \|v\|_2 = 1\} = \|A\|_2^2$.

3. $\|A\|_\infty = \max_i \sum_j |a_{ij}|$.

$$\|A\|_\infty = \sup\{\|Ax\|_\infty \mid \|x\|_\infty = 1\} = \sup\{\max_k |Ax|_k \mid \max_k |x_k| = 1\} =$$
$$= \sup\left\{ \max_k \left| \sum_i a_{ki}x_i \right| \;\middle|\; \max_i |x_i| = 1 \right\} = \max_k \sup\left\{ \left| \sum_i a_{ki}x_i \right| \;\middle|\; \max_i |x_i| = 1 \right\}.$$

Este supremo se alcanza cuando, para cada i, $x_i = 1$ si $a_{ki} > 0$ o $x_i = -1$ si $a_{ki} < 0$, con lo que $\sup\{|\sum_i a_{ki}x_i| \mid \max_i |x_i| = 1\} = |\sum_i |a_{ki}|| = \sum_i |a_{ki}|$, luego $\|A\|_\infty = \max_k \sum_i |a_{ki}|$.

4. Si A es normal, $\|A\|_2 = \rho(A)$.

La **norma euclídea**, $\|A\|_E \coloneqq \sqrt{\sum_{i,j} |a_{ij}|^2}$, es una norma matricial no subordinada a ninguna norma en $\mathbb{K}^n$, pero es más fácil de calcular que $\|\cdot\|_2$, y $\|A\|_2 \leq \|A\|_E \leq \sqrt{n}\|A\|_2$.

Si $A \in \mathcal{M}_n(\mathbb{K})$:

1. Toda norma matricial $\|\cdot\|$ en $\mathcal{M}_n(\mathbb{K})$ cumple $\rho(A) \leq \|A\|$.

   Sean λ un valor propio tal que $|\lambda| = \rho(A)$, $p \neq 0$ un vector propio de λ y $q \in \mathbb{K}^n$ tal que la matriz $pq^t \neq 0$. Entonces $\rho(A)\|pq^t\| = \|\lambda pq^t\| = \|(Ap)q^t\| = \|A(pq^t)\| \leq \|A\|\|pq^t\|$, y despejando, $\rho(A) \leq \|A\|$.

2. Para todo $\varepsilon > 0$ existe una norma matricial subordinada $\|\cdot\|$ tal que $\|A\| \leq \rho(A) + \varepsilon$.

   Sea U la matriz unitaria tal que $U^{-1}AU$ es triangular superior. Entonces la diagonal está formada por los valores propios $\lambda_1, \ldots, \lambda_n$, no necesariamente distintos, de A.

   Sea $D_\delta \coloneqq \operatorname{diag}(1, \delta, \ldots, \delta^{n-1})$ para $\delta > 0$, entonces $(UD_\delta)^{-1}A(UD_\delta) = D_\delta^{-1}U^{-1}AUD_\delta = D_{\delta^{-1}}U^{-1}AUD_\delta$, pero $(a_{ij})D_\delta = (\delta^j a_{ij})$ y $D_{\delta^{-1}}(a_{ij}) = (\delta^{-i}a_{ij})$, luego si $U^{-1}AU = (u_{ij})$, $D_\delta^{-1}U^{-1}AUD_\delta = D_{\delta^{-1}}(u_{ij})D_\delta = D_{\delta^{-1}}(\delta^j u_{ij}) = (\delta^{j-i}u_{ij})$.

   La diagonal no cambia, la matriz sigue siendo triangular superior y, para δ suficientemente pequeño, $\sum_{j=i+1}^n |\delta^{j-i}u_{ij}| < \varepsilon$ para cada i. Así, $\|(\delta^{j-i}u_{ij})\|_\infty = \max_i \sum_j \delta^{j-i}u_{ij} = \max_i(\lambda_i + \sum_{j=i+1}^n \delta^{j-i}u_{ij}) \leq \rho(A) + \varepsilon$. Tomando la norma $\|v\|_* \coloneqq \|(UD_\delta)^{-1}v\|_\infty$, la norma subordinada a esta cumple $\|A\|_* = \|(UD_\delta)^{-1}A(UD_\delta)\|_\infty \leq \rho(A) + \varepsilon$.

De aquí que $\rho(A) = \inf\{\|A\| \mid \|\cdot\| \text{ es una norma matricial en } \mathcal{M}_n(\mathbb{K})\}$.

Sea $B \in \mathcal{M}_n$, $\lim_k B^k = 0$ si y sólo si $\forall v \in \mathbb{K}^n, \lim_k B^k v = 0$, si y sólo si $\rho(B) < 1$, si y sólo si existe una norma subordinada tal que $\|B\| < 1$.

$[1 \implies 2]$ $0 \leq \lim_k \|B^k v\| \leq \lim_k \|B^k\|\|v\| = 0\|v\| = 0$, luego $\lim_k B^k v = 0$.

[$2 \implies 3$] Sea λ un valor propio de A y p un vector propio asociado, entonces $\lim_k B^k p = \lim_k \lambda^k p = p \lim_k \lambda^k = 0$, luego $|\lambda| < 1$.

[$3 \implies 4$] Por el teorema anterior, existe $\|\cdot\|$ tal que $\|B\| < \rho(B) + (1 - \rho(B)) = 1$.

[$4 \implies 1$] Sea $\|\cdot\|$ esta norma, $0 \leq \lim_k \|B^k\| \leq \lim_k \|B\|^k = 0$, luego $\lim_k B^k = 0$.

Toda norma matricial cumple $\lim_k \|B^k\|^{1/k} = \rho(B)$. **Demostración:** Sabemos que $\rho(B) \leq \|B\|$, y como $\rho(B) = \rho(B^k)^{1/k}$, queda $\rho(B) \leq \|B^k\|^{1/k}$ para todo k. Fijado ahora $\varepsilon > 0$, sea $B_\varepsilon := \frac{B}{\rho(B)+\varepsilon}$, se tiene $\rho(B_\varepsilon) < 1$, por lo que $\lim_k B_\varepsilon^k = 0$ y existe k_0 tal que, para $k \geq k_0$, $B_\varepsilon^k \leq 1$, pero entonces $\|B_\varepsilon^k\| = \frac{\|B^k\|}{(\rho(B)+\varepsilon)^k} \leq 1$ y $\|B^k\|^{1/k} \leq \rho(B) + \varepsilon$.

1.7. Análisis del error

Sean $A \in \mathcal{M}_{m\times n}$ invertible, $b \in \mathbb{K}^n \setminus 0$ y $\|\cdot\|$ una norma subordinada:

1. Considerando los sistemas $Ax = b$ y $A(x + \Delta x) = b + \Delta b$, $\frac{\|\Delta x\|}{\|x\|} \leq \|A\|\|A^{-1}\|\frac{\|\Delta b\|}{\|b\|}$.

   Es claro que $A\Delta x = \Delta b$ y por tanto $\Delta x = A^{-1}\Delta b$, con lo que $\|\Delta x\| \leq \|A^{-1}\|\|\Delta b\|$, y como también $\|b\| = \|Ax\| \leq \|A\|\|x\|$, podemos obtener la fórmula despejando.

2. Considerando los sistemas $Ax = b$ y $(A + \Delta A)(x + \Delta x) = b$, $\frac{\|\Delta x\|}{\|x+\Delta x\|} \leq \|A^{-1}\|\|\Delta A\|$ y $\frac{\|\Delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|\|\Delta A\|}{1-\|A^{-1}\|\|\Delta A\|}$.
   $(A + \Delta A)(x + \Delta x) = Ax + A\Delta x + \Delta A(x + \Delta x) = b + A\Delta x + \Delta A(x + \Delta x) = b$, luego $A\Delta x = -\Delta A(x + \Delta x)$ y por tanto $\Delta x = -A^{-1}\Delta A(x + \Delta x)$. Entonces $\|\Delta x\| \leq \|A^{-1}\|\|\Delta A\|\|x+\Delta x\|$, lo que nos da la primera desigualdad. A partir de aquí, $\|x+\Delta x\| \leq \|x\|+\|\Delta x\| \leq \|x\|+\|A^{-1}\|\|\Delta A\|\|x+\Delta x\|$ y por tanto $\|x+\Delta x\|(1-\|A^{-1}\|\|\Delta A\|) \leq \|x\|$, y despejando de esto y la primera desigualdad se obtiene la segunda.

Llamamos **número de condición** de A respecto a la norma $\|\cdot\|$ a $\mathrm{cond}A := \|A\|\|A^{-1}\|$, con lo que si $Ax = b$ y $A(x + \Delta x) = b + \Delta b$ entonces $\frac{\|\Delta x\|}{\|x\|} \leq \mathrm{cond}A\frac{\|\Delta b\|}{\|b\|}$, y si $Ax = b$ y $(A + \Delta)(x + \Delta x) = b$ entonces $\frac{\|\Delta x\|}{\|x+\Delta x\|} \leq \mathrm{cond}A\frac{\|\Delta A\|}{\|A\|}$. Estas desigualdades son las mejores posibles en el sentido de que se pueden encontrar $b, \Delta b \neq 0$ para los que se obtiene la igualdad en la primera desigualdad y $b \neq 0$ y $\Delta A \neq 0$ para los que se obtiene en la segunda.

Llamamos $\mathrm{cond}_p(A) := \|A^{-1}\|_p\|A\|_p$. Para toda $A \in \mathcal{M}_n$ invertible:

1. $\mathrm{cond}A \geq 1$.
2. $\mathrm{cond}A = \mathrm{cond}A^{-1}$.
3. $\forall \alpha \in \mathbb{K} \setminus 0, \mathrm{cond}(\alpha A) = \mathrm{cond}A$.
4. Sean M el mayor valor singular de A y m el menor, $\mathrm{cond}_2 A = \frac{M}{m}$.
5. Si A es normal, sean M el mayor valor propio de A y m el menor, $\mathrm{cond}_2 A = \rho(A)\rho(A^{-1}) = \frac{|\lambda_n(A)|}{|\lambda_1(A)|}$.
6. Sea U una matriz unitaria, $\mathrm{cond}_2 A = \mathrm{cond}_2(UA) = \mathrm{cond}_2(AU) = \mathrm{cond}_2(U^{-1}AU)$.

Sean A diagonalizable, P invertible con $D \coloneqq P^{-1}AP =: \operatorname{diag}(\lambda_i)$, $\|\cdot\|$ una norma con $\|\operatorname{diag}(d_1,\dots,d_n)\| = \max_i |d_i|$ para toda matriz diagonal y $D_i \coloneqq B(\lambda_i, \operatorname{cond}(P)\|\Delta A\|) \subseteq \mathbb{C}$,

$$\sigma(A+\Delta A) \subseteq \bigcup_{i=1}^{n} D_i.$$

Demostración: Sea $\lambda \neq \lambda_1,\dots,\lambda_n$, $D-\lambda I$ es invertible con inversa $\operatorname{diag}(\frac{1}{\lambda_1-\lambda},\dots,\frac{1}{\lambda_n-\lambda})$. Si λ es valor propio de $A+\Delta A$, $A+\Delta A-\lambda I$ no debe tener inversa, por lo que tampoco debe tener inversa $P^{-1}(A+\Delta A-\lambda I)P = P^{-1}AP - P^{-1}\lambda IP + P^{-1}\Delta AP = D-\lambda I + P^{-1}\Delta AP = (D-\lambda I)(I+(D-\lambda I)^{-1}P^{-1}\Delta AP) =: (D-\lambda I)(I+B)$, con lo que $I+B$ no debe ser invertible. Ahora bien, si $\|B\| < 1$,

$$(I+B)\sum_{k=0}^{n}(-1)^k B^k = \sum_{k=0}^{n}(-1)^k(B^k+B^{k+1}) = \sum_{k=0}^{n}(-1)^k B^k + \sum_{k=1}^{n+1}(-1)^{k-1}B^n =$$
$$= \sum_{k=0}^{n}(-1)^k B^k - \sum_{k=1}^{n+1}(-1)^k B^k = B^0 + (-1)^n B^{n+1} = I + (-1)^n B^{n+1},$$

pero $\lim_n B^{n+1} = 0$, luego

$$(I+B)\sum_{k=0}^{\infty}(-1)^k B^k = \lim_n (I+(-1)^n B^{n+1}) = I$$

y $\sum_{k=0}^{\infty}(-1)^k B^k$ es la inversa de $I+B$. Por tanto $\|B\| \geq 1$, luego $1 \leq \|(D-\lambda I)^{-1}P^{-1}\Delta AP\| \leq \|(D-\lambda I)^{-1}\|\|P^{-1}\|\|\Delta A\|\|P\| = \|(D-\lambda I)^{-1}\|\|\Delta A\|\operatorname{cond}(P)$, y como

$$\|(D-\lambda I)^{-1}\| = \max_k \left|\frac{1}{\lambda_k-\lambda}\right| = \frac{1}{\min_k |\lambda_k-\lambda|},$$

queda $1 \leq \frac{\|\Delta A\|\operatorname{cond}(P)}{\min_k |\lambda_k-\lambda|}$ y por tanto $\min_k |\lambda_k-\lambda| \leq \|\Delta A\|\operatorname{cond}(P)$.

Capítulo 2

Métodos directos

Algunos sistemas de ecuaciones lineales $Ax = b$ son fáciles de resolver:

1. Si A es diagonal no singular, para cada $k \in \{1, \ldots, n\}$ es $a_{kk}x_k = b_k$ y por tanto $x_k = \frac{b_k}{a_{kk}}$. La complejidad de resolverlo es $\Theta(n)$.

2. Si A es triangular superior no singular podemos usar el **método ascendente**: para cada k, $\sum_{i=k}^{n} a_{ki}x_i = b_k$ y por tanto $x_k = a_{kk}^{-1}\left(b_k - \sum_{i=k+1}^{n} a_{ki}x_i\right)$, lo que podemos usar para calcular las coordenadas desde x_n hasta x_1, «ascendiendo por el sistema». La complejidad es $\Theta(n^2)$.

3. Si A es triangular inferior no singular podemos usar el **método descendente**: para cada k, $\sum_{i=1}^{k} a_{ki}x_i = b_k$ y por tanto $x_k = a_{kk}^{-1}\left(b_k - \sum_{i=1}^{k-1} a_{ki}x_i\right)$, y podemos calcular las coordenadas desde x_1 hasta x_n, «descendiendo por el sistema». Como antes, la complejidad es $\Theta(n^2)$.

2.1. Factorización LU

Una **factorización LU** de una matriz $A \in \mathcal{M}_n$ es un par (L, U) formado por una matriz triangular inferior $L \in \mathcal{M}_n$ y una superior $U \in \mathcal{M}_n$ tales que $A = LU$. Dado un sistema de ecuaciones lineales $Ax = b$, si (L, U) es una factorización LU de A, $Ax = b \iff LUx = b$, y si L y U son no singulares, podemos obtener x resolviendo consecutivamente $Ly = b$ y $Ux = y$ con los métodos descendente y ascendente respectivamente.

El algoritmo 1 calcula una factorización LU en tiempo $\Theta(n^3)$ siempre que no se obtenga $p = 0$ en algún paso, y su validez se deriva de los comentarios al margen. Para determinar l_{kk} y u_{kk} hay varios criterios:

- **Criterio de Dootlittle**: $l_{kk} = 1$, $u_{kk} = p_k$.
- **Criterio de Crout**: $u_{kk} = 1$, $l_{kk} = p_k$.

Una matriz A no singular admite una factorización LU si y sólo si todos sus menores principales (submatrices cuadradas con las k primeras filas y columnas de A) son no singulares, si y sólo si el algoritmo dado termina con éxito.

Entrada: $A \coloneqq (a_{ij})$, matriz cuadrada de tamaño n.
Salida: Factorización $(L \coloneqq (l_{ij}), U \coloneqq (u_{ij}))$ de A, o error.
Inicializar L y U a 0;
para $k \leftarrow 1$ **a** n **hacer**
  // Hemos calculado las $k-1$ primeras columnas de L y filas de U.
  $p \leftarrow a_{kk} - \sum_{s=1}^{k-1} l_{ks}u_{sk}$; // $a_{kk} = \sum_{s=1}^{k} l_{ks}u_{sk}$
  si $p = 0$ **entonces devolver** error;
  sinó establecer l_{kk} y u_{kk} de forma que $l_{kk}u_{kk} = p$;
  para $i \leftarrow k+1$ **a** n **hacer**
    $u_{ki} \leftarrow l_{kk}^{-1}\left(a_{ki} - \sum_{s=1}^{k-1} l_{ks}u_{si}\right)$; // $a_{ki} = \sum_{s=1}^{k} l_{ks}u_{si}$
    $l_{ik} \leftarrow u_{kk}^{-1}\left(a_{ik} - \sum_{s=1}^{k-1} l_{is}u_{sk}\right)$; // $a_{ik} = \sum_{s=1}^{k} l_{is}u_{sk}$
  fin
fin

Algoritmo 1: Algoritmo de factorización LU.

$1 \implies 2]$ Sea $(L \coloneqq (l_{ij}), U \coloneqq (u_{ij}))$ esta factorización,

$$\det A = \det L \det U = \prod_{k=1}^{n} l_{kk} \prod_{k=1}^{n} u_{kk} \neq 0,$$

luego $l_{11}, \ldots, l_{nn}, u_{11}, \ldots, u_{nn} \neq 0$ y, si A_k es el menor principal de A de tamaño k, $\det A_k = \prod_{i=1}^{k} l_{ii} \prod_{i=1}^{k} u_{ii} \neq 0$.

$2 \implies 3]$ Sea A_k el menor principal de A de tamaño k, por el punto anterior de la prueba, $l_{kk}u_{kk} = \frac{\det A_k}{\det A_{k-1}}$. Si el algoritmo falla es porque $p = 0$ en alguna iteración k, luego $l_{kk}u_{kk} = \frac{\det A_k}{\det A_{k-1}} = 0$, con lo que si los menores $A_1, \ldots, A_{k-1}$ eran no singulares, A_k es singular.

$3 \implies 1]$ Obvio.

2.1.1. Factorización LDU

Una **factorización LDU** de $A \in \mathcal{M}_n$ es una terna (L, D, U) formada por una matriz triangular inferior $L \in \mathcal{M}_n$, una diagonal $D \in \mathcal{M}_n$ y una triangular superior $U \in \mathcal{M}_n$ tales que $A = LDU$ y las diagonales principales de L y U están formadas por unos.

Esta factorización, si existe, es única. **Demostración:** Si $A = L_1D_1M_1^t = L_2D_2M_2^t$ con $L_1, L_2, M_1, M_2 \in \mathcal{M}_n$ triangulares inferiores con unos en la diagonal y $D_1, D_2 \in \mathcal{M}_n$ diagonales, entonces L_2 y M_1^t son invertibles por tener determinante 1 y $L_2^{-1}L_1D_1 = D_2M_2^t(M_1^t)^{-1}$, pero como la matriz a la izquierda de la igualdad es triangular inferior y la de la derecha es triangular superior, ambas son diagonales y por tanto también son diagonales $L_2^{-1}L_1$ y $M_2^t(M_1^t)^{-1}$. Ahora bien, como L_2^{-1} es la traspuesta de la adjunta de L_2, los elementos de su diagonal son $\frac{1}{\det L_2} = 1$ y, como los de L_1 también, y ambas son triangulares inferiores, los de $L_2^{-1}L_1$ también lo son, luego $L_2^{-1}L_1 = I_n$ y $L_1 = L_2$. Análogamente $M_1 = M_2$, y finalmente $D_1 = L_2^{-1}L_1D_1 = D_2M_2^t(M_1^t)^{-1} = D_2$.

A partir de la factorización de Dootlittle $A = LU$, si U es no singular, la factorización LDU de A es $(L, D, \tilde{U})$ con $D \coloneqq \operatorname{diag}(u_{11}, \ldots, u_{nn})$ y $\tilde{U} \coloneqq (u_{ij}/u_{ii})_{ij}$. Así, si A es simétrica con $\det A \neq 0$ y admite una factorización LU con U no singular, existe una única factorización de la forma LDL^t donde L es triangular con unos en la diagonal y D es diagonal, pues sabemos que admite una factorización LDU $A = LDM^t$ y entonces $LDM^t = A = A^t = MDL^t$, pero como esta factorización es única, $L = M$.

2.1.2. Transformaciones de Gauss sin permutar filas

Sean $v \in \mathbb{K}^n$ y $k \in \{1, \ldots, n\}$ con $v_k \neq 0$, se tiene

$$Mv \coloneqq \begin{pmatrix} 1 & & & & & \\ & \ddots & & & 0 & \\ & & 1 & & & \\ & & -\frac{v_{k+1}}{v_k} & 1 & & \\ & 0 & \vdots & & \ddots & \\ & & -\frac{v_n}{v_k} & & & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ \vdots \\ v_k \\ v_{k+1} \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_k \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

M corresponde a la operación de «hacer ceros» bajo v_k multiplicando las filas $k+1$ hasta n por múltiplos de la fila k. M^{-1} es similar a M pero cambiando el signo a los elementos debajo de la diagonal. En efecto, sean M' la matriz descrita y

$$\tau \coloneqq \begin{pmatrix} 0 & \cdots & 0 & \frac{v_{k+1}}{v_k} \\ \vdots & & \vdots & \vdots \\ 0 & \cdots & 0 & \frac{v_n}{v_k} \end{pmatrix} \in \mathcal{M}_{(n-k)\times k},$$

entonces

$$MM' = \left(\begin{array}{c|c} I_k & 0 \\ \hline -\tau & I_{n-k} \end{array}\right) \left(\begin{array}{c|c} I_k & 0 \\ \hline \tau & I_{n-k} \end{array}\right) = \left(\begin{array}{c|c} I_k & 0 \\ \hline 0 & I_{n-k} \end{array}\right) = I_n.$$

Sean ahora $A \in \mathcal{M}_n$ con columnas $A_1, \ldots, A_n$, M_1 una matriz de la forma anterior tal que $M_1 A_1 = (a_{11}, 0, \ldots, 0)$ y

$$A^{(2)} \coloneqq M_1 A =: \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ 0 & a_{22}^{(2)} & \cdots & a_{2n}^{(2)} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{n2}^{(2)} & \cdots & a_{nn}^{(2)} \end{pmatrix}.$$

Si ahora llamamos M_2 a una matriz de la misma forma y tal que $M_2 A_2^{(2)} = (a_{12}, a_{22}^{(2)}, 0, \ldots, 0)$, es fácil ver que $M_2 A_1^{(2)} = A_1^{(2)}$. Definimos $A^{(3)}, \ldots, A^{(n)}$ y $M_3, \ldots, M_{n-1}$ de forma similar, y entonces $A^{(n)} = M_{n-1} \cdots M_1 A$ es triangular superior, suponiendo que se puede construir, lo que ocurre cuando $a_{11} \neq 0$ y, para $k \in \{2, \ldots, n-1\}$, $a_{kk}^{(k)} \neq 0$.

Sabemos que las transformaciones del tipo de $M_1, \ldots, M_n$ no afectan al determinante de la matriz ni el de sus menores, por lo que la matriz es triangulable por el método de eliminación de Gauss sin permutaciones de filas si y sólo si ninguno de sus menores principales hasta $n-1$ es singular, lo que para A no singular equivale a que sea factorizable LU.

En tal caso, sean $L \coloneqq M_1^{-1} \cdots M_{n-1}^{-1}$ y $U \coloneqq A^{(n)}$, entonces $LU = A$, siendo L triangular inferior con unos en la diagonal y U triangular superior, por lo que (L, U) es una factorización de Dootlittle.

2.1.3. Método de Gauss

Para evitar interrumpir una factorización LU al encontrar un cero en la diagonal, podemos intercambiar la fila que da problemas con una de la inferiores. Esto además interesa cuando encontramos un número pequeño para evitar la inestabilidad en los cálculos.

En el método con **elección del pivote parcial**, además de las matrices L y U, se construye una matriz P de permutación de filas de forma que $PA = LU$, con lo que resolver un sistema $Ax = b$ equivale a resolver $LUx = Pb$.

En el algoritmo de Gauss sin permutaciones de filas, se inicializa P a I_n y, al principio de cada etapa k, se busca el $i \in \{k, \ldots, n\}$ con mayor $|a_{ik}^{(k)}|$ y se intercambian las filas i y k tanto en $A^{(k)}$ como en las primeras $k-1$ columnas de L y en P.

En el método con **elección del pivote total**, se construye además una matriz Q de permutación de columnas de forma que $PAQ = LU$, y resolver $Ax = b$ equivale a resolver $LUy = Pb$ y calcular $x = Qy$.

En el algoritmo de Gauss sin permutaciones de filas, se inicializan P y Q a I_n y, al principio de cada etapa k, se busca el $(i, j) \in \{k, \ldots, n\}^2$ con mayor $|a_{ij}^{(k)}|$ y se intercambian las filas i y k en $A^{(k)}$, P y las primeras $k-1$ columnas de L, y las columnas j y k en $A^{(k)}$ y Q.

2.2. Sistemas con matrices especiales

2.2.1. Diagonal estrictamente dominante

Una matriz $A \coloneqq (a_{ij}) \in \mathcal{M}_n(\mathbb{C})$ tiene **diagonal estrictamente dominante** si para $k \in \{1, \ldots, n\}$,

$$|a_{kk}| > \sum_{\substack{j=1 \\ j \neq i}}^{n} |a_{kj}|.$$

Toda matriz con diagonal estrictamente dominante es no singular y admite una factorización LU. **Demostración:** Si $A \coloneqq (a_{ij})$ fuese singular, sus columnas serían linealmente dependientes y existiría $x \neq 0$ con $Ax = 0$. Sea k con $|x_k| = \max_i |x_i|$, como $\sum_j a_{ij} x_{ij} = 0$ para todo i,

$$a_{kk} x_k = -\sum_{\substack{j=1 \\ j \neq k}}^{n} a_{kj} x_j,$$

luego

$$|a_{kk}||x_k| \leq \sum_{\substack{j=1 \\ j \neq k}}^{n} |a_{kj}||x_j| \leq \sum_{\substack{j=1 \\ j \neq k}}^{} |a_{kj}||x_k|$$

y, despejando $|x_k|$, queda que A no es estrictamente dominante.# Como A tiene diagonal estrictamente dominante, $a_{11} > 0$ y se puede utilizar la transformación de Gauss M_1. Como

$B \coloneqq (b_{ij}) \coloneqq M_1 A$ tiene la misma primera fila que A, ceros bajo la diagonal en la primera columna y el resto de elementos son $b_{ij} = a_{ij} - \frac{a_{i1}}{a_{11}} a_{1j}$, para $i \in \{2, \dots, n\}$,

$$\begin{aligned}
|b_{ii}| &= \left| a_{ii} - \frac{a_{i1}}{a_{11}} a_{1i} \right| \geq |a_{ii}| - \frac{|a_{i1}|}{|a_{11}|} |a_{1i}| > \sum_{j \neq i} |a_{ij}| - \frac{|a_{i1}|}{|a_{11}|} |a_{1i}| \\
&= |a_{i1}| + \sum_{j \neq 1,i} \left| b_{ij} + \frac{a_{i1}}{a_{11}} a_{1j} \right| - \frac{|a_{i1}|}{|a_{11}|} |a_{1i}| \\
&\geq |a_{i1}| + \sum_{j \neq 1,i} |b_{ij}| - \sum_{j \neq 1,i} \frac{|a_{i1}|}{|a_{11}|} |a_{1j}| - \frac{|a_{i1}|}{|a_{11}|} |a_{1i}| \\
&= |a_{i1}| + \sum_{j \neq 1,i} |b_{ij}| - \frac{|a_{i1}|}{|a_{11}|} \sum_{j \neq 1} |a_{1j}| > |a_{i1}| + \sum_{j \neq 1,i} |b_{ij}| - \frac{|a_{i1}|}{|a_{11}|} |a_{11}| \\
&= \sum_{j \neq 1,i} |b_{ij}| = \sum_{j \neq i} |b_{ij}|.
\end{aligned}$$

Así, B es estrictamente dominante, por lo que podemos aplicarle la transformación de Gauss en la segunda columna, y por inducción podemos convertir A en triangular superior por el método de Gauss sin intercambio de filas, luego A admite una factorización LU.

Si A es estrictamente dominante, los cálculos para el método de Gauss sin intercambios de filas o columnas serán estables porque los pivotes no serán demasiado pequeños.

2.2.2. Matrices definidas positivas

$A \in \mathcal{M}_n(\mathbb{R})$ es **definida positiva** (**PD**, *positive definite*) si $\forall x \in \mathbb{R}^n \setminus 0, x^t A x > 0$. En tal caso:

1. A no es singular.

   Si lo fuera, las columnas serían linealmente dependientes y existiría $x \in \mathbb{R}^n \setminus 0$ con $Ax = 0$ y por tanto $x^t A x = 0\#$.

2. $\langle x, y \rangle_A \coloneqq y^* A x$ es un producto escalar en $\mathbb{R}^n$ y $\|x\|_A \coloneqq \sqrt{x^* A x}$ es una norma, la **norma euclídea asociada** a A.

3. Para $X \in \mathcal{M}_{n \times k}$ con rango k, $B \coloneqq X^t A X \in \mathcal{M}_k$ es PD.

   Para $x \in \mathbb{R}^k \setminus \{0\}$, $Xx \neq 0$, luego $x^t B x = x^t X^t A X x = (Xx)^t A X x > 0$.

4. Los menores principales de A son PD.

   Para cada $k \in \{1, \dots, n\}$, tomamos X como la submatriz de I_n formada por las k primeras columnas y queda que $X^t A X$, el menor principal de A de tamaño k, es PD.

5. Todos los elementos de la diagonal de A son positivos.

   Si $A =: (a_{ij})_{ij}$, para a_{kk} tomamos X como la columna k-ésima de I_n y queda que $X^t A X = (a_{kk})$ es PD, luego $1 a_{kk} 1 = a_{kk} > 0$.

6. A admite una factorización LDU con matriz diagonal PD.

   Todos los menores principales son PD y por tanto no singulares. Sea (L, D, U) la factorización, L es no singular, luego $L^{-1}A(L^{-1})^t = DU(L^{-1})^t$ es PD, pero al ser $U(L^{-1})^t$ triangular superior con unos en la diagonal, DM^tL^{-t} y $D =: (d_{ij})_{ij}$ tienen la misma diagonal, que tiene todos sus elementos positivos. Entonces, para $x \in \mathbb{R}^n \setminus \{0\}$, $x^tDx = \sum_k d_{kk}x_k^2 > 0$.

2.2.3. Matrices simétricas definidas positivas (SPD)

Sea $A \in \mathcal{M}_n(\mathbb{R})$ simétrica, A es definida positiva si y sólo si todos sus valores propios son positivos, si y sólo si todos sus menores principales tienen determinante positivo, si y sólo si admite una factorización LDU con diagonal definida positiva, si y sólo si existe una matriz triangular inferior con diagonal positiva L_C tal que $A = L_CL_C^t$, en cuyo caso esta es única.

$1 \implies 2]$ Sea p un vector propio, su valor propio asociado es el cociente de Rayleigh $R_A(p) = \frac{p^tAp}{p^tp} > 0$.

$2 \implies 1]$ Por ser A simétrica, existe O ortogonal con $D \coloneqq O^tAO$ diagonal. Sean $\lambda_1, \ldots, \lambda_n$ los elementos de esta diagonal, si $\lambda_1, \ldots, \lambda_n > 0$, es claro que D es PD, y por tanto $A = ODO^t$ también.

$1 \implies 3]$ Los menores principales de A son simétricos y PD, luego todos sus valores propios son positivos y por tanto el determinante es positivo.

$3 \implies 4]$ Como los menores principales son no singulares, A admite una factorización LDU (L, D, U), y los elementos de la diagonal de D, al ser cocientes de menores principales de A, son positivos.

$4 \implies 1]$ Sea (L, D, L^t) la factorización, como D es PD, $LDL^t = A$ también.

$4 \implies 5]$ Sea (L, D, U) la factorización, como D es PD, $\sqrt{D} \coloneqq \operatorname{diag}(\sqrt{D_{11}}, \ldots, \sqrt{D_{nn}})$ tiene diagonal positiva, y como A es simétrica y $U = L^t$, basta tomar $L_C \coloneqq L\sqrt{D}$ y entonces $L_CL_C^t = L\sqrt{D}\sqrt{D}^tL^t = L\sqrt{D}\sqrt{D}U = LDU = A$.

$5 \implies 4]$ Sean D la matriz diagonal formada por los cuadrados de los elementos de la diagonal de L_C y L el resultado de dividir cada columna de L_C por su elemento de la diagonal, entonces (L, D, L^t) es una factorización LDU de A. Si L_C no fuera única, podríamos obtener así factorizaciones LDU de A distintas.#

Cuando existe L triangular inferior con diagonal positiva tal que $A = LL^t$, llamamos a L el **factor L de Choleski** de A.

2.2.4. Matrices tridiagonales

$A \in \mathcal{M}_n$ es **banda** si existen enteros p y q tales que $a_{ij} = 0$ para $j - i \geq p$ o $j - i \leq -q$, y llamamos **ancho de banda** a $p + q - 1$. Si esto se cumple para $p = q = 2$, A es **tridiagonal**.

Dada una matriz tridiagonal

$$A = \begin{pmatrix} b_1 & c_1 & & & \\ a_2 & b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & a_{n-1} & b_{n-1} & c_{n-1} \\ & & & a_n & b_n \end{pmatrix},$$

si $\delta_0, \delta_1 \coloneqq 1$ y, para $k \in \{2, \dots, n\}$, $\delta_k \coloneqq b_k\delta_{k-1} - a_k c_{k-1}\delta_{k-2}$, entonces δ_k es el determinante del menor principal de orden k, y si todos los δ_k son no nulos,

$$A = \begin{pmatrix} 1 & & & & \\ a_2\frac{\delta_0}{\delta_1} & 1 & & & \\ & \ddots & \ddots & & \\ & & a_{n-1}\frac{\delta_{n-3}}{\delta_{n-2}} & 1 & \\ & & & a_n\frac{\delta_{n-2}}{\delta_{n-1}} & 1 \end{pmatrix} \begin{pmatrix} \frac{\delta_1}{\delta_0} & c_1 & & & \\ & \frac{\delta_2}{\delta_1} & c_2 & & \\ & & \ddots & \ddots & \\ & & & \frac{\delta_{n-1}}{\delta_{n-2}} & c_{n-1} \\ & & & & \frac{\delta_n}{\delta_{n-1}} \end{pmatrix}.$$

2.3. Factorización QR

Dado $v \in \mathbb{C}^n$, llamamos **matriz de Householder** de v a

$$H_v \coloneqq \begin{cases} I_n - \frac{2}{v^*v}vv^* & \text{si } v \neq 0, \\ I_n & \text{si } v = 0. \end{cases}$$

Si $a, v \in \mathbb{R}^n$, $H_v a$ es el vector simétrico de a respecto al subespacio $v^\perp$.

Demostración: Para $v = 0$, $v^\perp = \mathbb{C}^n$ y esto es obvio. Sea $v \neq 0$. Entonces

$$H_v a = a - \frac{2}{v^*v}vv^*a = a - \frac{2v^*a}{\|v\|^2}v = a - \frac{2\|v\|\|a\|\cos\alpha}{\|v\|^2}v = a - 2(\|a\|\cos\alpha)\frac{v}{\|v\|},$$

siendo α el ángulo entre a y v, pero $p \coloneqq (\|a\|\cos\alpha)\frac{v}{\|v\|}$ es la proyección de a en $\langle v\rangle$, luego $H_v a = a - 2p$ es la simetría de a sobre $v^\perp$.

H_v es unitaria. **Demostración:** Para $v = 0$ es obvio. Para $v \neq 0$, $(vv^*)^* = v^{**}v^* = vv^*$, luego

$$H_v H_v^* = \left(I_n - \frac{2}{v^*v}vv^*\right)\left(I_n - \frac{2}{v^*v}vv^*\right) = I_n - 4\frac{vv^*}{v^*v} + 4\frac{vv^*vv^*}{v^*vv^*v} = I_n - 4\frac{vv^*}{v^*v} + 4\frac{v\|v\|^2v^*}{\|v\|^2v^*v} = I_n,$$

con lo que $H_v^* = H_v^{-1}$.

Dados $a \in \mathbb{C}^n$ y α tal que $e^{i\alpha} = \operatorname{sign} a_1$, las matrices $A_\gamma \coloneqq H_{a+(\gamma,0,\dots,0)}$ con $\gamma = \pm\|a\|e^{i\alpha}$ cumplen $A_\gamma a = (\mp\|a\|e^{i\alpha}, 0, \dots, 0)$. **Demostración:** Para $a = 0$ es $\gamma = 0$ y $A_\gamma = H_0$, con lo que $A_\gamma a = I_n 0 = 0$. Sean $a \neq 0$, $e_1 \coloneqq (1, 0, \dots, 0)$ y $v_\gamma \coloneqq a + \gamma e_1$, entonces $v_\gamma^* a = (a^* + \overline{\gamma}e_1^*)a = \|a\|^2 + \overline{\gamma}a_1$ y $v_\gamma^* v_\gamma = (a^* + \overline{\gamma}e_1^*)(a + \gamma e_1) = \|a\|^2 + 2\mathrm{Re}(\overline{\gamma}a_1) + |\gamma|^2$, luego

$$\begin{aligned} H_v a &= a - 2\frac{\|a\|^2 + \overline{\gamma}a_1}{\|a\|^2 + 2\mathrm{Re}(\overline{\gamma}a_1) + |\gamma|^2}(a + \gamma e_1) \\ &= \left(1 - 2\frac{\|a\|^2 + \overline{\gamma}a_1}{\|a\|^2 + 2\mathrm{Re}(\overline{\gamma}a_1) + |\gamma|^2}\right)a - 2\gamma\frac{\|a\|^2 + \overline{\gamma}a_1}{\|a\|^2 + 2\mathrm{Re}(\overline{\gamma}a_1) + |\gamma|^2}e_1, \end{aligned}$$

pero para $\gamma = \pm\|a\|e^{i\alpha}$,

$$2\frac{\|a\|^2 + \overline{\gamma}a_1}{\|a\|^2 + 2\mathrm{Re}(\overline{\gamma}a_1) + |\gamma|^2} = 2\frac{\|a\|^2 \pm \|a\|e^{-i\alpha}a_1}{\|a\|^2 + 2\mathrm{Re}(\pm\|a\|e^{-i\alpha}a_1) + \|a\|^2\|e^{i\alpha}\|^2} = \frac{\|a\|^2 \pm \|a\|e^{-i\alpha}a_1}{\|a\|^2 \pm \|a\|e^{-i\alpha}a_1} = 1,$$

luego el coeficiente de a en la última expresión de $H_v a$ es $1-1=0$ y el de e_1 es $-\gamma = \mp\|a\|e^{i\alpha}$.

Una **factorización QR** de una matriz $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ es un par (Q,R) con $Q \in \mathcal{M}_m$ ortogonal y $R \in \mathcal{M}_{m\times n}$ triangular superior. El **método de Householder** para obtener esta factorización es el algoritmo 2, y consiste en usar matrices de Householder $H_1,\ldots,H_m$ para ir eliminando la parte inferior de las columnas de A haciendo $R \coloneqq H_m\cdots H_1 A$ y $Q \coloneqq (H_m\cdots H_1)^{-1} = H_1^{-1}\cdots H_m^{-1} = H_1^*\cdots H_m^*$.

Entrada: Matriz A de tamaño $m\times n$.
Salida: Factorización $(Q, R \coloneqq (r_{ij}))$ de A.
Inicializar $R \leftarrow A$ y $Q \leftarrow I_m$;
para $k \leftarrow 1$ **a** n **hacer**
  Obtener un vector v de tamaño $m-k+1$ tal que $H_v(r_{kk},\ldots,r_{mk}) = (x,0,\ldots,0)$ para algún x;
  $H \leftarrow H_{(0,\ldots,0,v_1,\ldots,v_{m-k+1})} = \left(\begin{array}{c|c} I_{k-1} & 0 \\ \hline 0 & H_v \end{array}\right)$;
  $Q \leftarrow QH^*$; // $QH^*HR = QR = A$
  $R \leftarrow HR$; // Anula $r_{k+1,k},\ldots,r_{m,k}$
fin

Algoritmo 2: Método de Householder de factorización QR

Si $A \in \mathcal{M}_{m\times n}$ tiene rango $n \le m$ y admite una factorización QR (Q,R), sean $A_1,\ldots,A_n$ las columnas de A y $Q_1,\ldots,Q_m$ las de Q, entonces $\mathrm{span}\{A_1,\ldots,A_k\} = \mathrm{span}\{Q_1,\ldots,Q_k\}$ para $k \in \{1,\ldots,n\}$, y $\mathrm{span}\{A_1,\ldots,A_n\}^\perp = \mathrm{span}\{Q_{n+1},\ldots,Q_m\}$. Además, si Q' es la submatriz de Q formada por sus n primeras columnas y R' es la submatriz de R formada por sus n primeras filas, entonces $A = Q'R'$.

Si (Q,R) es una factorización QR de $A \in \mathcal{M}_{m\times n}$, para resolver el sistema $Ax = b$, resolvemos $Rx = Q^*b = Qb$ por el método ascendente. Como las matrices de Householder son ortogonales, el condicionamiento del problema no varía.

En la práctica no se calculan las matrices de Householder, sino su acción sobre los vectores:

$$H_v b = b - 2\frac{v^*b}{\|v\|^2}v; \qquad b^*H_v = b^* - 2\frac{b^*v}{\|v\|^2}v^*.$$

Así, multiplicar una matriz de Householder de tamaño n por un vector (a la izquierda o la derecha) es $\Theta(n)$, con lo que multiplicar una matriz de Householder por una matriz de tamaño $n\times m$ es $\Theta(nm)$, y es fácil ver que la factorización QR de una matriz de tamaño $m\times n$ con el algoritmo indicado es $\Theta(nm(m+n))$.

2.4. Problemas de mínimos cuadrados

Dado un espacio vectorial normado $(E, \|\cdot\|)$, un subespacio G de E y $f \in E$, decimos que $g \in G$ es una **mejor aproximación** de f en G si

$$\|f - g\| = \inf_{h \in G} \|f - h\|.$$

Si G es de dimensión finita sobre $\mathbb{R}$ o $\mathbb{C}$, esta mejor aproximación existe. **Demostración:** Sea $K := \{g \in G \mid \|f - g\| \leq \|f\|\}$, $K \neq \emptyset$ porque $0 \in K$. K es cerrado y acotado al ser la bola $\overline{B}(f, \|f\|)$, luego por estar en $\mathbb{R}^n$ o $\mathbb{C}^n \cong \mathbb{R}^{2n}$, es compacto. Sea $(g_n)_n$ una sucesión en K tal que $\|f - g_n\| \to \inf_{h \in G} \|f - h\|$, por compacidad, $g \in K$, luego g es una mejor aproximación de f.

Un **espacio de Hilbert** es un espacio normado completo. Algunos son:

1. $\mathbb{R}^n$ con la norma euclídea.
2. $\mathcal{C}([a, b])$ con la norma dada por

$$\langle f, g \rangle := \int_a^b f(x) g(x) \omega(x) dx,$$

donde $\omega : [a, b] \to (0, +\infty)$ es continua.

Dado un espacio vectorial E con producto escalar, llamamos **ángulo** entre $f, g \in E \setminus 0$ al $\alpha \in [0, \pi]$ con $\langle f, g \rangle = \|f\| \|g\| \cos \alpha$, y decimos que $f, g \in E$ son **ortogonales**, $f \perp g$, si $\langle f, g \rangle = 0$. Entonces, $\forall f, g \in E$:

1. $\|f + g\|^2 = \|f\|^2 + \|g\|^2 + 2\langle f, g \rangle$.
2. **Teorema de Pitágoras:** $f \perp g \iff \|f + g\|^2 = \|f\|^2 + \|g\|^2$.
3. **Desigualdad de Cauchy-Schwartz:** $|\langle f, g \rangle| \leq \|f\| \|g\|$.
4. **Identidad del paralelogramo:** $\|f + g\|^2 + \|f - g\|^2 = 2\|f\|^2 + 2\|g\|^2$.

Sean E un espacio normado con producto escalar y $C \subseteq E$ no vacío, convexo y completo:

1. $\exists! g \in C : \|g\| = \min_{h \in C} \|h\|$.

   Sea $\alpha := \inf_{h \in C} \|h\|$, para $g, h \in C$, por la identidad del paralelogramo, $\frac{1}{4}\|g - h\|^2 = \frac{1}{2}\|g\|^2 + \frac{1}{2}\|h\|^2 - \frac{1}{4}\|g + h\|^2$, pero $\frac{1}{4}\|g + h\|^2 = \|\frac{g+h}{2}\|^2 \overset{C \text{ convexo}}{\geq} \alpha^2$, luego $\frac{1}{4}\|g - h\|^2 \leq \frac{1}{2}\|g\|^2 + \frac{1}{2}\|h\|^2 - \alpha^2$. Entonces, dada una sucesión $(g_n)_n$ de elementos de C con $\|g_n\| \to \alpha$, $\frac{1}{4}\|g_n - g_m\|^2 \leq \frac{1}{2}\|g_n\|^2 + \frac{1}{2}\|g_m\|^2 - \alpha^2 \to 0$ cuando $n, m \to \infty$, luego $(g_n)_n$ es de Cauchy y, por completitud, convergente hacia un cierto $g \in C$ que cumple $\|g\| = \lim_n \|g_n\| = \alpha$. Si hubiera otro $h \in C$ con $\|h\| = \alpha$, entonces $0 \leq \frac{1}{4}\|g - h\|^2 \leq \frac{1}{2}\alpha^2 + \frac{1}{2}\alpha^2 - \alpha^2 = 0$, luego $g = h$.

2. Para cada $f \in E$ existe una única mejor aproximación $g \in C$ de f en C.

   $f - C = \{f - h\}_{h \in C}$ es convexo y completo, luego existe un único $t \in C$ con $\|t\| = \min_{h \in f - C} \|h\| = \min_{h \in C} \|f - h\|$ y $f - t$ es la mejor aproximación buscada.

Sean E un espacio normado con producto escalar, $G \subseteq E$ un subespacio vectorial y $f \in E$, $g \in G$ es una mejor aproximación de f en G si y sólo si $f - g \perp G$, esto es, $\forall h \in G, f - g \perp h$.

$\Longrightarrow$] Dados $h \in G$ y $a > 0$, como $g - ah \in G$, $\|f-g\|^2 \leq \|f-(g-ah)\|^2 = \|f-g+ah\|^2 = \|f-g\|^2+a^2\|h\|^2+2a\langle f-g,h\rangle$, luego $\langle f-g,h\rangle \geq -\frac{a}{2}\|h\|^2$ y, haciendo $a \to 0$, $\langle f-g,h\rangle \geq 0$, y cambiando h por $-h$, $\langle f-g,-h\rangle = -\langle f-g,h\rangle \geq 0$, luego $\langle f-g,h\rangle = 0$.

$\Longleftarrow$] Para $h \in G$, como $f - g \perp g - h$, $\|f-h\|^2 = \|f-g\|^2 + \|g-h\|^2 \geq \|f-g\|^2$.

Si $G \subseteq E$ es de dimensión finita, $\{g_1, \ldots, g_m\}$ es un conjunto generador de G y $f \in E$, las tuplas $(x_1, \ldots, x_m)$ tales que $f \perp \sum_{k=1}^m x_k g_k$ son precisamente las soluciones del sistema de **ecuaciones normales**

$$\left\{ \sum_{j=1}^m \langle g_j, g_k\rangle x_j = \langle f, g_k\rangle \right\}_{k\in\{1,\ldots,m\}}.$$

En particular, si $(g_1, \ldots, g_m)$ es base de G, el sistema es compatible determinado, y si esta base es ortonormal, la proyección ortogonal de f en G es

$$\sum_{k=1}^m \langle f, g_k\rangle g_k,$$

por lo que el problema de buscar una mejor aproximación por mínimos cuadrados en un subespacio de un espacio de dimensión finita se reduce al de resolver un sistema de ecuaciones lineales conocida una base del subespacio o al de aplicar una fórmula conocida una base ortonormal de este.

Un **sistema lineal sobredeterminado** es uno con más ecuaciones que incógnitas en que la matriz de coeficientes tiene rango máximo. En general estos son incompatibles. No obstante, para $A \in \mathcal{M}_{m\times n}(\mathbb{R})$ con $n \leq m$ tal que $\operatorname{rg} A = n$ y $b \in \mathbb{R}^m$, existe $u \in \mathbb{R}^n$ tal que $\forall x \in \mathbb{R}^n, \|Au-b\| \leq \|Ax-b\|$, y este es la única solución de $A^tAx = A^tb$. **Demostración:** Sean $A_1, \ldots, A_n$ las columnas de A y $G \coloneqq \operatorname{span}(A_1, \ldots, A_n)$, la mejor aproximación $\sum_{k=1}^n A_k x_k$ de b en G existe y para cada $k \in \{1, \ldots, n\}$ cumple

$$(A_k)^tAx = \langle A_k, Ax\rangle = \left\langle A_k, \sum_{j=1}^n A_j x_j \right\rangle = \sum_{j=1}^n \langle A_k, A_j\rangle x_j = \langle A_k, b\rangle = (A_k)^t b,$$

luego cumple $A^tAx = A^tb$.

Para resolver este sistema por factorización LU, hacemos los productos A^tA ($\Theta(n^2m)$) y A^tb ($\Theta(nm)$), factorizamos ($\Theta(n^3)$) y resolvemos por los métodos ascendente y descendente ($\Theta(n^2)$), y el tiempo es $\Theta(n^2(m+n))$.

Por factorización QR, vemos que si (Q, R) es una factorización QR de A, $A^t = (QR)^t = R^tQ^t$, luego $A^tA = R^tQ^tQR = R^tR$ y solo tenemos que resolver $R^tRx = A^tb$. Así, hacemos el producto A^tb ($\Theta(nm)$), factorizamos calculando solo R ($\Theta(n^2m)$) y resolvemos por los métodos ascendente y descendente ($\Theta(n^2)$), y el tiempo es $\Theta(n^2m)$.

Capítulo 3

Métodos iterativos

Sean $A \in \mathcal{M}_{m\times n}(\mathbb{C})$ y $b \in \mathbb{C}^m$, un **método iterativo de resolución** del sistema lineal $Ax = b$ es un par (T, c) con $T \in \mathcal{M}_n$ y $c \in \mathbb{C}^n$ tal que la solución del sistema es el único punto fijo de $\Phi(x) \coloneqq Tx+c$. Si para todo $x \in \mathbb{C}^n$ la sucesión $(x_n)_n$ dada por $x_0 \coloneqq x$ y $x_{k+1} \coloneqq \Phi(x_k)$ converge hacia el punto fijo, (T, c) es **convergente**.

Sea $T \in \mathcal{M}_n$, $\rho(T) < 1$ si y sólo si, para cualesquiera $c, y \in \mathbb{C}^n$, la sucesión $x_0 \coloneqq y$, $x_{k+1} \coloneqq Tx_k + c$, converge.

$\Longrightarrow$] Entonces existe una norma matricial tal que $\|T\| < 1$, y si $\Phi(x) \coloneqq Tx+c$, $\|\Phi(x)-\Phi(y)\| = \|Tx - Ty\| = \|T(x - y)\| \leq \|T\|\|x - y\|$, luego Φ es contractiva y, por el teorema del punto fijo de Banach, la sucesión converge a un punto fijo.

$\Longleftarrow$] Sean $c, v \in \mathbb{C}^n$, x la solución de $x = Tx + c$, $y \coloneqq x - v$ y $(x_n)_n$ la sucesión del enunciado, entonces $x - x_0 = v = T^0v$, y si $x - x_k = T^kv$, entonces $x - x_{k+1} = (Tx+c)-(Tx_k+c) = T(x - x_k) = TT^kv = T^{k+1}v$. Por tanto $\lim_k T^kv = \lim_k(x - x_k) = 0$, y que esto ocurra para v arbitrario equivale a que $\rho(T) < 1$.

Dado un sistema lineal $Ax = b$, si $A = M - N$ con M fácil de invertir, entonces $(M^{-1}N, M^{-1}b)$ es un método iterativo de resolución de $Ax = b$, pues $Ax = Mx - Nx = b \iff Mx = Nx + b \iff x = M^{-1}Nx + M^{-1}b$. El **método iterativo de Richardson** para una matriz $A \coloneqq (a_{ij})$ sin ceros en la diagonal consiste en tomar como matriz fácil de invertir en lo anterior una matriz αI_n para un cierto $\alpha \in \mathbb{C}$.

3.1. Método de Jacobi

En adelante, $Ax = b$ es un sistema lineal tal que A no tiene ceros en la diagonal, D es la matriz diagonal de A, L la matriz formada por los elementos de A bajo la diagonal y U la formada por los elementos sobre la diagonal, de modo que $A = L + D + U$.

Para el método de Jacobi tomamos $M \coloneqq D$ y $N \coloneqq -(L + U)$, y nos queda el método iterativo $(T_J \coloneqq -D^{-1}(L + U), D^{-1}b)$.

Para calcular de forma eficiente, en cada iteración calculamos $r_k \coloneqq Ax_k - b$ y $x_{k+1} = x_k - D^{-1}r_k$, pues $D^{-1}r_k = D^{-1}(Dx_k+(L+U)x_k-b) = x_k-(-D^{-1}(L+U)x_k+D^{-1}b) = x_k-x_{k+1}$.

3.2. Método de Gauss-Seidel

Es como el de Jacobi pero, para calcular una coordenada de x_{k+1}, usamos las coordenadas anteriores, que ya habremos calculado, en vez de las correspondientes de x_k, pues serán una mejor aproximación. Si $A \in \mathcal{M}_n$, para i de 1 a n, calculamos

$$\tilde{r}_{ki} := \sum_{j=1}^{i-1} a_{ij} x_{(k+1)j} + \sum_{j=i}^{n} a_{ij} x_{kj} - b_i$$

y

$$x_{(k+1)i} := x_{ki} - \frac{\tilde{r}_{ki}}{a_{ii}} = \frac{1}{a_{ii}} \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_{(k+1)j} - \sum_{j=i+1}^{n} a_{ij} x_{kj} \right).$$

Esto es el método ($T_G := -(L+D)^{-1}U, (L+D)^{-1}b$), equivalente a tomar $M := L + D$ y $N := -U$. **Demostración:** Despejando,

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

esto es, $((L+D)x_{k+1})_i = b_i - (Ux_k)_i$, luego $(L+D)x_{k+1} = b - Ux_k$.

Tenemos que $\tilde{r}_{ki} = A\tilde{x}_{ki} - b$, donde $\tilde{x}_{ki} = (x_{(k+1)1}, \ldots, x_{(k+1)(i-1)}, x_{ki}, \ldots, x_{kn})$, por lo que podemos usar como condición de parada que $\tilde{r}_k$ sea lo suficientemente pequeño.

Teorema de P. Stein y R. L. Rosenberg (1948)**:** Si $A = D + L + U \in \mathcal{M}_n(\mathbb{R})$ con $L, U \leq 0$, $L, U \neq 0$ y L y U triangulares estrictamente inferior y superior, respectivamente, se da exactamente una de las siguientes afirmaciones:

1. $0 \leq \rho(T_G) \leq \rho(T_J) < 1$.
2. $\rho(T_J) = \rho(T_G) = 1$.
3. $1 < \rho(T_J) \leq \rho(T_G)$.

Por tanto, si el método de Jacobi converge, el de Gauss-Seidel lo hace más rápido, y si diverge, también.

Si $A \in \mathcal{M}_n(\mathbb{R})$ tiene diagonal estrictamente dominante, los métodos de Jacobi y Gauss-Seidel para resolver $Ax = b$ convergen y $\|T_G\|_\infty \leq \|T_J\|_\infty < 1$. **Demostración:** $(T_J)_{ij} = -(1-\delta_{ij})\frac{a_{ij}}{a_{ii}}$, luego $\|T_J\|_\infty = \max_i \sum_j |(T_J)_{ij}| = \max_i \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1$ y $\rho(T_J) < 1$. Sean $y \in \mathbb{R}^n$ y $z := T_G y$, con lo que $(L+D)z = -Uy$, y queremos ver que, para $i \in \{1, \ldots, n\}$, $|z_i| \leq \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| \|y\|_\infty$, y en particular $|z_i| \leq \|T_J\|_\infty \|y\|_\infty$. Para $i = 1$,

$$|a_{11}||z_1| = |(L+D)_{11} z_1| = |((L+D)z)_1| = |-Uy| = \left| \sum_{j=2}^{n} a_{1j} y_j \right| \leq \sum_{j=2}^{n} |a_{1j}||y_j| \leq \sum_{j=2}^{n} |a_{1j}| \|y\|_\infty.$$

Para $i > 1$, usando la fórmula con $b = 0$,

$$\begin{aligned} |z_i| &= \left| \frac{1}{a_{ii}} \left(-\sum_{j=1}^{i-1} a_{ij} z_j - \sum_{j=i+1}^{n} a_{ij} y_j \right) \right| \leq \frac{1}{|a_{ii}|} \left(\sum_{j=1}^{i-1} |a_{ij}||z_j| + \sum_{i=j+1}^{n} |a_{ij}||y_j| \right) \\ &\leq \frac{1}{|a_{ij}|} \left(\sum_{j=1}^{i-1} |a_{ij}| \|T_J\|_\infty + \sum_{j=1}^{i-1} |a_{ij}| \right) \|y\|_\infty \overset{\|T_J\|_\infty < 1}{\leq} \frac{1}{|a_{ij}|} \sum_{j \neq i} |a_{ij}| \|y\|_\infty. \end{aligned}$$

Por tanto $\|T_G y\|_\infty = \|z\|_\infty \leq \|T_J\|_\infty \|y\|_\infty$ y, tomando $y \coloneqq (1, \ldots, 1)_\infty$, $\|T_G\|_\infty = \text{máx}_i \sum_j |(T_G)$ $\|T_G y\|_\infty \leq \|T_J\|_\infty \|y\|_\infty = \|T_J\|_\infty$.

Si A es tridiagonal es $\rho(T_G) = \rho(T_J)^2$, con lo que los métodos de Jacobi y Gauss-Seidel convergen simultáneamente y, entonces, el de Gauss-Seidel converge más rápido. **Demostración:** Vemos primero que si $A : \mathbb{C} \to \mathcal{M}_n(\mathbb{C})$ es de la forma

$$A(\lambda) \coloneqq \begin{pmatrix} \lambda b_1 & c_1 & & & \\ \lambda^2 a_2 & \lambda b_2 & c_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda^2 a_{n-1} & \lambda b_{n-1} & c_{n-1} \\ & & & \lambda^2 a_n & \lambda b_n \end{pmatrix},$$

entonces $\det(A(\lambda)) = \det(A(1))$ para todo $\lambda \neq 0$. En efecto, sea $Q(\lambda) \coloneqq \text{diag}(\lambda, \lambda^2, \ldots, \lambda^n)$, es fácil ver que $A(\lambda) = Q(\lambda) A(1) Q(\lambda)^{-1}$. Los valores propios de T_J son los ceros de $p_J(\lambda) \coloneqq \det(-D^{-1}(L+U) - \lambda I_n)$, que son los mismos que los de $q_J(\lambda) \coloneqq \det(L+U+\lambda D)$. Los de T_G son los ceros de $p_G(\lambda) \coloneqq \det(-(L+D)^{-1}U - \lambda I_n)$, que son los de $q_G(\lambda) \coloneqq \det(U+\lambda L+\lambda D)$. Usando el resultado al principio, para $\lambda \neq 0$, $q_G(\lambda^2) = \det(\lambda^2 D + \lambda^2 L + U) = \det(\lambda^2 L + \lambda(\lambda D) + U) = \det(L + \lambda D + U) = q_J(\lambda)$, y para $\lambda = 0$ esto se cumple por continuidad. Así, λ es valor propio de T_J si y sólo si λ^2 lo es de T_G, de donde se obtiene el resultado.

3.3. Método de relajación

Se trata de encontrar un peso $\omega > 0$ para corregir las coordenadas de x_k poniendo

$$x_{(k+1)i} \coloneqq x_{ki} - \frac{\omega}{a_{ii}} \tilde{r}_{ki}$$

en el método de Gauss-Seidel. Entonces el método es $(T_R(\omega) \coloneqq (D + \omega L)^{-1}((1-\omega)D - \omega U), (D + \omega L)^{-1}\omega)$, que equivale a tomar $M \coloneqq \frac{1}{\omega} D + L$ y $N \coloneqq \frac{1-\omega}{\omega} D - U$. **Demostración:** Ahora es

$$a_{ii} x_{(k+1)i} = a_{ii} x_{ki} - \omega \left(\sum_{j=1}^{i-1} a_{ij} x_{(k+1)j} + \sum_{j=i}^{n} a_{ij} x_{kj} - b_i \right),$$

luego

$$a_{ii} x_{(k+1)i} + \omega \sum_{j=1}^{i-1} a_{ij} x_{(k+1)j} = (1-\omega) a_{ii} x_{ki} - \omega \sum_{j=i+1}^{n} a_{ij} x_{kj} - b_i,$$

con lo que $(D+\omega L)x_{k+1} = (1-\omega)D - \omega U - b$. Entonces, $(D+\omega L)^{-1}((1-\omega)D+\omega U) = (\frac{1}{\omega}D+\omega L)^{-1}(\frac{1-\omega}{\omega D}+U)$.

Teorema de Kahan (1958)**:** $\rho(T_R(\omega)) > |\omega - 1|$, y en particular el método de relajación solo puede ser convergente para $\omega \in (0,2)$. **Demostración:**

$$\det T_R(\omega) = \frac{\det((1-\omega)D - \omega U)}{\det(D+\omega L)} = \frac{\det((1-\omega)D)}{\det(D)} = (1-\omega)^n,$$

con lo que si $\lambda_1, \ldots, \lambda_n$ son los valores propios de $T_R(\omega)$, repetidos según sea su multiplicidad, $\rho(T_R(\omega))^n \geq \prod_{k=1}^n |\lambda_k| = |\det T_R(\omega)| = |1-\omega|^n$.

Si A es SPD, el método de relajación converge para $\omega \in (0,2)$. **Demostración:** Si $M \coloneqq \frac{1}{\omega}D + L$ y $N \coloneqq \frac{1-\omega}{\omega}D - U$, $A = M - N$ y $T_R(\omega) = M^{-1}N$. Además, como $L^t = U$, $M^t + N = \frac{2-\omega}{\omega}D$, que es SPD, y queremos ver que entonces $\rho(M^{-1}N) \leq \|M^{-1}N\|_A < 1$. En dimensión finita, $\|M^{-1}N\|_A = \max\{\|M^{-1}Nv\|_A \mid \|v\|_A = 1\}$. Sean $v \in \mathbb{R}^n$ con $\|v\|_A^2 = 1$ tal que $\|M^{-1}N\|_A = \|M^{-1}Nv\|_A$ y $w \coloneqq M^{-1}Av$, entonces $Mw = Av$, luego

$$\begin{aligned}
\|M^{-1}Nv\|_A^2 &= \langle AM^{-1}Nv, M^{-1}Nv\rangle = \langle AM^{-1}(M-A)v, M^{-1}(M-A)v\rangle = \\
&= \langle Av - AM^{-1}Av, v - M^{-1}Av\rangle = \\
&= \langle Av, v\rangle - \langle AM^{-1}Av, v\rangle - \langle Av, M^{-1}Av\rangle + \langle AM^{-1}Av, M^{-1}Av\rangle = \\
&= 1 - \langle M^{-1}Av, Av\rangle - \langle Mw, w\rangle + \langle Aw, w\rangle = \\
&= 1 - \langle w, Mw\rangle - \langle Mw, w\rangle + \langle (M-N)w, w\rangle = 1 - \langle M^t w, w\rangle - \langle Nw, w\rangle = \\
&= 1 - \langle (M^t + N)w, w\rangle < 1,
\end{aligned}$$

por ser $M^t + N$ definida positiva.

Si A es SPD y tridiagonal, los métodos de Jacobi, Gauss-Seidel y relajación para $0 < \omega < 2$ convergen. Además, el mínimo $\rho(T_R(\omega))$ para $\omega \in (0,2)$ se alcanza en

$$\omega = \omega_0 \coloneqq \frac{2}{1 + \sqrt{1 - \rho(T_J)^2}},$$

con lo que $\rho(T_R(\omega_0)) \leq \rho(T_G) < \rho(T_J)$, y se tiene $\rho(T_R(\omega_0)) = \omega_0 - 1$.

3.4. Método del descenso rápido

Si A es SPD, $x \in \mathbb{R}^n$ es solución de $Ax = b$ si y sólo si minimiza $g(x) \coloneqq x^tAx - 2x^tb$, y para $v \in \mathbb{R}^n \setminus 0$, el mínimo de $h(t) \coloneqq g(x+tv)$ es $\frac{v^t(b-Ax)}{v^tAv}$. **Demostración:** Sean $v \in \mathbb{R}^n \setminus \{0\}$ y $t \in \mathbb{R}$,

$$\begin{aligned}
g(x+tv) &= (x+tv)^tA(x+tv) - 2(x+tv)^tb = \\
&= x^tAx + tv^tAx + tx^tAv + t^2v^tAv - 2x^tb - 2tv^tb = g(x) - 2tv^t(b-Ax) + t^2v^tAv,
\end{aligned}$$

luego el mínimo de h es t_{xv} tal que $h'(t_{xv}) = 2t_{xv}v^tAv - 2v^t(b-Ax) = 0 \iff t_{xv} = \frac{v^t(b-Ax)}{v^tAv}$. En efecto, $h''(t_{xv}) = 2v^tAv > 0$. Así:

$\Longrightarrow$] Si $Ax = b$, $b - Ax = 0$ y $t_{xv} = 0$ para todo $v \neq 0$, luego x es el mínimo de g en cualquier dirección.

$\Longleftarrow$] Si x minimiza g, para cualquier dirección $v \neq 0$, $t_{xv} = 0$, luego $v^t(b - Ax) = 0$ y, en particular, $\|b - Ax\| = 0$ y $Ax = b$.

El vector gradiente

$$\nabla g(x) = \left(\frac{\partial g}{\partial x_1}(x), \ldots, \frac{\partial g}{\partial x_n}(x)\right)$$

es el de máximo crecimiento en x, y el **método del descenso rápido** consiste en partir de un x_0 y hacer $x_{k+1} \coloneqq x_k - \alpha \nabla g(x_k)$, donde α minimiza $g(x_{k+1})$. Se tiene $\nabla g(x) = 2(Ax - b)$. **Demostración:** $g(x) = \sum_{k=1}^{n} a_{kk}x_k^2 + 2\sum_{1 \leq i < j \leq n} a_{ij}x_j - 2\sum_{k=1}^{n} b_k x_k$, luego

$$\frac{\partial g}{\partial x_k}(x) = 2a_{kk}x_k + 2\left(\sum_{i=1}^{k-1} a_{ik}x_i + \sum_{j=k+1}^{n} a_{kj}x_j\right) - 2b_k = 2\left(\sum_{i=1}^{n} a_{ki}x_i - b_k\right)$$

y por tanto $\nabla g(x) = 2(Ax - b)$.

Este método es seguro, pero no se usa en la práctica por ser muy lento.

3.5. Método del gradiente conjugado

Sean A una matriz SPD de dimensión n, $(v_1, \ldots, v_n)$ una base de vectores de $\mathbb{R}^n$ ortogonal respecto a A, $y \in \mathbb{R}^n$ y las secuencias $(r_k)_{k=0}^n$ $(t_k)_{k=1}^n$ y $(x_k)_{k=0}^n$ dadas por

$$r_k \coloneqq b - Ax_k, \qquad t_k \coloneqq \frac{v_k^* r_{k-1}}{v_k^* A v_k}, \qquad x_0 \coloneqq y, \qquad x_{k+1} \coloneqq x_k + t_{k+1}v_{k+1},$$

entonces x_n es la solución del sistema $Ax = b$. **Demostración:** Tenemos $r_0 = b - Ax_0$, $r_1 = b - Ax_1 = b - A(x_0 + t_1 v_1) = b - Ax_0 - t_1 Av_1$, y por inducción $r_k = b - Ax_0 - t_1 Av_1 - \cdots - t_k Av_k = r_0 - t_1 Av_1 - \cdots - t_k Av_k$. Como los v_k son ortogonales, para $j \leq k$, $v_j^* r_k = v_j^* r_0 - t_j v_j^* A v_j$, y por definición de los t_j, $t_j v_j^* A v_j = v_j^* r_{j-1} = v_j^* r_0$, con lo que $v_j^* r_k = v_j^* r_0 - v_j^* r_0 = 0$ y $r_k \perp v_j$ en el producto escalar usual. En particular r_n es ortogonal a todos los v_j y por tanto $Ax_n - b = r_n = 0$.

El **método del gradiente conjugado**, mostrado en el algoritmo 3, calcula los términos de las secuencias a la vez que la base $(v_1, \ldots, v_n)$.

Se puede usar como condición de parada que γ sea suficientemente pequeño, y comprobamos que lo sea al terminar ya que de lo contrario tenemos inestabilidad en los cálculos.

Sea A SPD, llamamos **precondicionamiento** de A a una matriz C fácil de invertir tal que $\tilde{A} \coloneqq C^{-1}A(C^{-1})^t$ es SPD y $\operatorname{cond}_2 \tilde{A} < \operatorname{cond}_2 A$. Llamando $\tilde{x} \coloneqq C^t x$, el sistema $Ax = b$ es equivale al **sistema precondicionado** $(C^{-1}A(C^{-1})^t)\tilde{x} = C^{-1}b$.

Entrada: Matriz $A \in \mathcal{M}_n$ SPD de coeficientes, vector b de términos independientes y vector inicial x_0.
Salida: Solución x de $Ax = b$.

$x \leftarrow x_0$;
$v \leftarrow r \leftarrow b - Ax_0$;
$\gamma \leftarrow \|r_0\|^2$;
para $i \leftarrow 1$ **a** n **hacer**
    $y \leftarrow Av$;
    $t \leftarrow \frac{\gamma}{v \cdot y}$;
    $x \leftarrow x + tv$;
    $r \leftarrow r - ty$;
    $\beta \leftarrow \|r\|_2^2$;
    $s \leftarrow \frac{\beta}{\gamma}$;
    $\gamma \leftarrow \beta$;
    $v \leftarrow r + sv$;
fin

Algoritmo 3: Método del gradiente conjugado.

Capítulo 4

Valores y vectores propios

Teorema de los círculos de Gershgorin: El conjunto de valores propios de $A \in \mathcal{M}_n(\mathbb{C})$ está contenido en

$$\bigcup_{k=1}^{n} \overline{B}\left(a_{kk}, \sum_{\substack{j=1 \\ j\neq k}}^{n} |a_{kj}|\right).$$

Demostración: Si λ no está contenido en este conjunto, para cada k, $|a_{kk} - \lambda| > \sum_{j\neq k} |a_{kj}|$, luego $A - \lambda I$ tiene diagonal estrictamente dominante y por tanto es no singular y λ no es valor propio de A.

4.1. Método de la potencia o del cociente de Rayleigh

Sean $A \in \mathcal{M}_n(\mathbb{C})$ con valores propios $\lambda_1, \ldots, \lambda_n$ dispuestos tal que $|\lambda_1| \geq \cdots \geq |\lambda_n|$ y vectores propios respectivos $v_1, \ldots, v_n$ formando una base ortogonal de $\mathbb{C}^n$, y $p, y \in \mathbb{C}$, si $\langle x_0, v_1\rangle \neq 0$ y $\langle v_1, y\rangle \neq 0$, sean $(x_k)_k$ y $(r_k)_k$ las sucesiones dadas por $x_0 := p$, $x_{k+1} := Ax_k$ y $r_k := \frac{\langle x_{k+1}, y\rangle}{\langle x_k, y\rangle}$, entonces $(r_k)_k$ está bien definida y converge a λ_1, y $\frac{x_{2k}}{\|x_{2k}\|}$ converge a un múltiplo de v_1. **Demostración:** Sean $\phi(x) := \langle x, y\rangle$, $p =: \alpha_1 v_1 + \cdots + \alpha_n v_n$, se tiene $x_k = A^k p$, con lo que suponiendo $|\lambda_1| > |\lambda_2|$,

$$x_k = \alpha_1\lambda_1^k v_1 + \cdots + \alpha_n\lambda_n^k v_n = \lambda_1^k\left(\alpha_1 v_1 + \sum_{j=2}^{n}\left(\frac{\lambda_j}{\lambda_1}\right)^k \alpha_j v_j\right) =: \lambda_1^k(\alpha_1 v_1 + \varepsilon_k).$$

Es claro que $\varepsilon_k \to 0$, luego $\lim_{2k} \frac{x_{2k}}{\|x_{2k}\|} = \lim_k \frac{\alpha_1 v_1 + \varepsilon_{2k}}{\|\alpha_1 v_1 + \varepsilon_{2k}\|} = \frac{\alpha_1 v_1}{\|\alpha_1 v_1\|}$ y por ser ϕ lineal, como $\alpha_1 \neq 0$ y $\phi(v_1) \neq 0$,

$$\lim_k r_k = \lim_k \frac{\phi(x_{k+1})}{\phi(x_k)} = \lim_k \frac{\lambda_1^{k+1}\phi(\alpha_1 v_1 + \varepsilon_{k+1})}{\lambda_1^k \phi(\alpha_1 v_1 + \varepsilon_k)} = \lambda_1 \lim_k \frac{\alpha_1\phi(v_1) + \phi(\varepsilon_{k+1})}{\alpha_1\phi(v_1) + \phi(\varepsilon_{k+1})} = \lambda_1.$$

Es fácil ver cómo se generalizaría esto para cuando los $j \in \{1, \ldots, n\}$ primeros valores propios tienen igual valor absoluto.

El **método de la potencia** o **del cociente de Rayleigh** consiste en tomar $p, y \in \mathbb{C}$ arbitrarios en lo anterior, pues todavía no conocemos v_1, e ir construyendo $(x_k)_k$ y $(r_k)_k$ para obtener el valor propio de A con mayor valor absoluto.

En la práctica no se calcula $(x_k)_k$ directamente, pues puede tender a infinito o cero y esto da errores de condicionamiento. En su lugar se calcula $(y_k)_k$ dada por $y_0 \coloneqq \frac{x_0}{\|x_0\|}$ e $y_{k+1} \coloneqq \frac{Ay_k}{\|Ay_k\|}$, y entonces $r_k = \frac{\langle Ay_k, y\rangle}{\langle y_k, y\rangle}$. En efecto, si $y_k = \frac{x_k}{\|x_k\|}$, $y_{k+1} = \frac{Ay_k}{\|Ay_k\|} = \frac{Ax_k}{\|Ax_k\|} = \frac{x_{k+1}}{\|x_{k+1}\|}$, luego por inducción esto ocurre para todo k, y entonces, como $\|x_{k+1}\| = \|Ax_k\| = \|Ay_k\|\|x_k\|$,

$$r_k = \frac{\langle x_{k+1}, y\rangle}{\langle x_k, y\rangle} = \frac{\|x_{k+1}\|\langle y_{k+1}, y\rangle}{\|x_k\|\langle y_k, y\rangle} = \frac{\|x_{k+1}\|\langle Ay_k, y\rangle}{\|x_k\|\|Ay_k\|\langle y_k, y\rangle} = \frac{\langle Ay_k, y\rangle}{\langle y_k, y\rangle}.$$

Si A es invertible, el **método de la potencia inversa** consiste en aplicar el método de la potencia a A^{-1}, obteniendo el inverso del valor propio de A con menor valor absoluto, pues $Au = \lambda u \iff \lambda^{-1}u = A^{-1}u$. Para ello, se factoriza A y bien se obtiene A^{-1} resolviendo columna a columna o se resuelve en cada paso $Ax_{k+1} = x_k$.

Los valores propios de $A - \mu I$ son de la forma $\lambda - \mu$, siendo λ un valor propio de A, por lo que los métodos de la potencia y la potencia inversa sobre $A - \mu I$, llamados de la potencia y la potencia inversa **con desplazamiento**, nos darían respectivamente el valor propio más lejano y más cercano a μ.

4.2. Método de Jacobi

Sea $A \in \mathcal{M}_n(\mathbb{R})$ simétrica, y por tanto diagonalizable. Entonces el problema de encontrar los valores y vectores propios de una matriz se puede traducir en el de encontrar una matriz ortogonal que diagonalice A, que en el caso de $n = 2$ será un giro. El **método de Jacobi** consiste en construir una sucesión $(O_k)_k$ de giros en planos determinados por dos vectores de la base canónica de forma que $(A_k \coloneqq (O_1 \cdots O_k)^t A(O_1 \cdots O_k))_k$, que podemos obtener como $A_0 = A$ y $A_{k+1} = O_{k+1}^t A_k O_{k+1}$, converja a una matriz diagonal.

Sean $1 \le p < q \le n$, $\theta \in \mathbb{R}$, $A \coloneqq (a_{ij}) \in \mathcal{M}_n(\mathbb{R})$ simétrica,

$$O \coloneqq \begin{array}{c} \begin{array}{ccccccccccc} & & & p & & & & q & & & \\ & & & \downarrow & & & & \downarrow & & & \end{array} \\ \begin{pmatrix} 1 & & & & & & & & & & \\ & \ddots & & & & & & & & & \\ & & 1 & & & & & & & & \\ & & & \cos\theta & & & & \sin\theta & & & \\ & & & & 1 & & & & & & \\ & & & & & \ddots & & & & & \\ & & & & & & 1 & & & & \\ & & & -\sin\theta & & & & \cos\theta & & & \\ & & & & & & & & 1 & & \\ & & & & & & & & & \ddots & \\ & & & & & & & & & & 1 \end{pmatrix} \end{array} \begin{array}{l} \\ \\ \\ \\ \leftarrow p \\ \\ \\ \leftarrow q \\ \\ \\ \end{array},$$

y $B \coloneqq (b_{ij}) \coloneqq O^t A O$, entonces:

1. B es simétrica y cumple $\sum_{i,j} b_{ij}^2 = \sum_{i,j} a_{ij}^2$.

   $B^t = (O^tAO)^t = O^tA^tO = O^tAO = B$, luego B es simétrica. $B^tB = O^tA^tOO^tAO = O^tA^tAO$, luego $\operatorname{tr}(B^tB) = \operatorname{tr}(A^tA)$, pero

   $$\operatorname{tr}(A^tA) = \sum_j (A^tA)_j = \sum_{j,i} (A^*)_{ji} A_{ij} = \sum_{i,j} a_{ij}^2,$$

   y análogamente, $\operatorname{tr}(B^tB) = \sum_{i,j} b_{ij}^2$.

2. Si $a_{pq} \neq 0$, $b_{pq} = 0$ si y sólo si $\cot(2\theta) = \frac{a_{qq}-a_{pp}}{2a_{pq}}$, con lo que el valor de $\theta \in (-\frac{\pi}{4}, \frac{\pi}{4}] \setminus \{0\}$ que cumple esto es único, y para dicho valor, $\sum_k b_{kk}^2 = \sum_k a_{kk}^2 + 2a_{pq}^2$.

   Como

   $$\begin{pmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix},$$

   $$\begin{aligned} b_{pq} &= \cos\theta(a_{pp}\sin\theta + a_{pq}\cos\theta) - \sin\theta(a_{qp}\sin\theta + a_{qq}\cos\theta) \\ &= (a_{pp} - a_{qq})\sin\theta\cos\theta + a_{pq}(\cos^2\theta - \sin^2\theta) \\ &= \frac{a_{pp} - a_{qq}}{2}\sin(2\theta) + a_{pq}\cos(2\theta), \end{aligned}$$

   de donde se obtiene la primera parte del enunciado. Aplicando el punto 1 a las submatrices cuadradas de A y B con las filas y columnas p y q, $a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2 = b_{pp}^2 + b_{qq}^2 + 2b_{pq}^2$. Por la estructura de O, las columnas de AO son las de A excepto A_p y A_q, y dada $C \coloneqq (c_{ij}) \in \mathcal{M}_n(\mathbb{R})$, O^tC tiene las mismas filas que C salvo c_p y c_q. Entonces $b_{kk} = a_{kk}$ para $k \neq i, j$ pero $b_{pq}^2 = 0$, luego $\sum_k b_{kk}^2 = \sum_{k\neq i,j} b_{kk}^2 + b_{pp}^2 + b_{qq}^2 = \sum_{k\neq i,j} a_{kk}^2 + a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2 = \sum_k a_{kk}^2 + 2a_{pq}^2$.

3. Para el θ descrito en el apartado anterior, sean $x \coloneqq \frac{a_{qq}-a_{pp}}{2a_{pq}}$,

   $$t \coloneqq \begin{cases} -x + \sqrt{x^2+1} & \text{si } x \geq 0, \\ -x - \sqrt{x^2+1} & \text{si } x < 0, \end{cases}$$

   $c \coloneqq \frac{1}{\sqrt{1+t^2}}$ y $s \coloneqq \frac{t}{\sqrt{1+t^2}}$, para $i, j \neq p, q$,

   $$\begin{aligned} b_{pp} &= a_{pp} - ta_{pq}, & b_{qq} &= a_{qq} + ta_{pq}, & b_{pq} &= 0, \\ b_{pi} &= b_{ip} = ca_{ip} - sa_{iq}, & b_{qi} &= b_{iq} = sa_{ip} + ca_{iq}, & b_{ij} &= a_{ij}. \end{aligned}$$

   Sean $x \coloneqq \frac{a_{qq}-a_{pp}}{2a_{pq}}$ y $t \coloneqq \tan\theta$. Entonces $x = \cot 2\theta = \frac{\cos^2\theta - \sin^2\theta}{2\sin\theta\cos\theta} = \frac{1-\tan^2\theta}{2\tan\theta} = \frac{1-t^2}{2t}$, luego $t^2 - 2xt - 1 = 0$ y $t = \frac{2x \pm \sqrt{4x^2+4}}{2} = x \pm \sqrt{x^2+1}$, y como $|t| \leq 1$ porque $|\theta| \leq \frac{\pi}{4}$, queda el valor de t dado. Como $|\theta| \leq \frac{\pi}{4}$, $\cos\theta > 0$, y como $\tan^2\theta + 1 = \frac{1}{\cos^2\theta}$, $\cos\theta = c$

y $\sin\theta = s$. Entonces los casos b_{pi}, b_{qi}, b_{ip} y b_{iq} son obvios, y b_{pq} y b_{qp} vienen dados por el ejercicio anterior. Finalmente,

$$\begin{aligned}
b_{pp} &= c(a_{pp}c - a_{qp}s) - s(a_{qp}c - a_{qq}s) = a_{pp}c^2 + a_{qq}s^2 - 2csa_{pq} = \\
&= a_{pp} + s^2(a_{qq} - a_{pp}) - 2csa_{pq} = a_{pp} + \frac{t^2}{t^2+1}x2a_{pq} - \frac{2t}{t^2+1}a_{pq} = \\
&= a_{pp} + \frac{t(1-t^2)}{t^2+1}a_{pq} - \frac{2t}{t^2+1}a_{pq} = a_{pp} - ta_{pq}; \\
b_{qq} &= s(a_{pp}s + a_{pq}c) + c(a_{qp}s + a_{qq}c) = a_{pp}s^2 + a_{qq}c^2 + 2csa_{pq} = \\
&= a_{qq} + s^2(a_{pp} - a_{qq}) + 2csa_{pq} = a_{qq} - \frac{t^2}{t^2+1}x2a_{pq} + \frac{2t}{t^2+1}a_{pq} = \\
&= a_{qq} - \frac{t(1-t^2)}{t^2+1}a_{pq} + \frac{2t}{t^2+1}a_{pq} = a_{qq} + ta_{pq}.
\end{aligned}$$

Con esto, el método de Jacobi clásico es el algoritmo 4, que en cada iteración multiplica implícitamente A por la matriz de giro O que anula a_{pq}.

El **teorema de convergencia del método de Jacobi clásico** nos dice que, dada una matriz $A \in \mathcal{M}_n(\mathbb{R})$ simétrica, si para $k \geq 0$ llamamos A_k a la matriz A tras k iteraciones del bucle y Ω_k a la matriz Ω tras k iteraciones, ignorando la condición de parada y dejando A y Ω sin modificar si A es diagonal, $(A_k)_k$ converge a una matriz diagonal cuya diagonal está formada por los valores propios de A, y si además estos son distintos dos a dos, $(\Omega_k)_k$ converge a una matriz ortogonal cuyas columnas son los correspondientes vectores propios de A, en el mismo orden.

Demostración: Si algún A_k es diagonal, esto ya ocurre, pues Ω_k es ortogonal y $A_k = \Omega_k^t A \Omega_k$, por lo que supondremos que ningún A_k lo es. Por tanto A es de tamaño $n \geq 3$.

Vemos primero que, si $(x_k)_k$ es una sucesión acotada en un $\mathbb{R}$-espacio normado X de dimensión finita con una cantidad finita de puntos de acumulación $a_1, \ldots, a_M$ y

$$\lim_k \|x_{k+1} - x_k\| = 0,$$

entonces $(x_k)_k$ es convergente. Sea

$$\epsilon := \frac{1}{3}\min_{i\neq j}\|a_i - a_j\| > 0.$$

Existe un $k_0 \in \mathbb{N}$ tal que para $k \geq k_0$,

$$x_k \in \bigcup_{k=1}^{M} B(a_k, \epsilon),$$

pues en otro caso existiría una subsucesión $(x_{k_m})_m$ de $(x_k)_k$ tal que $x_{k_m} \notin \bigcup_{i=1}^{M} B(a_k, \epsilon)$, pero esta subsucesión está en un espacio acotado y, por tanto, tiene un punto de acumulación.#

Como $\lim_k \|x_{k+1} - x_k\| = 0$, existe $k_1 \geq k_0$ tal que para $k \geq k_1$, $\|x_{k+1} - x_k\| < \epsilon$. Sea i_0 tal que $x_{k_1} \in B(a_{i_0}, \epsilon)$, entonces $\|x_{k_1+1} - a_{i_0}\| \leq \|x_{k_1+1} - x_{k_1}\| + \|x_{k_1} - a_{i_0}\| < 2\epsilon$, y por la desigualdad triangular, $\|x_{k_1+1} - a_i\| > \varepsilon$ para $i \neq i_0$. Por tanto $x_{k_1+1} \in B(a_{i_0}, \varepsilon)$, y por

Entrada: Matriz simétrica real $A \coloneqq (a_{ij})$ de tamaño n y nivel de tolerancia a errores $e > 0$.

Salida: Vector λ de tamaño n que aproxima los valores propios de A y matriz ortogonal Ω de tamaño n cuyas columnas aproximan los correspondientes vectores propios.

$\Omega \leftarrow I_n$;

mientras $\sum_{1 \leq i < j \leq n} a_{ij} > e$ **hacer**

  // Elegimos p y q por el **criterio de Jacobi clásico**, y por la condición de parada elegida, $a_{pq} > 0$.

  Establecer $p < q$ tales que $|a_{pq}| = \text{máx}_{i<j} |a_{ij}|$;

  $x \leftarrow \frac{a_{qq} - a_{pp}}{2a_{pq}}$;

  si $x \geq 0$ **entonces** $t \leftarrow -x + \sqrt{x^2 + 1}$;

  sinó $t \leftarrow -x - \sqrt{x^2 + 1}$;

  $c \leftarrow \frac{1}{\sqrt{1+t^2}}$;

  $s \leftarrow \frac{t}{\sqrt{1+t^2}}$;

  $b_{pp} \leftarrow a_{pp} - ta_{pq}$;

  $b_{qq} \leftarrow a_{qq} + ta_{pq}$;

  $b_{pq}, b_{qp} \leftarrow 0$;

  para $i \neq p, q$ **hacer**

    $b_{pi}, b_{ip} \leftarrow ca_{ip} - sa_{iq}$;

    $b_{qi}, b_{iq} \leftarrow sa_{ip} + ca_{iq}$;

  fin

  para $i, j \neq p, q$ **hacer** $b_{ij} \leftarrow a_{ij}$;

  $A \leftarrow (b_{ij})$; // $A \leftarrow O^t AO$

  para $i \leftarrow 1$ **a** n **hacer**

    $o_{ip} \leftarrow c\omega_{ip} - s\omega_{iq}$;

    $o_{iq} \leftarrow s\omega_{ip} - c\omega_{iq}$;

    para $j \neq p, q$ **hacer** $o_{ij} \leftarrow \omega_{ij}$;

  fin

  $\Omega \leftarrow (o_{ij})$; // $\Omega \leftarrow \Omega * O$

fin

$\lambda \leftarrow$ diagonal de A;

Algoritmo 4: Método de Jacobi clásico.

inducción, $x_k \in B(a_{i_0}, \varepsilon)$ para todo $k \geq k_1$, con lo que solo hay un punto de acumulación, a_{i_0}, y entonces $\lim_k x_k = a_{i_0}$.

Para la primera parte del teorema, sean $A_k =: (a_{kij})_{ij}$ y $\varepsilon_k \coloneqq \sum_{i \neq j} (a_{kij})^2$. Dados los p y q de la iteración k, restringiéndonos a la submatriz de A formada por las filas y columnas p y q, la suma de los elementos de los 4 coeficientes se conserva tras el giro, luego $2a_{kpq}^2 + a_{kpp}^2 + a_{kqq}^2 = a_{(k+1)pp}^2 + a_{(k+1)qq}^2$. Además, la suma de los cuadrados de los elementos de A no cambia, y tampoco cambian los elementos de su diagonal distintos de p y q, y como ε_k es la suma de los cuadrados de los elementos de A_k fuera de la diagonal, $\varepsilon_{k+1} = \varepsilon_k - 2(a_{kpq})^2$.

Como $a_{pq} = \max_{i \neq j} |a_{ij}|$, $\varepsilon_k \leq n(n-1)a_{kpq}^2$, pues $n(n-1)$ es el número de elementos fuera de la diagonal principal. Así, $a_{kpq}^2 \geq \frac{\varepsilon_k}{n(n-1)}$ y

$$\varepsilon_{k+1} \leq \left(1 - \frac{2}{n(n-1)}\right) \varepsilon_k,$$

de donde $\lim_k \varepsilon_k = 0$ y los elementos de A_k fuera de la diagonal convergen a 0, y queda ver que los elementos de la diagonal también convergen.

Sea $D_k \coloneqq \operatorname{diag}(a_{k11}, \ldots, a_{knn})$. Si $(D_{k_m})_m$ una subsucesión de $(D_k)_k$, por continuidad, y como los elementos de A_{k_m} fuera de la diagonal convergen a 0, $\lim_m \det(\lambda I_n - A_{k_m}) = \det(\lambda I_n - D)$. Pero A_k y A son semejantes, luego tienen el mismo polinomio característico y, por tanto, este coincide con el de D. Así, los elementos de la diagonal de D son los valores propios de A y con las mismas multiplicidades. Por tanto, los puntos de acumulación de $(D_k)_k$ son las diagonales formadas por los valores propios de A en distinto orden, de las que hay un máximo de $n!$, y en particular $(D_k)_k$ tiene una cantidad finita de puntos de acumulación.

Tenemos que $\lim_k (D_{k+1} - D_k) = 0$. En efecto,

$$a_{(k+1)ii} - a_{kii} = \begin{cases} 0, & i \neq p, q; \\ -\tan\theta_k a_{kpq}, & i = p; \\ \tan\theta_k a_{kpq}, & i = q; \end{cases}$$

pero $|\tan\theta_k| \leq 1$ por ser $\theta_k \in (-\frac{\pi}{4}, \frac{\pi}{4}]$, y $|a_{kpq}| \leq \sqrt{\varepsilon_k} \to 0$. Además, $(D_k)_k$ está acotada, pues $\|D_k\|_E \leq \|A_k\|_E = \|A\|_E$. Aplicando la propiedad al principio, $(D_k)_k$ converge a una diagonal formada por los valores propios de A en algún orden, y por tanto $(A_k)_k$ también.

Para la segunda parte, sea $(\Omega_{k_m})_m$ una subsucesión de $(\Omega_k)_k$ que converge a un cierto punto acumulación P de $(\Omega_k)_k$, entonces $\Omega_{k_m}^t \to P^t$ e $I_n = \Omega_{k_m}^t \Omega_{k_m} \to P^t P$, luego $P^t P = I_n$ y P es ortogonal. Como $\Omega_k^t A \Omega_k \to D$, $P^t A P = D$, lo que implica que las columnas de P forman una base ortonormal de vectores propios asociados a los valores propios en D. Como estamos suponiendo que todos los valores propios son distintos, cada uno tiene un subespacio propio de dimensión 1 y hay exactamente dos vectores propios ortonormales, uno opuesto del otro, para cada valor propio, pudiendo escribir

$$P =: \begin{pmatrix} | & & | \\ \pm p_1 & \cdots & \pm p_n \\ | & & | \end{pmatrix},$$

con lo que los puntos de acumulación de $(\Omega_k)_k$ solo se diferencian en el signo de las columnas y por tanto hay un máximo de 2^n, en particular una cantidad finita.

Dada una matriz ortogonal O, $\|O\|_2 = 1$, y como todas las normas en $\mathcal{M}_n(\mathbb{R}) \cong \mathbb{R}^{n^2}$ son equivalentes, existe $\beta > 0$ tal que $\|O\|_E \leq \beta \|O\|_2 = \beta$. Por tanto $(\Omega_k)_k$ está acotada.

Si θ_k es tal que $t = \tan(2\theta_k)$ en la iteración k, en esta iteración,

$$\tan(2\theta_k) = \frac{2a_{kpq}}{a_{kqq} - a_{kpp}}.$$

Como cada $(a_{kii})_k$ converge a un valor propio, existe k_0 tal que, para $k \geq k_0$,

$$\min_{i \neq j} |a_{kii} - a_{kjj}| \geq \frac{1}{2} \min_{i \neq j} |\lambda_i - \lambda_j| =: M > 0,$$

con lo que $|a_{kqq} - a_{kpp}| \geq M$ y, como todos los elementos de A_k fuera de la diagonal principal tienden a 0, $(a_{kpq})_k$ tiende a cero (aunque p y q cambien según k), $\tan\theta_k \to 0$ y, como $|\theta_k| \leq \frac{\pi}{4}$, $\theta_k \to 0$, luego si O_k es el giro tal que $\Omega_{k+1} = \Omega_k O_k$, $O_k \to I_n$ y, por tanto, $\lim_k(\Omega_{k+1} - \Omega_k) = \lim_k(O_k - I_n)\Omega_k = 0$, pues $(\Omega_k)_k$ está acotada. Con esto, y aplicando la propiedad del principio, $(\Omega_k)_k$ converge a una matriz, cuyas columnas formaran una base ortonormal de vectores propios en el mismo orden que los valores propios de D.

4.3. Método QR

Dada una matriz $A \in \mathcal{M}_n$, definimos la sucesión $(A_k)_k$ como $A_0 \coloneqq A$ y $A_{k+1} \coloneqq R_k Q_k$, donde (Q_k, R_k) es la descomposición QR de A_k. Bajo ciertas condiciones, esta sucesión tiende a una matriz triangular superior, con los valores propios en la diagonal.

Para obtener una aproximación de los valores propios a partir de una aproximación $A_p \coloneqq (u_{ij})$ de dicha matriz, definimos una matriz $V \coloneqq (v_{ij}) \in \mathcal{M}_n$ dada por

$$v_{ij} \coloneqq \begin{cases} 1, & i = j; \\ 0, & i > j; \\ -\dfrac{1}{u_{ii} - u_{jj}} \displaystyle\sum_{k=i+1}^{j} u_{ik} v_{kj}, & i < j; \end{cases}$$

y los vectores propios son las columnas de $Q_1 \cdots Q_p V$.

Capítulo 5

Sistemas de ecuaciones no lineales

5.1. Iteración de punto fijo

CN

Dado un espacio métrico (X, d), $f : X \to X$ es **contractiva** si existe $c \in (0, 1)$ tal que $\forall x, y \in X, d(f(x), f(y)) \leq cd(x, y)$, y decimos que c es una **constante de contractividad** de f. [...] **Teorema del punto fijo de Banach:** Si (X, d) es completo y $f : X \to X$ es contractiva con constante c, existe un único punto fijo $\xi \in X$ y la sucesión dada por x_0 arbitrario y $x_{n+1} = f(x)$ converge a ξ con

$$d(x_n, \xi) \leq \frac{c^n}{1-c} d(x_0, x_1).$$

FVV1

Definimos la norma de una aplicación $L : (E, \|\cdot\|) \to (F, \|\cdot\|')$ como

$$\|L\| := [...] \sup\{\|L(x)\|'\}_{x \in E, \|x\| \leq 1}$$

[...] El **teorema del incremento finito** afirma que, sean $f : \Omega \subseteq \mathbb{R}^m \to \mathbb{R}^n$, $a, b \in \Omega$ con el segmento $[a, b] \subseteq \Omega$ y $M > 0$, si $\|df(x)\| \leq M$ para todo $x \in [a, b]$ se tiene $\|f(b) - f(a)\| \leq M\|b - a\|$.

Sean $\Omega \subseteq \mathbb{R}^n$ abierto, $f : \Omega \to \mathbb{R}^n$ $\mathcal{C}^1$ y $x \in \Omega$ un punto fijo de f con $\|df(x)\| < 1$, existe $\delta > 0$ tal que $f(\overline{B}(x, \delta)) \subseteq \Omega$, y toda sucesión $(x_n)_n$ dada por $x_0 \in \overline{B}(x, \delta)$ y $x_{k+1} := f(x_k)$ converge.

Sean $R := [a_1, b_1] \times \cdots \times [a_n, b_n]$, $f : R \to R$ diferenciable y $K \in (0, 1)$ tal que

$$\left| \frac{\partial f_i}{\partial x_j}(x) \right| \leq \frac{K}{n}$$

para cada $x \in R$ e $i, j \in \{1, \ldots, n\}$, entonces $\|df\|_\infty \leq K$.

La **aceleración de Gauss-Seidel** de una iteración de punto fijo consiste en considerar, en vez de $x_{k+1} \coloneqq g(x_k)$,

$$\begin{aligned} x_{(k+1)1} &:= g_1(x_{k1}, \ldots, x_{kn}), \\ x_{(k+1)2} &:= g_2(x_{(k+1)1}, x_{k2}, \ldots, x_{kn}), \\ &\vdots \\ x_{(k+1)n} &:= g_n(x_{(k+1)1}, \ldots, x_{(k+1)(n-1)}, x_{kn}). \end{aligned}$$

5.2. Método de Newton

Consideremos una ecuación $f(x) = 0$ con $f : \mathbb{R}^n \to \mathbb{R}^n$ diferenciable. Sustituyendo f por su polinomio de Taylor de primer orden en x_0 tenemos $f(x_0) + df(x_0)(x - x_0) = 0$, lo que nos da una aproximación de x.

Como **teorema**, si $f : \Omega \subseteq \mathbb{R}^n \to \mathbb{R}^n$ es $\mathcal{C}^2$, $\xi \in \Omega$ es un cero de f y $df(\xi)$ es no singular, existe $r > 0$ tal que, para todo $x \in B(\xi, r)$, la sucesión dada por $x_0 \coloneqq x$ y $x_{k+1} \coloneqq x_k - df(x_k)^{-1} f(x_k)$ converge a ξ, y existe $M > 0$ tal que

$$\|x_{k+1} - \xi\| \leq M \|x_k - \xi\|^2.$$

Demostración para $n = 2$**:** Queremos ver que $g(x) \coloneqq x - df(x)^{-1} f(x)$ es contractiva cerca de ξ, para lo que basta ver que es $\mathcal{C}^1$ y que $dg(\xi) = 0$. Como f es $\mathcal{C}^2$, $\det \circ df$ es continua y diferenciable en Ω, y en particular, en un entorno de ξ, las diferenciales tienen inversa

$$df(x)^{-1} \equiv \frac{1}{\det df(x)} \begin{pmatrix} \frac{\partial f_2}{\partial x_2}(x) & -\frac{\partial f_1}{\partial x_2}(x) \\ \frac{\partial f_2(x)}{\partial x_1}(x) & \frac{\partial f_1(x)}{\partial x_1}(x) \end{pmatrix} =: \begin{pmatrix} a_{11}(x) & a_{12}(x) \\ a_{21}(x) & a_{22}(x) \end{pmatrix},$$

y $a_{11}, a_{12}, a_{21}, a_{22}$ son $\mathcal{C}^1$. Entonces

$$g(x, y) \coloneqq \begin{pmatrix} x - a_{11}(x,y) f_1(x,y) - a_{12}(x,y) f_2(x,y) \\ y - a_{21}(x,y) f_1(x,y) - a_{22}(x,y) f_2(x,y) \end{pmatrix}.$$

Las derivadas parciales de g_1 son

$$\begin{aligned} \frac{\partial g_1}{\partial x_1}(x) &= 1 - \frac{\partial a_{11}}{\partial x_1}(x) f_1(x) - a_{11}(x) \frac{\partial f_1}{\partial x_1}(x) - \frac{\partial a_{12}}{\partial x_1}(x) f_2(x) - a_{12}(x) \frac{\partial f_2}{\partial x_1}(x), \\ \frac{\partial g_1}{\partial x_2}(x) &= -\frac{\partial a_{11}}{\partial x_2}(x) f_1(x) - a_{11}(x) \frac{\partial f_1}{\partial x_2}(x) - \frac{\partial a_{12}}{\partial x_2}(x) f_2(x) - a_{12}(x) \frac{\partial f_2}{\partial x_2}(x). \end{aligned}$$

Así, como $f(\xi) = 0$,

$$\begin{aligned} \frac{\partial g_1}{\partial x_1}(\xi) &= 1 - a_{11}(\xi) \frac{\partial f_1}{\partial x_1}(\xi) - a_{12}(\xi) \frac{\partial f_2}{\partial x_1}(\xi) \\ &= 1 - \frac{1}{\det df(x)} \left(\frac{\partial f_2}{\partial x_2}(\xi) \frac{\partial f_1}{\partial x_1}(\xi) - \frac{\partial f_1}{\partial x_2}(\xi) \frac{\partial f_2}{\partial x_1}(\xi) \right) = 1 - 1 = 0; \\ \frac{\partial g_1}{\partial x_2}(\xi) &= -a_{11}(\xi) \frac{\partial f_1}{\partial x_2}(\xi) - a_{12}(\xi) \frac{\partial f_2}{\partial x_2}(\xi) \\ &= -\frac{1}{\det df(x)} \left(\frac{\partial f_2}{\partial x_2}(\xi) \frac{\partial f_1}{\partial x_2}(\xi) - \frac{\partial f_1}{\partial x_2}(\xi) \frac{\partial f_2}{\partial x_2}(\xi) \right) = 0. \end{aligned}$$

Para g_2 es análogo, y solo queda ver el orden de convergencia. Dados un abierto conexo $C \subseteq \mathbb{R}^n$ y $f : C \to \mathbb{R}^n$, si existe $K > 0$ tal que para $x, y \in C$, $\|df(x) - df(y)\| \leq K\|x - y\|$, entonces, para $x, y \in C$,

$$\|f(x) - f(y) - df(y)(x - y)\| \leq \frac{K}{2}\|x - y\|^2.$$

En efecto, sea $\varphi : [0, 1] \to C$ dada por $\varphi(t) \coloneqq f(y + t(x - y))$, por la regla de la cadena, $\varphi'(t) = df(y + t(x - y))(x - y)$, por lo que

$$\begin{aligned}\|\varphi'(t) - \varphi'(0)\| &= \|(df(y + t(x - y)) - df(y))(x - y)\| \leq \\ &\leq \|df(y + t(x - y)) - df(y)\|\|x - y\| \leq Kt\|x - y\|^2,\end{aligned}$$

y como además

$$\Delta \coloneqq f(x) - f(y) - df(y)(x - y) = \varphi(1) - \varphi(0) - \varphi'(0) = \int_0^1 (\varphi'(t) - \varphi'(0))dt,$$

queda

$$\|\Delta\| \leq \int_0^1 \|\varphi'(t) - \varphi'(0)\|dt \leq K\|x - y\|^2 \int_0^1 t\, dt = \frac{K}{2}\|x - y\|^2.$$

Cuando esto se cumple,

$$\begin{aligned}\|x_{k+1} - \xi\| &= \|(x_k - \xi) - df(x_k)^{-1}f(x_k)\| = \|df(x_k)^{-1}(df(x_k)(x_k - \xi) - f(x_k))\| \leq \\ &\leq \|df(x_k)^{-1}\|\|df(x_k)(x_k - \xi) - f(x_k)\| \overset{f(\xi)=0}{=} \\ &= \|df(x_k)^{-1}\|\|f(\xi) - f(x_k) - df(x_k)(\xi - x_k)\| \leq \|df(x_k)^{-1}\|\frac{K}{2}\|x_k - \xi\|^2,\end{aligned}$$

y tomando $M \coloneqq \frac{K}{2}\sup_{x \in B(\xi,r)} \|df(x)^{-1}\|$ se obtiene la acotación.

5.3. Método de Broyden

El método de Newton requiere calcular la matriz diferencial en cada iteración ($\Theta(n^2)$ si se usa derivación numérica y evaluar f en un punto es $\Theta(1)$) y resolver un sistema lineal ($\Theta(n^3)$). Para reducir el orden de complejidad, podemos aproximar la diferencial con una matriz A_k tal que $A_k(x_k - x_{k-1}) = f(x_k) - f(x_{k-1})$, con lo que solo necesitamos la diferencial para obtener x_1. A cambio, pasamos de orden de convergencia cuadrático a supralineal.

Podemos usar

$$A_k \coloneqq A_{k-1} + \frac{1}{\|x_k - x_{k-1}\|_2^2} f(x_k)(x_k - x_{k-1})^t,$$

tomando $A_0 \coloneqq df(x_0)$. En efecto, como $x_k = x_{k-1} - A_{k-1}^{-1}f(x_{k-1})$, tenemos $A_{k-1}(x_k - x_{k-1}) = f(x_{k-1})$, y por tanto

$$\begin{aligned}&\left(A_{k-1} + \frac{1}{\|x_k - x_{k-1}\|_2^2} f(x_k)(x_k - x_{k-1})^t\right)(x_k - x_{k-1}) = \\ &\qquad = A_{k-1}(x_k - x_{k-1}) + f(x_k) = f(x_k) - f(x_{k-1}).\end{aligned}$$

Para calcular la inversa de A_k podemos usar la **fórmula de Sherman-Morrison**,

$$(A + vy^t)^{-1} = A^{-1} - \frac{1}{1 + y^t A^{-1} v}(A^{-1}v)(y^t A^{-1}),$$

suponiendo que $y^t A^{-1} x \neq -1$. Esto reduce el orden de complejidad de cada iteración a $\Theta(n^2)$ en el caso general, asumiendo que evaluar f en un punto es $O(n^2)$.

5.4. Método del descenso rápido

Entrada: Función $f : \mathbb{R}^n \to \mathbb{R}^n$, aproximación inicial $x \in \mathbb{R}^n$, margen de error $e > 0$ y tolerancia $t > 0$.
Salida: x tal que $f(x) \approx 0$, o error indicando que no se puede mejorar la aproximación.
Definir $g(x) := \sum_{k=1}^{n} f_k(x)^2$;
mientras $g(x) > e$ **hacer**
  $u \leftarrow \frac{\nabla g(x)}{\|\nabla g(x)\|}$;
  $\alpha \leftarrow 1$;
  mientras $g(x - \alpha u) > g(x)$ **hacer** $\alpha \leftarrow \frac{1}{2}\alpha$;
  si $|\alpha| < t$ **entonces devolver** error;
  // Interpolar en $(0, g(x)), (\frac{\alpha}{2}, g(x + \frac{\alpha}{2}u)), (\alpha, g(x + \alpha u))$.
  $g_0 \leftarrow g(x)$;
  $g_1 \leftarrow g(x + \frac{\alpha}{2}u)$;
  $g_2 \leftarrow g(x + \alpha u)$;
  $h_0 \leftarrow \frac{2(g_1 - g_0)}{\alpha}$; // Usamos diferencias divididas de Newton.
  $h_1 \leftarrow \frac{2(g_2 - g_1)}{\alpha}$;
  $h_{01} \leftarrow \frac{h_2 - h_1}{\alpha}$;
  // Obtener y usar el vértice de la parábola resultante.
  $\alpha_0 \leftarrow \frac{1}{2}(\frac{\alpha}{2} - \frac{h_0}{h_{01}})$;
  si $g(x + \alpha_0 u) < g_2$ **entonces** $x \leftarrow x + \alpha_0 u$;
  sinó $x \leftarrow x + \alpha u$;
fin

Algoritmo 5: Método del descenso rápido.

El método del descenso rápido o *steepest descent* es el algoritmo 5, y consiste en minimizar la función $g(x) \coloneqq \|f(x)\|_2^2$ desplazándonos, en cada iteración, en la dirección de mayor descenso en esta función. No es un método muy rápido, pero permite una estimación inicial más lejana que los métodos de Newton y Broyden, por lo que se suele usar para obtener una aproximación no muy fina que a su vez se usa como punto de partida para estos métodos.

Apéndice A

Octave

Los ficheros de código de Octave tienen extensión `.m` y pueden contener un *script* o una función. Los *scripts* contienen una secuencia de órdenes y se invocan con *`nombre`* (el nombre del fichero sin la extensión). Las funciones tienen forma:

```
function (val|[val,...]) = nombre(par, ...)
    stmt
    ...
endfunction
```

Donde *`nombre`* debe coincidir con el nombre del fichero, los parámetros (*`par`*) se pueden usar como variables, y los valores de retorno (*`val`*) se usan como variables inicialmente no definidas, de modo que su valor al final de la función es el valor devuelto. La función se invoca con la expresión *`nombre`*`(`*`arg`*`, ...)` si devuelve un solo valor o con la sentencia `[`*`var`*`, ...] =` *`nombre`*`(`*`arg`*`, ...)` si devuelve varios. Por defecto, Octave busca los *scripts* y funciones en el directorio actual (que se puede cambiar en la interfaz gráfica) y en la biblioteca estándar.

Las sentencias terminan con salto de línea o con `;` (normalmente seguido de salto de línea), pero si termina en salto de línea, el valor devuelto se imprime, bien como `ans =` *`valor`* si la sentencia es una expresión *`expr`* que devuelve un *`valor`*, o como *`A`* `=` *`valor`* si la sentencia es una asignación *`A`* `=` *`expr`*. Las variables se definen en su primera asignación.

Los comentarios empiezan por `%` o `#` y terminan al final de la línea.

Las variables de las funciones tienen ámbito local y las de los *scripts* tienen como ámbito el intérprete de comandos. Existe un ámbito global, y para indicar que una variable pertenece a este, se añade el comando `global` *`variable`* en cada ámbito en que se vaya a usar (en el intérprete, normalmente dentro de un *script*, y en las funciones).

A.1. Tipos de datos

A.1.1. Matrices

En Octave, los números (con sintaxis `[-+]?((\d+\.?|\d*\.\d+)([eE][-+]?\d+)?|[Ii]nf)` o ({número}\+)?{número}?i) representan matrices 1×1 de números de doble precisión, y las cadenas de caracteres (con sintaxis `'([^']|'')*'` o `"([^\\']|\\{escape})*"`) representan matrices fila de caracteres. En estas, `\n` indica un salto de línea y `\t` un tabulador.

La expresión `[`a_1`, ..., `a_p`]` concatena horizontalmente las matrices $a_1 \in \mathcal{M}_{m\times n_1}(S)$ hasta $a_p \in \mathcal{M}_{m\times n_p}(S)$ en una matriz en $\mathcal{M}_{m\times\sum_{k=1}^{p} n_k}(S)$, y la sintaxis `[`$a_{11}$`, ..., `$a_{1p_1}$`; ...; `$a_{q1}$`, ..., `$a_{qp_q}$`]` hace esto en cada parte, resultando en q matrices $b_k \in \mathcal{M}_{m_k\times n}(S)$, y las concatena verticalmente en una $\mathcal{M}_{\sum_{k=1}^{q} m_k\times n}(S)$.

Si una operación $*$ actúa sobre matrices, $.*$ actúa sobre cada elemento. Si se aplica a dos matrices de igual tamaño, se aplica a cada elemento de una con el correspondiente de la otra para obtener un elemento 1×1 que se sitúa en la posición correspondiente de la matriz de salida. Si una de las dos es de un elemento, se extiende a una de igual tamaño que la otra con todos los elementos iguales al original.

`A(x, y)` es la submatriz de *`A`* formada por las columnas con índice en el vector *`x`* y las filas con índice en el vector *`y`*, y *`A(x)`* convierte la matriz en un vector concatenando las traspuestas de sus columnas y toma los elementos del vector con índice en *`x`*. Los vectores de índice se pueden sustituir por `:` para tomar todas las filas o columnas. Los índices empiezan por 1.

`A(x, y) = expr` o *`A(x) = expr`* asigna los elementos de la submatriz a la izquierda del `=` a los de la devuelta por la expresión, que debe ser del mismo tamaño. Si la variable no existe, se crea, y si la submatriz indicada supone que *`A`* es más grande de lo que es, esta se amplía y se rellena con ceros.

A.1.2. Números

El operador `+` suma matrices numéricas de igual tamaño, `-` las resta y `*` multiplica matrices o una matriz por un escalar. Si *`a`* y *`b`* son escalares con $b \neq 0$, *`a/b`* es su cociente.

Llamamos vector a una matriz fila. Entonces *`a:b`* genera el vector $(a, a+1, \ldots, b)$ y *`a:t:b`* genera el vector $(a, a+t, \ldots, b)$. Cuando es posible, `A\B` devuelve una matriz X tal que $AX = B$. *`A'`* es A^*.

A.1.3. Booleanos

Los booleanos son números: 0 para falso, cualquier otro número (normalmente 1) para verdadero. Entonces `|` o `||` es el operador disyunción; `&` o `&&` es la conjunción. Los operadores de comparación `==`, `!=` (o `~=`), `<`, `<=`, `>` y `>=` hacen lo que se esperaría.

A.1.4. Listas

Las listas pueden contener elementos de distintos tipos. No se asigna a una lista, sino a sus elementos, a los que se hace referencia por *`lista{indice}`*. Los índices empiezan por 1. Si se asigna a un índice mayor que el tamaño de la lista, esta se amplía y se rellena con matrices numéricas 0×0.

A.1.5. Funciones anónimas

Una función anónima se expresa como `@(`*`par`*`, ...)` *`expr`*, puede asignarse y puede usarse como una función normal. Solo puede devolver un elemento.

A.2. Control de flujo

La sentencia

```
for var = vector
    stmt ...
endfor
```

evalúa los *stmt* una vez por cada elemento del *vector* con *var* tomando el valor del elemento. El *vector* también puede ser una lista, pero entonces devuelve listas de un elemento a las que hay que acceder con el índice 1.

```
while condición
    stmt ...
endwhile
```

evalúa los *stmt* repetidamente mientras la *condición* se cumpla, comprobándola al principio de cada iteración.

```
do
    stmt ...
until condición
```

evalúa los *stmt* repetidamente mientras la *condición* no se cumpla, comprobándola al principio de cada iteración salvo la primera.

```
if condición
    then-stmts
[else
    else-stmts]
endif
```

evalúa *condición* y, si se cumple, ejecuta *then-stmts*, y de lo contrario, si está, ejecuta *else-stmts*.

La sentencia `break` sale del bucle (`for` o `while`) más interno en que se encuentra, `continue` pasa a su siguiente iteración y `return` sale de la función.

```
try
    stmts
catch
    handle-stmts
end_try_catch
```

ejecuta los *stmts* y, si hay una excepción, la captura y ejecuta los *handle-stmts*.

A.3. Biblioteca estándar

La mayoría de funciones que reciben un número y devuelven otro también funcionan elemento a elemento con matrices.

`abs(`*z*`)` $|z|$.

`addpath(`*dir*`,...)` Añade los directorios indicados a la lista de rutas donde buscar *scripts* y funciones.

`and(`*x1*`,`*x2*`,...)` *x1* `&` *x2* `&` ...

`clc` Limpia la pantalla de la *shell*.

`clear` Borra todas las variables.

`clf` Vacía la ventana de gráficas, creándola si no existía ya.

`cond(`*A*`,`*p*`)` `norm(`*A*`,` *p*`)` `*` `norm(inv(`*A*`),` *p*`)`.

`cond(`*A*`)` `cond(`*A*`,2)`.

`contour(`*x*`,`*y*`,`*v*`,`*h*`,`*style*`)` Si a cada par formado por un elemento de *x* y el correspondiente de *y* se le asigna el elemento correspondiente de *v*, dibuja la curva de nivel (aproximada) de la gráfica a la altura *h*. El parámetro *style* es una cadena donde cada caracter indica una propiedad: `r` rojo, `b` azul, `m` magenta, etc.

`contour(`*x*`,`*y*`,`*z*`,`*h*`)` Como la anterior, con el estilo por defecto.

`diag(`*A*`,`*k*`)` Si *A* es vector, devuelve una matriz diagonal con elementos del vector en la diagonal $\{(i,j) \mid i+k=j\}$, y de lo contrario devuelve un vector con los elementos de dicha diagonal de *A*.

`diag(`*A*`)` `diag(`*A*`,0)`.

`disp(`*x*`)` Imprime el valor de *x*.

`dot(`*x*`,`*y*`)` Producto escalar hermitiano $\langle y, x\rangle$.

`[`*V*`,`*lambda*`]=eig(`*A*`)` Devuelve una matriz diagonal *lambda* en la que los elementos de las diagonal son los valores propios de *A* y una matriz *V* cuyas columnas son los vectores propios correspondientes.

`eps` Menor ϵ tal que $1+\epsilon > 1$ con la precisión de la máquina.

`error(`*text*`)` Lanza una excepción con un cierto texto.

`exp(`*z*`)` e^z.

`eye(`*n*`)` Matriz identidad de tamaño *n*.

`feval(`*función*`,...)` *función*`(...)`.

`figure(`*n*`)` Crea una ventana para gráficas identificada por *n* si esta no existe, y la usa para los dibujos posteriores.

`fminunc(`*f*`,`*x0*`)` Busca un mínimo local de *f* partiendo de una aproximación inicial *x0*.

`format` *opción* Cambia el formato en que se imprimen los números por defecto. Posibles valores son `short` o `long` para la cantidad de decimales (pocos o muchos), seguido opcionalmente por `e` para notación científica o `g` para elegir en cada caso si es mejor usar notación científica o normal.

`fprintf(`*fmt*`,...)` `printf(`*fmt*`,...)`.

`inv(`*A*`)` Inversa de la matriz cuadrada no singular *A*.

`isosurface(`*x*`,`*y*`,`*z*`,`*v*`,`*val*`)` Como `contour` para funciones de 3 variables (x, y, z).

`length(A)` `max(size(A))`.

`linspace(start,end,n)` Vector de n puntos equiespaciados de $start$ a end.

`[L,U,P,Q]=lu(A)` Descomposición de Gauss de A con elección de pivote total.

`[L,U,P]=lu(A)` Descomposición de Gauss de A con elección de pivote parcial.

`max(x,y)` máx$\{x, y\}$.

`[w,iw]=max(v)` Obtiene $w = \text{máx}\, v_i$ e iw con $w = v_{iw}$.

`max(A)` máx A_{ij}.

`[xx,yy]=meshgrid(x,y)` Si x es un vector de m elementos e y es un vector de y elementos, xx e yy son matrices $m \times n$ donde las filas de xx son copias de x y las columnas de yy son copias de y.

`min(x,y)` mín$\{x, y\}$.

`[w,iw]=min(v)` Obtiene $w = \text{mín}\, v_i$ e iw con $w = v_{iw}$.

`min(A)` mín A_{ij}.

`norm(A,p,opt)` Si `opt` es `"columns"` o `"cols"`, vector fila con la norma p de cada vector columna en A. Si es `"rows"`, vector columna con la norma p de cada vector fila en A.

`norm(A,p)` Norma p de A, matricial o vectorial según corresponda, donde p es un entero positivo o `Inf`.

`norm(A)` `norm(A,2)`.

`ones(m,n)` Matriz $m \times n$ formada por unos.

`or(x1,x2,...)` $x1$ `|` $x2$ `|` ...

`printf(fmt,...)` Imprime los argumentos con un cierto formato fmt, una cadena con caracteres normales que se representan a sí mismos y secuencias de escape que empiezan por `%`. Por ejemplo, `%f` imprime el siguiente argumento como número de punto flotante con varios decimales, `%e` también pero con notación científica, `%d` también pero con pocos decimales, `%s` como cadena y `%%` representa un caracter `%`.

`rand(m,n)` Matriz de m filas y n columnas con elementos aleatorios entre 0 y 1.

`rmpath(dir,...)` Elimina los directorios indicados de la lista de directorios donde buscar *scripts* y funciones.

`size(A)` Devuelve un vector con el número de elementos en cada dimensión de A. Las matrices tienen dos dimensiones, y las listas también con primera dimensión de tamaño 1.

`sign(z)` Si $z = 0$, 0, de lo contrario $\frac{z}{|z|}$.

`sort(v)` Copia del vector v con sus elementos en orden creciente.

`sortrows(A)` Copia de la matriz A con sus filas en orden de menor a mayor primer elemento, o segundo en caso de empate, etc.

`sqrt(z)` $\sqrt{z}$.

`sum(A)` $\sum A_{ij}$.

`[U,S,V]=svd(A)` Devuelve dos matriz ortogonales U y V y una diagonal S tales que $A = USV^*$.

`title(str)` Da un título a la gráfica.

`trace(A)` Traza de A.

`transpose(A)` A^t.

`tril(A,k)` Matriz como A pero con los elementos (i,j) con $j-i > k$ a 0.

`tril(A)` `tril(A,0)`, matriz triangular inferior.

`triu(A,k)` Matriz como A pero con los elementos (i,j) con $i-j > k$ a 0.

`triu(A)` `triu(A,0)`, matriz triangular superior.

`vander(c,n)` $(c_i^{n-j})_{ij} \in \mathcal{M}_{\texttt{size(c)}\times n}$.

`warning(text)` Muestra un texto a modo de alerta.

`xlabel(str)` Indica la leyenda del eje X de la gráfica.

`ylabel(str)` Indica la leyenda del eje Y de la gráfica.

`zeros(m,n)` Matriz nula de m filas y n columnas.

`zlabel(str)` Indica la leyenda del eje Z de la gráfica.