1 Vector Norms

Definition 1.1. *A vector norm on $\mathbb{C}^n$ is a linear map $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$, such that*

- $\|\mathbf{x}\| \geq 0$, $\forall \mathbf{x} \in \mathbb{C}^n$; *and* $\|\mathbf{x}\| = 0 \iff \mathbf{x} = \mathbf{0}_n$ *(positive definiteness).*
- $\|\alpha\,\mathbf{x}\| = |\alpha|\|\mathbf{x}\|$, $\forall \alpha \in \mathbb{C}$, $\forall \mathbf{x} \in \mathbb{C}^n$ *(absolute homogeneity).*
- $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ *(subadditivity or triangle inequality).*

Definition 1.2. *For a real number $p \geq 1$, the p-norm or L^p-norm of $\mathbf{x} \in \mathbb{C}^n$ is defined as*

$$\|\mathbf{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

The L^2-norm is known as the Euclidean norm.

Definition 1.3. *The L^∞-norm or uniform norm of $\mathbf{x} \in \mathbb{C}^n$ is defined as*

$$\|\mathbf{x}\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Theorem 1.4. *All the norms on $\mathbb{C}^n$ are equivalent, i.e., given two norms $\|\cdot\|_p$ and $\|\cdot\|_q$, there exist real constants $0 < c_1 \leq c_2 < \infty$, such that*

$$c_1\|\mathbf{x}\|_p \leq \|\mathbf{x}\|_q \leq c_2\|\mathbf{x}\|_p, \quad \forall \mathbf{x} \in \mathbb{C}^n.$$

Definition 1.5. *An inner product on $\mathbb{C}^n$ is a map $\langle \cdot, \cdot \rangle : \mathbb{C}^n \times \mathbb{C}^n \to \mathbb{C}$, such that*

- $\langle \mathbf{x}, \mathbf{x} \rangle \geq 0$, $\forall \mathbf{x} \in \mathbb{C}^n$; *and* $\langle \mathbf{x}, \mathbf{x} \rangle = 0 \iff \mathbf{x} = \mathbf{0}_n$.
- $\langle \mathbf{y}, \mathbf{x} \rangle = \overline{\langle \mathbf{x}, \mathbf{y} \rangle}$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{C}^n$.
- $\langle \alpha\,\mathbf{x}, \mathbf{y} \rangle = \alpha \langle \mathbf{x}, \mathbf{y} \rangle$, $\forall \alpha \in \mathbb{C}$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{C}^n$.
- $\langle \mathbf{x} + \mathbf{y}, \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{z} \rangle + \langle \mathbf{y}, \mathbf{z} \rangle$, $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{C}^n$.

Remark: from the second and third properties, it follows that $\langle \mathbf{x}, \alpha\,\mathbf{y} \rangle = \bar{\alpha} \langle \mathbf{x}, \mathbf{y} \rangle$, $\forall \alpha \in \mathbb{C}$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{C}^n$; and, from the second and fourth properties, $\langle \mathbf{x}, \mathbf{y} + \mathbf{z} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle + \langle \mathbf{x}, \mathbf{z} \rangle$. Therefore, an inner product on $\mathbb{C}^n$ is linear in the first component and antilinear or conjugate-linear in the second component; i.e., it is a sesquilinear map. There are some authors that define inner products on $\mathbb{C}^n$ to be linear in the second component and antilinear in the first one.

Definition 1.6. *The standard inner product on $\mathbb{C}^n$ is given by*

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^* \mathbf{x} = \sum_{i=1}^{n} x_i \overline{y_i}, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{C}^n,$$

where * denotes the conjugate transpose.

Definition 1.7. *Two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$ are orthogonal, if and only if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.*

Definition 1.8. *Given an inner product $\langle \cdot, \cdot \rangle$ on $\mathbb{C}^n$, the associated norm $\|\cdot\|$ is defined as*

$$\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}, \quad \forall \mathbf{x} \in \mathbb{C}^n.$$

Remark: The norm associated to the standard inner product on $\mathbb{C}^n$ is the L^2-norm or Euclidean norm:

$$\|\mathbf{x}\|_2 = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle} = \sqrt{\sum_{i=1}^{n} |x_i|^2}, \quad \forall \mathbf{x} \in \mathbb{C}^n.$$

Theorem 1.9 (Cauchy-Schwarz inequality)**.**

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|\|\mathbf{y}\|, \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{C}^n.$$

Moreover, the equality happens if and only if $\mathbf{x}$ and $\mathbf{y}$ are linearly independent.

Proof. Let $\mathbf{x}, \mathbf{y} \in \mathbb{C}^n$. If $\mathbf{y} = \mathbf{0}_n$, then $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent, and the theorem is trivially true. Assume $\mathbf{y} \neq \mathbf{0}_n$ and define

$$\lambda = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2} \in \mathbb{C};$$

then

$$\begin{aligned}
0 \leq \|\mathbf{x} - \lambda\,\mathbf{y}\|^2 &= \langle \mathbf{x} - \lambda\,\mathbf{y}, \mathbf{x} - \lambda\,\mathbf{y} \rangle \\
&= \langle \mathbf{x}, \mathbf{x} \rangle - \lambda \langle \mathbf{y}, \mathbf{x} \rangle - \bar{\lambda} \langle \mathbf{x}, \mathbf{y} \rangle + \lambda \bar{\lambda} \langle \mathbf{y}, \mathbf{y} \rangle \\
&= \|\mathbf{x}\|^2 - \lambda \overline{\langle \mathbf{x}, \mathbf{y} \rangle} - \bar{\lambda} \langle \mathbf{x}, \mathbf{y} \rangle + |\lambda|^2 \|\mathbf{y}\|^2 \\
&= \|\mathbf{x}\|^2 - \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\|\mathbf{y}\|^2} \overline{\langle \mathbf{x}, \mathbf{y} \rangle} - \frac{\overline{\langle \mathbf{x}, \mathbf{y} \rangle}}{\|\mathbf{y}\|^2} \langle \mathbf{x}, \mathbf{y} \rangle + \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^4} \|\mathbf{y}\|^2 \\
&= \|\mathbf{x}\|^2 - \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|^2}{\|\mathbf{y}\|^2}.
\end{aligned}$$

Therefore,

$$|\langle \mathbf{x}, \mathbf{y} \rangle|^2 \leq \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \iff |\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\| \|\mathbf{y}\|.$$

On the other hand, $|\langle \mathbf{x}, \mathbf{y} \rangle| = \|\mathbf{x}\| \|\mathbf{y}\| \iff \|\mathbf{x} - \lambda\,\mathbf{y}\|^2 = 0 \iff \mathbf{x} = \lambda\,\mathbf{y}$, i.e., the equality happens if and only if $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.

□

2 Matrix decompositions

Definition 2.1. *Let $\mathbb{C}^{m \times n}$ denote the vector space of all matrices of size $m \times n$ with entries in $\mathbb{C}$. A complex square matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ is normal if and only if*

$$\mathbf{A}^* \mathbf{A} = \mathbf{A}\,\mathbf{A}^*,$$

where * denotes the conjugate transpose. Therefore, a matrix is normal if and only if it commutes with its conjugate transpose.

Theorem 2.2. *An upper triangular matrix $\mathbf{U} \in \mathbb{C}^{n \times n}$ is normal if and only if it is a diagonal matrix.*

Proof. Let us prove it by induction on the matrix order. If $n = 1$, the result is trivial. Hence, assuming that the theorem holds for orders $\leq n - 1$, we have to prove it for n. Given an upper triangular matrix $\mathbf{U} \in \mathbb{C}^{n \times n}$, we can represent it as

$$\mathbf{U} = \begin{pmatrix} \tilde{\mathbf{U}} & \mathbf{u} \\ \mathbf{0}_{n-1}^T & a \end{pmatrix},$$

where $\tilde{\mathbf{U}} \in \mathbb{C}^{(n-1) \times (n-1)}$ is an upper triangular matrix, and $\mathbf{u} \in \mathbb{C}^{n-1}$ is a vector. Then, since $\mathbf{U}$ is normal,

$$\begin{aligned}
\mathbf{0}_{n \times n} = \mathbf{U}^* \mathbf{U} - \mathbf{U}\,\mathbf{U}^* &= \begin{pmatrix} \tilde{\mathbf{U}}^* & \mathbf{0}_{n-1} \\ \mathbf{u}^* & \bar{a} \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{U}} & \mathbf{u} \\ \mathbf{0}_{n-1}^T & a \end{pmatrix} - \begin{pmatrix} \tilde{\mathbf{U}} & \mathbf{u} \\ \mathbf{0}_{n-1}^T & a \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{U}}^* & \mathbf{0}_{n-1} \\ \mathbf{u}^* & \bar{a} \end{pmatrix} \\
&= \begin{pmatrix} \tilde{\mathbf{U}}^* \tilde{\mathbf{U}} - \tilde{\mathbf{U}}\,\tilde{\mathbf{U}}^* - \mathbf{u}\mathbf{u}* & \tilde{\mathbf{U}}^* \mathbf{u} - \bar{a}\mathbf{u} \\ \mathbf{u}^* \tilde{\mathbf{U}} - a\mathbf{u}^* & \mathbf{u}^* \mathbf{u} \end{pmatrix}.
\end{aligned}$$

On the one hand, $\mathbf{u}^* \mathbf{u} = 0$, so $\mathbf{u} = \mathbf{0}_n$. On the other hand, $\tilde{\mathbf{U}}^* \tilde{\mathbf{U}} - \tilde{\mathbf{U}}\,\tilde{\mathbf{U}}^* - \mathbf{u}\mathbf{u}^* = \tilde{\mathbf{U}}^* \tilde{\mathbf{U}} - \tilde{\mathbf{U}}\,\tilde{\mathbf{U}}^* = \mathbf{0}_{(n-1) \times (n-1)}$, i.e., $\tilde{\mathbf{U}}$ is normal, so, by the induction hypothesis, $\tilde{\mathbf{U}}$ is diagonal. Therefore, we conclude that $\mathbf{U}$ is diagonal.

Observe that the reciprocal is trivial, because diagonal matrices are trivially normal.

□

Remark: the theorem is also true for lower triangular matrices $\mathbf{L} \in \mathbb{C}^{n \times n}$, and the proof is identical.

Definition 2.3. *A complex square matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ is unitary if and only if its conjugate transpose $\mathbf{A}^*$ is also its inverse $\mathbf{A}^{-1}$, i.e., if and only if*

$$\mathbf{A}^* \mathbf{A} = \mathbf{A}\,\mathbf{A}^* = \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the identity matrix of order n.

Remarks: unitary matrices are trivially normal. The real analogue of a unitary matrix is an orthogonal matrix; in that case, ^{T} is used instead of *.

Definition 2.4. *A Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is a complex square matrix that is equal to its own conjugate transpose, i.e., $\mathbf{A} = \mathbf{A}^* \iff a_{ij} = \overline{a_{ji}}$, $\forall i, j \in \{1, \dots, n\}$. This is equivalent to $\mathbf{A}$ being self-adjoint, i.e., $\langle \mathbf{A}\,\mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{x}, \mathbf{A}\,\mathbf{y}\rangle$, $\forall \mathbf{x}, \mathbf{y} \in \mathbb{C}^n$.*

Remarks: Hermitian matrices are trivially normal. The real analogue of a Hermitian matrix is a symmetric matrix. Therefore, all the properties of Hermitian matrices apply immediately to symmetric matrices.

Theorem 2.5. *Hermitian matrices have real eigenvalues.*

Proof. Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a Hermitian matrix, λ an eigenvalue of $\mathbf{A}$, and $\mathbf{u} \neq \mathbf{0}$ an eigenvector of $\mathbf{A}$ associated to λ. Then,

$$\begin{aligned}\mathbf{A}\,\mathbf{u} = \lambda\,\mathbf{u} &\implies (\mathbf{A}\,\mathbf{u})^* = (\lambda\,\mathbf{u})^* \implies \mathbf{u}^*\mathbf{A}^* = \bar{\lambda}\,\mathbf{u}^* \implies \mathbf{u}^*\mathbf{A}^*\mathbf{u} = \bar{\lambda}\,\mathbf{u}^*\mathbf{u}.\\ &\implies \mathbf{u}^*\mathbf{A}\,\mathbf{u} = \bar{\lambda}\,\mathbf{u}^*\mathbf{u} \implies \mathbf{u}^*(\lambda\,\mathbf{u}) = \bar{\lambda}\,\mathbf{u}^*\mathbf{u} \implies \lambda\,\mathbf{u}^*\mathbf{u} = \bar{\lambda}\,\mathbf{u}^*\mathbf{u}\\ &\implies \lambda = \bar{\lambda} \implies \lambda \in \mathbb{R},\end{aligned}$$

where we have used that $\mathbf{A}^* = \mathbf{A}$, and that $\mathbf{u}^*\mathbf{u} > 0$.

□

Theorem 2.6. *Eigenvalues of Hermitian matrices associated to different eigenvalues are orthogonal.*

Proof. Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a Hermitian matrix, and $\mathbf{u} \neq \mathbf{0}$ and $\mathbf{v} \neq \mathbf{0}$ eigenvectors of $\mathbf{A}$ associated respectively to the eigenvalues λ and μ of $\mathbf{A}$, with $\lambda \neq \mu$. Then,

$$\begin{aligned}\mathbf{A}\,\mathbf{u} = \lambda\,\mathbf{u} &\implies (\mathbf{A}\,\mathbf{u})^* = (\lambda\,\mathbf{u})^* \implies \mathbf{u}^*\mathbf{A}^* = \lambda\,\mathbf{u}^* \implies \mathbf{u}^*\mathbf{A}^*\,\mathbf{v} = \lambda\,\mathbf{u}^*\mathbf{v}\\ &\implies \mathbf{u}^*\mathbf{A}\,\mathbf{v} = \lambda\,\mathbf{u}^*\mathbf{v} \implies \mathbf{u}^*(\mu\,\mathbf{v}) = \lambda\,\mathbf{u}^*\mathbf{v} \implies (\lambda - \mu)\mathbf{u}^*\mathbf{v} = 0\\ &\implies \mathbf{u}^*\mathbf{v} = 0,\end{aligned}$$

where we have used that $\mathbf{A}^* = \mathbf{A}$, and that λ and μ are real (see Theorem 2.5) and different.

□

Definition 2.7. *A skew-Hermitian or anti-Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is a complex square matrix whose conjugate transpose is equal to itself, but with all the entries being of opposite sign, i.e., $\mathbf{A}^* = -\mathbf{A} \iff a_{ij} = -\overline{a_{ji}}$, $\forall i, j \in \{1, \dots, n\}$.*

The real analogue of a skew-Hermitian matrix is a skew-symmetric or antisymmetric matrix.

Remark: skew-Hermitian matrices are trivially normal.

Theorem 2.8. *The eigenvalues of skew-Hermitian matrices are zero or purely imaginary.*

Proof. Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a skew-Hermitian matrix, λ an eigenvalue of $\mathbf{A}$, and $\mathbf{u} \neq \mathbf{0}$ an eigenvector of $\mathbf{A}$ associated to λ. Then,

$$\begin{aligned}\mathbf{A}\,\mathbf{u} = \lambda\,\mathbf{u} &\implies (\mathbf{A}\,\mathbf{u})^* = (\lambda\,\mathbf{u})^* \implies \mathbf{u}^*\mathbf{A}^* = \bar{\lambda}\,\mathbf{u}^* \implies \mathbf{u}^*\mathbf{A}^*\mathbf{u} = \bar{\lambda}\,\mathbf{u}^*\mathbf{u}.\\ &\implies -\mathbf{u}^*\mathbf{A}\,\mathbf{u} = \bar{\lambda}\,\mathbf{u}^*\mathbf{u} \implies -\mathbf{u}^*(\lambda\,\mathbf{u}) = \bar{\lambda}\,\mathbf{u}^*\mathbf{u} \implies -\lambda\,\mathbf{u}^*\mathbf{u} = \bar{\lambda}\,\mathbf{u}^*\mathbf{u}\\ &\implies -\lambda = \bar{\lambda} \implies \lambda + \bar{\lambda} = 0 \implies \Re(\lambda) = 0,\end{aligned}$$

where we have used that $\mathbf{A}^* = -\mathbf{A}$, and that $\mathbf{u}^*\mathbf{u} > 0$.

□

Theorem 2.9 (Schur decomposition)**.** *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a complex square matrix; then, there is a factorization of $\mathbf{A}$, called the Schur decomposition of $\mathbf{A}$, of the form*

$$\mathbf{A} = \mathbf{Q}\,\mathbf{U}\,\mathbf{Q}^*,$$

where $\mathbf{Q} \in \mathbb{C}^{n\times n}$ is a unitary matrix, and $\mathbf{U} \in \mathbb{C}^{n\times n}$ is an upper triangular matrix.

Proof. Let us prove it by induction on the matrix order. If $n = 1$, take $\mathbf{Q} = (1)$, $\mathbf{U} = (a_{11})$, and the result is trivially true. Hence, assuming that the theorem holds for orders $\leq n-1$, we have to prove it for n. Let $\mathbf{A}$ be a complex matrix of order n. Compute an eigenvalue λ_1 of $\mathbf{A}$ and an associated eigenvector $\mathbf{q}_1$ of unit length, i.e., such that $\mathbf{q}_1^* \mathbf{q}_1 = 1$. The central point is that the orthonormal set $\{\mathbf{q}_1\}$ can be expanded to an orthonormal basis $\mathbf{q}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n$ of $\mathbb{C}^n$. Let $\mathbf{Q}_1$ be the matrix whose columns are precisely $\mathbf{q}_1, \ldots, \mathbf{q}_n$. $\mathbf{Q}_1$ is obviously unitary, and

$$\begin{aligned}\mathbf{Q}_1^* \mathbf{A}\, \mathbf{Q}_1 &= \begin{pmatrix} \mathbf{q}_1^* \\ \vdots \\ \mathbf{q}_n^* \end{pmatrix} \mathbf{A} \begin{pmatrix} \mathbf{q}_1 & \ldots & \mathbf{q}_n \end{pmatrix} = \begin{pmatrix} \mathbf{q}_1^* \\ \vdots \\ \mathbf{q}_n^* \end{pmatrix} \begin{pmatrix} \lambda \mathbf{q}_1 & \mathbf{A}\,\mathbf{q}_2 & \ldots & \mathbf{A}\,\mathbf{q}_n \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & \mathbf{q}_1^* \mathbf{A}\,\mathbf{q}_2 & \ldots & \mathbf{q}_1^* \mathbf{A}\,\mathbf{q}_n \\ 0 & \mathbf{q}_2^* \mathbf{A}\,\mathbf{q}_2 & \ldots & \mathbf{q}_2^* \mathbf{A}\,\mathbf{q}_n \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \mathbf{q}_n^* \mathbf{A}\,\mathbf{q}_2 & \ldots & \mathbf{q}_n^* \mathbf{A}\,\mathbf{q}_n \end{pmatrix} = \begin{pmatrix} \lambda_1 & \mathbf{b}^T \\ \mathbf{0}_{n-1} & \tilde{\mathbf{A}} \end{pmatrix},\end{aligned}$$

where

$$\tilde{\mathbf{A}} = \begin{pmatrix} \mathbf{q}_2^* \mathbf{A}\,\mathbf{q}_2 & \ldots & \mathbf{q}_2^* \mathbf{A}\,\mathbf{q}_n \\ \vdots & \ddots & \vdots \\ \mathbf{q}_n^* \mathbf{A}\,\mathbf{q}_2 & \ldots & \mathbf{q}_n^* \mathbf{A}\,\mathbf{q}_n \end{pmatrix} \in \mathbb{C}^{(n-1)\times(n-1)}, \qquad \mathbf{b} = \begin{pmatrix} \mathbf{q}_1^* \mathbf{A}\,\mathbf{q}_2 \\ \vdots \\ \mathbf{q}_1^* \mathbf{A}\,\mathbf{q}_n \end{pmatrix} \in \mathbb{C}^{n-1}.$$

The order of $\tilde{\mathbf{A}}$ is $n-1$, so, by the induction hypothesis, there is a unitary matrix $\tilde{\mathbf{Q}} \in \mathbb{C}^{(n-1)\times(n-1)}$, and an upper-triangular matrix $\tilde{\mathbf{U}} \in \mathbb{C}^{(n-1)\times(n-1)}$, such that

$$\tilde{\mathbf{A}} = \tilde{\mathbf{Q}}\, \tilde{\mathbf{U}}\, \tilde{\mathbf{Q}}^* \iff \tilde{\mathbf{Q}}^*\, \tilde{\mathbf{A}}\, \tilde{\mathbf{Q}} = \tilde{\mathbf{U}}.$$

Let us define

$$\mathbf{Q} = \mathbf{Q}_1 \begin{pmatrix} 1 & \mathbf{0}_{n-1}^T \\ \mathbf{0}_{n-1} & \tilde{\mathbf{Q}} \end{pmatrix};$$

then

$$\begin{aligned}\mathbf{Q}^* \mathbf{A}\, \mathbf{Q} &= \begin{pmatrix} 1 & \mathbf{0}_{n-1}^T \\ \mathbf{0}_{n-1} & \tilde{\mathbf{Q}}^* \end{pmatrix} \mathbf{Q}_1^* \mathbf{A}\, \mathbf{Q}_1 \begin{pmatrix} 1 & \mathbf{0}_{n-1}^T \\ \mathbf{0}_{n-1} & \tilde{\mathbf{Q}} \end{pmatrix} \\ &= \begin{pmatrix} 1 & \mathbf{0}_{n-1}^T \\ \mathbf{0}_{n-1} & \tilde{\mathbf{Q}}^* \end{pmatrix} \begin{pmatrix} \lambda_1 & \mathbf{b}^T \\ \mathbf{0}_{n-1} & \tilde{\mathbf{A}} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}_{n-1}^T \\ \mathbf{0}_{n-1} & \tilde{\mathbf{Q}} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \mathbf{b}^T \\ \mathbf{0}_{n-1} & \tilde{\mathbf{Q}}^* \tilde{\mathbf{A}} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}_{n-1}^T \\ \mathbf{0}_{n-1} & \tilde{\mathbf{Q}} \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 & \mathbf{b}^T \tilde{\mathbf{Q}} \\ \mathbf{0}_{n-1} & \tilde{\mathbf{Q}}^* \tilde{\mathbf{A}}\, \tilde{\mathbf{Q}} \end{pmatrix} = \begin{pmatrix} \lambda_1 & \mathbf{b}^T \tilde{\mathbf{Q}} \\ \mathbf{0}_{n-1} & \tilde{\mathbf{U}} \end{pmatrix} = \mathbf{U} \iff \mathbf{A} = \mathbf{Q}\, \mathbf{U}\, \mathbf{Q}^*.\end{aligned}$$

□

Remark: since $\mathbf{Q}$ is unitary, $\mathbf{A} = \mathbf{Q}\, \mathbf{U}\, \mathbf{Q}^{-1}$, i.e., $\mathbf{A}$ and $\mathbf{U}$ are similar matrices, so they have the same spectrum, i.e., the same set of eigenvalues. Moreover, since $\mathbf{U}$ is triangular, the eigenvalues of $\mathbf{A}$ are precisely the diagonal entries of $\mathbf{U}$.

Theorem 2.10 (Spectral theorem for normal matrices)**.** *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$. The following statements are equivalent.*

(i) $\mathbf{A}$ *is normal.*

(ii) There is a unitary matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ and a diagonal matrix $\mathbf{D} \in \mathbb{C}^{n\times n}$ such that

$$\mathbf{A} = \mathbf{Q}\, \mathbf{D}\, \mathbf{Q}^*,$$

i.e., $\mathbf{A}$ *is unitarily diagonalizable.*

(iii) $\displaystyle\sum_{i=1}^{n}\sum_{j=1}^{n} |a_{ij}|^2 = \sum_{i=1}^{n} |\lambda_i|^2$*, where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{A}$ counting multiplicities.*

Proof. *(i)* $\implies$ *(ii)*: By Schur's theorem, $\mathbf{A} = \mathbf{Q}\, \mathbf{U}\, \mathbf{Q}^*$, for some unitary matrix $\mathbf{Q}$ and some upper-triangular matrix $\mathbf{U}$. Then, since $\mathbf{A}$ is normal,

$$\begin{aligned}\mathbf{A}\, \mathbf{A}^* = \mathbf{A}^* \mathbf{A} &\iff (\mathbf{Q}\, \mathbf{U}\, \mathbf{Q}^*)(\mathbf{Q}\, \mathbf{U}\, \mathbf{Q}^*)^* = (\mathbf{Q}\, \mathbf{U}\, \mathbf{Q}^*)^* (\mathbf{Q}\, \mathbf{U}\, \mathbf{Q}^*) \\ &\iff \mathbf{Q}\, \mathbf{U}\, \mathbf{Q}^* \mathbf{Q}\, \mathbf{U}^* \mathbf{Q}^* = \mathbf{Q}\, \mathbf{U}^* \mathbf{Q}^* \mathbf{Q}\, \mathbf{U}\, \mathbf{Q}^* \iff \mathbf{U}\, \mathbf{U}^* = \mathbf{U}^* \mathbf{U}.\end{aligned}$$

Therefore, $\mathbf{U}$ is normal, so, from Theorem 2.2, it is diagonal.
$(ii) \implies (i)$: It is trivial.
$(ii) \implies (iii)$:

$$\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2 = \mathrm{tr}(\mathbf{A}^*\mathbf{A}) = \mathrm{tr}((\mathbf{Q}\,\mathbf{D}\,\mathbf{Q}^*)^*(\mathbf{Q}\,\mathbf{D}\,\mathbf{Q}^*)) = \mathrm{tr}(\mathbf{Q}\,\mathbf{D}^*\mathbf{D}\,\mathbf{Q}^*) = \mathrm{tr}(\mathbf{D}^*\mathbf{D}) = \sum_{i=1}^{n}|\lambda_i|^2,$$

where we have used that $\mathbf{Q}\,\mathbf{D}^*\mathbf{D}\,\mathbf{Q}^* = \mathbf{D}^*\mathbf{D}$ are similar matrices, so they share the trace.
$(iii) \implies (ii)$: By Schur's theorem,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2 = \mathrm{tr}(\mathbf{A}^*\mathbf{A}) = \mathrm{tr}(\mathbf{Q}\,\mathbf{U}^*\mathbf{U}\,\mathbf{Q}^*) = \mathrm{tr}(\mathbf{U}^*\mathbf{U}) = \sum_{i=1}^{n}\sum_{j=1}^{n}|u_{ij}|^2$$

On the other hand,

$$\sum_{i=1}^{n}|\lambda_i|^2 = \sum_{i=1}^{n}|u_{ii}|^2.$$

Therefore, from (iii),

$$\sum_{i=1}^{n}|u_{ii}|^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}|u_{ij}|^2,$$

which implies $u_{ij} = 0$, whenever $i \neq j$, i.e., U is a diagonal matrix.

□

Remark: from (ii), $\mathbf{A}\,\mathbf{Q} = \mathbf{Q}\,\mathbf{D}$, i.e, the columns of $\mathbf{Q}$ are eigenvectors of $\mathbf{A}$, and the diagonal entries of $\mathbf{D}$ are the corresponding eigenvalues of $\mathbf{A}$. Furthermore, since $\mathbf{Q}$ is unitary ($\mathbf{Q}\,\mathbf{Q}^* = \mathbf{I}_n$), its columns are orthonormal. Therefore, (ii) is equivalent to stating that there is a base of $\mathbb{C}^n$ formed by orthonormal eigenvectors of $\mathbf{A}$.

Corollary 2.11. *A normal matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is Hermitian if and only if it has real eigenvalues.*

Proof. From Theorem 2.5, if a matrix $\mathbf{A}$ is Hermitian, it has real eigenvalues. On the other hand, if a matrix $\mathbf{A}$ is normal, the spectral theorem for normal matrices (Theorem 2.10) states that it can be factorized as $\mathbf{A} = \mathbf{Q}\,\mathbf{D}\,\mathbf{Q}^*$, for some unitary matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ and some diagonal matrix $\mathbf{D} \in \mathbb{C}^{n\times n}$. Since the eigenvalues of $\mathbf{A}$ are real, $\mathbf{D}$ is also real, and it follows trivially that $\mathbf{A}$ is Hermitian.

□

Theorem 2.12. *A Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is positive definite if and only if all its eigenvalues are positive.*

Proof. Let $\mathbf{A}$ be a Hermitian, positive-definite matrix. Let λ be an eigenvalue of $\mathbf{A}$, and $\mathbf{u}$ an eigenvector of $\mathbf{A}$ associated to λ; then

$$\mathbf{A}\,\mathbf{u} = \lambda\,\mathbf{u} \implies (\mathbf{A}\,\mathbf{u})^* = (\lambda\,\mathbf{u})^* \implies \mathbf{u}^*\mathbf{A}^* = \mathbf{u}^*\mathbf{A} = \lambda\,\mathbf{u}^* \implies \mathbf{u}^*\mathbf{A}\,\mathbf{u} = \lambda\,\mathbf{u}^*\mathbf{u} \implies \lambda = \frac{\mathbf{u}^*\mathbf{A}\,\mathbf{u}}{\mathbf{u}^*\mathbf{u}} > 0,$$

where we have used that λ is real (see Thereom 2.5).

Let $\mathbf{A}$ be a Hermitian matrix, with all its eigenvalues being positive. From Theorem 2.10 and Corollary 2.11, $\mathbf{A}$ admits a diagonalization of the form $\mathbf{A} = \mathbf{Q}\,\mathbf{D}\,\mathbf{Q}^*$, where $\mathbf{Q} \in \mathbb{C}^{n\times n}$ is a unitary matrix, and $\mathbf{D} \in \mathbb{R}^{n\times n}$ is a diagonal matrix with real entries. The diagonal entries of $\mathbf{D}$ are precisely the eigenvalues of $\mathbf{A}$, and we denote them as $\lambda_1, \ldots, \lambda_n$. On the other hand, the columns of $\mathbf{Q}$, which we denote $\mathbf{q}_1, \ldots, \mathbf{q}_n$, are the eigenvectors of $\mathbf{A}$ associated respectively to $\lambda_1, \ldots, \lambda_n$. Let $\mathbf{u} \in \mathbb{C}^n$; since $\mathbf{q}_1, \ldots, \mathbf{q}_n$ form an orthonormal basis of $\mathbb{C}^n$, $\mathbf{u}$ can be represented as a unique linear combination of them, i.e, there are $\alpha_1, \ldots, \alpha_n \in \mathbb{C}$, such that $\mathbf{u} = \alpha_1\mathbf{q}_1 + \ldots + \alpha_n\mathbf{q}_n$. Then,

$$\begin{aligned}\mathbf{u}^*\mathbf{A}\,\mathbf{u} &= (\alpha_1\mathbf{q}_1 + \ldots + \alpha_n\mathbf{q}_n)^*\mathbf{A}(\alpha_1\mathbf{q}_1 + \ldots + \alpha_n\mathbf{q}_n)\\ &= (\overline{\alpha_1}\mathbf{q}_1^* + \ldots + \overline{\alpha_n}\mathbf{q}_n^*)(\alpha_1\lambda_1\mathbf{q}_1 + \ldots + \alpha_n\lambda_n\mathbf{q}_n)\\ &= \lambda_1|\alpha_1|^2 + \ldots + \lambda_n|\alpha_n|^2 > 0,\end{aligned}$$

i.e., $\mathbf{A}$ is positive definite.

□

Corollary 2.13. *A Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is negative definite, if and only if all its eigenvalues are negative.*

Proof. We have trivially that a Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is negative definite, if and only if $-\mathbf{A}$ is positive definite. Therefore, it is enough to apply the theorem to $-\mathbf{A}$.

□

Definition 2.14. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a square matrix. The leading principal matrices of $\mathbf{A}$ are the square submatrices of order $k \in \{1, \ldots, n\}$ located at the upper left corner of $\mathbf{A}$. The leading principal minors of $\mathbf{A}$ are the determinants of the leading principal matrices of $\mathbf{A}$.*

Theorem 2.15 (Sylvester's criterion)**.** *A Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is positive definite, if and only if all its n leading principal minors are positive.*

Proof. Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a positive definite Hermitian matrix; then, $\mathbf{x}^*\mathbf{A}\,\mathbf{x} > 0$, $\forall \mathbf{x} \in \mathbb{C}^n$; in particular, if we take those $\mathbf{x}$ whose last entry is equal to zero, it follows that the leading $(n-1)\times(n-1)$ principal matrix of $\mathbf{A}$ is positive definite; if we take those $\mathbf{x}$ whose last two entries are equal to zero, it follows that the leading $(n-2)\times(n-2)$ principal matrix of $\mathbf{A}$ is positive definite; and so on. Therefore, if $\mathbf{A} \in \mathbb{C}^{n\times n}$ is positive definite, all its leading matrices are positive definite. Furthermore, since the eigenvalues of a positive-definite Hermitian matrix are real positive, we conclude that all the leading principal minors of $\mathbf{A}$ are positive.

Let us suppose now that the n leading principal minors of $\mathbf{A}$ are positive. In order to prove that $\mathbf{A}$ is positive definite, we use an induction argument on the matrix order. The case $n = 1$ is trivial. Let us assume that the theorem holds for orders $\leq n-1$, so we have to prove it for n. Let $\mathbf{A}$ be a Hermitian matrix of order n, such that its leading principal minors are positive; we have in particular $\det(\mathbf{A}) > 0$. Since $\det(\mathbf{A})$ is the product of the eigenvalues, if $\mathbf{A}$ is not positive definite, it must have at least two negative eigenvalues λ and μ. Let $\mathbf{u} \in \mathbb{C}^n$ and $\mathbf{v} \in \mathbb{C}^n$ be two eigenvalues associated respectively to λ and μ; since $\mathbf{A}$ is Hermitian, $\mathbf{u}$ and $\mathbf{v}$ are orthogonal. Take a linear combination of $\mathbf{u}$ and $\mathbf{v}$, $\mathbf{w} = \alpha\,\mathbf{u} + \beta\,\mathbf{v} \neq \mathbf{0}_n$, such that the last entry of $\mathbf{w}$ is zero; then, $\mathbf{u}^*\mathbf{A}\,\mathbf{v} = \mu\,\mathbf{u}^*\mathbf{v} = 0$, and $\mathbf{v}^*\mathbf{A}\,\mathbf{u} = \lambda\,\mathbf{u}^*\mathbf{v} = 0$, so

$$\mathbf{w}^*\mathbf{A}\,\mathbf{w} = (\bar{\alpha}\,\mathbf{u}^* + \bar{\beta}\,\mathbf{v}^*)\mathbf{A}\,(\alpha\,\mathbf{u} + \beta\,\mathbf{v}) = |\alpha|^2\mathbf{u}^*\mathbf{A}\,\mathbf{u} + |\beta|^2\mathbf{v}^*\mathbf{A}\,\mathbf{v} = \lambda|\alpha|^2\mathbf{u}^*\mathbf{u} + \mu|\beta|^2\mathbf{v}^*\mathbf{v} < 0.$$

Hence, the leading $(n-1)\times(n-1)$ principal submatrix of $\mathbf{A}$ is not positive definite, which is in contradiction with the induction assumption. Therefore, $\mathbf{A}$ must be positive definite.

□

Corollary 2.16. *A Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is negative definite, if and only if its leading principal minors of even order are positive, and its leading principal minors of odd order are negative.*

Proof. We have trivially that a Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is negative definite, if and only if $-\mathbf{A}$ is positive definite. Therefore, it is enough to apply the theorem to $-\mathbf{A}$.

□

Theorem 2.17 (Cholesky decomposition)**.** *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a Hermitian, positive-definite matrix; then, there is a unique factorization of $\mathbf{A}$, called the Choleksy decomposition of $\mathbf{A}$, of the form*

$$\mathbf{A} = \mathbf{L}\,\mathbf{L}^*,$$

where $\mathbf{L} \in \mathbb{C}^{n\times n}$ is a lower-triangular matrix with real and positive diagonal entries.

Proof. Let us prove it by induction on the matrix order. If $n = 1$, then $\mathbf{A} = (a_{ii})$, and since $\mathbf{A}$ is positive definite, $a_{11} > 0$ and the unique choice of $\mathbf{L}$ is $\mathbf{L} = (+\sqrt{a_{ii}})$.

Let us assume that the theorem holds for orders $\leq n-1$, so we have to prove it for n. Given a Hermitian matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, we can represent it as

$$\mathbf{A} = \begin{pmatrix} \tilde{\mathbf{A}} & \mathbf{b} \\ \mathbf{b}^* & c \end{pmatrix},$$

where $\tilde{\mathbf{A}} \in \mathbb{C}^{(n-1)\times(n-1)}$ is a Hermitian matrix, $\mathbf{b} \in \mathbb{C}^{n-1}$ is a vector and c is a real scalar. We have to find a lower-triangular matrix $\mathbf{L} \in \mathbb{C}^{n\times n}$, such that $\mathbf{A} = \mathbf{L}\,\mathbf{L}^*$. Let us represent $\mathbf{L}$ as

$$\mathbf{L} = \begin{pmatrix} \tilde{\mathbf{L}} & \mathbf{0}_{n-1} \\ \mathbf{x}^* & y \end{pmatrix},$$

where $\mathbf{x} \in \mathbb{C}^{n-1}$ is a vector and y is a scalar. Then,

$$\begin{pmatrix} \tilde{\mathbf{A}} & \mathbf{b} \\ \mathbf{b}^* & c \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{L}} & \mathbf{0}_{n-1} \\ \mathbf{x}^* & y \end{pmatrix} \begin{pmatrix} \tilde{\mathbf{L}}^* & \mathbf{x} \\ \mathbf{0}_{n-1}^T & \bar{y} \end{pmatrix} = \begin{pmatrix} \tilde{\mathbf{L}}\,\tilde{\mathbf{L}}^* & \tilde{\mathbf{L}}\,\mathbf{x} \\ \mathbf{x}^*\tilde{\mathbf{L}}^* & \mathbf{x}^*\mathbf{x} + |y|^2 \end{pmatrix}.$$

From Theorem 2.15, $\tilde{\mathbf{A}}$ is positive definite. Therefore, by the induction hypothesis, $\tilde{\mathbf{A}}$ has a Cholesky decomposition and we choose $\tilde{\mathbf{L}}$ to be the only lower-triangular matrix with real and positive diagonal entries, such that $\tilde{\mathbf{A}} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^*$. Remark that $\tilde{\mathbf{L}}$ is nonsingular, because $0 < \det(\tilde{A}) = |\det(\tilde{\mathbf{L}})|^2 \implies \det(\tilde{\mathbf{L}}) \neq 0$, so $\mathbf{b}$ is the unique solution of $\tilde{\mathbf{L}}\,\mathbf{x} = \mathbf{b}$. Finally, y satisfies $|y|^2 = c - \mathbf{x}^*\mathbf{x}$. Moreover, $0 < \det(\mathbf{A}) = |\det(\mathbf{L})|^2 = |\det(\tilde{\mathbf{L}})|^2|y|^2$, so $|y| > 0$, and $y \in \mathbb{R}^+$ is uniquely determined as $y = +\sqrt{c - \mathbf{x}^*\mathbf{x}}$. □

Although the previous proof can be used to compute $\mathbf{L}$, it is more efficient to compute its columns recursively. Indeed, it is immediate to compute the first column of $\mathbf{L}$:

$$a_{i1} = \sum_{p=1}^{n} l_{ip}\overline{l_{1p}} = l_{i1}\overline{l_{11}}, \Longrightarrow \begin{cases} l_{11} = +\sqrt{a_{11}} \in \mathbb{R}^+, \\ l_{i1} = \dfrac{a_{i1}}{l_{11}}, \quad i \in \{2, \ldots, n\}. \end{cases}$$

Then, assuming that the first $k-1$ columns of $\mathbf{L}$ are known, we compute the kth column:

$$a_{ik} = \sum_{p=1}^{n} l_{ip}\overline{l_{kp}} = \sum_{p=1}^{k} l_{ip}\overline{l_{kp}} = \sum_{p=1}^{k-1} l_{ip}\overline{l_{kp}} + l_{ik}\overline{l_{kk}} \Longrightarrow \begin{cases} l_{kk} = +\sqrt{a_{kk} - \displaystyle\sum_{p=1}^{k-1} |l_{kp}|^2} \in \mathbb{R}^+, \\ l_{ik} = \dfrac{a_{ik} - \displaystyle\sum_{p=1}^{k-1} l_{ip}\overline{l_{kp}}}{l_{kk}}, \quad i \in \{k+1, \ldots, n\}. \end{cases}$$

We repeat the procedure for $k \in \{2, \ldots, n\}$, computing first the diagonal entry l_{kk}, which is always chosen to be real and positive, then the remaining entries of the kth column of $\mathbf{L}$.

Respect to the computational cost of the Cholesky decomposition, given a Hermitian matrix $\mathbf{A}$ of order n, the algorithm requires the following operations:

- Square roots: n square roots, corresponding to the n diagonal entries of $\mathbf{L}$.
- Divisions: the computation of each nondiagonal entry of $\mathbf{L}$ requires one division, hence $(n^2 - n)/2$ divisions.
- Multiplications: each entry of the kth column of $\mathbf{L}$ requires $k-1$. Since the kth column of $\mathbf{L}$ has at most $n+1-k$ nonzero entries, there are $(n+1-k)(k-1)$ multiplications per column. Therefore, the total number of multiplications is

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

- Subtractions: the number of subtractions is the same as the the number of multiplications, hence $(n^3 - n)/6$.

Summarizing, if $n \gg 1$, then the number of square roots and of divisions is negligible in comparison with the number of multiplications and subtractions, which is in both cases approximately equal to $n^3/6$. Therefore, the total computational cost of the Cholesky decomposition is of approximately $n^3/6 + n^3/6 = n^3/3$ floating point operations or flops.

Definition 2.18. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$. A nonnegative scalar σ and a pair of vectors $\mathbf{u} \in \mathbb{C}^m$ and $\mathbf{v} \in \mathbb{C}^n$ are called respectively singular value and singular vectors of $\mathbf{A}$ if and only if they satisfy*

$$\begin{aligned} \mathbf{A}\,\mathbf{v} &= \sigma\,\mathbf{u}, \\ \mathbf{A}^*\mathbf{u} &= \sigma\,\mathbf{v}. \end{aligned}$$

Remark: singular vectors are usually normalized to have Euclidean length equal to one, i.e., $\mathbf{u}^*\mathbf{u} = 1$ and $\mathbf{v}^*\mathbf{v} = 1$.

Theorem 2.19 (Singular value decomposition)**.** *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$; then, there exists a factorization of $\mathbf{A}$, called the singular value decomposition (SVD) of $\mathbf{A}$, of the form*

$$\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*,$$

where $\mathbf{U} \in \mathbb{C}^{m\times m}$ *and* $\mathbf{V} \in \mathbb{C}^{n\times n}$ *are unitary matrices and* $\mathbf{\Sigma} \in \mathbb{R}^{m\times n}$ *is a diagonal matrix, with nonnegative real numbers on the diagonal. The diagonal entries of* $\mathbf{\Sigma}$ *are known as the singular values of* $\mathbf{A}$*. If the singular values are listed in descending order, there is uniqueness in the choice of* $\mathbf{\Sigma}$*, but not in the choice of* $\mathbf{U}$ *and* $\mathbf{V}$*.*

Remark: the term "singular value" relates to the distance between a matrix and the set of singular matrices.

Proposition 2.20. *Let* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *be a Hermitian matrix. If* $\lambda \geq 0$ *is an eigenvalue with associated eigenvector* $\mathbf{u}$*, then* λ *is also a singular value, and its associated singular vectors are both equal to* $\mathbf{u}$*. On the other hand, if* $\lambda < 0$ *is an eigenvalue with associated eigenvector* $\mathbf{u}$*, then* $|\lambda|$ *is a singular value, and its associated singular vectors are* $\mathbf{u}$ *and* $-\mathbf{u}$*.*

Proof. Trivial, from Definition 2.18. □

Theorem 2.21. *Let* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *be a normal matrix. Then, the singular values of* $\mathbf{A}$ *are the moduli of the eigenvalues of* $\mathbf{A}$*.*

Proposition 2.22. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$*. Then,* $\mathbf{A}^*\mathbf{A}$ *and* $\mathbf{A}\,\mathbf{A}^*$ *are Hermitian, nonnegative definite matrices, and* $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}^*\mathbf{A}) = \operatorname{rank}(\mathbf{A}\,\mathbf{A}^*)$*.*

Proof. Since $(\mathbf{A}^*\mathbf{A})^* = \mathbf{A}^*\mathbf{A}$, $\mathbf{A}^*\mathbf{A}$ is trivially Hermitian. Let us consider the singular value decomposition of $\mathbf{A}$, $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*$. Then, $\mathbf{A}^*\mathbf{A} = (\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*)^*\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^* = \mathbf{V}\,\mathbf{\Sigma}\,\mathbf{U}^*\mathbf{U}\,\mathbf{\Sigma}^T\mathbf{V}^* = \mathbf{V}(\mathbf{\Sigma}^T\mathbf{\Sigma})\mathbf{V}^*$. Therefore, $\mathbf{A}^*\mathbf{A}$ is orthogonally similar to the diagonal matrix $\mathbf{\Sigma}^T\mathbf{\Sigma}$, so the eigenvalues of $\mathbf{A}^*\mathbf{A}$ are precisely the diagonal entries of $\mathbf{\Sigma}^T\mathbf{\Sigma}$, which are real and nonnegative. Moreover, $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{\Sigma}) = \operatorname{rank}(\mathbf{\Sigma}^T\mathbf{\Sigma}) = \operatorname{rank}(\mathbf{A}^*\mathbf{A})$.

The proof is similar for $\mathbf{A}\,\mathbf{A}^*$; in that case, $\mathbf{A}\,\mathbf{A}^*$ is similar to $\mathbf{\Sigma}\,\mathbf{\Sigma}^T$, which is also a diagonal matrix with nonzero entries, and $\operatorname{rank}(\mathbf{\Sigma}\,\mathbf{\Sigma}^T) = \operatorname{rank}(\mathbf{\Sigma}^T\mathbf{\Sigma})$.

□

Corollary 2.23. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$*, with* $m \geq n$*; then, the singular values* $\sigma_1 \geq \ldots \geq \sigma_n$ *of* $\mathbf{A}$ *are the positive square roots of the eigenvalues of* $\mathbf{A}^*\mathbf{A}$*. Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$*, with* $m < n$*; then, the singular values* $\sigma_1 \geq \ldots \geq \sigma_m$ *of* $\mathbf{A}$ *are the positive square roots of the* m *largest eigenvalues of* $\mathbf{A}^*\mathbf{A}$*. In the latter case, since* $\operatorname{rank}(\mathbf{A}^*\mathbf{A}) = \operatorname{rank}(\mathbf{A}) \leq m < n$*, and* $\mathbf{A}^*\mathbf{A} \in \mathbb{C}^{n\times n}$*, there will be at least* $n - m$ *eigenvalues equal to zero that have to be discarded, when computing the singular values of* $\mathbf{A}$*.*

Reasoning in the same way, it is be possible to work with $\mathbf{A}\,\mathbf{A}^*$*, in order to obtain the singular values of* $\mathbf{A}$*.*

3 Matrix Norms

Definition 3.1. *A matrix norm on* $\mathbb{C}^{m\times n}$ *is a map* $\|\cdot\| : \mathbb{C}^{m\times n} \to \mathbb{R}$*, such that*

- $\|\mathbf{A}\| \geq 0$*,* $\forall \mathbf{A} \in \mathbb{C}^{m\times n}$*; and* $\|\mathbf{A}\| = 0 \iff \mathbf{A} = \mathbf{0}_{m\times n}$ *(positive definiteness).*
- $\|\alpha\,\mathbf{A}\| = |\alpha|\|\mathbf{A}\|$*,* $\forall \alpha \in \mathbb{C}$*,* $\forall \mathbf{A} \in \mathbb{C}^{m\times n}$ *(absolute homogeneity).*
- $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$*,* $\forall \mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times n}$ *(subadditivity or triangle inequality).*

Additionally, in the case of square matrices, the following property is usually required:

- $\|\mathbf{A}\,\mathbf{B}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$*,* $\forall \mathbf{A} \in \mathbb{C}^{m\times n}$*,* $\forall \mathbf{B} \in \mathbb{C}^{n\times l}$ *(submultiplicativity).*

  There are however norms that are not submultiplicative.

Definition 3.2. *Let us consider a vector norm* $\|\cdot\|$ *on* $\mathbb{C}^d$*, where* d *is the dimension of the vector space. An* $m \times n$ *matrix can be regarded as a linear operator from* $\mathbb{C}^n$ *to* $\mathbb{C}^m$*, and the corresponding induced norm or operator norm on the space* $\mathbb{C}^{m\times n}$ *is defined as*

$$\begin{aligned}\|\mathbf{A}\| &= \sup\{\|\mathbf{A}\,\mathbf{x}\| : \mathbf{x} \in \mathbb{C}^n,\ \textit{with } \|\mathbf{x}\| = 1\} \\ &= \sup\left\{\frac{\|\mathbf{A}\,\mathbf{x}\|}{\|\mathbf{x}\|} : \mathbf{x} \in \mathbb{C}^n,\ \textit{with } \mathbf{x} \neq \mathbf{0}_n\right\}.\end{aligned}$$

Observe that $\|\mathbf{I}\| = 1$ *trivially for any induced norm, where* $\mathbf{I}$ *is the identity matrix. Moreover, an induced is naturally submultiplicative:*

$$\|\mathbf{A}\| = \sup_{x\neq\mathbf{0}} \frac{\|\mathbf{A}\,\mathbf{x}\|}{\|\mathbf{x}\|} \Longrightarrow \|\mathbf{A}\,\mathbf{x}\| \leq \|\mathbf{A}\|\|\mathbf{x}\| \Longrightarrow \|\mathbf{A}\,\mathbf{B}\,\mathbf{x}\| \leq \|\mathbf{A}\|\|\mathbf{B}\,\mathbf{x}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|\|\mathbf{x}\|,$$
$$\Longrightarrow \frac{\|\mathbf{A}\,\mathbf{B}\,\mathbf{x}\|}{\|\mathbf{x}\|} \leq \|\mathbf{A}\|\|\mathbf{B}\|,\ \forall \mathbf{x}\neq\mathbf{0} \Longrightarrow \|\mathbf{A}\,\mathbf{B}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|,$$

for all $\mathbf{A}\in\mathbb{C}^{m\times n}$ *and* $\mathbf{B}\in\mathbb{C}^{n\times l}$*. An important case of induced norms is the norm induced by the* L^p*-norm for vectors:*

$$\|\mathbf{A}\|_p = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\,\mathbf{x}\|_p}{\|\mathbf{x}\|_p} = \sup_{\|\mathbf{x}\|_p=1} \|\mathbf{A}\,\mathbf{x}\|_p.$$

Remarks: Even if we are choosing the same vector-norm $\|\cdot\|$ on the departure space $\mathbb{C}^m$ and the arrival space $\mathbb{C}^n$, it would be possible to choose different norms. Regarding $\|\cdot\|_p$, this notation is also used for other types of matrix norms, so some confusion may arise. For example, among others, the entrywise matrix norm of $\mathbf{A}\in\mathbb{C}^{m\times n}$ is also denoted as $\|\mathbf{A}\|_p$: $\|\mathbf{A}\|_p = \|\operatorname{vec}(\mathbf{A})\|_p = (\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^p)^{1/p}$, where the operator vec puts the columns of $\mathbf{A}$ in one single column vector, i.e, $\mathbf{A}$ is regarded as a vector of length $m\times n$. On the other hand, the entrywise matrix norm with $p=2$ is especially important and it is called the Frobenius norm.

Definition 3.3. *Let* $\|\cdot\|_m$ *be a matrix norm, and* $\|\cdot\|_v$ *be a vector norm, where* $_m$ *and* $_v$ *stand respectively for matrix and vector.* $\|\cdot\|_m$ *and* $\|\cdot\|_v$ *are said to be compatible, if and only if*

$$\|\mathbf{A}\,\mathbf{x}\|_v \leq \|\mathbf{A}\|_m\|\mathbf{x}\|_v, \quad \forall \mathbf{A}\in\mathbb{C}^{m\times n}, \forall \mathbf{x}\in\mathbb{C}^n.$$

Observe that induced norms are compatible by definition.

Definition 3.4. *The Frobenius norm of* $\mathbf{A}\in\mathbb{C}^{m\times n}$ *is defined as*

$$\|\mathbf{A}\|_F = \|\operatorname{vec}(\mathbf{A})\|_2 = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{ij}|^2} = \sqrt{\operatorname{tr}(\mathbf{A}^*\mathbf{A})}.$$

Proposition 3.5. *The Frobenius norm is compatible with the Euclidean vector-norm:*

$$\|\mathbf{A}\,\mathbf{x}\|_2 \leq \|\mathbf{A}\|_F\|\mathbf{x}\|_2, \quad \forall \mathbf{A}\in\mathbb{C}^{m\times n}, \forall \mathbf{x}\in\mathbb{C}^n.$$

Proof. Let $\mathbf{a}_1,\ldots,\mathbf{a}_n\in\mathbb{C}^m$ denote the columns of $\mathbf{A}$, then

$$\|\mathbf{A}\,\mathbf{x}\|_2 = \left\|\sum_{j=1}^n x_j\mathbf{a}_j\right\|_2 \leq \sum_{j=1}^n |x_j|\|\mathbf{a}_j\|_2 \leq \|\mathbf{x}\|_2\sqrt{\sum_{j=1}^n \|\mathbf{a}_j\|_2^2} \leq \|\mathbf{x}\|_2\sqrt{\sum_{j=1}^n\sum_{i=1}^m |a_{ij}|^2} = \|\mathbf{A}\|_F\|\mathbf{x}\|_2,$$

where we have used the triangule inequality in the first inequality and the Cauchy-Schwarz theorem in the second inequality.

□

Proposition 3.6. *The Frobenius norm is submultiplicative.*

Proof. Let $\mathbf{A}\in\mathbb{C}^{m\times n}$, $\mathbf{B}\in\mathbb{C}^{n\times p}$, then

$$\|\mathbf{A}\,\mathbf{B}\|_F^2 = \sum_{i=1}^m\sum_{j=1}^p \left|\sum_{k=1}^n a_{ik}b_{kj}\right|^2 \leq \sum_{i=1}^m\sum_{j=1}^p \left(\sum_{k=1}^n |a_{ik}|^2\right)\left(\sum_{l=1}^n |b_{lj}|^2\right)$$
$$\leq \left(\sum_{i=1}^m\sum_{k=1}^n |a_{ik}|^2\right)\left(\sum_{l=1}^n\sum_{j=1}^p |b_{lj}|^2\right) = \|\mathbf{A}\|_F\|\mathbf{B}\|_F,$$

where we have used the Cauchy-Schwarz theorem in the inequality.

□

Definition 3.7. *Let* $\mathbf{A}\in\mathbb{C}^{n\times n}$ *be a complex matrix and* $\lambda_1,\ldots,\lambda_n\in\mathbb{C}$ *its eigenvalues. The spectral radius of* $\mathbf{A}$ *is defined as*

$$\rho(\mathbf{A}) = \max\{|\lambda_1|,\ldots,|\lambda_n|\}.$$

Theorem 3.8. *For any induced matrix norm* $\|\cdot\|_m$*,*

$$\rho(\mathbf{A}) \leq \|\mathbf{A}\|_m, \quad \forall \mathbf{A} \in \mathbb{C}^{n\times n}.$$

Proof. Let $\|\cdot\|_v$ be the vector norm that induces $\|\cdot\|_m$. Let $\mathbf{u} \in \mathbb{C}^n$ be an eigenvector of $\mathbf{A}$, with associated eigenvalue λ, then

$$\|\mathbf{A}\|_m \geq \frac{\|\mathbf{A}\,\mathbf{u}\|_v}{\|\mathbf{u}\|_v} = \frac{\|\lambda\,\mathbf{u}\|_v}{\|\mathbf{u}\|_v} = |\lambda|.$$

This inequality is true for every eigenvalue λ of $\mathbf{A}$, and, in particular, for the one with the largest modulus, which is by definition the spectral radius $\rho(\mathbf{A})$.

□

Corollary 3.9. *For any induced matrix norm* $\|\cdot\|_m$*, any matrix* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *and any natural number* k*,*

$$(\rho(\mathbf{A}))^k \leq \rho(\mathbf{A}^k) \leq \|\mathbf{A}^k\|_m \leq \|\mathbf{A}\|_m^k.$$

Proof. Since $\rho(\mathbf{A}) \leq \|\mathbf{A}\|_m$ for any matrix, it is in particular true for $\mathbf{A}^k$, i.e., $\rho(\mathbf{A}^k) \leq \|\mathbf{A}^k\|$, for all $k \in \mathbb{N}$. On the other hand, the induced matrix norms are submultiplicative, i.e., $\|\mathbf{A}\,\mathbf{B}\|_m \leq \|\mathbf{A}\|_m\|\mathbf{B}\|_m$, so in particular $\|\mathbf{A}^2\|_m \leq \|\mathbf{A}\|_m^2$, and, by induction, $\|\mathbf{A}^k\|_m \leq \|\mathbf{A}\|_m^k$, for all $k \in \mathbb{N}$.

In order to prove the first inequality, let $\|\cdot\|_v$ be the vector norm that induces $\|\cdot\|_m$, and let $u \neq \mathbf{0}$ be an eigenvector of $\mathbf{A}$ associated to the eigenvalue λ; then,

$$|\lambda|^k\|\mathbf{u}\|_v = \|\lambda^k\,\mathbf{u}\|_v = \|\mathbf{A}^k\mathbf{u}\|_v \leq \|\mathbf{A}^k\|_m\|\mathbf{u}\|_v \Longrightarrow |\lambda|^k \leq \|\mathbf{A}^k\|_v.$$

This inequality is true for every eigenvalue λ of $\mathbf{A}$, and, in particular, for the one with the largest modulus, which is by definition the spectral radius $\rho(\mathbf{A})$. Therefore, $(\rho(\mathbf{A}))^k \leq \|\mathbf{A}^k\|_m$, for all $k \in \mathbb{N}$.

□

Theorem 3.10. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$*. Then,*

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

Proof. Let $\mathbf{a}_1, \ldots \mathbf{a}_n \in \mathbb{C}^n$ denote the columns of $\mathbf{A}$, then, for all $\mathbf{x} \neq \mathbf{0}_n$,

$$\begin{aligned}\|\mathbf{A}\,\mathbf{x}\|_1 = \left\|\sum_{j=1}^{n} x_j\,\mathbf{a}_j\right\|_1 &\leq \sum_{j=1}^{n} \|x_j\,\mathbf{a}_j\|_1 = \sum_{j=1}^{n} |x_j|\|\mathbf{a}_j\|_1 \\ &\leq \left(\sum_{j=1}^{n} |x_j|\right) \max_{1\leq j\leq n} \|\mathbf{a}_j\|_1 = \|\mathbf{x}\|_1 \max_{1\leq j\leq n} \|\mathbf{a}_j\|_1\,;\end{aligned}$$

hence,

$$\frac{\|\mathbf{A}\,\mathbf{x}\|_1}{\|\mathbf{x}\|_1} \leq \max_{1\leq j\leq n} \|\mathbf{a}_j\|_1. \tag{1}$$

On the other hand, let k be the index where the maximum of $\|\mathbf{a}_j\|_1$ is achieved, i.e.,

$$\|\mathbf{a}_k\|_1 = \max_{1\leq j\leq n} \|\mathbf{a}_j\|_1.$$

Take $\mathbf{x} = \mathbf{e}_k$, where $\mathbf{e}_k$ is the vector with all its entries equal to 0, except for its kth entry, which is equal to one; then,

$$\|\mathbf{A}\,\mathbf{e}_k\|_1 = \|\mathbf{a}_k\|_1 = \|\mathbf{e}_k\|_1\|\mathbf{a}_k\|_1 = \|\mathbf{e}_k\|_1 \max_{1\leq j\leq n} \|\mathbf{a}_j\|_1;$$

hence,

$$\max_{1\leq j\leq n} \|\mathbf{a}_j\|_1 = \frac{\|\mathbf{A}\,\mathbf{e}_k\|_1}{\|\mathbf{e}_k\|_1}. \tag{2}$$

Putting together (1) and (2),

$$\frac{\|\mathbf{A}\,\mathbf{x}\|_1}{\|\mathbf{x}\|_1} \leq \max_{1\leq j\leq n} \|\mathbf{a}_j\|_1 = \frac{\|\mathbf{A}\,\mathbf{e}_k\|_1}{\|\mathbf{e}_k\|_1}, \quad \forall \mathbf{x} \neq \mathbf{0}_n,$$

i.e.,

$$\|\mathbf{A}\|_1 = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\,\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \max_{1\leq j\leq n} \|\mathbf{a}_j\|_1 = \frac{\|\mathbf{A}\,\mathbf{e}_k\|_1}{\|\mathbf{e}_k\|_1} \Longrightarrow \|\mathbf{A}\|_1 = \max_{1\leq j\leq n} \sum_{i=1}^{m} |a_{ij}|.$$

□

Theorem 3.11. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$. Then,*

$$\|\mathbf{A}\|_\infty = \|\mathbf{A}^*\|_1 = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|.$$

Proof. Let $\mathbf{a}_1, \dots \mathbf{a}_n \in \mathbb{C}^n$ denote the columns of $\mathbf{A}$, then, for all $\mathbf{x} \neq \mathbf{0}_n$,

$$\begin{aligned}\|\mathbf{A}\,\mathbf{x}\|_\infty &= \left\| \sum_{j=1}^{n} x_j\,\mathbf{a}_j \right\|_\infty = \max_{1\le i\le m} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \le \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}||x_j| \\ &\le \left(\max_{1\le j\le n} |x_j| \right) \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}| = \|\mathbf{x}\|_\infty \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|.\end{aligned}$$

hence,

$$\frac{\|\mathbf{A}\,\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \le \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|. \tag{3}$$

On the other hand, let p be the index where the maximum of $\sum_{j=1}^{n} |a_{ij}|$ is achieved, i.e.,

$$\sum_{j=1}^{n} |a_{pj}| = \sum_{j=1}^{n} |a_{ij}|.$$

Take $\mathbf{x} = \mathbf{1}_n$, where $\mathbf{1}_n \in \mathbb{R}^n$ is the vector having all its entries equal to 1, then

$$\|\mathbf{A}\cdot\mathbf{1}_n\|_\infty = \left\| \sum_{j=1}^{n} \mathbf{a}_j \right\|_\infty = \max_{1\le i\le m} \left| \sum_{j=1}^{n} a_{ij} \right| = \|\mathbf{1}_n\|_\infty \max_{1\le i\le m} \left| \sum_{j=1}^{n} a_{ij} \right|;$$

hence,

$$\max_{1\le i\le m} \left| \sum_{j=1}^{n} a_{ij} \right| = \frac{\|\mathbf{A}\cdot\mathbf{1}_n\|_\infty}{\|\mathbf{1}_n\|_\infty}. \tag{4}$$

Putting together (3) and (4),

$$\frac{\|\mathbf{A}\,\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \le \max_{1\le i\le m} \left| \sum_{j=1}^{n} a_{ij} \right| = \frac{\|\mathbf{A}\cdot\mathbf{1}_n\|_\infty}{\|\mathbf{1}_n\|_\infty}, \quad \forall \mathbf{x} \neq \mathbf{0}_n,$$

i.e.,

$$\|\mathbf{A}\|_\infty = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\,\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \max_{1\le i\le m} \left| \sum_{j=1}^{n} a_{ij} \right| = \frac{\|\mathbf{A}\cdot\mathbf{1}_n\|_\infty}{\|\mathbf{1}_n\|_\infty} \Longrightarrow \|\mathbf{A}\|_\infty = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}|.$$

Finally, from Theorem 3.10,

$$\|\mathbf{A}^*\|_1 = \max_{1\le i\le m} \sum_{j=1}^{n} |a_{ij}| \Longrightarrow \|\mathbf{A}\|_\infty = \|\mathbf{A}^*\|.$$

This concludes the proof.

□

Theorem 3.12. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$. Then,*

$$\|\mathbf{A}\|_2 = \sigma_1 = \sqrt{\rho(\mathbf{A}^*\mathbf{A})} = \sqrt{\rho(\mathbf{A}\,\mathbf{A}^*)} = \|\mathbf{A}^*\|_2,$$

where σ_1 is the largest singular value of $\mathbf{A}$, and ρ denotes the spectral radius of a matrix.

Proof. Let $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*$ be the singular value decomposition of $\mathbf{A}$. Let us observe first that, given $\mathbf{x} \in \mathbb{C}^n$, $\|\mathbf{x}\|_2 = 1 \iff \sqrt{\mathbf{x}^*\mathbf{x}} = 1 \iff \sqrt{\mathbf{x}^*\overline{\mathbf{V}}\,\overline{\mathbf{V}}^*\mathbf{x}} = 1 \iff \sqrt{(\overline{\mathbf{V}}^*\mathbf{x})^*\overline{\mathbf{V}}^*\mathbf{x}} = 1 \iff \|\mathbf{V}^*\mathbf{x}\|_2 = 1$. Therefore, denoting $\mathbf{y} = \mathbf{V}^*\mathbf{x}$,

$$\begin{aligned}\|\mathbf{A}\|_2 &= \sup_{\|\mathbf{x}\|_2=1} \|\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*\mathbf{x}\|_2 = \sup_{\|\mathbf{V}^*\mathbf{x}\|_2=1} \|\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*\mathbf{x}\|_2 = \sup_{\|\mathbf{y}\|_2=1} \|\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{y}\|_2 = \sup_{\|\mathbf{y}\|_2=1} \sqrt{(\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{y})^*\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{y}} \\ &= \sup_{\|\mathbf{y}\|_2=1} \sqrt{\mathbf{y}^*\mathbf{\Sigma}^T\,\mathbf{U}^*\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{y}} = \sup_{\|\mathbf{y}\|_2=1} \sqrt{\mathbf{y}^*\mathbf{\Sigma}^T\mathbf{\Sigma}\,\mathbf{y}}.\end{aligned}$$

If we denote $p = \min\{m, n\}$, $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{C}^n$, and $\sigma_i = \sigma_{ii}$ is the $i-th$ diagonal entry of $\mathbf{\Sigma}$, then $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$, and the last expression becomes

$$\|\mathbf{A}\|_2 = \max_{\|\mathbf{y}\|_2=1} \sqrt{\sigma_1^2|y_1|^2 + \ldots + \sigma_p^2|y_p|^2}.$$

On the one hand,

$$\max_{\|\mathbf{y}\|_2=1} \sqrt{\sigma_1^2|y_1|^2 + \ldots + \sigma_p^2|y_p|^2} \leq \sigma_1 \max_{\|\mathbf{y}\|_2=1} \sqrt{|y_1|^2 + \ldots + |y_p|^2} = \sigma_1.$$

On the other hand,

$$\max_{\|\mathbf{y}\|_2=1} \sqrt{\sigma_1^2|y_1|^2 + \ldots + \sigma_p^2|y_p|^2} \geq \max_{\|\mathbf{y}\|_2=1} \sqrt{\sigma_1^2|y_1|^2} = \sigma_1,$$

which is achieved at $\mathbf{y} = (1, 0, \ldots, 0)^T$. Since, from Corollary 2.23, the largest singular value of $\mathbf{A}$ is the square root of the largest eigenvector of $\mathbf{A}^*\mathbf{A}$ or $\mathbf{A}\,\mathbf{A}^*$, we conclude that

$$\|\mathbf{A}\|_2 = \sigma_1 = \sqrt{\rho(\mathbf{A}^*\mathbf{A})} = \sqrt{\rho(\mathbf{A}\,\mathbf{A}^*)} = \|\mathbf{A}^*\|_2.$$

□

Corollary 3.13. *If $\mathbf{A} \in \mathbb{C}^{n\times n}$ is normal,*

$$\|\mathbf{A}^k\|_2 = \|\mathbf{A}\|_2^k = (\rho(\mathbf{A}))^k, \quad \forall k \in \mathbb{N}.$$

Proof. From the spectral theorem for normal matrices (Theorem 2.10), $\mathbf{A}$ can be factorized as $\mathbf{A} = \mathbf{Q}\,\mathbf{D}\,\mathbf{Q}^*$, for some unitary matrix $\mathbf{Q} \in \mathbb{C}^{n\times n}$ and some diagonal matrix $\mathbf{D} \in \mathbb{C}^{n\times n}$. Therefore, $\mathbf{A}^2 = \mathbf{Q}\,\mathbf{D}\,\mathbf{Q}^*\mathbf{Q}\,\mathbf{D}\,\mathbf{Q}^* = \mathbf{Q}\,\mathbf{D}^2\,\mathbf{Q}^*$, and, by induction, $\mathbf{A}^k = \mathbf{Q}\,\mathbf{D}^k\mathbf{Q}^*$, for all $k \in \mathbb{N}$. Hence, $(\mathbf{A}^k)^*\mathbf{A}^k = \mathbf{Q}(\mathbf{D}^*\mathbf{D})^k\mathbf{Q}^*$, so

$$\|\mathbf{A}^k\|_2 = \sqrt{\rho((\mathbf{A}^k)^*(\mathbf{A}^k))} = \sqrt{\rho((\mathbf{D}^*\mathbf{D})^k)} = \left[\sqrt{\rho(\mathbf{D}^*\mathbf{D})}\right]^k = \|\mathbf{A}\|_2^k.$$

Furthermore, since $\rho(\mathbf{A}) = \rho(\mathbf{D}) = \rho(\mathbf{D}^*) = \rho(\mathbf{A}^*)$,

$$\|\mathbf{A}^k\|_2 = \sqrt{\rho((\mathbf{D}^*\mathbf{D})^k)} = \sqrt{\rho(\mathbf{D})^{2k}} = (\rho(\mathbf{D}))^k = (\rho(\mathbf{A}))^k.$$

□

Remark: the case $k = 1$ is especially important:

$$\|\mathbf{A}\|_2 = \rho(\mathbf{A}).$$

Theorem 3.14. *$\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_\infty$ and $\|\cdot\|_F$ are equivalent matrix norms. More precisely, for any $\mathbf{A} \in \mathbb{C}^{m\times n}$,*

$$\|\mathbf{A}\|_2 \leq \|\mathbf{A}\|_F \leq \sqrt{n}\|\mathbf{A}\|_2,$$
$$\frac{1}{\sqrt{n}}\|\mathbf{A}\|_\infty \leq \|\mathbf{A}\|_2 \leq \sqrt{m}\|\mathbf{A}\|_\infty,$$
$$\frac{1}{\sqrt{m}}\|\mathbf{A}\|_1 \leq \|\mathbf{A}\|_2 \leq \sqrt{n}\|\mathbf{A}\|_1.$$

Proof. Let $\mathbf{A} \in \mathbb{C}^{m\times n}$, then,

$$\|\mathbf{A}\|_2^2 = \rho(\mathbf{A}^*\mathbf{A}) \leq \mathrm{tr}(\mathbf{A}^*\mathbf{A}) = \|\mathbf{A}\|_F^2 \leq n\,\rho(\mathbf{A}^*\mathbf{A}) = n\|\mathbf{A}\|_2^2.$$

where we have used that the trace of a square matrix is equal to the sum of its eigenvalues.

In order to prove the equivalence between the $\|\cdot\|_2$ and the $\|\cdot\|_\infty$ matrix norms, let us observe first that, in the case of vectors,

$$\|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_2 \leq \sqrt{n}\|\mathbf{x}\|_\infty, \quad \forall \mathbf{x} \in \mathbb{C}^n.$$

This is so, because, given $\mathbf{x} = (x_1, \ldots, x_n)^T$, if k is the index of the component of $\mathbf{x}$ at which $\|\mathbf{x}\|_\infty = |x_k|$, then $|x_k| \leq (\sum_{j=1}^n |x_j|^2)^{1/2} \leq (n|x_k|)^{1/2}$. Hence,

$$\|\mathbf{A}\|_\infty = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\,\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} \leq \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\,\mathbf{x}\|_2}{\frac{1}{\sqrt{n}}\|\mathbf{x}\|_2} = \sqrt{n}\|\mathbf{x}\|_2,$$
$$\|\mathbf{A}\|_2 = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\,\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\sqrt{m}\|\mathbf{A}\,\mathbf{x}\|_\infty}{\|\mathbf{x}\|_\infty} = \sqrt{m}\|\mathbf{A}\|_\infty,$$

where we have used in the latter inequality that $\mathbf{A}\,\mathbf{x} \in \mathbb{C}^m$. Combining both inequalities, we conclude that

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

This last expression is also valid for $\mathbf{A}^* \in \mathbb{C}^{n\times m}$:

$$\frac{1}{\sqrt{m}}\|\mathbf{A}^*\|_\infty \leq \|\mathbf{A}^*\|_2 \leq \sqrt{n}\|\mathbf{A}^*\|_\infty.$$

Hence, bearing in mind that $\|\mathbf{A}^*\|_\infty = \|\mathbf{A}\|_1$ and that $\|\mathbf{A}^*\|_2 = \|\mathbf{A}\|_2$, the equivalence between the $\|\cdot\|_1$ and the $\|\cdot\|_2$ matrix-norms follows:

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

□

Theorem 3.15. *All the induced matrix p-norms are equivalent.*

Remark: the equivalence of two induced matrix norms is a corollary of the equivalence of their respective vector norms, as we have seen in Theorem 3.14.

4 The Moore-Penrose Inverse

Definition 4.1. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$. The pseudoinverse or Moore-Penrose Inverse of $\mathbf{A}$ is the only matrix $\mathbf{A}^+ \in \mathbb{C}^{n\times m}$ that satisfies the following properties:*

1. $\mathbf{A}\,\mathbf{A}^+\mathbf{A} = \mathbf{A}^+$
2. $\mathbf{A}^+\mathbf{A}\,\mathbf{A}^+ = \mathbf{A}$
3. $(\mathbf{A}\,\mathbf{A}^+)^* = \mathbf{A}\,\mathbf{A}^+$
4. $(\mathbf{A}^+\mathbf{A})^* = \mathbf{A}^+\mathbf{A}$

Definition 4.2. *Let $\mathbf{D} \in \mathbb{C}^{m\times n}$ be a diagonal matrix, i.e., such that $d_{ij} = 0$, for all $i \neq j$. The Moore-Penrose inverse of $\mathbf{D}$ is defined as the diagonal matrix $\mathbf{D}^+ \in \mathbb{C}^{n\times m}$, such that its diagonal entries are given as follows:*

$$d_{ii}^+ = \begin{cases} d_{ii}^{-1}, & d_{ii} \neq 0, \\ 0, & d_{ii} = 0. \end{cases}$$

Let $\mathbf{A} \in \mathbb{C}^{m\times n}$, and let $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*$ the singular value decomposition of $\mathbf{A}$. The Moore-Penrose inverse of $\mathbf{A}$ is defined as

$$\mathbf{A}^+ = \mathbf{V}\,\mathbf{\Sigma}^+\mathbf{U}^*.$$

Remark: if $\mathbf{A} \in \mathbb{C}^{n\times n}$ is nonsingular, we have trivially $\mathbf{A}^+ = \mathbf{A}^{-1}$.

Proposition 4.3. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$. If $\mathbf{A}$ has linearly independent columns (and, hence, $\mathbf{A}^*\mathbf{A}$ is nonsingular), the Moore-Penrose inverse of $\mathbf{A}$ is given by*

$$\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*.$$

On the other hand, if $\mathbf{A}$ has linearly independent rows (and, hence, $\mathbf{A}\,\mathbf{A}^*$ is nonsingular), the Moore-Penrose inverse of $\mathbf{A}$ is given by

$$\mathbf{A}^+ = \mathbf{A}^*(\mathbf{A}\,\mathbf{A}^*)^{-1}.$$

Proof. If $\mathbf{A}$ has linearly independent columns, then $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{A}^*\mathbf{A}) = n \leq m$. Let $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*$ be the singular value decomposition of $\mathbf{A}$, then

$$\begin{aligned}(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^* &= (\mathbf{V}\,\mathbf{\Sigma}^T\mathbf{U}^*\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*)^{-1}\mathbf{V}\,\mathbf{\Sigma}^T\mathbf{U}^* = (\mathbf{V}\,\mathbf{\Sigma}^T\mathbf{\Sigma}\,\mathbf{V}^*)^{-1}\mathbf{V}\,\mathbf{\Sigma}^T\mathbf{U}^* \\ &= \mathbf{V}(\mathbf{\Sigma}^T\mathbf{\Sigma})^{-1}\mathbf{V}^*\mathbf{V}\,\mathbf{\Sigma}^T\mathbf{U}^* = \mathbf{V}(\mathbf{\Sigma}^T\mathbf{\Sigma})^{-1}\mathbf{\Sigma}^T\mathbf{U}^* = \mathbf{V}\,\mathbf{\Sigma}^+\mathbf{U}^*,\end{aligned}$$

where we have used that $(\mathbf{\Sigma}^T\mathbf{\Sigma})^{-1}$ is a square, diagonal, nonsingular matrix whose diagonal matrices are precisely $\sigma_1^{-2}, \ldots, \sigma_n^{-2}$, so we have precisely $(\mathbf{\Sigma}^T\mathbf{\Sigma})^{-1}\mathbf{\Sigma}^T = \mathbf{\Sigma}^+$. Observe that $\mathbf{A}^+$ is a left inverse of $\mathbf{A}$, because $\mathbf{A}^+\mathbf{A} = \mathbf{I}$.

The proof is identical in the case that $\mathbf{A}$ has linearly independent rows. In that case, $\mathbf{A}^+$ is a right inverse of $\mathbf{A}$, because $\mathbf{A}\,\mathbf{A}^+ = \mathbf{I}$.

□

Theorem 4.4. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ and $\mathbf{b} \in \mathbb{C}^m$; then, the least-square solution of $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ is given by $\mathbf{x} = \mathbf{A}^+\mathbf{b} \in \mathbb{C}^n$; i.e.,*

- $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ *makes* $\|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2$ *minimum.*
- *Among the values of* $\mathbf{x}$ *for which the minimum of* $\|\mathbf{A}\,\mathbf{x}-\mathbf{b}\|_2$ *is achieved (in case of no uniqueness),* $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ *is the one with the smallest* $\|\cdot\|_2$ *norm.*

Proof. Let $\mathbf{A} \in \mathbb{C}^{m\times n}$, $m \geq n$, and let $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*$ be its singular value decomposition; then,

$$\|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2 = \|\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*\mathbf{x} - \mathbf{b}\|_2 = \|\mathbf{U}^*(\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*\mathbf{x} - \mathbf{b})\|_2 = \|\mathbf{\Sigma}\,\mathbf{V}^*\mathbf{x} - \mathbf{U}^*\mathbf{b}\|_2,$$

where we have used that the $\|\cdot\|_2$ is invariant by unitary matrices, i.e., $\|\mathbf{U}^*\mathbf{u}\|_2 = \|\mathbf{u}\|_2$, for all $\mathbf{u} \in \mathbb{C}^n$. Let us define $\mathbf{y} = \mathbf{V}^*\mathbf{x} \in \mathbb{C}^m$ and $\mathbf{z} = \mathbf{U}^*\mathbf{b} \in \mathbb{C}^m$; then, bearing in mind that $\operatorname{rank}(\mathbf{A}) = \operatorname{rank}(\mathbf{\Sigma}) = r \leq n \leq m$,

$$\begin{aligned}\|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2 = \|\mathbf{\Sigma}\,\mathbf{y} - \mathbf{z}\|_2 &= \|(\sigma_1 y_1, \ldots, \sigma_r y_r, 0, \ldots, 0)^T - (z_1 \ldots z_m)^T\|_2 \\ &= \sqrt{(\sigma_1 y_1 - z_1)^2 + \ldots + (\sigma_r y_r - z_r)^2 + z_{r+1}^2 + \ldots z_m^2}.\end{aligned}$$

Therefore, the minimum of $\|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2$ is achieved at $y_1 = z_1/\sigma_1, \ldots, y_r = z_r/\sigma_r$. On the other hand, the components $y_{r+1}, \ldots, y_m$ have no effect on $\|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2$, but we make them equal zero, in order that $\|\mathbf{y}\|_2$ is as small as possible. Summarizing,

$$\mathbf{y} = (z_1/\sigma_1, \ldots, z_r/\sigma_r, 0, \ldots, 0)^T \Longleftrightarrow \mathbf{y} = \mathbf{\Sigma}^+\mathbf{z} \Longleftrightarrow \mathbf{x} = \mathbf{V}\,\mathbf{\Sigma}^+\mathbf{U}^*\mathbf{b} = \mathbf{A}^+\mathbf{b},$$

and

$$\min_{\mathbf{x}\in\mathbb{C}^n} \|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2 = \sqrt{z_{r+1}^2 + \ldots z_m^2} = \|(0, \ldots, 0, z_{r+1}, \ldots, z_m)^T\|_2$$

□

Theorem 4.5. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ and $\mathbf{B} \in \mathbb{C}^{m\times p}$; then, the least-square solution of $\mathbf{A}\,\mathbf{X} = \mathbf{B}$ is given by $\mathbf{X} = \mathbf{A}^+\mathbf{B} \in \mathbb{C}^{n\times p}$; i.e.,*

- $\mathbf{X} = \mathbf{A}^+\mathbf{B}$ *makes* $\|\mathbf{A}\,\mathbf{X} - \mathbf{B}\|_F$ *minimum.*
- *Among the values of* $\mathbf{X}$ *for which the minimum of* $\|\mathbf{A}\,\mathbf{X}-\mathbf{B}\|_F$ *is achieved (in case of no uniqueness),* $\mathbf{X} = \mathbf{A}^+\mathbf{B}$ *is the one with the smallest* $\|\cdot\|_F$ *norm.*

Proof. The proof is similar to that of Theorem 4.4.

□

5 Condition number

Definition 5.1. *Given a square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, the condition number of $\mathbf{A}$ relative to the norm $\|\cdot\|$ is*

$$\kappa(\mathbf{A}) = \begin{cases} \|\mathbf{A}\|\|\mathbf{A}^{-1}\|, & \det(\mathbf{A}) \neq 0, \\ \infty, & \det(\mathbf{A}) = 0. \end{cases}$$

In the case of a rectangular matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$, the condition number of $\mathbf{A}$ relative to the norm $\|\cdot\|$ becomes

$$\kappa(\mathbf{A}) = \begin{cases} \|\mathbf{A}\|\|\mathbf{A}^+\|, & \operatorname{rank}(\mathbf{A}) = \min\{m, n\}, \\ \infty, & \operatorname{rank}(\mathbf{A}) < \min\{m, n\}, \end{cases}$$

where $\mathbf{A}^+$ is the pseudoinverse of $\mathbf{A}$.

Remarks: it is possible to find other not fully equivalent definitions of the condition number in the literature. For instance, there are authors that define $\kappa(\mathbf{A}) = \|\mathbf{A}\|\|\mathbf{A}^+\|$ for any $\mathbf{A}$, irrespectively of its size and rank, even when $\mathbf{A}$ is square and singular. On the other hand, in the case of rectangular matrices, some authors define $\kappa(\mathbf{A})$ only when $m > n$ and $\operatorname{rank}(\mathbf{A}) = n$, i.e., $\mathbf{A}$ is a full column rank matrix. However, the definition we are using has, among others, the advantage of being coherent with the expression of the condition number relative to the Euclidean norm, as given in Corollary 5.5.

Definition 5.2. *Given a square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, the condition number of $\mathbf{A}$ relative to the norm $\|\cdot\|$ is*

$$\kappa(\mathbf{A}) = \frac{\sup\limits_{\mathbf{x}\neq\mathbf{0}} \dfrac{\|\mathbf{A}\,\mathbf{x}\|}{\|\mathbf{x}\|}}{\inf\limits_{\mathbf{x}\neq\mathbf{0}} \dfrac{\|\mathbf{A}\,\mathbf{x}\|}{\|\mathbf{x}\|}}.$$

Proposition 5.3. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$. Then, Definition 5.1 and Definition 5.2 of $\kappa(\mathbf{A})$ are equivalent.*

Proof. If $\mathbf{A}$ is singular, then $\mathbf{A}\,\mathbf{x} = \mathbf{0}$, for some $\mathbf{x} \neq \mathbf{0}$, so $\inf_{\mathbf{x}\neq\mathbf{0}}(\|\mathbf{A}\,\mathbf{x}\|/\|\mathbf{x}\|) = 0$, and, hence, $\kappa(\mathbf{A}) = 0$. Conversely, if $\kappa(\mathbf{A}) = 0$, then there is some $\mathbf{x} \neq 0$ for which $\|\mathbf{A}\,\mathbf{x}\|/\|\mathbf{x}\| = 0$, so $\mathbf{A}\,\mathbf{x} = 0$, which implies that $\mathbf{A}$ is singular. On the other hand, if $\mathbf{A}$ is nonsingular,

$$\|\mathbf{A}^{-1}\| = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}^{-1}\mathbf{x}\|}{\|\mathbf{x}\|} = \sup_{\mathbf{A}\,\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}^{-1}\mathbf{A}\,\mathbf{x}\|}{\|\mathbf{A}\,\mathbf{x}\|} = \sup_{\mathbf{A}\,\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{x}\|}{\|\mathbf{A}\,\mathbf{x}\|} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{x}\|}{\|\mathbf{A}\,\mathbf{x}\|} = \sup_{\mathbf{x}\neq\mathbf{0}} \frac{1}{\dfrac{\|\mathbf{A}\,\mathbf{x}\|}{\|\mathbf{x}\|}} = \frac{1}{\inf\limits_{\mathbf{x}\neq\mathbf{0}} \dfrac{\|\mathbf{A}\,\mathbf{x}\|}{\|\mathbf{x}\|}},$$

where we have used that $\mathbf{A}$ maps $\mathbb{C}^n$ into itself. Therefore,

$$\kappa(\mathbf{A}) = \frac{\sup\limits_{\mathbf{x}\neq\mathbf{0}} \dfrac{\|\mathbf{A}\,\mathbf{x}\|}{\|\mathbf{x}\|}}{\inf\limits_{\mathbf{x}\neq\mathbf{0}} \dfrac{\|\mathbf{A}\,\mathbf{x}\|}{\|\mathbf{x}\|}} = \|\mathbf{A}\|\|\mathbf{A}^{-1}\|.$$

□

Lemma 5.4. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$, and $\operatorname{rank}(\mathbf{A}) = p$, where $p = \min\{m, n\}$. Then,*

$$\|\mathbf{A}^{+}\|_2 = \frac{1}{\sigma_p} = \frac{1}{\inf\limits_{\mathbf{x}\neq\mathbf{0}} \dfrac{\|\mathbf{A}\,\mathbf{x}\|_2}{\|\mathbf{x}\|_2}},$$

where σ_p is the smallest singular value of $\mathbf{A}$.

Proof. The proof is a corollary of Theorem 3.12, that stated that $\|\mathbf{A}\|_2 = \sigma_1$, where σ_1 is the largest singular value of $\mathbf{A}$. Let $\mathbf{A} = \mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*$ be the singular value decomposition of $\mathbf{A}$. Then $\|\mathbf{A}^{+}\|_2 = \|\mathbf{V}\,\mathbf{\Sigma}^{+}\mathbf{U}^*\|_2 = \|\mathbf{\Sigma}^{+}\|_2$. Now, bearing in mind that the singular values of $\mathbf{A}$ are $\sigma_1 \geq \ldots \geq \sigma_p > 0$, it follows that the singular values of $\mathbf{A}^{+}$ are $\sigma_p^{-1} \geq \ldots \geq \sigma_1^{1} > 0$. Hence, $\|\mathbf{A}^{+}\| = \sigma_p^{-1}$, which is the largest eigenvalue of $\mathbf{A}^{+}$. Finally,

$$\inf_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\,\mathbf{x}\|_2}{\|\mathbf{x}\|_2} = \inf_{\|\mathbf{x}\|_2=1} \|\mathbf{A}\,\mathbf{x}\|_2 = \inf_{\|\mathbf{x}\|_2=1} \|\mathbf{U}\,\mathbf{\Sigma}\,\mathbf{V}^*\mathbf{x}\|_2 = \inf_{\|\mathbf{x}\|_2=1} \|\mathbf{\Sigma}\|_2 = \sigma_p,$$

where we have reasoned as in the proof of Theorem 3.12.

□

Corollary 5.5. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$, then the condition number associated to the Euclidean norm is*

$$\kappa_2(\mathbf{A}) = \frac{\sigma_1}{\sigma_p} = \frac{\sup\limits_{\mathbf{x}\neq\mathbf{0}} \dfrac{\|\mathbf{A}\,\mathbf{x}\|_2}{\|\mathbf{x}\|_2}}{\inf\limits_{\mathbf{x}\neq\mathbf{0}} \dfrac{\|\mathbf{A}\,\mathbf{x}\|_2}{\|\mathbf{x}\|_2}},$$

where $p = \min\{m, n\}$, and σ_1 and σ_p are respectively the largest and smallest singular values of $\mathbf{A}$.

Remark that, in the case that $p < \min\{m, n\}$, we have $\sigma_p = 0$, so $\kappa_2(\mathbf{A}) = \infty$. Conversely, if $\kappa_2(\mathbf{A}) = \infty$, then $\sigma_p = 0$, so $\operatorname{rank}(\mathbf{A}) < p$. Therefore, this formula for the condition number associated to the Euclidean norm is equivalent to Definition 5.1 and Definition 5.2 for both square and rectangular matrices. Furthermore, if $\mathbf{A} \in \mathbb{C}^{n\times n}$ is a normal matrix, then

$$\kappa_2(\mathbf{A}) = \frac{|\lambda_1|}{|\lambda_n|},$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of $\mathbf{A}$, with $|\lambda_1| \geq \ldots \geq |\lambda_n|$.

Proposition 5.6. *Let $\mathbf{A} \in \mathbb{C}^{m\times n}$, then $\kappa(\mathbf{A}) \geq 1$.*

Proof. Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a square matrix. If $\mathbf{A}$ is singular, then $\kappa(\mathbf{A}) = \infty > 0$; if $\mathbf{A}$ is nonsingular, then

$$1 = \|\mathbf{I}\| = \|\mathbf{A} \cdot \mathbf{A}^{-1}\| \leq \|\mathbf{A}\|\|\mathbf{A}^{-1}\| = \kappa(\mathbf{A}).$$

On the other hand, let $\mathbf{A} \in \mathbb{C}^{m\times n}$ be a rectangular matrix, with $m > n$. If $\operatorname{rank}(\mathbf{A}) < n$, then $\kappa(\mathbf{A}) = \infty > 0$; if $\operatorname{rank}(\mathbf{A}) = n$, then $\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$; hence,

$$1 = \|\mathbf{I}\| = \|(\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{A}\| = \|\mathbf{A}^+\mathbf{A}\| \leq \|\mathbf{A}\|\|\mathbf{A}^+\| = \kappa(\mathbf{A}).$$

Finally, let $\mathbf{A} \in \mathbb{C}^{m\times n}$ be a rectangular matrix, with $m < n$. If $\operatorname{rank}(\mathbf{A}) < m$, then $\kappa(\mathbf{A}) = \infty > 0$; if $\operatorname{rank}(\mathbf{A}) = m$, then $\mathbf{A}^+ = \mathbf{A}^*(\mathbf{A}\,\mathbf{A}^*)^{-1}$; hence,

$$1 = \|\mathbf{I}\| = \|\mathbf{A}\,\mathbf{A}^*(\mathbf{A}\,\mathbf{A}^*)^{-1}\| = \|\mathbf{A}\,\mathbf{A}^+\| \leq \|\mathbf{A}^+\|\|\mathbf{A}\| = \kappa(\mathbf{A}).$$

□

Definition 5.7. *If $\kappa(\mathbf{A})$ is small, $\mathbf{A}$ is said to be well-conditioned; if $\kappa(\mathbf{A})$ is large, $\mathbf{A}$ is said to be ill-conditioned.*

Some other immediate properties of condition numbers:

- Let $\mathbf{P}$ be a permutation matrix, then $\kappa(\mathbf{P}) = 1$. This is true in particular for the identity matrix $\mathbf{I}$, i.e., $\kappa(\mathbf{I}) = 1$.
- Let $\mathbf{A} \in \mathbb{C}^{m\times n}$ and $\lambda \in \mathbb{C}$, then $\kappa(\lambda\,\mathbf{A}) = \kappa(\mathbf{A})$.
- Let $\mathbf{D} \in \mathbb{C}^{m\times n}$ be a diagonal matrix, then

$$\kappa_1(\mathbf{D}) = \kappa_2(\mathbf{D}) = \kappa_\infty(\mathbf{D}) = \frac{\max|d_{ii}|}{\min|d_{ii}|},$$

  where d_{ii}, $i \in \{1, \ldots, \min\{m, n\}\}$ are the diagonal entries of $\mathbf{D}$.

Therefore, the condition number of a matrix $\mathbf{A}$ offers a better measure of closeness to singularity than the determinant of $\mathbf{A}$.

5.1 Condition number and systems of linear equations

Lemma 5.8. *Let $\hat{\mathbf{x}}$ be an approximation of the solution of the system of linear equations $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, where $\mathbf{A} \in \mathbb{C}^{n\times n}$ is nonsingular, $\mathbf{b}, \mathbf{x} \in \mathbb{C}^n$. Let $\mathbf{r} = \mathbf{A}\,\hat{\mathbf{x}} - \mathbf{b}$ be the residual. Then, for any vector norm $\|\cdot\|_v$ and its induced matrix norm $\|\cdot\|_m$, the absolute error is bounded by*

$$\|\hat{\mathbf{x}} - \mathbf{x}\|_v \leq \|\mathbf{A}^{-1}\|_m\|\mathbf{r}\|_v.$$

Furthermore, if $\mathbf{x} \neq \mathbf{0}$ and $\mathbf{r} \neq \mathbf{0}$, then the relative error is bounded by

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|_v}{\|\mathbf{x}\|_v} \leq \kappa(\mathbf{A})\frac{\|\mathbf{r}\|_v}{\|\mathbf{b}\|_v},$$

where κ is the condition number relative to the norm $\|\cdot\|_v$.

Proof. Since $\mathbf{A}$ is nonsingular,

$$\mathbf{r} = \mathbf{A}\,\hat{\mathbf{x}} - \mathbf{b} = \mathbf{A}(\hat{\mathbf{x}} - \mathbf{x}) \Longrightarrow \hat{\mathbf{x}} - \mathbf{x} = \mathbf{A}^{-1}\mathbf{r} \Longrightarrow \|\hat{\mathbf{x}} - \mathbf{x}\|_v = \|\mathbf{A}^{-1}\mathbf{r}\|_v \leq \|\mathbf{A}^{-1}\|_m\|\mathbf{r}\|_v.$$

On the other hand,

$$\mathbf{b} = \mathbf{A}\,\mathbf{x} \Longrightarrow \|\mathbf{b}\|_v = \|\mathbf{A}\,\mathbf{x}\|_v \leq \|\mathbf{A}\|_m\|\mathbf{x}\|_v \Longleftrightarrow \frac{1}{\|\mathbf{x}\|_v} \leq \frac{\|\mathbf{A}\|_m}{\|\mathbf{b}\|_v};$$

therefore,

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|_v}{\|\mathbf{x}\|_v} \leq \frac{\|\mathbf{A}^{-1}\|_m\|\mathbf{r}\|_v}{\|\mathbf{x}\|_v} = \frac{\|\mathbf{A}^{-1}\|_m\|\mathbf{A}\|_m\|\mathbf{r}\|_v}{\|\mathbf{b}\|_v} = \kappa(\mathbf{A})\frac{\|\mathbf{r}\|_v}{\|\mathbf{b}\|_v}. \tag{5}$$

□

Remark: we can understand (5) from a perturbative point of view. Indeed, if we perturb the right-hand side of $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ by $\Delta\mathbf{b}$, then there is a $\Delta\mathbf{x}$, such that $\mathbf{A}(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b} + \Delta\mathbf{b}$. Hence, $\Delta\mathbf{x} = \hat{\mathbf{x}} - \mathbf{x}$ is the absolute error of $\mathbf{x}$, and $\Delta\mathbf{b} = \mathbf{A}\,\hat{\mathbf{x}} - \mathbf{b} = \mathbf{A}\,\Delta\mathbf{x} = \mathbf{r}$ is precisely the residual, so (5) becomes

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \kappa(\mathbf{A}) \frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|},$$

i.e., the relative error of $\mathbf{x}$ is bounded by the relative error of $\mathbf{b}$ times the condition number of $\mathbf{A}$. In fact, the condition number of $\mathbf{A}$ is the supremum of the relative error magnification factors, as the following theorem shows.

Lemma 5.9. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ be a square nonsingular matrix, then*

$$\kappa(\mathbf{A}) = \sup_{\substack{\mathbf{b}\neq\mathbf{0} \\ \Delta\mathbf{b}\neq\mathbf{0}}} \frac{\dfrac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|}}{\dfrac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}}, \tag{6}$$

where $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$, $\Delta\mathbf{x} = \mathbf{A}^{-1}\Delta\mathbf{b}$.

Proof.

$$\sup_{\substack{\mathbf{b}\neq\mathbf{0} \\ \Delta\mathbf{b}\neq\mathbf{0}}} \frac{\dfrac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|}}{\dfrac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}} = \sup_{\substack{\mathbf{b}\neq\mathbf{0} \\ \Delta\mathbf{b}\neq\mathbf{0}}} \left(\frac{\|\Delta\mathbf{x}\|}{\|\Delta\mathbf{b}\|} \cdot \frac{\|\mathbf{b}\|}{\|\mathbf{x}\|} \right) = \sup_{\substack{\mathbf{b}\neq\mathbf{0} \\ \Delta\mathbf{b}\neq\mathbf{0}}} \left(\frac{\|\mathbf{A}^{-1}\Delta\mathbf{b}\|}{\|\Delta\mathbf{b}\|} \cdot \frac{\|\mathbf{b}\|}{\|\mathbf{A}^{-1}\mathbf{b}\|} \right)$$

$$= \left(\sup_{\Delta\mathbf{b}\neq\mathbf{0}} \frac{\|\mathbf{A}^{-1}\Delta\mathbf{b}\|}{\|\Delta\mathbf{b}\|} \right) \left(\sup_{\mathbf{b}\neq\mathbf{0}} \frac{\|\mathbf{b}\|}{\|\mathbf{A}^{-1}\mathbf{b}\|} \right) = \|\mathbf{A}^{-1}\| \sup_{\mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{A}\,\mathbf{x}\|}{\|\mathbf{x}\|} = \|\mathbf{A}^{-1}\|\|\mathbf{A}\| = \kappa(\mathbf{A}).$$

□

Example. Let us consider the system of linear equations $\mathbf{A}\,\mathbf{x} = \mathbf{b}$:

$$\begin{pmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 4.1 \\ 9.7 \end{pmatrix} \Longrightarrow \mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

There is apparently nothing remarkable about $\mathbf{A}$. It is nonsingular, with $\det(\mathbf{A}) = 0.1 \neq 0$, and its inverse can be computed immediately, too:

$$\mathbf{A}^{-1} = \begin{pmatrix} -66 & 28 \\ 97 & -41 \end{pmatrix}.$$

Let us introduce now a small perturbation $\Delta\mathbf{b} = (0.01, 0)^T$ in the right-hand side, so $\hat{\mathbf{b}} = (4.11, 9.7)^T$; then,

$$\begin{pmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{pmatrix} \cdot \begin{pmatrix} \hat{x}_1 \\ \hat{x}_2 \end{pmatrix} = \begin{pmatrix} 4.11 \\ 9.7 \end{pmatrix} \Longrightarrow \tilde{\mathbf{x}} = \begin{pmatrix} 0.34 \\ 0.97 \end{pmatrix}$$

As we can see, the new solution $\hat{\mathbf{x}}$ has changed rather drastically. In order to quantify this, we will use for instance the 1-norm:

$$\|\Delta\mathbf{b}\|_1 = \left\| \begin{pmatrix} 0.01 \\ 0 \end{pmatrix} \right\|_1 = 0.01, \qquad \|\Delta\mathbf{x}\|_1 = \left\| \begin{pmatrix} 0.34 - 1 \\ 0.97 - 0 \end{pmatrix} \right\|_1 = 0.97 + 0.66 = 1.63.$$

Hence, bearing in mind that $\|\mathbf{b}\|_1 = 4.1 + 9.7 = 13.8$, and $\|\mathbf{x}\|_1 = 1$, the relative errors are

$$\frac{\|\Delta\mathbf{b}\|_1}{\|\mathbf{b}\|_1} = \frac{0.01}{13.8} = 7.24\ldots \cdot 10^{-4}, \qquad \frac{\|\Delta\mathbf{x}\|_1}{\|\mathbf{x}\|_1} = \frac{1.63}{1} = 1.63,$$

i.e., the ratio between both relative errors is $1.63/7.24\ldots \cdot 10^{-4} = 2249.4$. The explanation to this apparently pathological behavior is given by the condition number of $\mathbf{A}$. Indeed, from (6), $\kappa_1(\mathbf{A}) \geq 2249.4$. Let us compute exactly $\kappa_1(\mathbf{A})$:

$$\kappa_1(\mathbf{A}) = \|\mathbf{A}\|_1\|\mathbf{A}^{-1}\|_1 = \left\| \begin{pmatrix} 4.1 & 2.8 \\ 9.7 & 6.6 \end{pmatrix} \right\|_1 \left\| \begin{pmatrix} -66 & 28 \\ 97 & -41 \end{pmatrix} \right\|_1$$
$$= \max\{13.8, 9.4\} \cdot \max\{163, 69\} = 13.8 \cdot 163 = 2249.4.$$

Therefore, the maximum magnification factor is achieved for this choice of $\mathbf{b}$ and $\Delta\mathbf{b}$.

Theorem 5.10. *Let* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$*, and let* $\Delta\mathbf{A}$*,* $\Delta\mathbf{x}$ *and* $\Delta\mathbf{b}$ *be perturbations of, respectively,* $\mathbf{A}$*,* $\mathbf{x}$*,* $\mathbf{b}$*, such that* $(\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{b}+\Delta\mathbf{b}$*. If* $\mathbf{A}$ *is nonsingular, and* $\kappa(\mathbf{A})\|\Delta\mathbf{A}\|/\|\mathbf{A}\| < 1$ *in some induced matrix norm, then*

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(\mathbf{A})}{1-\kappa(\mathbf{A})\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}}\left(\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}+\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\right).$$

Proof. On the one hand, $\mathbf{A}(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{b}+\Delta\mathbf{b}-\Delta\mathbf{A}(\mathbf{x}+\Delta\mathbf{x})$; on the other hand, $\mathbf{A}\,\mathbf{x} = \mathbf{b}$. Subtracting both expressions,

$$\mathbf{A}\,\Delta\mathbf{x} = \Delta\mathbf{b}-\Delta\mathbf{A}(\mathbf{x}+\Delta\mathbf{x}) \Longleftrightarrow \Delta\mathbf{x} = \mathbf{A}^{-1}(\Delta\mathbf{b}-\Delta\mathbf{A}(\mathbf{x}+\Delta\mathbf{x})).$$

Therefore, taking norms,

$$\|\Delta\mathbf{x}\| = \|\mathbf{A}^{-1}(\Delta\mathbf{b}-\Delta\mathbf{A}(\mathbf{x}+\Delta\mathbf{x}))\| \leq \|\mathbf{A}^{-1}\|(\|\Delta\mathbf{b}\|+\|\Delta\mathbf{A}\|(\|\mathbf{x}\|+\|\Delta\mathbf{x}\|));$$

hence,

$$(1-\|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|)\|\Delta\mathbf{x}\| \leq \|\mathbf{A}^{-1}\|(\|\Delta\mathbf{b}\|+\|\Delta\mathbf{A}\|\|\mathbf{x}\|).$$

Dividing by $\|\mathbf{x}\|$ and using $\|\mathbf{b}\| = \|\mathbf{A}\,\mathbf{x}\| \leq \|\mathbf{A}\|\|\mathbf{x}\| \Longrightarrow \|\mathbf{x}\|^{-1} \leq \|\mathbf{A}\|/\|\mathbf{b}\|$,

$$\begin{aligned}(1-\|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\|)\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} &\leq \|\mathbf{A}^{-1}\|\left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{x}\|}+\|\Delta\mathbf{A}\|\right) \leq \|\mathbf{A}^{-1}\|\left(\frac{\|\mathbf{A}\|\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}+\|\Delta\mathbf{A}\|\right)\\ &= \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}+\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right) = \kappa(\mathbf{A})\left(\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}+\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}\right).\end{aligned}$$

Finally, bearing in mind that $\|\mathbf{A}^{-1}\|\|\Delta\mathbf{A}\| = \|\mathbf{A}^{-1}\|\|\mathbf{A}\|\|\Delta\mathbf{A}\|/\|\mathbf{A}\| = \kappa(\mathbf{A})\|\Delta\mathbf{A}\|/\|\mathbf{A}\|$, and that this quantity is smaller than one by hypothesis, we conclude that

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(\mathbf{A})}{1-\kappa(\mathbf{A})\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}}\left(\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}+\frac{\|\Delta\mathbf{b}\|}{\|\mathbf{b}\|}\right).$$

□

If $\kappa(\mathbf{A})\|\Delta\mathbf{A}\|/\|\mathbf{A}\| \ll 1$, we have that both the relative error of $\mathbf{b}$ and the relative error of $\mathbf{A}$ are approximately magnified by $\kappa(\mathbf{A})$.

Corollary 5.11. *Let* $\mathbf{A}\,\mathbf{x} = \mathbf{b}$*, and let* $\Delta\mathbf{A}$ *and* $\Delta\mathbf{x}$ *be perturbations of, respectively,* $\mathbf{A}$ *and* $\mathbf{x}$*, such that* $(\mathbf{A}+\Delta\mathbf{A})(\mathbf{x}+\Delta\mathbf{x}) = \mathbf{b}$*. If* $\mathbf{A}$ *is nonsingular, and* $\kappa(\mathbf{A})\|\Delta\mathbf{A}\|/\|\mathbf{A}\| < 1$ *in some induced matrix norm, then*

$$\frac{\|\Delta\mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\kappa(\mathbf{A})}{1-\kappa(\mathbf{A})\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}}\frac{\|\Delta\mathbf{A}\|}{\|\mathbf{A}\|}.$$

6 Iterative methods

Definition 6.1. *A matrix* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *is convergent, if and only if* $\lim_{k\to\infty}\mathbf{A}^k = \mathbf{0}_{n\times n}$.

Theorem 6.2. *A matrix* $\mathbf{A} \in \mathbb{C}^{n\times n}$ *is convergent, if and only if* $\rho(\mathbf{A}) < 1$.

Proof. Suppose that $\mathbf{A}$ is convergent. Let $\mathbf{u}$ be an eigenvector associated to the eigenvalue λ of $\mathbf{A}$, then

$$\lambda\,\mathbf{u} = \mathbf{A}\,\mathbf{u} \Longrightarrow \lambda^k\mathbf{u} = \mathbf{A}^k\mathbf{u} \Longrightarrow \lim_{k\to\infty}\lambda^k\mathbf{u} = \lim_{k\to\infty}\mathbf{A}^k\mathbf{u} = \mathbf{0} \Longrightarrow \lim_{k\to\infty}\lambda^k = 0 \Longrightarrow |\lambda| < 1.$$

Since this is true for any eigenvalue of $\mathbf{A}$, it follows that $\rho(\mathbf{A}) < 1$.

Suppose now that $\rho(\mathbf{A}) < 1$. From the well-known Jordan canonical form theorem, there is a nonsingular matrix $\mathbf{P} \in \mathbb{C}^{n\times n}$, and a block diagonal matrix $\mathbf{J} \in \mathbb{C}^{n\times n}$, such that $\mathbf{A} = \mathbf{P}\,\mathbf{J}\,\mathbf{P}^{-1}$. The Jordan matrix $\mathbf{J}$ is of the form

$$\mathbf{J} = \begin{pmatrix}\mathbf{J}_1 & & 0\\ & \ddots & \\ 0 & & \mathbf{J}_p\end{pmatrix},$$

and the blocks of $\mathbf{J}$ are of the form

$$\mathbf{J}_i = \begin{pmatrix}\lambda_i & 1 & & 0\\ & \lambda_i & \ddots & \\ & & \ddots & 1\\ 0 & & & \lambda_i\end{pmatrix},$$

where λ_i is an eigenvalue of $\mathbf{A}$. Therefore, $\mathbf{A}^k = \mathbf{P}\,\mathbf{J}^k\mathbf{P}^{-1}$, and $\mathbf{J}^k$ is also a block diagonal matrix, such that

$$\mathbf{J}^k = \begin{pmatrix} \mathbf{J}_1^k & & 0 \\ & \ddots & \\ 0 & & \mathbf{J}_p^k \end{pmatrix}.$$

It can be proved that, if $\mathbf{J}_i \in \mathbb{C}^{m_i \times m_i}$, then, for $k \geq m_i$,

$$\mathbf{J}_i^k = \begin{pmatrix} \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} & \binom{k}{2}\lambda_i^{k-2} & \cdots & \binom{k}{m_i-1}\lambda_i^{k-m_i+1} \\ & \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} & \cdots & \binom{k}{m_i-2}\lambda_i^{k-m_i+2} \\ & & \ddots & \ddots & \vdots \\ & & & \lambda_i^k & \binom{k}{1}\lambda_i^{k-1} \\ 0 & & & & \lambda_i^k \end{pmatrix}.$$

Therefore, if $\rho(\mathbf{A}) < 1$, then $|\lambda_i| < 1$, for all λ_i, so $\lim_{k\to\infty} \mathbf{J}_i^k = \mathbf{0}_{m_i \times m_i}$ and, hence, $\lim_{k\to\infty} \mathbf{J}^k = \mathbf{0}_{n\times n}$.

□

Corollary 6.3. *A matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is convergent, if $\|\mathbf{A}\| < 1$, for some matrix norm $\|\cdot\|$.*

In order to obtain a quick bound of the spectrum (i.e., the eigenvalues) of a square matrix without actually computing them, the following theorem can be very useful.

Theorem 6.4 (Gershgorin circle theorem I)**.** *Given a matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$, let $r_i = \sum_{j\neq i} |a_{ij}|$, $i \in \{1, \ldots, n\}$. Let $D(a_{ii}, r_i)$ be the closed disc in $\mathbb{C}$ (called the Gershgorin circ) centered at a_{ii} with radius r_i:*

$$D(a_{ii}, r_i) \equiv \{z \in \mathbb{C} \mid |z - a_{ii}| \leq r_i\}.$$

Then, every eigenvalue of $\mathbf{A}$ lies within at least one of the Gershgorin discs.

Proof. Let λ be an eigenvalue of $\mathbf{A}$. Let $\mathbf{u}$ be an eigenvector of $\mathbf{A}$ associated to λ, and normalized in such a way that one entry of $\mathbf{u}$, u_i, is equal to 1, and the other entries of $\mathbf{u}$ are of absolute value smaller than or equal to one, i.e., $u_i = 1$, $|u_j| = 1$, for $j \neq i$. Then,

$$\lambda\,\mathbf{u} = \mathbf{A}\,\mathbf{u} \Longrightarrow \lambda = \lambda u_i = \sum_{j=1}^{n} a_{ij}u_j = a_{ii}u_i + \sum_{j\neq i} a_{ij}u_j = a_{ii} + \sum_{j\neq i} a_{ij}u_j.$$

Therefore, applying the triangle inequality,

$$|\lambda - a_{ii}| = \left|\sum_{j\neq i} a_{ij}u_j\right| \leq \sum_{j\neq i} |a_{ij}||u_j| \leq \sum_{j\neq i} |a_{ij}| = r_i;$$

i.e., λ lies in the Gershgorin disc $D(a_{ii}, r_i)$.

□

Corollary 6.5. *A matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is diagonal, if and only if its Gershgorin discs coincide with its spectrum.*

Remark: one way to understand the Gershgorin circle theorem is that, if the off-diagonal entries of $\mathbf{A} \in \mathbb{C}^{n\times n}$ have small norms, then the eigenvalues of $\mathbf{A}$ will not be far from the diagonal entries of $\mathbf{A}$.

Corollary 6.6. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$; then, every eigenvalue of $\mathbf{A}$ lies within at least one of the Gershgorin discs corresponding to the columns of $\mathbf{A}$.*

Proof. Since $\mathbf{A}$ and $\mathbf{A}^T$ have the same spectrum, it is enough to apply Theorem 6.4 to $\mathbf{A}^T$. □

Definition 6.7. *A square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is strictly diagonally dominant, if and only if*

$$|a_{ii}| > \sum_{j\neq i} |a_{ij}|, \quad \forall i \in \{1, \ldots, n\}.$$

Unless otherwise stated, the expression "strictly diagonally dominant" is usually understood as "strictly diagonally dominant by rows". On the other hand, a square matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is strictly diagonally dominant by columns, if and only if

$$|a_{jj}| > \sum_{i\neq j} |a_{ij}|, \quad \forall j = 1, \ldots, n.$$

Corollary 6.8. *Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a real, symmetric matrix, and let λ be an eigenvalue of $\mathbf{A}$; then,*

$$\min_{1\le i\le n}\left\{a_{ii} - \sum_{j\neq i}|a_{ij}|\right\} \le \lambda \le \max_{1\le i\le n}\left\{a_{ii} + \sum_{j\neq i}|a_{ij}|\right\}.$$

Therefore, if $\mathbf{A}$ is strictly diagonal dominant, then it is nonsingular and positive definite.

Theorem 6.9 (Gershgorin circle theorem II)**.** *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$. If the union of k Gershgorin discs is disjoint from the union of the other $n-k$ Gershgorin discs, then the former and latter unions contain respectively k and $n-k$ eigenvalues of $\mathbf{A}$, counted with multiplicity.*

Proof. Let us define

$$\mathbf{B}(t) = (1-t)\mathbf{D} + t\,\mathbf{A}, \quad t \in [0,1],$$

where $\mathbf{D}$ is the diagonal matrix whose entries are equal to the diagonal entries of $\mathbf{A}$. Hence, the diagonal of $\mathbf{B}(t)$ is equal to the diagonal of $\mathbf{A}$, so the centers of the Gershgorin discs are the same, and the radii of the Gershgorin discs of $\mathbf{B}$ are t times the radii of the discs of $\mathbf{A}$. Therefore, the union of the corresponding k discs of $\mathbf{B}(t)$ is disjoint from the union of the remaining $n-k$ discs of $\mathbf{B}(t)$, for all $t \in [0,1]$. On the other hand, the discs are closed, so the distance between the two unions for $\mathbf{A}$ is $d > 0$, and the distance between the two unions for $B(t)$ is a nonincreasing function of t, so it is always at least d.

At $t = 0$, the theorem is trivially true: since $\mathbf{D} = \mathbf{B}(0)$ is a diagonal matrix, the discs are precisely eigenvalues of $\mathbf{D}$, so the k-disc and $n-k$-disc unions, being disjoint from each other, contain exactly k and $n-k$ eigenvalues, respectively. Let $\lambda(t)$ be an eigenvalue of $\mathbf{B}(t)$, such that $\lambda(0)$ lies in the k-disc union of $\mathbf{B}(0)$. Let $d(t)$ be the distance of $\lambda(t)$ to the $n-k$-disc union of $\mathbf{B}(t)$. Since $\lambda(t)$ is a **continuous** function of t, so is $d(t)$. When $t = 0$, we have trivially $d(0) \ge d > 0$. Suppose that, when $t = 1$, $\lambda(1)$ is in the $n-k$-disc union of $\mathbf{B}(1)$ (i.e., has "abandonned" the k-disc union); then, $d(1) = 0$. Therefore, since $d(t)$ is continuous, there is some $t_0 \in (0,1)$, such that $d(t_0) \in (0,d)$, which implies that $\lambda(t_0)$ is at neither of the unions, i.e., $\lambda(t_0)$ lies outside all the Gershgorin discs, which is a contradiction. Therefore, we conclude that $\lambda(1)$ lies in the union of the k discs. In exactly the same way, we conclude that, if $\lambda(0)$ lies in the $n-k$-disc union of $\mathbf{B}(0)$, then $\lambda(1)$ lies in the $n-k$-disc union of $\mathbf{B}(1)$. Summarizing, the k-disc and $n-k$-disc unions of $\mathbf{B}(0)$ have the same number of eigenvalues as the k-disc and $n-k$-disc unions of $\mathbf{A} = \mathbf{B}(1)$, i.e., k and $n-k$, respectively.

□

Remark: the key idea is that the eigenvalues of $\mathbf{B}(t)$, regarded as functions of t, change continuously with respect to t. This is indeed a consequence of a theorem that states that the roots of a complex polynomial change continuously when regarded as functions of the coefficients of the polynomial.

Corollary 6.10. *Let $\mathbf{A} \in \mathbb{C}^{n\times n}$. A disjoint Gershgorin disc of $\mathbf{A}$ contains exactly one eigenvalue of $\mathbf{A}$; moreover, this eigenvalue is real if the matrix $\mathbf{A}$ is real.*

Proof. If a Gershgorin disc D of $\mathbf{A}$ is disjoint, we define a 1-disc union formed only by it, and another $n-1$-disc union, formed by the other $n-1$ Gershgorin discs of $\mathbf{A}$. Then, the unions have, respectively, 1 and $n-1$ eigenvalues.

On the other hand, if $\mathbf{A}$ is real, it has a real characteristic polynomial. Therefore, if λ is an eigenvalue of $\mathbf{A}$, so is $\bar{\lambda}$. Suppose that $\lambda \in D$, where D is a disjoint Gershgorin disc of $\mathbf{D}$. Since $\mathbf{A}$ is real, the center of D is real, so the circle is symmetric to the real line; therefore, $\bar{\lambda}$ lies inside D, too. However, since there is one single eigenvalue inside D, it follows that $\lambda = \bar{\lambda}$, i.e., λ is real.

□

Remark: since $\mathbf{A}$ and $\mathbf{P}^{-1}\mathbf{A}\,\mathbf{P}$ have the same spectrum for any nonsingular matrix $\mathbf{P}$, it can be useful to apply the Gerschgorin theorem to $\mathbf{P}^{-1}\mathbf{A}\,\mathbf{P}$, for some conveniently chosen $\mathbf{P}$.

6.1 Formulation of the iterative methods

Let us consider the system of linear equations $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, where $\mathbf{A} \in \mathbb{C}^{n\times n}$ is a square nonsingular matrix, $\mathbf{b}, \mathbf{x} \in \mathbb{C}^n$. The idea of iterative methods is decomposing $\mathbf{A} = \mathbf{M} + \mathbf{N}$, where $\mathbf{M}$ is a square nonsingular matrix that is easier to invert than $\mathbf{A}$. Then,

$$\mathbf{A}\,\mathbf{x} = \mathbf{b} \iff (\mathbf{M}+\mathbf{N})\mathbf{x} = \mathbf{b} \iff \mathbf{x} = \mathbf{M}^{-1}(\mathbf{b} - \mathbf{N}\,\mathbf{x}); \tag{7}$$

hence, $\mathbf{x}$ solves $\mathbf{A}\,x = \mathbf{b}$, if and only if it solves $\mathbf{x} = \mathbf{M}^{-1}(\mathbf{b} - \mathbf{N}\,\mathbf{x})$. Therefore, if the following iterative scheme converges:

$$\mathbf{x}^{(k+1)} = \mathbf{M}^{-1}\left(\mathbf{b} - \mathbf{N}\,\mathbf{x}^{(k)}\right), \tag{8}$$

then,

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x},$$

where $\mathbf{x}$ is precisely the solution vector. In order to give a stopping criterion, we define the error and the residual at the kth iteration as

$$\mathbf{e}^{(k)} = \mathbf{x}^{(k)} - \mathbf{x}, \qquad \mathbf{r}^{(k)} = \mathbf{A}\,\mathbf{x}^{(k)} - \mathbf{b} = \mathbf{A}(\mathbf{x}^{(k)} - \mathbf{x}) = \mathbf{A}\,\mathbf{e}^{(k)},$$

respectively. Ideally, we would stop the iterative scheme (8), when $\|\mathbf{e}^{(k)}\| < \varepsilon$, in some norm $\|\cdot\|$, for $0 < \varepsilon \ll 1$. However, since $\mathbf{x}$ is unknown, we rather stop it when $\|\mathbf{r}^{(k)}\| < \varepsilon$. Indeed, $\|\mathbf{e}^{(k)}\| = \|\mathbf{A}^{-1}\mathbf{r}^{(k)}\| \leq \|\mathbf{A}^{-1}\|\|\mathbf{r}^{(k)}\|$, so $\|\mathbf{e}^{(k)}\|$ is bound by a multiple of $\|\mathbf{r}^{(k)}\|$. Let us mention that another possible stopping criterion is $\|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| < \varepsilon$.

In order to study the convergence of (8), we rewrite it as

$$\mathbf{x}^{(k+1)} = \mathbf{M}^{-1}\left(\mathbf{A}\,\mathbf{x}^{(k)} - \mathbf{r}^{(k)} - \mathbf{N}\,\mathbf{x}^{(k)}\right) = \mathbf{M}^{-1}\left(\mathbf{M}\,\mathbf{x}^{(k)} - \mathbf{r}^{(k)}\right) = \mathbf{x}^{(k)} - \mathbf{M}^{-1}\mathbf{r}^{(k)};$$

then, if we subtract $\mathbf{x}$ on both sides,

$$\mathbf{x}^{(k+1)} - \mathbf{x} = \mathbf{x}^{(k)} - \mathbf{x} - \mathbf{M}^{-1}\mathbf{r}^{(k)} \iff \mathbf{e}^{(k+1)} = \mathbf{e}^{(k)} - \mathbf{M}^{-1}\mathbf{r}^{(k)} \iff \mathbf{e}^{(k+1)} = (\mathbf{I} - \mathbf{M}^{-1}\mathbf{A})\mathbf{e}^{(k)}.$$

Hence, defining the iteration matrix $\mathbf{R}$ as

$$\mathbf{R} = \mathbf{I} - \mathbf{M}^{-1}\mathbf{A}, \tag{9}$$

we conclude that

$$\mathbf{e}^{(k)} = \mathbf{R}\,\mathbf{e}^{(k-1)} = \mathbf{R}^2\mathbf{e}^{(k-2)} = \ldots = \mathbf{R}^k\mathbf{e}^{(0)},$$

where $\mathbf{e}^{(0)} = \mathbf{x}^{(0)} - \mathbf{x}$ is the error in the initial vector.

Theorem 6.11. *The iterative scheme* (8) *converges for any initial data, if and only if its iteration matrix is a convergent matrix.*

Proof. If (8) is convergent, then

$$\lim_{k\to\infty} \mathbf{x}^{(k)} = \mathbf{x} \iff \lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}_n \iff \lim_{k\to\infty} \mathbf{R}^k\mathbf{e}^{(0)} = \mathbf{0}_n.$$

Since this must be true for any initial data, it implies that $\lim_{k\to\infty} \mathbf{R}^k = \mathbf{0}_{n\times n}$, i.e., $\mathbf{R}$ is convergent. Conversely, if $\mathbf{R}$ is convergent, then $\lim_{k\to\infty} \mathbf{R}^k = \mathbf{0}_{n\times n}$, so $\lim_{k\to\infty} \mathbf{e}^{(k)} = \mathbf{0}_n$, i.e., (8) is convergent. □

Remark: the spectral radius of $\mathbf{R}$ determines the convergence speed of the iterative speed. Indeed, if we take as $\mathbf{e}^{(0)}$ an eigenvector of $\mathbf{R}$ asociated to the dominating eigenvalue λ of $\mathbf{R}$, i.e., the eigenvalue of $\mathbf{R}$ with the largest modulus, then

$$\mathbf{R}\,\mathbf{e}^{(0)} = \rho(\mathbf{R})\mathbf{e}^{(0)} \Longrightarrow \mathbf{e}^{(k)} = (\rho(\mathbf{R}))^k\mathbf{e}^{(0)} \Longrightarrow \|\mathbf{e}^{(k)}\| = (\rho(\mathbf{R}))^k\|\mathbf{e}^{(0)}\|.$$

6.1.1 Jacobi method

Let us decompose $\mathbf{A}$ as $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$, where $\mathbf{L}$ is a lower-triangular matrix, $\mathbf{D}$ is a diagonal matrix and $\mathbf{U}$ is an upper-triangular matrix. In the Jacobi method, or Jacobi iterative method, we take $\mathbf{M} = \mathbf{D}$, $\mathbf{N} = \mathbf{L} + \mathbf{U}$ in the decomposition $\mathbf{A} = \mathbf{M} + \mathbf{N}$, so (7) becomes

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

Therefore the corresponding iterative scheme (i.e., the Jacobi iteration) is

$$\mathbf{x}^{(k+1)} = \mathbf{D}^{-1}\left(\mathbf{b} - (\mathbf{L} + \mathbf{U})\mathbf{x}^{(k)}\right),$$

or, equivalently,

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\right), \quad i = 1, \ldots, n. \tag{10}$$

Remark that $\mathbf{A}$ cannot have diagonal entries iqual to zero.

Theorem 6.12. *If $\mathbf{A} \in \mathbb{C}^{n\times n}$ is strictly diagonally dominant (either by rows or columns), then the Jacobi iterative method applied to $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ is convergent.*

Proof. We have to prove that the spectral radius of the iteration matrix $\mathbf{R}$, as defined in (9), is strictly smaller than one. In this case,

$$\mathbf{R} = \mathbf{I} - \mathbf{M}^{-1}\mathbf{A} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A} = \mathbf{D}^{-1}(\mathbf{D} - \mathbf{A}) = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}).$$

Hece, if $\mathbf{A}$ is strictly diagonally dominant by rows,

$$\rho(\mathbf{R}) \leq \|\mathbf{R}\|_\infty = \|\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\|_\infty = \max_{1\leq i\leq n} \sum_{\substack{j=1\\ j\neq i}}^{n} \left|\frac{a_{ij}}{a_{ii}}\right| = \max_{1\leq i\leq n} \left\{ \frac{1}{|a_{ii}|} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \right\} < 1.$$

On the other hand, if $\mathbf{A}$ is strictly diagonally dominant by columns,

$$\rho(\mathbf{R}) \leq \|\mathbf{R}\|_1 = \|\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U})\|_1 = \max_{1\leq j\leq n} \sum_{\substack{i=1\\ i\neq j}}^{n} \left|\frac{a_{ij}}{a_{jj}}\right| = \max_{1\leq j\leq n} \left\{ \frac{1}{|a_{jj}|} \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{ij}| \right\} < 1.$$

□

6.1.2 Gauss-Seidel method

Let us decompose $\mathbf{A}$ as $\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$, where $\mathbf{L}$ is a lower-triangular matrix, $\mathbf{D}$ is a diagonal matrix and $\mathbf{U}$ is an upper-triangular matrix. In the Gauss-Seidel method, or Gauss-Seidel iterative method, we take $\mathbf{M} = \mathbf{L} + \mathbf{D}$, $\mathbf{N} = \mathbf{U}$ in the decomposition $\mathbf{A} = \mathbf{M} + \mathbf{N}$, so (7) becomes

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

Therefore, the corresponding iterative scheme (i.e., the Gauss-Seidel iteration) is

$$\mathbf{x}^{(k+1)} = (\mathbf{L} + \mathbf{D})^{-1}\left(\mathbf{b} - \mathbf{U}\,\mathbf{x}^{(k)}\right),$$

or, equivalently,

$$x_i^{(k+1)} = \frac{1}{a_{ii}}\left(b_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\right), \quad i = 1, \ldots, n. \tag{11}$$

Theorem 6.13. *If $\mathbf{A} \in \mathbb{C}^{n\times n}$ is strictly diagonally dominant (either by rows or columns), then the Gauss-Seidel iterative method applied to $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ is convergent.*

Proof. We have to prove that the spectral radius of the iteration matrix $\mathbf{R}$, as defined in (9), is strictly smaller than one. In this case,

$$\mathbf{R} = \mathbf{I} - \mathbf{M}^{-1}\mathbf{A} = \mathbf{I} - (\mathbf{L} + \mathbf{D})^{-1}\mathbf{A} = (\mathbf{L} + \mathbf{D})^{-1}(\mathbf{L} + \mathbf{D} - \mathbf{A}) = -(\mathbf{L} + \mathbf{D})^{-1}\mathbf{U}.$$

Let us suppose first that $\mathbf{A}$ is strictly diagonally dominant by rows. Since $\rho(\mathbf{R}) \leq \|\mathbf{R}\|_\infty$, it is enough to prove that $\|\mathbf{R}\|_\infty < 1$. Let us take any $\mathbf{x} \in \mathbb{C}^n$, such that $\|\mathbf{x}\|_\infty = 1$, and define $\mathbf{y} = \mathbf{R}\,\mathbf{x} = -(\mathbf{L}+\mathbf{D})^{-1}\mathbf{U}\,\mathbf{x}$. Then, $(\mathbf{L}+\mathbf{D})\mathbf{y} = -\mathbf{U}\,\mathbf{x}$. We select the kth equation, where $k \in \{1, \ldots, n\}$ is such that $\|\mathbf{y}\|_\infty = |y_k|$:

$$a_{kk}y_k + \sum_{j=1}^{k-1} a_{kj}y_j = -\sum_{j=k+1}^{n} a_{kj}x_j. \tag{12}$$

On the one hand, the modulus of the left-hand side of (12) is lower-bounded by

$$\left|a_{kk}y_k + \sum_{j=1}^{k-1} a_{kj}y_j\right| \geq |a_{kk}||y_k| - \sum_{j=1}^{k-1} |a_{kj}||y_j| \geq \left(|a_{kk}| - \sum_{j=1}^{k-1} |a_{kj}|\right)\|\mathbf{y}\|_\infty, \tag{13}$$

where we have used that $|z+w| \geq |z| - |w|$, for all $z, w \in \mathbb{C}$. On the other hand, the modulus of the right-hand side of (12) is upper-bounded by

$$\left| -\sum_{j=k+1}^{n} a_{kj} x_j \right| \leq \sum_{j=k+1}^{n} |a_{kj}||x_j| \leq \left(\sum_{j=k+1}^{n} |a_{kj}| \right) \|\mathbf{x}\|_\infty = \sum_{j=k+1}^{n} |a_{kj}|. \tag{14}$$

Hence, putting (13) and (14) together,

$$\left(|a_{kk}| - \sum_{j=1}^{k-1} |a_{kj}| \right) \|\mathbf{y}\|_\infty \leq \sum_{j=k+1}^{n} |a_{kj}| \Longrightarrow \|\mathbf{y}\|_\infty \leq \frac{\displaystyle\sum_{j=k+1}^{n} |a_{kj}|}{|a_{kk}| - \displaystyle\sum_{j=1}^{k-1} |a_{kj}|} \leq \max_{1 \leq i \leq n} \frac{\displaystyle\sum_{j=i+1}^{n} |a_{ij}|}{|a_{ii}| - \displaystyle\sum_{j=1}^{i-1} |a_{ij}|} < 1,$$

where we have used that $\mathbf{A}$ is strictly diagonally dominant by rows:

$$|a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}| - \sum_{j=i+1}^{n} |a_{ij}| > 0 \Longrightarrow |a_{ii}| - \sum_{j=1}^{i-1} |a_{ij}| > \sum_{j=i+1}^{n} |a_{ij}| > 0, \quad \forall i \in \{1, \ldots, n\}.$$

Therefore,

$$\|\mathbf{R}\|_\infty = \sup_{\|\mathbf{x}\|_\infty = 1} \|\mathbf{R}\,\mathbf{x}\|_\infty = \sup_{\|\mathbf{x}\|_\infty = 1} \|\mathbf{y}(\mathbf{x})\|_\infty \leq \max_{1 \leq i \leq n} \frac{\displaystyle\sum_{j=i+1}^{n} |a_{ij}|}{|a_{ii}| - \displaystyle\sum_{j=1}^{i-1} |a_{ij}|} < 1.$$

Let us suppose now that $\mathbf{A}$ is strictly diagonally dominant by columns. The proof is very similar, so we omit many details. Since $\rho(\mathbf{R}) = \rho(\mathbf{R}^*) \leq \|\mathbf{R}^*\|_\infty$, it is enough to prove that $\|\mathbf{R}^*\|_\infty < 1$. Let us take any $\mathbf{x} \in \mathbb{C}^n$, such that $\|\mathbf{x}\|_\infty = 1$, and define $\mathbf{y} = \mathbf{R}^*\,\mathbf{x} = -(\mathbf{L}^* + \mathbf{D}^*)^{-1}\mathbf{U}^*\mathbf{x}$. Then, $(\mathbf{L}^* + \mathbf{D}^*)\mathbf{y} = -\mathbf{U}^*\mathbf{x}$. We select the kth equation, where $k \in \{1, \ldots, n\}$ is such that $\|\mathbf{y}\|_\infty = |y_k|$:

$$\overline{a_{kk}} y_k + \sum_{i=k+1}^{n} \overline{a_{ik}} y_i = -\sum_{i=1}^{k-1} \overline{a_{ik}} x_i.$$

Then, reasoning exactly as above, we get

$$\left(|a_{kk}| - \sum_{i=k+1}^{n} |a_{ik}| \right) \|\mathbf{y}\|_\infty \leq \sum_{i=1}^{k-1} |a_{ik}| \Longrightarrow \|\mathbf{y}\|_\infty \leq \frac{\displaystyle\sum_{i=1}^{k-1} |a_{ik}|}{|a_{kk}| - \displaystyle\sum_{i=k+1}^{n} |a_{ik}|} \leq \max_{1 \leq j \leq n} \frac{\displaystyle\sum_{i=1}^{j-1} |a_{ij}|}{|a_{jj}| - \displaystyle\sum_{i=j+1}^{n} |a_{ij}|} < 1,$$

where we have used that $\mathbf{A}$ is strictly diagonally dominant by rows. This enables us to conclude that $\|\mathbf{R}^*\|_\infty < 1$.

□

6.1.3 More on convergence of the Jacobi and Gauss-Seidel methods

The condition that, if $\mathbf{A}$ is strictly diagonally dominant, then the Jacobi method and the Gauss-Seidel method are convergent, is sufficient, but not necessary. For instance, both methods also converge if $\mathbf{A}$ is irreducibly diagonally dominant.

Definition 6.14. *A square matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ is (weakly) diagonally dominant, if and only if*

$$|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|, \quad \forall i \in \{1, \ldots, n\}.$$

As in the case of strictly diagonally dominant matrices, unless otherwise stated, by rows is understood. On the other hand, a square matrix $\mathbf{A} \in \mathbb{C}^{n \times n}$ is (weakly) diagonally dominant by columns, if and only if

$$|a_{jj}| \geq \sum_{i \neq j} |a_{ij}|, \quad \forall j \in \{1, \ldots, n\}.$$

Definition 6.15. *A matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is irreducible if and only if there is no permutation matrix ¶, such that*

$$\mathbf{P}\,\mathbf{A}\,\mathbf{P}^T = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{0} & \mathbf{A}_{22} \end{pmatrix},$$

where $\mathbf{A}_{11} \in \mathbb{C}^{k\times k}$ and $\mathbf{A}_{11} \in \mathbb{C}^{(n-k)\times(n-k)}$.

Equivalently, let $G(\mathbf{A})$ be the graph with vertices $\{1,\ldots,n\}$, and such that $i \to j$ is an arc of $G(\mathbf{A})$, if and only if $a_{ij} \neq 0$. Then, $\mathbf{A}$ is irreducible if and only if $G(\mathbf{A})$ is a connected graph, i.e., such that there is a path from any point to any other point of the graph.

Definition 6.16. *A matrix $\mathbf{A} \in \mathbb{C}^{n\times n}$ is irreducibly diagonally dominant (by rows) if and only if it is irreducible and diagonally dominant, and, the strict inequality happens at least for one row:*

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

In the case of irreducibly diagonally dominant by columns, the strict inequality happens at least for one column:

$$|a_{jj}| > \sum_{i\neq j} |a_{ij}|.$$

Theorem 6.17. *If $\mathbf{A} \in \mathbb{C}^{n\times n}$ is irreducibly diagonally dominant (either by rows or columns), then the Jacobi method applied to $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ is convergent.*

Theorem 6.18. *If $\mathbf{A} \in \mathbb{C}^{n\times n}$ is irreducibly diagonally dominant (either by rows or columns), then the Gauss-Seidel method applied to $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ is convergent.*

Theorem 6.19. *If $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $2\mathbf{D} - \mathbf{A}$ are symmetric and positive definite, with $\mathbf{D} \in \mathbb{R}^{n\times n}$ being the diagonal matrix formed by the diagonal entries of $\mathbf{A}$, then the Jacobi method applied to $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ is convergent.*

Remark: $2\mathbf{D} - \mathbf{A}$ has the same entries as $\mathbf{A}$, but with opposite sign, whenever $i \neq j$.

Theorem 6.20. *If $\mathbf{A} \in \mathbb{R}^{n\times n}$ is symmetric and positive definite, then the Gauss-Seidel method applied to $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ is convergent.*

6.1.4 Relaxation methods

Given an iterative method to solve $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, the corresponding relaxation method consists in first updating the kth approximation $\mathbf{x}^{(k)}$ of $\mathbf{x}$ to obtain $\hat{\mathbf{x}}^{(k+1)}$; then combining $\hat{\mathbf{x}}^{(k+1)}$ and $\hat{\mathbf{x}}^{(k)}$ to obtain the $(k+1)$th approximation $\mathbf{x}^{(k+1)}$:

$$\mathbf{x}^{(k)} \longrightarrow \hat{\mathbf{x}}^{(k+1)} \longrightarrow \mathbf{x}^{(k+1)} \equiv \omega\,\hat{\mathbf{x}}^{(k+1)} + (1-\omega)\mathbf{x}^{(k)}, \tag{15}$$

where the real parameter ω is called the relaxation parameter. On the one hand, when $\omega > 1$, we have an over-relaxation method, which can be used to speed up the convergence of a slow-converging scheme; on the other hand, when $\omega < 1$, we have an under-relaxation method, which can be used to make a diverging iterative scheme converge. The optimal choice of ω depends on the given problem and is hard to determine; hence, in practice, a few values of ω are usually chosen and, after a few iterations of the relaxed scheme, the most promising value of ω is kept.

Respect to the iterative methods that we know, the relaxed version of the Jacobi method is the Jacobi over-relaxation method (JOR), which arises after applying (15) to (10):

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}}\left(b_i - \sum_{j\neq i} a_{ij}x_j^{(k)}\right) + (1-\omega)x_i^{(k)}, \quad i = 1,\ldots,n, \tag{16}$$

with $\omega > 1$, or, in compact form,

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \omega\,\mathbf{D}^{-1}\left(\mathbf{b} - (\mathbf{L}+\mathbf{U})\mathbf{x}^{(k)}\right) + (1-\omega)\mathbf{x}^{(k)} \\ &= \mathbf{x}^{(k)} + \omega\,\mathbf{D}^{-1}\left(\mathbf{b} - \mathbf{A}\,\mathbf{x}^{(k)}\right).\end{aligned}$$

On the other hand, the relaxed version of the Gauss-Seidel method is the successive over-relaxation method (SOR), which arises after applying (15) to (11):

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

Therefore,

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

with $\omega > 1$, or, in compact form,

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

if and only if

$$\mathbf{x}^{(k+1)} = (\mathbf{D} + \omega\,\mathbf{L})^{-1} \left[\omega\,\mathbf{b} - (\omega\,\mathbf{U} + (\omega-1)\mathbf{D})\mathbf{x}^{(k)} \right].$$

Theorem 6.21. *Let* $\mathbf{A}$ *be real, symmetric and positive definite; then, the JOR method converges if* $0 < \omega < 2/\rho(\mathbf{D}^{-1}\mathbf{A})$.

Theorem 6.22 (Ostrowski and Reich)**.** *Let* $\mathbf{A}$ *be real, symmetric and positive definite; then, the SOR method converges, if and only if* $0 < \omega < 2$.

7 QR decomposition

Theorem 7.1 (QR decomposition)**.** *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$ *be a complex square matrix, with* $m \geq n$ *and* $\operatorname{rank}(A) = n$*; then, there is a unique factorization of* $\mathbf{A}$*, called the QR decomposition of* $\mathbf{A}$*, of the form*

$$\mathbf{A} = \mathbf{Q}\,\mathbf{R},$$

where $\mathbf{Q} \in \mathbb{C}^{m\times n}$ *is a matrix with orthonormal columns (i.e.,* $\mathbf{Q}^*\mathbf{Q} = \mathbf{I}_n$*), and* $\mathbf{R} \in \mathbb{C}^{n\times n}$ *is an upper triangular matrix, with real positive entries in the diagonal.*

Remarks:

- In this context, it is customary to call the upper triangular matrix $\mathbf{R}$, instead of $\mathbf{U}$
- There are authors that consider $\mathbf{Q} \in \mathbb{C}^{m\times m}$, and $\mathbf{R} \in \mathbb{C}^{m\times n}$, with the last $m-n$ rows of $\mathbf{R}$ being equal to zero; although, with that formulation, there is no uniqueness in the last $m-n$ columns of $\mathbf{Q}$. In fact, there exists a nonunique QR decomposition, with $\mathbf{Q} \in \mathbb{C}^{m\times m}$ and $\mathbf{R} \in \mathbb{C}^{m\times n}$, for every matrix $\mathbf{A}$, irrespectibly of its size and rank.
- There exist the related QL, RQ, LQ factorizations, with $\mathbf{L}$ being a lower triangular matrix.

Proof. Let us show uniqueness first. Suppose that $\mathbf{A}$ can be factorized as $\mathbf{A} = \mathbf{Q}_1\mathbf{R}_1 = \mathbf{Q}_2\mathbf{R}_2$, where $\mathbf{Q}_1^*\mathbf{Q}_1 = \mathbf{Q}_2^*\mathbf{Q}_2 = \mathbf{I}_n$, and $\mathbf{R}_1$ and $\mathbf{R}_2$ are upper triangular regular matrices with positive diagonal entries; then,

$$\mathbf{R}_1^*\mathbf{R}_1 = \mathbf{R}_1^*(\mathbf{Q}_1^*\mathbf{Q}_1)\mathbf{R}_1 = \mathbf{A}^*\mathbf{A} = \mathbf{R}_2^*(\mathbf{Q}_2^*\mathbf{Q}_2)\mathbf{R}_2 = \mathbf{R}_2^*\mathbf{R}_2 \iff (\mathbf{R}_2^{-1})^*\mathbf{R}_1^* = \mathbf{R}_2\mathbf{R}_1^{-1}.$$

Since the inverse of an upper triangular matrix is upper triangular, and the product of two upper triangular matrices is upper triangular, we conclude that $\mathbf{R}_2^{-1})^*\mathbf{R}_1^*$ is upper triangular. In the same way, we conclude that $\mathbf{R}_2\mathbf{R}_1^{-1}$ is lower triangular. Therefore, $(\mathbf{R}_2^{-1})^*\mathbf{R}_1^* = \mathbf{R}_2\mathbf{R}_1^{-1} = \mathbf{D}$, where $\mathbf{D}$ is a diagonal matrix. Then,

$$\begin{aligned}
&\mathbf{R}_2 = \mathbf{D}\,\mathbf{R}_1 \Longrightarrow r_{2,ii} = d_{ii}\,r_{1,ii}, \quad i = 1, \ldots, n,\\
&\mathbf{R}_1 = \mathbf{R}_2^*\mathbf{D} \Longrightarrow r_{1,ii} = d_{ii}\,r_{2,ii}^* = d_{ii}\,r_{2,ii}, \quad i = 1, \ldots, n,
\end{aligned}$$

where d_{ii}, $r_{1,ii}$ and $r_{2,ii}$ are, respectively, the diagonal entries of $\mathbf{D}$, $\mathbf{R}_1$, $\mathbf{R}_2$. Moreover, bearing in mind the positivity of the diagonal entries of $\mathbf{R}_1$ and $\mathbf{R}_2$,

$$d_{ii} = \frac{r_{2,ii}}{r_{1,ii}} = \frac{r_{1,ii}}{r_{2,ii}} \Longrightarrow r_{1,ii}^2 = r_{2,ii}^2 \Longrightarrow r_{1,ii} = r_{2,ii} \Longrightarrow d_{ii} = 1, \quad i = 1, \ldots, n,$$

i.e., $\mathbf{D} = \mathbf{I}_n$. Therefore,

$$\mathbf{R}_2\mathbf{R}_1^{-1} = \mathbf{I}_n \iff \mathbf{R}_1 = \mathbf{R}_2 \iff \mathbf{A}\,\mathbf{R}_1^{-1} = \mathbf{A}\,\mathbf{R}2^{-1} \iff \mathbf{Q}_1 = \mathbf{Q}_2.$$

In order to show the existence of the QR factorization, we use a constructive proof, i.e., a proof that constructs the factorization itself. Since $\text{rank}(\mathbf{A}) = n$, the columns $\{\mathbf{a}_1, \dots \mathbf{a}_n\}$ of $\mathbf{A}$ are linearly independent. Let $V = \text{span}(\{\mathbf{a}_1, \dots, \mathbf{a}_n\}) \subseteq \mathbb{R}^m$ be the vector space generated by $\{\mathbf{a}_1, \dots \mathbf{a}_n\}$; then, we construct an orthonormal base $\{\mathbf{v}_1, \dots \mathbf{v}_n\}$ of V, by applying the Gram-Schmidt process to $\{\mathbf{a}_1, \dots \mathbf{a}_n\}$:

$$\mathbf{u}_1 = \mathbf{a}_1 \Longrightarrow \mathbf{v}_1 = \frac{\mathbf{u}_1}{\sqrt{\langle \mathbf{u}_1, \mathbf{u}_1 \rangle}},$$
$$\mathbf{u}_j = \mathbf{a}_j - \sum_{k=1}^{j-i} \text{proj}_{\mathbf{u}_k}(\mathbf{a}_k) = \mathbf{a}_j - \sum_{k=1}^{j-i} \frac{\langle \mathbf{a}_j, \mathbf{u}_k \rangle}{\langle \mathbf{u}_k, \mathbf{u}_k \rangle} \mathbf{u}_k \Longrightarrow \mathbf{v}_j = \frac{\mathbf{u}_j}{\sqrt{\langle \mathbf{u}_j, \mathbf{u}_j \rangle}}, \quad j = 2, \dots, n,$$

where $\text{proj}_{\mathbf{u}}(\mathbf{a}) = (\langle \mathbf{a}, \mathbf{u} \rangle / \langle \mathbf{u}, \mathbf{u} \rangle)\mathbf{u}$ denotes the projection of $\mathbf{a}$ over $\mathbf{u}$. Therefore,

$$\text{span}(\{\mathbf{a}_1, \dots, \mathbf{a}_j\}) = \text{span}(\{\mathbf{v}_1, \dots, \mathbf{v}_j\}), \quad j = 1, \dots, n,$$

and, in particular, each $\mathbf{a}_j$ can be written as a unique linear combination of $\mathbf{v}_1, \dots, \mathbf{v}_j$, i.e., for each $\mathbf{a}_j$, there is a unique choice of nonzero scalars $\{\alpha_{1j}, \dots, \alpha_{jj}\}$, such that

$$\mathbf{a}_j = \alpha_{1j}\mathbf{v}_1 + \dots + \alpha_{jj}\mathbf{v}_j, \quad j = 1, \dots, n.$$

Finally, for each $j \in \{1, \dots, n\}$, we define

$$\begin{aligned} r_{ij} &\equiv \alpha_{ij}, \quad i < j, \\ r_{ij} &\equiv 0, \quad j > 1, \\ r_{jj} &\equiv |\alpha_{jj}|, \\ \mathbf{q}_j &\equiv \frac{\alpha_{jj}}{|\alpha_{jj}|}\mathbf{v}_j. \end{aligned}$$

Then, $\{\mathbf{q}_1, \dots, \mathbf{q}_n\}$ is an orthonormal basis, because, for all j, $|\alpha_{jj}/|\alpha_{jj}|| = 1$, and

$$\mathbf{a}_j = r_{1j}\mathbf{q}_1 + \dots + r_{jj}\mathbf{q}_j, \quad j = 1, \dots, n,$$

being the choice of the scalars r_{ij} unique, with r_{jj} real and positive. In compact matrix form,

$$\mathbf{A} = (\mathbf{a}_1 | \dots | \mathbf{a}_n) = (\mathbf{q}_1 | \dots | \mathbf{q}_n) \cdot \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1n} \\ & r_{22} & \dots & r_{2n} \\ & & \ddots & \vdots \\ 0 & & & r_{nn} \end{pmatrix} = \mathbf{Q}\,\mathbf{R},$$

where $\mathbf{Q} \in \mathbb{C}^{m \times n}$ is a matrix with orthonormal columns, and $\mathbf{R} \in \mathbb{C}^{n \times n}$ is an upper triangular matrix, with real positive diagonal entries. This concludes the proof.

□

Remarks: if $\text{rank}(A) = p < n$, a basis of $V = \text{span}(\{\mathbf{a}_1, \dots, \mathbf{a}_n\})$ consists of only p orthonormal vectors $\{\mathbf{q}_1, \dots, \mathbf{q}_p\}$. Therefore, since we need n orthonormal vectors to form the columns of $\mathbf{Q}$, we have to add $n - q$ vectors $\{\mathbf{q}_{p+1}, \dots, \mathbf{q}_n\}$, and this choice will not be unique, nor the choice of the entries of $\mathbf{R}$.

7.1 Numerical computation of the QR decomposition

There are three main techniques to compute the QR decomposition of a matrix $\mathbf{A} \in \mathbb{C}^{m \times n}$, with $m \geq n$, $n = \text{rank}(\mathbf{A})$:

- The Gram-Schmidt orthonormalization process. To apply it, we have to follow the steps described in the constructive proof of Theorem 7.1.
- Givens rotations.
- Householder reflectors.

7.1.1 Givens rotations

Given a vector $\mathbf{x} = (a, b)^T \in \mathbb{R}^2$, the matrix that rotates it θ degrees counterclockwise is

$$\mathbf{G} = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} = \begin{pmatrix} c & -s \\ s & c \end{pmatrix},$$

with $c = \cos(\theta)$, $s = \sin(\theta)$. Therefore, the idea is to choose $\mathbf{G}$, such that $\mathbf{G}\,\mathbf{x} = (\sqrt{a^2 + b^2}, 0)^T$. This can be achieved by taking

$$c = \frac{a}{\sqrt{a^2 + b^2}}, \quad s = -\frac{b}{\sqrt{a^2 + b^2}} \Longrightarrow \mathbf{G} = \begin{pmatrix} \frac{a}{\sqrt{a^2 + b^2}} & \frac{b}{\sqrt{a^2 + b^2}} \\ \frac{-b}{\sqrt{a^2 + b^2}} & \frac{a}{\sqrt{a^2 + b^2}} \end{pmatrix}.$$

Moreover, in the case that $\mathbf{x} = (a, b) \in \mathbb{C}^2$ is complex, $\mathbf{G}$ becomes

$$\mathbf{G} = \begin{pmatrix} c & -\bar{s} \\ s & \bar{c} \end{pmatrix} = \begin{pmatrix} \frac{\bar{a}}{\sqrt{|a|^2 + |b|^2}} & \frac{\bar{b}}{\sqrt{|a|^2 + |b|^2}} \\ \frac{-b}{\sqrt{|a|^2 + |b|^2}} & \frac{a}{\sqrt{|a|^2 + |b|^2}} \end{pmatrix} \Longrightarrow \mathbf{G} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \sqrt{|a|^2 + |b|^2} \\ 0 \end{pmatrix};$$

observe that $\mathbf{G}$ is unitary, so it preserves the Euclidean length of the vectors, and $\det(\mathbf{G}) = 1$.

In general, given a vector $\mathbf{u} \in \mathbb{C}^{n \times n}$, it is possible to carry out a rotation only on the ith and jth components, in such a way that the j-th component becomes zero, and the i-th component becomes $\sqrt{|u_i|^2 + |u_j|^2}$. This is achieved by the so-called Givens rotations.

Definition 7.2. *A Givens rotation is represented by a matrix $\mathbf{G} \in \mathbb{R}^{n \times n}$ of the form*

$$\mathbf{G}(i, j, \theta) = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \dots & \cos(\theta) & \dots & -\sin(\theta) & \dots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \dots & \sin(\theta) & \dots & \cos(\theta) & \dots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix} \begin{matrix} \\ \\ \leftarrow i \\ \\ \leftarrow j \\ \\ \\ \end{matrix}$$
$$\begin{matrix} & & \uparrow & & \uparrow & & \\ & & i & & j & & \end{matrix}$$

i.e., $\mathbf{G}$ is equal to the identity matrix $\mathbf{I}_n$, except for the four entries located at the intersection of the ith and jth rows and columns, with $1 \leq i < j \leq n$. Hence, given a vector $\mathbf{x} = (\mathbf{x}_1, \dots, x_n)^T \in \mathbb{R}^n$, it rotates the ith and jth entries of $\mathbf{x}$:

$$\mathbf{G}(i, j, \theta) \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_{i-1} \\ x_i \\ x_{i+1} \\ \vdots \\ x_{j-1} \\ x_j \\ x_{j+1} \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_{i-1} \\ \cos(\theta)x_i - \sin(\theta)x_j \\ x_{i+1} \\ \vdots \\ x_{j-1} \\ \sin(\theta)x_i + \cos(\theta)x_j \\ x_{j+1} \\ \vdots \\ x_n \end{pmatrix}.$$

As in the previous 2×2 example, it is straightforward to extend the definition of the Givens rotations to the complex case. More precisely, let $\mathbf{x} \in \mathbb{C}^n$, with $x_j \neq 0$; then, the Givens rotation matrix $\mathbf{G}$, such

that $[\mathbf{G}\,\mathbf{x}]_i = \sqrt{|x_i|^2 + |x_j|^2}$, $[\mathbf{G}\,\mathbf{x}]_j = 0$, and $[\mathbf{G}\,\mathbf{x}]_k = x_k$, for $k \notin \{i, j\}$, is given by

$$\mathbf{G} = \begin{pmatrix} 1 & \dots & 0 & \dots & 0 & \dots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \dots & \dfrac{\overline{x_i}}{\sqrt{|x_i|^2+|x_j|^2}} & \dots & \dfrac{\overline{x_j}}{\sqrt{|x_i|^2+|x_j|^2}} & \dots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \dots & \dfrac{-x_j}{\sqrt{|x_i|^2+|x_j|^2}} & \dots & \dfrac{x_i}{\sqrt{|x_i|^2+|x_j|^2}} & \dots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 & \dots & 1 \end{pmatrix} \Longrightarrow \mathbf{G} \cdot \begin{pmatrix} x_1 \\ \vdots \\ x_{i-1} \\ x_i \\ x_{i+1} \\ \vdots \\ x_{j-1} \\ x_j \\ x_{j+1} \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_{i-1} \\ \sqrt{|x_i|^2+|x_j|^2} \\ x_{i+1} \\ \vdots \\ x_{j-1} \\ 0 \\ x_{j+1} \\ \vdots \\ x_n \end{pmatrix}.$$

Computation of the QR decomposition by means of Givens rotations

In order to obtain the QR decomposition of a full column rank matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$, with $m \geq n$, we left-multiply it successively by Givens rotation matrices, making all the entries of the first column become zero, except for the first one; then, making all the entries of the second column become zero, except for the first two ones, etc. For instance, if $\mathbf{A} \in \mathbb{C}^{3\times 3}$ is a regular matrix, its QR decomposition can be computed in at most three steps (there can be different strategies):

$$\mathbf{A} = \begin{pmatrix} \times & \times & \times \\ \times & \times & \times \\ \times & \times & \times \end{pmatrix} \longrightarrow \mathbf{G}_1\mathbf{A} = \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ \times & \times & \times \end{pmatrix} \longrightarrow \mathbf{G}_2\mathbf{G}_1\mathbf{A} = \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & \times & \times \end{pmatrix}$$
$$\longrightarrow \mathbf{G}_3\mathbf{G}_2\mathbf{G}_1\mathbf{A} = \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & \times \end{pmatrix} = \mathbf{R}.$$

Finally, since the matrices $\mathbf{G}_k$ are unitary, and the product of unitary matrices is unitary, we conclude that $\mathbf{Q} = (\mathbf{G}_3\mathbf{G}_2\mathbf{G}_1)^{-1} = (\mathbf{G}_3\mathbf{G}_2\mathbf{G}_1)^* = \mathbf{G}_1^*\mathbf{G}_2^*\mathbf{G}_3^*$.

In general, using Givens rotations to compute the QR decomposition of a matrix $\mathbf{A}$ is particulary advisible when $\mathbf{A} \in \mathbb{C}^{m\times n}$ is a sparse matrix, i.e., the majority of its entries are zeros. Observe also that, if $m > n$, we obtain $\mathbf{Q} \in \mathbb{C}^{m\times m}$ and $\mathbf{R} \in \mathbb{C}^{m\times n}$, but, since the last $m-n$ rows of $\mathbf{R}$ consist only of zeros, it is safe to remove them, as well as the last $m-n$ columns of $\mathbf{Q}$.

7.1.2 Householder reflectors

Definition 7.3. *Given a vector $\mathbf{u} \in \mathbb{C}^n$, the Householder reflector associated to $\mathbf{u}$ is a matrix $\mathbf{H} \in \mathbb{C}^{n\times n}$ defined as*

$$\mathbf{H} = \mathbf{I}_n - \frac{2}{\mathbf{u}^*\mathbf{u}}\mathbf{u}\,\mathbf{u}^*.$$

Some authors ask $\mathbf{u}$ to be of unit length, i.e., $\|\mathbf{u}\|_2 = \sqrt{\mathbf{u}^*\mathbf{u}} = 1$; in that case,

$$\mathbf{H} = \mathbf{I}_n - 2\mathbf{u}\,\mathbf{u}^*.$$

Some properties of $\mathbf{H}$ are:

- $\mathbf{H}^* = \mathbf{H}$. This is evident from the definition.
- $\mathbf{H}^2 = \mathbf{I}_n$, because

$$\mathbf{H}^2 = \left(\mathbf{I}_n - \frac{2}{\mathbf{u}^*\mathbf{u}}\mathbf{u}\,\mathbf{u}^*\right)^2 = \mathbf{I}_n^2 + \frac{4}{(\mathbf{u}^*\mathbf{u})^2}\mathbf{u}\,\mathbf{u}^*\mathbf{u}\,\mathbf{u}^* - \frac{4}{\mathbf{u}^*\mathbf{u}}\mathbf{u}\,\mathbf{u}^* = \mathbf{I}_n.$$

  Therefore, bearing in mind the previous property, $\mathbf{H}^2 = \mathbf{H}^*\mathbf{H} = \mathbf{I}_n \iff \mathbf{H}^{-1} = \mathbf{H}^*$, i.e., $\mathbf{H}$ is unitary.

- $\mathbf{u}$ is an eigenvector of $\mathbf{H}$, with associated eigenvalue -1:

$$\mathbf{H}\,\mathbf{u} = \left(\mathbf{I}_n - \frac{2}{\mathbf{u}^*\mathbf{u}}\mathbf{u}\,\mathbf{u}^*\right)\mathbf{u} = \mathbf{u} - \frac{2}{\mathbf{u}^*\mathbf{u}}\mathbf{u}\,\mathbf{u}^*\mathbf{u} = \mathbf{u} - 2\mathbf{u} = -\mathbf{u}.$$

- Every $\mathbf{v}$ orthogonal to $\mathbf{u}$ is an eigenvector of $\mathbf{H}$, with associated eigenvalue 1:

$$\mathbf{H}\,\mathbf{v} = \left(\mathbf{I}_n - \frac{2}{\mathbf{u}^*\mathbf{u}}\mathbf{u}\,\mathbf{u}^*\right)\mathbf{v} = \mathbf{v} - 2\frac{\mathbf{u}^*\mathbf{v}}{\mathbf{u}^*\mathbf{u}}\mathbf{u} = \mathbf{v}.$$

Since $\dim(\mathrm{span}(\{\mathbf{u}\})) = 1$ and $\dim(\mathrm{span}(\{\mathbf{u}\})^\perp) = n-1$, where $\mathrm{span}(\{\mathbf{u}\})^\perp \subset \mathbb{C}^n$ denotes the orthogonal complement of $\mathrm{span}(\{\mathbf{u}\})$, we conclude from the last two properties that the only eigenvalues of $\mathbf{H}$ are -1, with multiplicity 1, and 1, with multiplicity $n-1$. In fact, the name reflector comes from the fact that $\mathbf{H}$, regarded as an linear operator, maps $\mathbf{x} \in \mathbb{C}^{n\times n}$ into its reflection with respect to the hyperplane orthogonal to $\mathbf{u}$ and containing the origin:

$$\mathbf{H}\,\mathbf{x} = \left(\mathbf{I}_n - \frac{2}{\mathbf{u}^*\mathbf{u}}\mathbf{u}\,\mathbf{u}^*\right)\mathbf{x} = \mathbf{x} - \frac{2}{\mathbf{u}^*\mathbf{u}}\mathbf{u}\,\mathbf{u}^*\mathbf{x} = \mathbf{x} - 2\frac{\mathbf{u}^*\mathbf{x}}{\mathbf{u}^*\mathbf{u}}\mathbf{u} = \mathbf{x} - 2\frac{\langle \mathbf{x},\mathbf{u}\rangle}{\langle \mathbf{u},\mathbf{u}\rangle}\mathbf{u} = \mathbf{x} - 2\,\mathrm{proj}_{\mathbf{u}}(\mathbf{x}),$$

so the mean between $\mathbf{x}$ and $\mathbf{H}\,x$ is orthogonal to $\mathbf{u}$, and the vector that joints $\mathbf{H}\,\mathbf{x}$ and $\mathbf{x}$ is parallel to $\mathbf{u}$:

$$\begin{cases} \dfrac{\mathbf{x}+\mathbf{H}\,\mathbf{x}}{2} = \mathbf{x} - \mathrm{proj}_{\mathbf{u}}(\mathbf{x}) \in \mathrm{span}(\{\mathbf{u}\})^\perp, \\ \mathbf{x} - \mathbf{H}\,\mathbf{x} = 2\,\mathrm{proj}_{\mathbf{u}}(\mathbf{x}) \in \mathrm{span}(\{\mathbf{u}\}); \end{cases}$$

in particular, if $\mathbf{x}$ is orthogonal to $\mathbf{u}$, it is reflected into itself; and if $\mathbf{x} = \mathbf{u}$, then its reflection is its opposite.

In order to apply the Householder reflectors to obtain the QR decomposition, we use the following lemma.

Lemma 7.4. *Let $\mathbf{x},\mathbf{y} \in \mathbb{C}^n$, such that $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2$, and $\mathbf{x}^*\mathbf{y} = \mathbf{y}^*\mathbf{x} \in \mathbb{R}$. Then, the Householder reflector associated to $\mathbf{u} = \mathbf{x} - \mathbf{y}$ maps $\mathbf{x}$ into $\mathbf{y}$, and $\mathbf{y}$ into $\mathbf{x}$.*

Proof.

$$\begin{aligned}\mathbf{H}\,\mathbf{x} &= \mathbf{x} - 2\frac{\mathbf{u}^*\mathbf{x}}{\mathbf{u}^*\mathbf{u}}\mathbf{u} = \mathbf{x} - 2\frac{(\mathbf{x}-\mathbf{y})^*\mathbf{x}}{(\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y})}(\mathbf{x}-\mathbf{y}) = \frac{\mathbf{x}(\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y}) - 2(\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^*\mathbf{x}}{(\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y})} \\ &= \frac{\mathbf{y}(\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y}) + (\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y}) - (\mathbf{x}-\mathbf{y})(\mathbf{x}-\mathbf{y})^*(2\mathbf{x})}{(\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y})} \\ &= \mathbf{y} - \frac{(\mathbf{x}-\mathbf{y})[(\mathbf{x}-\mathbf{y})^*(\mathbf{x}+\mathbf{y})]}{(\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y})} = \mathbf{y} - \frac{(\mathbf{x}-\mathbf{y})(\|\mathbf{x}\|_2^2 + \mathbf{x}^*\mathbf{y} - \mathbf{y}^*\mathbf{x} - \|\mathbf{y}\|_2^2)}{(\mathbf{x}-\mathbf{y})^*(\mathbf{x}-\mathbf{y})} = \mathbf{y},\end{aligned}$$

where we have used that $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2$, and that $\mathbf{x}^*\mathbf{y} = \mathbf{y}^*\mathbf{x}$. Therefore,

$$\mathbf{H}\,\mathbf{y} = \mathbf{H}\,\mathbf{H}\,\mathbf{x} = \mathbf{x}.$$

□

Remark: Observe that, if $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, the condition $\mathbf{x}^*\mathbf{y} = \mathbf{y}^*\mathbf{x} \in \mathbb{R}$ is always true.

Corollary 7.5. *Let $\mathbf{x} \in \mathbb{C}^n$, and $\mathbf{e}_k \in \mathbb{R}^n$ be the kth component of the canonical base of $\mathbb{R}^n$; then, the Householder reflector associated to $\mathbf{u} = \mathbf{x} + (x_k/|x_k|)\|\mathbf{x}\|_2\mathbf{e}_k$ maps $\mathbf{x}$ into $-(x_k/|x_k|)\|\mathbf{x}\|_2\mathbf{e}_k$.*

Proof. In this case, $\mathbf{y} = -(x_k/|x_k|)\|\mathbf{x}\|_2\mathbf{e}_k$. On the one hand, we have trivially $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2$. On the other hand, $\mathbf{x}^*\mathbf{y} = -\mathbf{x}^*((x_k/|x_k|)\|\mathbf{x}\|_2\mathbf{e}_k) = -(x_k/|x_k|)\|\mathbf{x}\|_2\overline{x_k} = -|x_k|\|\mathbf{x}\|_2 \in \mathbb{R}$, where we have use that $\mathbf{x}^*\mathbf{e}_k = \overline{x_k}$. □

Remarks: obviously, the Householder reflector associated to $\mathbf{u} = \mathbf{x} - (x_k/|x_k|)\|\mathbf{x}\|_2\mathbf{e}_k$ maps $\mathbf{x}$ into $(x_k/|x_k|)\|\mathbf{x}\|_2\mathbf{e}_k$. However, it is customary to take $\mathbf{u} = \mathbf{x} + (x_k/|x_k|)\|\mathbf{x}\|_2\mathbf{e}_k$, to minimize rounding errors; if $\mathbf{x} \in \mathbb{R}^n$, then $\mathbf{u} = \mathbf{x} + \mathrm{sign}(x_k)\|\mathbf{x}\|_2\mathbf{e}_k$. In general, the factor $2/(\mathbf{u}^*\mathbf{u})$ in the definition of $\mathbf{H}$ has a quite compact expression; if we denote $\sigma = (x_k/|x_k|)\|\mathbf{x}\|_2$, then $\mathbf{u} = \mathbf{x} + \sigma\,\mathbf{e}_k$, so

$$\begin{aligned}\frac{2}{\mathbf{u}^*\mathbf{u}} &= \frac{2}{(\mathbf{x}^* + \bar\sigma\,\mathbf{e}_k^T)(\mathbf{x} + \sigma\,\mathbf{e}_k)} = \frac{2}{\|\mathbf{x}\|_2^2 + \sigma\overline{x_k} + \bar\sigma x_k + |\sigma|^2} = \frac{2}{2|\sigma|^2 + 2|x_k|\sigma} \\ &= \frac{1}{|\sigma|(|x_k| + |\sigma|)} = \frac{1}{\|\mathbf{x}\|_2(|x_k| + \|\mathbf{x}\|_2)},\end{aligned}$$

and, in the case that $\mathbf{x}$ is real, $\sigma = \mathrm{sign}(x_k)\|\mathbf{x}\|_2$, so

$$\frac{2}{\mathbf{u}^*\mathbf{u}} = \frac{1}{\|\mathbf{x}\|_2(\mathrm{sign}(x_k)x_k + \|\mathbf{x}\|_2)} = \frac{1}{\mathrm{sign}(x_k)\|\mathbf{x}\|_2(x_k + \mathrm{sign}(x_k)\|\mathbf{x}\|_2)} = \frac{1}{\sigma\,u_k}.$$

Computation of the QR decomposition by means of Householder reflectors

Computing the QR decomposition of a full column rank matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$, with $m \geq n$, by means of the Householder reflectors is similar to computing it by using Givens rotations, but we make all the under-diagonal entries of a column become zero simultaneously. For instance, if $\mathbf{A} \in \mathbb{C}^{4\times 4}$ is a regular matrix, then

$$\mathbf{A} = \begin{pmatrix} \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \\ \times & \times & \times & \times \end{pmatrix} \longrightarrow \mathbf{H}_1\mathbf{A} = \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & \times & \times & \times \end{pmatrix} \longrightarrow \mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & \times & \times \end{pmatrix}$$

$$\longrightarrow \mathbf{H}_3\mathbf{H}_2\mathbf{H}_1\mathbf{A} = \begin{pmatrix} \times & \times & \times & \times \\ 0 & \times & \times & \times \\ 0 & 0 & \times & \times \\ 0 & 0 & 0 & \times \end{pmatrix} = \mathbf{R},$$

where

$$\mathbf{H}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & & & \\ 0 & & \tilde{\mathbf{H}}_2 & \\ 0 & & & \end{pmatrix}, \quad \mathbf{H}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & & \\ 0 & 0 & \tilde{\mathbf{H}}_3 & \end{pmatrix},$$

and $\tilde{\mathbf{H}}_2 \in \mathbb{C}^{3\times 3}$ and $\tilde{\mathbf{H}}_3 \in \mathbb{C}^{2\times 2}$ are respectively the Householder reflectors that act on the last three entries of the second column of $\mathbf{H}_1\mathbf{A}$, and on the last two entries of the third column of $\mathbf{H}_2\mathbf{H}_1\mathbf{A}$. Finally, since the matrices $\mathbf{H}_k$ are unitary and Hermitian, $\mathbf{Q} = (\mathbf{H}_3\mathbf{H}_2\mathbf{H}_1)^{-1} = (\mathbf{H}_3\mathbf{H}_2\mathbf{H}_1)^* = \mathbf{H}_1^*\mathbf{H}_2^*\mathbf{H}_3^* = \mathbf{H}_1\mathbf{H}_2\mathbf{H}_3$.

Observe that, as in the case of Givens rotations, if $m > n$, we can remove the last $m - n$ rows of $\mathbf{R}$ and the last $m - n$ columns of $\mathbf{Q}$.

7.1.3 Computation of the QR decomposition with column pivoting

Given a matrix $\mathbf{A} \in \mathbb{C}^{m\times n}$, with $m \geq n$, it is possible to modify slightly the Householder reflector technique, to obtain a factorization of the form $\mathbf{A}\,\mathbf{P} = \mathbf{Q}\,\mathbf{R}$, such that $|r_{11}| \geq \ldots \geq |r_{nn}|$, and $\mathbf{P}$ is a permutation matrix. More precisely, we first find a permutation matrix $\mathbf{P}_1$, such that the first column of $\mathbf{A}\,\mathbf{P}_1$ has the largest Euclidean norm $|r_{11}|$; then, we compute $\mathbf{H}_1$, in order that the under-diagonal elements of the first column of $\mathbf{H}_1\mathbf{A}\,\mathbf{P}_1$ equal zero. We proceed in a similar way with the other columns: we find a permutation matrix $\mathbf{P}_2$, such that the second column of $\mathbf{H}_1\mathbf{A}\,\mathbf{P}_1\mathbf{P}_2$, excluding its first element, has the largest Euclidean norm $|r_{22}|$, and we compute $\mathbf{H}_2$, in order that the under-diagonal elements of the second column of $\mathbf{H}_2\mathbf{H}_1\mathbf{A}\,\mathbf{P}_1\mathbf{P}_2$ equal zero, etc. If $\mathrm{rank}(\mathbf{A}) = r < n$,

$$\mathbf{H}_r \ldots \mathbf{H}_1\mathbf{A}\,\mathbf{P}_1 \ldots \mathbf{P}_r = \mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{pmatrix} \in \mathbb{C}^{m\times n},$$

where $\mathbf{R}_{11} \in \mathbb{C}^{r\times r}$ is upper triangular nonsingular, and $\mathbf{R}_{12} \in \mathbb{C}^{r\times(n-r)}$. Therefore, defining $\mathbf{Q} = \mathbf{H}_1 \ldots \mathbf{H}_r$, and $\mathbf{P} = \mathbf{P}_1 \ldots \mathbf{P}_r$, we have $\mathbf{A}\,\mathbf{P} = \mathbf{Q}\,\mathbf{R}$. On the other hand, if $\mathrm{rank}(\mathbf{A}) = n$,

$$\mathbf{H}_{n-1} \ldots \mathbf{H}_1\mathbf{A}\,\mathbf{P}_1 \ldots \mathbf{P}_{n-1} = \mathbf{R} = \begin{pmatrix} \mathbf{R}_{11} \\ \mathbf{0}_{(m-n)\times n} \end{pmatrix} \in \mathbb{C}^{m\times n},$$

where $\mathbf{R}_{11} \in \mathbb{C}^{n\times n}$ is upper triangular nonsingular. Here, $\mathbf{Q} = \mathbf{H}_1 \ldots \mathbf{H}_{n-1}$, and $\mathbf{P} = \mathbf{P}_1 \ldots \mathbf{P}_{n-1}$. As usual, if we remove the last zero rows of $\mathbf{R}$ and the corresponding columns of $\mathbf{Q}$, the factorization continues to be true.

The QR decomposition with pivoting improves the numerical accuracy and is especially useful when $\mathbf{A}$ is (nearly) rank deficient (i.e., it is not a full column rank matrix). For instance, if $\mathbf{A} \in \mathbb{C}^{3\times 3}$ has rank 2, the $\mathbf{R}$ matrix in the QR decomposition obtained by means of the Householder reflectors without pivoting might be

$$\mathbf{R} = \begin{pmatrix} \times & \times & \times \\ 0 & 0 & \times \\ 0 & 0 & \times \end{pmatrix} \text{ instead of } \mathbf{R} = \begin{pmatrix} \times & \times & \times \\ 0 & \times & \times \\ 0 & 0 & 0 \end{pmatrix}.$$

Therefore, the QR decomposition with column pivoting is a popular method to compute the rank of $\mathbf{A}$ and, hence, it constitutes a rank-revealing QR factorization.

8 More on least square problems

Let us consider the system of linear equations $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, with $\mathbf{A} \in \mathbb{C}^{m\times n}$, $\mathbf{b} \in \mathbb{C}^m$. In Theorem 4.4 of Section 4, we have proved that the least-square solution of that system is given by $\mathbf{x} = \mathbf{A}^+\mathbf{b}$. More precisely,

- $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ makes $\|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2$ minimum.
- Among the values of $\mathbf{x}$ for which the minimum of $\|\mathbf{A}\,\mathbf{x} - \mathbf{b}\|_2$ is achieved (in case of no uniqueness), $\mathbf{x} = \mathbf{A}^+\mathbf{b}$ is the one with the smallest $\|\cdot\|_2$ norm.

The most important case is when $\mathbf{A}$ is a full column rank matrix, i.e., $m \geq n$, and $\text{rank}(\mathbf{A}) = n$. Then, $\mathbf{A}^+ = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*$ (i.e., $\mathbf{A}^+ \equiv \mathbf{A}^{-1}$, if $m = n$), and there is uniqueness in the solution. Observe also that, when $\text{rank}(\mathbf{A}) = n$, computing $\mathbf{x} = \mathbf{A}^+\mathbf{b} = (\mathbf{A}^*\mathbf{A})^{-1}\mathbf{A}^*\mathbf{b}$ is equivalent to solving the so-called normal equations:

$$\mathbf{A}^*\mathbf{A}\,\mathbf{x} = \mathbf{A}^*\mathbf{b}, \tag{17}$$

which form a compatible determined system. On the other hand, the normal equations are equivalent to $\mathbf{A}^*(\mathbf{A}\,\mathbf{x} - \mathbf{b}) = 0$. Hence, $\mathbf{x}$ is a least-square solution of $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, if and only if the residual $\mathbf{A}\,\mathbf{x} - \mathbf{b}$ is orthogonal to the columns $\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$ of $\mathbf{A}$, i.e., if and only if $\mathbf{A}\,\mathbf{x}$ is the projection of $\mathbf{b}$ onto $V = \text{span}(\{\mathbf{a}_1, \ldots, \mathbf{a}_n\})$.

The previous ideas suggest different ways to find numerically the mean-square solution of $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, when $\text{rank}(\mathbf{A}) = n$:

- Compute $\mathbf{A}^+$, to obtain $\mathbf{x} = \mathbf{A}^+\mathbf{b}$. This is inefficient and should be avoided.
- Solve (17), by means of Gaussian elimination with pivoting.
- Compute the Cholesky factorization of $\mathbf{A}^*\mathbf{A}$, i.e., find a lower triangular matrix $\mathbf{L} \in \mathbb{C}$, with real and positive diagonal elements, and such that $\mathbf{A}^*\mathbf{A} = \mathbf{L}\,\mathbf{L}^*$. Then, transform (17) into $\mathbf{L}\,\mathbf{L}^*\mathbf{x} = \mathbf{A}^*\mathbf{b}$, and solve it in two steps:

$$\begin{aligned} \mathbf{L}\,\mathbf{y} = \mathbf{A}^*\mathbf{b} \quad &\longrightarrow \quad \text{obtain } \mathbf{y} \text{ using forward subtitution,} \\ \mathbf{L}^*\mathbf{x} = \mathbf{y} \quad &\longrightarrow \quad \text{obtain } \mathbf{x} \text{ using backward subtitution.} \end{aligned}$$

  This option is preferable to solving (17) using Gaussian elimination. In any case, $\mathbf{A}^*\mathbf{A}$ has often a very large condition number, so working directly on (17) may not be always advisable!

- Obtain the QR decomposition of $\mathbf{A}$, with $\mathbf{Q} \in \mathbb{C}^{m\times n}$, $\mathbf{R} \in \mathbb{C}^{n\times n}$, to transform (17) into $\mathbf{R}\,\mathbf{x} = \mathbf{Q}^*\mathbf{b}$:

$$\mathbf{A}^*\mathbf{A}\,\mathbf{x} = \mathbf{A}^*\mathbf{b} \Longleftrightarrow (\mathbf{Q}\,\mathbf{R})^*(\mathbf{Q}\,\mathbf{R})\,\mathbf{x} = (\mathbf{Q}\,\mathbf{R})^*\mathbf{b} \Longleftrightarrow \mathbf{R}^*\mathbf{Q}^*\mathbf{Q}\,\mathbf{R}\,\mathbf{x} = \mathbf{R}^*\mathbf{Q}^*\mathbf{b} \Longleftrightarrow \mathbf{R}\,\mathbf{x} = \mathbf{Q}^*\mathbf{b},$$

  and then solve $\mathbf{R}\,\mathbf{x} = \mathbf{Q}^*\mathbf{b}$, by using backward regression.

  A common implementation of this approach is to compute the QR decomposition of $\mathbf{A}$ via Householder reflections, but without obtaining explicitly $\mathbf{Q}$. More precisely, suppose that $\mathbf{H}_1$, applied to $\mathbf{A}$, makes all the under-diagonal entries of the first column become zero; then, we apply it to both sides of $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, to get $\mathbf{H}_1\mathbf{A}\,\mathbf{x} = \mathbf{H}_1\mathbf{b}$. We proceed similarly with the other columns: suppose that $\mathbf{H}_2$, applied to $\mathbf{H}_1\mathbf{A}$, makes all the under-diagonal entries of the second column become zero; then, we apply it to both sides of $\mathbf{H}_1\mathbf{A}\,\mathbf{x} = \mathbf{H}_1\mathbf{b}$, to get $\mathbf{H}_2\mathbf{H}_1\mathbf{A}\,\mathbf{x} = \mathbf{H}_2\mathbf{H}_1\mathbf{b}$; and so on. At the end of the process, we get

$$\mathbf{H}_{n-1}\ldots\mathbf{H}_1\mathbf{A}\,\mathbf{x} = \mathbf{R}\,\mathbf{x} = \mathbf{H}_{n-1}\ldots\mathbf{H}_1\mathbf{b},$$

  where $\mathbf{H}_{n-1}\ldots\mathbf{H}_1\mathbf{A} = \mathbf{R} \in \mathbb{C}^{m\times n}$ is upper triangular, and has its last $m - n$ rows equal to zero. Observe that it is not necessary to obtain explicitly $\mathbf{H}_1\ \ldots, \mathbf{H}_{n-1}$, because it is enough to compute its action on the matrices and vectors to which they are applied. Then,

$$\mathbf{R}\,\mathbf{x} = \begin{pmatrix} \mathbf{R}_{11} \\ \mathbf{0}_{(m-n)\times n} \end{pmatrix}\mathbf{x} = \mathbf{H}_{n-1}\ldots\mathbf{H}_1\mathbf{b} = \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \Longrightarrow \mathbf{R}_{11}^*\mathbf{R}_{11}\mathbf{x} = \mathbf{R}_{11}^*\mathbf{c} \Longrightarrow \mathbf{x} = \mathbf{R}_{11}^{-1}\mathbf{c},$$

  where $\mathbf{c} \in \mathbb{C}^n$, $\mathbf{d} \in \mathbb{C}^{m-n}$, and $\mathbf{R}_{11} \in \mathbb{C}^{n\times n}$ is upper-triangular nonsingular. Remark that we have ignored the last $m - n$ entries of $\mathbf{H}_{n-1}\ldots\mathbf{H}_1\mathbf{b}$.

8.1 Least-square problems for rank-defficient matrices

Suppose that $\mathbf{A} \in \mathbb{C}^{m\times n}$, with $m \geq n$, and $\text{rank}(A) = r < n$. In order to solve $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, we compute the QR decomposition of $\mathbf{A}$ with column pivoting. As in the case of full column rank matrices, we use Householder reflections, but without obtaining explicitly $\mathbf{Q}$, to transform $\mathbf{A}\,\mathbf{x} = \mathbf{b}$ into

$$\mathbf{H}_r \dots \mathbf{H}_1 \mathbf{A}\,\mathbf{P}_1 \dots \mathbf{P}_r \mathbf{P}_r^T \dots \mathbf{P}_1^T \mathbf{x} = \mathbf{R}\,\mathbf{P}^T \mathbf{x} = \mathbf{H}_r \dots \mathbf{H}_1 \mathbf{b},$$

where $\mathbf{H}_r \dots \mathbf{H}_1 \mathbf{A}\,\mathbf{P}_1 \dots \mathbf{P}_r = \mathbf{R}$ is upper triangular and has its $m - r$ columns equal to zero, and $\mathbf{P} = \mathbf{P}_1 \dots \mathbf{P}_r$. Hence,

$$\begin{aligned}
\mathbf{R}\,\mathbf{P}^T\mathbf{x} &= \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \mathbf{H}_{n-1} \dots \mathbf{H}_1 \mathbf{b} = \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \\
&\Longrightarrow \begin{pmatrix} \mathbf{R}_{11}^* & \mathbf{0}_{r\times(m-r)} \\ \mathbf{R}_{12}^* & \mathbf{0}_{(n-r)\times(m-r)} \end{pmatrix} \begin{pmatrix} \mathbf{R}_{11} & \mathbf{R}_{12} \\ \mathbf{0}_{(m-r)\times r} & \mathbf{0}_{(m-r)\times(n-r)} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{11}^* & \mathbf{0}_{r\times(m-r)} \\ \mathbf{R}_{12}^* & \mathbf{0}_{(n-r)\times(m-r)} \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ \mathbf{d} \end{pmatrix} \\
&\Longrightarrow \begin{pmatrix} \mathbf{R}_{11}^*\mathbf{R}_{11} & \mathbf{R}_{11}^*\mathbf{R}_{12} \\ \mathbf{R}_{12}^*\mathbf{R}_{11} & \mathbf{R}_{12}^*\mathbf{R}_{12} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \begin{pmatrix} \mathbf{R}_{11}^*\mathbf{c} \\ \mathbf{R}_{12}^*\mathbf{c} \end{pmatrix} \Longrightarrow \mathbf{R}_{11}\mathbf{y} + \mathbf{R}_{12}\mathbf{z} = \mathbf{c} \Longrightarrow \mathbf{y} = \mathbf{R}_{11}^{-1}(\mathbf{c} - \mathbf{R}_{12}\mathbf{z}) \\
&\Longrightarrow \mathbf{x} = \mathbf{P} \begin{pmatrix} \mathbf{R}_{11}^{-1}(\mathbf{c} - \mathbf{R}_{12}\mathbf{z}) \\ \mathbf{z} \end{pmatrix}
\end{aligned}$$

with $\mathbf{c}, \mathbf{y} \in \mathbb{C}^r$, $\mathbf{d}, \mathbf{z} \in \mathbb{C}^{m-r}$, $\mathbf{R}_{11} \in \mathbb{C}^{r\times r}$ upper-triangular nonsingular. In this case, there is no uniqueness, but the solution with the smallest Euclidean norm is obtained after taking $\mathbf{z} = \mathbf{0}_{m-r}$; this is called the basic solution:

$$\mathbf{x} = \mathbf{P} \begin{pmatrix} \mathbf{R}_{11}^{-1}\mathbf{c} \\ \mathbf{0}_{n-r} \end{pmatrix}.$$

Indeed, Matlab returns the basic solution of $\mathbf{A}\,\mathbf{x} = \mathbf{b}$, when typing `A\b`.

Observe that this technique can be applied with very little changes when $m < n$; and when $m \geq n$, and $\text{rank}(\mathbf{A}) = n$.

8.2 Data fitting

Given a (very) large collection of pairs $(x_i, y_i) \in \mathbb{R}^2$, $i = 1, \dots, m$, we are interesting in finding a function $\hat{y} = f(x)$, such that it enables us to approximate the second component of each pair by the first one (remark that it is customary to write $\hat{y}$ instead of y, because $f(x)$ gives an approximation). The most important case is the so-called linear regression. Suppose that we plot the pairs (x_i, y_i) and the resulting point cloud is roughly a straight line; then, we take $\hat{y} = f(x) = a\,x + b$, and, for a given x_i, the approximation to y_i, which we denote $\hat{y}_i$, is given by $\hat{y}_i = f(x_i) = a\,x_i + b$. There are different ways of determining a and b, but the most common one is to choose them in such a way that the sum of the squares of the errors is the smallest possible one. Hence, we have to minimize

$$E(a, b) = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{m} (a\,x_i + b - y_i)^2.$$

This is equivalent to finding the least-square solution of the following matrix problem:

$$\begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_m & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix},$$

which has no exact solution, unless there is a perfect linear relationship between x_i and y_i, for all i.

Suppose now that the point cloud does not have a linear shape, but an exponential one; then, we choose $\hat{y} = f(x) = a\,e^{b\,x}$, and we have to minimize

$$\sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{m} (a\,e^{b\,x_i} - y_i)^2. \tag{18}$$

However, this has not an equivalent in the form of a system of linear equations where a in b are the unknowns. To solve this issue, we apply the change of variable $u = x$, $v = \ln y$ to the original pairs (x_i, y_i), to get $(u_i, v_i) = (x_i, \ln y_i)$. Then, $\ln f(x) = \ln a + b\,x$, and, defining $c = \ln a$, we have $\hat{v} = f(u) = b\,u + c$. Hence, instead of (18), we rather minimize

$$\sum_{i=1}^{m} (\hat{v}_i - v_i)^2 = \sum_{i=1}^{m} (b\,u_i + c - v_i)^2,$$

which is equivalent to finding the least-square solution of

$$\begin{pmatrix} u_1 & 1 \\ \vdots & \vdots \\ u_m & 1 \end{pmatrix} \cdot \begin{pmatrix} b \\ c \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix}.$$

In general, there are a number of popular choices of $\hat{y} = f(x)$ that can be transformed to $\hat{v} = f(u) = a\,u + b$ by a simple change of variable. In what follows, we offer some examples:

- $\hat{y} = \dfrac{a}{x} + b \Longrightarrow u = \dfrac{1}{x},\ v = y.$
- $\hat{y} = \dfrac{d}{x + c} \Longrightarrow u = x\,y,\ v = y.$
- $\hat{y} = \dfrac{1}{a\,x + b} \Longrightarrow u = x,\ v = \dfrac{1}{y}.$
- $\hat{y} = \dfrac{x}{a\,x + b} \Longrightarrow u = \dfrac{1}{x},\ v = \dfrac{1}{y}.$
- $\hat{y} = a\,\ln x + b \Longrightarrow u = \ln x,\ v = y.$
- $\hat{y} = a\,e^{b\,x} \Longrightarrow u = x,\ v = \ln y.$
- $\hat{y} = \dfrac{1}{(a\,x + b)^2} \Longrightarrow u = x,\ v = \dfrac{1}{\sqrt{y}}.$
- $\hat{y} = a\,x\,e^{b\,x} \Longrightarrow u = x,\ v = \ln(y/x).$
- $\hat{y} = \dfrac{\lambda}{1 + a\,e^{b\,x}} \Longrightarrow u = x\ v = \ln\left(\dfrac{\lambda}{y} - 1\right).$

Linear combinations in least-square problems

Suppose that $\hat{y} = f(x)$ is a linear combination of n functions $f_j(x)$, $j = 1, \ldots, n$:

$$\hat{y} = f(x) = \sum_{j=1}^{n} c_j f_j(x).$$

Then, the problem consists in minimizing

$$E(c_1, \ldots, c_n) = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} (\hat{y}_i - y_i)^2 = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} c_j f_j(x) - y_i \right)^2,$$

which is equivalent to finding the least-square solution of

$$\begin{pmatrix} f_1(x_1) & \ldots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_n) & \ldots & f_n(x_n) \end{pmatrix} \cdot \begin{pmatrix} c_1 \\ \vdots \\ c_m \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_m \end{pmatrix}.$$

Multiple regression

Given a (very) large collection of triples $(x_i, y_i, z_i) \in \mathbb{R}^3$, $i = 1, \ldots, n$, we want to explain the third component as a linear relationship of the first two, i.e., $\hat{z}_i = f(x_i, y_i) = a\,x_i + b\,y_i + c$. Therefore, the problem consists in minimizing

$$E(a, b, c) = \sum_{i=1}^{m} e_i^2 = \sum_{i=1}^{m} (\hat{z}_i - z_i)^2 = \sum_{i=1}^{m} (a\,x_i + b\,y_i + c - z_i)^2,$$

which is equivalent to finding the least-square solution of

$$\begin{pmatrix} x_1 & y_1 & 1 \\ \vdots & \vdots & \vdots \\ x_m & y_m & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} z_1 \\ \vdots \\ z_m \end{pmatrix}.$$

Obviously, it is possible to consider higher dimensions, complex numbers, etc.