4

Solving linear systems

> *"We must admit with humility that, while number is purely a product of our minds, space has a reality outside our minds."*
>
> *Carl Friedrich Gauss (1777-1855)*

> *"Almost everything, which the mathematics of our century has brought forth in the way of original scientific ideas, attaches to the name of Gauss."*
>
> *Léopold Kronecker (1823-1891)*

The problem treated in this chapter is the numerical resolution systems of m linear equations in n variables, of the form

$$A\,x = b\,,$$

where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$ is the unknown solution vector. Such systems of equations can be encountered in almost every problem in scientific computing. Among the situations in which the numerical resolution of such problems is required appear notably the approximation of the solution of ordinary or partial differential equations by finite difference, finite element or finite volume methods (Chapters 5, 6, **??**). Even the solution of a nonlinear problem is usually obtained by solving a sequence of linear systems (Newton's method). Since the algorithms for solving linear systems are widely used in a large range of applications, the methods must then be *efficient*, *accurate*, *reliable* and *robust*.

Only in the case where the matrix A has full row and column rank, *i.e.,* $rank(A) = m = n$, does the linear system $A\,x = b$ have a *unique* solution for any right hand side b. When $rank(A) < n$, the system either has many solutions, is *undetermined* or has no solution, is *overdetermined*.

Two classes of methods for solving systems of linear equations are of interest: *direct* methods and *iterative* methods. In a direct method, the matrix of the initial linear system is transformed or factorized into a simpler form, involving diagonal or triangular matrices, using elementary transformations, which can be solved easily. The exact solution is obtained in a finite numer of arithmetic operations, if not considering numerical rounding errors. The most proeminent direct method is the *Gaussian elimination.* Iterative methods, on the other hand, compute a *sequence* of approximate solutions, which converges to the exact solution in the limit, *i.e.,* in practice until a desired accuracy is obtained.

During a long time, direct methods have been preferred to iterative methods for solving linear systems, mainly because of their simplicity and robustness. However, the emergence of conjugate gradient methods and Krylov subspace iterations provided an efficient alternative to direct solvers. Nowadays, iterative methods are almost mandatory in complex applications, notably because of memory and computational requirements that prohibit the use of direct methods. They usually involve a matrix-vector multiplication procedure that is cheap to compute on modern computer architectures. When the matrix A is very large and is composed of a majority of nonzero elements, the LU factorization for example would contains much more nonzero coefficients than the matrix A itself. Nonetheless, in some peculiar applications, very ill-conditioned matrices arise that may required a direct method for solving the problem at hand.

In the important case where the matrix A is *symmetric positive definite*, about half of the work can be spared. If only a fraction of the elements of A are nonzeroes, then the linear system is called *sparse*. Nowadays, without the knowledge and the exploitation of the *sparsity* of the matrix A, many application problems could not be solved.

Numerical methods for solving linear systems are a good illustration of the difference between analytical mathematics and "engineering" numerical analysis. Actually, significant progress in the design of algorithms have been obtained in the last decades, thanks to the advent of efficient computer architectures. Again, we face this intriguing context where some perfectibly theoretically sound methods reveal useless for computing the numerical solution.

In this chapter, we consider matrices with real or complex entries and therefore, we denote $\mathbb{K}$ a field that is either $\mathbb{R}$ or $\mathbb{C}$. In Section 1, the elementary properties of finite-dimensional vector spaces and matrix algebra are briefly reviewed, and statements may be given without proofs. When the matrix A is square dense and without apparent structure, the Gaussian elimination and LU factorization methods are likely to be the methods of choice, as will be seen in Section 2. Classical iterative algorithms and projection methods for solving sparse linear systems are presented in Section 3. Finally, a small section is devoted to methods for computing eigenvalues.

4.1 A linear algebra primer

Before introducing various methods for solving linear systems of the form $Ax = b$, we propose a brief overview of the main results in linear algebra suitable for our purpose. We refer the reader to classical courses in linear algebra and matrix computations for further details (see the bibliography section at the end of this chapter).

Since we refer frequently to *vectors* and *matrices*, let us recall some conventional notations. A lowercase letter like x will always denote a vector and its jth component will be written x_j. Vectors are almost always considered as column vectors. A matrix is denoted by an uppercase lettre like A and its (i,j)th element will be written a_{ij}. $\mathcal{M}_{m,n}(\mathbb{K})$ will denote the set of $m \times n$ (real or complex) matrices. Any matrix $A \in \mathcal{M}_{m,n}(\mathbb{K})$ may be defined by its columns $c_j \in \mathbb{K}^m$ as $A = [c_1|\dots|c_n]$. Given a matrix $A = (a_{ij})$, $(A^t)_{ij} = a_{ji}$ and $(A^*)_{ij} = \bar{a}_{ji}$ will denote the *transpose* of the matrix A and the *conjugate transpose* of A, respectively. If $A \in \mathcal{M}_{m,n}(\mathbb{K})$, then $|A| \in \mathcal{M}_{m,n}(\mathbb{K})$ denotes the matrix of absolute values of entries of A: $(|A|)_{ij} = |a_{ij}|$.

4.1.1 Basic notions

Definition 4.1.1. *A* vector space *over the field* $\mathbb{K}$ *is a nonempty set* V *in which* addition *and* mutiplication *are defined and such that for all vectors* $u, v \in V$ *and any scalars* $\alpha, \beta \in \mathbb{K}$*, the following properties must hold:*

1. *addition is* commutative *and* associative*;*
2. additive identity*:* $u + 0 = u$*, 0 is the* zero *vector;*
3. additive inverse*, for any* u*, there exists* $-u$ *such that* $u + (-v) = 0$*;*
4. *distributivity properties:*

$$\begin{aligned} \forall \alpha \in \mathbb{K}, \forall u, v \in V, &\quad \alpha(u+v) = \alpha u + \alpha v \\ \forall \alpha, \beta \in \mathbb{K}, \forall u \in V, &\quad (\alpha + \beta)u = \alpha u + \beta v\,. \end{aligned}$$

5. *assocative property:*

$$\forall \alpha, \beta \in \mathbb{K}, \forall u \in V, \quad (\alpha\beta)u = \alpha(\beta u)\,.$$

6. *scalar multiplication identity:* $1u = u$*.*

The set W of linear combinations of a system of p vectors of V, $\{u^1, \dots, u^p\}$, is a subspace of V called the *span* of the vector system and denoted by

$$W = \mathrm{span}\{u^1, \dots, u^p\} = \{w = \sum_{i=1}^{p} \alpha_i u^i\,,\ \text{with}\ \ \alpha_i \in \mathbb{K}\}\,.$$

A set of m vectors $\{u^1, \dots, u^m\}$ of V is called *linearly independent* if none of its element can be expressed as a linear combination of the other vectors, *i.e.* if the relation

$$\sum_{i=1}^{m} \alpha_i u^i = 0\,,$$

with $(\alpha_i)_{1 \le i \le m} \in \mathbb{K}$ implies that every $\alpha_i = 0$. Otherwise, it is called *linearly dependent*. A *basis* of V is then a linearly independent subset of V that spans V.

Let $u = (u_1, \dots, u_n)^t$ be a vector over a field $\mathbb{K}$. We denote $u^* \in \mathbb{K}^n$ the *adjunct* of the column vector u such that $u^* = (\bar{u}_1, \dots, \bar{u}_n)$ and the *transpose* of u the row vector $u^t = (u_1, \dots, u_n)$. We recall the definition of the matrix-vector and matrix-matrix products. Given a matrix $A \in \mathcal{M}_{m,n}(\mathbb{K})$ and a vector $u \in \mathbb{K}^n$, the vector $v = Au \in \mathbb{K}^m$ is such that:

$$v_i = \sum_{j=1}^{n} a_{ij}\, u_j\,, \quad i = 1, \dots, m\,.$$

and given $B \in \mathcal{M}_{n,p}(\mathbb{K})$ we define $C = AB \in \mathcal{M}_{m,p}(\mathbb{K})$ as:

$$c_{ij} = \sum_{k=1}^{n} a_{ik}\, b_{kj}\,, \quad i = 1, \dots, m \quad j = 1, \dots, p\,.$$

In the canonical basis of $\mathbb{K}^n$, the *dot product*, also known as the *scalar product*, of two vectors $u, v \in \mathbb{K}^n$ is denoted (u, v) and is the scalar defined as:

$$(u, v) = \sum_{i=1}^{n} u_i\, v_i^*\,.$$

We consider the case of *square matrices* A having n rows and n columns that belong to the set $\mathcal{M}_n(\mathbb{K})$, that is a noncommutative algebra for the multiplication. The neutral element is usually denoted I_n and is defined by its entries $(\delta_{i,j})_{1 \le i,j \le n}$, where $\delta_{i,j}$ is the Kronecker symbol. We recall that a square matrix $A \in \mathcal{M}_n(\mathbb{K})$ is said to be *invertible* (or nonsingular) if there exists a matrix $B \in \mathcal{M}_n(\mathbb{K})$ such that $AB = BA = I_n$. This *inverse* matrix B is denoted A^{-1}. In contrast, a noninvertible matrix is called *singular.*

Definition 4.1.2. *Suppose $A \in \mathcal{M}_n(\mathbb{K})$ is a square matrix.*

- *A is a* normal *matrix if and only if $AA^* = A^*A$,*
- *A is a* unitary *matrix if and only if $AA^* = A^*A = I_n$. Moreover, if $\mathbb{K} = \mathbb{R}$, A is a* orthogonal *matrix and $AA^t = A^tA = I_n$ and $A^t = A^{-1}$,*
- *A is a* Hermitian *matrix if and only if $A^* = A$. Moreover, if $\mathbb{K} = \mathbb{R}$, A is a* symmetric *or* self-adjoint *matrix and $A^t = A$.*

Proposition 4.1.1. *Every* Hermitian *matrix A is a* normal *matrix.*

Proof. Suppose $A \in \mathcal{M}_n(\mathbb{K})$ is such that $A^* = A$. Then, we have

$$AA^* = AA = A^*A\,.$$

□

In many applications, we will consider *sparse* matrices, *i.e.,* matrices primarily filled with zeros. In particular, a matrix $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{K})$ is a *lower triangular* matrix if and only if $a_{ij} = 0$, $1 \le i < j \le n$, A is a *upper triangular* matrix if and only if $a_{ij} = 0$, $1 \le j < i \le n$ and A is a *diagonal* matrix if $a_{ij} = 0$ for $i \neq j$. By contrast, a matrix A that is mainly populated by nonzeros if called a *dense* matrix.

Lemma 4.1.1. *Suppose L and U are two invertible lower triangular and upper triangular matrices in $\mathcal{M}_n(\mathbb{K})$, respectively. Then, L^{-1} (resp.U^{-1}) is a lower (resp. upper) triangular matrix.*

Proof. We show the first assertion only. Suppose $A = (a_{ij})$ is the inverse of $L = (l_{ij})$, we have then:

$$\delta_{i,j} = \sum_{k=1}^{n} a_{ik} l_{kj} = \sum_{k=1}^{j} a_{ik} l_{kj} \,, \qquad \forall\, 1 \leq i, j \leq n \,,$$

Hence, for $j = 1$ and $i > 1$ we have, $a_{i1} l_{11} = 0$ yielding to $a_{i1} = 0$. Likewise, for $j = 2$ and $i > 2$ we have $a_{i1} l_{12} + a_{i2} l_{22} = a_{i2} l_{22} = 0$ and thus $a_{i2} = 0$. By recurrence, we show that for every $j > i$, we have $a_{ij} = 0$. It is then easy to conclude that A is the inverse of L and is a lower triangular matrix. □

The *trace* of a square matrix $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{K})$ is the sum of its diagonal elements:

$$\operatorname{tr} A = \sum_{i=1}^{n} a_{ii} \,.$$

The *determinant* of a square matrix $A = (a_{ij}) \in \mathcal{M}_n(\mathbb{K})$ is defined as

$$\det A = \sum_{\sigma \in S_n} \varepsilon(\sigma) \prod_{i=1}^{n} a_{i\sigma(i)} \,,$$

where $\varepsilon(\sigma) = (-1)^{p(\sigma)}$ is the *signature* of the permutation σ, equal to 1 or -1, S_n is the set of all permutations of the set $\{1, 2, \ldots, n\}$ into itself and the number $p(\sigma)$ is the number of inversions in σ. Hence $\varepsilon(\sigma)$ is 1 if σ is even and -1 if σ is odd.

Lemma 4.1.2. *Given A and B two square matrices in $\mathcal{M}_n(\mathbb{K})$. Then*

1. $\det(AB) = (\det A)(\det B) = \det(BA)$;
2. $\det(A^t) = \det A$;
3. *A is invertible if and only if* $\det A \neq 0$.

The *kernel*, or null space, of a matrix $A \in \mathcal{M}_n(\mathbb{K})$ is the set denoted $\operatorname{Ker} A$ of vectors $x \in \mathbb{K}^n$ such that $Ax = 0$. The *range* of A is the set denoted $\operatorname{Im} A$ of vectors $y \in \mathbb{K}^n$ such that $y = Ax$, for $x \in \mathbb{K}^n$. By definition, the dimension of the space $\operatorname{Im} A$ is called the *rank* of A and is denoted by $\operatorname{rank} A$.

Lemma 4.1.3 (invertible matrix). *For any square matrix $A \in \mathcal{M}_n(\mathbb{K})$, the following assertions are equivalent:*

1. *A is invertible;*
2. $\operatorname{Ker} A = \{0\}$;
3. $\operatorname{Im} A = \mathbb{K}^n$;
4. *there exists $B \in \mathcal{M}_n(\mathbb{K})$ such that $AB = I_n$ and $BA = I_n$ (and $B = A^{-1}$).*

Lemma 4.1.4. *Let A and B be two invertible matrices in $\mathcal{M}_n(\mathbb{K})$. Then*

$$(AB)^{-1} = B^{-1}\,A^{-1}\,.$$

The *characteristic polynomial* of $A \in \mathcal{M}_n(\mathbb{K})$ is the polynomial $P_A(\lambda)$ of degree n defined on $\mathbb{K}$ by

$$P_A(\lambda) = \det(A - \lambda I_n)\,.$$

The *n roots* $\lambda_i \in \mathbb{K}$ of this polynomial, not necessarily distinct, are called the *eigenvalues* of A. A vector $x \in \mathbb{K}^n$, $x \neq 0$, such that $Ax = \lambda x$ is the *eigenvector* of A associated with the eigenvalue λ. There exists at least one such eigenvector. The set of eigenvalues of A is called the *spectrum* of A and is denoted Sp(A):

$$\mathrm{Sp}(A) = \{\lambda_i \in \mathbb{K}\,; 1 \leq i \leq n\,; \exists x_i \in \mathbb{K}^n \,:\, x_i \neq 0\,,\ Ax_i = \lambda_i x_i\}\,.$$

The *spectral radius* of A is the maximum of the moduli of the eigenvalues of A:

$$\rho(A) = \max_{1\leq i\leq n}\{|\lambda_i|\,;\ \lambda_i \in \mathrm{Sp}(A)\}\,.$$

Remark 4.1.1. 1. if $A \in \mathcal{M}_n(\mathbb{R})$, A may have complex eigenvalues;

2. the characteristic polynomial P_A is invariant under any basis change, *i.e.,* for any invertible matrix P we have:

$$\det(P^{-1}AP - \lambda I_n) = \det(A - \lambda I_n)\,;$$

3. the eigenvalues of a *Hermitian* matrix are all *real.*

The vector subspace defined by $R_\lambda = \mathrm{Ker}(A - \lambda I_n)$ is the *eigensubspace* associated with the eigenvalue λ. Moreover, the vector subspace

$$F_\lambda = \bigcup_{k\geq 1} \mathrm{Ker}(A - \lambda I_n)^k$$

is the *generalized eigenspace* associated with the eigenvalue λ.

Theorem 4.1.1 (Schur's theorem). *If $A \in \mathcal{M}_n(\mathbb{K})$ is a square matrix, then there exists a* unitary *matrix $U \in \mathcal{M}(\mathbb{K})$ such that $U^*\,A\,U$ is* triangular *and its diagonal entries are all eigenvalues of A.*

Corollary 4.1.1. *If $A \in \mathcal{M}_n(\mathbb{K})$ is a Hermitian matrix, then there exists a* unitary *matrix $U \in \mathcal{M}(\mathbb{K})$ such that $U^*\,A\,U$ is* diagonal *and its entries are all eigenvalues of A.*

Theorem 4.1.2 (Diagonalization). *$A \in \mathcal{M}_n(\mathbb{K})$ is* normal *if and only if there exists a* unitary *matrix $U \in \mathcal{M}(\mathbb{K})$ such that*

$$A = U\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)\,U^{-1}$$

where the λ_i are the eigenvalues of A.

Lemma 4.1.5. *For any matrix $A \in \mathcal{M}_{m,n}(\mathbb{K})$, the matrix A^*A is* Hermitian *and has real, nonnegative eigenvalues.*

The *singular* values of $A \in \mathcal{M}_{m,n}(\mathbb{K})$ are the *nonnegative* square roots of the n eigenvalues of A^*A. It is easy to see that the singular values of a *normal* matrix are the moduli of its eigenvalues.

Lemma 4.1.6 (SVD decomposition). *Let $A \in \mathcal{M}_{m,n}(\mathbb{K})$ be an arbitrary matrix, with $m > n$, having r positive singular values μ_i. Then, there exists two* unitary *matrices $U \in \mathcal{M}_n(\mathbb{K})$ and $V \in \mathcal{M}_n(\mathbb{K})$ and a* diagonal *matrix $\Sigma = \mathrm{diag}(\lambda_1, \dots, \lambda_r) \in \mathcal{M}_n(\mathbb{R})$, where $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_r \geq 0$ such that*

$$A = V\Sigma U^* \,.$$

If $m < n$, the singular value decomposition *(SVD) is defined by taking A^* and deduce the result by taking the adjoint.*

We observe that the rank of A is equal to r, the number of nonzero singular values of A.

The SVD decomposition has also a geometric interpretation. Let S^{n-1} be the unit sphere in $\mathbb{R}^n$, *i.e.,*

$$S^{n-1} = \{x = (x_1, \dots, x_n)^t \in \mathbb{R}^n \,;\, \sum_i x_i^2 = 1\} \,.$$

Then the image of the unit sphere S^{n-1} by a nonsingular matrix A is an ellipsoid centered at the origin of $\mathbb{R}^n$ with semiaxes $\mu_i v_i$, where v_i is the ith column of V.

The SVD can be used to define the *pseudoinverse* of a matrix $A \in \Updownarrow, \backslash(\mathbb{K})$. Let $A = V\Sigma U^*$ be the SVD factorization of A, then the pseudoinverse of A is the matrix $A^+ \in \mathcal{M}_{n,m}(\mathbb{K})$ defined by $A^+ = U\Sigma^+ V^*$, with Σ^+ is the pseudoinverse of Σ obtained by replacing every nonzero entry in Σ by its reciprocal (its inverse). We have the following properties.

Proposition 4.1.2. *Let $A \in \mathcal{M}_{m,n}(\mathbb{K})$ be an arbitrary matrix having r positive singular values μ_i. The following identities hold:*

$$\begin{aligned} A^+A &= U\Sigma^+\Sigma U^* = \sum_{i=1}^r u_i\, u_i^* \,; \\ AA^+ &= V\Sigma\Sigma^+ V^* = \sum_{i=1}^r v_i\, v_i^* \,; \qquad (4.1) \\ A &= \sum_{i=1}^r \mu_i v_i u_i^* \,, \qquad \textit{and} \qquad A^+ = \sum_{i=1}^r \frac{1}{\mu_i} u_i v_i^* \,. \end{aligned}$$

Furthemore, if A has maximal rank ($r = n \leq m$), its pseudoinverse is then given by:

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

4.1.2 Vector and matrix norms

Norms will be primarily used to measure errors in matrix computations. We recall the definition of a norm on a vector space $\mathbb{K}^n$ (see Chapter 1, Section 1.1.4).

Definition 4.1.3. *A mapping denoted by* $\| \cdot\| \ : \mathbb{K}^n \to \mathbb{R}$ *is called a* norm *if it satisfies all of the following:*

1. *for any* $x \in \mathbb{K}^n$, $\|x\| \geq 0$, *(*nonnegativity*);*
2. $\|x\| = 0$ *if and only if* $x = 0$, *(*nondegeneracy*)*
3. *for every* $\lambda \in \mathbb{K}$, $\|\alpha x\| = |\alpha|\|x\|$, *(*multiplicativity*);*
4. *for every* $x, y \in \mathbb{K}^n$, $\|x + y\| \leq \|x\| + \|y\|$, *(*triangle inequality*).*

We recall the definition of the *inner product.*

Definition 4.1.4. *Suppose* V *is a vector space over the field* $\mathbb{R}$*. An inner product on* V *is a* positive definite bilinear *form* $(\cdot, \cdot)_V : V \times V \to \mathbb{R}$ *that satisfies the following properties, for all vectors* $x, y \in V$

1. $(x, y)_V = (y, x)_V$, *(*symmetry*);*
2. $(x, x)_V > 0$ *and* $(x, x)_V = 0 \Leftrightarrow x = 0$, *(*positive definiteness*)*

Proposition 4.1.3. *Suppose* $(\cdot, \cdot)_V$ *is an inner product on* V*. Then*

$$\forall x \in V\,, \qquad \|x\|_V = (x, x)_V^{1/2}\,,$$

defines a norm on V*. Furthermore we have the Cauchy-Schwarz inequality*

$$\forall x, y \in V\,, \qquad |(x, y)_V| \;\leq\; \|x\|_V\, \|y\|_V\,,$$

with equality only if x *and* y *are* linearly dependent*.*

Proof. The Cauchy-Schwarz inequality is trivial to prove in the case $y = 0$. Suppose $(y, y) \neq 0$, thus posing $\alpha = (y, y)_V^{-1}(x, y)_V$ gives

$$0 \;\leq\; (x - \alpha y, x - \alpha y)_V = (x, x)_V - (y, y)_V^{-1}|(x, y)_V|^2\,,$$

which is true only if $|(x, y)_V|^2 \leq (x, x)_V(y, y)_V$ and the results follows. □

If $V = \mathbb{K}^n$ is endowed with a scalar or Hermitian product $(\cdot, \cdot)$ (and the subscript $(\cdot, \cdot)_V$ is omitted on purpose), then the mapping $x \mapsto (x, x)^{1/2}$ defines a *norm* on $\mathbb{K}^n$. The following norms are most important:

- the Euclidean norm: $\|x\|_2 = (\sum_{i=1}^{n} |x_i|^2)^{1/2}$;
- the ℓ^p-norms defined as: $\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$, $p \geq 1$;
- the ℓ^∞-norm: $\|x\|_\infty = \max_{1\leq i\leq n} |x_i|$,

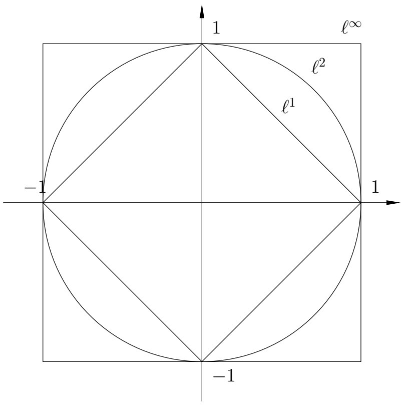

Fig. 4.1. *Unit disks of* $\mathbb{R}^2$ *for the norms* ℓ^1, ℓ^2 *and* ℓ^∞.

The unit disks of $\mathbb{R}^2$ for these norms are represented in Figure 4.1. A classical result about ℓ^p-norms is the *Hölder inequality*:

$$\|x^t y\|_1 \le \|x\|_p \|y\|_q\,, \quad \frac{1}{p}+\frac{1}{q}=1\,, \quad \forall x, y, \in \mathbb{K}^n\,.$$

and an important special case ($p = q = 2$) is the *Cauchy-Schwarz* inequality:

$$\|x^t y\|_1 \le \|x\|_2 \|y\|_2\,.$$

Theorem 4.1.3. *All norms are* equivalent *on* $\mathbb{K}^n$, i.e., *for all pairs of norms* $\|\cdot\|_\alpha$, $\|\cdot\|_\beta$ *on* $\mathbb{K}^n$, *there exist two constant* c_1 *and* c_2 *such that* $0 < c_1 \le c_2$, *and for all* $x \in \mathbb{K}^n$ *we have:*

$$c_1\|x\|_\alpha \le \|x\|_\beta \le c_2\|x\|_\alpha\,.$$

Example 4.1.1. For the previous ℓ^p-norms, we have the constants:

$$\|x\|_\infty \le \|x\|_p \le n^{1/p}\|x\|_\infty$$
$$\|x\|_2 \le \|x\|_1 \le \sqrt{n}\|x\|_2\,.$$

Since the space of matrices $\mathcal{M}_{m,n}(\mathbb{K})$ is a vector space *isomorphic* to $\mathbb{K}^{m\times n}$, the definition of a *matrix norm* shall be equivalent to the definition of a vector norm.

Definition 4.1.5. *A mapping* $\|\cdot\| : \mathcal{M}_{mn}(\mathbb{K}) \to \mathbb{K}$ *is a* matrix norm *if the following properties hold:*

1. $\|A\| \ge 0$, *and* $\|A\| = 0$ *if and only if* $A = 0$;
2. $\|\alpha A\| = |\alpha|\|A\|$, *for any scalar* α;
3. $\|A + B\| \le \|A\| + \|B\|$, *for all* $A, B \in \mathcal{M}_{m,n}(\mathbb{K})$.

Additionally, if $A, B \in \mathcal{M}_n(\mathbb{R})$, *the matrix norm satisfies*

4. $\|AB\| \le \|A\|\|B\|$ *and then the matrix norm is called a* submultiplicative *norm.*

The most frequently used matrix norms are

- the *Froebenius* (or *Schur*) norm, $\|A\|_F = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\right)^{1/2}$;
- the ℓ^p-norm, $\|A\|_p = \left(\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^p\right)^{1/p}$, for $p \ge 1$;
- the ℓ^∞-norm, $\|A\|_\infty = \max_{1\le i\le m, 1\le j\le n}|a_{ij}|$;

It is wise to define matrix norms that are subordinate to a vector norm in $\mathbb{K}^n$.

Definition 4.1.6 (induced norm). *Suppose* $\|\cdot\|$ *is a vector norm on* $\mathbb{K}^n$. *It induces the matrix norm* $\|\cdot\|$ *subordinate to the vector norm* $\|\cdot\|$ *defined by*

$$\|A\| = \sup_{x\in\mathbb{K}^n}\frac{\|Ax\|}{\|x\|}\,.$$

For convenience, we use the same notation for vector norms and matrix norms.

Proposition 4.1.4. *For any pair of vector norms* $\|\cdot\|_\alpha \in \mathbb{K}^m, \|\cdot\|_\beta \in \mathbb{K}^n$, *we have*

$$\|Ax\|_\beta \le \|A\|_{\alpha,\beta}\|x\|_\alpha$$

where $\|\cdot\|_{\alpha,\beta}$ *is a matrix norm defined by*

$$\|A\|_{\alpha,\beta} = \sup_{x\neq 0}\frac{\|Ax\|_\beta}{\|x\|_\alpha}\,,$$

that is subordinate to the vector norms.

Since the set $\{x \in \mathbb{K}^n\,;\ \|x\|_\alpha = 1\}$ is compact and $\|\cdot\|_\beta$ is continuous, it follows that

$$\|A\|_{\alpha,\beta} = \max_{\|x\|_\alpha=1}\|Ax\|_\beta = \|Ay\|_\beta$$

for some $y \in \mathbb{K}^n$ having a unit norm.

Proposition 4.1.5. *Froebenius and* ℓ^p*-norms satisfy the following properties, for* $A \in \mathcal{M}_{m,n}(\mathbb{K})$*:*

$$\|A\|_2 \le \|A\|_F \le \sqrt{n}\|A\|_2$$

$$\max_{i,j}|a_{ij}| \le \|A\|_2 \le \sqrt{mn}\max_{ij}|a_{ij}|$$

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

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

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

Theorem 4.1.4. *If $A \in \mathcal{M}_{m,n}(\mathbb{K})$, then there exists a unit ℓ^2-norm vector $z \in \mathbb{K}^n$ such that*

$$A^*Az = \mu^2 z \quad \textit{where} \quad \mu = \|A\|_2\,.$$

This result implies that $\|A\|_2^2$ is a *zero* of the polynomial

$$P(\lambda) = \det(A^*A - \lambda I_n)\,.$$

In particular, the ℓ^2-norm is the square root of the largest *eigenvalue* of A^*A:

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

Proof. Consider $\mathbb{K} = \mathbb{C}$. We have then

$$\|A\|_2^2 = \sup_{x\in\mathbb{C}^n} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \sup_{x\in\mathbb{C}^n} \frac{(Ax)^*(Ax)}{x^*x} = \sup_{x\in\mathbb{C}^n} \frac{x^*A^*Ax}{x^*x}\,.$$

We check that $(A^*A)^* = A^*(A^*)^* = A^*A$. Hence, the matrix A^*A is Hermitian and diagonalisable. There exists U, $U^*U = I_n$ and such that

$$UA^*AU = \mathrm{diag}(\mu_k)$$

where the values μ_k are the singular values of the matrix A and thus the eigenvalues of A^*A. It yields

$$\sup_{x\in\mathbb{C}^n} \frac{x^*A^*Ax}{x^*x} = \sup_{x\in\mathbb{C}^n} \frac{x^*U^*UA^*AU^*Ux}{x^*U^*Ux}$$

and by posing $y = Ux$ we have

$$\|A\|_2^2 = \sup_{y\in\mathbb{C}^n} \frac{y^*UA^*AU^*y}{y^*y} = \sup_{y\in\mathbb{C}^n} \frac{\sum_{k=1}^n \mu_k |y_k|^2}{\sum_{k=1}^n |y_k|^2}$$

where the μ_k are the eigenvalues of A^*A and by considering the vector $(0,\dots,|\mu_k|,\dots,0)^t$ we obtain $\|A\|_2^2 = \rho(AA^*)$. □

Corollary 4.1.2. *If $A \in \mathcal{M}_{m,n}(\mathbb{K})$, then $\|A\|_2 \le (\|A\|_1\|A\|_\infty)^{1/2}$.*

In the next paragraph, we consider square matrices only.

Definition 4.1.7 (convergence). *A sequence of matrices $(A_k)_{k\ge1}$ converges to a limit A for a matrix norm $\|\cdot\|$ if*

$$\lim_{k\to\infty} \|A_k - A\| = 0$$

and we write $A = \lim_{k\to\infty} A_k$.

Obviously, since $\mathcal{M}_n(\mathbb{K})$ is a vector space of finite dimension, all norms are equivalent and thus the notion of convergence does not depend on the norm considered. The following result gives a sufficient condition for a sequence of iterated powers of a matrix to converge to 0.

Lemma 4.1.7. *Suppose $A \in \mathcal{M}_n(\mathbb{C})$. The following assertions are equivalent:*

1. $\lim\limits_{n\to\infty} A^n = 0$,
2. $\lim\limits_{n\to\infty} A^n x = 0$, *for all* $x \in \mathbb{C}^n$;
3. $\rho(A) < 1$;
4. *there exists at least one subordinate norm such that* $\|A\| < 1$.

Proposition 4.1.6. *Let $A \in \mathcal{M}_n(\mathbb{K})$ be a matrix such that $\|A\|_p < 1$. Then, the matrix $I_n - A$ is nonsingular and we have*

$$(I_n - A)^{-1} = \sum_{n=0}^{+\infty} A^n \quad \textit{with} \quad \|(I_n - A)^{-1}\|_p \le \frac{1}{1 - \|A\|_p}\,.$$

Notice that $\|(I_n - A)^{-1} - I_n\|_p \le \|A\|_p/(1 - \|A\|_p)$. Hence if $\varepsilon \ll 1$ then $O(\varepsilon)$ perturbations in I_n induce $O(\varepsilon)$ pertubations in the inverse.

Lemma 4.1.8. *If A is nonsingular and $r = \|A^{-1}E\|_p < 1$, then $A + E$ is nonsingular and we have*

$$\|(A + E)^{-1} - A^{-1}\|_p \le \frac{\|E\|_p\,\|A^{-1}\|_p^2}{1 - r}\,.$$

4.1.3 Conditioning issues

We now turn to an important issue in numerical analysis, to see how numerically well-conditioned the problem at hand is. For instance, the *condition number* associated with the system $Ax = b$ provides a bound on the discrepency between the exact and the numerical solution, it is a measure of the *accuracy* of the computation, before considering *round-off* errors. It is indeed a property of a matrix.

Definition 4.1.8. *The quantity* $\mathrm{cond}(A) = \|A\|\cdot\|A^{-1}\|$ *is called the* condition number *of the matrix A with respect to the matrix norm $\|\cdot\|$ subordinate to the vector norm $\|\cdot\|$.*

Consider an invertible matrix $A \in \mathcal{M}_n(\mathbb{K})$ and a vector $b \in \mathbb{K}^n$. Let $x \in \mathbb{K}^n$ be the solution of the linear system $Ax = b$ given by $x = A^{-1}b$. Given a small perturbation δb of b, we denote $x + \delta x$ the solution of

$$A(x + \delta x) = b + \delta b\,.$$

For any vector norm $\|\cdot\|$ and its induced matrix norm $\|\cdot\|$, our goal is to bound the relative change $\|\delta x\|/\|x\|$ with respect to the relative error $\|\delta b\|/\|b\|$.By linearity and using the triangle inequality or vector norms, we write

$$A\delta x = \delta b \quad \Rightarrow \quad \|\delta x\| \le \|A^{-1}\|\|\delta b\|$$

and also

$$Ax = b \quad \Rightarrow \quad \|b\| \leq \|A\|\|x\|\,,$$

which is equivalent to

$$\frac{1}{\|x\|} \leq \|A\|\frac{1}{\|b\|}\,.$$

We can further rearrange these inequalities to

$$\frac{\|\delta x\|}{\|x\|} \leq \|A^{-1}\|\|A\|\frac{\|\delta b\|}{\|b\|} \leq \mathrm{cond}(A)\frac{\|\delta b\|}{\|b\|}\ .$$

We have the following properties.

Proposition 4.1.7. *Let $A \in \mathcal{M}_n(\mathbb{K})$ a square invertible matrix and $\|\cdot\|$ a matrix norm subordinate to a vector norm $\|\cdot\|$. Then we have*

1. $\mathrm{cond}(A^{-1}) = \mathrm{cond}(A)$*;*
2. $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$*, for any scalar $\alpha \in \mathbb{K}$;*
3. $\mathrm{cond}(I_n) = 1$*;*
4. $\mathrm{cond}(A) \geq 1$*;*
5. *a linear sytem $Ax = b$ is well-conditioned (resp. ill-conditioned) if* $\mathrm{cond}(A)$ *is low (resp. high).*

The condition number $\mathrm{cond}(A)$ defines the rate of changes in the solution x with respect to a change in the data b. It has a main drawback though. It involves $\|A^{-1}\|$ that is usually difficult to calculate. Nevertheless, for a normal matrix and the ℓ^2 matrix norm we have the following results.

Proposition 4.1.8. *Consider the ℓ^2 vector norm and its induced matrix norm.*

1. *if $A \in \mathcal{M}_n(\mathbb{K})$ then*
$$\mathrm{cond}_2(A) = \frac{\mu_{max}}{\mu_{min}}$$
where μ_{max}, μ_{min} are the maximal and minimal singular values of A, respectively.
2. *if $A \in \mathcal{M}_n(\mathbb{K})$ is a* Hermitian *matrix, we have*
$$\|A\|_2 = \rho(A)\,.$$
3. *if $A \in \mathcal{M}_n(\mathbb{R})$ is a* symmetric invertible *matrix, then*
$$\mathrm{cond}_2(A) = \frac{|\lambda_{max}|}{|\lambda_{min}|} = \rho(A)\rho(A^{-1})\,, \tag{4.2}$$
where λ_{max}, λ_{min} are maximal and minimal (by moduli) eigenvalues of A, respectively.
4. *for any* unitary *matrix U,* $\mathrm{cond}_2(U) = 1$*;*
5. *for any* unitary *matrix U,* $\mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(A)$*.*

Proof. We prove the identities 3 and 4.
Since A is a Hermitian matrix, there exists U such that $UU^* = I_n$ and $U^*AU = \mathrm{diag}(\lambda_i)$, and the $(\lambda_i)_i$ are the eigenvalues of A. We have

$$\|A\|_2^2 = \sup_{x\in\mathbb{K}^n} \frac{\|Ax\|_2^2}{\|x\|_2^2} = \sup_{x\in\mathbb{K}^n} \frac{(Ax)^*(Ax)}{x^*x}\,.$$

However,

$$\begin{aligned} x^*A^*Ax &= x^*UU^*A^*UU^*AUU^*x \\ &= (U^*x)^*(\mathrm{diag}(\lambda_i))^*(\mathrm{diag}(\lambda_i))U^*x \end{aligned}$$

and $U^*AU = (U^*AU)^*$. We pose $y = U^*x$ thus leading to

$$x^*A^*Ax = y^*\,\mathrm{diag}(\bar{\lambda}_i)\,\mathrm{diag}(\lambda_i)y\,,$$

and the results follows:

$$\|A\|_2^2 = \sup_{y\in\mathbb{K}^n} \frac{y^*\,\mathrm{diag}(|\lambda_i|^2)y}{y^*y} = \rho(A)^2\,.$$

Suppose $A \in \mathcal{M}_n(\mathbb{R})$ is a invertible symmetric matrix. We apply the previous result to the Hermitian inverse matrix A^{-1} to obtain

$$\|A^{-1}\|_2 = \rho(A^{-1}) = \frac{1}{|\lambda_{min}|}\,,$$

since $1/\lambda_i$ is an eigenvalue of A^{-1} and we deduce

$$\mathrm{cond}(A) = \|A\|_2\|A^{-1}\|_2 = \frac{|\lambda_{max}|}{|\lambda_{min}|}$$

that is the desired result. □

Remark 4.1.2. The identity (4.2) is optimal, as for any matrix norm we have

$$\mathrm{cond}(A) = \|A\|\|A^{-1}\| \geq \rho(A)\rho(A^{-1})\,.$$

Lemma 4.1.9. *Conditionings* cond_1, cond_2 *and* cond_∞ *are equivalent and we have, for any matrix* A

$$\begin{aligned} \frac{1}{n}\,\mathrm{cond}_2(A) \;&\leq\; \mathrm{cond}_1(A) \;\leq\; n\,\mathrm{cond}_2(A) \\ \frac{1}{n}\,\mathrm{cond}_\infty(A) \;&\leq\; \mathrm{cond}_2(A) \;\leq\; n\,\mathrm{cond}_\infty(A) \\ \frac{1}{\sqrt{n}}\,\mathrm{cond}_1(A) \;&\leq\; \mathrm{cond}_\infty(A) \;\leq\; n^2\,\mathrm{cond}_1(A)\,. \end{aligned}$$

Proof. The inequalities result from the equivalences between matrix and vector norms. □

4.2 Direct methods

The algorithms presented in this section are called *direct methods* because in the absence of rounding errors they would finally give the exact solution x of the problem $Ax = b$ after a finite number of elementary operations. The principle of direct methods is to find an invertible matrix M such that MA is an *upper triangular* matrix. This is called the *elimination procedure.* Hence, it remains to solve a linear system

$$MAx = Mb$$

using a *back-substitution* procedure, so called because the unknowns (the components of x) are computed in backward order, from x_n to x_1). Notice that in practice, the matrice M is not explicitly evaluated, only the matrix MA and the right-hand side vector Mb are calculated.

The matrix interpretation of the *Gauss pivoting* method is the *LU factorization* that shows that every invertible matrix can be decomposed as the product of a *lower triangular* matrix L by an *upper triangular* matrix U. This simplification can be further extended to the case of *symmetric positive definite* matrices, it is then called the *Cholesky factorization.*

Remark 4.2.1. We shall indicate that

1. the resolution of $Ax = b$, with $A \in \mathcal{M}_n(\mathbb{K})$ is not obtained by computing A^{-1} and then calculating $x = A^{-1}b$. The evaluation of A^{-1} is indeed equivalent to solving n linear systems

$$Ax_i = e_i\,, \qquad 1 \le i \le n\,,$$

where $(e_i)_{1\le i\le n}$ denote the basis vectors in $\mathbb{K}^n$.

2. if A is a *upper triangular* matrix, the resolution is trivial, we have then

$$\begin{cases} a_{11}x_1 + \cdots + a_{1n-1}x_{n-1} + a_{1n}x_n = b_1 \\ \qquad\qquad\vdots \\ a_{n-1n-1}x_{n-1} + a_{n-1n}x_n = b_{n-1} \\ a_{nn}x_n = b_n \end{cases}$$

and since $\prod_{i=1}^n a_{ii} = \det(A) \neq 0$, the system is solved by first computing x_n in the last equation and then x_{n-1} and so on, thus leading to

$$\begin{cases} x_n = a_{nn}^{-1}b_n \\ x_{n-1} = a_{n-1n-1}^{-1}(b_{n-1} - a_{n-1n}x_n) \\ \quad\vdots \\ x_1 = a_{11}^{-1}(b_1 - a_{12}x_2 - \cdots - a_{1n-1}x_{n-1} - a_{1n}x_n) \end{cases}$$

This *backward substitution* method requires $1 + 2 + \cdots + (n-1) = \frac{n(n-1)}{2}$ additions and $\frac{n(n-1)}{2}$ multiplications.

4.2.1 Cramer formulas

We consider a square linear system having n equations with n unknowns. We know that if the matrix $A \in \mathcal{M}_n(\mathbb{R})$ is nonsingular, then there exists a unique solution to the system $Ax = b$. The following proposition provides explicit formulas to compute the solution.

Proposition 4.2.1 (Cramer formulas). *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a nonsingular matrix. The solution $x = (x_1, \dots, x_n)^t$ to the linear system $Ax = b$ is given by its components*

$$x_i = \frac{\det(A_i)}{\det(A)}, \quad \forall\, 1 \le i \le n,$$

where $A_i = (a_1| \dots |a_{i-1}|b|a_{i+1}| \dots |a_n)$ is the matrix formed by replacing the ith column of A by the column vector b.

Proof. The system $Ax = b$ is equivalent to $\sum_{i=1}^n a_i x_i = b$, where $a_i \in \mathbb{R}^n$ represents the ith column in A. The components x_i are the entries of b in the basis formed by the columns a_i and thus we deduce that

$$\det(a_1| \dots |a_{i-1}|b|a_{i+1}| \dots |a_n) = x_i \det(A),$$

since the determinant is an alternate multilinear form. □

Nevertheless, having explicit formulas to calculate the solution x to the linear system is not always interesting, especially for large systems. Indeed, the number of multiplications required to compute the determinant of a square matrix A of order n is larger than $n!$. Consequently, more than $(n+1)!$ multiplications are needed to compute the solution of $Ax = b$. To give an idea of the computational cost, consider a system of $n = 10$ equations, it would require more than $4,000,000$ operations ! Hence, Cramer formulas are not used in practice because of their computational cost and we need to look for more efficient techniques of solving linear systems.

4.2.2 Gaussian elimination

A fundamental observation is that the following elementary operations can be performed on the system $Ax = b$ witout changing the set of solutions:

1. adding a multiple of the ith equation to the jth equation;
2. interchanging two equations (line swap);
3. multiplying one entire row by a nonzero scalar.

It is also possible to interchange two columns in A provided the corresponding interchanges are made in the components of the solution vector x.

Definition 4.2.1. *The first nonzero entry in $A \in \mathcal{M}_n(\mathbb{R})$ is called the* leading entry *(or the* pivot*) of the row. The matrix A is in* row echelon form *if*

1. all nonzero rows are above any zero rows;
2. the leading entry of a row i is strictly to the right of the leading entry of the row $i-1$.

The matrix A is in reduced row echelon form *if it is in row echelon form and*

3. every leading entry a_{ik} is 1 and a_{ik} is the only nonzero entry in this column k.

A system of linear equations $Ax = b$ is in row echelon if its augmented matrix $[A|b]$ (the entries in the last column of the matrix are the components of b, $a_{jn+1} = b_j$ and thus $[A|b] \in \mathcal{M}_{n,n+1}(\mathbb{R})$) is in row echelon form.

As mentioned, the idea behind the Gaussian elimination is to use the elementary operations to eliminate the unknowns in the system $Ax = b$, so as to obtain an equivalent triangular system, solved by backsubstitution.

We pose $A^{(1)} = A$ and $b^{(1)} = b$. For $i = 1, \ldots, n-1$, we calculate $A^{(i+1)}$ and $b^{(i+1)}$ such that the systems $A^{(i)}x = b^{(i)}$ and $A^{(i+1)}x = b^{(i+1)}$ are *equivalent*, where the matrices $A^{(i)}$ and $A^{(n)}$ have the following form

$$A^{(i)} = \begin{pmatrix} a_{11}^{(1)} & \ldots & \ldots & \ldots\ldots & a_{1n}^{(1)} \\ 0 & \ddots & & & \vdots \\ \vdots & & a_{ii}^{(i)} & & \vdots \\ \vdots & & a_{i+1i}^{(i)} & \ddots & \vdots \\ \vdots & & \vdots & \ddots & \vdots \\ 0 & \ldots & a_{ni}^{(i)} & \ldots\ldots & a_{nn}^{(i)} \end{pmatrix} \text{ and } A^{(n)} = \begin{pmatrix} a_{11}^{(n)} & \ldots & \ldots & \ldots & \ldots\ldots & a_{1n}^{(n)} \\ 0 & \ddots & & & & \vdots \\ \vdots & & a_{ii}^{(i)} & & & \vdots \\ \vdots & & & \ddots & & \vdots \\ \vdots & & & & \ddots & \vdots \\ 0 & \ldots & \ldots & \ldots & 0 & a_{nn}^{(n)} \end{pmatrix}$$

The two steps to solve the system $Ax = b$ are summarized as follows.

1. *Gaussian elimination*: using elementary operations, we modify $A^{(i)}$ and $b^{(i)}$ accordingly. Linear combinations of rows shall allow to cancel the entries of the ith column below the ith row in view of obtaining a upper triangular matrix. The two-steps algorithm is written as
   i) for $k \leq i$ and $j = 1, \ldots, n$, we pose $a_{kj}^{(i+1)} = a_{kj}^{(i)}$ and $b_k^{(i+1)} = b_k^{(i)}$;
   ii) for $k > i$, if $a_{ii}^{(i)} \neq 0$, we pose

$$\begin{cases} a_{kj}^{(i+1)} = a_{kj}^{(i)} - \dfrac{a_{ki}^{(i)}}{a_{ii}^{(i)}} a_{ij}^{(i)}\,, \quad \text{for } k = j, \ldots, n \\ b_k^{(i+1)} = b_k^{(i)} - \dfrac{a_{ki}^{(i)}}{a_{ii}^{(i)}} b_i^{(i)} \end{cases}$$

The matrix $A^{(i+1)}$ has the form defined hereabove and the two systems are equivalent. The diagonal entries $a_{11}, a_{22}^{(2)}, \ldots, a_{nn}^{(n)}$ which appear during th elmination procedure are called *pivotal elements*. Let A_k denote the kth leading principal submatrix of A. Since the determinant of a matrix does not change under row permutations, then

$$\det(A_k) = a_{11}^{(1)} \dots a_{kk}^{(k)}\,, \quad k = 1, \dots, n\,.$$

This implies that all pivotal elements $a_{ii}^{(i)}$, $1 \leq i \leq n$ in Gaussian elimination are nonzero if and only if $\det(A_k) \neq 0$, $k = 1, \dots, n$. In this case, after $(n-1)$ steps of elimination, we obtain the single equation

$$a_{nn}^{(n)} x_n = b_n^{(n)}\,.$$

We obtain finally an *upper triangular* system and we have also

$$\det(A) = A_{11}^{(1)} a_{22}^{(2)} \dots a_{nn}^{(n)}\,.$$

2. *backsubstitution*: denoting the upper triangular matrix $A^{(n)} = U$, the unknowns $(x_i)_{1 \leq i \leq n}$ can be computed as follows

$$x_n = \frac{b_n^{(n)}}{a_{nn}^{(n)}}\,, \quad \text{and then}$$
$$x_i = \frac{1}{a_{ii}^{(n)}} \left(b_i^{(i)} - \sum_{j=i+1}^{n} a_{ij}^{(n)} x_j \right)\,, \quad \text{for } i = n-1, \dots, 1\,.$$

Remark 4.2.2. Suppose that we face the case $a_{kk}^{(k)} = 0$ at step k in Gaussian elimination. If A is nonsingular, rhen in particular its first k columns are linearly independent, and so are the k columns of the reduced matrix. Hence, some $a_{ik}^{(k)}$, $i \leq k$ must be nonzeroes, say $a_{pk}^{(k)} \neq 0$. By interchanging rows k and p, this entry can be considered as *pivot* and the process can continue.

An important conclusion is given by the next proposition.

Proposition 4.2.2. *Any nonsingular matrix can be reduced to* triangular form *by Gaussian elimination (if appropriate row interchanges are made).*

Gaussian elimination requires $2/3n^3$ additions and multiplications and $n^2/2$ divisions. Thus, if $n = 10$ Gaussian elimination requires 700 operations (to be compared with Cramer formulas).

In principle, the reduced form of the matrix yields the rank of the matrix A. From the numerical point of view, *i.e.,* when performing floating point operations, it is in general necessary to perform row interchanges not only when a pivotal element is exactly zero, but also when its absolute value is *small*, to ensure the *numerical stability* of the subsequent operations. To this end, it is often interesting to chose between two strategies of Gaussian elimination.

1. in *partial pivoting*, the pivot is taken as the largest entry in the unreduced part of the column k: choose r as the smallest integer for which

   $$|a_{rk}^{(k)}| = \max_{k \le i \le n} |a_{ik}^{(k)}|\,,$$

   and interchanges rows k and r;
2. in *complete pivoting*, the element of the largest magnitude in the whole unreduced part of the matrix is chosen as pivot: choose r and s as the smallest integers for which

   $$|a_{rs}^{(k)}| = \max_{k \le i,j \le n} |a_{ij}^{(k)}|\,,$$

   and interchanges rows k and r, columns k and s.

The growth of the elements in the reduced matrix can be measured by the *growth ratio* defined as

$$g_n = \frac{\max_{i,j,k} |a_{ij}^{(k)}|}{\max_{i,j} |a_{ij}|}\,,$$

and for partial and complete pivoting, we have

$$|a_{ij}^{(k+1)}| \le |a_{ij}^{(k)}| + |\alpha_{ik}||a_{kj}^{(k)}| \le |a_{ij}^{(k)}| + |a_{kj}^{(k)}| \le 2\max_{i,j} |a_{ij}^{(k)}|\,,$$

with $\alpha_{i1} = \dfrac{a_{i1}}{a_{11}}$ and the bound $g_n \le 2^{n-1}$ follows by induction.

Theorem 4.2.1. *If Gaussian elimination is performed without pivoting, we have the following bounds on $g_n(A)$*

- *if A is nonsingular and* diagonally dominant*, i.e., $|a_{ii}| \ge \sum_{j \ne i} |a_{ij}|$, for all $i = 1, \dots, n$, then*

  $$g_n(A) \le 2\,.$$

- *if A is* symmetric *and* positive definite*, i.e., $A^t = A$ and $x^t A x > 0$, for all $x \ne 0$, then*

  $$g_n(A) \le 1\,.$$

- *if A is* totally positive *(*nonnegative*), i.e., if the determinant of every submatrix of A is positive (nonnegative), then*

  $$g_n(A) \le 1\,.$$

4.2.3 The LU factorization

For better understanding, we can rewrite the Gaussian elimination algorithm in a more abstract form, using matrix products only. Indeed, Gaussian elimination consists in decomposing the matrix A as the product of a lower triangular matrix L by an upper triangular matrix U: $A = LU$.

It can be observed that

$$A^{(n)} = L^{(n-1)}A^{(n-1)} = L^{(n-1)} \dots L^{(1)}A\,,$$

where $L^{(k)} = I_n - B^{(k)}$ with

$$B^{(k)} = \begin{pmatrix} 0 \dots 0 & 0 & 0 \dots 0 \\ \vdots \quad\ \vdots & \vdots & \vdots \quad\ \vdots \\ 0 \dots 0 & \alpha_{k+1}^{(k)} & 0 \dots 0 \\ \vdots \quad\ \vdots & \vdots & \vdots \quad\ \vdots \\ 0 \dots 0 & \alpha_n^{(k)} & 0 \dots 0 \end{pmatrix} \quad \text{with} \quad \alpha_i^{(k)} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\,, i = 1, \dots, n \tag{4.3}$$

as $A^{(k+1)} = L^{(k)}A^{(k)}$ and $b^{(k+1)} = L^{(k)}b^{(k)}$.

Lemma 4.2.1. *Let $B^{(k)} \in \mathcal{M}_n(\mathbb{R})$ be a square matrix of the form (4.3). Then, we have*

1. *$B^{(k)}B^{(l)} = 0$ for every $1 \leq k < l \leq n$;*
2. *for $L^{(k)} = I_n - B^{(k)}$, $L^{(k)}$ is* invertible *and $(L^{(k)})^{-1} = I_n + B^{(k)}$;*
3. *$L^{(k)}L^{(l)} = I_n - (B^{(k)} + B^{(l)})$, for every $1 \leq k < l \leq n$.*

Proof. Let $B^{(k)}$ and $B^{(l)} \in \mathcal{M}_n(\mathbb{R})$, we have for $1 \leq i, j \leq n$

$$(B^{(k)}B^{(l)})_{ij} = \sum_{m=1}^{n} (B^{(k)})_{im}(B^{(l)})_{mj}$$

but, for $m \neq k$, $(B^{(k)})_{im} = 0$ by assumption and thus

$$(B^{(k)}B^{(l)})_{ij} = (B^{(k)})_{ik}(B^{(l)})_{kj}\,.$$

Moreover, if $j \neq l$, we have $(B^{(l)})_{kj} = 0$ and thus $(B^{(k)}B^{(l)})_{ij} = 0$. If $j = l$, we have then

$$(B^{(k)}B^{(l)})_{il} = (B^{(k)})_{ik}(B^{(l)})_{kl}\,.$$

However, since $k < l$, the coefficient $(B^{(l)})_{kl}$ is null and thus $(B^{(k)}B^{(l)})_{il} = 0$. Finally, we obtain $(B^{(k)}B^{(l)} = 0$ for every $1 \leq k < l \leq n$. The remainder of the proof is more trivial. At first, we check that

$$(I_n - B^{(k)})(I_n + B^{(k)}) = I_n$$

and thus

$$(L^{(k)})^{-1} = (I_n - B^{(k)})^{-1} = I_n + B^{(k)}\,.$$

Then, we verify that, for $1 \leq k < l \leq n$,

$$L^{(k)}L^{(l)} = I_n - (B^{(k)} + B^{(l)})$$

and the results follows. □

Using this result, we write now

$$\begin{aligned}A &= (L^{(n-1)} \dots L^{(1)})^{-1} A^{(n)} \\ &= (L^{(1)})^{-1} \dots (L^{(n-1)})^{-1} A^{(n)} \\ &= (I_n + B^{(1)}) \dots (I_n + B^{(n-1)}) A^{(n)} \,.\end{aligned}$$

And, using a simiar property, changing $B^{(k)}$ in $-B^{(k)}$, we have

$$(I_n + B^{(1)}) \dots (I_n + B^{(n-1)}) = (I_n + B^{(1)} + \dots + B^{(n-1)})$$

thus leading to write

$$A = (I_n + B^{(1)} + \dots + B^{(n-1)}) A^{(n)} \,,$$

where the matrix $L = (I_n + B^{(1)} + \dots + B^{(n-1)})$ is a *lower triangular* matrix and the matrix $U = A^{(n)}$ is *upper triangular.*

We introduce the following result for the Gaussian elimination method without pivoting.

Lemma 4.2.2. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a nonsingular matrix, invertible. Suppose that all the principal submatrices of order 1 to $n-1$ are* invertible. *Then, there exists a unique* lower triangular *matrix L with unit diagonal coefficients and a* upper triangular *matrix U, invertible, such that*

$$A = LU \,.$$

Proof. We prove the uniqueness.
Suppose we have two decompositions $A = L_\alpha U_\alpha$ and $A = L_\beta U_\beta$ where L_α and L_β are lower triangular matrices with unit diagonal coefficients. Thus, L_α and L_β are invertible. Similarly, U_α and U_β are upper triangular invertible matrices. Then, we can write

$$L_\beta^{-1} L_\alpha = U_\beta U_\alpha^{-1} \,.$$

$L_\beta^{-1} L_\alpha$ is a lower triangular matrix with unit diagonal and $U_\beta U_\alpha^{-1}$ is an upper triangular matrix, it is thus diagonal and we have

$$L_\beta^{-1} L_\alpha = I_n \quad \text{and} \quad U_\beta U_\alpha^{-1} = I_n \,.$$

Hence, we have $L_\alpha = L_\beta$ and $U_\alpha = U_\beta$, which proves the uniqueness.

To apply the Gaussian elimination without pivoting, the coefficient $a_{ii}^{(i)}$ must be different of zero at each step. We prove the result by induction on nonzero eterminant for all submatrices.
We have $a_{11} \neq 0$ and thus $a_{11}^{(1)} \neq 0$.
We assume that $a_{kk}^{(k)} \neq 0$ for every $k = 1, \dots, i-1$. Using Gaussian elimination, we have then

$$A = A^{(1)} = (L^{(i-1)} \dots L^{(1)})^{-1} A^{(i)} \,.$$

By developing by blocks, we obtain

$$\begin{pmatrix} A_i & \cdot \\ \cdot & \cdot \end{pmatrix} = \begin{pmatrix} L_i & 0 \\ \cdot & \cdot \end{pmatrix} \begin{pmatrix} A_i^{(i)} & \cdot \\ \cdot & \cdot \end{pmatrix}$$

where A_i and $A_i^{(i)}$ are the principal submatrices of A and $A^{(i)}$, respectively and $L_i \in \mathcal{M}_i(\mathbb{R})$ is a lower triangular matrix with unit coefficients on its diagonal. Hence, we have

$$\det(A_i) = \det(L_i A_i^{(i)}) = \det(L_i) \det(A_i^{(i)}) = 1 \cdot a_{11}^{(1)} \dots a_{ii}^{(i)} \,.$$

By hypothesis, $\det(A_i) \neq 0$ and thus

$$\prod_{k=1}^{i} a_{kk}^{(k)} \neq 0 \quad \Rightarrow \quad a_{ii}^{(i)} \neq 0 \,.$$

We can thus apply a new step in Gaussian elimination. At the end, we will have

$$A = (L^{(1)})^{-1} L^{(1)} A^{(1)} = (L^{(1)})^{-1} A^{(2)} = \cdots = (I_n + B^{(1)} + \cdots + B^{(n-1)}) A^{(n-1)} \,.$$

We note $U = A^{(n-1)}$ and $L = (L^{(1)})^{-1} \dots (L^{(n-1)})^{-1}$ and thus we conclude to the result. □

Corollary 4.2.1. *Let $A \in \mathcal{M}_n(\mathbb{C})$ be a given* Hermitian positive definite *matrix. Then, A admits a LU decomposition.*

Proof. It consists in checking that the principal submatrices are invertible. These matrices are symmetric, positive definite and thus they are invertible. Using Lemma 4.2.2, we conclude about the LU decomposition of A. □

Remark 4.2.3. Suppose the problem $Ax = b$ need to be resolved for the same matrix A but for various data b, as it may happen when dealing with approximation methods. It is then important to avoid calculating the LU decomposition at each time, for efficiency purposes. Hence, the matrix $A^{(n)}$ must be computed once while the vectors $b^{(n)}$ and x must be evaluated for each right-hand side b. In practice, the matrices L and U are computed during the resolution of the first system and are kept in memory. The solution to every system $Ax = b$ is then obtained by solving two subproblems involving triangular matrices

1. find $y \in \mathbb{R}^n$, solving $Ly = b$;
2. find $x \in \mathbb{R}^n$, solving $Ux = y$.

We have the following result.

Theorem 4.2.2 (LU factorization). *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a given nonsingular matrix. Then, there exists a permutation matrix P such that the Gaussian elimination on the matrix PA can be carried out without pivoting giving the factorization*

$$PA = LU$$

where the pair of matrices (L, U), with L lower triangular with unit diagonal and U upper triangular, is uniquely determined.

Proof. cf. [Cia89]. □

In practice, the LU factorization of a matrix A can be achieved, if it exists, by setting the matrices L and U as follows

$$L = \begin{pmatrix} 1 & 0 & \dots & 0 \\ l_{21} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ l_{n1} & \dots & l_{nn-1} & 1 \end{pmatrix} \quad \text{and} \quad U = \begin{pmatrix} u_{11} & \dots & \dots & u_{1n} \\ 0 & u_{22} & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & u_{nn} \end{pmatrix}$$

and then, since L (resp. U) is lower (resp. upper) triangular, we write, for $1 \leq i, j \leq n$

$$a_{ij} = \sum_{k-1}^{n} l_{ik} u_{kj} = \sum_{k=1}^{m} l_{ik} u_{kj} \,,$$

where $m = \min(i, j)$. The entries l_{ij} and u_{ij} are easily calculated by reading in increasing order the columns of A (a complete description of the algorithm can be found in [All08], for instance).

Remark 4.2.4. 1. The LU factorization offers two advantages. It divides the solution of a linear system $Ax = b$ into two independent steps

a) the factorization $PA = LU$

b) the resolution of two systems: $Ly = Pb$ and $Ux = y$ for y and x, respectively. Indeed, we have

$$PAx = LUx = L(Ux) = Pb \,.$$

2. If the LU factorization of PA is known, then the solution x can be computed by solving the two systems $Ly = Pb$ and $Ux = y$ in $O(n^2)$ operations.
3. If the matrix A is *tridiagonal* then the number f operations required to solve the system $Ax = b$ involves $3(n-1)$ additions, $3(n-1)$ multiplications and $2n$ divisions [Cia89].

4.2.4 Cholesky method

The Cholesky method applied only to *symmetric* and *positive definite* matrices. We recall that a real matrix A is positive definite if all its eigenvalues are positive. If A is symmetric positive definite, it is invertible.

The Cholesky method[1] consists in determining a lower triangular matrix L such that

$$A = LL^t$$

so that solving the linear system $Ax = b$ becomes equivalent to solving two triangular systems

$$Ly = b \quad \text{and} \quad L^t x = y\,.$$

This decomposition allows also to compute the inverse matrix A^{-1} and the determinant, $\det(A) = \prod_{i=1}^{n} l_{ii}$. The matrix L appears sometimes as the *square root* of the matrix A.

Suppose the decomposition LL^t has been obtained, then we solve successively

1. the system $Ly = b$, written as follows

$$Ly = \begin{pmatrix} l_{11} & & 0 \\ \vdots & \ddots & \vdots \\ l_{n1} & \dots & l_{nn} \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

that can be rewritten componentwise as, for $i = 1, \dots, n$

$$\begin{aligned} l_{11}y_1 = b_1 \quad &\Rightarrow \quad y_1 = l_{11}^{-1} b_1 \\ l_{21}y_1 + l_{22}y_2 = b_2 \quad &\Rightarrow \quad y_2 = l_{22}^{-1}(b_2 - l_{21}y_1) \\ &\vdots \\ \sum_{j=1}^{n} l_{nj}y_j = b_n \quad &\Rightarrow \quad y_n = l_{nn}^{-1}\left(b_n - \sum_{j=1}^{n-1} l_{nj}y_j\right) \end{aligned}$$

that gives $(y_i)_{1\leq i \leq n}$.

2. the system $L^t x = y$ for x,

$$L^t x = \begin{pmatrix} l_{11} & & l_{1n} \\ & \ddots & \vdots \\ 0 & \dots & l_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

that gives $(x_i)_{1\leq i \leq n}$ using the classical *backward substitution*, as follows

[1] Named after the French mathematician André-Louis Cholesky (1875-1918).

$$\begin{aligned} l_{nn}x_n = y_n \quad &\Rightarrow \quad x_n = l_{nn}^{-1}y_n \\ &\vdots \\ \sum_{j-1}^{n} l_{j1}x_j = y_1 \quad &\Rightarrow \quad x_1 = l_{11}^{-1}\left(y_1 - \sum_{j=2}^{n} l_{j1}x_j\right) \end{aligned}$$

The existence and uniqueness of the LL^t decomposition of a symmetric and positive definite matrix A is given thereafter.

Theorem 4.2.3 (Cholesky factorization). *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a* symmetric *and* positive definite *matrix. There exists a unique real lower triangular matrix L, having positive diagonal entries $l_{ii} > 0$, for all $1 \le i \le n$, such that*

$$A = LL^t\,.$$

Proof. We have seen (Theorem 4.2.2) that there exists a pair of matrices (L, U) such that $PA = LU$. Here, we show that the decomposition can be obtained without any permutation. More precisely, Lemma (4.2.2) will allow us to conclude to the existence and uniqueness of the decomposition.

Since A is a positive definite matrix, all principal (symmetric) submatrices A_k are also positive definite and thus invertible. Assuming the LU decomposition of A, we write

$$\det(A_k) = \det((LU)_k) = \prod_{i=1}^{k} u_{ii}\,,$$

and we observe that $\det(A_k) > 0$ since all the eigenvalues of A are positive (and $det(A_k)$ correspond to the product of the eigenvalues of A_k), hence we have $u_{ii} > 0$, for all $1 \le i \le k$. By induction we show that $u_i > 0$ for all $1 \le i \le n$. Next, we introduce the real matrix $\Lambda = \mathrm{diag}(\sqrt{u_{ii}})$ in the LU factorization and we obtain

$$A = (L\Lambda)(\Lambda^{-1}U) = \tilde{L}\tilde{U} \qquad \text{with} \quad \tilde{L} = L\Lambda \quad \tilde{U} = \Lambda^{-1}U\,.$$

Furthermore, since A is symmetric, we write $\tilde{L}\tilde{U} = \tilde{U}^t\tilde{L}^t$ or, similarly

$$(\tilde{U}^t)^{-1}\tilde{L} = \tilde{L}^t\tilde{U}^{-1} \qquad \text{or} \qquad \tilde{L}(\tilde{U}^t)^{-1} = (\tilde{L})^{-1}\tilde{U}^t\,.$$

Since $\tilde{U}^t$ and $(\tilde{U}^t)^{-1}$ are lower triangular, $(\tilde{U}^t)^{-1}\tilde{L}$ is lower triangular. Likewise, $\tilde{L}^t\tilde{U}^{-1}$ is upper triangular. This matrix identity is only possible if both matrices are identical and equal to I_n

$$(\tilde{L}^t\tilde{U}^{-1})_{ii} = \sqrt{u_{ii}}\frac{1}{\sqrt{u_{ii}}} = 1$$

i.e., if $\tilde{L}^t\tilde{U}^{-1} = I_n$ or $\tilde{L}^t = \tilde{U}$. This proves the *existence* of (at least) one Cholesky factorization.

The *uniqueness* of the decomposition is a direct consequence of the uniqueness property of the LU decomposition. □

Remark 4.2.5. If the diagonal entries in the matrix L are not all positive, then the LL^t decomposition may not be unique.

The analysis of the factorization gives a practical algorithm for computing the matrix L. From the identity $A - LL^t$, we deduce

$$a_{ij} = (LL^t)_{ij} = \sum_{i=1}^{n} l_{ik} l_{jk} = \sum_{i=1}^{m} l_{ik} l_{kj}\,, \quad 1 \le i,j \le n$$

where $m = \min(i,j)$, and by noticing that $l_{pq} = 0$ for $1 \le p < q \le n$. The matrix A being symmetric, the previous identities must be satisfied for all $i \le j$, and the entries l_{ij} in L must be such that

$$a_{ij} = \sum_{k=1}^{i} l_{ik} l_{jk}\,, \qquad 1 \le i,j \le n\,.$$

Like for the LU factorization, the entries in L are computed by reading in increasing order the columns of A,

1. for $i = 1$, the first column of L is given by

$$\begin{aligned} a_{11} = l_{11} l_{11} \quad &\Rightarrow \quad l_{11} = \sqrt{a_1 1} \\ a_{12} = l_{11} l_{21} \quad &\Rightarrow \quad l_{21} = l_{11}^{-1} a_{12} \\ &\vdots \\ a_{1n} = l_{11} l_{n1} \quad &\Rightarrow \quad l_{n1} = l_{11}^{-1} a_{1n} \end{aligned}$$

2. for $i \ge 1$, we compute the column j of L, assuming the first $(j-1)$ columns of L have been previously computed.

$$\begin{aligned} a_{ii} = \sum_{k=1}^{i} l_{ik} l_{ik} \quad &\Rightarrow \quad l_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 \right)^{1/2} \\ a_{ii+1} = \sum_{k=1}^{i} l_{ik} l_{i+1k} \quad &\Rightarrow \quad l_{i+1i} = l_{ii}^{-1} \left(a_{ii+1} - \sum_{k=1}^{i-1} l_{ik} l_{i+1k} \right) \\ &\vdots \\ a_{in} = \sum_{k=1}^{n} l_{ik} l_{nk} \quad &\Rightarrow \quad l_{ni} = l_{ii}^{-1} \left(a_{in} - \sum_{k=1}^{i-1} l_{ik} l_{nk} \right)\,, \end{aligned}$$

and we know that

$$a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 > 0\,.$$

Remark 4.2.6. 1. The Cholesky method requires $n^3/6$ additions, $n^3/6$ multiplications, $n^2/2$ divisions and n square root computations.

2. The Cholesky method is also useful for computing the determinant of A, by noticing that
$$\det(A) = (\det(L))^2 .$$

Band matrices are commonly found in many applications (see Chapters 5, 6, **??**, for instance). Their peculiar structure reveals useful to spare memory and to improve computational efficiency during the factorization.

Definition 4.2.2. *A matrix $A \in \mathcal{M}_n(\mathbb{R})$ is said to be a* band *matrix if $a_{ij} = 0$ for $|i - j| > p$, $p \in \mathbb{N}$. The* bandwith *of A is then $2p + 1$.*

Proposition 4.2.3. *The Cholesky factorization preserves the* band structure *of the matrix A.*

Remark 4.2.7. Suppose the matrix A is nonsymmetric or is merely symmetric. We observe that, for A invertible,
$$Ax = b \Leftrightarrow A^t Ax = A^t b$$
since $\det(A^t) = \det(A) \neq 0$. A^tA is symmetric and positive definite. Indeed, given any matrix $B \in \mathcal{M}_n(\mathbb{R})$, B is symmetric if and only if $(Bx, y) = (x, By)$ and $(Bx, y) = (x, B^t y)$, for all $(x, y) \in \mathbb{R}^n \times \mathbb{R}^n$. Taking $B = A^tA$, we deduce that A^tA is symmetric. Moreover, since A is invertible, $(A^tAx, x) = (Ax, Ax) = |Ax|^2 > 0$, if $x \neq 0$. Hence, A^tA is symmetric and positive definite.

The Choleski method can be used to compute the solution x of $Ax = b$ when A is nonsymmetric. It consists in calculating A^tA and A^tb, then to solve $A^tAx = A^tb$ using the classical Choleski method. However, this factorization is more computationally expensive, it requires $O(4n^3/3)$ operations compared to $O(2n^3/3)$ operations for the LU factorization.

Remark 4.2.8. Regarding nonsquare systems of the form $Ax = b$, with $A \in \mathcal{M}_{m,n}(\mathbb{R})$, $b \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$, we shall notice that such systems have in general no solution, *i.e.,* there is more equations than unknowns and the systems are said to be *overdetermined.* Nevetheless, we look for x that solves the equations "at best". To this end, we define $f : \mathbb{R}^n \to \mathbb{R}$ by
$$f(x) = \|Ax - b\|_2^2$$
where $\|x\|_2 = (x, x)^{1/2}$ denotes the Euclidean norm on $\mathbb{R}^n$. Then we search for x solution of the *optimization* problem
$$f(x) \leq f(y) , \qquad \forall\, y \in \mathbb{R}^n .$$
We can develop $f(x)$ as follows

$$f(x) = (A^tAx, x) - 2(b, Ax) + (b, b)\,.$$

If there exists a solution to the optimization problem, it is given by the resolution of

$$A^tAx = A^tb$$

and we have seen that the Choleski method can be used to solve this system.

4.2.5 The QR decomposition

Another matrix factorization method, the QR decomposition, attempts to reduce the linear system $Ax = b$ to a triangular system. Here, the matrix A is factorized as the product of a *orthogonal* (unitary) matrix Q (such that $Q^t = Q^{-1}$ or $Q^* = Q^{-1}$) by an *upper triangular* matrix R. For any nonsingular matrix A, there exists an orthogonal matrix Q.

For solving the system $Ax = b$, the QR factorization determines an orthogonal matrix Q such that $Q^*A = R$ is upper triangular and compute Q^*b, and the solution x is then obtained by solving the triangular system $Rx = Q^*b$.

Theorem 4.2.4 (QR factorization). *Let $A \in \mathcal{M}_n(\mathbb{K})$ be a nonsingular matrix. There exists a unique pair (Q, R) of matrices, where Q is orthogonal and R is upper triangular with positive diagonal entries $r_{ii} > 0$, such that*

$$A = QR\,.$$

In order for us to prove this result, we need to introduce the *Gram-Schmidt orthogonalization* process. This process provides a convenient method for orthogonalizing a set of vector in a vector or an inner product space. We have the following statement, given here without proof.

Lemma 4.2.3 (Gram-Schmidt). *Consider a finite linearly independent set $S = \{x_1, \dots, x_k\}$ in $\mathbb{K}^n$, for $k \leq n$. Then, there exists an orthogonal set $S_o = \{y_1, \dot{;}y_k\}$ that spans the same k-dimensional subspace of $\mathbb{K}^n$ as S,* i.e.,

$$\text{span}\{y_1, \dots, y_k\} = \text{span}\{x_1, \dots, x_k\}\,.$$

Firstly, we define the *projection* operator, that projects a vector y orthogonally onto a vector x, by

$$\text{proj}_x(y) = \frac{(x, y)}{(x, x)}x = (x, y)\frac{x}{(x, x)}$$

where $(\cdot, \cdot)$ denotes the inner product. The Gram-Schmidt process is then defined by induction on k

$$\begin{aligned} y_1 &= x_1 & e_1 &= \frac{y_1}{\|y_1\|} \\ y_2 &= x_2 - \mathrm{proj}_{y_1}(x_2) & e_2 &= \frac{y_2}{\|y_2\|} \\ &\vdots & &\vdots \\ y_k &= x_k - \sum_{j=1}^{k-1} \mathrm{proj}_{y_k}(x_k) & e_k &= \frac{y_k}{\|y_k\|} \end{aligned}$$

At completion, $\{y_1, \dots, y_k\}$ is the system of orthogonal vectors and the set $\{e_1, \dots, e_k\}$ is *orthonormal*.

Proof (of Theorem 4.2.4). The uniqueness of the factorization is obtained by contradiction. Suppose there exist two decompositions

$$A = Q_1 R_1 \quad \text{and} \quad A = Q_2 R_2 \,.$$

Then $Q_2^* Q_1 = R_2 R_1^{-1}$ is upper triangular with positive diagonal entries. We have then

$$(Q_2^* Q_1)(Q_2^* Q_1)^* = I_n = LL^*$$

and we observe that $L = (Q_2^* Q_1)$ is a Cholesky factorization of I_n and thus $L = I_n$.

Since A is nonsingular, its column vectors $a_1, \dots, a_n$ form a basis of $\mathbb{R}^n$. We apply the Gram-Schmidt orthogonalization process to the $(a_i)_{1 \le i \le n}$. This yields an orthonormal basis $\{q_1, \dots, q_n\}$ defined by, for all $1 \le i \le n$

$$q_i = \frac{x_i}{\|x_i\|} \,, \quad \text{with} \quad x_i = a_i - \sum_{k=1}^{i-1} (q_k, a_i) q_k \,.$$

Then, posing $r_{ki} = (q_k, a_i)$ for $1 \le k \le i-1$, we deduce

$$a_i = \sum_{k=1}^{i} r_{ki} q_k \quad \text{and} \quad r_{ii} = \|a_i - \sum_{k=1}^{i-1} (q_k, a_i) q_k\| \, > \, 0 \,.$$

We denote by $Q = (q_1, \dots, q_n)$ the orthogonal matrix and by $R = (r_{ij})$ the upper triangular matrix, setting $r_{ki} = 0$ for $k > i$; we have then $A = QR$. □

Remark 4.2.9. The Gram-Schmidt algorithm for the QR factorization involves almost three times more operations than the Gaussian elimination. This drawback explains why it is not used in practice for solving square linear systems. However, it may be used for solving least square fitting problems involving rectangular matrices.

By extension, a rectangular matrix $A \in \mathcal{M}_{m,n}(\mathbb{R})$, with $m \ge n$, admits a QR decomposition if there exists an orthogonal matrix $Q \in \mathcal{M}_{m,m}(\mathbb{R})$ and an upper trapezoidal matrix $R \in \mathcal{M}_{m,n}(\mathbb{R})$ with rows $n+1$ to m all null, such that

$$A = Q \begin{pmatrix} R \\ 0 \end{pmatrix} .$$

Remark 4.2.10. On a computer, the Gram-Schmidt process gives vectors that are often not orthogonal because of rounding errors and therefore this process is said to be *numerically unstable* [Cia89]. Other orthogonalization methods are usually favored, like *Householder transformations* or *Given rotations* (see the references at the end of this chapter).

4.3 Iterative methods

The direct methods presented in the previous section are efficient, they provide the exact solution x (in the absence of rounding errors) of the linear system $Ax = b$. However, they require a large memory to store the matrix A. If the system results from the discretization of a partial differential equation, the matrix A is generally *sparse*, *i.e.,* contains a majority of zeroes, but can be very large. Iterative methods take advantage of the sparse structure of the matrix in memory and usually involve only matrix-vector products. Hence, a desired feature for a resolution method is to preserve at best the sparsity of the matrix A. This criterion is clearly discrediting most direct methods. In this regard, the LU decomposition *fills* the sparse structure of the matrix. If the bandwidth of each matrix L and U is $p+1$, the number of nonzero elements per line is then $2p+1$. However, we are considering here large sparse matrices $A \in \mathcal{M}_n(\mathbb{R})$, such that

$$m \ll 2p + 1 \ll n$$

where m denotes the maximal number of nonzero elements per line in A. In other words, the size required to store matrices L and U is much larger than the memory required to store the nonzero entries of A using a suitable data structure, like the CSR (Compressed Sparse Row) format.

4.3.1 General context

All algorithms considered in this section are called *iterative* because they compute a sequence $(x^{(k)})_{k\geq 1}$ of approximate solutions, given an initial data $x^{(0)}$, that converges under certain assumptions, toward the solution x of the problem $Ax = b$, when k tends to $+\infty$.

Since linear systems are considered, it would seem then reasonable to build the sequence of iterates of the general form

$$x^{(k+1)} = Bx^{(k)} + c \,, \quad k \geq 0 \,, \quad \text{for any } x^{(0)} \,. \tag{4.4}$$

Here, $B \in \mathcal{M}_n(\mathbb{R})$ and $c \in \mathbb{R}^n$ must be carefully chosen to ensure the convergence of the method. The next lemma gives a condition on the convergence of such method.

Lemma 4.3.1. *An iterative method of the form (4.4) converges to the solution x of $Ax = b$, for any choice of $x^{(0)}$, if and only if $\rho(B) < 1$.*

Proof. Having denoted $e^{(k)} = x^{(k)} - x$, the error at iteration k, we have

$$e^{(k)} = B^k e^{(0)}$$

and thanks to Lemma 4.1.7, it follows that $\lim\limits_{k\to\infty} B^k e^{(0)} = 0$ for any vector $e^{(0)}$ if and only if $\rho(B) < 1$.

Conversely, suppose that $\rho(B) > 1$, then there exists at least one eigenvalue $\lambda(B)| > 1$. Let $e^{(0)}$ be an eigenvector associated with λ, then $Be^{(0)} = \lambda e^{(0)}$ and $e^{(k)} = \lambda^k e^{(0)}$. This prevents $e^{(k)}$ from tending to 0 as $k \to \infty$. □

Remark 4.3.1. 1. An iterative method of the form (4.4) is not very useful in practice, because it requires the calculation of the inverse in $A^{-1}b$. Suppose the method is convergent. Then, by taking the limit in the induction relation, we obtain $x = Bx + c$ and, since $x = A^{-1}b$, this leads to set $c = (I_n - B)A^{-1}b$.

2. An iterative method of the form (4.4) is a special instance of iterative methods to find a *fixed point* for the mapping

$$F : x \in \mathbb{R}^n \to F(x) = (Bx + c) \in \mathbb{R}^n \,,$$

that is a *contraction* if $\|B\| < 1$, with respect to any matrix norm. By taking the matrix norm induced by the vector norm $\| \cdot \|$, we have

$$\|F(x) - F(y)\| \;\le\; \|B\| \|x - y\| \,.$$

Among the many iterative methods to solve linear systems, we restrict ourselves to a few algorithms that epitomize this class of methods.

4.3.2 Linear iterative methods

In this section, we consider only iterative methods in which the iterate $x^{(k+1)}$ is a function of $x^{(k)}$ only and not of $x^{(k-1)}, \dots x^{(1)}$, *i.e.,* $x^{(k+1)} = F(x^{(k)}$. A classical and general approach is based on a *regular decomposition* (or *splitting*) of the matrix A of the form

$$A = M - N \,,$$

where M is a nonsingular matrix, easy to invert in practice. We have then the set of equivalences

$$Ax = b \quad \Leftrightarrow \quad Mx = Nx + b \quad \Leftrightarrow \quad x = M^{-1}Nx + M^{-1}b \,.$$

We can then define the iterative method based on the splitting as follows

$$x^{(k+1)} = M^{-1}Nx^{(k)} + M^{-1}b\,, \quad k \geq 0\,, \quad \text{for any } x^{(0)}\,, \tag{4.5}$$

whose iteration matrix is $B = M^{-1}N = I_n - M^{-1}A$.

Remark 4.3.2. If the sequence $(x^{(k)})_{k\geq 1}$ converges to a limit x as $k \to \infty$, then by taking the limt in the induction relation we have $(M - N)x = Ax = b$.

Definition 4.3.1. *Let $A = M - N \in \mathcal{M}_n(\mathbb{K})$ with M nonsingular. The method of the form (4.5) is convergent if, for any $b \in \mathbb{K}^n$ and for any choice of $x^{(0)} \in \mathbb{K}^n$, the sequence $(x^{(k)})_{k\geq 1}$ converges to x, the solution of $Ax = b$ in $\mathbb{K}^n$. We have then*

$$\lim_{k\to\infty} \|x^{(k)} - x\| = 0\,,$$

where $\|\cdot\|$ denotes any norm in $\mathbb{K}^n$.

As previously, we consider the error at iteration $k + 1$, $e^{(k+1)} = x^{(k+1)} - x$. The method is convergent if $e^{(k+1)}$ tends to 0 as $k \to \infty$. However, since x is the unknown, $e^{(k+1)}$ cannot be evaluated. We would have come to the same conclusion by considering the *residual* $r^{(k+1)} = b - Ax^{(k+1)}$ instead of the error $e^{(k+1)}$. Nevertheless, we observe that

$$e^{(k+1)} = M^{(-1)}N(x^{(k)} - x) = M^{-1}Ne^{(k)} = Be^{(k)} = B^{k+1}e^{(0)}\,.$$

The next lemma provides convergence critera for the iterative method.

Lemma 4.3.2. *The iterative method (4.5) converges if and only if the* spectral radius *of B satisfies $\rho(B) < 1$.*

Proof. see Lemma 4.3.1. □

Jacobi, Gauss-Seidel and relaxation methods

In this section, we consider a few classical linear iterative methods, corresponding to different regular decompositions.

Suppose the diagonal entries a_{ii} of A are nonzero.

Definition 4.3.2. *The* Jacobi *iterative method is defined by the regular decomposition*

$$M = D\,, \qquad N = D - A = E + F\,,$$

where $D = \mathrm{diag}(a_{ii})$, and E denotes the lower triangular matrix of entries $e_{ij} = -a_{ij}$, if $i > j$ and F is the upper triangular matrix of entries $f_{ij} = -a_{ij}$ if $j > i$.

The iteration matrix B_J f the Jacobi method is then

$$B_J = M^{-1}N = D^{-1}(E+F) = I_n - D^{-1}A\,,$$

and the iterative algorithm is written as

$$Dx^{(k+1)} = (E+F)x^{(k)} + b\,, \quad k \geq 0\,, \qquad \text{for any } x^{(0)}\,.$$

Once $x^{(0)}$ has been chosen, the vector $x^{(k+1)}$ can be computed with the formula

$$a_{ii}x_i^{(k+1)} = b_i - \sum_{\substack{j>i \\ j\neq i}} a_{ij}x_j^{(k)}\,, \quad i = 1,\ldots,n\,. \tag{4.6}$$

The convergence of Jacobi algorithm is given by the following lemma.

Lemma 4.3.3. *For any initial vector $x^{(0)}$, the Jacobi method (4.6) converges if the matrix A is strictly diagonally dominant,* i.e., *if*

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

Proof. From Lemmas 4.1.7 and 4.3.2, we know that it is sufficient to find a subordinate norm $\|\cdot\|$ such that the matrix $B = D^{-1}(E+F)$ satisfies

$$\|B\| < 1\,.$$

Since A is strictly diagonally dominant, it is easy to show that

$$\sum_{j=1}^{n} |b_{ij}| = |a_{ii}|^{-1} \sum_{\substack{j=1 \\ j\neq i}}^{n} |a_{ij}| \; < \; 1\,,$$

and thus $\|B\|_\infty = \max\limits_{1\leq i\leq n} \sum\limits_{j=1}^{n} |b_{ij}| < 1$. □

Definition 4.3.3. *The* Gauss-Seidel *method is defined by the regular decomposition*

$$M = D - E\,, \qquad N = F\,,$$

where the matrices D, E and F are the same as with Jacobi method.

The iteration matrix B_{GS} is then

$$B_{GS} = M^{-1}N = (D-E)^{-1}F\,,$$

and the iterative algorithm can be written as

$$(D-E)x^{(k+1)} = Fx^{(k)} + b\,, \quad k \geq 0\,, \qquad \text{for any } x^{(0)}\,.$$

Once $x^{(0)}$ has been chosen, the vector $x^{(k+1)}$ can be computed with the formula

$$a_{ii}x_i^{(k+1)} = b_i - \sum_{j=i}^{i-1} a_{ij}x_j^{(k)} - \sum_{j=i+1}^{n} a_{ij}x_j^{(k)}\,, \quad i = 1, \dots, n\,. \tag{4.7}$$

And we have the following results for Gauss-Seidel method.

Proposition 4.3.1. *Let $A = M - N$ be a symmetric and positive definite matrix. If $M^t + N$ is symmetric and positive definite, then $\rho(M^{-1}N) < 1$.*

Lemma 4.3.4. *If A is symmetric and positive definite, then for any initial vector $x^{(0)}$, the Gauss-Seidel method converges to the solution x of $Ax = b$.*

Proof. We have to show that $M^t + N$ is symmetric and positive definite,

$$M^t + N = (D - E)^t = F = D - E^t + F\,,$$

and since A is symmetric, $E^t = F$ and thus $M^t + N = D$, that is symmetric and positive definite. The previous proposition allows to conclude. □

A variant consists in introducing a relaxation parameter ω in the Gauss-Seidel method to improve its convergence rate.

Definition 4.3.4. *Let $\omega \in \mathbb{R}^+$. The iterative* relaxation *method is defined by the regular decomposition*

$$M = \frac{D}{\omega} - E\,, \qquad N = \frac{1-\omega}{\omega}D + F\,,$$

where the matrices D, E and F are the same as with Jacobi method.

The iteration matrix B_ω is then

$$\begin{aligned} B_\omega = M^{-1}N &= \left(\frac{D}{\omega} - E\right)^{-1}\left(\frac{1-\omega}{\omega}D + F\right) \\ &= (I_n - \omega D^{-1}E)\left((1-\omega)I_n + \omega D^{-1}F\right)\,. \end{aligned}$$

and the iterative algorithm is written as

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

Remark 4.3.3. 1. The relaxation method is well defined if D is invertible;
2. if $\omega = 1$, the relaxation method coincides with the Gauss-Seidel method;
3. if $\omega > 1$ (resp. $\omega < 1$) it is called *over-relaxation* (resp. *under-relaxation*) method.

The convergence of the relaxation method results from the following property.

Proposition 4.3.2. *Let $A = M - N$ be the regular deomposition of a Hermitian positive definite matrix A, with M nonsingular. Then the matrix $M^* + N$ is Hermitian and, if $M^* + N$ is also positive definite, we have $\rho(M^{-1}N) < 1$.*

Lemma 4.3.5. *If A is a Hermitian positive definite matrix, then for any $\omega \in [0, 2[$ and any initial vector $x^{(0)}$, the relaxation method converges to x the solution of $Ax = b$.*

Proof. Matrix D is definite positive as A is positive definite. Hence, $\omega^{-1}D - E$ is nonsingular and, by noticing that $E^* = F$, we have

$$M^* + N = \omega^{-1}D -^* + \omega^{-1}(1-\omega)D + F = \omega^{-1}(2-\omega)D\,,$$

and we can conclude that $M^* + N$ is positive definite if and only if $\omega \in [0, 2[$ and the previous proposition leads to the result. □

Lemma 4.3.6. *For any matrix A, the spectral radius of B_ω is such that*

$$\rho(B_\omega) \geq |1-\omega|\,, \quad \textit{for all } \omega \neq 0\,,$$

and thus the relaxation method converges only if $\omega \in]0, 2[$.

Remark 4.3.4. The over-relaxation procedure can be also applied to Jacobi method and the algorithm (A.3) is then written as

$$x^{(k+1)} = x^{(k)} + \omega D^{-1}(b - Ax^{(k)})$$

and the vector $x^{(k+1)}$ is computed as

$$a_{ii}x_i^{(k+1)} = \omega \left(b_i - \sum_{\substack{j>i \\ j\neq i}}^{n} a_{ij}x_j^{(k)} \right) + (1-\omega)x_i^{(k)}\,, \quad i = 1, \ldots, n\,. \tag{4.8}$$

The corresponding iteration matrix is then

$$B_{J_\omega} = \omega B_J + (1-\omega)I_n\,.$$

Practical issues

When computing the sequence of approximate solutions, a practical stopping criterion is necessary. Since the solution vector x is unknown, it is of limited usefulness to evaluate the difference $\|x^{(k)} - x\|$ and to stop when the desired accuracy ε is reached. Since Ax is known, a simple convergence criterion could be $\|b - Ax^{(k)}\| \leq \varepsilon$. However, if $\|A^{-1}\|$ is large, we have then

$$\|x - x^{(k)}\| \leq \|A^{-1}\|\|b - Ax^{(k)}\| \leq \varepsilon\|A^{-1}\|\,,$$

and this term may not be small. For this reason, another criterion based on the residual is favored

$$\|b - Ax^{(k)}\| \leq \varepsilon\,\|b - Ax^{(0)}\| \qquad \Leftrightarrow \qquad \|r^{(k)}\| \leq \varepsilon\|r^{(0)}\|\,.$$

The computational cost of the iterative methods is at most $3/2n^2$ per iteration, which is favorable compared to direct methods if the number of iterations if small before n.

4.3.3 Gradient methods

In this section, we restrict ourselves to the case of real symmetric matrices, although the case of complex self-adjoint matrices can be handled quite similarly.

The linear iterative methods based on regular decomposition rely on parameters that may be difficult to set properly. In this regard, the *gradient method* (also called *steepest descent* method or *Richardson's* method) is defined as follows.

Definition 4.3.5. *Let $\alpha \in \mathbb{R}^*$. The* gradient method *is defined by the regular decomposition*

$$M = \alpha^{-1} I_n \qquad \textit{and} \qquad N = (\alpha^{-1} I_n - A)\,,$$

and the corresponding iterative algorithm can be written as

$$x^{(k+1)} = x^{(k)} + \alpha(b - Ax^{(k)})\,, \quad k \geq 1\,, \quad \text{for any } x^{(0)}\,. \tag{4.9}$$

Likewise, we have a convergence result for this method.

Lemma 4.3.7. *Let $(\lambda_i)_{1\leq i\leq n}$ denote the real eigenvalues of A and suppose $\lambda_i > 0$, for all $1 \leq i \leq n$. Then, the gradient method converges if and only if the parameter α is such that $0 < \alpha < 2/\lambda_{max}$.*
Moreover, the optimal parameter α_{opt} which minimizes the spectral radius of the iteration matrix $\rho(M^{-1}N)$ is given by

$$\alpha_{opt} = \frac{2}{\lambda_{min} + \lambda_{max}} \quad \textit{and} \quad \rho_{opt} = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{min} + \lambda_{max}} = \frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1}\,.$$

Proof. Since we assumed $0 < \alpha_{min} \leq \cdots \leq \lambda_{max}$, we can easily deduce that

$$-1 \;<\; 1 - \alpha\lambda_{max} \quad \Rightarrow \quad \alpha < \frac{2}{\lambda_{max}}\,.$$

The optimal value α_{opt} is obtained by considering the function $f : \lambda \mapsto |1 - \alpha\lambda|$. The analysis shows that f is decreasing on $]-\infty, 1/\alpha[$ and increasing on $]1/\alpha, +\infty[$. Hence, $\rho(M^{-1}N) = \max(|1 - \alpha\lambda_{min}|, |1 - \alpha\lambda_{max}|)$. On the other hand, the function $\alpha \in [0, 2/\lambda_{max} \mapsto \rho(M^{-1}N)$ admits a minimum at the point α_{opt} defined by $1 - \alpha_{opt} = \alpha_{opt}\lambda_{max} - 1$ (Figure 4.2). By substitution, the value ρ_{opt} s obtained. □

Next, we show another interpretation of the gradient method that will justify the denomination of projection method. To this end, we recall the following notions.

Definition 4.3.6. *Let f be a function from $\mathbb{R}^n$ into $\mathbb{R}$. The* gradient *of f at the point $x \in \mathbb{R}^n$ is defined by $\nabla f(x) = (Df(x))^t \in \mathbb{R}^n$ and we denote*

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

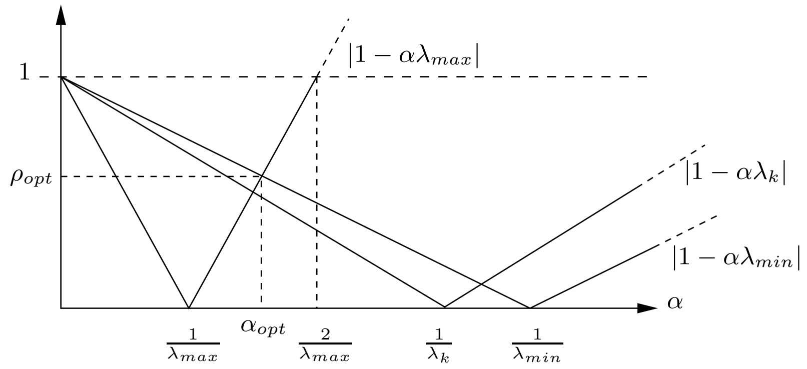

Fig. 4.2. *Spectral radius of $I_n - \alpha A$ as a function of α.*

For $(x,y) \in \mathbb{R}^n \times \mathbb{R}^n$, we have then

$$Df(x)y = \sum_{j=i}^{n} \frac{\partial f}{\partial x_j}(x) y_j = (\nabla f(x), y)\,.$$

And we have the following results.

Lemma 4.3.8 (Existence). *Let f be a function from $\mathbb{R}^n$ into $\mathbb{R}$ such that*

(i) f is continuous;
(ii) $f(x) \to +\infty$ when $\|x\| \to +\infty$

Then, there exists a point $x_0 \in \mathbb{R}^n$ such that $f(x_0) \le f(y)$ for all $y \in \mathbb{R}^n$.

Notice that this result does not hold if $f : E \to \mathbb{R}$, where E is a Banach space (*i.e.,* if the dimension of E is not finite). The point x_0 is called a *minimum* of f, or f is said to attain its minimum at x_0.

Definition 4.3.7 (Convexity). *Let f be a function from a vector space E into $\mathbb{R}$. The function f is called* convex *if, for any two points $(x,y) \in E^2$, $x \neq y$ and for $t \in [0,1]$*

$$f(tx + (1-t)y) \;\le\; tf(x) + (1-t)f(y)\,,$$

Furthemore, the function f is called strictly convex *if, for $t \in]0,1[$*

$$f(tx + (1-t)y) \;<\; tf(x) + (1-t)f(y)\,.$$

Corollary 4.3.1 (Existence and uniqueness). *Let f be a strictly convex function from $\mathbb{R}^n$ into $\mathbb{R}$ satisfying the hypothesis of Lemma 4.3.8. Then, there exists a unique point $x_0 \in \mathbb{R}^n$ such that*

$$f(x_0) \;=\; \inf_{y \in \mathbb{R}^n} f(y)\,.$$

We have the following characterization of a convex function.

Proposition 4.3.3. *Suppose E is a normed vector space on $\mathbb{R}$ and let consider a function $f \in C^1(E, \mathbb{R})$. Then, for any two points $(x, y) \in E^2$,*

1. *f is convex if and only if $f(y) \geq f(x) + Df(x)(y-x)$;*
2. *f is strictly convex if and only if $f(y) > f(x) + Df(x)(y-x)$, $x \neq y$.*

Proposition 4.3.4. *Let f be a C^1 continuous convex function from $\mathbb{R}^n$ into $\mathbb{R}$ and let $x_0 \in \mathbb{R}^n$. Then,*

$$f(x_0) = \inf_{y \in \mathbb{R}^n} f(y) \quad \Leftrightarrow \quad \nabla f(x_0) = 0\,.$$

Next, we introduce an optimization problem.

Minimization of a quadratic functional.

Let consider a symmetric matrix $A \in \mathcal{M}_n(\mathbb{R})$, a vector $b \in \mathbb{R}^n$ and the function f from $\mathbb{R}^n$ into $\mathbb{R}$ defined by

$$f(x) = \frac{1}{2}(Ax, x) - (b, x) = \frac{1}{2}\sum_{i,j=1}^{n} a_{ij}x_i x_j - \sum_{i=1}^{n} b_i x_i\,, \tag{4.10}$$

Then, $f \in C^\infty(\mathbb{R}^n, \mathbb{R})$. The evaluation of the gradient of f shows that

$$\nabla f(x) = \frac{1}{2}(Ax + A^t x) - b\,,$$

and hence, since A is symmetric we have

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

Indeed, computing the kth partial derivative of f yields

$$\begin{aligned}\frac{\partial f}{\partial x_k}(x) &= a_{kk}x_k + \frac{1}{2}\sum_{i \neq k} a_{ik}x_i + \frac{1}{2}\sum_{i \neq k} a_{ki}x_i - b_k \\ &= \sum_{i} a_{ik}x_i - b_k = (Ax - b)_k\,.\end{aligned}$$

Proposition 4.3.5. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a* symmetric *and* positive definite *matrix and $b \in \mathbb{R}^n$. Then, there exists a unique point $x_0 \in \mathbb{R}^n$ such that $f(x_0) \leq f(x)$ for any $x \in \mathbb{R}^n$, and x_0 is the unique* minimum *of f that is solution of the linear system $Ax = b$.*

Remark 4.3.5. The *quadratic form* f defined Equation (4.10), is often called the *energy* of system $Ax = b$. If x_0 is a solution of the system, then

$$f(y) = f(x_0 + (y - x_0)) = f(x_0) + \frac{1}{2}(A(y - x_0), y - x_0)\,, \quad \text{for all } y \in \mathbb{R}^n\,,$$

and we conclude that $f(y) > f(x_0)$ if $y \neq x_0$. This shows that x_0 is a minimizer of the functional f.

The following result provides an interpretation of the minimum as a projection on a vector subspace.

Corollary 4.3.2. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a real* symmetric *and* positive definite *matrix and let f be the function defined by (4.10). Let E be a vector subspace of $\mathbb{R}^n$. There exists a unique $x_0 \in E$ such that*

$$f(x_0) \leq f(x)\,, \quad \textit{for all } x \in E\,.$$

Moreover, x_0 is the unique vector of E such that

$$(Ax_0 - b, y) = 0\,, \quad \textit{for all } y \in E\,.$$

If we denote by P the matrix representation of the orthogonal projection on the vector subspace E in the canonical basis, we have

$$\min_{x\in E} f(x) = \min_{y\in\mathbb{R}^n} f(Py) \geq \min_{x\in\mathbb{R}^n} f(x)\,.$$

Hence, we have $f(Py) = \frac{1}{2}(P^tAPy, y) - (P^tb, y)$ and the matrix P^tAP is nonnegative. Moreover, P^tb belongs to $\mathrm{Im}(P^tAP)$ and the minimum of $f(Py)$ is attained by all x such that $P^tAPx = P^tb$. We can then enounce the following result.

Theorem 4.3.1. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a* symmetric *and* positive definite *matrix, and let f be the function defined on $\mathbb{R}^n$ by (4.10). Then,*

1. *$x \in \mathbb{R}^n$ is the* minimum *of f if and only if $\nabla f(x) = 0$,*
2. *suppose $\nabla f(x) \neq 0$ at $x \in \mathbb{R}^n$. Then, for any value $\alpha \in]0, 2/\rho(A)[$, we have*
$$f(x - \alpha\nabla f(x)) < f(x)\,.$$

Proof. (admitted here, see [All08] for instance). □

This result yields the definition of an iterative method to find the minimum of the function f. Namely, we define the sequence of points $(x_k)_{k\geq 1}$ such that the sequence $(f(k))_{k\geq 1}$ is decreasing

$$x_{k+1} = x_k - \alpha\nabla f(x_k) = x_k + \alpha(b - Ax_k)\,, \quad \text{given } x_0\,,$$

and we observe that this relation is exactly identical to the algorithm defined by Equation (4.9). In other words, solving a linear system $Ax = b$ for which the matrix A is symmetric and positive definite, is equivalent to minimizing a quadratic functional. This concept will be used to define optimization algorithms.

Gradient descent methods

Let consider $E = \mathbb{R}^n$ and $f \in C^0(E, \mathbb{R})$. We assume that there exists a point $x_0 \in \mathbb{R}^n$ such that $f(x_0) = \inf_{y \in E} f(y)$. The goal is to compute this point x_0.

Definition 4.3.8. *Let $f \in C^0(E, \mathbb{R})$ and $E = \mathbb{R}^n$. Let $x \in E$.*

1. *A vector $w \in E \backslash \{0\}$ is a* descent direction *at x if there exists $\alpha_0 > 0$ such that*
$$f(x + \alpha w) \leq f(x), \quad \textit{for all } \alpha \in [0, \alpha_0].$$
2. *A vector $w \in E \backslash \{0\}$ is a* strict descent direction *at x if there exists $\alpha_0 > 0$ such that*
$$f(x + \alpha w) \leq f(x), \quad \textit{for all } \alpha \in]0, \alpha_0].$$

Using this definition, we can suggest the following *descent method* for finding the point x_0 that realize $f(x_0) = \inf_{y \in E} f(y)$. It consist in constructing the sequence $(x^{(k)})_{k \geq 1}$ as follows

1. set $x^{(0)} = x_0 \in E$;
2. given $x^{(0)}, \ldots, x^{(k-1)}$
   a) find a strict descent direction w_k of $x^{(k)}$,
   b) define $x^{(k+1)} = x^{(k)} + \alpha_k w_k$, with α_k suitably chosen.

The next result explains how to chose the parameter α_k.

Proposition 4.3.6. *Let $E = \mathbb{R}^n$, $f \in C^1(E, \mathbb{R})$, $x \in E$ and $w \in E \backslash \{0\}$. We have the following properties*

(i) if w is a descent direction at x then $(w, \nabla f(x)) \leq 0$,
(ii)if $\nabla f(x) \neq 0$ then $w = -\nabla f(x)$ is a strict descent direction at x.

Proof. We introduce the function $\varphi \in C^1(\mathbb{R}, \mathbb{R})$ defined by

$$\varphi(\alpha) = f(x + \alpha w).$$

(i) We have then $\varphi'(\alpha) = (\nabla f(x + \alpha w), w)$. Since w is a descent direction at x, we have $\varphi(\alpha) < \varphi(0)$ for any $\alpha \in]0, \alpha_0]$ and thus

$$\frac{\varphi(\alpha) - \varphi(0)}{\alpha - 0} \leq 0, \quad \text{for all } \alpha \in]0, \alpha_0],$$

By passing to the limit, when $\alpha \to 0$, we deduce that $\varphi'(0) \leq 0$, *i.e.,* $(\nabla f(x), w) \leq 0$.
(ii) Let $w = -\nabla f(x) \neq 0$. The objective is to find $\alpha_0 > 0$ such that if $\alpha \in]0, \alpha_0]$ then $f(x + \alpha w) < f(x)$. This is equivalent to showing that $\varphi(\alpha) < \varphi(0)$. We have

$$\varphi'(0) = (\nabla f(x), w) = -|\nabla f(x)|^2 < 0.$$

Since φ' is continuous, there exists $\alpha_0 > 0$ such that, if $\alpha \in]0, \alpha_0]$, $\varphi'(\alpha) < 0$. However, if $\alpha \in]0, \alpha_0]$, then

$$\varphi(\alpha) - \varphi(0) = \int_0^\alpha \varphi'(t)dt \, < \, 0 \, ,$$

and thus $\varphi(\alpha) < \varphi(0)$ for any $\alpha \in]0, \alpha_0]$, that shows that w is a strict descent direction at x. □

And we deduce easily the iterative algorithm of Definition (4.3.5) for computing $x^{(k+1)}$ with $w_k = -\nabla f(x^{(k)}) = b - Ax^{(k)}$.

The main drawback of this iterative gradient method is that it involves a parameter α, chosen constant at each step. We have seen previously (Lemma 4.3.7) that the optimal parameter α_{opt} requires the knowledge of the smallest and the largest eigenvalues of A. Hence, this algorithm is of limited practical uselfulness. However, it can be improved by chosing at each step a different coefficient α_k that minimizes $f(x^{(k)} - \alpha \nabla f(x^{(k)}))$.

Definition 4.3.9. *The gradient method with* variable step size *for solving the linear system $Ax = b$ is defined by the algorithm*

$$x^{(k+1)} = x^{(k)} + \alpha_k (b - Ax^{(k)}) \, , \quad \textit{for any } x^{(0)} \, ,$$

where α_k is set as the minimum of the function $f(x^{(k)} - \alpha \nabla f(x^{(k)}))$, i.e.,

$$f(x^{(k)} - \alpha_k \nabla f(x^{(k)})) \leq f(x^{(k)} - \alpha \nabla f(x^{(k)})) \, , \quad \textit{for all } \alpha \geq 0 \, .$$

It remains to find a computational expression for α_k.

Lemma 4.3.9. *Let A be a symmetric and positive definite matrix. For the gradient algorithm with variable step, there exists a unique optimal step size α_k defined by*

$$\alpha_k = \frac{(w^{(k)}, w^{(k)})}{(w^{(k)}, Aw^{(k)})} \, , \quad \textit{where } w^{(k)} = b - Ax^{(k)} \, .$$

Proof. To find the parameter α_k, let us write the quadratic functional (4.10) for $x^{(k+1)}$:

$$\begin{aligned} f(x^{(k+1)}) &= \frac{1}{2}(A(x^{(k)} - \alpha_k w^{(k)}), x^{(k)} - \alpha_k w^{(k)}) - (b, x^{(k)} - \alpha_k w^{(k)}) \\ &= f(x^{(k)}) - \alpha_k (w^{(k)}, w^{(k)}) + \frac{1}{2}\alpha_k^2 (Aw^{(k)}, w^{(k)}) \, . \end{aligned}$$

Differentiating this function with respect to α_k and setting it to zero (to find the minimum), gives the expected value of α_k. □

Remark 4.3.6. 1. For a more general function f, there is usually no explicit formula for the parameter α_k. It can be obtained numerically by using Newton's method to find the roots of the gradient function.

2. The convergence of the gradient method can be quite slow (linear in general), especially if the condition number $\text{cond}_2(A)$ is large.

4.3.4 Projection methods

The research on unconstrained optimization problems had a positive impact on the improvement of gradient methods and effectively led to the conjugate gradient method and Krylov subspace methods[2]. We introduce first this important notion.

Definition 4.3.10. *Let $r \in \mathbb{R}^n$ and $A \in \mathcal{M}_n(\mathbb{R})$. We call* order-$k$ Krylov space *generated by the vector r and associated with the matrix A, the vector subspace of $\mathbb{R}^n$, denoted by $\mathcal{K}_k(A, r)$ (or simply by $\mathcal{K}_k$ if there is no ambiguity), spanned by the $k+1$ vectors $\{r, Ar, \dots, A^k r\}$,* i.e.,

$$\mathcal{K}_k(A,r) = \mathrm{span}(r, Ar, \dots, A^k r)\,.$$

By definition, $\dim(\mathcal{K}_k(A,r)) = k+1$.

Lemma 4.3.10. *The sequence of Krylov spaces $(\mathcal{K}_k)_{k\geq 0}$ is increasing,* i.e.,

$$\mathcal{K}_k \subset \mathcal{K}_{k+1}\,, \quad \textit{for all } k \geq 0\,.$$

Consider for instance the iterate $x^{(k+1)}$ of the gradient method

$$x^{(k+1)} = x^{(k)} + \alpha_k w^{(k)} = x^{(0)} + \sum_{j=0}^{k} \alpha_j w^{(j)}\,, \quad \text{with } w^{(k)} = (b - Ax^{(k)})\,.$$

The vector $w^{(k)}$ is called the *residual* and it can be seen that

1. $w^{(k)}$ belongs to the Krylov space $\mathcal{K}_k$ corresponding to the initial residual $w^{(0)}$;
2. $x^{(k+1)}$ belongs to the following affine space $W_k = [x^{(0)} + \mathcal{K}_k]$ defined as the set of vectors v such that $v - x^{(0)}$ belongs to $\mathcal{K}_k$, *i.e.,*

$$W_k = \left\{ v = x^{(0)} + y\,,\ \ y \in \mathcal{K}_k(A, w^{(0)}) \right\}\,.$$

The term $\sum_{j=0}^{k} \alpha_j w^{(j)}$ is a polynomial in A of degree less than k and the approximate solution x of $Ax = b$ is searched for in the space W_k. It seems wise to devise a method that searches for the approximate solution of the form

$$x^{(k+1)} = x^{(0)} + p_k(A) w^{(0)}\,, \tag{4.11}$$

where $p_K(\cdot)$ is a suitably defined polynomial, for instance such that $x^{(k+1)}$ represents the best approximation of x in W_k, in a sense that remains to be specified. A method that looks for the solution $x^{(k+1)} \in W_k$ of the form (4.11) is then called a *Krylov method.*

[2] Named after the Russian mathematician Alexei Krylov (1863-1945).

Proposition 4.3.7. *Let $(x^{(k)})_{k\geq 0}$ be a sequence in $\mathbb{R}^n$ and let $\mathcal{K}_k$ be the Krylov subspace generated by the residual vector $w^{(0)} = b - Ax^{(0)}$ and associated with A. If $x^{(k+1)} \in W_k = [x^{(0)} + \mathcal{K}_k]$, then $w^{(k+1)} \in \mathcal{K}_{k+1}$.*

Proof. If $x^{(k+1)} \in W_k$ there exists $(\alpha_j)_{0\leq j\leq k}$ such that

$$x^{(k+1)} = x^{(0)} + \sum_{j=0}^{k} \alpha_j A^j w^{(0)} \,,$$

and by multiplying by A and subtracting to b, we obtain

$$w^{(k+1)} = w^{(0)} - \sum_{j=0}^{k} \alpha_j A^{j+1} w^{(0)} \,.$$

Consequently, $w^{(k+1)} \in \mathcal{K}_{k+1}$ which is the expected result. □

Definition 4.3.11. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive definite matrix.*

1. *two vectors x and y in $\mathbb{R}^n\backslash\{0\}$ are A-*orthogonal*, or A-*conjugate*, if $(Ax, y) = (y, Ax) = 0$;*
2. *A collection $(w^{(0)}, \dots, w^{(p)})$ in $\mathbb{R}^n\backslash\{0\}$ is A-*conjugate* if $(w^{(i)}, Aw^{(j)}) = 0$ for any pair $(i, j) \in \{1, \dots, p\}^2$, $i \neq j$.*

Remark 4.3.7. Since A is symmetric positive definite, $(\cdot, A\cdot)$ defines a inner product

$$(x, y)_A = (Ax, y) = (x, Ay) \,,$$

and the corresponding norm $\|x\|_A = (Ax, x)^{1/2}$ is called the *energy norm*.

Proposition 4.3.8. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive definite matrix, $(w^{(0)}, \dots, w^{(p)})$ a collection of vectors in $\mathbb{R}^n$. Then,*

1. *if the set $(w^{(0)}, \dots, w^{(p)})$ is A-conjugate then the vectors are linearly independent;*
2. *if the case $p = n$, if the set $(w^{(0)}, \dots, w^{(n)})$ is A-conjugate then it forms a basis of $\mathbb{R}^n$.*

The Conjugate Gradient Method

In this section, we assume the matrices to be symmetric and positive definite. We pointed out that Krylov methods attempt to find the solution $x^{(k+1)}$ in the space W_k. Obviously, there are infinitely many possible choices for $x^{(k+1)}$ in the affine space W_k. To remove the ambiguity, we can chose $x^{(k+1)} \in W_k$ such that the residual $w^{(k+1)}$ is orthogonal to $\mathcal{K}_k$. This projection allows to define an iterative method.

The Hestenes-Stiefel *conjugate gradient method*[3] starts from an initial guess $x^{(0)}$ of the solution x and the corresponding residual $w^{(0)} = b - Ax^{(0)}$. And the A-conjugate directions of the method can be characterized as follows. At iteration $k+1$, it considers the Krylov subspace $\mathcal{K}_k$ of dimension $k+1$

$$\mathcal{K}_k = \mathrm{span}(w^{(0)}, \dots, A^k w^{(0)}) ,$$

and realizes the projection of $w^{(k)}$ on $\mathcal{K}_k$.

Definition 4.3.12 (Conjugate gradient method). *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive definite matrix, $b \in \mathbb{R}^n$. We define the conjugate gradient algorithm as follows*

1. **Initialization***:*
   $x^{(0)} \in \mathbb{R}^n$*, given*
   $w^{(0)} = b - Ax^{(0)}$*;*
   $d^{(0)} = w^{(0)}$*;*
2. **Iteration:**
   for $k \geq 0$*, until* $\|w^{(k)}\| < \varepsilon$*,*
   construct the sequence

$$\left.\begin{aligned} \alpha_k &= \frac{(w^{(k)}, d^{(k)})}{(d^{(k)}, Ad^{(k)})} \\ x^{(k+1)} &= x^{(k)} + \alpha_k d^{(k)} \\ w^{(k+1)} &= w^{(k)} - \alpha_k A d^{(k)} \end{aligned}\right\} \textit{compute new solution}$$

$$\left.\begin{aligned} \beta_k &= \frac{(Ad^{(k)}, w^{(k+1)})}{(Ad^{(k)}, d^{(k)})} \\ d^{(k+1)} &= w^{(k+1)} - \beta_k d^{(k)} \end{aligned}\right\} \textit{compute new direction}$$

It can be observed that the two parameters α_k and β_k can be also computed as

$$\alpha_k = \frac{\|w^{(k)}\|_2^2}{(d^{(k)}, Ad^{(k)})} , \quad \text{and} \quad \beta_k = \frac{\|w^{(k+1)}\|_2^2}{\|w^{(k)}\|_2^2} .$$

The conjugate gradient algorithm enjoys several orthogonality properties.

Proposition 4.3.9. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive definite matrix, $b \in \mathbb{R}^n$. Let $(x^{(k)})_{k \geq 0}$ be the sequence of approximations of the conjugate gradient method. Then*

1. *The Krylov space $\mathcal{K}_k = \mathrm{span}(w^{(0)}, \dots, A^k w^{(k)})$ is such that*

$$\mathcal{K}_k = \mathrm{span}(w^{(0)}, \dots, w^{(k)}) = \mathrm{span}(d^{(0)}, \dots, d^{(k)}) ,$$

[3] Named after the American mathematician Magnus Hestenes (1905-1991) and the Swiss mathematician Eduard Steifel (1909-1978), was first published in 1952.

2. *the sequence* $(w^{(k)})_{0\leq k\leq n-1}$ *is orthogonal,* i.e.,

$$(w^{(k)}, w^{(l)}) = 0 \quad \textit{for all}\; 0 \leq l < k \leq n-1\,,$$

furthermore, the residual $w^{(k)}$ *is orthogonal to all previous search directions* $(d^{(l)})_{0\leq l<k}$, i.e.,

$$(w^{(k)}, d^{(l)}) = 0\,, \quad \textit{for all}\; 0 \leq l < k \leq n-1\,,$$

3. *the sequence* $(d^{(k)})_{0\leq k\leq n-1}$ *is A-conjugate,* i.e.,

$$(Ad^{(k)}, d^{(l)}) = 0 \quad \textit{for all}\; 0 \leq l < k \leq n-1\,,$$

Proof. The definition of the conjugate gradient implies that $w^{(k+1)} \in \mathcal{K}_{k+1}$ and $w^{(k+1)} \perp \mathcal{K}_k$. Thus, $\mathcal{K}_k \subset \mathcal{K}_{k+1}$ and we have $\dim(\mathcal{K}_k) = k+1$. The family $(w^{(k)})_{0\leq k\leq n-1}$ is free and orthogonal and we deduce $\mathcal{K}_k = \operatorname{span}(w^{(0)}, \dots, w^{(k)})$. The family of directions $(d^{(k)})_{0\leq k\leq n-1}$ is also orthogonal for the inner product induced by A and forms a vector space of dimension $k+1$. By induction, we can show that $\operatorname{span}(d^{(0)}, \dots, d^{(k)}) \subset \mathcal{K}_k$. Suppose the assertion is true for $k-1$, we have then

$$d^{(k)} = w^{(k)} - \beta_k d^{(k-1)} \in \mathcal{K}_k \cup \mathcal{K}_{k-1} \subset \mathcal{K}_k\,,$$

and the results follows.
Suppose $(Ad^{(l)}, d^{(k)}) = 0$ for $l < k-1$. We have then

$$\begin{aligned}(Ad^{(l)}, d^{(k)}) &= (Ad^{(l)}, w^{(k)} - \beta_k d^{(k-1)}) \\ &= (Ad^{(l)}, w^{(k)}) = \alpha_l^{-1}(w^{(l)} - w^{(l+1)}, w^{(k)}) = 0\,.\end{aligned}$$

We shall notice that $\alpha_l \neq 0$. We deduce that the family $(d^{(k)})_{k\geq 0}$ is orthogonal fo the inner product $(A\cdot,\cdot)$. Since an orthogonal family of nonzero vectors is free then $\mathcal{K}_k = \operatorname{span}(d^{(0)}, \dots, d^{(k)})$. □

Lemma 4.3.11. *Let* $A \in \mathcal{M}_n(\mathbb{R})$ *be a symmetric positive definite matrix and* $b \in \mathbb{R}^n$. *The sequence* $(x^{(k)})_{0\leq k\leq p}$, *with* $p \leq n$, *defined by the conjugate gradient method is such that* $x^{(n)} = x_0$ *with* $Ax_0 = b$. *Furthermore, the algorithm converges to the solution in at most n iterations.*

Remark 4.3.8. The conjugate gradient method could have been defined by chosing $x^{(k+1)}$ in the affine space $[x^{(0)} + \mathcal{K}_k]$ that minimizes in this space the functional

$$f(x) = \frac{1}{2}(Ax, x) - (b, x)\,.$$

Indeed, if we define, for any $y \in \mathcal{K}_k$, the function

$$g(y) = f(x^{(0)} + y) = \frac{1}{2}(Ay, y) - (w^{(0)}, y) + f(x^{(0)})\,,$$

we observe that minimizing f on $[x^{(0)} + \mathcal{K}_k]$ is equivalent to minimizing g on the Krylov space $\mathcal{K}_k$. Thanks to Corollary 4.3.2, $g(y)$ admits a unique minimum in $\mathcal{K}_k$, denoted $(x^{(k+1)} - x^{(0)}$, that gives a unique $x^{(k+1)}$. And using the same Corollary, we have

$$(Ax^{(k+1)} - b, y) = 0\,, \quad \text{for all } y \in \mathcal{K}_k\,,$$

which is equivalent to writing that $w^{(k+1)}$ is orthogonal to $\mathcal{K}_k$. Hence, $x^{(k+1)}$ minimizes the energy functional (4.10) and from this property, it follows that the energy norm $\| \cdot \|_A$ in the conjugate gradient method is monotonically decreasing.

Practical issues

In theory, the convergence of the conjugate gradient is achieved when the residual $w^{(k)} = 0$ and then $x^{(k)}$ is the solution to the system $Ax = b$. However, this termination property is only valid in exact arithmetic. Indeed, because of rounding errors, this criterion may not be satisfied in practice. This explains why a parameter $\varepsilon \ll 1$ (in general chosen in the range $[10^{-8}, 10^{-4}]$) is introduced and the convergence is considered numerically achieved when

$$\|w^{(k)}\| \leq \varepsilon \, \|w^{(0)}\|\,.$$

Another issue concerns the choice of the initial vector $x^{(0)}$. Since there is in general no information on the solution to the system $Ax = b$, a typical choice is to set $x^{(0)} = 0$. But if a sequence of closely related problems, $x^{(0)}$ can be initialized to the solution of the previous resolution.

The conjugate gradient algorithm involves only the single matrix-vector product $Ad^{(k)}$ at each iteration, and for obvious efficiency reasons, the residual $w^{(k)}$ is computed using the induction formula and not via the theoretical formula $w^{(k)} = b - Ax^{(k)}$.

At this point, one may wonder whether the conjugate gradient shall be considered as a direct or as an iterative method. On the one hand, we know that $x^{(n-1)} = x$, the exact solution of $Ax = b$, but on the other hand, $x^{(k+1)}$ is computed from $x^{(k)}$ and from the previous descent directions $d^{(k)}$. Should it be regarded as a direct method, the number of operations in the worst case scenario $k = n - 1$ would then be in $O(n^3)$, more computationally expensive than Cholesky decomposition for instance. Clearly, the practice tends to consider the conjugate gradient method as an iterative method with an optimal convergence rate. We have the following result

Proposition 4.3.10. *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive definite matrix and $x \in \mathbb{R}^n$ be the exact soution of $Ax = b$. Let $(x^{(k)})_k$ be the sequence of approximate solutions obtained by the conjugate gradient algorithm. Then we have the following estimates*

$$\|x^{(k)} - x\|_2 \;\leq\; 2\,\mathrm{cond}_2(A)^{1/2} \left(\frac{\mathrm{cond}_2(A)^{1/2} - 1}{\mathrm{cond}_2(A)^{1/2} + 1}\right)^k \|x^{(0)} - x\|_2$$

$$\|x^{(k)} - x\|_A \;\leq\; 2 \left(\frac{\mathrm{cond}_2(A)^{1/2} - 1}{\mathrm{cond}_2(A)^{1/2} + 1}\right)^k \|x^{(0)} - x\|_A\,.$$

Proof. (see [GvL83], [Saa96]). □

It is interesting to notice that the approximation of x is strongly dependent on the number of iterations that are performed, *i.e.,* more iterations lead to a better solution. This confirms that the conjugate gradient is indeed an iterative method. Furthermore, the convergence rate is quadratic, *i.e.,* related to the square root of the condition number of A, and this is a much better result than with the gradient method. This also indicates that if $\mathrm{cond}_2(A)$ is close to 1 (optimal), then the convergence will be accelerated. Preconditioning is often a way of improving the efficiency of the conjugate gradient method (see next section). Finally, the convergence rate is often pessimistic asymptotically and the effective convergence is better. The effect of extremal eigenvalues tend to be eliminated as the number of iterations increases, thus leading to a *superlinear convergence.*

Preconditioned conjugate gradients

Suppose the matrix $A \in \mathcal{M}_n(\mathbb{R})$ is ill-conditioned and consider for instance $\varepsilon = \mathrm{cond}_2(A)^{-1} \ll 1$. At each iteration of the conjugate gradient, the error $\|x^{(k)} - x\|_A$ is then reduced by a factor $1 - 2\sqrt{\varepsilon}$ and the convergence rate is thus very slow. The idea of *preconditioning* is to replace the matrix A in the original system by another matrix, leading to the same solution, if its condition number is better than the condition number of A. Hence, the conjugate gradient method is expected to converge faster toward the solution.

Definition 4.3.13. *Let $Ax = b$ be the linear system to solve. Suppose P is a symmetric positive definite matrix, easy to invert and such that* $\mathrm{cond}_2(P^{-1}A)$ *is smaller than* $\mathrm{cond}_2(A)$*. The equivalent system $P^{-1}Ax = P^{-1}b$ is called a* preconditioned system.

Since A is symmetric and positive definite, the preconditioned matrix must also have this property. To this end, we take P symmetric positive definite and we consider a Choleski factorization $P = LL^t$. Posing $y = L^t x$ leads to solve the systems

$$L^{-1}AL^t y = L^{-1}b \quad \text{and} \quad (L^t)^{-1}y = x\,,$$

and this last system is easy to solve since L is upper triangular.

Consider $\tilde{A} \in \mathcal{M}_n(\mathbb{R})$ defined by $\tilde{A} = L^{-1}A(L^t)^{-1}$. The matrix $\tilde{A}$ is symmetric since

$$\tilde{A}^t = ((L^t)^{-1})^t A^t (L^{-1})^t = L^{-1} A (L^t)^{-1} = \tilde{A}\,,$$

and is positive definite as

$$(\tilde{A}x, x) = (L^{-1} A (L^t)^{-1} x, x) = (A (L^t)^{-1} x, (L^t)^{-1} x),\,,$$

and thus $(\tilde{A}x, x) > 0$ if $x \neq 0$. Thus, the conjugate gradient algorithm can be employed to solve the system $\tilde{A}y = L^{-1}b$, *i.e.,* to compute a sequence of approximate solutions $(y^{(k)})_{k\geq 0}$ such that

$$(\tilde{A}y^{(k)} - L^{-1}b, y) = 0\,, \quad \text{for all } y \in \mathcal{K}_k\,,$$

and then to deduce $(x^{(k)})_{k\geq 0}$ solution of the system. However, it is possible to compute directly the sequence $(x^{(k)})_{k\geq 0}$.

Definition 4.3.14 (Preconditioned conjugate gradient). *Let $A \in \mathcal{M}_n(\mathbb{R})$ be a symmetric positive definite matrix, $b \in \mathbb{R}^n$. The preconditioned conjugate gradient algorithm is defined as*

1. **Initialization***:*
$x^{(0)} \in \mathbb{R}^n$*, given*
$w^{(0)} = b - Ax^{(0)}$*;*
$s^{(0)} = (LL^t)^{-1} w^{(0)}$*;*
$d^{(0)} = s^{(0)}$*;*

2. **Iteration:**
for $k \geq 0$, until $\|w^{(k)}\| < \varepsilon$,
construct the sequence

$$\left.\begin{aligned} \alpha_k &= \frac{(w^{(k)}, d^{(k)})}{(d^{(k)}, Ad^{(k)})} \\ x^{(k+1)} &= x^{(k)} + \alpha_k d^{(k)} \\ w^{(k+1)} &= w^{(k)} - \alpha_k A d^{(k)} \\ s^{(k+1)} &= (LL^t)^{-1} w^{(k+1)} \end{aligned}\right\} \textit{compute new solution}$$

$$\left.\begin{aligned} \beta_k &= \frac{(Ad^{(k)}, s^{(k+1)})}{(Ad^{(k)}, d^{(k)})} \\ d^{(k+1)} &= s^{(k+1)} - \beta_k d^{(k)} \end{aligned}\right\} \textit{compute new direction}$$

We observe that the only new operation is the computation, at each iteration, of the term $s^{(k)} = (LL^t)^{-1} w^{(k)}$ that requires solving the linear system $(LL^t)s^{(k)} = w^{(k)}$. This procedure can be achieved in $O(n^2)$ operations in the worst case scenario. The error estimate is the same as for the unprecondtioned conjugate gradient, simply replacing the matrix A by $P^{-1}A$.

4.3.5 Nonsymmetric matrices: GMRes

We briefly mention another method for solving a linear system $Ax = b$, in the case where the matrix A is nonsymmetric. The GMRes[4] (Generalized Minimal Residual) method approximates the solution in a Krylov subspace with minimal residual. For the proofs and further analysis, the reader is refered to [Axe94], [Saa96], [QSS00].

Definition 4.3.15. *The* degree *of $r \in \mathbb{R}^n$ with respect to $A \in \mathcal{M}_n(\mathbb{R})$ is defined as the minimum degree of a non null polynomial p in A, for which $(p(A), r) = 0$.*

The dimension of the Krylov space $\mathcal{K}_k(A, r)$ is equal to the minimum between k and the degree of r with respect to A. Hence, the dimension of the Krylov subspaces is an increasing function of k.

Here, we show that at the iteration k, it is possible to construct a Krylov subspace of dimension k that minimizes the residual $w^{(k)}$. An orthonormal basis of $\mathcal{K}_k(A, w^{(k)})$ can be computed using *Arnoldi's algorithm*, for a fixed k.

Definition 4.3.16 (Arnoldi's algorithm). *Posing $v^{(1)} = w^{(0)}/\|w^{(0)}\|_2$, Arnoldi's algorithm generates the orthonormal basis for $\mathcal{K}_k(A, v^{(1)})$ using the Gram-Schmidt procedure. We have*

1. **Initialization***:*
$v^{(1)} = \dfrac{w^{(0)}}{\|w^{(0)}\|_2}$
2. **Iteration:**
for $j = 1, \dots, k$ compute

$$\begin{aligned} h_{ij} &= (Av^{(j)}, v^{(i)})\,, \qquad i = 1, 2, \dots, j \\ s^{(j)} &= Av^{(j)} - \sum_{i=1}^{j} h_{ij} v^{(i)}\,, \\ h_{j+1j} &= \|s^{(j)}\|\,, \quad \textit{if } h_{j+1j} = 0 \textit{ then exit.} \\ v^{(j+1)} &= \frac{s^{(j)}}{h_{j+1j}}\,. \end{aligned}$$

If $s^{(j)} = 0$, the process terminates by a *breakdown*.

Lemma 4.3.12. *Suppose $h_{j+1j} \neq 0$, $1 \leq j \leq k$. Then the family $(v^{(1)}, \dots, v^{(k)})$ is orthonormal and forms a basis of $\mathcal{K}_k(A, v^{(1)})$.*

[4] Y. Saad and M.H. Schultz (1986), GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems, *SIAM J. Sci. Stat. Comput.*, **7**: 856-869.

Denoting $V_k = [v^{(1)}|\dots|v^{(k)}] \in \mathcal{M}_{n,k}(\mathbb{R})$ the rectangular matrix whose columns are the vectors $v^{(j)}$, we have

$$V_k^t A V_k = H_k\,, \quad V_{k+1}^t A V_k = \hat{H}_k\,,$$

where $\hat{H}_k \in \mathcal{M}_{k+1,k}(\mathbb{R})$ is the upper *Hessenberg* matrix whose entries h_{ij} are given by the Arnoldi's algorithm and $H_k \in \mathcal{M}_k(\mathbb{R})$ is the restriction of $\hat{H}_k$ to the first k rows and k columns.

This algorithm will be used to solve the linear system $Ax = b$ by a Krylov method. Classically, we search for $x^{(k)} \in W_k$ as the vector that minimizes the norm of the residual $\|w^{(k)}\|_2$, *i.e.,* such that

$$\|Ax^{(k)} - b\|_2 = \min_{y\in W_k} \|Ay - b\|_2\,.$$

To this end, we write

$$x^{(k)} = x^{(0)} + V_k z^{(k)},\,,$$

where $z^{(k)}$ are coefficients to be determined. Hence, we have

$$w^{(k)} = w^{(0)} - AV_k z^{(k)} = w^{(0)} - V_{k+1}\hat{H}_k z^{(k)} = V_{k+1}(\|w^{(0)}\|_2 e_1 - \hat{H}_k z^{(k)})\,,$$

where e_1 is the first unit vector of $\mathbb{R}^{k+1}$.

Lemma 4.3.13. *Since the matrix V_{k+1} is orthogonal, the minimum of $w^{(k)}$ is characterized by*

$$\|\|w^{(0)}\|_2 e_1 - \hat{H}_k z^{(k)}\|_2 \le \|\|w^{(0)}\|_2 e_1 - \hat{H}_k y^{(k)}\|_2\,, \quad \textit{for all } y^{(k)} \in \mathbb{R}^k\,.$$

Hence, at each step k, $z^{(k)}$ is chosen in such a way to minimize the functional $\|\|w^{(0)}\|_2 e_1 - \hat{H}_k z^{(k)}\|_2$. This requires solving a linear least-squares problem of size k, at each iteration.

The GMRes method terminates in at most n iterations with the exact solution.

Proposition 4.3.11. *A breakdown occurs in the GMRes at iteration $k < n$ if and only if $x^{(k)} = x$, the solution of $Ax = b$.*

Practical issues

The GMRes algorithm requires storing a basis of the Krylov subspace in memory. Hene, the memory requirement increases linearly with the number of iterations k while the computational effort to orthogonalize $Av^{(k)}$ is proportional to n^2. Therefore, a maximal dimension of the Krylov space is fixed in practice, usually between 10 and 50. After k iterations, the GMRes algorithm is restarted with $x^{(0)} = x^{(k)}$ and a new Krylov subspace is constructed.

It has been observed in numerical experiments that the GMRes algorithm often exhibits a a*superlinear convergence*. This indicates that the rate of convergence improves at each iteration.

4.4 Methods for computing eigenvalues

In this section, we deal with approximations of the eigenvalues and eigenvectors of a matrix $A \in \mathcal{M}_n(\mathbb{C})$. This topic is of importance in many application fields, like structural dynamics for instance. Eigenvalues provide useful information about the evolution of a system governed by a matrix. They are also important in the analysis of the stability of many numerical methods. Two classes of method can be identified, *local* methods that compute only the extrema eigenvalues of A and *global* methods that provide the whole spectrum of A.

We recall that the eigenvalues and eigenvectors of A are solutions of the linear homogeneous system

$$(A - \lambda I_n)x = 0\,, \quad x \neq 0\,. \tag{4.12}$$

Hence, if λ is an eigenvalue of A then the matrix $A - \lambda I_n$ is a singular matrix. The eigenvalues are the roots of the *characteristic equation*

$$P_A(\lambda) = \det(A - \lambda I_n) = 0\,,$$

where $P_A(\lambda)$ denotes the characteristic polynomial of A. Thus, the matrix A has exactly n eigenvalues λ_i, counting multiple roots according to their multiplicities and then we have

$$P_A(\lambda) = \prod_{i=1}^{n}(\lambda_i - \lambda)\,.$$

The next result relates the spectral radius $\rho(A)$ to matrix norms and indicates that all the eigenvalues of A are enclosed in a circle a radius $\|A\|$ centered at the origin in the complex plane.

Lemma 4.4.1. *Let $A \in \mathcal{M}_n(\mathbb{C})$ and let $\|\cdot\|$ denote a matrix norm. Then*

$$\rho(A) \leq \|A\|\,, \quad \textit{or} \quad |\lambda| \leq \|A\|\,, \quad \textit{for all } \lambda \in \mathrm{Sp}(A)\,.$$

Proof. Suppose λ is an eigenvalue of A and $x \neq 0$ an associated eigenvector. We have then

$$|\lambda|\|x\| = \|\lambda x\| = \|Ax\| \leq \|A\|\|x\|\,,$$

and thus $|\lambda| \leq \| A\|$. □

The following theorem is used to locate eigenvalues of A.

Theorem 4.4.1 (Gershgorin's circle[5]). *Let $A \in \mathcal{M}_n(\mathbb{C})$. Then, all the eigenvalues of A lie in the union of the Gershgorin disks in the complex plane,* i.e.,

[5] Gerschgorin S., Über die Abgrenzung der Eigenwerte einer Matrix, *Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk*, **7**: 749-754, (1931).

$$\mathrm{Sp}(A) \subseteq \mathcal{S}_C = \bigcup_{i=1}^{n} D_i\,, \quad D_i = \{z \in \mathbb{C}\,;\ |z - a_{ii}| < r_i\}\,, \quad r_i = \sum_{j \neq i}^{n} |a_{ij}|\,.$$

Proof. Let λ be an eigenvalue of A and $x \neq 0$ a corresponding eigenvector. Then, for all $i = 1, \dots, n$

$$(\lambda - a_{ii})x_i = \sum_{j=1, i\neq j}^{n} a_{ij}x_j\,.$$

Introducing the $\|\cdot\|_\infty$ norm, we consider the index i such that $|x_i| = \|x\|_\infty$. Then, we have

$$|\lambda - a_{ii}| \leq \sum_{j=1}^{n} |a_{ij}| \frac{|x_j|}{|x_i|} \leq r_i\,.$$

This inequality implies the existence of a disk centered at a_{ii} of radius r_i that contains each and any eigenvalue λ. Hence, all eigenvalues lie in the union of these disks. □

Definition 4.4.1. *A matrix $A \in \mathcal{M}_n(\mathbb{C})$ is said to be* reducible *if there exists a permutation matrix P such that*

$$PAP^t = \begin{pmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{pmatrix}\,,$$

where A_{ii} are square matrices. Otherwise, A is called irreducible.

Proposition 4.4.1. *Let $A \in \mathcal{M}_n(\mathbb{C})$ be a irreducible matrix. Then each eigenvalue $\lambda \in \mathrm{Sp}(A)$ lies in the union of the Gershgorin disks (*i.e.*, cannot lie on the boundary of the union) unless its lies on the boundary of all Gershgorin disks.*

4.4.1 The power method

The *power method* if one of the oldest methods for approximating the eigenvalues and eigenvectors of a matrix and, more specifically, its eigenvalues λ_1 of largest module. Its interesting feature is that it does not decompose the matrix A and thus can be used with large sparse matrices.

Let $A \in \mathcal{M}_n(\mathbb{C})$ be a diagonalizable matrix and let $\Lambda \in \mathcal{M}_n(\mathbb{C})$ be the matrix of its eigenvectors $(x_i)_{1\leq i\leq n}$ associated with the eigenvalues $\lambda_1, \dots, \lambda_n$, supposed ordered as

$$|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|\,,$$

Here, λ_1 is called the *dominant* eigenvalue of A.

Definition 4.4.2. *Let $A \in \mathcal{M}_n(\mathbb{C})$ be a diagonalizable matrix. Given a starting vector unit norm $q^{(0)} \in \mathbb{C}^n$, the power method constructs the sequence of vectors $Aq^{(k)}$ using the recursive algorithm*

$$z^{(k)} = Aq^{(k-1)}$$
$$q^{(k)} = \frac{z^{(k)}}{\|z^{(k)}\|_2} .$$

By induction, we have then

$$q^{(k)} = \frac{A^k q^{(0)}}{\|A^k q^{(0)}\|_2} , \quad k \geq 1 ,$$

the initial vector $q^{(0)}$ can be expanded along the basis of the eigenvectors x_i of A

$$q^{(0)} = \sum_{j=1}^{n} \alpha_j x_j , \quad \alpha_j \in \mathbb{C} , \; j = 1, \dots, n ,$$

and we have then for all $k = 1, \dots, n$

$$q^{(k)} = \sum_{j=1}^{n} \lambda_j^k \alpha_j x_j = \lambda_1^k \left(\alpha_1 x_1 + \sum_{j=2}^{n} \left(\frac{\lambda_j}{\lambda_1} \right)^k \alpha_j \lambda_j \right) ,$$

and likewise, since $Ax_i = \lambda_i x_i$, we can write

$$A^k q^{(0)} = \alpha_1 \lambda_1^k \left(x_1 + \sum_{j=2}^{n} \frac{\alpha_j}{\alpha_1} \left(\frac{\lambda_j}{\lambda_1} \right)^k x_j \right) , \quad k = 1, \dots, n .$$

Since $|\lambda_j / \lambda_1| < 1$, for $j = 2, \dots, n$, we observe that

$$\frac{1}{\lambda_1^k} q^{(k)} = \alpha_1 x_1 + O \left(\left| \frac{\lambda_2}{\lambda_1} \right|^k \right) , \tag{4.13}$$

and the vector $q^{(k)}$ has an increasingly significant component in the direction of x_1 and thus will converge toward a limit vector which is he eigenvector associated with the dominating eigenvalue λ_1.

Lemma 4.4.2. *Let $A \in \mathcal{M}_n(\mathbb{C})$ be a diagonalizable matrix with eigenvalues such that $|\lambda_1| > |\lambda_2| \geq \cdots \geq |\lambda_n|$. If $\alpha_1 \neq 0$, there exists a constant $C > 0$ such that*

$$\|\tilde{q}^{(k)} - x_1\|_2 \leq C \left| \frac{\lambda_2}{\lambda_1} \right|^k , \quad k \geq 1 ,$$

where

$$\tilde{q}^{(k)} = \frac{q^{(k)} \|A^k q^{(0)}\|_2}{\alpha_1 \lambda_1^k} = x_1 + \sum_{j=2}^{n} \frac{\alpha_j}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k x_i , \quad k \geq 1 .$$

Convergence of the power method can be shown for any initial vector $q^{(0)}$ since it depends only on the assumption $\alpha_1 \neq 0$. However, the convergence rate (and thus the efficiency of the method) depends on the separation of the dominant eigenvalues, $|\lambda_2| / |\lambda_1| \ll 1$.

Remark 4.4.1. Because it provides only the dominant eigenvalue, the power method is of limited practical uselfulness. Nevertheless, specific applications require such algorithm. For example, it is used in *Google's PageRank algorithm*, which is an eigenvector of a matrix of order 2.7 billion[6].

4.4.2 Inverse iteration method

At the difference of the power method that converges always to the eigenvalue of largest module, the *inverse iteration* method[7] allows to chose the eigenvalue to converge to. More precisely, we look here for the eigenvalue of $A \in \mathcal{M}_n(\mathbb{C})$ which is *closest* to a given number $\mu \in \mathbb{C}$, $\mu \notin \mathrm{Sp}(A)$.

Firstly, we observe that the power method allows to compute the smallest eigenvalue in modulus of a nonsingular matrix $A \in \mathcal{M}_n(\mathbb{C})$, by applying the algorithm to A^{-1}. Indeed, from (4.13) we deduce that if the eigenvalues of A are such that

$$|\lambda_1| \geq | \lambda_2| \geq \cdots > |\lambda_n| \,,$$

and the power method is applied to A^{-1}, the vector $q^{(k)}$ will converge toward the eigenvector x_n corresponding to λ_n.

Let $\mu \notin \mathrm{Sp}(A)$ be a crude approximation of an eigenvalue λ of A. The power iteration can be applied to the matrix $B^{-1} = (A - \mu I_n)^{-1}$. The eigenvectors of $(A - \mu I_n)$ are those of A and the spectrum of this matrix if *shifted* by μ (called a *shift* for this reason). In other words, the eigenvalues ξ_i of the matrix B^{-1} are related to the eigenvalues of A by

$$\xi_i = \frac{1}{\lambda_i - \mu} \,, \qquad \lambda_i = \mu + \frac{1}{\xi_i} \,.$$

Suppose μ is closer to λ_m than any other eigenvalue of A, *i.e.,*

$$|\lambda_m - \mu| < |\lambda_i - \mu| \,, \quad \text{for all } i = 1, \dots, n \,,\ i \neq m \,,$$

then $\lambda_m - \mu$ is the smallest eigenvalue of $(A - \mu I_n)$ and likewise, ξ_m is the eigenvalue of B^{-1} with largest modulus. In the peculiar case $\mu = 0$, ξ_m is also the smallest eigenvalue of A.

Definition 4.4.3. *Let $A \in \mathcal{M}_n(\mathbb{C})$. Given a starting vector $q^{(0)} \in \mathbb{C}^n$, the inverse iteration algorithm constructs the sequence*

$$\begin{aligned} (A - \mu I_n) z^{(k)} &= q^{(k-1)} \\ q^{(k)} &= \frac{z^{(k)}}{\|z^{(k)}\|_2} \,. \end{aligned}$$

[6] Ipsen I., Wills R.M., Analysis and Computation of Google's PageRank, in *7th IMACS International Symposium on Iterative Methods in Scientific Computing*, Fields Institute, Toronto, Canada, 5-8 May (2005).

[7] Wielandt H., Das Iterationsverfahren bei nicht selbstadjungierten linearen Eigenwertaufgaben, *Math. Z.*, **50**: 93-143, (1944).

Notice that the matrix $(A - \mu I_n)$ is not inverted explicitly. At each iteration step k, a linear system must be solved. In the symmetric case, a factorization is performed at $k = 1$ on the matrix $A - \mu I_n = LDL^t$ where D is a block diagonal matrix and L is block lower triangular. In the unsymmetric case, the LU factorization is performed at step $k = 1$, so that at each step two linear systems are solved.

If the shift μ is sufficiently close to an eigenvalue λ_m,

$$|\lambda_m - \mu| \ll | \lambda_i - \mu| \,, \quad \text{for all } \lambda_i \neq \lambda_m \,,$$

then $(\lambda_m - \mu)^{-1}$ is a dominating eigenvalue of $(A - \mu I_n)$ and $q^{(k)}$ will converge quickly to the eigenvector x_i.

4.4.3 QR iteration

As its name tells, the QR iteration method[8] performs a QR decomposition (see Section 4.2.5), to write a matrix as a product of an orthogonal matrix Q and an upper triangular matrix R.

We consider here the case of nonsingular real matrices and assume that the matrices have distinct real eigenvalues, $|\lambda_1| > \cdots > |\lambda_n| > 0$.

Definition 4.4.4. *Let $A \in \mathcal{M}_n(\mathbb{R})$ such that $|\lambda_1| > \cdots > |\lambda_n| > 0$. The QR method constructs the sequence of matrices $(A_k)_{k \geq 1}$ with $A_1 = A$ and*

$$A_{k+1} = R_k Q_k \,,$$

where $Q_k R_k = A_k$ is the QR decomposition of A_k (see Section 4.2.5).

Since we have $A_{k+1} = Q_k^t (Q_k R_k) Q_k = Q_k^t (A_k) Q_k = Q_k^{-1}(A_k) Q_k$, we show by induction that

$$A_{k+1} = Q_k^t (A_k) Q_k = Q_k^t Q_{k-1}^t (A_{k-1}) Q_{k-1} Q_k = \cdots = (Q^{(k+1)})^t A Q^{(k+1)}$$

with $Q^{(k)} = Q_1 \dots Q_k$. And thus, $AQ^{(k)} = Q^{(k+1)} R_{k+1}$. In other words, every matrix A_k is orthogonally similar to the matrix A.

Lemma 4.4.3. *Let $A \in \mathcal{M}_n(\mathbb{R})$ such that $|\lambda_1| > \cdots > |\lambda_n|$. Then, the sequence of matrices $(A_k)_{k \geq 1}$ generated by the QR iteration algorithm converges to an upper triangular matrix whose diagonal entries are the eigenvalues of A. If the matrix A is symmetric, the sequence $(A_k)_{k \geq 1}$ tends to a diagonal matrix.*

Proof. (see [DB74], [All08], for instance). □

[8] Francis J.G.F., The QR Transformation, I, *The Computer Journal*, **4**(3): 265-271, (1961) and Kublanovskaya V.N., On some algorithms for the solution of the complete eigenvalue problem, *USSR Computational Mathematics and Mathematical Physics*, **3**: 637-657, (1961).

4.4.4 The Lanczos method

The Lanczos method[9] computes the eigenvalues of a real symmetric matrix A whose real eigenvalues are ordered as $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$. Actually, the Lanczos algorithm can be viewed as a simplified version of the Arnoldi algorithm (see Definition 4.3.16).

Definition 4.4.5 (Lanczos algorithm). *Given $w^{(0)} \in \mathbb{R}^n$ a nonzero vector and $\mathcal{K}_k(A, w^{(0)}$ the Krylov space spanned by $(w^{(0)}, \dots, A^k w^{(0)})$, the Lanczos algorithm generates a sequence of vectors $(v^{(j)})_{j\geq 1}$ by induction*

1. **Initialization***:*

$$v^{(0)} = 0\,, \quad v^{(1)} = \frac{w^{(0)}}{\|w^{(0)}\|}\,, \quad \beta_1 = 0\,,$$

2. **Iteration***:*
for $j = 2, \dots, k$, compute

$$\begin{aligned} s^{(j)} &= Av^{(j)} - \beta_j v^{(j-1)}\,, \\ \alpha_j &= (s^{(j)}, v^{(j)})\,, \\ s^{(j)} &= s^{(j)} - \alpha_j v^{(j)}\,, \\ \beta_{j+1} &= \|s^{(j)}\|\,, \quad \textit{if } \beta_{j+1} = 0 \textit{ then exit.} \\ v^{(j+1)} &= \frac{s^{(j)}}{\beta_{j+1}}\,. \end{aligned}$$

By comparison with Arnoldi's algorithm, we observe that $\alpha_j = h_{jj}$ and $\beta_j = h_{j-1j}$. The Lanczos algorithm generates a symmetric tridiagonal matrix

$$T_k = \begin{pmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{k-1} & \alpha_{k-1} & \beta_k \\ & & & \beta_k & \alpha_k \end{pmatrix}$$

and a matrix $V_k = (v^{(1)}|\dots|v^{(k)})$ with orthogonal columns spanning the Krylov space $\mathcal{K}_k(A, v^{(1)})$ such that

$$AV_k = V_k T_k + \beta - k + 1 v^{(k+1)} e_k^t\,,$$

where e_k is the kth vector of the canonical basis of $\mathbb{R}^k$. The interesting result is that the eigenvalues of T_k are eigenvalues of A as well.

[9] Lanczos C., An Iteration Method for the Solution of the Eigenvalue Problem of Linear Di?erential and Integral Operator, *J. Res. Nat. Bur. Stand.*, **45**: 255-282, (1950).

4.5 Exercises and Problems

Exercise 4.5.1 (Symmetric positive definite matrices).

1. Let $A \in \mathcal{M}_n(\mathbb{R})$ a symmetric matrix. Show that A is positive definite if and only if all its eigenvalues are stricly positive
2. Let $A \in \mathcal{M}_n(\mathbb{R})$ a symmetric and positive definite matrix. Show that we can define a matrix $M \in \mathcal{M}_n(\mathbb{R})$ symmetric and positive definite, such that $B^2 = A$.

Exercise 4.5.2 (Spectral radius).

1. What is the spectral radius of the matrix $A = \begin{pmatrix} a & 4 \\ 0 & a \end{pmatrix}$?

   Check that if $a \in]0,1[$ then $\rho(A) < 1$, but $\|A^p\|_2^{1/p}$ can be greater than 1.
2. Consider the matrix $A \in \mathcal{M}_n(\mathbb{R})$ defined by

$$A = \begin{pmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{pmatrix}$$

   Compute its gershorin disks and show that all eigenvalues λ of A are striclty positive.

Problem 4.5.1 (Linear system). Consider the linear system $Ax = b$ where

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 3 \end{pmatrix}, \quad b = \begin{pmatrix} 3 \\ 5 \end{pmatrix},$$

and the iterative method to solve it

$$x^{(k+1)} = B$$

References

[All08] Allaire G., Kaber S.M., *Numerical Linear Algebra*, Texts in Applied Mathematics, **55**, Springer-Verlag, New York, (2008).

[Axe94] Axelsson O., *Iterative Solution Methods*, Cambridge University Press, New York, (1994).

[Bel70] Bellman R., *Introduction to Matrix Analysis*, Mc Graw Hill, New York, (1970).

[Che05] Chen K., *Matrix preconditioning techniques and Applications*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, New York, (2005).

[Cia89] Ciarlet P.G., *Introduction to Numerical Linear Algebra and Optimisation*, Cambridge Texts in Applied Mathematics, Cambridge University Press, Cambridge, (1989).

[DB74] Dahlquist G., Björck A., *Numerical Methods*, Prentice Hall, Series in Automatic Computation, Englewood Cliffs, NJ, (1974).

[Dem97] Demmel J.W., *Applied Numerical Linear Algebra*, SIAM, Philadelphia, (1997).

[DER86] Duff I.S., Erisman A.M., Reid J.K., *Direct Methods for Sparse Matrices*, Oxford University Press, London, (1986).

[Fle80] Fletcher R., *Practical Methods of Optimization*, J. Wiley, New York, (1980).

[GvL83] Golub G.H., van Loan C.F., *Matrix Computations*, The John Hopkins University Press, Baltimore, 3rd edition, (1983).

[Hac94] ackbush W., *Iterative Solution of Large Sparse Systems of Equations*, Springer-verlag, new York, (1994).

[HK71] Hoffman K., Kunze, R., *Linear Algebra*, 2nd ed., Prentice Hall, Englewoods Cliffs, NJ, (1971).

[Hog07] Hogben L., *Handbook of Linear Algebra*, Discrete Mathematics and its Applications, L. Hogben ed., Chapman et al., CRC, Boca Raton, (2007).

[Kel95] Kelley C.T., *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, (1995).

[Kre05] Kressner D., *Numerical Methods for General and Structured Eigenvalue Problems*, Lecture Notes in Computational Science and Engineering, **46**, Springer, Berlin, (2005).

[Lan89] Lang S., *Linear Algebra*, Undergraduate Texts in Applied Mathematics, Springer-Verlag, New York, (1989).

[Lay03] Lay D.C., *Linear Algebra and its Applications*, 3rd ed., Addison Wesley Publishing Co., Reading, MA, (2003).

[Lax97] Lax P., *Linear Algebra*, John Wiley, New York, (1997).

[Ort87] Ortega J.M., *Matrix Theory, a Second Course*, Plenum, New York, (1987).

[QSS00] Quarteroni A., Sacco R., Saleri F., *Numerical Mathematics*, Texts in Applied Mathematics, **37**, Springer-Verlag, New York, (2000).

[Saa96] Saad Y., *Iterative Methods for Sparse Linear Systems*, PWS Publishing Company, Boston, (1996).

[Str80] Strang G., *Linear Algebra and its Applications*, 2nd ed., Academic Press Inc., New York, (1980).

[TB97] Trefethen L.N., Bau D., *Numerical Linear Algebra*, SIAM, Philadelphia, (1997).

[Var62] Varga R.S., *Matrix iterative analysis*, Prentice Hall, Englewood Cliffs, NJ, (1962).

[Wat02] Watkins D.S., *Fundamentals of Matrix Computations*, Pure and Applied Mathematics, John Wiley & Sons, New York, (2002).