Numerical Linear Algebra

Rao
Politecnico di Milano

Originally written in: **2024-09-16**

Last updated at: **2024-10-29**

Contents

1 Matrix Factorizations

1.1 Cholesky Factorization

Let $A \in \mathbb{R}^{n\times n}$ be a *symmetric and positive definite* (SPD) matrix. Then, there exists a unique upper triangular matrix $R \in \mathbb{R}^{n\times n}$ with positive diagonal entries such that:

$$A = R^T R \tag{1.1}$$

This factorization is called *Cholesky factorization.*

CHOLESKY FACTORIZATION

Let $r_{11} = \sqrt{a_{11}}$.

For $k = 2, \ldots$

$$r_{ij} = \frac{1}{r_{ii}}\left(a_{ij} - \sum_{k=1}^{i-1} r_{ki} r_{kj}\right)$$
$$r_{ii} = \sqrt{a_{ii} - \sum_{k=1}^{i-1} r_{ki}^2}$$

The computational cost of the Cholesky factorization is $O(n^3/3)$.

2 Norms

The essential notions of **size and distance** in a vector space are captured by norms. These are the *yardsticks* with which we measure approximations and convergence throughout numerical linear algebra.

2.1 Vector Norms

A norm is a function $\|\cdot\| : \mathbb{C}^m \to \mathbb{R}$ that assigns a real-valued length to each vector. In order to conform to a reasonable notion of length, a norm must satisfy the following three conditions. For all vectors x, y and for all scalars $\alpha \in \mathbb{C}$:

1. *Nonnegativity*: $\|\boldsymbol{x}\| \geq 0$ and $\|\boldsymbol{x}\| = 0$ if and only if $x = 0$.
2. *Triangle Inequality*: $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|$.
3. *Homogeneity*: $\|\boldsymbol{\alpha x}\| = |\alpha|\ \|\boldsymbol{x}\|$.

The above conditions allow for different notions of lenght, and at times it is useful to have the flexibility.

$$\begin{aligned} \|\boldsymbol{x}\|_1 &= \sum_{i=1}^{m} |x_i| \\ \|\boldsymbol{x}\|_2 &= \sqrt{\sum_{i=1}^{m} |x_i|^2} \\ \|\boldsymbol{x}\|_\infty &= \max_{\{1<i<m\}} |x_i| \\ \|\boldsymbol{x}\|_p &= \left(\sum_{i=1}^{m} |x_i|^p \right)^{\frac{1}{p}} \quad (p \leq 1 < \infty) \end{aligned} \tag{2.1}$$

Aside from the $p-$norms, the most useful norms are the *weighted p norms*, where each of the coordinates of a vector space is given its own weight. In general, given any norm $\|\cdot\|$, the *weighted p norm* is defined as:

$$\|\boldsymbol{x}\|_w = \|\boldsymbol{W}\boldsymbol{x}\| \tag{2.2}$$

Here W is the diagonal matrix with the ith diagonal entry is the weight $w_i \neq 0$. For example. a weighted 2-norm is specified as follows:

$$\|\boldsymbol{x}\|_W = \left(\sum_{i=1}^{m} |w_i x_i|^2 \right)^{\frac{1}{2}} \tag{2.3}$$

Thm Cauchy-Schwarz Inequality **theorem 2.1.1**

For any two vectors $x, y \in \mathbb{R}^n$, the following inequality holds:

$$|(x, y)| = |x^T y| \leq \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \tag{2.4}$$

Where strict equality holds if and only if x and y are linearly dependent.

We recall that the scalr product in $\mathbb{R}^n$ can be realyed to the p-norms by the H"older inequality:

$$|(x, y)| \leq \|\boldsymbol{x}\|_p \|\boldsymbol{y}\|_q \qquad \text{Where} \qquad \frac{1}{p} + \frac{1}{q} = 1 \tag{2.5}$$

Thm Norm continuity **theorem 2.1.2**

Any vector norm $\|\cdot\|$ defined on V is a continous function of its argument, namely, $\forall > 0, \exists C > 0$ such that if $\|\boldsymbol{x} - \hat{\boldsymbol{x}}\| \leq \varepsilon$ then $\|\boldsymbol{x}\| - \|\hat{\boldsymbol{x}}\| \leq C\varepsilon$, for any $x, \hat{x} \in V$.

Thm **theorem 2.1.3**

let $\|\cdot\|$ be a norm of $\mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$ be a matrix with n linearly independent columns. Then, the function $\|\cdot\|_{A^2}$ acting from $\mathbb{R}^n$ in to $\mathbb{R}$ defined as:

$$\|\boldsymbol{x}\|_{A^2} = \|\boldsymbol{A}\boldsymbol{x}\| \tag{2.6}$$

is a norm on $\mathbb{R}^n$.

Thm Convergence **theorem 2.1.4**

Let $\|\cdot\|$ be a norm in a finite dimensional space V. Then:

$$\lim_{k\to\infty} x^{(k)} = x \iff \lim_{k\to\infty} \|\boldsymbol{x}^{(k)} - \boldsymbol{x}\| = 0 \tag{2.7}$$

where $x \in V$ and $x^{(k)}$ is a sequence of vectors in V.

2.2 Matrix Norms

In dealing with a space of matrices, certain special norms are more useful thant the vector norms. These are the *induced matrix norms*, defined in terms of the behavior of a matrix as an operator between its normed domain and range spaces.

Def Matrix Norm **definition 2.2.1**

A *matrix norm* is a mapping $\|\cdot\| : \mathbb{R}^{m\times n} \to \mathbb{R}$ such that:

1. $\|\boldsymbol{A}\| \geq 0 \forall A \in \mathbb{R}^{m\times n}$ and $\|\boldsymbol{A}\| = 0$ if and only if $A = 0$.
2. $\|\boldsymbol{\alpha A}\| = |\alpha| \, \|\boldsymbol{A}\| \; \forall A \in \mathbb{R}^{m\times n}$ and $\alpha \in \mathbb{C}$.
3. $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\| \; \forall A, B \in \mathbb{R}^{m\times n}$(triangular inequality)

Def **definition 2.2.2**

We say that a matrix norm $\|\cdot\|$ is *compatible* or *consistent* with a vector norm $\|\cdot\|$ if:

$$\|\boldsymbol{Ax}\| \leq \|\boldsymbol{A}\| \, \|\boldsymbol{x}\| \qquad \forall x \in \mathbb{R}^n \tag{2.8}$$

More generally, given three norms, all denoted by $\|\cdot\|$, albeit defined on $\mathbb{R}^m, \mathbb{R}^n.\mathbb{R}^{m\times n}$, respectively, we say that they are consistent if if $\forall x \in \mathbb{R}^n, Ax = y \in \mathbb{R}^m$, we have that $\|\boldsymbol{y}\| \leq \|\boldsymbol{A}\| \, \|\boldsymbol{x}\|$.

Def Sub multiplicative **definition 2.2.3**

We say that a matrix norm $\|\cdot\|$ is *sub-multiplicative* if $\forall A \in \mathbb{R}^{n\times m}, \forall B \in \mathbb{R}^{m\times q}$ we have that

$$\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\| \, \|\boldsymbol{B}\| \tag{2.9}$$

Def Fronebius Norm **definition 2.2.4**

The norm

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2} = \sqrt{\mathrm{tr}(AA^H)} \tag{2.10}$$

is a matrix norm called the *Frobenius norm*. And it is compatible with the Euclidean vector norm $\|\cdot\|_2$. Indeed.

$$\|\boldsymbol{Ax}\|_2^2 = \sum_{i=1}^{m}|\sum_{j=1}^{n}a_{ij}x_j|^2 \leq \sum_{i=1}^{m}\sum_{j=1}^{n}|a_{ij}|^2\sum_{j=1}^{n}|x_j|^2 = \|\boldsymbol{A}\|_F^2\|\boldsymbol{x}\|_2^2 \tag{2.11}$$

Thm Induced Matrix Norm **theorem 2.2.1**

Let $\|\cdot\|$ be a vector norm. The function:

$$\|\boldsymbol{A}\| = \sup_{x\neq 0}\frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|} \tag{2.12}$$

is a matrix norm called *induced matrix norm or natural matrix norm.*

Proof: Check definition 2.2.1.

1. If $\|\boldsymbol{Ax}\| \geq 0$, then it follows that $\|\boldsymbol{A}\| = \sup_{\|\boldsymbol{x}\|=1}\|\boldsymbol{Ax}\| \geq 0$. Moreover,

$$\|\boldsymbol{A}\| = \sup_{x\neq 0}\frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|} = 0 \iff \|\boldsymbol{Ax}\| = 0 \forall x \neq 0 \tag{2.13}$$

and $Ax = 0 \forall x \neq 0$ if and only if $A = 0$; therefore, $\|\boldsymbol{A}\| = 0$ if and only if $A = 0$.

2. Given a scalar α, we have that:

$$\|\boldsymbol{\alpha A}\| = \sup_{x\neq 0}\frac{\|\boldsymbol{\alpha Ax}\|}{\|\boldsymbol{x}\|} = \sup_{x\neq 0}|\alpha|\,\frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|} = |\alpha|\sup_{x\neq 0}\frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|} = |\alpha|\,\|\boldsymbol{A}\| \tag{2.14}$$

3. Finally, triangular inequality holds. Indeed, by definition of supremum, if $x \neq 0$ then:

$$\frac{\|\boldsymbol{Ax}\|}{\|\boldsymbol{x}\|} \leq \|\boldsymbol{A}\| \Rightarrow \|\boldsymbol{Ax}\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{x}\| \tag{2.15}$$

So that, taking x with unit norm, one gets:

$$\|(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{x}\| \leq \|\boldsymbol{Ax}\| + \|\boldsymbol{Bx}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\| \tag{2.16}$$

from which it follows that $\|\boldsymbol{A}+\boldsymbol{B}\| = \sup_{\|\boldsymbol{x}\|=1}\|(\boldsymbol{A}+\boldsymbol{B})\boldsymbol{x}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$.

Relevant instances of induced matrix norms are the so-called *p-norms*:

$$\|\boldsymbol{A}\|_p = \sup_{\boldsymbol{x}\neq 0}\frac{\|\boldsymbol{Ax}\|_p}{\|\boldsymbol{x}\|_p} \tag{2.17}$$

The 1-norm(column sum norn):

$$\|\boldsymbol{A}\|_1 = \max_{j=1,\ldots,n}\sum_{i=1}^{m}|a_{ij}| \tag{2.18}$$

The infinity-norm(row sum norm):

$$\|\boldsymbol{A}\|_\infty = \max_{i=1,\dots,m} \sum_{j=1}^{n} |a_{ij}| \tag{2.19}$$

Moreover, we have $\|\boldsymbol{A}\|_1 = \|\boldsymbol{A^T}\|_\infty$ and, if A is self-adjoint or real sysmetric, then $\|\boldsymbol{A}\|_1 = \|\boldsymbol{A}\|_\infty$.

A special discussion is deserved by the *2-norm* or *spectral norm* for which the following theorem holds.

Thm Spectral Norm **theorem 2.2.2**

Let $\sigma_1(A)$ be the largest singular value of A. Then, the 2-norm of A is given by:

$$\|\boldsymbol{A}\|_2 = \sqrt{\rho(A^H A)} = \sqrt{\rho(A^H A)} == \sigma_1(A) \tag{2.20}$$

In particular, if A is hermitian (or real and symmetric), then $\|\boldsymbol{A}\|_2 = \rho(A)$.

Proof: Since $A^T A$ is hermitian, there exists a unitary matrix U such that

$$U^H A^H A U = \text{diag}(\mu_1, ..., \mu_n) \tag{2.21}$$

where μ_i are the positive eigenvalues of $A^H A$. Let $y = U^H x$, then:

$$\begin{aligned} \|\boldsymbol{A_2}\| &= \sup_{x\neq 0} \frac{\sqrt{(A^H A x, x)}}{\sqrt{(x,x)}} = \sup_{y\neq 0} \frac{\sqrt{(U^H A^H A U y, y)}}{\sqrt{(y,y)}} \\ &= \sup_{y\neq 0} \frac{\sqrt{\sum_{i=1}^n \mu_i \ |y_i|^2}}{\sqrt{\sum_{i=1}^n |y_i|^2}} = \sqrt{\max_{i=1,\dots,n} |\mu_i|} \end{aligned} \tag{2.22}$$

If A is hermitian, the same considerations as above apply directly to A. Finally, if A is unitary, we have

$$\|\boldsymbol{A}\|_2 = \sup_{x\neq 0} \frac{\|\boldsymbol{Ax}\|_2}{\|\boldsymbol{x}\|_2} = \sup_{x\neq 0} \frac{\|\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} = 1 \tag{2.23}$$

3 Principles of Numerical Mathematics

3.1 Well-posedness and Condition Number

Consider the following problem: find x such that:

$$F(x, d) = 0 \tag{3.1}$$

where F is a function of x and d. And three types of problems can be considered:

1. *direct problem*: given F and d, find x;
2. *inverse problem*: given F and x, find d;
3. *identification problem*: given x and d, find F.

Problems Eq. (3.1) are ***well-posed*** if it admits a *unique* solution, and the solution depends continuously on the data.

A problem which does not enjoy the property above is called ill posed or unstable and before undertaking its numerical solution it has to be regularized, that is, it must be suitably transformed into a well-posed problem.

Let D be the set of admissible data, i.e. the set of the values of d in correspondance of which problem Eq. (3.1) admits a unique solution. Continuous dependence on the data means that small perturbations on the data d of D yield "small" changes in the solution x.

Precisely, let $d \in D$ and denoted by δd a perturbation admissible in the sense that $d + \delta d \in D$ and by δx the corresponding change in the solution, in such a way that:

$$F(x + \delta x, d + \delta d) = 0 \tag{3.2}$$

Then, we require that:

$$\begin{gathered} \exists \eta_0 = \eta_0(d) > 0, \exists K_0 = K_0(d) \text{ such that} \\ \text{if } \|\boldsymbol{\delta d}\| \leqslant \eta_0 \text{ then } \|\boldsymbol{\delta x}\| \leqslant K_0 \, \|\boldsymbol{\delta d}\| \end{gathered} \tag{3.3}$$

The norms used for the data and for the solution may not coincide, whenever d and x represent variables of different kinds.

e.g. Wellposedness of Linear Systems **example 3.1.1**

Consider the problem of solving a linear system $Ax = b$. The problem is well-posed if it has below two properties:

1. The problem has a unique solution x.
2. The solution depends continously on the data.

The Eq. (3.3) is however more suitable to express in the following the concept of *numerical stability*, that is, the property that small perturbations on the data yield perturbations of the same order on the solution.

Def Condition Number **definition 3.1.1**

For problem Eq. (3.1), we define the *relative conditional number* to be:

$$K(d) = \sup \left\{ \frac{\frac{\|\boldsymbol{\delta x}\|}{\|x\|}}{\frac{\|\boldsymbol{\delta d}\|}{\|\boldsymbol{d}\|}}, \delta d \neq 0, d + \delta d \in D \right\} \tag{3.4}$$

Whenever $d = 0$ or $x = 0$, it is nessesary to consider the *absolute conditional number*:

$$K_{\text{abs}}(d) = \sup \left\{ \frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{\delta d}\|}, \delta d \neq 0, d + \delta d \in D \right\} \tag{3.5}$$

3.1.1 Absolute Condition Number

Let δx denote a small perturbation of x, and wirte $\delta f = f(x + \delta x, d) - f(x, d)$. The absolute condition number is then defined as:

$$K_{\text{abs}} = \lim_{\delta \to 0} \sup_{\|\boldsymbol{\delta x}\| \leq \delta} \frac{\|\boldsymbol{\delta f}\|}{\|\boldsymbol{\delta x}\|} \tag{3.6}$$

For most problems, the limit of the supremum in this formula can be interpreted as a supremum over all infinitesimal perturbations δx, and in the interest of readability, we shall generally write the formula simply as

$$K = \sup_{\delta x} \frac{\|\boldsymbol{\delta f}\|}{\|\boldsymbol{\delta x}\|} \tag{3.7}$$

with the understanding that δx and δf are infinitesimal.

If f is differentiable, we can evaluate the absolute condition number by means of the derivative of f. Let $J(x)$ be the matrix whose i, j entry is the partial derivative $\partial f_i / \partial x_j$, evaluated at x. The definition of derivative gives us, $\delta f \approx J(x)\delta x$, with equality in the limit $\|\boldsymbol{\delta x}\| \to 0$. The absolute condition number is then:

$$K = \|\boldsymbol{J}(\boldsymbol{x})\| \tag{3.8}$$

3.1.2 Relative Condition Number

When we are concerned with relative changes, we need the notion of relative condition. The *relative condition number* is defined as:

$$K = \lim_{\delta \to 0} \sup_{\|\boldsymbol{\delta x}\| \leq \delta} \left(\frac{\|\boldsymbol{\delta f}\|}{\|\boldsymbol{f}(\boldsymbol{x})\|} / \frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} \right) \tag{3.9}$$

or, assuming δx and δf are infinitesimal,

$$K = \sup_{\delta x} \frac{\frac{\|\boldsymbol{\delta f}\|}{\|\boldsymbol{f}(\boldsymbol{x})\|}}{\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|}} \tag{3.10}$$

If f is differentiable, we can express this equality in terms of the Jacobian matrix $J(x)$, as follows:

$$K = \frac{\|\boldsymbol{J}(\boldsymbol{x})\|}{\|\boldsymbol{f}\boldsymbol{x}\| \ / \ \|\boldsymbol{x}\|} \tag{3.11}$$

Problem Eq. (3.1) is called *ill-conditioned* if $K(d)$ is "big" for any admissible datum d (the precise meaning of "small" and "big" is going to change depending on the considered problem).

3.1.3 Codution of Matrix-Vector Multiplication

Now we come to one of the condition numbers of fundamental importance in numerical linear algebra.

Fix $A \in \mathbb{C}^{m \times n}$ and consider the problem of computing Ax from input x; that is, we are going to determine a condition number corresponding to perturbations of x but not A. Working directly from the definition of K, with $\|\cdot\|$ denoting an arbitrary vector norm and the corresponding induced matrix norm, we have:

$$K = \sup_{\delta x} \left(\frac{\|\boldsymbol{A}(\boldsymbol{x} + \boldsymbol{\delta x}) - \boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{A}\boldsymbol{x}\|} \right) / \left(\frac{\|\boldsymbol{\delta x}\|}{\|\boldsymbol{x}\|} \right) = \sup_{\delta x} \frac{\|\boldsymbol{A}\boldsymbol{\delta x}\|}{\|\boldsymbol{\delta x}\|} / \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} \tag{3.12}$$

that is,

$$K = \|\boldsymbol{A}\| \, \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{A}\boldsymbol{x}\|} \tag{3.13}$$

This is an exact formula for K, dependent on both A and x.

Suppose A is square and nonsingular. Then we can use the fact that $\|\boldsymbol{x}\| \, / \, \|\boldsymbol{A}\boldsymbol{x}\| \leq \|\boldsymbol{A}^{-1}\|$ to loosen Eq. (3.13) to a bound independent of x:

$$K \leq \|\boldsymbol{A}\| \, \|\boldsymbol{A}^{-1}\| \tag{3.14}$$

Or, one might write this as:

$$k = \alpha \, \|\boldsymbol{A}\| \, \|\boldsymbol{A}^{-1}\| \tag{3.15}$$

with

$$\alpha = \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{A}\boldsymbol{x}\|} / \, \|\boldsymbol{A}^{-1}\| \tag{3.16}$$

If $\|\cdot\| = \|\cdot\|_2$ this will occur whenever x is a multiple of a minimal right singular vector of A.

3.1.4 Condition number of a Matrix

The product $\|\boldsymbol{A}\| \, \|\boldsymbol{A}^{-1}\|$ comes up so often that it has its own name: it is the *condition number* of A:

$$K(A) = \|\boldsymbol{A}\| \, \|\boldsymbol{A}^{-1}\| \tag{3.17}$$

Thus, in this case the term *condition number* is attached to a matrix, not a problem. If $K(A)$ is small, A is said to be *well-conditioned*; if it is large, A is *ill-conditioned*. If A is singular, it is customary to write $K(A) = \infty$.

Note that if $\|\cdot\| = \|\cdot\|^2$, then $\|\boldsymbol{A}\| = \sigma_1$ and $A^{-1} = \frac{1}{\sigma_m}$. Thus

$$K(A) = \frac{\sigma_1}{\sigma_m} \tag{3.18}$$

In the 2-norm, and it is this formula that is generally used for computing 2-norm condition numbers of matrices.

For a rectangular matrix $A \in \mathbb{C}^{m \times n}$ of full rank, $m \geq n$, the condition numver is defined in terms of the **pseudoinverse**: $K(A) = \|\boldsymbol{A}\| \, \|\boldsymbol{A}^{+}\|$. Since A^+ is motivated by least squares problems, this definition is most useful in the case $\|\cdot\| = \|\cdot\|^2$, where we have

$$K(A) = \frac{\sigma_1}{\sigma_n} \tag{3.19}$$

3.1.5 Condition Number of a System of Equations

Specifically, let us hold b fixed and consider the behavior of the problem $A \to x = A^{-1}b$ when A is perturbed by infinitesimal δA. Then x must change by infinitesimal δx such that:

$$(A + \delta A)(x + \delta x) = 0 \tag{3.20}$$

Using the equality $Ax = b$ and dropping the doubly infinitesimal term $(\delta A)(\delta x)$, we obtain $(\sigma A)x + A(\sigma)x = 0$. that is, $\sigma x = -A^{-1}(\sigma A)x$. This equation implies $\|\boldsymbol{\sigma x}\| \leq \|\boldsymbol{A}^{-1}\| \, \|\boldsymbol{\sigma A}\| \, \|\boldsymbol{x}\|$, or equivalently:

$$\frac{\sigma x}{\|\boldsymbol{x}\|} / \frac{\sigma A}{\|\boldsymbol{A}\|} \leq \|\boldsymbol{A^{-1}}\| \, \|\boldsymbol{A}\| = K(A) \tag{3.21}$$

Thm **theorem 3.1.1**

Let b be fixed and consider the problem of computing $x = A^{-1}b$, where A is square and nonsingular. The condition number of this problem with respect to perturbations in A is

$$K(A) = \|\boldsymbol{A}\| \, \|\boldsymbol{A^{-1}}\| \tag{3.22}$$

3.2 Stability of Numerical Methods

We shall henceforth suppose the problem Eq. (3.1) to be well-posed and a numerical method for the approximate solution of Eq. (3.1) will consist, in general, of a sequence of approximate problems:

$$F(x_n, d_n) = 0 \tag{3.23}$$

depending on a certain parameter n (to be defined case by case). The undetstood expectation is that $x_n \to x$ as $n \to \infty$, that is, the sequence of approximate solutions **converges** to the exact solution.

For that, it is necessary that $d_n \to d$ and $F_n \to F$, as $n \to \infty$. Precisely, if the datum d of Eq. (3.1) is admissible for F_n, we say that Eq. (3.23) is consistent if:

$$F_{n(x,d)} = F_{n(x,d)} - F(x, d) \to 0 \text{ for } n \to \infty \tag{3.24}$$

3.2.1 Relations between Stability and Covergence

The concepts of stability and convergence are strongly connected.

Thm **theorem 3.2.1**

If problem Eq. (3.1) is well-posed, a *necessary* condition in order for the numerical problem Eq. (3.23) to be convergent is that it is stable.

4 Sparse matrices

4.1 Sparse matrices storage formats

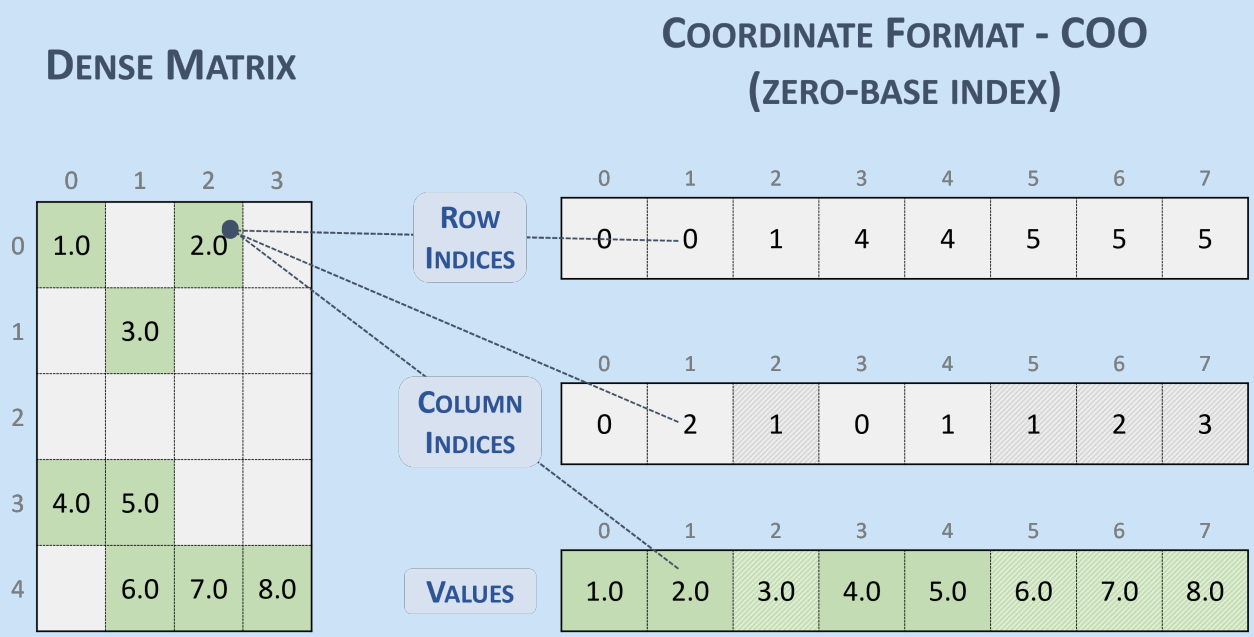
Sparse matrices are matrices that contain a large number of zero elements. The storage of these matrices can be optimized by using different formats. The most common formats are:

4.1.1 Coordinate format (COO)

The simplest storage scheme for sparse matrices is the so-called coordinate format. The data structure consists of three arrays:

1. `AA` - all the values of the nonzero elements of A in any order.
2. `JR` - the row indices of the nonzero elements of A.
3. `JC` - the column indices of the nonzero elements of A.

e.g. Coordinate format **example 4.1.1**



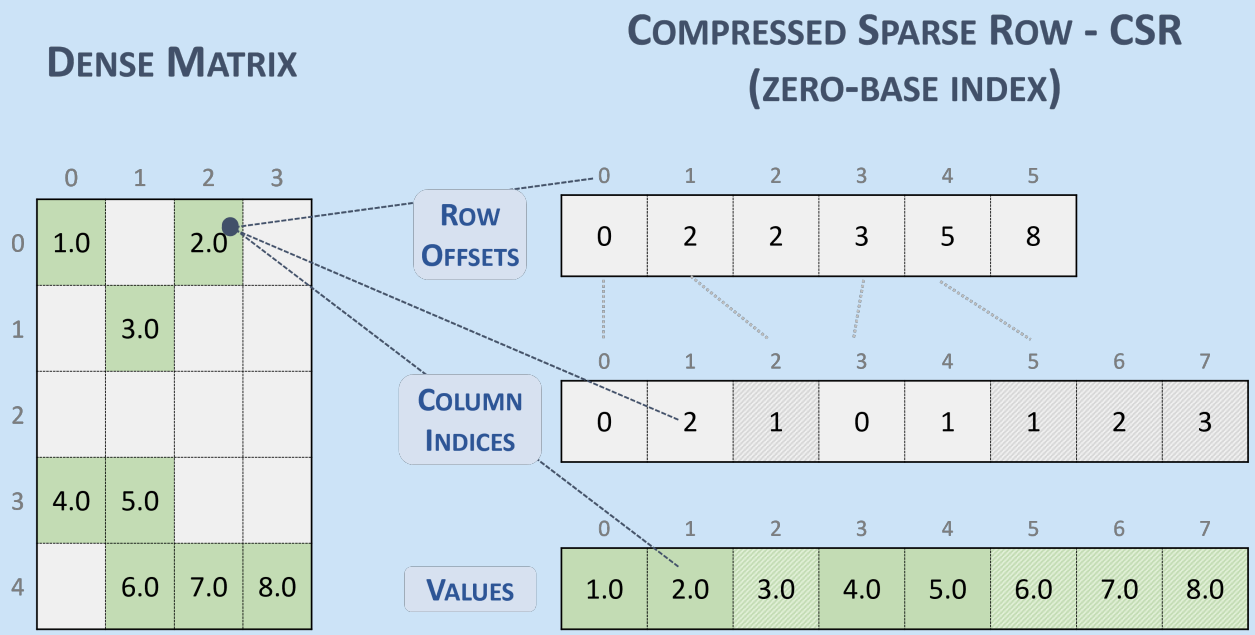
4.1.2 Compressed sparse row (CSR)

The CSR format is similar to COO, where the row indices are compressed and replaced by an array of offsets. The new data structure consists of three arrays:

1. `AA` - the real values a_{ij} sorted row by row, from row 1 to row n.
2. `JA` - the column indices of the nonzero elements of A.
3. `IA` - the row offsets. contains the pointers to the beginning of each row in the array A and JA. The content of IA is the position in the arrays AA and JA where the row i

starts. The length of IA is $n+1$, with $IA(n+1)$ containing the total number of nonzero elements in the matrix.

e.g. Compressed sparse row format **example 4.1.2**



To create a sparse matrix in the CSR format, we use the `csr_matrix` function, which is provided by the `scipy.sparse` module. Here is an example program:

```
import scipy.sparse as sp
from scipy import *

data = [1.0, 2.0, -1.0, 6.6, 1.4]
rows = [0, 1, 1, 3, 3]
cols = [1, 1, 2, 0, 4]

A = sp.csr_matrix((data, [rows, cols]), shape=(4, 5))
print(A)

>>> A.data
array([ 1. ,  2. , -1. ,  6.6,  1.4])
>>> A.indices
array([1, 1, 2, 0, 4], dtype=int32)
>>> A.indptr
array([0, 1, 3, 3, 5], dtype=int32)
```

5 Iterative methods for large linear systems

Given an $n \times n$ real matrix A and a real n-vector, the problem is: Find x belonging to R^n such that

$$Ax = b \tag{5.1}$$

where $\boldsymbol{x}$ is the exact solution of the linear system $A\boldsymbol{x} = \boldsymbol{b}$.

5.1 On the Convergence of Iterative Methods

The basic idea of iterative methods is to construct a sequence of vectors $\boldsymbol{x^k}$ that enjoy the property of *convergence*

$$\boldsymbol{x} = \lim_{k \to \infty} \boldsymbol{x^k} \tag{5.2}$$

In practice, the iterative process is stopped at the minimum value of n such that $\|\boldsymbol{x^{(n)}} - \boldsymbol{x}\| < \varepsilon$, where ε is a given tolerance and $\|\cdot\|$ is a suitable norm. However, since the exact solution is obviously not available, it is necessary to introduce suitable stopping criteria to monitor the convergence of the iteration.

To start with, we consider iterative methods of the form

$$\text{Given } \boldsymbol{x}^0, \boldsymbol{x}^{k+1} = B\boldsymbol{x^k} + \boldsymbol{f}, k \geq 0 \tag{5.3}$$

where B is an $n \times n$ square matrix called the *iteration matrix* and $\boldsymbol{f}$ is a vector that is obtained from the right-hand side $\boldsymbol{b}$.

having denoted by B an $n \times n$ square matrix called the iteration matrix and by $\boldsymbol{f}$ a vector that is obtained from the right hand side $\boldsymbol{b}$.

Def **definition 5.1.1**

An iterative method of the form Eq. (5.3) is said to be *convergent* with Eq. (5.2) if $\boldsymbol{f}$ and B are such that $\boldsymbol{x} = \boldsymbol{Bx} + \boldsymbol{f}$. Equivalently,

$$\boldsymbol{f} = (1 - B)A^{-1}\boldsymbol{b} \tag{5.4}$$

Having denoted by

$$\boldsymbol{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x} \tag{5.5}$$

the error at the k-th step of the iteration, the condition for convergence amounts to requiring that $\lim_{k \to \infty} \boldsymbol{e^k} = 0$ for any choice of the initial datum $\boldsymbol{x}^0$.

Thm **theorem 5.1.1**

Let Eq. (5.3) be a consistent method. Then, the sequence of vectors $\{x^k\}$ converges to the solution of Eq. (5.1) for any choice of $x^{(0)}$ iff $\rho(B) < 1$.

Proof. From Eq. (5.5) and the consistency assumption, the recursive relation $\boldsymbol{e^{k+1}} = Be^k$ is obtained. Therefore,

$$e^{(k)} = B^k e^{(0)}, \forall k = 0, 1, ... \tag{5.6}$$

Thus, thanks to Theorem 1.5, it follows that $\lim_{k\to\infty} B^k e^0 = 0$ for any $e^{(0)}$ iff $\rho(B) < 1$.

Def **definition 5.1.2**

Let B be the iteration matrix. We call:

1. $\|\boldsymbol{B^m}\|$ the *convergence factor* after m steps of the iteration.
2. $\|\boldsymbol{B}\|^{1/m}$ the *average convergence* factor after m steps;
3. $R_{m(B)} = -\frac{1}{m}\log\|\boldsymbol{B^m}\|$ the *average convergence rate* after m steps.

5.2 Linear Iterative Methods

A general technique to devise consistent linear iterative methods is based on an additive splitting of the matrix A of the form $A = P - N$, where P and N are two suitable matrices and P is nonsingular. For reasons that will be clear in the later sections, P is called *preconditioning matrix or preconditioner*.

Precisely, given $x^{(0)}$, one can compute $x^{(k)}$ for $k \geqslant$, solving the system:

$$Px^{(k+1)} = Nx^{(k)} + b \tag{5.7}$$

The iteration matrix of method Eq. (5.7) is $B = P^{-1}N$ and the vector $f = P^{-1}b$. Alternatively, the method can be written as:

$$x^{(k+1)} = x^{(k)} + P^{-1}r^{(k)} \tag{5.8}$$

where the residual $r^{(k)} = b - Ax^{(k)}$ is the vector that measures the error in the approximation $x^{(k)}$. Eq. (5.8) outlines the fact that a linear system, with coefficient matrix P, must be solved to update the solution at step $k + 1$. Thus P, besides being nonsingular, ought to be easily invertible, in order to keep the overall computational cost low.

5.2.1 Jacobi, Gauss-Seidel and Relaxation Methods

5.2.1.1 Jacobi Method and Over-Relaxation

If the diagonal entries of A are nonzero, we can single out in each equation the corresponding unknown, obtaining the equivalent linear system.

$$x_i = \frac{b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j}{a_{ii}}, i = 1, ..., n \tag{5.9}$$

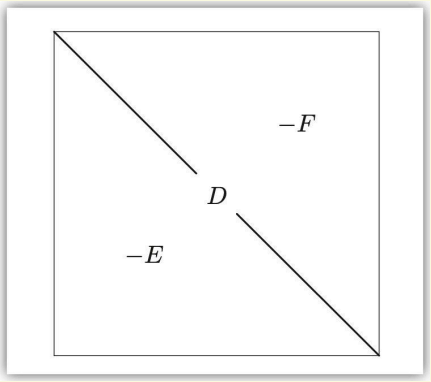
In the Jacobi method, once an arbitrarily initial guess $x^{(0)}$ is given, the solution is updated by the formula:

$$x_i^{(k+1)} = \frac{b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)}}{a_{ii}}, i = 1, ..., n \tag{5.10}$$

This amounts to performing the following splitting for A:

$$P = D, N = D - A = E + F$$

where D is the diagonal matrix of the diagonal entries of A, E is the lower triangular matrix, and F is the upper triangular matrix:



The iteration matrix of the Jacobi method is thus given by

$$B_j = D^{-1}(E + F) = I - D^{-1}A \tag{5.11}$$

A generalization of the Jacobi method is the over-relaxation method(or JOR), in which, having introduced a relaxation parameter ω, Eq. (5.10) is replaced by:

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \omega \frac{b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)}}{a_{ii}}, i = 1, ..., n \tag{5.12}$$

The corresponding iteration matrix is:

$$B_{j_w} = \omega B_j + (1 - \omega)I \tag{5.13}$$

This method is consistent if any $\omega \neq 0$ and for $\omega = 1$ it coincides with the Jacobi method.

5.2.1.2 The Gauss Seidel method

6 Numerical methods for overdetermined linear systems of equations

7 Solving large scale eigenvalue problems

7.1 Eigenvalues and Eigenvectors

Given a square matrix $A \in \mathbb{C}^{n\times n}$, find a scalar λ and a nonzero vector $x \in \mathbb{C}^n$ such that:

$$Ax = \lambda x \tag{7.1}$$

where:

1. The vector x is the *eigenvector*, And the scalar λ is the *eigenvalue*
2. The set of all the eigenvalues of a matrix A is called the *spectrum* of A, denoted by $\sigma(A)$.
3. The maximum modulus of all the eigenvalues is called the spectral radius of A and is denoted by $\rho(A) = \max_{\{\lambda\in\sigma(A)\}}|\lambda|$.

Remarks

1. The eigenvalues of a matrix are the roots of the characteristic polynomial $\det(A - \lambda I) = 0$.
2. From the Fundamental Theorem of Algebra, an $n \times n$ matrix has exactly n eigenvalues, counting multiplicities.
3. Each λ_i may be real but in general is a complex number
4. The eigenvalues $\lambda_1, \lambda_2, ..., \lambda_n$may not all have distinct values
5. Rayleigh quotient: $\lambda_i = \frac{x_i^H A x_i}{x_i^H x_i}$

7.2 The Power Method

Let $A \in \mathbb{C}^{n\times n}$ be a diagonalizable matrix and let $X \in \mathbb{C}^{n\times n}$ be the matrix of its eigenvectors $\boldsymbol{x_i}$. for $i = 1, ..., n$. Let us also suppose that the eigenvalues of A are ordered as

$$|\lambda_1| < |\lambda_2| \leq ... \leq |\lambda_n| \tag{7.2}$$

. Where λ_1 has algebraic multiplicity equal to 1. Under these assumptions, λ_1, is called the *dominant eigenvalue* of A.

Given an arbitrary initial vector $\boldsymbol{q}_0 \in \mathbb{C}^n$ with unitary Euclidean norm, consider for $k = 1, 2, ...$ the following iteration based on the computation of powers ofmatrices, commonly known as the *power method*:

$$\begin{aligned} \boldsymbol{z}^{(k)} &= A\boldsymbol{q}^{(k-1)} \\ \boldsymbol{q}^{(k)} &= \frac{\boldsymbol{z}^{(k)}}{\|\boldsymbol{z}^{(\boldsymbol{k})}\|} \\ \nu^{(k)} &= \boldsymbol{q}^{((k))^H} A\boldsymbol{q}^{(k)} \end{aligned} \tag{7.3}$$

THE POWER METHOD

$\boldsymbol{q}_0$ =some initial vector with $\|\boldsymbol{q_0}\| = 1$
For $k = 1, 2, ...$
| Apply A: $\boldsymbol{z}^{(k)} = A\boldsymbol{q}^{(k-1)}$

THE POWER METHOD

Normalize: $\boldsymbol{q}^{(k)} = \frac{\boldsymbol{z}^{(k)}}{\|\boldsymbol{z}^{(k)}\|}$
Compute Rayleigh quotient: $\nu^{(k)} = \boldsymbol{q}^{((k))^H} A\boldsymbol{q}^{(k)}$

Let us analyze the convergence of the power method. By induction on k, we have that:

$$\boldsymbol{q}^{(k)} = A^k \frac{\boldsymbol{q}^{(0)}}{\|\boldsymbol{A^k q^{(0)}}\|}, k \geq 1 \tag{7.4}$$

This relation explains the role played by the powers of A in the method. Because A is diagonalizable, its eigenvectors $\boldsymbol{x}_i$ form a basis of $\mathbb{C}^n$ and we can write:

$$\boldsymbol{q}^{(0)} = \sum_{i=1}^{n} \alpha_i \boldsymbol{x}_i \tag{7.5}$$

Moreover, since $A\boldsymbol{x}_i = \lambda_i \boldsymbol{x}_i$, we have:

$$\begin{aligned} A^k \boldsymbol{q}^{(0)} &= \sum_{i=1}^{n} \alpha_i \lambda_i^k \boldsymbol{x}_i \\ &= \alpha_1 \lambda_1^k \left(\boldsymbol{x}_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \boldsymbol{x}_i \right) \end{aligned} \tag{7.6}$$

Since $\left|\frac{\lambda_i}{\lambda_1}\right| < 1$ for $i = 2, ..., n$, as k increases the term $\sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1}\right)^k \boldsymbol{x}_i$ tends to assume an increasingly significant component in the direction of the eigenvector $\boldsymbol{x}_1$, while ts components in the other directions $\boldsymbol{x}_j$ decrease.

As $k \to \infty$, the vector $\boldsymbol{q}^{(k)}$ thus aligns itself along the direction of eigenvector $\boldsymbol{x}_1$, and the following error estimate holds at each step k.

Thm Convergence of the Power Method **theorem 7.2.1**

Let $A \in \mathbb{C}^{n \times n}$ be a diagonalizable matrix whose dominant eigenvalue is λ_1. Assuming that $\alpha_1 \neq 0$, there exists a constant $C > 0$ such that:

$$\|\tilde{\boldsymbol{q}}^{(\boldsymbol{k})} - \boldsymbol{x_1}\| \leq C \left(|\frac{\lambda_2}{\lambda_1}| \right)^k, k \geq 1 \tag{7.7}$$

where:

$$\tilde{\boldsymbol{q}}^{(k)} = \boldsymbol{x}_1 + \sum_{i=2}^{n} \frac{\alpha_i}{\alpha_1} \left(\frac{\lambda_i}{\lambda_1} \right)^k \boldsymbol{x}_i, k = 1, 2, ... \tag{7.8}$$

Estimate Eq. (7.7) expresses the convergence of the sequence of $\tilde{\boldsymbol{q}}^{(k)}$ towards the eigenvector $\boldsymbol{x}_1$ of A. Therefore the sequence of Rayleigh quotients

$$\tilde{\boldsymbol{q}}^{((k))^H} A \tilde{\boldsymbol{q}}^{(k)} / \|\tilde{\boldsymbol{q}}^{(\boldsymbol{k})}\| = \left(\boldsymbol{q}^{(k)}\right)^H A \boldsymbol{q}^{(k)} = \nu^{(k)} \tag{7.9}$$

will converge to the dominant eigenvalue λ_1 of A. As a consequence, and the convergence will be faster when the ratio $|\frac{\lambda_2}{\lambda_1}|$ is smaller.

7.3 Deflation

7.4 The Inverse Power Method

We look for an approximation of the eigenvalue of a matrix $A \in \mathbb{C}^{n\times n}$ which is *closest* to a given number $\mu \in \mathbb{C}$, where $\mu \notin \sigma(A)$. For this, the power iteration is applied to the matrix $\left(M_\mu\right)^{-1} = (A-\mu I)^{-1}$, yielding the so-called *inverse iteration* or *inverse power method*. The number μ is called the *shift* of the method.

The eigenvalues of M_u^{-1} are $\xi = (\lambda_i - \mu)^{-1}$, let us assume that there exists an integer m such that

$$|\lambda_m - \mu| < |\lambda_i - \mu| \tag{7.10}$$

Given an arbitrary initial vector $\boldsymbol{q}_0 \in \mathbb{C}^n$ with unitary Euclidean norm, for $k = 1, 2, ...$ the following sequence is constructed:

$$\begin{aligned} (A-\mu I)\boldsymbol{z}^{(k)} &= \boldsymbol{q}^{(k-1)} \\ \boldsymbol{q}^{(k)} &= \frac{\boldsymbol{z}^{(k)}}{\|\boldsymbol{z}^{(\boldsymbol{k})}\|} \\ \nu^{(k)} &= \boldsymbol{q}^{((k))^H} A\boldsymbol{q}^{(k)} \end{aligned} \tag{7.11}$$

THE INVERSE POWER METHOD

$\boldsymbol{q}_0$ =some initial vector with $\|\boldsymbol{q_0}\| = 1$
For $k = 1, 2, ...$
  Apply $(A-\mu I)$: $\boldsymbol{z}^{(k)} = (A-\mu I)\boldsymbol{q}^{(k-1)}$
  Normalize: $\boldsymbol{q}^{(k)} = \frac{\boldsymbol{z}^{(k)}}{\|\boldsymbol{z}^{(\boldsymbol{k})}\|}$
  Compute Rayleigh quotient: $\nu^{(k)} = \boldsymbol{q}^{((k))^H} A\boldsymbol{q}^{(k)}$

Notice that the eigenvectors of M_μ are the same as those of A since $M_\mu = X(\Lambda - \mu I_n)X^{-1}$, where $\Lambda = \text{diag}(\lambda_1, ..., \lambda_n)$. For this reason, the Rayleigh quotient in Eq. (7.11) is computed directly on the matrix A. The main difference with respect to Eq. (7.3) is that at each step k a linear system with coefficient matrix $M_\mu = A - \mu I$ *must be solved.*

7.5 QR Factorization