Numerical Linear Algebra

Rao

Politecnico di Milano

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

Last updated at: **2024-10-08**

Contents

1 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.

1.1 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*: $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$.
2. *Triangle Inequality*: $\|x + y\| \leq \|x\| + \|y\|$.
3. *Homogeneity*: $\|\alpha x\| = |\alpha| \; \|x\|$.

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

$$\begin{aligned} \|x\|_1 &= \sum_{i=1}^{m} |x_i| \\ \|x\|_2 &= \sqrt{\sum_{i=1}^{m} |x_i|^2} \\ \|x\|_\infty &= \max_{\{1<i<m\}} |x_i| \\ \|x\|_p &= \left(\sum_{i=1}^{m} |x_i|^p \right)^{\frac{1}{p}} (p \leq 1 < \infty) \end{aligned} \tag{1.1}$$

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:

$$\|x\|_w = \|Wx\| \tag{1.2}$$

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:

$$\|x\|_W = \left(\sum_{i=1}^{m} |w_i x_i|^2 \right)^{\frac{1}{2}} \tag{1.3}$$

Thm Cauchy-Schwarz Inequality **theorem 1.1.1**

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

$$|(x,y)| = |x^T y| \leq \|x\|_2 \|y\|_2 \tag{1.4}$$

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:

$$|(x,y)| \leq \|x\|_p \|y\|_q \qquad \text{Where} \qquad \frac{1}{p} + \frac{1}{q} = 1 \tag{1.5}$$

.

Thm **theorem 1.1.2**

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

Thm **theorem 1.1.3**

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:

$$\|x\|_{A^2} = \|Ax\| \tag{1.6}$$

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

Thm **theorem 1.1.4**

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

$$\lim_{k \to \infty} x^{(k)} = x \Longleftrightarrow \lim_{k \to \infty} \left\| x^{(k)} - x \right\| = 0 \tag{1.7}$$

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

1.2 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.

Given vector norms $\|\cdot\|_n$ and $\|\cdot\|_m$ on the domain and the range of $A \in \mathbb{C}^{m \times n}$, respectively, the induced matirx norm $\|A\|_{(m,n)}$ is the smallest number C for which the following inequality holds for all $x \in \mathbb{C}^n$:

$$\|Ax\|_{(m)} \leq C \|x\|_{(n)} \tag{1.8}$$

Def **definition 1.2.1**

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

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

Def **definition 1.2.2**

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

$$\|Ax\| \leq \|A\| \|x\| \qquad \forall x \in \mathbb{R}^n \tag{1.9}$$

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 $\|y\| \leq \|A\| \|x\|$.

Def **definition 1.2.3**

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

$$\|AB\| \leq \|A\| \|B\| \tag{1.10}$$

The norm

$$\|A\|_F = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2} = \sqrt{\mathrm{tr}(AA^H)} \tag{1.11}$$

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

$$\|Ax\|_2^2 = \sum_{i=1}^{m} | \sum_{j=1}^{n} a_{ij} x_j |^2 \leq \sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \sum_{j=1}^{n} |x_j|^2 = \|A\|_F^2 \|x\|_2^2 \tag{1.12}$$

Thm **theorem 1.2.1**

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

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} \tag{1.13}$$

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

Proof: Check definition 1.2.1.
1. If $\|Ax\| \geq 0$, then it follows that $\|A\| = \sup_{\|x\|=1}\|Ax\| \geq 0$. Moreover,

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = 0 \Longleftrightarrow \|Ax\| = 0 \forall x \neq 0 \tag{1.14}$$

and $Ax = 0 \forall x \neq 0$ if and only if $A = 0$; therefore, $\|A\| = 0$ if and only if $A = 0$.
2. Given a scalar α, we have that:

$$\|\alpha A\| = \sup_{x \neq 0} \frac{\|\alpha Ax\|}{\|x\|} = \sup_{x \neq 0} |\alpha| \frac{\|Ax\|}{\|x\|} = |\alpha| \sup_{x \neq 0} \frac{\|Ax\|}{\|x\|} = |\alpha| \ \|A\| \tag{1.15}$$

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

$$\frac{\|Ax\|}{\|x\|} \leq \|A\| \Rightarrow \|Ax\| \leq \|A\|\|x\| \tag{1.16}$$

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

$$\|(A+B)x\| \leq \|Ax\| + \|Bx\| \leq \|A\| + \|B\| \tag{1.17}$$

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

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

$$\|A\|_p = \sup_{\boldsymbol{x} \neq 0} \frac{\|A\boldsymbol{x}\|_p}{\|\boldsymbol{x}\|_p} \tag{1.18}$$

The 1-norm(column sum norn):

$$\|A\|_1 = \max_{j=1,\ldots,n}^{n} \sum_{i=1}^{m} |a_{ij}| \tag{1.19}$$

The infinity-norm(row sum norm):

$$\|A\|_\infty = \max_{i=1,\ldots,m}^{m} \sum_{j=1}^{n} |a_{ij}| \tag{1.20}$$

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

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

Thm **Spectral Norm** **theorem 1.2.2**

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

$$\|A\|_2 = \sqrt{\rho(A^H A)} = \sqrt{\rho(A^H A)} == \sigma_1(A) \tag{1.21}$$

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

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

$$U^H A^H A U = \operatorname{diag}(\mu_1, ..., \mu_n) \tag{1.22}$$

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

$$\begin{aligned} \|A_2\| &= \sup_{x \neq 0} \frac{\sqrt{(A^H A x, x)}}{\sqrt{(x, x)}} = \sup_{y \neq 0} \frac{\sqrt{(U^H A^H A U y, y)}}{\sqrt{(y, y)}} \\ &= \sup_{y \neq 0} \frac{\sqrt{\sum_{i=1}^n \mu_i \ |y_i|^2}}{\sqrt{\sum_{i=1}^n |y_i|^2}} = \sqrt{\max_{i=1,\ldots,n}^{n} |\mu_i|} \end{aligned} \tag{1.23}$$

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

$$\|A\|_2 = \sup_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \sup_{x \neq 0} \frac{\|x\|_2}{\|x\|_2} = 1 \tag{1.24}$$

2 Principles of Numerical Mathematics

2.1 Well-posedness and Condition Number of a Problem

Consider the following problem: find x such that:

$$F(x, d) = 0 \tag{2.1}$$

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. (2.1) 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. (2.1) 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.

e.g. **example 2.1.1**

For example, a *well-posed(well-conditioned)* problem is one with the property that all small perturbations of x lead to only small changes in $f(x)$. An *ill-posed(ill-conditioned)* problem is one for which small perturbations of x can lead to large changes in $f(x)$.

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:

$$F(x + \delta x, d + \delta d) = 0 \tag{2.2}$$

Then, we require that:

$$\begin{gathered} \exists \eta_0 = \eta_0(d) > 0, \exists K_0 = K_0(d) \text{ such that} \\ \text{if } \|\delta d\| \leqslant \eta_0 \text{ then } \|\delta x\| \leqslant K_0 \|\delta d\| \end{gathered} \tag{2.3}$$

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

The Eq. (2.3) 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 2.1.1**

For problem Eq. (2.1), we define the *relative conditional number* to be:

$$K(d) = \sup\left\{ \frac{\frac{\|\delta x\|}{\|x\|}}{\frac{\|\delta d\|}{\|d\|}}, \delta d \neq 0, d + \delta d \in D \right\} \tag{2.4}$$

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

$$K_{\text{abs}}(d) = \sup\left\{ \frac{\|\delta x\|}{\|\delta d\|}, \delta d \neq 0, d + \delta d \in D \right\} \tag{2.5}$$

2.1.1 Absolute Condition Number

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 $\|\delta x\| \to 0$. The absolute condition number is then:

$$K = \|J(x)\| \tag{2.6}$$

2.1.2 Relative Condition Number

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

$$K = \frac{\|J(x)\|}{\|fx\|/\|x\|} \tag{2.7}$$

Problem Eq. (2.1) 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).

2.2 Stability of Numerical Methods

We shall henceforth suppose the problem Eq. (2.1) to be well-posed and a numerical method for the approximate solution of Eq. (2.1) will consist, in general, of a sequence of approximate problems:

$$F(x_n, d_n) = 0 \tag{2.8}$$

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. (2.1) is admissible for F_n, we say that Eq. (2.8) is consistent if:

$$F_{n(x,d)} = F_{n(x,d)} - F(x,d) \to 0 \text{ for } n \to \infty \tag{2.9}$$

2.2.1 Relations between Stability and Covergence

The concepts of stability and convergence are strongly connected.

Thm **theorem 2.2.1**

If problem Eq. (2.1) is well-posed, a *necessary* condition in order for the numerical problem Eq. (2.8) to be convergent is that it is stable.

3 Sparse matrices

3.1 Sparse matrices storage formats

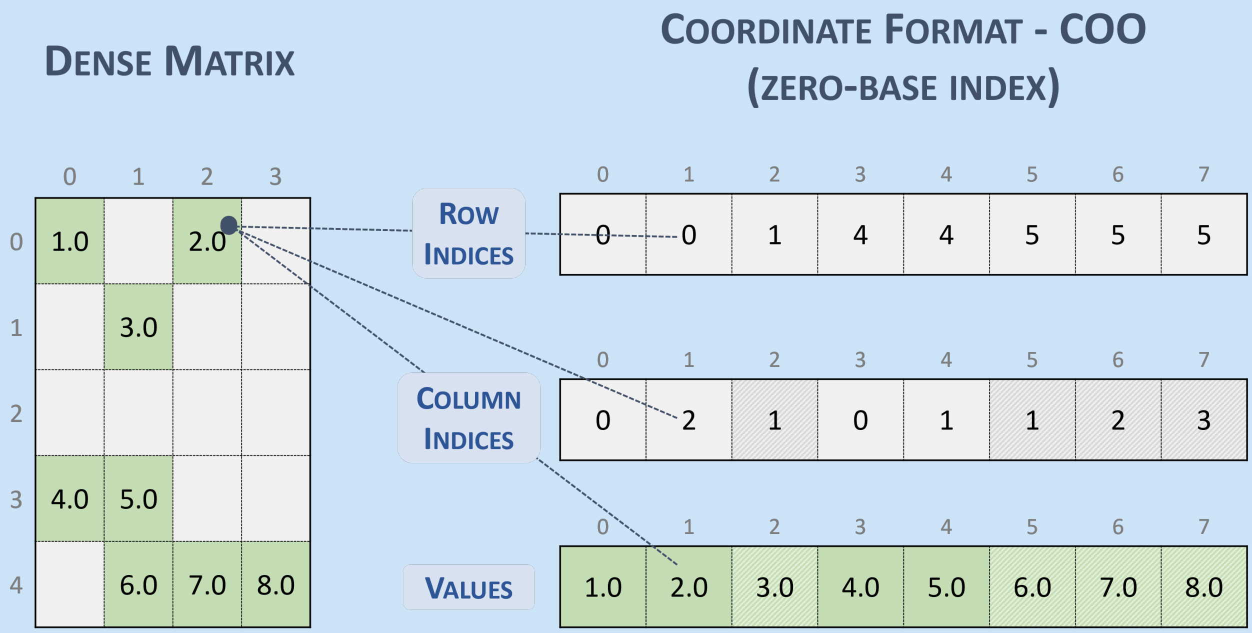
Sparse matrices are matrices that contain a large number of zero elements. The storage of these matrices can be optimized by using different formats. The most common formats are:

3.1.1 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 3.1.1



	0	1	2	3
0	1.0		2.0	
1		3.0		
2				
3	4.0	5.0		
4		6.0	7.0	8.0

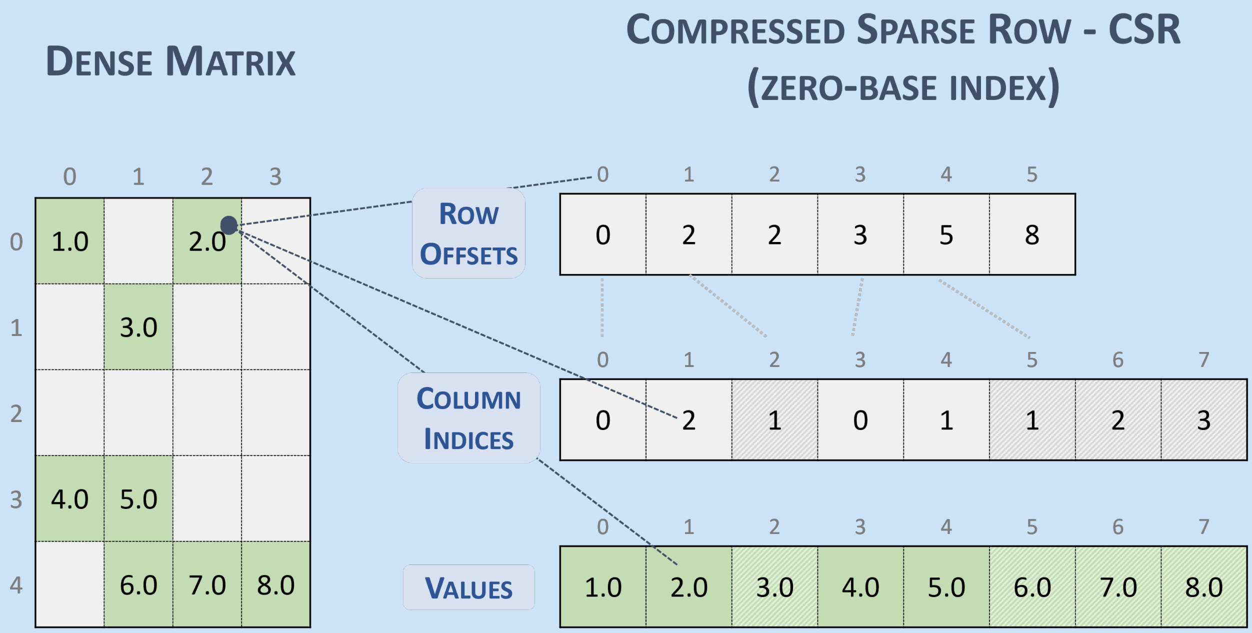
	0	1	2	3	4	5	6	7
ROW INDICES	0	0	1	4	4	5	5	5
COLUMN INDICES	0	2	1	0	1	1	2	3
VALUES	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0

3.1.2 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 3.1.2



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:

```python
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)
```

4 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

$$Ax = b \tag{4.1}$$

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

4.1 On the Convergence of Iterative Methods

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

$$\boldsymbol{x} = \lim_{k \to \infty} \boldsymbol{x}^{\boldsymbol{k}} \tag{4.2}$$

In practice, the iterative process is stopped at the minimum value of n such that $\left\|\boldsymbol{x}^{(\boldsymbol{n})} - \boldsymbol{x}\right\| < \varepsilon$, 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

$$\text{Given } \boldsymbol{x}^0, \boldsymbol{x}^{k+1} = B\boldsymbol{x}^{\boldsymbol{k}} + \boldsymbol{f}, k \geq 0 \tag{4.3}$$

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 4.1.1**

An iterative method of the form Eq. (4.3) is said to be *convergent* with Eq. (4.2) if $\boldsymbol{f}$ and B are such that $\boldsymbol{x} = \boldsymbol{B}\boldsymbol{x} + \boldsymbol{f}$. Equivalently,

$$\boldsymbol{f} = (1 - B)A^{-1}\boldsymbol{b} \tag{4.4}$$

Having denoted by

$$\mathbf{e}^{(k)} = \boldsymbol{x}^{(k)} - \boldsymbol{x} \tag{4.5}$$

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

Thm **theorem 4.1.1**

Let Eq. (4.3) be a consistent method. Then, the sequence of vectors $\{x^k\}$ converges to the solution of Eq. (4.1) for any choice of $x^{(0)}$ iff $\rho(B) < 1$.
Proof. From Eq. (4.5) and the consistency assumption, the recursive relation $\boldsymbol{e}^{k+1} = B\boldsymbol{e}^k$ is obtained. Therefore,

$$e^{(k)} = B^k e^{(0)}, \forall k = 0, 1, ... \tag{4.6}$$

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 4.1.2**

Let B be the iteration matrix. We call:

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

4.2 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:

$$Px^{(k+1)} = Nx^{(k)} + b \tag{4.7}$$

The iteration matrix of method Eq. (4.7) is $B = P^{-1}N$ and the vector $f = P^{-1}b$. Alternatively, the method can be written as:

$$x^{(k+1)} = x^{(k)} + P^{-1}r^{(k)} \tag{4.8}$$

where the residual $r^{(k)} = b - Ax^{(k)}$ is the vector that measures the error in the approximation $x^{(k)}$. Eq. (4.8) 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.

4.2.1 Jacobi, Gauss-Seidel and Relaxation Methods

4.2.1.1 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.

$$x_i = \frac{b_i - \sum_{\substack{j=1 \\ j\neq i}}^{n} a_{ij}x_j}{a_{ii}}, i = 1, ..., n \tag{4.9}$$

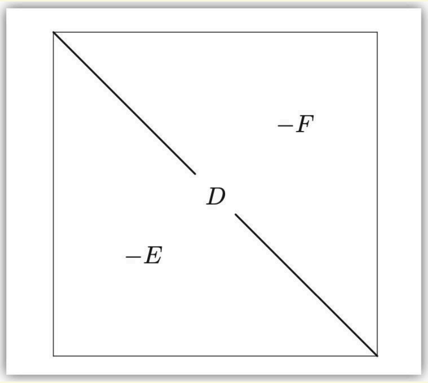
In the Jacobi method, once an arbitrarily initial guess $x^{(0)}$ is given, the solution is updated by the formula:

$$x_i^{(k+1)} = \frac{b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)}}{a_{ii}}, i = 1, ..., n \tag{4.10}$$

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

$$B_j = D^{-1}(E + F) = I - D^{-1}A \tag{4.11}$$

A generalization of the Jacobi method is the over-relaxation method(or JOR), in which, having introduced a relaxation parameter ω, Eq. (4.10) is replaced by:

$$x_i^{(k+1)} = (1 - \omega)x_i^{(k)} + \omega \frac{b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)}}{a_{ii}}, i = 1, ..., n \tag{4.12}$$

The corresponding iteration matrix is:

$$B_{j_w} = \omega B_j + (1 - \omega)I \tag{4.13}$$

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

4.2.1.2 The Gauss Seidel method

5 Numerical methods for overdetermined linear systems of equations