Chapter 3

linear Systems

The two fundamental problems of numerical matrix analysis are: the resolution of linear systems on the one hand and the calculation of eigenvalues and eigenvectors of matrices on the other.

This chapter focuses on solving a linear system. Given a matrix A and a vector b, find a solution to the system $Au = b$.

3.0.1 Perturbed Systems and Conditioning

Conditioning measures the influence of rounding errors on the solution of a given problem. It is highlighted by a slight perturbation of the initial data. It is a general notion that applies both to the roots of a polynomial with respect to the variation of its coefficients and to the eigenvalues or eigenvectors of a matrix with respect to the perturbation of its elements. Consider the following linear system $Ax = b$: We consider the linear system, where

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} u_1 \\ v_2 \\ v_3 \\ v_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 22 \\ 33 \\ 30 \end{pmatrix}$$

with solution $u = (1, 1, 1, 1)^t$.
We consider the perturbed system where the right-hand sides have been very slightly modified, the matrix remaining unchanged.

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} u_1 + \delta u_1 \\ v_2 + \delta v_2 \\ v_3 + \delta v_3 \\ v_4 + \delta v_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 22.3 \\ 33 \\ 30.9 \end{pmatrix}$$

with the solution $u = (9.2, -12.6, 4.5, -1.1)^t$.

On other hand, a relative error of order $\sim 1/200$ in the data leads to a relative error of ~ 10 in the result.

Error amplification ratio: ~ 2000! Now perturb the matrix elements slightly:

$$\begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.83 & 9 \\ 6.19 & 4.99 & 9 & 9.98 \end{pmatrix} \begin{pmatrix} u_1 + \Delta u_1 \\ u_2 + \Delta u_2 \\ u_3 + \Delta u_3 \\ u_4 + \Delta u_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix}$$

Solution:$(-81, 137, -34, 22)^t$.

Here again, very small variations in data (here the elements of matrix A) completely modify the result (the solution of the linear system). Given that matrix A of the system has a good shape, it is:

- symmetric
- determinant equals 1.

- it is invertible $A^{-1} = \begin{pmatrix} 10 & 7 & 8.1 & 7.2 \\ 7.08 & 5.04 & 6 & 5 \\ 8 & 5.98 & 9.83 & 9 \\ 6.19 & 4.99 & 9 & 9.98 \end{pmatrix}$

Analysis of results

- In the first case, we are given an invertible matrix A and the task is to compare the exact solutions: u and $u + \delta u$ of the systems $Au = b$ and $A(u + \delta u) = b + \delta b$.

  Let $\|.\|$ be any vector norm and $\|.\|$ a subordinate matrix norm, from equality $\delta u = A^{-1}\delta b$ and $b = Au$, we deduce $\|\delta u\| \leq \|A^{-1}\|\|\delta b\|$ and $\|b\| \leq \|A\|\|u\|$ so that the error in the result measured by the ratio $\frac{\|\delta u\|}{\|u\|}$ is increased according to the relative error $\frac{\|\delta b\|}{\|b\|}$ on data b in the following way

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

- In the second case, it is the matrix that varies, and the goal is to compare the exact solutions u and $u + \Delta u$ of systems $Au = b$ and $(A + \Delta A)(u + \Delta u) = b$, then:

$$\Delta u = -A^{-1}\Delta A(u + \Delta u)$$

$$\Rightarrow \|\Delta u\| \leq \|A^{-1}\| \, \|\Delta A\| \, \|u + \Delta u\|$$

$$\frac{\|\Delta u\|}{\|u + \Delta u\|} \leq \|A\| \, \|A^{-1}\| \, \frac{\|\Delta A\|}{\|A\|} = \text{cond}(A)\frac{\|\Delta A\|}{\|A\|}$$

  Thus, the relative error on the result, measured this time by the ratio $\frac{\|\Delta u\|}{\|u+\Delta u\|}$, is still bounded above in terms of the relative error on the data A.

Remarks 3.1

1. *The above reasoning is valid even if the matrix $(A + \Delta A)$ is singular, provided there exists a vector $(u + \Delta u)$ that is a solution of the second system.*
2. *If $\|\Delta A\|$ is "sufficiently small", one can hope that the ratio $\frac{\|\Delta u\|}{\|u+\Delta u\|}$ is a good approximation of the more "natural" relative error $\frac{\|\Delta u\|}{\|u\|}$.*

In both cases, we observe that the relative error on the result is bounded by the relative error on the data, multiplied by the number $\{\|A\|\|A^{-1}\|\}$. In other words, for the same relative error on the data, the corresponding relative error on the result can be larger when this number is larger; indeed, it will be shown that this number is optimal, the above inequalities being the "best" possible. These considerations lead us to the following definition.

Definition 3.1 *Let $\|\cdot\|$ be a subordinate matrix norm and A an invertible matrix. The number*

$$\mathcal{K}(A) = \|A\|\|A^{-1}\|,$$

*is called the **condition number** of the matrix A, relative to the considered matrix norm.*

Theorem 3.1 *For any matrix A,*

- $\mathcal{K}(A) \geq 1$
- $\mathcal{K}(A) = \mathcal{K}(A^{-1})$
- $\mathcal{K}(\alpha A) = \mathcal{K}(A)$ *for* $\alpha \neq 0$

Proof.

- Properties (1) result from the properties of matrix norms; in particular

$$AA^{-1} = I \Rightarrow 1 = \|I\| \leq \|A\|\|A^{-1}\|,$$

since the matrix norm is subordinate by hypothesis.

- $\mathcal{K}(A) = \mathcal{K}(A^{-1})$
- $\mathcal{K}(\alpha A) = \text{cond}(A)$ for $\alpha \neq 0$

■

Remarks 3.2

1. *For any matrix A,*

$$\mathcal{K}_2(A) = \frac{\mu_n(A)}{\mu_1(A)},$$

where $\mu_1(A) > 0$ and $\mu_n(A) > 0$ denote the smallest and largest singular values of the matrix A, respectively.

2. *If A is a normal matrix,*

$$\mathcal{K}_2(A) = \frac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|},$$

where the numbers $\lambda_i(A)$ are the eigenvalues of the matrix A.

3. *The condition number $\mathcal{K}_2(A)$ of a unitary or orthogonal matrix A is equal to* 1.

4. *The condition number $\mathcal{K}_2(A)$ is invariant under unitary transformation:*

$$UU^* = I \Rightarrow \mathcal{K}_2(A) = \mathcal{K}_2(AU) = \mathcal{K}_2(UA) = \mathcal{K}_2(U^*AU).$$

3.1 Review of direct methods

Introduction

The problem we will now consider is the numerical solution of a linear system $Au = b$, whose matrix A is invertible. The principle of the *direct methods* studied in this chapter boils down to the determination of an invertible matrix M such that *the matrix MA is upper triangular*, this corresponds to the elimination procedure. It then remains to solve the linear system

$$MAu = Mb,$$

using the *backward substitution* method (in *actual* computations, one does not explicitly compute the matrix M, but only the matrix MA and the vector Mb). This principle is the basis of *Gauss's method* for linear systems with "arbitrary" matrices and *Cholesky's method* for linear systems with symmetric positive definite matrices. Along the way, we study the computation of *determinants* of square matrices, the special case (very important in applications) of tridiagonal matrices (pointwise or blockwise), as well as the *Gauss-Jordan method*, which can be considered as a variant of Gauss's method, particularly well-suited for computing the inverse of a given matrix. Let us emphasize in this regard the futility of computing the inverse of a matrix for solving a linear system.
The matrix interpretation of Gauss's method is the *LU factorization* of a matrix. This very important result, notably due to its various applications in Matrix Numerical Analysis, shows that, up to possible row permutations, every invertible matrix can be written as the product of a *lower triangular matrix* L and an *upper triangular matrix* U. This factorization simplifies somewhat in the case of *symmetric positive definite matrices*; it then becomes the *Cholesky factorization*, the basis of the aforementioned Cholesky method. While Gauss's and Cholesky's methods are based on the factorization of the linear system's matrix into a product of a lower triangular matrix and an upper triangular matrix, the

Householder method is associated with the factorization of a matrix (not necessarily symmetric) into a product of an *orthogonal* matrix Q and an upper triangular matrix R. We describe this method in Section 3.1, where we show that such a *QR factorization* can be performed in a clever way using particular "elementary" orthogonal matrices, called Householder matrices.

Note that the QR factorization of a matrix will appear as an essential computational step in the "QR method" for approximating the eigenvalues of an arbitrary matrix (Chapter 4), and that Householder matrices will be used again for reducing a symmetric matrix to tridiagonal form.

Remarks concerning the solution of linear systems

Contrary to what a superficial analysis might suggest, the solution of a linear system $Au = b$, A: invertible matrix, is not obtained by computing the matrix A^{-1}, and then computing the vector $A^{-1}b$ (first remark). The computation of the matrix A^{-1} is indeed equivalent to solving the n linear systems (n: order of the matrix A):

$$Au_j = e_j, \quad 1 \leq j \leq n,$$

where e_j is the j-th vector of the canonical basis of K^n. In other words, such a method would replace the solution of one linear system (the given problem) by the solution of n linear systems, followed by the multiplication of the matrix A^{-1} by the vector b!

The methods we will study are based on the following second obvious remark: if the invertible matrix A is upper triangular (or lower triangular), the numerical solution of a linear system $Au = b$ is immediate; it is written indeed as

$$\begin{aligned} a_{11}u_1 + \ldots + a_{1,n-1}u_{n-1} + a_{1n}u_n &= b_1, \\ a_{n-1,n-1}u_{n-1} + a_{n-1,n}u_n &= b_{n-1}, \\ &\vdots \\ a_{nn}u_n &= b_n, \end{aligned}$$

and since $a_{11}a_{22}\ldots a_{nn} = \det(A) \neq 0$, one solves the system by computing u_n from the last equation, then u_{n-1} from the penultimate one, etc., which gives

$$\begin{aligned} u_n &= a_{nn}^{-1}b_n, \\ u_{n-1} &= a_{n-1,n-1}^{-1}(b_{n-1} - a_{n-1,n}u_n), \\ &\vdots \\ u_1 &= a_{11}^{-1}(b_1 - a_{12}u_2 - \ldots - a_{1,n-1}u_{n-1} - a_{1n}u_n). \end{aligned}$$

Each component u_i thus appearing as a linear function of b_i, b_{i+1}, ..., b_n, this shows incidentally that the inverse of a triangular matrix is a triangular matrix of the same type (upper or lower).

The method above, called the **backward substitution** method, thus requires in total

$$\begin{cases} 1 + 2 + \ldots + (n-1) = \dfrac{n(n-1)}{2} \text{ additions}, \\ 1 + 2 + \ldots + (n-1) = \dfrac{n(n-1)}{2} \text{ multiplications}, \\ n \text{ divisions}. \end{cases}$$

The backward substitution method extends to block triangular matrices. Thus, for example, the solution of the linear system

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ 0 & A_{22} & A_{23} \\ 0 & 0 & A_{33} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

reduces to the solution of the successive linear systems:

$$A_{33}u_3 = b_3, \quad A_{22}u_2 = b_2 - A_{23}u_3, \quad A_{11}u_1 = b_1 - A_{12}u_2 - A_{13}u_3.$$

Naturally, this assumes that one knows how to solve the linear systems whose matrices are the diagonal submatrices A_{ii}.

In the study of methods for solving linear systems (both direct and iterative), we will frequently use matrix notation. In particular, we will consider "intermediate" linear systems, say $Cu = d$, which we will "solve" in the form $u = C^{-1}d$, which seems to contradict the first remark stated above.

In fact, this is only a notational convenience, and if one looks more closely, one will realize that the matrix C is triangular, pointwise or blockwise, so that one solves the system $Cu = d$ using the backward substitution method, without explicitly computing the matrix C^{-1}. This is why it is sometimes said, with an obvious abuse of language, that such matrices C are "easy to invert".

3.1.1 Direct Methods (Gauss, Cholesky, Crout)

Principle: Find invertible M such that MA is upper triangular, then solve $MAu = Mb$.

- Method: Gauss (general).
- Method: Cholesky (symmetric positive definite).
- Method: LU factorization.
- Method:.....

3.1.2 QR Factorization via Householder

The goal of this section is to prove that, every matrix $A \in \mathbb{R}^{m\times n}, m > n$ has a factorisation of the form $A = QR$, where $Q \in \mathbb{R}^{m\times m}$ is Orthogonal and $R \in \mathbb{R}^{m\times n}$ is an upper triangular matrix.

Definition 3.2 *A Householder matrix $H = H(v) \in \mathbb{R}^{m\times n}$ is a matrix of the form*

$$H(v) = I - 2vv^t,$$

where $v \in \mathbb{R}^m$ satisfies either $\|v\|_2 = 1$ or $v = 0$.
A more general form of a Householder matrix is given by

$$H(v) = I - 2\frac{vv^t}{v^tv},$$

for an arbitrary vector $v \neq 0$.

Lemme 3.1 *let $H = H(v) \in \mathbb{R}^{m\times n}$ be a Householder matrix, then*

1. *H is symmetric.*
2. *$HH = I$ so that H is orthogonal.*
3. *$\det(H(v)) = -1$ if $v \neq 0$*

Proof. Not that: $vv^t = 1$, leads ■

Theorem 3.2
For any $A \in \mathbb{C}^{m\times n}$, there exists an orthogonal matrix $Q \in \mathbb{C}^{m\times m}$ and an upper triangular matrix $R \in \mathbb{C}^{m\times n}$, such that $A = QR$.
If A is invertible, the corresponding factorization $A = QR$ is unique.

Proof.

1. Let $u \in \mathbb{R}^n$. In the following, $\|\cdot\|$ and $\langle\cdot,\cdot\rangle$ denote the usual Euclidean norm and inner product on $\mathbb{R}^n$. If $u = 0$ (the zero vector) then $H(u) = I_n$ is symmetric and orthogonal.
   Now suppose $u \neq 0$. We first compute, using the properties of the transpose (linearity and transpose of a product),

$$H(u)^T = I_n^T - 2\frac{(uu^T)^T}{\|u\|^2} = I_n - 2\frac{(u^T)^T u^T}{\|u\|^2} = I_n - 2\frac{uu^T}{\|u\|^2} = H(u),$$

   so $H(u)$ is symmetric.

2. Next, we have

$$H(u)H(u)^T = H(u)^2 = \left(I_n - 2\frac{uu^T}{\|u\|^2}\right)^2$$

$$= I_n^2 - 2 \times I_n \times 2\frac{uu^T}{\|u\|^2} + \left(2\frac{uu^T}{\|u\|^2}\right)^2 \quad \text{because } I_n \text{ and } 2\frac{uu^T}{\|u\|^2} \text{ commute}$$

$$= I_n - 4\frac{uu^T}{\|u\|^2} + 4\frac{u(u^T u)u^T}{\|u\|^4}$$

$$= I_n - 4\frac{uu^T}{\|u\|^2} + 4\frac{uu^T}{\|u\|^2} \quad \text{because } u^T u = \|u\|^2$$

$$= I_n$$

   so $H(u)$ is orthogonal.
   We can better understand what $H(u)$ is by writing for all $v \in \mathbb{R}^n$,

$$H(u)v = v - 2\frac{uu^T v}{\|u\|^2} = v - 2\frac{u(u,v)}{\|u\|^2} = v - 2\frac{\langle v,u\rangle}{\|u\|^2}u = v - 2\left\langle v, \frac{u}{\|u\|}\right\rangle \frac{u}{\|u\|}.$$

   $H(u)$ is the matrix of the orthogonal reflection with respect to the hyperplane orthogonal to u. Here is a small diagram:

3. Let e be a unit vector in $\mathbf{R}^n$, $v \in \mathbf{R}^n$, and assume that v and e are not collinear. We have

$$H(v + \|v\|e)\, v = v - 2\frac{\langle v, v + \|v\|e\rangle}{\|v + \|v\|e\|^2}(v + \|v\|e).$$

   Now,

$$\|v + \|v\|e\|^2 = \|v\|^2 + 2\langle v, \|v\|e\rangle + \|\|v\|e\|\|^2$$
$$= 2\|v\|^2 + 2\|v\|\langle v, e\rangle \quad \text{because } \|e\| = 1$$
$$= 2\langle v, v + \|v\|e\rangle$$

   so

$$H(v + \|v\|e)\, v = v - (v + \|v\|e) = -\|v\|e.$$

   This can also be seen very clearly in a diagram... Draw it!
   Similarly, we show that $H(v - \|v\|e)\, v = \|v\|\, e$.

■

3.2 Construction of Householder matrix

In this section we give the steps how to construct the Householder matrix.

(a) Let $A \in \mathcal{M}_n(\mathbf{R})$. We look for a Householder matrix H such that the matrix HA has only zeros below the diagonal in its first column. Let v_1 be the first column vector of A and e_1 the first vector of the canonical basis. If v_1 is collinear with e_1, we set $H = I_n$ (nothing to do...). Otherwise, we take $H = H(v_1 + \|v_1\|e_1)$ or $H = H(v_1 - \|v_1\|e_1)$. We then have $Hv_1 = -\|v_1\|e_1$ or $Hv_1 = +\|v_1\|e_1$. In both cases, the first column of HA being the vector Hv_1, there are indeed zeros below the diagonal in this column.
Remark: How do we choose one possibility or the other? If doing it "by hand" on paper, do as you wish... If we want the diagonal coefficient $(HA)_{1,1}$ to be positive, we choose $H = H(v_1 - \|v_1\|e_1)$. The choice can also be guided by the desire to avoid dividing by numbers that are too small: since we must divide by $\|v_1 \pm \|v_1\|e_1\|^2$, if $(v_1)_1 = A_{1,1} < 0$, we rather choose $H = H(v_1 - \|v_1\|e_1)$ and if $(v_1)_1 = A_{1,1} > 0$, we rather choose $H = H(v_1 + \|v_1\|e_1)$.

(b) We construct the sequence of Householder matrices $(H^{(k)})_{1\leq k\leq n-1}$ and the sequence of matrices $(A^{(k)})_{0\leq k\leq n-1}$ by recurrence.

- **Step 0.** Set $A^{(0)} = A$;
- **Step 1.** Then set
  - $w_1 = v_1 - \|v_1\|e_1$ where v_1 is the first column vector of A (to fix ideas, we will always take the "$-$" sign);
  - $H^{(1)} = H(w_1)$ (if $v_1 = \|v_1\|e_1$, then $H^{(1)} = I_n$);
  - $A^{(1)} = H^{(1)}A^{(0)}$, then $A^{(1)}$ has zeros below the diagonal in its 1^{st} column.

$$A^{(1)} = \begin{pmatrix} \star & \star & \dots & \dots & \dots & \star \\ 0 & \star & \dots & \dots & \dots & \star \\ \vdots & \vdots & \dots & \dots & \dots & \vdots \\ \vdots & \vdots & \dots & \dots & \dots & \vdots \\ 0 & \star & \dots & \dots & \dots & \star \end{pmatrix}$$

- **Step** $k+1$**.** Let $0 \leq k \leq n-2$, assume we have constructed the matrices $A^{(0)}, \dots, A^{(k)}$, $H^{(1)}, \dots, H^{(k)}$ such that for all $1 \leq j \leq k$, $A^{(j)}$ has zeros below the diagonal in its first j columns, $H^{(j)}$ is a Householder matrix and $A^{(j)} = H^{(j)}A^{(j-1)}$.

$$\text{column numbers} \quad \begin{matrix} 1 & 2 & \dots & k & k+1 & \dots\dots & n \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \end{matrix}$$

$$A^{(k)} = \begin{pmatrix} \star & \star & \dots & \dots & \star & \dots & \dots & \star \\ 0 & \star & \dots & \dots & \vdots & \dots & \dots & \star \\ \vdots & 0 & \ddots & \dots & \vdots & \dots & \dots & \vdots \\ \vdots & \vdots & \ddots & \star & \vdots & \dots & \dots & \vdots \\ \vdots & \vdots & \dots & 0 & \star & \dots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & 0 & \star & \dots & \dots & \star \end{pmatrix}$$

Denote $A^{(k)} = (a_{ij}^{(k)})_{1\leq i,j\leq n}$ and $\widetilde{v}_{k+1} = (a_{i,k+1}^{(k)})_{k+1\leq i\leq n} \in \mathbf{R}^{n-k}$ ($\widetilde{v}_{k+1}$ corresponds to the elements in magenta in the matrix above).

Also denote $\widetilde{e}_{k+1} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbf{R}^{n-k}$.

We reason as in the first step but in dimension $n-k$.

Set $\widetilde{w}_{k+1} = \widetilde{v}_{k+1} - \|\widetilde{v}_{k+1}\|\widetilde{e}_{k+1}$, where $\|\widetilde{v}_{k+1}\|$ is the Euclidean norm of the vector $\widetilde{v}_{k+1}$ in $\mathbf{R}^{n-k}$.

Then set $\widetilde{H}^{(k+1)} = H(\widetilde{v}_{k+1} - \|\widetilde{v}_{k+1}\|\widetilde{e}_{k+1}) \in \mathcal{M}_{n-k}(\mathbf{R})$.

We have

$$\widetilde{H}^{(k+1)}\widetilde{v}_{k+1} = \|\widetilde{v}_{k+1}\|\widetilde{e}_{k+1} = \begin{pmatrix} \star \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbf{R}^{n-k}.$$

Then define $H^{(k+1)}$ by blocks

$$H^{(k+1)} = \left(\begin{array}{c|c} I_k & 0_{\mathcal{M}_{n-k}} \\ \hline 0_{\mathcal{M}_{n-k}} & \widetilde{H}^{(k+1)} \end{array}\right) = H(w_{k+1}), \quad \text{where } w_{k+1} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ \widetilde{w}_{k+1} \end{pmatrix} \in \mathbf{R}^n.$$

Finally, set $A^{(k+1)} = H^{(k+1)}A^{(k)}$: $A^{(k+1)}$ has zeros below the diagonal in its first $k+1$ columns. Thus the matrix $A^{(n-1)}$ is upper triangular.

(c) We have just described $A^{(n-1)} = H^{(n-1)} \dots H^{(1)}A$. But Householder matrices are symmetric and orthogonal, so for all $1 \leq j \leq n-1$, $H^{(j)}$ is invertible with inverse $H^{(j)}$, thus by setting $Q = (H^{(n-1)} \dots H^{(1)})^{-1} = H^{(1)} \dots H^{(n-1)}$ and $R = A^{(n-1)}$, we have

$$A = QR, \quad Q \text{ orthogonal and } R \text{ upper triangular,}$$

this is a QR decomposition of A. With the choice made, we also have all diagonal coefficients of R positive (except possibly the last one).

Operation Count.

To count the number of operations of this algorithm, we need to precisely count the number of operations to perform a matrix $\times$ vector product Hv with H a Householder matrix. As requested in the statement, we only count multiplications here.

Let $0 \leq k \leq n-2$. At step $k+1$ described above, we must:

- compute $\|\widetilde{v}_{k+1}\|$, where $\widetilde{v}_{k+1} \in \mathbf{R}^{n-k}$, so there are $n-k$ multiplications to do;
- compute $\widetilde{w}_{k+1} = \widetilde{v}_{k+1} - \|\widetilde{v}_{k+1}\|\widetilde{e}_{k+1}$, there are no multiplications (only one addition to compute the first component);
- compute the product $H^{(k+1)}A^{(k)}$: $H^{(k+1)}$ is a block matrix, we do not touch the first k columns of $A^{(k)}$ in this calculation, so it is about doing a product of matrices of size $n-k$, i.e., $n-k$ calculations of products of the type $\widetilde{H}^{(k+1)}v$ where v is a vector of $\mathbf{R}^{n-k}$; we write

$$\widetilde{H}^{(k+1)}v = v - 2\frac{\langle v, \widetilde{w}_{k+1}\rangle}{\|\widetilde{w}_{k+1}\|^2}\widetilde{w}_{k+1}.$$

For $\|\widetilde{w}_{k+1}\|^2$: $n-k$ multiplications are necessary; then for each product of the type $\widetilde{H}^{(k+1)}v$ (there are $n-k$ such quantities to compute):

- for $\langle v, \widetilde{w}_{k+1}\rangle$: $n-k$ multiplications are necessary;

- there is also one division and one multiplication by 2;
- a real product $\times$ vector: $n-k$ multiplications are necessary.

The number N_{op} of multiplications necessary for its implementation is therefore

$$N_{op} = \sum_{k=0}^{n-2} \Big((n-k) + (n-k) + (n-k)(n-k+2+n-k) \Big)$$

$$= \sum_{k=0}^{n-2} \Big(4(n-k) + 2(n-k)^2 \Big) = \sum_{l=2}^{n} (2l^2 + 4l)$$

Now $\sum_{l=1}^{n} l^2 = \frac{n(n+1)(2n+1)}{6} \underset{n\to+\infty}{\sim} \frac{n^3}{3}$ and $\sum_{l=1}^{n} l = \frac{n(n+1)}{2} = \underset{n\to+\infty}{o}(n^3)$ so $N_{op} \sim \frac{2n^3}{3}$.

Remark 3.1 *Remark: if we count both the number of multiplications and additions necessary as in exercise 6, we get $N_{op} \sim \frac{4n^3}{3}$. This algorithm is less costly than the Gram-Schmidt algorithm.*

Example 3.1 *Application. Consider the matrix* $A = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix}$.

Step 1. *The first column of A is* $v_1 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, *we have* $\|v_1\| = 1$.

Set $w_1 = v_1 - e_1 = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix}$ *then* $H^{(1)} = H(w_1) = I_3 - 2\frac{w_1 w_1^T}{\|w_1\|^2} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

and finally

$$A^{(1)} = H^{(1)} A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}$$

Step 2. *Set* $\tilde{v}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \in \mathbf{R}^2$, *we have* $\|\tilde{v}_2\| = \sqrt{2}$.

Set $\widetilde{w}_2 = \tilde{v}_2 - \sqrt{2}\tilde{e}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} - \sqrt{2} \begin{pmatrix} 1 \\ 0 \end{pmatrix} = \begin{pmatrix} 1-\sqrt{2} \\ 1 \end{pmatrix}$. *We have* $\|\tilde{w}_2\|^2 = (1-\sqrt{2})^2 + 1^2 = 2(2-\sqrt{2})$

and set $\widetilde{H}^{(2)} = H(\widetilde{w}_2) = I_2 - 2\frac{\widetilde{w}_2 \widetilde{w}_2^T}{\|\widetilde{w}_2\|^2} = \begin{pmatrix} 1 - \frac{(1-\sqrt{2})^2}{2-\sqrt{2}} & \frac{\sqrt{2}-1}{2-\sqrt{2}} \\ \frac{\sqrt{2}-1}{2-\sqrt{2}} & 1 - \frac{1}{2-\sqrt{2}} \end{pmatrix} \in \mathcal{M}_2(\mathbf{R})$.

And now we simplify. When we have fractions with square roots, we multiply by the conjugate expression...

For example, $\frac{\sqrt{2}-1}{2-\sqrt{2}} = \frac{(\sqrt{2}-1)(2+\sqrt{2})}{2^2-\sqrt{2}^2} = \frac{\sqrt{2}}{2}$. *We obtain* $\widetilde{H}^{(2)} = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$.

So we set $H^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}$ *then* $A^{(2)} = H^{(2)} A^{(1)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & -\frac{\sqrt{2}}{2} \end{pmatrix}$.

Conclusion: QR decomposition

The matrix $A^{(2)}$ is upper triangular, we set $R = A^{(2)}$.
We also set $Q = H^{(1)}H^{(2)}$: Q is orthogonal as a product of orthogonal matrices.
We have $A^{(2)} = H^{(2)}A^{(1)} = H^{(2)}H^{(1)}A$, so

$$A = (H^{(2)}H^{(1)})^{-1}A^{(2)} = (H^{(1)})^{-1}(H^{(2)})^{-1}A^{(2)} = H^{(1)}H^{(2)}A^{(2)} = QR$$

.

It remains to compute

$$Q = \begin{pmatrix} 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{pmatrix}.$$

And

$$R = A^{(2)} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \sqrt{2} & \frac{\sqrt{2}}{2} \\ 0 & 0 & -\frac{\sqrt{2}}{2} \end{pmatrix}.$$

3.3 Review of iterative methods

Introduction

This section gives only a brief overview of iterative methods for solving linear systems, for which we have limited ourselves to a few "exemplary" methods. These take the form

$$u_{k+1} = Bu_k + c, \quad k \geq 0, \, u_0 : \text{arbitrary vector},$$

the matrix B and the vector c being constructed from the given linear system $Au = b$ (the matrix B depends only on the matrix A).

We begin by establishing general *convergence* (sufficient conditions for $\lim_{k\to\infty} u_k = u$), and for comparing (evaluating the "rates of convergence") of iterative methods, then we describe in Section xxx some very commonly used methods: the Jacobi, Gauss-Seidel, and relaxation methods (of which the Gauss-Seidel method is a special case), pointwise or blockwise.

Finally, in Section xx, we give some precise results on *convergence* and *comparison of convergence rates*, applicable to the methods described above, notably in the case of linear systems whose matrix A is *symmetric positive definite and block tri-diagonal*. This special case is very important.

Generalities on Iterative Methods

Given an invertible matrix A and a vector b, we wish to compute the solution u of the linear system

$$Au = b.$$

Suppose we have found a matrix B and a vector c such that the matrix $I - B$ is invertible, and such that the unique solution of the linear system

$$u = Bu + c$$

is also the solution of $Au = b$.

The form of the system $u = Bu + c$ suggests defining an iterative method for solving the linear system $Au = b$. We take an arbitrary initial vector u_0, and define the sequence of vectors $(u_k)_{k=0}^{\infty}$ by

$$u_{k+1} = Bu_k + c, \quad k \geq 0.$$

We then say (a perfectly natural definition) that the iterative method is convergent if

$$\lim_{k\to\infty} u_k = u \text{ for every initial vector } u_0.$$

The following result gives the fundamental convergence criterion for iterative methods. Note that it involves only the matrix B, called the matrix of the considered iterative method.

Theorem 3.3 *The following propositions are equivalent:*

(1) The iterative method is convergent,

(2) $\varrho(B) < 1$,

(3) $\|B\| < 1$ *for at least one matrix norm* $\|\cdot\|$.

Proof. Saying that the method is convergent amounts to saying that

$$\lim_{k\to\infty} e_k = 0 \quad \text{for every vector } e_0 = u_0 - u,$$

where

$$e_k \stackrel{\text{def}}{=} u_k - u = B^k e_0, \quad k \geq 0,$$

is the k-th error vector.. ■

Remark 3.2 *The iterative method defined above is none other than the method of successive approximations, also called Picard's method, for finding a fixed point of the mapping*

$$f : v \in K^n \to f(v) = (Bv + c) \in K^n,$$

which is a contraction when property (3) is satisfied: the proof of Theorem 3.3 shows indeed that the matrix norm $\|\cdot\|$ *for which* $\|B\| < 1$ *can be chosen subordinate to a vector norm* $\|\cdot\|$*, so that*

$$\|f(v) - f(w)\| \leq \|B\|\|v - w\|.$$

How to choose between several convergent iterative methods for solving the same linear system $Au = b$*? To fix ideas, suppose the matrix* B *is normal. Then*

$$\|B^k e_0\|_2 \leq \|B^k\|_2 \|e_0\|_2 = (\varrho(B))^k \|e_0\|_2,$$

and this inequality is the best possible, in the sense that, for every integer $k \geq 0$*, there exists a vector* $e_0(k) \neq 0$ *for which it becomes an equality. In the case of normal matrices, the method is therefore faster when* $\varrho(B)$ *is smaller, since*

$$\sup_{\|e_0\|_2=1} \|B^k e_0\|_2^{1/k} = \varrho(B) \quad \textit{for all} \quad k \geq 0.$$

In the general case (arbitrary matrix B*, arbitrary vector norm), the conclusion is identical: asymptotically, the error vector* $e_k = B^k e_0$ *behaves "at worst" like* $(\varrho(B))^k$*, as specified by the following result.*

Theorem 3.4

1. *Let* $\|\cdot\|$ *be an arbitrary vector norm, and let* u *be such that*

   $$u = Bu + c.$$

   Consider the iterative method

   $$u_{k+1} = Bu_k + c, \quad k \geq 0.$$

   Then

   $$\lim_{k\to\infty} \left\{ \sup_{\|u_0 - u\|=1} \|u_k - u\|^{1/k} \right\} = \varrho(B).$$

2. *Let* $\|\cdot\|$ *be an arbitrary vector norm, and let* u *be such that*

   $$u = Bu + c = \bar{B}u + \bar{c}.$$

   Consider the iterative methods

   $$\bar{u}_{k+1} = \bar{B}\bar{u}_k + \bar{c}, \quad k \geq 0, \quad u_{k+1} = Bu_k + c, \quad k \geq 0,$$

   with $\varrho(B) < \varrho(\bar{B})$, $\quad u_0 = \bar{u}_0$*. Then, for any number* $\epsilon > 0$*, there exists an integer* $l(\epsilon)$ *such that*

   $$k \geq l \Rightarrow \sup_{\|u_0 - u\|=1} \left\{ \frac{\|\bar{u}_k - u\|}{\|u_k - u\|} \right\}^{1/k} \geq \frac{\varrho(\bar{B})}{\varrho(B) + \epsilon}.$$

Proof. Let $\|\cdot\|$ be the subordinate matrix norm. For every integer k, we can write

$$(\varrho(B))^k = \varrho(B^k) \leq \|B^k\| = \sup_{\|e_0\|=1} \|B^k e_0\|,$$

so that

$$\varrho(B) \leq \sup_{\|e_0\|=1} \|B^k e_0\|^{1/k} = \|B^k\|^{1/k},$$

and assertion (1) follows from Theorem 3.3. According to the same result, given $\epsilon > 0$, there exists an integer $l(\epsilon)$ such that

$$k \geq l \Rightarrow \sup_{\|e_0\|=1} \|B^k e_0\|^{1/k} \leq (\varrho(B) + \epsilon).$$

Furthermore, for every integer $k \geq l$, there exists a vector $e_0 = e_0(k)$ such that

$$\|e_0\| = 1 \quad \text{and} \quad \|\bar{B}^k e_0\|^{1/k} = \|\bar{B}^k\|^{1/k} \geq \varrho(\bar{B}),$$

and assertion (2) is thus proved. ■

Remarks 3.3 *The study of iterative methods therefore rests on the solution of the following two problems:*

1. *Given an iterative method with matrix B, determine if the method is convergent, i.e., if $\varrho(B) < 1$, or equivalently, if there exists a matrix norm such that $\|B\| < 1$ (Theorem 3.3).*
2. *Given two convergent iterative methods, compare them: the "faster" iterative method is the one whose matrix has the smaller spectral radius.*

Description of the Jacobi, Gauss-Seidel and Relaxation Methods

The methods we are going to describe are special cases of the following method: given a linear system $Au = b$, suppose we can write the invertible matrix A in the form of a decomposition $A = M - N$, where M is an invertible matrix and "easy to invert", i.e., in practice, diagonal or triangular, or else diagonal or triangular by blocks. Naturally, when we say that the matrix M is easy to invert, we do not mean that one actually computes the matrix M^{-1}. What we will have to compute is the solution of linear systems whose matrix is M.

We thus have the equivalences:

$$Au = b \iff Mu = Nu + b \iff u = M^{-1}Nu + M^{-1}b.$$

The last equation being of the form $u = Bu + c$, we associate with it the iterative method (following the considerations of the previous paragraph):

$$u_{k+1} = M^{-1}Nu_k + M^{-1}b, \quad k \geq 0, \quad u_0 : \text{arbitrary vector},$$

whose matrix is

$$B = M^{-1}N = I - M^{-1}A.$$

This shows incidentally that the matrix $(I - B) = M^{-1}A$ is invertible. In practice, one is thus led to solve the successive linear systems:

$$Mu_{k+1} = Nu_k + b, \quad k \geq 0.$$

In the case of the first two methods (Jacobi and Gauss-Seidel), the decomposition $A = M - N$ of the matrix $A = (a_{ij})$ is such that

$$(M)_{ij} = a_{ij} \text{ or } 0, \text{ depending on the values of the pair } (i, j),$$

which we will represent, with a (slight) abuse of notation, by the equality

$$A = \begin{pmatrix} & \\ -N & \\ M & \\ & \end{pmatrix}$$

to clearly show that the matrices M and N are disjoint. We thus see, in a very figurative way, that the corresponding iterative method consists roughly of inverting only the M part of A.

Intuitively, it seems that the more non-zero elements of the matrix A that M includes, the better the method should be (this will be useful for heuristically comparing the Jacobi and Gauss-Seidel methods introduced below), since if we consider the limit case where $M = A (so\ N = 0)$, we find the exact solution by solving the first equation of the iterative method. Other decompositions $A = M - N$ are not of the above type, in the sense that the matrices M and N overlap, as in the case of the relaxation method for example.

3.4 Iterative Methods for Solving Linear Systems

Iterative methods formally yield the solution x of a linear system after an infinite number of steps. At each step they require the computation of the residual of the system. In the case of a full matrix, their computational cost is therefore of the order of n^2 operations for each iteration, to be compared with an overall cost of the order of $\frac{2}{3}n^3$ operations needed by direct methods. Iterative methods can therefore become competitive with direct methods provided the number of iterations that are required to converge (within a prescribed tolerance) is either independent of n or scales sublinearly with respect to n.

Finally, we notice that, when A is ill-conditioned, a combined use of direct and iterative methods is made possible by preconditioning techniques that will be addressed in next section.

3.4.1 On the Convergence of Iterative Methods

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

$$x = \lim_{k \to \infty} x^{(k)}, \tag{3.1}$$

where x is the solution to (3.1). In practice, the iterative process is stopped at the minimum value of n such that $\|x^{(n)} - x\| < \varepsilon$, where ε is a fixed tolerance and $\|\cdot\|$ is any convenient vector 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 (see Section 4.6).

To start with, we consider iterative methods of the form

$$x^{(0)} \text{ given, } x^{(k+1)} = Bx^{(k)} + f, \quad k \geq 0, \tag{3.2}$$

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

Definition 3.3 *An iterative method of the form (4.2) is said to be consistent with (3.2) if f and B are such that $x = Bx + f$. Equivalently,*

$$f = (I - B)A^{-1}b.$$

Having denoted by

$$e^{(k)} = x^{(k)} - x \tag{4.3}$$

the error at the k-th step of the iteration, the condition for convergence (4.1) amounts to requiring that $\lim_{k \to \infty} e^{(k)} = 0$ for any choice of the initial datum $x^{(0)}$ (often called the initial guess).

Consistency alone does not suffice to ensure the convergence of the iterative method (4.2), as shown in the following example.

Example 3.2 *To solve the linear system $2Ix = b$, consider the iterative method*

$$x^{(k+1)} = -x^{(k)} + b,$$

which is obviously consistent. This scheme is not convergent for any choice of the initial guess. If, for instance, $x^{(0)} = 0$, the method generates the sequence $x^{(2k)} = 0, x^{(2k+1)} = b, k = 0, 1, \ldots$.

On the other hand, if $x^{(0)} = \frac{1}{2}b$ the method is convergent.

Jacobi, Gauss-Seidel and Relaxation Methods

In this section we consider some classical linear iterative methods.

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

$$x_i = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j \right], \quad i = 1, \ldots, n. \tag{3.1}$$

- In the Jacobi method, once an arbitrarily initial guess x^0 has been chosen, $x^{(k+1)}$ is computed by the formulae

$$x_i^{(k+1)} = \frac{1}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right], \quad i = 1, \ldots, n. \tag{3.2}$$

  This amounts to performing the following splitting for A

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

  where D is the diagonal matrix of the diagonal entries of A, E is the lower triangular matrix of entries $e_{ij} = -a_{ij}$ if $i > j$, $e_{ij} = 0$ if $i \leq j$, and F is the upper triangular matrix of entries $f_{ij} = -a_{ij}$ if $j > i$, $f_{ij} = 0$ if $j \leq i$. As a consequence, $A = D - (E + F)$.

  The iteration matrix of the Jacobi method is thus given by

$$B_J = D^{-1}(E + F) = I - D^{-1}A. \tag{3.3}$$

  A generalization of the Jacobi method is the over-relaxation method (or JOR), in which, having introduced a relaxation parameter ω, (3.2) is replaced by

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left[b_i - \sum_{\substack{j=1 \\ j \neq i}}^{n} a_{ij} x_j^{(k)} \right] + (1 - \omega) x_i^{(k)}, \quad i = 1, \ldots, n. \tag{3.4}$$

  The corresponding iteration matrix is

$$B_{J_\omega} = \omega B_J + (1 - \omega) I. \tag{3.5}$$

  The JOR method corresponds to

$$x^{(k+1)} = x^{(k)} + \omega D^{-1} r^{(k)}.$$

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

- The Gauss-Seidel method differs from the Jacobi method in the fact that at the $k+1$-th step the available values of $x_i^{(k+1)}$ are being used to update the solution, so that, instead of (3.2), one has

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

This method amounts to performing the following splitting for A

$$P = D - E, \quad N = F,$$

and the associated iteration matrix is

$$B_{GS} = (D - E)^{-1}F.$$

- Starting from Gauss-Seidel method, in analogy to what was done for Jacobi iterations, we introduce the successive over-relaxation method (or SOR method)

$$x_i^{(k+1)} = \frac{\omega}{a_{ii}} \left[b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{(k+1)} - \sum_{j=i+1}^{n} a_{ij} x_j^{(k)} \right] + (1 - \omega) x_i^{(k)}, \tag{3.6}$$

for $i = 1, \ldots, n$. The method (3.6) can be written in vector form as

$$(I - \omega D^{-1}E)x^{(k+1)} = [(1 - \omega)I + \omega D^{-1}F]x^{(k)} + \omega D^{-1}b \tag{3.7}$$

from which the iteration matrix is

$$B(\omega) = (I - \omega D^{-1}E)^{-1}[(1 - \omega)I + \omega D^{-1}F].$$

Multiplying by D both sides of (3.7) and recalling that $A = D - (E + F)$ yields the following form (4.7) of the SOR method

$$x^{(k+1)} = x^{(k)} + \left(\frac{1}{\omega} D - E \right)^{-1} r^{(k)}.$$

It is consistent for any $\omega \neq 0$ and for $\omega = 1$ it coincides with Gauss-Seidel method. In particular, if $\omega \in (0, 1)$ the method is called under-relaxation, while if $\omega > 1$ it is called over-relaxation.

Convergence Results for Jacobi and Gauss-Seidel Methods

There exist special classes of matrices for which it is possible to state a priori some convergence results for the methods examined in the previous section. The first result in this direction is the following.

Theorem 3.5 *If A is a strictly diagonally dominant matrix by rows, the Jacobi and Gauss-Seidel methods are convergent.*

Proof. Let us prove the part of the theorem concerning the Jacobi method, while for the Gauss-Seidel method we refer Quarteroni. Since A is strictly diagonally dominant by rows, $|a_{ii}| > \sum_{j=1}^{n} |a_{ij}|$ for $j \neq i$ and $i = 1, \ldots, n$. As a consequence,

$$\|B_J\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1, j\neq i}^{n} |a_{ij}|/|a_{ii}| < 1,$$

so that the Jacobi method is convergent. ■

Theorem 3.6

1. *If A and $2D - A$ are symmetric and positive definite matrices, then the Jacobi method is convergent and $\rho(B_J) = \|B_J\|_\Lambda = \|B_J\|_D$.*
2. *If A if symmetric positive definite, then the JOR method is convergent if $0 < \omega < 2/\rho(D^{-1}\Lambda)$.*
3. *If A is symmetric positive definite, the Gauss-Seidel method is monotonically convergent with respect to the norm $\|\cdot\|_\Lambda$.*

4. *if A is positive definite and tridiagonal, it can be shown that also the Jacobi method is convergent and $\rho(B_{GS}) = \rho^2(B_J)$. In this case, the Gauss-Seidel method is more rapidly convergent than the Jacobi method.*

Example 3.3 *Consider the 3×3 linear systems of the form $A_i x = b_i$, where b_i is always taken in such a way that the solution of the system is the unit vector, and the matrices A_i are*

$$A_1 = \begin{bmatrix} 3 & 0 & 4 \\ 7 & 4 & 2 \\ -1 & 1 & 2 \end{bmatrix}, \quad A_2 = \begin{bmatrix} -3 & 3 & -6 \\ -4 & 7 & -8 \\ 5 & 7 & -9 \end{bmatrix},$$

$$A_3 = \begin{bmatrix} 4 & 1 & 1 \\ 2 & -9 & 0 \\ 0 & -8 & -6 \end{bmatrix}, \quad A_4 = \begin{bmatrix} 7 & 6 & 9 \\ 4 & 5 & -4 \\ -7 & -3 & 8 \end{bmatrix}.$$

It can be checked that the Jacobi method does fail to converge for $A_1(\rho(B_J) = 1.33)$, while the Gauss-Seidel scheme is convergent. Conversely, in the case of A_2, the Jacobi method is convergent, while the Gauss-Seidel method fails to converge ($\rho(B_{GS}) = 1.1$). In the remaining two cases, the Jacobi method is more slowly convergent than the Gauss-Seidel method for matrix $A_3(\rho(B_J) = 0.44$ against $\rho(B_{GS}) = 0.018$, and the converse is true for $A_4(\rho(B_J) = 0.64$ while $\rho(B_{GS}) = 0.77$.

Convergence Results for the Relaxation Method

The following result provides a necessary condition on ω in order the SOR method to be convergent.

Theorem 3.7 *For any $\omega \in \mathbb{R}$ we have $\rho(B(\omega)) \geq |\omega - 1|$; therefore, the SOR method fails to converge if $\omega \leq 0$ or $\omega \geq 2$.*

Proof. If $\{\lambda_i\}$ denote the eigenvalues of the SOR iteration matrix, then

$$\left|\prod_{i=1}^{n} \lambda_i\right| = |\det[(1-\omega)I + \omega D^{-1}F]| = |1-\omega|^n.$$

Therefore, at least one eigenvalue λ_i must exist such that $|\lambda_i| \geq |1-\omega|$ and thus, in order for convergence to hold, we must have $|1-\omega| < 1$, that is $0 < \omega < 2$. ■

Assuming that A is symmetric and positive definite, the condition $0 < \omega < 2$, besides being necessary, becomes also sufficient for convergence.

Proposition 3.1 (Ostrowski) *If A is symmetric and positive definite, then the SOR method is convergent iff $0 < \omega < 2$. Moreover, its convergence is monotone with respect to $\|\cdot\|_A$.*

Remarks 3.4

1. *if A is strictly diagonally dominant by rows, the SOR method converges if $0 < \omega \leq 1$.*
2. *The results above show that the SOR method is more or less rapidly convergent, depending on the choice of the relaxation parameter ω. The question of how to determine the value ω_{opt} for which the convergence rate is the highest possible can be given a satisfactory answer only in special cases.*

Proposition 3.2 *If the matrix A enjoys the A-property and if B_J has real eigenvalues, then the SOR method converges for any choice of $x^{(0)}$ iff $\rho(B_J) < 1$ and $0 < \omega < 2$. Moreover,*

$$\omega_{opt} = \frac{2}{1 + \sqrt{1 - \rho(B_J)^2}} \tag{4.19}$$

and the corresponding asymptotic convergence factor is

$$\rho(B(\omega_{opt})) = \frac{1 - \sqrt{1 - \rho(B_J)^2}}{1 + \sqrt{1 - \rho(B_J)^2}}.$$

Mohamed Boudiaf University - M'sila
Faculty of Mathematics and Computer Science
Department of Mathematics

Academic Year: 2025-2026
Module: Matrix Numerical Analysis
3rd Year Applied Mathematics (S5)

TD Series: Linear Systems - Sensitivities

Exercise 01: Let the following matrices:

$$A = \begin{pmatrix} 5 & -4 & 2 \\ -1 & 2 & 3 \\ -2 & 1 & 0 \end{pmatrix}, \quad B = \begin{pmatrix} 3 & 6 & -1 \\ 3 & 1 & 0 \\ 2 & 4 & -7 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 7 & 3 \\ 4 & -2 & -2 \\ -2 & -1 & 1 \end{pmatrix}$$

1. Compute the condition number of matrix A and matrix B $(\text{cond}_1(\cdot), \text{cond}_2(\cdot))$ and $\text{cond}_\infty(\cdot)$.
2. What can be said about these matrices?

Exercise 02: We want to study a linear system $Ax = b$, where

$$A = \begin{pmatrix} 400 & -201 \\ -800 & 401 \end{pmatrix}, \quad b = \begin{pmatrix} 200 \\ -200 \end{pmatrix}$$

1. Compute the condition number of matrix A for the norms $\|\cdot\|_1, \|\cdot\|_\infty$. What do you think?
2. Compute the exact solution of the system $Ax = b$.
3. We perturb the matrix A, i.e.,
$$A = \begin{pmatrix} 401 & -201 \\ -800 & 401 \end{pmatrix}$$
   (a) Compute the solution of the system: $(A + \delta A)\bar{x} = b$.
4. Compute the residual vector $r = b - A\bar{x}$, then compute $\|r\|_\infty$
5. Compute the energy norm $\|\cdot\|_A$ for the vectors $\bar{x}$ and r.

Exercise 03
The goal of this exercise is to solve the linear system by the QR factorization method where matrices Q is orthogonal and R is upper triangular using Householder transformation (Method). Let the following linear system $AX = b$, such that

$$A = \begin{pmatrix} 4 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 2 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ -1 \\ 0 \end{pmatrix}$$

1. Determine matrices H_1 and H_2
2. Determine matrix R.
3. Determine matrix Q.
4. Determine the solution of the system.

Exercise 04
We want to solve the following linear system $AX = b$, such that

$$A = \begin{pmatrix} 1001 & 1000 \\ 1000 & 1001 \end{pmatrix}, \quad b = \begin{pmatrix} 2001 \\ 2001 \end{pmatrix}$$

1. Compute the condition number of matrix A, $k(A)$.
2. Find the solution of this system.
3. We perturb the right-hand side of the system $\delta b = (0, 1)$. Find the solution of this perturbed system.

Exercise 05
Let the following linear system $AX = b$, such that A =

1. Determine matrices H_1 and H_2
2. Determine matrix R.
3. Determine matrix Q.
4. Determine the solution of the system.

Exercise 06
In order to compare the Jacobi, Gauss-Seidel and SOR methods for solving the linear system $Ax = b$. Given the following linear system:

1. Compute the first 7 iterations for the Jacobi method taking an initial value $x^{(0)} = (1, 1, 1)^t$
2. Same question for the Gauss-Seidel method
3. Same question for the SOR method with $\omega = 1.25$
4. What can be concluded?