MDI 210

Numerical analysis and continuous optimization

Irène Charon Olivier Hudry

August 17, 2018

Contents

Chapter 1

Matrix Analysis - Generalities

1.1 Reminders of linear algebra

1.1.1 Adjoints

In the following, we consider $\mathbb{R}$ or $\mathbb{C}$ as a basic field. Let us first recall some definitions.

Given a vector x (usually represented in this manual by a column matrix), we call *adjoint* of x and we denote by x^* the transposed vector of the conjugate of x: if $x = \begin{pmatrix} x_1 \\ ... \\ x_i \\ ... \\ x_n \end{pmatrix}$, then $x^* = (\overline{x_1}, ..., \overline{x_i}, ..., \overline{x_n})$.

Remark. In $\mathbb{R}$, we have: $x^* = (x_1, ..., x_i, ..., x_n)$.

Le *Hermitian product* of two vectors x and y of dimension n is defined by: $(x, y) = \sum_{i=1}^{n} \overline{x_i} y_i$. If the vectors are represented by column vectors, we have: $(x, y) = x^* y$, where the product is the matrix product. If the vectors have real components, the Hermitian product becomes the *Euclidean scalar product*: $(x, y) = \sum_{i=1}^{n} x_i y_i = x^{\mathrm{t}} y$.

To a matrix A, we can associate its *adjoint matrix*, denoted by A^*, defined as follows: if $A = (a_{i,j})_{\substack{1 \leqslant i \leqslant n \\ 1 \leqslant j \leqslant p}}$, then $A^* = \overline{A^{\mathrm{t}}} = (\overline{a_{j,i}})_{\substack{1 \leqslant j \leqslant p \\ 1 \leqslant i \leqslant n}}$. We have: $(A^*)^* = A$.

If x and y are two column vectors having respectively n rows and p rows

and A a matrix with n rows and p columns, we can check the property: $(x, Ay) = (A^*x, y)$.

1.1.2 Kinds of matrices

A real square matrix A is said:

- *symmetric* if $A^{\mathrm{t}} = A$,
- *normal* if $AA^{\mathrm{t}} = A^{\mathrm{t}}A$,
- *orthogonal* if $AA^{\mathrm{t}} = A^{\mathrm{t}}A = I$, where I is the identity matrix.

In the following, the qualifiers "symmetric" and "orhogonal" only apply to real matrices.

A complex square matrix A is said:

- *Hermitian* if $A^* = A$,
- *normal* if $AA^* = A^*A$.
- *unitary* if $A^*A = AA^* = I$.

Note that a symmetric or Hermitian matrix is normal. The same applies to an orthogonal or unitary matrix. It is recalled that the eigenvalues of a symmetric real matrix or of a Hermitian matrix are real.

1.1.3 Spectrum of a matrix

An *eigenvalue* of a square matrix A is a scalar λ such that there exists a nonzero vector x satisfying: $Ax = \lambda x$. The vector x is then said *eigenvector* of A.

Let A be a square matrix. The *spectrum* of A is the set of eigenvalues of A. The *spectral radius* of A is the largest absolute value of the eigenvalues of A; it is denoted by $\rho(A)$.

1.1.4 Reduction of a matrix

Two square matrices are said to be *similar* if they are likely to represent the same linear application on two different bases. If A and B are two similar matrices, there is an invertible matrix P satisfying $A = P^{-1}BP$; the matrix P is called *change-of-basis matrix* indexmatrix!change-of-basis. A matrix is *diagonalizable* if it is similar to a diagonal matrix; this diagonal matrix consists of the eigenvalues of A counted with their order of multiplicity.

Any real symmetric matrix is similar to a real diagonal matrix.

A square matrix is invertible if and only if it does not have an eigenvalue equal to zero.

In fact we can also demonstrate, for the square matrices, the following results:

Theorem:

1. Let A be any square matrix; there exists a unit matrix U such that $U^{-1}AU$ is triangular.
2. Let A be a normal matrix; there is exists unitary matrix U such that $U^{-1}AU$ is diagonal.
3. Let A be a symmetric matrix; there exists an orthogonal matrix O such that $O^{-1}AO$ is diagonal.

Corollaries of the definitions and this theorem:

1. The absolute values of the eigenvalues of an orthogonal or unit matrix are equal to 1.
2. A Hermitian (resp. symmetric) or unitary matrix is diagonalizable by a unitary (orthogonal) change-of-basis matrix.
3. An orthogonal matrix O is diagonalizable by a matrix U, usually not real, unitary ($O = U^*DU$), the diagonal elements of D being of absolute value 1.

1.1.5 Singular values

The matrix A^*A is normal, so it is diagonalizable. It can easily be shown that its eigenvalues are positive or zero. The positive square roots of the eigenvalues of A^*A are called *singular values* of A.

The matrix A is invertible if and only if its singular values are all strictly positive.

Two matrices A and B are called *equivalent* if there are two invertible matrices U and V such that $B = U^{-1}AV$.

Let A be a square matrix; A is equivalent to a diagonal matrix whose diagonal consists of singular values of A. More precisely:

- if A is real, there are two orthogonal square matrices U and V and a diagonal matrix D consisting of singular values of A such that: $A = U^{\mathrm{t}}DV$;
- if A is complex, there are two unitary square matrices U and V and a diagonal matrix D consisting of singular values of A such that: $A = U^{*}DV$.

1.2 Norms

In the following, we will need the concepts of a vector norm and of a matrix norm.

Let $x = (x_i)_{1 \leqslant i \leqslant n}$ a vector. The three most common vector norms are:

- $||x||_1 = \sum_{i=1}^{n} |x_i|$ (norm 1)
- $||x||_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}}$ (norm 2, or Euclidean norm)
- $||x||_\infty = \max_{1 \leqslant i \leqslant n} |x_i|$ (infinity norm)

More generally: $||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ (p-norm). The proof that it is a norm uses the following inequalities:

- Hölder inequality: if p and q are two numbers checking $p > 1$ and equality $\frac{1}{p} + \frac{1}{q} = 1$ (which results in $q > 1$), then

$$\sum_{i=1}^{n} |x_i y_i| \leqslant \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}} \left(\sum_{i=1}^{n} |y_i|^q\right)^{\frac{1}{q}}.$$

For $p = q = 2$, it gives back the Cauchy-Schwarz inequality.

- Minkowski inequality:

$$\left(\sum_{i=1}^{n}|x_i+y_i|^p\right)^{\frac{1}{p}} \leqslant \left(\sum_{i=1}^{n}|x_i|^p\right)^{\frac{1}{p}} + \left(\sum_{i=1}^{n}|y_i|^p\right)^{\frac{1}{p}}.$$

In $\mathbb{R}^n$ and $\mathbb{C}^n$, all norms are *equivalent* (two norms $||\ ||$ and $||\ ||'$ are équivalent on a vector space E if there are two strictly positive constants C and C' such that, for any x in E: $C||x|| \leqslant ||x||' \leqslant C'||x||$).

We call $\mathcal{A}_n$ the ring of square matrices of order n with coefficients in $\mathbb{R}$; we use the same notation $\mathcal{A}_n$ for the ring of square matrices of order n with coefficients in $\mathbb{C}$. We call *matrix norm* an application from $\mathcal{A}_n$ to $\mathbb{R}^+$ denoted by $||\ ||$ which fulfils the following properties:

- for any matrices A of $\mathcal{A}_n$, $||A|| = 0 \Leftrightarrow A = 0$
- for any α of $\mathbb{R}$ (or $\mathbb{C}$) and for any A of $\mathcal{A}_n$, $||\alpha A|| = |\alpha|||A||$
- for any matrices A and B of $\mathcal{A}_n$, $||A+B|| \leqslant ||A|| + ||B||$
- for any matrices A and B of $\mathcal{A}_n$, $||A \times B|| \leqslant ||A|| \times ||B||$.

We can very easily build matrix norms from vectorial norms: they are then called *subordinate matrix norms*. For this, we can define $||A||$ by the following equivalent formulas:

$$||A|| = \sup_{x\neq 0}\frac{||Ax||}{||x||} = \sup_{||x||=1}||Ax|| = \sup_{0<||x||\leqslant 1}\frac{||Ax||}{||x||}.$$

We have: $||Ax|| \leqslant ||A||\ ||x||$.

The matrix norms subordinate to the most usual norms that we have described above are therefore, for $A = (a_{i,j})_{\substack{1\leqslant i\leqslant n\\ 1\leqslant j\leqslant n}}$:

- $||A||_1 = \sup_{x\neq 0}\dfrac{||Ax||_1}{||x||_1} = \max_{1\leqslant j\leqslant n}\displaystyle\sum_{i=1}^{n}|a_{ij}|$
- $||A||_2 = \sup_{x\neq 0}\dfrac{||Ax||_2}{||x||_2} = \sqrt{\rho(A^*A)} = ||A^*||_2$ where $\rho(A^*A)$ represents the largest absolute value of A^*A (spectral radius of A^*A)

- $||A||_\infty = \sup_{x\neq 0} \dfrac{||Ax||_\infty}{||x||_\infty} = \max_{1\leqslant i\leqslant n} \sum_{j=1}^{n} |a_{ij}|$.

The norm $||\ ||_2$ is invariant by unitary transformation: if U is a unitary matrix, that is, if U fulfils the relation $U^*U = I$, then we have

$$||A||_2 = ||AU||_2 = ||UA||_2 = ||U^*AU||_2.$$

If A is normal, that is, if A fulfils the relation $A^*A = AA^*$ (especially if A is Hermitian or symmetric), then $||A||_2 = \rho(A)$.

If A is unitary or orthogonal, $||A||_2 = 1$.

Remark. $||A||_1$ and $||A||_\infty$ are easy to compute but not $||A||_2$.

Theorem

- Let $||\ ||$ be a subordinate norm; let B satisfying $||B|| < 1$. So $I + B$ is invertible and $||(I + B)^{-1}|| \leqslant \dfrac{1}{1 - ||B||}$.
- If a matrix of the form $I + B$ is not invertible, then, for any norm, subordinate or not, $||B|| \geqslant 1$.

Example of a non-subordinate norm: the Euclidean norm

This norm is définied by: $||A||_E = \left(\sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}|^2\right)^{\frac{1}{2}} = \sqrt{\text{trace } (A^*A)}$ (remember that the trace of a matrix is the sum of its diagonal terms). The norm $||A||_E$ is invariant by unitary transformation; in other words, if $U^*U = I$, the $||A||_E = ||AU||_E = ||UA||_E = ||U^*AU||_E$. Moreover: $||A||_2 \leqslant ||A||_E \leqslant \sqrt{n}||A||_2$.

Theorem. Let $||\ ||$ be any norm (subordinate or not); we have: $\rho(A) \leqslant ||A||$ and, for every $\epsilon > 0$, there exists a subordinate norm $||\ ||_{A,\epsilon}$ satisfying $||A||_{A,\epsilon} \leqslant \rho(A) + \epsilon$.

1.2.1 Convergence of a sequence of vectors or a sequence of matrices

For a sequence (x^k) of vectors to converge, it is necessary and sufficient that the components of (x^k) converge. It is the same for a sequence of matrices.

In particular, we have the following theorem for the sequence of the powers of a matrix:

Theorem. Let B be a square matrix.

1. $\lim_{k\to\infty} B^k = 0 \Leftrightarrow \forall x, \lim_{k\to\infty} B^k x = 0 \Leftrightarrow \rho(B) < 1 \Leftrightarrow$ for at least one subordinate norm, $||B|| < 1$.

2. Let $||\ ||$ be any norm; then $\lim_{k\to\infty} ||B^k||^{\frac{1}{k}} = \rho(B)$.

Chapter 2

Problems of numerical analysis

The two main problems that we will study in the following of this course are the resolution of linear systems and computation of eigenvalues and eigenvectors of matrices. When applying numerical analysis methods to computational problems, there are two types of "quality" to consider. This is firstly the aspect called *complexity*, that is to say the number of elementary operations to perform to obtain a result, but also it is necessary to know if the solution is acceptable or not; in fact, two kinds of errors can be made: on the one hand, rounding errors due to the precision of the computations and, on the other hand, so-called truncation errors, when using iterative methods, while of course we stop after a finite number of iterations.

2.1 Errors

Rounding error: error due to coding where the number of digits representing a real is limited. If the number is coded on t bits for the significand, the error on the significand is upper bounded by 2^{-t}.

Truncation error: in iterative methods, computing the limit would require *a priori* an infinite number of iterations. Since computations are inevitably stopped after a k_0 number of iterations, we make a truncation error measured by $||x^\infty - x^{k_0}||$, where x^∞ represents the limit, x^{k_0} the result obtained at the k_0^{th} iteration and $||.||$ a given norm (in fact, x^∞ is unknown, which makes impossible to estimate the error).

2.2 Conditioning

In all that follows, we consider a linear system written in matrix form $Ax = b$. Before going into the details of the methods, which will be the subject of the next chapter, we will deal with an important parameter of a linear system: it is its *conditioning*, which is attached to the A matrix of the system. Most often, in practice, the coefficients of A, like the components of the vector b, are measurement results and are therefore tainted by a certain error. It is essential to see how a small modification of A or b affects, regardless of the method used, the supposedly exact solution of the system.

2.2.1 Conditioning of a linear system

Consider the following system:

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \text{ of solution } \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix}.$$

Consider now the perturbed system by slightly modifying the vector of the second member, the matrix remaining unchanged:

$$\begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix} \begin{pmatrix} x_1 + \delta x_1 \\ x_2 + \delta x_2 \\ x_3 + \delta x_3 \\ x_4 + \delta x_4 \end{pmatrix} = \begin{pmatrix} 32,1 \\ 22,9 \\ 33,1 \\ 30,9 \end{pmatrix} \quad \text{of solution} \quad \begin{pmatrix} 9,2 \\ -12,6 \\ 4,5 \\ -1,1 \end{pmatrix}.$$

It can be seen that a relative error of the order of $1/300$ on the second member results in a relative error of the order of 10 over several coordinates of the system solution, and therefore an amplification of the relative errors of the order of 3000.

Now consider slight modifications on the matrix with the system:

$$\begin{pmatrix} 10 & 7 & 8,1 & 7,2 \\ 7,08 & 5,04 & 6 & 5 \\ 8 & 5,98 & 9,89 & 9 \\ 6,99 & 4,99 & 9 & 9,98 \end{pmatrix} \begin{pmatrix} x_1 + \delta x_1 \\ x_2 + \delta x_2 \\ x_3 + \delta x_3 \\ x_4 + \delta x_4 \end{pmatrix} = \begin{pmatrix} 32 \\ 23 \\ 33 \\ 31 \end{pmatrix} \text{ of solution } \begin{pmatrix} -81 \\ 137 \\ -34 \\ 22 \end{pmatrix}.$$

We also note here that small variations of the elements of the matrix considerably modify the solution of the linear system.

Suppose that we consider the system: $A(x + \delta x) = b + \delta b$, all things being equal, and suppose the matrix A invertible. We see that we have $\delta x = A^{-1}\delta b$. If we then choose a matrix norm $||\ ||$ subordinate to a vector norm, we find $||\delta x|| \leqslant ||A^{-1}||\,||\delta b||$ and, in addition, $||b|| \leqslant ||A||\,||x||$ so that x has a relative error $\dfrac{||\delta x||}{||x||}$ which is upper bounded by $||A||\,||A^{-1}||\,\dfrac{||\delta b||}{||b||}$. We call *conditioning* of the matrix A (relative to the norm $||\ ||$) the quantity $||A||\,||A^{-1}||$, which we denote by $\text{cond}_{||\ ||}(A)$ or, more simply, $\text{cond}(A)$.

We could also prove that if we now add a small variation to the coefficients of A, so that this matrix becomes $A + \delta A$, then $\dfrac{||\delta x||}{||x + \delta x||}$ is upper bounded by $||A||\,||A^{-1}||\,\dfrac{||\delta A||}{||A||}$.

These two upper bounds prove the interest of conditioning. The more the conditioning of a matrix is close to 1, the more it is well conditioned (its conditioning is always greater than or equal to 1).

We have mentioned here only the conditioning of a matrix with respect to the resolution of a linear system. We will see later what conditioning is for a problem of eigenvalues. The same matrix can be badly conditioned as the matrix of a linear system and be well conditioned for the problem of finding eigenvalues, and *vice versa*.

The following theorem gives further information on the conditioning of a matrix in the sense of the systems.

Theorem: Let A be an invertible matrix. We then have:

1. $\text{cond}(A) \geqslant 1$
2. $\text{cond}(A) = \text{cond}(A^{-1})$
3. for any $\alpha \neq 0$, $\text{cond}(\alpha A) = \text{cond}(A)$
4. denoting by cond_2 the conditioning associated with $||\ ||_2$ and denoting respectively by $\mu_1(A)$ and $\mu_n(A)$ the smallest and the largest singular

values of A, $\mathrm{cond}_2(A) = \dfrac{\mu_n(A)}{\mu_1(A)}$

5. if A is normal (that is to say, fulfiled $AA^* = A^*A$), $\mathrm{cond}_2(A) = \dfrac{\max_i |\lambda_i(A)|}{\min_i |\lambda_i(A)|}$ where the $\lambda_i(A)$ represent the eigenvalues of A
6. if A is unitary or orthogonal, $\mathrm{cond}_2(A) = 1$
7. $\mathrm{cond}_2(A)$ is invariant by unitary or orthogonal transformation:
   if $UU^* = I$, then $\mathrm{cond}_2(A) = \mathrm{cond}_2(AU) = \mathrm{cond}_2(UA) = \mathrm{cond}_2(U^*AU)$,
   if $OO^{\mathrm{t}} = I$, then $\mathrm{cond}_2(A) = \mathrm{cond}_2(AO) = \mathrm{cond}_2(OA) = \mathrm{cond}_2(O^{\mathrm{t}}AO)$.

Let us compute, for example, the conditioning of the matrix used previously:

$$A = \begin{pmatrix} 10 & 7 & 8 & 7 \\ 7 & 5 & 6 & 5 \\ 8 & 6 & 10 & 9 \\ 7 & 5 & 9 & 10 \end{pmatrix}.$$

This matrix has for approximate eigenvalues:

$$\lambda_1 \approx 0,01015 < \lambda_2 \approx 0,8431 < \lambda_3 \approx 3,858 < \lambda_4 \approx 30,2887.$$

So we have: $\mathrm{cond}_2(A) = \dfrac{\lambda_4}{\lambda_1} \approx 2984$. The matrix A thus has a very bad conditioning, which explains the sensitivity to the errors of the linear systems defined with the matrix A.

As for all $\alpha \neq 0$, $\mathrm{cond}(\alpha A) = \mathrm{cond}(A)$, we cannot hope to reduce the conditioning of A by multiplying all its elements by the same number. On the other hand, it can be done by multiplying for example each line (and/or each column) by an appropriate coefficient; this is the problem of the *balancing of a matrix*, which can be stated as follows: given a matrix A, determine two invertible diagonal matrices D_1 and D_2 satisfying: $\mathrm{cond}(D_1AD_2) = \inf_{\Delta_1,\Delta_2 \text{ invertible diagonal}} \mathrm{cond}(\Delta_1 A \Delta_2)$.

We then solve $Ax = b$ in two steps:

- solving $D_1AD_2y = D_1b$
- solving $x = D_2y$.

In practice, conditioning is not a simple function of the elements of D_1 and D_2; we try instead to minimize the ratio between the largest and the smallest nonzero element of $A' = \Delta_1 A \Delta_2$. Set $E = \{(i,j) \text{ with } 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant n \text{ and } a_{ij} \neq 0\}$. We look for two matrices Δ_1 and Δ_2 diagonal and invertible which minimize the ratio:

$$\frac{\max_{(i,j)\in E}|a'_{ij}|}{\min_{(i,j)\in E}|a'_{ij}|}.$$

We consider now the case where A is a real matrix. Denoting by x_i the i^{th} element of the diagonal of Δ_1 and by y_i the i^{th} element of the diagonal of Δ_2, we have: $a'_{ij} = x_i a_{ij} y_j$. We go to logarithms by posing $\alpha_{ij} = \ln|a_{ij}|$, $u_i = \ln|x_i|$, $v_j = \ln|y_j|$. The problem becomes:

$$\text{minimize}_{\ u_i,v_j \text{ with } (i,j)\in E}\ [\max_{(i,j)\in E}(\alpha_{ij} + u_i + v_j) - \min_{(i,j)\in E}(\alpha_{ij} + u_i + v_j)],$$

what is rewritten as the following linear program (because, by a translation of the values, we can restrict ourselves to the solutions where the minimum on the u_i and v_j of $\alpha_{ij} + u_i + v_j$ is 0):

$$\begin{cases} \text{minimize } z \\ \text{with, for any } (i,j) \in E, \quad 0 \leqslant \alpha_{ij} + u_i + v_j \leqslant z \\ u_i \text{ and } v_j \text{ real.} \end{cases}$$

2.2.2 Conditioning of a problem of search of eigenvalues

In a search problem of eigenvalues, it is again important to know the influence of a small modification of the coefficients of the matrix A on the computed eigenvalues. This conditioning involves conditioning of the change-of-basis matrix from A to a diagonal shape, and not A directly. The following theorem makes it possible to define this new conditioning which will be denote $\Gamma(A)$.

Theorem: let A be a diagonalizable matrix and P a matrix such that $P^{-1}AP$ is diagonal with diagonal terms λ_i. Let $||\ ||$ be a matrix norm such that, for any diagonal matrix $\text{diag}(\delta_i)$:

$$||\text{diag}(\delta_i)|| = \max_i|\delta_i|.$$

So, for any matrix δA:

$$\text{spectrum}(A+\delta A) \subset \bigcup_{i=1}^{n} D_i,$$

with $D_i = \{z \in \mathbb{C} \text{ such that } |z-\lambda_i| \leqslant \text{cond}_{||\ ||}(P)||\delta A||\}$.

This means that, if A is diagonalizable, the δA perturbation globally leaves the eigenvalues in complex disks, centered on the old eigenvalues and radius $\text{cond}_{||\ ||}(P)||\delta A||$.

For A diagonalizable, the $\Gamma(A)$ conditioning relative to the search for eigenvalues is defined as the conditional minimum $\text{cond}_{||\ ||}(P)$ taken on matrices P such that $P^{-1}AP$ is diagonal. The theorem above indicates that, for A diagonalizable, we have the inclusion:

$$\text{spectrum}(A+\delta A) \subset \bigcup_{i=1}^{n} \{z \in \mathbb{C} \text{ such that } |z-\lambda_i| \leqslant \Gamma(A)||\delta A||\}.$$

Since a normal matrix is diagonalizable with a unitary change-of-basis matrix P, it has a $\Gamma(A)$ conditioning equal to 1 for $||\ ||_2$. This is therefore in particular the case for symmetric matrices. In the latter case, we have the theorem:

Theorem: Let A be a symmetric matrix and $B = A + \delta A$, where the δA perturbation is also symmetric. Let $\alpha_1 \leqslant \alpha_2 \leqslant ... \leqslant \alpha_n$ be the eigenvalues of A and $\beta_1 \leqslant \beta_2 \leqslant ... \leqslant \beta_n$ be the eigenvalues of B. Then, we have for $1 \leqslant i \leqslant n$: $|\alpha_i - \beta_i| \leqslant ||\delta A||_2$.

This theorem expresses that, if A and δA are both symmetric, each eigenvalue of $A+\delta A$ remains in a real interval centered on the old eigenvalue and of radius $||\delta A||_2$.

Chapter 3

Resolution of linear systems

3.1 Generalities

The problem we are interested in can be formulated as follows.

Problem: let $A = (a_{i,j})$ an invertible square matrix of dimension n, $x = (x_i)$ and $b = (b_i)$ two column vectors of dimension n; solve in x the system $Ax = b$.

Remarks.

1. Numerical methods of resolution do not generally use the computation of A^{-1}.

2. If A is in upper triangular form (the elements below the main diagonal are all null) with non-zero diagonal terms, then the resolution is easy. We start by solving the last relation, which is a linear equation in the only variable x_n, we transfer this value in the previous relation which becomes an equation in x_{n-1}, and we continue step by step up to x_1. This method, called *backward substitution* and summarized below, requires $n(n-1)/2$ additions, $n(n-1)/2$ multiplications and n divisions.

$$\begin{cases} a_{1,1}x_1 + \ \ldots\ + a_{1,n-1}x_{n-1} + a_{1,n}x_n = b_1 \\ \qquad\qquad \ldots \\ \qquad\qquad a_{n-1,n-1}x_{n-1} + a_{n-1,n}x_n = b_{n-1} \\ \qquad\qquad\qquad\qquad a_{n,n}x_n = b_n \end{cases}$$

$$\Rightarrow \begin{cases} x_n = \dfrac{b_n}{a_{n,n}} \\ x_{n-1} = \dfrac{b_{n-1} - a_{n-1,n}x_n}{a_{n-1,n-1}} \\ ... \\ x_1 = \dfrac{b_1 - a_{1,2}x_2 - ... - a_{1,n}x_n}{a_{1,1}} \end{cases}$$

3.2 Gauss method

Gauss method is used when the matrix A is invertible. The principle is as follows:

- Using linear combinations between the lines of A, we eliminate successively some variables of relations, to obtain a form $(MA)x = Mb$ where MA is an upper triangular matrix. Note that in fact we do not compute M, but we build directly MA and Mb.
- We solve $(MA)x = Mb$ by the backward substitutions method.

3.2.1 Elimination step

- We choose in the first column a coefficient $a_{i,1}$ different from 0; there is always one since the matrix is invertible. This element is the *pivot*.
- If the pivot is not in the first line, the line of the pivot is exchanged with the first line.
- By well-chosen linear combinations, obtained by subtracting the first line multiplied by the correct coefficient at each line, we eliminate all the terms of the column of the pivot located under the diagonal.
- We obtain then a matrix A' whose first column has only 0 under the first term which, it, is non-zero.
- We consider the matrix obtained by deleting the first row and the first column of A'. The process is repeated on this new matrix.
- We stop this process when the matrix obtained is of dimension 1.

- Relocating deleted rows and columns, we get a triangular matrix.

Remark. The determinant of A is obtained by the product of the pivots multiplied by $(-1)^p$, where p represents the number of times that the pivot was not on the diagonal.

Example 1

We consider the system:

$$\begin{cases} \mathbf{2}x_1 + x_2 - 3x_3 = 5 \\ 4x_1 + x_2 + 5x_3 = -1 \\ 10x_1 - 7x_2 + 13x_3 = -3 \end{cases}$$

After the first iteration, having chosen as pivot the value 2, in bold above, we obtain:

$$\begin{cases} 2x_1 + x_2 - 3x_3 = 5 \\ -\mathbf{1}x_2 + 11x_3 = -11 \\ -12x_2 + 28x_3 = -28 \end{cases}$$

After the second iteration (the pivot is the coefficient of the second row, second column and is equal to -1), we obtain:

$$\begin{cases} 2x_1 + x_2 - 3x_3 = 5 \\ -x_2 + 11x_3 = -11 \\ -104x_3 = 104 \end{cases}$$

We then apply a backward substitutions method and we successively obtain:

$x_3 = -1$, $x_2 = \dfrac{-11 - 11x_3}{-1} = 0$, $x_1 = \dfrac{5 - x_2 + 3x_3}{2} = 1$.

Remark. Since the rows have not been exchanged, the determinant of A is equal to the determinant of the matrix corresponding to the last system. So we have: $\det(A) = 2 \times (-1) \times (-104) = 208$.

Example 2

We consider the system:

$$\begin{cases} \mathbf{2}x_1 + x_2 - 3x_3 = -3 \\ 4x_1 + 2x_2 - x_3 = 4 \\ 6x_1 + 5x_2 + 8x_3 = 27 \end{cases}$$

After the first iteration, having chosen as pivot the value 2, in bold above, we obtain:

$$\begin{cases} 2x_1 + x_2 - 3x_3 = -3 \\ 0x_2 + 5x_3 = 10 \\ \mathbf{2}x_2 + 17x_3 = 36 \end{cases}$$

The pivot is now necessarily the coefficient of x_2 in the last line (of value 2); the second and third lines are exchanged; we obtain:

$$\begin{cases} 2x_1 + x_2 - 3x_3 = -3 \\ \mathbf{2}x_2 + 17x_3 = 36 \\ 0x_2 + 5x_3 = 10 \end{cases}$$

The coefficient of x_2 in the last line being zero, it remains only to perform the backward substitutions method:

$$x_3 = 10/5 = 2,\ x_2 = \frac{36 - 17x_3}{2} = 1,\ x_1 = \frac{-3 - x_2 + 3x_3}{2} = 1.$$

Remark. As the lines have been exchanged once, the determinant of A equals the determinant of the matrix corresponding to the last system multiplied by -1. So we have: $\det(A) = (-1) \times 2 \times 2 \times 5 = -20$.

3.2.2 Choice of the pivot

Because of rounding errors, the choice of the pivot is important; indeed, a pivot too small module can lead to bad solutions because of the division by the pivot. Two strategies are in fact possible.

- *Partiel pivot p*: we choose in the current column the term of largest module located under the diagonal or on this one.
- *Total pivot*: we choose the term of the largest module of the residual matrix, that is to say, if we are at the step $n - k + 1$, the matrix consisting of the k last lines and k last columns. This method is more expensive in time.

3.2.3 Complexity

The number of operations needed for the Gaussian method can be estimated; in the case where we do not choose the pivot, we do in all about $\frac{n^3}{3}$ additions,

as many multiplications, $\frac{n^2}{2}$ divisions and thus in total a number of arithmetic operations equivalent to $\frac{2n^3}{3}$.

3.2.4 Variant: Gauss-Jordan method

Compared to the Gaussian method, the only difference made by the Gauss-Jordan method is that in the elimination phase, the terms above the diagonal are also eliminated. This produces a diagonal matrix. This method is used in particular for computing the inverse of a matrix. We then solve simultaneously the n linear systems $Ax_j = e_j$, x_j being the column vector of the inverse matrix, the e_j constituting the vectors of the canonical basis of $\mathbb{R}^n$.

Example: the computation of the inverse of $A = \begin{pmatrix} 1 & -3 & 14 \\ 1 & -2 & 10 \\ -2 & 4 & -19 \end{pmatrix}$.

We solve the three systems:

$$\left\{\begin{array}{rcrcrcl|c|c} x_1 & - & 3x_2 & + & 14x_3 & = & 1 & 0 & 0 \\ x_1 & - & 2x_2 & + & 10x_3 & = & 0 & 1 & 0 \\ -2x_1 & + & 4x_2 & - & 19x_3 & = & 0 & 0 & 1 \end{array}\right.$$

First iteration (here, with partial pivot): we exchange the first and the third lines, which gives, with the pivot at the top left (in bold):

$$\left\{\begin{array}{rcrcrcl|c|c} \mathbf{-2}x_1 & + & 4x_2 & - & 19x_3 & = & 0 & 0 & 1 \\ x_1 & - & 2x_2 & + & 10x_3 & = & 0 & 1 & 0 \\ x_1 & - & 3x_2 & + & 14x_3 & = & 1 & 0 & 0 \end{array}\right.$$

The terms of the first column except the diagonal term are eliminated. We obtain:

$$\left\{\begin{array}{rcrcrcl|c|c} \mathbf{-2}x_1 & + & 4x_2 & - & 19\,x_3 & = & 0 & 0 & 1 \\ & & 0x_2 & + & 1/2\,x_3 & = & 0 & 1 & 1/2 \\ & - & x_2 & + & 9/2\,x_3 & = & 1 & 0 & 1/2 \end{array}\right.$$

Second iteration (now, with total pivot to illustrate this variant): the largest coefficient in absolute value being $9/2$, bottom right, we exchange the second and the third lines as well as the second and the third columns. We obtain:

$$\left\{\begin{array}{rcrcrcl|c|c} -2x_1 & - & 19x_3 & + & 4x_2 & = & 0 & 0 & 1 \\ & & \mathbf{9/2}\,x_3 & - & x_2 & = & 1 & 0 & 1/2 \\ & & 1/2\,x_3 & + & 0x_2 & = & 0 & 1 & 1/2 \end{array}\right.$$

The terms of the second column except the diagonal term are eliminated. We obtain:

$$\left\{\begin{array}{rrrrrcr|c|r} -2x_1 & & & - & 2/9\,x_2 & = & 38/9 & 0 & 28/9 \\ & & 9/2\,x_3 & - & x_2 & = & 1 & 0 & 1/2 \\ & & & & \mathbf{1/9}\,x_2 & = & -1/9 & 1 & 4/9 \end{array}\right.$$

Third and last iteration; the pivot can only be the element of the line not yet treated: it is the $1/9$ bottom right. We obtain:

$$\left\{\begin{array}{rrrrrcr|c|r} -2x_1 & & & & & = & 4 & 2 & 4 \\ & & 9/2 & x_3 & & = & 0 & 9 & 9/2 \\ & & & & 1/9\,x_2 & = & -1/9 & 1 & 4/9 \end{array}\right.$$

We can now solve the three systems immediately:

$$\left\{\begin{array}{rcr|r|r} x_1 & = & -2 & -1 & -2 \\ x_2 & = & -1 & 9 & 4 \\ x_3 & = & 0 & 2 & 1 \end{array}\right.$$

We deduce the inverse of A from these computations: $A^{-1} = \begin{pmatrix} -2 & -1 & -2 \\ -1 & 9 & 4 \\ 0 & 2 & 1 \end{pmatrix}$.

Since there are two exchanges of lines and one exchange of columns, the determinant of A is: $(-1)^3 \times (-2) \times \frac{9}{2} \times \frac{1}{9} = 1$.

3.3 LU factorization

We assume that we apply the Gaussian method and that, in the step k ($1 \leqslant k \leqslant n-1$), the pivot is always on the diagonal, that is that the term that appears in the cell (k,k) after $k-1$ steps is never equal to 0 (this is largely the general case).

Let k be an index verifying $1 \leqslant k \leqslant n$.

Denote by M_k the matrix of the system obtained after $k-1$ iterations, with $M_1 = A$; this matrix has values equal to 0 under the first $(k-1)$ terms of the diagonal (i.e. for any pair of indices (s,t) with $1 \leqslant t \leqslant k-1, s > t$) and, by hypothesis, $(M_k)_{k,k}$ is not equal to 0. The matrix M_n is upper triangular, we denote it U (for *upper*). We now assume $k \leqslant n-1$. For $1 \leqslant i \leqslant n$, let $\alpha_i = (M_k)_{i,k}$; thus, the k^e column of M_k (the pivot column) is:

$$\begin{pmatrix} \alpha_1 \\ ... \\ \alpha_k \neq 0 \\ ... \\ \alpha_n \end{pmatrix}.$$

Let E_k be the matrix that has 1 on the diagonal and elsewhere 0 except for $(E_k)_{i,k}$ with $i > k$ (part of the k^e column below the diagonal) and, for $i > k$, $(E_k)_{i,k} = -\frac{\alpha_i}{\alpha_k}$. This matrix E_k is therefore lower triangular and differs from the identity matrix only by its k^e column:

$$E_k = \begin{pmatrix} 1 & 0 & ... & & & & \\ 0 & 1 & 0 & ... & & & \\ & & ... & 0 & .. & & \\ & ... & 0 & 1 & 0 & ... & \\ & & & -\frac{\alpha_{k+1}}{\alpha_k} & 1 & 0 & ... \\ & & & ... & & & \\ & & & -\frac{\alpha_n}{\alpha_k} & 0 & & ...1 \end{pmatrix}.$$

We have by the choice of coefficients of E_k: $E_k M_k = M_{k+1}$.
With $M_1 = A$ and $M_n = U$, we obtain: $U = E_{n-1}E_{n-2}...E_1 A$.
Therefore: $A = E_1^{-1}E_2^{-1}...E_{n-1}^{-1}U$.
For $1 \leqslant k < i \leqslant n$, let $L_{i,k} = \frac{\alpha_i}{\alpha_k}$. We can easily check the equality:

$$E_k^{-1} = \begin{pmatrix} 1 & 0 & ... & & & & \\ 0 & 1 & 0 & ... & & & \\ & & ... & 0 & .. & & \\ & ... & 0 & 1 & 0 & ... & \\ & & & L_{k+1,k} & 1 & 0 & ... \\ & & & ... & & & \\ & & & L_{n,k} & 0 & & ...1 \end{pmatrix}.$$

and then

$$E_1^{-1}...E_{n-2}^{-1}E_{n-1}^{-1} = \begin{pmatrix} 1 & 0 & ... & & \\ L_{2,1} & 1 & 0 & ... & \\ L_{3,1} & L_{3,2} & 1 & 0 & ... \\ & & & ... & \\ L_{n,1} & L_{n,2} & \beta_{n,3} & ... & L_{n,n-1} & 1 \end{pmatrix}.$$

Denoting by L (for *lower*) the matrix above, we have: $A = LU$.

We thus obtain the so-called *LU factorization of A* : $A = LU$, with L lower triangular matrix with the value 1 on the diagonal and U upper triangular matrix.

Example

We treat again the example 1 of the Gaussian method with: $A = \begin{pmatrix} \mathbf{2} & 1 & -3 \\ 4 & 1 & 5 \\ 10 & -7 & 13 \end{pmatrix}$.

$M_1 = A$.

First iteration. The pivot is of value 2, in bold above;

- we have: $L_{2,1} = \dfrac{4}{2} = 2$, we multiply the first line by $L_{2,1} = 2$ before subtracting it from the second line.
- we have: $L_{3,1} = \dfrac{10}{2} = 5$, we multiply the first line by 5 before subtracting it from the third line.

We obtain the matrix:

$$M_2 = \begin{pmatrix} 2 & 1 & -3 \\ 0 & \mathbf{-1} & 11 \\ 0 & -12 & 28 \end{pmatrix}.$$

Second iteration. The pivot is of value -1, we have: $L_{3,2} = \dfrac{-12}{-1} = 12$, we multiply the second line by 12 before subtracting it from the third line. We obtain the matrix: $M_3 = \begin{pmatrix} 2 & 1 & -3 \\ 0 & -1 & 11 \\ 0 & 0 & -104 \end{pmatrix}$

Conclusion. We have: $U = \begin{pmatrix} 2 & 1 & -3 \\ 0 & -1 & 11 \\ 0 & 0 & -104 \end{pmatrix}$ and $L = \begin{pmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 5 & 12 & 1 \end{pmatrix}$.

The preceding considerations are based on the assumption that, in the application of the Gaussian method, the pivot is on the diagonal (see above). The following theorem gives a sufficient condition for this hypothesis to be satisfied.

Theorem of existence of the factorization LU
Let $A = (a_{ij})$ a square matrix (invertible) such that, for every k between 1 and n, the sub-matrix $\begin{pmatrix} a_{11} & \dots & a_{1k} \\ \vdots & & \vdots \\ a_{k1} & \dots & a_{kk} \end{pmatrix}$ is invertible. Then, the factorization $A = LU$ is possible (more precisely, the successive pivots can always be taken on the diagonal, without exchange of lines). Moreover, we can choose $L_{ii} = 1$ and the decomposition is then unique.

In fact, we can show that if the LU factorization fails (that is, if the pivots cannot always be chosen on the diagonal without exchanging lines), we can initially switch the rows of the matrix A to obtain a matrix A' for which the factorization LU is possible.

When several linear systems of the same matrix A have to be solved, the factorization LU is computed during the resolution of the first of these systems. The resolution of any later system $Ax = b$ results to the resolution of two systems of triangular matrices: the system $Ly = b$ and then the system $Ux = y$ (note that it is useless to know M explicitly, whose computation is not necessarily easy). Each system then takes only $n(n-1)$ additions, $n(n-1)$ multiplications and $2n$ divisions.

3.4 Cholesky method

The Cholesky method gives an interesting factorization in the case of symmetric positive definite matrices. In this case, one can choose an LU factorization with $U = L^{\mathrm{t}}$ while renouncing nevertheless to have diagonal terms all equal to 1 in L.

Theorem. Let A be a symmetric positive definite matrix. There is a lower triangular matrix B satisfying $A = BB^{\mathrm{t}}$. Moreover, we can impose that the diagonal elements of the matrix B are all strictly positive and the factorization $A = BB^{\mathrm{t}}$ is then unique.

In practice, we compute the matrix $B = \begin{pmatrix} b_{11} & 0 & \dots & & 0 \\ b_{21} & b_{22} & 0 & \dots & 0 \\ & & \dots & & \\ b_{n1} & b_{n2} & \dots & & b_{nn} \end{pmatrix}$ column

by column, from the equalities defining it: for $1 \leqslant i \leqslant j \leqslant n$,

$$a_{ij} = \sum_{k=1}^{i} b_{ik} b_{jk} = a_{ji}.$$

- For the first column, the formula gives
  - ⋆ $b_{11} = \sqrt{a_{11}}$
  - ⋆ pour $2 \leqslant i \leqslant n$, $b_{i1} = \dfrac{a_{i1}}{b_{11}}$.
- Pour $2 \leqslant j \leqslant n$,
  - ⋆ on the diagonal, $b_{jj} = \sqrt{a_{jj} - \sum_{k=1}^{j-1} {b_{jk}}^2}$
  - ⋆ for $j < i \leqslant n$, $b_{ij} = \dfrac{a_{ij} - \sum_{k=1}^{j-1} b_{ik} b_{jk}}{b_{jj}}$.

Remark.
1. The proof of the previous theorem would show that the b_{ij} thus obtained are well defined, thanks to the fact that A is positive definite.
2. The determinant of the A matrix can be easily computed:

$$\det(\mathrm{A}) = (b_{11} b_{22} ... b_{nn})^2.$$

A system $Ax = b$ then becomes $BB^{\mathrm{t}}x = b$. To solve the system, we solve $By = b$ and then $B^{\mathrm{t}}x = y$.

Complexity. In total (the factorization and the two resolutions), we do approximately $n^3/6$ additions, $n^3/6$ multiplications, $n^2/2$ divisions, n extractions of square roots, therefore, we do approximatively $n^3/3$ operations, that is to say, about half of the operations implemented by the Gaussian method. It is therefore advantageous to apply the method of Cholesky rather than the Gaussian method when A is symmetric definite positive.

Example.
Consider the following system:

$$\left\{ \begin{array}{rcrcrcr} 4x_1 & - & 2x_2 & & & = & 4 \\ -2x_1 & + & 2x_2 & + & 3x_3 & = & -8 \\ & & 3x_2 & + & 10x_3 & = & -20 \end{array} \right.$$

The corresponding A matrix is: $A = \begin{pmatrix} 4 & -2 & 0 \\ -2 & 2 & 3 \\ 0 & 3 & 10 \end{pmatrix}$. This matrix is symmetric definite positive. Indeed, let x be a vector of $\mathbb{R}^n$ represented by a column vector. We then have:

$$\begin{aligned} x^{\mathrm{t}}Ax &= 4{x_1}^2 - 4x_1x_2 + 2{x_2}^2 + 6x_2x_3 + 10{x_3}^2 \\ &= (2x_1 - x_2)^2 + (x_2 + 3x_3)^2 + {x_3}^2. \end{aligned}$$

Therefore, if x is not equal to zero, $x^{\mathrm{t}}Ax$ is a strictly positive real.

First step. We compute B such as $A = BB^{\mathrm{t}}$ with B upper triangular. The application of the preceding formulas gives:

$$\begin{aligned} b_{11} &= \sqrt{a_{11}} = 2 \\ b_{21} &= \frac{a_{21}}{b_{11}} = -1 \\ b_{31} &= \frac{a_{31}}{b_{11}} = 0 \\ b_{22} &= \sqrt{a_{22} - \sum_{k=1}^{1} {b_{2k}}^2} = \sqrt{2-1} = 1 \\ b_{32} &= \frac{a_{32} - \sum_{k=1}^{1} b_{3k}b_{2k}}{b_{22}} = \frac{3 - 0 \times (-1)}{1} = 3 \\ b_{33} &= \sqrt{a_{33} - \sum_{k=1}^{2} {b_{3k}}^2} = \sqrt{10 - 0 - 9} = 1. \end{aligned}$$

Therefore: $B = \begin{pmatrix} 2 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 3 & 1 \end{pmatrix}$ and $\det(A) = (2 \times 1 \times 1)^2 = 4$.

Second step. By the backward substitutions method, we solve the two systems $By = b$ and $B^{\mathrm{t}}x = y$.

The system $By = b$ is:

$$\left\{ \begin{array}{rcrcrcr} 2y_1 & & & & & = & 4 \\ -y_1 & + & y_2 & & & = & -8 \\ & & 3y_2 & + & y_3 & = & -20 \end{array} \right.$$

which has $y_1 = 2, y_2 = -6, y_3 = -2$ for solution.

The system ${}^{t}B x = y$ is:

$$\begin{cases} 2x_1 - x_2 = 2 \\ x_2 + 3x_3 = -6 \\ x_3 = -2 \end{cases}$$

which has $x_3 = -2, x_2 = 0, x_1 = 1$ for solution.

The solution of the system is therefore: $x = \begin{pmatrix} 1 \\ 0 \\ -2 \end{pmatrix}$.

If we had another system to solve with the same matrix A, only the second step would be applied.

Chapter 4

Eigenvalues and eigenvectors

Let us first remark that the search for the eigenvalues of a matrix, unlike the computation of its inverse, is a difficult problem. Given the polynomial $P(\lambda) = \lambda^n + a_1\lambda^{n-1} + ... + a_{n-1}\lambda + a_n$, define the matrix:

$$\begin{pmatrix} -a_1 & -a_2 & -a_3 & & ... & -a_{n-1} & -a_n \\ 1 & 0 & & & & & \\ 0 & 1 & 0 & & & & \\ & 0 & 1 & 0 & & & \\ & & & ... & & & \\ & & & 0 & 1 & 0 & \\ & & & & 0 & 1 & 0 \end{pmatrix}$$

Its characteristic polynomial is $(-1)^n P(\lambda)$; the matrix therefore has for eigenvalues the roots of P. According to Abel's theorem, it is impossible to compute the roots of any polynomial from degree 5 using a finite number of applications of the four usual arithmetic operations plus root extraction. If a search method of eigenvalues always converges in a finite number of these operations, it would be the same for the search for the roots of any polynomial equation, which is contrary to the result of Abel.

To compute an approximation of the eigenvalues of a matrix A, the basic idea is to look for a matrix similar to A, that is to say of the form $P^{-1}AP$, triangular or diagonal, and whose diagonal will consist of the eigenvalues of A. We will study in this chapter only one method, the method of Jacobi, which applies to the case of real symmetric matrices. Remember that the eigenvalues of such a matrix are real.

4.1 Jacobi method

Let A be a symmetric real matrix, let p and q be two indices fulfilling $\boldsymbol{p} < \boldsymbol{q}$ such that the element (non diagonal) a_{pq} is not equl to zero (if there is none, A is diagonal and the eigenvalues of A are precisely the values of the diagonal).

Let θ be a real number; we define a matrix Ω depending on θ. The matrix Ω differs from the identity matrix of order n only by the following four coefficients:

$$\Omega_{pp} = \Omega_{qq} = \cos\theta, \Omega_{pq} = \sin\theta, \Omega_{qp} = -\sin\theta.$$

The matrix Ω is represented below.

$$\Omega = \begin{pmatrix}
1 & 0 & & & & \dots & & & & 0 & 0 \\
0 & 1 & & & & \dots & & & & 0 & 0 \\
 & & \dots & & & \dots & & & \dots & & \\
 & & & \cos\theta & & \dots & & \sin\theta & & & \\
 & & & & 1 & \dots & & & & & \\
 & & & & & \dots & & & & & \\
 & & & & & \dots & 1 & & & & \\
 & & & -\sin\theta & & \dots & & \cos\theta & & & \\
 & & \dots & & & \dots & & & \dots & & \\
0 & 0 & & & & \dots & & & & 1 & 0 \\
0 & 0 & & & & \dots & & & & 0 & 1
\end{pmatrix}.$$

The matrix Ω is orthogonal. This is the rotation matrix of angle $-\theta$ in the plane defined by the p^{th} and q^{th} base vectors.

We set: $B = \Omega^{\text{t}} A \Omega$. The matrix B, also symmetric, is similar to the matrix A and thus admits the same eigenvalues as A. The following equalities are easily established:

$$\begin{cases}
\text{si } i \notin \{p,q\} \text{ and } j \notin \{p,q\}, \quad b_{ij} = b_{ji} = a_{ij} \\
\text{si } i \notin \{p,q\}, \quad b_{pi} = b_{ip} = a_{pi}\cos\theta - a_{qi}\sin\theta \\
\text{si } i \notin \{p,q\}, \quad b_{qi} = b_{iq} = a_{pi}\sin\theta + a_{qi}\cos\theta \\
b_{pp} = a_{pp}\cos^2\theta + a_{qq}\sin^2\theta - a_{pq}\sin 2\theta \\
b_{qq} = a_{pp}\sin^2\theta + a_{qq}\cos^2\theta + a_{pq}\sin 2\theta \\
b_{pq} = b_{qp} = a_{pq}\cos 2\theta + \dfrac{a_{pp} - a_{qq}}{2}\sin 2\theta.
\end{cases}$$

We notice the equivalence $b_{pq} = 0 \Leftrightarrow \cot 2\theta = \dfrac{a_{qq} - a_{pp}}{2a_{pq}}$ (where cot denotes the cotangent trigonometric function). We try to make $b_{pq} = 0$, so we choose θ to satisfy the formula above. There are four solutions in the interval $]-\pi, \pi]$, two successive solutions differing from $\pi/2$. So there is a unique solution in the interval $]-\frac{\pi}{4}, \frac{\pi}{4}]$, this is the chosen solution.

We set: $x = \dfrac{a_{qq} - a_{pp}}{2a_{pq}}, t = \tan\theta, s = \sin\theta, c = \cos\theta$. The following trigonometric relationships are recalled:

$$\cot 2\theta = \frac{\cos 2\theta}{\sin 2\theta} = \frac{\cos^2\theta - \sin^2\theta}{2\sin\theta\cos\theta} = \frac{1-t^2}{2t}.$$

We try to have: $x = \dfrac{1-t^2}{2t}$; so, t must fulfil the equation: $t^2 + 2xt - 1 = 0$. Since the product of the roots is -1 and θ is in the range $]-\frac{\pi}{4}, \frac{\pi}{4}]$, t is the root of the equation of smaller absolute value if $x \neq 0$, and is 1 if $x = 0$.

As we have $c > 0$, it comes $c = \dfrac{1}{\sqrt{1+t^2}}$ and $s = ct = \dfrac{t}{\sqrt{1+t^2}}$.

The coefficients of the B matrix can actually be computed by the following formulas, in which t, c and s are defined as above:

$$\begin{cases} \text{if } i \notin \{p,q\} \text{ and } j \notin \{p,q\}, b_{ij} = b_{ji} = a_{ij} \\ \text{if } i \notin \{p,q\}, \quad b_{pi} = b_{ip} = ca_{pi} - sa_{qi} \\ \text{if } i \notin \{p,q\}, \quad b_{qi} = b_{iq} = sa_{pi} + ca_{qi} \\ b_{pp} = a_{pp} - ta_{pq} \\ b_{qq} = a_{qq} + ta_{pq}. \end{cases}$$

One step of Jacobi method, summary

- We choose in the current matrix two indices p and q, with $\boldsymbol{p < q}$.
- We set $x = \dfrac{a_{qq} - a_{pp}}{2a_{pq}}$.
- We solve $t^2 + 2xt - 1 = 0$. We retain for t the root of the smallest absolute value if $x \neq 0$, and 1 if $x = 0$.
- We compute: $c = \dfrac{1}{\sqrt{1+t^2}}$ and $s = \dfrac{t}{\sqrt{1+t^2}}$.
- The new coefficients are computed using the above formulas that use c, s and t.

Remark. It is natural to wonder whether, in doing this transformation which has the merit of canceling non-diagonal elements, we do not risk, at the same

time, making non-null elements that were previously null. This is not only true, but even inevitable, since otherwise we would diagonalize the matrix A with about n^3 elementary operations, whereas we indicated in the initial remark that this was impossible. There is, however, good reason to hope, by reiterating the process, a convergence of the matrices B obtained to a diagonal matrix, as we will explain below.

Theorem. Let A be a real symmetric matrix and let B be the matrix obtained using the above method. We then have the relations:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} b_{ij}^2,$$

$$\sum_{i=1}^{n} a_{ii}^2 + 2a_{pq}^2 = \sum_{i=1}^{n} b_{ii}^2.$$

Proof of the theorem. The first relation results from the conservation of the norm $||\ ||_E$ by a unitary transformation. For the second, only the elements of the rows and columns p and q are modified. Therefore, diagonal elements other than a_{pp} and a_{qq} are invariant as well as their squares. We have:

$$\begin{aligned} b_{pp}^2 + b_{qq}^2 &= a_{pp}^2 + a_{qq}^2 + 2t^2 a_{pq}^2 + 2t a_{pq}(a_{qq} - a_{pp}) \\ &= a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2 + 2a_{pq}(t^2 a_{pq} + t(a_{qq} - a_{pp}) - a_{pq}). \end{aligned}$$

Now, the choice of t is that we have $t^2 + t\frac{a_{qq}-a_{pp}}{a_{pq}} - 1 = 0$.
Hence the result stated: $b_{pp}^2 + b_{qq}^2 = a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2$. $\Diamond$

This theorem shows that the weight of the matrix is displaced, during the iterations of the method of Jacobi, on the diagonal of the matrix and, consequently, that the non-diagonal elements have a weight which decreases. Moreover, it seems that in order to accelerate the convergence of the process, it is advantageous to choose as the pair (p, q) the indices of a non-diagonal element of maximum absolute value. It is indeed this choice that is often made.
Theorem.The sequence of the matrices obtained by the method of Jacobi is convergent and converges to a diagonal matrix containing the eigenvalues of A.

The Jacobi method also makes it possible to obtain an approximation of the eigenvectors of a matrix A, at least when the eigenvalues of A are

distinct. This is what the following theorem says.

Theorem. If all the eigenvalues of the A matrix are distinct, then the sequence of the Ω matrix products with the new Ω matrix on the right of the product at each step converges to an orthogonal matrix whose column vectors constitute an orthonormal set of eigenvectors of the matrix A.

Example 1. Let us apply Jacobi method to find approximations of eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 5 \end{pmatrix}$. There are only the coefficients $p = 1$ and $q = 2$ which are to be considered. With the previous notations, we have $x = 0$ and so $t = 1, s = c = \frac{\sqrt{2}}{2}$. Therefore, the matrix Ω is: $\Omega = \begin{pmatrix} \sqrt{2}/2 & \sqrt{2}/2 & 0 \\ -\sqrt{2}/2 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.
Applying the previous formulas gives:
- unchanged term (here, only one *a priori*): $b_{33} = a_{33} = 5$
- first line and first column, except diagonal:
$b_{12} = b_{21} = 0$
$b_{13} = b_{31} = ca_{13} - sa_{23} = 0$
- second row and second column, except diagonal:
$b_{23} = b_{32} = sa_{13} + ca_{23} = 0$
- diagonal terms that change *a priori*:
$b_{11} = a_{11} - ta_{12} = 1 - 2 = -1$
$b_{22} = a_{22} + ta_{12} = 1 + 2 = 3.$

We thus obtain: $B = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & 5 \end{pmatrix}$.

The matrix is diagonal, the Jacobi method converges here with one iteration (the example is very simple) and gives us the eigenvalues (exactly) as well as the eigenvectors of A. The eigenvalues of A are: -1, 3 and 5. The orthonormal basis of eigenvectors consists of vectors: $(\sqrt{2}/2, -\sqrt{2}/2, 0)^{\mathrm{t}}$, $(\sqrt{2}/2, \sqrt{2}/2, 0)^{\mathrm{t}}$, $(0, 0, 1)^{\mathrm{t}}$.

Example 2. Let us apply Jacobi method to find approximations of eigenvalues

and eigenvectors of the matrix $A = \begin{pmatrix} 1 & 2 & 4 \\ 2 & -3 & -1 \\ 4 & -1 & 7 \end{pmatrix}$.

First step.
We choose the largest absolute value of a non-diagonal coefficient: this is the value 4, with $p = 1, q = 3$.

We compute x: $x = \frac{7-1}{2\times 4} = \frac{3}{4}$.

We solve the equation $t^2 + 2xt - 1 = 0$, that is: $t^2 + \frac{3}{2}t - 1 = 0$, whose roots are $t = 1/2$ and $t = -2$. We keep the smallest root in absolute value: $t = 1/2$.

We compute c and s: $c = \dfrac{1}{\sqrt{1+t^2}} = \dfrac{2}{\sqrt{5}} = \dfrac{2\sqrt{5}}{5}$ and $s = tc = \dfrac{\sqrt{5}}{5}$.

Then we apply the formulas giving the coefficients of B, with of course $b_{13} = b_{31} = 0$:
b_{22} remains unchanged: $b_{22} = -3$

$$b_{12} = b_{21} = ca_{12} - sa_{32} = \frac{2\sqrt{5}}{5} \times 2 - \frac{\sqrt{5}}{5} \times (-1) = \sqrt{5}$$

$$b_{32} = b_{23} = sa_{12} + ca_{32} = \frac{\sqrt{5}}{5} \times 2 + \frac{2\sqrt{5}}{5} \times (-1) = 0$$

$$b_{11} = a_{11} - ta_{13} = 1 - \frac{1}{2} \times 4 = -1$$

$$b_{33} = a_{33} + ta_{13} = 7 + \frac{1}{2} \times 4 = 9.$$

We thus obtain the matrix B:

$$B = \begin{pmatrix} -1 & \sqrt{5} & 0 \\ \sqrt{5} & -3 & 0 \\ 0 & 0 & 9 \end{pmatrix}$$

and the change-of-basis matrix Ω_1:

$$\Omega_1 = \begin{pmatrix} 2\frac{\sqrt{5}}{5} & 0 & \frac{\sqrt{5}}{5} \\ 0 & 1 & 0 \\ -\frac{\sqrt{5}}{5} & 0 & 2\frac{\sqrt{5}}{5} \end{pmatrix} \approx \begin{pmatrix} 0,894 & 0 & 0,447 \\ 0 & 1 & 0 \\ -0,447 & 0 & 0,894 \end{pmatrix}.$$

Second step. We go back from the B matrix to move to a C matrix computed with Jacobi method.

We set: $p = 1, q = 2$. We compute x: $x = \frac{-3+1}{2\sqrt{5}} = -\frac{\sqrt{5}}{5}$.

We solve the equation: $t^2 - 2\frac{\sqrt{5}}{5}t - 1 = 0$ whose roots are: $t = \frac{\sqrt{5}}{5}(1+\sqrt{6})$ and $t = \frac{\sqrt{5}}{5}(1-\sqrt{6})$. The smallest root in absolute value is: $t = \frac{\sqrt{5}}{5}(1-\sqrt{6})$,

$t \approx -0,648$. Thus: $c = \dfrac{1}{\sqrt{1+t^2}} \approx 0,839$ and $s = ct \approx -0,544$. We then obtain:
$c_{33} = b_{33} = 9$
$c_{12} = c_{21} = 0$
$c_{11} = b_{11} - tb_{12} = -1 - \frac{\sqrt{5}}{5}(1-\sqrt{6})\sqrt{5} = -2+\sqrt{6}$
$c_{22} = b_{22} + tb_{12} = -3 + \frac{\sqrt{5}}{5}(1-\sqrt{6})\sqrt{5} = -2-\sqrt{6}$
$c_{13} = c_{31} = cb_{13} - sb_{23} = 0$
$c_{23} = c_{32} = sb_{13} + cb_{23} = 0.$

Therefore:

$$C = \begin{pmatrix} -2+\sqrt{6} & 0 & 0 \\ 0 & -2-\sqrt{6} & 0 \\ 0 & 0 & 9 \end{pmatrix}.$$

The approximate change-of-basis Ω_2 is given by:

$$\Omega_2 \approx \begin{pmatrix} 0,839 & -0,544 & 0 \\ 0,544 & 0,839 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Since the C matrix is diagonal, the method is complete. The eigenvalues of A are: $-2+\sqrt{6}$, $-2-\sqrt{6}$, 9.

An approximate orthonormal basis of eigenvectors is obtained by computing the product $\Omega_1\Omega_2$: $\Omega_1\Omega_2 \approx \begin{pmatrix} 0,75 & -0,486 & 0,447 \\ 0,544 & 0,839 & 0 \\ -0,375 & 0,243 & 0,894 \end{pmatrix}$.

Example 3. Let us apply Jacobi method to find approximations of eigenvalues and eigenvectors of the matrix $A = \begin{pmatrix} 1 & 2 & 4 \\ 2 & -3 & 0 \\ 4 & 0 & 7 \end{pmatrix}$.

We choose the largest absolute value of a non-diagonal coefficient. This is the value 4. We set: $p = 1, q = 3$.

As in example 2, the following is computed: $t = 1/2$, $c = \dfrac{2\sqrt{5}}{5}$ and

$s = \dfrac{\sqrt{5}}{5}$.

We have:

$b_{22} = -3$

$b_{13} = b_{31} = 0$

$b_{12} = b_{21} = ca_{12} - sa_{32} = \dfrac{2\sqrt{5}}{5} \times 2 - \dfrac{\sqrt{5}}{5} \times 0 = \dfrac{4\sqrt{5}}{5}$

$b_{32} = b_{23} = sa_{12} + ca_{32} = \dfrac{\sqrt{5}}{5} \times 2 + \dfrac{2\sqrt{5}}{5} \times 0 = \dfrac{2\sqrt{5}}{5}$

$b_{11} = a_{11} - ta_{13} = -1$

$b_{33} = a_{33} + ta_{13} = 9.$

We thus obtain: $B = \begin{pmatrix} -1 & 4\frac{\sqrt{5}}{5} & 0 \\ 4\frac{\sqrt{5}}{5} & -3 & 2\frac{\sqrt{5}}{5} \\ 0 & 2\frac{\sqrt{5}}{5} & 9 \end{pmatrix}$.

This example shows that coefficients can go from null to non-zero. Nevertheless, going from A to B, the weight of the matrix focused on the diagonal. The method should be continued to compute an approximation of the eigenvalues and eigenvectors of the A matrix.

Chapter 5

Linear programming: the simplex algorithm

5.1 Introduction

To illustrate what the *linear programming* (also called *linear optimization*) is, let us start with a simple example. With this example, we introduce some properties that will be useful for the *simplex algorithm*[1]. It was designed by G. Danzig from 1947[2] and remains since one of the main algorithms of linear optimization, even if other algorithms came then to compete, in particular the method of N. Karmakar[3].

A factory produces two kinds of products, p_1 and p_2, using two machines m_1 and m_2. It is assumed that the manufactured quantity of these products is not necessarily an integer, but only a positive real or zero. Each unit of

[1]The name of this method may seem a bit misleading. In geometry, a simplex of dimension d, or d-simplex, is the convex hull of $d+1$ points. Thus, a 1-simplex is a straight line segment, a 2-simplex is a triangle and a 3-simplex is a tetrahedron. The simplex method is not limited to simplexes, but more generally considers polyhedron. For this reason, perhaps the name of "polyhedron method" would have been more appropriate.

[2]G.B. Danzig, Linear Programming, in Problems for the Numerical Analysis of the Future, Proceedings of Symposium on Modern Calculating Machinery and Numerical Methods, UCLA, July 29-31, 1948. See also GB Danzig and MN Thapa, Linear programming 1: Introduction, 1997, and Linear Programming 2: Theory and Extensions, 2003, Springer-Verlag.

[3]N. Karmarkar (1984). A New Polynomial Time Algorithm for Linear Programming, Combinatorica, Vol. 4, nr. 4, p. 373-395.

product being manufactured must pass on both machines in any order and during the following times, expressed in minutes:

	p_1	p_2
m_1	30	20
m_2	40	10

The machine m_1 is available 6000 minutes per month and the machine m_2 is available 4000 minutes per month. The profit realized on a unit of the product p_1 is 400 €. The profit realized on a unit of the product p_2 is 200 €.

We want to find the monthly manufacturing plan that maximizes profit.

For this, let us call x_1 (respectively x_2) the number of units of the product p_1 (respectively p_2) to be produced monthly; we see that this problem can be expressed in the following form:

$$\text{Maximiser } z = 400x_1 + 200x_2$$
$$\text{avec les contraintes : } \begin{cases} 30x_1 + 20x_2 \leqslant 6000 \\ 40x_1 + 10x_2 \leqslant 4000 \\ x_1 \geqslant 0, x_2 \geqslant 0. \end{cases}$$

The problem being in two variables, it admits a graphical solution easy to implement, represented by the figure 5.1.

The points (x_1, x_2) that fulfil the constraints belong to the $OABC$ quadrilateral. Let λ be a real. The family of lines:

$$D_\lambda = \{(x_1, x_2) \text{ with } 400x_1 + 200x_2 = \lambda\}$$

is a family of parallel lines. Among those lines that have a non-empty intersection with the quadrilateral, it is the one that passes through B which corresponds to the largest value of λ: it meets the quadrilateral of the constraints at the point of coordinates (40 , 240). The optimal solution of our problem is therefore $x_1 = 40$, $x_2 = 240$ (et $z = 64\ 000$).

More generally, a problem of linear programming is a problem which can be formulated as follows:

$$\text{maximize a linear form of } n \text{ variables } x_1, ..., x_n\text{: } \sum_{j=1}^{n} c_j x_j$$

the variables being submitted:

- to m linear constraints: for $i \in \{1, 2, ..., m\}$, $\sum_{j=1}^{n} a_{ij} x_j \leqslant b_i$
- to n positivity constraints: for $j \in \{1, 2, ..., n\}$, $x_j \geqslant 0$.

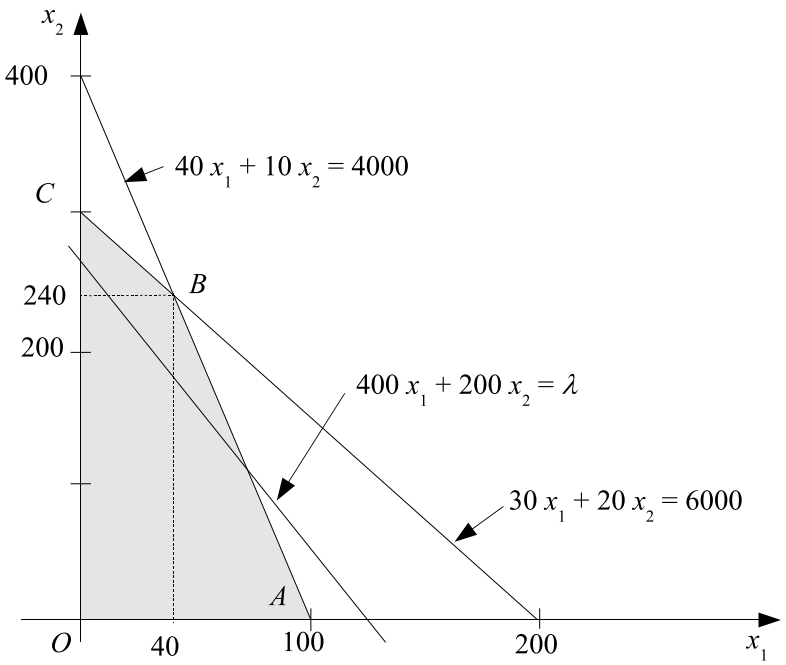

Figure 5.1: Illustration for the example.

This formulation is called the *standard form* of a linear programming problem. Other formulations of the problem to be solved can be considered. In all the following, we will not consider the case of strict inequalities (the domain defined by the constraints would not be closed, the problem might not admit an optimal solution, even if the function z is bounded from above on this domain). On the other hand, problems where it is a question of minimization, or for which there appear constraints of equality or non-strict inequality in the converse sens, or for which variables have other constraints than those to be positive or null (or are of non-constrained sign) can easily be stated in standard form, as specified by the following indications:

- to minimize a function f (linear or not) is to maximize $-f$, since we have the relation: minimum of $f = -$ maximum of $(-f)$;

- we transform an inequality of type "$\geqslant$" into an inequality of type "$\leqslant$" by multiplying it by -1;

- an equality $\sum_{j=1}^{n} a_{ij}x_j = b_i$ is the same as the two following inequalities: $\sum_{j=1}^{n} a_{ij}x_j \leqslant b_i$ and $\sum_{j=1}^{n}(-a_{ij})x_j \leqslant -b_i$;
- we replace a variable x constrained by the inequality $x \geqslant \alpha$ by the variable $x-\alpha$ which will have to be positive or null (if there are several constraints of this type, we only consider that with the highest value α and we eliminate the others);
- we replace a variable x constrained by the inequality $x \leqslant \beta$ by the variable $\beta - x$ which will have to be positive or null (if there are several constraints of this type, we only consider the one with the lowest value β and eliminate the others);
- we replace a variable constrained by the double inequality $\alpha \leqslant x \leqslant \beta$ by the variable $y = x - \alpha$ and we add the constraints $y \leqslant \beta - \alpha$ and $y \geqslant 0$ (if there are several constraints of this type involving the same variable x, we keep only the most restrictive interval);
- we write a variable x that is neither positive nor negative as the difference of two positive or zero variables: $x = x^+ - x^-$ with $x^+ \geqslant 0$ and $x^- \geqslant 0$.

We can then wonder if the proposed approach for the previous example is likely to be generalized to the resolution of any problem of linear programming. Since it is always possible to write a problem of linear programming (without constraint of strict inequality) in standard form, consider a problem set in this form:

$$\text{Maximize } z = \sum_{j=1}^{n} c_j x_j$$
$$\text{with the constraints: } \begin{cases} \text{for } i \in \{1, 2, ..., m\}, \sum_{j=1}^{n} a_{ij}x_j \leqslant b_i \\ \text{for } j \in \{1, 2, ..., n\}, x_j \geqslant 0. \end{cases}$$

The set of points of $\mathbb{R}^n$ of coordinates $x_1, ..., x_n$ fulfilling the $m + n$ previous constraints is a polyhedron called *constraints polyhedron* . This polyhedron is *convex*, that is, for every point M and any point P of the polyhedron, the segment $[M, P]$ is entirely contained in the polyhedron. Indeed, let $M = (x_1, ..., x_n)$ and $P = (y_1, ..., y_n)$ any two points of the constraints polyhedron; then, for any real λ verifying $0 \leqslant \lambda \leqslant 1$, it is easy to check that the point $\lambda M + (1 - \lambda)P$ (of coordinates $\lambda x_i + (1 - \lambda)y_i$) belongs to the polyhedron. The n-uplets $(x_1, ..., x_n)$ that fulfil the constraints are called *feasible*

solutions of the problem. These are the coordinates of the points located inside the constraints polyhedron which, in the example, was the quadrilateral $OABC$.

The development of the simplex method will show the following theorem (we can also be convinced of this result using the two-variable example given above):

Theorem 1. *We consider a linear programming problem whose constraints polyhedron is non-empty and whose function to be maximized is bounded from above on this polyhedron. Then the problem admits a maximum (which is finite) reached in at least one vertex of the constraints polyhedron.*

The idea of the simplex algorithm is to iteratively move from one vertex of the constraints polyhedron to an adjacent vertex following edges of the polyhedron so as to increase the value of the function to be optimized, until finding a vertex where the maximum is reached. It is thanks to the convexity of the polyhedron and to the linearity of the function to maximum that we can seek the maximum in a vertex of the polyhedron.

5.2 The simplex algorithm on an example

Let us apply the simplex algorithm to a more sophisticated example, to illustrate how it works.

A textile factory produces four types of textiles: kelsch, nanzouk, shantung and zenana. These textiles result from three main operations: spinning, weaving, dyeing. They are produced in varying length, measured here in kilometers. The production of one kilometer of textile requires a certain number of hours of spinning, weaving and dyeing, these numbers depending on the textile. In addition, the sale of these textiles brings some profit expressed in euros. This data are specified in the following table, for one kilometer of fabric:

	kelsch	nanzouk	shantung	zenana
spinning	2	4	5	7
weaving	1	1	2	2
dyeing	1	2	3	3
benefit	7	9	18	17

Each day, the company has 42 hours of spinning, 17 hours of weaving and 24 hours of dyeing. The aim is to establish a manufacturing plan so as to maximize the benefit (it is assumed that it is in stable state of manufacture and not in the initial phase where it is necessary to spin before weaving and weaving before dyeing).

Let us call x_1, x_2, x_3, x_4 the respective lengths of kelsch, nanzouk, shantung and zenana produced daily. The problem then admits the following modeling:

$$\begin{array}{l}\text{Maximize } z = 7x_1 + 9x_2 + 18x_3 + 17x_4\\ \text{with the constraints:}\\ \left\{\begin{array}{rcrcrcrcl} 2x_1 & + & 4x_2 & + & 5x_3 & + & 7x_4 & \leqslant & 42\\ x_1 & + & x_2 & + & 2x_3 & + & 2x_4 & \leqslant & 17\\ x_1 & + & 2x_2 & + & 3x_3 & + & 3x_4 & \leqslant & 24\\ x_1 \geqslant 0, x_2 \geqslant 0, x_3 \geqslant 0, x_4 \geqslant 0. & & & & & & & & \end{array}\right.\end{array}$$

We recognize a linear programming problem in standard form. We will solve this problem using the simplex algorithm that we will explain on this example.

We introduce three non-negative variables called *slack variables* x_5, x_6, x_7, which measure for each resource the difference between the quantity initially available and the quantity consumed by the manufacturing plan given by x_1, x_2, x_3 and x_4. We obtain what is called a *dictionary* (see below the general definition), the first for the resolution of this problem (there will be others):

$$\begin{array}{rcrcrcrcrcr} x_5 & = & 42 & - & 2x_1 & - & 4x_2 & - & 5x_3 & - & 7x_4\\ x_6 & = & 17 & - & x_1 & - & x_2 & - & 2x_3 & - & 2x_4\\ x_7 & = & 24 & - & x_1 & - & 2x_2 & - & 3x_3 & - & 3x_4\\ \hline z & = & & & 7x_1 & + & 9x_2 & + & 18x_3 & + & 17x_4 \end{array}$$

Dictionary I

The problem can now be written:

Maximize z with $x_k \geqslant 0$ for $1 \leqslant k \leqslant 7$.

The constraints polyhedron is limited in $\mathbb{R}^4$ by the hyperplanes of equations $x_k = 0$ for $1 \leqslant k \leqslant 7$.

In this dictionary, the variables x_5, x_6 and x_7 are expressed as linear

functions of the variables x_1, x_2, x_3 and x_4; we say that the variables x_5, x_6 and x_7 are currently the *basic variables* of the dictionary and the variables x_1, x_2 , x_3 and x_4 the *non-basic variables* of the dictionary[4]. The *basic solution* associated with the dictionary is the solution obtained by assigning the value 0 to all the non-basic variables; the values of the basic variables result from it.

In order to distinguish the functions and variables from the values of these functions and variables, we use the sign * when dealing with values: thus x^* will represent a value taken by the variable x. With this notation, the equalities $x_1^* = 0$, $x_2^* = 0$, $x_3^* = 0$, $x_4^* = 0$ give $x_5^* = 42$, $x_6^* = 17$ and $x_7^* = 24$. The seven variables having positive or null values in this basic solution, we say that this dictionary is *feasible* . We can notice that the point of coordinates (0, 0, 0, 0) is here a vertex of the constraints polyhedron; the basic solution associated with the dictionary then gives to z the value 0.

The following remark is the basis of the method: we consider the expression of z in the current dictionary; if, in this expression, a non-basic variable with a strictly positive coefficient is increased from 0 (other non-basic variables remaining zero), the value of z increases. In our example, choose the variable x_3 (we could also choose here any one of the other three non-basic variables). Keeping x_1, x_2 and x_4 to 0, we try to increase x_3 to the maximum, while retaining the property that the point M of $\mathbb{R}^4$ of coordinates $(0, 0, x_3, 0)$ remains in the polyhedron of the constraints (we move then on an edge of the constraints polyhedron from the vertex (0, 0, 0, 0)).

The constraints on increasing variable x_3 are:

$x_5 \geqslant 0$, which results in $x_3 \leqslant 8.4$;

$x_6 \geqslant 0$, which results in $x_3 \leqslant 8.5$;

$x_7 \geqslant 0$, which results in $x_3 \leqslant 8$.

The first hyperplane that the point M meets is therefore that of equation $x_7 = 0$: the point M then arrived at a new vertex of the constraints polyhedron, at the intersection of hyperplanes of equations $x_1 = 0$, $x_2 = 0$, $x_4 = 0$, $x_7 = 0$. We will compute a new dictionary by switching roles of x_3 and x_7, to iterate the process we have just employed. We use the equation of the dictionary I which gives x_7 to write x_3 as a function of x_1, x_2, x_4 and x_7; we then replace x_3 with this expression in the other equations of the dictionary.

[4]The definition 1 precise further what is a basis.

We thus obtain a second dictionary:

$$\begin{array}{rcrcrcrcrcr}
x_3 &=& 8 &-& \frac{1}{3}x_1 &-& \frac{2}{3}x_2 &-& x_4 &-& \frac{1}{3}x_7 \\
x_5 &=& 2 &-& \frac{1}{3}x_1 &-& \frac{2}{3}x_2 &-& 2x_4 &+& \frac{5}{3}x_7 \\
x_6 &=& 1 &-& \frac{1}{3}x_1 &+& \frac{1}{3}x_2 & & & + & \frac{2}{3}x_7 \\
\hline
z &=& 144 &+& x_1 &-& 3x_2 &-& x_4 &-& 6x_7
\end{array}$$

Dictionary II

We say that we have chosen x_3 as the *entering variable* and that x_7 was the *leaving variable* . The basic variables are now x_3, x_5 and x_6, and the non-basic variables x_1, x_2, x_4 and x_7. In the new basic solution, the function z is equal to 144, which is obtained by giving the value 0 to non-basic variables. We notice that we thus have a new feasible solution more interesting than that associated with the first dictionary.

In the new expression of the function z, we see that only the variable x_1 has a strictly positive coefficient: we introduce x_1 into the basis (x_1 is the entering variable), and we thus go by a new edge of the constraints polyhedron; we have the following limits on the possible increase of the value of x_1 from the null value, the other non-basic variables remaining at 0:

$x_3 \geqslant 0$, which results in $x_1 \leqslant 24$;

$x_5 \geqslant 0$, which results in $x_1 \leqslant 6$;

$x_6 \geqslant 0$, which results in $x_1 \leqslant 3$.

It is the third limit which is the most restrictive; x_6 leaves the basis, which leads to the following dictionary:

$$\begin{array}{rcrcrcrcrcr}
x_1 &=& 3 &+& x_2 & & & - & 3x_6 &+& 2x_7 \\
x_3 &=& 7 &-& x_2 &-& x_4 &+& x_6 &-& x_7 \\
x_5 &=& 1 &-& x_2 &-& 2x_4 &+& x_6 &+& x_7 \\
\hline
z &=& 147 &-& 2x_2 &-& x_4 &-& 3x_6 &-& 4x_7
\end{array}$$

Dictionnaire III

The basic solution associated with this new dictionary gives to z the value 147. Moreover, we see on the last line of this dictionary that, since the variables x_2, x_4, x_6, x_7 are positive or zero, the searched optimum of z is bounded from above by 147. The current basic solution provides us an optimal solution for the problem:

- three kilometers of kelsch, zero of nanzouk, seven of shantung and zero zenana must be made every day;

- all hours of weaving and dyeing are used, while there is one hour of spinning available;
- the maximum profit is equal to 147.

Remarks.
1. It turns out that the solution obtained here is integral whereas it was not imposed by the formulation of the problem. This is not general and the problems of integer linear programming (that is to say, linear programming problems for which variables must be integer) can be qualitatively more complicated.
2. The method consists, at each step, in choosing to enter in basis a variable whose coefficient in the function z to optimize is strictly positive. However, this does not always lead up to a strict increase of z. We will return to this phenomenon in the section on "degeneracy".
3. Finally, we had the chance to find, without any difficulty, a vertex of the constraints polyhedron or, in other words, a feasible dictionary available as starting point. Indeed, the "origin was feasible": when the variables $x_1, x_2, ..., x_n$ are equal to 0, the b_i being positive or zero, the values of the slack variable are positive or null. We will study further less favorable cases.

5.3 Definitions and terminology

Let us go back to some definitions. A linear programming problem is set in standard form if it is written in the form:

$$\text{maximize a linear form } z \text{ of } n \text{ variables } x_1, ..., x_n\text{: } z = \sum_{j=1}^{n} c_j x_j,$$

the variables verifying:

- m linear constraints: for $i \in \{1, 2, ..., m\}, \sum_{j=1}^{n} a_{ij} x_j \leqslant b_i$,
- n constraints of positivity: for $j \in \{1, 2, ..., n\}, x_j \geqslant 0$.

Any n-uplet of value $(x_1^*, ..., x_n^*)$ satisfying the constraints is a *feasible solution* . If a problem has feasible solutions, it is said feasible.

The function z is called *objective function*. The variables $x_1, ..., x_n$ are called *decision variables* or also *choice variable* or *main variable* or *initial*

variables; the variables $x_{n+1}, ..., x_{n+m}$ are called the *slack variables*. A solution $x_1^*, x_2^*, ..., x_{n+m}^*$ is feasible if and only if all its values are positive or zero; in other words: for $k \in \{1, 2, ..., n+m\}, x_k^* \geqslant 0$. A feasible solution that maximizes the objective function is called *optimal solution*. If a linear programming problem does not admit any feasible solution, it is said to be *infeasible* or *not feasible*. If a problem admits feasible solutions and the objective function can take arbitrarily large values, it is said to be feasible *unbounded* . So there are three types of problems:

- feasible and unbounded problems,
- feasible and bounded problems,
- infeasible problems.

A *dictionary* is a system of linear equations involving $x_1, ..., x_n, x_{n+1}, ..., x_{n+m}$ and z, and satisfying the following two properties:

- the equations constituting a dictionary must express in a unique way z and m of the $n+m$ variables $x_1, ..., x_{n+m}$ according to the n other variables and this, uniquely;
- the dictionary is equivalent to the dictionary defining the slack variables and the objective function, that is to say to the dictionary:

$$\begin{cases} x_{n+1} = b_1 - \sum_{j=1}^{n} a_{1j} x_j \\ ... \\ x_{n+i} = b_i - \sum_{j=1}^{n} a_{ij} x_j \\ ... \\ x_{n+m} = b_m - \sum_{j=1}^{n} a_{mj} x_j \\ \hline z = \sum_{j=1}^{n} c_j x_j \end{cases}$$

Definition 1. *A basis consists of m variables (basic variables or in basis variables) which can be written, uniquely and linearly, using the n other variables (non-basic variables), this expression being equivalent to the initial m equality constraints.*

A basis therefore defines a dictionary and vice versa. A basis being fixed, we obtain the *basic solution* associated with this basis or, which is the same

think, with the dictionary associated with this basis, by assigning the value 0 to all the non-basic variables. Geometrically, a solution that is both basic and feasible corresponds in fact to a vertex of the constraints polyhedron.

The aim of the simplex algorithm is to determine an optimal solution among the basic and feasible solutions (that is, among the vertices of the constraints polyhedron).

5.4 Summary of an iteration

To determine such an optimal feasible basic solution, we describe an iteration of the simplex algorithm in general. For this, we define two subsets of indices, J and I: J is the indices of the n non-basic variables in the current dictionary and I is the indices of the m basic variables. More precisely:

- $J \subset \{1, 2, ..., n+m\}$ with $|J| = n$ (initially, we pose $J = \{1, 2, ..., n\}$);
- $I = \{1, 2, ..., n+m\} \setminus J$;
- The current dictionary is described by the following equalities: for $i \in I$, $x_i = b'_i + \sum_{j \in J} a'_{ij} x_j$ and $z = z^* + \sum_{j \in J} c'_j x_j$; we suppose that the dictionary is feasible: for $i \in I$, $b'_i \geqslant 0$.

The current iteration is as follows:

- if all the coefficients c'_j are negative or null, the algorithm is finished: giving the value 0 to the non-basic variables, we obtain an optimal solution;
- otherwise:
  - ⋆ we choose a non-basic variable x_{j_0} with a strictly positive coefficient in z; it is the entering variable ; if there are several candidate variables to enter in basis, one can for example privilege the variable having the highest coefficient in z (first criterion of Danzig) or, what is generally more effective, privilege the variable giving the highest increase of z (second criterion of Danzig); we will see another choice in the next paragraph, in case of "degeneracy";

- ⋆ the leaving variable x_{i_0} is computed as the basic variable that most restricts the increase of x_{j_0}; for this, we consider, for $i \in I$ with $a'_{ij_0} < 0$, the ratios $\frac{-b'_i}{a'_{ij_0}}$: i_0 is the index for which this ratio is the smallest (if there are several candidate variables to get out of the basis, we can choose an arbitrary one, we will see a systematic choice in the following paragraph, again in case of "degeneracy");
- ⋆ we extract x_{j_0} from the current expression of x_{i_0};
- ⋆ we replace x_{j_0} by its new expression in z and in the expression of the other basic variables; the new current dictionary is obtained; we are ready to apply the following iteration.

Remark. When we go from the current dictionary to the next dictionary, we are sure that this one is feasible,by the choice of the leaving variable. In other words, we go from a feasible basic solution to another feasible basic solution. It is therefore useless to check this property when the new dictionary is obtained.

5.5 Degeneracy and cycling

Definition 2. *A feasible basic solution with one or more basic variables equal to 0 is called degenerate. A basis whose associated basic solution is degenerate is called degenerate .*

Example.

Consider the dictionary (not degenerate):

$$\begin{array}{rcrcrcrcr} x_4 & = & 1 & & & & & - & 2x_3 \\ x_5 & = & 3 & - & 2x_1 & + & 4x_2 & - & 6x_3 \\ x_6 & = & 2 & + & x_1 & - & 3x_2 & - & 4x_3 \\ \hline z & = & & & 2x_1 & - & x_2 & + & 8x_3 \end{array}$$

Choosing x_3 as entering variable, we see that the inequalities $x_4 \geqslant 0$, $x_5 \geqslant 0, x_6 \geqslant 0$ all three implie: $x_3 \leqslant 0.5$. Each of the three variables x_4, x_5, x_6 is therefore a candidate to leave the basis. If we choose x_4, we obtain as new dictionary:

$$\begin{array}{rcrcrcrcr}
x_3 & = & 0.5 & & & & & - & 0.5x_4 \\
x_5 & = & & - & 2x_1 & + & 4x_2 & + & 3x_4 \\
x_6 & = & & & x_1 & - & 3x_2 & + & 2x_4 \\
\hline
z & = & 4 & + & 2x_1 & - & x_2 & - & 4x_4
\end{array}$$

In the basic solution associated with this dictionary, x_5 and x_6 have the value zero. Due to the nullity of at least one of the basic variables, this basic solution is degenerate.

If we do an iteration from this dictionary, we see that, putting x_1 into basis (the only variable to have a positive coefficient in z), the inequality $x_5 \geqslant 0$ leads to $x_1 \leqslant 0$. The largest value attributable to x_1 is 0 and the value z^* will not increase during this iteration.

The disadvantage of these inevitable degenerate iterations is that they can induce a disastrous phenomenon for the convergence of the algorithm: cycling. We say that there is *cycling* when, after a finite number of iterations, we find a basis already met. In fact, because of the independance of the non-basic variables, as soon as we find the same partition of the $m+n$ variables in basic variables and non-basic variables, the dictionaries are the same (this situation is illustrated by the exercise 5.8.5).

Remark. Consider an iteration going from a dictionary D_1 to another dictionary D_2 with an entering variable x. We assume that the value of the function z in the basic solution associated to D_2 is equal to its value in the basic solution associated to D_1. This is only possible if x is null in the basic solutions associated to D_1 and D_2. Conséquently, no value of the variables changes during this iteration. If, during a sequence of dictionaries, the value of the function z does not increase, no variable changes; geometrically, we remain in the same vertex of the polyhedron, the edges that we are trying to follow are in fact of length zero.

Cycling can always be avoided by applying the rule of the lowest index (*Bland rule*[5]): when we have a choice on the entering variable or the leaving variable, we always choose the one of the lowest index. We will prove the correctness of this rule.

[5]R. G. Bland (1977), New finite pivoting rules for the simplex method, Mathematics of Operations Research, 2, 103-107.

Theorem 2 (Bland theorem)**.** *There is no cycling when, at any iteration performed from a degenerate dictionary, the entering and leaving variables are chosen as those of the lowest index among the candidate variables.*

Proof. Suppose that, applying the Bland rule, we find twice the same D_0 dictionary after a series of iterations having built the dictionaries D_0, D_1, ..., $D_k = D_0$; all these dictionaries are necessarily degenerate. We call *versatile variable* a variable that, during these iterations, is sometimes basic, sometimes non-basic (note that there are necessarily versatile variables when there is cycling); let t be the highest index of versatile variables. In the dictionaries suite D_0, D_1, ..., D_k, D_1, ..., D_k, it necessarily exists a dictionary D' in which x_t is leaving (that is, it is basic in D' and not in the following dictionary), and then a dictionary D'' where x_t is entering; let x_s be the variable that goes into basis when, starting from D', x_t leaves (x_s is not basis in D' but is in the following dictionary); x_s is versatile and so we have $s < t$.

In the dictionaries suite D_0, D_1, ..., D_k, D_1, ..., D_k, it necessarily exists a dictionary D' in which x_t is leaving (that is, it is basic in D' and not in the following dictionary), then a dictionary D'' where x_t is entering; let x_s be the variable that comes into basis when, starting from D', x_t leave (x_s is not basic in D' but is basic in the following dictionary); x_s is versatile and so we have $s < t$.

Denoting by I the set of indexes of the basic variables of D', we can write D' as:

$$\frac{\text{for } i \in I, x_i = b'_i - \sum_{j \notin I} a'_{ij} x_j}{z = z^* + \sum_{j \notin I} c'_j x_j}$$

Since the variable x_s is entering, we have $c'_s > 0$ and, since the Bland rule is used, we have, for $J \in J$ verifying $j < s$, $c'_j \leqslant 0$. Since the variable x_t is leaving in D', it comes $a'_{ts} > 0$.

The last line of D'' can be written as:

$$z = z^* + \sum_{k=1}^{n+m} c''_k x_k$$

where c''_k is zero if x_k is basic and $c''_t > 0$.

For any solution $(x^*_1, ..., x^*_{n+m})$ of the constraints system, we have, since

the value of z does not change during the cycle:

$$z^* + \sum_{j \notin I} c'_j x^*_j = z^* + \sum_{k=1}^{n+m} c''_k x^*_k.$$

If we define a particular solution of the constraints system by giving a null value to all non-basic variables in D' except x_s and any value x^*_s to x_s (the values of the other variables are then fully determined), the above equality becomes:

$$c'_s x^*_s = c''_s x^*_s + \sum_{i \in I} c''_i (b'_i - a'_{is} x^*_s)$$

or still:

$$\left(c'_s - c''_s + \sum_{i \in I} c''_i a'_{is} \right) x^*_s = \sum_{i \in I} c''_i b'_i.$$

This equality being true for any value x^*_s, it comes:

$$c'_s - c''_s + \sum_{i \in I} c''_i a'_{is} = 0.$$

Since it is x_t that is entering into D'' and not x_s while we have $s < t$, it is that we have $c''_s \leqslant 0$. As we have noticed the inequality $c'_s > 0$, there is an index r of I with $c''_r a'_{rs} < 0$.

By definition of r, the variable x_r was basic in D' and since c''_r is non-zero, it is not basic in D''. We deduce that x_r is a versatile variable, hence the inequality $r \leqslant t$.

Therefore, c''_t and a'_{ts} being positive, their product is also positive and r can not be equal to t, hence $r < t$.

Since x_t enters in the basis of D'' while we have $r < t$, x_r is not entering in D'' so we do not have $c''_r > 0$; therefore we have $a'_{rs} > 0$.

According to the remark above, all versatile variables keep the value zero during cycling. The variable x_r being versatile, it is equal to zero in the basic solution associated with D'. Consequently, we have $b'_r = 0$.

The variable x_r was therefore a candidate to leave the basis of D' as x_t and by choosing x_t, with $t > r$, we did not apply the rule of the lowest index, a contradiction. ◊

Remark. It is useless to apply Bland rule when the dictionary is not degenerate.

5.6 Complexity of the simplex algorithm

The complexity of an iteration comes mainly from updating the coefficients describing the dictionary. More precisely:

- check whether or not the last dictionary has been reached is $O(n)$;
- the determination of an entering variable (if any) is:
  - $\star$ in $O(n)$ if we apply the first criterion of Danzig;
  - $\star$ in $O(nm)$ if we apply the second criterion of Danzig;
  - $\star$ in $O(n)$ if we apply Bland rule;
- then the determination of the leaving variable is done in $O(m)$;
- finally, the computation of the coefficients of the new dictionary is done in $O(nm)$.

The complexity of an iteration is therefore in $O(nm)$. Now, Bland theorem shows that the number of iterations is bounded from above by the number of possible dictionaries. Since a dictionary is defined by a bipartition of the $n+m$ variables in n non-basic variables and m basic variables, the number of dictionaries is bounded from above by $\begin{pmatrix} n+m \\ n \end{pmatrix} = \begin{pmatrix} n+m \\ m \end{pmatrix}$. The complexity of the simplex algorithm can therefore be bounded from above by a function in $O\left(nm \begin{pmatrix} n+m \\ n \end{pmatrix}\right)$. We can see that this complexity can not be bounded from above by a polynomial in n and m (more in-depth studies can reduce this upper bound, but without getting an upper bound polynomial in n and m; for any dimension, V. Klee and G. Minty [6] have created instances, whose polyhedron is called *Klee-Minty cube* , for which the simplex algorithm has exponential complexity).

5.7 Search of a feasible dictionary

Here again we will use an example.

[6] Klee V, GJ Minty (1972) How good is the simplex algorithm? in O. Shisha, Inequalities III, Academic Press, New York -London, pp. 159-175.

Suppose we want to solve the following problem, written in standard form.

$$\begin{array}{l}\text{Maximize } z = x_1 - x_2 + x_3\\ \text{with the constraints:}\\ \left\{\begin{array}{l} 2x_1 - x_2 + 2x_3 \leqslant 4\\ 2x_1 - 3x_2 + x_3 \leqslant -5\\ -x_1 + x_2 - 2x_3 \leqslant -1\\ x_1 \geqslant 0, x_2 \geqslant 0, x_3 \geqslant 0.\end{array}\right.\end{array}$$

We introduce the following *auxiliary problem* (which we also write in standard form).

$$\begin{array}{l}\text{Maximize } w = -x_0\\ \text{with the constraints:}\\ \left\{\begin{array}{l} 2x_1 - x_2 + 2x_3 - x_0 \leqslant 4\\ 2x_1 - 3x_2 + x_3 - x_0 \leqslant -5\\ -x_1 + x_2 - 2x_3 - x_0 \leqslant -1\\ x_0 \geqslant 0, x_1 \geqslant 0, x_2 \geqslant 0, x_3 \geqslant 0.\end{array}\right.\end{array}$$

In a more general way, we obtain the auxiliary problem by adding x_0 to the b_i. This can be interpreted by considering that the resources are increased by x_0. It is obvious that if x_0 is big enough, the new resources become all positive or null. On the other hand, if the initial problem admits a feasible solution, we can take $x_0 = 0$. The question is to determine the smallest value to assign to x_0 for the problem to be feasible. We are thus led to minimize x_0, or to maximize $-x_0$.

Remark. It is possible to remove x_0 from the first members of the inequalities corresponding to a negative value of the second members, as it is done in the solution of the exercise 5.8.6.

The auxiliary problem admits feasible solutions since the solution $x_0^* = 5$, $x_1^* = x_2^* = x_3^* = 0$ is one. It is easy to see that the initial problem admits a feasible solution if and only if the auxiliary problem admits 0 for optimal value of the objective function. Moreover, if the auxiliary problem admits 0 as the optimal value, any optimal solution of the auxiliary problem gives a feasible solution of the initial problem, in "forgetting" x_0 (which is equal to 0 in this case).

We go back to the example and write the dictionary defining the slack variables of the auxiliary problem:

$$\begin{array}{rcrcrcrcrcr}
x_4 &=& 4 &-& 2x_1 &+& x_2 &-& 2x_3 &+& x_0 \\
x_5 &=& -5 &-& 2x_1 &+& 3x_2 &-& x_3 &+& x_0 \\
x_6 &=& -1 &+& x_1 &-& x_2 &+& 2x_3 &+& x_0 \\
\hline
w &=& & & & & & & & - & x_0
\end{array}$$

This dictionary is not feasible since by giving to the non-basic variables x_1, x_2, x_3, x_0 the value 0, the slack variables x_5 and x_6 take negative values. However, it can be transformed in a feasible dictionary in one iteration. Simply, we enter x_0 into the basis and choose for the leaving variable the variable that is "most negative" (here x_5).

We obtain:

$$\begin{array}{rcrcrcrcrcr}
x_0 &=& 5 &+& 2x_1 &-& 3x_2 &+& x_3 &+& x_5 \\
x_4 &=& 9 & & & - & 2x_2 &-& x_3 &+& x_5 \\
x_6 &=& 4 &+& 3x_1 &-& 4x_2 &+& 3x_3 &+& x_5 \\
\hline
w &=& -5 &-& 2x_1 &+& 3x_2 &-& x_3 &-& x_5
\end{array}$$

In this dictionary, x_2 is an entering variable. Determine the leaving variable:

$x_0 \geqslant 0$ gives $x_2 \leqslant \dfrac{5}{3}$;

$x_4 \geqslant 0$ gives in $x_2 \leqslant \dfrac{9}{2}$;

$x_6 \geqslant 0$ gives in $x_2 \leqslant 1$.

The variable x_6 leaves the basis. The following dictionary is then:

$$\begin{array}{rcrcrcrcrcr}
x_2 &=& 1 &+& 0.75x_1 &+& 0.75x_3 &+& 0.25x_5 &-& 0.25x_6 \\
x_0 &=& 2 &-& 0.25x_1 &-& 1.25x_3 &+& 0.25x_5 &+& 0.75x_6 \\
x_4 &=& 7 &-& 1.5x_1 &-& 2.5x_3 &+& 0.5x_5 &+& 0.5x_6 \\
\hline
w &=& -2 &+& 0.25x_1 &+& 1.25x_3 &-& 0.25x_5 &-& 0.75x_6
\end{array}$$

In the next step, let x_3 enter in basis: x_0 leaves to give the last dictionary of the auxiliary problem.

$$\begin{array}{rcrcrcrcrcr}
x_3 &=& 1{,}6 &-& 0{,}2x_1 &+& 0{,}2x_5 &+& 0{,}6x_6 &-& 0{,}8x_0 \\
x_2 &=& 2{,}2 &+& 0{,}6x_1 &+& 0{,}4x_5 &+& 0{,}2x_6 &-& 0{,}6x_0 \\
x_4 &=& 3 &-& x_1 & & & - & x_6 &+& 2x_0 \\
\hline
w &=& & & & & & & & - & x_0
\end{array}$$

We see that the initial problem has a feasible solution given by $x_1^* = 0$; $x_2^* = 2{,}2$; $x_3^* = 1,\ 6$. As mentioned above, because of the equivalence of the dictionaries, we deduce a feasible dictionary for the initial problem in "forgetting" x_0 and choosing as basic variables x_3, x_2, x_4 expressed above with x_1, x_5, x_6. It remains to express z with the same variables. Then we have as dictionary for z:

$$\begin{array}{rcrcrcrcr}
x_3 & = & 1{,}6 & - & 0{,}2x_1 & + & 0{,}2x_5 & + & 0{,}6x_6 \\
x_2 & = & 2{,}2 & + & 0{,}6x_1 & + & 0{,}4x_5 & + & 0{,}2x_6 \\
x_4 & = & 3 & - & x_1 & & & - & x_6 \\
\hline
z & = & -0{,}6 & + & 0{,}2x_1 & - & 0{,}2x_5 & + & 0{,}4x_6
\end{array}$$

We can now start from this dictionary, feasible, to determine the maximum of z by applying the simplex algorithm again. This method is known as *two-phases method*. In chapter **??**, we will see that for some problems where the origin is not feasible (because some of the b_i are negative) and where all the coefficients c_j are negative, we can use the problem called "dual", which makes it possible to solve only one problem instead of two. Such a problem is said to be *dual-feasible*.

5.8 Exercices

5.8.1 Exercice 1

Statement. Solve the following problem by the simplex method:

Maximize $z = 3x_1 + 2x_2 + 4x_3$
with constraints:

$$\begin{cases}
\begin{array}{rcrcrcl}
x_1 & + & x_2 & + & 2x_3 & \leqslant & 4 \\
2x_1 & & & + & 3x_3 & \leqslant & 5 \\
2x_1 & + & x_2 & + & 3x_3 & \leqslant & 7
\end{array} \\
x_1 \geqslant 0, x_2 \geqslant 0, x_3 \geqslant 0.
\end{cases}$$

Solution. Let us introduce the slack variables of the problem. We obtain as first dictionary:

$$\begin{array}{rcrcrcrcr}
x_4 & = & 4 & - & x_1 & - & x_2 & - & 2x_3 \\
x_5 & = & 5 & - & 2x_1 & & & - & 3x_3 \\
x_6 & = & 7 & - & 2x_1 & - & x_2 & - & 3x_3 \\
\hline
z & = & & & 3x_1 & + & 2x_2 & + & 4x_3
\end{array}$$

Since each of the three non-basic variables is a candidate to enter into the basis, let us look for the one whose increase from 0 increases the most the value of the objective function, which is currently 0 (second Danzig criterion). If x_1 goes into basis, as its increase is bounded by $5/2$, the objective function increases by $15/2$. If x_2 goes into basis, the objective function increases by 8. Finally, if it is x_3, the objective function increases by $20/3$. We therefore choose to enter the variable x_2. The basic variable x_4, which is the basic variable which constrains the most the increase of x_2, leaves the basis. We obtain the new dictionary:

$$\begin{array}{rcrcrcrcr} x_2 & = & 4 & - & x_1 & - & 2x_3 & - & x_4 \\ x_5 & = & 5 & - & 2x_1 & - & 3x_3 & & \\ x_6 & = & 3 & - & x_1 & - & x_3 & + & x_4 \\ \hline z & = & 8 & + & x_1 & & & - & 2x_4 \end{array}$$

Now, we no longer have the choice of the entering variable, since only x_1 has a positive coefficient in z, and x_5 leaves the basis. The new dictionary is the following:

$$\begin{array}{rcrcrcrcr} x_1 & = & \frac{5}{2} & - & \frac{3}{2}x_3 & & & - & \frac{1}{2}x_5 \\ x_2 & = & \frac{3}{2} & - & \frac{1}{2}x_3 & - & x_4 & + & \frac{1}{2}x_5 \\ x_6 & = & \frac{1}{2} & + & \frac{1}{2}x_3 & + & x_4 & + & \frac{1}{2}x_5 \\ \hline z & = & \frac{21}{2} & - & \frac{3}{2}x_3 & - & 2x_4 & - & \frac{1}{2}x_5 \end{array}$$

This dictionary is the last since there is no non-basic variable whose coefficient in z is strictly positive. The maximum of z is therefore $21/2$ and is obtained for the following values of the variables:

$$x_1^* = \frac{5}{2};\ x_2^* = \frac{3}{2};\ x_3^* = 0.$$

5.8.2 Exercice 2

Statement. Solve the following problem by the simplex method:
Q1. by entering in basis the variable of greatest coefficient in the objective function (first criterion of Danzig);

Q2. by entering in basis the variable whose increase, starting from 0, will increase more the objective function (second criterion of Danzig).

$$\begin{array}{l}\text{Maximize } z = 5x_1 + 6x_2 + 9x_3 + 8x_4\\ \text{with the constraints:}\\ \left\{\begin{array}{l} x_1 + 2x_2 + 3x_3 + x_4 \leqslant 5\\ x_1 + x_2 + 2x_3 + 3x_4 \leqslant 3\\ x_1 \geqslant 0, x_2 \geqslant 0, x_3 \geqslant 0, x_4 \geqslant 0.\end{array}\right.\end{array}$$

Solution. Let us introduce the slack variables of the problem. We obtain as first dictionary:

$$\begin{array}{rcrcrcrcrcr} x_5 &=& 5 &-& x_1 &-& 2x_2 &-& 3x_3 &-& x_4\\ x_6 &=& 3 &-& x_1 &-& x_2 &-& 2x_3 &-& 3x_4\\ \hline z &=& & & 5x_1 &+& 6x_2 &+& 9x_3 &+& 8x_4 \end{array}$$

Q1. According to the criterion retained here to enter a variable in basis, it is first the variable x_3 which enters in basis. The leaving variable is x_6. The new dictionary is as follows:

$$\begin{array}{rcrcrcrcrcr} x_3 &=& 1.5 &-& 0.5x_1 &-& 0.5x_2 &-& 1.5x_4 &-& 0.5x_6\\ x_5 &=& 0.5 &+& 0.5x_1 &-& 0.5x_2 &+& 3.5x_4 &+& 1.5x_6\\ \hline z &=& 13.5 &+& 0.5x_1 &+& 1.5x_2 &-& 5.5x_4 &-& 4.5x_6 \end{array}$$

If we still choose the entering variable of greateST coefficient, it is x_2. The leaving variable is then x_5. We obtain the dictionary below:

$$\begin{array}{rcrcrcrcrcr} x_2 &=& 1 &+& x_1 &+& 7x_4 &-& 2x_5 &+& 3x_6\\ x_3 &=& 1 &-& x_1 &-& 5x_4 &+& x_5 &-& 2x_6\\ \hline z &=& 15 &+& 2x_1 &+& 5x_4 &-& 3x_5 & & \end{array}$$

The variable x_4 now enters in basis and the x_3 variable leaves. We have now:

$$\begin{array}{rcrcrcrcrcr} x_4 &=& 0.2 &-& 0.2x_1 &-& 0.2x_3 &+& 0.2x_5 &-& 0.4x_6\\ x_2 &=& 2.4 &-& 0.4x_1 &-& 1.4x_3 &-& 0.6x_5 &+& 0.2x_6\\ \hline z &=& 16 &+& x_1 &-& x_3 &-& 2x_5 &-& 2x_6 \end{array}$$

Finally, the variable x_1 comes in basis and x_4 leaves. The last dictionary is:

$$\begin{array}{rcrcrcrcrcr} x_1 & = & 1 & - & x_3 & - & 5x_4 & + & x_5 & - & 2x_6 \\ x_2 & = & 2 & - & x_3 & + & 2x_4 & - & x_5 & + & x_6 \\ \hline z & = & 17 & - & 2x_3 & - & 5x_4 & - & x_5 & - & 4x_6 \end{array}$$

All the coefficients of z are negative or null: the basis $\{x_1, x_2\}$ is therefore optimal, with $x_1^* = 1$ and $x_2^* = 2$.

Q2. Consider now, using the table below, the four possibilities for choosing the entering variable:

entering variable	x_1	x_2	x_3	x_4
maximum increase of the variable	3	2.5	1.5	1
corresponding increase of z	15	15	13.5	8

The current criterion leads to choose x_1 or x_2. For example, let us enter x_1 (the conclusion will remain the same if we choose x_2 here); it is then the variable x_6 that leaves; the new dictionary is:

$$\begin{array}{rcrcrcrcrcr} x_1 & = & 3 & - & x_2 & - & 2x_3 & - & 3x_4 & - & x_6 \\ x_5 & = & 2 & - & x_2 & - & x_3 & + & 2x_4 & + & x_6 \\ \hline z & = & 15 & + & x_2 & - & x_3 & - & 7x_4 & - & 5x_6 \end{array}$$

Only the variable x_2 is candidate to enter in basis, the variable x_5 leaves; we obtain the same dictionary as above with the same conclusion.

5.8.3 Exercice 3

Statement. Give an example of a unbounded linear programming problem, written in standard form.

Solution. Consider the problem:

$$\text{Maximize } z = x_1 + x_2$$

$$\text{with the constraints: } \begin{cases} x_1 - x_2 \leqslant 1 \\ 2x_1 - 3x_2 \leqslant 2 \\ x_1 \geqslant 0, x_2 \geqslant 0 \end{cases}$$

Any solution of the form $x_1 = 0, x_2 = t \geqslant 0$ is feasible and for these values we have $z = t$. Since t is not bounded, the problem is not bounded.

5.8.4 Exercice 4

Statement. The same question as for exercise 3, but here we want an infeasible problem.

Solution. Consider the problem:

$$\text{Maximize } z = x_1 + x_2$$
$$\text{with the constraints: } \begin{cases} -2x_1 + 3x_2 \leqslant -4 \\ x_1 - x_2 \leqslant 1 \\ x_1 \geqslant 0, x_2 \geqslant 0. \end{cases}$$

Suppose the second constraint is satisfied, then we have$-x_1 + x_2 \geqslant -1$, whence: $-3x_1 + 3x_2 \geqslant -3$.

Since x_1 is positive, we get: $-2x_1 + 3x_2 \geqslant -3x_1 + 3x_2 \geqslant -3$, a contradiction with the first constraint. The problem is therefore infeasible.

5.8.5 Exercice 5

Statement. We want to apply the simplex algorithm to the dictionary below. Two strategies are considered when there are several variables candidate for entering or leaving the basis.

$$\begin{array}{rcrrrrrrrrr} x_5 & = & & - & 0.5x_1 & + & 5.5x_2 & + & 2.5x_3 & - & 9x_4 \\ x_6 & = & & - & 0.5x_1 & + & 1.5x_2 & + & 0.5x_3 & - & x_4 \\ x_7 & = & 1 & - & x_1 & & & & & & \\ \hline z & = & & & 10x_1 & - & 57x_2 & - & 9x_3 & - & 24x_4 \end{array}$$

Q1. In the case of a choice for an entering variable, we take the candidate variable with the highest coefficient in z (the first Danzig criterion) and, in the case of a choice for a leaving variable, we take the candidate variable of smallest index. What do we observe?

Q2. We apply the Bland rule: in case of choice for an entering or leaving variable, we take the candidate variable of smallest index. What do we observe?

Solution.
Q1. We start from the given dictionary.
We choose x_1 as entering variable and x_5 as leaving variable. After the first iteration:

$$\begin{array}{rcrrrrrrrrr}
x_1 & = & & & 11x_2 & + & 5x_3 & - & 18x_4 & - & 2x_5 \\
x_6 & = & & - & 4x_2 & - & 2x_3 & + & 8x_4 & + & x_5 \\
x_7 & = & 1 & - & 11x_2 & - & 5x_3 & + & 18x_4 & + & 2x_5 \\
\hline
z & = & & & 53x_2 & + & 41x_3 & - & 204x_4 & - & 20x_5
\end{array}$$

We choose x_2 as entering variable and x_6 as leaving variable. After the second iteration:

$$\begin{array}{rcrrrrrrrrr}
x_2 & = & & - & 0.5x_3 & + & 2x_4 & + & 0.25x_5 & - & 0.25x_6 \\
x_1 & = & & - & 0.5x_3 & + & 4x_4 & + & 0.75x_5 & - & 2.75x_6 \\
x_7 & = & 1 & + & 0.5x_3 & - & 4x_4 & - & 0.75x_5 & - & 2.75x_6 \\
\hline
z & = & & & 14.5x_3 & - & 98x_4 & - & 6.75x_5 & - & 13.25x_6
\end{array}$$

The variable x_3 is entering and x_1 is leaving. After the third iteration:

$$\begin{array}{rcrrrrrrrrr}
x_3 & = & & - & 2x_1 & + & 8x_4 & + & 1.5x_5 & - & 5.5x_6 \\
x_2 & = & & & x_1 & - & 2x_4 & - & 0.5x_5 & + & 2.5x_6 \\
x_7 & = & 1 & - & x_1 & & & & & & \\
\hline
z & = & & - & 29x_1 & + & 18x_4 & + & 15x_5 & - & 93x_6
\end{array}$$

We choose x_4 as entering variable and x_2 as leaving variable. After the fourth iteration:

$$\begin{array}{rcrrrrrrrrr}
x_4 & = & & & 0.5x_1 & - & 0.5x_2 & - & 0.25x_5 & + & 1.25x_6 \\
x_3 & = & & & 2x_1 & - & 4x_2 & - & 0.5x_5 & + & 4.5x_6 \\
x_7 & = & 1 & - & x_1 & & & & & & \\
\hline
z & = & & - & 20x_1 & - & 9x_2 & + & 10.5x_5 & - & 70.5x_6
\end{array}$$

We choose x_5 as entering variable and x_3 as leaving variable. After the fifth iteration:

$$\begin{array}{rcrrrrrrrrr}
x_5 & = & & & 4x_1 & - & 8x_2 & - & 2x_3 & + & 9x_6 \\
x_4 & = & & - & 0.5x_1 & + & 1.5x_2 & + & 0.5x_3 & - & x_6 \\
x_7 & = & 1 & - & x_1 & & & & & & \\
\hline
z & = & & & 22x_1 & - & 93x_2 & - & 21x_3 & + & 24x_6
\end{array}$$

We choose x_6 as entering variable and x_4 as leaving variable. After the sixth iteration:

$$\begin{array}{rcrrrrrrrrr}
x_5 & = & & - & 0.5x_1 & + & 5.5x_2 & + & 2.5x_3 & - & 9x_4 \\
x_6 & = & & - & 0.5x_1 & + & 1.5x_2 & + & 0.5x_3 & - & x_4 \\
x_7 & = & 1 & - & x_1 & & & & & & \\
\hline
z & = & & & 10x_1 & - & 57x_2 & - & 9x_3 & - & 24x_4
\end{array}$$

We find the original dictionary: we observe that there is cycling. It may be noted that the application of the second Danzig criterion instead of the first criterion for the choice of the entering variables does not avoid cycling either, since the steps that have just been performed are compatible with this criterion.

Q2. Bland rule gives the same five first iterations, but not the sixth. We repeat the previous computations after the fifth iteration:

$$\begin{array}{rcrcrcrcrcr}
x_5 & = & & & 4x_1 & - & 8x_2 & - & 2x_3 & + & 9x_6 \\
x_4 & = & & - & 0.5x_1 & + & 1.5x_2 & + & 0.5x_3 & - & x_6 \\
x_7 & = & 1 & - & x_1 & & & & & & \\
\hline
z & = & & & 22x_1 & - & 93x_2 & - & 21x_3 & + & 24x_6
\end{array}$$

We choose x_1 as entering variable (and not x_6) and x_4 as leaving variable. After the sixth iteration:

$$\begin{array}{rcrcrcrcrcr}
x_1 & = & & & 3x_2 & + & x_3 & - & 2x_4 & - & 2x_6 \\
x_5 & = & & & 4x_2 & + & 2x_3 & - & 8x_4 & + & x_6 \\
x_7 & = & 1 & - & 3x_2 & - & x_3 & + & 2x_4 & + & 2x_6 \\
\hline
z & = & & - & 27x_2 & + & x_3 & - & 44x_4 & - & 20x_6
\end{array}$$

We choose x_3 as entering variable and x_7 as leaving variable. After the seventh iteration:

$$\begin{array}{rcrcrcrcrcr}
x_3 & = & 1 & - & 3x_2 & + & 2x_4 & + & 2x_6 & - & x_7 \\
x_1 & = & 1 & & & & & & & - & x_7 \\
x_5 & = & 2 & - & 2x_2 & - & 4x_4 & + & 5x_6 & - & 2x_7 \\
\hline
z & = & 1 & - & 30x_2 & - & 42x_4 & - & 18x_6 & - & 2x_7
\end{array}$$

All coefficients in z are negative or null, the method stops. It can be seen that the application of the Bland rule has made it possible to avoid cycling.

5.8.6 Exercice 6

Statement.

Q1. We consider the problem below.

Maximize $z = 5x_1 + 3x_2$

with the constraints:
$$\begin{cases}
-4x_1 + 5x_2 \leqslant -10 \\
5x_1 + 2x_2 \leqslant 10 \\
3x_1 + 8x_2 \leqslant 12 \\
x_1 \geqslant 0, x_2 \geqslant 0.
\end{cases}$$

Show with the simplex algorithm that this problem does not admit any feasible solution.

Q2. We now consider the problem below (which differs from the preceding one only by a sign in the first constraint). Solve it using the two-phases method.

$$\text{Maximize } z = 5x_1 + 3x_2$$

$$\text{with the constraints: } \begin{cases} -4x_1 - 5x_2 \leqslant -10 \\ 5x_1 + 2x_2 \leqslant 10 \\ 3x_1 + 8x_2 \leqslant 12 \\ x_1 \geqslant 0, x_2 \geqslant 0. \end{cases}$$

Solution.

Both optimization problems in this exercise are in standard form. We find that in both cases, the solution obtained by setting the two variables x_1 and x_2 to zero is not feasible. The two-phase simplex method is used. The first phase begins with the writing of the auxiliary problem. For this, we can subtract a variable x_0 in the first three members of the inequalities, as for the example of the part 5.7; we can also subtract this variable x_0 from the first members of the inequalities with a negative second member. This is the method we choose here to illustrate it.

Q1. The auxiliary problem is then written:

$$\text{Maximize } w = -x_0$$

$$\text{with the constraints: } \begin{cases} -4x_1 + 5x_2 - x_0 \leqslant -10 \\ 5x_1 + 2x_2 \leqslant 10 \\ 3x_1 + 8x_2 \leqslant 12 \\ x_0 \geqslant 0, x_1 \geqslant 0, x_2 \geqslant 0. \end{cases}$$

We deduce the initial dictionary:

$$\begin{array}{rcrcrcrcr} x_3 & = & -10 & + & x_0 & + & 4x_1 & - & 5x_2 \\ x_4 & = & 10 & & & - & 5x_1 & - & 2x_2 \\ x_5 & = & 12 & & & - & 3x_1 & - & 8x_2 \\ \hline w & = & & - & x_0 & & & & \end{array}$$

This dictionary is not feasible, but we immediately obtain a feasible dictionary by choosing x_0 as entering variable and x_3 as leaving variable. We

obtain the dictionary below:

$$\begin{array}{rcrcrcrcr}
x_0 & = & 10 & - & 4x_1 & + & 5x_2 & + & x_3 \\
x_4 & = & 10 & - & 5x_1 & - & 2x_2 & & \\
x_5 & = & 12 & - & 3x_1 & - & 8x_2 & & \\
\hline
w & = & -10 & + & 4x_1 & - & 5x_2 & - & x_3
\end{array}$$

We choose x_1 as entering variable and x_4 as leaving variable; we obtain :

$$\begin{array}{rcrcrcrcr}
x_1 & = & 2 & - & \frac{2}{5}x_2 & & & - & \frac{1}{5}x_4 \\
x_0 & = & 2 & + & \frac{33}{5}x_2 & + & x_3 & + & \frac{4}{5}x_4 \\
x_5 & = & 6 & - & \frac{34}{5}x_2 & & & + & \frac{3}{5}x_4 \\
\hline
w & = & -2 & - & \frac{33}{5}x_2 & - & x_3 & - & \frac{4}{5}x_4
\end{array}$$

There is no more entering variable; the optimum of w is -2 and is therefore not zero: the studied problem does not admit a feasible solution.

Q2. In the same way as for the previous question, the auxiliary problem is written:

$$\text{Maximize } w = -x_0$$
$$\text{with the constraints: } \begin{cases} -4x_1 - 5x_2 - x_0 \leqslant -10 \\ 5x_1 + 2x_2 \leqslant 10 \\ 3x_1 + 8x_2 \leqslant 12 \\ x_1 \geqslant 0, x_2 \geqslant 0, x_0 \geqslant 0. \end{cases}$$

We obtain the following initial dictionary:

$$\begin{array}{rcrcrcrcr}
x_3 & = & -10 & + & x_0 & + & 4x_1 & + & 5x_2 \\
x_4 & = & 10 & & & - & 5x_1 & - & 2x_2 \\
x_5 & = & 12 & & & - & 3x_1 & - & 8x_2 \\
\hline
w & = & & - & x_0 & & & &
\end{array}$$

This dictionary is not feasible but, again, we immediately go to a feasible dictionary by choosing x_0 as entering variable and x_3 as leaving variable. We obtain the dictionary below:

$$\begin{array}{rcrcrcrcr}
x_0 & = & 10 & - & 4x_1 & - & 5x_2 & + & x_3 \\
x_4 & = & 10 & - & 5x_1 & - & 2x_2 & & \\
x_5 & = & 12 & - & 3x_1 & - & 8x_2 & & \\
\hline
w & = & -10 & + & 4x_1 & + & 5x_2 & - & x_3
\end{array}$$

We choose x_1 as entering variable and x_4 as leaving variable; we obtain :

$$\begin{array}{rcrcrcrcr}
x_1 & = & 2 & - & \frac{2}{5}x_2 & & & - & \frac{1}{5}x_4 \\
x_0 & = & 2 & - & \frac{17}{5}x_2 & + & x_3 & + & \frac{4}{5}x_4 \\
x_5 & = & 6 & - & \frac{34}{5}x_2 & & & + & \frac{3}{5}x_4 \\
\hline
w & = & -2 & + & \frac{17}{5}x_2 & - & x_3 & - & \frac{4}{5}x_4
\end{array}$$

We choose x_2 as entering variable and x_0 as leaving variable; we obtain :

$$\begin{array}{rcrcrcrcr}
x_2 & = & \frac{10}{17} & + & \frac{5}{17}x_3 & + & \frac{4}{17}x_4 & - & \frac{5}{17}x_0 \\
x_1 & = & \frac{30}{17} & - & \frac{2}{17}x_3 & - & \frac{5}{17}x_4 & + & \frac{2}{17}x_0 \\
x_5 & = & 2 & - & 2x_3 & - & x_4 & + & 2x_0 \\
\hline
w & = & & & & & & - & x_0
\end{array}$$

The optimum of the auxiliary problem is zero: the initial problem is feasible. We can now start the second phase of the method. To obtain a feasible dictionary of the initial problem, we use the last dictionary above, in which we delete the variable x_0 and we replace the function w by the function z expressed with the basic variables, that is, x_3 and x_4. We obtain the dictionary below:

$$\begin{array}{rcrcrcr}
x_2 & = & \frac{10}{17} & + & \frac{5}{17}x_3 & + & \frac{4}{17}x_4 \\
x_1 & = & \frac{30}{17} & - & \frac{2}{17}x_3 & - & \frac{5}{17}x_4 \\
x_5 & = & 2 & - & 2x_3 & - & x_4 \\
\hline
z & = & \frac{180}{17} & + & \frac{5}{17}x_3 & - & \frac{13}{17}x_4
\end{array}$$

We choose x_3 as entering variable and x_5 as leaving variable. The dictionary becomes:

$$\begin{array}{rcrcrcr} x_3 & = & 1 & - & \frac{1}{2}x_4 & - & \frac{1}{2}x_5 \\ x_2 & = & \frac{15}{17} & + & \frac{3}{34}x_4 & - & \frac{5}{34}x_5 \\ x_1 & = & \frac{28}{17} & - & \frac{4}{17}x_4 & + & \frac{1}{17}x_5 \\ \hline z & = & \frac{185}{17} & - & \frac{31}{34}x_4 & - & \frac{5}{34}x_5 \end{array}$$

This last dictionary is optimal; the optimal solution is therefore given by:

$x_1^* = \frac{28}{17}, x_2^* = \frac{15}{17}$ for the decision variables;

$x_3^* = 1, x_4^* = x_5^* = 0$ for the slack variables;

$z^* = \frac{185}{17}$ for the objective function.

5.8.7 Exercice 7

Statement. We consider a linear programming problem with a single constraint, defined by:

$$\text{Maximize } \sum_{j=1}^{n} u_j x_j \text{ with } \sum_{j=1}^{n} p_j x_j \leqslant P \text{ and } x_j \geqslant 0 \text{ for } 1 \leqslant j \leqslant n.$$

All the coefficients u_j and p_j as well as P are assumed to be strictly positive and we assume the variables ranked according to the decreasing values of the ratios u_j/p_j. Show that the variable x_1 is entering and that, by entering it in basis, we reaches the optimum of the objective in a single step. Express the optimum value of the objective according to the different coefficients.

Solution. Since the coefficient of the variable x_1 is, by hypothesis, positive, the variable x_1 is entering and we therefore exchange it with the unique variable in basis, which corresponds to the unique constraint, x_{n+1}. We had:

$$x_{n+1} = P - \sum_{j=1}^{n} p_j x_j$$

and after the exchange, we get:

$$x_1 = \frac{1}{p_1}\left(P - \sum_{j=2}^{n} p_j x_j - x_{n+1}\right).$$

By putting this value in the objective function, it comes:

$$z = \sum_{j=1}^{n} u_j x_j = \frac{u_1}{p_1}\left(P - \sum_{j=2}^{n} p_j x_j - x_{n+1}\right) + \sum_{j=2}^{n} u_j x_j$$

or:

$$z = \frac{u_1 P}{p_1} + \sum_{j=2}^{n}\left(u_j - \frac{u_1 p_j}{p_1}\right) x_j - \frac{u_1}{p_1} x_{n+1}.$$

Considering the numbering adopted, the coefficients of all variables that occur in the writing of z are negative or null. We thus found the maximum value of z in one iteration, and this one is equal to $\dfrac{u_1 P}{v_1}$: this corresponds to saturate the constraint with the variable for which the ratio u_j/p_j is maximum.

Chapter 6

Duality in linear programming

6.1 Definition of the dual problem

Remark

We only consider in this chapter **linear programming problems written in standard form**. To define the dual problem of any linear programming problem, we can put it in standard form before determining the dual problem, as indicated in the chapter 5. An example is given in exercise.

We therefore consider the problem (P):

$$\text{Maximize } z = \sum_{j=1}^{n} c_j x_j$$
$$\text{with the constraints } \begin{cases} \text{for } i \in \{1, 2, ..., m\}, \sum_{j=1}^{n} a_{ij} x_j \leqslant b_i \\ \text{for } j \in \{1, 2, ..., n\}, x_j \geqslant 0 \end{cases}$$

If there are m real y_i positive or zero such as, for any $j \in \{1, 2, ..., n\}$, $\sum_{i=1}^{m} a_{ij} y_i \geqslant c_j$, then we have, for any feasible solution $(x_1, ..., x_n)$ of (P):

$$\sum_{j=1}^{n} c_j x_j \leqslant \sum_{j=1}^{n} \left(\sum_{i=1}^{m} a_{ij} y_i \right) x_j = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} x_j \right) y_i \leqslant \sum_{i=1}^{m} b_i y_i.$$

Hence:

$$\sum_{j=1}^{n} c_j x_j \leqslant \sum_{i=1}^{m} b_i y_i$$

and this last quantity thus gives an upper bound of the objective function. The *dual problem* (D) of the problem (P) is:

$$\text{Minimize } \sum_{i=1}^{m} b_i y_i$$
$$\text{with the constraints } \begin{cases} \text{for } j \in \{1, 2, ..., n\}, \sum_{i=1}^{m} a_{ij} y_i \geqslant c_j \\ \text{for } i \in \{1, 2, ..., m\}, y_i \geqslant 0 \end{cases}$$

The problem (P) then takes the name of *primal problem*. We see that, for every feasible solution $y_1^*, ..., y_m^*$ of the dual problem (that is, satisfying the constraints of (D)), $\sum_{i=1}^{m} b_i y_i^*$ is an upper bound of the objective function of the primal problem.

Remark
It is easily established that the dual problem of (D) is (P).

6.2 Theorem of duality

From the definition of the dual problem, we immediately deduce the following proposition:

Proposition 3. *Let $(x_1^*, x_2^*, ..., x_n^*)$ be a feasible solution of the primal problem and $(y_1^*, y_2^*, ..., y_m^*)$ be a feasible solution of the dual problem. We have:*

$$\sum_{j=1}^{n} c_j x_j^* \leqslant \sum_{i=1}^{m} b_i y_i^*$$

Moreover, if the two above quantities are equal, then $x_1^*, x_2^*, ..., x_n^*$ constitute an optimal solution of the primal problem and $y_1^*, y_2^*, ..., y_m^*$ an optimal solution of the dual problem.

Application

Due to considerations about the dual problem, it is possible to check that we have found an optimal solution for a given problem. We will explain it on the problem dealt with in the first chapter 5. The problem (P) is:

$$\text{Maximize } z = 7x_1 + 9x_2 + 18x_3 + 17x_4$$
$$\text{with the constraints}$$
$$\begin{cases} 2x_1 + 4x_2 + 5x_3 + 7x_4 \leqslant 42 \\ x_1 + x_2 + 2x_3 + 2x_4 \leqslant 17 \\ x_1 + 2x_2 + 3x_3 + 3x_4 \leqslant 24 \\ x_1 \geqslant 0, x_2 \geqslant 0, x_3 \geqslant 0, x_4 \geqslant 0 \end{cases}$$

We had established that the optimum of this problem is $z^* = 147$ obtained for $x_1^* = 3, x_2^* = 0, x_3^* = 7, x_4^* = 0$. We want here to verify this result.

The dual problem (D) is:

$$\begin{array}{l}\text{Minimize } 42y_1 + 17y_2 + 24y_3\\ \text{with the constraints}\\ \left\{\begin{array}{l} 2y_1 + y_2 + y_3 \geqslant 7\\ 4y_1 + y_2 + 2y_3 \geqslant 9\\ 5y_1 + 2y_2 + 3y_3 \geqslant 18\\ 7y_1 + 2y_2 + 3y_3 \geqslant 17\\ y_1 \geqslant 0, y_2 \geqslant 0, y_3 \geqslant 0. \end{array}\right.\end{array}$$

Recall that in the last dictionary, the objective function was written:

$$z = 147 - 2x_2 - x_4 - 3x_6 - 4x_7.$$

We consider the values $y_1^* = 0, y_2^* = 3, y_3^* = 4$. These values are not chosen at random: they are the opposite of the coefficients respectively of x_5, x_6, x_7 in the above expression of z; we will justify this choice further.

We have: $42y_1^* + 17y_2^* + 24y_3^* = 147$.

Moreover, it is easy to check that the y_i^* satisfy the constraints of the dual problem, and thus constitute a feasible solution of the dual.

The above proposition allows us to affirm that the value 147 is the optimum of the primal problem: having found a feasible solution of the dual which gives to the objective function of the dual the value found for the objective function of the primal, we can affirm that we had found the maximum of the objective function of the primal and that we have also found the minimum of the objective function of the dual. This verification therefore constitutes a *certificate of optimality* of the solution found for the primal.

This proposal has the following corollary:

Proposition 4. *If the primal problem admits a feasible solution and is unbounded, the dual problem does not admit a feasible solution.*

Proof. Assume that the dual problem admits a feasible solution and let us denote by w^* the corresponding value of the objective function of the dual problem. The objective function of the primal problem is then bounded from above by w^*, which contradicts the hypothesis. ◊

Remarks
1. If (D) is feasible, (P) is either not feasible or feasible and bounded.
2. If (P) admits a feasible solution and is unbounded, (D) does not admit a feasible solution.
3. If (D) admits a feasible solution and is unbounded, (P) does not admit a feasible solution.
4. (P) and (D) cannot be simultaneously feasible and unbounded.
5. There are cases where (P) and (D) are simultaneously not feasible (see exercise 6.6.3).

The following theorem, sometimes called *fundamental theorem of duality*, generalizes the findings made above.

Theorem 5 (of duality)**.** *If the primal problem has an optimal solution* $x_1^*, x_2^*, ..., x_n^*$*, then the dual problem has an optimal solution* $y_1^*, y_2^*, ..., y_m^*$ *and* $\sum_{j=1}^{n} c_j x_j^* = \sum_{i=1}^{m} b_i y_i^*$ *(that is, the primal maximum is equal to the dual minimum).*

We will prove this fundamental theorem at the same time as the following proposition:

Proposition 6. *If the primal problem admits an optimal solution and if the expression of the objective function of the primal in the last dictionary obtained by the simplex method is written:*

$$z = z^* + \sum_{k=1}^{n+m} d_k x_k$$

(where x_{n+i} *represents the* i^{th} *slack variable), then an optimal solution of the dual problem is given by* $y_i^* = -d_{n+i}$*.*

Proof of the theorem of duality and of the proposition

Suppose that the primal has been solved by the simplex method. To the n initial variables of the problem we have added m slack variables $x_{n+1}, ..., x_{n+m}$. At the i^{th} primal constraint are associated the slack variable x_{n+i} and the variable y_i of the dual, which establishes a canonical link between x_{n+i} and y_i. Consider the expression of the objective function of the primal in the last dictionary of the primal simplex:

$$z = z^* + \sum_{k=1}^{n+m} d_k x_k.$$

The d_k are all negative or null (since this is the last dictionary) and the d_k associated with the basic variables are null.

Moreover, by définition of z we have $z^* = \sum_{j=1}^{n} c_j x_j^*$ and by définition of slack variables $x_{n+i} = b_i - \sum_{j=1}^{n} a_{ij} x_j$.

Let us set, for $i \in \{1, ..., m\}$, $y_i^* = -d_{n+i}$; we then have: $y_i^* \geqslant 0$. Moreover, distinguishing in z the slack variables from the others:

$$z = z^* + \sum_{j=1}^{n} d_j x_j - \sum_{i=1}^{m} \left(b_i - \sum_{j=1}^{n} a_{ij} x_j \right) y_i^*,$$

or:

$$z = z^* - \sum_{i=1}^{m} b_i y_i^* + \sum_{j=1}^{n} \left(d_j + \sum_{i=1}^{m} a_{ij} y_i^* \right) x_j.$$

But, by definition of z, we also have: $z = \sum_{j=1}^{n} c_j x_j$.

Because of the independence of the variables x_j, we deduce from these equalities:

$$\begin{cases} z^* = \sum_{i=1}^{m} b_i y_i^* \\ \text{for } j \in \{1, ..., n\}, c_j = d_j + \sum_{i=1}^{m} a_{ij} y_i^* \end{cases}$$

The d_j ($j \in \{1, ..., n+m\}$) being negative or null, we finally get:

$$\begin{cases} \text{for } j \in \{1, ..., n\}, \sum_{i=1}^{m} a_{ij} y_i^* \geqslant c_j \\ \text{for } i \in \{1, ..., m\}, y_i^* \geqslant 0. \end{cases}$$

The numbers $y_1^*, y_2^*, ..., y_m^*$ thus form a feasible solution of the dual problem which gives to the objective function of the dual problem the value z^*. The proposition at the beginning of this paragraph gives the conclusion. $\Diamond$

6.3 The complementary slackness theorem: a certificate of optimality

The example developed in the preceding paragraph gives a method to demonstrate the optimality of a solution of the primal problem but requires the knowledge of the last dictionary of the simplex method. We will see that we can also succeed in providing a *certificate of optimality* of the primal, knowing only the values x_1^*, ... x_n^* that give the maximum to the objective function of the primal.

Theorem 7 (Complementary slackness theorem). *A feasible solution $x_1^*, ..., x_n^*$ of the primal problem is optimal if and only there are numbers y_1^*, ..., y_m^* which fulfill the following:*

- *for $i \in \{1, ..., m\}$, if $\sum_{j=1}^{n} a_{ij} x_j^* < b_i$, then $y_i^* = 0$*
- *for $j \in \{1, ..., n\}$, if $x_j^* > 0$, then $\sum_{i=1}^{m} a_{ij} y_i^* = c_j$*

and is a feasible solution of the dual problem:

$$\begin{cases} \text{for } j \in \{1, ..., n\}, \sum_{i=1}^{m} a_{ij} y_i^* \geqslant c_j \\ \text{for } i \in \{1, ..., m\}, y_i^* \geqslant 0. \end{cases}$$

Moreover, these numbers y_1^*, ..., y_m^* constitute an optimal solution of the dual problem.

Before giving the proof of this theorem, we will apply it to the example of chapter I. The problem was written as:

Maximize $z = 7x_1 + 9x_2 + 18x_3 + 17x_4$
with the constraints:

$$\begin{cases} 2x_1 + 4x_2 + 5x_3 + 7x_4 \leqslant 42 \\ x_1 + x_2 + 2x_3 + 2x_4 \leqslant 17 \\ x_1 + 2x_2 + 3x_3 + 3x_4 \leqslant 24 \\ x_1 \geqslant 0, x_2 \geqslant 0, x_3 \geqslant 0, x_4 \geqslant 0. \end{cases}$$

Consider the statement:
"$x_1^* = 3, x_2^* = 0, x_3^* = 7, x_4^* = 0$ constitute an optimal solution of the primal".
It is easy to verify that these values define a feasible solution of the primal problem. We look for y_1^*, y_2^*, y_3^* satisfying:

$$\begin{cases} y_1^* = 0 \text{ since the first constraint of the problem "is not saturated"} \\ 2y_1^* + y_2^* + y_3^* = 7 \text{ since } x_1^* > 0 \\ 5y_1^* + 2y_2^* + 3y_3^* = 18 \text{ since } x_3^* > 0. \end{cases}$$

Using the nullity of y_1^*, we obtain:

$$\begin{cases} y_2^* + y_3^* = 7 \\ 2y_2^* + 3y_3^* = 18. \end{cases}$$

The resolution of this system gives $y_2^* = 3$, $y_3^* = 4$. We check that these values satisfy the constraints of the dual problem.

$$\begin{aligned} & 4y_1^* + y_2^* + 2y_3^* = 11 \geqslant 9 \\ \text{and} \quad & 7y_1^* + 2y_2^* + 3y_3^* = 18 \geqslant 17. \end{aligned}$$

The other two inequalities of the same type result from the system defining y_1^*,y_2^* and y_3^*. Finally y_1^*, y_2^*, y_3^* are positive or null.

The proposed solution for the primal is therefore optimal.

The values y_1^*, y_2^*, y_3^* constitute a feasible solution of the dual problem and gives the value 147 to the objective function of this problem. As the optimum value of the primal problem is also 147, y_1^*, y_2^*, y_3^* constitute an optimal solution of the dual problem.

Proof of the complementary slackness theorem.

The proof is immediately deducted from the result which we express and prove below.

If we know a feasible solution (x_j^*) of the primal and a feasible solution (y_i^*) of the dual, these solutions are optimal if and only if:

- for $j \in \{1, ..., n\}$, $x_j^* = 0$ or $\sum_{i=1}^{m} a_{ij} y_i^* = c_j$
- and, for $i \in \{1, ..., m\}$, $y_i^* = 0$ or $\sum_{j=1}^{n} a_{ij} x_j^* = b_i$.

Indeed, according to the duality theorem, if we know a feasible solution (x_j^*) of the primal and a feasible solution (y_i^*) of the dual, these solutions are optimal if and only if we have:

$$\sum_{j=1}^{n} c_j x_j^* = \sum_{i=1}^{m} b_i y_y^*.$$

However, we have the following inequalities (consequence of the feasibility of the solutions):

$$\sum_{j=1}^{n} c_j x_j^* \leqslant \sum_{j=1}^{n} \left(\sum_{i=1}^{m} a_{ij} y_i^* \right) x_j^* = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} x_j^* \right) y_i \leqslant \sum_{i=1}^{m} b_i y_i^*.$$

If, before summation, one of the inequalities was strict, it would be the same after summation. So we see that there is equality between the limits of this sequence if and only if we have:

- for $j \in \{1, ..., n\}$, $x_j^* = 0$ or $\sum_{i=1}^{m} a_{ij} y_i^* = c_j$
- and, for $i \in \{1, ..., m\}$, $y_i^* = 0$ or $\sum_{j=1}^{n} a_{ij} x_j^* = b_i$. $\Diamond$

If we can determine the y_i^* and one of the required inequalities (including the sign constraints) is not verfied, the solution is not optimal.

6.4 The economic significance of the dual

We will show here that knowledge of the solution of the dual problem can make it possible to take into account economic data.

We will consider that:

- b_i represents the total quantity of the resource i;
- a_{ij} represents the quantity of resource i consumed by the manufacture of a unit of product j;
- x_j represents the manufactured quantity of product j;
- c_j represents the value of a unit of the product j.

The relation at optimum: $z^* = \sum_{j=1}^{n} c_j x_j^* = \sum_{i=1}^{m} b_i y_i^*$ induces that y_i^* must represent the "unit value of the resource i". These dual variables y_i^* are often called *implicit price* . The value of y_i gives the maximum amount that one would be willing to pay to obtain an additional unit of the resource i.

The inequalities $\sum_{i=1}^{m} a_{ij} y_i \geqslant c_j$, $(j \in \{1, ..., n\})$ can be understood using the following schema: suppose that someone from outside the company wants to acquire the resources of the company; it must propose for the resources a price such that it is more interesting for the company to sell its resources than to manufacture the products itself (c_j is the expected profit on the product j) and well sure it wants to make this purchase of resources at a minimum price. Recall that the coefficient a_{ij} represents the quantity of the resource i required to produce a unit of the product j so that $\sum_{i=1}^{m} a_{ij} y_i$ represents the amount to be spent to acquire the resources necessary to manufacture a unit of the product j.

We will give a second interpretation using our example.

Problem

The tissue manufacturer of the previous chapter can have a few extra hours at an hourly price of t for its dyeing workers. Does he or not have an interest in using this possibility?

To solve this problem, we will express a theorem, which we will demonstrate after solving our problem.

Theorem 8. *We consider the problem (P):*

$$\text{Maximize } z = \sum_{j=1}^{n} c_j x_j$$
$$\text{with the constraints } \begin{cases} \text{for } i \in \{1, ..., m\}, \sum_{j=1}^{n} a_{ij} x_j \leqslant b_i \\ \text{for } j \in \{1, 2, ..., n\}, x_j \geqslant 0. \end{cases}$$

It is assumed that the optimal basis of (P) is not degenerate. For variations δb_i of b_i, we consider the problem (P_δ) defined by:

$$\text{Maximize } z = \sum_{j=1}^{n} c_j x_j$$
$$\text{with the constraints } \begin{cases} \text{for } i \in \{1, ..., m\}, \sum_{j=1}^{n} a_{ij} x_j \leqslant b_i + \delta b_i \\ \text{for } j \in \{1, 2, ..., n\}, x_j \geqslant 0. \end{cases}$$

We assume that the δb_i variations are small enough that the optimal basis of (P) is still feasible for (P_δ). The variation of the optimum value of the objective function of the linear program is then $\sum_{i=1}^{m} \delta b_i y_i^*$ where $(y_1^*, ..., y_m^*)$ is an optimal solution of the dual problem of (P).

Remark. In deepening the proof of the duality theorem, one would obtain that the non-degeneracy of the optimal basis of (P) implies the uniqueness of the solution of the dual problem.

For our problem, let us call u the number of extra hours for dyeing (with u small). The variation of the second member is (0, 0, u). The optimal solution of the dual problem is (0, 3, 4). The variation of the objective function is therefore $4u$. This is not a net benefit since it will cost him $t.u$ euros. We see that he has an interest in choosing this solution as soon as $t \leqslant 4$ (he is unlikely to convince his workers to work overtime at this price!). Here we find the interpretation of y_i^*: unit value of the resource.

Proof of the theorem. We consider the sequence of dictionaries obtained when we solve (P) by the simplex method. When we change b to $b + \delta b$, only the constants of the second members are changed. If the last dictionary remains feasible (that is, if the constants of the second members of the equalities expressing the basic variables remain positive or zero), then this dictionary remains optimal. We suppose we are in this case. The coefficients of the non-basic variables in the line expressing the function z being unchanged, the solution of the dual problem is unchanged. The common optimal value of the new primal and dual problems is: $\sum_{i=1}^{m}(b_i + \delta b_i)y_i^* = \sum_{i=1}^{m} b_i y_i^* + \sum_{i=1}^{m} \delta b_i y_i^*$

where $(y_1^*, ..., y_m^*)$ is the optimal solution to the dual problem of P. The variation of the optimum value of the objective function is $\sum_{i=1}^{m} \delta b_i y_i^*$. $\Diamond$

It is assumed that if the optimal basis of (P) is non degenerate, for reasons of continuity, there are non-zero variations of δb_i small enough to maintain the fact that the optimal basis of (P) remains feasible.

6.5 Dual-feasible problem dual-feasible

We consider a linear programming problem where the null solution is not feasible, but where the coefficients c_j of the objective function in the problem written in standard form are all negative or null. The use of the dual problem makes it possible to solve this problem without using the two-phase algorithm described in the first chapter.

Example

Consider the problem of linear programming:

$$\text{Minimize } x_1 + x_2$$
$$\text{with the constraints } \begin{cases} 3x_1 + x_2 \geqslant 4 \\ -7x_1 + x_2 \geqslant -7 \\ x_1 \geqslant 0, x_2 \geqslant 0 \end{cases}$$

whose writing in standard form is:

$$\text{Maximize } -x_1 - x_2$$
$$\text{with the constraints } \begin{cases} -3x_1 - x_2 \leqslant -4 \\ 7x_1 - x_2 \leqslant 7 \\ x_1 \geqslant 0, x_2 \geqslant 0. \end{cases}$$

The dual problem is written:

$$\text{Minimize } -4y_1 + 7y_2$$
$$\text{with the constraints } \begin{cases} -3y_1 + 7y_2 \geqslant -1 \\ -y_1 - y_2 \geqslant -1 \\ y_1 \geqslant 0, y_2 \geqslant 0. \end{cases}$$

or:

$$\text{Maximize } 4y_1 - 7y_2$$
$$\text{with the constraints } \begin{cases} 3y_1 - 7y_2 \leqslant 1 \\ y_1 + y_2 \leqslant 1 \\ y_1 \geqslant 0, y_2 \geqslant 0 \end{cases}$$

Slack variables now constitute a feasible basis: the simplex method requires only one phase. From the solution of the dual problem, we can deduce the solution of the primal problem.

6.6 Exercices

6.6.1 Exercice 1

Statement. We consider the problem:

$$\text{Maximize } z = 4x_1 + 3x_2$$
$$\text{with the constraints } \begin{cases} 5x_1 + 3x_2 \leqslant 30 \\ 2x_1 + 3x_2 \leqslant 24 \\ x_1 + 3x_2 \leqslant 18 \\ x_1 \geqslant 0, x_2 \geqslant 0. \end{cases}$$

Q1. Graphically solve this problem .
Q2. Use the complementary slackness theorem to prove that the graphical solution is correct.
Q3. The function z gives a profit in euros. It is planned to purchase one more unit from the first resource at a unit price of t. Until what value of t does this seem interesting?
Q4. We suppose we get a additional units of the first resource. Up to what value of a does the optimal basis of the initial problem remain feasible (then, this basis remains optimal)?

Solution.
Q1. We graphically represent the problem by the figure 6.1. The graphical solution is: $x_1^* = 3, x_2^* = 5$.

Q2. The solution obtained graphically is verified using the complementary slackness theorem. The solution $x_1^* = 3, x_2^* = 5$ is a feasible solution. We look for y_1^*, y_2^* and y_3^* fulfilling the conditions of the complementary slackness theorem.

- With $x_1^* = 3$ and $x_2^* = 5$, we have: $2x_1^* + 3x_2^* = 21 < 24$, which leads to: $y_2^* = 0$.

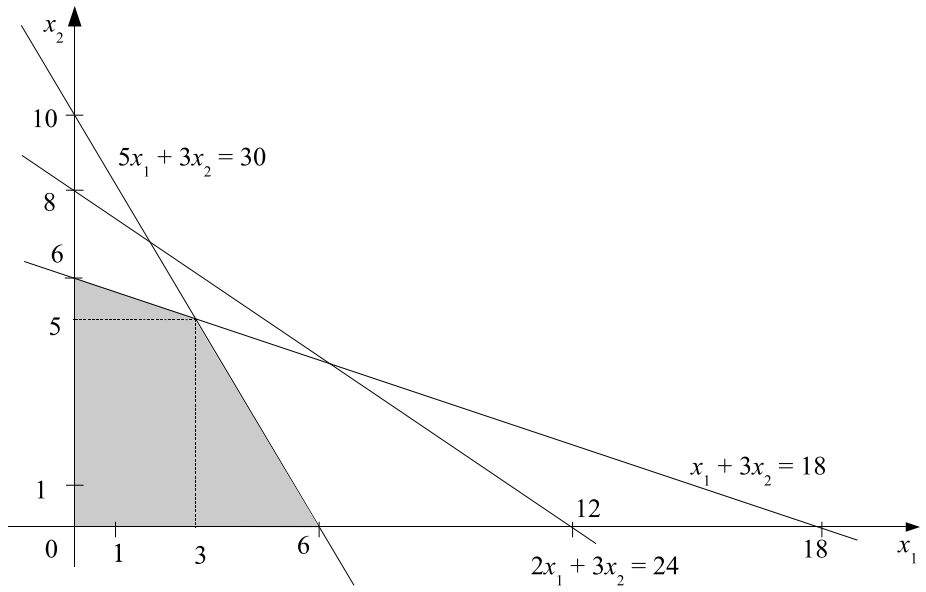

Figure 6.1: Graphical solution.

- Since x_1^* and x_2^* are not equal to zero, we must have:

$$\begin{cases} 5y_1^* + 2y_2^* + y_3^* = 4 \\ 3y_1^* + 2y_2^* + 3y_3^* = 3 \end{cases}.$$

With $y_2^* = 0$, the system above has the unique solution: $y_1^* = 3/4, y_3^* = 1/4$. It remains to check that the values $y_1^* = 3/4, y_2^* = 0, y_3^* = 1/4$ constitute a feasible solution of the dual problem, which is immediate. The solution determined graphically is indeed optimal.

Q3. The marginal value of the first resource is $3/4$. Getting one more unit from the first resource is interesting if its unit price is less than 0.75 euros.

Q4. We denote by x_3, x_4 and x_5 the three slack variables We notice that in the optimal basic solution of the initial problem, we have $x_3^* = 0$, $x_4^* = 3$, $x_5^* = 0$. Adding a to the first resource gives:

$$\begin{cases} 5x_1 + 3x_2 + x_3 = 30 + a \\ 2x_1 + 3x_2 + x_4 = 24 \\ x_1 + 3x_2 + x_5 = 18. \end{cases}$$

Using the numerical solution of the initial problem, let $x_1 = 3 + \delta x_1$, $x_2 = 5 + \delta x_2$, $x_4 = 3 + \delta x_4$. We have:

$$\begin{cases} 5\delta x_1 + 3\delta x_2 = a \\ 2\delta x_1 + 3\delta x_2 + \delta x_4 = 0 \\ \delta x_1 + 3\delta x_2 = 0. \end{cases}$$

This system has the solution: $\delta x_1 = \dfrac{a}{4}$, $\delta x_2 = -\dfrac{a}{12}$, $\delta x_4 = -\dfrac{a}{4}$.

The solution is feasible if and only if :

$$\begin{cases} 3 + \dfrac{a}{4} \geqslant 0 \\ 5 - \dfrac{a}{12} \geqslant 0 \\ 3 - \dfrac{a}{4} \geqslant 0 \end{cases}$$

that is, if and only if: $a \leqslant 12$.

6.6.2 Exercice 2

Statement. We propose $x_1^* = 0, x_2^* = \dfrac{4}{3}, x_3^* = \dfrac{2}{3}, x_4^* = \dfrac{5}{3}, x_5^* = 0$ as an optimal solution of the following problem:

$$\text{Maximize } z = 7x_1 + 6x_2 + 5x_3 - 2x_4 + 3x_5$$

$$\text{with the constraints } \begin{cases} x_1 + 3x_2 + 5x_3 - 2x_4 + 2x_5 \leqslant 4 \\ 4x_1 + 2x_2 - 2x_3 + x_4 + x_5 \leqslant 3 \\ 2x_1 + 4x_2 + 4x_3 - 2x_4 + 5x_5 \leqslant 5 \\ 3x_1 + x_2 + 2x_3 - x_4 - 2x_5 \leqslant 1 \\ x_1 \geqslant 0, x_2 \geqslant 0, x_3 \geqslant 0, x_4 \geqslant 0, x_5 \geqslant 0. \end{cases}$$

Is it correct ?

Solution.

The verification is as follows. We first examine whether the proposed solution is feasible.

- The proposed solution is positive or null.

- We check that it satisfies the other constraints and simultaneously identify saturated and non-saturated constraints.
  $x_1^* + 3x_2^* + 5x_3^* - 2x_4^* + 2x_5^* = 4$: saturated constraint.
  $4x_1^* + 2x_2^* - 2x_3^* + x_4^* + x_5^* = 3$: saturated constraint.
  $2x_1^* + 4x_2^* + 4x_3^* - 2x_4^* + 5x_5^* = 14/3 < 5$: fulfilled constraint but not saturated.
  $3x_1^* + x_2^* + 2x_3^* - x_4^* - 2x_5^* = 1$: saturated constraint.

- We write the equalities that the values y_i^* ($i = 1, 2, 3, 4$) must fulfill.
  - Since the third constraint is not saturated, $y_3^* = 0$.
  - Since $x_2^* > 0$, $3y_1^* + 2y_2^* + 4y_3^* + y_4^* = 6$.
  - Since $x_3^* > 0$, $5y_1^* - 2y_2^* + 4y_3^* + 2y_4^* = 5$.
  - Since $x_4^* > 0$, $-2y_1^* + y_2^* - 2y_3^* - y_4^* = -2$.

- We compute y_i^* ($i = 1, 2, 4$):

$$\begin{cases} 3y_1^* + 2y_2^* + y_4^* = 6 \\ 5y_1^* - 2y_2^* + 2y_4^* = 5 \\ -2y_1^* + y_2^* - y_4^* = -2 \end{cases}$$

  The solution of this system is: $y_1^* = y_2^* = y_4^* = 1$.

- We look at whether the y_i^* ($i = 1, 2, 3, 4$) constitute a feasible solution of the dual problem.
  - They are all positive or nul.
  - It remains to check the first and the fifth constraint of the dual problem since the other constraints are saturated by definition of y^*:
    $$y_1^* + 4y_2^* + 2y_3^* + 3y_4^* = 8 \geqslant 7$$
    $$2y_1^* + y_2^* + 5y_3^* - 2y_4^* = 1 < 3$$
    The last dual constraint is not verified: the current solution is not optimal.

We can notice that, if we now want to find the optimal solution, it would be wise to start from the basis x_2, x_3, x_4, x_8, where x_8 represents the third slack variable; this basis corresponds to the proposed solution.

6.6.3 Exercice 3

Statement. Give an example of a problem (P) such that neither the problem (P) nor the dual problem of (P) admit any feasible solution.

Solution. We consider the following problem (P):

$$\begin{array}{l}\text{Maximize } z = 2x_1 - x_2 \\ \left\{\begin{array}{l} x_1 - x_2 \leqslant 1 \\ -x_1 + x_2 \leqslant -2 \\ x_1 \geqslant 0, x_2 \geqslant 0 \end{array}\right.\end{array}$$

The dual (Q) of (P) is:

$$\begin{array}{l}\text{Minimize } w = y_1 - 2y_2 \\ \left\{\begin{array}{l} y_1 - y_2 \geqslant 2 \\ -y_1 + y_2 \geqslant -1 \\ y_1 \geqslant 0, y_2 \geqslant 0 \end{array}\right.\end{array}$$

It is easy to check that the problems (P) and (Q) do not admit any feasible solution.

6.6.4 Exercice 4

Statement.

Q1. We consider the problème (P) below :

$$\begin{array}{ll}\text{Minimize} & z = \sum_{j=1}^{n} c_j x_j \\ \text{with:} & \left\{\begin{array}{l} \text{for } i \in \{1, 2, ..., m\},\ \sum_{j=1}^{n} a_{ij} x_j \geqslant b_i \\ \text{for } i \in \{m+1, ..., m+p\},\ \sum_{j=1}^{n} a_{ij} x_j = b_i. \\ \text{for } j \in \{1, ..., n\},\ x_j \in \mathbb{R}. \end{array}\right.\end{array}$$

Show that the problem (Q) défined below is the dual problem of (P) :

$$\begin{array}{ll}\text{Maximize} & w = \sum_{i=1}^{m+p} b_i y_i \\ \text{with:} & \left\{\begin{array}{l} \text{for } j \in \{1, 2, ..., n\},\ \sum_{i=1}^{m+p} a_{ij} y_i = c_j \\ \text{for } i \in \{1, 2, ..., m\},\ y_i \geqslant 0 \\ \text{for } i \in \{m+1, ..., m+p\},\ y_i \in \mathbb{R}. \end{array}\right.\end{array}$$

Q2. Prove the following theorem (theorem of Farkas et Minkowski)[1] stipulating that the two propositions below are equivalent.

(i) Let $x \in \mathbb{R}^n$; if we have:

$$\begin{cases} \text{for } i \in \{1, 2, ..., m\}, \ \sum_{j=1}^{n} a_{ij} x_j \geqslant 0 \\ \text{for } i \in \{m+1, ..., m+p\}, \ \sum_{j=1}^{n} a_{ij} x_j = 0 \end{cases}$$

then : $\sum_{j=1}^{n} c_j x_j \geqslant 0$.

(ii) It exists $y \in \mathbb{R}^{m+p}$ fulfilling:

$$\begin{cases} \text{for } j \in \{1, 2, ..., n\}, \ \sum_{i=1}^{m+p} a_{ij} y_i = c_j \\ \text{for } i \in \{1, 2, ..., m\}, \ y_i \geqslant 0. \end{cases}$$

Solution.

Q1. We start by getting closer to the standard form:

$$\text{Maximize } \sum_{j=1}^{n} (-c_j x_j)$$
$$\text{with: } \begin{cases} \text{for } i \in \{1, 2, ..., m\}, \ \sum_{j=1}^{n} (-a_{ij} x_j) \leqslant -b_i \\ \text{for } i \in \{m+1, ..., m+p\}, \ \sum_{j=1}^{n} (-a_{ij} x_j) \leqslant -b_i \\ \text{for } i \in \{m+1, ..., m+p\}, \ \sum_{j=1}^{n} a_{ij} x_j \leqslant b_i \\ \text{for } j \in \{1, 2, ..., n\}, \ x_j \in \mathbb{R}. \end{cases}$$

Now let us reformulate the problem in standard form. For $j \in \{1, 2, ..., n\}$, we set: $x_j = x_j^1 - x_j^2$ with $x_j^1 \geqslant 0$ et $x_j^2 \geqslant 0$. We obtain:

$$\text{Maximize } \sum_{j=1}^{n} (-c_j x_j^1) + \sum_{j=1}^{n} c_j x_j^2$$
$$\text{with: } \begin{cases} \text{for } i \in \{1, 2, ..., m\}, \ \sum_{j=1}^{n} (-a_{ij} x_j^1) + \sum_{j=1}^{n} a_{ij} x_j^2 \leqslant -b_i \\ \text{for } i \in \{m+1, ..., m+p\}, \ \sum_{j=1}^{n} (-a_{ij} x_j^1) + \sum_{j=1}^{n} a_{ij} x_j^2 \leqslant -b_i \\ \text{for } i \in \{m+1, ..., m+p\}, \ \sum_{j=1}^{n} a_{ij} x_j^1 + \sum_{j=1}^{n} (-a_{ij} x_j^2) \leqslant b_i \\ \text{for } j \in \{1, 2, ..., n\}, \ x_j^1 \geqslant 0, \ x_j^2 \geqslant 0. \end{cases}$$

[1] G. Farkas, "Theorie der einfachen Ungleichungen", *Journal für die reine und angewandte Mathematik*, 124, 1902, 1–27. H. Minkowski, "Theorie der konvexen Körper, insbesondere Bergründung ihres Oberflächenbegriffs", *in Gesammelte Abhandlungen Hermann Minkowski II*, Teubner, Leipzig, 1911, 131–229. This theorem will be used in chapter 8.1 to establish the conditions of Karush, Kuhn and Tucker.

The dual problem is:

$$\text{Minimize } \sum_{i=1}^{m}(-b_i y_i) - \sum_{i=m+1}^{m+p} b_i y_i^1 + \sum_{i=m+1}^{m+p} b_i y_i^2$$
$$\text{with: } \begin{cases} \text{for } j \in \{1,2,...,n\}, \\ \quad \sum_{i=1}^{m}(-a_{ij} y_i) + \sum_{i=m+1}^{m+p}(-a_{ij} y_i^1) + \sum_{i=m+1}^{m+p} a_{ij} y_i^2 \geqslant -c_j \\ \text{for } j \in \{1,2,...,n\}, \\ \quad \sum_{i=1}^{m} a_{ij} y_i + \sum_{i=m+1}^{m+p} a_{ij} y_i^1 + \sum_{i=m+1}^{m+p}(-a_{ij} y_i^2) \geqslant c_j \\ \text{for } i \in \{1,2,...,m\},\ y_i \geqslant 0, \\ \text{for } i \in \{m+1,...,m+p\},\ y_i^1 \geqslant 0,\ y_i^2 \geqslant 0. \end{cases}$$

This can be rewritten:

$$\text{Maximize } \sum_{i=1}^{m} b_i y_i + \sum_{i=m+1}^{m+p} b_i (y_i^1 - y_i^2)$$
$$\text{avec : } \begin{cases} \text{for } j \in \{1,2,...,n\}, \\ \quad \sum_{i=1}^{m} a_{ij} y_i + \sum_{i=m+1}^{m+p} a_{ij} (y_i^1 - y_i^2) \leqslant c_j \\ \text{for } j \in \{1,2,...,n\}, \\ \quad \sum_{i=1}^{m} a_{ij} y_i + \sum_{i=m+1}^{m+p} a_{ij} (y_i^1 - y_i^2) \geqslant c_j \\ \text{for } i \in \{1,2,...,m\},\ y_i \geqslant 0 \\ \text{for } i \in \{m+1,...,m+p\},\ y_i^1 \geqslant 0,\ y_i^2 \geqslant 0. \end{cases}$$

By setting, for $i \in \{m+1,...,m+p\}$, $y_i = y_i^1 - y_i^2$, the variable y_i is unsigned and we can still write this dual problem as follows:

$$\text{Maximize } \sum_{i=1}^{m+p} b_i y_i$$
$$\text{with: } \begin{cases} \text{for } j \in \{1,2,...,n\},\ \sum_{i=1}^{m+p} a_{ij} y_i = c_j \\ \text{for } i \in \{1,2,...,m\},\ y_i \geqslant 0 \\ \text{for } i \in \{m+1,...,m+p\},\ y_i \in \mathbb{R}. \end{cases}$$

We obtain the problem (Q).

Q2. We use the previous question; we choose $b_i = 0$ for $i \in \{1,...,m+p\}$. The problems $(P)$and (Q) become (P_0) and (Q_0) defined by:

$$(P_0) \quad \begin{array}{l} \text{Minimize } z = \sum_{j=1}^{n} c_j x_j \\ \text{with: } \begin{cases} \text{for } i \in \{1,2,...,m\},\ \sum_{j=1}^{n} a_{ij} x_j \geqslant 0 \\ \text{for } i \in \{m+1,...,m+p\},\ \sum_{j=1}^{n} a_{ij} x_j = 0 \end{cases} \end{array}$$

and

$$(Q_0) \quad \begin{array}{l} \text{Maximize } w = 0 \\ \text{with: } \begin{cases} \text{for } j \in \{1,2,...,n\},\ \sum_{i=1}^{m+p} a_{ij} y_i = c_j \\ \text{for } i \in \{1,...,m\},\ y_i \geqslant 0 \\ \text{for } i \in \{m+1,...,m+p\},\ y_i \in \mathbb{R}. \end{cases} \end{array}$$

Note that the origin is realizable for (P_0) (and gives to z the value 0).

If proposition (i) is fulfilled, the problem (P_0) is bounded from below by 0 (in fact, its minimum is 0). According to the theorem of duality, the problem (Q_0) is feasible, which means that proposition (ii) is verified.

If proposition (ii) is fulfilled, the problem (Q_0) is feasible, of maximum 0. From duality theorem, the problem (P_0) is feasible of minimum 0, which means that proposition (i) is checked.

Chapter 7

Non linear optimization without constraint

7.1 Introduction

In this chapter, we are interested in optimizing functions defined on $\mathbb{R}^n$ and with values in $\mathbb{R}$.

Let f be a function from $\mathbb{R}^n$ to $\mathbb{R}$.

Definition 3. *We say that f reaches a global minimum (respectively maximum) in a point x^* of $\mathbb{R}^n$ if, for any $x \in \mathbb{R}^n$, we have $f(x) \geqslant f(x^*)$ (respectively $f(x) \leqslant f(x^*)$).*

Definition 4. *We say that f reaches a local minimum (respectively maximum) in a point x^* of $\mathbb{R}^n$ if there exists a ball B centered at x^* such that for any $x \in B$, we have $f(x) \geqslant f(x^*)$ (respectively $f(x) \leqslant f(x^*)$).*

We will first study the case $n = 1$, i.e. one-dimensional optimization, giving some optimization methods.

We will then go back to the general case to establish some theoretical results, in particular for the cases of quadratic functions and convex functions. We will detail some optimization methods: the descent methods, the conjugate gradient method and Newton method. One-dimensional optimization will often serve as a tool for multidimensional optimization.

Theorems and methods will be described for minimization. The case of maximization can be deduced directly since maximizing a function is minimizing its opposite: $\max_{x \in \mathbb{R}^n} f(x) = -\min_{x \in \mathbb{R}^n} -f(x)$.

7.2 One-dimensional optimization

We consider here a function f from $\mathbb{R}$ to $\mathbb{R}$ which we try to minimize.

7.2.1 Newton method

We assume f of class C ^{2}. Newton method consists of constructing a sequence (x_k) from a point x_0 as follows. In the point x_k, we approach f by:

$q(x) = f(x_k) + f'(x_k)(x - x_k) + \frac{1}{2}f''(x_k)(x - x_k)^2$.

We notice the following relation: $q'(x) = f'(x_k) + f''(x_k)(x - x_k)$.

If $f''(x_k) > 0$ (case where f is strictly convex around x_k), we set:

$$x_{k+1} = x_k - \frac{f'(x_k)}{f''(x_k)}$$

which is the point where q reaches its minimum ($q'(x_{k+1}) = 0$).

If we have $f''(x_k) \leqslant 0$, the method fails.

If f is of class C^3 and if x_0 is chosen close enough to a local minimum point x^* fulfilling $f''(x^*) > 0$, then the sequence (x_k) converges quadratically (see the definition in the part 7.7) to x^*. We will demonstrate this result in the case of the functions of several real variables (part 7.10).

7.2.2 Dichotomy for a differentiable function

Definition 5. *We say that a function is unimodal unimodal if there is a real x^* for which the function is strictly decreasing over $]-\infty, x^*]$ and strictly increasing on $[x^*, +\infty[$.*

The point x^* is then an global minimum of f.

It is assumed here that f is unimodal and differentiable. The point x^* is the only point where the derivative of f is equal to 0. The first step is to look for x_{min} and x_{max} such that $x_{min} < x^* < x_{max}$, so that we have the two relations $f'(x_{min}) < 0$ and $f'(x_{max}) > 0$.

After this first step, we put: $x = \frac{1}{2}(x_{min} + x_{max})$; if $f'(x) > 0$, we replace x_{max} by x, otherwise we replace x_{min} by x; the operation is repeated up to a stopping criterion to be specified.

The length of the interval being at each iteration divided by 2, we can show that the convergence is linear of rate 0.5 (see the definition in the part 7.7).

To determine x_{min} and x_{max}, a good method is the following (we assume that $f'(0)$ is not zero; otherwise, 0 is the solution of the problem!):

- define a step length $h > 0$
- if $f'(0) < 0$, do:
  - $\star$ $x_{min} \leftarrow 0$
  - $\star$ as long as $f'(h) < 0$, do:
    - $\triangleright$ $x_{min} \leftarrow h$
    - $\triangleright$ $h \leftarrow 2h$
  - $\star$ $x_{max} \leftarrow h$

  otherwise if $f'(0) > 0$, do:
  - $\star$ $h \leftarrow -h$
  - $\star$ $x_{max} \leftarrow 0$
  - $\star$ as long as $f'(h) > 0$, do:
    - $\triangleright$ $x_{max} \leftarrow h$
    - $\triangleright$ $h \leftarrow 2h$

  $x_{min} \leftarrow h$.

Remark. If f is not unimodal, the dichotomy is nevertheless applicable if we know x_{min} and x_{max} ($x_{min} < x_{max}$, $f'(x_{min}) < 0$, $f'(x_{max}) > 0$). It then converges to a local minimum that can be not global.

7.2.3 Quadratic interpolation

The method starts from the following principle: we first choose, using a preliminary algorithm, x_1, x_2 and x_3 fulfilling: $x_1 < x_2 < x_3$ as well as the inequalities $f(x_2) \leqslant f(x_1)$ and $f(x_2) \leqslant f(x_3)$. We approach f by a quadratic function q having the same values as f in x_1, x_2 and x_3:

$$q(x) = f(x_1)\frac{(x-x_2)(x-x_3)}{(x_1-x_2)(x_1-x_3)} + f(x_2)\frac{(x-x_1)(x-x_3)}{(x_2-x_1)(x_2-x_3)} + f(x_3)\frac{(x-x_1)(x-x_2)}{(x_3-x_1)(x_3-x_2)}.$$

The minimum of q is reached on $[x_1, x_3]$ at a point whose abscissa is easily expressed in terms of $x_1, x_2, x_3, f(x_1), f(x_2)$ and $f(x_3)$; we denote by x_4 this point. The update of the points x_1, x_2 and x_3 is done according to the following rules:

- if $f(x_4) \leqslant f(x_2)$
  - ⋆ if $x_4 \leqslant x_2$, the new triplet is (x_1, x_4, x_2)
    otherwise the new triplet is (x_2, x_4, x_3)

  otherwise
  - ⋆ if $x_4 \leqslant x_2$, the new triplet is (x_4, x_2, x_3)
    otherwise the new triplet is (x_1, x_2, x_4).

We can show that if f is fairly regular, the convergence is superlinear of order 1,3 (see definition in part 7.7).

7.2.4 Dichotomy without derivation for a unimodal function

It is assumed here that f is unimodal. Initially, using a preliminary algorithm, we choose a and b with $a < b$ and such that the minimum of f is reached between a and b. We then divide, using points d, c and e, the interval $[a, b]$ in four equal subintervals: $c = \dfrac{a+b}{2}$, $d = \dfrac{a+c}{2}$, $e = \dfrac{c+b}{2}$.

Comparing the values taken by f in a, b, c, d and e, we can eliminate two of the subintervals defined by these points and affirm that the minimum of f is reached in the union of two contiguous subintervals $[a_1, c_1]$ and $[c_1, b_1]$. The figure 7.1 illustrates such a case. We start again with the interval $[a_1, b_1]$. At each step, the length of the interval is divided by 2. The speed of convergence is linear.

7.3 Generalities for multidimensional optimization

We consider here functions f from $\mathbb{R}^n$ to $\mathbb{R}$. We try to determine the points where f reaches local or global extrema. For that, we need some definitions.

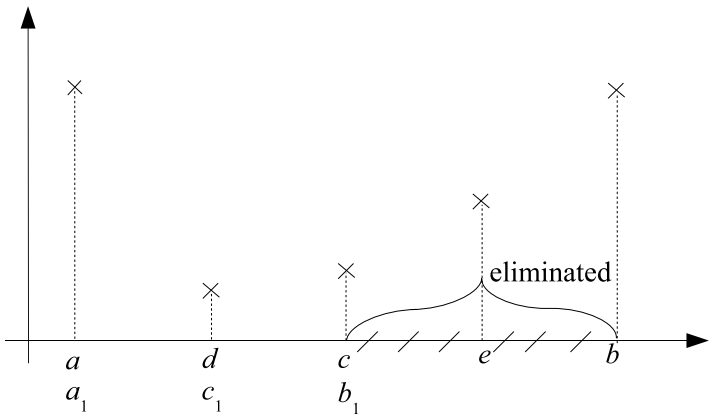

Figure 7.1: Dichotomy without derivation.

7.3.1 Notions of topology

Definition 6. *A part O of $\mathbb{R}^n$ is open, or is an open set, if, for any $x \in O$, there exists a ball of non-zero radius centered in x included within O.*

$\mathbb{R}^n$ is an open set. The empty set of $\mathbb{R}^n$ is open. Any product of open intervals of $\mathbb{R}$ is open.

Definition 7. *A part F of $\mathbb{R}^n$ is closed, or is a closed set, if its complementary is open.*

$\mathbb{R}^n$ is a closed set (so it is both open and closed). Any product of closed intervals of $\mathbb{R}$ is closed.

Definition 8. *A part K of $\mathbb{R}^n$ is compact, or is a compact set, if is closed and bounded.*

Theorem 9. *A continuous function f defined on a compact set K of $\mathbb{R}^n$ with real values reaches its bounds; in other words, there exists $x_1 \in K$ (respectively $x_2 \in K$) such that, for every $x \in K$, $f(x) \geqslant f(x_1)$ (respectively $f(x) \leqslant f(x_2)$): x_1 is a global minimum (respectively x_2 is a global maximum).*

Throughout this chapter, O denotes an open set of $\mathbb{R}^n$.

7.3.2 Gradient

Let f be a function from an open set O of $\mathbb{R}^n$ to $\mathbb{R}$ admitting at a point $x \in O$ first-order partial derivatives. We will set $x = (x_1, x_2, ..., x_n)^{\mathrm{t}}$ (the elements of $\mathbb{R}^n$ are assimilated to column vectors).

We name *gradient* of f at the point x and we denote by $\nabla f(x)$ the column vector:

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

If $F(x) = (f_1(x), ..., f_p(x))$ is a row vector where $f_1, ..., f_p$ are real functions of n real variables differentiable at the point x, then $\nabla F(x)$ is the matrix whose j^{th} column is $\nabla f_j(x)$.

The following formulas will be useful later: if A is a constant square matrix of order n, if $u(x)$ and $v(x)$ are two column vectors depending on x, then :

$$\nabla \left(u^{\mathrm{t}} A\right) = \nabla \left(u^{\mathrm{t}}\right) A$$
$$\nabla \left(u^{\mathrm{t}} v\right) = \nabla \left(u^{\mathrm{t}}\right) v + \nabla \left(v^{\mathrm{t}}\right) u.$$

If f admits continuous partial derivatives in x^0, we can apply the *Taylor formula to order 1*:

$$f(x) = f(x^0) + (x - x^0)^{\mathrm{t}}.\nabla f(x^0) + ||x - x^0||.\epsilon(x)$$

where $\epsilon(x)$ is a function that tends toward 0 when x tends toward x^0.

Remarks.
1. Suppose f of class C^1. If we consider the surface S of $\mathbb{R}^{n+1}$ of equation $x_{n+1} = f(x_1, ..., x_n)$, then the expression $x_{n+1} = f(x^0) + (x - x^0)^{\mathrm{t}}.\nabla f(x^0)$ gives the equation of the hyperplane tangent to S at the point $(x^0, f(x^0))$.
2. We will thereafter be interested in the variations of f in a direction d of $\mathbb{R}^n$ starting from a point x^0 of $\mathbb{R}^n$. For $s \in \mathbb{R}$, let $g(s) = f(x^0 + s.d)$. We then obtain:

$$g'(s) = d^{\mathrm{t}}.\nabla f(x^0 + s.d)$$
$$g'(0) = d^{\mathrm{t}}.\nabla f(x^0).$$

7.3.3 Hessian matrix

If now f admits second partial derivatives in x, we set:

$$\nabla^2 f(x) = \nabla \left(\nabla f(x)^{\mathrm{t}} \right),$$

that is:

$$\nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_2 \partial x_n}(x) \\ \dots & \dots & \dots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_n^2}(x) \end{pmatrix};$$

$\nabla^2 f$ is called the *Hessian matrix* of f.

If f is a function of class C^2 (in other words, f admits continuous second partial derivatives), the Hessian matrix of f is a symmetric matrix (Schwarz theorem).

If f is a function of class C^2 in x^0, we can write the Taylor formula of order 2:

$$f(x) = f(x^0) + (x - x^0)^{\mathrm{t}}.\nabla f(x^0) + \frac{1}{2}(x - x^0)^{\mathrm{t}} \nabla^2 f(x^0).(x - x^0) + ||x - x^0||^2.\epsilon(x),$$

where $\epsilon(x)$ is a function that tends towards 0 when x tends towards x^0.

7.4 Necessary condition and sufficient condition for local optimality

Suppose here that f is a function from $\mathbb{R}^n$ to $\mathbb{R}$ of class C^2.

We remind the following definitions:

Definition 9. *Let M be a real square symmetric matrix.*

- *M is positive semi-definite if, $\forall h \in \mathbb{R}^n, h^t.M.h \geqslant 0$,*
- *M is positive definite if, $\forall h \in \mathbb{R}^n \setminus \{0\}, h^t.M.h > 0$.*

A symmetric square real matrix is positive semi-definite if and only if its eigenvalues are positive or zero. It is positive definite if and only if its eigenvalues are strictly positive.

Theorem 10 (necessary condition of optimality)**.** *If f admits a local minimum in x^*, then:*

1. $\nabla f(x^*) = 0$

2. $\nabla^2 f(x^*)$ *is a positive semi-definite matrix.*

Proof. According to Taylor development of order 1 in x^*, we have:

$$f(x) = f(x^*) + (x - x^*)^{\mathrm{t}} \nabla f(x^*) + ||x - x^*|| \epsilon(x),$$

where $\epsilon(x)$ is a function that tends towards 0 when x tends towards x^*. In particular, by choosing $x = x^* - \theta.\nabla f(x^*)$, with $\theta \in \mathbb{R}$,we obtain :

$$f(x) - f(x^*) = -\theta ||\nabla f(x^*)||^2 + \theta \epsilon_1(\theta) = \theta \left(-||\nabla f(x^*)||^2 + \epsilon_1(\theta) \right),$$

where $\epsilon_1(\theta)$ is a function that tends towards 0 when θ tends towards 0. For θ positive, $f(x) - f(x^*)$ has the sign of $-||\nabla f(x^*)||^2 + \epsilon_1(\theta)$. If we have $\nabla f(x^*) \neq 0$, there exist in every neighborhood of x^* points x satisfying $f(x) < f(x^*)$ (for θ small positive, $f(x) - f(x^*$ has the sign of $-||\nabla f(x^*)||^2$, assuming this term is nonzero), a contradiction with the local optimality of x^*. Hence the result 1.

Suppose now that there exists $h \in \mathbb{R}^n$ such that we have the relation: $h^{\mathrm{t}} \nabla^2 f(x^*) h < 0$. We then have, according to Taylor development of order 2:

$$f(x^* + \theta h) - f(x^*) = \theta^2 \left(\frac{1}{2} h^t \nabla^2 f(x^*) h + \epsilon_2(\theta) \right),$$

where $\epsilon_2(\theta)$ is a function that tends towards 0 when θ tends towards 0.For θ small enough, the difference $f(x^* + \theta h) - f(x^*)$ would be negative, which contradicts the hypothesis on x^*. ◊

Theorem 11 (sufficient condition of optimality)**.** *If a function ffulfils in x^*:*

1. $\nabla f(x^*) = 0$
2. $\nabla^2 f(x^*)$ *is a positive definite matrix*

then f admits a local minimum in x^*.

Proof. The matrix $\nabla^2 f(x^*)$ being positive definite, there exists $a > 0$ such that:

$$\forall h \in \mathbb{R}^n, h^t \nabla^2 f(x^*) h \geqslant a ||h||^2.$$

Indeed, let us go on the sphere S of center 0 and radius 1 and set a to $a = \inf\{h^{\mathrm{t}} \nabla^2 f(x^*) h \text{ for } h \in S\}$. The sphere being a compact, the value a is

reached: $\exists\ h_0 \in S$ such that $a = h_0^{\mathrm{t}} \nabla^2 f(x^*) h_0 > 0$. It is easy to deduce the previous result. Let x be $\in \mathbb{R}^n$. Let us apply Taylor formula of order 2 by setting $h = x - x^*$:

$$f(x) - f(x^*) = f(x^* + h) - f(x^*) \geqslant ||h||^2 \left(\frac{a}{2} + \epsilon(h) \right),$$

where $\epsilon(h)$ is a function that tends towards 0 when h tends towards 0, which shows the theorem because, for h of fairly small norm, $\dfrac{a}{2} + \epsilon(h)$ has the sign of a, that is positive. We therefore have $f(x) \geqslant f(x^*)$ when x tends towards x^*: x^* is a local minimum of f. $\Diamond$

7.5 Quadratic functions

Let A be a symmetric matrix of order n, b a column vector of order n and c a real number. The function q from $\mathbb{R}^n$ to $\mathbb{R}$ defined by:

$$q(x) = c + b^{\mathrm{t}} x + \frac{1}{2} x^{\mathrm{t}} A x$$

is called *quadratic function* or also *quadratic form*.

Remark. The polynomial part of Taylor second-order development of a function f is the quadratic function q such that the surface of equation $x_{n+1} = q(x)$ is "the closest" to the surface of equation $x_{n+1} = f(x)$ near the considered point.

We have, using the formulas given in paragraph 1:

$$\nabla q(x) = \nabla(x^{\mathrm{t}}) b + \frac{1}{2} [\nabla(x^{\mathrm{t}}) A x + \nabla((Ax)^{\mathrm{t}}) x].$$

Now, $\nabla(x^{\mathrm{t}})$ is the identity matrix. We also have the following equalities:

$$\nabla((Ax)^{\mathrm{t}}) = \nabla(x^{\mathrm{t}} A^{\mathrm{t}}) = \nabla(x^{\mathrm{t}}) A^{\mathrm{t}} = A^{\mathrm{t}} = A.$$

Hence the expression of the gradient: $\nabla q(x) = b + Ax$.

Furthermore: $\nabla^2 q(x) = \nabla((\nabla q(x))^{\mathrm{t}}) = \nabla(b^{\mathrm{t}} + x^{\mathrm{t}} A^{\mathrm{t}}) = A^{\mathrm{t}} = A$. So we finally have: $\nabla^2 q(x) = A$.

Derivatives of order at least 3 of q are null. A quadratic function coincides with its Taylor development at order 2.

7.6 Convex functions

Definition 10. *A part of $\mathbb{R}^n$ is said to be convex if it contains any segment joining any two of its points.*

Definition 11. *We say that a function f defined on a convex part of $\mathbb{R}^n$ and with real values is convex if, for any x and any y of its domain of definition and for any λ of $]0,1[$, we have: $f(\lambda x + (1-\lambda)y) \leqslant \lambda f(x) + (1-\lambda)f(y)$. If this inequality is strict, we say that f is strictly convex.*

In all the part 7.6, we assume that f is defined on a convex open set O of $\mathbb{R}^n$.

Theorem 12. *If f is a convex function and admits partial derivatives, then f admits a global minimum in x^* if and only if we have $\nabla f(x^*) = 0$.*

Proof. According to the theorem 10, if x^* is a local minimum, we have $\nabla f(x^*) = 0$. Let us show the converse: if $\nabla f(x^*)$ is equal to 0, then f admits a global minimum in x^*. Let $x \in O$. For $s \in [0,1]$, let us set $g(s) = f(x^* + s(x - x^*))$. We have: $g(0) = f(x^*)$, $g(1) = f(x)$ and $g'(0) = (x - x^*)^{\mathrm{t}} \nabla f(x^*) = 0$. Moreover, we easily check that g is a convex function. Since the derivative of a convex function is increasing, we have, for $s \in [0,1]$, $g'(s) \geqslant 0$; g is increasing for $s \geqslant 0$, hence $g(1) \geqslant g(0)$: f admits a global minimum in x^*. ◊

Theorem 13. *If f is convex and has a local minimum in x^*, then f has a global minimum in x^*.*

Proof. If f has a local minimum in x^*, then $\nabla f(x^*) = 0$. If, furthermore, f is convex, the previous theorem leads to the conclusion that it admits a global minimum in x^*. ◊

We will admit the following theorem.

Theorem 14. *If f is twice differentiable with continuous second derivatives, the following propositions are equivalent:*

1. *f is convex.*
2. *For any x^0 of O, the tangent hyperplane at the point $(x^0, f(x^0))$ to the surface of equation $x_{n+1} = f(x)$ is below this surface; in other words, for every x of O, we have: $f(x) \geqslant f(x^0) + (\nabla f(x^0))^t (x - x^0)$.*
3. *For any x of O, $\nabla^2 f(x)$ is positive semi-definite.*

We deduce that a quadratic function $q(x) = \frac{1}{2}x^{\mathrm{t}}Ax + b^{\mathrm{t}}x + c$ is convex if and only if A is positive semi-definite. On the other hand, if A is positive definite, then q is strictly convex and admits a single global minimum.

7.7 Generalities on methods for optimization without constraint

It is assumed until the end of the chapter that $O = \mathbb{R}^n$.

Even if we are most of the time interested in global extrema, we will usually look for local extrema, even if we then examine (if possible) whether it is global extrema.

When we consider fractions in what follows, we will assume that the denominators are non-zero (the adaptations being immediate otherwise).

To determine a point where a function f reaches a local minimum, the methods very often consist in constructing a sequence x^0, x^1, ..., x^k, ... which must converge to a point x^* satisfying a necessary condition of optimality. This condition (often $\nabla f(x^*) = 0$) is usually not sufficient and the behavior of f in the neighborhood of x^* must be subject to additional study (which can include, among other things, the Hessian matrix of f at the point x^*).

7.7.1 Descent methods

The methods used are often descent methods; we call *descent method* any method where, at each stage, we set $x^{k+1} = x^k + s_k d^k$, where $s_k \in \mathbb{R}^+$ and d^k is a direction of $\mathbb{R}^n$ which fulfils $(d^k)^{\mathrm{t}}\nabla f(x^k) < 0$. This last condition means that $f(x^k + sd^k)$ has a negative derivative for $s = 0$: starting from x^k in the direction d^k, f decreases ("we descend"). The difference between the various descent methods is the choice of s_k and d^k, which must at least ensure $f(x^{k+1}) \leqslant f(x^k)$.

7.7.2 Speed of convergence

When the convergence of an algorithm has been established, an important quality of this algorithm is its *speed of convergence* .

- If we have $\dfrac{||x^{k+1} - x^*||}{||x^k - x^*||} \leqslant \alpha < 1$ for k large enough, convergence is said to be *linear of rate* α.

- If $\dfrac{||x^{k+1} - x^*||}{||x^k - x^*||}$ tends towards 0 when k tends towards infinity, we say that the *convergence* is *superlinear*.

- If $\dfrac{||x^{k+1} - x^*||}{||x^k - x^*||^\gamma}$ is bounded, with $\gamma > 1$, we say that the *convergence* is *superlinear of order* γ. In the case $\gamma = 2$, we say that the convergence is *quadratic*.

7.8 Gradient methods

7.8.1 Principle

It is a family of iterative methods that apply to differentiable functions and use the idea below.

Let d be a vector of $\mathbb{R}^n$ and x^k a point of $\mathbb{R}^n$ with $\nabla f(x^k) \neq 0$. We set, for $s \in \mathbb{R}$: $g(s) = f(x^k + sd)$.

We say that d is a *direction of descent* if $g'(0) < 0$. We saw the relationship $g'(0) = d^t \nabla f(x^k)$. Hence, denoting by θ the angle between $\nabla f(x^k)$ and d: $g'(0) = ||\nabla f(x^k)||||d|| \cos\theta$.

Assuming that d is unitary, $g'(0)$ is minimum if $\cos\theta = -1$, that is, if d is given by the opposite of the gradient: $d = -\dfrac{\nabla f(x^k)}{||\nabla f(x^k)||}$. This last direction gives what is called the *direction of steepest descent*. It is this choice that is made in the gradient methods.

7.8.2 Method of steepest descent with optimal step

The method of the steepest descent with optimal step is the most widely used gradient method. We choose here $d^k = -\nabla f(x^k)$ to have the steepest descent. We then set $g(s) = f(x^k - s\nabla f(x^k))$ and we compute s_k so as to minimize g for $s \geqslant 0$ (if such s_k exists). We are then reduced to a one-dimensional optimization problem. Let λ be a strictly positive constant. The algorithm of the steepest descent can be written in the following way:

- Choose a starting point x^0
- $k \leftarrow 0$
- repeat

- ⋆ $d^k \leftarrow -\nabla f(x^k)$
- ⋆ define the function g on $[0, +\infty[$ by $g(s) = f(x^k + sd^k)$
- ⋆ if g admits a global minimum s_k on the interval $[0, +\infty[$, $x^{k+1} \leftarrow x^k + s_k d^k$
  (we can also take for s_k the first local minimum starting from $s = 0$, especially in the case where g admits a local minimum but no global minimum)
  otherwise if g asymptotically tends towards $-\infty$, conclude that f has no finite minimum and stop
  otherwise (g is decreasing and tends asymptotically towards a finite limit), $x^{k+1} \leftarrow x^k + \lambda d^k$
- ⋆ $k \leftarrow k + 1$

as long as a given stop test is not fulfilled.

The stopping test can be for example:

- we have exhausted a number of iterations fixed in advance;
- the gradient is very small: $\sum_{i=1}^{n} (\frac{\partial f}{\partial x_i}(x^k))^2 \leqslant \epsilon$, where ϵ is a given parameter (we can of course consider another norm);
- the sequence x^k is "almost" stationary: $|f(x^{k+1}) - f(x^k)| \leqslant \epsilon$, where ϵ is a given parameter (we can of course consider another norm).

It can also be required that one of these tests be fulfilled on several iterations or that several tests are satisfied simultaneously. We can show that if $f(x)$ is a function of class C^1 that tends towards infinity when $||x||$ tends towards infinity, this algorithm converges to a stationary point (point where the gradient is equal to 0).

The disadvantage of this method is that the speed of convergence can be very low (linear with a rate close to 1). This slowness can be explained as follows: the equality $\frac{\delta}{\delta s}[f(x^k - s\nabla f(x^k))](s_k) = 0$ can be written : $[\nabla f(x^k)]^t \nabla f(x^{k+1}) = 0$; the successive directions of descent are orthogonal. In the figure 7.2, some contour lines and movements are represented. There is zig-zag convergence.

7.8.3 Method of steepest descent with fixed step

This method differs from the previous one in that we do not look for the minimum of the function f in the direction $-\nabla f(x^k)$ but that we do one

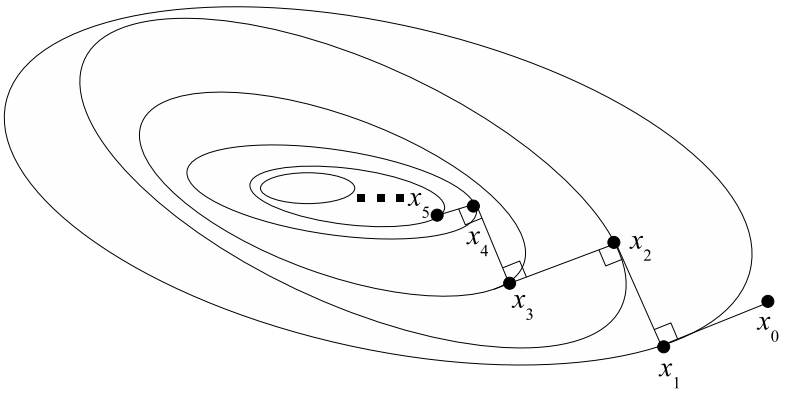

Figure 7.2: Zig-zag convergence.

step in that direction. The length of the step is fixed in advance. The algorithm can be written as follows:

- Set a constant λ strictly positive for the step length
- Choose a starting point x^0
- $k \leftarrow 0$
- repeat
  - $\star$ $d^k \leftarrow \dfrac{-\nabla f(x^k)}{||\nabla f(x^k)||}$
  - $\star$ $x^{k+1} \leftarrow x^k + \lambda d^k$
  - $\star$ $k \leftarrow k+1$

  as long as a given stop test is not fulfilled.

7.8.4 Accelerated method of steepest descent

The accelerated method of the steepest descent is a method of descent which is based on the method of the steepest descent with optimal step and which accelerates it.

Let p be a fixed integer. From a point x^k, we perform p iterations of the steepest descent method; we get a point y^k and we set $d^k = y^k - x^k$. The point x^{k+1} is the point where the function $f(x^k + sd^k)$ has a minimum for $s > 0$.

The figure 7.3 illustrates this method in the case $p = 2$. For $p = 1$, we apply the method of the steepest descent with optimal step.

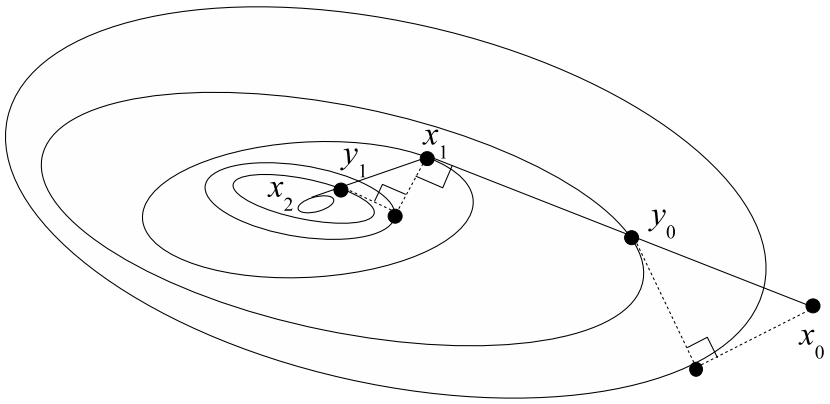

Figure 7.3: Accelerated convergence.

7.9 Conjugate gradients method

7.9.1 Case of a quadratic function

Let $q(x) = \frac{1}{2}x^{\mathrm{t}}Ax + b^{\mathrm{t}}x + c$ be a quadratic function, where A is a positive definite symmetric matrix.

The method consists, from a point x^0, of minimizing q according to n directions $d^0, d^1, ..., d^{n-1}$ mutually conjugated with respect to A that is to say satisfying: for $0 \leqslant i < j \leqslant n-1, (d^i)^{\mathrm{t}}Ad^j = 0$.

We consider such directions $d^0, d^1, ..., d^{n-1}$.

Having determined x^k, the point x^{k+1} is the point: $x^{k+1} = x^k + s_k d^k$ where s_k is chosen to minimize $q(x^k + s_k d^k)$.

So we have : $(d^k)^{\mathrm{t}}\nabla q(x^k + s_k d^k) = 0$ or: $(d^k)^{\mathrm{t}}[A(x^k + s_k d^k) + b] = 0$ from which we deduce: $s_k = -\dfrac{(d^k)^{\mathrm{t}}(Ax^k + b)}{(d^k)^{\mathrm{t}}Ad^k}$.

Lemma 15. *If $d^0, d^1, ..., d^{k-1}$ are mutually conjugated with respect to A, then we have, for every $i < k$: $(d^i)^t\nabla q(x^k) = 0$.*

Proof. We have:

$$\begin{aligned}(d^i)^{\mathrm{t}}\nabla q(x^k) &= (d^i)^{\mathrm{t}}(Ax^k+b)\\ &= (d^i)^{\mathrm{t}}\left[A\left(x^i+\sum_{j=i}^{k-1}s_j d^j\right)+b\right]\\ &= (d^i)^{\mathrm{t}}(Ax^i+b)+s_i(d^i)^{\mathrm{t}}Ad^i\\ &= 0 \text{ from the value of } s_i \text{ computed above.}\end{aligned}$$ ◊

Theorem 16. *If the directions $d^0, d^1, ..., d^{n-1}$ are mutually conjugated, the point x^n is the optimum of $q(x)$ on $\mathbb{R}^n$.*

Proof. Since the directions $d^0, d^1, ..., d^{n-1}$ are mutually conjugate, they form a basis of $\mathbb{R}^n$. According to the lemma 15, we have, for every i satisfying $0 \leqslant i \leqslant n-1$, $(d^i)^{\mathrm{t}}\nabla q(x^n) = 0$ whence $\nabla q(x^n) = 0$; with $\nabla^2 q(x^n) = A$ and A is positive definite, the theorem 11 permits to conclude. ◊

The Fletcher and Reeves method generates the directions d^i; it is explained below by setting: $g^k = \nabla q(x^k) = Ax^k + b$.

- Choose a starting point x^0.
- $d^0 \leftarrow -g^0$
- $s_0 \leftarrow -\dfrac{(d^0)^{\mathrm{t}}g^0}{(d^0)^{\mathrm{t}}Ad^0}$
- $x^1 \leftarrow x^0 + s_0 d^0$
- For k varying from 0 to $n-2$ do
  - $\star$ $b_k \leftarrow \dfrac{(d^k)^{\mathrm{t}}Ag^{k+1}}{(d^k)^{\mathrm{t}}Ad^k}$
  - $\star$ $d^{k+1} \leftarrow -g^{k+1} + b_k d^k$
  - $\star$ $s_{k+1} \leftarrow -\dfrac{(d^{k+1})^{\mathrm{t}}g^{k+1}}{(d^{k+1})^{\mathrm{t}}Ad^{k+1}}$.
  - $\star$ $x^{k+2} \leftarrow x^{k+1} + s_{k+1}d^{k+1}$

To justify the method, it is enough to check that $d^0, d^1, ...d^{n-1}$ are mutually conjugated. Let us show by induction on k that for $k \geqslant 0$, $d^0, d^1, ...d^k$

are mutually conjugated. There is nothing to check for $k = 0$. Suppose this is true for some k, $0 \leqslant k \leqslant n-2$. We then have for $k+1$:

$$\begin{aligned}(d^k)^{\mathrm{t}} A d^{k+1} &= (d^k)^{\mathrm{t}} A(-g^{k+1} + b_k d^k) \\ &= -(d^k)^{\mathrm{t}} A g^{k+1} + b_k (d^k)^{\mathrm{t}} A d^k = 0 \text{ according to the choice of } b_k.\end{aligned}$$

For $i < k$, $(d^{k+1})^{\mathrm{t}} A d^i = -(g^{k+1})^{\mathrm{t}} A d^i + b_k (d^k)^{\mathrm{t}} A d^i = -(g^{k+1})^{\mathrm{t}} A d^i$.

Now: $Ad^i = A\left(\dfrac{x^{i+1} - x^i}{s_i}\right) = \dfrac{Ax^{i+1} - Ax^i}{s_i} = \dfrac{g^{i+1} - g^i}{s_i}$.
On the other hand:

- si $i \geqslant 1$, $g^i = -d^i + b_{i-1} d^{i-1}$
- $g^0 = -d^0$.

According to the lemma and the recursion hypothesis, g^{k+1} is orthogonal to d^{i+1}, d^i and d^{i-1} ; Ad^i being a linear combination of these three vectors, $(g^{k+1})^{\mathrm{t}} A d^i = 0$, which shows the equality $(d^{k+1})^{\mathrm{t}} A d^i = 0$ for $i < k$.

It follows from the above that the recurrence assumption is true for $k+1$. Consequently, the directions d^0, ..., d^{n-1} are mutually conjugate.

Finally, we demonstrate a formula that will be useful in the following paragraph.
We have $g^{k+1} - g^k = A(x^{k+1} - x^k) = s_k A d^k$.
From which $(d^k)^{\mathrm{t}} A g^{k+1} = \dfrac{(g^{k+1} - g^k)^{\mathrm{t}} (g^{k+1})}{s_k}$.
As $g^k = -d^k + b_{k-1} d^{k-1}$, the lemma shows the equality $(g^{k+1})^{22^{\mathrm{t}}} g^k = 0$.

Hence, thanks to the lemma:
$b_k = \dfrac{(d^k)^{\mathrm{t}} A g^{k+1}}{(d^k)^{\mathrm{t}} A d^k} = \dfrac{1}{s_k} \dfrac{(g^{k+1})^{\mathrm{t}} g^{k+1}}{(d^k)^{\mathrm{t}} A d^k} = \dfrac{(g^{k+1})^{\mathrm{t}} g^{k+1}}{(d^k)^{\mathrm{t}} (g^{k+1} - g^k)} = -\dfrac{(g^{k+1})^{\mathrm{t}} g^{k+1}}{(d^k)^{\mathrm{t}} g^k}$.

Now: $(d^k)^{\mathrm{t}} g^k = (-g^k + b_{k-1} d^{k-1})^{\mathrm{t}} g^k = -(g^k)^{\mathrm{t}} g^k$ according to the lemma.
We deduce the result: $b_k = \dfrac{||g^{k+1}||^2}{||g^k||^2}$.

7.9.2 Case of any function

The Fletcher and Reeves algorithm for any function is:

- Choose a point x^0
- $d^0 \leftarrow -\nabla f(x^0)$
- $k \leftarrow 0$
- repeat
  - ⋆ choose s_k minimizing $f(x^k + sd^k)$ with respect to s
  - ⋆ $x^{k+1} \leftarrow x^k + s_k d^k$
  - ⋆ $b_k \leftarrow \dfrac{||\nabla f(x^{k+1})||^2}{||\nabla f(x^k)||^2}$
  - ⋆ $d^{k+1} \leftarrow -\nabla f(x^{k+1}) + b_k d^k$
  - ⋆ $k \leftarrow k + 1$

  until a stop test is fulfilled.

This method has the advantage of having a convergence speed much higher than that of conventional gradient algorithms.

7.10 Newton method

Suppose here that f is of class C ^{3}.

In the neighborhood of a point x^k, we approach f by the quadratic function q given by the Taylor formula of order 2:

$$q(x) = f(x^k) + (x - x^k)^{\mathrm{t}} \nabla f(x^k) + \frac{1}{2}(x - x^k)^{\mathrm{t}} \nabla^2 f(x^k)(x - x^k).$$

We can then choose for x^{k+1} the point, if it exists, which minimizes q; for this minimizing point q to exist, it is sufficient that $\nabla^2 f(x^k)$ be positive definite; it is then determined by the equation $\nabla q(x) = 0$, which is written:

$$\nabla f(x^k) + \nabla^2 f(x^k)(x - x^k) = 0,$$

from where :

$$x^{k+1} = x^k - [\nabla^2 f(x^k)]^{-1} \nabla f(x^k).$$

Proposition 17. *If x^0 is chosen sufficiently close to a local minimum x^* whose Hessian matrix of f is positive definite, then the sequence (x^k) has a quadratic convergence to x^*.*

Proof. We consider a vectorial norm $||\ ||$ and the matrix norm subordinate to it $||\ ||$ (see annex A). We will establish a sufficient condition on x^0 to ensure a quadratic convergence of the sequence (x^k) constructed from x^0 by Newton method. For this, we consider the following elements.

- We know that $\nabla^2 f(x^*)$ is positive definite. By continuity of the function $\nabla^2 f$, there exists a ball B_1 of center x^* and of radius r_1 on which $\nabla^2 f(x)$ is positive definite and therefore invertible. We then denote by M an upper bound of $||(\nabla^2 f)^{-1}||$ on B_1 (such an upper bound exists since $(\nabla^2 f)^{-1}$ is continuous and B_1 is a compact).

- Using the fact that f is of class C ^{3}, the Taylor formula with remainder under integral form shows that there exists a constant N strictly positive and a function $\phi(a,b)$ for which we have, if a and b are two points of B_1:

  $\star$ $\nabla f(b) = \nabla f(a) + \nabla^2 f(a)(b-a) + \phi(a,b)||b-a||^2$

  $\star$ $||\phi(a,b)|| \leqslant N.$ (1)

- We set $M' = MN$.

- We consider a real r simultaneously fulfilling $r \leqslant r_1$ and $r < \dfrac{1}{M'}$; we call B the ball centered in x^* and of radius r (note that B is included in B_1).

We assume that x^0 is in B. We will show by induction on k that the sequence (x^k) is entirely in B. This is true for $k = 0$ and we assume that it is true for $k \geqslant 0$.

We have: $\nabla f(x^*) - \nabla f(x^k) = \nabla^2 f(x^k)(x^* - x^k) + \phi(x^k, x^*)||x^* - x^k||^2$.

Using $\nabla f(x^*) = 0$ (consequence of the minimality of x^*):

$$\nabla^2 f(x^k)(x^k - x^*) = \nabla f(x^k) + \phi(x^k, x^*)||x^k - x^*||^2.$$

The matrix $\nabla^2 f(x^k)$ being invertible (since we have $||x^k - x^*|| \leqslant r_1$), we obtain, by multiplying on the left the two members by $[\nabla^2 f(x^k)]^{-1}$:

$$x^k - x^* = [\nabla^2 f(x^k)]^{-1}\nabla f(x^k) + [\nabla^2 f(x^k)]^{-1}\phi(x^k, x^*)||x^k - x^*||^2. \quad (2)$$

Moreover:

$$x^{k+1} - x^* = (x^{k+1} - x^k) + (x^k - x^*). \quad (3)$$

By construction:

$$x^{k+1} - x^k = -[\nabla^2 f(x^k)]^{-1}\nabla f(x^k). \quad (4)$$

Using the equalities (2), (3) and (4), we obtain:

$$\begin{aligned} x^{k+1} - x^* &= -[\nabla^2 f(x^k)]^{-1}\nabla f(x^k) \\ &\quad + [\nabla^2 f(x^k)]^{-1}\nabla f(x^k) + [\nabla^2 f(x^k)]^{-1}\phi(x^k, x^*)||x^k - x^*||^2 \\ &= [\nabla^2 f(x^k)]^{-1}\phi(x^k, x^*)||x^k - x^*||^2. \end{aligned}$$

Hence: $||x^{k+1}-x^*|| \leqslant ||[\nabla^2 f(x^k)]^{-1}||||\phi(x^k,x^*)||\,||x^k-x^*||^2$; consequently, exploiting the property (1) and the inequality $r \leqslant r_1$:

$||x^{k+1}-x^*|| \leqslant MN||x^k-x^*||^2 = M'||x^k-x^*||^2 = (M'||x^k-x^*||)||x^k-x^*||$.

As x^k is in the ball B, $||x^k - x^*|| \leqslant r < \dfrac{1}{M'}$; hence $M'||x^k - x^*|| < 1$; so we have : $||x^{k+1} - x^*|| \leqslant ||x^k - x^*||$; x^{k+1} is also in the ball B. We have established that the whole sequence (x^k) is in B.

Let us set $\alpha = M'||x^0 - x^*||$. We have $||x^0 - x^*|| \leqslant r < \dfrac{1}{M'}$, from which $\alpha < 1$.

We also have: $M'||x^{k+1}-x^*|| \leqslant (M'||x^k-x^*||)^2$; we obtain by recurrence:

$$M'||x^k - x^*|| \leqslant (M'||x^0 - x^*||)^{2^k} = \alpha^{2^k},$$

or:
$$||x^k - x^*|| \leqslant \frac{\alpha^{2^k}}{M'}.$$

The sequence x^k converges to x^*.

Finally, the inequality $||x^{k+1} - x^*|| \leqslant M'||x^* - x^k||^2$ shows that convergence is quadratic. ◊

7.11 Exercice

Statement. We are interested in the minimum of the function f defined on $\mathbb{R}^2$ by:

$$f(x,y) = e^{x+y} + x^2 + 2y^2.$$

Apply three iterations of the method of the steepest descent with optimal step.

Solution. The fonction f is of class C^∞. Let us start by determining the gradient and Hessian matrix of f:

$$\nabla f(x,y) = \begin{pmatrix} e^{x+y} + 2x \\ e^{x+y} + 4y \end{pmatrix}, \qquad \nabla^2 f(x,y) = \begin{pmatrix} e^{x+y} + 2 & e^{x+y} \\ e^{x+y} & e^{x+y} + 4 \end{pmatrix}.$$

The determinant of the Hessian matrix (product of eigenvalues) and its trace (sum of eigenvalues) being strictly positive, the eigenvalues of $\nabla^2 f(x,y)$ are strictly positive and $\nabla^2 f(x,y)$ is positive definite: f is therefore strictly convex.

We deduce that every local minimum is global, and a necessary and sufficient condition for x^* to be a minimum is $\nabla f(x^*) = 0$. As otherwise f tends towards infinity at infinity, f admits a global minimum, which is sought by the steepest descent with optimal step method.

We start from $P_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}; \nabla f(P_0) = \begin{pmatrix} 1 \\ 1 \end{pmatrix}; d_0 = \begin{pmatrix} -1 \\ -1 \end{pmatrix}$.

We look for s such that $g(s) = f(P_0 + sd_0)$ is minimum:

$g(s) = f(-s, -s) = e^{-2s} + 3s^2$.

We minimize g for example by dichotomy and we find $s = 0.216$. D'où :

$$P_1 = P_0 + 0.216 d_0 = \begin{pmatrix} -0.216 \\ -0.216 \end{pmatrix}; \nabla f(P_1) = \begin{pmatrix} 0.216 \\ -0.216 \end{pmatrix}; d_1 = \begin{pmatrix} -0.216 \\ 0.216 \end{pmatrix}.$$

We find that d_1 is orthogonal to d_0. We set

$$\begin{aligned} g(s) &= f(P_1 + sd_1) \\ &= f(-0.216(1+s), -0.216(1-s)) \\ &= e^{-2\times 0.216} + (0.216)^2[(1+s)^2 + 2(1-s)^2] \\ &= e^{-2\times 0.216} + (0.216)^2 h(s) \end{aligned}$$

with $h(s) = [(1+s)^2 + 2(1-s)^2] = 3s^2 - 2s + 3$.

The minimum of h is reached for $s = 1/3$. From which :

$P_2 = \begin{pmatrix} -0.216(1+1/3) \\ -0.216(1-1/3) \end{pmatrix} = \begin{pmatrix} -0.288 \\ -0.144 \end{pmatrix}; \nabla f(P_2) = \begin{pmatrix} 0.0732 \\ 0.0732 \end{pmatrix};$

$d_2 = \begin{pmatrix} -0.0732 \\ -0.0732 \end{pmatrix}$ (again d_2 is orthogonal to d_1...).

We now consider:

$$\begin{aligned} g(s) &= f(-0.288 - 0,0732s, -0.144 - 0,0732s) \\ &= e^{-0.432 - 0.1464s} + (0.288 + 0,0732s)^2 + 2(0.144 + 0,0732s)^2. \end{aligned}$$

We minimize g by dichotomy and we find $s = 0.2339$, hence:

$P_3 = \begin{pmatrix} -0.305 \\ -0.161 \end{pmatrix}$.

We can continue this way to have more precision.

Chapter 8

Non linear optimization with constraints

8.1 Generalities

We consider functions g_i $(1 \leqslant i \leqslant m)$ and h_j $(1 \leqslant j \leqslant p)$ defined on $\mathbb{R}^n$ and with real values. We set $I = \{1, ..., m\}$ and $J = \{1, ..., p\}$. We consider the set X of the elements of $\mathbb{R}^n$ satisfying:

$$\begin{cases} \text{for } i \in I, & g_i(x) \leqslant 0 \\ \text{for } j \in J, & h_j(x) = 0. \end{cases}$$

If we consider a continuous function from $\mathbb{R}^n$ to $\mathbb{R}$, the inverse image by this function of a closed set of $\mathbb{R}$ is a closed set of $\mathbb{R}^n$. Moreover, the intersection of closed sets of $\mathbb{R}^n$ is a closed set of $\mathbb{R}^n$. Consequently, if all the functions g_i and h_j are continuous, the set X is a closed set of $\mathbb{R}^n$.

We now consider a function f defined on an open set O of $\mathbb{R}^n$ (we often meet $O = \mathbb{R}^n$) and with real values. **We assume that O contains X.**

We consider the problem (P):

$$\text{minimize } f(x) \text{ for } x \in X.$$

Adaptations to a constrained maximization problem are immediate.

The conditions $g_i(x) \leqslant 0$ and $h_j(x) = 0$ are the *constraints* of the problem (P). Every x element of X is called *feasible solution* and X is the *feasible domain* (we also said *feasible set*). If, for $i \in I$ and for $x \in X$, we have $g_i(x) = 0$, we say that the constraint g_i is *saturated* in x.

We assume in this chapter that the functions g_i $(i \in I)$, h_j $(j \in J)$ and f are of class C ^{1} on O and that the feasible domain is not empty.

Theorem 18. *If the feasible domain is bounded, then the problem (P) admits at least one optimal solution.*

Proof. With the hypotheses, the feasible domain is a non-empty closed bounded set of $\mathbb{R}^n$, that is, a compact of $\mathbb{R}^n$. The result follows immediately from the theorem 9.

Definition 12. *The function f is said to be coercive if, for every real M, there exists r such that for $x \in O$ with $||x|| \geqslant r$, we have $f(x) \geqslant M$ (that is, the function f tends towards infinity if x tends towards infinity while remaining in O).*

Theorem 19. *If the function f is coercive, the problem (P) has at least one optimal solution.*

Proof. Let x be a feasible solution. Since f is coercive, there is a ball B of $\mathbb{R}^n$ centered on the origin such that for all x' in O and not in B, $f(x') > f(x)$. Since the set $B \cap X$ is a closed bounded set of $\mathbb{R}^n$, the function f reaches its minimum on $B \cap X$ at a point x^*. This point is also an optimal solution for the problem (P). $\Diamond$

Definition 13. *We say that a direction d is admissible at a point $x^0 \in X$ if there is a function ϕ from $\mathbb{R}$ to$\mathbb{R}^n$ such that:*

1. *$\phi(0) = x^0$*
2. *for all $t > 0$ small enough, $\phi(t) \in X$*
3. *the right derivative of ϕ at the point 0 is d.*

Let $x^0 \in X$. We denote by $A(x^0)$ the set of admissible directions in x^0; we set $I_0(x^0) = \{i \in I$ satisfying $g_i(x^0) = 0\}$.

Proposition 20. *If d is an admissible direction in x^0, then:*

1. *for $i \in I_0(x^0)$, $d^t \nabla g_i(x^0) \leqslant 0$*
2. *for $j \in J, d^t \nabla h_j(x^0) = 0$.*

Proof. Let ϕ be a function corresponding to the definition 13. We apply Taylor formula of order 1.

1. If $g_i(x^0) = 0$, we have: $g_i(\phi(t)) = td^t\nabla g_i(x^0) + t\epsilon(t)$ where $\epsilon(t) \to 0$ when $t \to 0$. For $t > 0$ small enough, $g_i(\phi(t)) \leqslant 0$ and then: $d^t\nabla g_i(x^0) + \epsilon(t) \leqslant 0$, which gives the result by passing to the limit when $t \to 0$.

2. We have $h_j(\phi(t)) = h_j(x^0) + td^t\nabla h_j(x^0) + t\epsilon(t)$ where $\epsilon(t) \to 0$ when $t \to 0$.

   For $t > 0$ small enough, $h_j(\phi(t)) = 0$ and $h_j(x^0) = 0$; we therefore have for $t > 0$ small enough: $d^t\nabla h_j(x^0) + \epsilon(t) = 0$, which gives the result by passing to the limit when $t \to 0$. $\Diamond$

We denote by $B(x^0)$ the set of directions d fulfilling:

- for $i \in I_0(x^0)$, $d^t\nabla g_i(x^0) \leqslant 0$
- $j \in J, d^t\nabla h_j(x^0) = 0$.

The proposition 20 is rewritten: $A(x^0) \subseteq B(x^0)$.

Definition 14. *We say that the constraints are qualified in $x^0 \in X$ if any direction in $B(x^0)$ is the limit of a sequence of directions of $A(x^0)$.*

The following propositions give sufficient conditions for constraints to be qualified.

Proposition 21. *If:*

- *fonctions g_i are convex,*
- *fonctions h_j are linear,*
- *it exists $\tilde{x} \in X$ with,*

  $$\text{for all } i \in I, g_i(\tilde{x}) < 0$$
  $$\text{for all } j \in J, h_j(\tilde{x}) = 0$$

then the constraints are qualified in every point of X.

Proposition 22. *We assume that for $j \in J$, the functions h_j are linear. If, at the point $x^0 \in X$, all the gradients*

- *$\nabla g_i(x^0)$ for $i \in I_0(x^0)$*

- $\nabla h_j(x^0)$ *for* $j \in J$

are linearly independent, then the constraints are qualified in x^0.

Before proving these propositions, two lemmas are established:

Lemma 23. *It is assumed that for* $j \in J$, *the fonctions* h_j *are linear. Let* $x^0 \in X$ *and* d *be a direction verifying:*

- *for* $i \in I_0(x^0), d^t \nabla g_i(x^0) < 0$
- *for* $j \in J$, $d^t \nabla h_j(x^0) = 0$.

Then d *is an admissible direction in* x^0.

Proof. For $t \geqslant 0$, we set: $\phi(t) = x^0 + td$. We have $\phi(0) = x^0$ and $\phi'(0) = d$, in other words, points 1 and 3 of the definition of an admissible direction are satisfied.

For $j \in J$, since the functions h_j are supposed to be linear, we can write: $h_j(\phi(t)) = h_j(x^0) + td^t \nabla h_j(x^0)$. By hypothesis on x^0, we have $h_j(x^0) = 0$ and, by hypothesis on d, we have $d^t \nabla h_j(x^0) = 0$. Hence $h_j(\phi(t)) = 0$.

Moreover, for $i \in I_0(x^0)$, we can write:

$g_i(\phi(t)) = g_i(x^0) + t(d^t \nabla g_i(x^0) + \epsilon(t))$, where $\epsilon(t) \to 0$ when $t \to 0$.

From $g_i(x^0) = 0$ and $d^t \nabla g_i(x^0) < 0$, we deduce $g_i(\phi(t)) \leqslant 0$ for t positive enough small.

Thus, the direction d is admissible. ◊

Lemma 24. *We assume that for* $j \in J$, *the functions* h_j *are linear. Let* $x^0 \in X$. *If there is a direction* $\tilde{d}$ *such that:*

- *for* $i \in I_0(x^0)$, $\tilde{d}^t \nabla g_i(x^0) < 0$
- *for* $j \in J, \tilde{d}^t \nabla h_j(x^0) = 0$,

then the constraints are qualified in x^0.

Proof. Let $d \in B(x^0)$ and $\tilde{d}$ verifying the assumptions of the lemma.

For $\lambda \in [0, 1[$, we set $d_\lambda = \lambda d + (1 - \lambda)\tilde{d}$.

For $i \in I_0(x^0)$: $d_\lambda^t \nabla g_i(x^0) = \lambda d^t \nabla g_i(x^0) + (1 - \lambda)\tilde{d}^t \nabla g_i(x^0) < 0$.

For $j \in J$: $d_\lambda^t \nabla h_j(x^0) = \lambda d^t \nabla h_j(x^0) + (1 - \lambda)\tilde{d}^t \nabla h_j(x^0) = 0$.

The lemma 23 indicates that for all $\lambda \in [0, 1[$, d_λ is a qualified direction. Considering a sequence λ_n of numbers tending towards 1 by lower values, we obtain a sequence d_{λ_n} of admissible directions that tends towards d: this

shows that the constraints are qualified in x^0. $\Diamond$

Proof of the proposition 21. Let $\tilde{x}$ which verifies the hypotheses of the proposition and x^0 a point of X.

Using the convexity of g_i, we have for $i \in I_0(x_0)$:

$0 > g_i(\tilde{x}) \geqslant g_i(x^0) + (\tilde{x} - x^0)^{\mathrm{t}} \nabla g_i(x^0)$.

Hence, as $g_i(x^0) = 0, (\tilde{x} - x^0)^{\mathrm{t}} \nabla g_i(x^0) < 0$. We set $\tilde{d} = \tilde{x} - x^0$; so we have $\tilde{d}^{\mathrm{t}} \nabla g_i(x^0) < 0$.

For $j \in J$, the fonctions h_j being linear: $0 = h_j(\tilde{x}) = h_j(x^0) + \tilde{d}^{\mathrm{t}} \nabla h_j(x^0)$; from which we deduce: $\tilde{d}^{\mathrm{t}} \nabla h_j(x^0) = 0$.

We use the lemma 24: the constraints are qualified in x^0. Since x^0 is arbitrary, we conclude that the constraints are qualified at every point of X. $\Diamond$

Proof of the proposition 22. We consider (Q) and (R), the two linear optimization problems describe below:

$$(Q) \qquad \begin{array}{l} \text{Maximize } z = \sum_{i \in I_0(x^0)} \lambda_i \\ \text{with } \left\{ \begin{array}{l} \sum_{j \in J} \mu_j \nabla h_j(x^0) - \sum_{i \in I_0(x^0)} \lambda_i \nabla g_i(x^0) = 0 \\ \text{for } i \in I_0(x^0), \lambda_i \geqslant 0 \\ \text{for } j \in J, \mu_j \in \mathbb{R}^n \end{array} \right. \end{array}$$

$$(R) \qquad \begin{array}{l} \text{Minimize } w = 0 \\ \text{with } \left\{ \begin{array}{l} \text{for } i \in I_0(x^0), d^{\mathrm{t}} \nabla g_i(x^0) \leqslant -1 \\ \text{for } j \in J, d^{\mathrm{t}} \nabla h_j(x^0) = 0. \\ d \in \mathbb{R}^n \end{array} \right. \end{array}$$

We can easily check that the problems (Q) and (R) are dual one to each other.

The problem (Q) is feasible since the null solution is feasible; we show that it is bounded from above by 0. Suppose that it can take a strictly positive value. Then, in this solution, at least one λ_i $(i \in I_0(x^0))$ is non-zero and the vectors $\nabla g_i(x^0)$ $(i \in I_0(x^0))$ and $\nabla h_j(x^0)$ $(j \in J)$ are linearly dependent, which is contrary to the hypothesis. The maximum of the problem (Q) is therefore 0.

We use the fondamental theorem of duality for linear optimization 5: the problem (R) is feasible. $\tilde{d}$ is a feasible solution of (R). Then just use the lemma 24 to conclude. $\Diamond$

Remark. We can show that the proposition 22 is still correct if we assume that the functions h_j are linear.

An example of point where constraints are not qualified.

We consider in $\mathbb{R}^2$ the domain represented in the figure 8.1 defined by:

$$\begin{cases} y \leqslant x^3 \\ x \leqslant 1 \\ y \geqslant 0. \end{cases}$$

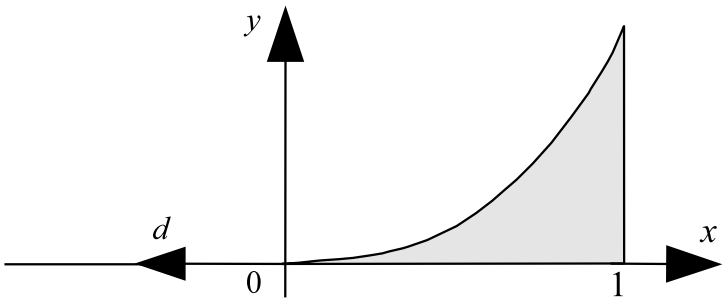

Figure 8.1: Unqualified constraints in (0, 0).

We set $g_1(x,y) = y - x^3$, $g_2(x,y) = x - 1$, $g_3(x,y) = -y$; the constraints can be written: $g_1(x,y) \leqslant 0$, $g_2(x,y) \leqslant 0$, $g_3(x,y) = -y$.

At point (0, 0), the constraints g_1 and g_3 are saturated.

$\nabla g_1(x,y) = \begin{pmatrix} -3x \\ 1 \end{pmatrix}$, $\nabla g_1(0,0) = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ and $\nabla g_3(0,0) = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$.

The direction $d = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$ fulfills $d^t \nabla g_1(0,0) = 0$ and $d^t \nabla g_3(0,0) = 0$, the direction d belongs to $B(0,0)$. However, the only admissible direction in (0, 0) is the direction $\begin{pmatrix} 1 \\ 0 \end{pmatrix}$. The direction d is not limit of a sequence of admissible directions.

Finally, we establish the following theorem:

Theorem 25. *We suppose that the problem admits a local minimum at a point x^* where the constraints are qualified. Then, if $d \in B(x^*)$:*

$$d^t \nabla f(x^*) \geqslant 0$$

(therefore, no admissible direction in x^* is a descent).

Proof. Let (d_k) be a sequence of admissible directions tending towards d and ϕ_k the function associated with d_k. Soit $t > 0$. We deduce:

$$f[\phi_k(t)] = f(x^*) + t d_k^{\mathrm{t}} \nabla f(x^*) + t\epsilon(t)$$

where $\epsilon(t) \to 0$ when $t \to 0$. If t is small enough: $f[\phi_k(t)] \geqslant f(x^*)$.

Then, we have: $t[d_k^{\mathrm{t}} \nabla f(x^*) + \epsilon(t)] \geqslant 0$ and so $d_k^{\mathrm{t}} \nabla f(x^*) + \epsilon(t) \geqslant 0$.

Going to the limit when t tends towards 0, we get $d_k^{\mathrm{t}} \nabla f(x^*) \geqslant 0$.

Going to the limit when k tends towards $+\infty$, we get $d^{\mathrm{t}} \nabla f(x^*) \geqslant 0$. $\Diamond$

8.2 Lagrange condition

We are interested here in the problem:

$$\begin{array}{l} \text{Minimize } f(x) \\ \text{with } \left\{ \begin{array}{ll} \text{for } j \in J, & h_j(x) = 0 \\ x \in \mathbb{R}^n & \end{array} \right. \end{array}$$

where the functions f and h_j $(j \in J)$ are of class C^1. The Lagrange condition, given in the following theorem, provides a necessary condition for an element of $\mathbb{R}^n$ to be a local minimum of (P).

Theorem 26. *Let x^* be a local minimum of the problem. It is assumed that the constraints are qualified in x^*. Then there are p real numbers μ_j $(j \in J)$ fulfilling:*

$$\nabla f(x^*) = \sum_{j \in J} \mu_j \nabla h_j(x^*).$$

Proof. Let us denote by E the subset of $\mathbb{R}^n$ generated by the vectors $\nabla h_j(x^*)$ $(j \in J)$ and $E^\perp$ the subspace orthogonal to E. We have:

$$\nabla f(x^*) = y + z. \text{ with } y \in E \text{ and } z \in E^\perp.$$

For $j \in J$, $(-z)^{\mathrm{t}} \nabla h_j(x^*) = 0$ since $-z$ belongs to $E^\perp$. Therefore, $-z$ belongs to $B(x^*)$; according to the theorem 25, it follows: $(-z)^{\mathrm{t}} \nabla f(x^*) \geqslant 0$.

But: $(-z)^{\mathrm{t}} \nabla f(x^*) = (-z)^{\mathrm{t}} y + (-z)^{\mathrm{t}} z = (-z)^{\mathrm{t}} z = -||z||^2$.

The inequality $-||z||^2 \geqslant 0$ results in $z = 0$. Hence the theorem. $\Diamond$

The theorem 27, a direct consequence of the theorem 29 demonstrated later, gives hypotheses for which the Lagrange condition is sufficient.

Theorem 27. *The Lagrange condition is sufficient when f is convex in an open set containing X and the h_j $(j \in J)$ are linear.*

8.3 Karush, Kuhn and Tucker conditions

We consider again the initial problem (P):

$$(P) \qquad \text{with} \quad \begin{cases} \text{Minimize } f(x) \\ \text{for } i \in I, g_i(x) \leqslant 0 \\ \text{for } j \in J, h_j(x) = 0 \\ x \in \mathbb{R}^n \end{cases}$$

where the functions f, g_i $(i \in I)$ and h_j $(j \in J)$ are assumed of class C^1. The following conditions, called *Karush, Kuhn and Tucker conditions*, give sufficient conditions of optimality and generalize the Lagrange condition:

Theorem 28. *It is assumed that the constraints are qualified in x^* and that x^* is a local minimum of the problem; then it exists:*

- $|I_0(x^*)|$ *positive or zero real numbers* λ_i *for* $i \in I_0(x^*)$
- p *real numbers* μ_j $(j \in J)$

satisfying:

$$\nabla f(x^*) = \sum_{j \in J} \mu_j \nabla h_j(x^*) - \sum_{i \in I_0(x^*)} \lambda_i \nabla g_i(x^*).$$

Remark. We notice that in the expression of $\nabla f(x^*)$, only the saturated constraints intervene.

Proof. First, we use the theorem 25.

The hypothesis $d \in B(x^*)$ can be written:

- for all $i \in I_0(x^*)$, $\sum_{k=1}^n \dfrac{\partial g_i}{\partial x_k}(x^*) d_k \leqslant 0$
- for all $j \in J$, $\sum_{k=1}^n \dfrac{\partial h_j}{\partial x_k}(x^*) d_k = 0$.

The conclusion $d^{\mathrm{t}} \nabla f(x^*) \geqslant 0$ in the case where x^* is a local minimum can be written: $\sum_{k=1}^n \dfrac{\partial f}{\partial x_k}(x^*) d_k \geqslant 0$. We set:

- for every $i \in I_0(x^*)$ and for every $k \in \{1, 2, ..., n\}$, $a_{ik} = -\dfrac{\partial g_i}{\partial x_k}(x^*)$
- for every $j \in J$ and every $k \in \{1, 2, ..., n\}$, $b_{jk} = \dfrac{\partial h_j}{\partial x_k}(x^*)$
- for every $k \in \{1, 2, ..., n\}$, $c_k = \dfrac{\partial f}{\partial x_k}(x^*)$.

With these notations, the theorem 25 can be written: for every $d \in \mathbb{R}^n$, if we have:

- for every $i \in I_0(x^*)$, $\sum_{k=1}^n a_{ik} d_k \geqslant 0$
- for every $j \in J$, $\sum_{k=1}^n b_{jk} d_k = 0$,

then $\sum_{k=1}^n c_k d_k \geqslant 0$.

We use the Farkas and Minkowski theorem which is the object of the exercise 6.6.4: it exists λ_i for $i \in I_0(x^*)$ and μ_j for $j \in J$ fulfilling:

- for $k \in \{1, ..., n\}$, $\sum_{i \in I_0(x^*)} a_{ik} \lambda_i + \sum_{j \in J} b_{jk} \mu_j = c_k$
- for $i \in I_0(x^*)$, $\lambda_i \geqslant 0$.

The first line can be written:

- for $k \in \{1, ..., n\}$, $\sum_{j \in J} \mu_j \dfrac{\partial h_j}{\partial x_k}(x^*) - \sum_{i \in I_0(x^*)} \lambda_i \dfrac{\partial g_i}{\partial x_k}(x^*) = \dfrac{\partial f}{\partial x_k}(x^*)$

or finally: $\sum_{j \in J} \mu_j \nabla h_j(x^*) - \sum_{i \in I_0(x^*)} \lambda_i \nabla g_i(x^*) = \nabla f(x^*)$.

With the positivity of the λ_i, we get the statement of Karush, Kuhn and Tucker theorem. ◊

We illustrate below two cases where there is no equality constraint; the conditions of Karush, Kuhn and Tucker then express that it is necessary that $\nabla f(x^*)$ be decomposed on $\{-\nabla g_i(x^*)(i \in I_0(x^*)\}$ with positive or zero coefficients.

Illustration.

⋆ *Case $n = 2, p = 0$ and one saturated inequality constraint*

We assume that we are in $\mathbb{R}^2$; we denote by x_1 and x_2 the coordinates of a point. We assume that only the constraint $g(x_1, x_2) \leqslant 0$ is saturated at the

point (x_1^*, x_2^*): $g(x_1^*, x_2^*) = 0$. The vector $-\nabla g(x_1^*, x_2^*)$ is perpendicular to the curve of equation $g(x_1, x_2) = 0$ and directed to the interior of the domain. If the vector $\nabla f(x_1^*, x_2^*)$ makes a non-zero angle with $-\nabla g(x_1^*, x_2^*)$, we can find a direction of descent for f directed to the interior of the domain and the point (x_1^*, x_2^*) is not a local minimum.

The figure 8.2 illustrates this case; we remind that a direction is a direction of descent if it makes at an obtuse angle with $\nabla f(x_1^*, x_2^*)$ and that an admissible direction makes an acute angle with $-\nabla g(x_1^*, x_2^*)$.

Only the case where $\nabla f(x_1^*, x_2^*)$ makes a null angle with $-\nabla g(x_1^*, x_2^*)$ is compatible with the local optimality of (x_1^*, x_2^*).

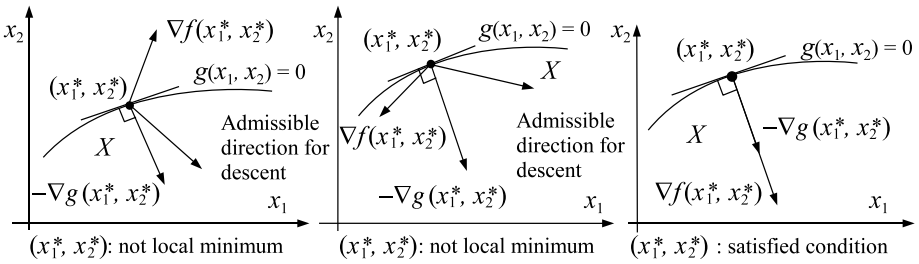

Figure 8.2: In the plane, only one saturated constraintDans le plan, une seule contrainte saturée.

⋆ *Case* $n = 2, p = 0$ *and two saturated inequality constraints.*

On the figure 8.3, so that no direction of descent is directed towards the domain, it is necessary that the vector $\nabla f(x_1^*, x_2^*)$ is in the sector formed by $-\nabla g_1(x_1^*, x_2^*)$ and $-\nabla g_2(x_1^*, x_2^*)$. This is the case in the figure.

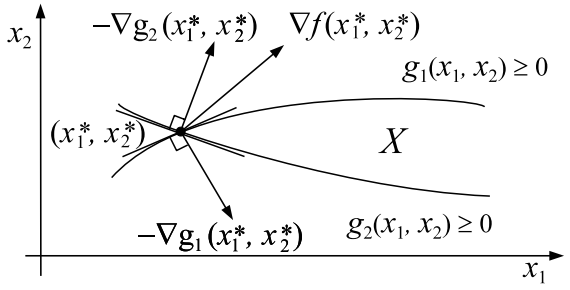

Figure 8.3: In the plane, two saturated constraints.

The theorem 29 gives hypotheses for which the Karush, Kuhn and Tucker conditions are sufficient for a local minimum.

Theorem 29. *It is assumed that the constraints are qualified at a point x^*. The Karush, Kuhn and Tucker conditions at the point x^* are sufficient to have a local minimum if there is a neighborhood of x^* in which we simultaneously have the functions f and g_i $(i \in I_0(x^*))$ convex and functions h_j linear $(j \in J)$.*

Proof. Suppose there are positive or zero real numbers λ_i $(i \in I_0(x^*))$ and real numbers μ_j $(1 \leqslant j \leqslant p)$ such that:

$$\nabla f(x^*) = \sum_{j \in J} \mu_j \nabla h_j(x^*) - \sum_{i \in I_0(x^*)} \lambda_i \nabla g_i(x^*).$$

We consider a ball B centered in x^* in which the functions f and g_i $(i \in I_0(x^*))$ are convex and the functions h_j linear $(j \in J)$. Let $x \in B \cap X$, we will show the inequality $f(x) \geqslant f(x^*)$, which will prove the theorem.

The convexity of f on B induces: $f(x) \geqslant f(x^*) + (x - x^*)^{\mathrm{t}} \nabla f(x^*)$. Using the Karush, Kuhn and Tucker conditions:

$f(x) \geqslant f(x^*) + \sum_{j \in J} \mu_j (x - x^*)^{\mathrm{t}} \nabla h_j(x^*) - \sum_{i \in I_0(x^*)} \lambda_i (x - x^*)^{\mathrm{t}} \nabla g_i(x^*)$.
For $j \in J$: $(x - x^*)^{\mathrm{t}} \nabla h_j(x^*) = h_j(x) - h_j(x^*) = 0$.
Let $i \in I_0(x^*)$; the function g_i being convex on B:

$g_i(x) \geqslant g_i(x^*) + (x - x^*)^{\mathrm{t}} \nabla g_i(x^*)$.
So we have:

$(x - x^*)^{\mathrm{t}} \nabla g_i(x^*) \leqslant g_i(x) - g_i(x^*)$.
We have $\lambda_i \geqslant 0$ and moreover, by hypothesis, $g_i(x^*) = 0$ and $g_i(x) \leqslant 0$, hence:

$\lambda_i (x - x^*)^{\mathrm{t}} \nabla g_i(x^*) \leqslant 0$.
We finally get $f(x) \geqslant f(x^*)$: f admits a local minimum at the point x^*. $\Diamond$

8.4 Descent method

In this part 8.4, we consider in the following problem:

$$\begin{array}{l} \text{Minimize } f(x) \\ \text{with, for } 1 \leqslant i \leqslant m,\ g_i(x) \leqslant 0. \end{array}$$

To try to solve this problem, we choose a starting point $x^0 \in X$ and we build iteratively a sequence x^k of X satisfying $f(x^{k+1}) < f(x^k)$, until the obtained value is considered to be satisfying.

When we are in x^k, we look for a descent direction d that does not make "immediately" go out of X. We then seek, by moving in the direction d, a point x^{k+1} of X better than x^k (for example, minimizing $f(x^k + sd)$ for $s > 0$, with the constraint that $x^k + sd$ belongs to X, if we know how to solve this new problem). We start again from x^{k+1} as long as a certain stopping criterion is not fulfilled.

To choose d, we can solve the problem:

$$\text{Minimize } d^t \nabla f(x^k)$$
$$\text{with } \begin{cases} d^t \nabla g_i(x^k) \leqslant 0 \text{ for any } i \text{ such that } g_i(x^k) = 0 \\ d^t d = 1. \end{cases}$$

The vector d is normed so as to choose, among the possible directions, a direction that maximizes the angle with $\nabla f(x^k)$. This gives the direction $d \in B(x^k)$ of greatest descent. We can replace the condition that d be of norm 1 by a condition imposing that d is of norm bounded from above by a fixed constant.

We can choose to replace the condition $d^t d = 1$ by the condition: $-1 \leqslant d_i \leqslant 1$ $(1 \leqslant i \leqslant n)$ to have a linear problem; in this case, the selected direction will not be exactly the direction of greatest descent.

The method, as just described, may encounter difficulties. Consider the example shown in figure 8.4. Any move in the direction d leeds to go out of X. We need a projection procedure so that x^{k+1} is in X, a procedure that can be schematically represented by the figure 8.5. Note, however, that this projection is not useful, for example, in the case where the constraints are linear.

Another possibility to overcome this difficulty is to replace the constraints $d^t \nabla g_i(x^k) \leqslant 0$ by $d^t \nabla g_i(x^k) \leqslant -\epsilon$, where ϵ is a positive parameter. So, instead of accepting a direction d which starts tangentially to the level surface $g_i(x) = 0$, which is allowed by the constraint $d^t \nabla g_i(x^k) \leqslant 0$, we impose to d to "enter", at least locally, in the half-space of equation $g_i(x) < 0$. The difficulty then lies in the choice of ϵ.

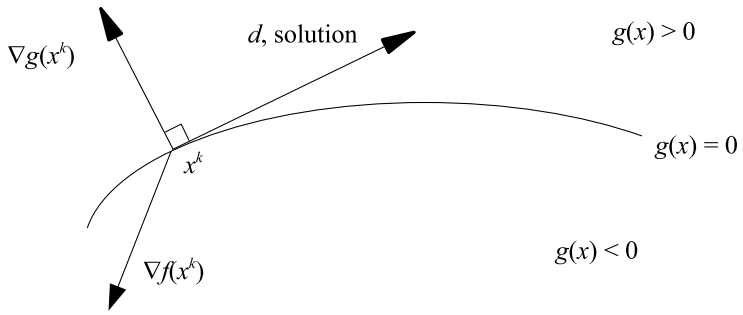

Figure 8.4: Exit of the feasible domain.

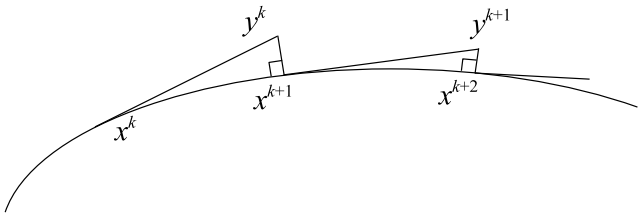

Figure 8.5: Projection on the feasible domain.

8.5 Case of convex functions

8.5.1 Généralities

It is assumed throughout this part that the definition domain O of f is a convex open set of $\mathbb{R}^n$ and that:

- the fonctions g_i $(1 \leqslant i \leqslant m)$ are convex on O,
- the fonctions h_j $(1 \leqslant j \leqslant p)$ are linear on O,
- the fonction f is convex on its domain of definition O.

Remarks.
1. If g is a convex function on O, the set of x satisfying $g(x) \leqslant 0$ is convex.
2. If h is a linear function on O, the set of x fulfilling $h(x) \leqslant 0$ is a hyperplane, and therefore is convex (and concave).
3. The intersection of convex domains being convex, the feasible domain X of the problem (P), defined by:

$$X = \{x \in O \text{ fulfilling } g_i(x) \leqslant 0 \text{ for } 1 \leqslant i \leqslant m \text{ and } h_j(x) = 0 \text{ for } 1 \leqslant j \leqslant p\}$$

is convex.

Theorem 30. *If f is strictly convex, the problem (P) admits at most an optimal solution.*

Proof. Suppose there are in X two optimal solutions x and y; we have in particular $f(x) = f(y)$. Let us put $z = \dfrac{x+y}{2}$. The convexity of X implies that $z \in X$ and the strict convexity of f implies the inequality $f(z) < \dfrac{f(x)+f(y)}{2} = f(x)$, a contradiction with the supposed optimality of x. ◊

From theorems 18 and 30, we deduce:

Theorem 31. *If the feasible domain is bounded and if f is strictly convex, the problem (P) admits a unique optimal solution.*

From theorem 19 and 30, we deduce:

Theorem 32. *If f is strictly convex and coercive, the problem (P) admits a unique optimal solution.*

Theorem 33. *Suppose we have a local minimum at a point x^* where the constraints are qualified. Then the problem (P) admits a global minimum at the point x^*.*

Proof. Let $x \in X$.

We define a function ψ on the interval $[0, 1]$ by: $\psi(t) = f[x^* + t(x - x^*)]$.

We have: $\psi(0) = f(x^*)$ and $\psi(1) = f(x)$. In addition, the convexity of f results in the convexity of ψ.

Otherwise: $\psi'(0) = (x - x^*)^{\mathrm{t}} \nabla f(x^*)$.

The direction $d = x - x^*$ belongs to $B(x^*)$ because, if it were not, we would go out the feasible domain following the direction d from x^*. The theorem 25 shows the inequality $\psi'(0) \geqslant 0$. Since the function ψ is convex on $[0, 1]$, $\psi'(0) \geqslant 0$ implies $\psi(1) \geqslant \psi(0)$, that is: $f(x) \geqslant f(x^*)$. Therefore, x^* is a global minimum of (P). ◊

We can now give conditions for the of Karush, Kuhn and Tucker conditions be sufficient for a global minimum by leaning on the theorems 29 and 33.

Theorem 34. *It is assumed that the assumptions fixed at the beginning of the part 8.5.1 are fulfilled. If the Karush, Kuhn, and Tucker conditions are satisfied at a point x^* where the constraints are qualified, then x^* is a global minimum of (P). Moreover, if f is strictly convex, x^* is the only point where (P) reaches the global minimum.*

Proof. According to theorem 29, the problem (P) reaches a local minimum in x^*. According to the theorem 33, x^* is a global minimum of (P). If, moreover, f is strictly convex, theorem 30 makes possible to conclude that x^* is the only global minimum of (P). $\Diamond$

8.5.2 Linearisation: introduction

We consider a function f defined on an open set O of $\mathbb{R}^n$ with real values and of class C^1. We are interested in the problem of finding the minimum of f on a convex closed domain X included in O. We will "linearize" f, that is, approach f by its Taylor development of order 1, and that in a sequence of points built using this linearization. This leads to a simple algorithm described below, whose limitations, however, will be noted.

We consider the following algorithm:

- $x^0 \leftarrow$ any point of X
- $k \leftarrow 0$
- repeat
  - $\star$ $x^{k+1} \leftarrow$ a point that minimizes $f(x^k) + (x - x^k)^{\mathrm{t}} \nabla f(x^k)$ on X
  - $\star$ $k \leftarrow k + 1$

  until a stop test to be specified is fulfilled.

Remarks.
1) The point x^{k+1} is also a point that minimizes $x^{\mathrm{t}} \nabla f(x^k)$ on X since this function differs from the function $f(x^k) + (x - x^k)^{\mathrm{t}} \nabla f(x^k)$ by a constant.
2) If the domain X is a polyhedron, the determination of x^{k+1} is a linear optimization problem.

Let us apply this algorithm to the following problem in $\mathbb{R}^2$, denoted by (P_0):

$$(P_0) \qquad \begin{array}{l} \text{Minimize } f(x_1, x_2) = (x_1 - 3)^2 + (x_2 - 5)^2 \\ \text{with the constraints: } \begin{cases} x_2 - 2x_1 \leqslant 0 \\ 2x_1 + x_2 - 20 \leqslant 0 \\ -2x_1 + 3x_2 - 4 \leqslant 0 \\ x_2 \geqslant 0. \end{cases} \end{array}$$

Figure 8.6 illustrates this problem. The objective function is constant on circles centered on the point C of coordinates (3, 5). We can therefore anticipate that it is minimum for the point of the domain closest to the center of C, that is, the point $(\frac{49}{13}, \frac{50}{13})$. We denote by X the feasible domain, grayed on the figure.

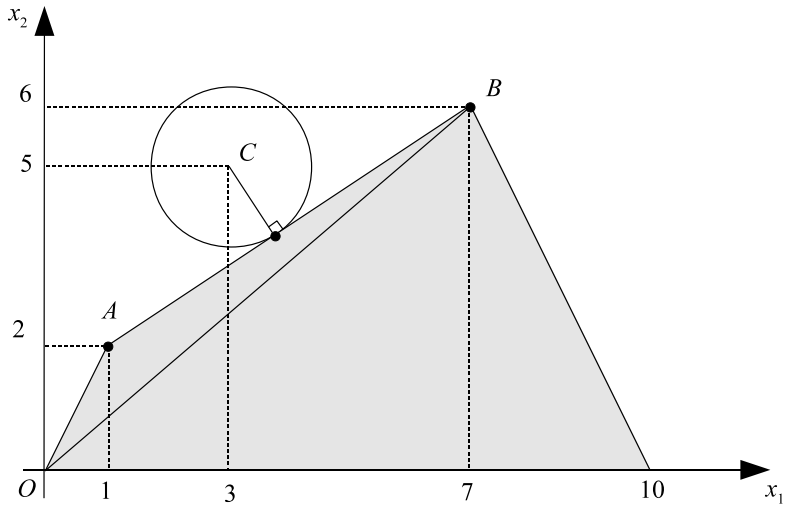

Figure 8.6: Cycling with linearisation.

We have: $\nabla f(x_1, x_2) = \begin{pmatrix} 2x_1 - 6 \\ 2x_2 - 10 \end{pmatrix}$.

We start the algorithm from the origin O, with $\nabla f(0,0) = \begin{pmatrix} -6 \\ -10 \end{pmatrix}$. We are looking for the minimum on X of $-6x_1 - 10x_2$. The considerations developed in the chapter **??** show that the minimum is reached in one of the four vertices of X; it is easy to show that this is the point $B = (7.6)$.

We start again from the vertex B, with $\nabla f(7,6) = \begin{pmatrix} 8 \\ 2 \end{pmatrix}$. We are looking for the minimum of $8x_1 + 2x_2$ on X. It is reached in one of the four vertices of X; we check that it is the point $A = (1,2)$.

We start again from A, with $\nabla f(1,2) = \begin{pmatrix} -4 \\ -6 \end{pmatrix}$. We are looking for the minimum of $-4x_1 - 6x_2$ on X. We find the vertex B. If we continue the method, we alternate between A and B: the method does not converge.

8.5.3 Linéarisation: Frank and Wolfe method

Frank and Wolfe method applies in the case where X is convex and compact.

The algorithm is as follows:

- $x^0 \leftarrow$ any point of X
- $k \leftarrow 0$
- repeat
  - $\star$ $\tilde{x}^k \leftarrow$ a point that minimize $x^{\mathrm{t}} \nabla f(x^k)$ sur X
  - $\star$ $x^{k+1} \leftarrow$ a point that minimizes f on the segment $[x^k, \tilde{x}^k]$
  - $\star$ $k \leftarrow k+1$

  until a stop test to be specified is satisfied.

In particular, the following proposition shows that the method stops if, for a certain index k, we have $x^{k+1} = x^k$.

Proposition 35. *If, in Franck and Wolfe algorithm, we have $x^{k+1} = x^k$, then the problem admits a global minimum in x^k.*

Proof. Let $x \in X$. The convexity of the function f implies:

$$f(x) - f(x^k) \geqslant (x - x^k)^{\mathrm{t}} \nabla f(x^k).$$

The choice of $\tilde{x}^k$ gives: $x^{\mathrm{t}} \nabla f(x^k) \geqslant (\tilde{x}^k)^{\mathrm{t}} \nabla f(x^k)$. From which $f(x) - f(x^k) \geqslant (\tilde{x}^k - x^k)^{\mathrm{t}} \nabla f(x^k)$.

Let us set also, for $t \in [0,1]$: $\phi(t) = f(x^k + t(\tilde{x}^k - x^k))$.

The minimum of f on the segment $[x^k, \tilde{x}^k]$ is obtained in x^{k+1}, that is to say in x^k. The function ϕ thus reaches its minimum for $t = 0$; in addition, f being convex, the function ϕ is unimodal. This results in: $\phi'(0) \geqslant 0$. Moreover: $\phi'(0) = (\tilde{x}^k - x^k)^{\mathrm{t}} \nabla f(x^k)$. We get: $f(x) - f(x^k) \geqslant 0$; x^k is thus a global minimum of f on X. $\Diamond$

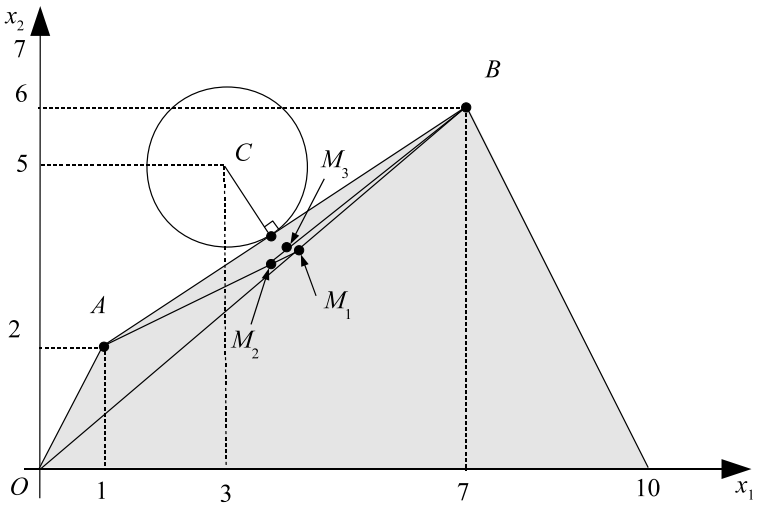

Figure 8.7: Resolution of (P_0) by the Frank and Wolfe method.

Let us apply this method to the previous problem (P_0) from the origin O: $(0,\ 0)$. The progress of the algorithm is illustrated by the figure 8.7.

We have: $\nabla f(x_1, x_2) = \begin{pmatrix} 2x_1 - 6 \\ 2x_2 - 10 \end{pmatrix}$.

Step 1.

We start from the point $M_0 = (0, 0)$.

We look for the minimum on X of $(x_1, x_2)^{\mathrm{t}} \nabla f(0, 0) = -6x_1 - 10x_2$. The point that reaches this minimum is the point $B = (7.6)$.

We look for the minimum of f on the segment $[O, B]$. We parametrize this segment by $x_1 = 7t, x_2 = 6t$ $(0 \leqslant t \leqslant 1)$ and we set: $\phi_1(t) = f(7t, 6t)$, which gives:

$$\begin{aligned} \phi_1(t) &= (7t - 3)^2 + (6t - 5)^2 \\ &= 49t^2 - 42t + 9 + 36t^2 - 60t + 25 \\ &= 85t^2 - 102t + 34. \end{aligned}$$

Therefore: $\phi_1'(t) = 170t - 102.$

The minimum of ϕ_1 is obtained for $t_1 = \dfrac{3}{5} = 0.6$.

The function f thus reaches its minimum on the segment $[O, B]$ at the point: $M_1 = (7 \times 0.6, 6 \times 0.6) = (4.2, 3.6)$.

Step 2.

We start from the point $M_1 = (4.2, 3.6)$.

We look for the minimum on X of $(x_1, x_2)^{\mathrm{t}} \nabla f(4.2, 3.6) = 2.4x_1 - 2.8x_2$. The point that reaches this minimum is the point $A = (1, 2)$.

The function $l_2(x_1, x_2) = 2.4x_1 - 2.8x_2$ has its minimum on X at the point $A = (1, 2)$.

We look for the minimum of f on the segment $[M_1, A]$.

We parametrize this segment by: $\begin{cases} x_1 = t + 4.2(1-t) \\ x_2 = 2t + 3.6(1-t) \end{cases} (0 \leqslant t \leqslant 1)$

or: $\begin{cases} x_1 = -3.2t + 4.2 \\ x_2 = -1.6t + 3.6 \end{cases} (0 \leqslant t \leqslant 1)$.

We set: $\phi_2(t) = f(-3.2t + 4.2\,; -1.6t + 3.6)$, which gives:

$$\begin{aligned} \phi_2(t) &= (-3.2t + 1.2)^2 + (-1.6t - 1.4)^2 \\ &= 10.24t^2 - 7.68t + 1.44 + 2.56t^2 + 4.48t + 1.96 \\ &= 12.8t^2 - 3.2t + 3.4. \end{aligned}$$

Hence $\phi_2'(t) = 25.6t - 3.2$.

The minimum of ϕ_2 is obtained for $t = \dfrac{3.2}{25.6} = 0.125$.

The function f thus reaches its minimum on the segment $[M_1, A]$ at the point $M_2 = (-0.125 \times 3.2 + 4.2, -0.125 \times 1.6 + 3.6) = (3.8, 3.4)$.

Step 3.

We start from the point $M_2 = (3.8, 3.4)$.

We look for the minimum on X of $(x_1, x_2)^{\mathrm{t}} \nabla f(3.8, 3.4) = 1.6x_1 - 3.2y_2$. The point that reaches this minimum is the point $B = (7, 6)$.

We look for the minimum of f on the segment $[M_2 B]$.

We parametrize this segment by: $\begin{cases} x_1 = 3.8t + 7(1-t) \\ x_2 = 3.4t + 6(1-t) \end{cases} (0 \leqslant t \leqslant 1)$

or: $\begin{cases} x_1 = -3.2t + 7 \\ x_2 = -2.6t + 6 \end{cases} (0 \leqslant t \leqslant 1)$.

We set: $\phi_3(t) = f(-3.2t + 7, -2.6t + 6)$, which gives:

$$\begin{aligned} \phi_3(t) &= (-3.2t + 4)^2 + (-2.6t + 1)^2 \\ &= 10.24t^2 - 25.6t + 16 + 6.76t^2 - 5.2t + 1 \\ &= 17t^2 - 30.8t + 17. \end{aligned}$$

Hence: $\phi_3'(t) = 34t - 30.8$.

The minimum of ϕ_3 is obtained for $t = \dfrac{30.8}{34} \simeq 0.9059$.

The function f thus reaches its minimum on the segment $[M_2, B]$ at the point $M_3 = (-\dfrac{30.8}{34} \times 3.2 + 7, -\dfrac{30.8}{34} \times 2.6 + 6) \simeq (4.10, 3.64)$.

We can continue this way. The theorem 36 below will show that the sequence of points M_k converges to the global minimum of the problem.

However, we could have chosen another starting point. For example, let us restart the algorithm by starting at the point $M'_0 = A$.

Step 1.

We start from the point $A = (1, 2)$.

We look for the minimum on X of $(x_1, x_2)^{\mathrm{t}} \nabla f(1, 2) = -4x_1 - 6x_2$. The point that reaches this minimum is the point $B = (7, 6)$.

We look for the minimum of f on the segment $[A, B]$.

We parametrize this segment by: $\begin{cases} x_1 = t + 7(1-t) \\ x_2 = 2t + 6(1-t) \end{cases} (0 \leqslant t \leqslant 1)$

or $\begin{cases} x_1 = -6t + 7 \\ x_2 = -4t + 6 \end{cases} (0 \leqslant t \leqslant 1)$.

We set: $\phi_1(t) = f(-6t + 7, -4t + 6)$, which gives:

$$\begin{aligned} \phi_1(t) &= f(-6t + 7, -4t + 6) \\ &= 36t^2 - 48t + 16 + 16t^2 - 8t + 1 \\ &= 52t^2 - 56t + 17. \\ \text{Hence: } \phi'_1(t) &= 104t - 56. \end{aligned}$$

The minimum of ϕ_1 is obtained for $t = \dfrac{56}{104} = \dfrac{7}{13} \simeq 0.538$.

The function f thus reaches its minimum on the segment $[A, B]$ at the point $M'_1 = (-\dfrac{7}{13} \times 6 + 7, -\dfrac{7}{13} \times 4 + 6) = (\dfrac{49}{13}, \dfrac{50}{13}) \simeq (3.769, 3.846)$.

Step 2.

We start from the point $M'_1 = (\dfrac{49}{13}, \dfrac{50}{13})$.

We look for the minimum on X of:

$$(x_1, x_2)^{\mathrm{t}} \nabla f(\frac{49}{13}, \frac{50}{13}) = \frac{20}{13} x_1 - \frac{30}{13} x_2 = \frac{10}{13}(2x_1 - 3x_2).$$

The point that reaches this minimum is the point $B = (7, 6)$.

The previous function is constant on the segment $[A, B]$ (its value is $-\frac{40}{3}$) and its minimum on X is obtained on this whole segment. Any point of this segment can be chosen as a minimum; we choose the point M_1'.

We look for the minimum of f on the segment $[M_1', M_1']$, reduced to a point: the sequence of points (M_k') is stationary (in M_1') and the algorithm ends. We can immediately check that the condition of Karush, Kuhn and Tucker are fulfilled at the point M_1'. The condition being sufficient here, the optimum has been obtained.

Theorem 36. *Let f be a function defined on a convex open set O of $\mathbb{R}^n$ with real values, of class C^1; we assume that f is strictly convex and that X is a compact convex polyhedron of $\mathbb{R}^n$ included in O. Frank and Wolfe method applied to the problem (P) of minimization of f over X converges to the single global minimum of (P).*

Remark. X being a polyhedron, we can limit ourselves, when we look for $\tilde{x}^k$, to the vertices of X.

We first establish the following lemma.

Lemma 37. *With the hypotheses of the theorem, let $(x^k)_{k\in\mathbb{N}}$ be a sequence of points of X such that the sequence $f(x^k)_{k\in\mathbb{N}}$ is decreasing. Suppose that the sequence $(x^k)_{k\in\mathbb{N}}$ has a subsequence $(x^k)_{k\in K\subset\mathbb{N}}$ converging to a global minimum of f. Then the sequence $(x^k)_{k\in\mathbb{N}}$ also converges to this global minimum.*

Proof of the lemma. Let x^* be the limit of the subsequence $(x^k)_{k\in K}$; assume that the sequence $(x^k)_{k\in\mathbb{N}}$ does not converge to x^*. Then, there exists $\epsilon > 0$ such that, for every $N \in \mathbb{N}$, there exists $k \geqslant N$ with $||x^k - x^*|| \geqslant \epsilon$; we deduce that there exists an infinite subsequence $(x^k)_{k\in U\subset\mathbb{N}}$ of the sequence $(x^k)_{k\in\mathbb{N}}$ satisfying, for all $k \in U, ||x^k - x^*|| \geqslant \epsilon$. Since X is a compact, we can extract from the sequence $(x^k)_{k\in U}$ a convergent subsequence $(x^k)_{k\in V\subset U}$; let y^* be the limit of this subsequence. By going to the limit, we have $||y^* - x^*|| \geqslant \epsilon$ and so $y^* \neq x^*$.

Since the function f is continuous, the sequences $(f(x^k))_{k\in K}$ and $(f(x^k))_{k\in V}$ are respectively convergent to $f(x^*)$ and $f(y^*)$. The point x^* giving a global minimum of f, we have: $f(y^*) \geqslant f(x^*)$. Suppose we have $f(y^*) > f(x^*)$. There exists $k_0 \in K$ such that $f(x^{k_0}) < f(y^*)$. Let $k \in V$ satisfy $k \geqslant k_0$.

Since the sequence $f(x^k)$ is decreasing, $f(x^k) \leqslant f(x^{k_0})$, this which implies: $f(x^k) < f(y^*)$, a contradiction with the fact that the sequence $(f(x^k))_{k\in V}$ converges by decreasing to $f(y^*)$.

We thus have: $f(y^*) = f(x^*)$, which is impossible since the function f has a single global minimum on X (see the theorem 31). $\Diamond$

Proof of the theorem. The theorem 31 already shows the existence and uniqueness of a global minimum. If the sequence built by the method becomes stationary after a finite number of steps, the proposition 35 shows that this stationary point is the global minimum.

We assume that this is not the case and we denote by $(x^k)_{k\in\mathbb{N}}$ the sequence built by Frank and Wolfe method. Since this sequence is in X which is compact, it has a convergent subsequence $(x^k)_{k\in K\subset\mathbb{N}}$; denote by x^* the limit of this sequence. The sequence $(\tilde{x}^k)_{k\in K}$ obtained by the linearization takes each of its values in one of the vertices of the polyhedron; the number of vertices of the polyhedron being finite, we can extract from the sequence $(x^k)_{k\in K}$ a subsequence $(x^k)_{k\in H\subset K}$ such that the sequence $(\tilde{x}^k)_{k\in H}$ is constant; we name $\tilde{x}$ this constant value.

The sequence $(x^k)_{k\in H}$ converges to x^*; let us show that x^* constitutes the global minimm of the problem (P).

Let $t \in [0,1]$ and $k \in H$. The linearization in x^k takes its minimum in $\tilde{x}^k = \tilde{x}$; the method then looks for the minimum of f on the segment $[x^k, \tilde{x}^k] = [x^k, \tilde{x}]$ and obtains the point x^{k+1}; we have:

$f(x^k + t(\tilde{x} - x^k)) \geqslant f(x^{k+1})$.

Let $h(k) \in H$ be among the indices $h \in H$ satisfying $h \geqslant k+1$; the sequence $f(x^k)_{k\in\mathbb{N}}$ being decreasing by construction of x^k:

$f(x^k + t(\tilde{x} - x^k)) \geqslant f(x^{h(k)+1})$.

By going to the limit for k in H that tends to infinity, we get:

$$f(x^* + t(\tilde{x} - x^*)) \geqslant f(x^*).$$

We write the Taylor formula of the function $t \to f(x^* + t(\tilde{x} - x^*))$ in the neighborhood of 0, at order 1:

$$f(x^* + t(\tilde{x} - x^*)) = f(x^*) + t(\tilde{x} - x^*)^{\mathrm{t}}\nabla f(x^*) + t\epsilon(t)$$

where $\epsilon(t)$ tends towards 0 when t tends towards 0.

Using the inequality obtained above, it comes:

$$t(\tilde{x} - x^*)^{\mathrm{t}}\nabla f(x^*) + t\epsilon(t) \geqslant 0.$$

or, for $t > 0$: $(\tilde{x} - x^*)^{\mathrm{t}}\nabla f(x^*) + \epsilon(t) \geqslant 0$.

When t tends towards 0, we obtain: $(\tilde{x} - x^*)^{\mathrm{t}}\nabla f(x^*) \geqslant 0$, that we can write:

$$\tilde{x}^{\mathrm{t}} \nabla f(x^*) \geqslant (x^*)^{\mathrm{t}} \nabla f(x^*). \quad (1)$$

Let $k \in H$ and $x \in X$. By construction of $\tilde{x}^k = \tilde{x}$, we have:

$$x^{\mathrm{t}} \nabla f(x^k) \geqslant \tilde{x}^{\mathrm{t}} \nabla f(x^k).$$

By going to the limit for k in H that tends to infinity, we get:

$$x^{\mathrm{t}} \nabla f(x^*) \geqslant \tilde{x}^{\mathrm{t}} \nabla f(x^*). \quad (2)$$

Using the inequalities (1) and (2), we obtain:

$$x^{\mathrm{t}} \nabla f(x^*) \geqslant (x^*)^{\mathrm{t}} \nabla f(x^*),$$

or: $(x - x^*)^{\mathrm{t}} \nabla f(x^*) \geqslant 0$.

The function f being convex, it comes: $(x - x^*)^{\mathrm{t}} \nabla f(x^*) \leqslant f(x) - f(x^*)$.

We now have $f(x) - f(x^*) \geqslant 0$, which shows that x^* is the optimal solution of the problem (P).

The lemma 37 shows that $(x^k)_{k \in \mathbb{N}}$ also converges to x^*. $\Diamond$

8.5.4 Linéarisation : Kelley cutting-plane method

The Kelley cutting-plane method searches the minimum of f on a convex domain. The problem is solved using linear optimization.

Remark. Consider the problem:

$$\begin{array}{l} \text{Minimize } f(x) \\ \text{with, for } 1 \leqslant i \leqslant m,\ g_i(x) \leqslant 0 \end{array}$$

where the functions f and $g_i, 1 \leqslant i \leqslant m$, are convex. This problem can be solved by adding a real variable y in the following form:

$$\begin{array}{l} \text{Minimize } y \\ \text{with the constraints: } \left\{ \begin{array}{l} \text{for } 1 \leqslant i \leqslant m,\ g_i(x) \leqslant 0 \\ f(x) - y \leqslant 0. \end{array} \right. \end{array}$$

Indeed, at the optimum, we will have $y = f(x)$ since y only intervenes in the constraint $f(x) - y \leqslant 0$. Minimizing y is equivalent to minimizing $f(x)$.

We thus reduce the problem to the optimization of a linear function on a convex domain.

On the basis of the above remark, Kelley cutting-plane method will be described as the minimization of a linear function on a domain defined by inequalities of type $g_i(x) \leqslant 0$ where the functions g_i are convex:

$$(P) \qquad \begin{array}{l} \text{Minimize } f(x) \\ \text{with, for } 1 \leqslant i \leqslant m,\ g_i(x) \leqslant 0 \end{array}$$

where the function f is linear and the functions $g_i, 1 \leqslant i \leqslant m$, are convex.

We denote by X the feasible domain .

At each step k, the algorithm includes X in a polyhedron Q_k and looks for the minimum of f in Q_k; we thus resolve a linear optimization problem. We make sure to always have $Q_{k+1} \subset Q_k$. The step k is described below.

Initialization of the algorithm.

We choose any point x^0 in $\mathbb{R}^n$.

We consider the polyhedron Q_0 defined by:

$$\text{for } 1 \leqslant i \leqslant m, g_i(x^0) + (x - x^0)^{\mathrm{t}} \nabla g_i(x^0) \leqslant 0.$$

Let $x \in X$ and i be between 1 and m. We have $g_i(x) \leqslant 0$ and, since the function g_i is convex:

$$g_i(x) \geqslant g_i(x^0) + (x - x^0)^{\mathrm{t}} \nabla g_i(x^0).$$

So : $g_i(x^0) + (x - x^0)^{\mathrm{t}} \nabla g_i(x^0) \leqslant 0$. Therefore we have : $X \subset Q_0$.

We solve the problem of minimizing the function f on the domain Q_0. This is a linear optimization problem that we know how to solve with the simplex algorithm.

If the linear optimization problem is not bounded, the initialization encounters a problem. We then make the domain bounded by choosing a "large" value M and adding the $2n$ inequalities $-M \leqslant x_i \leqslant M$ for $1 \leqslant i \leqslant n$. If, during the method generating the sequence x^k, a point of the sequence saturates one of these added constraints, we start from the initialization with a greater value of M, for example by doubling the previous value of M. We can notice that, if we know at first that X is bounded, it may be interesting to add at the beginning such constraints by ensuring that X is strictly included in the block defined by the added constraints. One can also, when the initialization has led to an unbounded problem, try another starting point.

Suppose now that the linear optimization problem admits a solution, denoted it by x^1. We finish the initialization by $k \leftarrow 1$.

Description of the step k ($k \geqslant 1$).

The previous step defined a new point x^k which gives the optimum of the problem on the polyhedron Q_{k-1}.

If x^k belongs to X, x^k gives the solution of the problem since the inclusion $X \subset Q_{k-1}$ gives the following relations: $f(x^k) = \min_{Q_{k-1}} f \leqslant \min_X f \leqslant f(x^k)$.

The algorithm ends. Suppose that x^k does not belong to X.

There is at least one of the m constraints of the problem that is not satisfied. We choose a constraint, g_{i_k}, which maximizes $g_i(x^k)$: for $1 \leqslant i \leqslant m$, $g_i(x^k) \leqslant g_{i_k}(x^k)$, which results in $g_{i_k}(x^k) > 0$.

We consider the inequality $g_{i_k}(x^k) + (x - x^k)^{\mathrm{t}} \nabla g_{i_k}(x^k) \leqslant 0$ which defines a half-space E_k. As for initialization, the X domain is included in the hyperplan E_k. On the other hand, the point x^k does not belong to E_k: the hyperplane which has the equation $g_{i_k}(x^k) + (x - x^k)^{\mathrm{t}} \nabla g_{i_k}(x^k) = 0$ separates the point x^k from the domain X.

We consider the polyhedron $Q_k = Q_{k-1} \cap E_k$.

We solve the problem of minimizing the function f on the domain Q_k. The problem is bounded from below since the minimum on Q_k is bounded from below by the minimum on Q_{k-1}. Let x^{k+1} be a point that reaches the minimum. We end the step by incrementing k by 1.

Illustration.

We consider in $\mathbb{R}^2$ the problem:

$$\begin{array}{l} \text{Minimize } f(x_1, x_2) = -x_1 + x_2 \\ \text{with } \left\{ \begin{array}{l} x_1^2 + (x_2 + 3)^2 \leqslant 10 \\ x_1^2 + (x_2 - 2)^2 \leqslant 5. \end{array} \right. \end{array}$$

We set: $g_1(x_1, x_2) = x_1^2 + (x_2 + 3)^2 - 10$ and $g_2(x_1, x_2) = x_1^2 + (x_2 - 2)^2 - 5$. We deduce:

$$\nabla g_1(x_1, x_2) = \begin{pmatrix} 2x_1 \\ 2x_2 + 6 \end{pmatrix}, \ \nabla g_2(x_1, x_2) = \begin{pmatrix} 2x_1 \\ 2x_2 - 4 \end{pmatrix}.$$

The domain is illustrated by the figure 8.8.

We try to initialize the method from the point $O = (0, 0)$. We obtain:

$g_1(0,0) + ((x_1, x_2) - (0,0))^{\mathrm{t}} \nabla g_1(0,0) = 6x_2 - 1$

and $g_2(0,0) + ((x_1, x_2) - (0,0))^{\mathrm{t}} \nabla g_2(0,0) = -4x_2 - 1$.

The domain Q_1 is then defined by: $x_2 \leqslant 1/6$ and $x_2 \geqslant 1/4$; the function f is not bounded from below in this domain.

We try another starting point, the point $M_0 = (2, -1)$.

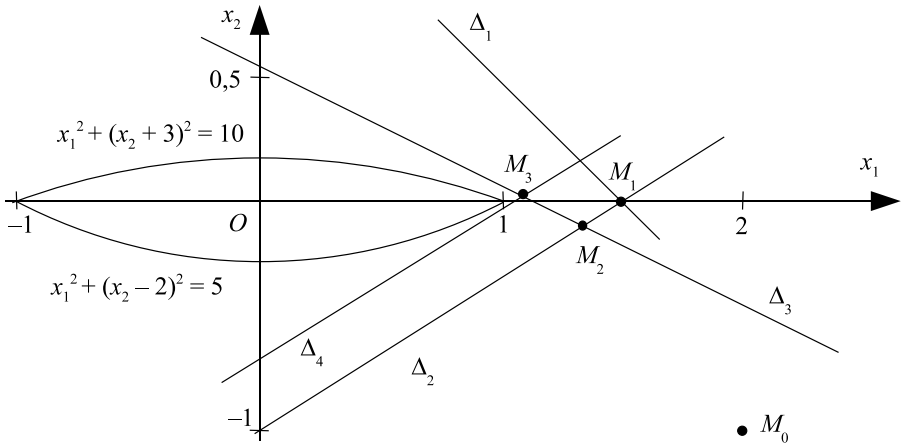

Figure 8.8: Kelley cutting-plane method.

Initialization.

$g_1(2,-1) + ((x_1,x_2)-(2,-1))^{\mathrm{t}}\nabla g_1(2,-1) = 4x_1 + 4x_2 - 6$

$g_2(2,-1) + ((x_1,x_2)-(2,-1))^{\mathrm{t}}\nabla g_2(2,-1) = 4x_1 - 6x_2 - 6.$

We denote by Δ_1 the line of equation $4x_1 + 4x_2 - 6 = 0$ and Δ_2 the line of equation $4x_1 - 6x_2 - 6 = 0$.

We then solve the problem:

$$\text{Minimize } f(x_1,x_2) = -x_1 + x_2$$
$$\text{with } \begin{cases} 4x_1 + 4x_2 \leqslant 6 \\ 4x_1 - 6x_2 \leqslant 6. \end{cases}$$

The solution is $M_1 = (\frac{3}{2}, 0)$.

Step 1.

We have: $g_1(M_1) = g_2(M_1) = \dfrac{5}{4}$.

Both constraints are violated at the point M_1 and take the same value. One or the other can be chosen as the most violated constraint. We choose g_1. We have: $\nabla g_1(\frac{3}{2}, 0) = \begin{pmatrix} 3 \\ 6 \end{pmatrix}$

$$g_1(\frac{3}{2},0) + ((x_1 - \frac{3}{2}, x_2))^{\mathrm{t}}\nabla g_1(\frac{3}{2},0) = \frac{5}{4} + 3x_1 - \frac{9}{2} + 6x_2$$
$$= 3x_1 + 6x_2 - \frac{13}{4}.$$

We denote by Δ_3 the line of equation $3x_1 + 6x_2 = \dfrac{13}{4}$.

We then solve the problem:

$$\text{Minimize } f(x_1, x_2) = -x_1 + x_2$$
$$\text{with } \begin{cases} 4x_1 + 4x_2 \leqslant 6 \\ 4x_1 - 6x_2 \leqslant 6 \\ 3x_1 + 6x_2 \leqslant \dfrac{13}{4}. \end{cases}$$

The solution is at the intersection of Δ_2 and Δ_3, which is the point $M_2 = (\dfrac{37}{28}, -\dfrac{5}{42}) \simeq (1.3214, -0.1190)$.

Étape 2.

We continue with the approximate values to simplify the computations. We have $g_1(M_2) \simeq -0.0462$ and $g_2(M_2) \simeq 1.2362$.

Both constraints are violated at the point M_2 and the most violated constraint is g_2, which is retained. We have:

$$\nabla g_2(1.3214, -0.1190)) = \begin{pmatrix} 2.6428 \\ -4,2380 \end{pmatrix}.$$

Hence:

$g_2(M_2)(x_1 - 1.3214, x_2 + 0.1190)^{\mathrm{t}} \nabla g_2(M_2) = 2.6428x_1 - 4.238x_2 - 2.7603.$

We denote by Δ_4 the line of equation $2.6428x_1 - 4.238x_2 = 2.7603$.

We then solve the problem:

$$\text{Minimize } f(x_1, x_2) = -x_1 + x_2$$
$$\text{with } \begin{cases} 4x_1 + 4x_2 \leqslant 6 \\ 4x_1 - 6x_2 \leqslant 6 \\ 3x_1 + 6x_2 \leqslant 3.25 \\ 2.6428x_1 - 4.238x_2 \leqslant 2.7603 \end{cases}$$

The solution is at the intersection of the lines Δ_3 and Δ_4, that is to say the pointt $M_3 \simeq (1.057, 0.013)$.

We stop here the algorithm. The optimum of the problem is at the point (1, 0), this résult can be verified with the conditions of Karush, Kuhn and Tucker, which are sufficient here. It follows that the sequence of points approaches the point which gives the optimum of the problem.

We finally establish the theorem below.

Theorem 38. *If the problem (P) admits a minimum at a finite distance, then any accumulation point of the sequence (x^k) generated by Kelley cutting-plane method is a solution of the problem (P).*

Proof. We place ourselves in the case where the method does not find the solution after a finite number of steps.

Let $\tilde{x}$ be an accumulation point of the sequence $(x^k)_{k\in\mathbb{N}}$. We can extract from this sequence a subsequence $(x^k)_{k\in K\subset\mathbb{N}}$ converging towards $\tilde{x}$.

Let x^* be a point of X reaching the minimum.

For any $k \in K$, we have $f(x^k) \leqslant f(x^*)$ since $X \subset Q_k$. Going to the limit when $k \in K$ tends to infinity, we get: $f(\tilde{x}) \leqslant f(x^*)$. We will show that $\tilde{x} \in X$.

We denote by i_k the index of the constraint retained in the step k, that is to say a most violated constraint in this step. The sequence $(i_k)_{k\in K}$ takes all its values in the finite set $\{1, 2, ..., m\}$. We can thus extract a constant subsequence $(i_h)_{h\in H\subset K}$, we denote by i_0 the constant.

Let $h \in H$. We denote by $r(h)$ the smallest index r of H satisfying $r > h$. The point $x^{r(h)}$ is in the half space E_h:

$g_{i_0}(x^h) + (x^{r(h)} - x^h)^{\mathrm{t}}\nabla g_{i_0}(x^h) \leqslant 0$.

Hence, using the fact that the constraint g_{i_0} is not satisfied by x^h:

$0 < g_{i_0}(x^h) \leqslant -(x^{r(h)} - x^h)^{\mathrm{t}}\nabla g_{i_0}(x^h) \leqslant ||(x^{r(h)} - x^h||||\nabla g_{i_0}(x^h)||$.

When $h \in H$ tends towards infinity, $r(h)$ also tends to infinity.

Moreover, when $h \in H$ tends towards infinity, $x^h \to \tilde{x}$ and $x^{r(h)} \to \tilde{x}$. Hence: $||x^{r(h)} - x^h|| \to 0$.

Furthermore, $\nabla g_{i_0}(x^h) \to \nabla g_{i_0}(\tilde{x})$, which implies that $||\nabla g_{i_0}(x^h)||$ is bounded. Therefore :

$||x^{r(h)} - x^h||||\nabla g_{i_0}(x^h)|| \to 0$, doù $g_{i_0}(x^h) \to 0$.

As $g_{i_0}(x^h) \to g_{i_0}(\tilde{x})$, we have: $g_{i_0}(\tilde{x}) = 0$. The point $\tilde{x}$ satisfies the constraint i_0.

Now let $i \in \{1, 2, ..., m\}$. The constraint g_{i_0} being the most violated, for $h \in H, g_i(x^h) \leqslant g_{i_0}(x^h)$. When $h \in H$ tends to infinity, we get: $g_i(\tilde{x}) \leqslant 0$. The point $\tilde{x}$ satisfies all the constraints, it belongs to X. As the inequality $f(\tilde{x}) \leqslant f(x^*)$ has been shown, $\tilde{x}$ gives the minimum of f on X. $\Diamond$

Remark. The points generated by Kelley cutting-plane method provide lower bounds of the problem whereas Frank and Wolfe method gives upper bounds.

8.6 Exercice

Statement. We are interested in the optimization problem defined on $\mathbb{R}^2$ as follows:

$$\text{Minimize } 2x_1^2 + x_2^4$$
$$\text{with the constraints } \begin{cases} x_1 \geqslant 1 \\ x_1 + ax_2 \geqslant a + 1 \end{cases}$$

where a is a real parameter.

Q1. For what values of a can we say that the minimum is reached at the point (1, 1)? For these values, is the minimum also reached at another point?

Q2. Solve the following problem using the descent method from point (1, 1):

$$\text{Minimize } 2x_1^2 + x_2^4$$
$$\text{with the constraints } \begin{cases} x_1 \geqslant 1 \\ 2x_1 + x_2 \geqslant 3. \end{cases}$$

The result will be indicated in an interval of length 0.00001.

Solution. Q1. Let:

$f(x_1, x_2) = 2x_1^2 + x_2^4$,
$g_1(x_1, x_2) = 1 - x_1$
$g_2(x_1, x_2) = a + 1 - x_1 - ax_2$.

The problem can be written:

$$\text{Minimize } f(x_1, x_2)$$
$$\text{with the constraints :}$$
$$\begin{cases} g_1(x_1, x_2) \leqslant 0 \\ g_2(x_1, x_2) \leqslant 0. \end{cases}$$

The Hessian matrix $\nabla^2 f(x_1, x_2) = \begin{pmatrix} 4 & 0 \\ 0 & 12x_2^2 \end{pmatrix}$ is positive: the function f is convex on $\mathbb{R}^2$ and it is strictly convex on the open half-plane defined by $x_1 > 0$, domain that contains the feasible domain.

On the other hand, a linear function is convex (and indeed also concave): g_1 and g_2 are convex. In addition, the interior of the feasible domain is non-empty, it contains for example the point (2, 1). The proposition 21 shows that the constraints are qualified at every point of $\mathbb{R}^2$. Therefore, according to the theorem 34, the point (1, 1) is a global minimum of the problem if and only if the Karush, Kuhn and Tucker conditions are fulfilled at the point (1, 1). At this point the two constraints are saturated; to say that the Karush, Kuhn and Tucker conditions are fulfilled is to show that there are two real numbers positive or zero λ_1 and λ_2 such as:

$\nabla f(1,1) = -\lambda_1 \nabla g_1(1,1) - \lambda_2 \nabla g_2(1,1)$.

Now, we have:

$\nabla f(x_1, x_2) = \begin{pmatrix} 4x_1 \\ 4x_2^3 \end{pmatrix}$, $\nabla f(1,1) = \begin{pmatrix} 4 \\ 4 \end{pmatrix}$,

$\nabla g_1(1,1) = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$, $\nabla g_2(1,1) = \begin{pmatrix} -1 \\ -a \end{pmatrix}$.

Let us find coefficients λ_1 and λ_2 satisfying:

$\begin{pmatrix} 4 \\ 4 \end{pmatrix} = \lambda_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \lambda_2 \begin{pmatrix} 1 \\ a \end{pmatrix}$,

what is written:

$$\begin{cases} \lambda_1 + \lambda_2 = 4 \\ a\lambda_2 = 4. \end{cases}$$

For there to be a solution, it is necessary and sufficient to have $a \neq 0$ and then:

$$\lambda_1 = 4\left(1 - \frac{1}{a}\right), \lambda_2 = \frac{4}{a}.$$

The Karush, Kuhn and Tucker conditions are fulfilled if and only if λ_1 and λ_2 are positive or null, that is, if and only if we have $a \geqslant 1$.

The theorem 30 shows that there is at most a global minimum. When there is a global minimum, it is a unique global minimum.

Q2. The problem of this question corresponds to the initial problem with $a = 1/2$: the minimum is not reached at the point (1,1).

Let us apply the descent method starting from point (1,1): let us search for the admissible direction of greatest descent for f at this point (see figure8.9).

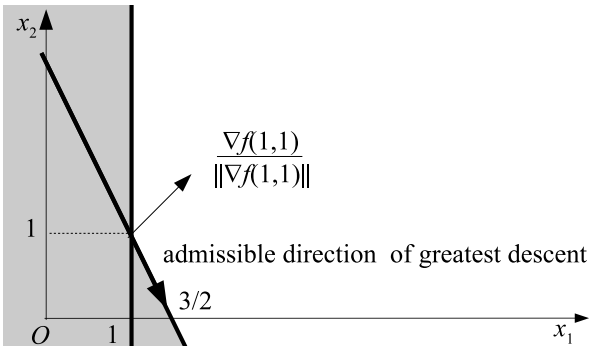

Figure 8.9: Graphic solution.

We can see graphically that the direction $d = (1, -2)$ is appropriate. Then let us find the minimum of $\phi(s) = f[(1,1) + s(1,-2)]$ for s positive; by noticing that we do not go out of the domain:

$\phi(s) = 2(1+s)^2 + (1-2s)^4$.

Hence, $\phi'(s) = 4(1+s) - 8(1-2s)^3$.

The function ϕ is convex because f is convex. So we try to have $\phi'(s) = 0$ (or nearly), which we do by a dichotomous type method:
$\phi'(0) = -4 < 0$, $\phi'(1/2) = 6 > 0$, $\phi'(0.25) > 0$, $\phi'(0.125) > 0$, $\phi'(0.06) < 0$, $\phi'(0.09) < 0$, $\phi'(0.1) > 0$, $\phi'(0.095) > 0$, $\phi'(0.0925) > 0$, $\phi'(0.092) > 0$, $\phi'(0.091) < 0$, $\phi'(0.0915) > 0$, $\phi'(0.0913) < 0$, $\phi'(0.0914) < 0$, $\phi'(0.09145) > 0$, $\phi'(0.09142) > 0$, $\phi'(0.09141) > 0$:

$$0.09140 < s_{min} < 0.09141.$$

The minimum of f in the direction d is reached at the point (1.0914, 0.8172).

Only the constraint g_2 is saturated at this point. Let us see if Karush, Kuhn and Tucker conditions are now satisfied:

$$\nabla f(1.0914, 0.8172) = \begin{pmatrix} 4.3656 \\ 2.18296 \end{pmatrix}$$

$$\nabla g_2(1.0914, 0.8172) = \begin{pmatrix} -1 \\ -1/2 \end{pmatrix} = \frac{-1}{4.3656} \begin{pmatrix} 4.3656 \\ 2.18296 \end{pmatrix}.$$

Therefore, $\nabla f(1.0914, 0.8172)$ and ∇g_2 are collinear of opposite directions: Karush, Kuhn, and Tucker conditions are fulfilled and the point (1.0914, 0.8172) is the global minimum of the problem. condition

Appendix A

Norm

Let $x = (x_i)_{1 \leqslant i \leqslant n}$ a vector. The three most common vector norms are:

- $||x||_1 = \sum_{i=1}^{n} |x_i|$ (norm 1)
- $||x||_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}}$ (norm 2, or Euclidean norm)
- $||x||_\infty = \max_{1 \leqslant i \leqslant n} |x_i|$ (infinity norm)

More generally: $||x||_p = (\sum_{i=1}^{n} |x_i|^p)^{\frac{1}{p}}$ (p-norm).

In $\mathbb{R}^n$ and $\mathbb{C}^n$, all norms are *equivalent* (two norms $||\ ||$ and $||\ ||'$ are équivalent on a vector space E if there are two strictly positive constants C and C' such that, for all x in E: $C||x|| \leqslant ||x||' \leqslant C'||x||$).

We can also use matrix norms. We call $\mathcal{A}_n$ the ring of the square matrices of order n with coefficients in $\mathbb{R}$ or $\mathbb{C}$. We call *matrix norm* an application from $\mathcal{A}_n$ to $\mathbb{R}^+$ denoted by $||\ ||$ which fulfils the following properties:

- for all matrices A of $\mathcal{A}_n$, $||A|| = 0 \Leftrightarrow A = 0$
- for all α of $\mathbb{R}$ (or $\mathbb{C}$) and for all A of $\mathcal{A}_n$, $||\alpha A|| = |\alpha|||A||$
- for all matrices A and B of $\mathcal{A}_n$, $||A + B|| \leqslant ||A|| + ||B||$
- for all matrices A and B of $\mathcal{A}_n$, $||A \times B|| \leqslant ||A|| \times ||B||$.

We can very easily build matrix norms from vectorial norms: they are then called *subordinate matrix norms*. For this, we can define $||A||$ by the following equivalent formulas:

$$||A|| = \sup_{x\neq 0}\frac{||Ax||}{||x||} = \sup_{||x||=1}||Ax|| = \sup_{0<||x||\leqslant 1}\frac{||Ax||}{||x||}.$$

We have: $||Ax|| \leqslant ||A||\ ||x||$.

The matrix norms subordinate to the most usual norms that we have described above are therefore, for $A = (a_{i,j})_{\substack{1\leqslant i\leqslant n\\ 1\leqslant j\leqslant n}}$:

- $||A||_1 = \sup_{x\neq 0}\dfrac{||Ax||_1}{||x||_1} = \max_{1\leqslant j\leqslant n}\displaystyle\sum_{i=1}^{n}|a_{ij}|$
- $||A||_2 = \sup_{x\neq 0}\dfrac{||Ax||_2}{||x||_2} = \sqrt{\rho(A^*A)} = ||A^*||_2$ where $\rho(A^*A)$ represents the largest absolute value of A^*A (spectral radius of A^*A)
- $||A||_\infty = \sup_{x\neq 0}\dfrac{||Ax||_\infty}{||x||_\infty} = \max_{1\leqslant i\leqslant n}\displaystyle\sum_{j=1}^{n}|a_{ij}|$.

The norm $||\ ||_2$ is invariant by unitary transformation: if U is a unitary matrix, that is, satisfies $U^*U = I$, then we have

$$||A||_2 = ||AU||_2 = ||UA||_2 = ||U^*AU||_2.$$

If A is normal, that is, if A fulfils the relation $A^*A = AA^*$ (especially if A is Hermitian or symmetric), then $||A||_2 = \rho(A)$.

If A est unitary or orthogonal, $||A||_2 = 1$.

Remark. $||A||_1$ and $||A||_\infty$ are easy to compute but not $||A||_2$.

Theorem

- Let $||\ ||$ be a subordinate norm; let B satisfying $||B|| < 1$. So $I + B$ is invertible and $||(I+B)^{-1}|| \leqslant \dfrac{1}{1-||B||}$.
- If a matrix of the form $I + B$ is not invertible, then, for any norm, subordinate or not, $||B|| \geqslant 1$.

Example of a non-subordinate norm: the Euclidean norm

This norm is définied by: $||A||_E = \left(\sum_{i=1}^{n}\sum_{j=1}^{n}|a_{ij}|^2\right)^{\frac{1}{2}} = \sqrt{\text{trace } (A^*A)}$ (remember that the trace of a matrix is the sum of its diagonal terms). The norm $||A||_E$ is invariant by unitary transformation; in other words, if $U^*U = I$, then $||A||_E = ||AU||_E = ||UA||_E = ||U^*AU||_E$.
Moreover: $||A||_2 \leqslant ||A||_E \leqslant \sqrt{n}||A||_2$.