Optimization for beginners

Fabio Camilli Paola Loreti

Sapienza Università di Roma, Facoltà di Ingegneria dell'Informazione,
Dipartimento di Scienze di Base e Applicate per l'Ingegneria.

Introduction

Contents

CHAPTER 1

Prerequisities

1. Vectors

Vector spaces were introduced in the 17th century in the calculus of solutions of systems of linear equations. More abstract treatment was formulated firstly by Giuseppe Peano (1858-1932). Although here it is given an abstract formulation, the only applications we will consider involve Euclidean vector spaces, matrices and systems of linear equations.

Let V be a set endowed with an inner operation $+ : V \times V \to V$ and a product by scalars $\cdot : \mathbb{R} \times V \to V$ (the symbol $\cdot$ is usually omitted in the explicit computations). The triad $(V, +, \cdot)$, or shortly V, is a *real vector space* if the following properties for the operations are verified (for any $u, v \in V$ and $\lambda, \mu \in \mathbb{R}$)

- $(V, +)$ is a group, i.e. it satisfies
  - Associativity of addition: $u + (v + w) = (u + v) + w$.
  - Commutativity of addition: $v + w = w + v$.
  - Identity element of addition: There exists an element $\underline{0} \in \mathcal{V}$, called the zero vector, such that $v + \underline{0} = v$ for all $v \in V$.
  - Inverse elements of addition: For any $v \in V$, there exists an element $w \in V$, called the additive inverse of v, such that $v + w = \underline{0}$. The additive inverse is denoted by $-v$.
- Distributivity of scalar multiplication with respect to vector addition: $\lambda(v + w) = \lambda v + \lambda w$.
- Distributivity of scalar multiplication with respect to field ($\mathbb{R}$) addition: $(\lambda + \mu)v = \lambda v + \mu v$.
- Compatibility of scalar multiplication with field ($\mathbb{R}$) multiplication: $\lambda(\mu v) = (\lambda\mu)v$.
- Identity element of scalar multiplication: $1v = v$, where 1 denotes the multiplicative identity in $\mathbb{R}$.

Note that the previous definition can be given by replacing $\mathbb{R}$ with any field $\mathbb{K}$. The elements of V are called *vectors*.

Example 1. Example of vector spaces are:

- The set with one element $\{\underline{0}\}$ with trivial operations is a (real) vector space.
- The field $\mathbb{R}$ with its operations of sum and product is itself a real vector space.
- The set $\mathbb{R}^N$ is a real vector space with the following operations. Given $x = (x_1, x_2, \ldots, x_N)$ and $y = (y_1, y_2, \ldots, y_N)$ in $\mathbb{R}^N$ and $\lambda \in \mathbb{R}$, one can define

$$(x + y) = (x_1 + y_1, x_2 + y_2, \ldots, x_N + y_N)$$

$$\lambda x = (\lambda x_1, \lambda x_2, \dots, \lambda x_N).$$

- The set $C([a,b])$ of the continuous functions in an interval $[a,b]$ with the operations

$$(f+g)(x) = f(x) + g(x), \qquad (\lambda f)(x) = \lambda f(x)$$

Since the only inner operation in a vector space is the sum, the key word for vector space theory is *linearity*. Given $v_1, v_2, \dots, v_k$ a linear combination of vectors is a vector

$$\lambda_1 v_1 + \lambda_2 v_2 + \lambda_k x_k,$$

where $\lambda_1, \lambda_2, \dots \lambda_k \in \mathbb{R}$ are called the *coefficients* of the linear combination.

DEFINITION. A finite non-empty set of vectors is said *linearly dependent* if there exists a linear combination of them with not all zero coefficients such that

$$\lambda_1 v_1 + \lambda_2 v_2 + \cdots + \lambda_k x_k = \underline{0}$$

A set of vectors which is not linearly dependent is called *linearly independent*.

Note that being linearly dependent does not depend on the order of the vectors. Moreover if the set includes the null vector, than it is certainly a set of linear dependent vectors, since, being for instance $v_1 = \underline{0}$, one can write

$$1v_1 + 0v_2 + \cdots + 0x_k = \underline{0}.$$

Moreover, if the set has only one element v, then it is linear dependent if and only if $v = \underline{0}$. It the set is $\{v_1, v_2\}$, then it is linear dependent if and only if either $v_1 = \underline{0}$, or $v_2 = \underline{0}$, or v_2 is multiple of v_1, i.e. $v_2 = \lambda v_1$ for some $l \in \mathbb{R}$.

Given V and vectors $v_1, v_2, \dots, v_k$ in V, consider the set of all the possible linear combinations of these vector

$$U = \{\lambda_1 v_1 + \lambda_2 v_2 + \lambda_k v_k \text{ such that } \lambda_i \in \mathbb{R}, i = 1 \dots k\}.$$

Since the sum of two elements of U and the product by a scalar of an element of U still belongs to U, then with the same operations defined in V, U itself a vector space, contained in V, and it is said a *vector subspace* of V. It is the subspace *generated by* the vectors $v_1, v_2, \dots, v_k$, and the set $\{v_1, v_2, \dots, v_k\}$ is said to be a set of generators of U. If it happens that $U = V$ then it is a set of generators for V. Thus

DEFINITION. A set of linearly independent vectors of V which generates V is said a *basis* of V

Note that given a set of generators, one can chose a subset of it, which is maximal with respect to the condition of being independent. Then it is still a set of generators, hence it is a basis for V. The definition of a set of generators can be extended to infinite sets, so it may happen that a vector space has a basis with infinite elements. In any case, a vector space has an infinite amount of bases, but it can be proven that all of them have the same cardinality, which is defined as the *dimension* of the vector space. Any set of vectors which exceed the dimension of the space is necessarily a set of dependent vectors.

EXAMPLE 2. The *canonical* (or *standard*) basis in $\mathbb{R}^N$ is given by $e_1 = (1, 0, \dots, 0)$, $e_2 = (0, 1, \dots, 0)$, $e_N = (0, 0, \dots, 1)$ and the dimension of $\mathbb{R}^N$ is N.

REMARK. We note that any part of the theory of vector spaces can be developed in the infinite dimensional case, with slight care. A typical example of infinite dimensional (real) vector space is the set of polynomials in one variable $a_0+a_1x+\cdots+a_nx^n$ and a basis is $\{1, x, x^2, \dots\}$. On the other side, the space of continuous (differentiable) functions on a fixed domain has no numerable basis.

PROPOSITION 1.1. *A set of vectors $\{v_1, v_2, \dots, v_N\}$ is a basis of V if and only if any vector $v \in V$ can be expressed as a linear combination of the vectors $v_1, v_2, \dots, v_N$, in a unique way, up to the order.*

By the above proposition, given a basis $\{v_1, v_2, \dots, v_N\}$ of V, any $v \in V$ is bi-univocally associated to the set $(\lambda_1, \dots, \lambda_N)$, said the set of the *coordinates* of v, of its coefficients with respect to the basis (to avoid ambiguity the bases are chosen as ordered sets).

EXAMPLE 3. If $x = (x_1, x_2, \dots, x_N) \in \mathbb{R}^N$, its coordinates with respect to the canonical basis of $\mathbb{R}^N$ are the component of the vector x itself.

DEFINITION. Let V and W be real vector spaces, a map $L : V \to W$ is said a *linear transformation* if it is compatible with the operations in V and W respectively, that is if

$$L(\lambda u + \mu v) = \lambda L(u) + \mu L(v), \qquad \forall u, v \in V, \quad \forall \lambda, \mu \in \mathbb{R}.$$

In particular, If $W = \mathbb{R}$, L is said a *linear functional.*

It is simple to see that the set

$$\mathcal{L}(V, W) = \{L : V \to W \text{ linear}\}$$

with the natural operations of sum and multiplication by scalars is a vector space. The set of all the linear functionals from $V \to \mathbb{R}$ is called the *dual space* of V and it is denoted by $V^\star$.

EXAMPLE 4. In the next we find a basis of $(\mathbb{R}^N)^\star$, the dual space of $\mathbb{R}^N$. For every $i = 1, 2, \dots, N$ we consider the functional

$$e^i : \mathbb{R}^N \to \mathbb{R}, \quad e^i(x) = x \cdot e_i = x_i,$$

which selects the i-th coordinate of the vector x with respect to the canonical basis of $\mathbb{R}^N$. In particular, for any $i, j = 1, \dots, N$ we have

$$\boxed{e^i(e_j) = \delta_{ij}},$$

with

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j. \end{cases} \tag{1.1}$$

With this notation it is easily proven that the e^i's are linearly independent vectors of $(\mathbb{R}^N)^\star$. Consider now a linear functional

$$L : \mathbb{R}^N \to \mathbb{R},$$

$$L(x) = L(x_1 \cdot e_1 + x_2 \cdot e_2 + \cdots + x_N e_N) = x_1 L(e_1) + x_2 L(e_2) + \cdots + x_N L(e_N)$$

Setting

$$a_i = L(e_i) \in \mathbb{R},\ i = 1, 2, \ldots, N$$

it is

$$L(x) = x_1 a_1 + x_2 a_2 + \cdots + x_N a_N,$$

thus L is univocally determined by its values on the vectors $e_1, \ldots, e_N$ of the canonical basis, that is the scalars $a_1, \ldots, a_N$. Moreover, recalling the definition of the e^i's, one can write L as a linear combination of $e^1, e^2, \ldots, e^N$,with coefficients $a_1, a_2, \ldots, a_N$. In fact,

$$L(x) = x_1 a_1 + x_2 a_2 + \cdots + x_N a_N = a_1 e^1(x) + a_2 e^2(x) + \ldots a_N e^N(x).$$

Thus $e^1, e^2, \ldots, e^N$ generate $(\mathbb{R}^N)^\star$, hence they are a basis for $(\mathbb{R}^N)^\star$, which is said the *canonical (dual) basis* of $(\mathbb{R}^N)^\star$. This relation gives an explicit one-to-one correspondence between $(\mathbb{R}^N)^\star$ and $\mathbb{R}^N$.

More generally, it is possible to show that there is a one to one correspondence between the vector space of linear transformations $\mathcal{L}(N, M) = \{L : \mathbb{R}^N \to \mathbb{R}^M, \text{linear}\}$ and the set of $M \times N$ matrices as we are going to see later.

2. Matrices

We denote by $\mathcal{S}^{M\times N}$ the set of matrices with M rows and N columns. A matrix with the same number of rows and columns is said a *square* matrix.

Given two matrices with real (or complex) coefficients $P = (p_{ij})$ and $Q = (q_{ij}) \in \mathcal{S}^{M\times N}$, $\lambda \in \mathbb{R}$ then we define

$$R = P + Q \in \mathcal{S}^{M\times N} \text{ and } \Leftrightarrow r_{ij} = p_{ij} + q_{ij} \qquad i = 1, \ldots, M\ j = 1, \ldots, N$$
$$S = \lambda Q \in \mathcal{S}^{M\times N} \Leftrightarrow s_{ij} = \lambda q_{ij} \qquad i = 1, \ldots, M\ j = 1, \ldots, N$$

The null matrix in $\mathcal{S}^{M\times N}$ is the matrix with all null entries, i.e.

$$\begin{pmatrix} 0 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & 0 \end{pmatrix} \tag{1.2}$$

With those operations, $\mathcal{S}^{M\times N}$ is a real (complex) vector space. We introduce a multiplication between matrices. Given $Q \in \mathcal{S}^{M\times P}$and $P \in \mathcal{S}^{P\times N}$ we set

$$R = QP \in \mathcal{S}^{M\times N} \Leftrightarrow r_{ij} = \sum_{k=1}^{P} q_{ik} p_{kj} \qquad i = 1, \ldots, M \quad j = 1, \ldots, N$$

It is important to observe that the previous definition is well posed if and only the number of columns of first term of the multiplication is equal to the number of rows of the second term.

Remark. If $N = M = P$, the matrices P and Q can be multiplied in both the orders, but the product is not commutative. Take for instance

$$\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 4 & 6 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & 2 \\ 7 & 4 \end{pmatrix}.$$

The following holds:

$$\begin{aligned} Q(P_1 + P_2) &= QP_1 + QP_2, && \forall Q \in \mathcal{S}^{M\times P},\ P_1, P_2 \in \mathcal{S}^{P\times N}, \\ Q(\lambda P) &= \lambda(QP), && \forall \lambda \in \mathbb{R},\ Q \in \mathcal{S}^{M\times P},\ P \in \mathcal{S}^{P\times N}. \end{aligned}$$

The *unit (square) matrix* $I = (\delta_{i,j}) \in \mathcal{S}^{N\times N}$

$$\begin{pmatrix} 1 & 0 & 0 \\ \vdots & \ddots & \vdots \\ 0 & 0 & 1 \end{pmatrix}, \tag{1.3}$$

it is the neutral element for the product of matrices, that is

$$QI = IQ = Q, \quad \forall Q \in \mathcal{S}^{N\times N}.$$

The *transposition* is a linear transformation from $\mathcal{S}^{M\times N}$ to $\mathcal{S}^{N\times M}$ which associates to $Q \in \mathcal{S}^{M\times N}$ the matrix $Q^T \in \mathcal{S}^{M\times N}$ whose elements are

$$q^T_{ij} = q_{ji} \qquad i = 1, \dots, M \quad j = 1, \dots, N.$$

Given $Q, R \in \mathcal{S}^{M\times N}$, $\lambda \in \mathbb{R}$, we have

$$\begin{aligned} &(Q^T)^T = Q \\ &(Q + R)^T = Q^T + R^T \\ &(QR)^T = R^T Q^T \\ &(\lambda Q)^T = \lambda Q^T \end{aligned}$$

Definition. A matrix $Q \in \mathcal{S}^{N\times N}$ is said

i) *symmetric* if $Q = Q^T$.
ii) *orthogonal* if $QQ^T = Q^T Q = I$

Remark. For matrices with complex entries one can also consider the complex coniugate of the entries of the matrices. Given $Q \in \mathcal{S}^{M\times N}$, we consider the *conjugate transpose* $Q^\star \in \mathcal{S}^{N\times M}$ whose elements are

$$q^\star_{ij} = \overline{q}_{ji} \qquad i = 1, \dots, M \quad j = 1, \dots, N$$

(observe that if Q has real entries, then $Q^\star = Q^T$). For $Q, R \in \mathcal{S}^{M\times N}$ complex matrices, $\lambda \in \mathbb{C}$, we have

$$\begin{aligned} &(Q^\star)^\star = Q \\ &(Q + R)^\star = Q^\star + R^\star \\ &(QR)^\star = R^\star Q^\star \\ &(\lambda Q)^\star = \overline{\lambda} Q^\star \end{aligned}$$

A complex matrix $Q \in \mathcal{S}^{N\times N}$ is said

i) *hermitian* if $Q = Q^\star$;
ii) *unitary* if $QQ^\star = Q^\star Q = I$.

DEFINITION. A matrix $F \in \mathcal{S}^{N\times N}$ is said *invertible* (or nonsingular) if there exists a matrix $G\in \mathcal{S}^{N\times N}$ such that

$$FG = GF = I.$$

The *inverse* G is denoted by F^{-1}.

Not all square matrices have inverses, but if F is invertible also F^{-1} is invertible. We recall some properties. Take $F, G \in \mathcal{S}^{N\times N}$ invertible matrices.

$$(F^{-1})^{-1} = F$$
$$(F^T)^{-1} = (F^{-1})^T$$
$$(FG)^{-1} = G^{-1}F^{-1}$$

Let Q be a square matrix with coefficients in a field $\mathbb{K}$. The *determinant* $\det(Q)$ of Q is scalar associated to the matrix Q which helps in distinguish invertible from non invertible matrices. It can be computed using the *Leibniz formula*, which can be considered as the definition of the determinant (although there is a non computational way to define it):

$$\det(Q) = \sum_{\sigma\in\mathcal{S}_n} \operatorname{sgn}(\sigma) \prod_{i=1}^{N} q_{i,\sigma(i)},$$

where sgn is the sign function of permutations (+1 and -1 for even and odd permutation, respectively) in the permutation group $\mathcal{S}_N$.

EXAMPLE 5.

- If $N = 1$, $Q = q \in \mathbb{R}$, then $\det(Q) = q$;
- If $N = 2$ then $\mathcal{S}_2 = \{(12), (21)\}$ and

$$\det(Q) = \prod_{i=1}^{2} q_{i,\sigma_1(i)} - \prod_{i=1}^{2} q_{i,\sigma_2(i)} = q_{1,1}q_{2,2} - q_{1,2}q_{2,1}.$$

- By Leibniz formula, it is easy to see that if Q is the null matrix, then $\det(Q) = 0$. If $Q = I$, then $\det(Q) = 1$. Moreover if Q is a *diagonal matrix* (with all zero on the entries $q_{i,j}$ with different indices) or *triangular* (that is, if $q_{i,j} = 0$ for any $i < j$, or for any $i > j$), then the determinant is the product of the entries on the diagonal.

PROPOSITION 1.2. *The following properties hold*

$$\det(F^T) = \det(F)$$
$$\det(FG) = \det(F)\det(G)$$
$$\det(F^{-1}) = \frac{1}{\det(F)}$$
$$\det(\lambda F) = \lambda^N \det(F),\ \forall\lambda \in \mathbb{R}.$$

An easier way to compute the determinant of a matrix is given by following rule, due to Laplace,

$$\det(Q) = \begin{cases} q_{11} & \text{if} \quad N = 1 \\ \sum_{j=1}^{N} q_{ij}\Delta_{ij} & \text{if} \quad N > 1 \end{cases} \tag{1.4}$$

where $\Delta_{ij} = (-1)^{i+j}\det(Q_{ij})$ and $\det(\mathrm{Q})_{ij}$ is the determinant of the submatrix $\in \mathcal{S}^{N-1\times N-1}$ obtained by removing from Q its i-th row and j-th column.

It can be shown that

THEOREM 1.3. *A matrix is invertible if and only if its determinant is different from zero.*

3. Linear transformations

Given a matrix $Q \in \mathcal{S}^{M\times N}$, consider the transformation $L_Q : \mathbb{R}^N\mathbb{R}^M$ defined by

$$L_Q(x) = Qx, \ \forall x \in \mathbb{R}^N,$$

where x is intended as a matrix $N \times 1$. Explicitly:

(1.5)

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \dots a_{1,M} \\ \dots & \dots & \dots \\ a_{M,1} & a_{M,2} & \dots a_{M,N} \end{pmatrix}\begin{pmatrix} x_1 \\ \dots \\ x_N \end{pmatrix} = \begin{pmatrix} a_{1,1}x_1 + a_{1,2}x_2 \cdots + a_{1,N}x_N \\ \dots\dots\dots \\ a_{M,1}x_1 + a_{M,2}x_2 + \cdots + a_{M,N}x_N \end{pmatrix}.$$

Since the multiplication of matrices commutes with the sum and the product with scalars, it is easily seen that L_Q is a *linear transformation*.

Note that if $x = e_i$, an element of the canonical basis of $\mathbb{R}^N$, then $L_Q(e_i)$ is the vector q^i given by the i-th column of the matrix Q. By abuse of notation, we identify L_Q with Q. Indeed, it can be shown that there is a one to one correspondence between the vector space of linear transformations $\mathcal{L}(N, M) = \{L : \mathbb{R}^N \to \mathbb{R}^M, \text{linear}\}$ and the vector spaces of matrices $\mathcal{S}^{M\times N}$. In fact, given a linear transformation L, the corresponding matrix have elements $a_{ij} = \xi_i \cdot L(e_j)$ with ξ_i the canonical bases in $\mathbb{R}^N$ $= 1, \dots M$, $j = 1, \dots, N$. This correspondence respects the structure of vector spaces, so that it is actually an *isomorphism* of vector spaces. For $M = 1$ this is the same as the isomorphism between $\mathbb{R}^N$ and its dual space $(\mathbb{R}^N)^\star$.

Composition of linear transformations correspond to product of matrices, i.e. $L_Q \circ L_P = L_{QP}$.

It is worthwhile to note that given two different bases of a finite dimensional vector space V, then there exist a linear transformation which brings the vectors of one basis on the vectors of the other basis, and vice versa, so that it is an invertible transformation. To this transformation it corresponds an invertible square matrix, the inverse of which relates the coordinates of the first basis the the coordinates of the second.

Two (square) matrices Q and Q' associated to the same linear transformation of V in itself are said *similar*. Analytically, two (square) matrices Q and Q' are s *similar* if there exists an invertible matrix C (the matrix of the change of basis in V) such that $Q' = C^{-1}QC$.

The *image* of Q is

$$\mathcal{R}(Q) = \{y \in \mathbb{R}^N \text{ such that } \exists x \in \mathbb{R}^M : y = Qx\}$$

The linearity implies that $\mathcal{R}(Q)$ is a vector subspace of $\mathbb{R}^N$. Let $\{q^1, q^2, \ldots, q^N\}$ be the columns of the matrix, they are vectors of $\mathbb{R}^M$. Since $L_Q(e_i) = q^i$, for $i = 1, \ldots, N$ and the set e_i generates $\mathbb{R}^N$, then $\mathcal{R}(Q)$ is a subspace of $\mathbb{R}^M$ generated by $\{q^1, q^2, \ldots, q^N\}$. We define the *rank* of Q as the dimension $rk(Q)$ of the vector space $\mathcal{R}(Q)$, that is the maximum number of linearly independent column vectors of Q.

The *kernel* of the matrix Q is a vector subspace of $\mathbb{R}^N$ defined by

$$\mathcal{N}(Q) = \{x \in \mathbb{R}^N : Qx = \underline{0}\}$$

It .

PROPOSITION 1.4. *We have*

i) $rk(Q) = rk(Q^T)$
ii $rk(Q) + \dim(\ker(Q)) = N$

THEOREM 1.5. *Let Q be a square real $N \times N$ matrix . Then the following statements are equivalent:*

i) *Q is invertible.*
ii) *$det\ Q \neq 0$*
iii) *$rk(Q) = N$.*
iv) *The vectorial equation $Qx = \underline{0}$ has only the trivial solution $x = \underline{0}$*
v) *The vectorial equation $Qx = b$ has exactly one solution for each $b \in \mathbb{R}^N$.*

Moreover

PROPOSITION 1.6.

$$\mathcal{N}(Q) = \mathcal{R}^{\perp}(Q^T)$$

PROOF. If $x \in \mathcal{N}(Q)$, then $(Q^T y) \cdot x = yQx = 0, \forall y \in \mathbb{R}^N$, hence $x \in \mathcal{R}^{\perp}(Q^T)$.

If $x \in \mathcal{R}^{\perp}(Q^T)$, then $0 = Q^T y \cdot x = yQx, \forall y$, hence $x \in \mathcal{N}(Q)$. □

4. System of linear equations.

A solution of a linear system $Ax = b$ is an assignment of values to the variables $x_1, x_2, ..., x_N$ such that $Ax = b$. A linear system may behave in any one of three possible ways:

(1) The system has infinitely many solutions.
(2) The system has a single unique solution.
(3) The system has no solution.

.....

5. Scalar products

Let $x = (x_1, \dots, x_N), y = (y_1, \dots, y_N) \in \mathbb{R}^N$. The *usual scalar product* in $\mathbb{R}^N$ is a real number defined by

$$x \cdot y = x_1 y_1 + \cdots + x_N y_N.$$

The scalar product is a bilinear form as a function in the vectorial variables x and y, that is:

$$\text{for any } x, y, z \in \mathbb{R}^N,\ \lambda \in \mathbb{R}$$
$$x \cdot y = y \cdot x, \qquad (x+y)\cdot z = x \cdot z + y \cdot z, \qquad \lambda x \cdot y = \lambda x \cdot y.$$

Moreover, the scalar product is *positive definite*:

$$x \cdot x \geq 0 \quad \text{and} \quad x \cdot x = 0 \iff x = 0.$$

We can define the *modulus* or *norm* of $x \in \mathbb{R}^N$ associated to the standard scalar product:

$$\|x\| = \sqrt{x \cdot x} = \sqrt{x_1^2 + x_2^2 + \dots x_N^2}.$$

It holds

$$\|\lambda x\| = |\lambda|\, \|x\|.$$

We have the Cauchy-Schwartz inequality connecting scalar product and norms of vectors:

$$|x \cdot y| \leq \|x\| \|y\|,$$

with equality if and only if the vectors are proportional, i.e $y = \lambda x$. It follows the *triangular inequality*

$$\|x + y\| \leq \|x\| + \|y\|.$$

6. Symmetric matrices

A symmetric matrix Q is said *nonnegative* (respectively, *nonpositive*) if the quadratic form $x^T Q x$ is positive (respectively, negative) semi-definite, i.e. if

$$x^T Q x = \sum_{i,j=1}^{N} q_{i,j} x_i x_j \geq 0 \qquad (\text{respectively, } x^T Q x \leq 0,) \quad \forall x \in \mathbb{R}^N$$

and *positive* (respectively, *negative*) if the quadratic form $x^T Q x$ is positive (respectively, negative) definite

$$x^T Q x = \sum_{i,j=1}^{N} q_{i,j} x_i x_j\, 0 \qquad (x^T Q x < 0) \quad \forall x \in \mathbb{R}^N, x \neq 0.$$

Example 6. An example of positive matrix is

$$I = \begin{pmatrix} 1 & \dots & 0 \\ \vdots & 1 & \vdots \\ 0 & \dots & 1 \end{pmatrix} \tag{1.6}$$

since $x^T I x = x_1^2 + \cdots + x_N^2 > 0$ for $x = (x_1, \dots, x_N) \neq 0$. An example of nonnegative matrix is

$$Q = \begin{pmatrix} 2 & 0 \\ 0 & 0 \end{pmatrix} \tag{1.7}$$

while an indefinite matrix is

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & -2 \end{pmatrix} \tag{1.8}$$

If we restrict the analysis to 2×2 matrices, then the have the following

PROPOSITION 1.7. *Let*

$$Q = \begin{pmatrix} q_{11} & q_{12} \\ q_{21} & q_{22} \end{pmatrix} \tag{1.9}$$

a symmetric matrix. Then

$$\begin{cases} \textit{If } \det Q > 0 \textit{ and } \begin{cases} q_{11} > 0 & \textit{then } Q \textit{ is positive} \\ q_{11} < 0 & \textit{then } Q \textit{ is negative} \end{cases} \\ \textit{If } \det Q = 0 \textit{ and } \begin{cases} q_{11},\, q_{22} \geq 0 & \textit{then } Q \textit{ is nonnegative} \\ q_{11},\, q_{22} \leq 0 & \textit{then } Q \textit{ is nonpositive} \end{cases} \\ \textit{If } \det Q < 0, \textit{ then } Q \textit{ is indefinite} \end{cases}$$

REMARK. We can introduce a partial order in the set of symmetric matrices in the following way

$$Q \leq P \quad \Longleftrightarrow \quad P - Q \geq 0.$$

7. Exercises

1. Show that

$$\|x + y\|^2 = \|x\|^2 + 2x \cdot y + \|y\|^2 \qquad \text{for all } x, y \in \mathbb{R}^N,$$

2. Show that

$$xy \leq \frac{x^2}{2} + \frac{y^2}{2}, \qquad \text{for all } x, y \in \mathbb{R}$$

3. Show that

$$xy \leq \epsilon x^2 + \frac{y^2}{4\epsilon}, \qquad \text{for all } x, y \in \mathbb{R}, \epsilon > 0$$

4. Show that

$$|x \cdot y| \leq \frac{\epsilon}{2}\|x\|^2 + \frac{\|y\|^2}{2\epsilon}, \qquad \text{for all } x, y \in \mathbb{R}^N, \epsilon > 0.$$

Hint:

$$0 \leq \|x \pm \epsilon y\|^2 = \|x\|^2 \pm 2\epsilon x \cdot y + \epsilon^2 \|y\|^2,$$

5. Using exercise 4, show the Cauchy-Schwartz inequality

$$|x \cdot y| \leq \|x\|\|y\| \qquad \text{for all } x, y \in \mathbb{R}^N, \tag{1.10}$$

6. Show that equality in (1.10) holds if and only if the vectors are proportional.

8. Inequalities

Given N real numbers x_1, $x_2, \cdots x_N$, we define their *arithmetic mean* as

$$M_a = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$$

and their *geometric mean* as

$$M_g = \sqrt[N]{x_1 \cdot x_2 \cdots x_N} = \sqrt[N]{\prod_{i=1}^{N} x_i}$$

THEOREM 1.8 (Mean Inequality). *Given N real positive numbers x_1, $x_2, \cdots x_N$*

$$M_g = \sqrt[N]{\prod_{i=1}^{N} x_i} \leq \frac{\sum_{i=1}^{N} x_i}{N} = M_a.$$

Equality holds if and only if x_1, $x_2, \cdots x_N$ are equal.

PROOF. By induction

i) If $N = 1$ the inequality is true since

$$M_a = x_1 = M_g.$$

ii) Assuming the inequality true at step $N - 1$

$$M'_g = \sqrt[N-1]{\prod_{i=1}^{N-1} x_i} \leq \frac{\sum_{i=1}^{N-1} x_i}{N-1} = M'_a,$$

then we have

$$M_a = \frac{(N-1)}{N} M'_a + \frac{x_N}{N} = \left(M'_a + \frac{(x_N - M'_a)}{N} \right),$$

$$\frac{M_a}{M'_a} = \left(1 + \frac{x_N - M'_a}{M'_a} \frac{1}{N} \right)$$

hence

$$\left(\frac{M_a}{M'_a} \right)^N = \left(1 + \frac{x_N - M'_a}{M'_a} \frac{1}{N} \right)^N$$

To apply Bernoulli's inequality (i.e. $(1+q)^N \geq 1 + Nq$ for $q > -1$) we need $-M'_a + x_N \geq -NM'_a$ that is

$$(N-1)M'_a + x_N \geq 0,$$

which is true since $M'_a, x_N \geq 0$. Then

$$\left(\frac{M_a}{M'_a} \right)^N \geq \left(1 + \frac{x_N - M'_a}{M'_a} \right) = \frac{x_N}{M'_a}$$

$$(M_a)^N \geq x_N (M'_a)^{N-1},$$

and by the inductive step

$$(M_a)^N \geq x_N (M'_g)^{N-1} = x_1 \cdot x_2 \ldots x_N = \prod_{i=1}^{N} x_i.$$

□

We define *conjugate exponents* two positive real numbers p and q such that

$$\frac{1}{p}+\frac{1}{q}=1 \tag{1.11}$$

THEOREM 1.9 (Young Inequality). *For all conjugate exponents $p,q\in\mathbb{R}$ and any nonnegative real numbers x,y*

$$xy\le\frac{x^p}{p}+\frac{y^q}{q} \tag{1.12}$$

PROOF. Consider first the case $p,q\in\mathbb{Q}$. Then $p=\frac{n}{m}$ with $m,n\in\mathbb{N}$ with $m<n$ and

$$q=\frac{n}{n-m}.$$

Then by taking

$$x_1=x_2=\cdots=x_m=|x|^p$$

$$x_{m+1}=\cdots=x_N=|y|^q$$

in (1.8), we get the inequality (1.12). For $p,q\in\mathbb{R}$, we get the inequality by the density of $\mathbb{Q}$ in $\mathbb{R}$. □

REMARK. An alternative proof of (1.12) can be done by using the convexity of the function $x\to e^x$. In fact

$$xy=e^{\ln x+\ln y}=e^{\frac{1}{p}\ln x^p+\frac{1}{q}\ln y^q}\le\frac{1}{p}e^{\ln x^p}+\frac{1}{q}e^{\ln x^q}=\frac{x^p}{p}+\frac{x^q}{q}$$

THEOREM 1.10 (Hölder inequality). *For any conjugate exponents $p,q\in[1,+\infty)$ and for any $x,y\in\mathbb{R}^N$ we have*

$$|x\cdot y|\le\|x\|_p\|x\|_q. \tag{1.13}$$

For the proof of the inequality we refer to [**?**]. In particular, for $q=p=2$ (1.13) gives the Cauchy-Schwartz inequality (1.10).

THEOREM 1.11 (Minkowski inequality). *For any $p\in[1,+\infty)$ and for any $x,y\in\mathbb{R}^N$ we have*

$$\|x+y\|_p\le\|x\|_p+\|y\|_p. \tag{1.14}$$

An immediate consequence of (1.14) is

COROLLARY 1.12. *For any $p\in[1,+\infty)$,*

$$\|x\|_p=\big(\sum_{i=1}^N|x_i|^p\big)^{\frac{1}{p}}\qquad x\in\mathbb{R}^N \tag{1.15}$$

defines a norm in $\mathbb{R}^N$.

8.1. Exercises. **1**. Show that the sequence (x_N) defined by

$$x_N = \Big(1 + \frac{1}{N}\Big)^N, \quad N = 1, 2, \dots.$$

is bounded and increasing.

Solution. Applying the inequality (1.8)

$$\sqrt[n+1]{a_1 \dots a_{N+1}} \le \frac{a_1 + \dots + a_{N+1}}{N+1}$$

with

$$a_1 = \dots = a_n = 1 + \frac{1}{N} \quad \text{and} \quad a_{N+1} = 1,$$

we obtain

$$\sqrt[N+1]{\Big(1 + \frac{1}{N}\Big)^N} \le \frac{N+2}{N+1} = 1 + \frac{1}{N+1}.$$

This is equivalent to $x_N \le x_{N+1}$. Hence (x_N) is increasing.

2. Show that

$$\sqrt[\frac{N(N+1)}{2}]{\prod_{i=1}^{N} \frac{1}{i^i}}$$

is the minimum value of the function

$$f(x_1, x_2, \dots, x_N) = \frac{x_1}{x_2} + \sqrt{\frac{x_2}{x_3}} + \dots + \sqrt[N]{\frac{x_N}{x_1}}, \qquad x_1, x_2, \quad x_N > 0$$

.

3. The harmonic mean of N real positive numbers x_1, $x_2, \cdots x_N$ is

$$M_h = N\Big(\frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_N}\Big)^{-1} = N\Big(\sum_{i=1}^{N} \frac{1}{x_i}\Big)^{-1}$$

Prove by induction the following

THEOREM 1.13. *Given N real positive numbers x_1, $x_2, \cdots x_N$*

$$M_h = N\Big(\sum_{i=1}^{N} \frac{1}{x_i}\Big)^{-1} \le \sqrt[N]{\prod_{i=1}^{N} x_i} = M_q$$

Equality holds if and only if x_1, $x_2, \cdots x_N$ are equal.

4. Using (1.8) solve the minimization problem: find the minimum of

$$f(x_1, x_2, \cdots x_N) = \frac{x_1 + x_2 + \dots + x_N}{N},$$

under the condition $x_i \ge 0$ i$= 1, \dots, N$ and $x_1 x_2 \cdots x_N = 1$.

5. Using (1.13) solve the minimization problem: find the minimum of

$$f(x_1, x_2, \cdots x_N) = \frac{1}{x_1} + \frac{1}{x_2} + \dots + \frac{1}{x_N}$$

under the condition $N \ge 2$ $x_i > 0$ i$= 1, \dots, N$ and $x_1 x_2 \cdots x_N = 1$.

CHAPTER 2

Optimization in $\mathbb{R}^N$

We denote by $B(x,r)$ the *ball* of center $x \in \mathbb{R}^N$ and radius $r > 0$, that is $B(x,r) = \{y \in \mathbb{R}^N : \|y - x\| \le r\}$,

DEFINITION. Given a set Ω, we say that $x \in \mathbb{R}^N$ is an accumulation point for Ω if there exists a sequence (x_n) such that $x_n \in \Omega$, $x_n \neq x$ and $\lim_{n\to\infty} \|x_n - x\| = 0$. Equivalently, if for any $\delta > 0$, there exists $y \neq x$ such that $y \in \Omega \cap B(x,\delta)$.

1. Liminf and limsup

Given a sequence (x_n) we define

$$\liminf_{n\to+\infty} x_n = \lim_{n\to+\infty} \big(\inf_{m\ge n} x_m \big) \qquad \limsup_{n\to+\infty} x_n = \lim_{n\to+\infty} \big(\sup_{m\ge n} x_m \big)$$

Observe that in the set of the extended real number $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$, the lim inf and lim sup always exist. Moreover the sequence (x_n) admits limit ℓ if only if $\limsup_{n\to+\infty} x_n = \liminf_{n\to+\infty} x_n = \ell$.

EXAMPLE 7. The sequence $x_n = (-1)^n$ does not admit limit, while

$$\liminf_{n\to+\infty} x_n = -1 \qquad \limsup_{n\to+\infty} x_n = 1.$$

Given a function $f : \Omega \to \mathbb{R}$ and an accumulation point x of Ω, we define the liminf and limsup of function f at x as

$$\liminf_{y\to x} f(y) = \sup_r \inf_{y\in B(x,r)\setminus\{x\}} f(y) \tag{2.1}$$

$$\limsup_{y\to x} f(y) = \inf_r \sup_{y\in B(x,r)\setminus\{x\}} f(y) \tag{2.2}$$

DEFINITION. We say that a function

i) $f : \Omega \to \mathbb{R} \cup \{+\infty\}$ is *lower semi-continuous* (lsc) at $x \in \Omega$ if for any sequence $(x_j) \subset \Omega$ converging to x as $j \to +\infty$ we have

$$f(x) \le \liminf_j f(x_j).$$

Moreover f is lsc in Ω if it is lsc at x for all $x \in \Omega$.

ii) A $f : \Omega \to \mathbb{R} \cup \{-\infty\}$ is *upper semi-continuous* (usc) at $x \in \Omega$ if for any sequence $(x_j) \subset \Omega$ converging to x we have

$$f(x) \ge \limsup_j f(x_j).$$

Moreover f is usc in Ω if it is usc at x for all $x \in \Omega$.

iii) $f : \Omega \to \mathbb{R}$ is continuous at x if for any sequence $(x_j) \subset \Omega$ converging to x we have

$$f(x) = \lim_j f(x_j).$$

Moreover f is continuous in Ω if it is continuous at x for all $x \in \Omega$

Notice that f is lsc if and only if $f(x) = \liminf_{y\to x} f(y)$, $\forall x \in \Omega$. Moreover a function $f$is continuous at x (in Ω) if and only is lower and upper semi-continuous is continuous at x (in Ω).

REMARK. The continuity of f at x is equivalent to

$$f(x + h) = f(x) + \omega(h)$$

with $\omega(h) \to 0$ as $h \to 0$.

.

2. Compact sets and Weierstrass theorem

DEFINITION. A set of $\Omega \subset \mathbb{R}^N$ is said *compact* if every open cover has a finite subcover, i.e. for every arbitrary collection $\{U_\alpha\}_{\alpha\in A}$ such that $\Omega = \cup_{\alpha\in A} U_\alpha$, then there exists a finite set $B \subset A$ such that $\Omega = \cup_{\alpha\in B} U_\alpha$

PROPOSITION 2.1. *For any set of $\Omega \subset \mathbb{R}^N$, the following three conditions are equivalent:*

i) *Every open cover has a finite subcover.*
ii) *Every sequence in the set has a convergent subsequence, whose limit point belongs to Ω (i.e., Ω is sequentially compact).*
iii) *Ω is closed and bounded.*

EXAMPLE 8. Example of compact sets in $\mathbb{R}^N$ are: closed N-balls, N-spheres, the Cantor set.

DEFINITION. We say that

i) $f(x_0)$ is the *minimum* (respectively, *maximum*) of f in Ω iff

$$f(x_0) \le f(x), \quad \forall x \in \Omega \qquad \text{(respectively, } f(x_0) \ge f(x)\ \forall x \in \Omega) \ .$$

ii) $f(x_0)$ is a *relative minimum* (respectively, *relative maximum*) of f in Ω iff there exists $\delta > 0$ such that

$$f(x_0) \le f(x), \quad \forall x \in \Omega \cap B_\delta(x_0) \qquad \text{(respectively, } f(x_0) \ge f(x)\ \forall x \in \Omega \cap B_\delta(x_0))$$

An important problem in the applications is to establish the extremal value of a function. The Weierstrass' theorem gives a condition for the existence of minima and maxima.

THEOREM 2.2 (Weierstrass). *If $f : \Omega \subset \mathbb{R}^N \to \mathbb{R}$ is lower semi-continuous (respectively, upper semi-continuous) and Ω is a compact set, then there exists the minimum (respectively, maximum) of f in Ω.*

In particular, a continuous function admits maximum and minimum on a compact set.

Observe that without the compactness of the set Ω, existence of extremals is not guarantee. For example $\arctan x$ is continuous and bounded in $\mathbb{R}$, but it does not admit maximum and minimum in $\mathbb{R}$. On the other side existence of extremal can happen even if the hypothesis of previous theorem are not satisfied.

2.1. Minima in an unbounded set. We introduce a simple condition to guarantee the existence of a minimum in an unbounded set.

THEOREM 2.3. *Let $f : \mathbb{R}^N \to \mathbb{R}$ continuous, and*

$$\lim_{|x|\to+\infty} f(x) = +\infty,$$

there there exists x^* such that $f(x^*) = \inf_{\mathbb{R}^N} f(x)$

EXAMPLE 9. The problem is to minimize

$$f(x) = \frac{1}{2}Qx \cdot x + px, \quad \text{in } \mathbb{R}^N$$

with Q a $N \times N$ positive definite matrix, and $p \in \mathbb{R}^N$.
Then f is continuous in all the space $\mathbb{R}^N$,and, for any $x \neq 0$,we have

$$Qx \cdot x = Q\frac{x}{|x|}|x|\frac{x}{|x|}|x| = |x|^2 Q\frac{x}{|x|}\frac{x}{|x|} \geq 2c|x|^2,$$

with c a positive constant. Then, by Cauchy Schwarz inequality

$$f(x) \geq c|x|^2 - |p||x|,$$

then the assumptions of (2.3) are verified and we can conclude that the minimum exists. By the second order sufficient conditions(2.8) the unique (by strong convexity) minimum point is given by $x = -Q^{-1}p$.

THEOREM 2.4. *Let $K \neq \emptyset$ a closed, unbounded subset of $\mathbb{R}^N$. Let $f : K \to \mathbb{R}$ continuous, and*

$$\lim_{|x|\to+\infty,\, x\in K} f(x) = +\infty,$$

then there there exists x^* such that $f(x^*) = \inf_K f(x)$

PROOF. To see this, fix $x_0 \in K$, and say $m = f(x_0)$. Then we can fix a δ_0 such that for all $\delta > \delta_0$, $\quad f(x) > m$. The set $K_0 = \{x \in K \text{ such that } |x| \leq \delta_0$ is closed and bounded, by Wierstrass theorem there exists a minimum m^* on K_0. Since $x_0 \in K$, $m \leq m^*$. This shows that m^* is the global minimum, since x_0 was arbitrarily chosen in K. □

3. Necessary and sufficient conditions for extremals

By $Df(x)$ we denote the *gradient* of f evaluate at x, i.e.

$$Df(x) = \big(\frac{\partial f}{\partial x_1}(x), \frac{\partial f}{\partial x_2}(x), \ldots, \frac{\partial f}{\partial x_N}(x)\big).$$

We also recall the definition of *directional derivate*: if $v \in \mathbb{R}^N$ and $|v| = 1$ the directional derivatives of f with respect to the direction v is

$$D_v f(x) = \lim_{t\to 0} \frac{f(x+tv) - f(x)}{t}. \tag{2.3}$$

In particular for $v = e_i$, $D_v f(x) = \frac{\partial f}{\partial x_i}(x)$.
The *differentiability* of f at x means that the gradient exists and

$$f(x) = f(x_0) + Df(x_0) \cdot (x - x_0) + o(\|x - x_0\|), \tag{2.4}$$

where the Landau symbol $o(h)$ means a real function such that $\lim_{h\to 0} \frac{o(h)}{|h|} = 0$. It is easy to see that if f is differentiable at x, then

$$D_v f(x) = Df(x) \cdot v$$

THEOREM 2.5 (First-Order Necessary Condition). *Let $\Omega \subset \mathbb{R}^N$ and $x_0 \in \Omega$ a minimizer of f in Ω. If f is differentiable at x_0, then*

$$Df(x_0) \cdot (x - x_0) \geq 0, \qquad \forall x \text{ such that } x \in \Omega. \tag{2.5}$$

PROOF. Follows by (2.4). □

We say that a point x_0 is in the interior of Ω if there exists $\delta > 0$ such that $B(x_0, \delta) \subset \Omega$. By Theorem 2.5 it immediately follows the *Fermat theorem*

COROLLARY 2.6 (Fermat). *Let $\Omega \subset \mathbb{R}^N$ and $x_0 \in \Omega$ a minimizer of f in Ω. If f is differentiable at x_0 and x_0 is in the interior of Ω, then*

$$Df(x_0) = 0.$$

If f admits second order partial derivates in Ω, we associate to f its *Hessian matrix*

$$D^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_N} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial^2 f}{\partial x_N \partial x_1} & \frac{\partial^2 f}{\partial x_N \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_N \partial x_N} \end{pmatrix} \tag{2.6}$$

THEOREM 2.7 (Second-Order Necessary Condition). *Let $\Omega \subset \mathbb{R}^N$ and $f \in C^2(\Omega)$. If $x_0 \in \Omega$ is a local minimum point (respectively, local maximum point) of f in the interior of Ω, then*

$$D^2 f(x_0) \geq 0 \qquad \textit{(respectively } D^2 f(x_0) \leq 0\textit{)}.$$

PROOF. If f is a $C^2(\Omega)$ and x_0 is in the interior of Ω, by second order Taylor's expansion at x_0 we have

$$f(x) = f(x_0) + Df(x_0) \cdot (x - x_0) + \frac{1}{2}(x - x_0)^T D^2 f(x_0)(x - x_0) + o(\|x - x_0\|^2) \tag{2.7}$$

for $x \to x_0$. The statement follows immediately observing that if x_0 is a local minimum point, then $Df(x_0) = 0$ and therefore

$$0 \leq f(x) - f(x_0) \leq \frac{1}{2}(x - x_0)^T D^2 f(x_0)(x - x_0)$$

for x close to x_0. □

THEOREM 2.8 (Second-Order Sufficient Condition). *Let $\Omega \subset \mathbb{R}^N$ and $f \in C^2(\Omega)$. If $x_0 \in \Omega$ is in the interior of Ω and*

$$Df(x_0) = 0, \qquad D^2 f(x_0) > 0, \qquad \textit{(respectively } D^2 f(x_0) < 0\textit{)}$$

then x_0 is a (strict) local minimum (respectively, local maximum) of f in Ω.

PROOF. The proof follows immediately by (2.7) □

REMARK. If $N = 2$, then for $x = (x_1, x_2)$

$$D^2 f(x) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2}(x) & \frac{\partial^2 f}{\partial x_1 \partial x_2}(x) \\ \frac{\partial^2 f}{\partial x_2 \partial x_1}(x) & \frac{\partial^2 f}{\partial x_2^2}(x) \end{pmatrix} \tag{2.8}$$

By Prop.1.7, we get the 2^{nd} order necessary conditions

$$D^2 f(x_0) \geq 0 \iff det(D^2 f) \geq 0 \text{ and } \frac{\partial^2 f}{\partial x_1^2}(x_0) \geq 0,$$

$$D^2 f(x_0) \leq 0 \iff det(D^2 f) \geq 0 \text{ and } \frac{\partial^2 f}{\partial x_1^2}(x_0) \leq 0,$$

and 2^{nd} order sufficient ones

$$D^2 f(x_0) > 0 \iff det(D^2 f) > 0 \text{ and } \frac{\partial^2 f}{\partial x_1^2}(x_0) > 0$$

$$D^2 f(x_0) < 0 \iff det(D^2 f) > 0 \text{ and } \frac{\partial^2 f}{\partial x_1^2}(x_0) < 0$$

4. Necessary and sufficient conditions in weak form

In many applications, the requirement of differentiability is a too strong assumption which is not satisfied by the data of the problem. The following lemma is the basis for the introduction of a *weak* notion of gradient.

LEMMA 2.9. *Assume $f : \mathbb{R}^N \to \mathbb{R}$ is a continuous function and it is differentiable at x_0. Then there exists $\phi \in C^1(\mathbb{R}^N)$ such that $f(x_0) = \phi(x_0)$ and $f - \phi$ has a strict local minimum at x_0.*

PROOF. Replacing u by $v(x) = f(x + x_0) - f(x_0) - Df(x_0) \cdot x$ we may assume w.l.o.g.

$$x_0 = 0, \quad f(0) = Df(0) = 0.$$

Then

$$f(x) = |x| f_1(x),$$

with $f_1(x) : \mathbb{R}^N \to \mathbb{R}$ a continuous function such that $f_1(0) = 0$. Then we set

$$f_2(|x|) = \inf_{|y| \leq |x|} |f_1(y)|$$

We have $f_2 : [0, +\infty) \to [0, +\infty)$, continuous and non increasing. Define

$$\phi(x) = \int_{|x|}^{|2x|} f_2(t)dt - |x|^2. \tag{2.9}$$

By (2.9) $\phi(0) = D\phi(0) = 0$ ($|\phi(x)| \leq |x| f_2(|x|) - |x|^2$), and for any $x \neq 0$ the gradient can be computed

$$D\phi(x) = \frac{2x}{|x|} f_2(|2x|) - \frac{x}{|x|} f_2(|x|) - 2x.$$

Moreover if $x \neq 0$

$$f(x) - \phi(x) \geq |x|^2 > 0 = f(0) - \phi(0),$$

hence ϕ has the required properties. □

DEFINITION. Given a function $f : \Omega \to \mathbb{R}$ and $x \in \Omega$,

i) the *super-differential* of f in x is the *set*

$$D^+f(x) := \{p \in \mathbb{R}^N \ : \ f(x+h) \leq f(x) + ph + o(|h|), \quad h \to 0\}.$$

ii) the *sub-differential* of f in x is the *set*

$$D^-f(x) := \{p \in \mathbb{R}^N \ : \ f(x+h) \geq f(x) + ph + o(|h|), \quad h \to 0\}.$$

The notions of super-differential and sub-differential allow us to generalize some basic results about differentiable functions. For example we have

PROPOSITION 2.10. *Let $f : \Omega \to \mathbb{R}$ be a function and $x \in \Omega$.*
(i) If f has a local maximum in x, then $0 \in D^+f(x)$.
(ii) If f has a local minimum in $x \in \Omega$, then $0 \in D^-f(x)$.

PROOF. If f has a local maximum at $x \in \Omega$, then $f(x+h) - f(x) \leq 0$ for every h, close to zero. Hence

$$f(x+h) \leq f(x) + 0 \cdot h + o(|h|)$$

for $h \to 0$ and thus $0 \in D^+f(x)$. The other case is similar. □

The weak form of Theorem 2.5 is the following

THEOREM 2.11 (Weak First-Order Necessary Condition). *Let $\Omega \subset \mathbb{R}^N$ and $x_0 \in \Omega$ be a minimizer of f in Ω. If f is lsc at x_0 , then there exists a function $\phi \in C^1(\Omega)$ such that $f(x_0) = \phi(x_0)$, x_0 is minimizer for ϕ in Ω and*

$$D\phi(x^*) = 0, \ \forall x \text{ such that } x - x^* \in \Omega.$$

PROOF. Replacing u by $v(x) = f(x + x_0) - f(x_0) - p \cdot x$ we may assume

$$x_0 = 0, \ f(0) = p = 0.$$

Then the condition $0 \in D^-f(x)$ is equivalent to

$$f(x) \geq |x| f_1(x),$$

with $f_1(x) : \mathbb{R}^N \to \mathbb{R}$ and $f_1 \to 0$ as $x \to 0$. Then we set

$$f_2(|x|) = \inf_{|y| \leq |x|} |f_1(y)|$$

We have $f_2 : [0, +\infty) \to [0, +\infty)$ $f_2 \leq f_1$, f_2 non increasing and

$$f(x) \geq |x| f_2(x),$$

near 0. Then we take and

$$\phi(x) = \int_{|x|}^{2|x|} f_2(t)dt - |x|^2. \tag{2.10}$$

By (2.9) $\phi(0) = D\phi(0) = 0$ ($|\phi(x) \leq |x| f_2(|x|) - |x|^2$), and for any $x \neq 0$ the gradient can be computed

$$D\phi(x) = \frac{2x}{|x|} f_2(|2x|) - \frac{x}{|x|} f_2(|x|) - 2x.$$

Moreover if $x \neq 0$

$$f(x) - \phi(x) \geq |x|^2 > 0 = f(0) - \phi(0),$$

hence ϕ has the required properties. □

An important tool to obtain this type of results is the Dini's Theorem.

5. Dini's theorem

Let $f : I \subset \mathbb{R}^2 \to \mathbb{R}$ be a function defined in an open set I and consider the equation

$$f(x, y) = 0. \tag{2.11}$$

In many applications, it is important to transform the implicit relation between x and y given by f in an explicit one of the type $y = \phi(x)$. The *Dini's Implicit Function Theorem* states that this is possible under some mild assumptions on f and its partial derivates.

THEOREM 2.12. *Let $f : I \subset \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ be a C^1 function in the open set I and $(x_0, y_0) \in I$ where $x_0 = (x_1^0, \dots, x_N^0)$ and $y_0 \in \mathbb{R}$. If*

- $f(x_0, y_0) = 0$
- $f_y(x_0, y_0) \neq 0$,

there exists $\delta, k > 0$ such that defined $I_1 = \{x \in \mathbb{R}^N : |x - x_i^0| \leq \delta, i = 1, \dots, N\}$, $I_2 = (y_0 - k, y_0 + k)$, then for any $x \in I_1$, there exists a unique $y \in I_2$ such that (2.11) is satisfied. It is therefore defined a function $\phi : I_1 \to I_2$ such that

- $y_0 = \phi(x_0)$;
- $f(x, \phi(x)) = 0$ *for any* $x \in I_1$;
- $\phi \in C^1(I_1)$ *and*

$$\phi_{x_i}(x) = -\frac{f_{x_i}(x, \phi(x))}{f_y(x, \phi(x))} \qquad \forall x \in I_1,\ i = 1, \dots, N. \tag{2.12}$$

PROOF. For simplicity we give a direct proof for the case $N = 2$. The general case can be proved by a fixed point argument.

We assume w.l.o.g. that $f_y(x_0, y_0) > 0$, hence there exists $R = [x_0 - h, x_0 + h] \times [y_0 - k, y_0 + k]$ such that $f_y(x, y) > 0$ for any $(x, y) \in R$. It follows that for any fixed $\bar{x} \in [x_0 - h, x_0 + h]$, the function $f(\bar{x}, y)$, $y \in [y_0 - k, y_0 + k]$ is strictly increasing.
In particular, for $\bar{x} = x_0$, since $f(x_0, y_0) = 0$, we have $f(x_0, y_0 - k) < 0$ and $f(x_0, y_0 + k) > 0$. By the continuity of the functions $f(\cdot, y_0 \pm k)$, we can determine $\delta > 0$ such that $f(x, y_0 - \delta) < 0$ and $f(x, y_0 + \delta) > 0$ for any $x \in (x_0 - \delta, x_0 + \delta)$.
We conclude that for any $x \in (x_0 - \delta, x_0 + \delta)$, the function $f(x, \cdot)$ for $y \in [y_0 - k, y_0 + k]$ is strictly increasing and such that $f(x, y_0 - k) f(x, y_0 + k) < 0$. By the intermediate value theorem, for any $x \in (x_0 - \delta, x_0 + \delta)$, there exists a unique $y \in (y_0 - k, y_0 + k)$ such that $f(x, y) = 0$.
Now define the function $\phi : I_1 \to I_2$ by associating to x the unique y such that $f(x, y) = 0$. Hence $y_0 = \phi(x_0)$ and $g(x, \phi(x)) = 0$ for any $x \in I_1$. Let

us show that ϕ is C^1 and (2.14). Given $x \in I_1$, consider an increment Δx such that $x + \Delta x \in I_1$ and set $\Delta\phi = \phi(x + \Delta x) - \phi(x)$. We have

$$f(x + \Delta x, \phi(x) + \Delta\phi) - f(x, \phi(x)) = 0.$$

By Lagrange formula we have that there exist $\theta \in [0, 1]$ such that

$$f_x(x + \theta\Delta x, \phi(x) + \theta\Delta\phi)\Delta x + f_y(x + \theta\Delta x, \phi(x) + \theta\Delta\phi)\Delta\phi = 0.$$

If Δx is sufficiently small in such a way that $(x + \Delta x, \phi(x) + \Delta\phi) \in R$, then $f_y(x + \theta\Delta x, \phi(x) + \theta\Delta\phi) \neq 0$ and therefore

$$\Delta\phi = -\frac{f_x(x + \theta\Delta x, \phi(x) + \theta\Delta\phi)}{f_y(x + \theta\Delta x, \phi(x) + \theta\Delta\phi)}\Delta x. \tag{2.13}$$

Since $f \in C^1$ and R is compact, there exist m, $M > 0$ such that

$$f_y(x, y) \geq m,\ |fx(x, y)| \leq M|qquad\forall(x, y) \in R.$$

Hence by (2.13) we get

$$|\Delta\phi| \leq \frac{M}{m}|\Delta x|$$

and therefore $\lim_{\Delta x\to 0} \Delta\phi = 0$, giving the continuity of ϕ in I_1. For $\Delta x \neq 0$, by (2.13) we get

$$\frac{\Delta\phi}{\Delta x} = -\frac{f_x(x + \theta\Delta x, \phi(x) + \theta\Delta\phi)}{f_y(x + \theta\Delta x, \phi(x) + \theta\Delta\phi)}$$

and therefore for $\Delta x \to 0$ we get the derivability of ϕ and (2.11) □

REMARK. Exchanging the role of x and y, we get that if $f(x_0, y_0) = 0$ and $f_x(x_0, y_0) \neq 0$, then there exists a neighborhood J_1 of y_0 and a function ψ such that $y_0 = \psi(x_0)$, $g(\psi(y), y) = 0$ for any $y \in J_1$, $\psi \in C^1(J_1)$ and $\psi'(x) = -\frac{f_y(\psi(y),y)}{f_x(\psi(y),y)}$ for all $y \in J_1$.

REMARK. If $f \in C^k(I)$, then it is possible to prove that $\phi \in C^k(I_1)$. By applying the chain rule for the relation $f(x, \phi(x)) = 0$ it is possible to deduce a formula for all the derivatives of the function ϕ. For $n = 2$ we get

$$\phi''(x) = -\frac{f_{xx}f_y^2 - 2f_{xy}f_xf_y + f_{yy}f_y^2}{f_y^3}.$$

Since in general the function ϕ is not known explicitly, with the aid of the formulas for its derivatives, we can write a Taylor expansion of the function ϕ near the point x_0.

We conclude this section with a vectorial version of the Dini's implicit function theorem. We recall that for a function $f : \mathbb{R}^N \to \mathbb{R}^M$, $f = (f_1, \ldots, f_M)$, the Jacobian matrix $Jf(x_0)$ at x_0 is the $M \times N$ matrix defined by

$$Jf(x_0) = \left.\frac{\partial(f_1, \ldots, f_M)}{\partial(x_1, \ldots, x_N)}\right|_{x_0}$$

$$= \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(x_0) & \frac{\partial f_1}{\partial x_2}(x_0) & \cdots & \frac{\partial f_1}{\partial x_N}(x_0) \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial f_M}{\partial x_1}(x_0) & \frac{\partial f_M}{\partial x_2}(x_0) & \cdots & \frac{\partial f_M}{\partial x_N}(x_0) \end{pmatrix}$$

THEOREM 2.13. *Let $f : I \subset \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}^M$ be a C^1 function in the open set I and $(x_0, y_0) \in I$ where $x_0 = (x_1^0, \dots, x_N^0)$ and $y_0 = (y_1^0, \dots, y_M^0)$. If*

- *$f(x_0, y_0) = 0$*
- *$det[J_y f]|_{(x_0,y_0)} \neq 0$ ($J_y f$ is the Jacobian of f with respect to y)*

there exist open neighborhoods A_{x_0} of x_0 in $\mathbb{R}^N$ and B_{y_0} of y_0 in $\mathbb{R}^M$ such that for $x \in A_{x_0}$, the equation $f(x, y) = 0$ admits a unique solution $y \in B_{y_0}$. It is therefore defined a function $\phi : A_{x_0} \to B_{y_0}$ such that

- *$y_0 = \phi(x_0)$;*
- *$f(x, \phi(x)) = 0$ for any $x \in A_{x_0}$;*
- *$\phi \in C^1(A_{x_0}, B_{y_0})$ and*

$$\frac{\partial \phi_i}{\partial x_j}(x) = -\frac{det\left(\frac{\partial(f_1,\dots,f_i,\dots,f_M)}{\partial(y_1,\dots,x_j,\dots,y_M)}\right)_{(x,\phi(x))}}{det\left(\frac{\partial(f_1,\dots,f_M)}{\partial(y_1,\dots,y_M)}\right)_{(x,\phi(x))}} \tag{2.14}$$

for all $x \in A_{x_0}$, $i = 1, \dots, N$, $j = 1, \dots, M$.

6. Extremals under state constraints

As we have seen in the previous sections, we have necessary and sufficient conditions to study extremals in the interior of a given set. In this section we look for (necessary) conditions for extremals on the boundary of a set. Let $N = 2$ and consider a function $g \in C^1(I)$, where I is an open set. Define the set

$$Z = \{(x, y) \in \mathbb{R}^2 : g(x, y) = 0\}$$

and assume that $Z \neq \emptyset$. Let $f : I \to \mathbb{R}$ be a C^1 function and assume that there exists a point $(x_0, y_0) \in Z$ such that $g_y(x_0, y_0) \neq 0$ and (x_0, y_0) is a local minimum [maximum] point of f relatively to the constraint Z, i.e. there exists a neighborhood U of (x_0, y_0) such that $f(x_0, y_0) \leq f(x, y)$ $[f(x_0, y_0) \geq f(x, y)]$ for any $(x, y) \in U \cap Z$.
Arguing formally, by the implicit function theorem, we have that for some function $\phi \in C^1(I_1)$, with I_1 a suitable neighborhood of x_0, the constraint Z can be described near x_0 as the set $\{(x, y) : x \in I_1,\ y = \phi(x)\}$ and moreover x_0 is an extremal point of $f(x, \phi(x))$ in I_1. Hence by the chain rule and $y_0 = \phi(x_0)$ we get

$$f_x(x_0, y_0) + f_y(x_0, y_0)\phi'(x_0) = 0. \tag{2.15}$$

On the other hand, by Dini's theorem

$$\phi'(x_0) = -\frac{g_x(x_0, y_0)}{g_y(x_0, y_0)}.$$

Substituting in (2.15), we get the necessary condition for the extremum

$$f_x(x_0, y_0) - f_y(x_0, y_0)\frac{g_x(x_0, y_0)}{g_y(x_0, y_0)} = 0.$$

If also $g_x(x_0, y_0) \neq 0$ then setting

$$\lambda^* = -\frac{f_y(x_0, y_0)}{g_y(x_0, y_0)} = -\frac{f_x(x_0, y_0)}{g_x(x_0, y_0)},$$

the *conditions* for f having an extremal point under the constraint Z are

$$\begin{cases} Df(x_0, y_0) + \lambda^* Dg(x_0, y_0) = 0, \\ g(x_0, y_0*) = 0 \end{cases} \tag{2.16}$$

The previous formal approach to the existence of extreme under constraints will be made rigorous in the next *Lagrange Multiplier Theorem.*

More generally, we consider the problem of minimizing [maximizing] a function subject to some constraints, i.e. the problem to minimize [maximize] f for when the variable x satisfy a constraint $g(x) = 0$ for some function $g : \mathbb{R}^N \to \mathbb{R}^M$, $M < N$.

THEOREM 2.14. *Let I an open subset of $\mathbb{R}^N$, $f : I \to \mathbb{R}$, $g : I \to \mathbb{R}^M$, C^1 functions in I and $x_0 \in I$. If there exists an open neighborhood U of x_0 in $\mathbb{R}^N$ such that*

$$f(x) \leq f(x_0) \quad [f(x) \geq f(x_0)] \qquad \forall x \in U \cap \{x \in I : g(x) = 0\}$$

then there exist μ, $\lambda = (\lambda_1, \ldots, \lambda_M)$, not both zero, such that

$$\begin{cases} \mu \frac{\partial f}{\partial x_i}(x_0) + \sum_{j=1}^M \lambda_j \frac{\partial g_j}{\partial x_i}(x_0) = 0, \quad i = 1, \ldots, M \\ g_i(x_0) = 0, \quad i = 1, \ldots, M \end{cases} \tag{2.17}$$

REMARK. It is worthwhile to remark that the Lagrange conditions (2.17) are *necessary* but *not sufficient* for x_0 being an extremal point of f under the constraint $g(x) = 0$. For sufficient conditions as usual is necessary to introduce inequality involving the second order derivatives of f and g.

COROLLARY 2.15. *Under the same assumptions of Theorem 2.14, if $Jg(x)$ (the Jacobian matrix of g) has rank M at x_0, then there exists a unique $\lambda \in \mathbb{R}^M$ such that*

$$\begin{cases} \frac{\partial f}{\partial x_i}(x_0) + \sum_{j=1}^M \lambda_j \frac{\partial g_j}{\partial x_i}(x_0) = 0, \quad i = 1, \ldots, M \\ g_i(x_0) = 0, \quad i = 1, \ldots, M \end{cases} \tag{2.18}$$

REMARK. Note that the first condition in (2.18) can be reformulated by saying that the function $F(x) = f(x) + \sum_{j=1}^M \lambda_j g_j(x)$ satisfies $DF(x_0) = 0$. Moreover observe that (2.18) is a system of $N + M$ equation in the $N + M$ unknowns $(x_1^0, \ldots, x_N^0)$, the coordinate of the extremal point, and $(\lambda_1, \ldots, \lambda_M)$, the Lagrange multiplier.

PROOF OF THEOREM 2.14. Let us consider the application

$$\begin{aligned} &\phi : I \times \mathbb{R} \to \mathbb{R}^{M+1} \\ &(x, u) \to \phi(x, u) = (f(x) - f(x_0) + u, g_1(x), \ldots, g_M(x)). \end{aligned}$$

The map ϕ is C^1 in $I \times \mathbb{R}$ and $\phi(x_0, 0) = 0$. We claim that the Jacobian matrix of the function $F(x) = (f(x), g_1(x), \ldots, g_m(x))$ cannot have rank $M + 1$.In fact, assume for example that

$$det \left\{ \frac{\partial(f, g_1, \ldots, g_M)}{\partial(x_1, \ldots, x_{M+1})} \right\}_{x_0} \neq 0.$$

Then, defined $\tilde{f}(x,u) = f(x) - f(x_0) + u$, since

$$det\left\{\frac{\partial(f,g_1,\dots,g_M)}{\partial(x_1,\dots,x_{M+1})}\right\}_{x_0} = det\left\{\frac{\partial(\tilde{f},g_1,\dots,g_M)}{\partial(x_1,\dots,x_{M+1})}\right\}_{x_0} \neq 0$$

by the Dini's Theorem 2.13 we get that in the equation $\phi(x,u) = 0$ we can be explicit the variable $(x_1,\dots,x_{M+1})$ with respect to the other variables in a neighborhood of $(x_0,0)$.

If we therefore set $\xi = (x_{M+2},\dots,x_N)$, $\eta = (x_1,\dots,x_{M+1})$, $\xi_0 = (x^0_{M+2},\dots,x^0_N)$ and $\eta = (x^0_1,\dots,x^0_{M+1})$, we get that there exist $\delta > 0$ and open neighborhoods A of ξ_0 in $\mathbb{R}^{N-M-1}$ and B of η_0 in $\mathbb{R}^{M+1}$ and a function $\phi : A \times (-\delta,\delta) \to \mathbb{R}^{M+1}$ such that for any $(\xi,u) \in A \times (-\delta,\delta)$, $\phi(x,u) \in B$ and

$$\begin{aligned} &f(\phi(\xi,u),\xi) - f(x_0) + u = 0 \\ &g_i(\phi(\xi,u),\xi) = 0 \qquad i = 1,\dots,M \end{aligned}$$

If we take A and B sufficiently small in such a way that $A \times B \subset U$, then $(\phi(\xi,u),\xi) \in U$ and

$$f(\phi(\xi,u),\xi) = f(x_0) - u \begin{cases} < f(x_0), & \text{if } u > 0; \\ > f(x_0), & \text{if } u < 0. \end{cases}$$

giving a contradiction to the existence of an extremal point in x_0. Hence the matrix $\left\{\frac{\partial(f,g_1,\dots,g_M)}{\partial(x_1,\dots,x_N)}\right\}_{x_0}$ cannot have full rank. Therefore the vectors $(\frac{\partial f}{\partial x_1}(x_0),\dots,(\frac{\partial f}{\partial x_N}(x_0),\ (\frac{\partial g_i}{\partial x_1}(x_0),\dots,(\frac{\partial g_i}{\partial x_N})(x_0),\ i = 1,\dots,M$ given by its rows are linear dependent in $\mathbb{R}^N$ and we get the first condition in (2.17), the second one expressing the condition that x_0 is on the constraint. □

Proof of Corollary 2.15. Consider the homogeneous linear system in the unknown $z = (z_1,\dots,z_M)$

$$\sum_{j=1}^{M} z_j(\frac{\partial g_j}{\partial x_i})(x_0) = 0 \qquad i = 1,\dots,N \tag{2.19}$$

Since$[Jg(x_0)]$ has rank M, then the linear system admits only the null solution. If $\mu = 0$ in (2.17), we should have also $\lambda = 0$ and therefore a contradiction since μ and λ cannot be both null. Hence we can divided by μ the first equation in (2.17) to get the first equation in (2.18). Uniqueness of λ follows again by $det[Jg(x_0)] \neq 0$. □

Example 10 (Minima in an affine set). Given $x, b \in \mathbb{R}^N$ and an $N \times N$ matrix A, f real and convex in $\mathbb{R}^N$

$$\text{Minimize } f(x); \text{ such that } Ax = b.$$

The condition $Df(x^*) \in \mathcal{R}(A^T)$ can be obtained following the optimality conditions (see (2.5))

$$Df(x^*)(x - x^*) \geq 0, \qquad \forall x \text{ verifying } Ax = b.$$

Then $x = x^* + v$, $v \in \mathcal{N}(A)$ we get $Df(x^*)v = 0$, $v \in \mathcal{N}(A)$. Since $\mathcal{N}(A) = \mathcal{R}(A^T)$, the above condition means $Df(x^*) \in \mathcal{R}(A^T)$. Applying

the Lagrange condition (2.17) we find that there exists $\lambda \in \mathbb{R}^N$ such that

$$\begin{cases} Df(x^*) + A^T\lambda = 0, \\ Ax^* = b, \end{cases} \tag{2.20}$$

giving again $Df(x^*) \in \mathcal{R}(A^T)$.

CHAPTER 3

Convex functions

1. Convex set

DEFINITION. A set $\Omega \subset \mathbb{R}^N$ is said *convex* if for any x and $y \in \Omega$,

$$\lambda x + (1-\lambda)y \in \Omega \qquad \text{for any } \lambda \in [0,1].$$

The previous definition says that if Ω contains x and y, then it contains the segment of vertices x and y. A ball $B_R(x)$ is a convex set, while the annulus $B_R(x) \setminus B_r(x)$ is not a convex set.

Given $y_1, \dots, y_k \in \mathbb{R}^N$ and k non negative numbers $\lambda_1, \lambda_2, \dots, \lambda_k$ such that $\sum_{k=1}^N \lambda_k = 1$, we consider the convex combination

$$\overline{y} = \lambda_1 y_1 + \lambda_2 y_2 + \dots \lambda_k y_k.$$

DEFINITION. Given a set Ω, the convex hull of Ω, $\mathrm{co}(\Omega)$ is smallest convex set containing Ω. Equivalently, $\mathrm{co}(\Omega)$ is the set obtained by all the possible convex combinations of points in Ω, i.e.

$$\mathrm{co}(\Omega) = \{\sum_i \lambda_i x_i : \, x_i \in \Omega, \lambda_i \geq 0, \sum_i \lambda_i = 1\}$$

A fundamental theorem, due to *Caratheodory*, say that the convex hull of a set can be obtained by taking the all the convex combinations of a finite number of points in Ω. The following result say that two disjoint convex sets can be always separated by an hyperplane

THEOREM 3.1 (Separation theorem). *Let C_1 and C_2 be two convex sets $\subset \mathbb{R}^N$ such that* $\mathrm{int}(C_1) \cap C_2 = \emptyset$ *(where* int *denotes the set of the interior points). Then there exists $p \in \mathbb{R}^N$, $p \neq 0$, such that*

$$py \geq px, \qquad \forall x \in C_1, y \in C_2 \tag{3.1}$$

2. Convex and concave functions

DEFINITION. Let C be an open, convex. A function $f : C \to \mathbb{R}$ is said to be

i) *convex* if

$$\lambda f(x) + (1-\lambda) f(y) \geq f(\lambda x + (1-\lambda)y) \quad \forall x, y \in C, \lambda \in [0,1]. \tag{3.2}$$

Moreover is said *strictly convex* if a strict inequality holds in (3.3) for $\lambda \in (0,1)$

ii) *concave* if $-f$ is convex, i.e.

$$\lambda f(x) + (1-\lambda) f(y) \leq f(\lambda x + (1-\lambda)y) \quad \forall x, y \in C, \lambda \in [0,1]. \tag{3.3}$$

Note that affine functions in $\mathbb{R}^N$ are convex and concave, while an example of strictly convex function is $f(x) = \|x\|^2$. The function in $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f(x) = \begin{cases} |x|^2, & x \geq 0, \\ |x| & x < 0 \end{cases} \tag{3.4}$$

is convex, but not strictly convex.

For regular functions we have characterizations of convexity by means of the derivatives of the function.

PROPOSITION 3.2. *Let $f : C \to \mathcal{R}$, where C is a convex subset of $\mathbb{R}^N$. Then:*

i) *If $f \in C^1(C)$, then f is convex in C if and only if*

$$f(y) - f(x) \geq Df(x)(y - x) for any x, y \in C. \tag{3.5}$$

ii) *If $f \in C^2(C)$, then f is convex in C if and only if $D^2 f(x) \geq 0$ for any $x \in C$.*

PROOF. □

In optimization theory, it is important to establish the uniqueness of minimum points.

PROPOSITION 3.3. *Let $C \subset \mathbb{R}^N$ be a convex set and $f : C \to \mathbb{R}$ be a strictly convex function. Then every local minimum point is global.*

PROOF. Assume by contradiction that there exist x^* local minimum point which is not global. Then then there exists $\delta > 0$ such that

$$f(x^*) \leq f(x), \qquad \forall x \in C \cap B_\delta(x^*)$$

and, for some $\hat{x}$, $f(\hat{x}) < f(x^*)$. Since $\hat{x} \neq x^*$, we take $\lambda \in (0, \min\{1, \frac{1}{|\hat{x} - x^\star|}\}$ and $x_\lambda = \lambda\hat{x} + (1 - \lambda)x^*$. Then

$$f(x_\lambda) \leq \lambda f(\hat{x}) + (1 - \lambda)f(x^*) < \lambda f(x^*) + (1 - \lambda)f(x^*) = f(x^*),$$

contradicting the assumption that x^* is a local minimum point. □

EXAMPLE 11. Let $D \neq \emptyset$ a closed, compact subset of $\mathbb{R}^N$ and $x \in \mathbb{R}^N$. Consider the following constrained minimization problem:

$$\text{Find } \hat{y} \in D \text{ such that } \|\hat{y} - x\| = \inf_{y \in D} \|\hat{y} - x\|.$$

As an application of the Weierstrass theorem, we see that the minimum of $f(y) = \|\hat{y} - x\|$ on D is attained, say at $y^* \in D$. We define the distance of x from D the quantity

$$d_D(x) = \|x - y^*\|.$$

Then, $\forall x \in \mathbb{R}^N$ there exists $\hat{y}$ minimizing the distance of the point x to the set D. Moreover it is possible to see that $\forall x \in \mathbb{R}^N \setminus D$ the minimum is attained on the boundary, i.e.

$$\min_{y \in D} \|y - x\| = \min_{y \in \partial D} \|y - x\|.$$

In general the point realizing the minimum is not unique. But if D is convex, then Prop. 3.3 says that for any $x \in \mathbb{R}^N$, there is a unique $y_x \in D$ such that $d_D(x) = \|x - y_x\|$. The point y_x is called the *projection* of x on the set D.

A property of convex function is the following

PROPOSITION 3.4. *Let C be an open, convex set. Then $f : C \to \mathbb{R}$ is convex (respectively concave) if and only if*

$$f(x) = \sup_{i\in\mathcal{I}} g_i(x), \qquad (\textit{respectively, } f(x) = \inf_{i\in\mathcal{I}} g_i(x)) \tag{3.6}$$

with $g_i(x)$ affine functions.

A subset of convex functions is the set of function strongly convex, that is

DEFINITION. Let C open, convex set. A function $f : C \to \mathbb{R}$ is said to be *strongly convex* if there exists $c > 0$ such that

$$\lambda f(x) + (1-\lambda)f(y) \geq f(\lambda x + (1-\lambda)y) - c\lambda(1-\lambda)|x-y|^2 \tag{3.7}$$

$\forall x, y \in C$, $\lambda \in [0,1]$.

The following proposition gives a helpful tool to show property by perturbation

PROPOSITION 3.5. *The function f is strongly convex on C (with constant c) if and only if the function $f - \frac{1}{2}c|x|^2$ is convex*

THEOREM 3.6 (Jensen inequality). *Let $f : C \to \mathbb{R}$ be a convex function of the convex set C, then f is convex if and only if*

$$f\left(\sum_{i=1}^{p} \lambda_i x_i\right) \leq \sum_{i=1}^{p} \lambda_i f(x_i),$$

for any finite subset $\{x_1, \dots, x_p\} \subset C$, and for any $\lambda_i \geq 0$, and $\sum_{i=1}^{p} \lambda_i = 1$.

2.1. Legendre-Fenchel transform. Let $f : \mathbb{R}^N \to \mathbb{R}$ be a convex function, satisfying the super-linearity condition

$$\lim_{|x|\to+\infty} \frac{f(x)}{|x|} = +\infty. \tag{3.8}$$

Then, the Legendre-Fenchel transform of f is defined by

$$f^*(x) = \sup_{y\in\mathbb{R}^N} \big[x \cdot y - f(y)\big] \qquad x \in \mathbb{R}^N \tag{3.9}$$

THEOREM 3.7. *Let C be an open, convex set, $f : C \to \mathbb{R}$ convex and satisfying* (3.8)*, then*

i) *for any x, there exists $y = y(x)$ such that the* sup *in* (3.9) *is attained.*
ii) *f^* is convex and $\lim_{|x|\to+\infty} \frac{f^*(x)}{|x|} = +\infty$.*
iii) *$f^{**} = f$*

3. Inf- and sup convolutions

DEFINITION. Let f continuous

Compute the inf- and sup convolutions in the case
$f(x) = 1 - |x|$ and $f(x) = |x| - 1$

4. First Order Condition for convex function

THEOREM 3.8. *Let $f \in C^1(C)$ be a convex function in the open set C. Given a convex set $K \subset K$ we have*

$$f(x^*) = \min_{x\in K} f(x) \Leftrightarrow x^* \in K \, and \, Df(x^*)(x - x^*) \geq 0, \, \forall x \in K \tag{3.10}$$

PROOF. By (3.5)

$$f(x) - f(x^*) \geq Df(x^*)(x - x^*) \qquad \text{for any } x \in K$$

by $Df(x^*)(x - x^*) \geq 0$ for any $x \in K$, we get $f(x^*) \leq f(x)$ for any $x \in K$. Viceversa, assume that x^* is a minimum point, take $x \neq x^* \in K$ and $x_\lambda = x^* + \lambda(x - x^*)$ with $\lambda \in (0, 1)$, then

$$\frac{f(x^* + \lambda(x - x^*)v) - f(x^*)}{\lambda} \geq 0$$

As $\lambda \to 0$ we have $Df(x^*)(x - x^*) \geq 0$. □

EXAMPLE 12. We explicit conditions to find the minimum of $f(x)$ in the convex set $\{x \geq 0\}$. By (3.10) to find the minimum point x^* we impose

$$x^* \geq 0 \qquad Df(x^*)(x - x^*) \geq 0, \forall x \geq 0. \tag{3.11}$$

Taking $x = x^* + e^i$ we get $\frac{\partial f}{\partial x_i}(x^*) \geq 0$. Since $x^* \geq 0$, by (3.10) we get $Df(x^*)x \geq Df(x^*)x^* \geq 0$ and therefore for $x = 0$ $Df(x^*)x^* = 0$. By the non negativity of each factor, $f_{x_i}(x^*)x_i^* = 0$ for $i = 1, \ldots, N$. Then by (3.11) we have the *complementary conditions*

$$x^* \geq 0 \qquad \frac{\partial f}{\partial x_i}(x^*) \geq 0, \;\; f_{x_i}(x^*)x_i^* = 0, \quad i = 1, \ldots, N. \tag{3.12}$$

4.1. Signed distance function. Given a set $D \neq \emptyset$, closed we define

$$d_+(x) = \begin{cases} -d(x, \partial D) & \text{if } x \in \text{Int(D)}, \\ 0 & \text{if } x \in \partial \text{D} \\ d(x, \partial D) & \text{if } x \in C(\text{Int(D)}) \end{cases} \tag{3.13}$$

The set D is described by $\{x \in \mathbb{R}^N \text{ such that } d_+(x) \leq 0\}$

CHAPTER 4

Optimization problems with unilateral and bilateral constraints

In section 6 we studied the problem of minimizing a function subject to bilateral constraints, i.e. the problem to minimize f for when the variable x satisfy a constraint $g(x) = 0$ for some function $g : \mathbb{R}^N \to \mathbb{R}^M$. In this section we generalize this problem by considering, besides the bilateral constraints, also unilateral ones, i.e. constraints defined by inequalities.

The problem is therefore the following:

Given $f : \mathbb{R}^N \to \mathbb{R}$ and $g : \mathbb{R}^N \to \mathbb{R}^M$, $h : \mathbb{R}^N \to \mathbb{R}^P$, find

$$\begin{aligned} \min \quad \{f(x) : x \in \mathbb{R}^N \text{s.t. } & g_i(x) \geq 0,\ i = 1, \dots, M, \\ & h_i(x) = 0,\ i = 1, \dots, P\} \end{aligned} \tag{4.1}$$

1. Necessary condition and the Fritz John's theorem

In this section we consider a generalization for problem with unilateral constraints of the Lagrange Multipliers Theorem 2.14. The proof is due to Fritz John and it is based on a penalization technique.

THEOREM 4.1. *Let I an open subset of $\mathbb{R}^N$, $f : I \to \mathbb{R}$, $g : I \to \mathbb{R}^M$, $h : I \to \mathbb{R}^P$, C^1 functions in I and $x_0 \in I$. If there exists an open neighborhood U of x_0 in $\mathbb{R}^N$ such that*

$$f(x_0) \leq f(x) \quad [f(x) \geq f(x_0)] \qquad \forall x \in U \cap \{x \in I :\ g(x) \leq 0,\ h(x) = 0\}$$

then there exist λ_0, $\lambda = (\lambda_1, \dots, \lambda_M)$ and $\mu = (\mu_1, \dots, \mu_P)$ such that

i)

$$\begin{cases} \lambda_0 \frac{\partial f}{\partial x_i}(x_0) + \sum_{j=1}^{M} \lambda_j \frac{\partial g_j}{\partial x_i}(x_0) + \sum_{j=1}^{P} \mu_j \frac{\partial h_j}{\partial x_i}(x_0) = 0,\ i = 1, \dots, N \\ \lambda_i g_i(x_0) = 0,\ i = 1, \dots, M,\ (\lambda_0, \lambda) \geq 0,\ (\lambda_0, \lambda, \mu) \neq 0 \\ g(x_0) \leq 0,\ h(x_0) = 0 \end{cases} \tag{4.2}$$

ii) *In any neighborhood of x_0, there exists x such that*

$$\begin{aligned} \lambda_i g_i(x) > 0 \qquad & \forall\ i\ s.t.\ \lambda_i > 0 \\ \mu_i h_i(x) > 0 \qquad & \forall\ i\ s.t.\ \mu_i \neq 0 \end{aligned} \tag{4.3}$$

PROOF. We consider the proof for the case of a minimum point, the case of a maximum point point being analogous By the definition of constrained minimum point and the continuity of f, g and h we can consider $\delta > 0$ such

that for any $x \in B(x_0,\delta) \cap \{x \in I : g(x) \le 0,\, h(x) = 0\}$

$$\begin{aligned} &f(x_0) \le f(x) \\ &g_i(x) < 0 \qquad \text{if } g_i(x_0) < 0 \end{aligned}$$

We introduce the penalized functional

$$\mathcal{F}_k(x) = f(x) + \frac{1}{2}\|x - x_0\|^2 + \frac{k}{2}\left(\sum_{i_1}^{M} g_i^+(x)^2 + \sum_{i_1}^{P} h_i(x)^2\right)$$

where $g_i(x)^+ = \max\{g_i(x), 0\}$ whose square is a C^1 function with gradient $2g_i^+(x)Dg_i(x)$. Consider the optimization problem

$$\min_{x \in B(x_0,\delta)} \mathcal{F}_k(x) \tag{4.4}$$

By Weierstrass' Theorem (see Theorem 2.2), there exists $x_k \in B(x_0,\delta)$ of minimum for $\mathcal{F}_k$ in $\overline{B(x_0,\delta)}$. In particular we have

$$\mathcal{F}_k(x_k) \le F_k(x_0) = f(x_0) \tag{4.5}$$

(recall that $g_i(x_0) \le 0$ and $h_i(x_0) = 0$). Moreover, by compactness, the sequence $\{x_k\}_{k\in\mathbb{N}}$ converges up to a subsequence to a point $x^+ \in B(x_0,\delta)$. By (4.5)

$$\sum_{i=1}^{M} g_i^+(x_k)^2 + \sum_{i=1}^{P} h_i(x_k)^2 \le \frac{2}{k}\left(f(x_0) - f(x_k) - \frac{1}{2}\|x_k - x_0\|^2\right)$$

and by the continuity of g_i, h_i we get for $k \to \infty$

$$\sum_{i=1}^{M} g_i^+(x^*)^2 + \sum_{i=1}^{P} h_i(x^*)^2 \le 0$$

and therefore

$$g_i(x^*) \le 0,\ i = 1,\dots,M, \text{ and } h_i(x^*) = 0,\ i = 1,\dots,P. \tag{4.6}$$

Moreover by (4.5), we get

$$f(x_k) + \frac{1}{2}\|x_k - x_0\|^2 \le \mathcal{F}_k(x_k) \le f(x_0)$$

and passing to the limit for $k \to \infty$ we get

$$f(x^*) + \frac{1}{2}\|x^* - x_0\|^2 \le f(x_0). \tag{4.7}$$

By (4.6), $x^* \in \{x \in I : g(x) \le 0,\, h(x) = 0\}$ and therefore $f(x^*) \ge f(x_0)$. Hence (4.8) we conclude that $\|x^* - x_0\|^2 = 0$ and therefore $x^* = x_0$. Since $x_k \to x_0$, we can conclude that for k sufficiently large $x_k \in B(x_0,\delta)$ and by Fermat's Theorem 2.6

$$\begin{aligned} \frac{\partial \mathcal{F}_k}{\partial x_i}(x_k) = \frac{\partial f}{\partial x_i}(x_k) + (x_k - x_0) + \sum_{j=1}^{M} k g_j^+(x_k)\frac{\partial g_j}{\partial x_i}(x_k) \\ + \sum_{j=1}^{P} k h_j(x_k)\frac{\partial h_j}{\partial x_i}(x_k) = 0, \qquad i = 1,\dots,N \end{aligned} \tag{4.8}$$

Define L^k, $\lambda_0^k \in \mathbb{R}$, $\lambda^k \in \mathbb{R}^M$, $\mu^k \in \mathbb{R}^P$ by

$$L^k = \left(1 + \sum_{j=1}^{M} (kg_j^+(x_k))^2 + \sum_{j=1}^{P} (kh_j(x_k))^2\right)^2,$$
$$\lambda_0^k = \frac{1}{L^k}, \qquad \lambda_i^k = \frac{kg_i^+(x_k)}{L^k}, \qquad \mu_i^k = \frac{kh_i(x_k)}{L^k}$$

then

$$|(\lambda_0^k, \lambda^k, \mu^k)|^2 = \left(\frac{1}{L^k}\right)^2 + \sum_{j=1}^{M} \left(\frac{kg_j^+(x_k)}{L^k}\right)^2 + \sum_{j=1}^{M} \left(\frac{kh_j(x_k)}{L^k}\right)^2 =$$
$$= \left(\frac{1}{L^k}\right)^2 \left(1 + \sum_{j=1}^{M} \left(kg_j^+(x_k)\right)^2 + \sum_{j=1}^{M} (kh_j(x_k))^2\right) = 1$$

By compactness the sequence $(\lambda_0^k, \lambda^k, \mu^k)$ converges, up to a subsequence, for $k \to \infty$ to a vector $(\lambda_0, \lambda, \mu)$ such that $|(\lambda_0, \lambda, \mu)| = 1$. Dividing (4.8) by L^k, we get

$$\lambda_0^k \frac{\partial f}{\partial x_i}(x_k) + \frac{(x_k - x_0)}{L^k} + \sum_{j=1}^{M} \lambda_j^k \frac{\partial g_j}{\partial x_i}(x_k) + \sum_{j=1}^{P} \mu_j^k \frac{\partial h_j}{\partial x_i}(x_k) = 0 \tag{4.9}$$

and recalling that, up to a subsequence, $x_k \to x_0$ and $(\lambda_0^k, \lambda^k, \mu^k) \to (\lambda_0, \lambda, \mu)$ we get the first condition in (4.2). Since $\lambda_0^k, \lambda^k \geq 0$ we also get at the limit $\lambda_0, \lambda \geq 0$.

Let i be such that $g_i(x_0) < 0$, then $g_i(x_k) < 0$ and therefore

$$\lambda_i^k = \max\{g_i(x_k), 0\} = 0 \tag{4.10}$$

Hence if $g_i(x_0) < 0$, we get $\lambda_i g(_i(x_0) = 0$. Taking into account for the other indices i, $g_i(x_0) = 0$, we can write $\lambda_i g_i(x_0) = 0$ for any $i = 1, \ldots, M$ obtaining in such a way all the conditions in (4.2).

To prove (4.3), observe that is $\lambda_i > 0$, then $\lambda_i^k > 0$ for k sufficiently large. Hence by (4.10) we get $g_i(x_k) > 0$ for such k. In a similar way if $\mu_i \neq 0$, then for k sufficiently large μ_i^k and therefore $h_i(x_k)$ must have same sign of μ_i. Therefore $\mu_i^k h_i(x_k) > 0$. It follows that for k sufficiently large we can find x_k such that

$$\lambda_i g_i(x_k) > 0 \qquad \forall\ i \text{ s.t. } \lambda_i > 0$$
$$\mu_i h_i(x_k) > 0 \qquad \forall\ i \text{ s.t. } \mu_i \neq 0$$

giving (4.3). □

Remark. If Jg and Jh are the Jacobian matrices of g and h, then (4.2) can be rewritten in matrix notation as

$$\begin{cases} \lambda_0 Df(x_0) + [Jg(x_0)]^T \lambda + [Jh(x_0)]^T \mu = 0 \\ \lambda g(x_0) = 0,\ (\lambda_0, \lambda) \geq 0,\ (\lambda_0, \lambda, \mu) \neq 0 \\ g(x_0) \leq 0,\ h(x_0) = 0 \end{cases} \tag{4.11}$$

REMARK. The previous theorem contains as a particular case Theorem 2.14. As already said for this latter theorem, conditions (4.2) are necessary but not sufficient conditions for x_0 being an extremal point.

It is worthwhile to observe that if the coefficient λ_0 multiplying the term Df is null, conditions (4.2) could be not very useful since they do not involve the function f. In Corollary 2.15, adding a condition about the rank of the matrix Jg, it was possible to avoid this degenerate case. We are therefore interested to find conditions in such a way that in (4.2) $\lambda_0 > 0$.

COROLLARY 4.2. *Under the same assumption of Theorem 4.1, define* $I*(x_0) = \{i \in \{1, \dots, M\} : g(x_0) = 0\}$ *and assume that that the* $\#(I^(x_0)+P$ *vectors* $\{Dg_i(x),\ i \in I^*(x_0)\}$, $\{Dh_i(x_0),\ i = 1, \dots, M\}$ *are linearly independent. Then there exist* $\lambda = (\lambda_1, \dots, \lambda_M)$ *and* $\mu = (\mu_1, \dots, \mu_P)$ *such that*

$$
\begin{cases}
\frac{\partial f}{\partial x_i}(x_0) + \sum_{j=1}^{M} \lambda_j \frac{\partial g_j}{\partial x_i}(x_0) + \sum_{j=1}^{P} \mu_j \frac{\partial h_j}{\partial x_i}(x_0) = 0,\ i = 1, \dots, N \\
\lambda_i g_i(x_0) = 0,\ i = 1, \dots, M, \\
g(x_0) \le 0,\ h(x_0) = 0,\ \lambda \ge 0
\end{cases}
\tag{4.12}
$$

PROOF. By Theorem 4.1 we know that there exist λ_0, λ and μ, not all null, such that conditions (4.2) are satisfied. We claim that $\lambda_0 \neq 0$. Assume by contradiction that $\lambda_0 = 0$, then recalling that $\lambda_i = 0$ if $g_i(x_0) < 0$, we get

$$
\sum_{j \in I^*(x_0)} \lambda_j \frac{\partial g_j}{\partial x_i}(x_0) + \sum_{j=1}^{P} \mu_j \frac{\partial h_j}{\partial x_i}(x_0) = 0 \qquad i = 1, \dots, N.
$$

By the linear independence of the vectors, we get $\lambda = 0$ and $\mu = 0$. Hence $\lambda_0 \neq 0$ and we can divided by λ_0 in the first condition of (4.2) to (4.12). □

Another important regular case is the following

COROLLARY 4.3. *Under the same assumption of Theorem 4.1, assume that the function* h *is linear and* g_i, $i = 1, \dots, M$ *are concave in* x_0. *Then there exist* $\lambda = (\lambda_1, \dots, \lambda_M)$ *and* $\mu = (\mu_1, \dots, \mu_P)$ *such that* (4.12) *holds*

PROOF. In a neighborhood of x_0 we can write

$$
\begin{aligned}
h_i(x) &= h_i(x_0) + Dh_i(x_0)(x - x_0) \\
g_i(x) &\le \lambda g_i(x_0) + Dg_i(x_0)(x - x_0) \qquad i \in I^*(x_0)
\end{aligned}
$$

where $I^*(x_0)$ as in Corollary 4.2. Hence

$$
\begin{aligned}
\sum_{i=1}^{P} \mu_i h_i(x) + \sum_{i=1}^{M} \lambda_i g_i(x) \le \sum_{i=1}^{P} \mu_i h_i(x_0) + \sum_{i \in I^*(x_0)} \lambda_i g_i(x_0) \\
+ \left(\sum_{i=1}^{P} \mu_i Dh_i(x_0) + \sum_{i \in I^*(x_0)} \lambda_i Dg_i(x_0) \right) (x - x_0)
\end{aligned}
$$

If in (4.2) we assume by contradiction that $\lambda_0 = 0$ and we recall that $\lambda_i = 0$ for $i \notin I^*(x_0)$, by the previous inequality we get

$$
\sum_{i=1}^{P} \mu_i h_i(x) + \sum_{i=1}^{M} \lambda_i g_i(x) \le 0.
$$

But since $\lambda \neq 0$ if $\lambda_0 = 0$, then there exists x in a neighborhood of x_0 such that

$$\sum_{i=1}^{P} \mu_i h_i(x) + \sum_{i=1}^{M} \lambda_i g_i(x) > 0$$

and therefore a contradiction to $\lambda_0 = 0$. □

REMARK. The conditions in (4.12) are known as the *Kurush-Kuhn-Tucker (KKT)* necessary conditions

REMARK. The Lagrangian $L : \mathbb{R}^N \times \mathbb{R}^M_+ \times \mathbb{R}^p$ associated to the optimization (4.1) is given by

$$L(x, \lambda, \nu) = f(x) + \lambda g(x) + \nu h(x), \tag{4.13}$$

with $\lambda, \nu \in \mathbb{R}^M_+ \times \mathbb{R}^P$. The KKT conditions can be reformulated as follows

$$\begin{cases} \frac{\partial L}{\partial x_i}(x_0, \lambda, \mu) = 0,\ i = 1, \dots, N \\ \\ \lambda_i g_i(x_0) = 0,\ i = 1, \dots, M, \\ \\ g(x_0) \leq 0,\ h(x_0) = 0,\ \lambda \geq 0 \end{cases} \tag{4.14}$$

EXAMPLE 13. The following example shows that the conditions (4.12) are necessary, but not sufficient for the existence of an extremal point. Consider (4.1) with

$$\begin{cases} f(x_1, x_2) = x_1 x_2 \\ g^1(x_1, x_2) = -x_1 - x_2 + 3 \geq 0 \\ g^2(x_1, x_2) = -x_2 + x_1 \leq 0. \end{cases} \tag{4.15}$$

The Karush-Kuhn-Tucker conditions for $x^0 = (x_1, x_2)$ are

$$\begin{cases} \lambda_1 \geq 0, \quad \lambda_2 \geq 0 \\ f_{x_1}(x^0) + \lambda_1 g^1_{x_1}(x^0) + \lambda_2 g^2_{x_1}(x^0) = 0, \\ f_{x_2}(x^0) + \lambda_1 g^1_{x_2}(x^0) + \lambda_2 g^2_{x_2}(x^0) = 0, \\ \lambda_1 g^1(x^0) + \lambda_2 g^2(x^0) = 0 \\ g^1(x^0) \leq 0, \qquad g^2(x^0) \leq 0 \end{cases} \tag{4.16}$$

Since

$$\begin{aligned} g^1_{x_1}(x_1, x_2) &= -1 \qquad g^1_{x_2}(x_1, x_2) = -1 \\ g^2_{x_1}(x_1, x_2) &= 1 \qquad g^2_{x_2}(x_1, x_2) = -1 \end{aligned}$$

and the conditions becomes

$$\begin{cases} \lambda_1 \geq 0, \quad \lambda_2 \geq 0 \\ x^0_2 - \lambda_1 + \lambda_2 = 0, \\ x^0_1 - \lambda_1 - \lambda_2 = 0, \\ \lambda_1(-x^0_1 - x^0_2 - 3) + \lambda_2(-x_2 + x_1) = 0 \\ g^1(x^0) \leq 0, \qquad g^2(x^0) \leq 0 \end{cases} \tag{4.17}$$

Since $\lambda_1,\ \lambda_2 \neq 0$ are not both null, the only solution is $x^0_1 = x^0_2 = 1$, $\lambda_1 = 1$ and $\lambda_2 \neq 0$, which is not a local minimizer since in the direction $(1, 1)$ the function increases, while in the direction $(-1, 1)$ decreases. Note that f is not convex.

2. Sufficient conditions in the convex case

In some particular case, the KKT condition becomes (4.14) becomes also sufficient.

THEOREM 4.4. *Besides the assumptions in Theorem 4.1, assume that the functions f and g_i, $i = 1 \dots, M$ are convex and that $h(x) = Ax - b$, i.e. the bilateral constraints are linear. Then x_0 is a solution of* (4.1) *if and only if it satisfies* (4.14).

Moreover, if f is strictly convex, x_0 is the unique solution of (4.1).

PROOF. Since $\lambda \geq 0$ for any $x \in \{x \in I : g(x) \leq 0,\, h(x) = 0\}$,

$$f(x) \geq f(x) + \lambda g(x) + \mu h(x).$$

Moreover by linearity of h, convexity of f and g_i and $\lambda_0 \geq 0$ we have

$$\begin{aligned} h(x) &= h(x_0) + Jh(x_0)(x - x_0) \\ f(x) &\geq f(x_0) + Df(x_0)(x - x_0) \\ \lambda g(x) &\geq \lambda g(x_0) + \lambda Jg(x_0)(x - x_0) \end{aligned}$$

Hence, by (4.11)

$$\begin{aligned} f(x) &\geq f(x) + \lambda g(x) + \mu h(x) \geq f(x_0) + Df(x_0)(x - x_0) \\ &+\lambda g(x_0) + \lambda Jg(x_0)(x - x_0) + \mu h(x_0) + \mu Jh(x_0)(x - x_0) \\ &\geq f(x_0) + \left(Df(x_0) + Jg(x_0)^T\lambda + Jh(x_0)^T\mu\right)(x - x_0) = f(x_0) \end{aligned}$$

for any admissible x, hence x_0 is a minimum point.

If f is strictly convex, a similar calculation gives $f(x) > f(x_0)$ for any admissible x, showing that x_0 is the unique global minimum. □

3. Examples of constrained problems

In this section we analyze some examples of constrained problems writing explicitly the corresponding optimality conditions.

3.1. Linear constraints. An important case is the one with linear constraints, i.e.

$$\min\{f(x) : x \in \mathbb{R}^N \text{s.t. } Ax(x) \geq b,\, i = 1, \dots, M\} \tag{4.18}$$

where A is a $M \times N$ matrix and $b \in \mathbb{R}^M$.

$$L(x, \lambda) = f(x) + \lambda(b - Ax) \tag{4.19}$$

with $\lambda \in \mathbb{R}^M_+$. The KKT conditions can be reformulated as follows

$$\begin{cases} Ax_0 \geq b \\ \\ D_x L(x_0, \lambda_0) = Df(x_0) - A^T\lambda_0 = 0 \\ \lambda(b - Ax_0) = 0, \quad \lambda \geq 0 \end{cases} \tag{4.20}$$

If f is convex, then by Theorem 4.4 the conditions in (4.20) are also sufficient

3.2. Non negative constraints. We consider a problem of the type

$$\min\{f(x) : x \in \mathbb{R}^N \text{s.t. } x \geq 0,\ i = 1, \dots, M\} \tag{4.21}$$

i.e. the constraints are linear with Aidentity matrix and $b = 0$. By (4.20) we get

$$\begin{aligned} Df(x_0) - \lambda &= 0 \\ x_0 \geq 0,\ \lambda \geq 0,\ \lambda x &= 0 \end{aligned}$$

hence $\lambda_i = \frac{\partial f}{\partial x_i}(x_0)$ and therefore

$$\begin{aligned} \frac{\partial f}{\partial x_i}(x_0) &\geq 0 \quad \text{if } x_{0,i} = 0 \\ \frac{\partial f}{\partial x_i}(x_0) &= 0 \quad \text{if } x_{0,i} > 0 \end{aligned}$$

3.3. Box constraints. We consider a problem of the type

$$\min\{f(x) : x \in \mathbb{R}^N \text{s.t. } a_i \leq x_i \leq b_i,\ i = 1, \dots, N\} \tag{4.22}$$

where a, $b \in \mathcal{R}^N$ with $a_i < b_i$ (note that the constraints are linear). We consider the Lagrangian

$$L(x, \lambda) = f(x) + \lambda(a - x) + \nu(x - b)$$

By (4.12), we get

$$\begin{aligned} Df(x_0) - \lambda + \nu &= 0 \\ (a - x_0)\lambda = 0,\ (x_0 - b)\nu &= 0,\ (\lambda, \nu) \geq 0 \end{aligned}$$

Set

$$J_a = \{j : x_{0,j} = a_j\},\ J_b = \{j : x_{0,j} = b_j\},\ J_a = \{j : a_j < x_{0,j} < b_j\}$$

If $j \in J_a$, then $x_j < b_j$, hence $\nu_j = 0$. It follows that

$$\frac{\partial f}{\partial x_j}(x_0) = \lambda_j \geq 0.$$

Similarly, if $j \in J_b$, $\lambda_j = 0$ and

$$\frac{\partial f}{\partial x_j}(x_0) = -\nu_j \leq 0.$$

If $j \in J_0$, then $\lambda_j = \nu_j = 0$ and therefore

$$\frac{\partial f}{\partial x_j}(x_0) = 0$$

The necessary conditions for optimality are

$$\begin{aligned} \frac{\partial f}{\partial x_j}(x_0) &\geq 0 \quad \text{if } x_{0,j} = a_j \\ \frac{\partial f}{\partial x_j}(x_0) &\leq 0 \quad \text{if } x_{0,j} = b_j \\ \frac{\partial f}{\partial x_j}(x_0) &= 0 \quad \text{if } a_j < x_{0,j} < b_j. \end{aligned}$$

These condition are also sufficient if f is convex.

4. Slater condition and saddle points of the Lagrangian

We consider an optimization problem with only unilateral constraints

$$\min\{f(x) : x \in \mathbb{R}^N \text{s.t. } g_i(x) \leq 0,\ i = 1, \dots, M\} \tag{4.23}$$

DEFINITION. Assume that the function $g : \Omega \subset \mathbb{R}^N \to \mathbb{R}^M$ are C^1 and convex in the open convex set Ω containing the constraints set

$$U = \{x \in \mathbb{R}^N : g(x) \leq 0\}.$$

The set U is told to satisfy the *Slater condition* if there exists $x \in U$ such that $g(x) < 0$.

Consider the Lagrangian associated to the optimization problem (4.23), i.e.

$$L(x, \lambda) = f(x) + \lambda g(x) \tag{4.24}$$

The next theorem establish a link between solution of problems (4.23) and saddle points of the Lagrangian

THEOREM 4.5. *Assume that $g : \mathbb{R}^N \to \mathbb{R}^M$ satisfies the Slater condition in the set U and $f : \mathbb{R}^N \to \mathbb{R}$ is convex. Then x_0 is a solution of the optimization problem* (4.23) ***if and only*** *it is a saddle point for the Lagrangian, i.e.*

$$L(x_0, \lambda_0) = \min_{x \in \mathbb{R}^N} \max_{\lambda \in \mathbb{R}^M_+} L(x, \lambda) = \max_{\lambda \in \mathbb{R}^M_+} \min_{x \in \mathbb{R}^N} L(x, \lambda) \tag{4.25}$$

PROOF. Assume that x_0 solves the minimum problem and consider the sets

$$K(x) = \{(t_0, t) \in \mathbb{R} \times \mathbb{R}^M,\ t_0 \geq f(x),\ t \geq g(x)\}$$
$$K = \cup_{x \in \mathbb{R}^N} K(x)$$
$$S = \{(s_0, s) \in \mathbb{R} \times \mathbb{R}^M,\ s_0 \leq f(x_0),\ s \leq 0\}.$$

Claim 1 K is convex.

PROOF. Take (t_0, t) and $(t_0', t') \in K$. Then for some some x, x' we have

$$t_0 \geq f(x),\ t \geq g(x), \qquad t_0' \geq f(x'),\ t' \geq g(x').$$

Take for $0 \leq \lambda \leq 1$, then $\lambda t_0 + (1-\lambda)t_0' \geq \lambda f(x) + (1-\lambda) f(x')$. By the convexity of f

$$\lambda t_0 + (1-\lambda)t_0' \geq f(\lambda x + (1-\lambda)x')$$

and similarly

$$\lambda t_0 + (1-\lambda)t_0' \geq g(\lambda x + (1-\lambda)x')$$

This shows that $\lambda t_0 + (1-\lambda)t_0' \in K(\lambda x + (1-\lambda)x') \subset K$. □

Claim 2 S is convex.

PROOF. Take (s_0, s) and $(s_0', s') \in S$ and for $0 \leq \lambda \leq 1$ $\lambda(s_0, s) + (1-\lambda)(s_0', s')$. Then

$$\lambda s_0 + (1-\lambda)s_0' \geq f(\lambda x_0 + (1-\lambda)x_0) = f(x_0),$$

and

$$\lambda s + (1-\lambda)s' \leq 0.$$

□

Claim 3 $\mathrm{Int(S)} \cap \mathrm{K} = \emptyset$

PROOF. Recall that $\mathrm{Int(S)}\{(\mathrm{s_0, s}) \in \mathbb{R} \times \mathbb{R}^{\mathrm{M}}, \mathrm{s_0} < \mathrm{f(x_0)}, \mathrm{s} < 0\}$. If the claim 3 is not true, then there exists $(s_0, s) \in \mathrm{Int(S)} \cap \mathrm{K}$. Hence there exists $\hat{x} \in \mathbb{R}^N$ such that

$$f(\hat{x}) \leq s_0 < f(x_0), \qquad g(\hat{x}) \leq s < 0.$$

contradicting the assumption that x_0 is a minimum point. □

Then we apply the Separation's Theorem, see (3.1), to get the existence of a non null vector $(p_0, p) \in \mathbb{R} \times \mathbb{R}^M$ such that

$$p_0 s_0 + ps \geq p_0 t_0 + pt, \text{ for any } (s_0, s) \in S,\ (t_0, t) \in K \tag{4.26}$$

We have

Claim 4 $p_0 \leq 0$, $p \leq 0$.

PROOF. Indeed, arguing by contradiction, if $p_0 > 0$, taking $s = 0$, $x = 0$, $t_0 = f(0)$, $t = g(0)$ in (4.26) we get

$$p_0 s_0 \geq g(0)p + f(0)p_0 \qquad \forall s_0 \leq f(x_0)$$

As s_0 goes to $-\infty$ the first hand side goes to $-\infty$ while the second is greater than a constant. Hence a contradiction.

Assume that $p_k > 0$ for some $k \in \{1, \ldots, M\}$. By taking $(s_0, s) = (f(x_0), -te_k)$ and $(t_0, t) = (f(0), g(0))$ we have

$$-tp_k \geq -f(x_0)p_0 + p_0 f(0) + pg(0).$$

As t goes to $+\infty$ we get a contradiction. □

Claim 5 $p_0 < 0$.

PROOF. Indeed for $(t_0, t) = (f(x), g(x))$, x any real number and $(s_0, s) = (f(x_0), 0)$ we get

$$p_0 f(x_0) \geq p_0 f(x) + pg(x). \tag{4.27}$$

if $p_0 = 0$ then $0 \geq pg(x)$. Since $p \leq 0$ this means $g(x) \geq 0\ \forall x \in \{x \in I : g(x) \leq 0\}$ and therefore a contradiction to the Slater condition. □

Define $\lambda_0 = \frac{p}{p_0}$. We now verify that (x_0, λ_0) is a saddle point for the Lagrangian.

Claim 6 $\lambda_0 g(x_0) = 0$

PROOF. By (4.27)

$$p_0 f(x_0) \geq p_0 f(x) + pg(x) \quad \forall x \in \mathbb{R}^N.$$

Since p_0 is negative, we have

$$f(x_0) \leq f(x) + \lambda_0 g(x) \quad \forall x \in \mathbb{R}^N.$$

Taking $x = x_0$, we get $0 \leq \lambda_0 g(x)$ hence

$$0 \leq \lambda_0 g(x) \tag{4.28}$$

Since the other inequality is true by $x_0 \in \{x \in I : g(x) \leq 0\}$ and therefore $g(x_0) \leq 0$ and $\lambda_0 \geq 0$, we get

$$\lambda_0 g(x_0) = 0$$

□

Next, we observe that $\lambda g(x_0) \leq 0$ per ogni $\lambda \in \mathbb{R}^M_+$. Then

$$f(x_0) + \lambda g(x_0) \leq f(x_0) + \lambda_0 g(x_0) \leq f(x) + \lambda_0 g(x)$$

and therefore (x_0, λ_0) is a saddle point for the Lagrangian.
To prove the reverse implication, assume that (x_0, λ_0) is a saddle point for the Lagrangian. Then

$$f(x_0) + \lambda g(x_0) \leq f(x_0) + \lambda_0 g(x_0) \leq f(x) + \lambda_0 g(x), \qquad \forall x \in \mathbb{R}^N, \lambda \in \mathbb{R}^M_+$$

This implies

$$\lambda g(x_0) - \lambda_0 g(x_0) = g(x_0)(\lambda - \lambda_0) \leq 0, \qquad \forall \lambda \in \mathbb{R}^M$$

By taking $\lambda = \lambda_0 + e_i$, $i = 1, \ldots, M$ in the previous inequality we get $g(x_0) \leq 0$, hence $x_0 \in U$.
Moreover by taking $\lambda = 0$ since $\lambda_0 \geq 0$, we get $\lambda_0 g(x_0) \geq 0$ then $\lambda_0 g(x_0) = 0$. Since $\lambda_0 g(x) \leq 0$ for all $x \in U$, we have

$$f(x_0) = f(x_0) + \lambda_0 g(x_0) \leq f(x) + \lambda_0 g(x) \leq f(x), \qquad \forall x \in U, \lambda \in \mathbb{R}^M_+$$

and therefore t x_0 is solution of the minimum problem (4.23). □

Bibliography

[1] C.D. Pagani, S. Salsa Analisi Matematica VOL II, Masson.
[2] N. Fusco, P. Marcellini C. Sbordone, Analisi Matematica due, LIGUORI
[3] E.K.P. Chong, S. H. Zak, An Introduction to optimization
[4] http://gol.dsi.unifi.it/users/sciandrone/$cap7\_8PM06$.pdf
[5] http://www.mat.uniroma1.it/people/capuzzo/didattica/dispense/ottimizzazione05.pdf
[6] Hardy, Littlewood, Polya Inequalities.