Notes on Advanced Optimization

Regina S. Burachik
# Contents

## 1 Introduction
1.1 Aims of this course .................................................. 1
1.2 Mathematical Formulation of an Optimization Problem and Examples ............. 3
   1.2.1 Curve Fitting ............................................... 3
   1.2.2 A transportation problem ................................... 4
   1.2.3 Location of facilities ....................................... 5
   1.2.4 Hanging Chain ............................................. 5
1.3 Types of Optimization Problems ..................................... 7
   1.3.1 Continuous and Discrete Optimization Problems ....................... 7
   1.3.2 Differentiable and Convex Optimization Problems .................... 8
   1.3.3 Global and Local Optimization ................................ 8
   1.3.4 Nonconvex Problems ....................................... 8

## 2 Theory
2.1 Elements of Convex Analysis ....................................... 12
   2.1.1 Convex sets .............................................. 12
   2.1.2 Convex Hull ............................................ 14
   2.1.3 Convex Functions ....................................... 15
   2.1.4 Operations preserving convexity of functions ....................... 18
2.2 Dierentiability Properties of Convex Functions .......................... 18
   2.2.1 Gradients and Hessians ................................... 19
   2.2.2 One-sided directional derivatives ........................... 22
2.3 Epigraphs and Level Sets ......................................... 25
   2.3.1 Support of sets at boundary points .......................... 28
2.4 Topological Properties of Convex Sets ................................ 29
2.5 Separation results ............................................... 31
2.6 Subgradients and Subdierentials .................................... 35
2.7 Existence of solutions of Optimization Problems .......................... 39
2.8 Optimality Conditions ........................................... 41
   2.8.1 Necessary Optimality Conditions ................................ 42
   2.8.2 Sufficient Optimality Conditions .......................... 44
   2.8.3 Saddle Points ........................................... 45
2.9 Optimality conditions for Constrained Problems .......................... 46
   2.9.1 Equality Constraints ...................................... 46
   2.9.2 Equality and Inequality Constraints ........................ 52
   2.9.3 Sensitivity Analysis for Constrained problems ................... 56
2.10 Optimality Conditions for Convex Nondierentiable Problems ............... 60
2.11 Maximization of a Convex Function ................................ 72
3 Methods

3.1 Methods for Unconstrained Problems
3.1.1 Newton’s Method
3.1.2 Implementation of Newton’s Method
3.1.3 Quadratic Convergence of Newton Method
3.1.4 Steepest Descent Method
3.1.5 Steepest Descent with Armijo Rule
3.1.6 Convergence of Steepest Descent Methods
3.1.7 Quasi-Newton Methods

3.2 Methods for Constrained Problems
3.2.1 Penalty Methods
3.2.2 The Exact Penalty Method
3.2.3 Barrier Methods

A Topology of $\mathbb{R}^n$

A.1 Notation
A.2 Norms and Scalar Product
A.3 Scalar Product
A.4 Balls in $\mathbb{R}^n$
A.5 Boundedness
A.6 Open and closed sets
A.7 Convergence of sequences
A.8 Calculus in $\mathbb{R}^n$
A.8.1 Continuous Functions
A.8.2 Differentiable Functions
A.9 Useful facts from Linear Algebra
A.9.1 Matrices and Linear transformations
A.9.2 Orthogonal subspaces
A.9.3 Square Matrices and Eigenvalues
To my teachers, who determined what I am.
To my students, who determine what I will be.

Acknowledgments

I am indebted to the School of Mathematics and Statistics of the University of South Australia, which provided me with the environment that made the writing of these notes possible. I am also grateful to my students Neil Tatcher, David Sambell, and Stephen J. Maher, who carefully read these notes, and made important suggestions for improvements.
Chapter 1

Introduction

1.1 Aims of this course

The broad discipline of Optimization includes the following general areas:

- **Modelling:** Formulation of a given problem as an optimization problem;
- **Theory:** Study of (i) existence and uniqueness of solutions, and (ii) characterization of solutions of optimization problems;
- **Methods:** Development and convergence analysis of algorithms for solving optimization problems;
- **Implementation:** Transcription of the algorithm to the medium of a suitable technical device (e.g., a digital computer). For instance, on a digital computer, implementation of a method consists of writing a code corresponding to the method and running it on the computer.

Our main focus will be on the second and third areas above; that is, with the Theory and Methods available for important kinds of optimization problems. In fact, the other two areas strongly rely on Theory and Methods. Namely, if we know which problems we can tackle and how, then we know the kind of mathematical models we can pose and efficiently solve. If our theoretical knowledge becomes stronger, then we can improve our model accordingly. Finally, we note that the implementation is merely a transcription of the theory and methods into some computer language or software.

At the end of this course, you should be able to:

(a) Identify the main features of a given optimization problem (i.e., whether the problem is constrained or not, differentiable or not, convex or not, etc.).

(b) Decide whether or not the problem has solutions, and whether or not these solutions can be found analytically.

(c) Become aquainted with numerical algorithms able to approximate the solution

(d) Analyze the convergence of certain optimization techniques.

The first two items are purely theoretical, and focus on necessary and sufficient conditions for optimality. These conditions are very important for the following reasons. First, they may allow us, in some special cases, to obtain the solution analytically. The latter situation not only provides us with a solution itself, it also allows us to describe a solution as a function of the data. This information is much more valuable than just having a solution. Second, most of the numerical
methods for solving optimization problems are based on optimality conditions. At each step of these algorithms, we check the optimality conditions at the current iterate. If they hold, we stop (we have a “solution”). Otherwise, we improve the current iterate. Thus, optimality conditions give us the background and the motivation for the methods, which are the second part of the course.

Items (b) and (c) focus on the analysis of algorithms for solving different kinds of optimization problems. This analysis is essential for choosing the correct solution technique for a given problem. Each technique has its advantages and drawbacks, and we must be aware of them a priori. Furthermore, it is important for us to understand deeply the convergence properties of the techniques available; because a deep understanding of the state-of-the-art techniques is essential for us to obtain further (and needed) improvements of these techniques.

We will also study other things which are not directly related with optimality conditions and optimization methods. This includes some basic mathematical knowledge, as well as some elements of convex analysis. This material is important in its own, because it is the theoretical background necessary for understanding and developing the new optimization tools that certainly will appear in the decades to come.

For example, 15-20 years ago, the Simplex method was the main tool for solving Linear Programming problems. We know today that some interior point methods are, for some kinds of problems, much more efficient than the Simplex method. We should therefore be well prepared with a strong mathematical background in order to cope with the further (and usually unexpected) changes to come.

In Section 1.2 we define a general optimization problem. We also give some examples and model them as specific optimization problems. In Section 1.3 we describe the different types of problems we are concerned with. Section 2.1 is a brief introduction to Convex sets and functions. The classical subject of existence of solutions of constrained and unconstrained problems is treated in Section 2.7. Optimality Conditions for unconstrained and constrained problems is studied in Section 2.8. In Section 3.1 we analyze two important families of techniques for unconstrained optimization: Newton’s method and Steepest descent method. The constrained case is studied in Section 3.2, where we focus on Penalty and Barrier Methods. Finally, Appendices A and B contain basic and necessary background on Analysis and Linear Algebra, respectively.
1.2 Mathematical Formulation of an Optimization Problem and Examples

One of the most common problems we may face is the one of making the best decision according to some criteria, amongst a given set of possible decisions. For example, factories minimize cost, investors minimize risks or maximize return. Not only human decisions depend on optimizing some given criteria. Nature itself also optimizes. Indeed, physical systems tend to minimize energy. Therefore, to master the mathematical tools involved in the study of optimization is essential both for decision science and for the better understanding of physical systems.

The main ingredients of any optimization problem are:

1. The variables of our problem. These are, for instance, all the possible decisions we can make.
2. The objective function of our problem. This function measures how “good” a given decision is. This objective can be the cost, the return, or the energy of the system.

The goal is to find those variables which give us the optimal (or best, in certain sense) value of the objective function. This best possible value is called the optimal value of the optimization problem.

When the possible decisions we can take span the whole space of variables, we say that our problem is unconstrained. Otherwise, we have a constrained problem. The restrictions on the variables are called constraints, and are usually expressed through equalities and/or inequalities involving the variables. Typically, a constraint set \( C \subseteq \mathbb{R}^n \) given by equalities and inequalities is the set of all points \( x \in \mathbb{R}^n \) such that

\[
g_1(x) \leq 0, \quad \cdots, \quad g_m(x) \leq 0, \\
\vdots \quad \cdots \quad \vdots \\
h_1(x) = 0, \quad \cdots, \quad h_r(x) = 0.
\]

(1.2.1)

All points verifying the constraints of a given problem are called feasible points.

Identifying the variables, the objective function, and the constraints of a given real-life problem is known as modelling. The construction of a model is, as we saw before, the first step (and a very important one) in the optimization process. We note that modelling requires a great deal of creativity and it can be much more difficult than solving the resulting mathematical model.

Let us consider some simple real-life examples, and model them.

1.2.1 Curve Fitting

This example has been taken from [5, page 2]. Suppose you are given \( M \) observed points \((t_1, s_1), \ldots, (t_M, s_M)\) in the plane, and you look for the straight line in the plane that best fits the given data. Let us express an arbitrary straight line as \( y(t) := mt + b \), so our decision variables are pairs \((m, b) \in \mathbb{R}^2\). Since there is no restriction on the kind of straight line we want, our problem is unconstrained. Our requirement on the line is for it to be as close as possible to the given points, so we must minimize the deviation of the line from these points. Let us measure the “deviation” of the line from each point as

\[d_i := (mt_i + b - s_i)^2, \quad \text{where } i = 1, \ldots, M.\]

The total deviation is then \(\sum_{i=1}^{M} (mt_i + b - s_i)^2\). This is the objective we want to minimize. Therefore, our optimization problem becomes

Minimize \(\sum_{i=1}^{M} (mt_i + b - s_i)^2\),

subject to \((m, b) \in \mathbb{R}^2\).
A solution of the above problem is known as a “Least square error solution”. Alternatively, we can choose to measure the deviation as

\[ d_i := |mt_i + b - s_i|, \quad \text{where } i = 1, \ldots, M, \]

then the problem becomes

\[
\text{Minimize } \sum_{i=1}^{M} |mt_i + b - s_i|,
\]

subject to \((m, b) \in \mathbb{R}^2\).

A solution of this problem is known as the “Least total error solution”. The solutions given by each of these optimization problems are different, and in some specific cases we may prefer one of the solutions over the other. Note also that these two problems have very different properties. The first one is the minimization of a differentiable (moreover, quadratic) function, while the second problem is not differentiable. Therefore, they require different mathematical tools for their solution. Hopefully, we will be able to solve both problems along this course.

1.2.2 A transportation problem

This example has been taken from [8, page 4]. A chemical company has two factories \( F_1 \) and \( F_2 \) and \( R_1, \ldots, R_m \) retail outlets. Each factory produces a different chemical substance. Factory \( F_1 \) can produce \( c_1 \) tons of the first chemical product, and factory \( F_2 \) can produce \( c_2 \) tons of the second chemical product. We say that \( c_i \) is the capacity of factory \( F_i \). Each retail outlet \( R_j \) has a known weekly demand of \( d_j \) tons. The demand \( d_j \) is the sum of the demands of both products for the retail \( R_j \).

The price for (or the cost of) shipping one ton of the product from \( F_1 \) to \( R_j \) is \( p_{1j} \). Similarly, the price of shipping one ton of the second product to \( R_j \) is \( p_{2j} \).

We want to find out how much product we must ship weekly from both factories to each retail outlet, so as to satisfy the demand and minimize the shipping cost.

The variables of this problem are the tons of product 1 shipped to each outlet \( R_j \), and the tons of product 2 shipped to each outlet \( R_j \). Let us denote by \( x_{1j} \) the tons of product 1 shipped to \( R_j \), and \( x_{2j} \) the tons of product 2 shipped to \( R_j \). Therefore, our variables are \((x_{1j}, x_{2j}) \in \mathbb{R}^m \times \mathbb{R}^m\) (recall that \( m \) is the number of outlets). A natural requirement on these variables is that \( x_{1j} \geq 0 \) and \( x_{2j} \geq 0 \) for all \( j = 1, \ldots, m \). The total cost of shipping a given decision vector \((x_{1j}, x_{2j})\) is \( \sum_{j=1}^{m}(p_{1j}x_{1j} + p_{2j}x_{2j}) \). This is our objective function.

Because we want to satisfy the weekly demand of each retail \( R_j \), we must impose the inequality

\[ x_{1j} + x_{2j} \geq d_j, \quad \text{for all } j = 1, \ldots, m. \]

On the other hand, we cannot exceed the capacity of each factory, which is expressed as

\[ \sum_{j=i}^{m} x_{1j} \leq c_1, \quad \sum_{j=i}^{m} x_{2j} \leq c_2. \]

Altogether, we obtain the Linear Programming problem:

\[
\text{Minimize } \sum_{i,j} p_{ij} x_{ij},
\]

subject to \( x_{1j} + x_{2j} \geq d_j, \quad \text{for } j = 1, \ldots, m, \)

\[ \sum_{j=i}^{m} x_{ij} \leq c_i, \quad \text{for } i = 1, 2, \]

\[ x_{ij} \geq 0, \quad \text{for } i = 1, 2 \text{ and } j = 1, \ldots, m. \]
1.2.3 Location of facilities

This example was taken from [1, page 2]. We want to find the best possible location of certain facility, so that it can easily be reached from some fixed locations. For example, we want to decide where to build a school (or a fire station, or a hospital) such that some specific regions (having fixed locations in the city) are able to use that facility. The variables of our problem are therefore the coordinates \((x, y) \in \mathbb{R}^2\). Because we move in the city along perpendicular lines (i.e., streets), a usual way for modelling the distance between two given locations \((a, b)\) and \((x, y)\) in the city is the \textit{rectangular} distance (or \(L_1\)-distance), defined as

\[
d_1((a, b), (x, y)) := |a - x| + |b - y|.
\]

Our criteria can be to minimize the maximum of the rectangular distances between the facility and the given points \((a_1, b_1), \ldots, (a_m, b_m)\). This gives the objective function

\[
f(x, y) := \max_{j=1,\ldots,m} |a_j - x| + |b_j - y|.
\]

Therefore our minimization problem becomes

\[
\begin{align*}
\text{Minimize} & \quad \max_{j=1,\ldots,m} |a_j - x| + |b_j - y|, \\
\text{subject to} & \quad (x, y) \in \mathbb{R}^2.
\end{align*}
\]

This kind of problem is called a minimax problem. It can be rewritten in a nicer way if we introduce more variables. Let us define the additional variable \(z := \max_{j=1,\ldots,m} |a_j - x| + |b_j - y|\). The constraints upon \(z\) are

\[
z \geq |a_j - x| + |b_j - y|, \quad \text{for all } j = 1, \ldots, m.
\]

Altogether, we can re-write our problem as

\[
\begin{align*}
\text{Minimize} & \quad z, \\
\text{subject to} & \quad z \geq |a_j - x| + |b_j - y|, \quad \text{for all } j = 1, \ldots, m.
\end{align*}
\]

Now, the variables are \((x, y, z)\) and the objective function is simply \(g(x, y, x) = z\). As we see, this problem has inequality constraints, which are nondifferentiable. By adding more variables, it can be reformulated as a linear programming problem.

1.2.4 Hanging Chain

This example has been taken from [7, Example 4, Section 10.4] A chain is suspended from two hooks that are at height 0. The distance between the hooks is denoted as \(D\). The chain itself consists of \(M\) identical links. We assume that both length and weight of each link is 1 unit. We want to find the shape of the chain at equilibrium. In order to find this shape we minimize its potential energy, which is a function of the position of the links. In order to determine our variables, objective function and constraints, we use the following facts.

1. The potential energy \(P\) of the chain consists of the sum of the potential energy of each link \(P_i\) of the chain, where \(i = 1, \ldots, M\).

2. The potential energy of a body is its weight times its vertical height (from some reference point).
We take the top of the chain as zero height, and the positive height of each link is measured downwards, up to the center of that link. The displacement of the link \( i \) in the horizontal direction is denoted as \( x_i \), and its vertical displacement is denoted as \( y_i \). Because the length of the link is 1, we have \( 1 = x_i^2 + y_i^2 \). See Figure 1.2.1. The symmetry of each link implies that its center of mass is located at the center of the link. Therefore, the contribution of potential energy \( P_1 \) from the first link on the left (we call it link 1) is given by its weight times its height (the latter is the distance between its center of mass and the top of the chain)

\[
P_1 = (1/2)y_1.
\]

Indeed, the weight of the link is 1 and its height is \((1/2)y_1\) because its center of mass is in the center of the link. Similarly, the contribution of potential energy \( P_2 \) from the second link is

\[
P_2 = y_1 + (1/2)y_2.
\]

This allows us to write down the contribution of potential energy \( P_i \) of link \( i \) as

\[
P_i = y_1 + y_2 + \cdots + y_{i-1} + (1/2)y_i.
\]

Our variables are the vertical displacements \( y_1, \ldots, y_M \) of each of the links (which represent all possible shapes of the chain). Our objective function is the total potential energy of the chain, which is the sum of all the terms \( P_i \); that is,

\[
P(y_1, \ldots, y_M) := \sum_{i=1}^{M} (y_1 + y_2 + \cdots + y_{i-1} + (1/2)y_i) = \sum_{i=1}^{M} (M - i + \frac{1}{2})y_i.
\]

The total vertical displacement of the chain is zero, so we have the constraint

\[
\sum_{i=1}^{M} y_i = 0,
\]

and the total horizontal displacement of the chain is \( D \), so we have the constraint

\[
\sum_{i=1}^{M} \sqrt{1 - y_i^2} = D.
\]

Altogether, we obtain the optimization problem

Minimize \( \sum_{i=1}^{M} (M - i + \frac{1}{2})y_i \),

subject to \( \sum_{i=1}^{M} y_i = 0 \), \( \sum_{i=1}^{M} \sqrt{1 - y_i^2} = D \).

This problem is differentiable with equality constraints, and soon we will be able to solve it.
1.3 Types of Optimization Problems

The previous examples suggest that there are many different types of optimization problems. We have seen already that optimization problems can be constrained or unconstrained, according to whether or not there is some restriction on the variables. The type of problem we describe next depends on the nature of the variables.

1.3.1 Continuous and Discrete Optimization Problems

The variables in the model problem of section 1.2.2 are continuous, in the sense that any pair of real coordinates \((x_{ij}, x_{kj})\) constitutes an admissible choice. More precisely, if for every feasible point \(x\) we can find feasible points arbitrarily close to \(x\), such a variable is a continuous variable of the problem. When all the variables of the problem verify the latter property, we have a Continuous Optimization Problem. A typical continuous optimization problem is one in which the variables are vectors with real-valued components. In particular, every unconstrained problem is continuous. If a variable of the problem can only take values in some discrete set (i.e., a set of isolated points), then we say that such a variable is a discrete variable. Suppose that, in the model problem of section 1.2.2, the factory stores its weekly production in barrels of a fixed volume (e.g., if the product is not solid but liquid). These barrels must be full when shipped, so that the transportation is done efficiently. This means that the tons of product \(x_{ij}\) shipped from factory \(F_i\) to retail outlet \(R_j\) is an integer \(n_{ij}\) (number of barrels) times a fixed volume \(v\) (the volume of each barrel). The volume \(v\) is constant, thus we have \(x_{ij} = n_{ij}v\). Therefore, our new decision variables are the integer numbers \(n_{ij}\). In this case, the model problem of section 1.2.2 is a discrete problem. The model problem of section 1.2.3 is a continuous problem, because the variables are two dimensional vectors of real coordinates. However, if in this model we would have the additional requirement that our facility must be located at a corner of the block (e.g., traffic lights), then the problem becomes discrete, because the admissible choices are isolated points determined by each corner of the city. If some of the variables of a given problem are continuous and others are discrete, we have a Mixed Integer Programming Problem. For instance, if the product of factory \(F_1\) in the model problem
of section 1.2.2 is a liquid substance which is shipped in full barrels and the factory $F_2$ produces
a solid substance which can be shipped in any format and quantity, then the problem becomes a
mixed-integer programming problem.

Continuous problems are usually easier to solve, because they can be studied using the powerful
tools of calculus. More precisely, from the values of the objective function and its derivative at
a fixed variable $x$ (which can be a candidate for a solution), we can deduce the behaviour of the
objective function at all points in a neighborhood around $x$. The behaviour of the objective around
$x$ is relevant only for continuous problems because only in this case we can approach $x$ through
feasible points. On the other hand, discrete problems are intrinsically more complex and have a
combinatorial flavour. In these problems, the value of the objective function at a given feasible
point $x$ cannot give us any information on the behaviour of the objective function at points “next”
to $x$. However, the powerful tools developed for solving continuous optimization problems are also
used for solving discrete ones. Indeed, some algorithms for solving discrete problems consist of
solving a sequence of continuous optimization problems. These continuous optimization problems
are called continuous relaxations of the original problem.

1.3.2 Differentiable and Convex Optimization Problems

We concentrate our study on continuous optimization problems. Amongst these, we will study
Differentiable Problems and Convex Problems. Differentiable Problems are those in which the
constraint set has the form given in (1.2.1) and the functions $g_1,\ldots,g_m,h_1,\ldots,h_r$ are differ-
entiable. Problems with a constraint set as in (1.2.1) are called Nonlinear Programming
Problems. In the particular case in which the objective function $f$, as well as the constraint functions
$g_1,\ldots,g_m,h_1,\ldots,h_r$ are all linear, we have a Linear Programming Problem.

The theory and the methods available for the solution of differentiable problems are both rich
and powerful, as we see in this course. Naturally, the main tools for analyzing these problems come
from differentiable calculus.

Convex problems also have rich and powerful (as well as beautiful) theory and methods. In
these problems both the objective function and the constraint set are convex. A typical constraint
set of a convex problem has the form given in (1.2.1), where the functions $g_1,\ldots,g_m$ are convex
and the functions $h_1,\ldots,h_r$ are affine. In particular, every Linear Programming problem is a
convex problem. Using Convex Analysis, we will be able to characterize all solutions even when
the problem is nondifferentiable. Moreover, convex problems can be efficiently (both in theoretical
and in the practical meaning of the word) solved numerically. The latter is not the case for general
nonconvex problems.

1.3.3 Global and Local Optimization

The most efficient optimization methods find a local minimizer; that is, a point $x$ such that the
value of the objective function at $x$ is smaller than the value at every point in some neighborhood
of $x$. A global minimizer is a point $x$ such that the value of the objective function at $x$ is smaller
than the value at every other point of the space. Clearly, global solutions are much more difficult
to find than local ones. However, global solutions are necessary (or at least desirable) in many
real-life applications. It is clear that every global solution is local, but the converse is in general
not true. However, if the problem is convex, then it is easy to show that every local solution is a
global one. We point out that local optimization methods can be used for finding global solutions.
Indeed, some global methods consist of solving a sequence of local optimization problems.

1.3.4 Nonconvex Problems

Solving nonconvex and nondifferentiable problems can be extremely difficult, and sometimes im-
possible. However, the theory of convex optimization provides very helpful tools in these difficult
cases.

For example, if we have a nonconvex problem and we want to solve it by some optimization technique, we need a starting point. This point can be obtained by approximating our original, nonconvex problem by a convex one. We can solve this approximated problem, and its exact solution provides the desired starting point.

Methods for finding solutions of nonconvex and nondifferentiable problems are called *global optimization methods*. These methods usually require an estimate of the *optimal value* (that is, the value of the objective function at the solution) of the problem. This optimal value is, in most cases, unknown. Standard methods for estimating this optimal value are based on convex optimization.
Chapter 2

Theory
2.1 Elements of Convex Analysis

2.1.1 Convex sets

Convex sets and convex functions play a central role in Optimization Theory. We start with convex sets. Given two points \( x, y \in \mathbb{R}^n \), the line segment joining \( x \) and \( y \) is the set of all averages of \( x \) and \( y \), i.e., the set given by
\[
\{ \lambda x + (1 - \lambda)y \mid 0 \leq \lambda \leq 1 \}.
\]
This set is denoted by \([x, y]\). Line segments are the basic tools for defining convex sets.

**Definition 2.1.1** Given a set \( C \subset \mathbb{R}^n \), we say that \( C \) is convex if and only if, for every pair of points \( u, v \in C \), and every \( \lambda \in [0, 1] \), it holds that \( \lambda u + (1 - \lambda)v \in C \). In other words, the whole line segment joining \( u \) and \( v \) is contained in \( C \). The point \( \lambda u + (1 - \lambda)v \) is called a convex combination of \( u \) and \( v \).

The definition above asserts that a set is convex when all convex combinations of two points of the set remain in the set. This result can be easily extended to any finite number of points.

**Exercise 2.1.1** Prove that \( C \subset \mathbb{R}^n \) is convex if and only if for every \( m \in \mathbb{N} \) and every set of \( m \) points \( \{x_1, \ldots, x_m\} \subset C \) we have that
\[
\sum_{i=1}^{m} \lambda_i x_i \in C,
\]
where \( \sum_{i=1}^{m} \lambda_i = 1 \) and \( \lambda_i \geq 0 \) for all \( i = 1, \ldots, m \).

**Remark 2.1.1** As we mentioned in Definition 2.1.1, an expression of the form \( \lambda x + (1 - \lambda)y \) with \( \lambda \in [0, 1] \) is called a convex combination of \( x \) and \( y \). In the same way, an expression of the form \( \sum_{i=1}^{m} \lambda_i x_i \) with \( \sum_{i=1}^{m} \lambda_i = 1 \) and \( \lambda_i \geq 0 \) for all \( i = 1, \ldots, m \) is called a convex combination of the \( m \) points \( \{x_1, \ldots, x_m\} \). Note that, when we keep the requirement \( \sum_{i=1}^{m} \lambda_i = 1 \), and we drop the requirement \( \lambda_i \geq 0 \) for all \( i = 1, \ldots, m \), the expression \( \sum_{i=1}^{m} \lambda_i x_i \) is called an affine combination of the \( m \) points \( \{x_1, \ldots, x_m\} \) (see Definition A.9.31 in the Appendix). Moreover, if no requirement is made on the coefficients \( \lambda_i \), we get a linear combination of the \( m \) points \( \{x_1, \ldots, x_m\} \). So, we see that a convex set is “closed” for convex combinations of its points, while affine (linear) sets are “closed” for affine (linear) combinations of its points. These definitions imply readily that subspaces and affine manifolds must be convex.

![Diagram of convex and nonconvex sets](image)

**Figure 2.1.1:** Examples of convex and nonconvex sets.

In Figure 2.1.1, the set (a) includes the interior of the surface. Hence the sets (a) and (c) are convex, while (b) is not.
Example 2.1.1 We list here some important examples of convex sets. All items from (a)-(f) are affine sets (See Definition A.9.31).

(a) **Subspaces.** Since subspaces are closed for linear combinations of its elements, then they must also be closed for convex combinations of them.

(b) Let $A \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. For a given $x \in \mathbb{R}^n$, we say that $x \geq 0$ when every coordinate of $x$ is nonnegative, i.e., when $x_i \geq 0$ for all $i = 1, \ldots, n$. The sets

$$\{x \in \mathbb{R}^n \mid Ax = b\}, \{x \in \mathbb{R}^n \mid Ax = b, x \geq 0\}, \{x \in \mathbb{R}^n \mid Ax \leq b, x \geq 0\},$$

are convex. Which of them is affine?

(c) **Line.** Let $0 \neq v \in \mathbb{R}^n$ and take a fixed $x_0 \in \mathbb{R}^n$. Then the line with direction $v$ passing through $x_0$ is denoted by $L(v, x_0)$ and defined as the set

$$L(v, x_0) := \{x \in \mathbb{R}^n \mid x = \lambda v + x_0, \text{ for all } \lambda \in \mathbb{R}\}$$

(d) **Halfline or Ray.** Let $v \in \mathbb{R}^n$ and $x_0 \in \mathbb{R}^n$ as in (a). The halfline with direction $v$ and vertex $x_0$ is denoted by $L_+(v, x_0)$ and defined as the set

$$L_+(v, x_0) := \{x \in \mathbb{R}^n \mid x = \lambda v + x_0, \text{ for all } \lambda \geq 0\}.$$  

A halfline is also called a ray.

(e) **Hyperplane.** Let $0 \neq a \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$. Then the set

$$H(a, \gamma) := \{x \in \mathbb{R}^n \mid \langle a, x \rangle = \gamma\}$$

convex. The set $H(a, \gamma)$ is called the hyperplane defined by $a$ and $\gamma$. The vector $a$ is a normal of the hyperplane.

(f) **Halfspace.** Take $0 \neq a \in \mathbb{R}^n$, $\gamma \in \mathbb{R}$, and $H(a, \gamma)$ the hyperplane defined by $a$ and $\gamma$. The sets

$$H(a, \gamma)^+ := \{x \in \mathbb{R}^n \mid \langle a, x \rangle \geq \gamma\} \text{ and } H(a, \gamma)^- := \{x \in \mathbb{R}^n \mid \langle a, x \rangle \leq \gamma\}$$

are convex. The sets $H(a, \gamma)^+$ and $H(a, \gamma)^-$ are called the halfspaces defined by $H(a, \gamma)$.

(g) **Balls.** The closed and open balls $B(x_0, r)$ and $B[x_0, r]$ are convex (see Section A.4).

Exercise 2.1.2 Prove the following assertions.

(a) Let $\{C_i\}_{i \in I}$ be an arbitrary collection of convex sets. Then $\bigcap_{i \in I} C_i$ is convex.

(b) Let $C \subset \mathbb{R}^n$ be a convex set and $a \in \mathbb{R}$. Then the set $aC := \{ax \mid x \in C\}$ is convex.

(c) Let $C_1, C_2 \subset \mathbb{R}^n$ be convex sets. Then the set $C_1 + C_2 := \{x_1 + x_2 \mid x_1 \in C_1, x_2 \in C_2\}$ is convex.

(d) Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be an affine function, i.e., $F(x) = L(x) + b$, with $L : \mathbb{R}^n \to \mathbb{R}^m$ linear and $b \in \mathbb{R}^m$. Then for every convex set $C \subset \mathbb{R}^n$, the set $F(C) := \{F(y) \mid y \in C\} \subset \mathbb{R}^m$ is convex. If $G : \mathbb{R}^m \to \mathbb{R}^n$ is an affine function, then the set

$$G^{-1}(C) := \{z \in \mathbb{R}^m \mid G(z) \in C\},$$

is convex.
(e) Let $C \subset \mathbb{R}^n$. Prove that the following statements are equivalent.

(i) $C$ is an affine set

(ii) For every $x_0 \in C$, the set $C - x_0 := \{z - x_0 : z \in C\}$ is a subspace.

(iii) There exists $x_0 \in C$ such that the set $C - x_0$ is a subspace.

(f) Let $C_1 \subset \mathbb{R}^n$ and $C_2 \subset \mathbb{R}^m$ be two convex sets. Then the Cartesian product of these sets, denoted by $C_1 \times C_2$, and defined as

$$C_1 \times C_2 := \{(x_1, x_2) \in \mathbb{R}^n \times \mathbb{R}^m | x_1 \in C_1, x_2 \in C_2\},$$

is convex.

(g) Show with an example that the union of convex sets may not be convex. For which families $\{C_i\}_{i \in I}$ you can say that $\bigcup_i C_i$ is convex?

(h) Prove that a single point is convex. Prove that the empty set is convex.

(i) Which of the sets in Example 2.1.1 have a convex boundary? Which of the sets in Example 2.1.1 have a convex and nonempty interior? (See Section A.6).

(j) A set $C \subset \mathbb{R}^n$ is called a cone if for every $x \in C$ and $\alpha \geq 0$ we have $\alpha x \in C$. Prove that the intersection of cones is a cone. Prove that the union of cones is also a cone. Prove that the sum of cones is also a cone. If $C$ is a cone, then for any given $a \in \mathbb{R}$, we have that the set $aC$ is a cone.

Prove that the following statements are equivalent.

(a) $C$ is a convex cone (that is, a cone which is also convex),

(b) $C + C \subset C$ (that is, for every $x, y \in C$, we have $x + y \in C$).

Give examples of nonconvex cones. Prove that an affine set is a cone if and only if it is a subspace. When is a halfspace a cone? When is the intersection of halfspaces a cone? (Hint: draw pictures!)

2.1.2 Convex Hull

We pointed out before that the intersection of convex sets is convex. This fact is instrumental in order to define the smallest convex set containing a fixed set $A \subset \mathbb{R}^n$. We define next this set.

**Definition 2.1.2** Let $A \subset \mathbb{R}^n$. The convex hull of $A$, denoted by $\text{co}(A)$ is the intersection of all convex sets containing $A$; that is,

$$\text{co}(A) = \cap_{M \supseteq A, M \text{ convex}} M.$$

**Example 2.1.2** Let $\{x_0, \ldots, x_p\} \subset \mathbb{R}^n$. The $p+1$ points $\{x_0, x_1, \ldots, x_p\}$ are affinely independent if and only if the $p$ vectors $\{x_1 - x_0, \ldots, x_p - x_0\}$ are linearly independent. A $k$-simplex is the convex hull of $k+1$ affinely independent points. The dimension of a $k$-simplex is $k$. In Figure 2.1.2, we depicted a 1-simplex, a 2-simplex and a 3-simplex in $\mathbb{R}^3$.

There is a much simpler way of expressing $\text{co}(A)$ than the one given in Definition 2.1.2.

**Theorem 2.1.1** Let $A \subset \mathbb{R}^n$. Then $\text{co}(A)$ is the set of all (finite) convex combinations of elements of $A$. 

Proof. Denote by $B$ the set of all finite convex combinations of elements of $A$. With this notation, the conclusion of the theorem can be restated as $B = \text{co}(A)$.

We want to prove first that $B \subset \text{co}(A)$. Indeed, take $z \in B$. By definition of $B$ there exist \( \{a_1, \ldots, a_p\} \subset A \) and $\lambda_1, \ldots, \lambda_p$ convex coefficients (i.e., $\sum_{i=1}^{p} \lambda_i = 1$ and all $\lambda_j \geq 0$) such that

\[
z = \sum_{i=1}^{p} \lambda_i a_i.
\]

Note that $A \subset \text{co}(A)$, with $\text{co}(A)$ a convex set. Hence, for all $i = 1, \ldots, p$ we have $a_i \subset \text{co}(A)$. Using now the convexity of $\text{co}(A)$ we conclude that every convex combination of the $a_i$s must also remain in $\text{co}(A)$. In particular, $z \in \text{co}(A)$. This gives $B \subset \text{co}(A)$. In order to prove the opposite inclusion, it is enough to show that $B$ is convex. Indeed, in the latter situation we would have $B$ convex and $B \supset A$, so $B \supset \text{co}(A)$ by definition of $\text{co}(A)$. Thus, let us show that $B$ is convex. Take $x, y \in B$, so

\[
x = \sum_{i=1}^{p} \lambda_i x_i \quad \text{and} \quad y = \sum_{i=1}^{s} \beta_i y_i,
\]

where $\{x_1, \ldots, x_p, y_1, \ldots, y_s\} \subset A$ and the coefficients $\{\lambda_1, \ldots, \lambda_p\}, \{\beta_1, \ldots, \beta_s\}$ are nonnegative and verify

\[
\sum_{i=1}^{p} \lambda_i = 1 \quad \text{and} \quad \sum_{i=1}^{s} \beta_i = 1.
\]

Fix $\alpha \in (0, 1)$, the convex combination

\[
z_0 := \alpha x + (1 - \alpha)y = \sum_{i=1}^{p} (\alpha \lambda_i) x_i + \sum_{i=1}^{s} [(1 - \alpha)\beta_i] y_i.
\]

It is easy to check that $\sum_{i=1}^{p} (\alpha \lambda_i) + \sum_{i=1}^{s} [(1 - \alpha)\beta_i] = 1$. Therefore $z_0 \in B$ by definition of $B$. This proves the convexity of $B$. \[\square\]

2.1.3 Convex Functions

Now we can give the definition of convex functions. It is theoretically important to deal with functions which can attain the value $+\infty$. This justifies the following definition.

Definition 2.1.3 Consider a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. The domain of $f$ is the set

\[
\text{dom } f := \{x \in \mathbb{R}^n \mid f(x) < +\infty\}.
\]

Therefore, we can think of $f$ as a function which is finite on its domain, and it is identically equal to $+\infty$ at points outside the set $\text{dom } f$.

Exercise 2.1.3 Let $f_1, f_2 : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be two functions. What is $\text{dom } (f_1 + f_2)$?
Definition 2.1.4 A function $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is said to be convex if and only if for every pair of points $x, y \in \text{dom } f$, and every $\lambda \in [0, 1]$, it holds that

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y).$$

In other words, the value of $f$ on a convex combination of $x$ and $y$ is less than or equal to the same convex combination of the values $f(x)$ and $f(y)$ (see Figure 2.1.3). Note that the inequality above is always true when $x$ or $y$ do not belong to $\text{dom } f$.

Definition 2.1.5 A function $g : \mathbb{R}^n \to \mathbb{R}$ is said to be a concave function if and only if $-g$ is convex.

Exercise 2.1.4 Prove that, if $f$ is convex, then $\text{dom } f$ must be a convex set.

Exercise 2.1.5 (a) Let $F : \mathbb{R}^n \to \mathbb{R}$ be an affine function. Then $F$ is convex.

(b) The function $\|x\| := (\sum_{i=1}^n |x_i|^2)^{1/2}$ (i.e., the Euclidean norm in $\mathbb{R}^n$) is convex.

(c) Prove that $F : \mathbb{R}^n \to \mathbb{R}$ is an affine function if and only if both $F$ and $-F$ are convex.

Exercise 2.1.6 Let $C \subset \mathbb{R}^n$ be a set. Prove that the following statements are equivalent.

(a) The set $C$ is convex.

(b) The function $\delta_C : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ defined as

$$\delta_C(x) := \begin{cases} 0 & x \in C, \\ +\infty & x \notin C, \end{cases}$$

is a convex function.

Exercise 2.1.7 Exercise 2.1.2(d) cannot be extended to a general convex function $F : \mathbb{R}^n \to \mathbb{R}$. Indeed, consider the set $C := [0, \infty)$ and the convex function $F(t) := \begin{cases} 10 & t = 0, \\ -\log (t + 1) & t > 0. \end{cases}$

Verify that $F$ and $C$ are convex, while the set $F(C)$ is not convex.

Lemma 2.1.1

(a) A differentiable function $\varphi : \mathbb{R} \to \mathbb{R}$ is convex if and only if its derivative $\varphi'$ is increasing.

(b) A twice differentiable function $\varphi : \mathbb{R} \to \mathbb{R}$ is convex if and only if $\varphi''(t) \geq 0$ for every $t \in \mathbb{R}$. 


Proof. (a) Assume that \( \varphi \) is convex, let us prove that \( \varphi' \) is increasing. First, we will show that for all \( t, s \in \mathbb{R} \) it holds
\[
\varphi(t) \geq \varphi(s) + \varphi'(s)(t-s). \tag{2.1.1}
\]
The expression above is precisely the Gradient Inequality in \( \mathbb{R} \) (see Theorem 2.2.2). Note that (2.1.1) holds (as an equality) when \( t = s \). So we can assume that \( s > t \). Because \( \varphi \) is convex, we can write for \( \alpha \in (0, 1] \)
\[
\varphi(t + \alpha(s - t)) = \varphi(\alpha s + (1 - \alpha)t) \leq \alpha \varphi(s) + (1 - \alpha)\varphi(t),
\]
which can be re-written as
\[
\frac{\varphi(t + \alpha(s - t)) - \varphi(t)}{\alpha(s - t)} \leq \frac{\varphi(s) - \varphi(t)}{(s - t)}.
\]
We used in the expression above the fact that \( \alpha \in (0, 1] \) and \( s > t \). Taking limit in the above expression for \( \alpha \to 0^+ \) we get
\[
\varphi'(t) \leq \frac{\varphi(s) - \varphi(t)}{(s - t)}.
\]
Note that \( s > t \), so the inequality above yields (2.1.1). The proof of (2.1.1) for the case \( s < t \) can be obtained in a similar way, and is left as an exercise. Now we use (2.1.1) to prove that \( \varphi' \) is increasing. By (2.1.1) we can write
\[
\varphi(t) - \varphi(s) \geq \varphi'(s)(t-s)
\]
\[
\varphi(s) - \varphi(t) \geq \varphi'(t)(s-t),
\]
summing up both expressions and re-arranging the resulting inequality we get
\[
0 \leq (\varphi'(s) - \varphi'(t))(s - t),
\]
which means that \( \varphi' \) is increasing. Assume now that \( \varphi' \) is increasing, and let us prove that \( \varphi \) is convex. Fix \( s, t \) two different points and \( \alpha \in (0, 1) \) (note that the inequality in the definition of convex function trivially holds when \( s = t \) or \( \alpha \) is equal to \( 0 \) or \( 1 \)). Assume that \( t < s \) and call \( t_\alpha := at + (1 - \alpha)s \). The Mean Value Theorem for \( \varphi \) gives us the equalities
\[
\varphi(t_\alpha) = \varphi(t) + \varphi'(\theta)(t_\alpha - t)
\]
\[
\varphi(t_\alpha) = \varphi(s) + \varphi'(\eta)(t_\alpha - s), \tag{2.1.2}
\]
where \( t \leq \theta \leq t_\alpha \leq \eta \leq s \). Because \( \varphi' \) is increasing, we know that
\[
\varphi'(\theta) \leq \varphi'(\eta). \tag{2.1.3}
\]
Multiplying the first equality in (2.1.2) by \( \alpha \), the second one by \( 1 - \alpha \), and summing up both equalities, we obtain
\[
\varphi(t_\alpha) = \alpha \varphi(t) + (1 - \alpha)\varphi(s) + \alpha(1 - \alpha)(s - t)[\varphi'(\theta) - \varphi'(\eta)].
\]
From (2.1.3), we see that the last term on the right-hand side is nonpositive. This yields
\[
\varphi(t_\alpha) = \alpha \varphi(t) + (1 - \alpha)\varphi(s) + \alpha(1 - \alpha)(s - t)[\varphi'(\theta) - \varphi'(\eta)]
\]
\[
\leq \alpha \varphi(t) + (1 - \alpha)\varphi(s),
\]
which proves the convexity of \( \varphi \). Let us prove now (b). Assume that \( \varphi \) is twice differentiable. Statement (b) now follows from the classical fact that a function from \( \mathbb{R} \) to \( \mathbb{R} \) is increasing if and only if its derivative is nonnegative. □

From Lemma 2.1.5, we see how to find out whether a (differentiable or twice differentiable) function \( \varphi : \mathbb{R} \to \mathbb{R} \) is convex or not. We can use these facts for determining whether a (differentiable or twice differentiable) function \( f : \mathbb{R}^n \to \mathbb{R} \) is convex or not. The one-dimensional result can be used for the \( n \)-dimensional case thanks to the next fact, which we state as an exercise.
2.1.4 Operations preserving convexity of functions

We list next some operations which preserve the convexity of functions.

**Lemma 2.1.2** Assume that $f, g : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ are convex, then $af + bg$ is convex for all $a, b \geq 0$.

**Proof.** The proof follows directly from the definitions and is left as an exercise. $\square$

2.2 Differentiability Properties of Convex Functions

We start this section by recalling the concept and some ideas related with a differentiable function. Let $S \subset \mathbb{R}^n$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}$. Consider the interior of the set $S$, which we denote as $S^o$ (see Definition A.6.25). We say that $f$ is differentiable at $x_0 \in S^o$ is there exists a vector $\nabla f(x_0)$, called the gradient vector of $f$ at $x_0$, and a function $R : \mathbb{R}^n \rightarrow \mathbb{R}$ such that

$$f(x) = f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle + R(x-x_0) \quad \text{for every } x \in S,$$  

(2.2.4)

where $\lim_{x \rightarrow x_0} \frac{R(x-x_0)}{\|x-x_0\|} = 0$. Because we approach the point $x_0$ from all possible directions in the space, the point $x_0$ must be in the interior of $S$. We say that $f$ is differentiable on an open subset $S' \subset S$ if it is differentiable at every point $x \in S'$. The representation (2.2.4) of $f(x)$ around the point $x_0$ is called a first-order expansion of $f$ around (or at) $x_0$. The first two terms of this representation; that is, the linear function

$$l(x) := f(x_0) + \langle \nabla f(x_0), x - x_0 \rangle,$$  

(2.2.5)

is called a first-order (Taylor series) approximation of $f$ around (or at) $x_0$. The last term in expression (2.2.4) is called the remainder term of the Taylor expansion.

Let $f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\}$ be a function which is differentiable at a point $x_0 \in \text{dom} f$ and let $v \in \mathbb{R}^n$ be a nonzero direction. Then the limit

$$\frac{\partial}{\partial v} f(x_0) = \lim_{t \rightarrow 0} \frac{f(x_0 + tv) - f(x_0)}{t},$$  

(2.2.6)

is well defined an is called the directional derivative of $f$ in the direction $v$ at the point $x_0$. This directional derivative is denoted by $\frac{\partial}{\partial v} f(x_0)$ and it holds that $\frac{\partial}{\partial v} f(x_0) = \langle \nabla f(x_0), v \rangle$. If $f$ is not differentiable at $x_0$, then we might have to compute the side-limits (i.e., for $t \rightarrow 0^+$ or $t \rightarrow 0^-$) of the expression above (see (2.2.16)).

**Exercise 2.2.8** Take $f : \mathbb{R}^n \rightarrow \mathbb{R}$. For every fixed $0 \neq v \in \mathbb{R}^n$ we can define the function $\varphi(t) := f(x_0 + tv)$. Prove that

1. If $f$ is differentiable at $x_0 \in \mathbb{R}^n$ then $\varphi$ is differentiable at 0 and in this case we have $\varphi'(0) = \nabla f(x_0)^T v = \langle \nabla f(x_0), v \rangle$.

2. If $f$ is twice differentiable at $x_0 \in \mathbb{R}^n$ then $\varphi$ is twice differentiable at 0 and $\varphi''(0) = v^T \nabla^2 f(x_0) v = \langle \nabla^2 f(x_0)v, v \rangle$.

Hint: For (1), compute $\varphi'(0) = \lim_{t \rightarrow 0} \frac{\varphi(t) - \varphi(0)}{t}$ using the definition of $\varphi$, and the Taylor expansion (2.2.4). For (2), use the equality $\varphi''(0) = \frac{\partial}{\partial v} \frac{\partial}{\partial u} f(x_0)$ and the fact that for every pair of nonzero directions $u, w \in \mathbb{R}^n$ we have that $\frac{\partial}{\partial w} \frac{\partial}{\partial u} f(x_0) = \langle \nabla^2 f(x_0)w, v \rangle$. 

**Definition 2.2.6** A function \( f : \mathbb{R}^n \to \mathbb{R} \cup \{ +\infty \} \) is said to be strictly convex if and only if, for every pair of different points \( x, y \in \text{dom} \ f \), and every \( \lambda \in (0, 1) \), it holds that
\[
f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y).
\]
In other words, the value of \( f \) on a nontrivial (that is, \( \lambda \neq 0, 1 \)) convex combination of \( x \) and \( y \) is strictly less than the same convex combination of the values \( f(x) \) and \( f(y) \).

**Exercise 2.2.9** Prove that an affine function \( F : \mathbb{R}^n \to \mathbb{R} \) is not strictly convex.

**Lemma 2.2.3** Let \( f : \mathbb{R}^n \to \mathbb{R} \cup \{ +\infty \} \). For every fixed \( 0 \neq v \in \mathbb{R}^n \) and every \( x_0 \in \mathbb{R}^n \), consider the function \( \varphi(t) := f(x_0 + tv) \). Then,

(i) If \( f \) is convex, then \( \varphi \) is convex.

(ii) Conversely, if for every fixed \( 0 \neq v \in \mathbb{R}^n \) and every \( x_0 \in \mathbb{R}^n \) we have that \( \varphi(t) := f(x_0 + tv) \) is convex, then \( f \) must be convex.

(iii) Same as in (i) and (ii), but replacing “convex” by “strictly convex” everywhere.

**Proof.** We prove (i) and (ii). The proof of (iii) is very similar to (i) and (ii), and is left as an exercise. Proof of (i). Case 1: \( C = \mathbb{R}^n \). Assume that \( f \) is convex, we want to prove that \( \varphi(t) := f(x_0 + tv) \) is convex. Let \( t_1, t_2 \in \text{dom} \varphi \) and fix \( a \in [0, 1] \).
\[
\varphi(at_1 + (1 - a)t_2) = f(x_0 + (at_1 + (1 - a)t_2)v)
\]
\[
= f(a(x_0 + t_1v) + (1 - a)(x_0 + t_2v)) \leq af(x_0 + t_1v) + (1 - a)f(x_0 + t_2v)
\]
\[
= a\varphi(t_1) + (1 - a)\varphi(t_2),
\]
therefore \( \varphi \) is convex. Let us prove now (ii). Assume that for every fixed \( 0 \neq v \in \mathbb{R}^n \) and every \( x_0 \in \mathbb{R}^n \) we have that \( \varphi(t) := f(x_0 + tv) \) is convex, and let us prove that \( f \) is convex. Fix \( x, y \in \text{dom} \ f \) and \( a \in [0, 1] \). We must prove that
\[
f(ax + (1 - a)y) \leq af(x) + (1 - a)f(y).
\]
(2.2.7)
We have
\[
a(x + (1 - a)y) = y + a(x - y).
\]
Take now \( x_0 = y \) and \( v = (x - y) \) in the definition of \( \varphi \). We know that \( \varphi(t) := f(y + t(x - y)) \) is convex, with \( \varphi(1) = f(x) \) and \( \varphi(0) = f(y) \). Therefore,
\[
f(ax + (1 - a)y) = \varphi(a) = \varphi(a 1 + (1 - 0) 0) \leq a\varphi(1) + (1 - a)\varphi(0) = af(x) + (1 - a)f(y),
\]
so (2.2.7) holds. We established (2.2.7) for arbitrary \( x, y \in \text{dom} \ f \) and \( a \in [0, 1] \), therefore \( f \) is convex. \now

### 2.2.1 Gradients and Hessians

We start this section with an important corollary of Lemma 2.2.3.

**Corollary 2.2.1** Let \( f : \mathbb{R}^n \to \mathbb{R} \) be twice differentiable on an open set \( C \subset \mathbb{R}^n \). Then \( f \) is convex on \( C \) if and only if \( \nabla^2 f(x) \) is positive semidefinite for all \( x \in C \).

**Proof.** The function \( f \) is convex if and only if for every \( x_0 \in \mathbb{R}^n \) and every \( 0 \neq v \in \mathbb{R}^n \) we have that \( \varphi(t) := f(x_0 + tv) \) is convex. Because \( f \) is twice differentiable, we must have \( \varphi \) twice differentiable as well. Using also the fact that \( \varphi \) is convex, we have \( \varphi''(t) \geq 0 \) for all \( t \in \mathbb{R} \). But \( \varphi''(t) = \langle \nabla^2 f(x_0 + tv)v, v \rangle \geq 0 \) for all \( t \in \mathbb{R} \). If \( t = 0 \), we get \( \langle \nabla^2 f(x_0)v, v \rangle \geq 0 \). Altogether, we conclude that for every \( x_0 \) and \( v \) we will have that \( \langle \nabla^2 f(x_0)v, v \rangle \geq 0 \). Equivalently, \( \nabla^2 f(x_0) \) is positive semidefinite for every \( x_0 \in \mathbb{R}^n \). \now
Exercise 2.2.10 Consider the function \( f : \mathbb{R}^2 \to \mathbb{R} \cup \{+\infty\} \) defined as
\[
f(x_1, x_2) := \begin{cases} 
eq x_1 x_2 & \text{if } x_1 \geq 0, x_2 \geq 0, \\ +\infty & \text{otherwise.} \end{cases}
\]
Use the previous result to prove that \( f \) is not convex.

Exercise 2.2.11 Let \( f : \mathbb{R}^2 \to \mathbb{R} \) be a twice differentiable function. Prove that there exists a \( \lambda \in \mathbb{R} \) such that \( g : \mathbb{R}^2 \to \mathbb{R} \) defined as \( g(x) := f(x) + (\lambda/2)\|x\|^2 \) is convex. Hint: compute the Hessian of \( g \) and prove that for \( \lambda \) large enough, this Hessian is positive definite.

The following result is essential in the development and analysis of optimality conditions for convex optimization. It is known as the Gradient Inequality.

Theorem 2.2.2 Let \( f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \).

(a) The function \( f \) is convex if and only if
\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \tag{2.2.8}
\]
for every \( y \in \mathbb{R}^n \) and every \( x \) such that \( f \) is differentiable at \( x \) (i.e., for every \( x \in \text{dom} f \) such that \( \nabla f(x) \) exists).

(b) The function \( f \) is strictly convex if and only if the inequality (2.2.8) is strict whenever \( y \neq x \).

Proof. (a) If \( y = x \), (2.2.8) trivially holds. Fix now \( x, y \) such that \( x \neq y \). Since \( f \) is differentiable at \( x \), all directional derivatives exist and they are obtained as \( \partial f(x)/\partial w = \langle \nabla f(x), w \rangle \). Hence
\[
\partial f(x)/\partial (y - x) = \lim_{\alpha \to 0^+} \frac{f(x + \alpha(y - x)) - f(x)}{\alpha} = \langle \nabla f(x), (y - x) \rangle. \tag{2.2.9}
\]
Assume that \( f \) is convex, so for all \( \alpha \in [0, 1] \),
\[
f(x + \alpha(y - x)) \leq \alpha f(y) + (1 - \alpha)f(x),
\]
from which
\[
\frac{f(x + \alpha(y - x)) - f(x)}{\alpha} \leq f(y) - f(x),
\]
for all \( \alpha \in (0, 1] \). Taking now limit for \( \alpha \to 0^+ \) and using (2.2.9) we get inequality (2.2.8). For the proof of the converse, assume that inequality (2.2.8) holds. Fix \( x, y \in \mathbb{R}^n \) and \( \alpha \in [0, 1] \). Call \( z := \alpha x + (1 - \alpha)y \). Use (2.2.8) twice, to get
\[
\begin{align*}
f(x) & \geq f(z) + \langle \nabla f(z), x - z \rangle, \\
f(y) & \geq f(z) + \langle \nabla f(z), y - z \rangle. \tag{2.2.10}
\end{align*}
\]
Multiply the first inequality by \( \alpha \), the second by \( (1 - \alpha) \), and add them to get
\[
\alpha f(x) + (1 - \alpha)f(y) \geq f(z) = f(\alpha x + (1 - \alpha)y),
\]
yielding convexity of \( f \). The part of (b) which assumes that the strict inequality holds, and establishes strict convexity of \( f \), proceeds in a way similar to the second part of (a) (where you assume strict inequalities in (2.2.10)), and is left to the reader. Let us now prove that, if the function is strictly convex, then the strict inequality holds in (2.2.8) whenever \( y \neq x \). A strictly convex function is in particular convex, so we know by part (a) of this theorem that for every \( y \in \text{dom} f \) and every \( x \) such that \( \nabla f(x) \) exists, we must have that (2.2.8) holds. Assume that there exist \( y, x \) such that the inequality is not strict, so we must have
\[
f(y) = f(x) + \langle \nabla f(x), y - x \rangle. \tag{2.2.11}
\]
Because $f$ is strictly convex, for any $a \in (0, 1)$ we can write

$$f(ay + (1-a)x) = f(x + a(y-x)) < af(y) + (1-a)f(x). \quad (2.2.12)$$

Using (2.2.8), and combining (2.2.12) with (2.2.11) we obtain

$$f(x) + a((\nabla f(x), y-x)) \leq f(x + a(y-x)) < f(x) + a((\nabla f(x), y-x)),$$

which is a contradiction. Therefore the strict inequality must hold for all pairs $y, x$ such that $y \in \text{dom } f$ and $\nabla f(x)$ exists. \qed

**Remark 2.2.2** The result above has important applications in convex optimization. One of them is the fact that inequality (2.2.8) asserts that the linear approximation $l(x) := f(x) + (\nabla f(x), y-x)$ bounds $f$ from below. Therefore, if we have a constrained convex optimization problem of the kind

$$(P) \quad \min_{x \in C} f(x),$$

where the set $C \subset \mathbb{R}^n$ is convex and bounded (see Definition A.5.21), then the minimum of the linear function $l$ in $C$ can be easily computed and provides a lower bound for the optimal value of problem (P). Inequality (2.2.8) can also be useful for approximating the constraint set. For instance, if the constraint set $C := \{ y \in \mathbb{R}^n : g_i(y) \leq 0, \ i = 1, \ldots, m \}$, where the functions $g_i : \mathbb{R}^n \to \mathbb{R}$ are convex and differentiable, then the set $\tilde{C} := \{ y : g_i(x) + (\nabla g_i(x), y-x) \leq 0, \ i = 1, \ldots, m \}$ is a polyhedral (that is, an intersection of a finite number of half-spaces) approximation of $C$. Inequality (2.2.8) for the $g_i$s implies that $\tilde{C} \supset C$ and therefore $\tilde{C}$ is called an outer polyhedral approximation of $C$.

**Exercise 2.2.12** Consider $C$ and $\tilde{C}$ as in the previous remark. Prove that $\tilde{C} \supset C$.

**Corollary 2.2.2** Assume that $\varphi : \mathbb{R} \to \mathbb{R}$ is differentiable. Then $\varphi$ is strictly convex if and only if $\varphi'$ is strictly increasing.

**Proof.** From Theorem 2.2.2(b), we have that $\varphi$ is strictly convex if and only if for every $s \neq t$ we have

$$\varphi(s) > \varphi(t) + \varphi'(t)(s-t),$$

$$\varphi(t) > \varphi(s) + \varphi'(s)(t-s),$$

adding up the inequalities above and rearranging the resulting expression we obtain

$$(\varphi'(t) - \varphi'(s))(t-s) > 0,$$

whenever $s \neq t$. Equivalently, $\varphi'$ is strictly increasing. \qed

**Exercise 2.2.13** Let $f : \mathbb{R}^n \to \mathbb{R}$. Prove that $f$ is strictly convex if and only if for every fixed $0 \neq v \in \mathbb{R}^n$ and every $x_0 \in \mathbb{R}^n$ it holds that $\varphi'$ is strictly increasing, where $\varphi(t) := f(x_0 + tv)$. Hint: Use the previous corollary and Theorem 2.2.2(b).

The next result extends Lemma 2.1.1(a) to functions from $\mathbb{R}^n$ to $\mathbb{R}$.

**Theorem 2.2.3** Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable (i.e., $\nabla f(x)$ exists for every $x \in \mathbb{R}^n$). The function $f$ is convex if and only if

$$\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0. \quad (2.2.13)$$

for every $x, y \in \mathbb{R}^n$. 

Proof. Use the Gradient inequality (see Theorem 2.2.2(a)) to write

\[
\begin{align*}
  f(x) &\geq f(y) + \langle \nabla f(y), x - y \rangle, \\
  f(y) &\geq f(x) + \langle \nabla f(x), y - x \rangle,
\end{align*}
\]

adding up the inequalities above and rearranging the resulting expression we obtain (2.2.13). Conversely, assume that (2.2.13) holds. In order to prove that \(f\) is convex we use Lemma 2.2.3(ii). Take arbitrary \(x_0 \in \mathbb{R}^n\) and \(0 \neq v \in \mathbb{R}^n\) and let us show that \(\varphi(t) := f(x_0 + tv)\) is convex. The latter is true if and only if \(\varphi'\) is increasing. Note that \(\varphi'(t) = \langle \nabla f(x_0 + tv), v \rangle\). Using also (2.2.13) we have

\[
0 \leq \langle \nabla f(x_0 + tv) - \nabla f(x_0 + sv), (x_0 + tv) - (x_0 + sv) \rangle = (t - s)\langle \nabla f(x_0 + tv) - \nabla f(x_0 + sv), v \rangle = (t - s)(\varphi'(t) - \varphi'(s)),
\]
equivalently, \(\varphi'\) is increasing. By Lemma 2.2.3(ii), we conclude that \(f\) is convex. \(\square\)

Exercise 2.2.14 Let \(f : \mathbb{R}^n \to \mathbb{R}\) be differentiable (i.e., \(\nabla f(x)\) exists for every \(x \in \mathbb{R}^n\)). Prove that \(f\) is strictly convex if and only if

\[
\langle \nabla f(x) - \nabla f(y), x - y \rangle > 0.
\]

for every \(x, y \in \mathbb{R}^n\) such that \(x \neq y\).

Remark 2.2.3 A function \(G : \mathbb{R}^n \to \mathbb{R}^n\) which verifies

\[
\langle G(x) - G(y), x - y \rangle \geq 0 \quad \forall \; x, y \in \mathbb{R}^n,
\]
is called a monotone mapping. If the inequality above is strict whenever \(x \neq y\), we say that \(G\) is a strictly monotone mapping. Therefore, Theorem 2.2.3 asserts that the gradient of a convex function is a monotone mapping, and Exercise 2.2.14 asserts that the gradient of a strictly convex function is a strictly monotone mapping. This extends the fact that the derivative of a (strictly) convex function from \(\mathbb{R}\) to \(\mathbb{R}\) is (strictly) increasing.

2.2.2 One-sided directional derivatives

Directional derivatives (see (2.2.6)) are important for checking whether or not a given point is a local solution of an optimization problem. They are also important for devising efficient optimization methods. However, when \(f\) is not differentiable, the directional derivative (as defined in (2.2.6)) might not exist. In this situation, we may only be able to analyze the one sided-directional derivative of \(f\) at the point \(x_0\) in the direction \(v\). This derivative, denoted as \(f'(x_0, v)\), represents the speed of \(f\) at the point \(x_0\), in the direction of “positive” \(v\). More precisely, let \(S \subset \mathbb{R}^n\) and \(f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}\). Fix \(x_0 \in \text{dom } f\) such that \(x_0 + tv \in \text{dom } f\) for all \(t > 0\) sufficiently small. The one-sided directional derivative of \(f\) at the point \(x_0\) in the direction \(v\) is defined as

\[
f'(x_0, v) := \lim_{t \to 0^+} \frac{f(x_0 + tv) - f(x_0)}{t},
\]

(compare this limit with the one in (2.2.6)). This limit is often called (for short) the directional derivative of \(f\) at the point \(x_0\) in the direction \(v\), however it should be understood that it is just one of the side limits required in (2.2.6).

Example 2.2.3 Let \(f : \mathbb{R} \to \mathbb{R}\) defined as

\[
f(x) := \begin{cases} 
-\sqrt{x} & \text{if } x \geq 0, \\
0 & \text{if } x < 0.
\end{cases}
\]
In \( \mathbb{R} \) we can only consider two directions: 1 and \(-1\). Let us compute \( f'(0, 1) \) and \( f'(0, -1) \).

\[
f'(0, 1) = \lim_{t \to 0^+} \frac{f(0 + t) - f(0)}{t} = \lim_{t \to 0^+} \frac{-\sqrt{t}}{t} = -\infty,
\]
and

\[
f'(0, -1) = \lim_{t \to 0^+} \frac{f(0 - t) - f(0)}{t} = 0.
\]

See Figure 2.2.4.

![Figure 2.2.4: Directional derivative of \( f \) in direction \( v = 1 \) does not exist.](image)

When \( f \) is a convex function, then the directional derivative exists in the interior of the domain. The proof of this result relies on a classical fact from Analysis, which we state next.

**Lemma 2.2.4** Let \( R > 0 \) and take a function \( q : [-R, R] \to \mathbb{R} \). Assume \( q \) is increasing and bounded below on the interval \((0, R]\). Then

\[
\inf_{\lambda > 0} q(\lambda), \tag{2.2.17}
\]

exists. Moreover, in this situation we have

\[
\inf_{\lambda > 0} q(\lambda) = \lim_{\lambda \to 0^+} q(\lambda). \tag{2.2.18}
\]

**Proof.** It is clear that the infimum exists because \( q \) is bounded below. The second statement follows from the fact that \( q \) decreases as \( \lambda \) decreases to \( 0^+ \). \( \Box \)

**Lemma 2.2.5** Let \( x_0 \in \text{dom } f \) and assume that there exists \( R > 0 \) such that \( x_0 + tv \in \text{dom } f \) for every \( t \in [-R, R] \). Then, the directional derivative \( f'(x_0, v) \) exists.

**Proof.** The proof is performed in steps. First, we prove that the function \( q : (0, R] \to \mathbb{R} \) defined by

\[
q(\lambda) := \frac{f(x_0 + \lambda v) - f(x_0)}{\lambda}
\]
is increasing; that is, whenever \( 0 < a_1 < a_2 \leq R \) we have \( q(a_1) \leq q(a_2) \). Indeed, the convexity of \( f \) yields

\[
f(x_0 + a_1 v) = f((a_1/a_2)(x_0 + a_2 v) + (1 - (a_1/a_2))x_0)
\leq (a_1/a_2)f(x_0 + a_2 v) + (1 - (a_1/a_2))f(x_0),
\]
which becomes, after elementary rearrangements,

\[
q(a_1) = \frac{f(x_0 + a_1 v) - f(x_0)}{a_1} \leq \frac{f(x_0 + a_2 v) - f(x_0)}{a_2} = q(a_2).
\]
Therefore $q$ is increasing, as claimed. Second, let us prove that $q$ is bounded below in $(0, R]$, so that we can use Lemma 2.2.4. Fix $a \in (0, 1)$. By assumption, we have that $x_0 - Rv \in \text{dom } f$, and the convexity of $f$ implies that

$$f(x_0) = f\left(\frac{a}{\alpha+1}(x_0 - Rv) + \frac{1}{\alpha+1}(x_0 + a(Rv))\right) \leq \frac{a}{\alpha+1}f(x_0 - Rv) + \frac{1}{\alpha+1}f(x_0 + (aR)v).$$

Rearranging the above expression gives

$$\frac{f(x_0) - f(x_0 - Rv)}{R} \leq \frac{f(x_0 + (aR)v) - f(x_0)}{aR} = q(aR),$$

for all $a \in (0, 1)$. Changing to the variable $\lambda := aR$ in the expression above we see that for every $\lambda \in (0, R)$

$$\frac{f(x_0) - f(x_0 - Rv)}{R} \leq \frac{f(x_0 + \lambda v) - f(x_0)}{\lambda} = q(\lambda).$$

Taking infimum in this expression for all $\lambda \in (0, R)$ we get

$$\frac{f(x_0) - f(x_0 - Rv)}{R} \leq \inf_{R>\lambda>0} q(\lambda) \quad (2.2.19)$$

By Lemma 2.2.4, the infimum in the right-hand side of the expression above exists. Because $q$ decreases as $\lambda$ decreases to $0^+$ (this is the argument in the proof of (2.2.18)), we have that

$$\inf_{R>\lambda>0} q(\lambda) = \lim_{\lambda \to 0^+} q(\lambda) = f'(x_0, v),$$

which exists because it coincides with the right-hand side of the expression in (2.2.19). The second equality in the expression above follows from (2.2.16) and the definition of $q$. □

**Exercise 2.2.15** Consider the functions $f_1, f_2, f_3 : \mathbb{R} \to \mathbb{R}$ defined as $f_1(x) := (1/2)|x|$, $f_2(x) := (x - 1)^2$, and $f_3(x) := (1/2)((1/2)x + 1)$. Take now

$$g(x) := \max \{f_1(x), f_2(x), f_3(x)\},$$

see Figure 2.2.5. Note that $g$ is not differentiable at the points $x_0 = 1/4$ and $x_1 = 2$. Compute the directional derivatives $g'(x_0, 1), g'(x_0, -1), g'(x_1, 1)$ and $g'(x_1, -1)$.

![Figure 2.2.5: g is the maximum of three functions.](image-url)
2.3 Epigraphs and Level Sets

Given a real-valued function, we can always associate to it its level sets. These sets will have special properties, depending on the properties of the given function.

**Definition 2.3.7** Let \( f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) be a continuous function. The *closed level set* of \( f \) is defined as

\[
S_f[\alpha] := \{ x \mid f(x) \leq \alpha \},
\]

and the *strict level set* of \( f \) is defined as

\[
S_f(\alpha) := \{ x \mid f(x) < \alpha \}.
\]

In the same way, the *level curve* of \( f \) is defined as

\[
\partial S_f(\alpha) := \{ x \mid f(x) = \alpha \}.
\]

**Remark 2.3.4** Note that the sets \( S_f[\alpha], S_f(\alpha) \) and \( \partial S_f(\alpha) \) are all contained in \( \text{dom } f \). Moreover, because \( f \) is a continuous function, we have that the sets

\[
S_f[\alpha] = f^{-1}((-\infty, \alpha]) , \quad \partial S_f(\alpha) = f^{-1}(\alpha),
\]

are closed. Indeed, they are the inverse images of the closed sets \((-\infty, \alpha]\) and \(\{\alpha\}\), respectively. By similar arguments, the set

\[
S_f(\alpha) = f^{-1}((-\infty, \alpha)),
\]

is open.

In Figure 2.3.6, we have that \( S_f[\alpha] = [x_1, x_2] \cup [x_3, x_4], S_f(\alpha) = (x_1, x_2) \cup (x_3, x_4), \) and \( \partial S_f(\alpha) = \{x_1, x_2, x_3, x_4\} \). As we can see in this figure, the sets \( S_f[\alpha] \) and \( S_f(\alpha) \) are not convex. However, when \( f \) is convex, the closed and strict level sets are always convex.

**Proposition 2.3.1** Assume that \( f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) is convex, then the sets \( S_f[\alpha] \) and \( S_f(\alpha) \) are convex.

**Proof.** The proof follows directly from the definitions and is left as an exercise.

**Exercise 2.3.16** Consider an optimization problem of the kind

\[
(P) \quad \min_{x \in C} f(x),
\]

where \( C \subset \mathbb{R}^n \) and \( f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) continuous. A solution of \( (P) \) is a point \( x^* \) such that \( x^* \in C \) and \( f(x) \geq f(x^*) \) for every \( x \in C \). Denote by \( P^* := \inf_{x \in C} f(x) \).

(i) Prove that the set of solutions of \( (P) \) is the level set \( S_{\tilde{f}}[P^*] \), where \( \tilde{f} : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) is defined as

\[
\tilde{f}(x) := \begin{cases} 
 f(x) & \text{if } x \in C, \\
 +\infty & \text{if } x \notin C.
\end{cases}
\]

(ii) Prove that \( S_f[P^*] = \emptyset \) if \( P^* = -\infty \).

(iii) The converse of (ii) does not hold. Find an example in which \( S_f[P^*] = \emptyset \) with \( P^* \in \mathbb{R} \).

**Remark 2.3.5** The converse of Proposition 2.3.1 is not true. More precisely, \( S_f[\alpha] \) and \( S_f(\alpha) \) can be convex for every \( \alpha \in \mathbb{R} \), for a nonconvex \( f \). See Figure 2.3.7.
Exercise 2.3.17 When possible, draw the strict and closed level sets of the following functions:

(a) \( f : \mathbb{R}^2 \to \mathbb{R} \) given by \( f(x) = x_1 + x_2 \).
(b) \( f : \mathbb{R}^2 \to \mathbb{R} \) given by \( f(x) = x_1^2 + x_2^2 \).
(c) \( f : \mathbb{R} \to \mathbb{R} \) given by \( f(x) = \sqrt{|x|} \).
(d) \( f : \mathbb{R} \to \mathbb{R} \) given by \( f(x) = \sin x \).
(e) \( f : \mathbb{R}^2 \to \mathbb{R} \) given by \( f(x) = e^{-x_1} + e^{-x_2} \).
(f) \( f : \mathbb{R}^3 \to \mathbb{R} \) given by \( f(x) = x_1 + x_2 + x_3 \).

Which of these level sets is convex for all \( \alpha \in \mathbb{R} \)? Which of these functions is convex? Which of these level sets are bounded? When possible, draw also the level curves of all these functions.

In Classical Calculus, it is usual to look at the graph of \( f \), denoted by \( \text{Graph}(f) \) and defined as:

\[
\text{Graph}(f) := \{(x,f(x)) \in \mathbb{R}^n \times \mathbb{R} \mid x \in \mathbb{R}^n \text{ such that } f(x) \in \mathbb{R}\}.
\]

However, when studying convex functions, the set of points \((x, \alpha) \in \mathbb{R}^n \times \mathbb{R}\) located above the graph of \( f \) is more useful. This set deserves a definition.
Definition 2.3.8 Given $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$, the epigraph of $f$ is denoted by $Epi f$ and defined as:

$$Epi f := \{(x, \alpha) \in \mathbb{R}^n \times \mathbb{R} | f(x) \leq \alpha\}.$$ 

If $(x, \alpha) \in Epi f$, then we must have $x \in \text{dom } f$. In Figure 2.3.8 we depicted some epigraphs. As we mentioned above, the level sets of $f$ may be convex, while $f$ is not. This pathology does not happen with the epigraph. In fact it holds that convexity of the epigraph is equivalent to the convexity of the function (see Proposition 2.3.2). In Figure 2.3.8 (a) and (d) we have epigraphs of nonconvex functions, while the other two correspond to convex functions.

![Figure 2.3.8: Some epigraphs.](image)

Remark 2.3.6 If $x_0 \in \text{dom } f$, then the vertical ray $\{(x_0, f(x_0) + r) : r \geq 0\} \subset Epi f$.

Proposition 2.3.2 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. Then $Epi f$ is convex if and only if $f$ is convex.

Proof. Assume first that $f$ is convex, and take two elements $(x_1, a_1), (x_2, a_2) \in Epi f$, so that $f(x_1) \leq a_1$ and $f(x_2) \leq a_2$. Then we can write for all $\alpha \in [0, 1]$

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2) \leq \alpha a_1 + (1 - \alpha)a_2.$$

The above inequality means that

$$(\alpha x_1 + (1 - \alpha)x_2, \alpha a_1 + (1 - \alpha)a_2) = \alpha(x_1, a_1) + (1 - \alpha)(x_2, a_2) \in Epi f,$$

so $Epi f$ is convex. Conversely, assume that $Epi f$ is convex. Fix $\alpha \in [0, 1]$ and $x_1, x_2 \in \text{dom } f$. We want to show that

$$f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2).$$

Because the points $(x_1, f(x_1)), (x_2, f(x_2)) \in Epi f$ and $Epi f$ is convex, we have

$$\alpha(x_1, f(x_1)) + (1 - \alpha)(x_2, f(x_2)) = (\alpha x_1 + (1 - \alpha)x_2, \alpha f(x_1) + (1 - \alpha)f(x_2)) \in Epi f.$$ 

By Remark 2.3.6 this means that $f(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha f(x_1) + (1 - \alpha)f(x_2)$. Hence $f$ is convex. □
Exercise 2.3.18 Draw all epigraphs of the functions of Exercise 2.3.17(a)-(e). Check which of these epigraphs is convex, and use Proposition 2.3.2 for determining which of the functions are convex.

Proposition 2.3.3 Let \( g_i : \mathbb{R}^n \to \mathbb{R} \cup \{ +\infty \} \) for \( i \in I \) be a family of convex functions. For every \( x \in \mathbb{R}^n \) define the function \( g : \mathbb{R}^n \to \mathbb{R} \cup \{ +\infty \} \) as

\[
g(x) := \sup_{i \in I} g_i(x).
\]

Then,

(a) \( \text{dom } g = \cap_{i \in I} \text{dom } g_i \),

(b) \( S_g[\alpha] = \cap_{i \in I} S_{g_i}[\alpha] \),

(c) \( \text{Epi } g = \cap_{i \in I} \text{Epi } g_i \),

(d) If all the \( g_i \)s are convex, then \( g \) is convex.

Proof. (Sketch of the Proof.) The proof of item (d) follows from item (c), Proposition 2.3.2 and Exercise 2.1.2. For (a), note that \( x \in \text{dom } g \) if and only if \( g(x) < +\infty \). This fact is equivalent to \( g_i(x) < +\infty \) for all \( i \in I \). In other words, \( x \in \cap_{i \in I} \text{dom } g_i \). Item (b) follows from (a). For Item (c), use that

\[
g(x) \leq a \iff g_i(x) \leq a \forall i \in I.
\]

\( \square \)

Exercise 2.3.19 Fill all the gaps in the proof of Proposition 2.3.3.

2.3.1 Support of sets at boundary points

Let \( S \) be a nonempty subset of \( \mathbb{R}^n \) and fix \( x_0 \) a boundary point of \( S \) (see Definition A.6.25 and Figure A.6.2 for a definition and graphic illustration of boundary points). Given a nonzero vector \( a \in \mathbb{R}^n \), the hyperplane with normal \( a \) passing through the point \( x_0 \) is

\[
H := \{ x \in \mathbb{R}^n : \langle x - x_0, a \rangle = 0 \}. \tag{2.3.20}
\]

With the notation of Example 2.1.1(e), we have that \( H = H(a, \langle x_0, a \rangle) \). The hyperplane \( H \) defined in (2.3.20) is a supporting hyperplane of \( S \) at the point \( x_0 \) if either

\[
S \subset H^+ := \{ x \in \mathbb{R}^n : \langle x - x_0, a \rangle \geq 0 \}, \tag{2.3.21}
\]

or else,

\[
S \subset H^- := \{ x \in \mathbb{R}^n : \langle x - x_0, a \rangle \leq 0 \}. \tag{2.3.22}
\]

The first inclusion means that \( \langle x - x_0, a \rangle \geq 0 \) for every \( x \in S \), and the second one means that \( \langle x - x_0, a \rangle \leq 0 \) for every \( x \in S \). When both inclusions hold simultaneously, it means that \( S \subset H \), and in this case \( H \) is an improper supporting hyperplane of \( S \) at \( x_0 \). Otherwise; that is, when \( S \not\subset H \), then \( H \) is called a proper supporting hyperplane of \( S \) at \( x_0 \). In Figure 2.3.9, we see examples of proper supporting hyperplanes of \( S \), and an improper supporting hyperplane of \( S' \).

Note that the smoothness of the boundary implies that the supporting hyperplane is unique at the points \( x_2, x_3, x_5 \). At \( x_1, x_4 \), the kink in the boundary allows the existence of an infinite number of supporting hyperplanes. Note that the same hyperplane supports the points \( x_2, x_3, \) and \( H \) supports an infinite number of boundary points of \( S \).
Remark 2.3.7 Fix a nonzero vector $a \in \mathbb{R}^n$. Consider the function $g : \mathbb{R}^n \to \mathbb{R}$ defined by $g(x) = \langle x, a \rangle$. Let $S \subset \mathbb{R}^n$ and take $x_0$ a boundary point of $S$. Then inclusion (2.3.21) is equivalent to

$$\inf_{x \in S} g(x) \geq g(x_0).$$

Hence, if $x_0 \in S$, we must have that (2.3.21) holds if and only if $x_0$ is a minimizer of $g$ over $S$. In a similar way, it can be checked by the reader that, if $x_0 \in S$, then (2.3.22) holds if and only if $x_0$ is a maximizer of $g$ over the $S$.

Exercise 2.3.20 Consider the epigraph of the function $g$ given in Exercise 2.2.15. In which points of this epigraph we have an infinite number of supporting hyperplanes? Can you find a relationship between the one-sided directional derivatives and the normal to the supporting hyperplanes at the points $x_0, x_1$?

2.4 Topological Properties of Convex Sets

Before going ahead into this section, the reader is kindly prompted to consult Section A.6, where we recall the basic definitions of: open sets, closed sets, closure and interior of a set, and boundary points. All these notions are essential for our complete understanding of convex sets and their properties.

We start this section by proving that the line segment joining a point in the interior of a convex set and a point in the closure of the set is contained in the interior of the set.

Theorem 2.4.4 Let $C \subset \mathbb{R}^n$ be a convex set with nonempty interior. Fix $x_1 \in \overline{C}$ and $x_2 \in C^o$. Then, for all $a \in (0, 1)$ we have

$$ax_1 + (1 - a)x_2 \in C^o.$$

Proof. Because $x_2 \in C^o$, there exists $r > 0$ such that $B(x_2, r) \subset C$. Fix $a \in (0, 1)$ and call $y := ax_1 + (1 - a)x_2$. We want to show that $y \in C^o$. In order to do this, we have to find $r_0 > 0$ such that $B(y, r_0) \subset C$. We claim that the latter holds for $r_0 := (1 - a)r$. Let us prove this claim. Take $z \in B(y, r_0)$. This means (see Section A.4) that

$$r_1 := r_0 - \|z - y\| = (1 - a)r - \|z - y\| > 0.$$

In particular, we get

$$\frac{1}{(1 - a)}\|z - y\| = r - \frac{r_1}{(1 - a)}.$$  (2.4.23)
Because \( x_1 \in C \) we know that 
\[ B(x_1, r_1) \cap C \neq \emptyset, \]
so we can take \( z_1 \in B(x_1, r_1) \cap C \). We get that
\[ \frac{a}{1-a} \| x_1 - z_1 \| < \frac{a r_1}{1-a}. \] (2.4.24)

Consider now the point \( z_2 := \frac{z - az_1}{1-a} \).

Then,
\[ \| z_2 - x_2 \| = \| \frac{z - az_1}{1-a} - x_2 \| = \| \frac{z - az_1 - (y - ax_1)}{1-a} \| \\
\leq \frac{1}{1-a} \| z - y + a(x_1 - z_1) \| \\
< r - \frac{r_1}{1-a} + \frac{a r_1}{1-a} = r - (1-a)r_1 < r, \]
where we used (2.4.23) and (2.4.24) in the second inequality. The above expression yields \( z_2 \in B(x_2, r) \subset C \) and hence \( x_2 \in C \). The definition of \( z_2 \) implies that
\[ z = az_1 + (1-a)z_2. \]

Since both \( z_1, z_2 \in C \) we conclude that \( z \in C \), as we wanted. \( \square \)

An important consequence of the previous theorem is the fact that the interior of a convex set is convex.

**Corollary 2.4.3** If \( C \subset \mathbb{R}^n \) is convex, then \( C^o \) is also convex.

**Proof.** Take \( x_1, x_2 \in C^o \) and \( a \in (0, 1) \). Because \( x_1 \in C^o \subset C \), we can apply Theorem 2.4.4 to the points \( x_1, x_2 \), where \( x_1 \in C \) and \( x_2 \in C^o \), to conclude that \( ax_1 + (1-a)x_2 \in C^o. \) \( \square \)

The closure of a convex set is also convex. The proof of this fact is much simpler. The following simple exercise will be used.

**Exercise 2.4.21** Let \( F \subset \mathbb{R}^n \). A point \( x \in \overline{F} \) if and only if there exists a sequence \( \{x_k\} \subset F \) such that \( x_k \to x \).

**Proposition 2.4.4** If \( C \subset \mathbb{R}^n \) is convex, then \( \overline{C} \) is also convex.

**Proof.** Fix \( x, y \in \overline{C} \) and \( a \in [0, 1] \). By Exercise 2.4.21 we know that there exist sequences \( \{x_k\}, \{y_k\} \subset C \) such that \( x_k \to x \) and \( y_k \to y \). Because \( x_k, y_k \in C \) for every \( k \), and \( C \) is convex by assumption, we must have that \( ax_k + (1-a)y_k \in C \). This fact, together with the assumptions on \( \{x_k\}, \{y_k\} \) yield \( ax + (1-a)y \in \overline{C} \). Therefore, \( \overline{C} \) is convex. \( \square \)

The next result proves that taking the closure of a convex set produces the same result as taking the closure of its interior.

**Corollary 2.4.4** Assume \( C \subset \mathbb{R}^n \) is convex and such that \( C^o \neq \emptyset \). Then \( \overline{C} = C^o \).

**Proof.** Because \( C^o \subset C \), it always holds that \( \overline{C^o} \subset \overline{C} \). So it is enough to show that \( \overline{C^o} \supset \overline{C} \). Take \( x \in \overline{C} \) and fix any \( y \in C^o \). Define now the sequence \( \{z_n\} \) as \( z_n := (1/n)x + (1 - (1/n))y. \) From Theorem 2.4.4 we know that \( z_n \in C^o \) for all \( n \). Moreover, the sequence \( \{z_n\} \subset C^o \) converges to \( x \). Using now Exercise 2.4.21 we conclude that \( x \in \overline{C^o} \), as we wanted. \( \square \)

**Corollary 2.4.5** Assume \( C \subset \mathbb{R}^n \) is convex and such that \( C^o \neq \emptyset \). Then \( (\overline{C})^o = C^o \).
Proof. Because $C \subset \overline{C}$, taking interior in both sides we obtain $C^o \subset (\overline{C})^o$. So it is enough to show that $(\overline{C})^o \subset C^o$. Take $x \in (\overline{C})^o$. We must show that $x \in C^o$. Because $x \in (\overline{C})^o$, there exists $r > 0$ such that $B(x, r) \subset C$. Take $x_0 \in C^o$ (recall that $C^o \neq \emptyset$). If $x_0 = x$ then $x \in C^o$ as we wanted. So assume that $x \neq x_0$. Take the point $z = x + \frac{r(x-x_0)}{2\|x-x_0\|}$. Then we have that $z \in B(x, r) \subset \overline{C}$. Moreover, it is a simple calculation to check that

$$x = \frac{\|x-x_0\|}{\|x-x_0\| + r/2} z + \frac{r/2}{\|x-x_0\| + r/2} x_0.$$

Because $z \in \overline{C}$ and $x_0 \in C^o$, we can apply Theorem 2.4.4 to conclude that $x \in C^o$. This completes the proof. □

Exercise 2.4.22 Let $V \subset \mathbb{R}^n$. Prove that $\partial V = \overline{V} \cap (V^o)^c$. Hint: use the definitions.

Proposition 2.4.5 Assume $C \subset \mathbb{R}^n$ is convex. Then $\partial C = \partial(\overline{C})$.

Proof. Using Exercise 2.4.22 for the sets $C$ and $\overline{C}$ yields the equalities $\partial C = \overline{C} \cap (C^o)^c$ and $\partial(\overline{C}) = \overline{C} \cap (C^o)^c$. Using Corollary 2.4.5 in the second equality yields $\partial(\overline{C}) = \overline{C} \cap (C^o)^c = \partial C$, where we also used the first equality. The proof is complete. □

2.5 Separation results

In this section we establish main separation properties of convex sets. In order to prove our first theorem, we need the following useful property concerning optimization problems in which the objective function is strictly convex. Consider an optimization problem of the form

$$(P) \quad \min_{x \in C} f(x),$$

where $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $C \subset \mathbb{R}^n$. Assume that $C \cap \text{dom} f \neq \emptyset$. The value

$$\inf_{x \in C} f(x) \in [-\infty, +\infty),$$

is called the optimal value of $(P)$. The set of solutions of problem $(P)$ is denoted $\text{Argmin}_C f$ and its elements are called minimizers or optimal solutions of $(P)$.

Proposition 2.5.6 Consider the optimization problem $(P)$ where where $f$ is strictly convex. Assume that $C \cap \text{dom} f \neq \emptyset$. Then $\text{Argmin}_C f$ has at most one element.

Proof. It is enough to prove that, if $\text{Argmin}_C f \neq \emptyset$, then $\text{Argmin}_C f$ has only one element. Assume that $\text{Argmin}_C f$ has two different elements $x, y$, and call $f^* := \inf_{x \in C} f(x)$ the optimal value of $(P)$. Because $C \cap \text{dom} f \neq \emptyset$ we know that $f^* < +\infty$. If $f^* = -\infty$, then we have $\text{Argmin}_C f = \emptyset$ ($f$ never attains the value $-\infty$). Therefore, we can assume that $f^* \in \mathbb{R}$. The fact that $x, y \in \text{Argmin}_C f$ yield $f^* = f(x) = f(y)$. Because $f$ is strictly convex, we can write

$$f\left(\frac{x + y}{2}\right) < (1/2) f(x) + (1/2) f(y) = (1/2) f^* + (1/2) f^* = f^*,$$

and therefore $f\left(\frac{x + y}{2}\right) < f^*$ with $\frac{x + y}{2} \in C$. This contradicts the definition of $f^*$ (we cannot have a point in $C$ with an objective value smaller than the optimal value). This contradiction implies that $\text{Argmin}_C f$ must have only one element. □

Exercise 2.5.23 Find a strictly convex function $f$ and a convex set $C$ such that problem $(2.7.31)$ has no solutions.
Our first result states that we can always “project” a point into a convex set.

**Theorem 2.5.5 (Projection theorem)** Let $C \subset \mathbb{R}^n$ be a nonempty closed and convex set. Fix $z \notin C$. Then,

(a) there exists a unique $\bar{x} \in C$ such that $\bar{x}$ is the closest element to $z$ in $C$, in other words, $\bar{x}$ is the unique solution of the optimization problem

$$(P) \quad \min_{x \in C} \|x - z\|^2,$$

(b) the point $\bar{x}$ is characterized by the inequality

$$\langle x - \bar{x}, z - \bar{x} \rangle \leq 0,$$

for all $x \in C$.

In this situation, the unique point $\bar{x}$ verifying (b) is called the projection of $z$ onto $C$ and it is denoted as $P_C(z)$.

**Proof.** The proof of (a) uses Weierstrass Theorem (see Theorem 2.7.12). Because $C$ is nonempty we can take a fixed element in $C$, say $\bar{x}$, and consider the set

$$\bar{C} := \{x \in C : \|x - z\| \leq \|\bar{x} - z\|\}.$$  

The set $\bar{C}$ is closed, convex and bounded (Exercise 2.5.24). Using also the fact that $f(x) := \|x - z\|^2$ is a continuous function, we can quote Weierstrass Theorem to conclude that the optimization problem

$$(P_1) \quad \min_{x \in \bar{C}} \|x - z\|^2,$$

has at least one solution. Moreover, by strict convexity of $f$ we can use Proposition 2.5.6 to conclude that the solution of $(P_1)$ is unique. In order to complete the proof of (a), we need to prove that problems $(P)$ and $(P_1)$ have the same set of solutions. Because $C \supset \bar{C}$, every solution of $(P)$ will also be a solution of $(P_1)$ (Exercise 2.5.25). Let us then prove the converse. Take $\bar{x}$ the unique solution of $(P_1)$. We have to check that

$$f(\bar{x}) \leq f(x),$$

for all $x \in C$. By definition of $\bar{x}$, the above inequality holds when $x \in \bar{C}$. Take now $x \notin \bar{C}$, this yields

$$f(x) = \|x - z\|^2 \geq \|\bar{x} - z\|^2 \geq \|\bar{x} - z\|^2 = f(\bar{x}),$$

where the second inequality is using that $\bar{x} \in \bar{C}$ and so $f(\bar{x}) \geq f(\bar{x})$. This proves our claim that problems $(P)$ and $(P_1)$ have the same set of solutions. Consequently, problem $(P)$ has a unique solution, given by $\bar{x}$. We must have $\bar{x} = x$. Now let us prove (b). We have to show that a given point $\hat{x}$ is the solution of $(P)$ if and only if the inequality in (b) holds for $\hat{x}$. Assume first that the inequality in (b) holds for some $\hat{x}$. Then, for every $x \in C$ we can write

$$\|z - x\|^2 = \|z - \hat{x}\|^2 + \|\hat{x} - x\|^2 + 2\langle z - \hat{x}, \hat{x} - x \rangle \geq \|z - \hat{x}\|^2 + \|\hat{x} - x\|^2 \geq \|z - \hat{x}\|^2,$$

where the inequality in (b) has been used in the first inequality. The above expression implies that $\hat{x}$ is the unique solution of problem $(P)$. Therefore we must have $\hat{x} = \bar{x}$. Now let us prove that, if $\bar{x}$ is
the solution of (P), then \( \bar{x} \) verifies the inequality in (b). Fix \( x \in C \) and take \( x' := \bar{x} + a(x - \bar{x}) \in C \), where \( a \in (0, 1) \). Because \( \bar{x} \) solves (P), we have that

\[
\|z - x'\|^2 = \|z - \bar{x} - a(x - \bar{x})\|^2 \geq \|z - \bar{x}\|^2,
\]

which can be re-arranged as

\[
a\|\bar{x} - x\|^2 - 2\langle z - \bar{x}, x - \bar{x} \rangle \geq 0.
\]

Letting \( a \to 0^+ \) we obtain \( \langle z - \bar{x}, x - \bar{x} \rangle \leq 0 \) for every \( x \in C \). This is precisely the inequality in (b). \( \square \)

Exercise 2.5.24 Prove that the set \( \tilde{C} \) used in Theorem 2.5.5 is closed, convex and bounded.

Exercise 2.5.25 With the notations of the proof of Theorem 2.5.5. Prove that (i) \( C \supset \tilde{C} \), and that (ii) every solution of (P) will also be a solution of (P1).

Definition 2.5.9 Let \( C_1, C_2 \subseteq \mathbb{R}^n \) be two nonempty sets. We say that a hyperplane \( H = H(a, \gamma) \) separates \( C_1 \) and \( C_2 \) whenever \( C_1 \subseteq H^+ \) and \( C_2 \subseteq H^- \). In other words, when

\[
\langle x, a \rangle \geq \gamma \quad \text{and} \quad \langle x', a \rangle \leq \gamma \quad \forall x \in C_1, \forall x' \in C_2.
\]

The separation above is said to be proper if \( C_1 \cup C_2 \not\subseteq H \). We say that a hyperplane \( H = H(a, \gamma) \) strongly separates \( C_1 \) and \( C_2 \) if there exists \( \varepsilon > 0 \) such that

\[
\langle x, a \rangle \leq \gamma < \gamma + \varepsilon \leq \langle x', a \rangle \quad \forall x \in C_1, \forall x' \in C_2.
\]

We say that a hyperplane \( H = H(a, \gamma) \) strictly separates \( C_1 \) and \( C_2 \) if

\[
\langle x, a \rangle < \gamma < \langle x', a \rangle \quad \forall x \in C_1, \forall x' \in C_2.
\]

The next separation theorem, which uses the previous result, is the most important separation result, because it generates all other separation results.

Theorem 2.5.6 (Separation of a closed set and a point not in the set) Let \( C \) be a closed and convex set and fix \( z \not\in C \). Then, there exists \( (a, \gamma) \in \mathbb{R}^n \times \mathbb{R} \), with \( \|a\| = 1 \), such that

\[
\langle a, x \rangle \leq \gamma < \langle a, z \rangle,
\]

for all \( x \in C \).

Proof. If \( C = \emptyset \), then (2.5.25) holds for some \( (a, \gamma) \in \mathbb{R}^n \times \mathbb{R} \) (otherwise, there would be an element in \( C \) for which (2.5.25) is not true). Assume that \( C \neq \emptyset \). Because \( C \) is closed and convex and \( z \not\in C \) we can apply the projection theorem (Theorem 2.5.5) and obtain the projection of \( \bar{x} \) onto the set \( C \), which by item (b) of the same theorem is characterized by the inequality

\[
\langle z - \bar{x}, x - \bar{x} \rangle \leq 0,
\]

for all \( x \in C \). Because \( \bar{x} \in C \), the set \( C \) is closed and \( z \not\in C \), we must have \( \bar{x} - z \neq 0 \) (is the same true when \( C \) is not closed?). Therefore, the unit vector (see the definition of unit vector in Section A.3) \( a := \frac{z - \bar{x}}{\|z - \bar{x}\|} \) is well defined. Defining also the real number \( \gamma := \langle \bar{x}, a \rangle \), and using the inequality (2.5.26) we obtain

\[
\langle a, x \rangle \leq \langle a, \bar{x} \rangle = \gamma,
\]

for all \( x \in C \). Hence, the left-hand side of (2.5.25) has been proved. In order to check the right-hand side of (2.5.25), use the definition of \( a \) and \( \gamma \) to obtain

\[
\langle a, z \rangle - \gamma = \left( \frac{z - \bar{x}}{\|z - \bar{x}\|}, z \right) - \langle \bar{x}, \frac{z - \bar{x}}{\|z - \bar{x}\|} \rangle
\]

\[
= \left( \frac{z - \bar{x}}{\|z - \bar{x}\|}, z - \bar{x} \right) = \|z - \bar{x}\| > 0,
\]
which establishes the right-hand side of (2.5.25). □

Recall that a boundary point $x$ is characterized by the fact that every ball around $x$ intersects both the set and its complement. See Definition A.6.21 for more details and properties of boundary points.

We prove next that a convex set has a supporting hyperplane at every boundary point. In the proof of the next result we use the following fact, regarding the boundary of a set.

**Theorem 2.5.7 (Supporting hyperplane at boundary points)** Let $C$ be a convex set and fix $\bar{x} \in \partial C$. Then, there exists $a \in \mathbb{R}^n$, with $\|a\| = 1$, such that
\[
\langle a, x - \bar{x} \rangle \leq 0,
\]
for all $x \in C$.

**Proof.** Using Proposition 2.4.5, we have that $\partial C = \partial(\overline{C})$, and hence $\bar{x} \in \partial(\overline{C})$. Using the definition of a boundary point (see Definition A.6.25) we have that $B[\bar{x}, 1/n] \cap (\overline{C})^c \neq \emptyset$, for every $n \in \mathbb{N}$. Take $y_n \in B[\bar{x}, 1/n] \cap (\overline{C})^c$ for every $n$. Then the sequence $\{y_n\}$ verifies that $y_n \notin \overline{C}$ and $y_n \to \bar{x}$. In other words, the sequence $\{y_n\}$ converges (see Section A.7) towards $\bar{x}$ from “outside” the set $\overline{C}$. Because $y_n \notin \overline{C}$ and $\overline{C}$ is a closed set, by Theorem 2.5.6 (applied to the set $\overline{C}$ and the point $y_n \notin \overline{C}$), there exists $a_n \in \mathbb{R}^n$ such that $\|a_n\| = 1$ and
\[
\langle a_n, x \rangle < \langle a_n, y_n \rangle,
\]
for all $x \in \overline{C}$. By Theorem A.7.43, there exists a convergent subsequence of $\{a_n\}$, with limit point $a$ such that $\|a\| = 1$. Using now (2.5.28) for that subsequence, and taking limits, we obtain $\langle a, x \rangle \leq \langle a, \bar{x} \rangle$ for all $x \in \overline{C}$. The proof is complete. □

**Corollary 2.5.6** Let $C$ be a convex set and assume $z \notin C^o$. Then, there exists $a \in \mathbb{R}^n$, with $\|a\| = 1$, such that
\[
\langle a, x - z \rangle \leq 0,
\]
for all $x \in C$.

**Proof.** If $z \notin C^o$, then either $z \in \partial C$, or $z \notin \overline{C}$. If $z \in \partial C$, then we can apply Theorem 2.5.7. If $z \notin \overline{C}$ then we can use Theorem 2.5.6. □

Now we are in conditions of separating two disjoint convex sets.

**Theorem 2.5.8 (Separation of two convex sets)** Let $C_1, C_2$ be two convex sets such that $C_1 \cap C_2 = \emptyset$. Then, there exists $a \in \mathbb{R}^n$, with $\|a\| = 1$, such that
\[
\langle a, x \rangle \leq \langle a, x' \rangle,
\]
for all $x \in C_1$ and all $x' \in C_2$. Equivalently,
\[
\sup_{x \in C_1} \langle a, x \rangle \leq \inf_{x' \in C_2} \langle a, x' \rangle.
\]

**Proof.** Consider the set $C := C_1 - C_2 := \{x - y : x \in C_1, y \in C_2\}$. From Exercise 2.1.2(b)(c), we know that $C$ is convex. Moreover, $0 \notin C$. In particular, $0 \notin C^o$. The conclusion of the theorem now follows from Corollary 2.5.6. □

If one of the sets is bounded, we can strongly separate them. In order to establish this stronger separation result, we need the following lemma.
Lemma 2.5.6 Assume that $C_1, C_2 \subset \mathbb{R}^n$ are closed sets, and suppose that $C_1$ is bounded. Then $C_1 - C_2$ is closed.

Proof. In order to show that $C_1 - C_2$ is closed, we will show that $\overline{C_1 - C_2} \subset C_1 - C_2$. Take $z \in C_1 - C_2$. From Exercise 2.4.21, we know that there exists a sequence $\{z_k\} \subset C_1 - C_2$ such that $z_k \to z$. Because $z_k \in C_1 - C_2$, there exist $x_k \in C_1$ and $y_k \in C_2$ such that $z_k = x_k - y_k$. Note that $\{x_k\} \subset C_1$ and $C_1$ is bounded and closed. By Theorem A.7.43 we know that there exists a subsequence $\{x_{k_j}\} \subset \{x_k\}$ which converges to some $x \in C_1$. Altogether,

$$y_{k_j} - x = (y_{k_j} - x_{k_j}) + (x_{k_j} - x) \to z + 0 = z,$$

and hence $y_{k_j} \to x - z$. Because $\{y_{k_j}\} \subset C_2$ and $C_2$ is closed we have that $x - z \in C_2$. Hence, $z = x - (x - z) \in C_1 - C_2$, as we wanted. This proves that $C_1 - C_2$ is closed. □

Theorem 2.5.9 (Strong separation of two convex sets) Let $C_1, C_2$ be two closed and convex sets such that $C_1 \cap C_2 = \emptyset$. Then, there exist $a \in \mathbb{R}^n$ and $\varepsilon > 0$, with $\|a\| = 1$, and such that

$$(a, x) + \varepsilon \leq \langle a, x' \rangle,$$

for all $x \in C_1$ and all $x' \in C_2$. Equivalently,

$$\sup_{x \in C_1} \langle a, x \rangle + \varepsilon \leq \inf_{x' \in C_2} \langle a, x' \rangle$$

Proof. Consider the set $C := C_1 - C_2$. As in the previous theorem we know that $0 \notin C$. Moreover, by Lemma 2.5.6, $C$ is closed. Now we can use Theorem 2.5.6 to conclude that there exists $(p, \varepsilon) \in \mathbb{R}^n \times \mathbb{R}$, with $\|p\| = 1$, such that

$$\langle p, x \rangle \geq \varepsilon > \langle a, 0 \rangle = 0,$$

(2.5.29)

for all $x \in C$. This implies that $\varepsilon > 0$ and using also the fact that $C = C_1 - C_2$ we get from the left-hand inequality in (2.5.29) that $\langle p, x_1 \rangle \geq \langle p, x_2 \rangle + \varepsilon$ for all $x_1 \in C_1$ and all $x_2 \in C_2$. The proof is complete. □

2.6 Subgradients and Subdifferentials

Definition 2.6.10 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. A vector $w \in \mathbb{R}^n$ is said to be a subgradient of $f$ at the point $x_0 \in \text{dom } f$ when

$$f(y) \geq f(x_0) + \langle w, y - x_0 \rangle,$$

(2.6.30)

for all $y \in \mathbb{R}^n$. The set of all such vectors is called the subdifferential of $f$ at $x_0$, and it is denoted as $\partial f(x_0)$. If $x_0 \notin \text{dom } f$ we define $\partial f(x_0) := \emptyset$.

Remark 2.6.8 When $f$ is differentiable at $x_0$, then Theorem 2.2.2 implies that $\partial f(x_0)$ is not empty and $\nabla f(x_0) \in \partial f(x_0)$. We will see below (see Corollary 2.6.7) that, in this situation, $\nabla f(x_0)$ is the only element of the set $\partial f(x_0)$. Even when the function is not differentiable at $x_0$, the subgradients play a similar role as the gradient plays for differentiable functions. In Figure 2.6.10 we see that each subgradient of $f$ at $x_0$ defines a linear approximation of $f$ at $x_0$ which stays below the function $f$ everywhere. Inequality (2.2.8) states this property for $\nabla f(x_0)$, and Corollary 2.6.7 proves that, when $f$ is differentiable, inequality (2.2.8) holds for a given vector if and only if that vector is $\nabla f(x_0)$. 
The following example introduces the concept of normal cone, which is the cone of all the directions that move “outwards” the set $C$, from the point $x_0 \in C$. Suppose that we want to optimize a function over the constraint set $C$. In this case it is essential to know whether $f$ grows or not when we move inwards $C$, from a given point $x_0$ in the boundary of $C$. For this reason, normal cones are instrumental for devising optimality conditions for convex optimization problems (see Theorem 2.10.26).

Exercise 2.6.26 Let $C \subset \mathbb{R}^n$ be a closed and convex set. Let $\delta C$ be the function defined in Exercise 2.1.6. Prove that

$$\partial \delta_C(x_0) = \begin{cases} \{ w \in \mathbb{R}^n : \langle w, y - x_0 \rangle \leq 0, \forall y \in C \} & \text{if } x_0 \in C, \\ \emptyset & \text{if } x_0 \not\in C. \end{cases}$$

Prove also that this set is a cone (see Exercise 2.1.2(j) for the definition of cone). The set $\partial \delta_C(x_0)$ is called normal cone of $C$ at the point $x_0$ and denoted by $N_C(x_0)$ (see Figure 2.6.11).

Exercise 2.6.27 Let $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be a convex function. Prove that $\partial f(x_0)$ is a closed and convex set. If $f$ is also continuous everywhere, then the set $\partial f(x_0)$ is bounded. Hint for the second statement: Assume that there is a sequence $\{w_k\} \subset \partial f(x_0)$ which is unbounded. Consider the sequence $\{y_k\}$ defined by $y_k := x_0 + \frac{w_k}{\|w_k\|}$. Check that $\{y_k\}$ is bounded. Use (2.6.30) for $y = y_k$. Take limits and obtain a contradiction using the fact that a continuous function maps bounded sets on bounded sets.

Exercise 2.6.28 Find a convex function $f : \mathbb{R} \to \mathbb{R}$ such that $\partial f(0) = [0, 1]$. Find a convex function $f : \mathbb{R} \to \mathbb{R}$ such that $\partial f(0) = [0, +\infty)$. Give an example of a function for which $\partial f(0) = \emptyset$.

Recall from Calculus that a vector $v$ is the gradient of $f$ at a point $x_0$ if and only if $(v, -1)$ is the normal to a hyperplane tangent to the graph of $f$ at the point $(x_0, f(x_0))$. Therefore, we can
“see” whether a function is differentiable or not by checking that its graph has a unique tangent plane at the point \((x_0, f(x_0))\). Subgradients have a similar geometric interpretation. A function \(f\) has a subgradient \(w\) at a point \(x_0\) if and only if the epigraph of \(f\) has a nonvertical supporting hyperplane at the boundary point \((x_0, f(x_0))\) (see Figure 2.6.12). This is a consequence of the fact that in this situation \((w, -1)\) is an (outward) normal to the epigraph of \(f\) at the boundary point \((x_0, f(x_0))\). Hence, a convex function is subdifferentiable (that is, \(f\) has a subgradient) at the point \(x_0\) if we can place a nonvertical supporting hyperplane at the point \((x_0, f(x_0))\) in \(\text{Epi } f\). This fact is formally proved next.

**Theorem 2.6.10** Let \(f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}\) be a convex function and let \(x_0 \in \text{dom } f\). The following statements are equivalent.

(a) \(\partial f(x_0) \neq \emptyset\),

(b) the set \(\text{Epi } f\) has a nonvertical supporting hyperplane at \((x_0, f(x_0))\).

**Proof.** Assume (a) holds and fix \(w \in \partial f(x_0)\). From (2.6.30) we can write for every \(y \in \text{dom } f\)

\[
0 \geq \langle w, y - x_0 \rangle + (-1)[f(y) - f(x_0)] = \langle (w, -1), (y - x_0, [f(y) - f(x_0)]) \rangle = \langle (w, -1), (y, f(y)) - (x_0, f(x_0)) \rangle.
\]

\(^1\)A hyperplane in \(\mathbb{R}^{n+1}\) is nonvertical when its normal \((v, t) \in \mathbb{R}^{n+1}\) is such that \(t \neq 0\).
Figure 2.6.12: Subdifferentiability and supporting hyperplanes to the epigraph.

Note that we are using here the fact that the scalar product is $\mathbb{R}^{n+1}$ has the form

$$\langle (u, a), (v, b) \rangle = \langle u, v \rangle + a \cdot b,$$

for all $(u, a), (v, b) \in \mathbb{R}^{n+1}$. Because for every $(y, \alpha) \in \text{Epi } f$ we have $\alpha \geq f(y)$ the above expression yields

$$0 \geq \langle (w, -1), (y, \alpha) - (x_0, f(x_0)) \rangle,$$

for every $(y, \alpha) \in \text{Epi } f$. This means that $\text{Epi } f$ is contained in the half-space $H^-$ defined by the hyperplane

$$H := \{ (x, t) \in \mathbb{R}^{n+1} : \langle (w, -1), (y, \alpha) - (x_0, f(x_0)) \rangle = 0 \}.$$

So $H$ is a supporting hyperplane of $\text{Epi } f$ at $(x_0, f(x_0))$. This hyperplane is clearly nonvertical because the last coordinate of its normal $(w, -1)$ is nonzero. Conversely, suppose there exists a nonvertical supporting hyperplane $H$ of $\text{Epi } f$ at $(x_0, f(x_0))$. Take a normal $(u, a) \in \mathbb{R}^{n+1}$ of this hyperplane. Because $H$ is nonvertical we know that $a \neq 0$. Without loss of generality assume that $\text{Epi } f \subset H^+$. This means that for every $(y, \alpha) \in \text{Epi } f$ we have

$$0 \leq \langle (u, a), (y, \alpha) - (x_0, f(x_0)) \rangle = \langle u, y - x_0 \rangle + a[\alpha - f(x_0)].$$

Because we can make $\alpha \to +\infty$ we must have $a > 0$ (we know that $a \neq 0$). Dividing the expression above by $a$ we get for every $(y, \alpha) \in \text{Epi } f$

$$0 \leq \langle (u/a), y - x_0 \rangle + [\alpha - f(x_0)].$$

In other words, for every $y \in \text{dom } f$ and every $\alpha \geq f(y)$ we can write

$$\alpha \geq f(x_0) + \langle (-u/a), y - x_0 \rangle.$$

The above expression holds, in particular, when $\alpha = f(y)$, so we get for every $y \in \text{dom } f$

$$f(y) \geq f(x_0) + \langle (-u/a), y - x_0 \rangle.$$
This yields \((-u/a) \in \partial f(x_0)\). The proof is complete. □

Recall from Calculus (and from (2.2.6)) that the directional derivative of a differentiable function is the projection of the gradient on the given direction. When the function \(f\) is convex, we can also relate the (one-sided) directional derivative with the subgradients of \(f\).

**Theorem 2.6.11** Let \(f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}\) be a convex function and let \(x_0 \in \text{dom } f\). The following statements are equivalent.

(a) \(w_0 \in \partial f(x_0)\),
(b) For every direction \(0 \neq v \in \mathbb{R}^n\) we have that \(f'(x_0, v)\) exists and \(f'(x_0, v) \geq \langle v, w_0 \rangle\).

**Proof.** Assume first that \(w_0 \in \partial f(x_0)\). So we can write, for every \(0 \neq v \in \mathbb{R}^n\)

\[
f'(x_0, v) = \lim_{t \to 0^+} \frac{f(x_0 + tv) - f(x_0)}{t} \geq \lim_{t \to 0^+} \frac{t(w_0, v)}{t} = \langle w_0, v \rangle,
\]
as wanted. Conversely, assume that (b) holds, and suppose that (a) is not true. This means that there exists \(\tilde{y} \in \text{dom } f\) such that \(f(\tilde{y}) < f(x_0) + \langle w_0, \tilde{y} - x_0 \rangle\). Note that in this case we must have \(\tilde{y} \neq x_0\). Define now the function \(q : (0, 1) \to \mathbb{R}\) as

\[
q(t) := \frac{f(x_0 + t(\tilde{y} - x_0)) - f(x_0)}{t}.
\]

As in the first part of the proof of Lemma 2.2.5, it holds that \(q\) is increasing on \((0, 1]\). This fact, together with (b) for \(v := \tilde{y} - x_0 \neq 0\), yields

\[
(w_0, \tilde{y} - x_0) \leq f'(x_0, \tilde{y} - x_0) = \lim_{t \to 0^+} q(t) = \inf_{t>0} q(t) \leq q(1) = f(\tilde{y}) - f(x_0) < \langle w_0, \tilde{y} - x_0 \rangle,
\]

where we used the assumption on \(\tilde{y}\) in the last inequality. The above expression entails a contradiction and therefore we must have \(w_0 \in \partial f(x_0)\). The proof is complete. □

**Corollary 2.6.7** Assume that \(f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}\) is convex and differentiable at \(x_0\). Then \(\partial f(x_0) = \{\nabla f(x_0)\}\).

**Proof.** From (2.2.8) and the definition of subdifferential we know that \(\nabla f(x_0) \in \partial f(x_0)\). Assume now that there exists \(w_0 \neq \nabla f(x_0)\) such that \(w_0 \in \partial f(x_0)\). Then for every \(0 \neq v \in \mathbb{R}^n\) we can write

\[
\langle \nabla f(x_0), v \rangle = f'(x_0, v) \geq \langle w_0, v \rangle,
\]

where the first equality follows from the fact that \(f\) is differentiable at \(x_0\) (and hence all directional derivatives exist and are obtained by projecting the given direction on the gradient), and the second equality is using Theorem 2.6.11. The above expression implies that

\[
0 \geq \langle w_0 - \nabla f(x_0), v \rangle,
\]

for every \(v \in \mathbb{R}^n\). Choosing \(v := w_0 - \nabla f(x_0) \neq 0\) we obtain \(w_0 = \nabla f(x_0)\), a contradiction. Therefore we must have \(\partial f(x_0) = \{\nabla f(x_0)\}\). The proof is complete. □

### 2.7 Existence of solutions of Optimization Problems

We focus here on some basic theoretical facts which guarantee existence of solutions of optimization problems.

Let \(f : \mathbb{R}^n \to \mathbb{R}\) be a continuous function and \(C \subset \mathbb{R}^n\). We say that \(x^*\) is a solution of the minimization problem

\[
\min_{x \in C} f(x),
\]

(2.7.31)
provided that \( x^* \in C \) and \( f(x^*) \leq f(x) \) for all \( x \in C \). In such a case, we say that a minimum exists. When \( C \) is the whole space; that is, \( C = \mathbb{R}^n \), we have an unconstrained problem. If \( C \) is a proper subset of \( \mathbb{R}^n \), we have a constrained problem.

We start by recalling the main result available for constrained problems.

**Theorem 2.7.12 (Weierstrass Theorem)** Let \( C \) be a nonempty, compact set, and let \( f : C \to \mathbb{R} \) be continuous on \( C \). Then, problem (2.7.31) has a solution.

**Proof.** We know by Theorem A.8.47 that the set \( f(C) \subset \mathbb{R} \) is compact, so it is a bounded set. In particular, there exists \( \alpha \in \mathbb{R} \) such that \( \alpha := \inf \{ f(x) \mid x \in C \} \). Define the sequence of sets \( \{C_k\} \) in the following way: \( C_k := \{ x \in C \mid \alpha \leq f(x) \leq \alpha + (1/k) \} \) for all \( k \in \mathbb{N} \). By definition of infimum, \( C_k \neq \emptyset \) and so there exists a sequence of points \( x_k \in C_k \) for all \( k \). So the sequence \( \{x_k\} \) is contained in \( C \) and by Bolzano-Weierstrass Theorem (see Theorem A.7.43) there exists a point \( x^* \in C \), which is an accumulation point of \( \{x_k\} \). Using now the fact that \( \alpha \leq f(x_k) \leq \alpha + (1/k) \) for all \( k \), the fact that there is a subsequence of \( \{x_k\} \) converging to \( x^* \), and continuity of \( f \), we get \( f(x^*) = \alpha \). Now the definition of \( \alpha \) implies that \( x^* \) is a minimizer of \( f \) in \( C \). \( \Box \)

**Remark 2.7.9** Note that Theorem 2.7.12 asserts also the existence of solutions for the maximization problem

\[
\max_{x \in C} h(x),
\]

where \( C \) is a compact set. Indeed, this is just a consequence of applying Theorem 2.7.12 to \( f = -h \).

We need more instances under which the conclusion of Theorem 2.7.12 is valid. In Theorem 2.7.12 we used the compactness of \( C \). When the set \( C \) is not compact, an alternative assumption can be posed on the objective function. This assumption is **coercivity**.

**Definition 2.7.11** We say that \( f : \mathbb{R}^n \to \mathbb{R} \) is coercive when \( \lim_{\|x\| \to \infty} f(x) = \infty \). In other words, a function \( f \) is coercive when the values of \( f \) tend to infinity whenever the norm of \( x \) tends to infinity.

Note that \( f \) is a coercive function if and only if for every \( M \in \mathbb{R} \) there exists \( R > 0 \) such that

\[
\|x\| > R \implies f(x) > M.
\]

As a consequence, the values of \( f \) cannot remain bounded over an unbounded set.

**Exercise 2.7.29** Prove that the following functions are coercive.

(i) \( f(x) = \sum_{i=1}^{n} x_i^2 = \|x\|^2 \).

(ii) \( f(x_1, x_2) = x_1^4 + x_2^4 - 3x_1x_2 \).

(iii) \( f(x) = e^{\|x\|^2} - (\|x\|^2)^{100} \).

Prove that the following functions are not coercive.

(i) \( f : \mathbb{R}^n \to \mathbb{R} \) defined by \( f(x) = c^T x \) (i.e., a linear function).

(ii) \( f(x_1, x_2) = x_1^2 + x_2^2 - 2x_1x_2 \).

(iii) \( f(x_1, x_2) = e^{x_1} + e^{x_2} - (x_1^2 + x_2^2)^{100} \).

**Example 2.7.4** Let \( A \in \mathbb{R}^{n \times n} \) be a symmetric matrix and take \( f(x) = x^TAx \). Prove that \( A \) is positive definite if and only if \( f \) is coercive. Hint: Use the fact that \( A \) is positive definite if and only if there exists \( \lambda > 0 \) such that \( x^TAx \geq \lambda \|x\|^2 \) for all \( x \in \mathbb{R}^n \). For the converse, use Theorem A.9.52. Prove that, if \( f \) is coercive, then all eigenvalues must be positive.
**Theorem 2.7.13** Let $C$ be a nonempty subset of $\mathbb{R}^n$ and let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuous function. Assume that one of the following two conditions holds:

(i) $C$ is closed and $f$ is coercive.

(ii) There exists $\alpha \in \mathbb{R}$ such that

$$S_f[\alpha] \cap C = \{ z \in C \mid f(z) \leq \alpha \},$$

is nonempty and compact.

Then, there exists $\bar{z} \in C$ such that $f(\bar{z}) = \inf_{z \in C} f(z)$. In other words, problem (2.7.31) has a solution.

**Proof.** Assume we are in case (i). Let $\{z^k\} \subset C$ be such that

$$\lim_{k \to \infty} f(z^k) = \inf_{z \in C} f(z).$$ \hfill (2.7.32)

Note that the right hand side of the expression above is not $+\infty$ because the set $C$ is nonempty. Since $f$ is coercive and the above limit is not $+\infty$, we can conclude that the sequence $\{z^k\}$ must be bounded. Therefore, by Theorem A.7.43 (Bolzano-Weierstrass Theorem), the sequence $\{z^k\}$ has a subsequence $\{z^{k_j}\}$ which is convergent to some $\bar{z}$. Since $C$ is closed and $\{z^{k_j}\} \subset C$ we must have $\bar{z} \in C$. Using now the continuity of $f$, we get $\lim_{j \to \infty} f(z^{k_j}) = f(\bar{z})$. Combining this with (2.7.32), we conclude that $\bar{z}$ is a solution of problem (2.7.31).

Assume now (ii). Note that in this case $\alpha \geq \inf_{z \in C} f(z)$. Indeed, if $\alpha < \inf_{z \in C} f(z)$, then the set in (ii) must be empty, a contradiction with our assumption. If $\alpha = \inf_{z \in C} f(z)$, then the set in (ii) is exactly the set of solutions of problem (2.7.31). Since by assumption this set is nonempty, we are done. Assume now that $\alpha > \inf_{z \in C} f(z)$, and consider a sequence $\{z^k\} \subset C$ as in (2.7.32). Note that, for $k$ sufficiently large, say for $k \geq k_0$, we must have $f(z^k) \leq \alpha$. By assumption (ii), for $k \geq k_0$, $\{z^k\}$ is bounded. The proof now proceeds like in part (i) and is left to the reader. $\square$

### 2.8 Optimality Conditions

Let $\mathbb{R}^n$ be the $n$-dimensional Euclidean space, and take $f$ a real-valued function defined on $\mathbb{R}^n$. The **Unconstrained Minimization problem** can be formally stated as

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad x \in \mathbb{R}^n.
\end{align*}
\]  \hfill (2.8.33)

The function $f$ is called the **objective function** or **cost function**. To solve the problem above consists of finding $\bar{x}$ such that

$$f(\bar{x}) \leq f(x),$$  \hfill (2.8.34)

for all $x \in \mathbb{R}^n$. Such a point is called a **global minimizer** or **global solution**. A global minimizer is said to be **strict** when the inequality above is strict for every $x \neq \bar{x}$. Property (2.8.34) is merely algebraic and, in principle, very difficult to verify, since in order to check it, we need to compare the values of $f(\bar{x})$ and $f(x)$ for all $x \in \mathbb{R}^n$! A less-demanding requirement is to find a point $\bar{x}$ for which inequality (2.8.34) is true for all points $x$ in a given neighborhood of $\bar{x}$. This point is called a **local minimizer**. More precisely, $\bar{x}$ is a **local minimizer** when there exists $r > 0$ such that

$$f(\bar{x}) \leq f(x),$$  \hfill (2.8.35)

for all $x \in B(\bar{x}, r)$. In Figure 2.8.13, $x_1$ is a (nonstrict) local minimizer, $x_2$ is a strict local minimizer and $x_3$ is a global (and also strict) minimizer.
A linear programming problem with a polyhedron for a feasible set falls into the category of continuous problems. However, it has some combinatorial features as well. One of this combinatorial features is asserted by the Fundamental Theorem of Linear Programming: The solutions of the Linear Program can be found by searching among the finite set of vertices of the polyhedron representing the feasible region. This is the basic idea of the SIMPLEX Method. On the other hand, the nonlinear nature of this problem allows to develop nonlinear programming approaches as well, giving rise to Interior Point Methods.

2.8.1 Necessary Optimality Conditions

The way to find a solution of (2.8.34), is to determine a set of “candidates”, which contains all possible solutions of the problem. Within this set, we have to search for those points which are true solutions. The set of “candidates” is obtained using the so-called Necessary Optimality Conditions. These conditions are called necessary because they are forcibly satisfied by every minimum, but a point that satisfies them is not necessarily a minimum.

Among the necessary optimality conditions the most important one is known as Fermat necessary condition. This optimality condition is based in the simple fact that, when the functional value at $x^*$ is (locally) maximum or minimum, then all directional derivatives at this point must be zero. Indeed, if there exists one directional derivative which is positive (or negative), we should have an increase (or decrease) of the function along this direction. This contradicts the fact that $f$ has a local extreme at $x^*$.

**Definition 2.8.12** Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable on an open set $D \subset \mathbb{R}^n$. A point $x^* \in D$ is called a stationary or critical point of $f$ when $\nabla f(x^*) = 0$. A point satisfying the latter equality is said to verify first order optimality conditions for problem (2.8.33).

In order to state our optimality conditions, we need Taylor’s Formula.

**Theorem 2.8.14 (Taylor’s Formula)** Fix $x^*, x \in \mathbb{R}^n$ and assume that $f : \mathbb{R}^n \to \mathbb{R}$ has continuous first and second partial derivatives on some open set containing the line segment $[x, x^*]$. Then there exists $z \in [x, x^*]$ such that

$$f(x) = f(x^*) + \nabla f(x^*)^T (x - x^*) + \frac{1}{2} (x - x^*)^T \nabla^2 f(x^*)(x - x^*) + o(||x - x^*||^2),$$

where $\lim_{x \to x^*} \frac{o(||x - x^*||^2)}{||x - x^*||^2} = 0.$
The next result states that every solution of (2.8.34) (and hence also (2.8.35)) is a stationary point.

**Theorem 2.8.15 (Fermat necessary optimality conditions)** Assume that \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is continuously differentiable on an open set \( D \subset \mathbb{R}^n \). Then,

(i) If \( x^* \in D \) is a local minimizer of \( f \), then it must verify \( \nabla f(x^*) = 0 \).

(ii) If \( f \) is twice continuously differentiable, then we also have \( \nabla^2 f(x^*) \) is positive semidefinite.

**Proof.** (i) Assume that \( x^* \) is a local minimum of \( f \) and suppose that \( \nabla f(x^*) \neq 0 \). Then there exists \( v \neq 0 \) such that \( \nabla f(x^*)^T v < 0 \). By continuity of \( \nabla f(\cdot) \) there exists \( r > 0 \) such that \( \nabla f(z)^T v < 0 \) for every \( z \in B(x^*, r) \). Fix now \( \lambda \in (0, r/\|v\|) \). Then \( x^* + tv \in B(x^*, r) \) for all \( t \in (0, \lambda) \). So \( \nabla f(x^* + tv)^T v < 0 \) for all \( t \in (0, \lambda) \). By the mean value theorem for \( f : \mathbb{R}^n \rightarrow \mathbb{R} \), there exists \( t \in (0, \lambda) \) such that

\[
 f(x^* + \lambda v) = f(x^*) + \lambda \nabla f(x^*)^T v < f(x^*).
\]

Since \( x^* + \lambda v \in B(x^*, r) \), the above inequality contradicts the fact that \( f(z) \geq f(x^*) \) for every \( z \in B(x^*, r) \). Hence we must have \( \nabla f(x^*) = 0 \). For proving (ii), we use (2.8.36), the equality \( \nabla f(x^*) = 0 \), and the fact that \( x^* \) is an unconstrained minimizer to obtain

\[
 f(x^*) \leq f(x) = f(x^*) + \frac{1}{2}(x - x^*)^T \nabla^2 f(x^*)(x - x^*) + o(\|x - x^*\|^2) \quad \text{for every } x \in \mathbb{R}^n,
\]

where \( \lim_{x \rightarrow x^*} \frac{o(\|x - x^*\|^2)}{\|x - x^*\|^2} = 0 \). So we get for every \( x \in \mathbb{R}^n \)

\[
 0 \leq f(x) - f(x^*) = \|x - x^*\|^2(\frac{1}{2\|x - x^*\|^2}(x - x^*)^T \nabla^2 f(x^*)(x - x^*) + \frac{o(\|x - x^*\|^2)}{\|x - x^*\|^2}).
\]

Assume, for contradiction purposes, that for some \( d \in \mathbb{R}^n \) we have \( d^T \nabla^2 f(x^*)d < 0 \). Let \( t \in \mathbb{R} \) and consider \( x(t) := td + x^* \). Using \( x := x(t) \) in the above expression we obtain

\[
 0 \leq f(x(t)) - f(x^*) = t^2\|d\|^2(\frac{1}{2}d^T \nabla^2 f(x^*)d + \frac{o(t^2\|d\|^2)}{t^2}).
\]

The properties of the function \( o(\cdot) \) imply that for \( t \) small enough, the expression between brackets will be negative (because \( d^T \nabla^2 f(x^*)d < 0 \)), which contradicts the left hand side of the expression. \( \square \)

Fermat necessary conditions (Theorem 2.8.15) are sufficient for convex differentiable problems.

**Corollary 2.8.8** Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) be differentiable and convex. Then \( x^* \) is a stationary point of \( f \) if and only if it is a global minimum.

**Proof.** If \( x^* \) is a global minimum, then it is a stationary point by the previous theorem. Conversely, assume that \( \nabla f(x^*) = 0 \), by (2.2.8) we get

\[
 f(y) \geq f(x^*) + \langle \nabla f(x^*), y - x \rangle = f(x^*),
\]

and hence \( x^* \) is a global minimizer. \( \square \)

We will see that this result can be extended to nondifferentiable convex functions.
2.8.2 Sufficient Optimality Conditions

We point out that a stationary point can be a minimum, a maximum, or neither. Only when the function is known to be convex, we can say that the point is in fact a global minimum. In Figure 2.8.13, the points $x_4$ and $x_5$ are stationary points, which are maximum, while $x_6$ is a stationary point which is an inflection point (a point in which the sign of the second derivative changes). Without convexity, we can only determine the nature of a stationary point when the function is twice continuously differentiable. In this case, we need to use the second order necessary conditions. The proof of the result below follows directly from Taylor’s Formula and the definition of stationary point.

**Theorem 2.8.16** Suppose that $x^*$ is a stationary point of $f : \mathbb{R}^n \to \mathbb{R}$. Assume also that $f$ is twice continuously differentiable on $\mathbb{R}^n$. Then

(a) If $(x-x^*)^T \nabla^2 f(z)(x-x^*) \geq 0$ for all $x \in \mathbb{R}^n$ and all $z \in [x, x^*]$, then $x^*$ is a global minimizer.

(b) If $(x-x^*)^T \nabla^2 f(z)(x-x^*) > 0$ for all $x \in \mathbb{R}^n$ such that $x \neq x^*$ and all $z \in [x, x^*]$, then $x^*$ is a strict global minimizer.

(c) If $(x-x^*)^T \nabla^2 f(z)(x-x^*) \leq 0$ for all $x \in \mathbb{R}^n$ and all $z \in [x, x^*]$, then $x^*$ is a global maximizer.

(d) If $(x-x^*)^T \nabla^2 f(z)(x-x^*) < 0$ for all $x \in \mathbb{R}^n$ such that $x \neq x^*$ and all $z \in [x, x^*]$, then $x^*$ is a strict global maximizer.

As we see, the sign of $(x-x^*)^T \nabla^2 f(z)(x-x^*)$ is important for determining whether a stationary point is a maximum or a minimum. This sign is difficult to know a priori, unless the matrix $\nabla^2 f(z)$ has special properties. A main tool for determining whether $\nabla^2 f(z)$ is positive semidefinite or not is Theorem A.9.53.

Combining Theorem 2.8.16 with Definition A.9.37, we obtain sufficient optimality conditions. In items (c) and (c') below, we are using the fact that, when $\nabla^2 f(x^*)$ is positive (or negative) definite, then the same holds for $\nabla^2 f(z)$, where $z$ is in some neighborhood of $x^*$.

**Theorem 2.8.17 (Second order sufficient optimality conditions)** Suppose that $x^*$ is a stationary point of $f : \mathbb{R}^n \to \mathbb{R}$. Assume also that $f$ has continuous first and second partial derivatives on $\mathbb{R}^n$. Then

(a) If $\nabla^2 f(z)$ is positive semidefinite for all $z \in \mathbb{R}^n$, then $x^*$ is a global minimizer.

(b) If $\nabla^2 f(z)$ is positive definite for all $z \in \mathbb{R}^n$, then $x^*$ is a strict global minimizer.

(c) If $\nabla^2 f(x^*)$ is positive definite, then $x^*$ is a strict local minimizer.

(a') If $\nabla^2 f(z)$ is negative semidefinite for all $z \in \mathbb{R}^n$, then $x^*$ is a global maximizer.

(b') If $\nabla^2 f(z)$ is negative definite for all $z \in \mathbb{R}^n$, then $x^*$ is a strict global maximizer.

(c') If $\nabla^2 f(x^*)$ is negative definite, then $x^*$ is a strict local maximizer.

Corollary 2.2.1 and item (a) above provide an alternative proof of Theorem 2.8.8 when $f$ is twice continuously differentiable.

**Exercise 2.8.30** For the following functions, find all stationary points, and determine whether they are (local, global, strict) maximum, minimum, or neither.

(i) $f(x_1, x_2, x_3) = x_1^2 + x_2^2 + x_3^2 - x_1x_2 + x_2x_3 - x_1x_3 - x_1 + x_2,$

(ii) $f(x_1, x_2, x_3) = e^{x_1-x_2} + e^{x_2-x_1} + e^{x_1^2 + \frac{1}{2}x_3^2},$
Prove that this function cannot have global extrema.

2.8.3 Saddle Points

Recall that a horizontal point of inflection of a real-valued function is a critical point at which the second derivative changes sign. In the case of a function of several variables, these points are called saddle points, and are stationary points at which the Hessian is indefinite. Indeed, assume that \( f \in \mathcal{C}^2 \) and suppose that \( x^* \) is a stationary point such that \( \nabla^2 f(x^*) \) is indefinite. Then by Definition A.9.37, there exists \( x, y \in \mathbb{R}^n \) such that \( x^T \nabla^2 f(x^*) x > 0 \) and \( y^T \nabla^2 f(x^*) y < 0 \). Using the continuity of the second derivatives of \( f \) we conclude that there exists \( r > 0 \) such that

\[
x^T \nabla^2 f(x^* + tx) x > 0 \quad \text{and} \quad y^T \nabla^2 f(x^* + ty) y < 0,
\]

for all \( t \) with \(|t| < r\). Define now \( \varphi(t) := f(x^* + tx) \) and \( \psi(t) := f(x^* + ty) \). Since \( x^* \) is a stationary point, we must have \( \varphi'(0) = 0 \) and \( \psi'(0) = 0 \). Also, \( \varphi''(0) = x^T \nabla^2 f(x^* + tx) x > 0 \) and \( \psi''(0) = y^T \nabla^2 f(x^* + ty) y < 0 \). This readily implies that 0 is a strict local minimum of \( \varphi \) and a strict local maximum of \( \psi \). Therefore, if we move from \( x^* \) in the directions \( x \) or \( -x \), the values of \( f \) increase, and if we move from \( x^* \) in the directions \( y \) or \( -y \), the values of \( f \) decrease. This is the reason for which such a point is called a saddle-point. Let us define formally this concept.

**Definition 2.8.13** A point \( x^* \) is a saddle point of \( f : \mathbb{R}^n \to \mathbb{R} \) when there exist two vectors \( x, y \) such that the function \( \varphi(t) := f(x^* + tx) \) has at \( t = 0 \) a strict local minimizer and \( \psi(t) := f(x^* + ty) \) has at \( t = 0 \) a strict local maximizer.

The discussion preceding the definition is now formally stated.

**Theorem 2.8.18** Assume that \( f \in \mathcal{C}^2 \) and \( x^* \) is a stationary point such that \( \nabla^2 f(x^*) \) is indefinite, then \( x^* \) is a saddle point of \( f \).

**Exercise 2.8.32** Consider the function \( f(x_1, x_2, x_3) = a x_1^2 + x_1 + b x_2^3 + x_2 + c x_3^2 + x_3 \). Choose values of \( a, b, c \in \mathbb{R} \) not simultaneously zero such that:

(i) \( f \) has no saddle points, and two stationary points, which are both minima.

(ii) \( f \) has exactly one saddle point and one minimum.

Prove that this function cannot have global extrema.

**Exercise 2.8.33** Consider the function \( f(x_1, x_2) = \frac{x_1^2}{4} + \frac{x_2^4}{6} - x_1 x_2 \). Compute all its critical points, and determine whether they are maximum, minimum or saddle points. In the case of extreme points, say whether they are local, global, strict or not strict.

Combining the preceding results, we see that a given critical point is an extreme point when the Hessian has a defined sign, or a saddle point when the Hessian is indefinite. When the Hessian has a semi-defined sign (i.e., it is positive or negative semi-definite), there is nothing we can say. The following exercise illustrates this situation.
Exercise 2.8.34 Consider the function \( f(x_1, x_2) = (x_1 - x_2^2)(x_1 - 2x_2^2) \). Verify that

(i) \( x^* = (0, 0) \) is the unique stationary point.

(ii) The Hessian is positive semi definite at \( x^* \), and \( x^* \) not a local minimum. (Hint: follow the values of \( f \) along the curve \( (t, \sqrt{\frac{2t}{3}}) \))

(iii) \( x^* \) is a local minimum of the function \( \varphi_d(t) := f(td) \), for all \( 0 \neq d \in \mathbb{R}^n \).

Consider now the function \( f(x_1, x_2) = (x_1 - x_2^2)^2 \). Verify that

(i) \( f \) has infinite stationary points.

(ii) The Hessian is positive semi definite at every stationary point.

(iii) All stationary points of \( f \) are global minima.

2.9 Optimality conditions for Constrained Problems

2.9.1 Equality Constraints

According to the structure of the set \( C \), we can state different kinds of Optimality Conditions. The simplest case is the one in which \( C \) is given by equality constraints:

\[
C := \{ x \in \mathbb{R}^n \mid h_1(x) = 0, \ldots, h_m(x) = 0 \},
\]

where \( h_1, \ldots, h_m : \mathbb{R}^n \to \mathbb{R} \) are continuously differentiable functions. With this set of constraints, our problem becomes

\[
\begin{align*}
\text{Minimize} & \quad f(x) \\
\text{subject to} & \quad h_i(x) = 0, \quad i = 1, \ldots, m.
\end{align*}
\]

(2.9.37)

We assume that the functions \( f, h_1, \ldots, h_m : \mathbb{R}^n \to \mathbb{R} \) are continuously differentiable. For convenience, denote by

\[
h(x) := \begin{pmatrix}
    h_1(x) \\
    \vdots \\
    h_m(x)
\end{pmatrix},
\]

the function from \( \mathbb{R}^n \) to \( \mathbb{R}^m \) which has each \( h_i(x) \) as a coordinate. The gradient matrix of \( h \) is the matrix given by

\[
\nabla h(x) := \begin{bmatrix}
    \nabla h_1(x) & \cdots & \nabla h_m(x)
\end{bmatrix} \in \mathbb{R}^{n \times m},
\]

while the Jacobian matrix of \( h \), is the matrix given by

\[
J_h(x) := \begin{pmatrix}
    \nabla h_1(x)^T \\
    \vdots \\
    \nabla h_m(x)^T
\end{pmatrix} \in \mathbb{R}^{m \times n},
\]

so \( J_h(x)^T = \nabla h(x) \).

In order to prove our main theorem, we need the following tools. Given a fixed vector \( \bar{x} \), define the subspace

\[
V(\bar{x}) := \{ y \in \mathbb{R}^n \mid \nabla h_i(\bar{x})^T y = 0, \quad i = 1, \ldots, m \}.
\]

The subspace \( V(\bar{x}) \) is formed by all directions which are tangent to the constraint set at the point \( \bar{x} \). Therefore, it is called the subspace of tangent feasible directions at the point \( \bar{x} \). In Figure 2.9.14, the points \( \bar{x}, \hat{x} \) verify \( h(\bar{x}) = h(\hat{x}) = 0 \), where \( h(x) = (h_1(x), h_2(x))^T \). The gradients at \( \bar{x} \) are linearly independent, and hence (because we are in \( \mathbb{R}^2 \)), \( V(\bar{x}) = \{(0, 0)\} \). The gradients at \( \hat{x} \) are linearly dependent, so \( V(\hat{x}) \neq \{(0, 0)\} \).
Exercise 2.9.35 Consider the function $h : \mathbb{R}^3 \to \mathbb{R}^3$ defined by $h(x_1, x_2, x_3) = (x_1^3 - x_2 + x_3^2, x_2, x_1 + x_2 + x_3)$. Compute $V(0, 0, 0)$ and $V(-1, 0, 1)$.

A point $x$ is called regular for problem (2.9.37) when the gradients $\{\nabla h_1(x), \ldots, \nabla h_m(x)\} \subset \mathbb{R}^n$ are linearly independent. This condition forces $m \leq n$.

Theorem 2.9.19 (Lagrange Multiplier Theorem-Necessary Optimality Conditions) Assume that the functions $f, h_1, \ldots, h_m : \mathbb{R}^n \to \mathbb{R}$ are continuously differentiable. Let $x^*$ be a regular point which is a local minimizer of problem (2.9.37). Then there exists a unique vector $\lambda^* \in \mathbb{R}^m$, called Lagrange Multiplier vector, such that

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) = 0.$$  \hfill (2.9.38)

If the functions $f, h_1, \ldots, h_m : \mathbb{R}^n \to \mathbb{R}$ are twice continuously differentiable, we have that

$$y^t \left( \nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla^2 h_i(x^*) \right) y \geq 0,$$  \hfill (2.9.39)

for all $y \in V(x^*)$.

Proof. We will approximate the original problem (2.9.37) by a sequence of unconstrained problems, in such a way that the solutions of these problems converge to the solution $x^*$ of our original problem. Let $F_k : \mathbb{R}^n \to \mathbb{R}$ be the function defined by

$$F_k(x) := f(x) + \frac{k}{2} ||h(x)||^2 + \frac{\alpha}{2} ||x - x^*||^2,$$

where $x^*$ is a local minimizer of problem (2.9.37) and $\alpha$ is an arbitrary positive number. So take $r > 0$ such that $f(y) \geq f(x^*)$ for all $y \in B[x^*, r]$ satisfying the equality constraints $h(y) = 0$. Now let $x^k$ be a solution of the auxiliary problem

$$\min F_k(x)$$

s.t. $x \in B[x^*, r].$ \hfill (2.9.40)

By Theorem 2.7.12, this problem always has a solution. Our first step is to show that the sequence $\{x^k\}$ converges to $x^*$. By definition of $x^k$, we have that

$$F_k(x^k) = f(x^k) + \frac{k}{2} ||h(x^k)||^2 + \frac{\alpha}{2} ||x^k - x^*||^2 \leq F_k(x^*) = f(x^*),$$  \hfill (2.9.41)
where we are also using the fact that \( h(x^*) = 0 \). The sequence \( \{x^k\} \) is contained in the compact set \( B[x^*, r] \), so by Theorem A.7.43 it has at least one accumulation point \( \bar{x} \in B[x^*, r] \). Let \( \{x^{k_j}\} \subset \{x^k\} \) be a subsequence converging to \( \bar{x} \). We claim that

\[
\lim_{j \to \infty} \|h(x^{k_j})\| = 0. \tag{2.9.42}
\]

Indeed, we know by continuity of \( h \) that the above limit is \( \|h(\bar{x})\| \). Assume, for contradiction purposes, that \( \|h(\bar{x})\| > 0 \). In this case, we must have

\[
\lim_{j \to \infty} \frac{k_j}{2} \|h(x^{k_j})\| = +\infty. \tag{2.9.43}
\]

Note that the numerical sequences \( \{f(x^{k_j})\} \) and \( \{\|x^{k_j} - x^*\|^2\} \) converge to \( f(\bar{x}) \) and \( \|\bar{x} - x^*\|^2 \) respectively. Using this fact in (2.9.41) for \( k = k_j \) and taking limits for \( j \to \infty \) we conclude that if (2.9.43) is true then the left-hand side of (2.9.41) must converge to \( +\infty \), a contradiction with the fact that it is bounded above by \( f(x^*) \). So our claim is true and (2.9.42) holds. As a consequence, the accumulation point \( \bar{x} \) satisfies \( \bar{x} \in B[x^*, r] \) and \( h(\bar{x}) = 0 \), and using the fact \( x^* \) is a solution over \( B[x^*, r] \) we conclude that

\[
f(x^*) \leq f(\bar{x}). \tag{2.9.44}
\]

From (2.9.41), we get

\[
f(x^k) + \frac{\alpha}{2} \|x^k - x^*\|^2 \leq f(x^*).
\]

Using this inequality for \( k = k_j \) and taking limits for \( j \to \infty \), we conclude that

\[
f(\bar{x}) + \frac{\alpha}{2} \|\bar{x} - x^*\|^2 \leq f(x^*) \leq f(\bar{x}),
\]

where we are also using (2.9.44). This readily implies that \( \bar{x} = x^* \) and hence the whole sequence \( \{x^k\} \) converges to \( x^* \). So for large enough \( k \), the sequence \( \{x^k\} \) will be in the interior of the ball \( B[x^*, r] \). Therefore, for large enough \( k \), \( x^k \) is an unconstrained minimizer of problem (2.9.40). So the necessary conditions given in Theorem 2.8.15 hold for \( F_k \) at \( x^k \). In other words,

\[
\nabla F_k(x^k) = 0 \quad \text{and} \quad \nabla^2 F_k(x^k) \text{ is positive semi-definite.} \tag{2.9.45}
\]

The first statement in (2.9.45) means that

\[
\nabla f(x^k) + k \nabla h(x^k)h(x^k) + \alpha(x^k - x^*) = 0. \tag{2.9.46}
\]

Since the sequence \( \{x^k\} \) converges to \( x^* \) and the gradient matrix \( \nabla h(x^*) \) has rank \( m \), we must have that \( \nabla h(x^*) \) also has rank \( m \) for large enough \( k \). This fact, together with Exercise A.9.68, implies that the matrix \( \nabla h(x^k)\nabla h(x^k)\) is positive definite for large enough \( k \). In particular, because \( \nabla h(x^k)^T \nabla h(x^k) \) is invertible the matrix \( (\nabla h(x^k)^T \nabla h(x^k))^{-1} \nabla h(x^k)^T \) is well-defined. Multiplying by this matrix the left-hand side of the above equality and rearranging the resulting expression, we get

\[
kh(x^k) = -((\nabla h(x^k)^T \nabla h(x^k))^{-1} \nabla h(x^k)^T) \left( \nabla f(x^k) + \alpha(x^k - x^*) \right).
\]

By taking limit in this expression for \( k \to \infty \), we see that

\[
\lim_{k \to \infty} kh(x^k) = -((\nabla h(x^*)^T \nabla h(x^*))^{-1} \nabla h(x^*)^T) \nabla f(x^*) =: \lambda^*.
\]

Using the definition of \( \lambda^* \) and taking limits in (2.9.46) we get

\[
\nabla f(x^*) + \nabla h(x^*)\lambda^* = 0.
\]
This proves the existence of a Lagrange multiplier vector verifying (2.9.38). In order to prove (2.9.39), we need to use the second order information in (2.9.45). So we know that for large enough \( k \) the matrix

\[
\nabla^2 F_k(x^k) = \nabla^2 f(x^k) + k \sum_{i=1}^{m} \nabla^2 h_i(x^k)h_i(x^k) + k\nabla h(x^k)\nabla h(x^k)^t + \alpha I,
\]

is positive semi-definite for all \( \alpha > 0 \). Now fix \( y \in V(x^*) \). Define the vectors

\[
y^k := y - \nabla h(x^k)(\nabla h(x^k)^t\nabla h(x^k))^{-1}\nabla h(x^k)^ty.
\]

It is a matter of simple algebra to check that \( y^k \in V(x^k) \). We claim now that the sequence \( \{y^k\} \) converges to \( y \). Indeed,

\[
\lim_{k \to \infty} y^k = y - \nabla h(x^*)(\nabla h(x^*)^t\nabla h(x^*))^{-1}\nabla h(x^*)^ty = y,
\]

where we are using the fact that \( y \in V(x^*) \), which means that \( \nabla h(x^*)^ty = 0 \). So the claim is true and \( \{y^k\} \) converges to \( y \). Since \( \nabla h(x^k)^ty^k = 0 \) we can write

\[
0 \leq y^k \nabla^2 F_k(x^k)y^k = y^k \left( \nabla^2 f(x^k) + \sum_{i=1}^{m} \nabla^2 h_i(x^k)[kh(x^k)]_i \right) y^k + \alpha\|y^k\|^2.
\]

By definition of \( \lambda^* \), we have that \( \lim_{k \to \infty} kh_i(x^k) = \lambda_i^*_k \). So, taking limits we get

\[
0 \leq y^t \left( \nabla^2 f(x^*) + \sum_{i=1}^{m} \nabla^2 h_i(x^*)\lambda_i^* \right) y + \alpha\|y\|^2,
\]

for all \( y \in V(x^*) \). But this inequality holds for all \( \alpha > 0 \). So we conclude that

\[
0 \leq y^t \left( \nabla^2 f(x^*) + \sum_{i=1}^{m} \nabla^2 h_i(x^*)\lambda_i^* \right) y,
\]

for all \( y \in V(x^*) \), as we wanted to prove. \( \square \)

**Remark 2.9.10** Observe the analogy between the optimality conditions presented in Theorem 2.8.15 and those stated in Theorem 2.9.19. The latter one has also first and second order conditions, but the constraints are now incorporated in those conditions. More precisely, the Lagrangian function \( \mathcal{L} : \mathbb{R}^{n+m} \to \mathbb{R} \) associated with problem (2.9.37) is defined as

\[
\mathcal{L}(x, \lambda) := f(x) + \sum_{i=1}^{m} \lambda_i h_i(x).
\]

With this definition, Theorem 2.9.19 asserts that: If \( x^* \) is a regular point and a local solution of (2.9.37), then we must have

\[
\nabla_x \mathcal{L}(x^*, \lambda^*) = 0, \quad \nabla_\lambda \mathcal{L}(x^*, \lambda^*) = h(x^*) = 0,
\]

which are the first order conditions on the Lagrangian, as well as

\[
y^t \nabla^2_{xx} \mathcal{L}(x^*, \lambda^*) y \geq 0, \quad \text{for all} \ y \in V(x^*),
\]

which are the second order conditions on the Lagrangian, imposed on the subspace \( V(x^*) \). The first order conditions represent \( n + m \) equations with \( n + m \) unknowns, the coordinates \( x^* \) and \( \lambda^* \). In this sense, the Lagrangian represents an elegant way of transforming our constrained \( n \)-dimensional problem, into an unconstrained \( n + m \)-dimensional one. So, we simplify the structure of the problem, at the cost of adding more variables to it.
We state below the second order sufficient optimality conditions for the equality constrained problem.

**Theorem 2.9.20 (Second Order Sufficient Conditions-Equality Constraints)** Assume that the functions \( f, h_1, \ldots, h_m : \mathbb{R}^n \to \mathbb{R} \) are twice continuously differentiable and let \( x^* \in \mathbb{R}^n \) and \( \lambda^* \in \mathbb{R}^m \) be such that
\[
\nabla_x L(x^*, \lambda^*) = 0, \quad \nabla_{\lambda} L(x^*, \lambda^*) = 0,
\]
\[
y^t \nabla^2_{xx} L(x^*, \lambda^*) y > 0 \quad \text{for all } y \neq 0 \text{ with } \nabla h(x^*)^t y = 0.
\]

Then \( x^* \) is a strict local minimum of \( f \) subject to \( h(x) = 0 \). In fact, there exist scalars \( \gamma > 0 \) and \( \varepsilon > 0 \) such that
\[
f(x) \geq f(x^*) + (\gamma/2) \| x - x^* \|^2, \quad \text{for all } x \text{ with } h(x) = 0 \text{ and } \| x - x^* \| < \varepsilon.
\]

**Proof.** First note that \( x^* \in C \) because \( \nabla_\lambda L(x^*, \lambda^*) = h(x^*) = 0 \). Assume that the last statement of the theorem is not true, so there exists a sequence \( \{x^k\} \) such that \( h(x^k) = 0 \) and \( x^k \to x^* \) for all \( k \). Letting \( k \to \infty \) in the expression above and using the continuity of the gradients \( \nabla h_i(\cdot) \), we get \( 0 = \nabla h_i(x^*)^t y_i \) for all \( i \). We will now show that \( y^t \nabla^2_{xx} L(x^*, \lambda^*) y \leq 0 \), thus coming to a contradiction. Call \( \delta_k := \| x^k - x^* \| \). By Taylor’s Formula
\[
0 = h_i(x^k) - h_i(x^*) = \delta_k \nabla h_i(x^*)^t y_i + (1/2) \delta^2_k (y_i^t \nabla^2 h_i(x^*) + \eta_{ik}(x^k - x^*)) y_i,
\]
for some \( \eta_{ik} \in (0, 1) \). Letting \( k \to \infty \) in the expression above and using the continuity of the gradients \( \nabla h_i(\cdot) \), we get \( 0 = \nabla h_i(x^*)^t y_i \) for all \( i \). We will now show that \( y^t \nabla^2_{xx} L(x^*, \lambda^*) y \leq 0 \), thus coming to a contradiction. Call \( \delta_k := \| x^k - x^* \| \). By Taylor’s Formula
\[
0 = h_i(x^k) - h_i(x^*) = \delta_k \nabla h_i(x^*)^t y_i + (1/2) \delta^2_k (y_i^t \nabla^2 h_i(x^*) + \eta_{ik}(x^k - x^*)) y_i,
\]
for some \( \eta_{ik}, \xi_k \in (0, 1) \). Multiply the first expression by \( \lambda_i^* \) and sum over all possible indexes \( i \). Then add the resulting expression to the second inequality to get
\[
(1/k) \delta^2_k \geq \delta_k (\nabla f(x^*)^t + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*))^t y_k
\]
\[
+ (1/2) \delta^2_k (y_k^t (\nabla^2 f(x^*) + \eta_{ik}(x^k - x^*))) + \sum_{i=1}^m \lambda_i^* \nabla^2 h_i(x^*)^t \eta_{ik}(x^k - x^*)) y_k,
\]
Since \( \nabla f(x^*)^t + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*) = 0 \) the expression above becomes
\[
(2/k) \geq (y_k^t (\nabla^2 f(x^*) + \xi_k(x^k - x^*))) + \sum_{i=1}^m \lambda_i^* \nabla^2 h_i(x^*)^t \eta_{ik}(x^k - x^*)) y_k,
\]
Taking now limits for \( k \to \infty \) we get \( 0 \geq y^t \nabla^2_{xx} L(x^*, \lambda^*) y \), contradicting our hypothesis. \( \square \)

**Example 2.9.5 (An illustration from producer theory)** Let us consider a cost-minimisation problem faced by a firm, which uses two inputs \( x_2 \) and \( x_2 \) to produce a single output \( y \) through the production function \( y = g(x_1, x_2) \). Assume that the production function for this case is \( g(x_1, x_2) = x_1 x_2 \). The unit prices of \( x_1 \) and \( x_2 \) are \( w_1 \geq 0 \) and \( w_2 \geq 0 \) respectively. The firm wishes to find the cheapest input combination \( (x_1, x_2) \) for producing a given amount \( \bar{y} > 0 \) of output \( y \). The feasible set of inputs is then defined as
\[
C := \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 x_2 = \bar{y}, x_1 \geq 0, x_2 \geq 0 \}.
\]
Since the price of each input combination \((x_1, x_2)\) is \(f(x_1, x_2) := w_1x_1 + w_2x_2\), the firm’s optimisation problem is
\[
\min f(x_1, x_2) = w_1x_1 + w_2x_2 \\
\text{subject to } (x_1, x_2) \in C
\]
We will assume that both prices \(w_1, w_2\) are strictly positive. If one of them is negative, then the minimisation problem stated above has no solution, because the objective function would be unbounded below. If some of these prices is zero, then there is no point in optimising the corresponding unit, because the amount would have no influence in the final price. Our next step is to reduce the problem to an equality constrained one. Note that, if \(x_1 = 0\) or \(x_2 = 0\) then the production function \(g(x_1, x_2) = x_1x_2 = 0\) and we want \(\tilde{y} > 0\). So the inputs \(x_1\) and \(x_2\) should be positive. Therefore we can actually work with the constraint set
\[
C_0 := \{(x_1, x_2) \in \mathbb{R}^2 | x_1x_2 = \tilde{y}\},
\]
and work out the inactive constraints \(x_1 > 0\) and \(x_2 > 0\) separately. Now we can pose the Lagrangian
\[
L(x_1, x_2, \lambda) = f(x_1, x_2) + \lambda [\tilde{y} - g(x_1, x_2)] = w_1x_1 + w_2x_2 + \lambda [\tilde{y} - x_1x_2],
\]
and the optimality conditions in (2.9.47) give the equalities
\[
w_1 - \lambda x_2 = 0 \\
w_2 - \lambda x_1 = 0 \\
\tilde{y} - x_1x_2 = 0.
\]
From the first equation we must have \(\lambda \neq 0\). We solve the first and second equations for \(x_1\) and \(x_2\) respectively, and we plug the results on the third equation. Choosing positive inputs \(x_1\) and \(x_2\) gives \(x^* = (\sqrt{w_2\tilde{y}/w_1}, \sqrt{w_1\tilde{y}/w_2})\) and \(\lambda^* = \sqrt{w_1w_2/\tilde{y}}\). In order to prove that \(x^*\) is in fact a minimum, let us check the second order sufficient conditions. The Hessian of the Lagrangian is
\[
\nabla^2 L(x^*, \lambda^*) = \nabla^2 f(x^*) + \lambda^* \nabla^2 g(x^*) \\
= \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} + \lambda^* \begin{pmatrix} 0 & -1 \\ -1 & 0 \end{pmatrix} \\
= \begin{pmatrix} 0 & -\lambda^* \\ -\lambda^* & 0 \end{pmatrix}.
\]
By Theorem 2.9.20, \(x^*\) is a strict minimizer if \(g' \nabla^2 L(x^*, \lambda^*) y > 0\) for every nonzero \(y \in V(x^*)\). The subspace \(V(x^*)\) is given by
\[
V(x^*) = \{y \in \mathbb{R}^2 | \nabla g(x^*)' y = 0\}.
\]
A few simple algebraic steps prove that
\[
V(x^*) = \{t(-w_2, w_1) | t \in \mathbb{R}\}.
\]
For \(y = t(-w_2, w_1) \in V(x^*)\) with \(t \neq 0\) we have \(g' \nabla^2 L(x^*, \lambda^*) y = 2t^2 \lambda^* w_1w_2 > 0\). Hence \(x^*\) is a strict minimum for the minimisation problem.

**Remark 2.9.11** The reduction to equality constrained problems should be done with caution, and in some cases it may not lead to a solution. In order to illustrate this fact, assume that the production function of the previous example were modified to \(g(x_1, x_2) = x_1 + x_2\). A little reflection then shows that, in this case, the solution of the problem is given by
\[
\begin{cases} 
(y, 0) & \text{if } w_1 < w_2 \\
(0, \tilde{y}) & \text{if } w_2 < w_1,
\end{cases}
\]
and that any pair element in \( C \) is a solution if \( w_1 = w_2 \). Therefore, the nonnegativity constraints matter in a nontrivial sense. If you ignore these constraints and find the stationary points of the Lagrangian, you obtain the system

\[
\begin{align*}
    w_1 + \lambda &= 0 \\
    w_2 + \lambda &= 0 \\
    x_1 + x_2 &= 0,
\end{align*}
\]

which has a solution only when \( w_1 = w_2 \). We need the results of the next section for tackling this problem.

### 2.9.2 Equality and Inequality Constraints

Consider a constraint set defined by both equality and inequality constraints. More precisely, consider the problem

\[
\begin{align*}
    \min f(x) \\
    \text{subject to } h_1(x) &= 0, \ldots, h_m(x) = 0, \\
    g_1(x) &\leq 0, \ldots, g_r(x) \leq 0,
\end{align*}
\]

where \( f, h_i, g_j \) are continuously differentiable from \( \mathbb{R}^n \) to \( \mathbb{R} \). We can write problem (2.9.49) more succinctly as

\[
\begin{align*}
    \min f(x) \\
    \text{subject to } h(x) &= 0, \\
    g(x) &\leq 0,
\end{align*}
\]

where \( h : \mathbb{R}^n \rightarrow \mathbb{R}^m \) and \( g : \mathbb{R}^n \rightarrow \mathbb{R}^r \) are the functions

\[
    h(x) := (h_1(x), \ldots, h_m(x)), \quad g(x) := (g_1(x), \ldots, g_r(x)).
\]

The optimality conditions that can be established for this case are called Karush-Kuhn-Tucker conditions and extend the results of the Lagrange multiplier Theorem. Before stating this result, let us define the Lagrangian associated with Problem (2.9.50).

**Definition 2.9.14** The Lagrangian associated with Problem (2.9.50) is the function \( L : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^r \rightarrow \mathbb{R} \) given by

\[
L(x, \lambda, \mu) := f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) + \sum_{j=1}^{r} \mu_j g_j(x).
\]

Recall that an inequality constraint \( g_j \) is called active at the point \( x \) when \( g_j(x) = 0 \). Denote by \( A(x) := \{ j \mid g_j(x) = 0 \} \) the set of indexes corresponding to active inequality constraints at \( x \). The set \( A(x) \) allows us to express the inequality constrained problem as an equality constrained one. This is shown in the following lemma.

**Lemma 2.9.7** Assume \( x^* \) is a local solution of the constrained problem \( (P) \)

\[
\begin{align*}
    \min f(x) \\
    h(x) &= 0 \\
    g(x) &\leq 0,
\end{align*}
\]

where \( h : \mathbb{R}^n \rightarrow \mathbb{R}^m \) and \( g : \mathbb{R}^n \rightarrow \mathbb{R}^r \). Then \( x^* \) is also a local solution of the equality constrained problem \( (P^*) \):

\[
\begin{align*}
    \min f(x) \\
    h(x) &= 0 \\
    g_j(x) &= 0, \quad \forall j \in A(x^*),
\end{align*}
\]

where \( A(x^*) \) is the set of indexes \( j \) such that \( g_j(x^*) = 0 \).
Proof. Consider the set \( U := \{ z \in \mathbb{R}^n / \text{such that } g_j(z) < 0, j \notin A(x^*) \} \). Since all \( g_j \) are continuous the set \( U \) is open. Note that the definition of \( A(x^*) \) implies that \( x^* \) belongs to \( U \). So there exists a radius \( r_0 > 0 \) such that \( B(r_0, x^*) \subset U \). On the other hand, \( x^* \) is a local solution of \((P)\), and hence there exists a radius \( r_1 > 0 \) such that for all \( x \in B(x^*, r_1) \) satisfying \( h(x) = 0 \) and \( g(x) \leq 0 \) it holds that \( f(x) \geq f(x^*) \). Now take \( r = \min \{ r_0, r_1 \} \). We have that for all \( x' \in B(x^*, r) \) such that \( h(x') = 0 \) and \( g_j(x') = 0 \) for all \( j \in A(x^*) \) it holds
\[
x' \in U \text{ and hence } g_j(x') < 0 \text{ for all } j \notin A(x^*).
\]
Hence \( x' \) is feasible for \((P)\) so that \( f(x') \geq f(x^*) \). This means that \( x^* \) is a local solution of problem \((P^*)\). The proof is complete.

Remark 2.9.12 Note that Lemma 2.9.7 says that the inequality constrained problem, at a given local solution, behaves as an equality constrained problem, where the equality constraints are the active ones at the given local solution. The crucial part of the proof is the fact that the set \( U \) is open.

A point \( x \) is called regular for problem \((2.9.50)\) when the set of all equality constraint gradients \( \{ \nabla h_1(x), \ldots, \nabla h_m(x) \} \), together with all active inequality constraint gradients \( \{ \nabla g_j(x) \}_{j \in A(x)} \), is linearly independent. We also say that a point is regular when the problem has no equality constraints and all the gradients corresponding to active inequality constraints \( \{ \nabla g_j(x) \}_{j \in A(x)} \) are linearly independent.

Theorem 2.9.21 (Karush-Kuhn-Tucker Necessary conditions- Inequality Constraints)
Let \( x^* \) be a local minimizer of Problem \((2.9.50)\), where all functions involved \( f, h_i, g_j \) are continuously differentiable. Assume that \( x^* \) is regular. Then there exist unique Lagrange multiplier vectors \( \lambda^* \in \mathbb{R}^m \) and \( \mu^* \in \mathbb{R}^r \) such that
\[
\nabla_x L(x^*, \lambda^*, \mu^*) = 0,
\mu_j^* \geq 0, \quad j = 1, \ldots, r,
\mu_j^* = 0, \quad j \notin A(x^*). \tag{2.9.51}
\]
If in addition, all functions involved \( f, h_i, g_j \) are twice continuously differentiable, it holds that
\[
y^t \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) y \geq 0, \tag{2.9.52}
\]
for all \( y \in \mathbb{R}^m \) such that \( \nabla h_i(x^*)^t y = 0 \) for all \( i = 1, \ldots, m \) and \( \nabla g_j(x^*)^t y = 0 \) for all \( j \in A(x^*) \).

Proof. The idea of the proof relies on the case of equality constraints. Indeed, if \( x^* \) is a regular local minimizer of Problem \((2.9.50)\), then by Lemma 2.9.7 it is also a local minimizer of the equality constrained problem
\[
\min f(x)
\text{subject to } h(x) = 0, g_j(x) = 0 \forall j \in A(x^*).
\]
Therefore, using the regularity assumption and Theorem 2.9.19, there exist Lagrange multipliers \( \lambda_1, \ldots, \lambda_m \) and \( \mu_j \) for all \( j \in A(x^*) \) such that
\[
\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*) + \sum_{j \in A(x^*)} \mu_j \nabla g_j(x^*) = 0.
\]
Defining \( \mu_j = 0 \) whenever \( j \notin A(x^*) \), we get
\[
\nabla f(x^*) + \sum_{i=1}^m \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^r \mu_j \nabla g_j(x^*) = 0.
\]
We have to prove now that \( \mu_j \geq 0 \) for all \( j \in A(x^*) \). The proof is analogous to the proof of the necessary conditions for the equality constrained case. Indeed, define

\[
g^+_j(x) := \max \{0, g_j(x)\}.
\]

Note that \( g^+_j(x) \) is continuous but not differentiable. However, the function \( (g^+_j(x))^2 \) is always differentiable, with gradient \( 2g^+_j(x)\nabla g_j(x) \). So for all \( \alpha > 0 \) the function

\[
F_k(x) := f(x) + (k/2)||h(x)||^2 + (k/2) \sum_{j=1}^{r}(g^+_j(x))^2 + (\alpha/2)||x - x^*||^2,
\]

is differentiable. Let \( r > 0 \) be such that \( f(x) \geq f(x^*) \) for all \( x \in B[x^*,r] \) such that \( h(x) = 0 \) and \( g(x) \leq 0 \). Consider the minimisation problem

\[
\min_{x \in B[x^*,r]} F_k(x) = (2.9.53)
\]

As in the proof of Theorem 2.9.19, the problem above has a solution \( x^k \). Using an argument similar to the one used in Theorem 2.9.19 we get that \( x^k \) converges to \( x^* \). This implies that for large enough \( k \), \( x^k \) is in the interior of the ball \( B[x^*,r] \), and hence they can be seen as unconstrained minimizers of problem (2.9.53). Therefore they verify the first order necessary conditions for optimality

\[
\nabla F_k(x^k) = 0.
\]

This condition can be re-written as

\[
\nabla f(x^k) + k \left[ \nabla h(x^k) \nabla g(x^k) \right] \left( \begin{array}{c} h(x^k) \\ g^+(x^k) \end{array} \right) + \alpha(x^k - x^*) = 0.
\]

At this point, we use the regularity assumption as in Theorem 2.9.19 and conclude that the Lagrange multipliers verify

\[
\lambda^*_i = \lim_{k \to \infty} k h_i(x^k), \quad i = 1, \ldots, m,
\]

\[
\mu^*_j = \lim_{k \to \infty} k g^+_j(x^k), \quad j = 1, \ldots, r.
\]

Since \( g^+_j(x^k) \geq 0 \), we obtain \( \mu^*_j \geq 0 \) for \( j = 1, \ldots, r \). \( \square \)

Note that condition \( \mu^*_j = 0, \ j \notin A(x^*) \) can also be written as

\[
\mu^*_j g_j(x^*) = 0, \ \text{for all} \ j = 1, \ldots, r,
\]

and is called \textit{complementary slackness condition} or simply \textit{complementarity condition}. It means that, whenever a constraint is \textit{inactive} at \( x^* \) (i.e., \( g_j(x^*) < 0 \)), then it is not important for the problem, and its corresponding multiplier must be zero.

**Exercise 2.9.36** Complete all technical details of the proof above.

**Exercise 2.9.37** Prove that when \( g : \mathbb{R}^n \to \mathbb{R} \) is differentiable the gradient of \( (g^+(x))^2 \) is \( 2g^+(x)\nabla g(x) \). Assume now that \( g \) is twice continuously differentiable. Is it true that \( (g^+_j(x))^2 \) is twice continuously differentiable?

**Example 2.9.6 (Example 2.9.5 revisited)** Assume now that the production function in Example 2.9.5 is \( g(x_1, x_2) = x_1 + x_2 \). The constraint set is

\[
C := \{(x_1, x_2) \in \mathbb{R}^2 \mid x_1 + x_2 = \bar{y}, \ x_1 \geq 0, \ x_2 \geq 0\}.
\]
As mentioned before, the inequality constraints are essential for this production function. So the Lagrangian is

\[ L(x, \lambda, \mu_1, \mu_2) = w_1x_1 + w_2x_2 + \lambda [y - x_1 - x_2] + \mu_1(-x_1) + \mu_2(-x_2), \]

and the Kuhn-Tucker conditions become

\[
\begin{align*}
    w_1 - \lambda - \mu_1 &= 0, \\
    w_2 - \lambda - \mu_2 &= 0, \\
    \mu_1 &\geq 0, \mu_2 \geq 0, \\
    \mu_1x_1 &= 0, \mu_2x_2 = 0, \\
    y - (x_1 + x_2) &= 0.
\end{align*}
\]

(2.9.54)

If \( \mu_1 > 0 \), we must have \( x_1 = 0 \) and \( x_2 = y > 0 \). Note that in this case we must have \( \mu_2 = 0 \), \( w_2 = \lambda \) and hence \( \mu_1 = w_1 - w_2 > 0 \). So the price of \( x_1 \) is higher, and this forces the optimal input to be \((0, y)\). A similar argument leads to the optimal solution \((y, 0)\) when \( \mu_2 > 0 \). The remaining case, when \( \mu_1 = \mu_2 = 0 \) is when the prices \( w_1 = w_2 \). In this case every nonnegative pair such that \( x_1 + x_2 = y \) is an optimal solution.

Now let us look at the convex case, i.e., the case in which both the objective function and the constraint set are convex. For the constraint set of problem (2.9.50) to be convex, we need every equality constraint \( h_i \) to be an affine function. Indeed, assume \( h_i \) is not affine and take two points \( x, y \) such that \( h_i(x) = h_i(y) = 0 \). If \( h_i \) is not affine on the line defined by these two points, there exists \( \lambda \in \mathbb{R} \) such that

\[ h_i(\lambda x + (1 - \lambda)y) \neq \lambda h_i(x) + (1 - \lambda)h_i(y) = 0. \]

So the convex combination does not satisfy the constraint \( h_i \). Therefore, we can guarantee that problem (2.9.50) has a convex structure when \( f \) and all \( g_j \) are convex, and all the equality constraints \( h_i \) are affine. By Theorem 2.9.21, we know that every local minimum of problem (2.9.50) satisfies the optimality conditions (2.9.51). The result below proves that in the convex case, conditions (2.9.51) also imply that \( x^* \) is a solution of problem (2.9.50).

**Theorem 2.9.22 (Karush-Kuhn-Tucker-Convex Case)** Assume that \( f \) and all \( g_j \) are convex, and that all equality constraints \( h_i \) are affine. Let \( x^*, \lambda^* \in \mathbb{R}^m \) and \( \mu^* \in \mathbb{R}^r \) be such that

\[
\begin{align*}
    \nabla_x L(x^*, \lambda^*, \mu^*) &= 0, \\
    \nabla_{\lambda} L(x^*, \lambda^*, \mu^*) &= 0, \\
    \mu_j^* &\geq 0, \quad j = 1, \ldots, r, \\
    \mu_j^* g_j(x^*) &= 0, \quad j = 1, \ldots, r, \\
    g_j(x^*) &\leq 0, \quad j = 1, \ldots, r, \\
    h_i(x^*) &= 0, \quad i = 1, \ldots, m.
\end{align*}
\]

(2.9.55)

Then \( x^* \) is a solution of (2.9.50).

**Proof.** Assume we have \( x^*, \lambda^* \in \mathbb{R}^m \) and \( \mu^* \in \mathbb{R}^r \) satisfying (2.9.55). In other words, we have

\[
\begin{align*}
(2.9.56 - a) \quad &\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(x^*) = 0 \\
(2.9.56 - b) \quad &\mu_j^* \geq 0, \\
(2.9.56 - c) \quad &\mu_j^* g_j(x^*) = 0, \quad j = 1, \ldots, r, \\
(2.9.56 - d) \quad &g_j(x^*) \leq 0, \quad j = 1, \ldots, r, \\
(2.9.56 - e) \quad &h_i(x^*) = 0, \quad i = 1, \ldots, m.
\end{align*}
\]

Let \( x \) be any feasible point. Since the functions \( h_i \) are affine we have

\[ h_i(x^*) + \alpha \nabla h_i(x^*)^t(x - x^*) = h_i(x^* + \alpha(x - x^*)) = \alpha h_i(x) + (1 - \alpha)h_i(x^*) = 0, \]
where we used the fact that both $x$ and $x^*$ are feasible. So for $i = 1, \ldots, m$ we get $\nabla h_i(x^*)^t(x - x^*) = 0$. Hence
\[
\sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*)^t(x - x^*) = 0. \quad (2.9.57)
\]
It is clear from the complementarity condition (2.9.56-c) that for all $j \notin A(x^*)$ we must have $\mu_j^* = 0$. Now take $j \in A(x^*)$. By the gradient inequality for the functions $g_j$ with $j \in A(x^*)$ we have
\[
g_j(x^*) + \alpha \nabla g_j(x^*)^t(x - x^*) \leq g_j(x^* + \alpha(x - x^*)) \leq \alpha g_j(x) + (1 - \alpha)g_j(x^*) \leq 0 = g_j(x^*),
\]
where we used that $g_j(x^*) = 0$ and $g_j(x) \leq 0$. Therefore, for all $j \in A(x^*)$ we get $\nabla g_j(x^*)^t(x - x^*) \leq 0$. Hence
\[
\sum_{j=1}^{r} \mu_j^* \nabla g_j(x^*)^t(x - x^*) \leq 0. \quad (2.9.58)
\]
Adding (2.9.57) and (2.9.58) we get
\[
\sum_{i=1}^{m} \lambda_i^* \nabla h_i(x^*)^t(x - x^*) + \sum_{j=1}^{r} \mu_j^* \nabla g_j(x^*)^t(x - x^*) \leq 0.
\]
Combining the above expression with (2.9.56-a) we conclude that
\[
\nabla f(x^*)^t(x - x^*) \geq 0,
\]
for every feasible point $x$. Using now the gradient inequality for $f$, we get
\[
f(x) \geq f(x^*) + \nabla f(x^*)^t(x - x^*) \geq f(x^*),
\]
for every feasible point $x$. Thus $x^*$ is a solution of (2.9.50). \hfill \Box

**Exercise 2.9.38** Consider the problem
\[
\min x_1x_2 \\
\text{s.t. } x_1 - x_2 - 2 \leq 0 \\
x_2 \leq 0.
\]
Prove that $x^* = (1, -1)$ is a strict local minimizer (it is in fact a global solution, check the level curves of the objective function).

We give in Figure 2.9.15 a chart that summarises how to use the optimality conditions for solving each of the different minimisation problems we have just seen.

### 2.9.3 Sensitivity Analysis for Constrained problems

Recall that, for Linear Problems, sensitivity analysis studies the dependency of both the solution and the optimal values under small perturbations on the data. It is a classical fact of Linear Programming that the sensitivity of a given linear problem is closely related with the dual variables (i.e., the Lagrange multipliers) associated to the problem. A very similar situation can be extended to the Nonlinear case, where the optimal solution of the perturbed problem can be shown to be a continuously differentiable function of the perturbation. Moreover, the rate of change of the optimal value of the perturbed problem is expressed in terms of the Lagrange multipliers associated with the problem. We will consider the equality constrained problem in our analysis. However, when inequalities are present, we can transform them into equalities by using Valentine’s trick: Given an inequality constraint $g : \mathbb{R}^n \to \mathbb{R}$
\[
g(x) \leq 0 \iff \exists t \in \mathbb{R} : g(x) = -t^2 \iff G(x, t) = g(x) + t^2 = 0.
\]
In this way we transform all inequalities into equalities, at the cost of adding an artificial variable \( t \). Note that \( G(x, t) \) inherits all differentiability properties enjoyed by \( g \). Moreover, \( G \) is convex as a function of \((x, t)\) if and only if \( g \) is a convex function of \( x \).

So we can consider the equality-only constrained problem

\[
\begin{align*}
\min & \quad f(x) \\
\text{s.t.} & \quad h(x) = 0,
\end{align*}
\]

where \( h : \mathbb{R}^n \to \mathbb{R}^m \). Let \( x^* \) be a local minimizer of (2.9.59). We look at the following perturbed versions of problem (2.9.59).

\[
\begin{align*}
\min & \quad f(x) \\
\text{s.t.} & \quad h(x) = p,
\end{align*}
\]

where \( p \in \mathbb{R}^m \) is the perturbation parameter. Note that \( p = 0 \) represents our original problem (2.9.59). For a given \( p \), the local solutions of (2.9.60) are denoted as \( x^*(p) \). The optimal value of problem (2.9.60) also depends on the perturbation parameter \( p \) and is given by \( f(x^*(p)) \). We want to study the properties of the functions \( x(\cdot) \) and \( v(\cdot) := f(x(\cdot)) \) for \( p \) in a small neighborhood of \( p = 0 \). These properties are established in the following theorem, where we assume that the original local minimizer verifies the second order sufficient conditions of Theorem 2.9.20.

In order to prove Theorem 2.9.24, we need a standard result from Analysis: The Inverse Function Theorem, which we state below. Its proof can be found, e.g., in W. Rudin, *Principles of Mathematical Analysis*. 

---

Figure 2.9.15: Using optimality conditions.
Theorem 2.9.23 (Inverse Function Theorem) Let $F : \mathbb{R}^n \to \mathbb{R}^n$ be a $C^k$ function (i.e., with continuous derivatives up to $k$-th order) and $x^* \in \mathbb{R}^n$ be such that the Jacobian matrix $J_F(x^*)$ is invertible. Then

(i) There is an open neighborhood $U$ of $x^*$ such that $y^* := F(x^*)$ is an interior point of the image $V := F(U)$ and the restriction of $F$ to $U$, $F : U \to V$ has an inverse $G := F^{-1}$.

(ii) $G$ is also a $C^k$ function, and $J_G(y^*) = J_F(x^*)^{-1}$.

Now we are in conditions to state and prove the sensitivity result.

Theorem 2.9.24 Assume that $x^*$ is a regular local minimizer of problem (2.9.59). Suppose that there exists $\lambda^* \in \mathbb{R}$ such that

\[
\begin{align*}
\nabla_x L(x^*, \lambda^*) &= 0, \\
h(x^*) &= 0, \\
y^T \nabla^2_x L(x^*, \lambda^*) y &> 0, \text{ for all nonzero } y \in V(x^*).
\end{align*}
\]

Then there exists a neighborhood $U \subset \mathbb{R}^m$ of $p = 0$ and a continuously differentiable function $x : U \to \mathbb{R}^n$ such that $x(0) = x^*$ and

(i) $x(p)$ is a strict local minimizer of problem (2.9.60),

(ii) $v(p) = f(x(p))$ verifies $\nabla v(0) = -\lambda^* t$.

Proof. The Lagrangian associated with problem (2.9.59) is

\[L(x, \lambda) = f(x) + \sum_{i=1}^m \lambda_i h_i(x).\]

The gradient and Hessian of $L$ with respect to the variables $(x, \lambda)$ are

\[
\nabla L(x, \lambda) = \begin{pmatrix}
\nabla_x L(x, \lambda) \\
\nabla_\lambda L(x, \lambda)
\end{pmatrix} = \begin{pmatrix}
\nabla f(x) + \sum_{i=1}^m \lambda_i \nabla h_i(x) \\
h(x)
\end{pmatrix} \in \mathbb{R}^{n+m}
\]

and

\[
\nabla^2 L(x, \lambda) = \begin{pmatrix}
\nabla^2_x L(x, \lambda) & \nabla_\lambda (\nabla_x L(x, \lambda)) \\
\nabla_\lambda (\nabla_x L(x, \lambda)) & \nabla^2_\lambda L(x, \lambda)
\end{pmatrix}
\]

\[
= \begin{pmatrix}
\nabla^2 f(x) + \sum_{i=1}^m \lambda_i \nabla^2 h_i(x) & \nabla h_1(x) & \ldots & \nabla h_m(x) \\
\n\nabla h_1(x)^t & 0 & \ldots & 0 \\
\vdots & 0 & \ldots & 0 \\
\n\nabla h_m(x)^t & 0 & \ldots & 0
\end{pmatrix} \in \mathbb{R}^{(n+m)\times(n+m)}
\]
By Exercise 2.9.39 and the regularity assumption we have that the matrix $\nabla^2 L(x^*, \lambda^*)$ is invertible. By the Inverse Function Theorem, Theorem 2.9.23 applied to $F := \nabla L : \mathbb{R}^{n+m} \to \mathbb{R}^{n+m}$ we have that

(i) There is an open neighborhood $U \times W \subset \mathbb{R}^n \times \mathbb{R}^m$ of $(x^*, \lambda^*)$ such that $(0, 0) = \nabla L(x^*, \lambda^*)$ is an interior point of the image $\nabla L(U \times W)$ and the restriction of $\nabla L$ to $U \times W$, $\nabla L : U \times W \to \nabla L(U \times W)$ has an inverse $(\nabla L)^{-1}$.

(ii) $\nabla L^{-1}$ is a $C^1$-function on $\nabla L(U \times W)$.

Note that $F^{-1}(0, 0) = (x^*, \lambda^*)$ so that the set $\nabla L(U \times W)$ is in fact a neighborhood of zero. Take $U_0 \subset \mathbb{R}^n$ and $W_0 \subset \mathbb{R}^m$ such that $(0, 0) \in U_0 \times W_0 \subset \nabla L(U \times W)$. Now consider the function $F^{-1}(0, \cdot) : W_0 \to U \times W$, obtained by restricting $F^{-1}$ to the set $\{0\} \times W_0$. We pointed out before that $W_0$ is a neighborhood of zero. Furthermore, for each $p \in W_0$, the image $F^{-1}(0, p) = (x(p), \lambda(p))$ verifies the necessary optimality conditions for problem (2.9.60). Indeed, $F^{-1}(0, p) = (x(p), \lambda(p))$ if and only if $\nabla L(x(p), \lambda(p)) = (0, p) = (\nabla_x L(x(p), \lambda(p)), h(x(p)))$, which means that $(x(p), \lambda(p))$ verifies the first order necessary conditions for problem (2.9.60). We claim now that there exists a smaller neighborhood of zero $W_1 \subset W_0$ such that for every $p \in W_1$, the sufficient optimality conditions are verified by the corresponding optimal solution $(x(p), \lambda(p))$. In other words, we will prove that the point $(x(p), \lambda(p))$ is a strict local solution of (2.9.60) for every $p \in W_1$. Indeed, if the claim is false, then there exists a sequence $p_k \to 0 \in \mathbb{R}^m$ and a sequence $d_k \in V(x(p_k))$ with $\|d_k\| = 1$ such that $d_k^T \nabla^2 L(x(p_k), \lambda(p_k))d_k \leq 0$. Since $p_k \to 0$ and the function $F^{-1}(0, \cdot)$ is continuous, we have $F^{-1}(0, p_k) = (x(p_k), \lambda(p_k)) \to F^{-1}(0, 0) = (x^*, \lambda^*)$. Since the Hessian of the Lagrangian $\nabla^2 L$ is continuous we must have

$$\nabla^2 L(x(p_k), \lambda(p_k)) \to \nabla^2 L(x^*, \lambda^*).$$

(2.9.61)

By Bolzano-Weierstrass Theorem there exists a convergent subsequence $\{d_{k_i}\}$, with nonzero limit $d$. By the definition of $d_k$ and (2.9.61) we must have $d^T \nabla^2 L(x^*, \lambda^*)d \leq 0$, which contradicts the assumption of $(x^*, \lambda^*)$. Therefore the claim is true and there exists a neighborhood of zero $W_1$ such that for every $p \in W_1$ the point $(x(p), \lambda(p))$ is a strict local solution of (2.9.60). This completes the proof of item (i). In order to prove (ii), we use the chain rule to differentiate the functions $v(p) = f(x(p))$ and $h(x(p)) = p$. to get

$$\nabla v(0) = \nabla f(x(0))^T J_x(0) = \nabla f(x^*) J_x(0) \quad \text{and} \quad J_h(x(0)) J_x(0) = J_h(x^*) J_x(0) = I$$

where we used the fact that $x(0) = x^*$. So the first order necessary conditions $\nabla f(x^*) + \sum_{i=1}^m \lambda_i \nabla h_i(x^*) = 0$ yield

$$\nabla v(0) = (- \sum_{i=1}^m \lambda_i \nabla h_i(x^*)) J_x(0) = - \lambda^T J_h(x^*) J_x(0) = - \lambda^T I = - \lambda^T,$nabla v(0) = (- \sum_{i=1}^m \lambda_i \nabla h_i(x^*)) J_x(0) = - \lambda^T J_h(x^*) J_x(0) = - \lambda^T I = - \lambda^T,$$

which completes the proof of (ii). $\square$

**Exercise 2.9.39** Let $C \in \mathbb{R}^{m \times n}$ be symmetric and $A \in \mathbb{R}^{m \times n}$ such that $y^T Cy \neq 0$ for every nonzero element $y \in N(A)$. Then the matrix

$$Q := \begin{pmatrix} C & A^T \\ A & 0 \end{pmatrix}$$

is invertible if and only if the rows of $A$ are linearly independent.

Hint: First assume that $Q$ is invertible and suppose the conclusion is not true. Note that the rows of a matrix $A \in \mathbb{R}^{m \times n}$ are linearly dependent if and only if there exists a nonzero $v \in \mathbb{R}^m$ such that $A^Tv = 0$. Then prove that $(0, v) \in N(Q)$, contradicting the assumption on $Q$. Conversely, assume that the rows of $A$ are linearly independent. Prove that $(x, v) \in \mathbb{R}^{n+m}$ is in the null space of $Q$ if and only if $x \in N(A)$ and $Cx = -A^Tv$. Prove that, if $x = 0$, then by the assumption on $A$ we must have $v = 0$. If $x \neq 0$, use the assumption on $C$ to obtain a contradiction.
2.10 Optimality Conditions for Convex Nondifferentiable Problems

A convex problem is an optimization problem in which both the objective function and the constraint set are convex. Some authors define as a convex problem one of the form

$$\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad g_i(x) \leq 0, \quad i = 1, \ldots, m \\
& \quad \langle a_i, x \rangle = b_i, \quad j = 1, \ldots, p,
\end{align*}$$

(2.10.62)

where \( f, g_i : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) are convex functions, \( \{a_1, \ldots, a_p\} \subset \mathbb{R}^n \) and \( \{b_1, \ldots, b_p\} \subset \mathbb{R} \). The equality constraints are affine so that the constraint set remains convex. Note that, when the problem is convex in the sense of our definition (i.e., when the objective function and the constraint set are convex), then it can always be formulated as in 2.10.62. Indeed, let \( f \) be the convex objective and \( C \) the convex constraint set. Then our minimization problem can be formulated as

$$\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad \delta_C(x) \leq 0,
\end{align*}$$

where \( \delta_C \) is defined as in 2.1.6.

Recall that, for a differentiable convex problem, every local solution is global (see Corollary 2.8.8). We prove this next for the nondifferentiable case. Consider a general convex problem of the form

$$\begin{align*}
\text{minimize} & \quad f(x) \\
\text{subject to} & \quad x \in C.
\end{align*}$$

(2.10.63)

**Theorem 2.10.25** Assume that \( \emptyset \neq C \subset \mathbb{R}^n \) is convex and \( f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \) is convex. Let \( \bar{x} \in C \) be a local solution of (2.10.63), i.e., there exists \( r > 0 \) such that

$$f(\bar{x}) \leq f(y) \quad \text{for every } y \in C \cap B(\bar{x}, r).$$

Then,

(i) \( \bar{x} \) is a (global) solution of (2.10.63).

(ii) If, in addition, either \( \bar{x} \) is a strict local minimizer or \( f \) is strictly convex, then the global solution \( \bar{x} \) is unique.

**Proof.** Assume that (i) is not true. Then there exists \( \bar{x} \in C \) such that \( f(\bar{x}) < f(\bar{x}) \). This inequality, together with (2.10.64) yields \( \bar{x} \notin B(\bar{x}, r) \). In other words, \( \|\bar{x} - \bar{x}\| > r \) and hence \( \lambda := \frac{r}{\|\bar{x} - \bar{x}\|} < 1 \). For this fixed choice of \( \lambda \) define \( z := \lambda \bar{x} + (1 - \lambda)\bar{x} \). Note that \( z \in C \) and \( \|z - \bar{x}\| = r \), so \( z \in C \cap B(\bar{x}, r) \) By (2.10.64) we must have \( f(z) \geq f(\bar{x}) \). Use also convexity to write

$$f(\bar{x}) \leq f(\bar{x} + \lambda(\bar{x} - \bar{x})) = f(z) \leq \lambda f(\bar{x}) + (1 - \lambda)f(\bar{x}) \leq f(\bar{x}),$$

where we also used the assumption on \( \bar{x} \). The expression above entails a contradiction, so (i) must hold. For proving (ii), we assume that \( \bar{x} \) is a strict local minimizer. The case of \( f \) strictly convex is a consequence of Proposition 2.5.6 and the details of the proof are left to the reader. If \( \bar{x} \) is a strict local minimizer, then for some \( r > 0 \) we have

$$f(\bar{x}) < f(y) \quad \text{for every } y \in C \cap B(\bar{x}, r).$$

(2.10.65)

By part (i) of this theorem, we know that \( \bar{x} \) is a global solution of (2.10.63). Assume there exists a global solution \( \bar{x} \) of (2.10.63) with \( \bar{x} \neq \bar{x} \). Then we must have \( f(\bar{x}) = f(\bar{x}) \) (both are global...
solutions, so they must have the same (optimal) objective value). By convexity we also have for every \( a \in [0, 1] \)
\[
f(\bar{x}) \leq f(a\bar{x} + (1-a)\bar{x}) \leq af(\bar{x}) + (1-a)f(\bar{x}) = f(\bar{x}).
\]
Hence \( f(a\bar{x} + (1-a)\bar{x}) = f(\bar{x}) \) for every \( a \in [0, 1] \). Taking \( a := \frac{r}{\|x-\bar{x}\|} < 1 \) and using also (2.10.65) we get
\[
f(\bar{x}) < f(a\bar{x} + (1-a)\bar{x}),
\]
a contradiction. Therefore \( \bar{x} \) is the unique global solution of (2.10.63).

We recall now the concepts of indicator function and normal cone, introduced in Exercise 2.6.26. We state formally their definition.

**Definition 2.10.15** Let \( C \subset \mathbb{R}^n \) be a closed and convex set. The function \( \delta_C : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\} \)
is called the indicator function of the set \( C \) and is defined as
\[
\delta_C(x) := \begin{cases} 0 & x \in C, \\ +\infty & x \notin C, \end{cases}
\]
Recall that
\[
\partial \delta_C(x_0) = \begin{cases} \emptyset & \text{if } x_0 \in C, \\ \{w \in \mathbb{R}^n : \langle w, y - x_0 \rangle \leq 0, \forall y \in C\} & \text{if } x_0 \notin C. \end{cases}
\]
The set \( \partial \delta_C(x_0) \) is called normal cone of \( C \) at the point \( x_0 \) and is denoted by \( N_C(x_0) \) (see also Figure 2.6.11).

The following lemma gives us an intuitive connection between normal cone at boundary points and supporting hyperplanes of a set at a boundary point.

**Lemma 2.10.8** Let \( C \subset \mathbb{R}^n \) be a convex set. Then,

(a) \( N_C(x_0) = \{0\} \) if and only if \( x_0 \in C^\circ \).

(b) \( N_C(x_0) \supseteq \{0\} \) if and only if \( x_0 \) belongs to the boundary of \( C \).

**Proof.** (a) Assume that \( N_C(x_0) = \{0\} \). From Definition 2.10.15 we must have \( x_0 \in C \). If \( x_0 \) belongs to the boundary of \( C \), we know by Theorem 2.5.7 that there is a supporting hyperplane of \( C \) at \( x_0 \). In other words, there exist \( 0 \neq w \in \mathbb{R}^n \) and \( \gamma \in \mathbb{R} \) such that
\[
\langle w, z \rangle \leq \gamma \quad \text{for all } z \in C \quad \text{and} \quad \langle w, x_0 \rangle = \gamma.
\]
We can combine the above two expressions to get
\[
\langle w, z - x_0 \rangle \leq 0 \quad \text{for all } z \in C,
\]
so \( 0 \neq w \in N_C(x_0) \), contradicting \( N_C(x_0) = \{0\} \). Therefore \( x_0 \in C^\circ \). Conversely, fix \( x_0 \in C^\circ \). And assume there exists \( 0 \neq w \in N_C(x_0) \). Because \( w \in N_C(x_0) \) we must have
\[
\langle w, z - x_0 \rangle \leq 0,
\]
for every \( z \in C \). Using now that \( x_0 \in C^\circ \), we can find \( r > 0 \) such that \( x_0 + rw \in C \). Taking \( z = x_0 + rw \in C \) in (2.10.66) we get \( r\|w\|^2 \leq 0 \), which yields \( w = 0 \), a contradiction. So we must have \( N_C(x_0) = \{0\} \). Item (b) follows from (a). Indeed, assume that \( x_0 \) belongs to the boundary of \( C \). Then we cannot have \( N_C(x_0) = \{0\} \), because the latter equality and item (a) yields \( x_0 \in C^\circ \), a contradiction. Conversely, if \( N_C(x_0) \supseteq \{0\} \), then we cannot have \( x_0 \in C^\circ \), because the latter inclusion and (a) yield \( N_C(x_0) = \{0\} \).

**Exercise 2.10.40** Prove the following statements.
(1) If \( C = \{ x_0 \} \subset \mathbb{R}^n \) then \( N_C(x_0) = \mathbb{R}^n \).

(2) Let \( C = B(x_0, r) \) and fix \( y \in B(x_0, r) \). Then \( \| y - x_0 \| = r \) (i.e., \( y \) belongs to the boundary of \( B(x_0, r) \)) if and only if \( N_C(y) = \{ \lambda(y - x_0) : \lambda \geq 0 \} \). If \( y \) is such that \( \| y - x_0 \| < r \) (i.e., \( y \) belongs to the interior of \( B(x_0, r) \)) then \( N_C(y) = \{ 0 \} \) (see Figure 2.10.16).

Hint: See Lemma 2.10.8.

If we look at \( N_C(\cdot) \) as a mapping from \( \mathbb{R}^n \) to \( \mathbb{R}^n \), we see that it maps a point \( x \in \mathbb{R}^n \) into a set \( N_C(x) \subset \mathbb{R}^n \). More generally, for every convex function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{ +\infty \} \) we have that the mapping \( \partial f(\cdot) \) maps a point \( x \in \mathbb{R}^n \) into the subset \( \partial f(x) \subset \mathbb{R}^n \) (which might be the empty set). These mappings are therefore called point-to-set mappings.

**Definition 2.10.16** We say that a mapping \( A \) is a point-to-set mapping when, for every \( x \in \mathbb{R}^n \) we have \( A(x) \subset \mathbb{R}^n \). This situation is denoted as \( A : \mathbb{R}^n \rightharpoonup \mathbb{R}^n \). The set \( \{ x \in \mathbb{R}^n : A(x) \neq \emptyset \} \) is called the domain of the mapping \( A \).

We established in Lemma 2.10.8 the relationship between normals at a boundary point \( x_0 \) of a set \( C \) and supporting hyperplanes of \( C \) at \( x_0 \). When the set \( C \) is the epigraph of a convex function \( f \), then the supporting hyperplanes of \( C \) at a given boundary point \( (x_0, f(x_0)) \) are also related with subgradients of \( f \) at \( x_0 \). We have seen in Theorem 2.6.10 that the existence of a subgradient \( w \in \partial f(x_0) \) is equivalent to having a nonvertical supporting hyperplane of \( \text{Epi} f \) at \( (x_0, f(x_0)) \). However, a nonvertical supporting hyperplane might not exist at certain boundary points of the epigraph. In the epigraph depicted in Figure 2.10.17, we see that the only supporting hyperplanes at the points \((-1,0)\) and \((1,0)\) are vertical, so the function has no subgradients at \( x = 1, -1 \).

**Exercise 2.10.41** Find the function \( f : \mathbb{R} \rightarrow \mathbb{R} \cup \{ +\infty \} \) whose epigraph is depicted in Figure 2.10.17.

The following lemma proves that a nonvertical supporting hyperplane always exists when \( x_0 \in (\text{dom } f)^0 \). In other words, a pathological situation as depicted in Figure 2.10.17 can only happen at the boundary of \( \text{dom } f \).

**Lemma 2.10.9** Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{ +\infty \} \) be convex and fix \( x_0 \in (\text{dom } f)^0 \). Then there exists a nonvertical supporting hyperplane of \( \text{Epi} f \) at \( (x_0, f(x_0)) \). In the latter situation, \( \partial f(x_0) \neq \emptyset \).
Proof. Because $\text{Epi } f$ is convex, we know by Theorem 2.5.7 that there exists $(0, 0) \neq (w_0, a_0) \in \mathbb{R}^{n+1}$ such that the hyperplane

$$H := \{(y, t) \in \mathbb{R}^{n+1} : \langle w_0, y \rangle + a_0 t = b_0 \},$$

is a supporting hyperplane of $\text{Epi } f$ at $(x_0, f(x_0))$. In other words, we have that $(x_0, f(x_0)) \in Epi f \cap H$ and

$$Epi f \subset H^+ := \{(y, t) \in \mathbb{R}^{n+1} : \langle w_0, y \rangle + a_0 t \geq b_0 \}.$$

Then, we must have $a_0 \geq 0$. Indeed, note that $(x_0, f(x_0) + \gamma) \in Epi f$ for every $\gamma \geq 0$. Using now the inclusion $Epi f \subset H^+$ we get

$$\langle w_0, x_0 \rangle + a_0(f(x_0) + \gamma) = \langle w_0, x_0 \rangle + a_0f(x_0) + a_0\gamma \geq b_0,$$

for all $\gamma \geq 0$. Letting $\gamma \to +\infty$ we see that $a_0 \not< 0$, so we must have $a_0 \geq 0$. Now let us prove that, if $x_0 \in (\text{dom } f)^\circ$, then $a_0 > 0$ (i.e., $H$ is not vertical). Suppose that $a_0 = 0$. Because $H$ is a supporting hyperplane, we know that $(w_0, a_0) \neq (0, 0)$, so we must have $w_0 \neq 0$. From the inclusion $Epi f \subset H^+$ we get

$$\langle w_0, y \rangle \geq b_0,$$

for every $y \in \text{dom } f$. Because $(x_0, f(x_0)) \in Epi f \cap H$ we must have $\langle w_0, x_0 \rangle = b_0$, which, together with the previous expression gives

$$\langle w_0, y - x_0 \rangle \geq 0,$$

for every $y \in \text{dom } f$. Using now that $x_0 \in (\text{dom } f)^\circ$, we conclude that there exists $r > 0$ such that $B(x_0, r) \subset \text{dom } f$. In particular, the point $(x_0 - r \frac{w_0}{\|w_0\|}) \in B(x_0, r)$ and hence $(x_0 - r \frac{w_0}{\|w_0\|}) \in \text{dom } f$. Using the above inequality for $y = (x_0 - r \frac{w_0}{\|w_0\|}) \in \text{dom } f$ yields

$$\langle w_0, -r \frac{w_0}{\|w_0\|} \rangle = -r\|w_0\| \geq 0.$$

This implies that $w_0 = 0$, a contradiction. Hence, we must have $a_0 > 0$. Let us prove now that, in the latter situation, $\partial f(x_0) \neq \emptyset$. Indeed, we have just proved that there exists a nonvertical supporting hyperplane of $Epi f$ at the point $(x_0, f(x_0))$. Now use Theorem 2.6.10 to conclude that $\partial f(x_0) \neq \emptyset$. □
Corollary 2.10.9 Let \( f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\} \) be convex and fix \( x_0 \in \text{dom} f \). If \( f \) is continuous at \( x_0 \), then \( x_0 \in (\text{dom} f)' \) and \( (\text{Epi} f)' \neq \emptyset \). In the latter situation, \( \partial f(x_0) \neq \emptyset \).

Proof. Note that the last statement follows directly from the first statement and the previous lemma. So let us prove the first statement. Assume that \( f \) is continuous at \( x_0 \in \text{dom} f \). Because \( f(x_0) \in \mathbb{R} \) and \( f \) is continuous at \( x_0 \), there exists \( r > 0 \) such that for every \( y \in B(x_0, r) \) we have \( f(y) < f(x_0) + 1 \). Therefore, \( B(x_0, r) \subset \text{dom} f \). This yields \( x_0 \in (\text{dom} f)' \). Moreover, we have

\[
B(x_0, r) \times (f(x_0) + 1, +\infty) \subset \text{Epi} f.
\]

Indeed, fix \((y, t) \in B(x_0, r) \times (f(x_0) + 1, +\infty)\), then \( f(y) < f(x_0) + 1 < t \). So \( f(y) < t \) and hence \((y, t) \in \text{Epi} f \). Because \( B(x_0, r) \times (f(x_0) + 1, +\infty) \) is an open set, we conclude that \( (\text{Epi} f)' \neq \emptyset \). The proof is complete. \( \square \)

Many basic tools of classical analysis, such as continuity concepts, can be extended to the more general setting of point-to-set mappings. Recall from Exercise 2.1.2 that the sum of two arbitrary sets \( C_1 \) and \( C_2 \) is defined as

\[
C_1 + C_2 := \{x_1 + x_2 \mid x_1 \in C_1, x_2 \in C_2\}.
\]

In particular, if \( C_1 = \emptyset \), we must have \( C_1 + C_2 = \emptyset \).

Remark 2.10.13 Let \( A, B : \mathbb{R}^n \rightrightarrows \mathbb{R}^n \) be two point-to-set mappings. The following statements are equivalent.

(a) There exists \( w \in \mathbb{R}^n \) such that \( w \in -A(x) \cap B(x) \),

(b) \( 0 \in A(x) + B(x) \).

Indeed, recall that the set \(-A(x) = \{-z \in \mathbb{R}^n : z \in A(x)\} \), so (a) means that

\[
-w \in A(x), \quad w \in B(x).
\]

Therefore \( 0 = -w + w \in A(x) + B(x) \), so (b) holds. Conversely, if (b) holds, then there exists \( w_1 \in A(x) \) and \( w_2 \in B(x) \) such that \( 0 = w_1 + w_2 \). Hence \(-w_1 = w_2 \in -A(x) \cap B(x) \) and so (a) holds for \( w = w_2 \).

We state and prove next the (nondifferentiable) optimality conditions for convex problems. In the same way as in the differentiable case, the optimality conditions for convex problems are both necessary and sufficient.

Theorem 2.10.26 (Characterization of solutions of nondifferentiable convex problems)
Assume that \( 0 \neq C \subset \mathbb{R}^n \) is convex and \( f : \mathbb{R}^n \to \mathbb{R} \) is convex. A point \( \bar{x} \) is a solution of (2.10.63) if and only if \( \bar{x} \in C \) and there exists \( w \in \partial f(\bar{x}) \) such that

\[
\langle w, y - \bar{x} \rangle \geq 0, \quad \text{for all} \ y \in C. \tag{2.10.67}
\]

The inequality above is equivalent to the inclusion

\[
0 \in \partial f(\bar{x}) + N_C(\bar{x}), \tag{2.10.68}
\]

where \( N_C \) is the normal cone to \( C \) at the point \( \bar{x} \in C \) (see Exercise 2.6.26).
Proof. The last statement of the theorem follows from Remark 2.10.13. Indeed, inequality (2.10.67) holds if and only if \(-w \in N_C(x)\). Because \(w \in \partial f(x)\), item (a) of Remark 2.10.13 holds for \(x := \tilde{x}\) and the point-to-set mappings \(A := N_C(\cdot)\), \(B := \partial f(\cdot)\). Because inclusion (2.10.68) is precisely item (b) of Remark 2.10.13, we obtain the desired equivalence. So let us prove now that every solution of (2.10.63) is characterized by (2.10.67). If (2.10.67) holds, then \(\tilde{x} \in C\). The subgradient inequality now implies that \(\tilde{x}\) is a solution. Indeed, by (2.2.8) we have

\[
f(y) \geq f(\tilde{x}) + \langle w, y - \tilde{x} \rangle \geq f(\tilde{x}), \quad \text{for all } y \in C,
\]

where we used (2.10.67) in the rightmost inequality. Because \(\tilde{x} \in C\), the above expression implies that \(\tilde{x}\) is a solution of (2.10.63). Conversely, assume now that \(\tilde{x}\) is a solution of (2.10.63) and consider the two following sets.

\[
\begin{align*}
\Omega_1 & := \{(y - \tilde{x}, a) \in \mathbb{R}^n \times \mathbb{R} : y \in \mathbb{R}^n, f(y) - f(\tilde{x}) < a\} \\
\Omega_2 & := \{(y - \tilde{x}, a) \in \mathbb{R}^n \times \mathbb{R} : y \in C, a \leq 0\}.
\end{align*}
\]

It is an exercise for the reader to prove that \(\Omega_1\) and \(\Omega_2\) are convex. Moreover, \(\Omega_1 \cap \Omega_2 = \emptyset\). Indeed, if there exists \((y - \tilde{x}, a) \in \Omega_1 \cap \Omega_2\) we have

\[
0 \geq a > f(y) - f(\tilde{x}),
\]

so \(f(y) < f(\tilde{x})\) with \(y \in C\). This contradicts the fact that \(\tilde{x}\) is a solution of (2.10.63). Therefore, \(\Omega_1\) and \(\Omega_2\) are convex and disjoint. By Theorem 2.5.8, there exists \((u, b) \in \mathbb{R}^n \times \mathbb{R}\), with \(\|(u, b)\| = 1\), such that

\[
\langle u, y - \tilde{x} \rangle + ab \leq \langle u, y' - \tilde{x} \rangle + a'b,
\]

for all \((y - \tilde{x}, a) \in \Omega_1\), \((y' - \tilde{x}, a') \in \Omega_2\). Using now the definitions of \(\Omega_1\) and \(\Omega_2\) we get that

\[
\langle u, y - \tilde{x} \rangle + ab \leq \langle u, y' - \tilde{x} \rangle + a'b, \tag{2.10.69}
\]

for all \(y \in \mathbb{R}^n\), \(f(y) - f(\tilde{x}) < a\), \(y' \in C\), \(a' \leq 0\). Choosing \(y' := \tilde{x} \in C\) and \(a' = 0\) in the right-hand side yields

\[
\langle u, y - \tilde{x} \rangle + ab \leq 0, \tag{2.10.70}
\]

for all \(y \in \mathbb{R}^n\), \(f(y) - f(\tilde{x}) < a\). Choosing \(y := \tilde{x}\) and an arbitrary \(a > 0\) in (2.10.70) gives \(ab \leq 0\) for every \(a > 0\). Hence \(b \leq 0\). We claim that \(b < 0\). Note that \(f\) is finite everywhere, so \(\text{dom } f = \mathbb{R}^n\). This implies that for every fixed \(y \in \mathbb{R}^n\) we can find \(a > 0\) large enough such that \((y - \tilde{x}, a) \in \Omega_1\). Assume now that \(b = 0\), we will arrive to a contradiction. From (2.10.70) we can write

\[
\langle u, y - \tilde{x} \rangle \leq 0,
\]

for all \((y - \tilde{x}, a) \in \Omega_1\). Because \(\text{dom } f = \mathbb{R}^n\) we can take any arbitrary \(y \in \mathbb{R}^n\) in the inequality above. Take \(y = u + \tilde{x}\) in the last inequality to get \(u = 0\). Altogether \(\|(u, b)\| = 1\), contradicting the fact that \(\|(u, b)\| = 1\). So we must have \(b < 0\). Divide the left-hand side of (2.10.70) by \(-b > 0\) to get

\[
\langle -\frac{u}{b}, y - \tilde{x} \rangle - a \leq 0, \tag{2.10.71}
\]

whenever \(y \in \mathbb{R}^n\) and \(f(y) - f(\tilde{x}) < a\). For every fixed \(y \in \mathbb{R}^n\) and every fixed \(\varepsilon > 0\), take \(a := f(y) - f(\tilde{x}) + \varepsilon > f(y) - f(\tilde{x})\). So we can use (2.10.71) for this choice of \(y\) and \(a\), to get

\[
\langle -\frac{u}{b}, y - \tilde{x} \rangle - [f(y) - f(\tilde{x}) + \varepsilon] \leq 0,
\]

which rewrites as

\[
f(y) + \varepsilon \geq f(\tilde{x}) + \langle -\frac{u}{b}, y - \tilde{x} \rangle.
\]
Because $\varepsilon > 0$ is arbitrary we can take $\varepsilon \downarrow 0$ to conclude that
\[ f(y) \geq f(\bar{x}) + \langle -\frac{u}{b}, y - \bar{x} \rangle, \quad \text{for all } y \in \mathbb{R}^n. \]
This yields $\nabla f(\bar{x}) = -\frac{u}{b}$. Choosing now $y := \bar{x}$, $a > 0$ and $a' = 0$ in (2.10.69) gives
\[ ab \leq \langle u, y' - \bar{x} \rangle, \]
for all $a > 0$ and all $y' \in C$. Divide this expression by $-b > 0$ to get
\[ -a \leq \langle -\frac{u}{b}, y' - \bar{x} \rangle = \langle w, y' - \bar{x} \rangle, \]
for all $a > 0$ and all $y' \in C$. Because the above inequality holds for all $a > 0$ we conclude that
\[ 0 \leq \langle w, y' - \bar{x} \rangle, \]
for all $y' \in C$. So (2.10.67) holds for $w := -\frac{u}{b} \in \partial f(\bar{x})$. The proof is complete. \qed

The next corollary is a consequence of Theorem 2.10.26 and Corollary 2.10.10.

**Corollary 2.10.10** Assume that $\emptyset \neq C \subset \mathbb{R}^n$ is convex and open, and let $f : \mathbb{R}^n \to \mathbb{R}$ be convex. A point $\bar{x} \in C$ is a solution of (2.10.63) if and only if $0 \in \partial f(\bar{x})$.

**Proof.** By Theorem 2.10.26, $\bar{x}$ is a solution of (2.10.63) if and only if $\bar{x} \in C$ and there exists $w \in \partial f(\bar{x})$ such that
\[ \langle w, y - \bar{x} \rangle \geq 0, \quad \text{for all } y \in C. \]
Assume, for contradiction purposes, that $0 \notin \partial f(\bar{x})$. Hence, the element $w \in \partial f(\bar{x})$ for which the above expression holds must be nonzero. Because $C$ is an open set and $\bar{x} \in C$, there exists $r > 0$ such that $B(\bar{x}, r) \subset C$. In particular, the point $y := \bar{x} - r\frac{w}{\|w\|} \in B(\bar{x}, r)$ and is well defined because $w \neq 0$. The above inequality for this choice of $y$ yields
\[ -r\|w\| \geq 0, \]
so $w = 0$, a contradiction. Therefore we must have $0 \in \partial f(\bar{x})$. Conversely, if $0 \in \partial f(\bar{x})$, then (2.2.8) gives
\[ f(y) \geq f(\bar{x}) + \langle 0, y - \bar{x} \rangle = f(\bar{x}), \]
for all $y \in \mathbb{R}^n$, which implies that $\bar{x}$ is a global minimizer of $f$. Because $\bar{x} \in C$, we conclude that $\bar{x}$ is a solution of (2.10.63). \qed

The next corollary is a consequence of Theorem 2.10.26 and Corollary 2.10.10.

**Corollary 2.10.11** Assume that $\emptyset \neq C \subset \mathbb{R}^n$ is convex, and let $f : \mathbb{R}^n \to \mathbb{R}$ be convex and differentiable. A point $\bar{x} \in C$ is a solution of (2.10.63) if and only if $\langle \nabla f(\bar{x}), y - \bar{x} \rangle \geq 0$ for all $y \in C$. If, additionally, $C$ is open, then $\bar{x}$ is a solution of (2.10.63) if and only if $\nabla f(\bar{x}) = 0$.

**Exercise 2.10.42** Prove Corollary 2.10.11.

Theorem 2.10.26 and its corollaries have important applications for devising methods for solving problem (2.10.63). Indeed, assume that a current iterate $x$ is not a solution of problem (2.10.63), then we must have $\langle \nabla f(x), y - x \rangle < 0$ for some $y \in C$. We want to move to a point $x^+$ such that $f(x^+) < f(x)$. From Taylor’s development we have for $t \geq 0$
\[ f(x + t(y - x)) = f(x) + t\langle \nabla f(x), y - x \rangle + o(t) = f(x) + t\langle \nabla f(x), y - x \rangle + \frac{o(t)}{t}, \]
where \( \lim_{t \to 0^+} \frac{\alpha(t)}{t} = 0 \). Because \( \langle \nabla f(x), y - x \rangle < 0 \), we have that for \( t \) small enough the expression between brackets must be negative, so for \( t \) small enough we have \( f(x + t(y - x)) < f(x) \). In other words, the direction \( d := y - x \) is a good direction for getting a smaller value of \( f \). A common procedure is to choose \( t \) according to the rule

\[
t^+ \in \text{Argmin}_{t \geq 0, x + t(y - x) \in C} f(x + t(y - x)),
\]

and take \( x^+ := x + t^+(y - x) \). This procedure is called the Method of feasible directions.

**Example 2.10.7** Let us model the curve fitting problem of section 1.2.1 using the “least square deviation” objective function. Therefore, we want to solve the optimization problem

Minimize \( \sum_{i=1}^{M} (m t_i + b - s_i)^2 \),

subject to \((m; b) \in \mathbb{R}^2\).

Because the problem is unconstrained and the objective function is strictly convex and coercive, a solution exists and it is unique. Suppose that we got the measurements \((0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 0)\). In this case, the objective function becomes

\[
f(m, b) = \sum_{i=0}^{4} (i m + b - i)^2 + (5m + b)^2,
\]

Because \( f \) is convex and differentiable, we can quote Corollary 2.8.8 and find the solution by computing the unique stationary point of \( f \). So we want to find \((m, b)\) such that

\[
\nabla f(m, b) = \left[ \frac{\sum_{i=0}^{4} 2(i m + b - i)i + 10(5m + b)}{\sum_{i=0}^{4} 2(i m + b - i) + 2(5m + b)} \right] = \left[ 0 \right].
\]

It is left to the reader to check that the unique solution of this linear system is \((m, b) = (2/7, 20/21)\). The solution we found, which is depicted in Figure 2.10.18 indicates that the “noisy” measurement \((5, 0)\) has strong influence on the solution.

![Figure 2.10.18: The least square solution](image-url)
Assume we model the same problem by looking for the “Least total error solution”. In this case, the problem takes the form

\[
\text{Minimize } \sum_{i=1}^{M} |mt_i + b - s_i| = \sum_{i=0}^{4} |i m + b - i| + |5m + b|,
\]

subject to \((m, b) \in \mathbb{R}^2\).

Define \(g(m, b) := \sum_{i=0}^{3} |i m + b - i| + |5m + b|\). Because we have in our case \(C = \mathbb{R}^n\), we can apply Corollary 2.10.10 to conclude that a solution \((m^*, b^*)\) should verify \((0, 0) \in \partial g(m^*, b^*)\). In order to compute the set \(\partial g(m, b)\) we need the following theorem, which asserts that the subdifferential of a sum of convex functions at a given point is, under certain assumptions, equal to the sum of each of the subdifferentials at that point.

In Figure 2.10.19, we have that \(f_1(x) = |x|\) and \(f_2(x) = x - 1\) when \(x \geq 0\) and \(f_2(x) = x^2 - 1\) when \(x \leq 0\). So that \((f_1 + f_2)(x) = x^2 - x - 1\) when \(x \leq 0\) and \((f_1 + f_2)(x) = 2x - 1\). We see by direct calculation that \(\partial f_1(0) = [-1, 1]\), \(\partial f_2(0) = [0, 1]\), and \(\partial (f_1 + f_2)(0) = [-1, 2]\). So we have in this case that \(\partial (f_1 + f_2)(0) = \partial f_1(0) + \partial f_2(0)\).

![Figure 2.10.19: \([-1, 2] = \partial (f_1 + f_2)(0) = \partial f_1(0) + \partial f_2(0) = [-1, 1] + [0, 1]\).](image)

Our next theorem proves that the property depicted in Figure 2.10.19 holds under some assumptions on \(f_1\) and \(f_2\).

**Theorem 2.10.27** Assume that \(f_1, f_2 : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}\) are convex.

(i) It always holds that

\[
\partial f_1(x) + \partial f_2(x) \subset \partial (f_1 + f_2)(x), \text{ for all } x \in \text{dom } f_1 \cap \text{dom } f_2.
\]

(ii) If there exists a point in \(\text{dom } f_1 \cap \text{dom } f_2\) at which \(f_1\) is continuous, then

\[
\partial f_1(x) + \partial f_2(x) = \partial (f_1 + f_2)(x),
\]

for all \(x \in \text{dom } f_1 \cap \text{dom } f_2\).

**Proof.** Let us prove (i). Fix \(\bar{x} \in \text{dom } f_1 \cap \text{dom } f_2\). If any of the sets \(\partial f_1(\bar{x})\) or \(\partial f_2(\bar{x})\) is empty, the the sum \(\partial f_1(\bar{x}) + \partial f_2(\bar{x})\) must also be empty (the sum of empty set with any set must be the empty set). Therefore (2.10.72) trivially holds in this case. So we can assume that both sets \(\partial f_1(\bar{x})\) and \(\partial f_2(\bar{x})\) are nonempty. In other words, there exist \(v_1 \in \partial f_1(\bar{x})\) and \(v_2 \in \partial f_2(\bar{x})\). From the definition of subdifferential we can write for all \(y \in \mathbb{R}^n\)

\[
\begin{align*}
 f_1(y) &\geq f_1(\bar{x}) + \langle v_1, y - \bar{x} \rangle, \\
 f_2(y) &\geq f_2(\bar{x}) + \langle v_2, y - \bar{x} \rangle,
\end{align*}
\]
and summing up both inequalities we obtain for all \( y \in \mathbb{R}^n \)
\[
(f_1 + f_2)(y) = f_1(y) + f_2(y) \geq f_1(\bar{x}) + f_2(\bar{x}) + \langle v_1 + v_2, y - \bar{x} \rangle,
\]
\[
= (f_1 + f_2)(\bar{x}) + \langle (v_1 + v_2), y - \bar{x} \rangle,
\]
and hence \( v_1 + v_2 \in \partial(f_1 + f_2)(\bar{x}) \). We have proved that \( \partial f_1(\bar{x}) + \partial f_2(\bar{x}) \subset \partial(f_1 + f_2)(\bar{x}) \), and the proof of (i) is complete. Let us prove now (ii). In view of (i), it is enough to prove that
\[
\partial f_1(\bar{x}) + \partial f_2(\bar{x}) \supset \partial(f_1 + f_2)(\bar{x}),
\]
(2.10.73)
for every \( \bar{x} \in \text{dom} f \cap \text{dom} g \). There are two cases to consider. The first case is \( \partial(f_1 + f_2)(\bar{x}) = \emptyset \). In this case, (2.10.73) trivially holds. Let us consider now the second case: \( \partial(f_1 + f_2)(\bar{x}) \neq \emptyset \) and fix \( v_0 \in \partial(f_1 + f_2)(\bar{x}) \). Consider the functions
\[
g_1(x) = f_1(x + \bar{x}) - f_1(\bar{x}) - \langle x, v_0 \rangle,
\]
and
\[
g_2(x) = f_2(x + \bar{x}) - f_2(\bar{x}).
\]
Note that \( g_1(0) = 0 = g_2(0) \), \( \text{dom} g_1 = \{ z - \bar{x} : z \in \text{dom} f_1 \} = \text{dom} f_1 - \bar{x} \), and \( \text{dom} g_2 = \{ z - \bar{x} : z \in \text{dom} f_2 \} = \text{dom} f_2 - \bar{x} \). It can also be verified that
(a) If \( v_0 \in \partial(f_1 + f_2)(\bar{x}) \), then \( 0 \in \partial(g_1 + g_2)(0) \).
(b) If \( 0 \in \partial g_1(0) + \partial g_2(0) \) then \( v_0 \in \partial f_1(\bar{x}) + \partial f_2(\bar{x}) \).

We prove next item (b), and leave (a) as an exercise. Assume that \( 0 \in \partial g_1(0) + \partial g_2(0) \), so there exists \( w \in \partial g_1(0) \) such that \( -w \in \partial g_2(0) \) (so that the sum of both is zero). This means that for every \( x \in \mathbb{R}^n \) we can write
\[
g_1(x) \geq \langle w, x \rangle, \quad \text{and} \quad g_2(x) \geq \langle -w, x \rangle,
\]
where we also used that \( g_1(0) = 0 = g_2(0) \). Using now the definitions of \( g_1 \) and \( g_2 \) we can rewrite the above inequalities as
\[
f_1(x + \bar{x}) - f_1(\bar{x}) - \langle x, v_0 \rangle \geq \langle w, x \rangle,
\]
\[
f_2(x + \bar{x}) - f_2(\bar{x}) \geq \langle -w, x \rangle,
\]
for every \( x \in \mathbb{R}^n \). Calling \( y := x + \bar{x} \) and noting that \( y \in \mathbb{R}^n \) is arbitrary we can rewrite the above expressions as
\[
f_1(y) \geq f_1(\bar{x}) + \langle v_0 - w, y - \bar{x} \rangle,
\]
\[
f_2(y) \geq f_2(\bar{x}) + \langle w, y - \bar{x} \rangle,
\]
so \( v_0 - w \in \partial f_1(\bar{x}) \) and \( w \in \partial f_2(\bar{x}) \) which gives \( v_0 = [v_0 - w] + w \in \partial f_1(\bar{x}) + \partial f_2(\bar{x}) \), as claimed.

Our next step is to prove that \( 0 \in \partial g_1(0) + \partial g_2(0) \). If this claim is true, then item (b) yields \( v_0 \in \partial f_1(\bar{x}) + \partial f_2(\bar{x}) \) and because \( v_0 \in \partial f_1 + f_2(\bar{x}) \) is arbitrary we get \( \partial(f_1 + f_2)(\bar{x}) \subset \partial f_1(\bar{x}) + \partial f_2(\bar{x}) \), as wanted. So, let us prove the claim that \( 0 \in \partial g_1(0) + \partial g_2(0) \). Using (a) and the assumption that \( v_0 \in \partial(f_1 + f_2)(\bar{x}) \), we have \( 0 \in \partial(g_1 + g_2)(0) \), and using also \( g_1(0) = g_2(0) = 0 \) it follows that
\[
(g_1 + g_2)(x) \geq (g_1 + g_2)(0) = 0,
\]
(2.10.74)
for every \( x \in \mathbb{R}^n \). Because \( f_1 \) is continuous at a point, then \( g_1 \) must also be continuous at a point. Corollary 2.10.9 yields \( (\text{Epi } g_1)^o \neq \emptyset \) and \( (\text{dom } g_1)^o \neq \emptyset \). Call \( C_1 := \text{Epi } g_1 \) and \( C_2 := \{(y,t) \in \mathbb{R}^n \times \mathbb{R} \mid t \leq -g_2(y) \} \). By (2.10.74), \( (C_1)^o \cap C_2 = \emptyset \). Indeed, assume that \( (y,t) \in (C_1)^o \cap C_2 \). Note that \( (y,t) \in (C_1)^o \subset \text{Epi } g_1 \), so we have \( g_1(y) \leq t \). If \( g_1(y) = t \), then \( (y,t) = (y,g_1(y)) \) which we
know is a boundary point of $Epi \ g_I$. This contradicts the fact that $(y, t) \in (C_1)^o = (Epi \ g_I)^o$. So we must have $g_1(y) < t$. Therefore, we can write
\[
g_1(y) < -g_2(y),
\]
where we used the inclusion $(y, t) \in C_2$ in the rightmost inequality. The above expression implies that $g_1(y) + g_2(y) < 0$, contradicting (2.10.74). Therefore, we have $(C_1)^o \cap C_2 = \emptyset$. Using now the separation Theorem 2.5.8 for the disjoint sets $(C_1)^o, C_2 \subset \mathbb{R}^{n+1}$, there exists $(p, a) \in \mathbb{R}^{n+1}$, $(p, a) \neq (0, 0)$ such that
\[
\langle p, y \rangle + a t \leq \langle p, y' \rangle + a t',
\]
for all $(y, t) \in (Epi \ g_I)^o = (C_1)^o$ and all $(y', t')$ such that $t' \leq -g_2(y')$. First note that the left inequality in (2.10.75) can be extended to the whole set $C_1$ by taking limits. So (2.10.75) holds for all $(y, t) \in Epi \ g_I$ and all $(y', t')$ such that $t' \leq -g_2(y')$. By letting $t \to +\infty$ in (2.10.75) we see that $a \leq 0$. We claim that $a < 0$. Indeed, if $a = 0$ then $p \neq 0$, because $(p, a) \neq (0, 0)$. Using (2.10.75) for $(y', t') = (0, 0) \in C_2$ we get $\langle p, y \rangle \leq 0$ for all $y \in dom \ g_1$ and using (2.10.75) for $(y, t) = (0, 0) \in Epi \ g_I$ we get $\langle p, y' \rangle \geq 0$ for all $y' \in dom \ g_2$. Hence the convex sets dom $g_1$ and dom $g_2$ are separated by the nontrivial hyperplane \{$x \in \mathbb{R}^n | \langle p, x \rangle = 0$\}. By assumption, $g_1$ is continuous at some point $z \in dom \ g_1 \cap dom \ g_2$, so we can invoke Corollary 2.10.9 to conclude that $z \in (dom \ g_1)^o \cap dom \ g_2$. Because $z \in dom \ g_1 \cap dom \ g_2$, we must have $\langle p, z \rangle = 0$. Since $z \in (dom \ g_1)^o$, there exists $\delta > 0$ such that $z + \delta p \in dom \ g_1$. Thus, $0 \geq \langle p, z + \delta p \rangle = \delta \|p\|^2 > 0$, a contradiction. Therefore, it holds that $a < 0$ and without any loss of generality, we can assume that $a = -1$ (otherwise, divide (2.10.75) by $-a > 0$). Fix $x \in dom \ g_1 \cap dom \ g_2$. We note now that:

(c) The leftmost inequality in (2.10.75) for $(x, g_1(x)) \in Epi \ g_I$ and $(0, 0) \in C_2$ yields $p \in \partial \ g_1(0)$, while the rightmost one for $(x, -g_2(x)) \in C_2$ and $(0, 0) \in Epi \ g_I$ gives $-p \in \partial \ g_2(0)$.

Therefore, $0 = p + [-p] \in \partial \ g_1(0) + \partial \ g_2(0)$. The claim is established. Now we invoke (b) for concluding that $v_0 \in \partial f_1(x_0) + \partial f_2(x_0)$, completing the proof.

**Exercise 2.10.43** Prove statements (a) and (c) in the proof of Theorem 2.10.27.

**Remark 2.10.14** The fact that Theorem 2.10.27(ii) may not hold in general can be illustrated by considering $f_1$ and $f_2$ as the indicator functions of two balls $B_1$ and $B_2$ intersecting only at a boundary point $z$ (see Figure 2.10.16). Note that $\delta B_1 + \delta B_2 = \delta B_1 \cap B_2$. It holds that $\mathbb{R}^n = N_{(z)}(z) = \partial \delta B_1 \cap B_2(z) = \partial (\delta B_1 + \delta B_2)(z) \supseteq \partial B_1(z) + \partial B_2(z) = N_{C_1}(z) + N_{C_2}(z)$. We must require extra conditions for the opposite inclusion in (2.10.72) to hold. The most common of these conditions is the one used in Theorem 2.10.27(ii).

**Remark 2.10.15** Note that Theorem 2.10.27(ii) can be extended to a finite sum of $m$ convex functions $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ such that $\cap_{i=1}^m (\text{dom } f_i)^o \neq \emptyset$.

**Example 2.10.8** Let us go back to the nondifferentiable problem at the end of Example 2.10.7. Let us compute the set $\partial (\sum_{i=0}^4 |i m + b - i| + |5 m + b|)$. Call $g_i(m, b) := |i m + b - i|$ for $i = 0, 1, 2, 3, 4$ and $g_5(m, b) := |5 m + b|$. So the objective function of the nondifferentiable problem becomes $g(m, b) := \sum_{i=0}^5 g_i(m, b)$. Note that all $g_i$s are everywhere continuous, so dom $g_i = \mathbb{R}^n$ for $i = 0, 1, 2, 3, 4$ and hence from Theorem 2.10.27(ii) and Remark 2.10.15 we can write
\[
\partial g(m, b) = \partial \left( \sum_{i=0}^4 |i m + b - i| + |5 m + b| \right) = \sum_{i=0}^5 \partial g_i(m, b).
\]

It can be checked for $i = 0, 1, 2, 3, 4$ that
\[
\partial g_i(m, b) := \begin{cases} (i, 1) & \text{ if } i m + b - i > 0, \\ (-i, -1) & \text{ if } i m + b - i < 0, \\ \{\lambda(i, 1) : \lambda \in [-1, 1]\} & \text{ if } i m + b - i = 0. 
\end{cases}
\]
Note that the case \( i \, m + b - i = 0 \) gives the convex combination of the other two cases. We also have

\[
\partial g_5(m, b) := \begin{cases} 
(5, 1) & \text{if } 5 \, m + b > 0, \\
(-5, -1) & \text{if } 5 \, m + b < 0, \\
\{ \lambda(5, 1) : \lambda \in [-1, 1] \} & \text{if } 5 \, m + b = 0.
\end{cases}
\]

We claim that

(I) \((0, 0) \in \partial g(1, 0)\),

(II) \((0, 0) \notin \partial g(2/7, 20/21)\)

Item (I) means that the solution found by solving the nondifferentiable problem is the one which ignores the “noisy” data \((5, 0)\). See Figure 2.10.20, where both solutions are depicted. Item (II) means that the solution found by using the least square error is not a minimizer of \(g\). We will prove here (I), and leave the proof of (II) as an exercise. In order to establish (I), note that for \((m, b) = (1, 0)\) we have \(i \, m + b - i = 0\) for all \(i = 1, \ldots, 4\) so \(\lambda_i(i, 1) \in \partial g_i(1, 0)\) for every \(\lambda_i \in [-1, 1]\). Because \(5 \, m + b = 5 > 0\) we have \(\partial g_5(1, 0) = (5, 1)\). So the optimality condition \((0, 0) \in \partial g(1, 0) = \sum_{i=0}^4 \partial g_1(1, 0) + \partial g_5(1, 0)\) can be written as

\[
(0, 0) = \sum_{i=0}^4 \lambda_i(i, 1) + (5, 1),
\]

for some \(\lambda_i \in [-1, 1]\). The reader can now check that the above equality holds for the choices \(\lambda_0 = \lambda_1 = 1/4\) and \(\lambda_3 = \lambda_4 = -3/4\). These values of \(\lambda_i\) are obtained by looking at the convex set \(\Omega := \{ \sum_{i=0}^4 \lambda_i(i, 1) : \lambda_i \in [-1, 1] \}\) which is the shaded area (including the boundary) depicted in Figure 2.10.21. The optimality condition above then becomes \(- (5, 1) \in \Omega\). Indeed, \(- (5, 1) \in \Omega\) because it is the convex combination of the points \((-7, -2), (1, 2) \in \Omega\).

**Exercise 2.10.44** With the notation of Example 2.10.8, prove the following statements.

(i) The points \((-7, -2), (1, 2)\) belong to the set \(\Omega\) and \((-5, 1)\) is a convex combination of them.

(ii) The point \((0, 0) \notin \partial g(2/7, 20/21)\). Hint: check that \(g(2/7, 20/21) > g(1, 0)\).
2.11 Maximization of a Convex Function

We consider here the problem
\[ \max_{x \in C} h(x), \tag{2.11.76} \]
where \( C \subset \mathbb{R}^n \) and \( h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\} \) is convex. Note that this is not a convex problem, because it is equivalent to
\[ \min_{x \in C} (-h)(x), \]
which is the minimization of a concave function over a convex set.

We see from Figure 2.11.22 that all solutions of problem (2.11.76), if they exist, they must lie in the boundary of \( C \). We prove this fact next.

**Theorem 2.11.28** Assume that \( h : \mathbb{R}^n \rightarrow \mathbb{R} \cup \{+\infty\} \) is not constant on \( C \) and problem (2.11.76) has solutions. Then all the solutions are at the boundary of \( C \).

**Proof.** Fix \( \hat{x} \in C \) a solution of problem (2.11.76). Because \( h \) is not constant over \( C \), there exists \( \hat{x} \in C \) such that \( h(\hat{x}) < h(\hat{x}) \). By Definition A.6.25(i) we have that \( C = C^o \cup \partial C \). If \( C^o = \emptyset \) then \( C = \partial C \) and therefore the conclusion of the theorem trivially holds. So we can assume that \( C^o \neq \emptyset \). Take \( z \in C^o \). We claim that \( z \) cannot solve (2.11.76). In order to prove the claim, we will show that \( h(z) < h(\hat{x}) \). Because \( z \in C^o \) there exists \( r > 0 \) such that \( B(z, r) \subset C \). Take \( y := z + r \frac{\hat{x} - z}{\|\hat{x} - z\|} \).
Note that \( \|y - z\| = r \) so \( y \in B(z, r) \subset C \). Moreover, let \( a := \frac{1}{1 + \|z - y\|} \in (0, 1) \). So \( 1 - a = \frac{r}{r + \|z - y\|} \).

With these definitions, it is easy to check that \( z = ay + (1 - a)\tilde{x} \). Using now the convexity of \( h \) we get
\[
h(z) \leq ah(y) + (1 - a)h(\tilde{x}) < ah(\tilde{x}) + (1 - a)h(\tilde{x}) = h(\tilde{x}),
\]
where we also used the fact that \( \tilde{x} \) solves (2.11.76) and the assumption on \( \tilde{x} \). The above expression implies that \( z \) cannot be a solution of (2.11.76). The proof is complete. \( \square \)
Chapter 3

Methods

3.1 Methods for Unconstrained Problems

We have seen until now optimality conditions for finding the unconstrained minimum of a given function. In practice, to check these conditions may be difficult or even impossible. In fact, to locate all stationary points we have to solve \textit{exactly} the nonlinear system of equations $\nabla f(x) = 0$. To analyze the sign of the Hessian at these solutions is even more difficult in practice. In these cases, the way of solving the problem is to produce a sequence $\{x_k\}$ (by some computational routine) that (under some reasonable conditions) converges to a minimum of $f$. Clearly, the iterative method makes sense only when the computation of each iterate $x_k$ is considerably easier that the computation of the minimizers of $f$. Typically, iterative methods start at an initial guess $x_0$ and recursively find a point $x_{k+1}$ using the information of the previous point $x_k$. Based in the point $x_k$, the method chooses

1. a nonzero \textit{search direction} $p_k$, and
2. a (positive) \textit{step-size} $\alpha_k$.

Then the next iterate is given by

$$x_{k+1} := x_k + \alpha_k p_k.$$  \hfill (3.1.1)

The particular way of choosing the search direction and the step-size is what characterizes a given method. Usually, the direction $p_k$ and the step size $\alpha_k$ are chosen in such a way that

$$f(x_{k+1}) = f(x_k + \alpha_k p_k) < f(x_k).$$ \hfill (3.1.2)

In other words, when searching for a minimum of $f$, we ask the next point to decrease the value of $f$. These methods are called \textit{monotone methods} or \textit{methods of iterative descent}. Not for every direction $p \in \mathbb{R}^n$ it is possible to find a step size $\alpha$ such that (3.1.2) holds.

**Lemma 3.1.10** Let $f : \mathbb{R}^n \to \mathbb{R}$ be $C^2$ and fix $0 \neq p \in \mathbb{R}^n$. Assume that $\nabla f(x)^T p \neq 0$. Then the following assertions are equivalent.

(a) There exists $\alpha > 0$ such that $f(x + tp) < f(x)$ for all $t \in (0, \alpha]$.

(b) $\nabla f(x)^T p < 0$.

**Proof.** For a fixed $\lambda > 0$, Taylor’s formula gives

$$f(x + \lambda p) = f(x) + \lambda \nabla f(x)^T p + \frac{\lambda^2}{2} p^T \nabla^2 f(z) p = f(x) + \lambda \left[ \nabla f(x)^T p + \frac{\lambda}{2} p^T \nabla^2 f(z) p \right]$$
for some $z \in [x_0, x_0 + \lambda p]$. If (b) holds, then there exists $\alpha > 0$ small enough such that the expression between brackets is negative for all $\lambda \in (0, \alpha)$. This yields

$$f(x + \lambda p) < f(x),$$

for all $\alpha > \lambda > 0$. Assume now that (a) holds. Using (a) and Taylor’s formula, we can write

$$f(x + tp) = f(x) + t \left[ \nabla f(x)^T p + \frac{t}{2} p^T \nabla^2 f(z)p \right] < f(x),$$

for all $t \in (0, \alpha]$ and for some $z \in [x, x + tp]$. This implies that $\left[ \nabla f(x)^T p + \frac{t}{2} p^T \nabla^2 f(z)p \right] < 0$ for all $t \in (0, \alpha]$.

Let us define formally such directions.

**Definition 3.1.17** Let $f : \mathbb{R}^n \to \mathbb{R}$. A nonzero vector $p \in \mathbb{R}^n$ is called a descent direction for $f$ at $x$ when $\nabla f(x)^T p < 0$.

**Example 3.1.9** Let $f : \mathbb{R}^n \to \mathbb{R}$ and take $x \in \mathbb{R}^n$ such that $\nabla f(x) \neq 0$. Then the direction $p_1 := -\nabla f(x)$ is a descent direction. Assume now that $\nabla f(x) \neq 0$ and that $\nabla^2 f(x)$ is positive definite, then $p_2 := -\left[ \nabla^2 f(x) \right]^{-1} \nabla f(x)$ is a descent direction.

All existing iterative methods are derived from two basic methods which have historical importance. These are Newton’s method and the Steepest Descent method (also called Cauchy method). Both of them have good as well as bad features. We will start by studying these two methods, and then we will derive from them some known methods.

### 3.1.1 Newton’s Method

Suppose we want to minimize a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$. Taylor’s formula gives

$$f(x_0 + p) = f(x_0) + \nabla f(x_0)^T p + \frac{1}{2} p^T \nabla^2 f(z)p,$$

for some $z \in [x_0, x_0 + p]$. The idea of Newton’s method is to approximate the above expression by the quadratic function

$$m_f(p) = f(x_0) + \nabla f(x_0)^T p + \frac{1}{2} p^T \nabla^2 f(x_0)p \approx f(x_0 + p),$$

where the unknown $z$ of Taylor’s formula is replaced by the given point $x_0$. If this quadratic approximation $m_f$ is close to the original $f$, then it is reasonable to expect that the minimum of $m_f$ is close to the minimum of $f$, with the advantage that the computation of the minimum of a quadratic function is considerably easier! Indeed, the minimum $p^*$ of $m_f$ must verify the first order optimality condition and hence $0 = \nabla m_f(p^*) = \nabla f(x_0) + \nabla^2 f(x_0)p^*$, so $p^* = -\left[ \nabla^2 f(x_0) \right]^{-1} \nabla f(x_0)$, provided the matrix $\left[ \nabla^2 f(x_0) \right]^{-1}$ exists. Newton’s method thus takes the next iterate $x_1 := x_0 + p^* = x_0 - \left[ \nabla^2 f(x_0) \right]^{-1} \nabla f(x_0)$. Formally, given an initial iterate $x_0$, the pure Newton method is defined as: Given $x_k$, take $x_{k+1}$ as

$$x_{k+1} = x_k - \left[ \nabla^2 f(x_k) \right]^{-1} \nabla f(x_k). \quad (3.1.3)$$

So, we see that Newton method chooses $p^k = -\left[ \nabla^2 f(x_k) \right]^{-1} \nabla f(x_k)$ and $\alpha_k = 1$ in (3.1.1). The step is called pure because of the choice $\alpha_k = 1$. The direction $p^k := -\left[ \nabla^2 f(x_k) \right]^{-1} \nabla f(x_k)$ is called the Newton direction. From Example 3.1.9, we see that the Newton direction is a descent direction when the Hessian $\left[ \nabla^2 f(x_k) \right]$ is positive definite.

The preceding discussion implies that particularly good steps are expected when the original function is quadratic.
Theorem 3.1.29 Let $A \in \mathbb{R}^{n \times n}$ be positive definite, $b \in \mathbb{R}^n$ and $a \in \mathbb{R}$. Then the quadratic function
\[ f(x) := a + b^T x + \frac{1}{2} x^T A x, \quad (3.1.4) \]
is strictly convex and has a unique global minimizer at $x^* := -A^{-1} b$. Moreover, for every initial point $x_0 \in \mathbb{R}^n$, Newton’s method reaches this minimum in only one step, that is, $x_1 = x^*$.

Proof. We leave the first statement of the theorem as an exercise. As for the last one, note that
\[ x_1 = x_0 - [\nabla^2 f(x_0)]^{-1} \nabla f(x_0) = x_0 - A^{-1} (Ax_0 + b) = -A^{-1} b = x^*. \]
The point $x^*$ is a global minimum because the Hessian $\nabla^2 f(x^*) = A$ is positive definite. □

When the function $f$ is not as in Theorem 3.1.29, the sequence generated by Newton’s method may not converge, and if converges, it may not converge to a minimum. For instance, if $f$ is quadratic with negative definite matrix $A$ in (3.1.4), then Newton’s sequence converges to a global maximum in one step (the proof is the same as the one given in Theorem 3.1.29, but the point $x_1 = x^*$ is a global maximum because the Hessian of $f$ is always the negative definite matrix $A$).

If the function $f$ is such that the Hessian $\nabla^2 f(x_k)$ is not defined for some iterate $x_k$, then the $k+1$-step in (3.1.3) cannot be performed. Moreover, even when this Hessian is well defined for all values of $k$, the Newton direction $p_k$ may not be a descent direction. We study these pathologies in the following section.

Exercise 3.1.45 Construct the Newton sequence for the following functions

1. $f(x_1, x_2) = x_1^4 + 2x_1^2 x_2^2 + x_2^4$ with the choices $x_0 = (1, 1)$ and $x_0 = (1, 0)$.
2. $f(t) = t^4 - 32t^2$ and $t_0 = 1$.
3. $f(t) = t^4 - 32t^2$ and $t_0 = 3$.
4. $f(t) = 2/3|t|^{3/2}$ and $t_0 = 1$.

In case (1), find $x_{k+1}$ by solving the linear system
\[ \nabla^2 f(x_k) (x_{k+1} - x_k) = -\nabla f(x_k). \]

3.1.2 Implementation of Newton’s Method

Formula (3.1.3) is never used for finding the next iterate $x_{k+1}$, because the computation of the inverse of the Hessian is usually very expensive. Instead, given the current iterate $x_k$ and the current Hessian $\nabla^2 f(x_k)$, the following linear system is solved on the unknown $x_{k+1}$:
\[ \nabla^2 f(x_k) (x_{k+1} - x_k) = -\nabla f(x_k). \]

This system can be solved for Gaussian elimination, or by LU or Choleski factorization when the Hessian is positive definite. Ideally, close to the minimum $x^*$ of $f$, the Hessian is likely to be positive definite, and hence the Hessian will still be positive definite in certain neighborhood of the point $x^*$. Hence, if the minimum of $f$ is such that $\nabla^2 f(x^*)$ is positive definite, Newton’s method will be well defined in certain neighborhood of the point $x^*$, and Newton’s direction will be a descent direction on that neighborhood. Observe that, even for strictly convex functions, we may not have a positive definite matrix at $x^*$ (consider the function $f : \mathbb{R} \rightarrow \mathbb{R}$ defined as $f(t) = t^2$). If we are far from the minimum, then convergence of the method may not occur. Check this fact in Exercise 3.1.45(2) and (4).
When using Newton method close to a critical point, the rapidity with which it converges is not an accidental feature. It can be proved that, when close enough to a critical point \( x^* \), there exists a constant \( C > 0 \) such that
\[
\| x_{k+1} - x^* \| < C \| x_k - x^* \|^2,
\]
for all \( k \) large enough. This important property is called quadratic convergence. We prove this fact in the next section.

### 3.1.3 Quadratic Convergence of Newton Method

In order to prove quadratic convergence of Newton’s method, we make use of some standard facts of Numerical Calculus. The norm of a matrix \( A \in \mathbb{R}^{n \times n} \) is defined as:
\[
\| A \| := \max_{\| x \| \leq 1} \| Ax \|.
\]
This readily implies that
\[
\| Ax \| \leq \| A \| \| x \|.
\]
It is well-known that the Fundamental Theorem of Integral Calculus states that
\[
\varphi(a) - \varphi(b) = \int_0^1 \varphi'(b + t(a - b))(a - b)dt = \int_a^b \varphi'(t)dt.
\]
We need to extend this fact to functions from \( \mathbb{R}^n \) to \( \mathbb{R}^m \). We define first the integral of a vector-valued function.

**Definition 3.1.18** A function \( G : \mathbb{R} \to \mathbb{R}^m \) can be described as \( G(t) := (g_1(t), \ldots, g_m(t))^T \), where \( g_i : \mathbb{R} \to \mathbb{R} \) for \( i = 1, \ldots, n \). Then \( \int_0^1 G(t)dt \) is the vector of \( \mathbb{R}^m \) with components
\[
\left[ \int_0^1 G(t)dt \right]_i := \int_0^1 g_i(t)dt.
\]
for all \( i = 1, \ldots, m \).

**Example 3.1.10** Fix \( x, y \in \mathbb{R}^n \) and take a twice continuously differentiable \( f : \mathbb{R}^n \to \mathbb{R} \). Define \( G : \mathbb{R} \to \mathbb{R}^n \) as \( G(t) := \nabla^2 f(x + t(y - x))(y - x) \). Then \( \int_0^1 \nabla^2 f(x + t(y - x))(y - x)dt \in \mathbb{R}^n \) has for components
\[
\left[ \int_0^1 \nabla^2 f(x + t(y - x))(y - x)dt \right]_i := \int_0^1 \sum_{j=1}^n \frac{\partial^2 f(x + t(y - x))}{\partial x_i \partial x_j} (x_j - y_j)dt,
\]
for all \( i = 1, \ldots, n \).

Now we can state and prove the Fundamental Theorem of Calculus for vector-valued functions.

**Theorem 3.1.30** Let \( G : \mathbb{R}^n \to \mathbb{R}^n \) continuously differentiable on a convex set \( D \) and given by \( G(x) := (g_1(x), \ldots, g_m(x))^T \). Then for all \( x, y \in D \), it holds that
\[
G(y) - G(x) = \int_0^1 J_G(x + t(y - x))(y - x)dt. \tag{3.1.5}
\]
Proof. Fix $x, y \in D$ and for $i = 1, \ldots, n$ define the real-valued functions $\varphi_i(t) := g_i(x + t(y - x))$ for $t \in [0, 1]$. We can now apply the Fundamental Theorem of Integral Calculus to each of these real-valued functions, to get

$$\varphi_i(1) - \varphi_i(0) = \int_0^1 \varphi_i'(t) dt. \tag{3.1.6}$$

A simple calculation shows that

$$\varphi_i'(t) = \nabla g_i(x + t(y - x))^T(x - y) = \sum_{j=1}^n \frac{\partial g_i(x + t(y - x))}{\partial x_j}(x_j - y_j).$$

Combine this fact with (3.1.6) to get (3.1.5). □

It is well-known for a continuous real-valued function $\varphi : \mathbb{R} \to \mathbb{R}$ that

$$| \int_0^1 \varphi(t) dt | \leq \int_0^1 | \varphi(t) | dt.$$

This result can as well be extended to vector-valued functions.

**Theorem 3.1.31** Let $G : \mathbb{R} \to \mathbb{R}^n$ be continuous. Then

$$\| \int_a^b G(t) dt \| \leq \int_a^b \| G(t) \| dt. \tag{3.1.7}$$

**Proof.** We give bellow a sketch of the proof. Fix $\varepsilon > 0$ arbitrary and use the classical definition of Riemann integral to find a partition $a < t_0 < t_1 < \cdots < t_p < b$ of $[a, b]$ which verifies

$$\| \int_a^b G(t) dt - \sum_{i=1}^p G(t_i)(t_i - t_{i-1}) \| < \varepsilon.$$

Hence,

$$\| \int_a^b G(t) dt \| < \varepsilon + \sum_{i=1}^p \| G(t_i)(t_i - t_{i-1}) \| \leq \varepsilon + \sum_{i=1}^p \| G(t_i) \| (t_i - t_{i-1}) < 2\varepsilon + \int_a^b \| G(t) \| dt.$$

Since this inequality holds for all $\varepsilon > 0$, we conclude (3.1.7).

□

Finally, we need an often-used estimate.

**Theorem 3.1.32** Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously twice differentiable such that its Hessian verifies a Lipschitz condition in some convex set $D$, in other words:

$$\| \nabla^2 f(u) - \nabla^2 f(v) \| \leq C \| u - v \|,$$

for all $u, v \in D$. Then, for all $x, y \in D$,

$$\| \nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x) \| \leq \frac{C}{2} \| x - y \|^2.$$

**Proof.** Apply Theorem 3.1.30 to $G(x) := \nabla f(x)$ to get

$$\nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x) = \int_0^1 \left[ \nabla^2 f(x + t(y - x)) - \nabla^2 f(x) \right] (y - x) dt.$$

Taking norms in both sides and using Theorem 3.1.31, we can write

$$\| \nabla f(y) - \nabla f(x) - \nabla^2 f(x)(y - x) \| \leq \| y - x \| \int_0^1 \| \nabla^2 f(x + t(y - x)) - \nabla^2 f(x) \| dt \leq \frac{C}{2} \| y - x \|^2,$$

where we used the Lipschitz assumption on the Hessian and the fact that $\int_0^1 tdT = 1/2$. □

We have now all the tools we need to establish quadratic convergence of Newton's method.
Theorem 3.1.33 (Local Convergence of Newton’s Method) Let \( f : \mathbb{R}^n \to \mathbb{R} \) be three times continuously differentiable and assume that, at the local minimum \( x^* \), the Hessian \( \nabla^2 f(x^*) \) is positive definite. Then, if started sufficiently close to \( x^* \), the sequence generated by Newton’s method converges to \( x^* \). The order of convergence is at least two.

**Proof.** Our first step is to prove that there exists a neighborhood of \( x^* \) such that, if we choose \( x_0 \) in that neighborhood, then all subsequent iterates exist and remain within that neighborhood. In other words, there exists a radius \( r > 0 \) such that, whenever \( x_0 \in B[x^*, r] \), it holds that \( x_k \) is well defined and belongs to \( B[x^*, r] \) for all \( k \). Second, we will prove that, when \( x_0 \in B[x^*, r] \), then
\[
\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} \leq C_0.
\]

Since the Hessian \( \nabla^2 f(x^*) \) is positive definite, we can find \( R_1 \in (0, 1) \) such that for all \( z \in B[x^*, R_1] \), the corresponding Hessian \( \nabla^2 f(z) \) is also positive definite. Define the function \( G : \mathbb{R}^n \to \mathbb{R}^n \) as \( G(x) := x - [\nabla^2 f(x)]^{-1} \nabla f(x) \). Differentiating \( G \) formally we get
\[
J_G(x) = I - J_{\nabla^2 f(x)}^{-1} \nabla f(x) = -J_{\nabla^2 f(x)}^{-1} \nabla^2 f(x).
\]

Hence,
\[
J_G(x^*) = -J_{\nabla^2 f(x^*)}^{-1} \nabla f(x^*) = 0 \quad \text{and} \quad G(x^*) = x^*,
\]
where we used the fact that \( \nabla f(x^*) = 0 \). (For a rigorous computation of \( J_G(x) \), see James Ortega, *Numerical Analysis, a second course*, page 146). By definition of Jacobian, there exists \( \delta \in (0, R_1) \) such that
\[
\frac{\|G(x) - G(x^*) - J_G(x^*)(x - x^*)\|}{\|x - x^*\|} < R_1,
\]
for all \( x \in B[x^*, \delta] \). So, if \( x_k \in B[x^*, \delta] \), we can write
\[
\frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|^2} = \frac{\|G(x_k) - G(x^*)\|}{\|x_k - x^*\|} = \frac{\|G(x_k) - G(x^*) - J_G(x^*)(x_k - x^*)\|}{\|x_k - x^*\|} < R_1.
\]
But this yields \( \|x_{k+1} - x^*\| < R_1 \|x_k - x^*\| < R_1 \delta < \delta \), where we used that \( R_1 < 1 \) in the last inequality. Hence the first claim holds for \( r < \delta \).

From the above expression a quite stronger conclusion can be obtained, namely, that
\[
\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0,
\]
whenever \( x_0 \in B[x^*, r] \).

Indeed, noting that the whole sequence remains in \( B[x^*, \delta] \) and using iteratively that \( \|x_{k+1} - x^*\| < R_1 \|x_k - x^*\| \), we get
\[
\|x_{k+1} - x^*\| < R_1 \|x_k - x^*\| < (R_1)^2 \|x_{k-1} - x^*\| < \cdots < (R_1)^{k+1} \|x_0 - x^*\|,
\]
and using again that \( R_1 < 1 \), we obtain that \( \lim_{k \to \infty} x_k = x^* \). Using once more the definition of Jacobian and the fact that \( \lim_{k \to \infty} \|x_k - x^*\| = 0 \), we can write
\[
\lim_{k \to \infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = \lim_{k \to \infty} \frac{\|G(x_k) - G(x^*)\|}{\|x_k - x^*\|} = \lim_{k \to \infty} \frac{\|G(x_k) - G(x^*) - J_G(x^*)(x_k - x^*)\|}{\|x_k - x^*\|} = 0,
\]
which proves (3.1.10) for \( r < \delta \). We proceed now to prove quadratic convergence in \( B[x^*, r] \). Take \( r < \delta \) such that for all \( z \in B[x^*, r] \) it holds \( \|\nabla^2 f(z)^{-1}\| < \alpha \).

Note that the Hessian has continuously differentiable entries (f is \( C^3 \)), and hence it can be proved that it is Lipschitz in the ball \( B[x^*, r] \). In other words, there exists a constant \( C_1 > 0 \) such that
\[
\|\nabla^2 f(u) - \nabla^2 f(v)\| \leq C_1 \|u - v\|,
\]
for all \( u, v \in B[x^*, r] \). We know that \( x_k \in B[x^*, r] \) for all \( k \), hence we can use Theorem 3.1.32 to get

\[
\|x_{k+1} - x^*\| = \|x_k - \left[ \nabla^2 f(x_k) \right]^{-1} \nabla f(x_k) - x^* + \left[ \nabla^2 f(x_k) \right]^{-1} \nabla f(x^*) \|
\]

\[
= \| \left[ \nabla^2 f(x_k) \right]^{-1} (\nabla f(x_k) - \nabla f(x^*) - \left[ \nabla^2 f(x_k) \right] (x^* - x_k)) \|
\]

\[
= \| \left[ \nabla^2 f(x_k) \right]^{-1} (\nabla f(x_k) - \nabla f(x^*) - \left[ \nabla^2 f(x_k) \right] (x^* - x_k)) + (\nabla f(x^*) - \nabla^2 f(x_k)) (x^* - x_k) \|
\]

\[
\leq \| \left[ \nabla^2 f(x_k) \right]^{-1} (\nabla f(x_k) - \nabla f(x^*) - \left[ \nabla^2 f(x_k) \right] (x^* - x_k)) + \left[ \nabla^2 f(x_k) \right]^{-1} ((\nabla^2 f(x^*) - \nabla^2 f(x_k)) (x^* - x_k)) \|
\]

\[
\leq \frac{\alpha C_1}{2} \|x^* - x_k\|^2 + \alpha C_1 \|x^* - x_k\|^2 = \frac{3\alpha C_1}{2} \|x^* - x_k\|^2.
\]

This proves our second claim for \( C_0 := \frac{3\alpha C_1}{2} \). \( \square \)

Property (3.1.10) is called superlinear convergence. Hence the above proof establishes that, when \( f \) is twice continuously differentiable, superlinear convergence holds, provided we start close enough to the solution. When the function is three times differentiable and we start close enough to the solution, the convergence of Newton’s sequence is quadratic. We summarize below the most important features of Newton’s Method.

**Drawbacks:**

- We assure convergence only when starting in a certain region around the solution, this region is called the **convergence region**.

- It requires expensive computations at each iteration. Namely, to perform a step of Newton’s method, we need to solve a linear system of equations. In particular, the Hessian has to be computed at each current point.

**Advantages:**

- If we start within the convergence region, then the method has quadratic convergence, so it converges very fast to the solution.

The Steepest Descent method, which we describe in the next section, has features that can be seen as complementary to those outlined above.

### 3.1.4 Steepest Descent Method

This method is also called Cauchy method, because it was introduced by Cauchy in 1847. Since then it became the most frequently used thanks to the fact that it does not require expensive calculations. Assume we want to minimize \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) a continuously differentiable function and suppose that \( \nabla f(x) \neq 0 \). We mentioned in Example 3.1.9 that the direction \( p := -\frac{\nabla f(x)}{\|\nabla f(x)\|} \) is a descent direction. In fact, along this direction the function has its most rapid decrease. To check this property, observe that, by Cauchy-Swartz inequality, for all \( 0 \neq d \in \mathbb{R}^n \) such that \( \|d\| = 1 \) we have that

\[
-\|\nabla f(x)\| \leq \langle \nabla f(x), d \rangle \leq \|\nabla f(x)\|.
\]

It is direct to check that the extreme values are obtained for \( d = \pm \frac{\nabla f(x)}{\|\nabla f(x)\|} \). Since the directional derivative of \( f \) along the direction \( d \) is precisely \( \langle \nabla f(x), d \rangle \), we get that the minimum value of this directional derivative is obtained for \( d = -\frac{\nabla f(x)}{\|\nabla f(x)\|} \). For this reason, \( d \) is called the **steepest descent direction** of \( f \) at the point \( x \). We proceed now to define formally this method.
Definition 3.1.19 Let \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) be a continuously differentiable function and fix \( x_0 \in \mathbb{R}^n \). Then the Steepest Descent sequence \( \{x_k\} \) with initial point \( x_0 \) for minimizing \( f \) is defined by the formula:

\[
x_{k+1} := x_k - \alpha_k \nabla f(x_k),
\]

where the step-size \( \alpha_k > 0 \) is the value of \( \alpha \geq 0 \) that minimizes the function \( f \) in the direction \(-\nabla f(x^{(k)})\). In other words, \( \alpha_k > 0 \) is the solution of the unidimensional minimization problem

\[
\min_{\alpha \geq 0} f(x_k - \alpha \nabla f(x^{(k)})).
\]

Exercise 3.1.46 Consider the function \( f : \mathbb{R}^2 \rightarrow \mathbb{R} \) defined as \( f(x_1, x_2) = x_1^2 - x_1 x_2 + 1/2 x_2^2 \), with initial point \( x^{(0)} = (1, 0) \). Find the first four iterates of the steepest descent sequence. Find the unconstrained minimizer of \( f \) using the second order optimality conditions, and draw the trajectory of the sequence toward the solution.

As we see from the exercise above, the path toward the solution has directions that are successively orthogonal to one another. This feature is characteristic of steepest descent, as we see in the theorem below.

Theorem 3.1.34 The steepest descent method moves in perpendicular steps. More precisely, assume that \( \{x_k\} \) is the steepest descent sequence with initial point \( x_0 \). Suppose also that \( \nabla f(x^{(k)}) \neq 0 \) for all \( k \) (i.e., for all \( k \), \( x^{(k)} \) is not a critical point of \( f \)). Then it holds that

\[
(x^{(k+1)} - x^{(k)})^T (x^{(k+2)} - x^{(k+1)}) = 0.
\]

Proof.

Define \( \varphi_k(t) := f(x^{(k)} - t \nabla f(x^{(k)})) \), by definition of the step-size for steepest descent, we have that \( x^{(k+1)} = x^{(k)} - t_k \nabla f(x^{(k)}) \), where \( t_k \) solves the minimization problem

\[
\min_{t \geq 0} \varphi_k(t).
\]

Since \( t = 0 \) is not a solution of this problem (why?), the minimizer is attained at a \( t_k > 0 \). Then it must be \( \varphi'(t_k) = 0 \). In other words,

\[
0 = \varphi'(t_k) = - \left[ \nabla f(x^{(k)} - t_k \nabla f(x^{(k)}) \right]^T \nabla f(x^{(k)}) = - \nabla f(x^{(k+1)}) \nabla f(x^{(k)}).
\]

But \( (x^{(k+1)} - x^{(k)})^T (x^{(k+2)} - x^{(k+1)}) = t_k t_{k+1} \nabla f(x^{(k)})^T \nabla f(x^{(k+1)}) \), so the proof is complete. \( \square \)

The theorem above helps to understand better the behavior of the steepest descent method. In Figure 1.3, we depicted the level sets of a quadratic function of two variables. Observe that the gradient of \( f \) at a boundary point \( x^k \) of this level set is always perpendicular to the level curve \( \{ x | f(x) = f(x^k) \} \). The gradients at each current point are drawn as solid arrows, while the trajectory of the method is indicated by dotted arrows. The direction of steepest descent \(-\nabla f(x^{(k)})\) is the best possible direction at the point \( x^{(k)} \), however, at the next iterate \( x^{k+1} \), it becomes perpendicular to the best direction at the new point, \(-\nabla f(x^{k+1})\). This feature makes this method very slow for practical applications.

Exercise 3.1.47 Compute the first three iterates of the steepest descent sequence for \( f(x_1, x_2) = \frac{x_1^2 + x_2^2}{2} \), starting at \( x^{(0)} = (\frac{\sqrt{2}}{2}, \frac{1}{\sqrt{3}})^T \). Draw the trajectory of the iterates towards the minimizer of \( f \). Calculate the distance \( \| x^{(k)} - x^* \| \) for \( k = 0, 1, 2, 3 \). Compute the Newton iterates and compare the behavior of both methods in this case.
From the above exercise, we see that Newton can be drastically faster than Steepest descent. However, we mentioned before that Newton’s direction was not always a descent direction. This disadvantage is not shared by the steepest descent method, as we see next.

**Theorem 3.1.35** Assume that \( \{x^{(k)}\} \) is the steepest descent sequence for \( f \) and that \( \nabla f(x^{(k)}) \neq 0 \). Then \( f(x^{(k+1)}) < f(x^{(k)}) \).

**Proof.** From the proof of Theorem 3.1.34, we see that

\[
\varphi'(0) = -\left[ \nabla f\left(x^{(k)}\right) \right]^T \nabla f(x^{(k)}) = -\|\nabla f(x^{(k)})\|^2 < 0.
\]

But this implies that there exists \( \bar{t} > 0 \) such that \( \varphi(\bar{t}) < \varphi(0) = f(x^{(k)}) \). Since \( t_k \) is the minimizer of \( \varphi \), we must have

\[
f(x^{(k+1)}) = \varphi(t_k) \leq \varphi(\bar{t}) < \varphi(0) = f(x^{(k)}),
\]

as we wanted to prove. \( \square \)

As we see from Exercise 3.1.47, one problem of this method is that the steps are too long. A better strategy would be to use this direction, but choose a step-size without using the minimization (3.1.12). To solve this minimization problem at each step makes the step-size choice unnecessarily expensive. Moreover, why should we solve exactly the minimization problem in (3.1.12), if we only want an approximation of the solution? In order to overcome this drawback, there are other rules for choosing the step-size, called **Successive Step-size Selection**. In the simplest example of this kind of selection, an initial step-size \( s \) is chosen, and if \( s \) does not produce descent of \( f \), i.e., if

\[
f(x^{(k)}) - s \nabla f(x^{(k)}) \geq f(x^{(k)}),
\]

then certain fraction of \( s \) is chosen, say \((1/2)s\), and so on, until a satisfactory decrease is obtained. However, this choice of the step-size does not guarantee convergence to a solution, since the improvement on \( f \) may not be substantial enough. In other words, a successive step-size selection should be designed in such a way that it ensures a substantial improvement of the values of \( f \). An example of such efficient successive step-size selection is the so-called **Armijo Rule**, which we describe next. According to the way in which the step-size is chosen, we will call method (3.1.11) the **Steepest Descent Method with Armijo Rule**.

### 3.1.5 Steepest Descent with Armijo Rule

In this variant of Steepest Descent, there is a fixed initial step-size \( s \), and two fractions \( \sigma, \beta \in (0, 1) \). Choose \( m_k \) the first natural number for which the inequality

\[
f(x^{(k)}) - \beta^m s \nabla f(x^{(k)}) < f(x^{(k)}) - \sigma \beta^m s \|\nabla f(x^{(k)})\|^2,
\]

holds. In other words, we increase \( m = 0, 1, 2, \ldots \) until the inequality above is satisfied for the first time, and we call this value \( m_k \). The Armijo test (3.1.13) not only requires a decrease of the values of \( f \), but it requires this decrease to be smaller than certain fraction of \(-s \|\nabla f(x^{(k)})\|^2\). A natural question now is: Why is it possible to find such an integer \( m \)? The answer is in the following theorem.

**Theorem 3.1.36** Assume that \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is twice continuously differentiable and that \( \nabla f(x^{(k)}) \neq 0 \) for all \( k \). Fix \( \delta \in (0, 1) \). Then there exists \( r > 0 \) small enough such that

\[
f(x^{(k)}) - r \nabla f(x^{(k)}) < f(x^{(k)}) - \delta r \|\nabla f(x^{(k)})\|^2,
\]

holds.
Proof. Take \( r > 0 \) such that
\[
r \left[ \nabla f(x^{(k)})^T \nabla^2 f(x^{(k)}) \nabla f(x^{(k)}) \right] < 2(1 - \delta) \| \nabla f(x^{(k)}) \|^2.
\]
Note that this inequality holds for all \( r > 0 \) if \( \left[ \nabla f(x^{(k)})^T \nabla^2 f(x^{(k)}) \nabla f(x^{(k)}) \right] \leq 0 \), and for
\[
r < \frac{2(1 - \delta) \| \nabla f(x^{(k)}) \|^2}{\nabla f(x^{(k)})^T \nabla^2 f(x^{(k)}) \nabla f(x^{(k)})}
\]
when \( \nabla f(x^{(k)})^T \nabla^2 f(x^{(k)}) \nabla f(x^{(k)}) > 0 \). Replacing this value of \( r \) in Taylor formula we get
\[
f(x^{(k)} - r \nabla f(x^{(k)})) = f(x^{(k)}) - r \| \nabla f(x^{(k)}) \|^2 + \frac{r^2}{2} \nabla f(x^{(k)})^T \nabla^2 f(x^{(k)}) \nabla f(x^{(k)})
< f(x^{(k)}) - \delta r \| \nabla f(x^{(k)}) \|^2,
\]
which completes the proof. \( \square \)

Taking \( r = \beta^m s \) and \( \delta = \sigma \) in Theorem 3.1.36, we see that for \( m \) large enough, inequality (3.1.13) is verified.

An important advantage of both variants of the steepest descent method over Newton’s method is that they always converge to a stationary point of \( f \), no matter how far the initial guess \( x^{(0)} \) is from the solution.

### 3.1.6 Convergence of Steepest Descent Methods

We prove below that the Steepest descent method with minimization rule and with Armijo rule have the property that every accumulation point is a stationary point of \( f \).

**Theorem 3.1.37** Let \( f : \mathbb{R}^n \to \mathbb{R} \) be continuously differentiable, and assume that the sequence \( \{x^{(k)}\} \) is generated by steepest descent method with minimization rule or by steepest descent method with Armijo rule. Then every accumulation point of \( \{x^{(k)}\} \) is a stationary point of \( f \).

**Proof.** Let \( \{x^{(k)}\} \) be the sequence generated by steepest descent method with Armijo rule, and call \( \alpha_k := \beta^{m_k} s \). By definition of \( m_k \), we know that
\[
f(x^{(k+1)}) = f(x^{(k)} - \beta^{m_k} s \nabla f(x^{(k)})) < f(x^{(k)}) - \sigma \beta^{m_k} s \| \nabla f(x^{(k)}) \|^2.
\]
But the definition of \( m_k \) also tells us that \( m = m_k - 1 \) does not verify (3.1.13). In other words,
\[
f(x^{(k)} - \beta^{m_k-1} s \nabla f(x^{(k)})) \geq f(x^{(k)}) - \sigma \beta^{m_k-1} s \| \nabla f(x^{(k)}) \|^2,
\]
or, equivalently
\[
f(x^{(k)} - (\alpha_k/\beta) \nabla f(x^{(k)})) \geq f(x^{(k)}) - \sigma (\alpha_k/\beta) \| \nabla f(x^{(k)}) \|^2.
\]
Now take \( \bar{x} \) an accumulation point of \( \{x^{(k)}\} \), and let \( \{x^{(k_j)}\} \) be a subsequence converging to \( \bar{x} \). In order to arrive at a contradiction, assume \( \nabla f(\bar{x}) \neq 0 \). By definition of Armijo rule, the sequence of real numbers \( \{f(x^{(k)})\} \) is decreasing, and hence convergent. Using also (3.1.15)
\[
0 = \lim_{k \to \infty} f(x^{(k)}) - f(x^{(k+1)}) \geq \lim_{k \to \infty} \sigma \alpha_k \| \nabla f(x^{(k)}) \|^2.
\]
This implies that \( \lim_{k \to \infty} \alpha_k \| \nabla f(x^{(k)}) \|^2 = 0 \), and in particular
\[
\lim_{j \to \infty} \alpha_{k_j} \| \nabla f(x^{(k_j)}) \|^2 = 0.
\]
Using that $f$ is continuously differentiable, and that the subsequence $\{x^{(k_j)}\}$ converges to $\bar{x}$, we get

$$\lim_{j \to \infty} \|\nabla f(x^{(k_j)})\|^2 = \|\nabla f(\bar{x})\|^2 > 0.$$

Combining this fact with (3.1.17), we conclude that $\lim_{j \to \infty} \alpha_{k_j} = 0$. Now rearrange inequality (3.1.16) to obtain

$$\frac{f(x^{(k)} + (-\alpha_k/\beta)\nabla f(x^{(k)})) - f(x^{(k)})}{(-\alpha_k/\beta)} \leq \sigma \|\nabla f(x^{(k)})\|^2.$$

By the mean value theorem, there exists $t_k \in (0, 1)$ such that

$$f(x^{(k)} + (-\alpha_k/\beta)\nabla f(x^{(k)})) - f(x^{(k)}) = (-\alpha_k/\beta) \left[ \nabla f(x^{(k)}) + t_k(-\alpha_k/\beta)\nabla f(x^{(k)})^T \nabla f(x^{(k)}) \right].$$

Combining the last two expressions we can write

$$\nabla f(x^{(k)}) + t_k(-\alpha_k/\beta)\nabla f(x^{(k)})^T \nabla f(x^{(k)}) \leq \sigma \|\nabla f(x^{(k)})\|^2.$$ (3.1.18)

Observe that $t_k$ is bounded, $\lim_{j \to \infty} \alpha_{k_j} = 0$ and $\lim_{j \to \infty} \nabla f(x^{(k_j)}) = \nabla f(\bar{x})$. Altogether, these facts yield

$$\lim_{j \to \infty} x^{(k_j)} + t_{k_j}(-\alpha_{k_j}/\beta)\nabla f(x^{(k_j)}) = \bar{x},$$

which gives $\lim_{j \to \infty} \nabla f(x^{(k_j)}) + t_{k_j}(-\alpha_{k_j}/\beta)\nabla f(x^{(k_j)}) = \nabla f(\bar{x})$. Now taking limits in (3.1.18) for $k = k_j$ and $j \to \infty$ we get

$$\|\nabla f(\bar{x})\|^2 \leq \sigma \|\nabla f(\bar{x})\|^2,$$

which contradicts the fact that $\sigma < 1$. Hence we must have $\nabla f(\bar{x}) = 0$. Now let us prove the same result for the sequence $\{\tilde{x}^k\}$ generated by steepest descent with minimization rule.

Take again $\bar{x}$ an accumulation point of $\{\tilde{x}^k\}$, and let $\{\tilde{x}^{k_j}\}_j$ be a subsequence converging to $\bar{x}$. In order to arrive at a contradiction, assume $\nabla f(\bar{x}) \neq 0$. For each fixed $k$, consider the point $x^k$ obtained by performing an Armijo step from $\tilde{x}^k$, and let $\alpha_k$ be the corresponding step-size. By the minimization rule, we have that $f(x^k) \geq f(\tilde{x}^{k+1})$ and that the sequence of real numbers $\{f(\tilde{x}^k)\}$ is decreasing. So

$$0 = \lim_{k \to \infty} f(\tilde{x}^k) - f(\tilde{x}^{k+1}) \geq \lim_{k \to \infty} f(\tilde{x}^k) - f(x^k) \geq \lim_{k \to \infty} -\sigma \alpha_k \|\nabla f(\tilde{x}^k)\|^2.$$

In a similar way as in the first part of the proof, we conclude that $\lim_{j \to \infty} \alpha_{k_j} = 0$. Proceeding now as in (3.1.18) above, we obtain

$$\nabla f(\tilde{x}^{k_j}) + t_{k_j}(-\alpha_{k_j}/\beta)\nabla f(\tilde{x}^{k_j})^T \nabla f(\tilde{x}^{k_j}) \leq \sigma \|\nabla f(\tilde{x}^{k_j})\|^2,$$

for some $t_{k_j} \in (0, 1)$. Taking again limits for $j \to \infty$ we arrive at the same contradiction $\sigma \geq 1$. Hence $\bar{x}$ must be a stationary point.

We summarize below the most important features of both variants of the Steepest Descent Method.

**Drawbacks:**
- Convergence is slow.
- When using (3.1.13), computation of the iterates is not expensive.

As we see, these features complement those of Newton method.
3.1.7 Quasi-Newton Methods

Quasi-Newton methods try to avoid the drawbacks of both Newton and Steepest descent, while keeping computations cheap. They have the general form

\[ x^{k+1} := x^k + \alpha_k d_k, \quad d_k := -(D_k)^{-1} \nabla f(x^k), \]

where \( D_k \) is positive definite matrix. Ideally, \( D_k \) approximates the Hessian of \( f \) at \( x^k \), so that the step is as close as possible to Newton’s. The use of an approximation of the Hessian of \( f \) avoids the heavy computations of Newton method, while preserving fast convergence. Besides from the positive definiteness, which ensures that the search direction is a descent direction, we need to establish a requirement on \( D_k \) which guarantees some proximity to the Hessian of \( f \). This condition is called the secant condition relative to \( r_f(x^k) \) and is as follows

\[ D_{k+1}(x^{k+1} - x^k) = \nabla f(x^{k+1}) - \nabla f(x^k). \]

Another requirement on the sequence of matrices \( \{D_k\} \) is that the update from \( D_k \) to \( D_{k+1} \) should be as simple as possible. More details on Quasi-Newton methods can be found, for instance, in the books [6, 3].

3.2 Methods for Constrained Problems

A constrained problem is in general more difficult to solve than an unconstrained one. In order to find the local solutions we need to solve the corresponding optimality conditions, which is usually a complicated nonlinear system of equations. In this case, an iterative method like for instance Newton’s method may be necessary to find the approximate solutions of the system given by the optimality conditions. In other words, we may not be able to solve the optimality conditions directly. This justifies the idea of replacing the original, constrained problem by an unconstrained one, in which the new objective function incorporates the constraints. This is the idea of the so-called Penalty and Barrier Methods. Once the constrained problem is conveniently reformulated as an unconstrained one, then usual techniques for unconstrained optimization (e.g., Newton’s method or steepest descent method) can be applied. Consider problem (2.9.50) with constraint functions \( h: \mathbb{R}^n \to \mathbb{R}^m \) and \( g: \mathbb{R}^n \to \mathbb{R}^r \) so

\[ h(x) := \begin{pmatrix} h_1(x) \\ \vdots \\ h_m(x) \end{pmatrix}, \quad g(x) := \begin{pmatrix} g_1(x) \\ \vdots \\ g_r(x) \end{pmatrix}. \]

The feasible region is therefore given by

\[ \mathcal{F} := \{ x \in \mathbb{R}^n \mid h(x) = 0, \ g(x) \leq 0 \}. \]

Both Penalty and Barrier method modify the objective function reflecting the feasible set in the new objective function. The Penalty Method generates a sequence of (possibly unfeasible) points \( \{x_k\} \) which approach the feasible region. Under additional assumptions, the sequence generated by the penalty method converges to a solution of the original problem. The Barrier Method generates a sequence of feasible points which, under suitable conditions, converges to a solution of the original constrained problem. We start by describing the Penalty Method.

3.2.1 Penalty Methods

The main ingredient in a penalty method is the penalty function, which is appended to the original objective function. The reason for the terminology is the fact that it imposes a high price for points far from the feasible set.
**Definition 3.2.20** We say that the function \( p : \mathbb{R}^n \to \mathbb{R} \) is a penalty function for the feasible set \( F \) of problem (2.9.50) if

\[
p(x) = 0 \text{ for all } x \in F \text{ and } p(x) > 0 \text{ for all } x \notin F.
\]

It is easy to use the constraint functions \( h \) and \( g \) for constructing penalty functions, which we do in the following example.

**Example 3.2.11** Let \( F \) be the feasible set of problem (2.9.50). Consider \( g_j^+(x) := \max\{0, g_j(x)\} \) for all \( j = 1, \ldots, r \). Note that a point \( \bar{x} \) is infeasible if and only if there exists \( j \in \{1, \ldots, r\} \) or \( i \in \{1, \ldots, m\} \) such that \( g_j^+(\bar{x}) > 0 \) or \( |h_i(\bar{x})| > 0 \). This observation proves that the following functions are penalty functions for the feasible set \( F \).

(i) \( p_1(x) := \sum_{i=1}^m |h_i(x)| + \sum_{j=1}^r g_j^+(x) \)

(ii) \( p_2(x) := \sum_{i=1}^m (h_i(x))^2 + \sum_{j=1}^r (g_j^+(x))^2 \)

An important difference between \( p_1 \) and \( p_2 \) is that \( p_1 \) is not differentiable, while \( p_2 \) is. However, in general, \( p_2 \) is not twice differentiable.

**Exercise 3.2.48** Determine which of the functions below is a penalty function for \( F \).

(i) \( q_1(x) := \sum_{i=1}^m (h_i(x))^k + \sum_{j=1}^r g_j^+(x) \), with \( k \in \mathbb{N} \) an odd number.

(ii) \( q_2(x) := \sum_{i=1}^m \ln((h_i(x))^2 + 1) + \sum_{j=1}^r (g_j^+(x))^3 \)

(iii) \( q_3(x) := \sum_{i=1}^m \exp h_i(x) + \sum_{j=1}^r \exp g_j^+(x) \)

(iv) \( q_4(x) := \sum_{i=1}^m \ln \frac{1}{1 + (h_i(x))^2} + \sum_{j=1}^r g_j^+(x) \)

In case the function is not a penalty function, modify it so that it becomes a penalty function.

**Exercise 3.2.49** Assume that \( p_1 \) and \( p_2 \) are penalty functions for \( F \). Prove the statements below.

(i) \( q(x) := \alpha_1 p_1(x) + \alpha_2 p_2(x) \) is a penalty function for \( F \) for every \( \alpha_1 > 0, \alpha_2 > 0 \).

(ii) \( q(x) := cp_1(x)p_2(x) \) is a penalty function for \( F \) for every \( c > 0 \).

(iii) \( q(x) := (p_1(x))^k + (p_2(x))^k \) is a penalty function for \( F \) for every \( k > 0 \).

(iv) Let \( \phi_1, \phi_2 : \mathbb{R} \to \mathbb{R} \) be two increasing functions such that \( \phi_1(0) = \phi_2(0) = 0 \). Then \( q(x) := \phi_1(p_1(x)) + \phi_2(p_2(x)) \) is a penalty function for \( F \).

The penalty method chooses a parameter \( c > 0 \) and replaces problem (2.9.50) by the (unconstrained) problem

\[
(P_c) \quad \min_{x \in \mathbb{R}^n} p_c(x) = f(x) + cp(x)
\]

subject to \( x \in \mathbb{R}^n \). 

Note that for ”big” values of \( c > 0 \), a solution \( x_c \) of \( (P_c) \) is forced to have ”small” \( p(x_c) \). In other words, by increasing the value of \( c \), the solutions of each \( (P_c) \) become closer and closer to the feasible set. A remarkable fact is that, when a sequence generated by this procedure converges, the limit is a solution of the original problem.

**Theorem 3.2.38** Let \( p : \mathbb{R}^n \to \mathbb{R} \) be a continuous penalty function for \( F \). Let the sequence of parameters \( \{c_k\} \) be chosen such that \( 0 < c_k \) for all \( k \in \mathbb{N} \) and \( \lim_{k \to \infty} c_k = +\infty \). Assume that every problem \( (P_{c_k}) \) has a global minimizer \( x_k \). If there exists \( x^* = \lim_{k \to \infty} x_k \) then \( x^* \) is a solution of problem (2.9.50).
Proof. Take $\bar{x} \in \mathcal{F}$. Then we know that $p(\bar{x}) = 0$. Note also that $c_k p(x_k) \geq 0$ for all $k$. Therefore,
\[
 f(x_k) \leq f(x_k) + c_k p(x_k) \leq f(\bar{x}) + c_k p(\bar{x}) = f(\bar{x}),
\]
where we used the fact that $x_k$ is a solution of $(P_{c_k})$ in the second inequality. Using the above inequality and continuity of $f$ we get
\[
 f(x^*) = \lim_{k \to \infty} f(x_k) \leq f(\bar{x}) \quad \text{and} \quad 0 \leq c_k p(x_k) \leq f(\bar{x}) - f(x_k).
\] (3.2.20)

Since $\lim_{k \to \infty} f(\bar{x}) - f(x_k) = f(\bar{x}) - f(x^*)$ the right hand inequality in (3.2.20) yields $\lim_{k \to \infty} p(x_k) = 0 = p(x^*)$ where we used in the second equality the fact that $p$ is continuous. So $x^* \in \mathcal{F}$. Because $\bar{x} \in \mathcal{F}$ is arbitrary, the left inequality in (3.2.20) implies that $x^*$ is a solution of problem (2.9.50).

Exercise 3.2.50 Assume that the sequence of parameters $\{c_k\}$ in problem $(P_{c_k})$ is chosen such that $0 < c_k < c_{k+1}$ for all $k \in \mathbb{N}$. Recall that the objective function of problem $(P_{c_k})$ is given by $p_{c_k}(x) := f(x) + c_k p(x)$. Denote by $x_k$ a solution of $(P_{c_k})$. Prove the following inequalities:
\[
p_{c_k}(x_k) \leq p_{c_{k+1}}(x_{k+1}), \quad p(x_k) \geq p(x_{k+1}), \quad f(x_k) \leq f(x_{k+1}).
\]

Hint: For the second inequality: Note that, by definition of $x_k$ and $x_{k+1}$ we have $f(x_k) + c_{k+1} p(x_k) \geq f(x_{k+1}) + c_{k+1} p(x_{k+1})$ and $f(x_{k+1}) + c_k p(x_{k+1}) \geq f(x_k) + c_k p(x_k)$. Sum up these inequalities and simplify the resulting expression. For the rightmost inequality, use the middle one.

3.2.2 The Exact Penalty Method

We described above a procedure which generates a sequence of problems $(P_{c_k})$ whose solutions $\{x_k\}$ may converge to a solution of the original problem. There is an important drawback of this approach, and it is the fact that the problems $(P_{c_k})$ become more and more ill-behaved the bigger is $c_k$. Furthermore, for large values of $c_k$, the penalty term becomes predominant and little importance is given to the objective function $f$, which carries an essential piece of information on the original problem. Therefore, a key question concerning penalty methods is the following: is there some penalty function for which there exists $c > 0$ such that the solutions of $(P_c)$ are also solutions of the original problem (2.9.50)? This question has in fact a positive answer, and it allows us to provide an alternative formulation of the original problem as an unconstrained one, such that the solution of this reformulation automatically solves our original problem. This in turn produces a second question: how can we find the parameter $c$ which does the job? We will answer to these questions in Theorem 3.2.40. Penalty functions with this desired property are called exact penalty functions. The parameter $c^*$ which provides the reformulation of the original problem as the unconstrained problem $(P_{c^*})$ is called exact penalty parameter. An important fact is that the value of the exact penalty parameter $c^*$ is closely connected with the Karush-Kuhn-Tucker multipliers. Let the penalty function be the function $p_1$ given in Example 3.2.11(i), so that the corresponding subproblems are
\[
(P_c) \quad \min_{x \in \mathbb{R}^n} f(x) + c \left( \sum_{i=1}^m |h_i(x)| + \sum_{j=1}^r g_j^+(x) \right)
\]

We pointed out before that $p_1$ is not everywhere differentiable. However, it offers the remarkable advantage described above. In other words, $p_1$ is an exact penalty function. Therefore, there exists $c^* > 0$ such that the solutions of $(P_{c^*})$ solve our original problem. Recall that an inequality constraint $g_j$ is called active at the point $x$ when $g_j(x) = 0$. Denote by $A(x) := \{ j \mid g_j(x) = 0 \}$ the set of indexes corresponding to active inequality constraints at $x$. We say that a point $x^*$ is regular when the gradients of the equality constraints $\{ \nabla h_i(x^*) \}$ together with the gradients of the active inequality constraints $\{ \nabla g_j(x^*) \}_{j \in A(x^*)}$ is a linearly independent set.
Because we will use them in the next theorem, we recall now the second order sufficient conditions for problem (2.9.50). Let $V(x^*)$ be the cone considered in Theorem 2.9.21, in other words,

$$V(x^*) := \{ y \in \mathbb{R}^n : \nabla h_i(x^*)^T y = 0 \text{ for all } i = 1, \ldots, m \text{ and } \nabla g_j(x^*)^T y = 0 \text{ for all } j \in A(x^*) \},$$

which is the subspace of tangent feasible directions for problem (2.9.50) at the point $x^*$. We also need to recall the cone of critical directions. Let $x \in \mathbb{R}^n$ and let

$$T(x) = \{ y \in \mathbb{R}^n : y^T \nabla h_i(x) = 0 \text{ for all } i, \ y^T \nabla g_j(x) \leq 0 \text{ for all } j \in A(x) \text{ and } y^T \nabla f(x) \leq 0 \}.$$ 

The set $T(x)$ is called the cone of critical directions from $x$. This cone, instead of $V(x)$, is the one considered in the following theorem. In the next exercise the relation between these two cones is made precise for $x$ a KKT point.

**Exercise 3.2.51** Let $x^*$ be a Karush-Kuhn-Tucker point for problem (2.9.50), i.e.,

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0,$$

$$\mu^*_j \geq 0, \ j = 1, \ldots, r,$$

$$\mu^*_j = 0, \ j \notin A(x^*)$$

$$h(x^*) = 0, \ g(x^*) \leq 0.$$

Then we always have that $V(x^*) \subset T(x^*)$. If, additionally we have that $\mu^*_j > 0$ for all $j \in A(x^*)$ (this condition is called strict complementarity) then $T(x^*) \subset V(x^*)$.

**Theorem 3.2.39** Assume $f, h_1, \ldots, h_m, g_1, \ldots, g_r$ are twice continuously differentiable. Let $x^* \in \mathbb{R}^m$ and $(\lambda^*, \mu^*) \in \mathbb{R}^m \times \mathbb{R}^r$ such that the first order necessary conditions are met, i.e.,

$$\nabla_x L(x^*, \lambda^*, \mu^*) = 0,$$

$$\mu^*_j \geq 0, \ j = 1, \ldots, r,$$

$$\mu^*_j = 0, \ j \notin A(x^*)$$

$$h(x^*) = 0, \ g(x^*) \leq 0.$$  \hspace{1cm} (3.2.21)

Assume that $y^T \nabla^2_{xx} L(x^*, \lambda^*, \mu^*) y > 0$ for all nonzero $y \in T(x^*)$. Then $x^*$ is a strict local minimum of problem (2.9.50). More precisely, there exists $\gamma > 0$ and $\varepsilon > 0$ such that

$$f(x) \geq f(x^*) + (\gamma/2) \| x - x^* \|^2,$$

for all $x$ such that $h(x) = 0, \ g(x) \leq 0$ and $\| x - x^* \| < \varepsilon$.

**Exercise 3.2.52** Prove Theorem 3.2.39 following the similar steps as those in Theorem 2.9.20.

**Theorem 3.2.40 (Exact Penalty Theorem)** Let $x^*$ be a local minimizer of Problem (2.9.50). Assume that $x^*$ is regular and satisfies the sufficient optimality conditions for strict minimum given in Theorem 3.2.39 for the multipliers $(\lambda^*, \mu^*)$. Suppose also that $\mu^*_j > 0$ for all $j \in A(x^*)$ (i.e., strict complementarity holds). Then $x^*$ is also a local minimizer of the problem

$$\mathcal{P}_c \min_{x \in \mathbb{R}^n} f(x) + c \left( \sum_{i=1}^m |h_i(x)| + \sum_{j=1}^r g_j^+(x) \right)$$

whenever $c > 0$ satisfies

$$c > \max_{i=1, \ldots, m, j \in A(x^*)} \{ |\lambda^*_i|, \mu^*_j \}.$$ \hspace{1cm} (3.2.22)
Proof. Note that Lemma 2.9.7 implies that problem (2.9.50) can be reformulated as
\[
\min f(x) \text{ s.t. } h(x) = 0, \quad g_j(x) = 0 \forall j \in A(x^*).
\]
So we can assume that all constraints are equality constraints. In other words, our original problem is
\[
\min f(x) \text{ s.t. } h(x) = 0.
\]
For this problem our assumption on \( c \) simplifies to
\[
c > \max_{i=1,\ldots,m} \{ |\lambda_i^*| \}.
\]
So there exists \( a > 0 \) such that
\[
c - |\lambda_i^*| > a > 0 \quad \forall \ i = 1, \ldots, m. \tag{3.2.23}
\]
Now consider its parameterized problem
\[
(P(p)) \quad \min f(x) \text{ s.t. } h(x) = p.
\]
By Exercise 3.2.51, the assumption that \( \mu_i^* > 0 \) for all \( j \in A(x^*) \) implies that \( V(x^*) = T(x^*) \).
Hence, if \( x^* \) verifies the assumptions of Theorem 3.2.39, then it also verifies the assumptions of the Sensitivity Theorem 2.9.24. The latter theorem asserts that the optimal value function
\[
v(p) = \min_{x \in \mathbb{R}^n} \{ f(x) \mid h(x) = p \}
\]
is continuously differentiable with gradient \( \nabla v(0) = -\lambda^* \), so for a small perturbation \( p \neq 0 \) we can write the first order Taylor development of \( v \) around zero as
\[
v(p) = v(0) + \nabla v(0)^T p + o(p), \tag{3.2.24}
\]
where \( \lim_{p \to 0} \frac{o(p)}{||p||_1} = 0 \) and \( ||p||_1 := \sum_{i=1}^m |p_i| \) is the \( l_1 \) norm of \( p \). Therefore for \( p \) small enough we have
\[
(a + \frac{o(p)}{||p||_1}) > 0, \tag{3.2.25}
\]
where \( a \) is as in (3.2.23). Summing up at both sides of (3.2.24) the term \( \sum_{i=1}^m c |p_i| \) we obtain
\[
v(p) + \sum_{i=1}^m c |p_i| = v(0) + \nabla v(0)^T p + \sum_{i=1}^m c |p_i| + o(p)
\geq v(0) - \sum_{i=1}^m |\lambda_i^*| |p_i| + \sum_{i=1}^m c |p_i| + o(p)
= v(0) + \sum_{i=1}^m |p_i| \left( c - |\lambda_i^*| + \frac{o(p)}{||p||_1} \right) \tag{3.2.26}
\geq v(0) + \sum_{i=1}^m |p_i| \left( a + \frac{o(p)}{||p||_1} \right)
\geq v(0) = f(x^*),
\]
where we used the fact that \( st \geq -|s||t| \) for all \( s, t \in \mathbb{R} \) in the first inequality, (3.2.23) in the second inequality, and (3.2.25) in the third one. Note that (3.2.26) holds for \( ||p||_1 \) sufficiently small. So there exists \( \delta > 0 \) such that (3.2.26) holds for every \( ||p||_1 < \delta \). Assume that \( x^* \) is not a local solution of \((P_c)\). Therefore, there exists a sequence \( \{x_k\} \) such that \( x_k \to x^* \) and
\[
p_c(x_k) < p_c(x^*). \tag{3.2.27}
\]
Consider the sequence \( \{p_k\} \) defined by \( p_k := h(x_k) \), then \( p_k = h(x_k) \to h(x^*) = 0 \). So there exists \( k_0 \) such that \( ||p_k||_1 < \delta \) for every \( k \geq k_0 \). Because \( h(x_k) = p_k \) we have
\[
v(p_k) = \min_{x \in \mathbb{R}^n} \{ f(x) \mid h(x) = p_k \} \leq f(x_k).
\]
Fix $k \geq k_0$, using (3.2.26) we have
\[
 p_c(x^*) = f(x^*) + c\|h(x^*)\|_1 = v(0) \leq v(p_k) + c\|p_k\|_1 
\leq f(x_k) + c\|h(x_k)\|_1 = p_c(x_k),
\]
contradicting (3.2.27). \qed

Unfortunately, the exact penalty parameter $\max_{i=1,\ldots,m} \{|\lambda_i^*|\}$ is not known in advance, since it requires the solution of the Karush-Kuhn-Tucker optimality conditions. What is done in practice is to increase $c$ until we pass this unknown threshold. This procedure does not require the parameter $c$ to go to infinity, but it may also produce ill-conditioned penalty problems ($P_c$) when the exact penalty parameter is large. Note that this is the case when the norm of the gradient $\|\nabla v(0)\|$ is large, which means that our original problem is very sensitive to perturbations.

**Remark 3.2.16** The penalty function $p_2$ given in Example 3.2.11(ii) is in general not exact and hence the penalty method generated by $p_2$ requires the penalty parameter to go to $+\infty$. Let us illustrate this fact for the equality constrained problem
\[
 \min f(x) \quad h(x) = 0.
\]
The KKT conditions for this problem are
\[
 \nabla f(\bar{x}) + \sum_{i=1}^m \lambda_i \nabla h_i(\bar{x}) = 0, \quad h(\bar{x}) = 0. \tag{3.2.28}
\]
Assume that the constraints of our original problem are not superfluous, i.e., that every solution $\bar{x}$ of the original problem is not an unconstrained minimum of $f$. This implies in particular that $\nabla f(\bar{x}) \neq 0$. We consider the quadratic penalized subproblems
\[
 \min f(x) + c_k \sum_{i=1}^m (h_i(x))^2 \quad x \in \mathbb{R}^n,
\]
where $f, h$ are assumed to be continuously differentiable. The optimality conditions on the solutions $x_k$ become
\[
 \nabla f(x_k) + 2c_k \sum_{i=1}^m \nabla h_i(x_k) h_i(x_k) = 0,
\]
or, equivalently
\[
 (1/c_k) \nabla f(x_k) + 2 \sum_{i=1}^m \nabla h_i(x_k) h_i(x_k) = 0, \tag{3.2.29}
\]
If the sequence $x_k$ converges to an optimal point $\bar{x}$ we must have
\[
 \nabla f(x_k) \to \nabla f(\bar{x}), \quad \nabla h_i(x_k) \to \nabla h_i(\bar{x}), \quad \text{for all } i = 1, \ldots, m, \quad \text{and } h(x_k) \to h(\bar{x}) = 0.
\]
We can always assume that the sequence $\{c_k\}$ is strictly increasing. Hence it either tends to infinity or it tends to its supremum $\bar{c} > 0$. We claim that it must tend to infinity. Indeed, if this is not the case, taking limit for $k \to \infty$ in (3.2.29) we get $\frac{1}{\bar{c}} \nabla f(\bar{x}) = 0$ which gives $\nabla f(\bar{x}) = 0$, contradicting our assumption on $\bar{x}$. Therefore we must have $c_k \to \infty$ for the quadratic penalty method.

When the problem is not convex, we need strict inequality in (3.2.22), because this allows us to use the fact (3.2.23). However, if the problem is convex, we only need the weaker assumption $c \geq \max_{i=1,\ldots,m,j\in A(x^*)} \{|\lambda_i^*|, \mu_j^*\}$. We prove this fact below.
Theorem 3.2.41 Assume that problem (2.9.50) is convex, i.e., f, g are convex and all h_i are affine. Suppose also that f, g are differentiable and fix x* a solution of problem (2.9.50) with Lagrange multipliers (λ^*, μ^*) ∈ ℝ^{m+r}. Then x* is a minimum of p_c whenever c ≥ max_{i=1,...,m,j∈A(x*)} {|λ_i^*|, μ_j^*}.

Proof. The Lagrangian associated with problem (2.9.50) is

\[ L(x, λ, μ) = f(x) + \sum_{i=1}^{m} λ_i h_i(x) + \sum_{j=1}^{r} μ_j g_j(x). \]

Since x* a solution of problem (2.9.50), we have that

\[ \nabla_x L(x^*, λ^*, μ^*) = 0 \quad \text{and} \quad L(x^*, λ^*, μ^*) = f(x^*) = p_c(x^*). \]  

(3.2.30)

Under the convexity of the data, the function \( L(\cdot, λ^*, μ^*) \) is convex. Therefore, the stationary points are unconstrained minima. This allows us to write

\[ p_c(x^*) = L(x^*, λ^*, μ^*) \leq L(x, λ^*, μ^*) = f(x) + \sum_{i=1}^{m} λ_i^* h_i(x) + \sum_{j=1}^{r} μ_j^* g_j(x) \]

\[ \leq f(x) + \sum_{i=1}^{m} |λ_i^*| |h_i(x)| + \sum_{j=1}^{r} μ_j^* (g_j^+(x)) \]

\[ \leq f(x) + c \left( \sum_{i=1}^{m} |h_i(x)| + \sum_{j=1}^{r} (g_j^+(x)) \right) \]

\[ = p_c(x). \]

We used the RHS of (3.2.30) in the first equality and the fact that \( μ_j^* g_j(x) \leq μ_j^* (g_j^+(x)) \) in the last inequality. The expression above implies that \( x^* \) is a global minimum of \( p_c \).

As an illustration of the exact penalty parameter and its relation with the Lagrange multipliers, let us look at the example below, taken from [3, page 434].

Example 3.2.12 Consider the following problem

\[ \min x_1 \quad \text{s.t.} \quad x_1^2 + x_2^2 = 1. \]

Posing the optimality conditions for the equality constrained case yields \( x^* = (-1, 0)^t \) and \( λ^* = 1/2 \) (check this). Therefore the optimal value is \( f(x^*) = -1 \). Now let us pose the penalized problem, with penalization function \( p_1 \):

\[ (P_c) \quad \min x_1 + c |x_1^2 + x_2^2 - 1| \]

\[ x \in \mathbb{R}^n \]

The minimum value \( M(c) \) of this problem is given by

\[ M(c) := \min \{ M_1, M_0 \}, \]

where \( M_0 := \inf_{\{x \mid x_1^2 + x_2^2 \leq 1\}} p_c(x), \) \( M_1 := \inf_{\{x \mid x_1^2 + x_2^2 > 1\}} p_c(x) \). Note that the infimum possible first coordinate of the region \( \{x \mid x_1^2 + x_2^2 \leq 1\} \) is clearly \(-1\). So \( M_0 = -1 \) and so

\[ M(c) = \min \{ M_1, -1 \}. \]

Let us compute \( M_1 \). The corresponding problem is

\[ \min x_1 + c |x_1^2 + x_2^2 - 1| \]

\[ -x_1^2 - x_2^2 + 1 \leq 0. \]
The constraints allow to simplify the objective function, so we have

\[
\begin{align*}
(P) \quad & \min_{x} x_1 + c(x_1^2 + x_2^2 - 1) \\
& -x_1^2 - x_2^2 + 1 \leq 0.
\end{align*}
\]

Note that, if every solution of \((P)\) lies in the boundary of the constraint set, then its optimal value must be \(-1\) and the solution must be \(x^* = (-1, 0)^t\). Indeed, in this case the penalized term vanishes and we are left with the original equality constrained problem. So in this case both the optimal value and the solution of \((P)\) will coincide with those of the original problem. So every solution with different optimal value must be in the interior of the constraint set. Thus let us assume that \(x^* = (-1, 0)^t\) lies in the interior of the constraint set. The Lagrangian for \((P)\) is

\[
\mathcal{L}(x, \mu) = x_1 + c(x_1^2 + x_2^2 - 1) + \mu(-x_1^2 - x_2^2 + 1) \quad \text{and the KKT conditions become}
\]

\[
\nabla_x \mathcal{L}(x, \mu) = \begin{pmatrix} 0 \\
2cx_1 - 2\mu x_1 \\
2cx_2 - 2\mu x_2 \end{pmatrix}, \quad \mu(x_1^2 + x_2^2 - 1) = 0, \mu \geq 0.
\]

Since we are looking for solutions with \(-x_1^2 - x_2^2 + 1 < 0\) the complementarity conditions imply \(\mu = 0\). This gives \(x(c) = (-\frac{1}{\sqrt{c}}, 0)^t\). The constraint \(-x_1(c)^2 - x_2(c)^2 + 1 = -\frac{1}{c} + 1 < 0\) will be satisfied only if \(c < 1/2\). In this case we have \(M_1 = \frac{\sqrt{c}}{2} - c < -1\). Indeed, if \(c > 1/2\) then \((c - (1/2))^2 > 0\) which rewrites as \(\frac{1}{4}c + c > 1\), or, equivalently, \(M_1 = \frac{\sqrt{c}}{2} - c < -1\). In this case, we will have \(M(c) < -1\) and the solution of \((P_c)\) will be different than \(x^* = (-1, 0)^t\). However, if \(c \geq 1/2\) then all solutions of the KKT system will verify the equality constraint \(x_1^2 + x_2^2 = -1\). Indeed, if \(\mu > 0\) then by the complementarity condition the solution will be in the boundary. If \(\mu = 0\) then \(x(c) = (x_1(c), x_2(c))^t = (-\frac{1}{\sqrt{c}}, 0)^t\) and the condition \(c \geq 1/2\) implies \(x_1(c)^2 + x_2(c)^2 = \frac{1}{c} \leq 1\). Since \(x(c)\) must also verify the constraint \(x_1(c)^2 + x_2(c)^2 \geq 1\), we have equality also in this case. As noted before, in this case we must have \(x^* = (-1, 0)^t\). Hence, the solution of \((P_c)\) for \(c \geq 1/2\) is the solution of our original problem.

When the penalty function is not exact, then we must make \(c \to \infty\). Let us illustrate this fact with the previous example for the quadratic penalty function.

**Example 3.2.13** We have the problem

\[
\min_{x} x_1 \\
\text{s.t. } x_1^2 + x_2^2 = 1.
\]

The quadratic penalized problem, with penalization function \(p_2\) is given by

\[
(P_c) \quad \min_{x} x_1 + c(x_1^2 + x_2^2 - 1)^2 =: q_c(x) \\
x \in \mathbb{R}^n
\]

The unconstrained optimality conditions yield

\[
1 + 4cx_1(x_1^2 + x_2^2 - 1) = 0 \quad \text{and} \quad 4cx_2(x_1^2 + x_2^2 - 1) = 0. \quad (3.2.31)
\]

Call \(x(c) = (x_1(c), x_2(c))\) a solution of \((3.2.31)\). A quick look at these equations tells us that \(x(c)\) must verify \(x_2(c) = 0\) so that we have \(x(c) = (x_1(c), 0)\). Using this fact in the LHS of \((3.2.31)\) we see that \(x_1(c)\) is a root of the third degree polynomial \(\phi_c(t) := t^3 - t + \frac{1}{2c}\). Recall that a third degree polynomial of real coefficients has either one real root or three real roots. One of these roots is always less than or equal to \(-1\), because \(\phi_c(-1) = \frac{1}{2c} > 0\) and \(\phi_c(t) \downarrow -\infty\) when \(t \to -\infty\). From the study of the derivatives we can check that for \(c < \frac{2\sqrt{3}}{3}\) the polynomial \(\phi_c(t)\) has only one real root, which by the previous fact must be less than or equal to \(-1\). If \(c = \frac{2\sqrt{3}}{3}\) then there is another real root of multiplicity two, which is equal to \(1/\sqrt{3}\). If \(c > \frac{3\sqrt{3}}{8}\) (which will eventually hold since \(c \to \infty\)), there are three different real roots, one of them is the one which is less than or equal to
−1 and the other two are strictly positive. For \( c \to \infty \) we have \( \phi_c(t) \to t^3 - t = t(t^2 - 1) \) so the positive roots tend to 0 and 1, and the negative one tends to −1. For \( c \) large enough, we should choose \( x_1(c) \) as the negative root, since this choice of \( x_1(c) \) will give us the smallest objective value. Indeed, it is clear that \( q_c(x_1,0) \geq 0 \) if \( x_1 \geq 0 \). We claim that when \( x_1 < -1 \) and \( c > \frac{1}{16} \) then \( q_c(x_1,0) < 0 \). Indeed, if \( x_1 < -1 \) then \( x_1^2 > 1 \) and from the LHS of (3.2.31) we see that
\[
(x_1^2 - 1) = \frac{-1}{4cx_1}.
\]
Using this equality and the fact that \( x_1^2 > 1 \) in the expression of \( q_c \) we get \( q_c(x_1,0) = x_1 + \frac{1}{16cx_1^2} < -1 + \frac{1}{16} < 0 \) whenever \( c > \frac{1}{16} \). Hence for \( c \) large enough our solution \( x(c) = (x_1(c),0) \) must have \( x_1(c) \) the root of \( \phi_c(t) \) which is less than or equal to −1 and as pointed out before when \( c \to \infty \) this root converges to −1. Therefore \( x(c) \to x^* \) for \( c \to \infty \). Note by the way that the optimal \( x^* \) is only attained in the limit, and hence no exact penalty parameter can be found.

**Remark 3.2.17** Graphing the functions \( q_c \) for increasing values of \( c \) shows that the minima become and more sharp.

**Exercise 3.2.53** Check that the two positive roots \( a_c, b_c \) of \( \phi_c(t) = t^3 - t + \frac{1}{4c} \) verify \( 0 < a_c \leq b_c \).

### 3.2.3 Barrier Methods

Barrier functions are defined in a subset of the original constraint set \( \mathcal{F} \). Let us start by defining this set.

\[
\mathcal{F}_< := \{ x \in \mathbb{R}^n \mid h(x) = 0, \text{ and } g_j(x) < 0 \text{ for all } j = 1, \ldots, r \}.
\]

A function \( \rho : \mathcal{F}_< \to \mathbb{R} \) is called a **barrier function** for \( \mathcal{F} \) when for all \( \bar{x} \in \mathcal{F} \setminus \mathcal{F}_< \) such that there exists a sequence \( \{x_k\} \subset \mathcal{F}_< \) with \( x_k \to \bar{x} \) we must have \( \rho(x_k) \to \infty \). In other words, whenever \( \bar{x} \in \mathcal{F} \) verifies some of the inequality constraints as equality, and there is a sequence in \( \mathcal{F}_< \) converging to \( \bar{x} \), we must have the values of the barrier function tending to infinity. Now consider the problem

\[
(B_r) \quad \min_{x \in \mathbb{R}^n} f(x) + r \rho(x).
\]

The parameter \( r > 0 \) is called the **barrier parameter**. The **barrier method** assumes that \( \mathcal{F}_< \) is not empty and that there exists a global minimizer of problem \((B_r)\). This method generates a sequence \( \{x(r)\} \) of solutions of \((B_r)\), with \( r \to 0 \). In general, the solution of our original problem will be a boundary point of the feasible set. This is the most relevant situation, because it means that the constraints are not superfluous. If the barrier method generates a sequence \( \{x_k\} \) that converges to the solution, then we must have \( \rho(x_k) \to \infty \). Since \( r_k \downarrow 0 \), the **barrier term** \( r \rho(x) \) becomes more and more ill-behaved.

**Example 3.2.14** Let us give now two of the most usual examples of barrier functions for a feasible set given by inequality constraints only, i.e., a feasible set of the form

\[
\mathcal{F} = \{ x \in \mathbb{R}^n \mid g(x) \leq 0 \}.
\]

(i) The **Logarithmic Barrier function**, also called Frisch’s barrier function, defined as

\[
\rho_1(x) := -\sum_{j=1}^{r} \ln (-g_j(x)).
\]

(ii) The **inverse** barrier function, given by

\[
\rho_2(x) := -\sum_{j=1}^{r} \frac{1}{g_j(x)}.
\]
Exercise 3.2.54 Prove that the barrier functions given in Example 3.2.14 are continuous in $F_<$ and verify the barrier property. Prove that the barrier functions are convex when the $g_j$ are convex.

Exercise 3.2.55 Assume that $\rho_1$ and $\rho_2$ are barrier functions for $F$. Prove the statements below.

(i) $q(x) := \alpha_1 \rho_1(x) + \alpha_2 \rho_2(x)$ is a barrier function for $F$ for every $\alpha_1 > 0$, $\alpha_2 > 0$.

(ii) $q(x) := c \rho_1(x) \rho_2(x)$ is a barrier function for $F$ for every $c > 0$.

(iii) $q(x) := (\rho_1(x))^k + (\rho_2(x))^k$ is a barrier function for $F$ for every $k > 0$.

A similar convergence result to that of the penalty method holds.

Theorem 3.2.42 Assume the feasible set $F$ verifies that every point $x \in F \setminus F_<$ can be approximated by a sequence in $F_<$ in $\mathbb{R}^n$. In other words, for all $x \in F \setminus F_<$ there exists a sequence $\{z_j\} \subset F_<$ such that $z_j \to x$. Suppose also that the barrier function $\rho$ is continuous in the set $F_<$. Take the barrier parameters $\{r_k\}$ such that $0 < r_{k+1} < r_k$ and $r_k \to 0$. If the global minimizers $x_k$ of $(B_{r_k})$ converge to some $\bar{x}$, then $\bar{x}$ is a solution of problem (2.9.50).

Proof. Since the feasible set is closed and $x_k$ converges $\bar{x}$, we must have $\bar{x} \in F$. We must show that $f(x) \geq f(\bar{x})$ for every $x \in F$. Let us start by proving this fact when $x \in F_<$. In this case the definition of $x_k$ gives

$$f(x_k) + r_k \rho(x_k) \leq f(x) + r_k \rho(x).$$

If $\bar{x} \in F_<$, then by continuity of $\rho$ we have $\rho(x_k) \to \rho(\bar{x}) \in \mathbb{R}$. Combining this with the assumption $r_k \to 0$ yields $f(\bar{x}) \leq f(x)$. If $\bar{x} \notin F_<$ then by definition of barrier function we have $\rho(x_k) \to \infty$ and hence $\rho(x_k) \geq \rho(x)$ for $k$ large enough. Using this fact in the above inequality yields

$$f(x_k) \leq f(x),$$

for $k$ large enough. Taking now limit for $k \to \infty$ and using the assumption that $x_k$ converges to $\bar{x}$ we get $f(\bar{x}) \leq f(x)$ also when $\bar{x} \notin F_<$. Now take $x \in F \setminus F_<$. By our assumption, there exists a sequence $\{z_j\} \subset F_<$ such that $z_j \to x$. Using the previous argument for each fixed $z_j$ we have $f(z_j) \leq f(\bar{x})$. Using now the continuity of $f$ and the fact that $z_j \to x$ we get $f(\bar{x}) \leq f(x)$. The proof is complete. □

Let us illustrate the barrier method with a problem similar to Example 3.2.13.

Example 3.2.15 We have now the inequality constrained problem

$$\min x_1 \quad \text{s.t.} \quad x_1^2 + x_2^2 \leq 1.$$

It is clear that the solution of this problem is, as in Example 3.2.13, $x^* = (-1, 0)^t$. The logarithmic barrier method applied to this problem generates a sequence $x(r)$ which solves the problems

$$(B_r) \quad \min x_1 - r \ln (1 - x_1^2 - x_2^2)$$

$x \in \mathbb{R}^n$

The unconstrained optimality conditions yield

$$1 + \frac{2r x_1}{1 - x_1^2 - x_2^2} = 0 \quad \text{and} \quad \frac{2r x_2}{1 - x_1^2 - x_2^2} = 0.$$ (3.2.32)

A solution $x(r) = (x_1(r), x_2(r))$ of the system above must verify $x_2(r) = 0$ and $x_1(r)^2 - 2r x_1(r) - 1 = 0$. The last equality gives

$$x_1(r) = r \pm \sqrt{r^2 + 1}.$$
Note that the objective function of the barrier problem must have \(1 - x_1^2 - x_3^2 > 0\). Since \(x_2 = 0\) this means that we must have \(1 - x_1^2 > 0\). But this inequality only holds for \(x_1(r) = r - \sqrt{r^2 + 1}\) since in this case we have \(1 - x_1^2 = 2r(\sqrt{r^2 + 1} - r) > 0\). Therefore, the solution of our system must be \(x_1(r) = r - \sqrt{r^2 + 1} \rightarrow -1\) when \(r \downarrow 0\). Hence as in Example 3.2.13 we again have that the sequence \(x(r) \rightarrow (-1, 0)\). Also in this case the minimizers of the barrier problems never coincide with the solution of the original problem. Moreover, the barrier function is not even defined at the solution.

**Exercise 3.2.56** [3, page 383] Consider the inequality constrained problem

\[
\begin{align*}
\min & \quad \tfrac{1}{2}(x_1^2 + x_2^2) \\
\text{s.t.} & \quad 2 - x_1 \leq 0.
\end{align*}
\]

Using the KKT optimality conditions, check that the solution is \(x^* = (2,0)^t\). Prove that the solution \(x(r)\) of the barrier problems

\[
(B_r) \quad \min \frac{1}{2}(x_1^2 + x_2^2) - r \ln (x_1 - 2) \quad x \in \mathbb{R}^n
\]

is given by \(x(r) = 1 + \sqrt{1 + r}\) and \(x_2(r) = 0\). Hint: pose the optimality conditions for the convex problem \(B_r\) and obtain \(x_1(r)\) as the roots of a quadratic polynomial. Choose the only root which verifies \(x_1 > 2\).
Appendix A

Topology of $\mathbb{R}^n$

We collect here some essential definitions and results from Analysis and Linear Algebra.

A.1 Notation

If $S$ is a set and $x$ is an element of $S$ we write $x \in S$. We denote by $\mathbb{R}$ the real line and by $\mathbb{R}^n$ the set of $n$-dimensional vectors with coordinates in $\mathbb{R}$. The coordinates of an $n$ dimensional vector $x \in \mathbb{R}^n$ is indicated as $x_i$ for $i = 1, \ldots, n$. Vectors in $\mathbb{R}^n$ are seen as column vectors, and the transpose of a column vector $x$ is denoted as $x^t$, and is a row vector. The set of $n \times m$ matrices of real coordinates is denoted by $\mathbb{R}^{n \times m}$, and each of the $n \cdot m$ coordinates or components of a matrix $A \in \mathbb{R}^{n \times m}$ is written $A(i, j)$ or $a_{ij}$ with $i = 1, \ldots, n$ and $j = 1, \ldots, m$. The transpose of the matrix $A \in \mathbb{R}^{n \times m}$ is denoted by $A^t$ and verifies $A^t(i, j) = A(j, i)$. If $w \in \mathbb{R}^n$, the notation $w \geq 0$ indicates that $w_i \geq 0$ for all $i = 1, \ldots, n$. The notation $w > 0$ indicates that $w_i > 0$ for all $i = 1, \ldots, n$. If $w, v \in \mathbb{R}^n$, the notations $w \geq v$ or $w > v$ mean that $w - v \geq 0$ or $w - v > 0$.

A.2 Norms and Scalar Product

Norms are functions which are used for measuring the distance between points or sets in $\mathbb{R}^n$. We say that a function $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}$ is a norm if it satisfies the following properties:

N-1) $\|x\| \geq 0$ for all $x \in \mathbb{R}^n$,
N-2) $\|\lambda x\| = |\lambda| \|x\|$ for all $x \in \mathbb{R}^n$ and all $\lambda \in \mathbb{R}$,
N-3) $\|x + y\| \leq \|x\| + \|y\|$ for all $x, y \in \mathbb{R}$.

Property (3) above is known as the triangle inequality. The following are examples of norms in $\mathbb{R}^n$. The Euclidean norm is defined as

$$\|x\|_2 := \left( \sum_{i=1}^{n} |x_i|^2 \right)^{\frac{1}{2}},$$

where $x = (x_1, \ldots, x_n)^t$. Note that the definition implies $\|x\| = 0$ if and only if $x = 0$ (take $\lambda = 0$ in N-2). The $L_\infty$ norm on $\mathbb{R}^n$ is given by

$$\|x\|_\infty := \max_{i=1, \ldots, n} |x_i|,$$

while the $L_1$ norm is defined as

$$\|x\|_1 := \sum_{i=1}^{n} |x_i|.$$
The norm we will mostly use is the Euclidean norm. We will see that norms are used almost everywhere in our analysis. A simple and natural instance in which we use norms is when the original problem is too difficult to be solved directly. It is common in this case to approximate it by a simpler problem, which we know how to solve. However, we need to estimate the distance between the solution of the approximated problem and the solution of the original one. This distance is the absolute error of the approximation. We will often work with nonzero directions \(d \in \mathbb{R}^n\). Since we are only interested in the direction the vector is pointing to, we can always take \(|d| = 1\). In this case we say that \(d\) is a unit vector. Many other examples of the use of norms will be met along the course.

### A.3 Scalar Product

The **scalar product** is an operation between two vectors in \(\mathbb{R}^n\). It is called **scalar** because the result of the operation is a real number. We will see below that this number is the cosine of the angle between both vectors, when they are unit vectors (i.e., vectors of norm 1). To know the angle between vectors is essential in optimization. For instance, when devising a method for minimising a function \(f : \mathbb{R}^n \to \mathbb{R}\), we often start with a guess \(x_0\), and then make a step towards a “better” point \(x_1\), from \(x_1\) we go to an even better point \(x_2\) and so on. However, when we stand at a given point in the space, we have to choose a “good” direction among the infinite possible ones. The “good” directions are generally chosen among the directions along which \(f\) decreases. Those directions are defined by the fact that they make an acute angle with the gradient of \(f\) at the current point. Hence we need the scalar product for choosing a good “search” direction.

Given \(x, y \in \mathbb{R}^n\), the **scalar product** between them is denoted by \(\langle x, y \rangle\) and defined as

\[
\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i.
\]

In particular, we have that \(\langle x, x \rangle = \|x\|^2\). The following are important properties of the scalar product.

- **SP-1)** \(\langle \lambda x, y \rangle = \langle x, \lambda y \rangle = \lambda \langle x, y \rangle\) for all \(x, y \in \mathbb{R}^n\) and all \(\lambda \in \mathbb{R}\),
- **SP-2)** \(\langle x, z + y \rangle = \langle x, z \rangle + \langle x, y \rangle\) for all \(x, y, z \in \mathbb{R}^n\),

Note that property SP-1 implies that \(\langle x, 0 \rangle = \langle 0, y \rangle = 0\) for all \(x, y \in \mathbb{R}^n\). Even though the scalar product is defined using the coordinates of the vectors, its value does not depend on the particular coordinate system. Namely,

\[
\langle x, y \rangle = \cos \alpha \|x\| \|y\|, \tag{A.3.1}
\]

where \(\alpha\) is the angle between the vectors \(x\) and \(y\). Hence, a positive scalar product indicates an acute angle between \(x\) and \(y\), while obtuse angles have negative scalar product. Note that perpendicularity is then represented by

\[
\langle x, y \rangle = 0,
\]

because it corresponds to the case in which \(\alpha = \pi/2\). We say in this case that the vectors \(x, y\) are **orthogonal**. A direct consequence of (A.3.1) is the famous Cauchy-Schwartz inequality:

\[
|\langle x, y \rangle| \leq \|x\| \|y\|.
\]

Let us prove that (A.3.1) is true. Equality (A.3.1) is trivial when one of the vectors is zero, because in this case both sides of the equation are zero. Assume now that both vectors are not zero, and consider the two-dimensional subspace \(\Pi\) defined by \(x\) and \(y\). Assume first that \(\|y\| = 1\). Within the subspace \(\Pi\), consider the base \(\{y, y_0\}\), where \(y_0\) belongs to \(\Pi\) and is orthogonal to \(y\) (see Figure
A.3.1. Call $d_1$ and $d_2$ the projections of $x$ onto $y$ and $y_0$ respectively, so that $x = d_1y + d_2y_0$. Hence

$$\langle x, y \rangle = \langle d_1y + d_2y_0, y \rangle = d_1\langle y, y \rangle + d_2\langle y_0, y \rangle = d_1,$$

where we used the properties SP-1, SP-2 as well as the facts $\langle y, y \rangle = \|y\| = 1$ (by assumption on $y$) and $\langle y_0, y \rangle = 0$. Using that $\cos \alpha = d_1/\|x\|$ we get

$$\langle x, y \rangle = \cos \alpha \|x\|,$$

so (A.3.1) holds when $\|y\| = 1$. When $\|y\| \neq 1$ we repeat the same argument with $y' := y/\|y\|$ as the unit vector. The proof of this case is left as an exercise.

An important and very useful property is the Parallelogram Law, which states that

$$\|x + y\|^2 + \|x - y\|^2 = 2(\|x\|^2 + \|y\|^2).$$

Indeed, we have that

$$\|x + y\|^2 = \langle x + y, x + y \rangle = \|x\|^2 + \|y\|^2 + 2\langle x, y \rangle$$

and

$$\|x - y\|^2 = \langle x - y, x - y \rangle = \|x\|^2 + \|y\|^2 - 2\langle x, y \rangle,$$

and the law follows summing up both expressions above.

### A.4 Balls in $\mathbb{R}^n$

Whenever we move in the space, it is important to know whether we are “close” or not to a given point. More generally, we may want to study the whole set of points which are at a specific distance from a given point. This is done by using norms. Fix $x_0 \in \mathbb{R}^n$, then the closed unit ball around $x_0$ is defined as

$$B[x_0, 1] := \{ y \in \mathbb{R}^n \mid \|x_0 - y\| \leq 1 \},$$

while

$$B(x_0, 1) := \{ y \in \mathbb{R}^n \mid \|x_0 - y\| < 1 \},$$

is called the open unit ball around $x_0$. In general, for every $r > 0$, we define

$$B[x_0, r] := \{ y \in \mathbb{R}^n \mid \|x_0 - y\| \leq r \}, \quad B(x_0, r) := \{ y \in \mathbb{R}^n \mid \|x_0 - y\| < r \}$$

the corresponding closed and open balls around $x_0$ of radius $r$. 

![Figure A.3.1: A plane defined by $x, y$](image-url)
A.5 Boundedness

A subset $A \subset \mathbb{R}^n$ is bounded when there is a number $r > 0$ such that for all $x \in A$ we must have $\|x\| \leq r$. This property can be easily expressed using balls.

**Definition A.5.21** We say that a set $A \subset \mathbb{R}^n$ is bounded when there exist $R > 0$ such that $A \subset B[0,R]$.

**Exercise A.5.57** Prove that a set $C \neq \emptyset$ if and only if there exists $\alpha > 0$ such that $C \cap \{x \mid \|x\|^2 \leq \alpha\} \neq \emptyset$.

A.6 Open and closed sets

Balls are essential for defining the nature of a given subset $A \subset \mathbb{R}^n$. More precisely, if $a \in A$ is such that there is a whole ball around $a$ which remains inside $A$, then we say that this point is an interior point of $A$. Note that this statement is equivalent to say that there exists (a small enough) $r > 0$ such that $B(a,r) \subset A$. When such an $r > 0$ doesn’t exist, then $a$ must be a boundary point of $A$. See Figure A.6.2, where $a$ is an interior point and $a'$ is a boundary point. Concepts like interior, boundary and closure are essential for describing the nature of a set.

A given subset of $\mathbb{R}^n$ is said to be open when around every point $x_0$ of the set we can find a ball centered at $x_0$, which is totally included in the set. The formal definition follows.

**Definition A.6.22** Given a set $V \subset \mathbb{R}^n$, we say that $V$ is open when for every $x_0 \in V$, there exists $r > 0$ such that $B(x_0,r) \subset V$.

**Exercise A.6.58** Prove that an open ball is open.

**Exercise A.6.59** Prove the following statements.

$(O_1)$ $\emptyset, \mathbb{R}^n$ are open,

$(O_2)$ arbitrary union of open sets is open,

$(O_3)$ finite intersection of open sets is open.

**Definition A.6.23** Let $A \subset \mathbb{R}^n$. The complement of $A$ is the set $A^c := \{x \in \mathbb{R}^n \mid x \notin A\}$.

Complements of open sets are called closed sets.

**Definition A.6.24** Given a set $F \subset \mathbb{R}^n$, we say that $F$ is closed when $F^c$ is open.

**Exercise A.6.60** Prove that a closed ball is closed.
Exercise A.6.61 Prove the following statements.
(C1) $\emptyset, \mathbb{R}^n$ are closed,
(C2) arbitrary intersection of closed sets is closed,
(C3) finite union of closed sets is closed.

Definition A.6.25 Given a set $V \subset \mathbb{R}^n$, the closure of $V$ is denoted as $\overline{V}$ and is given by
$$\overline{V} := \{ y \in \mathbb{R}^n \mid \text{for all } r > 0, \text{ it holds } B(y,r) \cap V \neq \emptyset \}.$$  

The interior of $V$ is denoted as $V^\circ$ and is given by
$$V^\circ := \{ y \in V \mid \text{there exists } r > 0, \text{ such that } B(y,r) \subset V \}.$$  

The boundary of $V$ is denoted as $\partial V$ and is given by
$$\partial V := \{ y \in \mathbb{R}^n \mid \text{for all } r > 0, \text{ it holds } B(y,r) \cap V \neq \emptyset \text{ and } B(y,r) \cap V^c \neq \emptyset \}.$$  

It is clear that $\overline{V}$ and $\partial V$ are closed sets, while $V^\circ$ is open. For arbitrary $V$, it holds that
(i) $\overline{V} = V^\circ \cup \partial V$,
(ii) $V^\circ \subset V$,
(iii) $\overline{V} \supset V$,
(iv) $F$ is closed if and only if $\overline{F} = F$,
(v) $V$ is open if and only if $V^\circ = V$.

Exercise A.6.62 Prove assertions (i)-(v).

A.7 Convergence of sequences

Every minimisation algorithm generates a sequence of vectors which, ideally, gets closer and closer to a solution of the original problem. This property of “getting closer and closer” to a certain point is expressed mathematically by the concept of convergence. Let us start by defining a sequence in $\mathbb{R}^n$.

A sequence $\{x_k\}_k \subset \mathbb{R}^n$ is a function $x(\cdot) : \mathbb{N} \to \mathbb{R}^n$. A subsequence of the sequence $\{x_k\}_k \subset \mathbb{R}^n$ is obtained by restricting the function $x(\cdot)$ to an infinite subset of $J \subset \mathbb{N}$. For simplicity, we denote the whole sequence $x(\cdot)$ by its image $\{x_k\}_k$, and the subsequence corresponding to the index set $J \subset \mathbb{N}$ as $\{x_k\}_{k \in J}$.

The theorems in sections A.7 and A.8.1 are rather classical and can be found in any good textbook on real-valued analysis.

Definition A.7.26 We say that the sequence $\{x_k\}_k \subset \mathbb{R}^n$ converges to $\bar{x}$ if for every $\varepsilon > 0$ there exists $k_0 \in \mathbb{N}$ such that $\|x_k - \bar{x}\| < \varepsilon$ for all $k \geq k_0$. We denote this situation as $\lim_k x_k = \bar{x}$. We say that the sequence $\{x_k\}_k \subset \mathbb{R}^n$ has a subsequence which converges to $\bar{x}$ when there exists an infinite subset $J \subset \mathbb{N}$ such that the subsequence $\{x_k\}_{k \in J}$ converges to $\bar{x}$. The latter situation is denoted as $\lim_{k \in J} x_k = \bar{x}$. Points which are limits of subsequences of $\{x_k\}_k \subset \mathbb{R}^n$ are called accumulation points of $\{x_k\}_k$.

Exercise A.7.63 Prove that a convergent sequence is always bounded. Prove that the converse is not true.
Example A.7.16 The sequence \( \{x_n\} \) given by \( x^n := (n \frac{1-(-1)^n}{2}, \cos 1/n) \) is unbounded, because the odd terms have first component \((x^n)_1 = n\), so the whole sequence cannot be placed inside a ball. The even terms of the sequence are given by \( x^n := (0, \cos 1/n) \), so they are bounded (contained in the closed unit ball). Moreover, the even terms converge to the vector \((0, 1)\).

Exercise A.7.64 Prove that a bounded sequence \( \{x_k\}_k \) converges to \( \bar{x} \) if and only if it has \( \bar{x} \) as its unique accumulation point.

A crucial result concerning bounded sequences is the following.

Theorem A.7.43 (Bolzano-Weierstrass Theorem) Every bounded sequence has a convergent subsequence.

Exercise A.7.65 Prove that a set \( C \) is unbounded (i.e., not bounded) if and only if it contains a sequence \( \{x_k\}_k \) such that \( \lim_k \|x_k\| = \infty \).

Closedness of a set can be expressed in terms of convergent sequences.

Theorem A.7.44 A set \( F \) is closed if and only if it contains the limits of all its convergent sequences. In other words, \( F \) is closed if and only if for every sequence \( \{x_k\}_k \subset F \) and converging to \( \bar{x} \), it holds that \( \bar{x} \in F \).

Definition A.7.27 A set is called compact when it is closed and bounded.

Example A.7.17 The simplest example of a compact sets are sets of a single element \( \{x\} \). More generally, every set which has a finite number of elements is compact. A convergent sequence, together with its limit, is also a compact set.

We can now combine the definition above with the Bolzano-Weierstrass Theorem to get the following important fact about compact sets.

Theorem A.7.45 A set \( K \) is compact if and only if every sequence \( \{x_k\}_k \subset K \) has a convergent subsequence, which converges to some \( \bar{x} \in K \).

A.8 Calculus in \( \mathbb{R}^n \)

A.8.1 Continuous Functions

We will always deal with continuous functions, since it doesn’t make sense to minimize a function which is not continuous. Let \( U \subseteq \mathbb{R}^n \) and \( V \subseteq \mathbb{R}^m \), and consider a function \( f : U \to V \).

Definition A.8.28 We say that \( f \) is continuous at \( x \in U \) if and only if for every \( \varepsilon > 0 \) there exists \( \delta > 0 \) such that for all \( y \in U \) verifying \( \|y - x\| < \delta \) we have \( \|f(y) - f(x)\| < \varepsilon \). We say that \( f \) is continuous on \( U \) when it is continuous at every point of \( U \).

Continuity of functions can be stated in terms of sequences.

Theorem A.8.46 A function \( f : U \to V \) is continuous at \( x \in U \) if and only if for every sequence \( \{x_k\}_k \subset U \) and converging to \( x \), it holds
\[
\lim_k f(x_k) = f(x).
\]
It is well-known that continuity is preserved by composition and sum of functions. If \( f, g \) are continuous at \( x \) and \( g(x) \neq 0 \), then also \( f/g \) is continuous at \( x \).

Continuous functions combine well with compact sets.

Theorem A.8.47 If \( f : \mathbb{R}^n \to V \) is continuous and \( K \subseteq \mathbb{R}^n \) is compact, then \( f(K) := \{f(x) | x \in K\} \) is compact.
A.8.2 Differentiable Functions

Let $D \subseteq \mathbb{R}^n$ be an open set and consider a function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$. We say that $f$ is differentiable at $x \in D$ if there is an $m \times n$ matrix $J_f(x)$ such that

$$\lim_{\|h\| \to 0} \frac{f(x + h) - f(x) - Df(x)h}{\|h\|} = 0. \quad (A.8.2)$$

The matrix $J_f(x)$ is called the Jacobian of $f$ at $x$. We say that $f$ is differentiable on $D$ when it is differentiable at every point of $D$. Important and useful examples of these derivatives will be shown in the next section.

Gradient, Hessian and Jacobian

Even though definition (A.8.2) is not constructive (i.e. it doesn’t give an explicit expression of $J_f(x)$), we can give explicit expressions for $J_f(x)$ in terms of partial derivatives. Assume first that the function $f$ is real valued (i.e., $m = 1$ in the definition above). In this case the matrix $J_f(x)$ becomes a vector, which is called the gradient of $f$ at $x$. We can express this vector by means of the partial derivatives of $f$. The partial derivatives of $f$ represent the variation of $f$ along the canonical directions $\{e^1, \ldots, e^n\}$ of $\mathbb{R}^n$. Let us define formally these derivatives. Let $D \subseteq \mathbb{R}^n$ be an open set and consider a function $f : D \subseteq \mathbb{R}^n \to \mathbb{R}$. The partial derivatives $\frac{\partial f(x)}{\partial x_i}, \ldots, \frac{\partial f(x)}{\partial x_n}$ of $f$ at $x$ are defined as

$$\frac{\partial f(x)}{\partial x_i} := \lim_{t \to 0} \frac{f(x + te^i) - f(x)}{t}, \quad \text{for } i = 1, \ldots, n,$$

where $e^i := (0, \ldots, 1, 0, \ldots, 0)^t$. The set of $n$ vectors $\{e^1, \ldots, e^n\}$ is called the canonical or standard basis of $\mathbb{R}^n$.

The gradient of $f$ at $x \in D$ is the vector consisting of the partial derivatives as coordinates. In other words, $\nabla f(x) = (\frac{\partial f(x)}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n})$.

Once we can express the derivative of $f$ in terms of partial derivatives, we would like to do the same for constructing a “second derivative” of $f$. In order to do this, we look at each partial derivative $\frac{\partial^2 f(x)}{\partial x_i \partial x_j} : \mathbb{R}^n \to \mathbb{R}$, which is in turn a real valued function, and we compute again its partial derivatives at $x$. We obtain in this way $n^2$ “second derivatives”, denoted as $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ for $i, j = 1, \ldots, n$. The $n \times n$ matrix $\nabla^2 f(x)$ consisting of these coefficients is the Hessian of $f$ at $x$. More precisely, the Hessian of $f$ at $x$ is the matrix $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$ such that $[\nabla^2 f(x)]_{ij} := \frac{\partial^2 f(x)}{\partial x_i \partial x_j}$. If the partial derivatives $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ are continuous for every $i$, then we say that $f$ is $C^1$. If the second derivatives $\frac{\partial^2 f(x)}{\partial x_i \partial x_j}$ are continuous for all $i, j$, we say that $f$ is $C^2$. It is a well-known fact that the Hessian is a symmetric matrix when $f$ is $C^2$.

Now consider $f : D \subseteq \mathbb{R}^n \to \mathbb{R}^m$. Then $f$ can be written in terms of its coordinates $\{f_1, \ldots, f_m\}$ as $f(x) = (f_1(x), \ldots, f_m(x))^t$. The first derivative of $f$ can then be expressed in terms of the gradients of each $f_i$. Since we have $m$ of these coordinate functions, we get $m$ gradients, yielding now an $m \times n$ matrix of partial derivatives. This matrix is the Jacobian of $f$, and it is expressed explicitly as

$$J_f(x) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \cdots & \frac{\partial f_1(x)}{\partial x_n} \\ \frac{\partial f_2(x)}{\partial x_1} & \cdots & \frac{\partial f_2(x)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1} & \cdots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \nabla f_1(x) \\ \nabla f_2(x) \\ \vdots \\ \nabla f_m(x) \end{pmatrix}.$$ 

If each coordinate function $f_i$ is $C^1$, we say that $f$ is $C^1$. 


A.9 Useful facts from Linear Algebra

Definition A.9.29 Let \( \{v_1, \ldots, v_n\} \subset \mathbb{R}^n \), consider the expression
\[
v := \alpha_1 v_1 + \cdots + \alpha_n v_n.
\]

1) If \( \alpha_1, \alpha_2, \ldots, \alpha_n \in \mathbb{R} \), then \( v \) is called a linear combination of the vectors \( v_1, \ldots, v_n \).

2) If \( \alpha_1 + \alpha_2 + \cdots + \alpha_n = 1 \), then \( v \) is called an affine combination of the vectors \( v_1, \ldots, v_n \).

3) If \( \alpha_1 + \alpha_2 + \cdots + \alpha_n = 1 \) and \( \alpha_i \geq 0 \) for all \( i \), then \( v \) is called a convex combination of the vectors \( v_1, \ldots, v_n \).

Definition A.9.30 A set of vectors \( \{v_1, \ldots, v_n\} \subset \mathbb{R}^n \) is said to be linearly independent when every linear combination resulting in the zero vector must have \( \alpha_i = 0 \) for all \( i \). In other words, whenever \( 0 = \alpha_1 v_1 + \cdots + \alpha_n v_n \), we must have \( \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0 \). Otherwise, the set is said to be linearly dependent. A set of vectors \( \{v_1, \ldots, v_p\} \subset \mathbb{R}^n \) is said to generate or span \( \mathbb{R}^n \) when every element of \( \mathbb{R}^n \) can be expressed as a linear combination of \( \{v_1, \ldots, v_p\} \). We denote this situation by \( \text{span}\{v_1, \ldots, v_p\} = \mathbb{R}^n \). A set of vectors is said to be a basis of \( \mathbb{R}^n \) if it is linearly independent and generates \( \mathbb{R}^n \).

The following theorem is well known.

Theorem A.9.48 Every independent set can be extended to a basis of \( \mathbb{R}^n \). Every basis of \( \mathbb{R}^n \) has \( n \) elements.

Definition A.9.31 A subset \( S \subset \mathbb{R}^n \) is called a subspace if it is nonempty and closed for the basic operations of the linear space. In other words, if

1. \( 0 \in S \),
2. For all \( x, y \in S \), it holds \( x + y \in S \),
3. For all \( x \in S \) and \( \lambda \in \mathbb{R} \), it holds \( \lambda x \in S \).

The dimension of \( S \) is the number of elements of a basis of \( S \). A subset \( V \in \mathbb{R}^n \) is called an affine set if it is closed for affine combinations. In other words, whenever \( x, y \in V \), then \( \lambda x + (1-\lambda)y \in V \) for all \( \lambda \in \mathbb{R} \).

Remark A.9.18 It can be seen that every affine space is a translation of a subspace. In other words, for every \( V \) affine space, there exists a subspace \( S \) and a point \( x_0 \in V \) such that \( V = x_0 + S := \{x_0 + y \mid y \in S\} \). The dimension of \( V \) is the dimension of the subspace \( S \). The subspace \( S \) is unique and it is called the parallel subspace to \( V \).

Definition A.9.32 Given a subset \( S \subset \mathbb{R}^n \) we can define its Affine Hull, denoted by \( \text{Aff}[S] \), as the intersection of all affine spaces containing \( S \). In other words, the set \( \text{Aff}[S] \) is the smallest affine space containing \( S \).

Example A.9.18 If \( S \subset \mathbb{R}^n \) is such that \( S^o \neq \emptyset \), then \( \text{Aff}[S] = \mathbb{R}^n \). If \( S \) is an affine space, then \( \text{Aff}[S] = S \). If \( S = \{a, b\} \) two different points in \( \mathbb{R}^n \), then \( \text{Aff}[S] \) is the line passing through these two points.

Example A.9.19 Suppose that \( S = \{(x,0) \in \mathbb{R}^2 : x^2 \leq 1\} = (-1,1) \times \{0\} \). Then \( \text{Aff}[S] = \mathbb{R} \times \{0\} \).
A.9.1 Matrices and Linear transformations

A function \( f : \mathbb{R}^n \to \mathbb{R}^m \) is said to be linear if it preserves the basic operations between linear spaces. More precisely, if

1. \( f(x + y) = f(x) + f(y) \) for all \( x, y \in \mathbb{R}^n \),
2. \( f(\lambda x) = \lambda f(x) \) for all \( x \in \mathbb{R}^n \) and all \( \lambda \in \mathbb{R} \).

Two important subspaces are associated with a linear function.

**Definition A.9.33** Given \( f : \mathbb{R}^n \to \mathbb{R}^m \) a linear function, the null space or kernel of \( f \) is the subspace of \( \mathbb{R}^n \) given by

\[
N(f) := \{ x \in \mathbb{R}^n \mid f(x) = 0 \},
\]

and the range of \( f \) is the subspace of \( \mathbb{R}^m \) given by

\[
R(f) := \{ y \in \mathbb{R}^m \mid \text{there exists } x \in \mathbb{R}^n \text{ such that } f(x) = y \}.
\]

For defining a linear function, it is enough to know its values in a basis.

**Theorem A.9.49** Every linear transformation is uniquely determined by its values at a basis of vectors of its domain.

As a consequence of the theorem above, every linear function \( f : \mathbb{R}^n \to \mathbb{R}^m \) can be associated with an \( m \times n \) matrix.

**Definition A.9.34** (Matrix associated with a linear function) Let \( f : \mathbb{R}^n \to \mathbb{R}^m \) be a linear function and fix two bases \( \mathcal{B}_n := \{ v_1, \ldots, v_n \} \) and \( \mathcal{B}_m := \{ w_1, \ldots, w_m \} \) of \( \mathbb{R}^n \) and \( \mathbb{R}^m \), respectively. Then we can write

\[
\begin{align*}
 f(v_1) &= a_{11}w_1 + a_{21}w_2 + \cdots + a_{m1}w_m, \\
 f(v_2) &= a_{12}w_1 + a_{22}w_2 + \cdots + a_{m2}w_m, \\
 & \vdots \quad \vdots \quad \vdots \\
 f(v_n) &= a_{1n}w_1 + a_{2n}w_2 + \cdots + a_{mn}w_m.
\end{align*}
\]

The \( mn \) coefficients \( a_{ij} \) obtained above are unique and they define the \( m \times n \) matrix

\[
M(f) := \begin{pmatrix}
 a_{11} & a_{12} & \cdots & a_{1n} \\
 a_{21} & a_{22} & \cdots & a_{2n} \\
 \vdots & \vdots & \ddots & \vdots \\
 a_{m1} & a_{m2} & \cdots & a_{mn}
\end{pmatrix}.
\]

The matrix \( M(f) \) is called the matrix representation of \( f \) with respect to the given bases. Of course, the coefficients of the matrix \( M(f) \) depend on the choice of the bases \( \mathcal{B}_n \) and \( \mathcal{B}_m \).

**Exercise A.9.66** Assume that the bases \( \mathcal{B}_n \) and \( \mathcal{B}_m \) in Definition A.9.34 are the canonical ones. Prove that \( x \in N(f) \) if and only if \( M(f)x = 0 \). Prove that \( R(f) \) is the subspace spanned by the columns of \( M(f) \).

We observed above that, to every linear transformation \( f \), we can associate a matrix. Conversely, every matrix \( A \in \mathbb{R}^{m \times n} \) can be associated with the linear transformation \( f_A : \mathbb{R}^n \to \mathbb{R}^m \) given by \( f_A(x) := Ax \). Therefore we can define null space and range of the matrix \( A \), as the null space and range of the transformation \( f_A \).
Definition A.9.35 Given a matrix $A \in \mathbb{R}^{m \times n}$, the null space of $A$ is the subspace of $\mathbb{R}^n$ given by

$$N(A) := \{ x \in \mathbb{R}^n \mid Ax = 0 \}.$$ 

The image or range of the matrix $A$ is the subspace of $\mathbb{R}^m$ given by

$$R(A) := \{ y \in \mathbb{R}^m \mid \exists x \in \mathbb{R}^n \text{ with } Ax = y \}.$$ 

The column rank of $A$ is the maximum number of columns of $A$ which are linearly independent. Analogously, the row rank of $A$ is the maximum number of rows of $A$ which are linearly independent. It is a simple fact that these numbers are equal, and its common value is called the rank of $A$. The column space of $A$ is the subspace (of $\mathbb{R}^m$) spanned by the columns of $A$. Analogously, the row space of $A$ is the subspace (of $\mathbb{R}^n$) spanned by the rows of $A$. Given $A \in \mathbb{R}^{m \times n}$, the transpose of $A$ is an $n \times m$ matrix denoted by $A^t$ and defined as:

$$[A^t]_{ij} := a_{ji}, \text{ for all } i = 1, \ldots, n, j = 1, \ldots, m.$$ 

Remark A.9.19 Note that the matrix $A$ becomes the matrix associated with $f_A$ when the bases are the canonical ones. Therefore, by Exercise A.9.66, $R(A)$ is the subspace of $\mathbb{R}^m$ spanned by the columns of $A$.

A.9.2 Orthogonal subspaces

We say that $x$ is orthogonal to $y$ when $x^t y = 0$. We say that a subspace $S_1$ is orthogonal to a subspace $S_2$ when $x^t y = 0$ for all $x \in S_1$ and all $y \in S_2$. We denote this situation by $S_1 \perp S_2$, $S_1 = S_2^\perp$ or $S_1^\perp = S_2$.

Definition A.9.36 We say that two subspaces $S_1, S_2 \subset \mathbb{R}^n$ are orthogonal complements of each other if and only if

1. their sum spans the whole space $\mathbb{R}^n$,
2. $S_1 \perp S_2$.

We denote the situation above as $S_1 \oplus \perp S_2 = \mathbb{R}^n$.

Remark A.9.20 Given a subspace $S \subset \mathbb{R}^n$, we can point out to the following facts.

(I) There exists a unique subspace $S^\perp$ such that $S \oplus \perp S^\perp = \mathbb{R}^n$.

(II) $(S^\perp)^\perp = S$.

The inclusion $(S^\perp)^\perp \subset S$ follows from (I) and is left as an exercise. The opposite inclusion $(S^\perp)^\perp \supset S$ follows directly from the definitions.

Theorem A.9.50 (Basic Theorem of Linear Algebra) Let $A \in \mathbb{R}^{m \times n}$ be a matrix of rank $r$. Then

1. $\dim N(A) + r = n$,
2. $\dim R(A^t) = \dim R(A) = r$,
3. $\dim N(A^t) + r = m$,
4. $N(A) \oplus \perp R(A^t) = \mathbb{R}^n$, in other words $N(A)^\perp = R(A^t)$,
5. $N(A^t) \oplus \perp R(A) = \mathbb{R}^m$, in other words $N(A) = R(A^t)^\perp$. 

A.9.3 Square Matrices and Eigenvalues

The basic result regarding square matrices is the following theorem, which characterises nonsingular matrices.

**Theorem A.9.51** The following statements are equivalent for a matrix \( A \in \mathbb{R}^{n \times n} \).

1. \( N(A) = 0 \),
2. \( R(A) = \mathbb{R}^n \),
3. For every choice of \( b \in \mathbb{R}^n \), the linear system \( Ax = b \) has a unique solution,
4. \( \det A \neq 0 \).
5. \( A \) is nonsingular.
6. \( A^t \) is nonsingular.
7. There exists a unique matrix \( B \in \mathbb{R}^{n \times n} \) such that \( AB = BA = I \).
8. The rows of \( A \) are linearly independent.
9. The columns of \( A \) are linearly independent.

In the above situation, the matrix \( B \) is called the inverse of \( A \). We denote the inverse of \( A \) as \( A^{-1} \).

A central question in Numerical Analysis and Optimization, and a very important problem in applied mathematics, is the study of eigenvectors and eigenvalues of a square matrix. More precisely, let \( A \in \mathbb{R}^{n \times n} \) and consider the problem of finding a nonzero vector \( x \) (in \( \mathbb{R}^n \) or \( \mathbb{C}^n \)) and a scalar \( \lambda \) (in \( \mathbb{R} \) or \( \mathbb{C} \)) such that

\[ Ax = \lambda x. \]

The vector \( x \) is called an eigenvector of \( A \), and \( \lambda \) its corresponding eigenvalue. The set of all eigenvalues of \( A \) is called the spectrum of \( A \). It is well known that the eigenvalues of \( A \) are the solutions of the equation \( \det (A - \lambda I) = 0 \). The polynomial \( \det (A - \lambda I) \) is called the characteristic polynomial of \( A \). Since the characteristic polynomial of \( A \) has degree \( n \) on \( \lambda \), \( A \) has at most \( n \) different eigenvalues, each of them associated with an eigenvector.

**Definition A.9.37** Let \( A \in \mathbb{R}^{n \times n} \). We say that \( A \) is symmetric when \( a_{ij} = a_{ji} \) for all \( i = 1, \ldots, n \) and all \( j = 1, \ldots, m \). We say that \( A \) is positive semidefinite whenever \( x^tAx \geq 0 \) for all \( x \in \mathbb{R}^n \). If we have \( x^tAx > 0 \) for all nonzero \( x \in \mathbb{R}^n \), we say that \( A \) is a positive definite matrix. We say that a square matrix is orthogonal when \( A^t = A^{-1} \). We will also say a given matrix is indefinite when there exists \( x, y \in \mathbb{R}^n \) such that \( x^tAx > 0 \) and \( y^tAy < 0 \). We say that \( A \) is negative semidefinite when \( x^tAx \leq 0 \) for all \( x \in \mathbb{R}^n \). and negative definite when \( x^tAx < 0 \) for all \( x \neq 0, x \in \mathbb{R}^n \).

**Exercise A.9.67**

1. Show that the matrix

\[ A = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \]

is not positive definite.

2. Show that the matrix

\[ B = \begin{pmatrix} \alpha & -1 \\ -1 & \alpha \end{pmatrix} \]

is positive definite for all \( \alpha > 1 \), positive semidefinite for \( \alpha = 1 \), negative definite for all \( \alpha < -1 \), negative semidefinite for all \( \alpha = -1 \), and indefinite for \( \alpha = 0 \).
Exercise A.9.68 Let $A \in \mathbb{R}^{m \times n}$. Prove that the matrix $AA^t \in \mathbb{R}^{m \times m}$ is always symmetric and positive semidefinite. Prove that a matrix $A \in \mathbb{R}^{m \times n}$ has rank $m$ if and only if the $m \times m$ matrix $AA^t$ is nonsingular. Consequently, if $A \in \mathbb{R}^{m \times m}$ is nonsingular, then $AA^t \in \mathbb{R}^{m \times m}$ is symmetric and positive definite. Hint for the second statement: assume that $A$ has rank $m$. Therefore, the null space of $A^t$ must be zero (why?). Then check that the null space of $AA^t$ must also be zero (indeed, a vector $z \in \mathbb{R}^m$ belongs to the null space of $AA^t$ if and only if $AA^tz = 0$. Pre-multiplying by $z$ we get $0 = z^tAA^tz = (A^tz)^t(A^tz) = \|A^tz\|^2$. In other words, $A^tz = 0$ and since the null space of $A^t$ is zero we must have $z = 0$). This proves that $AA^t$ is nonsingular. For the converse, prove that if the null space of $AA^t$ is zero, then the null space of $A^t$ must be zero, and this gives rank of $A$ equal to $m$.

Exercise A.9.69 Let $A \in \mathbb{R}^{n \times n}$. Prove that
i) $A$ is singular if and only if it has an eigenvalue that is equal to zero.
ii) The eigenvalues of a triangular matrix are equal to its diagonal entries.
iii) The eigenvalues of $A + cI$ are equal to $\lambda + c$, where $\lambda$ is an eigenvalue of $A$.
iv) The eigenvalues of $A^k$ are equal to $\lambda^k$, where $\lambda$ is an eigenvalue of $A$.
v) If $A$ is nonsingular, all eigenvalues of $A$ are nonzero (see item (i)) and the eigenvalues of $A^{-1}$ are equal to $\lambda^{-1}$, where $\lambda$ is an eigenvalue of $A$.
v) The eigenvalues of $A$ and $A^t$ are the same.

Let us recall some basic facts concerning symmetric matrices.

Theorem A.9.52 Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Then
1) All eigenvalues of $A$ are real.
2) $A$ has a set of linearly independent and mutually orthogonal eigenvectors.
3) $A$ is positive semidefinite if and only if all eigenvalues are nonnegative.
4) $A$ is positive definite if and only if all eigenvalues are positive.

Definition A.9.38 Given a symmetric positive definite matrix $A$, the positive square root of $A$ is the unique positive definite matrix $B$ such that $B^2 = A$.

In Optimization, the “sign” of a matrix (i.e., whether the matrix is positive definite, negative definite or indefinite) can be crucial. This sign determines, for instance, whether a given function is or not convex (see Corollary 2.2.1), or whether a given stationary point is or not a minimum (see Theorem 2.8.16). Therefore, a test for checking the sign of a matrix becomes very useful. We state it here. We first the definition of principal minor of a matrix.

Suppose $A \in \mathbb{R}^{n \times n}$ is symmetric. Define $\Delta_k$ to be the determinant of the upper left-hand corner $k \times k$ sub-matrix of $A$ for $1 \leq k \leq n$. In other words, if

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix},$$

then for all $1 \leq k \leq n$ consider the sub-matrix of $A$ given by

$$M_k := \begin{pmatrix} a_{11} & \cdots & a_{1k} \\ \vdots & \ddots & \vdots \\ a_{k1} & \cdots & a_{kk} \end{pmatrix},$$
and let $\Delta_k$ be its determinant. The value $\Delta_k$ is called the $k$-th principal minor of $A$. The test can be formulated as follows.

**Theorem A.9.53** Let $A \in \mathbb{R}^{n \times n}$ be symmetric and let $\Delta_k$ be the $k$-th principal minors of $A$ for $1 \leq k \leq n$. Then

1. $A$ is positive definite if and only if $\Delta_k > 0$ for all $1 \leq k \leq n$.
2. $A$ is negative definite if and only if $(-1)^k \Delta_k > 0$ for all $1 \leq k \leq n$. In other words, $\Delta_1 < 0, \Delta_2 > 0, \Delta_3 < 0, \ldots, (-1)^n \det A > 0$.
3. $A$ is positive semidefinite if $\Delta_k > 0$ for all $1 \leq k \leq n - 1$ and $\Delta_n = 0$.
4. $A$ is negative semidefinite if $(-1)^k \Delta_k > 0$ for all $1 \leq k \leq n - 1$ and $\Delta_n = 0$.

**Remark A.9.21** It is clear that the converse of items (3) and (4) of Theorem A.9.53 do not hold. The zero matrix is both positive and negative semidefinite, and all the principal minors are zero.
Bibliography


