Optimization
(SF1811 / SF1831 / SF1841)

Amol Sasane and Krister Svanberg

Department of Mathematics, Royal Institute of Technology, Stockholm

Contents

Part 1. Linear programming

Part 2. Quadratic optimization

Part 3. Nonlinear optimization

Part 4. Some Linear Algebra

Chapter 1

Introduction

The basic object of study in an optimization problem is a *real-valued* function f defined on a set $\mathcal{F}$:

$$f : \mathcal{F} \to \mathbb{R},$$

and the *optimization problem* is to determine a $\widehat{x} \in \mathcal{F}$ that minimizes f, that is,

$$\text{Find } \widehat{x} \in \mathcal{F} \text{ such that for all other } x\text{'s from } \mathcal{F},\ f(\widehat{x}) \leq f(x).$$

Depending on the nature of $\mathcal{F}$, there are various types of courses, for example combinatorial optimization, calculus of variations, stochastic optimization and so on. In this course $\mathcal{F}$ will be a subset of $\mathbb{R}^n$.

1.1. The basic problem

In other words, we will consider the following problem in this course:

$$\left\{ \begin{array}{ll} \text{minimize} & f(x), \\ \text{subject to} & x \in \mathcal{F}, \end{array} \right. \tag{1.1}$$

where

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix},$$

is the vector of variables and takes values in $\mathbb{R}^n$, $\mathcal{F}$ is a given subset of $\mathbb{R}^n$, and f is a given real-valued function which is defined (at least) on $\mathcal{F}$. The function f is called the *objective function* and $\mathcal{F}$ is called the *feasible set.*

Note that there is no loss of generality in considering only minimization problems, since a maximization problem for f on $\mathcal{F}$ is a minimization problem for $-f$ on $\mathcal{F}$. (Why?)

The course is subdivided into three main parts, depending on the nature of f and $\mathcal{F}$:

(1) Linear programming.

(2) Quadratic optimization.

(3) Nonlinear optimization.

The difficulty level increases as one goes down the above list.

1.1.1. Linear programming. If the objective function is a linear function and the feasible set is given by a bunch of linear inequalities, then the corresponding optimization problem (1.1) is called *linear programming.* Thus the general linear programming problem has the following form:

$$\begin{cases} \text{minimize} & c^\top x, \\ \text{subject to} & Ax \geq b, \end{cases}$$

where $c \in \mathbb{R}^n$ and $b \in \mathbb{R}^m$ are given vectors, and $A \in \mathbb{R}^{m \times n}$ is a given matrix. The inequality "$\geq$" above means that this inequality holds component-wise. Thus there are m scalar inequalities in $Ax \geq b$. This problem is a special case of (1.1), where

$$f(x) = c^\top x \quad \text{and} \quad \mathcal{F} = \{x \in \mathbb{R}^n : Ax \geq b\}.$$

1.1.2. Quadratic optimization. If the objective function is a quadratic function and the feasible set is given by a bunch of linear inequalities, then the corresponding optimization problem (1.1) is called *quadratic optimization.* Thus the general quadratic optimization problem has the following form:

$$\begin{cases} \text{minimize} & \frac{1}{2} x^\top H x + c^\top x, \\ \text{subject to} & Ax \geq b, \end{cases}$$

where $c \in \mathbb{R}^n$, $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $H \in \mathbb{R}^{n \times n}$ is a symmetric matrix. This problem is a special case of (1.1), where

$$f(x) = \frac{1}{2} x^\top H x + c^\top x \quad \text{and} \quad \mathcal{F} = \{x \in \mathbb{R}^n : Ax \geq b\}.$$

1.1.3. Nonlinear optimization. The *nonlinear optimization problem* has the following form:

$$\begin{cases} \text{minimize} & f(x), \\ \text{subject to} & g_i(x) \leq 0, \; i = 1, \dots, m, \end{cases}$$

where f and $g_1, \dots, g_m$ are given functions from $\mathbb{R}^n$ to $\mathbb{R}$. These functions will be assumed to be continuously differentiable, and at least one of them will be assumed to be nonlinear (otherwise, we will have a linear programming problem). The feasible set in this case is given by

$$\mathcal{F} = \{x \in \mathbb{R}^n : g_i(x) \leq 0, \; i = 1, \dots, m\}.$$

1.2. Minimum of a subset of $\mathbb{R}$

Definition 1.1. Let S be a subset of $\mathbb{R}$.

(1) An element $u \in \mathbb{R}$ is said to be an *upper bound of* S if for all $x \in S$, $x \leq u$. If the set of all upper bounds of S is not empty, then S is said to be *bounded above.*

(2) An element $l \in \mathbb{R}$ is said to be a *lower bound of* S if for all $x \in S$, $l \leq x$. If the set of all lower bounds of S is not empty, then S is said to be *bounded below.*

Example 1.2.

(1) The set $S = \{x \in \mathbb{R} : 0 \leq x < 1\}$ is bounded above and bounded below. Any real number y satisfying $1 \leq y$ (for instance 1, 2, 100) is an upper bound of S, and any real number z satisfying $z \leq 0$ (for instance 0, -1) is a lower bound of S.

(2) The set $S = \{n : n \in \mathbb{N}\}$ is not bounded above. Although it is bounded below (any real number $x \leq 1$ serves as a lower bound), it has no upper bound, and so it is not bounded above.

(3) The set[1] $S = \{(-1)^n : n \in \mathbb{N}\}$ is bounded above and bounded below. It is bounded above by 1 and bounded below by -1. More generally, any finite set S is bounded above and below.

(4) The set $S = \left\{\frac{1}{n} : n \in \mathbb{N}\right\}$ is bounded above and bounded below. Any real number x satisfying $1 \leq x$ is an upper bound, and 0 is a lower bound.

(5) The sets $\mathbb{Z}$ and $\mathbb{R}$ are neither bounded above nor bounded below. Indeed, this follows from the inequality $z < z + 1$.

(6) The set $\emptyset$ is bounded above and is bounded below. (Why?) ◊

We now introduce the notions of a least upper bound (also called supremum) and a greatest lower bound (also called infimum) of a subset S of $\mathbb{R}$.

Definition 1.3. Let S be a subset of $\mathbb{R}$.

(1) An element $u_* \in \mathbb{R}$ is said to be a *least upper bound of* S (or a *supremum of* S) if
  (a) u_* is an upper bound of S, and
  (b) if u is an upper bound of S, then $u_* \leq u$.

(2) An element $l_* \in \mathbb{R}$ is said to be a *greatest lower bound of* S (or an *infimum of* S) if
  (a) l_* is a lower bound of S, and
  (b) if l is a lower bound of S, then $l \leq l_*$.

Example 1.4. If $S = \{x \in \mathbb{R} : 0 \leq x < 1\}$, then the supremum of S is 1 and the infimum of S is 0.

Clearly 1 is an upper bound of S.

Now we show that if u is another upper bound, then $1 \leq u$. Suppose not, that is, $u < 1$. Then we have

$$0 \leq u < \frac{u+1}{2} < 1, \tag{1.2}$$

where the first inequality is a consequence of the facts that u is an upper bound of S and $0 \in S$, while the last two inequalities follow using $u < 1$. From (1.2), it follows that the number $\frac{u+1}{2}$ satisfies $0 < \frac{u+1}{2} < 1$, and so it belongs to S. The middle inequality in (1.2) above then shows that u cannot be an upper bound for S, a contradiction. Hence 1 is a supremum.

Next we show that this is the only supremum, since if u_* is another supremum, then in particular u_* is also an upper bound, and the above argument shows that $1 \leq u_*$. But $1 < u_*$ is not possible as 1 is an upper bound, and as u_* is a supremum, u_* must be less than or equal to 1. So it follows that $u_* = 1$.

Similarly one can show that the infimum of S is 0. ◊

In the above example, there was a unique supremum and infimum of the set S. In fact, this is always the case and we have the following result.

Theorem 1.5. *If the least upper bound of a subset S of $\mathbb{R}$ exists, then it is unique.*

Proof. Suppose that u_* and u_*' are two least upper bounds of S. Then in particular u_* and u_*' are also upper bounds of S. Now since u_* is a least upper bound of S and u_*' is an upper bound of S, it follows that

$$u_* \leq u_*'. \tag{1.3}$$

Furthermore, since u_*' is a least upper bound of S and u_* is an upper bound of S, it follows that

$$u_*' \leq u_*. \tag{1.4}$$

From (1.3) and (1.4), we obtain $u_* = u_*'$. □

[1] Note that this set is simply the finite set $\{-1, 1\}$.

Thus it makes sense to talk about *the* least upper bound of a set. The least upper bound of a set S (if it exists) is denoted by

$$\sup S$$

(the abbreviation of 'supremum of S'). Similarly, the infimum of a set S (if it exists) is also unique, and is denoted by

$$\inf S.$$

When the supremum and the infimum of a set belong to the set, then we give them special names, namely the maximum and minimum, respectively, of that set.

Definition 1.6.

(1) If $\sup S \in S$, then $\sup S$ is called a *maximum of* S, denoted by $\max S$.

(2) If $\inf S \in S$, then $\inf S$ is called a *minimum of* S, denoted by $\min S$.

Example 1.7.

(1) If $S = \{x \in \mathbb{R} : 0 \leq x < 1\}$, then $\sup S = 1 \notin S$ and so $\max S$ does not exist. But $\inf S = 0 \in S$, and so $\min S = 0$.

(2) If $S = \{n : n \in \mathbb{N}\}$, then $\sup S$ does not exist, $\inf S = 1$, $\max S$ does not exist, and $\min S = 1$.

(3) If $S = \{(-1)^n : n \in \mathbb{N}\}$, then $\sup S = 1$, $\inf S = -1$, $\max S = 1$, $\min S = -1$.

(4) If $S = \left\{\frac{1}{n} : n \in \mathbb{N}\right\}$, then $\sup S = 1$ and $\max S = 1$. It can be shown that $\inf S = 0$. So $\min S$ does not exist.

(5) For the sets $\mathbb{Z}$ and $\mathbb{R}$, sup, inf, max, min do not exist.

(6) For the set $\emptyset$, sup, inf, max, min do not exist. $\Diamond$

In the above examples, we note that if S is nonempty and bounded above, then its supremum exists. In fact this is a fundamental property of the real numbers, called the *least upper bound property* of the real numbers, which we state below:

$$\boxed{\text{If } S \text{ is a nonempty subset of } \mathbb{R} \text{ having an upper bound, then } \sup S \text{ exists.}}$$

Remark 1.8. In Exercise 1.14 below, given a nonempty set S of $\mathbb{R}$, we define $-S = \{-x : x \in S\}$. One can show that if a nonempty subset S of $\mathbb{R}$ is bounded below, then $-S$ is bounded above and so $\sup(-S)$ exists, by the least upper bound property. The negative of this supremum, namely $-\sup(-S)$, can then be shown to serve as the greatest lower bound of S (this is precisely the content of Exercise 1.14). Thus the real numbers also have the 'greatest lower bound property': If S is a nonempty subset of $\mathbb{R}$ having an lower bound, then $\inf S$ exists.

In fact one can define the infimum and supremum of *every* subset S of $\mathbb{R}$ in the extended real line, that is, the set $\mathbb{R}$ together with the symbols $+\infty$ and $-\infty$.

Definition 1.9. Let $S \subset \mathbb{R}$.

(1) If S is not bounded above, then $\sup S = +\infty$.

(2) If S is not bounded below, then $\inf S = -\infty$.

(3) If $S = \emptyset$, then[2] $\sup \emptyset = -\infty$ and $\inf \emptyset = +\infty$.

Exercise 1.10. Provide the following information about the set S

(1) Does $\max S$ exist? If yes, what is it?

(2) Does $\min S$ exist? If yes, what is it?

[2] this makes sense, since *every* real number serves as an upper bound (lower bound).

where S is given by:

(1) $(0,1]$

(2) $[0,1]$

(3) $(0,1)$

(4) $\{\frac{1}{n} : n \in \mathbb{Z} \setminus \{0\}\}$

(5) $\{-\frac{1}{n} : n \in \mathbb{N}\}$

(6) $\left\{\frac{n}{n+1} : n \in \mathbb{N}\right\}$

(7) $\{x \in \mathbb{R} : x^2 \leq 2\}$

(8) $\{0, 2, 10, 2010\}$

(9) $\{(-1)^n \left(1 + \frac{1}{n}\right) : n \in \mathbb{N}\}$

(10) $\{x^2 : x \in \mathbb{R}\}$

(11) $\{\frac{x^2}{1+x^2} : x \in \mathbb{R}\}$.

Exercise 1.11. Determine whether the following statements are TRUE or FALSE.

(1) If u is an upper bound of a subset S of $\mathbb{R}$, and $u' < u$, then u' is not an upper bound for S.

(2) If u_* is the least upper bound of a subset S of $\mathbb{R}$, and ϵ is any positive real number, then $u_* - \epsilon$ is not an upper bound of S.

(3) Every subset of $\mathbb{R}$ has a maximum.

(4) Every subset of $\mathbb{R}$ has a supremum which is a real number.

(5) For every set that has a maximum, the maximum belongs to the set.

Exercise 1.12. Let A and B be subsets of $\mathbb{R}$ such that $A \subset B$. Prove that $\sup A \leq \sup B$.

Exercise 1.13. Let A and B be nonempty subsets of $\mathbb{R}$. Define $A + B = \{x + y : x \in A \text{ and } y \in B\}$. Prove that $\sup(A + B) \leq \sup A + \sup B$.

Exercise 1.14. Let S be a nonempty subset of real numbers that is bounded below. Let $-S$ denote the set of all real numbers $-x$, where x belongs to S. Prove that $\inf S$ exists and $\inf S = -\sup(-S)$.

1.3. Optimal value and optimal solutions

Consider again the central problem in this course, which we label as (P):

$$(P) : \left\{ \begin{array}{ll} \text{minimize} & f(x), \\ \text{subject to} & x \in \mathcal{F}, \end{array} \right. \tag{1.5}$$

The set of values of the function f with domain $\mathcal{F}$ is

$$S := \{f(x) : x \in \mathcal{F}\}.$$

S is a subset of $\mathbb{R}$, and by the previous section, it always has an infimum (which can possibly be $-\infty$ or $+\infty$). We have the following definitions.

Definition 1.15.

(1) The *optimal value* of the problem (P) is $\inf S = \inf\limits_{x \in \mathcal{F}} f(x)$.

(2) A vector $x \in \mathbb{R}^n$ is called a *feasible solution* to the problem (P) if $x \in \mathcal{F}$.

(3) A vector $\widehat{x} \in \mathbb{R}^n$ is called an *optimal solution* to the problem (P) if $\widehat{x} \in \mathcal{F}$ and for all $x \in \mathcal{F}$, $f(\widehat{x}) \leq f(x)$.

If $\widehat{x}$ is an optimal solution to (P), then $f(\widehat{x}) = \min S = \min\{f(x) : x \in \mathcal{F}\} = \min\limits_{x \in \mathcal{F}} f(x)$.

It can happen that the problem (P) has optimal value which is a real number, even though there is no optimal solution. For example, consider the problem (P) for $f : \mathcal{F} \to \mathbb{R}$, when $f(x) = x$

and the feasible set is $\mathcal{F} = (0, 1]$. In this case, the optimal value of (P) is 0, but there is no optimal solution.

1.4. A useful result from real analysis

Through out these notes, the distance used in $\mathbb{R}^n$ will be the one given by the *Euclidean norm.* That is, if

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n,$$

then $\|x\| = \sqrt{x_1^2 + \cdots + x_n^2}$. Thus if $x, y \in \mathbb{R}^n$, the *distance between x and y* is defined as $\|x - y\|$.

A function f from $\mathcal{F}$ $(\subset \mathbb{R}^n)$ to $\mathbb{R}^n$ is said to be *continuous at* $x_0 \in \mathcal{F}$ if for every $\epsilon > 0$, there exists a $\delta > 0$ such that for all $x \in \mathcal{F}$ satisfying $\|x - x_0\| < \delta$, we have that $\|f(x) - f(x_0)\| < \epsilon$. A function f from $\mathcal{F}$ $(\subset \mathbb{R}^n)$ to $\mathbb{R}^n$ is said to be *continuous* if for each $x_0 \in \mathcal{F}$, f is continuous at x_0.

A subset $\mathcal{F}$ of $\mathbb{R}^n$ is called *bounded* if there exists a $R > 0$ such that for all $x \in \mathcal{F}$, $\|x\| \le R$. A subset $\mathcal{F}$ of $\mathbb{R}^n$ is said to be *open* if for each $x_0 \in \mathcal{F}$, there is a $r = r(x_0) > 0$ such that the open ball

$$B(x_0, r) := \{x \in \mathbb{R}^n : \|x - x_0\| < r\}$$

is contained in $\mathcal{F}$. A subset $\mathcal{F}$ of $\mathbb{R}^n$ is said to be *closed* if its complement is open. A closed and bounded subset $\mathcal{F}$ of $\mathbb{R}^n$ is called *compact.*

The following result from real analysis is very useful in optimization.

Theorem 1.16. *Suppose that K is a nonempty compact subset of $\mathbb{R}^n$ and that $f : K \to \mathbb{R}$ is a continuous function. Then*

$$S := \{f(x) : x \in K\}$$

is a nonempty bounded subset of $\mathbb{R}$, and so $\sup S$, $\inf S$ exist. Moreover, they are attained, that is, there exist points $x_1, x_2 \in K$ such that

$$\begin{aligned} f(x_1) &= \max_{x \in K} f(x) = \sup_{x \in K} f(x), \\ f(x_2) &= \min_{x \in K} f(x) = \inf_{x \in K} f(x). \end{aligned}$$

Proof. See for example, Rudin [**R**]. □

Exercise 1.17. Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function such that

$$\lim_{x \to +\infty} f(x) = 0 = \lim_{x \to -\infty} f(x).$$

Show that f must have a (global) maximum or a minimum on $\mathbb{R}$. Give examples to show that it can happen that the function has a maximum and no minimum, and the function has a minimum and no maximum.

Exercise 1.18. In $\mathbb{R}^n$, is the unit sphere $S^{n-1} := \{x \in \mathbb{R}^n : \|x\| = 1\}$ compact? Justify your answer.

Exercise 1.19. $(*)$ Prove that there is a constant C such that if p is any real polynomial of degree 2010, then

$$|p(0)| \le C \int_{-1}^{1} |p(x)| dx.$$

Hint: View the set of polynomials of degree 2010 as a subset of $\mathbb{R}^{2011}$. Consider the continuous function $p \mapsto \frac{|p(0)|}{\int_{-1}^{1} |p(x)| dx}$ on the unit sphere in $\mathbb{R}^{2011}$.

Part 1

Linear programming

Chapter 2

Introduction to linear programming

2.1. What is linear programming?

2.1.1. The problem. Linear programming is a part of the subject of optimization, where the function $f : \mathcal{F} \to \mathbb{R}$ is *linear*, and the set $\mathcal{F}$ is described by *linear* equalities and/or inequalities. Thus, the function f has the form

$$f(x) = c_1x_1 + \cdots + c_nx_n = c^\top x,$$

where $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$ (taking values in $\mathbb{R}^n$) is the variable, and $c = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}$ ($\in \mathbb{R}^n$) is fixed.

The set $\mathcal{F}$ is the set of points in $\mathbb{R}^n$ that satisfy a bunch of linear inequalities[1]:

$$a_{i1}x_1 + \cdots + a_{in}x_n \geq b_i \quad i \in \mathcal{I}.$$

So the *linear programming problem* is: given such an f, minimize (or maximize) f, that is, find a $\widehat{x} \in \mathcal{F}$ such that for all $x \in \mathcal{F}$, $f(\widehat{x}) \leq f(x)$.

2.1.2. Why the name 'linear programming'? To use the adjective 'linear' is obvious, since the function f is linear and the set $\mathcal{F}$ is described by *linear* equalities/inequalities.

But why does one use the word 'programming'? There is a historical reason behind this. The problem arose in the 1940s in an allocation problem in USA's army. And there every $x \in \mathcal{F}$ corresponded to, and was referred to, as a (military) 'program'. So the problem of finding which program $\widehat{x}$ optimized f, was referred to as 'linear programming', and the name has stuck. The history of the problem and principal contributors are shown in the table below:

Kantorovich	1939	Production/transportation planning
Koopmans	WW II	Solution to transportation problems
Dantzig	1947	Simplex method
Khachiyan	1979	Polynomial complexity algorithm
Karmarkar	1984	Polynomial complexity algorithm

In this course, we will study the simplex method, which is a widely used method for solving linear programming problems.

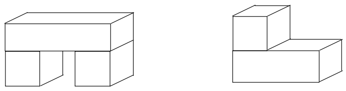
[1] or equalities. But an equality $a_{i1}x_1 + \cdots + a_{in}x_n = b_i$ can be considered to be a pair of inequalities, namely, $a_{i1}x_1 + \cdots + a_{in}x_n \geq b_i$ and $-(a_{i1}x_1 + \cdots + a_{in}x_n) \geq -b_i$

2.1.3. Why study linear programming? The reason is that the need arises in applications. Although linear functions are very simple, linear programming problems arise frequently in practice, for example in economics, networks, scheduling and so on. We will see a very simplified example from production planning in the next section, but other applications will be met along the way. In particular, we will study network flow problems in greater detail in due course. However, now we will begin with simple example.

2.2. An example

2.2.1. The problem. We own a furniture company that produces two kinds of furniture: tables and chairs. It produces these two types of furniture from two types of parts: big parts and small parts. Big parts and small parts look like this:

In order to manufacture a table, one big part and two small parts are used, while to manufacture a chair, one small part and one big part are used:



A table sells for SEK 400 and a chair for SEK 300. Assume that the furniture company has 200 big parts and 300 small parts.

The question is: How many tables and chairs should our company produce so as to maximize its profit?

If the company produces T number of tables and C number of chairs, then the corresponding profit is $400T + 300C$. And we want to maximize this. But there are constraints on T and S. In order to make T tables, we need T big parts and $2T$ small parts, while to make C chairs, we need C big parts and C small parts. So totally we need $T + C$ big parts, which must be less than or equal to 200 and totally we need $2T+C$ small parts, which must be less than or equal to 300. Also the number of chairs and tables cannot be negative. Thus we arrive at the following constraints:

$$\begin{array}{rcl} T + C & \leq & 200 \\ 2T + C & \leq & 300 \\ T & \geq & 0 \\ C & \geq & 0. \end{array}$$

So we have the following problem: maximize $f : \mathcal{F} \to \mathbb{R}$, where $f(T, C) = 400T + 300C$, and

$$\mathcal{F} = \left\{ (T, C) \;\middle|\; \begin{array}{rcl} T + C & \leq & 200 \\ 2T + C & \leq & 300 \\ T & \geq & 0 \\ C & \geq & 0 \end{array} \right\}.$$

So we see that we have a problem of the type described in Section 2.1.

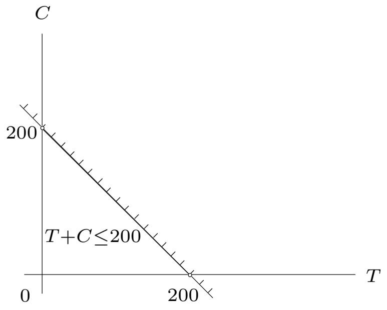


Figure 1. The half plane of points (T, C) in $\mathbb{R}^2$ satisfying $T + C \leq 200$. (The shaded region above the line denotes the set of 'deleted' points, that is, those *not* satisfying the inequality $T + C \leq 200$.)

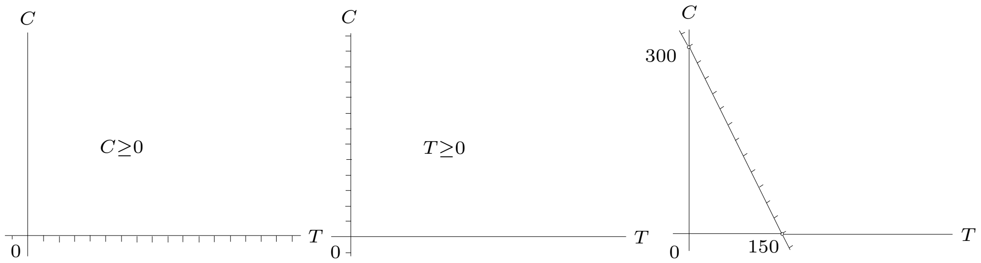


Figure 2. The half planes of points (T, C) in $\mathbb{R}^2$ satisfying $C \geq 0$ and $T \geq 0$, and $2T + C \leq 300$ respectively.

2.2.2. What does the set $\mathcal{F}$ look like? The set of points (T, C) satisfying the inequality $T + C \leq 200$ lie in a half plane as shown in Figure 1.

Similarly, each of the other inequalities describe the half planes depicted in Figure 2.

If *all* the constraints must be satisfied, then we get the *intersection* of all these half planes, namely $\mathcal{F}$ is the following convex polygon shown in Figure 3.

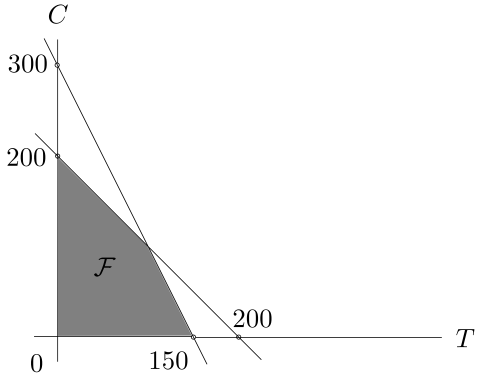


Figure 3. The convex polygon $\mathcal{F}$ (the intersection of the four half planes in Figures 1 and 2).

2.2.3. What elementary calculus tells us. From elementary calculus, we know that the derivative of the function f, (that is, the gradient ∇f) must be zero at an interior maximizer $\widehat{x}$. But the gradient of the function at an $x \in \mathcal{F}$ is $\nabla f(x) = \begin{bmatrix} 400 & 300 \end{bmatrix}$, which is never zero. So the only conclusion we arrive at based on the calculus we have learnt so far, is that if there is a maximizer $\widehat{x} \in \mathcal{F}$, then it must lie on the boundary of $\mathcal{F}$. So this doesn't seem to help much. But now we will see that it is possible to give a graphical solution to the problem.

2.2.4. A graphical solution. Now we will see that it is possible to give a graphical solution to the problem. In order to do this, let us first fix a profit, say P, and look at all the pairs (T, C) that give this profit, that is, (T, C) in $\mathbb{R}^2$ that satisfy $400T + 300C = P$. This represents a straight line in the $\mathbb{R}^2$ plane, which is perpendicular to the line joining the origin and the point $(400, 300)$. For different values of P, we get different lines, which are parallel to each other. For example, if $P = 0$, we get the line ℓ_1 passing through the origin, and if $P = 60000$, we get the line ℓ_2. We see that as P increases, the line moves upwards, and so the profit is maximized when the line is as high as possible, while simultaneously intersecting the set $\mathcal{F}$. This line is labelled by $\ell_{\max}$, and $f(T, C)$ is maximized at the corner point E of the set $\mathcal{F}$, as shown in the Figure 4. The point E is a common point for the lines $T + C = 200$ and $2T + C = 300$, and so E corresponds to the point $(T, C) = (100, 100)$. The maximum profit is thus given by $f(100, 100) = 400 \cdot 100 + 300 \cdot 100 = 70000$ SEK. So we have solved the problem graphically.

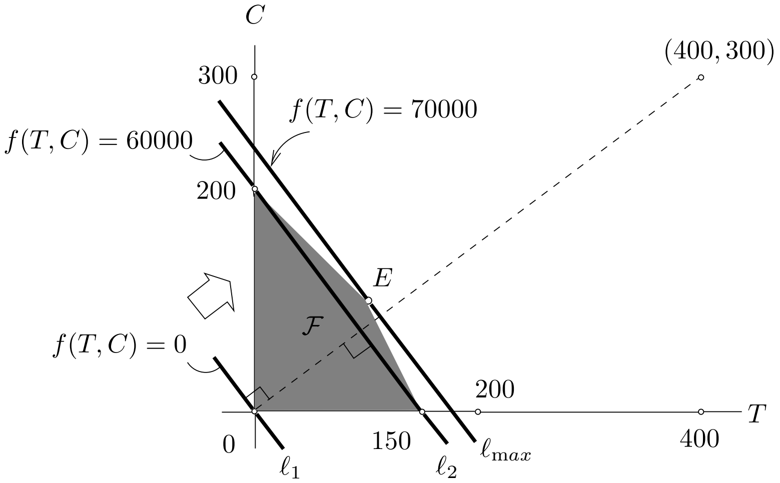
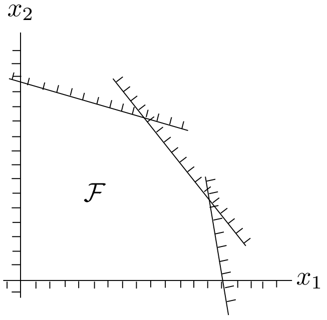


Figure 4. The function $(T, C) \mapsto f(T, C) = 400T + 300C$ is maximized at the extreme point E of the convex polygon $\mathcal{F}$. The arrow shows the direction in which the lines $f(T, C) = P$ move as the P increases.

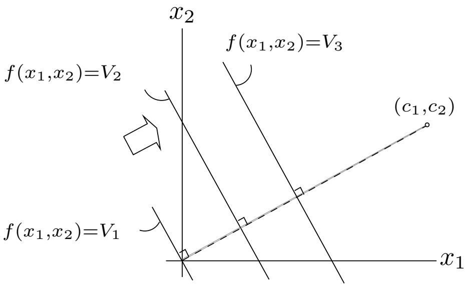
2.2.5. The general case in $\mathbb{R}^2$. More generally in $\mathbb{R}^2$, an inequality of the type $a_{i1}x_1 + a_{i2}x_2 \geq b_i$ $(i \in \mathcal{I})$ determines a half plane, and so the set $\mathcal{F}$ described by $a_{i1}x_1 + a_{i2}x_2 \geq b_i$, $i \in \mathcal{I}$ is again an intersection of half planes, and so it describes a convex polygon.



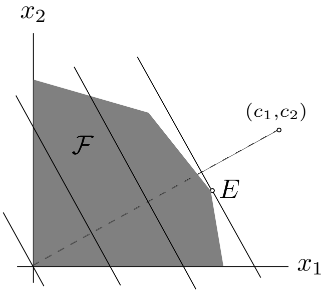
The level sets of the function f to be optimized are straight lines that are perpendicular to the vector c:

$$f(x_1, x_2) = c^\top x = c_1x_1 + c_2x_2 = V,$$

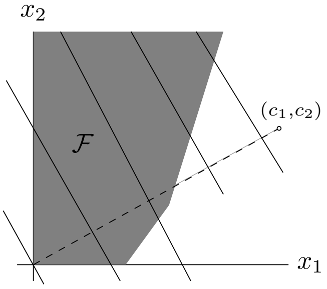
that is they are perpendicular to the line joining $(0, 0)$ and (c_1, c_2); see the following figure.



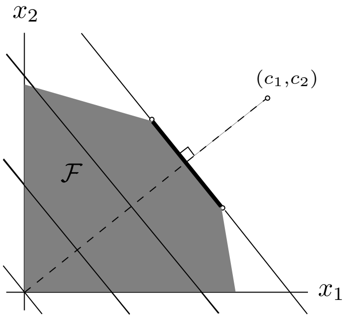
Thus by a reasoning similar to our specific example, we see that the function f is maximized or minimized again at a corner point or an extreme point of the convex polygon, as shown below.



Of course, it may happen that there is no extremizer at all as shown in the following figure, where the set $\mathcal{F}$ is unbounded.

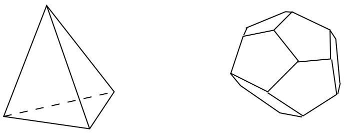


It may also happen that there are infinitely many extremizers; see the figure below, where the vector c is perpendicular to one of the sides of the convex polygon $\mathcal{F}$.



But in any case, we notice that *if there is an extremizer, then there is an extreme point of the convex polygon $\mathcal{F}$ that is an extremizer.* So it suffices to check the extreme points of the convex polygon $\mathcal{F}$.

2.2.6. How does one work in $\mathbb{R}^n$? Generally in $\mathbb{R}^n$, the constraints $a_{i1}x_1 + \cdots + a_{in}x_n \geq b_i$, $i \in \mathcal{I}$ describe half spaces, and so the set $\mathcal{F}$ described by a bunch of these describes a 'convex polytope', just like in $\mathbb{R}^2$, where we obtained a convex polygon. Examples of convex polytopes in $\mathbb{R}^3$ are shown below:



Since the function to be maximized or minimized is linear, once again, the level sets $f(x) = C$ (where C is a constant) are hyperplanes that move parallel to each other. So the function is maximized or minimized at an 'extreme point' of the convex polytope $\mathcal{F}$.

But what exactly do we mean by an 'extreme point of a convex polytope' $\mathcal{F}$? And how do we calculate these? We will learn to determine the extreme points of $\mathcal{F}$ by means of linear algebra. This is the content of the Chapter 4. We will also learn in this chapter that in a linear programming problem, it suffices to check the extreme points of $\mathcal{F}$.

However, in actual applications, this number of extreme points can be terribly large, and calculating all extreme points is not a viable option. There is a way out. Instead of first calculating all extreme points and then checking the values of the function at each of these extreme points, one follows the algorithm shown in Figure 5. This is called the *Simplex Method*, and we will learn this in Chapter 5.

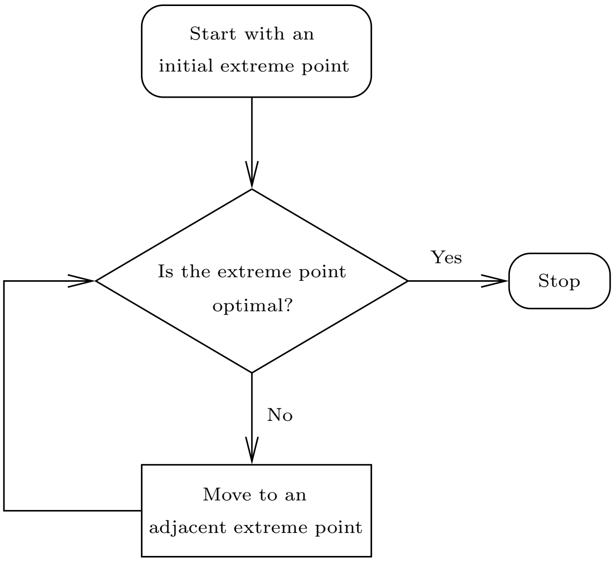

Start with an initial extreme point
Is the extreme point optimal?
Yes
Stop
No
Move to an adjacent extreme point

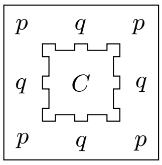
Figure 5. The simplex method.

But first, in the next chapter, we will learn about the *standard form* of the linear programming problem. This is a linear programming problem having a particular form. We will see all types of linear programming problems can be converted to an equivalent linear programming problem in the standard form. In the sequel, we will then learn to solve the linear programming problem in the standard form alone.

Exercise 2.1. Solve the following linear programming problem graphically:

$$\left\{\begin{array}{ll} \text{maximize} & 2x_1 + 5x_2 \\ \text{subject to} & 0 \leq x_1 \leq 4, \\ & 0 \leq x_2 \leq 6, \\ & x_1 + x_2 \leq 8. \end{array}\right.$$

Exercise 2.2. How should one position 28 guards according to the symmetric arrangement shown below around the castle C so as to have the maximum number of guards on each side? Here p and q denote numbers of guards. Pose this as a linear programming problem and solve it graphically.



Chapter 3

The standard form

We saw in the previous chapter that a linear programming problem is an optimization problem in which the function to be optimized is linear and the domain of the function is described by linear inequalities. Depending of the particular application at hand, the exact form of these constraints may differ. However, we will learn in this chapter that it is always possible to convert the given linear programming problem to an equivalent form, called the *standard form*, given below:

$$\text{Minimize } f : \mathcal{F} \to \mathbb{R}$$

where

$$\begin{aligned} f(x) &= c^\top x \;\; (x \in \mathcal{F}) \quad \text{and} \\ \mathcal{F} &= \{x \in \mathbb{R}^n : Ax = b \text{ and } x \geq 0\}. \end{aligned}$$

Here $c \in \mathbb{R}^n$ is a fixed vector, $A \in \mathbb{R}^{m \times n}$ is a fixed matrix and $b \in \mathbb{R}^m$ is a fixed vector. Thus A, b, c are given.

The vector inequality $x \geq 0$ is simply an abbreviation of the n inequalities for its components, that is, $x_1 \geq 0, \ldots, x_n \geq 0$. Thus written out, the linear programming problem in the standard form is:

$$\left\{ \begin{array}{rl} \text{minimize} & c_1x_1 + \cdots + c_nx_n \\ \text{subject to} & a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ & \quad\vdots \\ & a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \\ & x_1 \geq 0, \ \ldots, \ x_n \geq 0. \end{array} \right. \tag{3.1}$$

In the next sections, we will see how various seemingly different linear programming problems can be rephrased as equivalent linear programming problems in the standard form.

3.1. Slack variables

Consider the problem

$$(LP) : \left\{ \begin{array}{rl} \text{minimize} & c_1x_1 + \cdots + c_nx_n \\ \text{subject to} & a_{11}x_1 + \cdots + a_{1n}x_n \leq b_1 \\ & \quad\vdots \\ & a_{m1}x_1 + \cdots + a_{mn}x_n \leq b_m \\ & x_1 \geq 0, \ \ldots, \ x_n \geq 0. \end{array} \right. \tag{3.2}$$

In this case the set $\mathcal{F}$ is determined entirely by linear inequalities. The problem may be alternatively expressed as

$$(LP'):\left\{\begin{array}{ll}\text{minimize} & c_1x_1+\cdots+c_nx_n\\ \text{subject to} & a_{11}x_1+\cdots+a_{1n}x_n+y_1=b_1\\ & \quad\vdots\\ & a_{m1}x_1+\cdots+a_{mn}x_n+y_m=b_m\\ & x_1\geq 0,\ \dots,\ x_n\geq 0,\\ & y_1\geq 0,\ \dots,\ y_m\geq 0.\end{array}\right. \tag{3.3}$$

The newly introduced nonnegative variables y_i convert the inequalities

$$a_{i1}x_1+\cdots+a_{in}x_n\leq b_i$$

to equalities

$$a_{i1}x_1+\cdots+a_{in}x_n+y_i=b_i.$$

The variables y_i are referred to as *slack variables.* By considering the new problem as one having the $n+m$ unknowns $x_1,\dots,x_n,y_1,\dots,y_m$, the problem takes the standard form. The new $m\times(n+m)$ matrix that now describes the linear equalities in the constraints has the special form

$$\left[\begin{array}{cc} A & I\end{array}\right].$$

(Thus the columns have been partitioned into two parts, the first n columns are the columns of the original matrix A, and the last m columns are the columns of the $m\times m$ identity matrix I.)

Example 3.1. Let us revisit the example we had looked at in Section 2.2. With x_1 representing T and x_2 representing C, we had the following problem:

$$\begin{array}{rl}\text{maximize} & 400x_1+300x_2\\ \text{subject to} & x_1+x_2\leq 200\\ & 2x_1+x_2\leq 300\\ & x_1\geq 0,\ x_2\geq 0.\end{array}$$

The problem is not in the standard form. In order to put it into the standard form, we introduce slack variables, so that the problem takes the form:

$$\begin{array}{rl}\text{minimize} & -400x_1-300x_2\\ \text{subject to} & x_1+x_2+y_1=200\\ & 2x_1+x_2+y_2=300\\ & x_1\geq 0,\ x_2\geq 0,\ y_1\geq 0,\ y_2\geq 0.\end{array}$$

Thus we have

$$\begin{array}{rcl} A &=& \begin{bmatrix}1&1&1&0\\2&1&0&1\end{bmatrix},\\ b &=& \begin{bmatrix}200\\300\end{bmatrix},\\ c &=& \begin{bmatrix}-400\\-300\\0\\0\end{bmatrix}.\end{array}$$

The problem is now in standard form. ◊

Suppose we have an inequality of the type $a_{i1}x_1+\cdots+a_{in}x_n\geq b_i$, which is the same as $-a_{i1}x_1-\cdots-a_{in}x_n\leq -b_i$. Then it can be converted into $-a_{i1}x_1-\cdots-a_{in}x_n+y_i=-b_i$ with the introduction of the slack variable $y_i\geq 0$.

So by the introduction of slack variables, any set of linear inequalities can be converted to the standard form if the unknown variables are all nonnegative.

But what happens if one or more of the unknown variables are *not* restricted to be nonnegative? We will see a method of handling this below.

3.2. Free variables

Suppose for example, that the restriction $x_1 \geq 0$ is absent. So x_1 is free to take on any real value.

We can then write $x_1 = u_1 - v_1$, where we demand that $u_1 \geq 0$ and $v_1 \geq 0$. If we substitute $u_1 - v_1$ for x_1 everywhere in (3.1), we observe two things: the linearity of the objective function, and the linearity of the constraints is preserved, and moreover, all variables $u_1, v_1, x_2, \dots, x_n$ are now required to be nonnegative. The problem is now expressed in the $n+1$ variables $u_1, v_1, x_2, \dots, x_n$.

Example 3.2. Consider the problem

$$\begin{array}{rl} \text{minimize} & x_1 + 3x_2 + 4x_3 \\ \text{subject to} & x_1 + 2x_2 + x_3 = 5 \\ & 2x_1 + 3x_2 + x_3 = 6 \\ \text{and} & x_2 \geq 0,\ x_3 \geq 0. \end{array}$$

Since x_1 is unrestricted, we set $x_1 = u_1 - v_1$, where $u_1 \geq 0$ and $v_1 \geq 0$. Substituting this for x_1 everywhere, we obtain the new problem:

$$\begin{array}{rl} \text{minimize} & u_1 - v_1 + 3x_2 + 4x_3 \\ \text{subject to} & u_1 - v_1 + 2x_2 + x_3 = 5 \\ & 2u_1 - 2v_1 + 3x_2 + x_3 = 6 \\ \text{and} & u_1 \geq 0,\ v_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0, \end{array}$$

and this is in the standard form. ♢

Exercise 3.3. Convert the following problems to standard form:

$$(1) \left\{ \begin{array}{rl} \text{minimize} & x + 2y + 3z \\ \text{subject to} & 2 \leq x + y \leq 3 \\ & 4 \leq x + z \leq 5 \\ \text{and} & x \geq 0,\ y \geq 0,\ z \geq 0. \end{array} \right.$$

$$(2) \left\{ \begin{array}{rl} \text{minimize} & x + y + z \\ \text{subject to} & x + 2y + 3z = 10 \\ \text{and} & x \geq 1,\ y \geq 2,\ z \geq 1. \end{array} \right.$$

$$(3) \left\{ \begin{array}{rl} \text{minimize} & |x| + |y| + |z| \\ \text{subject to} & x + 2y = 1 \\ & x + z = 1. \end{array} \right.$$

(See Example 3.7.)

3.3. Some examples

In this section, we list some of the classical application areas where linear programming problems arose. The domain of applications is forever expanding, and no one can tell what new applications might arise in the future. So by no means is the choice of examples complete.

Example 3.4 (The diet problem)**.** How can we determine the most economical diet that satisfies the basic minimum nutritional requirements for good health? Such a problem might be one faced for example by the dietician of an army.

We assume that n different foods are available in the market (for example, spinach, sausages, peas, etc.), and that the jth food sells at a price c_j per unit. In addition, there are m basic nutritional ingredients (carbohydrates, protein, vitamins, etc.). To achieve a balanced diet, each individual must receive at least b_i units of the ith nutrient per day. Finally, we assume that each unit of food j contains a_{ij} units of the ith nutrient.

If we denote by x_j the number of units of food j in the diet, then the problem is to select the x_j's to minimize the total cost, namely, $c_1x_1 + \cdots + c_nx_n$, subject to the nutritional constraints

$$\begin{array}{rcl} a_{11}x_1 + \cdots + a_{1n}x_n & \geq & b_1 \\ & \vdots & \\ a_{m1}x_1 + \cdots + a_{mn}x_n & \geq & b_m \end{array}$$

and the nonnegativity constraints $x_1 \geq 0$, ..., $x_n \geq 0$ on the food quantities. ◇

Example 3.5 (The transportation problem)**.** Quantities $s_1, \dots, s_\ell$, respectively, of a certain product are to be shipped from each of ℓ locations (sources) and received in amounts $d_1, \dots, d_k$, respectively, at each of k destinations. Associated with the shipping of a unit product from the source i to the destination j is a unit shipping cost c_{ij}. We want to determine the amounts x_{ij} to be shipped between each source-destination pair (i, j) so that the shipping requirements are satisfied and the transportation cost is minimized.

We set up an array as shown below:

$$\begin{array}{|ccc|c} x_{11} & \dots & x_{1k} & s_1 \\ \vdots & & \vdots & \vdots \\ x_{\ell 1} & \dots & x_{\ell k} & s_\ell \\ \hline d_1 & \dots & d_k & \end{array}$$

The ith row in this array defines the variables associated with the ith source, while the jth column in this array defines the variables associated with the jth destination. The problem is to select nonnegative x_{ij} in this array so that the sum across the ith row is s_i, the sum down the jth column is d_j, and the transportation cost

$$\sum_{j=1}^{k} \sum_{i=1}^{\ell} c_{ij}x_{ij}$$

is minimized. It is assumed that

$$\sum_{i=1}^{\ell} s_i = \sum_{j=1}^{k} d_j,$$

that is, that the total amount shipped is equal to the total amount received.

Thus we arrive at the following linear programming problem:

$$\begin{array}{rl} \text{minimize} & \displaystyle\sum_{j=1}^{k} \sum_{i=1}^{\ell} c_{ij}x_{ij} \\ \text{subject to} & \displaystyle\sum_{j=1}^{k} x_{ij} = s_i \text{ for } i = 1, \dots, \ell, \\ & \displaystyle\sum_{i=1}^{\ell} x_{ij} = d_j \text{ for } j = 1, \dots, k, \\ & x_{ij} \geq 0 \text{ for } i = 1, \dots, \ell,\ j = 1, \dots, k. \end{array}$$

This is a problem in ℓk variables, and the problem is in standard form, with A being a $(\ell + k) \times (\ell k)$ matrix. ◇

Example 3.6 (The manufacturing problem)**.** A factory is capable of having n different production activities, each of which produces various amounts of m commodities. Each activity can be operated at any level $x_i \geq 0$ but when operated at the unity level the ith activity costs c_i and yields a_{ji} units of the jth commodity. Assuming linearity of the production, if we are given m numbers $b_1, \ldots, b_m$ describing the output requirements of the m commodities, and we wish to minimize the production costs, we obtain the following linear programming problem:

$$\begin{array}{rl} \text{minimize} & c_1x_1 + \cdots + c_nx_n \\ \text{subject to} & a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ & \quad\vdots \\ & a_{m1}x_1 + \cdots + a_{mn}x_n = b_m \\ & x_1 \geq 0, \ \ldots, \ x_n \geq 0. \end{array}$$

$\diamondsuit$

Example 3.7 (Line fitting)**.** An experiment results in m observation points, which are pairs of real numbers:

$$(x_1, y_1), \ \ldots, \ (x_m, y_m).$$

(For example, the x_i's might be the blood pressures of patients and the y_i's might be the corresponding drug dosages given to cure the patient.) It is desired to find a line

$$y = \sigma x + c$$

so that the maximum of all the vertical distances of the observation points to the line is minimized; see Figure 1.

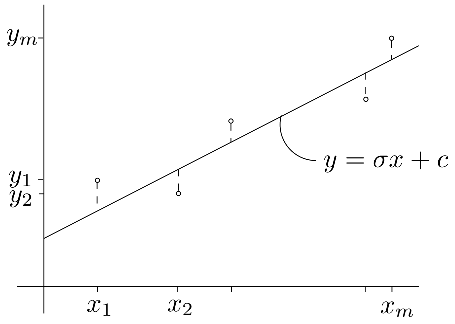


Figure 1. Line fitting through observational data points.

The problem is that of finding the constants σ and c so that the maximum of the m numbers

$$|\sigma x_1 + c - y_1|, \ \ldots, \ |\sigma x_m + c - y_m|$$

is minimized. We can write this as a linear programming problem in the three variables w, σ and c as follows:

$$\begin{array}{rl} \text{minimize} & w \\ \text{subject to} & w \geq \ \ \sigma x_i + c - y_i \text{ for } i = 1, \ldots, m, \\ & w \geq -(\sigma x_i + c - y_i) \text{ for } i = 1, \ldots, m. \end{array}$$

(Why?) This a linear programming problem.

Suppose now that instead we would like to determine σ and c such that the *sum* of the m vertical distances between the line and the given points is minimized, that is, we want to minimize the sum

$$|\sigma x_1 + c - y_1| + \cdots + |\sigma x_m + c - y_m|.$$

We can also formulate this as a linear programming problem in the variables σ, c and $v_1, \dots, v_m$ as follows:

$$\begin{array}{rl} \text{minimize} & v_1 + \cdots + v_m \\ \text{subject to} & v_i \geq \ \ \sigma x_i + c - y_i \text{ for } i = 1, \dots, m, \\ & v_i \geq -(\sigma x_i + c - y_i) \text{ for } i = 1, \dots, m. \end{array}$$

(Why?) This a linear programming problem. ♢

Exercise 3.8. A company manufactures three products called A, B, C. The manufacturing process consists of two phases, called Cutting and Pressing. (Imagine a paper mill.) Each product goes through both these phases.

The department of Cutting, can be used for a maximum of 8 hours per day. Moreover, it has the following capacities for each of the products:

Product	Capacity (in units of product per hour)
A	2000
B	1600
C	1100

The production in the department of Cutting can be switched between the products A,B,C smoothly (so negligible time is wasted).

The department of Pressing, can be used for a maximum of 8 hours per day. Moreover, it has the following capacities for each of the products:

Product	Capacity (in units of product per hour)
A	1000
B	1500
C	2400

The production in the department of Pressing can be switched between the products A,B,C smoothly (so negligible time is wasted).

The profit made per manufactured unit of the products in a day are given as follows:

Product	Profit (in SEK per unit of the product per day)
A	12
B	9
C	8

The company now wants to determine how many units of each product should be produced in a day to make the total profit as large as possible, within the capacity constraints of its two production departments of Cutting and Pressing. Formulate this as a linear programming problem.

Exercise 3.9. A cider company produces four types of cider: Apple, Pear, Mixed and Standard. Every hectoliter of each type of cider requires a certain number of working hours p for production, and a certain number of hours q for packaging. Also the profit v (in units of SEK/hectoliter of cider sold) made for each of these ciders is different depending on the type. These numbers p, q and v for the four types of ciders are specified below:

Cider type	p	q	v
Apple	1.6	1.2	196
Pear	1.8	1.2	210
Mixed	3.2	1.2	280
Standard	5.4	1.8	442

In a week the cider company can spend 80 hours on production and 40 hours on packaging. Also, the company has decided that the Apple cider shall constitute at least 20% of the total volume of cider produced, while the Pear cider shall constitute at most 30% of the total volume of cider produced.

The company wants to decide how much of each sort of cider it should produce in a week so as to maximize its profit under the constraints described above. Formulate this as a linear programming problem.

Exercise 3.10. In a certain city there is a subway line with 12 stations. One year ago, a careful survey of the number of commuters between different pairs of stations was made. In particular, for each pair (i, j) with $i \neq j$ and $i, j \in \{1, \dots, 12\}$, the average number r_{ij} of commuters per day that use the subway to go between station i to station j (that is, enter at station i and exit at station j) was recorded.

As one year has passed since this survey, it is reasonable to expect that these numbers r_{ij} have changed, since many people have changed their residence or place of work in the meantime. So one would like to update this survey. But we don't want to repeat the careful survey done earlier. So suppose we do the following now: for every $i \in \{1, \dots 12\}$ we record the average number p_i of commuters per day that enter the subway at station i, and we record also the average number q_i of commuters per day that leave the subway at station i.

Now we want to replace the old numbers r_{ij} with new numbers x_{ij} that are consistent with the observations p_i and q_j, while differing "as little as possible" from the old numbers r_{ij}. Formulate this as a linear programming problem. Take

$$\max_{ij} |x_{ij} - r_{ij}|$$

as a measure of how much the numbers x_{ij} differ from the numbers r_{ij}.

Exercise 3.11. A factory F has agreed to supply quantities q_1, q_2, q_3 tonnes of a certain product to a customer C at the end of three consecutive months. In each month, the factory can manufacture at most a tonnes of the product, and the cost of manufacturing is c SEK/tonne. But the factory can also use "overtime", and then it can produce an additional maximum of b tonnes per month, but with a manufacturing cost for overtime of d SEK/tonne. It is given that $a > b$ and $d > c$.

The surplus quantities of the product manufactured in a month, but not delivered at the end of the month, can be stored for delivery in another month. The storage cost is s SEK/tonne per month.

If the factory does not supply the agreed quantity each month, then they can deliver the missing quantity at a later month, but no later than the third (=last) month. The agreed fee for being late is f SEK/tonne per month.

At the beginning of month 1, the storage is empty, and we want the storage to be empty at the end of the third month. It is given that $q_1 + q_2 + q_3 < 3a + 3b$.

The company wants to plan its production so that its total cost is minimized. Formulate this as a linear programming problem.

Hint: For each month j, introduce variables for the amount produced with "normal" working time, amount produced with overtime, amount delivered to the customer at the end of the month, amount stored in that month, and amount owed to the customer at the beginning of that month.

Exercise 3.12. Assume that $a_1, \dots, a_m$ are given nonzero vectors in $\mathbb{R}^3$ and that $b_1, \dots, b_m$ are given positive numbers. Let

$$\mathbb{P} = \{x \in \mathbb{R}^3 : a_i^\top x \leq b_i, \; i = 1, \dots, m\}.$$

One can think of $\mathbb{P}$ as a region in $\mathbb{R}^3$ whose "walls" are formed by the planes

$$P_i = \{x \in \mathbb{R}^3 : a_i^\top x = b_i\}, \quad i = 1, \dots, m.$$

Suppose that we want to find the center and the radius of the largest sphere contained in $\mathbb{P}$. Formulate this as a linear programming problem. Use the fact that the distance $d(y, P_i)$ of a point $y \in \mathbb{R}^3$ to the plane P_i is given by the formula

$$d(y, P_i) = \frac{|b_i - a_i^\top y|}{\|a_i\|}.$$

(For a derivation of this formula for the distance of a point to the plane, see Exercise 10.7.)

Chapter 4

Basic feasible solutions and extreme points

Recall that the linear programming problem in the standard form is:

$$(P): \quad \begin{cases} \text{minimize} & c^\top x \\ \text{subject to} & Ax = b \\ \text{and} & x \geq 0, \end{cases}$$

where

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^n$$

is the variable, and $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$ are given.

In this chapter we will see that solving this problem amounts to searching for an optimal solution amongst a *finite* number of points in $\mathbb{R}^n$. These points will be called basic feasible solutions, and they can be computed by linear algebraic calculations. Moreover we will see that these basic feasible solutions really correspond to "corners" or "extreme points" of the feasible set.

4.1. Definitions and the standing assumptions

We will begin with a few definitions.

Definition 4.1.

(1) We call the set $\mathcal{F} = \{x \in \mathbb{R}^n : Ax = b \text{ and } x \geq 0\}$ the *feasible set* of the linear programming problem (P).

(2) A point $x \in \mathcal{F}$ is called a *feasible point* of the linear programming problem (P).

(3) A point $\widehat{x} \in \mathbb{R}^n$ is called an *optimal solution* of the linear programming problem (P) if
   (a) $\widehat{x} \in \mathcal{F}$, and
   (b) for all $x \in \mathcal{F}$, $c^\top \widehat{x} \leq c^\top x$.

Exercise 4.2. Does every linear programming problem in standard form have a nonempty feasible set? If "yes", provide a proof. If "no", give a specific counterexample.

Does every linear programming problem in standard form (assuming a nonempty feasible set) have an optimal solution? If "yes", provide a proof. If "no", give a specific counterexample.

4.1.1. Standing assumptions. We will make the following assumption in the linear programming problem (P):

$$\boxed{A \text{ has rank } m, \text{ that is, } A \text{ has independent rows.}}$$

This has, among others, the following consequences:

(1) $m \leq n$.

(2) The columns of A span $\mathbb{R}^m$. Thus given any vector $b \in \mathbb{R}^m$, we can always be sure that there is at least one $x \in \mathbb{R}^n$ such that $Ax = b$ (although we can't be sure that this x is feasible, since we are not guaranteed in general that such an x will also satisfy the constraint $x \geq 0$).

We make this assumption first of all to avoid trivialities and difficulties of a nonessential nature. Without this assumption, we will have to worry about whether or not $b \in \operatorname{ran} A$ for the solvability of $Ax = b$. Also, if in the original problem some of the rows of A are linearly dependent, we can eliminate those that can be expressed as a linear combination of the other remaining ones, without changing the feasible set. In this manner we can arrive at a matrix A for which the rows *are* linearly independent. So this assumption does not really restrict the class of problems we can solve.

Note that under the above assumption, if in addition we have $n = m$, then the matrix A is a square matrix which is invertible. So the equation $Ax = b$ has precisely one solution, namely $x = A^{-1}b$. Again the problem of optimization becomes a trivial one, since the feasible set is either empty (if it is not the case that $x = A^{-1}b \geq 0$) or has just one point! So in addition to the assumption that the rank of A is m, it is reasonable to also assume in the sequel that

$$\boxed{m < n.}$$

4.2. Basic solutions and feasible basic solutions

In this section we will learn how to calculate "basic feasible solutions" to $Ax = b$. It turns out that the solution to the linear programming problem (P) can be found among these (finitely many!) basic feasible solutions. We will see this later. But now, we will first learn how one calculates these basic feasible solutions.

Let us denote by $a_1, \dots, a_n$ the n columns of A. Thus:

$$A = \begin{bmatrix} a_1 & \dots & a_n \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

Assume that we select m independent columns $a_{\beta_1}, \dots, a_{\beta_m}$ from the n columns of A. Then these chosen columns form a basis for $\mathbb{R}^m$. We have the following notation and terminology:

(1) The tuple $\beta = (\beta_1, \dots, \beta_m)$ is called the *basic index tuple*.

(2) Let A_β be the $m \times m$ matrix of the chosen columns, that is,

$$A_\beta = \begin{bmatrix} a_{\beta_1} & \dots & a_{\beta_m} \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

The matrix A_β is called the *basic matrix* (corresponding to β).

(3) Let

$$x_\beta := \begin{bmatrix} x_{\beta_1} \\ \vdots \\ x_{\beta_m} \end{bmatrix}$$

be the vector of variables corresponding to the chosen columns of A. We call x_β the *basic variable vector*, and we call its components, namely the variables $x_{\beta_1}, \dots, x_{\beta_m}$ the *basic variables*.

We collect the $\ell = n - m$ columns of A that are left over (and which did not go into the basic matrix A_β) into a matrix A_ν. Similarly, the left over components of the variable vector x, which did not go into the basic variable vector x_β, are collected to form a vector x_ν. Thus:

$$A_\nu = \left[\begin{array}{ccc} a_{\nu_1} & \dots & a_{\nu_\ell} \end{array} \right] \in \mathbb{R}^{m \times \ell} \quad \text{and} \quad x_\nu := \left[\begin{array}{c} x_{\nu_1} \\ \vdots \\ x_{\nu_\ell} \end{array} \right]$$

We refer to the tuple $\nu = (\nu_1, \dots, \nu_\ell)$ as the *non-basic index tuple.* The components x_{ν_i} in the vector x_ν are called *non-basic variables.* Similarly, if the vector

$$v = \left[\begin{array}{c} v_1 \\ \vdots \\ v_n \end{array} \right] \in \mathbb{R}^n,$$

then we will use the notation v_β, v_ν to mean the vectors

$$v_\beta = \left[\begin{array}{c} v_{\beta_1} \\ \vdots \\ v_{\beta_n} \end{array} \right] \in \mathbb{R}^m, \quad v_\nu = \left[\begin{array}{c} v_{\nu_1} \\ \vdots \\ v_{\nu_\ell} \end{array} \right] \in \mathbb{R}^\ell.$$

Example 4.3. We revisit Example 3.1. Let

$$A = \left[\begin{array}{cccc} 1 & 1 & 1 & 0 \\ 2 & 1 & 0 & 1 \end{array} \right] \in \mathbb{R}^{2 \times 4} \quad \text{and} \quad b = \left[\begin{array}{c} 200 \\ 300 \end{array} \right] \in \mathbb{R}^2.$$

The system $Ax = b$ can be written as

$$x_1 \underbrace{\left[\begin{array}{c} 1 \\ 2 \end{array} \right]}_{a_1} + x_2 \underbrace{\left[\begin{array}{c} 1 \\ 1 \end{array} \right]}_{a_2} + x_3 \underbrace{\left[\begin{array}{c} 1 \\ 0 \end{array} \right]}_{a_3} + x_4 \underbrace{\left[\begin{array}{c} 0 \\ 1 \end{array} \right]}_{a_4} = \left[\begin{array}{c} 200 \\ 300 \end{array} \right].$$

Suppose that we choose a_3 and a_2 (which are linearly independent). Then $\beta_1 = 3$, $\beta_2 = 2$, and so $\beta = (3, 2)$. Also,

$$A_\beta = \left[\begin{array}{cc} a_3 & a_2 \end{array} \right] = \left[\begin{array}{cc} 1 & 1 \\ 0 & 1 \end{array} \right] \quad \text{and} \quad x_\beta = \left[\begin{array}{c} x_3 \\ x_2 \end{array} \right].$$

Finally, we have $\nu_1 = 1$, $\nu_2 = 4$, $\nu = (1, 4)$,

$$A_\nu = \left[\begin{array}{cc} a_1 & a_4 \end{array} \right] = \left[\begin{array}{cc} 1 & 0 \\ 2 & 1 \end{array} \right] \quad \text{and} \quad x_\nu = \left[\begin{array}{c} x_1 \\ x_4 \end{array} \right].$$

With this β, the basic variables are x_3 and x_2, while the non-basic variables are x_1 and x_4. ♢

For a chosen basis of $\mathbb{R}^m$ from columns of A, and with corresponding index tuples β and ν, the equation $Ax = b$ is the same as

$$A_\beta x_\beta + A_\nu x_\nu = b, \tag{4.1}$$

since

$$A_\beta x_\beta + A_\nu x_\nu = \sum_{i=1}^{m} x_{\beta_i} a_{\beta_i} + \sum_{i=1}^{\ell} x_{\nu_i} a_{\nu_i} = \sum_{i=1}^{n} x_i a_i = Ax = b.$$

Suppose that all non-basic variables are set to 0, that is, $x_\nu = 0$. Then (4.1) gives a unique solution for the basic variables, namely

$$x_\beta = A_\beta^{-1} b.$$

This corresponds to a feasible solution for the problem (P) iff $x_\beta \geq 0$. In light of this, we give the following definitions.

Definition 4.4. Suppose β is a basic index tuple.

(1) A *basic solution* corresponding to β is a solution x to $Ax = b$ such that $A_\beta x_\beta = b$ and $x_\nu = 0$.

(2) A *basic feasible solution* corresponding to β is a basic solution x such that $x_\beta \geq 0$.

(3) A basic feasible solution x such that none of the components of x_β are zero, is called a *non-degenerate* basic feasible solution. (Thus all the components x_{β_i} are positive.)

(4) A basic feasible solution x such that at least one of the components of x_β is zero, is called a *degenerate* basic feasible solution. (Thus all the components x_{β_i} are nonnegative and at least one of them is zero.)

Example 4.5. Consider Example 4.3, where

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 2 & 1 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{2\times 4} \quad \text{and} \quad b = \begin{bmatrix} 200 \\ 300 \end{bmatrix} \in \mathbb{R}^2.$$

Let $\beta = (3, 2)$. For a basic solution corresponding to β, we must then have $x_1 = x_4 = 0$, and

$$A_\beta x_\beta = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x_3 \\ x_2 \end{bmatrix} = \begin{bmatrix} 200 \\ 300 \end{bmatrix} = b,$$

and so

$$x_\beta = \begin{bmatrix} x_3 \\ x_2 \end{bmatrix} = \begin{bmatrix} -100 \\ 300 \end{bmatrix}.$$

Hence

$$x = \begin{bmatrix} 0 \\ 300 \\ -100 \\ 0 \end{bmatrix}$$

is a basic solution corresponding to β. It is not a basic feasible solution, since it is not the case that $x_\beta \geq 0$ (indeed, $x_3 = -100 < 0$).

On the other hand, if we choose $\beta = (1, 2)$, then the basic solution corresponding to β must have $x_3 = x_4 = 0$, and

$$A_\beta x_\beta = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 200 \\ 300 \end{bmatrix} = b,$$

and so

$$x_\beta = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 100 \\ 100 \end{bmatrix}.$$

Hence

$$x = \begin{bmatrix} 100 \\ 100 \\ 0 \\ 0 \end{bmatrix}$$

is a basic solution corresponding to β. It is also a basic feasible solution, since $x_\beta \geq 0$. Moreover, it is a non-degenerate basic feasible solution, since all the components of x_β are positive. ◊

Example 4.6. Now suppose that

$$A = \begin{bmatrix} 3 & 2 & 1 & 1 \\ 2 & 1 & 3 & 1 \end{bmatrix} \in \mathbb{R}^{2\times 4} \quad \text{and} \quad b = \begin{bmatrix} 5 \\ 5 \end{bmatrix} \in \mathbb{R}^2.$$

Let $\beta = (2, 4)$. For a basic solution corresponding to β, we must then have $x_1 = x_3 = 0$, and

$$A_\beta x_\beta = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_2 \\ x_4 \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \end{bmatrix} = b,$$

and so

$$x_\beta = \begin{bmatrix} x_2 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 5 \end{bmatrix}.$$

Hence

$$x = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 5 \end{bmatrix}$$

is a basic solution corresponding to β. It is also a basic feasible solution, since $x_\beta \geq 0$. But it is a degenerate basic feasible solution since at least one of the components of x_β is 0 ($x_2 = 0$). $\Diamond$

Exercise 4.7. Consider the system $Ax = b$, where

$$A = \begin{bmatrix} 2 & -1 & 2 & -1 & 3 \\ 1 & 2 & 3 & 1 & 0 \end{bmatrix}, \quad b = \begin{bmatrix} 14 \\ 5 \end{bmatrix}.$$

Check if the system has basic solutions. If yes, find all basic solutions and all basic feasible solutions.

4.3. The fundamental theorem of linear programming

In this section we will learn the (at first somewhat surprising[1]) result which says that if the linear programming problem has a solution, then there is a basic feasible solution which is an optimal solution[2]. So this means that it is enough to just search for an optimal solution among the (finitely many) basic feasible solutions, since the minimum value (if it exists) is always achieved at such a solution.

Theorem 4.8 (Fundamental theorem of linear programming). *Consider the linear programming problem (P).*

(1) *If there exists a feasible solution, then there exists a basic feasible solution.*

(2) *If there exists an optimal solution, then there exists an optimal basic feasible solution.*

Proof. (1) Suppose that x is a feasible solution and that it has k positive components corresponding to the index tuple $(\gamma_1, \dots, \gamma_k)$ and the rest of the components are 0. Then with the usual notation

$$x_{\gamma_1} a_{\gamma_1} + \cdots + x_{\gamma_k} a_{\gamma_k} = b. \tag{4.2}$$

We now consider the two possible cases:

$\underline{1}^\circ$ $a_{\gamma_1}, \dots, a_{\gamma_k}$ are linearly independent. Then $k \leq m$, since the rank of A is m. If $k = m$, then the solution x is basic, and we are done. Suppose on the other hand that $k < m$. Since the rank of A is m, in addition to our k columns $a_{\gamma_1}, \dots, a_{\gamma_k}$, we can find extra $m - k$ columns of A (from the remaining $n - k$ columns) so that these m columns form a basis for $\mathbb{R}^m$. Hence we now see that the solution x is a degenerate basic feasible solution corresponding to our construction of the m independent columns.

$\underline{2}^\circ$ $a_{\gamma_1}, \dots, a_{\gamma_k}$ are linearly dependent. Then there are k scalars $y_{\gamma_1}, \dots, y_{\gamma_k}$, not all zeros, such that

$$y_{\gamma_1} a_{\gamma_1} + \cdots + y_{\gamma_k} a_{\gamma_k} = 0. \tag{4.3}$$

[1] We will see later that these basic feasible solutions correspond geometrically to corner points of the set $\mathcal{F}$, and so this result is then something we would expect.

[2] A basic feasible solution which is an optimal solution will henceforth be referred to as an *optimal basic feasible solution.*

We may assume that at least one of the y_i is positive (otherwise, we can multiply (4.3) by -1 to ensure this). Now we multiply (4.3) by a scalar ϵ and subtract the resulting equation from (4.2) to obtain

$$(x_{\gamma_1} - \epsilon y_{\gamma_1})a_{\gamma_1} + \cdots + (x_{\gamma_k} - \epsilon y_{\gamma_k})a_{\gamma_k} = b.$$

Set y to be be the vector in $\mathbb{R}^n$ whose $j = \gamma_i$th entry is y_{γ_i} and for j not any of the γ_is, the entry is 0. With this notation, we see that $A(x - \epsilon y) = b$ for all ϵ. Now let

$$\epsilon_* = \min\left\{\frac{x_i}{y_i} : y_i > 0\right\} > 0.$$

Then the components of $x - \epsilon_* y$ are all nonnegative, and at least one amongst the components $x_{\gamma_1} - \epsilon_* y_{\gamma_1}, \dots, x_{\gamma_k} - \epsilon_* y_{\gamma_k}$ is 0. So we have now obtained a feasible solution $x - \epsilon_* y$ with at most $k-1$ positive components. We can now repeat this process if necessary until we get either that all the components of our solution x are zero[3], or the nonzero components of our solution correspond to linearly independent columns of A. In the former case, our zero solution is a (degenerate) basic feasible solution, and we are done. In the latter case, we are in Case 1°, and so this completes the proof of part (1).

(2) Now suppose that x is an optimal solution. We proceed in the same manner as above. So just as before, suppose that x has k positive components corresponding to the index tuple $(\gamma_1, \dots, \gamma_k)$ and the rest of the components are 0. We consider the two cases as above.

The argument in Case 1° is precisely the same as before, and the same x is an optimal feasible basic solution (possibly degenerate).

In the second case, we proceed similarly, but we must also ensure that $x - \epsilon_* y$ is optimal. We will do this by showing that $c^\top y = 0$. Indeed, then we have $c^\top(x - \epsilon_* y) = c^\top x$, and the optimality of $x - \epsilon_* y$ follows from the optimality of x.

Assume, on the contrary, that $c^\top y \neq 0$. Now we choose r to be the real number which has the same sign as $c^\top y$ and such that

$$|r| := \min\left\{\left|\frac{x_i}{y_i}\right| : y_i \neq 0\right\} > 0.$$

Then we claim that the vector $x - ry$ is feasible. Indeed,

(1) if $y_i = 0$, then $x_i - ry_i = x_i - r0 = x_i \geq 0$;

(2) if $y_i > 0$, then $x_i - ry_i \geq x_i - \dfrac{x_i}{y_i} y_i = 0$;

(3) if $y_i < 0$, then $x_i - ry_i \geq x_i + \dfrac{|x_i|}{|y_i|} y_i = x_i + \dfrac{x_i}{-y_i} y_i = 0$.

Then we obtain $c^\top x > c^\top x - rc^\top y = c^\top(x - ry)$. But this contradicts the optimality of x. Hence $c^\top y = 0$.

So we arrive at the conclusion that $x - \epsilon_* y$ is optimal. But now the proof is completed exactly in the same way as the rest of the proof of Case 2° in part (1). □

Exercise 4.9. Suppose that x_0 is a feasible solution to the linear programming problem (P) in the standard form, where A has rank m. Show that there is a feasible solution x to (P) that has at most $m+1$ positive components and such that the objective function has the same value, that is, $c^\top x = c^\top x_0$. *Hint:* Add the constraint $c^\top x = c^\top x_0$.

[3] this can happen if the $a_{\gamma_1}, \dots, a_{\gamma_k}$ were all zero to begin with

4.3.1. Sufficiency of checking basic feasible solutions. The fundamental theorem of linear programming reduces the task of solving the problem (P) to that of searching solutions among the basic feasible solutions. But the number of basic feasible solutions is finite! After all, once we choose m independent columns of the matrix A, there can be at most one basic feasible solution, and the number of ways of selecting m columns from n ones is itself finite, given by

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}.$$

But in applications, this number can be terribly large. For example if $m = 5$ and $n = 50$, then

$$\binom{n}{m} = \binom{50}{5} = 2118760.$$

This would potentially be the number of basic feasible solutions to be checked for optimality. So a more efficient method is needed, and there is indeed such a method, called the *simplex method*.

In the simplex method, we don't calculate all basic feasible solutions like crazy. Instead, once we have a basic feasible solution (which corresponds to a "corner" of $\mathcal{F}$), we calculate a next one (which corresponds to the basic feasible solution of an "adjacent corner") by noticing in what direction the function $x \mapsto c^\top x$ decreases most rapidly. In this manner, we efficiently reach the optimal solution, without having to go through all the basic feasible solutions. We will learn this method in the next chapter, but first we will convince ourselves that basic feasible solutions do correspond to extreme (or corner) points of $\mathcal{F}$.

4.4. Geometric view of basic feasible solutions

In order to see that basic feasible solutions do correspond to corner points of $\mathcal{F}$, we must first of all explain what we mean by corner points. We do this below.

4.4.1. Convex sets and extreme points.

Definition 4.10. A set $C \subset \mathbb{R}^n$ is called *convex* if for all $x, y \in C$ and all $t \in (0,1)$, we have that $(1-t)x + ty \in C$.

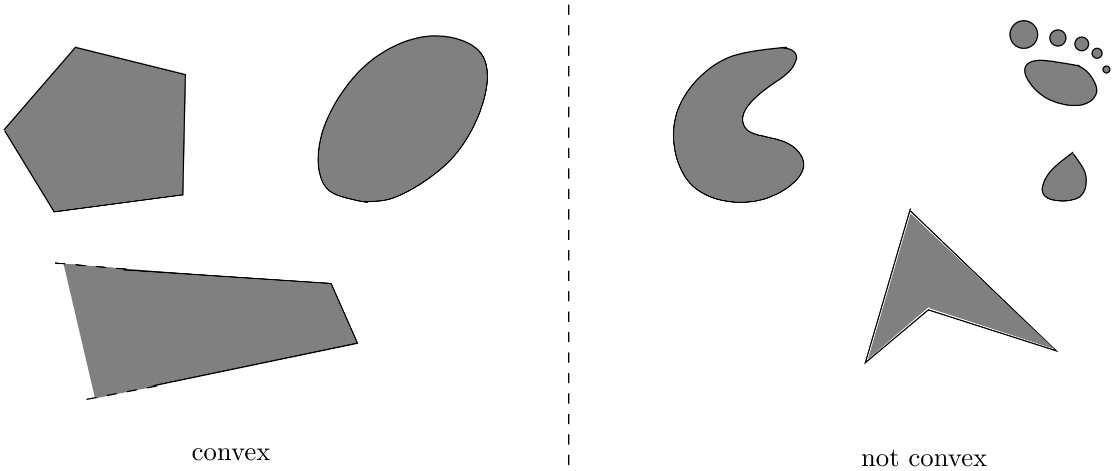


Figure 1. Convex and nonconvex sets.

Thus a set C is convex if for every pair of points x and y in C, the line segment joining x and y is also in C.

Example 4.11.

(1) $\mathbb{R}^n$ is convex.

(2) $\emptyset$ is convex. (Why?)

(3) $B(a,r) = \{x \in \mathbb{R}^n : \|x-a\| \le r\}$ is convex.

(4) $S(a,r) = \{x \in \mathbb{R}^n : \|x-a\| = r\}$ is not convex.

(5) $\mathcal{F} = \{x \in \mathbb{R}^n : Ax = b \text{ and } x \ge 0\}$ is convex.

Exercise 4.12. Let $(C_i)_{i\in I}$ be a family of convex sets in $\mathbb{R}^n$. Prove that their intersection $C = \bigcap_{i\in I} C_i$ is convex as well.

Exercise 4.13. Let $C \subset \mathbb{R}^n$ be a convex set. Show that for all $n \in \mathbb{N}$ and all $x_1, \dots, x_n \in C$, there holds that $\dfrac{x_1 + \cdots + x_n}{n} \in C$. *Hint:* Use induction.

Definition 4.14. Let C be a convex set. A point $x \in C$ is called an *extreme point* of C if there are no two distinct points $y, z \in C$ such that $x = (1-t)y + tz$ for some $t \in (0,1)$.

Example 4.15.

(1) The convex set $\mathbb{R}^n$ has no extreme points.

(2) The convex set $\emptyset$ has no extreme points.

(3) The set of extreme points of the convex set $B(a,r)$ is $S(a,r)$.

We will now see that the set of extreme points of the convex set $\mathcal{F} = \{x \in \mathbb{R}^n : Ax = b \text{ and } x \ge 0\}$ is precisely the set of basic feasible solutions.

4.4.2. Basic feasible solutions=extreme points of $\mathcal{F}$.

Theorem 4.16. *Let $\mathcal{F} = \{x \in \mathbb{R}^n : Ax = b$ and $x \ge 0\}$ and let $x \in \mathcal{F}$. Then x is an extreme point of $\mathcal{F}$ iff x is a basic feasible solution of (P).*

Proof. (If) Let x be a basic feasible solution corresponding to the basic index tuple $(\beta_1, \dots, \beta_m)$. Then $x_{\beta_1} a_{\beta_1} + \cdots + x_{\beta_m} a_{\beta_m} = b$, where $a_{\beta_1}, \dots, a_{\beta_m}$ are linearly independent. Suppose that x can be expressed as a convex combination of points $y, z \in \mathcal{F}$, that is, $x = \alpha y + (1-\alpha) z$ for some $\alpha \in (0,1)$. Since all the components of x, y, z are nonnegative, and since $\alpha \in (0,1)$, it follows that the components of y and z corresponding to indices not in the basic index tuple must all be zero. Since we know that $y, z \in \mathcal{F}$, we can conclude that $y_{\beta_1} a_{\beta_1} + \cdots + y_{\beta_m} a_{\beta_m} = b$ and $z_{\beta_1} a_{\beta_1} + \cdots + z_{\beta_m} a_{\beta_m} = b$. But by the linear independence of $a_{\beta_1}, \dots, a_{\beta_m}$, it follows that $y = z$ $(= x)$. So x is an extreme point of $\mathcal{F}$.

(Only if) Let x be an extreme point of $\mathcal{F}$, having k positive components corresponding to the index tuple $(\gamma_1, \dots, \gamma_k)$ and the rest of the components are 0. With the usual notation, we have $x_{\gamma_1} a_{\gamma_1} + \cdots + x_{\gamma_k} a_{\gamma_k} = b$. We will show that x is a basic feasible solution, by showing that the vectors $a_{\gamma_1}, \dots, a_{\gamma_k}$ are linearly independent. We will do this by contradiction. Suppose that there are scalars $y_{\gamma_1}, \dots, y_{\gamma_k}$, not all zeros, such that $y_{\gamma_1} a_{\gamma_1} + \cdots + y_{\gamma_k} a_{\gamma_k} = 0$. Set y to be be the vector in $\mathbb{R}^n$ whose $j = \gamma_i$th entry is y_{γ_i} and for j not any of the γ_is, the entry is 0. Since $x_{\gamma_1}, \dots, x_{\gamma_k}$ are all positive, we can choose[4] a positive ϵ such that $x + \epsilon y \ge 0$ as well as $x - \epsilon y \ge 0$. Then the vectors $x + \epsilon y$ and $x - \epsilon y$ belong to $\mathcal{F}$ (why?) and they are distinct (why?). Since $x = \frac{1}{2}(x+\epsilon y) + \frac{1}{2}(x - \epsilon y)$, we arrive at the conclusion that x is *not* an extreme point of $\mathcal{F}$, a contradiction. So $a_{\gamma_1}, \dots, a_{\gamma_k}$ are linearly independent, and hence x is a basic feasible solution. $\Box$

This theorem sheds some light on the nature of the convex set $\mathcal{F}$.

[4] For example, $\epsilon = \min \left\{ \left| \dfrac{x_i}{y_i} \right| : y_i \ne 0 \right\}$ works.

Corollary 4.17. *$\mathcal{F}$ has only finitely many extreme points.*

Proof. There are finitely many basic feasible solutions to (P). □

Corollary 4.18. *If the convex set $\mathcal{F}$ is nonempty, and it has at least one extreme point.*

Proof. The set $\mathcal{F}$ being nonempty simply means that there is a feasible solution. But then by the fundamental theorem of linear programming, we know that there must be a basic feasible solution. By the theorem above, we know that this basic feasible solution is an extreme point of $\mathcal{F}$, and so we obtain the desired conclusion. □

Corollary 4.19. *If there is an optimal solution to (P), then there is an optimal solution to (P) which is an extreme point of $\mathcal{F}$.*

Proof. Again this follows from the fundamental theorem of linear programming and the theorem above. □

We now consider some examples.

Example 4.20. Let $\mathcal{F} = \{x \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 1,\ x_1, x_2, x_3 \geq 0\}$. Thus $n = 3$, $m = 1$, $A = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix}$ and $b = [1]$. The three basic feasible solutions are

$$\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

corresponding to $\beta = 1$, $\beta = 2$ and $\beta = 3$, respectively. These points are the extreme points of the triangle (the convex set $\mathcal{F}$) shown in Figure 2. ◊

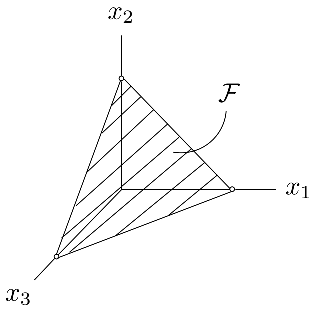


Figure 2. $\mathcal{F}$ and its extreme points.

Example 4.21. Let

$$\mathcal{F} = \{x \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 1,\ 2x_1 + 3x_2 = 1,\ x_1, x_2, x_3 \geq 0\}.$$

Thus $n = 3$, $m = 2$,

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 0 \end{bmatrix} \quad \text{and} \quad b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

The three basic solutions are

$$\begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}, \quad \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{2} \end{bmatrix}, \quad \begin{bmatrix} 0 \\ \frac{1}{3} \\ \frac{2}{3} \end{bmatrix},$$

corresponding to $\beta = (1, 2)$, $\beta = (1, 3)$ and $\beta = (2, 3)$, respectively. Of these, the first one is not feasible. So we have two basic feasible solutions. And these are the extreme points of the line segment (the convex set $\mathcal{F}$) shown in Figure 3. ◊

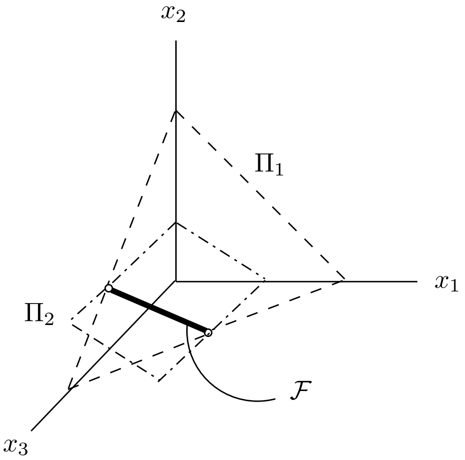


Figure 3. $\mathcal{F}$ and its extreme points. $\mathcal{F}$ is the intersection of the hyperplanes Π_1 (given by $x_1 + x_2 + x_3 = 1$) and Π_2 ($2x_1 + 3x_2 = 1$) and the half spaces $x_1 \geq 0$, $x_2 \geq 0$, $x_3 \geq 0$.

Example 4.22. Consider Example 4.3 again, where

$$A = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 2 & 1 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{2\times 4} \quad \text{and} \quad b = \begin{bmatrix} 200 \\ 300 \end{bmatrix} \in \mathbb{R}^2.$$

In this case

$$\mathcal{F} = \left\{ x \in \mathbb{R}^4 : \begin{array}{r} x_1 + x_2 + x_3 = 200, \\ 2x_1 + x_2 + x_4 = 300, \\ x_1, x_2, x_3, x_4 \geq 0 \end{array} \right\}.$$

There are $\binom{4}{2} = 6$ possible basic solutions, and they are

$$\begin{bmatrix} 0 \\ 300 \\ -100 \\ 0 \end{bmatrix}, \begin{bmatrix} 100 \\ 100 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 200 \\ 0 \\ 100 \end{bmatrix}, \begin{bmatrix} 150 \\ 0 \\ 50 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 200 \\ 300 \end{bmatrix}, \begin{bmatrix} 200 \\ 0 \\ 0 \\ -100 \end{bmatrix},$$

corresponding to

$$\beta = (2,3),\ \beta = (1,2),\ \beta = (2,4),\ \beta = (1,3),\ \beta = (3,4),\ \beta = (1,4),$$

respectively. Of these, there are four feasible basic solutions (all of the above except the first one and the last one). Moreover, these basic feasible solutions are all non-degenerate. We cannot of course draw a picture in $\mathbb{R}^4$, but the projection of $\mathcal{F}$ in the (x_1, x_2)-plane is shown in Figure 4. ♢

Exercise 4.23. Let C be a convex set in $\mathbb{R}^n$ and C' be a convex set in $\mathbb{R}^m$. Suppose that $T \in \mathbb{R}^{m\times n}$ is such that the map $x \mapsto Tx : \mathbb{R}^n \to \mathbb{R}^m$ establishes a one-to-one correspondence between C and C' (that is, for every $c \in C$, first of all $Tc \in C'$, and moreover, for every $c' \in C'$, there is a unique $c \in C$ such that $Tc = c'$). Show that T also establishes a one-to-one correspondence between the extreme points of C and C'.

Exercise 4.24. Consider the two linear programming problems (LP) and (LP') in (3.2) and (3.3), respectively, considered in Section 3.1. Show that there is a one-to-one correspondence between the extreme points of the feasible sets of these two problems.

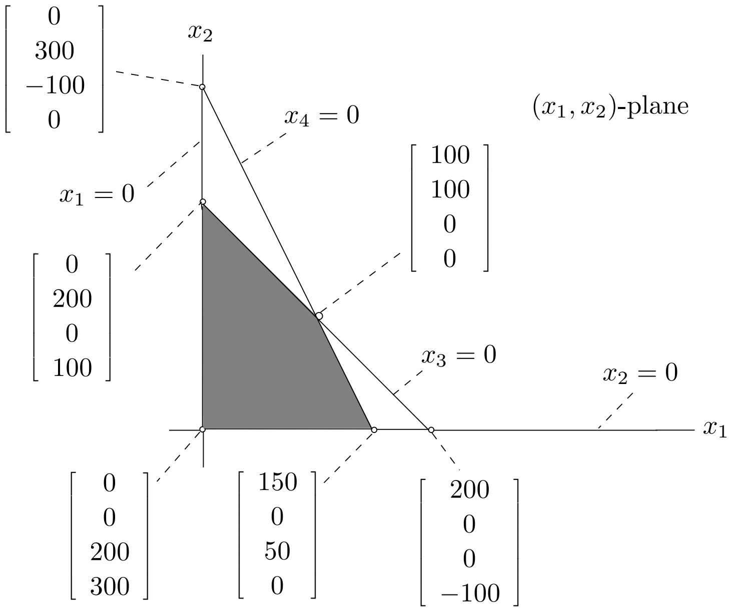

$\begin{bmatrix} 0 \\ 300 \\ -100 \\ 0 \end{bmatrix}$
x_2
(x_1, x_2)-plane
$x_4 = 0$
$x_1 = 0$
$\begin{bmatrix} 100 \\ 100 \\ 0 \\ 0 \end{bmatrix}$
$\begin{bmatrix} 0 \\ 200 \\ 0 \\ 100 \end{bmatrix}$
$x_3 = 0$
$x_2 = 0$
x_1
$\begin{bmatrix} 0 \\ 0 \\ 200 \\ 300 \end{bmatrix}$
$\begin{bmatrix} 150 \\ 0 \\ 50 \\ 0 \end{bmatrix}$
$\begin{bmatrix} 200 \\ 0 \\ 0 \\ -100 \end{bmatrix}$

Figure 4. Projection of $\mathcal{F}$ and of its four extreme points in the (x_1, x_2)-plane. We have also shown the other two (non-feasible) basic solutions.

Chapter 5

The simplex method

The idea behind the simplex method is to proceed from one basic feasible solution to the next (that is one extreme point of the constraint set $\mathcal{F}$ to a new one) in such a way so as to continually decrease the value of the objective function, until a minimum is reached.

The results of the previous chapter guarantee that it is enough to consider only basic feasible solutions in our search for an optimal feasible solution. The main point in the simplex method is that it is an *efficient* way of searching among the basic feasible solutions.

Why the name "simplex method"? The word "simplex" is used to describe a convex polytope. Since we are moving between corners of a convex polytope $\mathcal{F}$, the name makes sense.

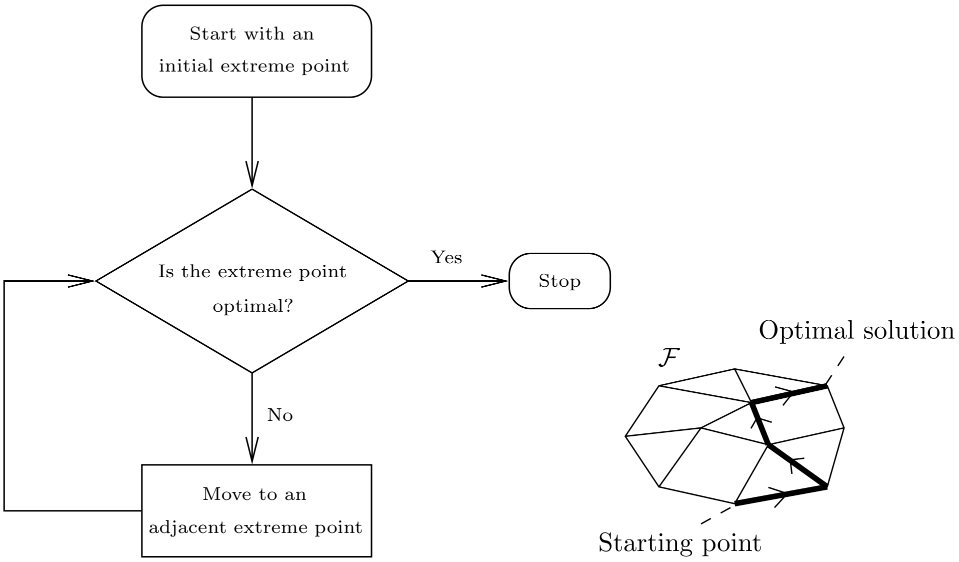


Figure 1. The simplex algorithm begins at a starting extreme point and moves along edges of the polytope until it reaches the extreme point which is the optimum solution.

5.1. Preliminaries

Before we learn the simplex method in the subsequent sections, in this section we will make a few observations that will lead to the simplex algorithm. Suppose that we have chosen a basic index tuple β. Then associated with this β, we introduce the following notation.

Notation	element of	definition
$\overline{b}$	$\mathbb{R}^m$	$A_\beta \overline{b} = b$
$\overline{a}_j$ $(j = 1, \dots, n)$	$\mathbb{R}^m$	$A_\beta \overline{a}_j = a_j$
y	$\mathbb{R}^m$	$A_\beta^\top y = c_\beta$
r	$\mathbb{R}^n$	$r = c - A^\top y$
$\overline{z}$	$\mathbb{R}$	$\overline{z} = c_\beta^\top \overline{b} = y^\top A_\beta \overline{b} = y^\top b$

In particular, $r_\beta^\top = c_\beta^\top - y^\top A_\beta$ and $r_\nu^\top = c_\nu^\top - y^\top A_\nu$. We also introduce the variable z by $z = c^\top x$. With the help of the above notation, we can express z as a function of the non-basic variable x_ν under the constraint $Ax = b$ as follows:

$$\begin{aligned} z &= c^\top x \\ &= c_\beta^\top x_\beta + c_\nu^\top x_\nu \\ &= y^\top A_\beta x_\beta + c_\nu^\top x_\nu \\ &= y^\top (b - A_\nu x_\nu) + c_\nu^\top x_\nu \\ &= y^\top b + (c_\nu^\top - y^\top A_\nu) x_\nu \\ &= \overline{z} + r_\nu^\top x_\nu. \end{aligned}$$

Thus we obtain $z = c^\top x = \overline{z} + r_\nu^\top x_\nu = \overline{z} + \sum_{i=1}^{\ell} r_{\nu_i} x_{\nu_i}$.

The components r_{ν_i} of the vector r_ν are called the *reduced costs* for the nonbasic variables. In the basic solution $x_\nu = 0$, and so $x_\beta = \overline{b}$ and $z = \overline{z}$.

Theorem 5.1. *Suppose that $\overline{b} \geq 0$ and $r_\nu \geq 0$. Then the basic feasible solution x with $x_\beta = \overline{b}$ and $x_\nu = 0$ is an optimal solution to the linear programming problem (P).*

Proof. The linear programming problem (P) can be rewritten as

$$\begin{array}{rl} \text{minimize} & \overline{z} + r_\nu^\top x_\nu \\ \text{subject to} & A_\beta x_\beta + A_\nu x_\nu = b \\ \text{and} & x_\beta \geq 0 \text{ and } x_\nu \geq 0. \end{array}$$

Let $\widetilde{x}$ be a feasible solution to the problem. Then we have in particular that $\widetilde{x}_\nu \geq 0$. Together with the assumption that $r_\nu \geq 0$, this yields that the cost corresponding to $\widetilde{x}$ is at least $\overline{z}$:

$$c^\top \widetilde{x} = \overline{z} + r_\nu^\top \widetilde{x}_\nu \geq \overline{z}.$$

But $\overline{z}$ is precisely the cost corresponding to the basic feasible solution x with $x_\beta = \overline{b}$ and $x_\nu = 0$. Hence this basic feasible solution is optimal. □

So this result tells us when to stop. If in our algorithm (for moving amongst the basic feasible solutions) we reach a basic feasible solution corresponding to a β for which $\overline{b} \geq 0$ and $r_\nu \geq 0$, we have got an optimal feasible solution!

Now we see how we actually go from a current basic feasible solution to another one if the condition $r_\nu \geq 0$ is not satisfied.

So we suppose that for our current basic feasible solution $r_{\nu_q} < 0$ for some q. In order to find a better basic feasible solution, we let x_{ν_q} be a new basic variable, and proceed as follows.

Let $x_{\nu_q} = t$, where t increases from 0. (Note that in our current basic feasible solution, the value of the variable x_ν is 0.) Meanwhile we keep the other non-basic variables still at 0. Thus $x_{\nu_i} = 0$ for all $i \neq q$. Then the cost function is simply equal to

$$z = \overline{z} + r_{\nu_q} t,$$

with the constraint $A_\beta x_\beta + t a_{\nu_q} = b$. The constraint can be rewritten as $A_\beta(x_\beta + t\overline{a}_{\nu_q} - \overline{b}) = 0$, and since A_β is invertible, it follows that $x_\beta + t\overline{a}_{\nu_q} - \overline{b} = 0$, and so

$$x_\beta = \overline{b} - t\overline{a}_{\nu_q}.$$

Then we have two possible cases:

$\boxed{1^\circ}$ $\overline{a}_{\nu_q} \leq 0$. Then t can increase unboundedly while satisfying the constraint $x_\beta \geq 0$. Thus we have found a "ray" in the set $\mathcal{F}$, and this ray is defined by

$$x_\beta(t) = \overline{b} - t\overline{a}_{\nu_q} \text{ and } x_\nu(t) = te_q,$$

where e_q is the standard basis vector for $\mathbb{R}^\ell$ with the qth component 1 and all others 0. Then for every $t \geq 0$, this gives a feasible solution $x(t)$ to the problem, with the corresponding cost

$$z(t) := \overline{z} + r_{\nu_q} t$$

with (recall!) $r_{\nu_q} < 0$. Hence if we let $t \nearrow +\infty$, then we see that the cost $z(t) \searrow -\infty$. This implies that the linear programming problem (P) has no optimal solution.

$\boxed{2^\circ}$ $\neg[\overline{a}_{\nu_q} \leq 0]$. Suppose that the vector $\overline{a}_{\nu_q}$ has at least one component that is positive. Let

$$\overline{b} = \begin{bmatrix} \overline{b}_1 \\ \vdots \\ \overline{b}_m \end{bmatrix} \quad \text{and} \quad \overline{a}_{\nu_q} = \begin{bmatrix} \overline{a}_{1,\nu_q} \\ \vdots \\ \overline{a}_{m,\nu_q} \end{bmatrix}.$$

Then for each i with $\overline{a}_{i,\nu_q} > 0$, the t can at most be $\frac{\overline{b}_i}{\overline{a}_{i,\nu_q}}$ for feasibility (so that $x_\beta \geq 0$). Indeed, if $t > \frac{\overline{b}_i}{\overline{a}_{i,\nu_q}}$, then

$$x_{\beta_i} = \overline{b}_i - t\overline{a}_{i,\nu_q} < 0,$$

which renders the $x(t)$ to be not feasible. Hence the maximum that t can increase is given by

$$t_{\max} = \min\left\{ \frac{\overline{b}_i}{\overline{a}_{i,\nu_q}} \;:\; \overline{a}_{i,\nu_q} > 0 \right\}.$$

(This is because if $t > t_{\max}$, then at least one basic variable takes a negative value, making it not feasible.) Let $p \in \{1, \dots, m\}$ be an index for which

$$t_{\max} = \frac{\overline{b}_p}{\overline{a}_{p,\nu_q}} \text{ with } \overline{a}_{p,\nu_q} > 0.$$

So we have that when t (that is, the variable x_{ν_q}) has increased from 0 to $t_{\max}$, x_{β_p} has become 0. Thus we have found a new feasible basic solution, where x_{ν_q} has become a new basic variable, while x_{β_p} has become a new non-basic variable (with value 0). We update the basic index tuple β to the new basic index tuple obtained by replacing β_p in β by ν_q.

If $t_{\max} > 0$, then the new basic feasible solution gives rise to a cost which is strictly smaller than the previous basic feasible solution: indeed this is because $r_{\nu_q} < 0$ and so

$$\overline{z} + r_{\nu_q} t_{\max} < \overline{z}.$$

Having $t_{\max} > 0$ is guaranteed for example when the previous basic feasible solution is nondegenerate, since then all $\overline{b}_i$s are positive.

There is one point that we have not yet checked, namely if the columns of A corresponding to the updated β are linearly independent, that is, if

$$a_{\beta_1}, \dots, a_{\beta_{p-1}}, a_{\nu_q}, a_{\beta_{p+1}}, \dots, a_{\beta_m}$$

are linearly independent in $\mathbb{R}^m$. We prove this below. The crucial observation is that $\overline{a}_{p,\nu_q} > 0$.

We have $A_\beta \overline{a}_{\nu_q} = a_{\nu_q}$, and so we can express a_{ν_q} as a linear combination of $a_{\beta_1}, \dots, a_{\beta_m}$:

$$a_{\nu_q} = \sum_{i=1}^{p-1} \overline{a}_{i,\nu_q} a_{\beta_i} + \underbrace{\overline{a}_{p,\nu_q}}_{>0} a_{\beta_p} + \sum_{i=p+1}^{m} \overline{a}_{i,\nu_q} a_{\beta_i}. \tag{5.1}$$

Suppose that there are scalars $\alpha_1, \dots, \alpha_m$ such that

$$\alpha_1 a_{\beta_1} + \cdots + \alpha_{p-1} a_{\beta_{p-1}} + \alpha_p a_{\nu_q} + \alpha_{p+1} a_{\beta_{p+1}} + \cdots + \alpha_m a_{\beta_m} = 0.$$

Then using (5.1), we obtain

$$\sum_{i=1}^{p-1} (\alpha_i + \alpha_p \overline{a}_{i,\nu_q}) a_{\beta_i} + (\alpha_p \overline{a}_{p,\nu_q}) a_{\beta_p} + \sum_{i=p+1}^{m} (\alpha_i + \alpha_p \overline{a}_{i,\nu_q}) a_{\beta_i} = 0.$$

By the linear independence of $a_{\beta_1}, \dots, a_{\beta_m}$, we obtain that for all $i \neq p$, $\alpha_i + \alpha_p \overline{a}_{i,\nu_q} = 0$ and $\alpha_p \overline{a}_{p,\nu_q} = 0$. Since $\overline{a}_{p,\nu_q} > 0$, this last equality gives $\alpha_p = 0$, and then we obtain from the other equalities that the α_is are zero also when $i \neq p$. So we have obtained that $\alpha_1 = \cdots = \alpha_m = 0$, proving the desired independence.

5.2. The simplex algorithm

We consolidate the observations made in the previous section to obtain the simplex method for solving the linear programming problem (P).

Here is the *simplex method*:

(1) Given is a partition of the variables, represented via the index tuples β and ν, corresponding to a basic feasible solution x. Calculate the vectors $\overline{b}, y, r_\nu$:

$$A_\beta \overline{b} = b, \quad A_\beta^\top y = c_\beta, \quad r_\nu = c_\nu - A_\nu^\top y.$$

(Since x is a basic feasible solution, $\overline{b} \geq 0$.)

(2) $\underline{1}$° If $r_\nu \geq 0$, then the algorithm **terminates**, and the basic feasible solution defined via $x_\beta = \overline{b}$ and $x_\nu = 0$ is an optimal solution to the linear programming problem (P).

$\underline{2}$° If $\neg[r_\nu \geq 0]$, then choose a q such that r_{ν_q} is the most negative component of r_ν, and calculate the vector $\overline{a}_{\nu_q}$: $A_\beta \overline{a}_{\nu_q} = a_{\nu_q}$.

(3) $\underline{1}$° If $\overline{a}_{\nu_q} \leq 0$, then the algorithm **terminates**, and the problem has no optimal solution.

$\underline{2}$° If $\neg[\overline{a}_{\nu_q} \leq 0]$, then calculate $t_{\max}$ and determine a p:

$$t_{\max} = \min \left\{ \frac{\overline{b}_i}{\overline{a}_{i,\nu_q}} \ : \ \overline{a}_{i,\nu_q} > 0 \right\},$$

and $p \in \{1, \dots, m\}$ is an index for which

$$\overline{a}_{p,\nu_q} > 0 \text{ and } t_{\max} = \frac{\overline{b}_p}{\overline{a}_{p,\nu_q}}.$$

Update the index vectors β and ν by interchanging ν_q and β_p, and go to Step (1).

The algorithm is also illustrated in the form of a flow chart shown in Figure 2.

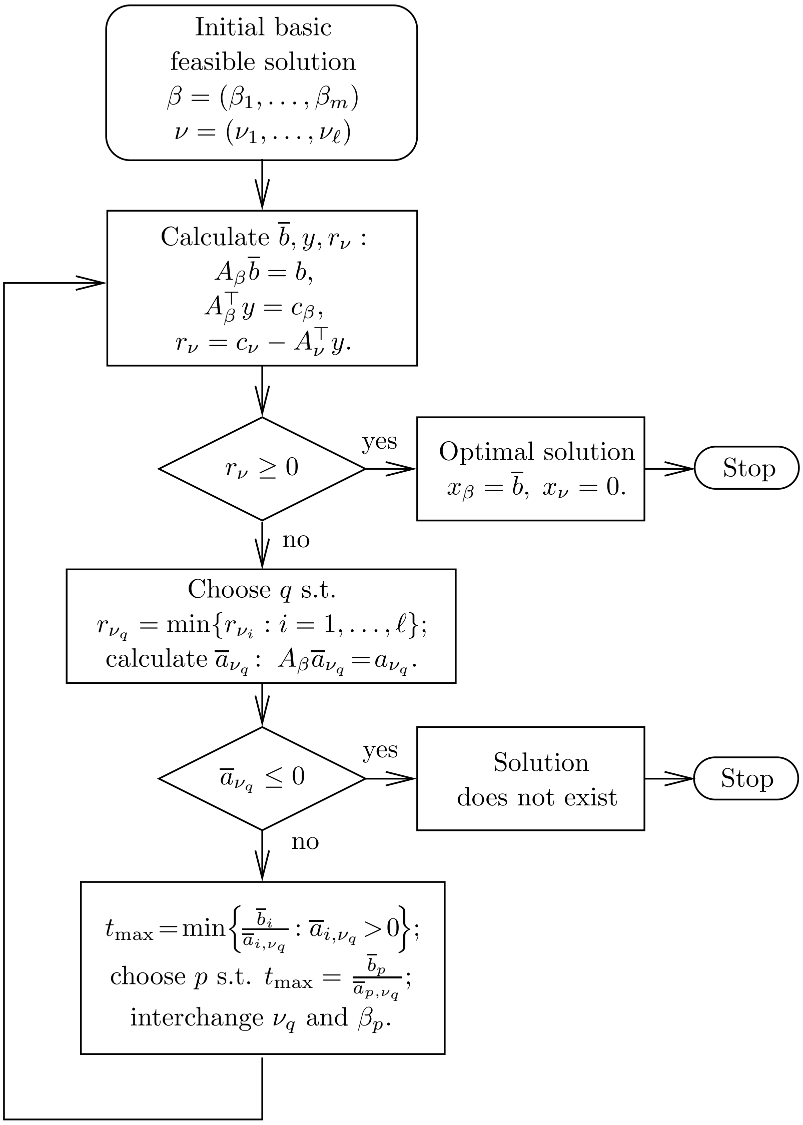

Initial basic
feasible solution
$\beta = (\beta_1, \dots, \beta_m)$
$\nu = (\nu_1, \dots, \nu_\ell)$
Calculate $\overline{b}, y, r_\nu$:
$A_\beta \overline{b} = b,$
$A_\beta^\top y = c_\beta,$
$r_\nu = c_\nu - A_\nu^\top y.$
$r_\nu \geq 0$
yes
Optimal solution
$x_\beta = \overline{b},\ x_\nu = 0.$
Stop
no
Choose q s.t.
$r_{\nu_q} = \min\{r_{\nu_i} : i = 1, \dots, \ell\};$
calculate $\overline{a}_{\nu_q}$: $A_\beta \overline{a}_{\nu_q} = a_{\nu_q}.$
$\overline{a}_{\nu_q} \leq 0$
yes
Solution
does not exist
Stop
no
$t_{\max} = \min\left\{\frac{\overline{b}_i}{\overline{a}_{i,\nu_q}} : \overline{a}_{i,\nu_q} > 0\right\};$
choose p s.t. $t_{\max} = \frac{\overline{b}_p}{\overline{a}_{p,\nu_q}};$
interchange ν_q and β_p.

Figure 2. The simplex method.

When solving *large* problems with the simplex method, namely problems with perhaps thousands of constraints and even more number of variables, it is necessary to keep certain things in mind. Among others, one should bear in mind that when one is solving equations involving the basic matrix A_β, that this matrix differs from the previous basic matrix in just one column. One way to use this is that with every basic matrix change, one should "update the LU-factors" of the basic matrix. We will not get into the implementation aspects of how one goes about doing this here, but one can read about this for example in [**GNS**, §7.5.2].

Each loop in which one goes from Step (1) to Step (4) (namely from one basic feasible solution to a new one) is referred to as an *iteration*. It can be shown (although we will not prove this here) that after each iteration, the new basic feasible solution obtained is "adjacent" to the previous one. Recall that basic feasible solutions corresponded to extreme points of $\mathcal{F}$. Two extreme points x_1, x_2 are said to be *adjacent* if for each chosen point on the segment S joining them (that is,

$S = \{\alpha x_1 + (1-\alpha)x_2 \ : \ \alpha \in [0,1]\}$), this point cannot be written as a convex combination[1] of points not on S; see Figure 3.

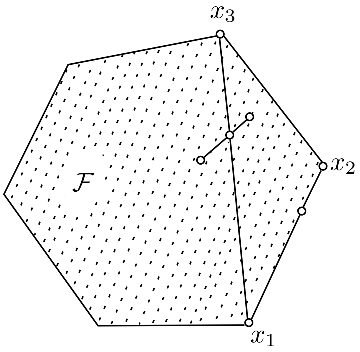


Figure 3. Amongst the extreme points x_1, x_2, x_3 of $\mathcal{F}$, x_1 and x_2 are adjacent, while x_1 and x_3 aren't.

5.3. How do we find an initial basic feasible solution?

In Step (1) of the simplex method, we have assumed that we have a basic feasible solution to begin with. And at the end of Step (3), we have shown that we obtain a new basic feasible solution, and so it is safe to go to Step (1) again. But right at the beginning, how do we *start* with a basic feasible solution?

The point is that starting with a basic solution is no trouble at all. After all, we can just choose any m independent columns of A, and form the corresponding index tuple β and so on. But we are of course not guaranteed that the x constructed in this manner is *feasible*, that is, it is also such that $x \geq 0$.

A brute force way to tackle this is to start calculating all possible (at most $\binom{n}{m}$) such basic solutions, and start with the simplex method the moment we have found a basic solution that is also feasible. But this is not efficient and so we need something more practical.

In this section we study a way of starting with a basic feasible solution rather than using the brute force method above. Our method will be to consider an auxiliary linear programming problem first.

We assume that in our standard form of the linear programming problem (P), each component of b is nonnegative. This can be ensured by multiplying some of the equations in $Ax = b$ by -1 if necessary.

The key observation is the following. Consider the linear programming problem

$$(P'): \left\{ \begin{array}{rl} \text{minimize} & y_1 + \cdots + y_m \\ \text{subject to} & \begin{bmatrix} A & I_m \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = b \\ \text{and} & x \geq 0, \ y \geq 0. \end{array} \right.$$

Then this problem has an obvious basic feasible solution, namely

$$\begin{bmatrix} 0 \\ b \end{bmatrix}. \tag{5.2}$$

[1] A point v is a *convex combination* of $v_1, \dots, v_k$ if $v = \lambda_k v_1 + \cdots + \lambda_k v_k$ for some scalars $\lambda_1, \dots, \lambda_k \geq 0$.

In order to find a basic feasible solution to our linear programming problem (P), we associate with (P) the artificial linear programming problem (P'). Since this associated artificial problem has the obvious basic feasible solution given by (5.2), we have no trouble starting the simplex method for (P'). It turns out that an optimal feasible solution to (P') with objective value 0 yields a staring basic feasible solution to (P)!

Theorem 5.2. *The linear programming problem (P) has a basic feasible solution iff the associated artificial linear programming problem (P') has an optimal feasible solution with objective value* 0.

Proof. (Only if) Suppose that (P) has a basic feasible solution x. Then the vector $\begin{bmatrix} x \\ 0 \end{bmatrix}$ is a basic feasible solution for (P'), and the associated cost is 0. But since the cost of the problem (P) is always nonnegative, it follows that this is in fact an optimal feasible solution for (P').

(If) Suppose that (P') has an optimal feasible solution $\begin{bmatrix} x \\ y \end{bmatrix}$ with objective value 0. But the cost of (P') associated with this solution is $y_1 + \cdots + y_m = 0$. Since $y \geq 0$, it follows that $y = 0$. But using

$$\begin{bmatrix} A & I_m \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = b$$

and the fact that $y = 0$, we obtain that $Ax = b$. Moreover, we know that $x \geq 0$. So this x is a feasible solution to (P). By the Fundamental Theorem of Linear Programming, we conclude that there must also exist a basic feasible solution to (P). □

Note that from the proof of the 'If' part of the above result, we see that as yet we do not actually have a way of constructing a basic feasible solution to (P). So how do we actually go about finding an initial basic feasible solution for (P)?

The answer is the following algorithm:

(1) We first set up the associated artificial linear programming problem (P').

(2) For (P'), we use the simplex method to find an optimal basic feasible solution[2], starting from the basic feasible solution

$$\begin{bmatrix} 0 \\ b \end{bmatrix}.$$

We then have the following two possible cases:

$\underline{1}^\circ$ There is an optimal solution for (P') with a positive objective value. Then the problem (P) has no basic feasible solution.

$\underline{2}^\circ$ There is an optimal basic feasible solution for (P') with objective value 0. This solution must have the form

$$\begin{bmatrix} x \\ 0 \end{bmatrix}.$$

If all the y_is are non-basic variables, then it follows that the basic ones are a subset of the components of x and the rest of the components of x are zero. So it follows that the x is not just a feasible solution, but in fact a basic feasible solution for (P). If some of the y_is are basic variables (degenerate case), we can first exchange these basic variables with non-basic x_i variables (which are also 0), to obtain a optimal basic feasible solution to (P') where the basic variables involve components of x only. At this stage, as in the previous paragraph, it follows that the x a basic feasible solution for (P).

[2] It can be shown that an optimal feasible solution for (P') exists, but will not prove this.

5.4. An example

Let us revisit Example 4.3, and solve it using the simplex method. Recall that we have $n = 4$, $m = 2$,

$$\begin{aligned} A &= \begin{bmatrix} 1 & 1 & 1 & 0 \\ 2 & 1 & 0 & 1 \end{bmatrix}, \\ b &= \begin{bmatrix} 200 \\ 300 \end{bmatrix}, \\ c &= \begin{bmatrix} -400 \\ -300 \\ 0 \\ 0 \end{bmatrix}. \end{aligned}$$

We start with x_3 and x_4 as the initial basic variables.

First iteration.

(1) As x_3 and x_4 are basic variables, we have $\beta = (3, 4)$ and $\nu = (1, 2)$. Thus the basic matrix

$$A_\beta = \begin{bmatrix} a_3 & a_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix},$$

while

$$A_\nu = \begin{bmatrix} a_1 & a_2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 2 & 1 \end{bmatrix}.$$

The basic variables take the value $x_\beta = \overline{b}$ at the initial basic solution, where $\overline{b}$ is determined by $A_\beta \overline{b} = b$, that is,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \overline{b} = \begin{bmatrix} 200 \\ 300 \end{bmatrix},$$

and so

$$\overline{b} = \begin{bmatrix} 200 \\ 300 \end{bmatrix}.$$

Note that this gives a *feasible* basic solution since $\overline{b} \geq 0$. This basic feasible solution is

$$\begin{bmatrix} 0 \\ 0 \\ 200 \\ 300 \end{bmatrix}.$$

We now determine the simplex multipliers (components of y) by solving $A_\beta^\top y = c_\beta$, that is,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} y = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

and so

$$y = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

The reduced costs for the non basic variables (components of r_ν) are determined by solving

$$r_\nu = c_\nu - A_\nu^\top y,$$

that is,

$$r_\nu = \begin{bmatrix} -400 \\ -300 \end{bmatrix} - \begin{bmatrix} 1 & 2 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -400 \\ -300 \end{bmatrix}.$$

(2) Since $\neg[r_\nu \geq 0]$, we must now choose q such that r_{ν_q} is the most negative component of r_ν. Since $r_{\nu_1} = r_1 = -400 < 0$ and $r_{\nu_2} = r_2 = -300 < 0$, we choose $q = 1$. (Thus x_1 becomes a new basic variable.)

We must also determine the vector $\overline{a}_{\nu_q} = \overline{a}_1$ by $A_\beta \overline{a}_1 = a_1$, that is,

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \overline{a}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

and so

$$\overline{a}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

(3) Since $\neg[\overline{a}_1 \leq 0]$, we must now determine $t_{\max}$ and p. (Recall that $t_{\max}$ is the largest the new basic variable x_1 can grow.) We have

$$t_{\max} = \min \left\{ \frac{\overline{b}_i}{\overline{a}_{i,\nu_q}} \; : \; \overline{a}_{i,\nu_q} > 0 \right\} = \min \left\{ \frac{200}{1}, \frac{300}{2} \right\} = 150,$$

while $p \in \{1, \dots, m\} = \{1, 2\}$ is an index for which

$$\overline{a}_{p,\nu_q} > 0 \text{ and } t_{\max} = \frac{\overline{b}_p}{\overline{a}_{p,\nu_q}},$$

and so we see that $p = 2$. So the basic variable $x_{\beta_p} = x_{\beta_2} = x_4$ leaves the set of basic variables. Hence we have that the new basic index tuple is $\beta = (3, 1)$ and the new non-basic index tuple is $\nu = (2, 4)$, and this stage, we have arrived at a new basic feasible solution. So we shall now begin with the second iteration.

Second iteration.

(1) Now $\beta = (3, 1)$ and $\nu = (2, 4)$. Thus the basic matrix

$$A_\beta = \begin{bmatrix} a_3 & a_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix},$$

while

$$A_\nu = \begin{bmatrix} a_2 & a_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

The basic variables take the value $x_\beta = \overline{b}$, where $\overline{b}$ is determined by $A_\beta \overline{b} = b$, that is,

$$\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} \overline{b} = \begin{bmatrix} 200 \\ 300 \end{bmatrix},$$

and so

$$\overline{b} = \begin{bmatrix} 50 \\ 150 \end{bmatrix}.$$

As expected, this gives a basic feasible solution, given by

$$\begin{bmatrix} 150 \\ 0 \\ 50 \\ 0 \end{bmatrix}.$$

We now determine the simplex multipliers (components of y) by solving $A_\beta^\top y = c_\beta$, that is,

$$\begin{bmatrix} 1 & 0 \\ 1 & 2 \end{bmatrix} y = \begin{bmatrix} 0 \\ -400 \end{bmatrix},$$

and so

$$y = \begin{bmatrix} 0 \\ -200 \end{bmatrix}.$$

The reduced costs for the non-basic variables (components of r_ν) are determined by solving

$$r_\nu = c_\nu - A_\nu^\top y,$$

that is,

$$r_\nu = \begin{bmatrix} -300 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ -200 \end{bmatrix} = \begin{bmatrix} -100 \\ 200 \end{bmatrix}.$$

(2) Since $\neg[r_\nu \geq 0]$, we must now choose q such that r_{ν_q} is the most negative component of r_ν. Since $r_{\nu_1} = r_2 = -100 < 0$, we choose $q = 1$. (Thus x_2 becomes a new basic variable.)

We must also determine the vector $\overline{a}_{\nu_q} = \overline{a}_2$ by $A_\beta \overline{a}_2 = a_2$, that is,

$$\begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} \overline{a}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix},$$

and so

$$\overline{a}_2 = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}.$$

(3) Since $\neg[\overline{a}_2 \leq 0]$, we must now determine $t_{\max}$ and p. (Now $t_{\max}$ is the largest the new basic variable x_2 can grow.) We have

$$t_{\max} = \min\left\{ \frac{\overline{b}_i}{\overline{a}_{i,\nu_q}} \; : \; \overline{a}_{i,\nu_q} > 0 \right\} = \min\left\{ \frac{50}{1/2}, \frac{150}{1/2} \right\} = 100,$$

while $p \in \{1, \dots, m\} = \{1, 2\}$ is an index for which

$$\overline{a}_{p,\nu_q} > 0 \text{ and } t_{\max} = \frac{\overline{b}_p}{\overline{a}_{p,\nu_q}},$$

and so we see that $p = 1$. So the basic variable $x_{\beta_p} = x_{\beta_1} = x_3$ leaves the set of basic variables. Hence we have that the new basic index tuple is $\beta = (2, 1)$ and the new non-basic index tuple is $\nu = (3, 4)$, and this stage, we have arrived at a new basic feasible solution. So we shall now begin with the third iteration.

Third iteration.

(1) Now $\beta = (2, 1)$ and $\nu = (3, 4)$. Thus the basic matrix

$$A_\beta = \begin{bmatrix} a_2 & a_1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix},$$

while

$$A_\nu = \begin{bmatrix} a_3 & a_4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

The basic variables take the value $x_\beta = \overline{b}$, where $\overline{b}$ is determined by $A_\beta \overline{b} = b$, that is,

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} \overline{b} = \begin{bmatrix} 200 \\ 300 \end{bmatrix},$$

and so

$$\overline{b} = \begin{bmatrix} 100 \\ 100 \end{bmatrix}.$$

This gives yet again a basic feasible solution, given by

$$\begin{bmatrix} 100 \\ 100 \\ 0 \\ 0 \end{bmatrix}.$$

We now determine the simplex multipliers (components of y) by solving $A_\beta^\top y = c_\beta$, that is,

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix} y = \begin{bmatrix} -300 \\ -400 \end{bmatrix},$$

and so

$$y = \begin{bmatrix} -200 \\ -100 \end{bmatrix}.$$

The reduced costs for the non-basic variables (components of r_ν) are determined by solving

$$r_\nu = c_\nu - A_\nu^\top y,$$

that is,

$$r_\nu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -1000 \\ -100 \end{bmatrix} = \begin{bmatrix} 1000 \\ 100 \end{bmatrix}.$$

(2) Since $[r_\nu \geq 0]$, the program terminates, and this basic feasible solution is optimal.

The optimal cost is

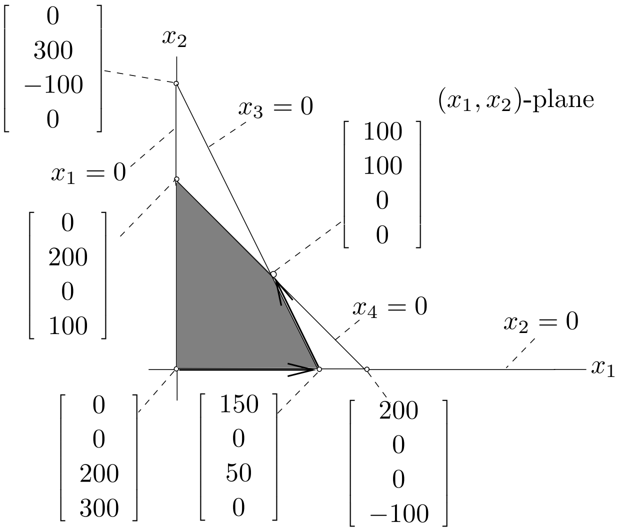
$$c^\top x = \begin{bmatrix} -400 & -300 & 0 & 0 \end{bmatrix} \begin{bmatrix} 100 \\ 100 \\ 0 \\ 0 \end{bmatrix} = -70000.$$

(Note that in Example 3.1, we had converted the original maximization problem from Section 2.2 into a minimization problem, and so for our original maximization problem, the maximum profit is 70000, as seen already in Section 2.2.)

Starting with the initial basic index tuple $\beta = (3, 4)$, the sequence of basic vectors created by the simplex method is:

$$\begin{bmatrix} 0 \\ 0 \\ 200 \\ 300 \end{bmatrix} \longrightarrow \begin{bmatrix} 150 \\ 0 \\ 50 \\ 0 \end{bmatrix} \longrightarrow \begin{bmatrix} 100 \\ 100 \\ 0 \\ 0 \end{bmatrix}$$

This is illustrated in Figure 4.



$\begin{bmatrix} 0 \\ 300 \\ -100 \\ 0 \end{bmatrix}$
x_2
$x_3 = 0$
(x_1, x_2)-plane
$\begin{bmatrix} 100 \\ 100 \\ 0 \\ 0 \end{bmatrix}$
$x_1 = 0$
$\begin{bmatrix} 0 \\ 200 \\ 0 \\ 100 \end{bmatrix}$
$x_4 = 0$
$x_2 = 0$
x_1
$\begin{bmatrix} 0 \\ 0 \\ 200 \\ 300 \end{bmatrix}$
$\begin{bmatrix} 150 \\ 0 \\ 50 \\ 0 \end{bmatrix}$
$\begin{bmatrix} 200 \\ 0 \\ 0 \\ -100 \end{bmatrix}$

Figure 4. The path through the basic feasible solutions generated by the simplex method starting with the initial basic index tuple $\beta = (3, 4)$.

5.5. Problems with the simplex method

Although the simplex method works very well, it is good to keep in mind possible problems associated with the algorithm. We discuss two main problems. One is the issue called *cyclicity* and the other is about the *computational complexity.*

5.5.1. Termination of the simplex method and cyclicity. So far we have not discussed whether the simplex method terminates in a finite number of iterations. It turns out that if there exist degenerate basic feasible solutions, then it can happen that the simplex algorithm cycles between degenerate solutions and hence never terminates. It can be shown that if all the basic feasible solutions are non-degenerate, then the simplex algorithm terminates after a finite number of iterations.

Theorem 5.3. *If all of the basic feasible solutions are non-degenerate, then the simplex algorithm terminates after a finite number of iterations.*

Proof. If a basic feasible solution is non-degenerate, then it has exactly m positive components. In this case,

$$t_{\max} = \min\left\{ \frac{\overline{b}_i}{\overline{a}_{i,\nu_q}} : \overline{a}_{i,\nu_q} > 0 \right\} > 0.$$

So the new basic feasible solution gives rise to a cost which is strictly smaller than the previous basic feasible solution. Therefore, at each iteration, the objective value decreases, and consequently a basic feasible solution that has appeared once can never reappear. But we know that there are only finitely many basic solutions, and hence finitely many basic feasible solutions. So the algorithm terminates after a finite number of iterations. □

Cycling resulting from degeneracy is not a frequent occurrence in practice. But the fact that it could happen has led the development of methods to avoid cycling. We will not study these here in this first introductory course.

5.5.2. Computation complexity of the simplex method. A natural question that the user of an algorithm asks is:

> "As the size of the input to an algorithm increases,
> how does the running time change?"

Roughly speaking, the computational complexity of an algorithm is this relationship between the amount of time or the number of steps that it takes to solve the problem as a function of the size of the input.

The simplex method is very efficient in practice. Although the total number of basic feasible solutions could be as large as $\binom{n}{m}$, it is rare that one needs to perform as many iterations. Nevertheless, there are examples of linear programming problems which require $2^n - 1$ steps in order to find the solution. Thus the worst case behaviour is bad, since it is exponential in n.

This bad worst-case scenario of the simplex method has led to the search for other more efficient polynomial time algorithms for solving linear programming problems. An example of one such method is an interior point algorithm of Karmarkar. Its main feature is that the optimal extreme points are not approached by following the edges but rather by moving within the interior of the polyhedron. However, we will not study this here.

Exercise 5.4. Consider the following linear programming problem:

$$\begin{array}{rl} \text{minimize} & -3x_1 + 4x_2 - 2x_3 + 5x_4 \\ \text{subject to} & x_1 + x_2 - x_3 - x_4 \leq 8, \\ & x_1 - x_2 + x_3 - x_4 \leq 4, \\ & x_1, x_2, x_3, x_4 \geq 0. \end{array}$$

Transform the problem to standard form using two slack variables, x_5 and x_6. Solve this problem using the simplex method. Start with the introduced slack variables as the basic variables.

Suppose that the objective coefficient corresponding to x_4 is changed from 5 to 2. Use the simplex method to solve this modified problem. Start from the final solution found in the previous part of this exercise.

Exercise 5.5. Consider the following set of constraints:

$$\begin{aligned}
&x_1 + 2x_2 + 3x_3 + 4x_4 = 10,\\
&2x_1 + 3x_2 + 4x_3 + 5x_4 = 12,\\
&x_1, x_2, x_3, x_4 \geq 0.
\end{aligned}$$

To find out systematically whether of not there exists a feasible solution, we consider the following linear programming problem (LP) with two 'artificial' variables x_5 and x_6:

$$(LP)\begin{cases}
\text{minimize} & x_5 + x_6,\\
\text{subject to} & x_1 + 2x_2 + 3x_3 + 4x_4 + x_5 = 10,\\
& 2x_1 + 3x_2 + 4x_3 + 5x_4 + x_6 = 12,\\
& x_1, x_2, x_3, x_4, x_5, x_6 \geq 0.
\end{cases}$$

Solve the problem (LP) using the simplex method, and find the optimal cost. Based on this, decide whether or not there is a vector satisfying the original set of constraints.

Exercise 5.6. Consider the following linear programming problem:

$$\begin{aligned}
\text{minimize}\quad & 4x_1 + 3x_2 + 2x_3 + 3x_4 + 4x_5,\\
\text{subject to}\quad & 4x_1 + 3x_2 + 2x_3 + x_4 = 5,\\
& x_2 + 2x_3 + 3x_4 + 4x_5 = 3,\\
& x_1, x_2, x_3, x_4, x_5 \geq 0.
\end{aligned}$$

Use the simplex method to find an optimal solution. Start with x_1 and x_5 as basic variables.

The optimal solution is not unique. Find another optimal solution than the one obtained in the previous part.

Exercise 5.7. Consider the following linear programming problem:

$$\begin{aligned}
\text{minimize}\quad & x_1 + 5x_2 + 2x_3,\\
\text{subject to}\quad & x_1 + x_2 \geq 2,\\
& x_1 + x_3 \geq 2,\\
& x_2 + x_3 \geq 2,\\
& x_1, x_2, x_3 \geq 0.
\end{aligned}$$

Use the simplex method to find an optimal solution and the optimal value. Start with the basic solution with $x_1 = x_2 = x_3 = 1$.

Exercise 5.8. Consider the following linear programming problem:

$$\begin{aligned}
\text{maximize}\quad & q^\top x,\\
\text{subject to}\quad & Px \leq b,\\
& x \geq 0,
\end{aligned}$$

where $P = \begin{bmatrix} 1 & -1 & 1 \\ 1 & 1 & -1 \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, and $q = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$.

Use the simplex method to solve this problem. Start with the slack variables as basic variables.

Find two vectors $x_0 \in \mathbb{R}^3$ and $d \in \mathbb{R}^3$ such that with $x(t) := x_0 + td$, $t \in \mathbb{R}$, there holds that:

(1) $x(t)$ is a feasible solution for each $t > 0$, and

(2) $q^\top x(t) \to +\infty$ as $t \to +\infty$.

Exercise 5.9. Consider the linear programming problem

$$(LP): \begin{cases}
\text{minimize} & c^\top x,\\
\text{subject to} & Ax = b,\\
& x \geq 0,
\end{cases}$$

where $A = \begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 \\ 2 & 2 & 2 & 1 & 0 & 2 \\ 2 & 0 & 2 & 1 & 2 & 1 \end{bmatrix}$, $b = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}$, and $c = \begin{bmatrix} 2 \\ 1 \\ 1 \\ 1 \\ 2 \\ 1 \end{bmatrix}$.

Suppose that x_1, x_3, x_5 are chosen as basic variables. Find the corresponding basic solution and verify that it is feasible. Using the simplex method determine a new (better) basic feasible solution and check that it is optimal.

Exercise 5.10. We are given the following five vectors in $\mathbb{R}^3$:

$$a_1 = \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}, \quad a_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad a_3 = \begin{bmatrix} -1 \\ 0 \\ -1 \end{bmatrix}, \quad a_4 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad b = \begin{bmatrix} 2 \\ 3 \\ 6 \end{bmatrix}.$$

We want to find out if there exist nonnegative scalars x_j such that

$$b = x_1 a_1 + x_2 a_2 + x_3 a_3 + x_4 a_4.$$

To this end, we consider the following linear programming problem in the seven variables formed by $x = (x_1, x_2, x_3, x_4)$, $y = (y_1, y_2, y_3)$:

$$(LP): \left\{ \begin{array}{ll} \text{minimize} & y_1 + y_2 + y_3, \\ \text{subject to} & Ax + Iy = b, \\ & x \geq 0 \text{ and } y \geq 0, \end{array} \right.$$

where $A = \begin{bmatrix} a_1 & a_2 & a_3 & a_4 \end{bmatrix}$.

Show that $x = (2, 5, 0, 0)$, $y = (0, 0, 1)$ is an optimal solution to (LP). Do there exist nonnegative scalars x_j such that $b = x_1 a_1 + x_2 a_2 + x_3 a_3 + x_4 a_4$?

Chapter 6

Duality theory

This chapter deals with central theoretical results for linear programming, namely the duality theorem and the complimentarity theorem.

First of all we must define the so-called *dual* problem corresponding to the given linear programming problem. The coupling between this dual problem and the original (called *primal*) problem is most significant if the primal problem has the following form:

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{n} c_j x_j \\ \text{subject to} \quad & \sum_{j=1}^{n} a_{ij} x_j \geq b_j, \quad i = 1, \dots, m, \\ & x_j \geq 0, \quad j = 1, \dots, n, \end{aligned}$$

or written in a more compact form,

$$\begin{aligned} \text{minimize} \quad & c^\top x \\ \text{subject to} \quad & Ax \geq b, \\ & x \geq 0. \end{aligned}$$

So in this chapter, we will start with this form of the linear programming problem, which is often referred to as the *canonical form.*

In many (perhaps most) applications of linear optimization, one has the constraint that $x \geq 0$, that is, that the variables must be nonnegative. Such constraints can obviously be absorbed in the constraints $Ax \geq b$ (by making A taller), but there are advantages to consider the constraint $x \geq 0$ being separate from $Ax \geq b$. Firstly, the constraint $x \geq 0$ is so simple that it would be wasteful to treat it as a part of general linear inequalities; calculations can be made less heavy if one utilizes the special structure of the inequalities $x \geq 0$. Secondly, as suggested above, the duality and complimentarity theory is more significant if $x \geq 0$ is handled separately.

Consequently, in this chapter, we will consider the linear programming problem in the form

$$(P): \begin{cases} \text{minimize} & c^\top x \\ \text{subject to} & Ax \geq b, \\ & x \geq 0, \end{cases} \tag{6.1}$$

where $x \in \mathbb{R}^n$ is the variable vector, and $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$ are given. We refer to this linear programming problem as *the primal problem (P).*

The m inequalities given by $Ax \geq b$ are called *general* while the n inequalities given by $x \geq 0$ are called *simple.* The set of all $x \in \mathbb{R}^n$ which satisfy all the constraints in P is called the *feasible set of (P)*, and is denoted by $\mathcal{F}_P$:

$$\mathcal{F}_P = \{x \in \mathbb{R}^n \mid Ax \geq b \text{ and } x \geq 0\}.$$

A difference with the previous few chapters is that no we do not make any special assumptions on the matrix A. Thus any of the cases $m > n$, $m = n$ or $m < n$ are possible. Moreover, we do not assume that A has linearly independent rows or columns.

6.1. The dual linear programming problem

The following linear programming problem is called *the dual problem* to the above problem (P):

$$(D) : \begin{cases} \text{maximize} & b^\top y \\ \text{subject to} & A^\top y \leq c, \\ & y \geq 0, \end{cases} \tag{6.2}$$

where $y \in \mathbb{R}^m$ is the variable vector, and the A, b, c are the same as in (P). We refer to this linear programming problem as *the dual problem (D).* A visual mnemonic is shown below:

$$\begin{array}{c|ccccc}
 & x_1 & \dots & x_n & \geq & 0 \\
\hline
y_1 & a_{11} & \dots & a_{1n} & \geq & b_1 \\
\vdots & \vdots & & \vdots & & \vdots \\
y_m & a_{m1} & \dots & a_{mn} & \geq & b_m \\
\geq & \leq & & \leq & & \\
0 & c_1 & \dots & c_n & & \boxed{\text{min/max}}
\end{array}$$

The n inequalities given by $A^\top y \leq c$ are called *general* while the m inequalities given by $y \geq 0$ are called *simple.* The set of all $y \in \mathbb{R}^m$ which satisfy all the constraints in (D) is called the *feasible set of (D)*, and is denoted by $\mathcal{F}_D$:

$$\mathcal{F}_D = \{y \in \mathbb{R}^m \mid A^\top y \leq c \text{ and } y \geq 0\}.$$

Here are a few additional definitions:

(1) The point $x \in \mathbb{R}^n$ is called a *feasible solution to (P)* if $x \in \mathcal{F}_P$.

(2) The point $\widehat{x} \in \mathbb{R}^n$ is called an *optimal solution to (P)* if $\widehat{x} \in \mathcal{F}_P$ and for all $x \in \mathcal{F}_P$, $c^\top \widehat{x} \leq c^\top x$.

(3) The point $y \in \mathbb{R}^m$ is called a *feasible solution to (D)* if $y \in \mathcal{F}_D$.

(4) The point $\widehat{y} \in \mathbb{R}^m$ is called an *optimal solution to (D)* if $\widehat{y} \in \mathcal{F}_D$ and for all $y \in \mathcal{F}_D$, $b^\top \widehat{y} \geq b^\top y$.

6.2. The duality theorem

The following inequality is fundamental.

Proposition 6.1. *For every $x \in \mathcal{F}_P$ and every $y \in \mathcal{F}_D$, $c^\top x \geq b^\top y$.*

Proof. If $x \in \mathcal{F}_P$ and $y \in \mathcal{F}_D$, then observing that $x^\top A^\top y = y^\top Ax$, we obtain

$$\begin{array}{rcl} c^\top x - b^\top y & = & x^\top c - x^\top A^\top y + y^\top Ax - y^\top b \\ & = & x^\top (c - A^\top y) + y^\top (Ax - b) \geq 0, \end{array}$$

since $x \geq 0$, $c - A^\top y \geq 0$, $y \geq 0$ and $Ax - b \geq 0$. □

An immediate consequence of the above inequality is the following optimality condition:

Corollary 6.2. *If $\widehat{x} \in \mathcal{F}_P$, $\widehat{y} \in \mathcal{F}_D$ and $c^\top \widehat{x} = b^\top \widehat{y}$, then $\widehat{x}$ and $\widehat{y}$ are optimal for (P) and for (D), respectively.*

Proof. For every $x \in \mathcal{F}_P$ and $y \in \mathcal{F}_D$, we have

$$c^\top x \geq b^\top \widehat{y} = c^\top \widehat{x} \geq b^\top y.$$

In particular, we have obtained $c^\top x \geq c^\top \widehat{x}$ and $b^\top \widehat{y} \geq b^\top y$, giving the desired optimalities of $\widehat{x}$ and $\widehat{y}$ for (P) and (D), respectively. □

The following important theorem is proved in the appendix to this chapter in Section 6.7.

Theorem 6.3 (The duality theorem)**.**

(1) *If both $\mathcal{F}_P \neq \emptyset$ and $\mathcal{F}_D \neq \emptyset$, then there exists at least one optimal solution $\widehat{x}$ to (P) and there exists at least one optimal solution $\widehat{y}$ to (D). Moreover, $c^\top \widehat{x} = b^\top \widehat{y}$.*

(2) *If $\mathcal{F}_P \neq \emptyset$, but $\mathcal{F}_D = \emptyset$, then for every $\rho \in \mathbb{R}$, there exists an $x \in \mathcal{F}_P$ such that $c^\top x < \rho$. One then says that the optimal value of (P) is $-\infty$. In this case neither (P) nor (D) has an optimal solution.*

(3) *If $\mathcal{F}_D \neq \emptyset$, but $\mathcal{F}_P = \emptyset$, then for every $\rho \in \mathbb{R}$, there exists a $y \in \mathcal{F}_D$ such that $b^\top y > \rho$. One then says that the optimal value of (D) is $+\infty$. In this case neither (P) nor (D) has an optimal solution.*

(4) *Finally, it can happen that both $\mathcal{F}_P = \emptyset$ and $\mathcal{F}_D = \emptyset$. (That is neither (P) nor (D) have any feasible solutions.)*

As a direct consequence of this theorem, we get the converse to Corollary 6.2 above.

Corollary 6.4. *If $\widehat{x} \in \mathcal{F}_P$, $\widehat{y} \in \mathcal{F}_D$ are optimal solutions to (P) and (D), respectively then $c^\top \widehat{x} = b^\top \widehat{y}$.*

6.3. The complimentarity theorem

From Corollaries 6.2 and 6.4, it follows that $\widehat{x}$ and $\widehat{y}$ are optimal solutions to (P) and (D) iff $\widehat{x} \in \mathcal{F}_P$, $\widehat{y} \in \mathcal{F}_D$ and $c^\top \widehat{x} = b^\top \widehat{y}$. We shall now give an alternative (and often more useful) criterion for determining when two solutions to (P) and (D) are also optimal solutions to (P) and (D).

If $x \in \mathbb{R}^n$, then we set $s = Ax - b$. Then $x \in \mathcal{F}_P$ is equivalent to $x \geq 0$ and $s \geq 0$.

If $y \in \mathbb{R}^m$, then we set $r = c - A^\top y$. Then $y \in \mathcal{F}_D$ is equivalent to $y \geq 0$ and $r \geq 0$.

With the help of these notations, the *complementarity theorem* can be formulated as follows:

Theorem 6.5 (The complimentarity theorem)**.** *$x \in \mathbb{R}^n$ is an optimal solution to (P) and $y \in \mathbb{R}^m$ is an optimal solution to (D) iff*

$$\begin{aligned} &x_j \geq 0, \;\; r_j \geq 0, \;\; x_j r_j = 0 \;\; \textit{for } j = 1, \dots, n, \\ &y_i \geq 0, \;\; s_i \geq 0, \;\; y_i s_i = 0 \;\; \textit{for } i = 1, \dots, m, \end{aligned} \tag{6.3}$$

where $s := Ax - b$ and $r := c - A^\top y$.

Rephrased in somewhat briefer form: $x \in \mathbb{R}^n$ is an optimal solution to (P) and $y \in \mathbb{R}^m$ is an optimal solution to (D) iff

$$Ax \geq b, \;\; A^\top y \leq c, \;\; x \geq 0, \;\; y \geq 0, \;\; y^\top (Ax - b) = 0, \;\; x^\top (c - A^\top y) = 0.$$

Proof. First we observe that whenever $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$, one has:

$$\begin{aligned} c^\top x - b^\top y &= x^\top c - x^\top A^\top y + y^\top Ax - y^\top b \\ &= x^\top (c - A^\top y) + y^\top (Ax - b) \\ &= x^\top r + y^\top s \\ &= \sum_{j=1}^{n} x_j r_j + \sum_{i=1}^{m} y_i s_i. \end{aligned}$$

(If) Suppose that the inequalities (6.3) are satisfied. Then $x \in \mathcal{F}_P$ (since $x \geq 0$ and $s \geq 0$) and $y \in \mathcal{F}_D$ (since $y \geq 0$ and $r \geq 0$). Moreover,

$$c^\top x - b^\top y = \sum_{j=1}^{n} x_j r_j + \sum_{i=1}^{m} y_i s_i = 0,$$

where the last equality holds since all $x_j r_i = 0$ and all $y_i s_i = 0$. Thus $c^\top x = b^\top y$, and so by Corollary 6.2, x and y are optimal solutions to (P) and (D), respectively.

(Only if) Now suppose that x and y are optimal solutions to (P) and (D), respectively. Since x is a feasible solution to (P), we must have $x \geq 0$ and $s \geq 0$, and similarly, since y is a feasible solution to (D), we must have $y \geq 0$ and $r \geq 0$. Moreover, by Corollary 6.4, we have $c^\top x = b^\top y$, which gives:

$$0 = c^\top x - b^\top y = \sum_{j=1}^{n} x_j r_j + \sum_{i=1}^{m} y_i s_i.$$

But note that each term in the sum above is nonnegative (indeed, $x_j \geq 0$, $r_j \geq 0$ for all j and $y_i \geq 0$, $s_i \geq 0$ for all i). So the only way their sum can be zero is when each term is zero. Hence all $x_j r_j = 0$ and all $y_i s_i = 0$. □

As in the last part of the proof of the 'only if' part of the above theorem, we note that if we *know* that $x, r, s, y \geq 0$, then the condition

$$x_j r_j = 0 \text{ for all } j \quad \text{and} \quad y_i s_i = 0 \text{ for all } i$$

in the theorem is equivalent with

$$x^\top r = 0 \text{ and } y^\top s = 0.$$

In words, the complimentarity theorem says that the necessary and sufficient condition for a feasible solution to (P) and a feasible solution to (D) to also be *optimal* solutions to (P) and (D), respectively, is that the following two things must hold:

(1) for each $j \in \{1, \dots, n\}$, either the jth *simple* inequality in problem (P) is an equality[1], or the jth *general* inequality in problem (D) is an equality[2],

(2) for each $i \in \{1, \dots, m\}$, either the ith *simple* inequality in problem (D) is an equality[3], or the ith *general* inequality in problem (P) is an equality[4].

[1] that is, $x_j = 0$
[2] that is, $r_j = 0$
[3] that is, $y_i = 0$
[4] that is, $s_i = 0$

6.4. Dual problem (general form)

Consider the linear programming problem in the following general form

$$(P):\left\{\begin{array}{ll} \text{minimize} & c_1^\top x_1 + c_2^\top x_2 \\ \text{subject to} & A_{11}x_1 + A_{12}x_2 \geq b_1, \\ & A_{21}x_1 + A_{22}x_2 = b_2, \\ & x_1 \geq 0,\ x_2 \text{ is free.} \end{array}\right. \tag{6.4}$$

Here

$$\begin{aligned} &c_1 \in \mathbb{R}^{n_1}, \\ &c_2 \in \mathbb{R}^{n_2}, \\ &b_1 \in \mathbb{R}^{m_1}, \\ &b_2 \in \mathbb{R}^{m_2}, \\ &A_{11} \in \mathbb{R}^{m_1 \times n_1}, \\ &A_{12} \in \mathbb{R}^{m_1 \times n_2}, \\ &A_{21} \in \mathbb{R}^{m_2 \times n_1}, \\ &A_{22} \in \mathbb{R}^{m_2 \times n_2} \end{aligned}$$

are given. The variables $x_1 \in \mathbb{R}^{n_1}$ and $x_2 \in \mathbb{R}^{n_2}$. To say that x_2 is "free" simply means that it is not constrained to be nonnegative (unlike x_1).

We will now transform the problem in the canonical form (6.1). To this end, we write the equality constraint $A_{21}x_1 + A_{22}x_2 = b_2$ as a pair of inequality constraints:

$$\begin{array}{rcl} A_{21}x_1 + A_{22}x_2 & \geq & b_2, \\ -A_{21}x_1 - A_{22}x_2 & \geq & -b_2. \end{array}$$

We write the free variable x_2 as a difference of two (constrained) nonnegative variables: $x_2 = v_2 - v_3$, where $v_2 \geq 0$ and $v_3 \geq 0$. Thus the problem now takes the following form:

$$(P):\left\{\begin{array}{lrcl} \text{minimize} & c_1^\top x_1 + c_2^\top v_2 - c_2^\top v_3 & & \\ \text{subject to} & A_{11}x_1 + A_{12}v_2 - A_{12}v_3 & \geq & b_1, \\ & A_{21}x_1 + A_{22}v_2 - A_{22}v_3 & \geq & b_2, \\ & -A_{21}x_1 - A_{22}v_2 + A_{22}v_3 & \geq & -b_2, \\ & x_1 \geq 0,\ v_2 \geq 0,\ v_3 \geq 0. & & \end{array}\right. \tag{6.5}$$

This is indeed a problem in the canonical form (6.1), with

$$A = \left[\begin{array}{rrr} A_{11} & A_{12} & -A_{12} \\ A_{21} & A_{22} & -A_{22} \\ -A_{21} & -A_{22} & A_{22} \end{array}\right],\quad b = \left[\begin{array}{r} b_1 \\ b_2 \\ -b_2 \end{array}\right],\quad c = \left[\begin{array}{r} c_1 \\ c_2 \\ -c_2 \end{array}\right],$$

and

$$x = \left[\begin{array}{c} x_1 \\ v_2 \\ v_3 \end{array}\right].$$

For writing down the dual problem, we introduce the variable vector

$$y = \left[\begin{array}{c} y_1 \\ u_2 \\ u_3 \end{array}\right],$$

and note that

$$A^\top = \left[\begin{array}{rrr} A_{11}^\top & A_{21}^\top & -A_{21}^\top \\ A_{12}^\top & A_{22}^\top & -A_{22}^\top \\ -A_{12}^\top & -A_{22}^\top & A_{22}^\top \end{array}\right].$$

Since the dual of the problem (6.1) is given by (6.2), the dual of the above problem (6.5) is the following:

$$(D):\left\{\begin{array}{lrcl} \text{maximize} & b_1^\top y_1 + b_2^\top u_2 - b_2^\top u_3 & & \\ \text{subject to} & A_{11}^\top y_1 + A_{21}^\top u_2 - A_{21}^\top u_3 & \leq & c_1, \\ & A_{12}^\top y_1 + A_{22}^\top u_2 - A_{22}^\top u_3 & \leq & c_2, \\ & -A_{12}^\top y_1 - A_{22}^\top u_2 + A_{22}^\top u_3 & \leq & -c_2, \\ & y_1 \geq 0,\ u_2 \geq 0,\ u_3 \geq 0. & & \end{array}\right. \tag{6.6}$$

Here the inequalities

$$\begin{array}{rcl} A_{12}^\top y_1 + A_{22}^\top u_2 - A_{22}^\top u_3 & \leq & c_2, \\ -A_{12}^\top y_1 - A_{22}^\top u_2 + A_{22}^\top u_3 & \leq & -c_2, \end{array}$$

can be replaced by the equality

$$A_{12}^\top y_1 + A_{22}^\top u_2 - A_{22}^\top u_3 = c_2.$$

Furthermore we replace the difference between the vectors u_2 and u_3 by the vector y_2, that is, $y_2 = u_2 - u_3$, which is not constrained. Thus we arrive at the following problem in the variable vectors $y_1 \in \mathbb{R}^{m_1}$ and $y_2 \in \mathbb{R}^{m_2}$:

$$(D):\left\{\begin{array}{ll} \text{maximize} & b_1^\top y_1 + b_2^\top y_2 \\ \text{subject to} & A_{11}^\top y_1 + A_{21}^\top y_2 \leq c_1, \\ & A_{12}^\top y_1 + A_{22}^\top y_2 = c_2, \\ & y_1 \geq 0,\ y_2 \text{ is free.} \end{array}\right. \tag{6.7}$$

This constitutes the dual problem to (6.4).

6.5. Dual problem (standard form)

Now consider the linear programming problem in the standard form

$$(P):\left\{\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & Ax = b, \\ & x \geq 0. \end{array}\right. \tag{6.8}$$

This is a special case of (6.4), with

$$A_{21} = A,\ \ c_1 = c,\ \ b_2 = b,\ \ x_1 = x,$$

while A_{11}, A_{12}, A_{22}, c_2, b_1, x_2 are "empty" (that is, they are absent). Thus the dual problem to (6.8) is the dual of (6.4) in this special case (with $A_{21} = A$ etc.), that is (with y_2 now denoted simply by y),

$$(D):\left\{\begin{array}{ll} \text{maximize} & b^\top y \\ \text{subject to} & A^\top y \leq c. \end{array}\right. \tag{6.9}$$

This is the dual problem to (6.8).

Assume that we have solved the given problem of the form (6.8) with the simplex method described earlier, and suppose that the algorithm was terminated owing to $r_\nu \geq 0$. With the notation used earlier, then there holds that

$$A_\beta \overline{b} = b,\ \ \overline{b} \geq 0,\ \ A_\beta^\top y = c_\beta, \quad \text{and} \quad A_\nu^\top y \leq c_\nu.$$

(The last inequality follows since $r_\nu = c_\nu - A_\nu^\top y \geq 0$.)

Let the vector $x \in \mathbb{R}^n$ be defined by $x_\beta = \overline{b}$ and $x_\delta = 0$. Then $x \geq 0$ and $Ax = A_\beta x_\beta + A_\nu x_\nu = A_\beta \overline{b} = b$ (so that x is the basic feasible solution to the primal problem (6.8) corresponding to basic index tuple β).

Furthermore, $A^\top y \leq c$, since $A_\beta^\top y = c_\beta$ and $A_\nu^\top y \leq c_\nu$. This means that the vector $y \in \mathbb{R}^m$ is a feasible solution to the dual problem (6.9).

Finally, we have

$$c^\top x = c_\beta^\top x_\beta + c_\nu^\top x_\nu = c_\beta^\top x_\beta = (A_\beta^\top y)^\top \overline{b} = y^\top A_\beta \overline{b} = y^\top b = b^\top y.$$

So we have the following:

(1) x is a feasible solution to the primal problem.

(2) y is a feasible solution to the dual problem.

(3) $c^\top x = b^\top y$.

Combining these observations it follows that y is an optimal solution to the dual problem (6.9)!

Thus when we solve the primal problem (6.8) with the simplex method, we have also determined an optimal solution y to the dual problem (6.9).

6.6. An economic interpretation

Consider the example from Section 2.2. Suppose that there is a rival furniture manufacturer (let us call them IKEA) who also produce tables and chairs, and whose raw material is identical to what we use, namely the big parts and small parts considered earlier. Suppose IKEA wants to expand their production and are interested in buying our resources (that is, the number of big and small parts we have got). The question IKEA asks themselves is: "What is the lowest we can pay to get the resources?"

To study this problem, we introduce the variables

$$\begin{aligned} y_1 &= \text{ the price at which IKEA offers to buy 1 big part,} \\ y_2 &= \text{ the price at which IKEA offers to buy 1 small part,} \end{aligned}$$

and let w be the total price IKEA offers us for the 200 big parts and 300 small parts we own. Thus

$$w = 200 \cdot y_1 + 300 \cdot y_2.$$

IKEA knows that in order for us to accept their offer, they should set the price high enough so that we make at least as much money by selling them our raw materials as we would with our optimal production plan. For example, we sell a table for SEK 400, and one table needs one big part and two small parts. If we sell one big part and two small parts to IKEA, we would make SEK $1 \cdot y_1 + 2 \cdot y_2$, and so it is sensible that IKEA chooses the y_1 and y_2 so that

$$1 \cdot y_1 + 2 \cdot y_2 \geq 400.$$

Similarly, if one considers a chair which we sell for SEK 300 versus the amount obtained by selling the raw materials (one big and one small part) to IKEA, we obtain the inequality

$$1 \cdot y_1 + 1 \cdot y_2 \geq 300.$$

And of course the prices y_1 and y_2 should be nonnegative. Consequently, IKEA is faced with the following optimization problem:

$$\begin{array}{rl} \text{minimize} & 200y_1 + 300y_2 \\ \text{subject to} & y_1 + 2y_2 \geq 400 \\ & y_1 + y_2 \geq 300 \\ & y_1 \geq 0,\ y_2 \geq 0. \end{array}$$

One can check that this is the dual problem to our original problem, namely,

$$\begin{array}{rl} \text{maximize} & 400 \cdot T + 300 \cdot C \\ \text{subject to} & T + C \leq 200 \\ & 2T + C \leq 300 \\ & T \geq 0,\ C \geq 0. \end{array}$$

In general, the variables in dual problem can be interpreted as fictitious prices associated with our resources (constraints) in the original primal problem. And the optimal solution we have for the original primal problem can then corresponds to an optimal solution for the dual problem, where we use the limited resources in such a manner so as to minimize the costs associated with using them. Although we have merely hinted upon this economic interpretation by means of the example above, we will not go into further detail here.

Exercise 6.6. A computer program of unknown quality has been downloaded by a user from the net to solve linear programming problems of the type

$$\begin{array}{rl} \text{minimize} & c^\top x, \\ \text{subject to} & Ax = b, \\ & x \geq 0. \end{array}$$

The user tests the program with the following data:

$$A = \begin{bmatrix} 3 & 2 & 1 & 3 & 3 & 2 \\ 2 & 4 & 2 & 1 & 2 & 1 \\ 1 & 2 & 3 & 2 & 3 & 3 \end{bmatrix}, \quad b = \begin{bmatrix} 14 \\ 16 \\ 10 \end{bmatrix}, \quad c = \begin{bmatrix} 2 \\ 3 \\ 2 \\ 2 \\ 3 \\ 2 \end{bmatrix}.$$

The program then outputs that $x = (3, 2, 1, 0, 0, 0)$ is an optimal solution and $y = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$ is the optimal solution to the dual problem.

Check that the output of the program is correct.

Exercise 6.7. Consider the formulation of the linear programming problem in Exercise 5.4 in the standard form (with 6 variables). Write down the dual linear programming problem. Visualize the feasible set to the dual problem in a figure with the dual variables y_1 and y_2 on the axes. What happens to this figure when the objective function coefficient of x_4 in the primal problem is changed from a 5 to a 2? Can you explain your answer?

Exercise 6.8. Find the dual to the problem (D) given by (6.7). What do you observe?

Exercise 6.9. Verify that the following two linear programming problems are each others duals.

$$(P): \left\{ \begin{array}{rl} \text{minimize} & x_3, \\ \text{subject to} & -x_1 + 2x_2 + x_3 \geq 0, \\ & 3x_1 - 4x_2 + x_3 \geq 0, \\ & x_1 + x_2 = 1, \\ & x_1, x_2 \geq 0, \;\; x_3 \text{ free.} \end{array} \right\}$$

$$(D): \left\{ \begin{array}{rl} \text{maximize} & y_3, \\ \text{subject to} & -y_1 + 3y_2 + y_3 \leq 0, \\ & 2y_1 - 4y_2 + y_3 \leq 0, \\ & y_1 + y_2 = 1, \\ & y_1, y_2 \geq 0, y_3 \text{ free.} \end{array} \right\}$$

The problem (P) has been solved and the optimal solution obtained is $x = (0.6, 0.4, -0.2)$. Use this information to obtain an optimal solution y to the dual problem.

Exercise 6.10. Formulate the dual problem to the linear programming problem considered in Exercise 5.6. Find an optimal solution to it using the result found in Exercise 5.6. Illustrate the feasible set and the optimal solution graphically in a figure with the y_1 and y_2 variables on the two axes.

Exercise 6.11. Formulate the dual problem to the linear programming problem considered in Exercise 5.7. Find an optimal solution to it using the result found in Exercise 5.7. Verify that the objective values for the primal and dual problems are equal.

Exercise 6.12. Formulate the dual problem to the linear programming problem considered in Exercise 5.8. Determine if it has any feasible solutions. Can you explain your answer?

Exercise 6.13. We know that the following two linear programming problems (P) and (D) are duals of each other.

$$(P): \left\{ \begin{array}{cl} \text{minimize} & c^\top x, \\ \text{subject to} & Ax \geq b, \\ & x \geq 0. \end{array} \right\}$$

$$(D): \left\{ \begin{array}{cl} \text{maximize} & b^\top y, \\ \text{subject to} & A^\top y \leq c, \\ & y \geq 0. \end{array} \right\}$$

Study the connection between the optimal values (which can possibly be $+\infty$ or $-\infty$) of the problems (P) and (D) in each of the following cases listed below. In each case,

$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}.$$

In each case, solve the problems (P) and (D) graphically, calculate their optimal values, and if there are any, their optimal solutions.

(1) $b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, c = \begin{bmatrix} -2 \\ 2 \end{bmatrix}.$

(2) $b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, c = \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$

(3) $b = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, c = \begin{bmatrix} -2 \\ 2 \end{bmatrix}.$

(4) $b = \begin{bmatrix} -1 \\ 1 \end{bmatrix}, c = \begin{bmatrix} 2 \\ 2 \end{bmatrix}.$

(5) $b = \begin{bmatrix} -1 \\ -1 \end{bmatrix}, c = \begin{bmatrix} 2 \\ -2 \end{bmatrix}.$

Exercise 6.14. Formulate the dual problem to the linear programming problem considered in Exercise 5.9. Find an optimal solution to it using the result found in Exercise 5.9. Verify that the objective values for the primal and dual problems are equal.

Exercise 6.15. Consider the problem described in Exercise 5.10. Is there a vector $z \in \mathbb{R}^3$ such that $b^\top z > 0$ and $a_j^\top z \leq 0$ for all j? If so, find such a z. *Hint:* Consider the dual programming problem to the linear programming problem set up in Exercise 5.10.

Exercise 6.16. Suppose that $A \in \mathbb{R}^{n \times n}$ is a matrix with the property $A^\top = -A$, that $c \in \mathbb{R}^{n \times 1}$, and that the following linear programming problem has a feasible solution:

$$\left\{ \begin{array}{cl} \text{minimize} & c^\top x \\ \text{subject to} & Ax \geq -c, \\ & x \geq 0. \end{array} \right.$$

Conclude that the problem has an optimal solution. What is the optimal objective function value of this problem?

6.7. Appendix

In order to prove Theorem 6.3, we will use *Farkas' lemma*, which is a result interesting in its own right.

6.7.1. Farkas' lemma.

Lemma 6.17 (Farkas' lemma). *Suppose that the* $m+1$ *vectors* $q, p_1, \dots, p_m$ *in* $\mathbb{R}^n$ *are given. Then* exactly one *of the following two systems in* $x \in \mathbb{R}^n$ *and* $y \in \mathbb{R}^m$ *have at least one solution:*

$$(\text{L}): \left\{ \begin{array}{l} q^\top x < 0, \\ p_1^\top x \geq 0, \\ \vdots \\ p_m^\top x \geq 0. \end{array} \right. \qquad (\text{R}): \left\{ \begin{array}{l} q = y_1 p_1 + \cdots + y_m p_m, \\ y_1 \geq 0, \\ \vdots \\ y_m \geq 0. \end{array} \right.$$

If we introduce the $m \times n$ matrix P with rows $p_1^\top, \dots, p_m^\top$, then the above result says that exactly one of the following two systems has at least one solution:

$$\text{(L)} : \left\{ \begin{array}{l} q^\top x < 0, \\ Px \geq 0. \end{array} \right. \qquad \text{(R)} : \left\{ \begin{array}{l} q = P^\top y, \\ y \geq 0. \end{array} \right.$$

The proof can be given using the following:

Lemma 6.18. *Let*

$$K = \left\{ v \in \mathbb{R}^n \ : \ v = \sum_{i=1}^{m} y_i p_i, \ y_i \geq 0 \textit{ for all } i \right\}.$$

For every $q \in \mathbb{R}^n$, there is a point $r \in K$ that is closest to q.

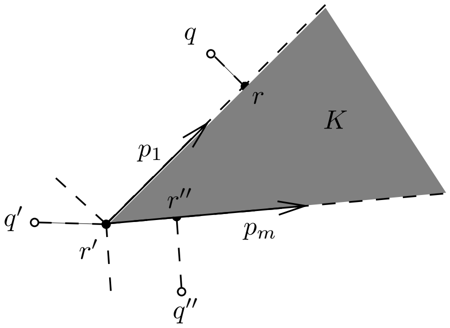


Figure 1. For $q, q', q'' \in \mathbb{R}^n$, r, r', r'', respectively, are the closest points in K.

Proof. We prove the claim using induction on m.

The claim does hold if $m = 1$, since the map $t \mapsto \|q - tp_1\|$ from $[0, +\infty)$ to $\mathbb{R}$ is continuous and as $t \to \infty$, $\|q - tp_1\| \to \infty$ (avoiding the trivial case $p_1 = 0$). Thus the function assumes a minimum on $[0, +\infty)$, say at $t_0 \in [0, +\infty)$. (Why?) Then $t_0 p_1$ belongs to K and is closest to q.

So let us make the induction hypothesis that we have proved the claim for some $m \geq 1$. We want to prove it for $m + 1$ given points. Note that $m + 1 \geq 2$. Suppose the points q and $p_1, \dots, p_{m+1}$ are given. By the induction hypothesis, for each $j = 1, \dots, m + 1$, the convex set

$$K_j = \left\{ v \in \mathbb{R}^n \ : \ v = \sum_{i=1}^{j-1} y_i p_i + \sum_{i=j+1}^{m+1} y_i p_i, \ y_i \geq 0 \text{ for all } i \right\}$$

has a closest point r_j to q. Now we consider the following three only possible cases:

$\underline{1}^\circ$ $q \in K$. Then we can simply choose $r = q$.

$\underline{2}^\circ$ $q \notin K$, but q belongs to the span of $p_1, \dots, p_{m+1}$. Let r be the point closest to q amongst $r_1, \dots, r_{m+1}$. Then this r belongs to K (since it is in one of the K_js, which in turn are all subsets of K). We will now also show that r is closest to q amongst all the points of K, that is, $\|q - r\| \leq \|q - v\|$ for all $v \in K$. Write

$$\begin{array}{rcl} q &=& \alpha_1 p_1 + \cdots + \alpha_{m+1} p_{m+1}, \\ v &=& y_1 p_1 + \cdots + y_{m+1} p_{m+1}, \end{array}$$

where $y_1, \dots, y_{m+1} \geq 0$, and $\alpha_1, \dots, \alpha_{m+1} \in \mathbb{R}$. Set

$$t = \min \left\{ \frac{y_j}{y_j - \alpha_j} : \alpha_j < 0 \right\}.$$

There must be at least one j where $\alpha_j < 0$, since q lies outside K. Then $0 \leq t < 1$ and the minimum is attained at some index i. So we have $y_j + t(\alpha_j - y_j) = t\alpha_j + (1-t)y_j \geq 0$ for all j and $t\alpha_i + (1-t)y_i = 0$. Then[5] $tq + (1-t)v \in K_i$, so that

$$\begin{aligned} \|q - r\| &\leq \|q - r_i\| \\ &\leq \|q - (tq + (1-t)v)\| = (1-t)\|q - v\| \\ &\leq \|q - v\|. \end{aligned}$$

$\underline{3}^\circ$ $q \notin K$, and q does not belong to the span of $p_1, \dots, p_{m+1}$. Choose an orthonormal basis $e_1, \dots, e_\ell$ for the span of the vectors $p_1, \dots, p_{m+1}$. Set $q' = (q, e_1)e_1 + \cdots + (q, e_\ell)e_\ell$. Then this q' does belong to the span of $p_1, \dots, p_{m+1}$, and by the previous two cases, we know that there is a point r in K closest to q'. But $(q - q', e_1) = \cdots = (q - q', e_\ell) = 0$. (Why?) Hence $q - q'$ is orthogonal to the span of $p_1, \dots, p_{m+1}$, and so

$$\|q - r\|^2 = \|q - q'\|^2 + \|q' - r\|^2 \leq \|q - q'\|^2 + \|q' - v\|^2 = \|q - v\|^2$$

for all $v \in K$. Consequently, r is also the closest point in K to q.

This completes the proof. □

Proof of Farkas' lemma; Lemma 6.17: Suppose that the right hand side system (R) *has* a solution $y \in \mathbb{R}^m$. Then for all $x \in \mathbb{R}^n$ that satisfies $p_i^\top x \geq 0$, $i = 1, \dots, m$, there holds that

$$q^\top x = \sum_{i=1}^m \underbrace{y_i}_{\geq 0} \underbrace{p_i^\top x}_{\geq 0} \geq 0,$$

which implies that the left hand system (L) has no solution.

Now suppose that the left hand system (L) has no solution. Choose r in K closest to q as in Lemma 6.18. We first show that

$$(p_j, r - q) \geq 0 \ \ (j = 1, \dots, m) \quad \text{and} \quad (r, r - q) \leq 0.$$

For if there is a i such that $(p_i, r - q) < 0$ then for a sufficiently small $t > 0$, we would have

$$\|q - (r + tp_i)\|^2 = \|q - r\|^2 + 2t(p_i, r - q) + t^2\|p_i\|^2 < \|q - r\|^2,$$

contradicting the choice of r, because $r + tp_i$ belongs to K.

Similarly, if $(r, r - q) > 0$, then for a sufficiently small $t \in (0, 1)$,

$$\|q - (r - tr)\|^2 = \|q - r\|^2 - 2t(r, r - q) + t^2\|r\|^2 < \|q - r\|^2,$$

contradicting the choice of r, because $r - tr = (1-t)r$ belongs to K.

But since (L) has no solution, by taking $x = r - q$, we must have $q^\top x \geq 0$, that is, $(q, r - q) \geq 0$. Combining this with our earlier observation that $(r, r - q) \leq 0$, we obtain

$$\|r - q\|^2 = (r - q, r - q) = (r, r - q) - (q, r - q) \leq 0.$$

Hence $q = r \in K$. In other words, $q = y_1 p_1 + \cdots + y_m p_m$ for some $y_1, \dots, y_m \geq 0$. So the system (R) has a solution. □

There are many different variants of Farkas' lemma. Lemmas 6.19 and 6.20 given below will be used to obtain the duality results we have learnt in this chapter.

[5] Geometrically, the segment joining q and v meets the side K_i of K.

Lemma 6.19. *Given $c \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m\times n}$, then exactly one of the following systems in $v \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ has at least one solution.*

$$\text{(L1)}: \begin{cases} c^\top v < 0, \\ Av \geq 0, \\ v \geq 0. \end{cases} \qquad \text{(R1)}: \begin{cases} A^\top y \leq c, \\ y \geq 0. \end{cases}$$

Proof. If in Farkas' lemma, we replace the x by v, the q by c and the matrix P with the matrix

$$\begin{bmatrix} A \\ I_n \end{bmatrix},$$

then the system (L) in Farkas' lemma is precisely the system (L1). The corresponding right hand side system (R) in Farkas' lemma becomes the following one in $y \in \mathbb{R}^m$ and $r \in \mathbb{R}^n$:

$$\begin{cases} A^\top y + I_n r = c, \\ y \geq 0, \\ r \geq 0. \end{cases}$$

But this is equivalent to the right hand side system (R1) above. Thus the claim follows from Farkas' lemma. □

Lemma 6.20. *Given $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m\times n}$, then exactly one of the following systems in $u \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$ has at least one solution.*

$$\text{(L2)}: \begin{cases} b^\top u > 0, \\ A^\top u \leq 0, \\ u \geq 0. \end{cases} \qquad \text{(R2)}: \begin{cases} Ax \geq b, \\ x \geq 0. \end{cases}$$

Proof. If in Farkas' lemma, we replace the x by u, the q by $-b$ and the matrix P with the matrix

$$\begin{bmatrix} -A^\top \\ I_m \end{bmatrix},$$

then the system (L) in Farkas' lemma is precisely the system (L2). The corresponding right hand side system (R) in Farkas' lemma becomes the following one in $x \in \mathbb{R}^n$ and $s \in \mathbb{R}^m$:

$$\begin{cases} -Ax + I_m s = -b, \\ x \geq 0, \\ s \geq 0. \end{cases}$$

But this is equivalent to the right hand side system (R2) above. Thus the claim follows from Farkas' lemma. □

6.7.2. Proof of Theorem 6.3. Before proving the duality theorem, we observe that each case in the duality theorem is actually possible.

Lemma 6.21. *Each of the following cases are possible:*

(1) $\mathcal{F}_P \neq \emptyset$ *and* $\mathcal{F}_D \neq \emptyset$.

(2) $\mathcal{F}_P \neq \emptyset$ *and* $\mathcal{F}_D = \emptyset$.

(3) $\mathcal{F}_P = \emptyset$ *and* $\mathcal{F}_D \neq \emptyset$.

(4) $\mathcal{F}_P = \emptyset$ *and* $\mathcal{F}_D = \emptyset$.

Proof. We give four examples for each of the cases. In each example, $m = 2$, $n = 2$, and

$$A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}.$$

The vectors b and c will be different, depending on which case we consider.

b	c	$\mathcal{F}_P$	$\mathcal{F}_D$
$\begin{bmatrix} 1 \\ -2 \end{bmatrix}$	$\begin{bmatrix} -1 \\ 2 \end{bmatrix}$	$\neq \emptyset;\ \widehat{x} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} \in \mathcal{F}_P$	$\neq \emptyset;\ \widehat{y} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \in \mathcal{F}_D$
$\begin{bmatrix} 1 \\ -2 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -2 \end{bmatrix}$	$\neq \emptyset;\ x = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \in \mathcal{F}_P$	$\emptyset$
$\begin{bmatrix} -1 \\ 2 \end{bmatrix}$	$\begin{bmatrix} -1 \\ 2 \end{bmatrix}$	$\emptyset$	$\neq \emptyset;\ y = \begin{bmatrix} 0 \\ 2 \end{bmatrix} \in \mathcal{F}_D$
$\begin{bmatrix} -1 \\ 2 \end{bmatrix}$	$\begin{bmatrix} 1 \\ -2 \end{bmatrix}$	$\emptyset$	$\emptyset$

(In the first case, $\widehat{x} \in \mathcal{F}_P$ and $\widehat{y} \in \mathcal{F}_D$ are optimal for P and D, respectively, since we also have $c^\top \widehat{x} = -2 = b^\top \widehat{y}$.) □

In particular, the last claim of the lemma above gives the last claim in the duality theorem.

Lemma 6.22. *If $\mathcal{F}_P \neq \emptyset$ and $\mathcal{F}_D = \emptyset$, then $\{c^\top x : x \in \mathcal{F}_P\}$ is not bounded below.*

Proof. Since $\mathcal{F}_D = \emptyset$, it follows from Lemma 6.19 that there exists a solution $v \in \mathbb{R}^n$ to the left system (L1), that is, $c^\top v < 0$, $Av \geq 0$ and $v \geq 0$. Fix an $x \in \mathcal{F}_P$, and let $x(t) = x + tv$, $t \in \mathbb{R}$. For all $t > 0$, there holds that $Ax(t) \geq b$, $x(t) \geq 0$ and $c^\top x(t) = c^\top x + tc^\top v$. So we see that $x(t) \in \mathcal{F}_P$ for all $t > 0$, and that $c^\top x(t) \to -\infty$ as $t \to +\infty$. Consequently $\{c^\top x : x \in \mathcal{F}_P\}$ is not bounded below. □

Lemma 6.23. *If $\mathcal{F}_P = \emptyset$ and $\mathcal{F}_P \neq \emptyset$, then $\{b^\top y : y \in \mathcal{F}_D\}$ is not bounded above.*

Proof. Since $\mathcal{F}_P = \emptyset$, it follows from Lemma 6.20 that there exists a solution $u \in \mathbb{R}^m$ to the left system (L2), that is, $b^\top u > 0$, $A^\top u \leq 0$ and $u \geq 0$. Fix a $y \in \mathcal{F}_D$, and let $y(t) = y + tu$, $t \in \mathbb{R}$. For all $t > 0$, there holds that $A^\top y(t) \leq c$, $y(t) \geq 0$ and $b^\top y(t) = b^\top y + tb^\top u$. So we see that $y(t) \in \mathcal{F}_D$ for all $t > 0$, and that $b^\top y(t) \to +\infty$ as $t \to +\infty$. Consequently $\{b^\top y : y \in \mathcal{F}_D\}$ is not bounded above. □

Lemma 6.24. *If $\mathcal{F}_P \neq \emptyset$ and $\mathcal{F}_D \neq \emptyset$, then there is at least one optimal solution $\widehat{x}$ to (P) and at least one optimal solution $\widehat{y}$ to (D). Moreover, $c^\top \widehat{x} = b^\top \widehat{y}$, which means that the optimal values in (P) and (D) are the same.*

Proof. The result will follow from Corollary 6.2 if we manage to show that the following system in the variables $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ has at least one solution:

$$\begin{array}{rcl} Ax & \geq & b, \\ x & \geq & 0, \\ A^\top y & \leq & c, \\ y & \geq & 0, \\ c^\top x - b^\top y & = & 0. \end{array}$$

But this is equivalent to the following system of equalities and inequalities in the variables $x \in \mathbb{R}^n$, $y \in \mathbb{R}^m$, $s \in \mathbb{R}^m$ and $r \in \mathbb{R}^n$ having at least one solution:

$$\begin{array}{rcl} -Ax + Is & = & -b, \\ A^\top y + Ir & = & c, \\ c^\top x - b^\top y & = & 0, \\ x & \geq & 0, \\ y & \geq & 0, \\ s & \geq & 0, \\ r & \geq & 0. \end{array}$$

This system can also be written compactly as $P^\top z = q$ and $z \geq 0$, where

$$P^\top = \begin{bmatrix} -A & 0 & I_m & 0 \\ 0 & A^\top & 0 & I_n \\ c^\top & -b^\top & 0 & 0 \end{bmatrix}, \quad q = \begin{bmatrix} -b \\ c \\ 0 \end{bmatrix}, \quad z = \begin{bmatrix} x \\ y \\ s \\ r \end{bmatrix}.$$

From Farkas' lemma, the system $P^\top z = q$ and $z \geq 0$ has at least one solution iff the system $q^\top w < 0$ and $Pw \geq 0$ has no solution, and in our case

$$P = \begin{bmatrix} -A^\top & 0 & c \\ 0 & A & -b \\ I_m & 0 & 0 \\ 0 & I_n & 0 \end{bmatrix}, \quad q^\top = \begin{bmatrix} -b^\top & c^\top & 0 \end{bmatrix}, \quad w = \begin{bmatrix} u \\ v \\ t \end{bmatrix}.$$

We shall show that the following equivalent system in the variables $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$ and $t \in \mathbb{R}$ does not have a solution:

$$\left.\begin{array}{rcl} A^\top u & \leq & tc, \\ Av & \geq & tb, \\ u & \geq & 0, \\ v & \geq & 0, \\ c^\top v - b^\top u & < & 0. \end{array}\right\} \qquad (6.10)$$

We will show this in two steps. First we will show that (6.10) cannot have a solution with $t > 0$. Next we will show that it cannot have a solution with $t \leq 0$.

Suppose first that $t > 0$. Then every solution u and v to the first four (of the five) constraints in (6.10) satisfies

$$\begin{array}{rcl} t(c^\top v - b^\top u) & = & v^\top (tc) - u^\top (tb) \\ & = & v^\top (tc) - v^\top A^\top u + u^\top A v - u^\top (tb) \\ & = & v^\top (tc - A^\top u) + u^\top (Av - tb) \\ & \geq & 0 + 0 = 0, \end{array}$$

that is, $c^\top v - b^\top u \geq 0$. So the final (fifth) constraint in (6.10) cannot be satisfied.

Now suppose that $t \leq 0$. Since $\mathcal{F}_P \neq \emptyset$ and $\mathcal{F}_D \neq \emptyset$, there exists a $\overline{x} \in \mathbb{R}^n$ and $\overline{y} \in \mathbb{R}^m$ such that $A\overline{x} \geq b$, $\overline{x} \geq 0$, $A^\top \overline{y} \leq c$ and $\overline{y} \geq 0$. Proposition 6.1 implies that $c^\top \overline{x} - b^\top \overline{y} \geq 0$. Now we have

$$\begin{array}{rcl} c^\top v - tb^\top \overline{y} & = & v^\top c - v^\top A^\top \overline{y} + \overline{y}^\top A v - t\overline{y}^\top b \\ & = & v^\top (c - A^\top \overline{y}) + \overline{y}^\top (Av - tb) \geq 0 + 0 = 0. \end{array}$$

Moreover,

$$\begin{array}{rcl} tc^\top \overline{x} - b^\top u & = & \overline{x}^\top tc - \overline{x}^\top A^\top u + u^\top A \overline{x} - u^\top b \\ & = & \overline{x}^\top (tc - A^\top u) + u^\top (A\overline{x} - b) \geq 0 + 0 = 0. \end{array}$$

Adding the inequalities $c^\top v - tb^\top \overline{y} \geq 0$ and $tc^\top \overline{x} - b^\top u \geq 0$ gives

$$c^\top v - b^\top u \geq \underbrace{-t}_{\geq 0} \underbrace{(c^\top \overline{x} - b^\top \overline{y})}_{\geq 0} \geq 0.$$

So once again the final (fifth) constraint in (6.10) cannot be satisfied. This completes the proof. □

Chapter 7

Network flow problems

There are a number of linear programming problems that have a special structure. One such special problem is the network flow problem. We will study these in this chapter. They are important for two reasons:

(1) They represent broad classes of problems frequently met in applications.

(2) They have an associated rich theory, which provides important insight.

By a “network flow problem”, we mean an example of linear programming, namely that of finding the minimum cost flow in a network. This gives a way of handling many types of linear programming applications. We begin with some basic concepts from graph theory.

A *network* is a pair (N, E), were N is a finite set of *nodes*, and E is a set of *directed edges* between pairs of nodes. The nodes in N are numbered from 1 to m, where m is the number of nodes. See Figure 1. A directed edge which goes from node i to node j is denoted by (i, j) and is drawn as an edge with an arrow going from i to j. See Figure 2. There can be two edges between two given nodes, namely (i, j) and (j, i), and we consider them as different directed edges. See Figure 4. Let n denote the number of elements in E.

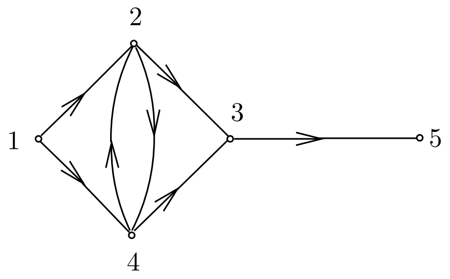


Figure 1. A network with $m = 5$ nodes and the directed edges $(1, 2), (1, 4), (2, 4), (4, 2), (2, 3), (4, 3), (3, 5)$.

Figure 2. A directed edge between nodes i and j.

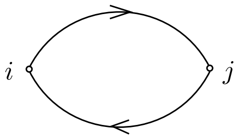


Figure 3. Two directed edges between nodes i and j.

We will assume throughout that the network is *connected*, which means that it is not the case that there is a bunch of separated networks with no connection between them. More precisely there is a path[1] between any pair of vertices x and y.

Let x_{ij} denote the *flow* in the edge $(i, j) \in E$. To every edge $(i, j) \in E$, we are given a *cost of flow*, denoted by c_{ij}. Let $x \in \mathbb{R}^n$ and $c \in \mathbb{R}^n$ be the vectors with components x_{ij} and c_{ij}, respectively, for $(i, j) \in E$, arranged in a certain order.

The node is called

(1) a *source* if the network flow is added at this node (from "outside" the network);

(2) a *sink* if the flow is absorbed at this node (to be sent outside the network);

(3) an *intermediate node* if it is neither a source nor a sink.

At each node, one has the *flow balance*, that is, inflow=outflow. Here the inflow at a node is the sum of all the flows from all the directed edges into this node, together with the flow supplied to this node from outside the network (if it happens to be a source node). On the other hand the outflow at a node is the sum of all the flows to all the directed edges out of this node, together with the flow to the outside from this node if it happens to be a sink node.

Let b_i denote the amount of flow supplied to the network from the outside at node i. For sources $b_i > 0$, while for sinks $b_i < 0$. For intermediate nodes, $b_i = 0$. We assume that

$$\sum_{i=1}^{m} b_i = 0. \tag{7.1}$$

For making the subsequent discussion concrete, we will consider an example of a network, with $m = 5$ nodes and $n = 7$ directed edges given by:

$$E = \{(1, 2), (1, 4), (2, 4), (4, 2), (2, 3), (4, 3), (3, 5)\}.$$

We have shown the network in Figure 1 above. We assume that the nodes 1 and 2 are source nodes, with flows from the outside equal to 40 and 35, respectively, while the nodes 3 and 5 are sink nodes, with flows outside being 20 and 55, respectively.

Suppose that we arrange the edges of the network in some order. For example,

$$(1, 2),\ (1, 4),\ (2, 4),\ (4, 2),\ (2, 3),\ (4, 3),\ (3, 5). \tag{7.2}$$

Corresponding to such an order of the edges, it is possible to write down the *incidence matrix* $\widetilde{A} \in \mathbb{R}^{m\times n}$ of the network, defined by

$$a_{ik} = \begin{cases} 1 & \text{if the } k\text{th edge starts at node } i, \\ -1 & \text{if the } k\text{th edge ends at node } i, \\ 0 & \text{otherwise.} \end{cases}$$

[1] a sequence of undirected edges $(x, p_1), (p_1, p_2), \dots, (p_{k-1}, p_k), (p_k, y)$

Thus with the order of edges we had chosen above for our example, the incidence matrix is

$$\widetilde{A}=\begin{array}{c} \text{nodes}\backslash\text{edges} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array}\begin{array}{c} \begin{array}{ccccccc} (1{,}2) & (1{,}4) & (2{,}4) & (4{,}2) & (2{,}3) & (4{,}3) & (3{,}5) \end{array} \\ \left[\begin{array}{ccccccc} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 1 \\ 0 & -1 & -1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{array}\right] \end{array}. \tag{7.3}$$

The *network flow problem* corresponding to this network is the following:

$$(NFP):\left\{\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & \widetilde{A}x=\widetilde{b}, \\ & x\geq 0, \end{array}\right.$$

where we use the order in (7.2) for the components of x and c

$$x=\left[\begin{array}{c} x_{12} \\ x_{13} \\ x_{24} \\ x_{42} \\ x_{23} \\ x_{43} \\ x_{35} \end{array}\right] \quad \text{and} \quad c=\left[\begin{array}{c} c_{12} \\ c_{13} \\ c_{24} \\ c_{42} \\ c_{23} \\ c_{43} \\ c_{35} \end{array}\right],$$

while the matrix $\widetilde{A}$ is given by (7.3), and $\widetilde{b}$ is given by

$$\widetilde{b}=\left[\begin{array}{c} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \end{array}\right]=\left[\begin{array}{c} 40 \\ 35 \\ -20 \\ 0 \\ -55 \end{array}\right].$$

The matrix $\widetilde{A}$ has linearly dependent rows, since the sum of all the rows is the zero row. This is always true for any incidence matrix. (Why?) So the last row of $\widetilde{A}$ is minus the sum of the other rows of $\widetilde{A}$. Also, owing to our assumption (7.1), we also have that the last component of $\widetilde{b}$ is minus the sum of the other components of $\widetilde{b}$. This means that we can remove the last equation in the system $\widetilde{A}x=\widetilde{b}$, without changing the feasible set (since the last equation is satisfied automatically whenever the others are).

So in the sequel, we will write the network flow problem in the following form:

$$(NFP):\left\{\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & Ax=b, \\ & x\geq 0, \end{array}\right.$$

where A is the $(m-1)\times n$ matrix obtained from $\widetilde{A}$ by deleting the last row, and the vector $b\in\mathbb{R}^{m-1}$ is obtained from $\widetilde{b}$ by deleting the last component, that is:

$$A=\begin{array}{c} \text{nodes}\backslash\text{edges} \\ 1 \\ 2 \\ 3 \\ 4 \end{array}\begin{array}{c} \begin{array}{ccccccc} (1{,}2) & (1{,}4) & (2{,}4) & (4{,}2) & (2{,}3) & (4{,}3) & (3{,}5) \end{array} \\ \left[\begin{array}{ccccccc} 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & 1 \\ 0 & -1 & -1 & 1 & 0 & 1 & 0 \end{array}\right] \end{array},$$

and

$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} = \begin{bmatrix} 40 \\ 35 \\ -20 \\ 0 \end{bmatrix}.$$

The matrix A has linearly independent rows, since if

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix}$$

is such that $y^\top A = 0$, then we have that

$$\begin{aligned} y_1 - y_2 &= 0, \\ y_1 - y_4 &= 0, \\ y_2 - y_4 &= 0, \\ y_2 - y_3 &= 0, \\ y_3 - y_4 &= 0, \\ y_3 &= 0. \end{aligned}$$

The last equation and the second last one yield $y_4 = 0$, which in turn with the second and third equation yields $y_1 = y_2 = 0$. Thus $y = 0$.

We shall now see how the simplex method can be simplified when applied to the network flow problem (NFP).

Since $A \in \mathbb{R}^{(m-1)\times n}$ has linearly independent rows, every basic solution has $m-1$ basic variables and $n-(m-1) = n-m+1$ non-basic variables. The corresponding basic matrix A_β has the size $(m-1) \times (m-1)$, which in our example is 4×4.

There is a nice interpretation of basic solutions to the network flow problem (NFP), based on the notion of a spanning tree. A subset T of edges of a network is called a *spanning tree* if

(1) every node of the network touches at least one edge in this subset T,

(2) the network made out of the edges from the subset T is connected, and

(3) there is no "loop" formed by the edges in T (neglecting the directions of the edges).

An example of a spanning tree for our network from Figure 1 is $T = \{(1,2),(2,4),(4,3),(3,5)\}$; see Figure 4.

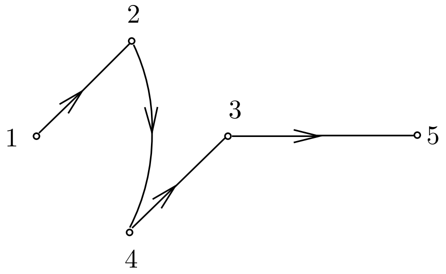


Figure 4. The tree $T = \{(1,2),(2,4),(4,3),(3,5)\}$.

The connection between spanning trees and basic variables is due to the following theorem, which we will not prove here.

Theorem 7.1. *Consider the matrix $A \in \mathbb{R}^{(m-1)\times n}$ in the network flow problem (NFP). A set of $m-1$ columns of A is linearly independent iff the corresponding $m-1$ edges form a spanning tree for the network.*

Given a spanning tree, and a corresponding basic matrix A_β, it is easy to the determine directly in the network which values the basic variables must have without solving the system $A_\beta x_\beta = b$. We illustrate the procedure by means of our example. Consider for example the spanning tree $T = \{(1,2),(2,4),(4,3),(3,5)\}$ from Figure 4. Then the corresponding basic matrix A_β and the non-basic matrix A_ν are given, respectively, by

$$\begin{array}{c} \text{nodes\textbackslash edges} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} (1,2) & (2,4) & (4,3) & (3,5) \end{array} \\ \left[\begin{array}{cccc} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \\ 0 & -1 & 1 & 0 \end{array}\right] \end{array} \text{ and } \begin{array}{c} \text{nodes\textbackslash edges} \\ 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{ccc} (1,4) & (4,2) & (2,3) \end{array} \\ \left[\begin{array}{ccc} 1 & 0 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \\ -1 & 1 & 0 \end{array}\right] \end{array}.$$

The three non-basic variables x_{14}, x_{42} and x_{23} are all 0 in the basic solution.

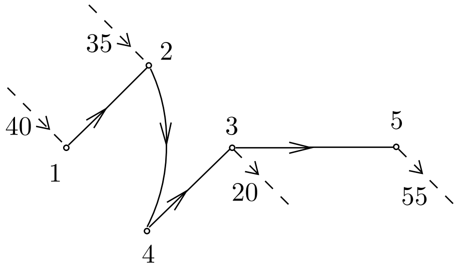


Figure 5. Flow balance in the tree T.

The values of the basic variables can be calculated in the following way (see Figure 5):

$x_{12} = 40$, since the flow balance holds at node 1.

$x_{24} = 40 + 35 = 75$, since the flow balance holds at node 2.

$x_{43} = 75$, since the flow balance holds at node 4.

$x_{35} = 55$, since the flow balance holds at node 3.

We see that the result is a basic feasible solution, since all the basic variables are nonnegative, but this is not guaranteed to happen with every spanning tree. We shall later indicate how one can systematically determine a basic feasible solution for starting the simplex method.

Assume that we have found a basic feasible solution. The next step is to find if it is optimal, by calculating the reduced costs r_{ij} for the non-basic variables. First we calculate the vector $y \in \mathbb{R}^{m-1}$ via $y^\top A_\beta = c_\beta^\top$. In our case, with $y^\top = \begin{bmatrix} y_1 & y_2 & y_3 & y_4 \end{bmatrix}$, this gives:

$$\begin{aligned} y_1 - y_2 &= c_{12}, \\ y_2 - y_4 &= c_{42}, \\ y_4 - y_3 &= c_{43}, \\ y_3 &= 0. \end{aligned}$$

If we introduce $y_5 = 0$, then these equations can be written compactly as $y_i - y_j = c_{ij}$ for all the edges $(i,j) \in T$, that is, for all the basic edges.

Every scalar y_i corresponds to a node in the network. The values of these scalars can be determined directly from the network in the following manner (see Figure 6):

First set $y_5 = 0$, by definition.

The basic edge $(3,5)$ then gives $y_3 - y_5 = c_{35}$, and so $y_3 = c_{35}$.

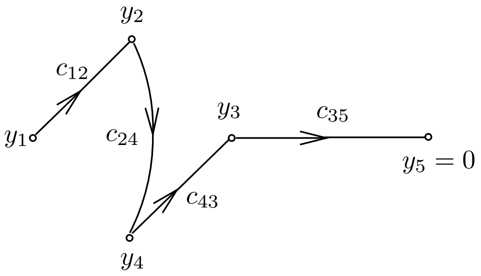


Figure 6. Flow balance in the tree T.

The basic edge $(4, 3)$ gives $y_4 - y_3 = c_{43}$, and so $y_4 = c_{43} + c_{35}$.

The basic edge $(2, 4)$ gives $y_2 - y_4 = c_{24}$, and so $y_2 = c_{24} + c_{43} + c_{35}$.

The basic edge $(1, 2)$ gives $y_1 - y_2 = c_{12}$; $y_1 = c_{12} + c_{24} + c_{43} + c_{35}$.

For example if

$$c = \begin{bmatrix} c_{12} \\ c_{14} \\ c_{24} \\ c_{42} \\ c_{23} \\ c_{43} \\ c_{35} \end{bmatrix} = \begin{bmatrix} 2 \\ 5 \\ 2 \\ 2 \\ 1 \\ 1 \\ 2 \end{bmatrix},$$

then we obtain $y_5 = 0$, $y_3 = 2$, $y_4 = 3$, $y_2 = 5$, $y_1 = 7$.

The next step is to calculate the reduced costs for the non-basic variables, that is,

$$r_\nu^\top = c_\nu^\top - y^\top A_\nu.$$

In our case, this gives:

$$\begin{aligned} r_{14} &= c_{14} - (y_1 - y_4) = 5 - (7 - 3) = 1, \\ r_{42} &= c_{42} - (y_4 - y_2) = 2 - (3 - 5) = 0, \\ r_{23} &= c_{23} - (y_2 - y_3) = 1 - (5 - 2) = -2. \end{aligned}$$

Since we had set $y_5 = 0$ earlier, the above equations can be written compactly as follows:

$$r_{ij} = c_{ij} - y_i + y_j \quad \text{for all } (i, j) \in E \setminus T,$$

that is, for all the non-basic edges.

If $r_\nu \geq 0$, then the basic feasible solution is optimal. On the other hand, if there is at least one non-basic variable for which $r_{ij} < 0$, then we let one of these non-basic variables to become a new basic variable. In our example, we have that $r_{23} = -2 < 0$ (and the other $r_{ij} \geq 0$), which means that we set $x_{23} = t$ and let the t increase from 0, while the other non-basic variables remain at 0. With this, the basic variables are functions of t.

How the basic variables change can be determined as follows (see Figure 7):

$x_{12} = 40$, since the flow balance holds at node 1.

$x_{24} = 75 - t$, since the flow balance holds at node 2.

$x_{43} = 75 - t$, since the flow balance holds at node 4.

$x_{35} = 55$, since the flow balance holds at node 3.

We see that the "new" basic edge $(2, 3)$ together with some of the other "tree edges" (that is, some of the edges corresponding to the basic variables) form a loop in the network. Every basic

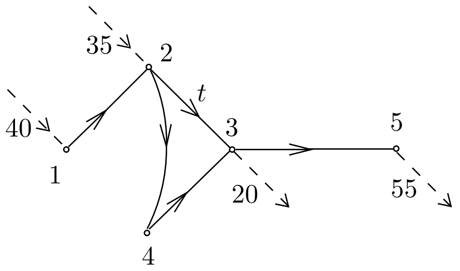


Figure 7. Flow balance with $x_{23} = t$.

variable in this loop changes either by $-t$ or by $+t$. (In our considered example, it just so happens that the basic variables all change by $-t$, but this is not always guaranteed; it can be the case that they change by $+t$ too.) The basic variables which do not lie in the loop are independent of t.

The basic variable which goes out, is the one among those that have been changed by $-t$, which first goes to 0 when t increases. In our example, t can increase till 75. Then x_{24} as well as x_{43} become 0. We can choose either to go out of the list of basic variables. So if we decide upon x_{24} to go out, then we have the new basic variables corresponding to the tree

$$T_{\text{new}} = \{(1,2),(2,3),(4,3),(3,5)\},$$

and the new (degenerate) basic feasible solution is given by

$$\begin{aligned}&x_{12} = 40, \;\; x_{23} = 75, \;\; x_{43} = 0, \;\; x_{35} = 55 \;\; \text{(basic variables);}\\ &x_{14} = 0, \;\; x_{24} = 0, \;\; x_{42} = 0 \;\; \text{(non-basic variables).}\end{aligned}$$

This completes one iteration in the simplex method, and we can now begin the second iteration, where we have to repeat the steps above. The scalars y_i can be calculated as follows:

First we set $y_5 = 0$, by definition.

The basic edge $(1,2)$ gives $y_3 - y_5 = c_{35}$, and so $y_3 = c_{35} = 2$.

The basic edge $(4,3)$ gives $y_4 - y_3 = c_{43}$, and so $y_4 = c_{43} + c_{35} = 3$.

The basic edge $(2,3)$ gives $y_2 - y_3 = c_{23}$, and so $y_2 = c_{23} + y_3 = 3$.

The basic edge $(1,2)$ gives $y_1 - y_2 = c_{12}$, and so $y_1 = c_{12} + y_2 = 5$.

The next step is to calculate the reduced costs for the non-basic variables using $r_{ij} = c_{ij} - y_i + y_j$, which gives:

$$\begin{aligned} r_{14} &= c_{14} - y_1 + y_4 = 5 - 5 + 3 = 3,\\ r_{24} &= c_{24} - y_2 + y_4 = 2 - 3 + 3 = 2,\\ r_{42} &= c_{42} - y_4 + y_2 = 2 - 3 + 3 = 2.\end{aligned}$$

Since all $r_{ij} \geq 0$, this basic feasible solution is optimal.

We note that when the simplex method is applied to the network flow problem (NFP) as above, the calculations comprise just additions and subtractions. This implies that if all the b_i and c_{ij} are whole numbers, then in the solution above, we just add or subtract whole numbers, and so no rounding off errors ever occur. Furthermore, the resulting x_{ij} are also whole numbers. Hence the optimal solution found in this manner will have integral components, although we did not explicitly demand this!

Finally, we indicate how one can determine an initial basic feasible solution in slick manner to the network flow problem (NFP). First we grow the network by introducing an *extra node* which we give number $m+1$. So in our example, now we will have 6 nodes. Next introduce m *extra edges* as follows:

(1) For each source node, introduce an edge *from* the source node *to* the extra node.

(2) For each sink node, introduce an edge *to* the sink node *from* the extra node.

(3) For each intermediate node, introduce either an edge from the intermediate node to the extra node or an edge to the intermediate node from the extra node.

In our example, the extra edges $(1,6), (2,6), (6,3), (6,5), (4,6)$ have been introduced; see Figure 8.

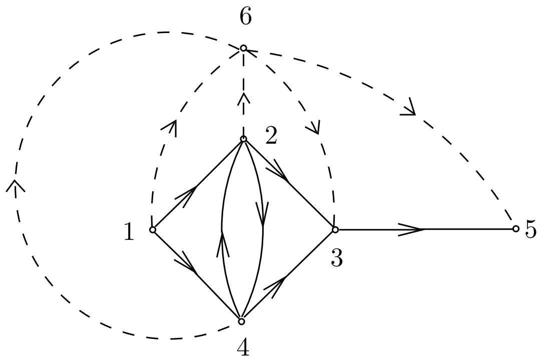


Figure 8. The network with the extra node and extra edges.

Now consider a network flow problem for this extended network, where the cost coefficients c_{ij} for the extra edges are chosen to be a "large enough" number M, so that the flows through these extra edges are very expensive, while the original edges in the network continue to have the original cost coefficients.

A basic feasible solution to this extended network flow problem is obtained by choosing the basic variables to be the ones corresponding to the extra edges (which form a spanning tree for the extended network). The basic variables values are given by $x_{i,(m+1)} = b_i \geq 0$ in the edges *to* the extra node, and $x_{m+1,i} = -b_i \geq 0$ in the edges *from* the extra node. In our example, we obtain the basic variable values to be the following: $x_{16} = 40$, $x_{26} = 35$, $x_{63} = 20$, $x_{65} = 55$, $x_{46} = 0$.

Then one can apply the simplex method (as described above) on this extended network flow problem. Since the extra edges are very expensive as compared to the original edges, the simplex method automatically sees to it that the flow to all the extra edges is 0 (if this is possible), and so the flow is transferred to the edges in the original network. The optimal solution to the extended network flow problem is then an optimal solution also to the original problem.

Exercise 7.2. Consider the linear programming problem

$$\begin{array}{ll} \text{minimize} & c^\top x \\ \text{subject to} & \widetilde{A}x = b, \\ & x \geq 0, \end{array}$$

where

$$\widetilde{A} = \left[\begin{array}{ccc|ccc|ccc} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \\ \hline -1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 \\ 0 & -1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 0 & 0 & -1 \end{array}\right], b = \begin{bmatrix} 3 \\ 5 \\ 7 \\ -2 \\ -4 \\ -9 \end{bmatrix},$$

and $c = \begin{bmatrix} 2 & 3 & 4 & 3 & 3 & 4 & 3 & 2 & 4 \end{bmatrix}^\top$.

Deduce from the special form of the matrix $\widetilde{A}$ that the problem is a network flow problem, and draw the corresponding network.

Verify that $\widehat{x} := \begin{bmatrix} 2 & 0 & 1 & 0 & 0 & 5 & 0 & 4 & 3 \end{bmatrix}^\top$ is an optimal solution to the problem.

Exercise 7.3. A given directed network has the node set N and directed edge set E given by:

$$\begin{aligned} N &= \{1,2,3,4,5,6\} \\ E &= \{(1,2),(1,3),(2,3),(2,4),(3,4),(3,5),(4,5),(4,6),(5,6)\}. \end{aligned}$$

The network has two source nodes: node 1 with supply 25 units, and node 2 with supply 10 units. The network also has two sink nodes: node 5 with a demand of 15 units, and node 6 with a demand of 20 units. The nodes 3 and 4 are intermediate nodes, with neither a supply nor a demand. The cost c_{ij} of the flow in the edge (i,j) (in units of 10^3 SEK per unit flow) is given by:

$c_{12} = 3,\ c_{13} = 2,\ c_{23} = 1,\ c_{24} = 4,\ c_{34} = 4,\ c_{35} = 4,$

$c_{45} = 1,\ c_{46} = 2,\ c_{56} = 3.$

Find a flow with minimum cost that fulfils the constraints on the supply and demand as specified above. Start with the following basic feasible solution:

$$x_{12} = 10,\ x_{13} = x_{35} = 15,\ x_{24} = 20,\ x_{46} = 20,$$

and the other x_{ij} are zeros. Find the optimal cost.

Exercise 7.4. A company has two factories F1 and F2, and three big customers C1, C2, C3. All transport from the factories to the customers go through the company's reloading terminals, T1 and T2. Since the factories, terminals and customers are spread out over the country, the transport costs between different points is different. The transportation costs from the factories to the terminals and from the terminals to the customers (in units of 100 SEK/tonne) are given in the following two tables:

	T1	T2
F1	7	6
F2	4	5

	C1	C2	C3
T1	6	7	7
T2	6	9	5

The demands of the company's product for each of the customers in a specific week is 200 tonnes, while the company's supply in the same week is 300 tonnes in each of the two factories. The head of the company's transport division has proposed the following transport plan, in unit tonnes:

	T1	T2
F1	0	300
F2	200	100

	C1	C2	C3
T1	0	200	0
T2	200	0	200

You have been hired as an optimization expert, and you need to decide whether the proposed plan is optimal from the point of view of minimizing the transportation cost. If the proposed plan is not optimal, then you have been asked to provide an optimal plan. What is your answer?

Exercise 7.5. The linear optimization problem stated below in the variables x_{ik} and z_{kj} can be interpreted as a network flow problem with I source nodes, K intermediate nodes, J sink nodes, an edge from every source node to every intermediate node (corresponding to the flows x_{ik}), and an edge from every intermediate node to every sink node (corresponding to the flows z_{kj}).

$$\begin{aligned} \text{minimize} \quad & \sum_{i=1}^{I}\sum_{k=1}^{K} p_{ik}x_{ik} + \sum_{k=1}^{K}\sum_{j=1}^{J} q_{kj}z_{kj} \\ \text{subject to} \quad & \sum_{k=1}^{K} x_{ik} = s_i \text{ for } i = 1,\dots,I, \\ & -\sum_{i=1}^{I} x_{ik} + \sum_{j=1}^{J} z_{kj} = 0 \text{ for } k = 1,\dots,K, \\ & -\sum_{k=1}^{K} z_{kj} = -d_j \text{ for } j = 1,\dots,J, \\ & x_{ik} \geq 0,\ z_{kj} \geq 0 \text{ for all } i,j,k. \end{aligned}$$

here s_i, d_j, p_{ik} and q_{kj} are positive numbers such that

$$\sum_{i=1}^{I} s_i = \sum_{j=1}^{J} d_j.$$

Assume specifically that we have the following data given:

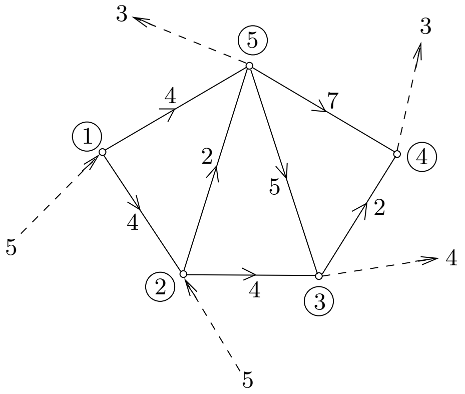
$$\begin{aligned} &I = J = K = 2, \\ &s_1 = 30,\ s_2 = 20, \\ &d_1 = 40,\ d_2 = 10, \\ &p_{11} = 5,\ p_{12} = 2,\ p_{21} = 3,\ p_{22} = 2, \\ &q_{11} = 5,\ q_{12} = 5,\ q_{21} = 7,\ q_{22} = 6. \end{aligned}$$

Show that the following solution to the problem is optimal:

$$\begin{aligned} &x_{11} = 0,\ x_{12} = 30,\ x_{21} = 20,\ x_{22} = 0, \\ &z_{11} = 20,\ z_{12} = 0,\ z_{21} = 20,\ z_{22} = 10. \end{aligned}$$

Calculate the optimal cost.

Exercise 7.6. Consider the minimum cost of flow problem for the network shown below.



We have numbered the 5 nodes. The nodes 1 and 2 are source nodes with a supply of 5 units each, while the nodes 3, 4 and 5 are sink nodes with demands of 4, 3 and 3 units, respectively. Beside each directed edge we have indicated the cost c_{ij} per unit flow.

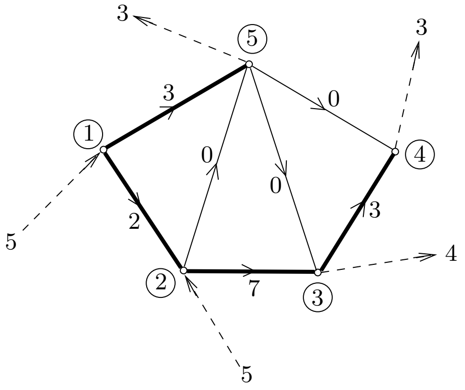
(1) Write the incidence matrix A for the network, with the following order of the edges:

$$(1, 2),\ (1, 5),\ (2, 3),\ (2, 5),\ (3, 4),\ (5, 3),\ (5, 4).$$

(Above, the notation (i, j) means the directed edge from node i to node j.)

Let x_{ij} denote the flow from node i to node j. Specify the constraints on variables x_{ij} of the linear programming problem associated with this network flow problem.

(2) Show that the solution in the figure below is optimal. We have indicated the flow x_{ij} beside each edge (i, j).



(3) Suppose that c_{53} changes from 5 to 3. Verify that the solution in part (a) is no longer optimal. Hence determine an optimal solution to the new problem (with $c_{53} = 3$). Start from the solution given in part (a).

Part 2

Quadratic optimization

Chapter 8

Convex optimization: basic properties

In a certain sense an optimization problem is well-posed if the feasible set is a convex set and the objective function to be minimized is convex. In this chapter we establish some basic properties of this type of problems.

We have already seen what is a convex *set* means in Section 4.4.1. Now we define convex *functions.*

8.1. Convex and strictly convex functions

Definition 8.1. Let $C \subset \mathbb{R}^n$ be a convex set. A function $f : C \to \mathbb{R}$ is said to be *convex* if for all $x, y \in C$ and all $t \in (0, 1)$,

$$f((1-t)x + ty) \leq (1-t)f(x) + tf(y).$$

An often useful equivalent form of this inequality is the following:

$$f(x + t(y-x)) \leq f(x) + t(f(y) - f(x)).$$

If the left hand side is strictly less than the right hand side for all distinct x and y in C, then f is called *strictly convex.*

Geometrically, the definition says that every linear interpolation of a convex function lies above the graph of the function; see Figure 1 in the case when $C = \mathbb{R}$.

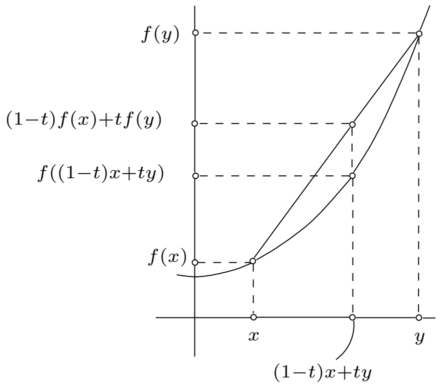


Figure 1. A convex function.

Exercise 8.2. Show that $f : \mathbb{R} \to \mathbb{R}$ given by $f(x) = x^2$ $(x \in \mathbb{R})$ is convex using the definition.

A slick way of proving convexity of smooth functions from $\mathbb{R}$ to $\mathbb{R}$ is to check if f'' is nonnegative; see Exercise 8.3 below.

Exercise 8.3. Prove that if $f : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable and $f''(x) \geq 0$ for all $x \in \mathbb{R}$, then f is convex. Moreover, show that the condition that $f''(x) > 0$ for all x guarantees strict convexity.

Exercise 8.4. Show that the $f : \mathbb{R} \to \mathbb{R}$ is a convex function, where f is given by:

(1) $f(x) = x$.
(2) $f(x) = x^2$.
(3) $f(x) = e^x$.
(4) $f(x) = e^{-x}$.
(5) $f(x) = |x|$.

In which of these cases is the function strictly convex?

Exercise 8.5. Show that if f is convex on the convex set $C \subset \mathbb{R}^n$, and $r \in \mathbb{R}$, then the set

$$K := \{x \in C : f(x) \leq r\}$$

is a convex subset of $\mathbb{R}^n$.

Exercise 8.6. Let $C \subset \mathbb{R}^n$ be a convex set, and let $f : C \to \mathbb{R}$ be a function. Define the *epigraph* of f by

$$U(f) = \bigcup_{x \in C} \{x\} \times (f(x), +\infty) \subset C \times \mathbb{R}.$$

This is the 'region above the graph of f'. Show that if f is convex, then $U(f)$ is a convex subset of $C \times \mathbb{R}$.

Exercise 8.7. Let $C \subset \mathbb{R}^n$ be a convex set and $f : C \to \mathbb{R}$ be a convex function. Show that for all $n \in \mathbb{N}$ and all $x_1, \dots, x_n \in C$, there holds that

$$f\left(\frac{x_1 + \cdots + x_n}{n}\right) \leq \frac{f(x_1) + \cdots + f(x_n)}{n}.$$

Exercise 8.8. Determine which of the following statements are TRUE. If the statement is FALSE, then you should give a counterexample, and if the statement is TRUE, then give a reason why. Let C_1, C_2 be convex subsets of $\mathbb{R}^n$ such that $C_1 \subset C_2$. Let $f : C_1 \to \mathbb{R}$ and $F : C_2 \to \mathbb{R}$ be functions such that $F|_{C_1} = f$, that is, for all $x \in C_1$, $F(x) = f(x)$.

(i) If F is convex, then f is convex.
(ii) If f is convex, then F is convex.

Exercise 8.9. Let C be a convex set subset of $\mathbb{R}^n$.

(1) Suppose that $(f_\alpha)_{\alpha \in I}$ be a family of convex functions on C such that $\sup\limits_{\alpha \in I} f_\alpha(x) < +\infty$ for all $x \in C$. Show that the function f defined by $f(x) = \sup\limits_{\alpha \in I} f_\alpha(x)$ $(x \in C)$, is convex. Prove that the set $K = \{x \in C : f_\alpha(x) \leq 0,\ \alpha \in I\}$ is a convex subset of $\mathbb{R}^n$.
(2) If $f_1, \dots, f_n$ are n convex functions on C and $\alpha_1, \dots, \alpha_n$ are nonnegative numbers, then s defined by $s(x) = \alpha_1 f_1(x) + \cdots + \alpha_n f_n(x)$ $(x \in C)$ is convex.

8.2. Convex optimization

Let $\mathcal{F} \subset \mathbb{R}^n$ be a given convex set and let $f : \mathcal{F} \to \mathbb{R}$ be a given convex function. The *convex optimization problem* is the following:

$$(CO) : \left\{ \begin{array}{ll} \text{minimize} & f(x), \\ \text{subject to} & x \in \mathcal{F}. \end{array} \right.$$

The function f is called the *objective function* for the problem (CO). The set $\mathcal{F}$ is called the *feasible set* for the problem (CO). An element $x \in \mathbb{R}^n$ is said to be a *feasible solution* for the problem (CO) if $x \in \mathcal{F}$. An element $\widehat{x} \in \mathbb{R}^n$ is said to be an *optimal feasible solution* for the problem (CO) if $\widehat{x} \in \mathcal{F}$ and for all $x \in \mathcal{F}$, $f(\widehat{x}) \leq f(x)$.

8.3. Set of optimal solutions

For a convex optimization problem (CO), each of the following three alternatives can hold:

1° The set of optimal solutions is empty.

2° The set of optimal solutions is nonempty, and consists of only one element $\widehat{x}$.

3° The set of optimal solutions is nonempty, and consists of more than one element.

Example 8.10. Let $\mathcal{F} = \mathbb{R}$.

1° If $f : \mathcal{F} \to \mathbb{R}$ is given by $f(x) = x$, then the set of optimal solutions of the problem (CO) for this f is empty.

2° If $f : \mathcal{F} \to \mathbb{R}$ is given by $f(x) = x^2$, then the set of optimal solutions of the problem (CO) for this f is nonempty, and consists of only one element $\widehat{x}$, namely $\widehat{x} = 0$.

3° If $f : \mathcal{F} \to \mathbb{R}$ is given by $f(x) = 0$, then the set of optimal solutions of the problem (CO) for this f is nonempty, and consists of more than one element. The set of optimal solutions is in fact $\mathcal{F} = \mathbb{R}$.

Lemma 8.11. *If there are more than one optimal solutions to the problem (CO), then there are infinitely many solutions, and moreover, the set of all optimal solutions is a convex set.*

Proof. If $\widehat{x}$ and $\widehat{y}$ are distinct optimal solutions in $\mathcal{F}$, then for all $t \in (0,1)$, $(1-t)\widehat{x} + t\widehat{y} \in \mathcal{F}$ is also an optimal solution: for all $x \in \mathcal{F}$,

$$f((1-t)\widehat{x} + t\widehat{y}) \le (1-t)f(\widehat{x}) + tf(\widehat{y}) \le (1-t)f(x) + tf(x) = f(x).$$

Thus there are infinitely many optimal solutions, since for distinct $t, t' \in (0,1)$,

$$(1-t)\widehat{x} + t\widehat{y} = \widehat{x} + t(\widehat{y} - \widehat{x}) \ne \widehat{x} + t'(\widehat{y} - \widehat{x}) = (1-t')\widehat{x} + t'\widehat{y}.$$

The optimality of $(1-t)\widehat{x} + t\widehat{y}$ demonstrated above also shows that the set of optimal solutions is convex. □

Lemma 8.12. *In the problem (CO), if f is strictly convex and $\mathcal{F}$ is convex, then the problem (CO) has* at most one *optimal solution.*

Proof. Let $\widehat{x}$ and $\widehat{y}$ be distinct optimal solutions in $\mathcal{F}$. We will proceed as in the previous lemma. By the strict convexity of f, we have for all $t \in (0,1)$,

$$f((1-t)\widehat{x} + t\widehat{y}) < (1-t)f(\widehat{x}) + tf(\widehat{y}) \le (1-t)f(\widehat{x}) + tf(\widehat{x}) = f(\widehat{x}),$$

contradicting the optimality of $\widehat{x}$. □

8.4. Feasible directions and descent directions

Given a feasible point $x \in \mathcal{F}$, one often wants to know in what directions one can move without immediately ending up out of $\mathcal{F}$. Also we want to know in which directions the objective function decreases, that is, the directions in which the graph of the function slopes downwards.

Definition 8.13. A vector $d \in \mathbb{R}^n$ is called a *feasible direction at $x \in \mathcal{F}$* if there exists an $\epsilon > 0$ such that $x + td \in \mathcal{F}$ for all $t \in (0, \epsilon)$.

A vector $d \in \mathbb{R}^n$ is called a *descent direction for f at $x \in \mathcal{F}$* if there exists an $\epsilon > 0$ such that $f(x + td) < f(x)$ for all $t \in (0, \epsilon)$.

A vector $d \in \mathbb{R}^n$ is called a *feasible descent direction for f at $x \in \mathcal{F}$* if d is *both* a feasible direction at x and a descent direction for f at x, that is, if there exists an $\epsilon > 0$ such that $x+td \in \mathcal{F}$ and $f(x + td) < f(x)$ for all $t \in (0, \epsilon)$.

8.5. Feasible descent directions and optimality

The following result is very useful in order to determine whether or not a given point $\widehat{x}$ is an optimal solution to the problem (CO).

Theorem 8.14. *A point $\widehat{x} \in \mathcal{F}$ is an optimal solution to the problem (CO) iff there does* not *exist a feasible descent direction for f at $\widehat{x}$.*

Proof. (Only if) Suppose that there is a feasible descent direction d for f at $\widehat{x}$. Then there is an $\epsilon > 0$ such that $x + td \in \mathcal{F}$ and $f(\widehat{x} + td) < f(\widehat{x})$ for all $t \in (0, \epsilon)$, which implies that $\widehat{x}$ is not optimal.

(If) Suppose that $\widehat{x} \in \mathcal{F}$ is not an optimal solution to the problem (CO). Then there exists a $y \in \mathcal{F}$ such that $f(y) < f(\widehat{x})$. We will show that $d := y - \widehat{x}$ is a feasible descent direction for f at $\widehat{x}$. Let $x(t) := \widehat{x} + td$ for all $t \in (0, 1)$. Since $\mathcal{F}$ is convex, it follows that $x(t) \in \mathcal{F}$ for all $t \in (0, 1)$. Hence d is a feasible direction at $\widehat{x}$ (take $\epsilon = 1$!). Also, from the convexity of f, we have

$$\begin{aligned} f(x(t)) &= f((1-t)\widehat{x} + ty) \\ &\leq (1-t)f(\widehat{x}) + tf(y) \\ &< (1-t)f(\widehat{x}) + tf(\widehat{x}) \\ &= f(\widehat{x}), \end{aligned}$$

for all $t \in (0, 1)$, showing that d is a descent direction for f at $\widehat{x}$. □

Chapter 9

Quadratic optimization: no constraints

Definition 9.1. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called a *quadratic function* if

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x + c_0 \quad (x \in \mathbb{R}^n), \tag{9.1}$$

where $H \in \mathbb{R}^{n\times n}$ is a symmetric matrix, $c \in \mathbb{R}^n$, $c_0 \in \mathbb{R}$.

Example 9.2. The function f given by $f(x_1, x_2, x_3) = -2x_1x_2 + x_3 + 1$, is a quadratic function since it has the form (9.1), with

$$H = \begin{bmatrix} 0 & -2 & 0 \\ -2 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad c = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad c_0 = 1.$$

$\Diamond$

Definition 9.3. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quadratic function.

A point $\widehat{x} \in \mathbb{R}^n$ is said to be a *minimizer of* f if for all $x \in \mathbb{R}^n$, $f(\widehat{x}) \le f(x)$.

f is said to be *bounded from below* if there exists a $l \in \mathbb{R}$ such that $f(x) \ge l$ for all $x \in \mathbb{R}^n$.

Example 9.4. The function f given by $f(x_1, x_2, x_3) = -2x_1x_2 + x_3 + 1$, is not bounded from below. Indeed, for $t > 0$, we have $f(t, t, 0) = -2t^2 + 1$, and so as $t \nearrow +\infty$, $f(t, t, 0) \to -\infty$. $\Diamond$

If f is not bounded from below, then there is no minimizer of f, since for any $\widehat{x} \in \mathbb{R}^n$, there is a $x \in \mathbb{R}^n$ such that $f(x) < f(\widehat{x})$ (otherwise with $l := f(\widehat{x})$, f would be bounded from below!).

9.1. The Taylor expansion of a quadratic function

Lemma 9.5. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quadratic function, given by*

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x + c_0 \quad (x \in \mathbb{R}^n),$$

where $H \in \mathbb{R}^{n\times n}$ is a symmetric matrix, $c \in \mathbb{R}^n$, $c_0 \in \mathbb{R}$. Then for all $x \in \mathbb{R}^n$, all $d \in \mathbb{R}^n$ and all $t \in \mathbb{R}$,

$$f(x + td) = f(x) + t(Hx + c)^\top d + \frac{1}{2}t^2 d^\top Hd.$$

Proof. This is a straightforward calculation:

$$\begin{array}{rcl} f(x+td) & = & \frac{1}{2}(x+td)^\top H(x+td)+c^\top(x+td)+c_0 \\ & = & \frac{1}{2}(x^\top Hx+td^\top Hx+tx^\top Hd+t^2d^\top Hd)+c^\top x+tc^\top d+c_0 \\ & = & \left[\frac{1}{2}x^\top Hx+c^\top x+c_0\right]+t\left(\frac{1}{2}d^\top Hx+\frac{1}{2}x^\top Hd+c^\top d\right)+\frac{1}{2}t^2d^\top Hd \\ & = & f(x)+t(Hx+c)^\top d+\frac{1}{2}t^2d^\top Hd. \end{array}$$

This completes the proof. □

In particular, with $t=1$ and $d=y-x$ in the above, we obtain that for all $x,y\in\mathbb{R}^n$,

$$f(y)=f(x)+(Hx+c)^\top(y-x)+\frac{1}{2}(y-x)^\top H(y-x). \tag{9.2}$$

Remark 9.6. It is not hard to verify that the *gradient of f at x* is given by

$$\nabla f(x)=\left[\begin{array}{ccc}\frac{\partial f}{\partial x_1}(x) & \dots & \frac{\partial f}{\partial x_n}(x)\end{array}\right]=(Hx+c)^\top.$$

The *Hessian of f at x* is the derivative of the gradient function at x, and can be identified with the $n\times n$ matrix whose entry in the ith row and jth column is equal to

$$\frac{\partial^2 f}{\partial x_i\partial x_j}(x),$$

where i,j range from 1 to n, and it can be checked that this is equal to H. Since the Hessian does not change with x, all further derivatives of f are identically 0. Hence (9.2) is just the Taylor expansion of the quadratic function f, and all terms beyond order 2 are zero.

9.2. Convex and strictly convex quadratic functions

When is a quadratic function convex? When is it strictly convex? The following lemma answers these questions.

Lemma 9.7. *Let $f:\mathbb{R}^n\to\mathbb{R}$ be a quadratic function, given by*

$$f(x)=\frac{1}{2}x^\top Hx+c^\top x+c_0 \quad (x\in\mathbb{R}^n),$$

where $H\in\mathbb{R}^{n\times n}$ is a symmetric matrix, $c\in\mathbb{R}^n$, $c_0\in\mathbb{R}$. Then

(1) *f is convex iff H is positive semi-definite.*

(2) *f is strictly convex iff H is positive definite.*

Proof. For all $x,y\in\mathbb{R}^n$ and all $t\in(0,1)$, we have

$$\begin{array}{rcl} (1-t)f(x)+tf(y)-f((1-t)x+ty) & = & f(x)+t(f(y)-f(x))-f(x+t(y-x)) \\ & = & f(x)+t\left((Hx+c)^\top(y-x)+\frac{1}{2}(y-x)^\top H(y-x)\right) \\ & & -\left(f(x)+t(Hx+c)^\top(y-x)+\frac{1}{2}t^2(y-x)^\top H(y-x)\right) \\ & = & \frac{1}{2}(t-t^2)(y-x)^\top H(y-x). \end{array}$$

Consequently, we have for all $x, y \in \mathbb{R}^n$ and all $t \in (0,1)$,

$$(1-t)f(x) + tf(y) - f((1-t)x + ty = \frac{1}{2}t(1-t)(y-x)^\top H(y-x). \tag{9.3}$$

(1) If f is convex, then (9.3) implies that for all $x, y \in \mathbb{R}^n$ and all $t \in (0,1)$, there holds that $\frac{1}{2}t(1-t)(y-x)^\top H(y-x) \geq 0$. In particular with $x = 0$ and $t = \frac{1}{2}$, we obtain that for all $y \in \mathbb{R}^n$, $y^\top Hy \geq 0$, and so H is positive semi-definite.

Conversely, if H is positive semi-definite, then it follows that $(y-x)^\top H(y-x) \geq 0$ for all $x, y \in \mathbb{R}^n$. Also for all $t \in (0,1)$, clearly $\frac{1}{2}t(1-t) > 0$. Hence for all $x, y \in \mathbb{R}^n$ and all $t \in (0,1)$, $\frac{1}{2}t(1-t)(y-x)^\top H(y-x) \geq 0$. But now from (9.3), we obtain $f((1-t)x+ty) \leq (1-t)f(x)+tf(y)$ for all $x, y \in \mathbb{R}^n$ and all $t \in (0,1)$, showing that f is convex.

(2) Suppose f is strictly convex. Let $y \neq 0 =: x$ and $t = \frac{1}{2}$. Then by the strict convexity of f, $f((1-t)x + ty) < (1-t)f(x) + tf(y)$, and so by (9.3), $\frac{1}{8}y^\top Hy > 0$. But the choice of $y \neq 0$ was arbitrary, and so H is positive definite.

Conversely, let H be positive definite. Then for $y \neq x$, we have $(y-x)^\top H(y-x) > 0$. Also for all $t \in (0,1)$, clearly $\frac{1}{2}t(1-t) > 0$. Hence for all $x, y \in \mathbb{R}^n$ with $x \neq y$ and all $t \in (0,1)$, $\frac{1}{2}t(1-t)(y-x)^\top H(y-x) > 0$. But now from (9.3), we obtain $f((1-t)x+ty) < (1-t)f(x)+tf(y)$ for all $x, y \in \mathbb{R}^n$ with $x \neq y$ and all $t \in (0,1)$, showing that f is strictly convex. □

Exercise 9.8. Show that in each of the following two cases, $f : \mathbb{R}^3 \to \mathbb{R}$ is convex. In which of the cases is f strictly convex?

(1) $f(x) = x_1^2 + 2x_2^2 + 5x_3^2 + 3x_2x_3$.

(2) $f(x) = 2x_1^2 + x_2^2 + x_3^2 - 2x_1x_2 + 2x_1x_3$.

Exercise 9.9. For which values of a is the function $f : \mathbb{R}^2 \to \mathbb{R}$, given by $f(x) = x_1^2 + 2x_2^2 + 2ax_1x_2$ convex? Strictly convex?

9.3. Descent directions

For all quadratic functions (convex as well as non-convex), the following holds.

Lemma 9.10. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quadratic function, given by*

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x + c_0 \quad (x \in \mathbb{R}^n),$$

where $H \in \mathbb{R}^{n\times n}$ is a symmetric matrix, $c \in \mathbb{R}^n$, $c_0 \in \mathbb{R}$.

If $x \in \mathbb{R}^n$ and $d \in \mathbb{R}^n$ are such that $(Hx + c)^\top d < 0$, then d is a descent direction at x.

Proof. Let $t > 0$. We have

$$f(x+td) = f(x) + t(Hx+c)^\top d + \frac{1}{2}t^2 d^\top Hd = f(x) + \frac{1}{2}t(2(Hx+c)^\top d + td^\top Hd) < f(x)$$

for all $t > 0$ such that

$$t(d^\top Hd) < \underbrace{-2(Hx+c)^\top d}_{>0}.$$

Note that this is guaranteed for all $t > 0$ if $d^\top Hd \leq 0$. On the other hand, if $d^\top Hd > 0$, then this is guaranteed for all small enough $t > 0$ (in fact for all $0 < t < \frac{-2(Hx+c)^\top d}{d^\top Hd}$). This shows that d is a descent direction for f at x. □

For a *convex* quadratic function, also the converse of the above result holds.

Lemma 9.11. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a convex quadratic function, given by*

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x + c_0 \quad (x \in \mathbb{R}^n),$$

where $H \in \mathbb{R}^{n\times n}$ is a symmetric positive semi-definite matrix, $c \in \mathbb{R}^n$, $c_0 \in \mathbb{R}$.

The vector $d \in \mathbb{R}^n$ is a descent direction at $x \in \mathbb{R}^n$ iff $(Hx + c)^\top d < 0$.

Proof. In light of the previous lemma, we only need to show that 'only if' part. That is, we want to show that if $d \in \mathbb{R}^n$ is a descent direction at $x \in \mathbb{R}^n$, then $(Hx + c)^\top d < 0$. Equivalently, we will prove that if $(Hx + c)^\top d \geq 0$, then d is *not* a descent direction at x. Suppose therefore that $(Hx + c)^\top d \geq 0$. Then for all $t \geq 0$, we have

$$f(x + td) = f(x) + t\underbrace{(Hx + c)^\top d}_{\geq 0} + \frac{1}{2}t^2\underbrace{d^\top Hd}_{\geq 0} \geq f(x),$$

since H is positive semi-definite. This shows that d cannot be a descent direction for f at x. □

Exercise 9.12. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quadratic function, given by

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x + c_0 \quad (x \in \mathbb{R}^n),$$

where $H \in \mathbb{R}^{n\times n}$ is a symmetric matrix, $c \in \mathbb{R}^n$, $c_0 \in \mathbb{R}$. Show that if the vector $d \in \mathbb{R}^n$ is a descent direction at $x \in \mathbb{R}^n$, then $(Hx + c)^\top d \leq 0$.

9.4. Minimizing non-convex quadratics on $\mathbb{R}^n$

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a quadratic function, given by

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x + c_0 \quad (x \in \mathbb{R}^n),$$

where $H \in \mathbb{R}^{n\times n}$ is a symmetric matrix, $c \in \mathbb{R}^n$, $c_0 \in \mathbb{R}$. Suppose that the quadratic function f is not convex, that is H is not positive semi-definite. Then there exists a vector $d \in \mathbb{R}^n$ such that $d^\top Hd < 0$. Define $x(t) = td$, where $t \in \mathbb{R}$. We have

$$f(x(t)) = f(td) = \frac{1}{2}t^2 d^\top Hd + tc^\top d + c_0.$$

Since $d^\top Hd < 0$, as $t \to +\infty$, we have $f(x(t)) \to -\infty$. This means that f is not bounded below, and so there is no minimizer of f.

In light of the discussion in this section, we will assume in the rest of this chapter that f is convex, that is,

$$\boxed{H \text{ is positive semi-definite.}}$$

9.5. Minimizing convex quadratics on $\mathbb{R}^n$

Now suppose that H is positive semi-definite, that is, f is convex. Note that since there are no constraints, the feasible set $\mathcal{F} = \mathbb{R}^n$, which is convex. So we have a convex optimization problem, namely

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}x^\top Hx + c^\top x + c_0, \\ \text{subject to} & x \in \mathbb{R}^n. \end{array} \tag{9.4}$$

Theorem 9.13. *Let H be positive semi-definite. The point $\widehat{x} \in \mathbb{R}^n$ is an optimal solution to* (9.4) *iff $H\widehat{x} = -c$.*

Proof. The problem is a convex optimization problem. So by Theorem 8.14, $\widehat{x}$ is an optimal solution to the problem (9.4) iff there does not exist a feasible descent direction for f at $\widehat{x}$. But the feasible set is $\mathbb{R}^n$, and so every vector $d \in \mathbb{R}^n$ is feasible at $\widehat{x}$. Also by Lemma 9.11, $d \in \mathbb{R}^n$ is a descent direction for f at $\widehat{x}$ iff $(H\widehat{x}+c)^\top d < 0$. Combining these facts, we conclude that $\widehat{x}$ is an optimal solution to the problem (9.4) iff for every $d \in \mathbb{R}^n$, $(H\widehat{x}+c)^\top d \geq 0$. But $(H\widehat{x}+c)^\top d \geq 0$ for every $d \in \mathbb{R}^n$ iff $H\widehat{x}+c=0$. (Why?) Consequently, $\widehat{x}$ is an optimal solution iff $H\widehat{x} = -c$. □

So the above result implies that if H is positive semi-definite, then every minimizer of f is a solution to the system $Hx = -c$. This solution has at least one solution iff $-c \in \operatorname{ran} H$. So if $-c \notin \operatorname{ran} H$, then there is no minimizer of f. The following result says that one can say more.

Theorem 9.14. *Suppose that H is positive semi-definite and that $-c \notin \operatorname{ran} H$. Then there is a vector $d \in \mathbb{R}^n$ such that $f(td) \to -\infty$ as $t \nearrow +\infty$. (That is, f is not bounded from below.)*

Proof. Since H is symmetric, the two subspaces $\ker H$ and $\operatorname{ran} H$ are orthogonal to each other[1], and so the vector $-c \in \mathbb{R}^n$ can be uniquely decomposed as $-c = d + p$, where $d \in \ker H$ and $p \in \operatorname{ran} H$. The fact that $-c \notin \operatorname{ran} H$ implies that $d \neq 0$, and so

$$c^\top d = -(d+p)^\top d = -d^\top d - \underbrace{p^\top d}_{=0} = -d^\top d = -\|d\|^2 < 0.$$

Also $Hd = 0$. Thus

$$f(td) = \frac{1}{2}t^2 d^\top H d + t c^\top d + c_0 = 0 - t\|d\|^2 + c_0.$$

So $f(td) \to -\infty$ as $t \nearrow +\infty$. □

9.5.1. The strictly convex case. If H is positive definite, then the system $Hx = -c$ has a *unique* solution (since every positive definite matrix is invertible). Thus there is a unique $\widehat{x} \in \mathbb{R}^n$ which is optimal solution to (9.4), given by $\widehat{x} = -H^{-1}c$.

Exercise 9.15. Find a symmetric $H \in \mathbb{R}^3$ such that

$$\frac{1}{2}x^\top Hx = (x_1 - x_2)^2 + (x_2 - x_3)^2 + (x_3 - x_1)^2, \quad x \in \mathbb{R}^3.$$

Is H positive semi-definite? What is the kernel of H?

Let $c \in \mathbb{R}^3$ be given, and consider the problem of minimizing f on $\mathbb{R}^3$, where

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x, \quad x \in \mathbb{R}^3.$$

Show that there exists a vector $v \in \mathbb{R}^3$ such that:

$$[f \text{ has at least one minimizer}] \quad \Leftrightarrow \quad [v^\top c = 0].$$

Find a vector c such that f has at least one minimizer. Find a vector c such that f is not bounded from below.

Exercise 9.16. Let L_1, L_2 be two given lines in parametric form in $\mathbb{R}^3$ as follows:

$$L_1 = \{x \in \mathbb{R}^3 : x = a + \alpha \cdot u, \text{ for some } \alpha \in \mathbb{R}\},$$
$$L_2 = \{x \in \mathbb{R}^3 : x = b + \beta \cdot u, \text{ for some } \beta \in \mathbb{R}\},$$

where a, b and u, v are fixed given vectors in $\mathbb{R}^3$. We also assume that the direction vectors u and v of the lines are normalized, that is, $u^\top u = v^\top v = 1$, and that they are not parallel (and so $u^\top v < 1$).

We would like to connect these lines with a thread having the shortest length, that is, we would like to determine points $\widehat{x} \in L_1$ and $\widehat{y} \in L_2$ such that the when the taut thread is tied at these two points, then its length is the smallest possible one.

(1) Formulate the given problem as a quadratic optimization problem in two variables (α and β).

(2) Show that the problem is convex.

(3) Suppose that $u^\top v = 0$. Find the optimal $\widehat{x}$ and $\widehat{y}$ in terms of the given data.

[1] See Exercise 24.2.

9.6. Summary

Let $f(x) = \frac{1}{2}x^\top Hx + c^\top x + c_0$, where H is symmetric. Then:

$\widehat{x}$ is a minimizer of f iff H is positive semi-definite and $H\widehat{x} = -c$. In particular, if H is positive definite, then there is a unique minimizer of f, given by $\widehat{x} = -H^{-1}c$.

If H is positive semi-definite and $-c \notin \operatorname{ran} H$, then f is not bounded from below and f does not have a minimizer.

If H is not positive semi-definite, then f is not bounded from below and f does not have a minimizer.

As a corollary, we have obtained the following interesting property of quadratic functions: f has a minimizer iff it is bounded below.

Chapter 10

Quadratic optimization: equality constraints

In this chapter we will consider the following quadratic optimization problem with linear equality constraints:

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}x^\top Hx + c^\top x + c_0, \\ \text{subject to} & Ax = b. \end{array} \tag{10.1}$$

Here $A \in \mathbb{R}^{m\times n}$, $H \in \mathbb{R}^{n\times n}$ is symmetric, $b \in \mathbb{R}^m$, $c \in \mathbb{R}^n$ and $c_0 \in \mathbb{R}$. The vector $x \in \mathbb{R}^n$ is the vector of variables. The feasible set is $\mathcal{F} = \{x \in \mathbb{R}^n : Ax = b\}$, and the objective function is the function f, given by

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x + c_0.$$

If the system $Ax = b$ does not have any solution (that is, $b \notin \operatorname{ran} A$), then the feasible set $\mathcal{F}$ is empty, and the optimization problem is trivial.

Also, if the system $Ax = b$ has exactly one solution, then this solution is also the unique optimal solution to the problem (10.1), since there is no other feasible solution which is better! Thus this case is trivial as well.

So the only interesting case is when the system $Ax = b$ has many different solutions, which is equivalent to the condition that $b \in \operatorname{ran} A$ and $\ker A \neq \{0\}$. This is fulfilled for example if the matrix A has more columns than rows, that is, $n > m$, and the columns of A span $\mathbb{R}^m$. Indeed, then $\ker A$ has dimension $n - m > 0$, and then $\operatorname{ran} A = \mathbb{R}^m$ implies that the system $Ax = b$ has many different solutions for any given $b \in \mathbb{R}^m$.

Thus in the remainder of the chapter, it will be assumed that

$$\boxed{b \in \operatorname{ran} A, \text{ and } \ker A \neq \{0\}.}$$

10.1. Representation of the feasible solutions

Let $\overline{x} \in \mathbb{R}^n$ be *a* feasible solution, that is, $A\overline{x} = b$. The other feasible solutions x are then characterized by $x - \overline{x} \in \ker A$: Indeed, first of all if $x \in \mathcal{F}$, then $Ax = b$, and so we have that $A(x - \overline{x}) = Ax - A\overline{x} = b - b = 0$, that is, $x - \overline{x} \in \ker A$. On the other hand, if $x - \overline{x} \in \ker A$, then $A(x - \overline{x}) = 0$, and so $Ax = A(x - \overline{x} + \overline{x}) = A(x - \overline{x}) + A\overline{x} = 0 + b = b$.

Now let k be the dimension of $\ker A$, and let $z_1, \dots, z_k$ form a basis for $\ker A$. Define the $n \times k$ matrix Z as follows:

$$Z = \begin{bmatrix} z_1 & \dots & z_k \end{bmatrix}.$$

Then $x - \overline{x} \in \ker A$ iff $x - \overline{x} = Zv$ for some $v \in \mathbb{R}^k$, and so we obtain the following representation of the feasible solutions:

$$x \in \mathcal{F} \quad \text{iff} \quad x = \overline{x} + Zv \text{ for some } v \in \mathbb{R}^k.$$

Since $z_1, \dots, z_k$ are linearly independent, it follows that for every $x \in \mathcal{F}$, there is a unique $v \in \mathbb{R}^k$.

10.2. When is the problem convex?

The feasible set is always convex. Indeed, suppose that $x, y \in \mathcal{F}$ and $t \in (0,1)$. Then $Ax = b$ and $Ay = b$, and so

$$A((1-t)x + ty) = (1-t)Ax + tAy = (1-t)b + tb = b.$$

Thus $(1-t)x + ty \in \mathcal{F}$ as well.

We will now give a necessary and sufficient condition for f to be convex on $\mathcal{F}$.

Lemma 10.1. *Let $\mathcal{F} = \{x \in \mathbb{R}^n : Ax = b\}$, and $f : \mathcal{F} \to \mathbb{R}$ be given by*

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x + c_0 \quad (x \in \mathcal{F}).$$

Then f is convex iff $Z^\top HZ$ is positive semi-definite. (Here Z is a matrix of the type described in the previous section.)

Proof. (If) Let $Z^\top HZ$ be positive semi-definite, and let $x, y \in \mathcal{F}$, $t \in (0,1)$. Then $x - y \in \ker A$, and so $x - y = Zv$ for some $v \in \mathbb{R}^k$. With the same calculation as done earlier in the beginning of the proof of Lemma 9.7, we see that

$$(1-t)f(x) + tf(y) - f((1-t)x + ty) = \frac{1}{2}(t - t^2)(y - x)^\top H(y - x).$$

Thus we have $(1-t)f(x) + tf(y) - f((1-t)x + ty) = \frac{1}{2}t(1-t)v^\top(Z^\top HZ)v \geq 0$. Hence f is convex.

(Only if) Suppose that f is convex. Let $v \in \mathbb{R}^k$. Take $x = \overline{x}$ and $y = \overline{x} + Zv$, where $\overline{x} \in \mathbb{R}^n$ is such that $A\overline{x} = b$. Set $t = \frac{1}{2}$. Then we have

$$0 \leq (1-t)f(x) + tf(y) - f((1-t)x + ty) = \frac{1}{2}(t - t^2)(y - x)^\top H(y - x) = \frac{1}{8}v^\top Z^\top HZv,$$

and so $v^\top(Z^\top HZ)v \geq 0$. But the choice of $v \in \mathbb{R}^k$ was arbitrary, and this means that $Z^\top HZ$ is positive semi-definite. □

A sufficient (but not necessary[1]) condition for $Z^\top HZ$ to be positive semi-definite is that H is positive semi-definite, that is, that f is convex in the whole space.

One can also show the following in the same manner as the proof of Lemma 10.1 above (more or less simply by swapping ≥ 0 by > 0), and we leave this verification as an exercise.

Exercise 10.2. Let $\mathcal{F} = \{x \in \mathbb{R}^n : Ax = b\}$ and let $f : \mathcal{F} \to \mathbb{R}$ be given by

$$f(x) = \frac{1}{2}x^\top Hx + c^\top x + c_0 \quad (x \in \mathcal{F}).$$

Show that f is strictly convex iff $Z^\top HZ$ is positive definite. (With Z as described previously.)

In the rest of this chapter we make the standing assumption that the problem (10.1) is convex, that is,

$$\boxed{Z^\top HZ \text{ is positive semi-definite.}}$$

[1] For example, take $Z = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $H = \begin{bmatrix} 0 & 0 \\ 0 & -1 \end{bmatrix}$.

10.3. Optimal solution by the nullspace method

Based on the representation of feasible solutions given in Section 10.1, we can replace x by $\overline{x}+Zv$ (where $\overline{x}$ is fixed, and $v \in \mathbb{R}^k$ is unconstrained). Then the objective function becomes a quadratic function in the new variable vector v:

$$\begin{aligned} f(\overline{x}+Zv) &= f(\overline{x}) + (H\overline{x}+c)^\top Zv + \frac{1}{2}(Zv)^\top H(Zv) \\ &\qquad \text{(using Lemma 9.5 with } t=1 \text{ and } d=Zv\text{)} \\ &= f(\overline{x}) + (Z^\top(H\overline{x}+c))^\top v + \frac{1}{2}v^\top(Z^\top HZ)v. \end{aligned}$$

The problem (10.1) is thus equivalent to the following *unconstrained* problem in v:

$$\begin{array}{rl} \text{minimize} & f(\overline{x}) + (Z^\top(H\overline{x}+c))^\top v + \frac{1}{2}v^\top(Z^\top HZ)v, \\ \text{subject to} & v \in \mathbb{R}^k. \end{array} \tag{10.2}$$

Since $Z^\top HZ$ is positive semi-definite, it follows from Theorem 9.13 that $\widehat{v} \in \mathbb{R}^k$ is an optimal solution to the problem (10.2) iff

$$(Z^\top HZ)\widehat{v} = -Z^\top(H\overline{x}+c).$$

Consequently, $\widehat{x} \in \mathcal{F}$ is an optimal solution to the problem (10.1) iff

$$(Z^\top HZ)\widehat{v} = -Z^\top(H\overline{x}+c) \quad \text{and} \quad \widehat{x} = \overline{x} + Z\widehat{v}.$$

In the special case that $Z^\top HZ$ is positive definite (and not just positive *semi*-definite), the system $(Z^\top HZ)\widehat{v} = -Z^\top(H\overline{x}+c)$ has a unique solution $\widehat{v}$, and then $\widehat{x} = \overline{x} + Z\widehat{v}$ is the unique optimal solution to the problem (10.1).

10.4. Optimal solution by the Lagrange method

There is another way to solve the problem (10.1), which does not involve the null space matrix Z. Sometimes, this leads to a more efficient method than the one considered in the previous section. Furthermore, this alternative method is important for generalizations of the theory and it can also give important insights in specific applications.

The set of feasible descent directions for f at a given point $x \in \mathcal{F}$ can be characterized in a simple and explicit way for the (convex) problem (10.1).

Lemma 10.3. *Consider the problem* (10.1) *where f is convex. Then $d \in \mathbb{R}^n$ is a feasible descent direction for f at $x \in \mathcal{F}$ iff $d \in \ker A$ and $(Hx+c)^\top d < 0$.*

Proof. (If) Suppose first that $x \in \mathcal{F}$, $d \in \ker A$ and $(Hx+c)^\top d < 0$. For every $t \in \mathbb{R}$, $A(x+td) = Ax + tAd = Ax = b$, and so d is a feasible direction at x. Now let $t > 0$. We have

$$f(x+td) = f(x) + t(Hx+c)^\top d + \frac{1}{2}t^2 d^\top Hd = f(x) + \frac{1}{2}t(2(Hx+c)^\top d + td^\top Hd) < f(x)$$

for all $t > 0$ such that

$$t(d^\top Hd) < \underbrace{-2(Hx+c)^\top d}_{>0}.$$

Note that this is guaranteed for all $t > 0$ if $d^\top Hd \le 0$. On the other hand, if $d^\top Hd > 0$, then this is guaranteed for all small enough $t > 0$ (in fact for all $0 < t < \frac{-2(Hx+c)^\top d}{d^\top Hd}$). This shows that d is a descent direction for f at x.

(Only if) Suppose now that $x \in \mathcal{F}$ and that d is a feasible descent direction for f at x. If $d \notin \ker A$, then $Ad \neq 0$. Hence for all $t \neq 0$, $A(x+td) = Ax + tAd \neq b$, which means that d is not a feasible direction at x, a contradiction. So we conclude that $d \in \ker A$. Now suppose that $(Hx+c)^\top d \geq 0$. For all $t \geq 0$, we have

$$f(x+td) = f(x) + t\,\underbrace{(Hx+c)^\top d}_{\geq 0} + \frac{1}{2}t^2\,\underbrace{d^\top Hd}_{\geq 0} \geq f(x),$$

since $(Hx+c)^\top d \geq$ and $d^\top Hd \geq 0$ (convexity of f!). But this means that d is not a feasible descent direction for f at x, a contradiction. So $(Hx+c)^\top d < 0$. □

Lemma 10.4. *Consider the problem* (10.1)*, where f is convex. Then $\widehat{x} \in \mathcal{F}$ is an optimal solution iff $(H\widehat{x}+c)^\top d = 0$ for all $d \in \ker A$.*

Proof. Theorem 8.14 and the previous lemma yield that a point $\widehat{x}$ is an optimal solution to the problem (10.1) iff

$$\text{for all } d \in \ker A, \quad (H\widehat{x}+c)^\top d \geq 0,$$

which in turn is satisfied iff

$$\text{for all } d \in \ker A, \quad (H\widehat{x}+c)^\top d = 0$$

(since $d \in \ker A \Leftrightarrow -d \in \ker A$). □

Theorem 10.5. *Consider the problem* (10.1)*, where f is convex. Then $\widehat{x} \in \mathbb{R}^n$ is an optimal solution iff $A\widehat{x} = b$ and there exists a $u \in \mathbb{R}^m$ such that $H\widehat{x} + c = A^\top u$.*

Proof. By the previous lemma, $\widehat{x} \in \mathbb{R}^n$ is an optimal solution to the problem (10.1) iff $A\widehat{x} = b$ and $H\widehat{x} + c \in (\ker A)^\perp$. So the result follows using the fact that $(\ker A)^\perp = \operatorname{ran} A^\top$. □

Remark 10.6. The above result says that $\widehat{x} \in \mathbb{R}^n$ is an optimal solution to the (convex) problem (10.1) iff $\widehat{x}$ is the "x-part" of a solution to the system

$$\begin{bmatrix} H & -A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} x \\ u \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix}. \tag{10.3}$$

Why is this method referred to as the *Lagrange* method? This comes from the fact that (10.3) can be viewed as a special case of the more general *Lagrange conditions* for nonlinear optimization with equality constraints, and u is then a vector of *Lagrange multipliers*.

Exercise 10.7. Let a be a nonzero vector in $\mathbb{R}^n$ and b be a nonnegative number. Consider the hyperplane $P := \{x \in \mathbb{R}^n : a^\top x = b\}$. Let $y \in \mathbb{R}^n$ be given. Suppose we want to find the distance $d(y,P)$ of the point y to the plane P, where $d(y,P) := \inf_{x \in P} \|x-y\|$.

Formulate a quadratic optimization problem subject to linearity constraints that enables one to find $d(y,P)$. Solve this quadratic optimization problem, and prove that the unique point $\widehat{x}$ in P that is closest to y is given by

$$\widehat{x} = \frac{b - a^\top y}{\|a\|^2} a + y.$$

Also show that $d(y,P) = \dfrac{|b - a^\top y|}{\|a\|}$.

Exercise 10.8. Let f be given by $f(x) = (x_1-x_2)^2 + (x_2-x_3)^2 + (x_3-x_1)^2$.

Let $A = \begin{bmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{bmatrix}$ and $b = \begin{bmatrix} 10 \\ 14 \end{bmatrix}$.

(1) Find *one* solution $\overline{x}$ to the system of equations $Ax = b$.

(2) Determine a basis for $\ker A$.

(3) Find an optimal solution $\widehat{x}$ to the problem

$$\begin{cases} \text{minimize} & f(x) \\ \text{subject to} & Ax = b. \end{cases}$$

using the nullspace method.

Exercise 10.9. Consider the quadratic optimization problem

$$\left\{\begin{array}{ll} \text{minimize} & \frac{1}{2}x^\top Hx, \\ \text{subject to} & Ax = b \end{array}\right.$$

where $H = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 1 & 0 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 2 & 1 \end{bmatrix}$, $A = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{bmatrix}$, $b = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}$.

(1) Show that a feasible solution for the problem is given by $\overline{x} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \end{bmatrix}^\top$.

(2) Show that the vectors z_1, z_2 form a basis for $\ker A$, where $z_1 = \begin{bmatrix} 0 & -1 & 0 & 1 & 0 \end{bmatrix}^\top$ and $z_2 = \begin{bmatrix} 1 & 0 & -1 & 0 & 1 \end{bmatrix}^\top$.

(3) Find an optimal solution $\widehat{x}$ to the problem.

Exercise 10.10. Let $A = \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \end{bmatrix}$ and $q = \begin{bmatrix} 4 \\ 2 \\ 0 \\ -2 \end{bmatrix}$.

(1) Consider the problem of determining a vector x in the kernel of A which is closest to q, that is,

$$(P1) : \left\{\begin{array}{ll} \text{minimize} & \|x - q\|^2 \\ \text{subject to} & x \in \ker A. \end{array}\right.$$

Find an optimal x.

(2) Next consider the problem of determining a vector x in the range of $A^\top$ which is closest to q, that is,

$$(P2) : \left\{\begin{array}{ll} \text{minimize} & \|x - q\|^2 \\ \text{subject to} & x \in \text{ran } (A^\top). \end{array}\right.$$

Find an optimal x.

10.5. Summary

Consider the quadratic optimization problem

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}x^\top Hx + c^\top x + c_0, \\ \text{subject to} & Ax = b. \end{array}$$

Let $\overline{x}$ be *a* solution of the system $Ax = b$, and let Z be a matrix whose columns form a basis for $\ker A$. Suppose that $Z^\top HZ$ is positive semi-definite. Then the following are equivalent:

(1) $\widehat{x}$ is an optimal solution to this problem.

(2) $\widehat{x} = \overline{x} + Z\widehat{v}$, where $\widehat{v}$ satisfies $(Z^\top HZ)\widehat{v} = -Z^\top(H\overline{x} + c)$.

(3) There exists a u such that $\begin{bmatrix} H & -A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \widehat{x} \\ u \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix}$.

10.6. Some remarks

10.6.1. What if *f* is not convex? We have throughout assumed that the quadratic optimization problem (10.1) is convex, that is, that $Z^\top HZ$ is positive semi-definite. If this is not the case, then the objective function in (10.2) is not bounded below and it does not have a minimizer. Thus the original problem (10.1) also does not have an optimal solution if $Z^\top HZ$ is not positive semi-definite.

10.6.2. Is there always an optimal solution? Even if $Z^\top HZ$ is positive semi-definite, it is not guaranteed that there is an optimal solution to the problem (10.2). A necessary and sufficient condition for the existence of at least one optimal solution to (10.2), and thereby also to the problem (10.1), is that the system $(Z^\top HZ)v = -Z^\top(H\overline{x} + c)$ should have at least one solution v, that is, $-Z^\top(H\overline{x} + c) \in \operatorname{ran}(Z^\top HZ)$.

10.6.3. Nullspace method versus Lagrange method. Suppose for simplicity that the rows of A are linearly independent. Then the dimension of $\ker A$ is $n - m$, and the system $(Z^\top HZ)v = -Z^\top(H\overline{x}+c)$ has the size $(n-m) \times (n-m)$, while the system (10.3) has the size $(n+m) \times (n+m)$.

If for instance $n \approx 2m$, then $n + m \approx 3(n - m)$, so that the "Lagrange system" has about three times as many equations (and unknowns) as the "nullspace system".

If on the other hand, $m \ll n$, then both system are of about the same size, but then the Lagrange method has the advantage that we don't need the matrix Z. Moreover, the *sparsity* (that is a large number of zeroes) in H and A is used more efficiently in the Lagrange method, since Z is typically dense even if A is sparse.

So which of the two methods is best depends on the problem at hand, and it is safest to master them both!

Chapter 11

Least-squares problems

Consider the system of equations $Ax = b$, where $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$ and $x \in \mathbb{R}^n$. In many applications, $b \notin \operatorname{ran} A$, and then the system does not have a solution. Nevertheless, one might want to find an $x \in \mathbb{R}^n$ which is "closest" in satisfying $Ax = b$. A natural measure used to determine if x "almost" satisfies $Ax = b$ is to see how small the error $\|Ax - b\|^2$ is. Hence we arrive at the following *least-squares problem*:

$$\left\{\begin{array}{ll} \text{minimize} & \frac{1}{2}(Ax-b)^\top(Ax-b), \\ \text{subject to} & x \in \mathbb{R}^n. \end{array}\right. \tag{11.1}$$

Thus we want to minimize the square of the length of the "error vector" $Ax - b$. The factor $\frac{1}{2}$ is introduced in order to simplify some of the expressions that occur in this chapter.

11.1. A model fitting example

Let us see an instance where a problem of the type discussed above appears naturally.

Suppose that s is a quantity that depends on the variable t, that is, $s = g(t)$, where g is not entirely known. Suppose moreover that based on some given measured data, we want to estimate the function g. The measured data consists of m given points:

$$(t_1, s_1), \dots, \ (t_m, s_m),$$

where s_i is a measurement of s for $t = t_i$, that is, s_i is the measurement of $g(t_i)$. A common approach to estimate g is then to do a parameterization of the form

$$g(t) \approx \alpha_1\varphi_1(t) + \cdots + \alpha_n\varphi_n(t) = \sum_{j=1}^{n} \alpha_j\varphi_j(t), \tag{11.2}$$

where the φ_j are given "basis functions", and the α_j are unknown coefficients. The basis functions can be for instance, polynomials $\varphi_j(t) = t^{j-1}$, or trigonometric functions $\varphi_j(t) = \sin\frac{2\pi jt}{T}$, or something else, depending on the context. Ideally, we would like to choose the coefficients so that

$$\sum_{j=1}^{n} \alpha_j\varphi_j(t_i) = s_i \quad \text{for all } i \in \{1, \dots, m\}. \tag{11.3}$$

However, in practice, it typically happens that

(1) the number of measurements is larger than the number of coefficients α_j, that is, $m > n$,

(2) the approximation in (11.2) is not exact, and

(3) the measurements of s contain measurement noise.

As a consequence of these, we cannot solve the system (11.3). Instead, we seek coefficients α_j that make the following "error" as small as possible:

$$\sum_{i=1}^{m}\left(\sum_{j=1}^{n}\alpha_j\varphi_j(t_i)-s_i\right)^2. \tag{11.4}$$

By defining

$$A=\begin{bmatrix}\varphi_1(t_1) & \dots & \varphi_n(t_1)\\ \vdots & & \vdots\\ \varphi_1(t_m) & \dots & \varphi_n(t_m)\end{bmatrix},\quad b=\begin{bmatrix}s_1\\ \vdots\\ s_m\end{bmatrix},\quad x=\begin{bmatrix}\alpha_1\\ \vdots\\ \alpha_n\end{bmatrix},$$

the system (11.3) can be written as $Ax=b$, while the problem of minimizing (11.4) can be written in the form (11.1) above.

11.2. Existence of optimal solutions

Consider the least-squares problem (11.1). Let f be the objective function, that is,

$$f(x)=\frac{1}{2}(Ax-b)^\top(Ax-b)=\frac{1}{2}x^\top A^\top Ax-b^\top Ax+\frac{1}{2}b^\top b.$$

We see that f is a quadratic function of the form (9.1), namely,

$$f(x)=\frac{1}{2}x^\top Hx+c^\top x+c_0$$

with $H=A^\top A$ and $c=-A^\top b$.

A nice fact associated with the least-squares problem (11.1) is that there always exists at least one minimizer $\widehat{x}$ of f. Indeed, according to Theorem 9.13, the quadratic function f has at least one minimizer if H is positive semi-definite and $c\in\operatorname{ran}H$. In our special case above, $H=A^\top A$ is indeed positive semi-definite and $c=-A^\top b=A^\top(-b)\in\operatorname{ran}A^\top=\operatorname{ran}(A^\top A)=\operatorname{ran}H$. (From Exercise 24.2, it follows that $\operatorname{ran}A^\top=\operatorname{ran}A^\top A$.)

11.3. Normal equations

By Theorem 9.13, x is a minimizer for f iff $Hx=-c$, and in our special case, this becomes:

$$A^\top Ax=A^\top b. \tag{11.5}$$

This system is called the set of *normal equations* for the least-squares problem (11.1).

Since we have already seen that the least-squares problem (11.1) always has at least one solution, it follows that the system (11.5) always has at least one solution.

Even if it so happens that there are infinitely many solutions to the normal equations, Ax is the same for any solution x. Indeed, if x_1 and x_2 are two solutions, that is, if $A^\top Ax_1=A^\top b=A^\top Ax_2$, then $A^\top A(x_1-x_2)=0$. This means that with $y:=A(x_1-x_2)$,

$$y^\top y=(x_1-x_2)^\top\underbrace{A^\top A(x_1-x_2)}_{=0}=(x_1-x_2)^\top 0=0,$$

and so $y=0$. Thus $Ax_1=Ax_2$.

11.4. Geometric interpretation

There is a natural geometric interpretation of the least-squares problem (11.1) and of the corresponding normal equations (11.5).

The problem (11.1) can be interpreted as the problem of deciding the point in the subspace $\operatorname{ran} A$ that lies closest to the point $b \in \mathbb{R}^m$, since it can be written equivalently as

$$\left\{ \begin{array}{ll} \text{minimize} & \frac{1}{2}\|y-b\|^2, \\ \text{subject to} & y \in \operatorname{ran} A. \end{array} \right. \tag{11.6}$$

See Figure 1.

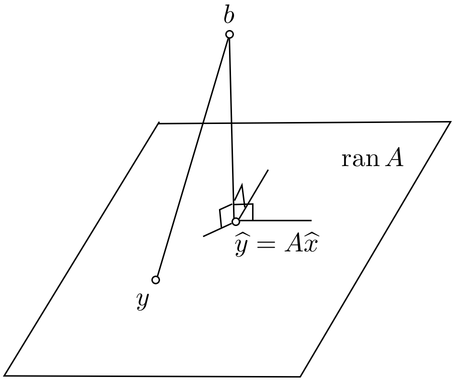


Figure 1. Geometric interpretation of the least squares problem (11.1) and of the corresponding normal equations (11.5).

The normal equations (11.5) can be written as $A^\top(b - Ax) = 0$, which is equivalent to saying that $b - Ax \in \ker A^\top$, which in turn is equivalent to $b - Ax \in (\operatorname{ran} A)^\perp$. Thus the normal equations say that the "error vector" $b - Ax$ should be *orthogonal* to $\operatorname{ran} A$. See Figure 1.

So the optimal solution $\widehat{y}$ $(= A\widehat{x})$ to the problem (11.6) is determined by $\widehat{y} \in \operatorname{ran} A$ and $b - \widehat{y} \in (\operatorname{ran} A)^\perp$, that is $b - \widehat{y}$ is orthogonal to $\operatorname{ran} A$. As we have already seen above (at the end of the previous section), the point $\widehat{y} = A\widehat{x}$ is unique, even if $\widehat{x}$ is not.

11.5. Case 1°: A has independent columns

If A has linearly independent columns, then $A^\top A$ is invertible. Thus the normal equations (11.5) have a *unique* solution $\widehat{x}$.

11.6. Case 2°: A has dependent columns

If A has linearly dependent columns, then $A^\top A$ is *not* invertible, and then there are infinitely many solutions to the normal equations (11.5). A common way of selecting *one* amongst these solutions is to take one which is the "shortest", that is, one with the least norm.

Let $\overline{x}$ be *a* solution to (11.5). Then any x is another solution to (11.5) iff $Ax = A\overline{x}$. We have already shown the 'only if' part earlier. The 'if' part can be checked as follows: if $Ax = A\overline{x}$, then

$$A^\top Ax = A^\top(Ax) = A^\top(A\overline{x}) = A^\top b.$$

Thus the problem of determining the least-norm solution to the normal equations can be written as:

$$\left\{ \begin{array}{ll} \text{minimize} & \frac{1}{2}\|x\|^2, \\ \text{subject to} & Ax = A\overline{x}. \end{array} \right. \tag{11.7}$$

This is a problem of the form (10.1), with $H = I$, $c = 0$, $c_0 = 0$, $A = A$ and $b = A\overline{x}$.

According to Theorem 10.5, $\widehat{x}$ is an optimal solution to the problem (11.7) iff there exists a u such that

$$\begin{aligned} I\widehat{x} - A^\top u &= 0, && (11.8)\\ A\widehat{x} &= A\overline{x}. && (11.9) \end{aligned}$$

The first equation (11.8) says that $\widehat{x} = A^\top u$, and if we substitute this in the second equation (11.9) above, we obtain $AA^\top u = A\overline{x}$. Since $A\overline{x} \in \operatorname{ran} A = \operatorname{ran}(AA^\top)$, the system $AA^\top u = A\overline{x}$ always has at least one solution $\widehat{u}$. And then $\widehat{x} := A^\top \widehat{u}$ clearly satisfies (11.8) and (11.9). Thus $\widehat{x}$ found in this manner is an optimal solution to the problem (11.7).

Even if it so happens that there are infinitely many solutions u to $AA^\top u = A\overline{x}$, it turns out that the vector $A^\top u$ is the same. Indeed, if u_1, u_2 are such that $AA^\top u_1 = A\overline{x} = AA^\top u_2$, then $AA^\top(u_1 - u_2) = 0$. So with $v := A^\top(u_1 - u_2)$,

$$v^\top v = (u_1 - u_2)^\top \underbrace{AA^\top(u_1 - u_2)}_{=0} = (u_1 - u_2)^\top 0 = 0,$$

and so $v = 0$. Thus $A^\top u_1 = A^\top u_2$.

This implies that $\widehat{x} = A^\top \widehat{u}$ is the *unique* optimal solution to the problem (11.7), even if the system $AA^\top u = A\overline{x}$ does not have a unique solution.

11.7. Optimal solution in terms of the pseudo inverse

Suppose we want to determine the least-norm solution to the least-squares problem as described in the previous section. Then one can proceed as follows. Suppose that A has the *singular value decomposition*

$$A = USV^\top,$$

where the $m \times r$ matrix U has orthogonal columns and the $n \times r$ matrix V has orthogonal rows, that is, $U^\top U = I$ and $V^\top V = I$, while the $r \times r$ matrix S is diagonal with strictly positive diagonal elements. (See Section 11.9 for a proof of this.) Since $A^\top = VSU^\top$, we obtain

$$A^\top A = VS^2V^\top \text{ and } AA^\top = US^2U^\top.$$

So the normal equations (11.5) take the form $VS^2V^\top x = VSU^\top b$, which is equivalent to the system $V^\top x = S^{-1}U^\top b$. If $\overline{x}$ is *a* solution to $V^\top x = S^{-1}U^\top b$, then the system $AA^\top u = A\overline{x}$ takes the form

$$US^2U^\top u = USV^\top \overline{x},$$

which is equivalent to $SU^\top u = V^\top \overline{x}$. If $\widehat{u}$ is *a* solution to the system $SU^\top u = V^\top \overline{x}$, then the least-norm solution to the least-squares problem (11.1) is thus given by

$$\widehat{x} = A^\top \widehat{u} = VSU^\top \widehat{u} = VV^\top \overline{x} = VS^{-1}U^\top b = A^+ b,$$

where the matrix $A^+ := VS^{-1}U^\top$ is called the *pseudo inverse* of A.

Exercise 11.1. Verify that $AA^+A = A$ and $A^+AA^+ = A^+$.

Exercise 11.2. Let $A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$ and $b = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix}$, where b_1 and b_2 are given numbers.

(1) Determine *all* optimal solutions x to the following problem:

$$(P1): \left\{ \begin{array}{ll} \text{minimize} & \|Ax - b\|^2 \\ \text{subject to} & x \in \mathbb{R}^2. \end{array} \right.$$

(2) Let $X(b)$ be the set of all optimal solutions to the problem $(P1)$ above. Determine the unique optimal solution to the following problem:

$$(P2):\left\{\begin{array}{ll}\text{minimize} & \|x\|^2\\ \text{subject to} & x\in X(b).\end{array}\right.$$

(3) Let $\widehat{x}(b)$ denote the optimal solution to the problem $(P2)$ above. Show that $\widehat{x}(b)=A^{+}b$ for a certain matrix A^{+}. Find A^{+}.

(4) Let $\epsilon>0$. Determine the unique optimal solution to the following problem:

$$(P3):\left\{\begin{array}{ll}\text{minimize} & \|Ax-b\|^2+\epsilon\|x\|^2\\ \text{subject to} & x\in\mathbb{R}^2.\end{array}\right.$$

(5) Let $\widetilde{x}(b,\epsilon)$ denote the optimal solution to the problem $(P3)$ above. Show that $\widetilde{x}(b,\epsilon)=\widetilde{A}_\epsilon$ for a certain matrix $\widetilde{A}_\epsilon$ (which has entries depending on ϵ). Prove that as $\epsilon\to 0$, each entry of $\widetilde{A}_\epsilon$ goes to the corresponding entry of A^{+}.

Exercise 11.3. Let U,V be subsets of $\mathbb{R}^4$, given by $U=\{u\in\mathbb{R}^4 : Ru=p\}$ and $V=\{v\in\mathbb{R}^4 : Sv=q\}$, where

$$R=\begin{bmatrix}1&0&0&1\\0&1&0&1\\0&0&1&1\end{bmatrix},\ S=\begin{bmatrix}1&1&0&0\\1&0&1&0\\1&0&0&1\end{bmatrix},\ p=\begin{bmatrix}1\\1\\1\end{bmatrix},\ q=\begin{bmatrix}2\\2\\2\end{bmatrix}.$$

Determine the distance $d(U,V)$ between U and V, that is, the smallest possible distance between $u\in U$ and $v\in V$. Also find points $\widehat{u}\in U$ and $\widehat{v}\in V$ for which the distance between $\widehat{u}$ and $\widehat{v}$ is $d(U,V)$.

Exercise 11.4. Consider the optimization problem in the variable $x\in\mathbb{R}^2$:

$$\text{minimize } \frac{1}{2}(Ax-b)^\top(Ax-b),$$

where $A=\begin{bmatrix}2&-1\\-1&2\\1&1\end{bmatrix}$ and $b=\begin{bmatrix}2\\1\\4\end{bmatrix}$. Find an optimal solution.

Exercise 11.5. A civil engineer is assigned the task of determining the heights above sea level of three hills, H_1, H_2, H_3. He stands at sea level and measures the heights (in meters) of H_1, H_2, H_3 as 1236, 1941, 2417, respectively. Then to check his work, he climbs hill H_1 and measures the height of H_2 above H_1 as 711m, and the height of H_3 above H_1 as 1177m. Noting that these latter measurements are not consistent with those made at sea level, he climbs hill H_2, and measures the height of H_3 to be 474m above H_2. Again he notes the inconsistency of this measurement with those made earlier. As he drives back to his office, he suddenly remembers his days as a student in the Optimization course, and he decides to solve a quadratic optimization problem associated with this problem by considering the problem of mimimizing the least squares error associated with the measurements. Compute the optimal solution for him so that he can keep both hands on the steering wheel.

11.8. Summary

The least-squares problem is the problem of minimizing

$$\frac{1}{2}\|Ax-b\|^2=\frac{1}{2}(Ax-b)^\top(Ax-b).$$

A vector $\widehat{x}$ is an optimal solution to the least-squares problem iff $\widehat{x}$ is a solution to the normal equations

$$A^\top Ax=A^\top b,$$

which always has at least one solution.

If the columns of A are linearly independent, then the normal equations have a unique solution.

If the columns of A are linearly dependent, then the normal equations have infinitely many solutions. Amongst these, there is a unique one with least norm. The least norm solution $\widehat{x}$ is given by $\widehat{x}=A^\top\widehat{u}$, where $\widehat{u}$ is a solution to $AA^\top u=A\overline{x}$. Here $\overline{x}$ is an arbitrary solution to the normal equations.

11.9. Appendix: singular value decomposition

In this appendix, we will prove the following result.

Theorem 11.6 (Singular value decomposition)**.** *Let $A \in \mathbb{R}^{m\times n}$. Then there exists an integer r and there exist matrices $U \in \mathbb{R}^{m\times r}$, $S \in \mathbb{R}^{r\times r}$ and $V \in \mathbb{R}^{n\times r}$ such that*

(1) $A = USV^\top$,

(2) $U^\top U = I$ *and* $V^\top V = I$,

(3) *S is a diagonal matrix with positive diagonal entries.*

Definition 11.7. Let $A \in \mathbb{R}^{m\times n}$ and let $\lambda_1, \dots, \lambda_n$ be the eigenvalues of $A^\top A$. Then $\sqrt{\lambda_1}, \dots, \sqrt{\lambda_n}$ are called the *singular values of* A.

Proof of Theorem 11.6. Suppose that $\sigma_1, \dots, \sigma_r$ are the nonzero singular values of A. This means that $\sigma_j = 0$ for $j > r$. By the definition of singular values, $\sigma_1^2, \dots, \sigma_n^2$ are the eigenvalues of $A^\top A$. Let us denote the corresponding basis[1] of orthogonal eigenvectors of $A^\top A$ by $v_1, \dots, v_n$, that is, $A^\top A v_j = \sigma_j^2 v_j$, $j = 1, \dots, n$.

The vectors $w_j := \frac{1}{\sigma_j} A v_j$, $j = 1, \dots, r$, form an orthonormal system. Indeed, we have

$$(Av_j, Av_k) = (A^\top A v_j, v_k) = (\sigma_j^2 v_j, v_k) = \sigma_j^2 (v_j, v_k) = \begin{cases} 0 & \text{if } j \neq k, \\ \sigma_j^2 & \text{if } j = k, \end{cases}$$

since $v_1, \dots, v_r$ is an orthonormal system. This proves the claim.

If $j \in \{1, \dots, r\}$, then $Av_j = \sigma_j w_j = \sigma_j w_j v_j^\top v_j = \displaystyle\sum_{k=1}^{r} \sigma_k w_k v_k^\top v_j = \left(\sum_{k=1}^{r} \sigma_k w_k v_k^\top \right) v_j$.

On the other hand, if $j \in \{r+1, \dots, n\}$, then $Av_j = 0 = \displaystyle\sum_{k=1}^{r} \sigma_k w_k v_k^\top v_j = \left(\sum_{k=1}^{r} \sigma_k w_k v_k^\top \right) v_j$.

So for all v_j, $j = 1, \dots, n$, we have $Av_j = \left(\displaystyle\sum_{k=1}^{r} \sigma_k w_k v_k^\top \right) v_j$.

Since $v_1, \dots, v_n$ forms a basis for $\mathbb{R}^n$, it follows that for all $x \in \mathbb{R}^n$, we have

$$Ax = \left(\sum_{k=1}^{r} \sigma_k w_k v_k^\top \right) x,$$

that is, $A = \displaystyle\sum_{k=1}^{r} \sigma_k w_k v_k^\top = USV^\top$, where

$$\begin{aligned} U &:= \begin{bmatrix} w_1 & \dots & w_r \end{bmatrix} \in \mathbb{R}^{m\times r}, \\ S &:= \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \in \mathbb{R}^{r\times r}, \\ V &:= \begin{bmatrix} v_1 & \dots & v_r \end{bmatrix} \in \mathbb{R}^{n\times r}. \end{aligned}$$

The relations $U^\top U = I$ and $V^\top V = I$ follow from the orthonormality of the systems $w_1, \dots, w_r$ and $v_1, \dots, v_r$, respectively. □

[1] The existence of such a basis follows from the Spectral Theorem applied to the symmetric matrix $A^\top A$.

Part 3

Nonlinear optimization

Chapter 12

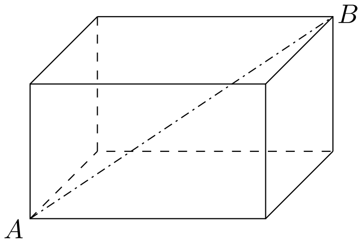
Introduction

We will now consider the following very general optimization problem in $\mathbb{R}^n$:

$$\begin{aligned} \text{minimize} \quad & f(x), \\ \text{subject to} \quad & x \in \mathcal{F}, \end{aligned}$$

where $x \in \mathbb{R}^n$ is a vector of variables, $\mathcal{F}$ is a given subset of $\mathbb{R}^n$, and f is a real-valued function that is defined (at least) on the set $\mathcal{F}$. The function f is called the *objective function* and $\mathcal{F}$ is called the *feasible set.*

Example 12.1. Consider for example, that we want to determine what the length, height and breadth of a box should be so that the total surface area of the box's six faces is as small as possible, but so that the box's volume is at least 100 cubic decimeters and the box's spatial diagonal is at least 9 decimeters.



If we denote the length, height and breadth of the box by x_1, x_2 and x_3, respectively, then the problem can be formulated as follows:

$$\begin{aligned} \text{minimize} \quad & 2x_1x_2 + 2x_2x_3 + 2x_3x_1, \\ \text{subject to} \quad & x_1x_2x_3 - 100 \geq 0, \\ & x_1^2 + x_2^2 + x_3^2 - 9^2 \geq 0, \\ & x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0. \end{aligned}$$

Here $x = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}^\top$, the objective function f is given by $f(x) = 2x_1x_2 + 2x_2x_3 + 2x_3x_1$, and the feasible set is

$$\mathcal{F} = \left\{ x \in \mathbb{R}^3 : \begin{array}{c} x_1x_2x_3 - 100 \geq 0, \\ x_1^2 + x_2^2 + x_3^2 - 9^2 \geq 0, \\ x_1 \geq 0,\ x_2 \geq 0,\ x_3 \geq 0 \end{array} \right\}.$$

$\Diamond$

The example above falls under a special subclass of optimization problems called *nonlinear optimization*[1]. In this class of problems, we will assume that the objective function f is continuously differentiable, and the feasible set is described by a set of constraints of the type

$$\begin{aligned} g_i(x) &\leq 0 \text{ (inequality constraints) and/or} \\ h_i(x) &= 0 \text{ (equality constraints),} \end{aligned}$$

where g_i and h_i are given continuously differentiable functions. The nonlinear optimization problem thus has the following form:

$$\begin{array}{rl} \text{minimize} & f(x), \\ \text{subject to} & g_i(x) \leq 0, \;\; i = 1, \ldots, m_1, \\ & h_j(x) = 0, \;\; j = 1, \ldots, m_2, \end{array}$$

where at least one of the functions f, $g_1, \cdots g_{m_1}$, $h_1, \ldots, h_{m_2}$ are nonlinear. (Otherwise we would just have a *linear* programming problem, which we have already learnt to solve in Part I of this course!) The feasible set is given by

$$\left\{ x \in \mathbb{R}^n : \begin{array}{l} g_i(x) \leq 0 \text{ for } i = 1, \ldots, m_1, \text{ and} \\ h_j(x) = 0 \text{ for } j = 1, \ldots, m_2 \end{array} \right\}.$$

It is not unusual that one has the special case when there are no constraints and that $\mathcal{F} = \mathbb{R}^n$. In this case, we say that it is a nonlinear optimization problem without constraints, or that it is an *unconstrained* nonlinear optimization problem. Otherwise it is called a *constrained* nonlinear optimization problem.

[1] or *nonlinear programming*

Chapter 13

The one variable case

In this chapter f will be a real-valued function of a real variable x, that is, $f : \mathbb{R} \to \mathbb{R}$. We will assume that the (first) derivative f' and the second derivative f'' exist and are continuous on $\mathbb{R}$:

$$\boxed{f' \text{ and } f'' \text{ exist and are continuous.}}$$

Definition 13.1. A point $\widehat{x} \in \mathbb{R}$ is called a *local minimizer* of f if there exists a $\delta > 0$ such that for all $x \in \mathbb{R}$ that satisfy[1] $|x - \widehat{x}| < \delta$, we have $f(\widehat{x}) \le f(x)$. A point $\widehat{x} \in \mathbb{R}$ is called a *global minimizer* of f if for all $x \in \mathbb{R}$, $f(\widehat{x}) \le f(x)$. See Figure 1.

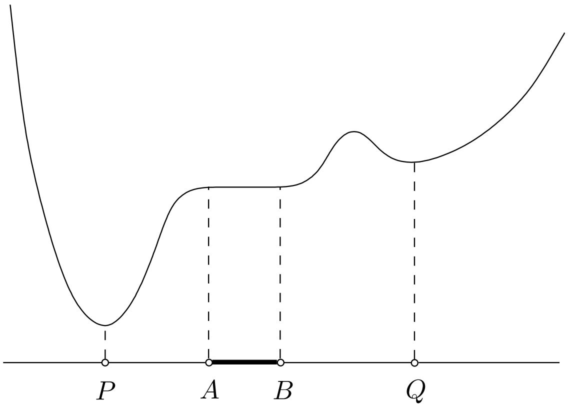


Figure 1. The point P is a global minimizer. The point Q and all points in the interior of the line segment AB are all local minimizers.

It is obvious that every global minimizer is also a local minimizer, but it can happen (for non-convex functions) that there exist local minimizers which are not global minimizers.

Recall that the derivative of f at a point $\widehat{x}$ is by definition

$$f'(\widehat{x}) = \lim_{x \to \widehat{x}} \frac{f(x) - f(\widehat{x})}{x - \widehat{x}},$$

which means that for every $\epsilon > 0$, there exists a $\delta > 0$ such that for all $x \in \mathbb{R}$ with $0 < |x - \widehat{x}| < \delta$, there holds that

$$\left| \frac{f(x) - f(\widehat{x})}{x - \widehat{x}} - f'(\widehat{x}) \right| < \epsilon,$$

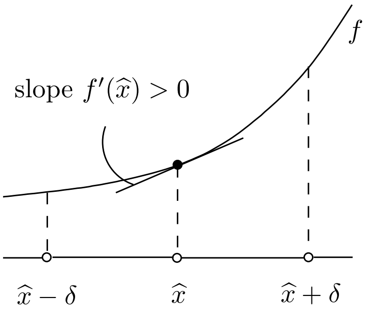
or equivalently that $-\epsilon < \dfrac{f(x) - f(\widehat{x})}{x - \widehat{x}} - f'(\widehat{x}) < \epsilon$.

[1] equivalently, $\widehat{x} - \delta < x < \widehat{x} + \delta$

Lemma 13.2. *Let $f'(\widehat{x}) > 0$. Then there exists a $\delta > 0$ such that:*

(1) *for all $x \in (\widehat{x}, \widehat{x} + \delta)$, $f(\widehat{x}) < f(x)$,*

(2) *for all $x \in (\widehat{x} - \delta, \widehat{x})$, $f(x) < f(\widehat{x})$.*



Proof. Let $\epsilon = \frac{1}{2} f'(x) > 0$. Then there is a $\delta > 0$ such that for all $x \in \mathbb{R}$ satisfying $0 < |x - \widehat{x}| < \delta$,

$$-\epsilon < \frac{f(x) - f(\widehat{x})}{x - \widehat{x}} - f'(\widehat{x}) < \epsilon,$$

and in particular,

$$\frac{1}{2} f'(x) < \frac{f(x) - f(\widehat{x})}{x - \widehat{x}}. \tag{13.1}$$

For all $0 < x - \widehat{x} < \delta$, (13.1) gives $0 < \frac{1}{2} f'(x)(x - \widehat{x}) < f(x) - f(\widehat{x})$, and so $f(\widehat{x}) < f(x)$.

On the other hand, if $-\delta < x - \widehat{x} < 0$, then from (13.1), $0 > \frac{1}{2} f'(x)(x - \widehat{x}) > f(x) - f(\widehat{x})$, and so $f(x) < f(\widehat{x})$. □

By replacing f by $-f$ in the above result, we obtain the following:

Lemma 13.3. *Let $f'(\widehat{x}) < 0$. Then there exists a $\delta > 0$ such that:*

(1) *for all $x \in (\widehat{x}, \widehat{x} + \delta)$, $f(\widehat{x}) > f(x)$,*

(2) *for all $x \in (\widehat{x} - \delta, \widehat{x})$, $f(x) > f(\widehat{x})$.*

Theorem 13.4. *If $\widehat{x}$ is a local minimizer of f, then $f'(\widehat{x}) = 0$.*

Proof. By the previous two lemmas, we know that if $f'(\widehat{x}) > 0$, then $\widehat{x}$ is not a local minimizer of f (because for example $f(\widehat{x} - \frac{\delta}{2^n}) < f(\widehat{x})$ for all $n \in \mathbb{N}$ and some $\delta > 0$), and also if $f'(\widehat{x}) < 0$, then $\widehat{x}$ is not a local minimizer of f. So the only remaining case is that $f'(\widehat{x}) = 0$. □

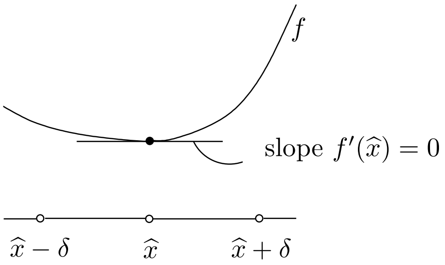
Recall *Taylor's formula*: if $x, \widehat{x} \in \mathbb{R}$ then

$$f(x) = f(\widehat{x}) + f'(\widehat{x})(x - \widehat{x}) + \frac{1}{2} f''(\xi)(x - \widehat{x})^2, \tag{13.2}$$

for some ξ between x and $\widehat{x}$, that is, $\xi = \widehat{x} + \theta \cdot (x - \widehat{x})$ for some $\theta \in (0, 1)$. The exact value of θ depends on what f is and what x and $\widehat{x}$ are. But we will not need to know this. It suffices to know that $\theta \in (0, 1)$ and that ξ lies between x and $\widehat{x}$.

In Theorem 13.4 we learnt that the vanishing of the derivative at a point is a *necessary* condition for that point to be a local minimizer. Now we will see that this condition is also *sufficient* if in addition we also have that the second derivative at that point is positive. In fact we then have a "strict" local minimum at that point.

Lemma 13.5. *If $f'(\widehat{x}) = 0$ and $f''(\widehat{x}) > 0$, then there exists a $\delta > 0$ such that for all $x \in \mathbb{R}$ such that $0 < |x - \widehat{x}| < \delta$, we have $f(x) > f(\widehat{x})$.*



Proof. Since f'' is continuous, and since $f''(\widehat{x}) > 0$, there exists a $\delta > 0$ such that $f''(x) > 0$ for all x in the interval $(\widehat{x} - \delta, \widehat{x} + \delta)$. (Why?)

But for $x \in (\widehat{x} - \delta, \widehat{x} + \delta)$, we have by Taylor's formula that

$$f(x) = f(\widehat{x}) + f'(\widehat{x})(x - \widehat{x}) + \frac{1}{2} f''(\xi)(x - \widehat{x})^2 = f(\widehat{x}) + 0 + \frac{1}{2} \underbrace{f''(\xi)}_{>0}(x - \widehat{x})^2 > f(\widehat{x})$$

if $\widehat{x} \neq x$. (Note that $\xi = \widehat{x} + \theta \cdot (x - \widehat{x})$ lies between x and $\widehat{x}$ and so it lies in the interval $(\widehat{x} - \delta, \widehat{x} + \delta)$; but we know that on this interval f'' takes positive values.) □

By replacing f by $-f$ in the above result, we obtain the following.

Lemma 13.6. *If $f'(\widehat{x}) = 0$ and $f''(\widehat{x}) < 0$, then there exists a $\delta > 0$ such that for all $x \in \mathbb{R}$ such that $0 < |x - \widehat{x}| < \delta$, we have $f(x) < f(\widehat{x})$.*

Theorem 13.7.

(1) *A necessary (but not sufficient) condition for $\widehat{x}$ to be a local minimizer of f is that $f'(\widehat{x}) = 0$ and $f''(\widehat{x}) \geq 0$.*

(2) *A sufficient (but not necessary) condition for $\widehat{x}$ to be a local minimizer of f is that $f'(\widehat{x}) = 0$ and $f''(\widehat{x}) > 0$.*

Proof. (1) If $\widehat{x}$ is a local minimizer, then by Theorem 13.4, $f'(\widehat{x}) = 0$. Also, Lemma 13.6 shows that if $f''(\widehat{x}) < 0$, then $\widehat{x}$ cannot be a local minimizer (for example because $f(\widehat{x} + \frac{\delta}{2^n}) < f(\widehat{x})$ for all $n \in \mathbb{N}$ and some $\delta > 0$). So $f''(\widehat{x}) \geq 0$.

The fact that this condition is not sufficient can be seen by considering the example $f(x) = x^3$ ($x \in \mathbb{R}$) and $\widehat{x} = 0$. Then $f'(\widehat{x}) = f'(0) = 3x^2|_{x=0} = 0$ and $f''(\widehat{x}) = f''(0) = 6x|_{x=0} = 0 \geq 0$, but $\widehat{x} = 0$ is not a local minimizer. See Figure 2.

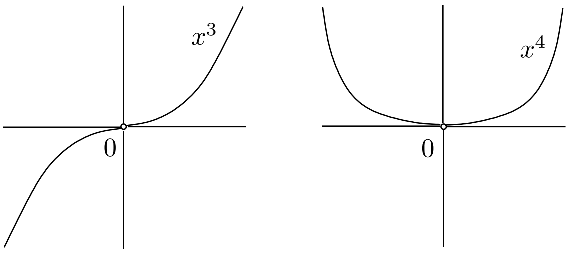


Figure 2. 0 is not a local minimizer for $x \mapsto x^3$, but is a global minimizer for $x \mapsto x^4$.

(2) Now if $f'(\widehat{x}) = 0$ and $f''(\widehat{x}) > 0$, then from Lemma 13.5, it follows that $\widehat{x}$ is a local minimizer of f.

The fact that this condition is not necessary can be seen by considering the example $f(x) = x^4$ $(x \in \mathbb{R})$ and $\widehat{x} = 0$. Then $\widehat{x} = 0$ is a global minimizer of f, and although there holds that $f'(\widehat{x}) = f'(0) = 4x^3|_{x=0} = 0$, we have that $f''(\widehat{x}) = f''(0) = 12x^2|_{x=0} = 0$ and so $f''(\widehat{x})$ is not positive in this case. See Figure 2. □

Exercise 13.8. Let $f : \mathbb{R} \to \mathbb{R}$ be given by $f(x) = e^{x^2}$. Show that 0 is a global minimizer of f.

Exercise 13.9. Consider $g : \mathbb{R} \to \mathbb{R}$ given by $g(x) = 3x^4 - 4x^3 + 1$. Find all local minimizers of g.

Chapter 14

The multivariable case

In this chapter, we will consider the problem of minimizing a given multivariable function *without* any constraints, that is, the problem of the form

$$\begin{array}{rl} \text{minimize} & f(x), \\ \text{subject to} & x \in \mathbb{R}^n, \end{array}$$

where f is a given real-valued function on $\mathbb{R}^n$. We shall assume henceforth that

$$\boxed{f \text{ is twice continuously differentiable.}}$$

To say that f is twice continuously differentiable means that the n partial derivatives of f, namely the functions

$$\frac{\partial f}{\partial x_j} \quad (j = 1, \dots, n)$$

all exist, and are continuous in $\mathbb{R}^n$, and moreover their n^2 partial derivatives, namely

$$\frac{\partial^2 f}{\partial x_i \partial x_j} \quad (i, j \in \{1, \dots, n\})$$

all exist and are continuous in $\mathbb{R}^n$.

Then one can define the *gradient of f at $x \in \mathbb{R}^n$* to be the following row vector:

$$\nabla f(x) = \left[\begin{array}{ccc} \frac{\partial f}{\partial x_1}(x) & \dots & \frac{\partial f}{\partial x_n}(x) \end{array} \right].$$

Furthermore, the *Hessian*[1] *of f at $x \in \mathbb{R}^n$* is defined as the $n \times n$ symmetric matrix $F(x)$, which has

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x)$$

as the entry in the ith row and jth column, that is,

$$F(x) = \left[\begin{array}{ccc} \frac{\partial^2 f}{\partial x_1 \partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_1 \partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1}(x) & \dots & \frac{\partial^2 f}{\partial x_n \partial x_n}(x) \end{array} \right].$$

[1] This was introduced in the 19th century by the German mathematician Ludwig Otto Hesse and it was later named after him. (Hesse himself had used the term "functional determinants".)

Note that $F(x)$ is symmetric, since by our assumption that the partial derivatives of f are continuous, we have that

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x).$$

Definition 14.1. A point $\widehat{x} \in \mathbb{R}^n$ is called a *local minimizer* of f if there exists a $\delta > 0$ such that for all $x \in \mathbb{R}^n$ that satisfy $\|x - \widehat{x}\| < \delta$, we have $f(\widehat{x}) \leq f(x)$. A point $\widehat{x} \in \mathbb{R}^n$ is called a *global minimizer* of f if for all $x \in \mathbb{R}^n$, $f(\widehat{x}) \leq f(x)$.

It is obvious that every global minimizer is also a local minimizer, but it can happen (for non-convex functions) that there exist local minimizers which are not global minimizers.

Let $\widehat{x} \in \mathbb{R}^n$ be a given point. Suppose we want to determine whether or not $\widehat{x}$ is a minimizer for f. In order to do so, let us see how the objective function changes along a line passing through $\widehat{x}$. Thus let us take a nonzero vector $d \in \mathbb{R}^n$, and let

$$x(t) = \widehat{x} + td, \quad t \in \mathbb{R},$$

This defines a line in $\mathbb{R}^n$ (in parametric form) passing through $\widehat{x}$ and having direction d. In particular, $x(0) = \widehat{x}$. See Figure 1.

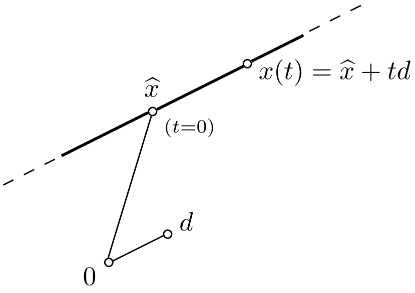


Figure 1. The line passing through $\widehat{x}$ having direction d.

We will study the objective function f along this line, and so we define the function φ of *one* variable by

$$\varphi(t) = f(x(t)) = f(\widehat{x} + td) \quad (t \in \mathbb{R}).$$

Since f is twice differentiable on $\mathbb{R}$, so is φ. Indeed, by the chain rule, one has that

$$\varphi'(t) = \nabla f(x(t))d \quad \text{and} \quad \varphi''(t) = d^\top F(x(t))d. \tag{14.1}$$

In particular,

$$\varphi'(0) = \nabla f(\widehat{x})d \quad \text{and} \quad \varphi''(0) = d^\top F(\widehat{x})d. \tag{14.2}$$

The number $\varphi'(0) = \nabla f(\widehat{x})d$ is called the *directional derivative of f at $\widehat{x}$ in the direction d.*

Lemma 14.2. *If $\nabla f(\widehat{x})d < 0$, then there exists a $\delta > 0$ such that for all $t \in (0, \delta)$, $f(\widehat{x}+td) < f(\widehat{x})$.*

Proof. This follows immediately by an application of Lemma 13.3 to the function φ. □

Lemma 14.3. *If $\widehat{x}$ is a local minimizer of f, then for all $d \in \mathbb{R}^n$, $t = 0$ is a local minimizer of the function φ given by*

$$\varphi(t) = f(x(t)) = f(\widehat{x} + td) \quad (t \in \mathbb{R}). \tag{14.3}$$

Proof. Suppose that $\delta > 0$ is such that for all $x \in \mathbb{R}^n$ satisfying $\|x-\widehat{x}\| < \delta$, we have $f(x) \geq f(\widehat{x})$. Define $\epsilon = \delta/\|d\|$. Then for all $t \in \mathbb{R}$ such that $|t| < \epsilon$, we have that

$$\|x(t) - \widehat{x}\| = \|\widehat{x} + td - \widehat{x}\| = \|td\| = |t|\|d\| < \epsilon\|d\| = \frac{\delta}{\|d\|}\|d\| = \delta.$$

Consequently, for such t's, $\varphi(t) = f(x(t)) \geq f(\widehat{x}) = \varphi(0)$. □

Observe that the converse to the above result does not hold[2]! Even if for every vector $d \in \mathbb{R}^n$, there holds that $t = 0$ is a local minimizer for the function φ defined via (14.3), it can happen that $\widehat{x}$ is not a local minimizer of f. This is illustrated in the following example.

Example 14.4. Let $n = 2$. Consider the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x) = (x_2 - x_1^2)(x_2 - 3x_1^2) \quad (x \in \mathbb{R}^2).$$

Take $\widehat{x} = 0$.

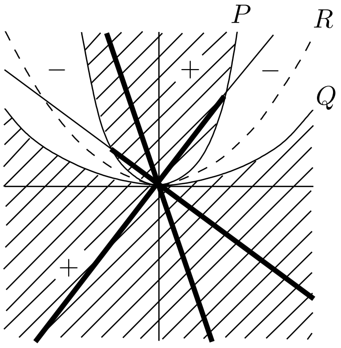


Figure 2. The parabolas P and Q given by $x_2 = x_1^2$ and $x_2 = 3x_1^2$, respectively.

Note that in the (x_1, x_2)-plane, $f(x) = 0$ precisely on the two parabolas P and Q shown in Figure 2. Also, the function f is positive above P and below Q, while it is negative between P and Q. With this information, we see that 0 cannot be a local minimizer of f. After all, we can take points x between the two parabolas P and Q, which are arbitrarily close to 0, but for which $f(x) < 0 = f(0)$. (For example, points x of the type $x = (\epsilon, 2\epsilon^2)$ on the dotted parabola R between P and Q shown in Figure 2, with ϵ small enough.)

On the other hand, if we fix any direction d, we see that as we approach 0 along this line, eventually we lie in the region where f is positive; see Figure 2. This shows that φ does have a minimum at 0. ◊

Theorem 14.5. *A necessary (but not sufficient) condition for $\widehat{x}$ to be a local minimizer of f is that $\nabla f(\widehat{x}) = 0$ and that $F(\widehat{x})$ is positive semi-definite.*

Proof. From Lemma 14.3, we know that for every $d \in \mathbb{R}^n$, the corresponding φ defined by (14.3) has a local minimum at $t = 0$. But by the first half of Theorem 13.7, it follows that $\varphi'(0) = 0$ and $\varphi''(0) \geq 0$. From (14.2), it follows that for all $d \in \mathbb{R}^n$, $\varphi'(0) = \nabla f(\widehat{x})d = 0$ and $\varphi''(0) = d^\top F(\widehat{x})d \geq 0$. Consequently, $\nabla f(\widehat{x}) = 0$ and that $F(\widehat{x})$ is positive semi-definite. □

We had seen an example of the non-sufficiency claim above in Theorem 13.7. Another instance illustrating this is our Example 14.4 above; see the exercise below.

Exercise 14.6. Calculate $\nabla f(0)$ and the Hessian $F(0)$ in Example 14.4.

[2] if $n > 1$

In order to derive *sufficient* conditions for $\widehat{x}$ to be a local minimizer of f, we will first give the multi-variable analogue of Taylor's formula (13.2).

Let $x \in \mathbb{R}^n$. We want to compare $f(x)$ with $f(\widehat{x})$, without explicitly calculating $f(x)$. Set $d = x - \widehat{x}$ and let

$$\varphi(t) := f(\widehat{x} + td) = f(\widehat{x} + t(x - \widehat{x})).$$

Then $\varphi(0) = f(\widehat{x})$ and $\varphi(1) = f(x)$. A special case of (the one-variable) Taylor's formula (13.2) is:

$$\varphi(1) = \varphi(0) + \varphi'(0) + \frac{1}{2}\varphi''(\theta) \quad \text{for some } \theta \in (0, 1).$$

This gives, using (14.1) and (14.2), that

$$f(x) = f(\widehat{x}) + \nabla f(\widehat{x})d + \frac{1}{2}d^\top F(\widehat{x} + \theta d)d \quad \text{for some } \theta \in (0, 1).$$

Equivalently, for some $\theta \in (0, 1)$,

$$f(x) = f(\widehat{x}) + \nabla f(\widehat{x})(x - \widehat{x}) + \frac{1}{2}(x - \widehat{x})^\top F(\widehat{x} + \theta(x - \widehat{x}))(x - \widehat{x}). \tag{14.4}$$

This is the multi-variable analogue of (13.2).

We will need the following result.

Lemma 14.7. *If the Hessian $F(\widehat{x})$ of f at $\widehat{x}$ is positive definite, that is, if for all nonzero $d \in \mathbb{R}^n$, $d^\top F(\widehat{x})d > 0$, then there exists a $\delta > 0$ such that for all $x \in \mathbb{R}^n$ satisfying $\|x - \widehat{x}\| < \delta$, $F(x)$ is positive definite.*

Proof. Consider the compact set $K = \{d \in \mathbb{R}^n : \|d\| = 1\}$. The continuous function $d \mapsto d^\top F(\widehat{x})d$ has a minimum value m on K, and since $F(\widehat{x})$ is positive definite, $m > 0$. Let $\epsilon := \frac{m}{2n^2} > 0$. Since the maps $x \mapsto F_{ij}(x)$ are all continuous, there exists a $\delta > 0$ such that for all $x \in \mathbb{R}^n$ satisfying $\|x - \widehat{x}\| < \delta$, $\max\limits_{i,j} |F_{ij}(x) - F_{ij}(\widehat{x})| < \epsilon$.

For $x \in \mathbb{R}^n$ satisfying $\|x - \widehat{x}\| < \delta$, and any $d \in K$, we then have

$$\begin{aligned} d^\top F(x)d &= d^\top F(\widehat{x})d + d^\top(F(x) - F(\widehat{x}))d \geq m - |d^\top(F(x) - F(\widehat{x}))d| \\ &= m - \left|\sum_{i=1}^n d_i \sum_{j=1}^n (F_{ij}(x) - F_{ij}(\widehat{x}))d_j\right| \\ &\geq m - \sum_{i=1}^n |d_i| \sum_{j=1}^n |F_{ij}(x) - F_{ij}(\widehat{x})||d_j| \\ &\geq m - \sum_{i=1}^n |d_i| \sum_{j=1}^n \epsilon|d_j| = m - \epsilon \sum_{i=1}^n |d_i| \sum_{j=1}^n |d_j| \\ &\geq m - \epsilon \sum_{i=1}^n 1 \sum_{j=1}^n 1 = m - \epsilon n^2 \\ &= m - \frac{m}{2n^2} \cdot n^2 = m - \frac{m}{2} = \frac{m}{2} > 0. \end{aligned}$$

But every nonzero $d \in \mathbb{R}^n$ can be written as $d = \|d\|\overline{d}$, where $\overline{d} \in K$. Hence by the above, for x's satisfying $\|x - \widehat{x}\| < \delta$, we have

$$d^\top F(x)d = \|d\|^2 \overline{d}^\top F(x)\overline{d} > \|d\|^2 \frac{m}{2} > 0.$$

This completes the proof. □

Theorem 14.8. *A sufficient (but not necessary) condition for $\widehat{x}$ to be a local minimizer of f is that $\nabla f(\widehat{x}) = 0$ and $F(\widehat{x})$ is positive definite.*

Proof. Suppose that $\nabla f(\widehat{x}) = 0$ and $F(\widehat{x})$ is positive definite. By Lemma 14.7, there exists a $\delta > 0$ such that for all $x \in \mathbb{R}^n$ satisfying $\|x - \widehat{x}\| < \delta$, $F(x)$ is positive definite. For these x's, there also holds that $F(\widehat{x} + \theta(x - \widehat{x}))$ is positive definite for every $\theta \in (0, 1)$, since

$$\|(\widehat{x} + \theta(x - \widehat{x})) - \widehat{x}\| = \|\theta(x - \widehat{x}))\| = \theta\|x - \widehat{x}\| \le \|x - \widehat{x}\| < \delta.$$

Using (14.4), we have for all $x \in \mathbb{R}^n$ satisfying $\|x - \widehat{x}\| < \delta$ that

$$f(x) = f(\widehat{x}) + \underbrace{\nabla f(\widehat{x})}_{=0}(x - \widehat{x}) + \frac{1}{2}(x - \widehat{x})^\top F(\widehat{x} + \theta(x - \widehat{x}))(x - \widehat{x}) \ge f(\widehat{x}),$$

with equality iff $x = \widehat{x}$. Thus $\widehat{x}$ is a (strict) local minimizer of f. □

We had seen an example of the non-necessity claim above in Theorem 13.7. Yet another example is given in the exercise below.

Exercise 14.9. Let $f(x) = x_1^4 + x_2^4$ ($x \in \mathbb{R}^2$). Clearly $0 \in \mathbb{R}^2$ is a global minimizer. Calculate $\nabla f(0)$ and the Hessian $F(0)$.

Exercise 14.10. Check that $\widehat{x} := \begin{bmatrix} 2 & 1 \end{bmatrix}^\top$ is a strict local minimizer of f, where f is given by $f(x_1, x_2) = x_1^3 - 12x_1x_2 + 8x_2^3$.

Exercise 14.11. Find all global minimizers for the function g on $\mathbb{R}^2$ given by

$$g(x_1, x_2) = x_1^4 - 12x_1x_2 + x_2^4, \quad (x_1, x_2) \in \mathbb{R}^2.$$

Chapter 15

Convexity revisited

An optimization problem is in a certain sense "well-posed" if the objective function which should be minimized is a *convex* function and the feasible region over which the minimization is to take place is a *convex* set. One of the many nice properties possessed by such problems is that every local optimal solution is a global optimal solution. In this chapter we will list a few important properties of convex functions.

The following result shows that for a continuously differentiable convex function, every tangent plane to the function lies *below* the graph of the function. Thus every (first order) approximation by a linear map of a convex function gives an *under*estimate.

Theorem 15.1. *Suppose that $C \subset \mathbb{R}^n$ is a given convex set and that f is continuously differentiable on C. Then f is convex on C iff*

$$\text{for all } x, y \in C, \quad f(y) \geq f(x) + \nabla f(x)(y - x). \tag{15.1}$$

Proof. (Only if) Suppose that f is convex. Let $x, y \in C$. By the convexity of f, for $t \in (0,1)$, $f(x + t(y - x)) \leq f(x) + t(f(y) - f(x))$, that is,

$$\frac{f(x + t(y - x)) - f(x)}{t} \leq f(y) - f(x)$$

for all $t \in (0, 1)$. Passing the limit as $t \searrow 0$ (and by the chain rule), $\nabla f(x)(y - x) \leq f(y) - f(x)$.

(If) Suppose that (15.1) holds. Let $u, v \in C$ and let $t \in (0,1)$. let $x := (1 - t)u + tv \in C$ and $y := v \in C$. Then by (15.1), we obtain

$$f(v) \geq f(x) + \nabla f(x)(v - ((1 - t)u + tv)) = f(x) + (1 - t)\nabla f(x)(v - u). \tag{15.2}$$

Using the same x, but now with $y = u$, we obtain from (15.1) that

$$f(u) \geq f(x) + \nabla f(x)(u - ((1 - t)u + tv)) = f(x) + t\nabla f(x)(u - v). \tag{15.3}$$

Multiplying (15.2) by t (> 0), multiplying (15.3) by $1 - t$ (> 0), and by adding the results, we obtain $(1 - t)f(u) + tf(v) \geq f(x) = f((1 - t)u + tv)$. Hence f is convex. □

We had seen that if f is a function of one variable such that $f''(x) \geq 0$ for all x, then f is convex. Observe that the condition $f''(x) \geq 0$ for all x means that f' is an increasing function, that is, $f'(y) \geq f'(x)$ whenever $y \geq x$. Equivalently, $(f'(y) - f'(x))(y - x) \geq 0$ for all x, y. In fact a stronger result is true.

Theorem 15.2. *Suppose that f is a continuously differentiable function on a convex set $C \subset \mathbb{R}^n$. Then f is convex on C iff*

$$\text{for all } x, y \in C, \quad (\nabla f(y) - \nabla f(x))(y - x) \geq 0. \tag{15.4}$$

Proof. (Only if) Suppose that f is convex. Let $x, y \in C$. By Theorem 15.1, we have that $f(y) \geq f(x)+\nabla f(x)(y-x)$. By interchanging x and y, we also obtain $f(x) \geq f(y)+\nabla f(y)(x-y)$. Adding the two inequalities we have now obtained yields (15.4).

(If) Suppose that (15.4) holds. Let $u, v \in C$. Define the function $\varphi(t) = f((1-t)u+tv)$, $t \in [0,1]$. Then $\varphi(1) = f(v)$, $\varphi(0) = f(u)$, and $\varphi'(t) = \nabla f((1-t)u+tv)(v-u)$. By the mean value theorem applied to φ, we obtain that

$$\frac{\varphi(1)-\varphi(0)}{1-0} = \varphi'(\theta)$$

for some $\theta \in (0,1)$, that is, $f(v) = f(u)+\nabla f(w)(v-u)$, where $w := u+\theta(v-u)$. By (15.4), we have $(\nabla f(w)-\nabla f(u))(w-u) \geq 0$. But $w-u = \theta(v-u)$, and so we obtain $(\nabla f(w)-\nabla f(u))(v-u) \geq 0$, or equivalently, $\nabla f(w)(v-u) \geq \nabla f(u)(v-u)$. Consequently,

$$f(v) = f(u)+\nabla f(w)(v-u) \geq f(u)+\nabla f(u)(v-u).$$

By Theorem 15.1, it follows that f is convex. □

In the case of twice differentiable functions, we also have the following test for convexity.

Theorem 15.3. *Suppose that f is a twice continuously differentiable function on a convex set $C \subset \mathbb{R}^n$. Then f is convex on C iff*

$$\text{for all } x, y \in C,\quad (y-x)^\top F(x)(y-x) \geq 0, \tag{15.5}$$

where $F(x)$ denotes the Hessian of f at x.

Proof. (Only if) Suppose that f is convex. Let $x, y \in C$ be such that $(y-x)^\top F(x)(y-x) < 0$. Let $d := y-x$. Then $d^\top F(x)d < 0$. Since f is twice continuously differentiable, it follows that the map $t \mapsto d^\top F(x+td)d$ is continuous, and so there exists an $\epsilon \in (0,1)$ such that for all $t \in [0,\epsilon]$, $d^\top F(x+td)d < 0$. Now let $u := x$ and $v = x+\epsilon d$. By Taylor's formula,

$$f(v) = f(u)+\nabla f(u)(v-u)+\frac{1}{2}(v-u)^\top F(w)(v-u),$$

where $w = u+\theta(v-u)$ for some $\theta \in (0,1)$. But $v-u = \epsilon d$ and so $w = u+\theta(v-u) = x+\theta\epsilon d = x+td$, where $t \in (0,\epsilon)$. Hence

$$\frac{1}{2}(v-u)^\top F(w)(v-u) = \frac{1}{2}\epsilon^2 d^\top F(x+td)d < 0.$$

Consequently,

$$f(v) = f(u)+\nabla f(u)(v-u)+\frac{1}{2}(v-u)^\top F(w)(v-u) < f(u)+\nabla f(u)(v-u),$$

which by Theorem 15.1 implies that f is not convex, a contradiction. So (15.5) holds.

(If) Suppose that (15.5) holds. Let $u, v \in C$. By Taylor's formula,

$$f(v) = f(u)+\nabla f(u)(v-u)+\frac{1}{2}(v-u)^\top F(w)(v-u),$$

where $w = u+\theta(v-u)$ for some $\theta \in (0,1)$. But $v-w = (1-\theta)(v-u)$, and so

$$\frac{1}{2}(v-u)^\top F(w)(v-u) = \frac{1}{2}\frac{1}{(1-\theta)^2}(v-w)^\top F(w)(v-w) \geq 0,$$

where the last inequality follows from (15.5). Hence we have

$$f(v) = f(u)+\nabla f(u)(v-u)+\frac{1}{2}(v-u)^\top F(w)(v-u) \geq f(u)+\nabla f(u)(v-u).$$

By Theorem 15.1, f is convex. □

From this result, it follows that a *sufficient* condition for f to be convex on C is that the Hessian $F(x)$ is positive semi-definite for all $x \in C$. But now consider the following example.

Example 15.4. Consider the convex set $C := \{x \in \mathbb{R}^2 : x_1 = x_2\}$; see Figure 1.

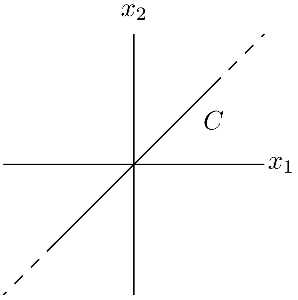


Figure 1. $C^\circ = \emptyset$.

Let $f(x) = x_1 x_2$ ($x \in \mathbb{R}^2$). Then the Hessian is given by

$$F(x) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad x \in \mathbb{R}^2,$$

and this is a constant matrix which is *not* positive semi-definite. Indeed, if we take $d = \begin{bmatrix} 1 & -1 \end{bmatrix}^\top$, then $d^\top F(x) d < 0$.

However, the condition (15.5) is satisfied. Indeed, if $x, y \in C$, then

$$(y-x)^\top F(x)(y-x) = 2(y_1 - x_1)(y_2 - x_2) = 2(y_1 - x_1)^2 \geq 0.$$

So the function is convex. $\Diamond$

The convex set C in the above example was "thin", and it had no "interior" points. See Figure 1. We will now see that if this is not the case, then the point-wise positive-definiteness of $F(x)$ on C is enough to guarantee that (15.5) holds, which in turn guarantees the convexity of f.

Definition 15.5. Let $S \subset \mathbb{R}^n$. A point $y \in S$ is called an *interior point of* S, if there exists a $\epsilon > 0$ such that the *ball* with centre y and radius ϵ is contained in S, that is, for all $x \in \mathbb{R}^n$ such that $\|x - y\| < \epsilon$, we have that $x \in S$. See Figure 2. The *interior of* S, denoted by S°, is the set of all interior points of S.

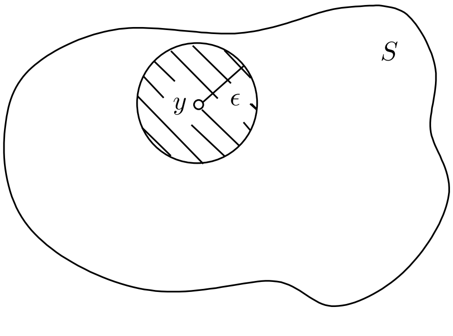


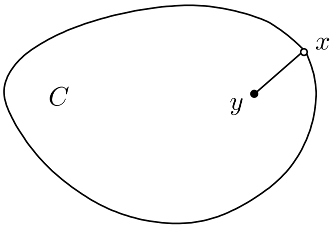
Figure 2. y is an interior point of S.

Exercise 15.6. Find the interior of each of the following subsets of $\mathbb{R}^2$:

$$\{x \in \mathbb{R}^2 : x_1 = x_2\}, \quad \mathbb{R}^2, \quad \emptyset, \quad \{x \in \mathbb{R}^2 : x_1^2 + x_2^2 \leq 1\}.$$

First we will prove the following result about convex sets and their interiors.

Lemma 15.7. *Let $C \subset \mathbb{R}^n$ be a convex set such that $C^\circ \neq \emptyset$. If $x \in C$ and $y \in C^\circ$, then $x + t(y - x) \in C^\circ$ for all $t \in (0, 1]$.*



In particular, it follows that C° is a convex set as well. Note that we do not demand that $x \in C^\circ$ in the lemma.

Proof. Let $x \in C$, $y \in C^\circ$, $t \in (0, 1]$ and set $u := x + t(y - x)$. We must show that $u \in C^\circ$. Since $y \in C^\circ$, there exists an $\epsilon > 0$ such that $v \in C$ for all $v \in \mathbb{R}^n$ satisfying $\|v - y\| < \epsilon$.

We will show that the ball with centre u and radius ϵt is contained in C, which implies that $u \in C^\circ$. Let $w \in \mathbb{R}^n$ satisfy $\|w - u\| < \epsilon t$. We want to show $w \in C$. Let $v := x + \frac{1}{t}(w - x)$. Then $w = x + t(v - x)$, and if we show that $v \in C$, then we will obtain that $w \in C$. We have

$$v - y = x - y + \frac{1}{t}(w - x) = \frac{1}{t}(x - u) + \frac{1}{t}(w - x) = \frac{1}{t}(w - u),$$

and so $\|v - y\| = \frac{1}{t}\|w - u\| < \frac{1}{t}\epsilon t = \epsilon$. Hence $v \in C$. □

Theorem 15.8. *Let $C \subset \mathbb{R}^n$ be a convex set having a nonempty interior, and $f : C \to \mathbb{R}$ be twice continuously differentiable on C. Then f is convex iff*

$$\text{for all } x \in C, \;\; F(x) \text{ is positive semi-definite,} \tag{15.6}$$

where $F(x)$ denotes the Hessian of f at x.

Proof. (Only if) Suppose that f is convex. Suppose that $F(x)$ is *not* positive semi-definite for some $x \in C$. Thus there exists a $d \in \mathbb{R}^n$ such that $d^\top F(x) d < 0$. Let y be an interior point of C. Since f is twice continuously differentiable, the map $t \mapsto d^\top F(x + t(y - x))d$ is continuous, and so there exists a $t \in (0, 1)$, small enough such that $d^\top F(x + t(y - x))d < 0$. Let $u := x + t(y - x)$. Then $d^\top F(u)d < 0$. By Lemma 15.7, $u \in C^\circ$, which implies that there exists a $\gamma > 0$ such that $v := u + \gamma d \in C$. Then we have $u \in C$, $v \in C$ and $\gamma d = v - u$. Thus $(v - u)^\top F(u)(v - u) = \gamma^2 d^\top F(u)d < 0$. By Theorem 15.3, we arrive at the contradiction that f is not convex.

(If) Suppose that (15.6) holds. But then (15.5) holds for all $x, y \in C$, and by Theorem 15.3, it follows that f is convex. □

A special case of the above result is the case when $C = \mathbb{R}^n$.

Corollary 15.9. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function on $\mathbb{R}^n$. Then f is convex iff the Hessian $F(x)$ of f at x is positive semi-definite for all $x \in \mathbb{R}^n$.*

Finally, we end this chapter with an application of Theorem 15.1 to optimization.

Theorem 15.10. *Suppose that the function f is convex and continuously differentiable on $\mathbb{R}^n$. Then $\widehat{x} \in \mathbb{R}^n$ is a global minimizer of f iff $\nabla f(\widehat{x}) = 0$.*

Proof. (Only if) Suppose that $\widehat{x} \in \mathbb{R}^n$ is a global minimizer of f. Then $\widehat{x} \in \mathbb{R}^n$ is a local minimizer of f. By Theorem 14.5, it follows that $\nabla f(\widehat{x}) = 0$.

(If) Let $\nabla f(\widehat{x}) = 0$. By Theorem 15.1, for all $x \in \mathbb{R}^n$, $f(x) \geq f(\widehat{x}) + \nabla f(\widehat{x})(x - \widehat{x}) = f(\widehat{x})$. Hence $\widehat{x} \in \mathbb{R}^n$ is a global minimizer of f. □

Exercise 15.11. Show that in each of the following cases, the function f is convex:

(1) $f(x_1, x_2) = \log(e^{a_1 x_1} + e^{a_2 x_2})$, where the a_1, a_2 are real constants.

(2) $f(x_1, x_2) = \dfrac{x_1^2}{x_2}$, for $x_2 > 0$.

(3) $f(x) = -\sqrt{x_1 x_2}$, for $x_1 > 0$ and $x_2 > 0$.

Exercise 15.12.

(1) Let $I \subset \mathbb{R}$ be an interval.
  (a) Let $f : I \to \mathbb{R}$ be an increasing convex function, and let $g : C \to I$ be a convex function on the convex set $C \subset \mathbb{R}^n$. Prove that $f \circ g : C \to \mathbb{R}$ is a convex function on C.
  (b) If on the other hand $f : I \to \mathbb{R}$ is a decreasing convex function, and $g : C \to I$ is a *concave*[1] function on the convex set $C \subset \mathbb{R}^n$, then show that $f \circ g : C \to \mathbb{R}$ is a convex function on C.

(2) Show that if $g : C \to \mathbb{R}$ is concave and positive, then $\dfrac{1}{g}$ is convex.

(3) Show that in each of the following cases, the function f is convex:

$$\text{(a) } f(x) = \log\left(\sum_{i=1}^{n} e^{a_i x_i}\right)$$

$$\text{(b) } f(x) = \sqrt{\sum_{i=1}^{n} x_i^2}$$

$$\text{(c) } f(x) = -\left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} \quad (x_i > 0).$$

Exercise 15.13. Prove the *arithmetic mean-geometric mean inequality* for positive numbers $x_1, \dots, x_n$:

$$\frac{x_1 + \cdots + x_n}{n} \geq \sqrt[n]{x_1 \dots x_n}.$$

Exercise 15.14. Let f, g be two given real-valued convex functions on $\mathbb{R}^n$, and consider the following convex optimization problem (in the variable $x \in \mathbb{R}^n$), which we denote by (P):

$$(P) : \begin{cases} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0. \end{cases}$$

This exercise is about determining an upper bound on the optimal value of (P) by solving an associated linear programming problem.

Given K points $x^{(1)}, \dots, x^{(K)}$ in $\mathbb{R}^n$, consider the following linear programming problem (in the variables $w_1, \dots, w_K$), which we denote by (LP):

$$(LP) : \begin{cases} \text{minimize } \displaystyle\sum_{k=1}^{K} w_k f(x^{(k)}) \\ \text{subject to } \displaystyle\sum_{k=1}^{K} w_k g(x^{(k)}) \leq 0, \ \sum_{k=1}^{K} w_k = 1, \ w_k \geq 0 \ (k = 1, \dots, K). \end{cases}$$

Suppose that $\widehat{x}$ is an optimal solution to (P) and that $\widehat{w} = \begin{bmatrix} \widehat{w}_1 & \dots & \widehat{w}_K \end{bmatrix}^\top$ is an optimal basic solution to (LP).

(1) Show that $\displaystyle\sum_{k=1}^{K} \widehat{w}_k f(x^{(k)}) \geq f(\widehat{x})$.
  (So the optimal value to (LP) gives an upper bound for the optimal value of (P).)

[1] that is, $-g$ is convex

(2) Assume that one of the given points $x^{(k_*)}$ is an optimal solution to (P). Find an optimal solution to (LP), and show that the optimal values to (P) and (LP) are then the same.

Exercise 15.15. For which real values of a is the following function convex on $\mathbb{R}^3$?

$$f(x_1,x_2,x_3) = x_1^2 + 5x_2^2 + ax_3^2 + 2x_1x_2 + 4x_2x_3 + x_2^4, \quad (x_1,x_2,x_3) \in \mathbb{R}^3.$$

Chapter 16

Newton's method

In this chapter we assume that

$$\boxed{\begin{array}{l} f : \mathbb{R}^n \to \mathbb{R} \text{ is twice continuously differentiable, and} \\ \text{its Hessian } F(x) \text{ at } x \text{ is positive definite for all } x \in \mathbb{R}^n. \end{array}}$$

This implies that f is (strictly) convex with a global minimizer $\widehat{x} \in \mathbb{R}^n$ characterized by $\nabla f(\widehat{x}) = 0$.

In this chapter we will learn how one can determine $\widehat{x}$ numerically using *Newton's method.* This method is iterative, and so it suffices to describe how from an iteration point $x^{(k)}$, one generates the next iteration point $x^{(k+1)}$. The user can choose the starting point $x^{(1)}$ as best as possible.

The basic idea behind this method is the following. Given a starting point, we construct a quadratic approximation to the objective function of second order, that is, the first and second derivatives of the quadratic approximation match the respective ones of the original function at the starting point. We then minimize this approximation, instead of the original objective function. We use the minimizer of the approximation as the starting point in the next step, and repeat the procedure. If the objective function is quadratic, then the approximation is exact, and the method yields the true minimizer in just one step. If, on the other hand, the objective function is not quadratic, then the approximation will produce only an estimate of the position of the true minimizer. Figure 1 illustrates this idea.

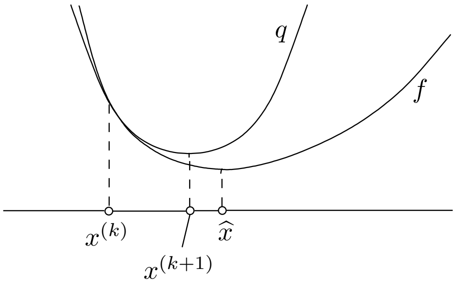


Figure 1. The quadratic approximation q of f at $x^{(k)}$ is used to determine the estimate $x^{(k+1)}$ of the true minimizer $\widehat{x}$ of f.

Given the iteration point $x^{(k)}$, we can first calculate the gradient $\nabla f(x^{(k)})$ and the Hessian $F(x^{(k)})$.

If $\nabla f(x^{(k)}) = 0$, then we have found the sought after minimizer $\widehat{x}$, and we terminate our search.

Suppose that $\nabla f(x^{(k)}) \neq 0$. The second order Taylor-approximation of the function f at the point $x^{(k)}$, expressed in terms of the vector $d = x - x^{(k)} \in \mathbb{R}^N$ is given by

$$f(x^{(k)} + d) \approx f(x^{(k)}) + \nabla f(x^{(k)})d + \frac{1}{2}d^\top F(x^{(k)})d.$$

The right hand side above is a strictly convex quadratic function, which is minimized by the unique solution to the following system in the unknown $d \in \mathbb{R}^n$:

$$F(x^{(k)})d = -(\nabla f(x^{(k)}))^\top. \tag{16.1}$$

Denote the unique solution to this by $d^{(k)}$. Since $\nabla f(x^{(k)}) \neq 0$, also $d^{(k)} \neq 0$. Furthermore, the directional derivative of f at $x^{(k)}$ in the direction of $d^{(k)}$ is

$$\nabla f(x^{(k)})d^{(k)} = -(d^{(k)})^\top F(x^{(k)})d^{(k)} < 0.$$

Thus in the direction $d^{(k)}$, f decreases (see Lemma 14.2) that is, it is a descent direction for f at $x^{(k)}$. Note that in order to arrive at this conclusion, we have used the fact that $F(x^{(k)})$ is positive definite.

The natural candidate for the next iteration point is $x^{(k)} + d^{(k)}$, which minimizes the quadratic Taylor approximation of f at $x^{(k)}$.

In the special case when $n = 1$ (so that f is a function of one variable), the system of equations (16.1) collapses to just the single equation

$$f''(x^{(k)})d = -f'(x^{(k)}),$$

and then

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

Exercise 16.1. Using Newton's method, find a minimizer (up to, say, two decimal places) of f given by

$$f(x) = x^2 - \sin x, \quad x \in \mathbb{R}.$$

Start with $x^{(0)} = 1$.

Exercise 16.2. We want to use Newton's method for finding a minimizer of the function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(x_1, x_2) = x_1^4 + 2x_1^2x_2^2 + x_2^4.$$

Show that if the current iterate x is of the form (a, a) with $a \neq 0$, then the next iterate is $(\frac{2}{3}a, \frac{2}{3}a)$. Based on this observation, what do you think $\widehat{x}$ is?

Exercise 16.3. Let n be a given (large) integer. Consider the function f given by

$$f(x) = \sum_{j=1}^{n}(x_j^4 - x_j^3 + x_j^2 - x_j),$$

where $x = (x_1, \dots, x_n) \in \mathbb{R}^n$.

(1) Show that f is convex.

(2) Suppose want to use Newton's method to find a minimizer of f in $\mathbb{R}^n$. Suppose we start from $x^{(1)} = (1, \dots, 1) \in \mathbb{R}^n$. Perform one iteration of Newton's algorithm, and find $x^{(2)}$.

Chapter 17

Nonlinear least squares problem: Gauss-Newton

In this chapter we shall consider the so-called nonlinear least-squares problem. This problem arises in many different applications, among others when we want to fit a mathematical model to given measurement data. Here one needs to determine the values of certain parameters in the model, and this must be done so that the difference between the model and the measured data is made as small as possible, that is, one wants to minimize the sum of the squares of the differences.

If the parameters enter the model linearly, then one obtains a linear least-squares problem. This is a relatively simple type of problem, and we have already considered this in Part II of this course; see Chapter 11.

If the parameters enter the model in a nonlinear manner, which is not unusual, then one obtains a nonlinear least-squares problem. This type of problem has the form

$$\text{minimize } f(x) = \frac{1}{2}\sum_{i=1}^{m}(h_i(x))^2, \tag{17.1}$$

where $h_1, \dots, h_m$ are given functions from $\mathbb{R}^n$ to $\mathbb{R}$. The factor $\frac{1}{2}$ is introduced so that some of the expressions are simplified below. Usually m is significantly larger than n, that is, the number of functions h_i is significantly larger than the number of variables x_j. In the context of model fitting, this corresponds to the fact that the number of measured observations is significantly larger than the number of parameters to be determined.

An often relevant interpretation of the optimization problem (17.1) is that we actually want to solve the nonlinear equation system

$$\left\{\begin{array}{rcl} h_1(x) & = & 0, \\ & \vdots & \\ h_m(x) & = & 0, \end{array}\right. \tag{17.2}$$

but since the number of equations (m) is larger than the number of variables (n), typically this system has no solution. Then it is natural to ask for a solution x which fails to satisfy (17.2) "as little as possible", for example, an optimal solution to the problem (17.1). (Note that if the system (17.2) has a solution $\widehat{x}$, then $\widehat{x}$ is a global optimal solution to the problem (17.1).)

Here is a concrete example of a nonlinear least-squares problem.

Example 17.1. Suppose that we want to determine the coordinates (x_1, x_2) of a point P_0 by measuring the distances $b_1, \dots, b_m$ from P_0 to m reference points $P_1, \dots, P_m$ with known coordinates.

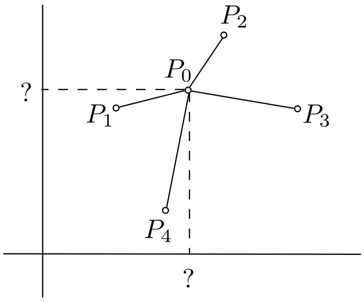


Figure 1. Estimating the coordinates of P_0.

Consider the special case when $m = 4$, and the points P_1, P_2, P_3, P_4 have the following coordinates:

$$\begin{array}{rcl} P_1 & \equiv & (-40, 30) \\ P_2 & \equiv & (40, 30), \\ P_3 & \equiv & (-30, -40) \\ P_4 & \equiv & (30, -40). \end{array}$$

Suppose that the measured distances of P_0 to the points P_1, P_2, P_3, P_4, are equal to $b_1 = 51$, $b_2 = 52$, $b_3 = 48$, $b_4 = 49$, respectively. Ideally, we want to determine x_1, x_2 such that

$$\begin{array}{rcl} h_1(x) & := & \sqrt{(x_1+40)^2 + (x_2-30)^2} - 51 = 0, \\ h_2(x) & := & \sqrt{(x_1-40)^2 + (x_2-30)^2} - 52 = 0, \\ h_3(x) & := & \sqrt{(x_1+30)^2 + (x_2+40)^2} - 48 = 0, \\ h_4(x) & := & \sqrt{(x_1-30)^2 + (x_2+40)^2} - 49 = 0. \end{array}$$

But since the measured distances b_i's are not exact owing to measurement errors, one doesn't really want an $x = (x_1, x_2)$ that satisfies the above four equations exactly. Instead, we consider the least-squares problem

$$\text{minimize } f(x) = \frac{1}{2}\Big((h_1(x))^2 + (h_2(x))^2 + (h_3(x))^2 + (h_4(x))^2\Big).$$

Since the functions h_i's are nonlinear functions of x_1 and x_2, this is a nonlinear least-squares problem. ♢

Even if in principle one could use Newton's method for minimizing f given by (17.1), it is most often both simpler and more efficient to use the so-called Gauss-Newton method, which uses the special structure that the problem has. This method can be interpreted in two different ways. We shall give both these interpretations here, and begin with the technically easier one.

17.1. First interpretation

The method is iterative, so it is enough to describe how one moves from an iteration point $x^{(k)}$ to the next iteration point $x^{(k+1)}$.

We linearize each function h_i at the iteration point $x^{(k)}$, that is, approximate each h_i by its first order Taylor polynomial at $x^{(k)}$:

$$h_i(x) \approx h_i(x^{(k)}) + \nabla h_i(x^{(k)})(x - x^{(k)}), \quad i = 1, \dots m.$$

With $d := x - x^{(k)}$, that is, $x = x^{(k)} + d$, we can rewrite the above as

$$h_i(x^{(k)} + d) \approx h_i(x^{(k)}) + \nabla h_i(x^{(k)})d, \quad i = 1, \dots m.$$

Let $h(x) \in \mathbb{R}^m$ be the column with components $h_1(x), \dots, h_m(x)$, and let $\nabla h(x)$ be the $m \times n$ matrix with the rows $\nabla h_1(x), \dots, \nabla h_m(x)$, that is,

$$h(x) = \begin{bmatrix} h_1(x) \\ \vdots \\ h_m(x) \end{bmatrix} \quad \text{and} \quad \nabla h(x) = \begin{bmatrix} \dfrac{\partial h_1}{\partial x_1}(x) & \dots & \dfrac{\partial h_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial h_m}{\partial x_1}(x) & \dots & \dfrac{\partial h_m}{\partial x_n}(x) \end{bmatrix}.$$

Then the objective function can be written compactly as

$$f(x) = \frac{1}{2}\|h(x)\|^2,$$

while the first order approximations above can be written as

$$h(x^{(k)} + d) \approx h(x^{(k)}) + \nabla h(x^{(k)})d.$$

The corresponding approximation of the objective function f is then given by

$$\begin{aligned} f(x^{(k)} + d) &= \frac{1}{2}\|h(x^{(k)} + d)\|^2 \\ &\approx \frac{1}{2}\|h(x^{(k)}) + \nabla h(x^{(k)})d\|^2 \\ &= \frac{1}{2}\|A^{(k)}d - b^{(k)}\|^2, \end{aligned} \tag{17.3}$$

where we introduce the the matrix $A^{(k)} := \nabla h(x^{(k)})$ and the vector $b^{(k)} := -h(x^{(k)})$. In the Gauss-Newton method, we minimize the right hand side of (17.3) in the variable vector $d \in \mathbb{R}^n$:

$$\text{minimize } \frac{1}{2}\|A^{(k)}d - b^{(k)}\|^2. \tag{17.4}$$

But this is a linear least-squares problem, which we have learnt to solve in Part II of this course. It has a solution given by the normal equations $(A^{(k)})^\top A^{(k)}d = (A^{(k)})^\top b$, that is,

$$(\nabla h(x^{(k)}))^\top \nabla h(x^{(k)})d = -(\nabla h(x^{(k)}))^\top h(x^{(k)}). \tag{17.5}$$

Let $d^{(k)}$ be a solution to these normal equations.

Then the next iteration point is given by

$$x^{(k+1)} = x^{(k)} + d^{(k)}.$$

If the columns of $\nabla h(x^{(k)})$ are linearly independent (which is usually the case, since $m > n$), then the matrix $(\nabla h(x^{(k)}))^\top \nabla h(x^{(k)})$ is positive definite, and so there is a unique solution to the normal equations.

17.2. Second interpretation

The second interpretation of the Gauss-Newton's method stems from the observation that the gradient and Hessian of the objective function $f(x) = \frac{1}{2}\|h(x)\|^2$ can be written in the following form:

$$\nabla f(x) \quad = \quad h(x)^\top \nabla h(x), \tag{17.6}$$

$$F(x) \quad = \quad (\nabla h(x))^\top \nabla h(x) + \sum_{i=1}^{m} h_i(x) H_i(x), \tag{17.7}$$

where $H_i(x)$ denotes the Hessian of the function h_i at x.

Let $x^{(k)}$ be the current iteration point. If $F(x^{(k)})$ is positive definite, then one can use Newton's method, that is, determine a direction $d^{(k)}$ via the equation system:

$$F(x^{(k)}) = -\nabla f(x^{(k)}). \tag{17.8}$$

Furthermore, in many cases it is also possible to do the approximation

$$F(x^{(k)}) \approx \nabla(h(x^{(k)}))^\top \nabla h(x^{(k)}), \tag{17.9}$$

and ignore the term

$$\sum_{i=1}^{m} h_i(x^{(k)}) H_i(x^{(k)}).$$

For example, in model fitting problems, it is reasonable that every $h_i(x)$ is "rather small", at least after a couple of iterations, if the model fits well to the data. It can also be the case that the functions h_i are "almost linear", so that the second derivatives are small. If we use the approximation (17.9) in (17.8), then we obtain the equation

$$(\nabla h(x^{(k)}))^\top \nabla h(x^{(k)}) d = -(\nabla h(x^{(k)}))^\top h(x^{(k)}), \tag{17.10}$$

which is the same as the equation (17.5) obtained in the preceding section.

An important advantage of the Gauss-Newton method (17.5), when compared with Newton's method (17.8), is that one doesn't need to calculate any second derivatives.

We now revisit Example 17.1 considered at the beginning of this chapter and solve it using the Gauss-Newton method.

Example 17.2. We start with

$$x^{(1)} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Then

$$h(x^{(1)}) = \begin{bmatrix} -1 \\ -2 \\ 2 \\ 1 \end{bmatrix},$$

and $f(x^{(1)}) = 5$. Also,

$$\nabla h(x^{(1)}) = \begin{bmatrix} 0.8 & -0.6 \\ -0.8 & -0.6 \\ 0.6 & 0.8 \\ -0.6 & 0.8 \end{bmatrix}.$$

Thus

$$(\nabla h(x^{(1)}))^\top \nabla h(x^{(1)}) = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

and

$$(\nabla h(x^{(1)}))^\top h(x^{(1)}) = \begin{bmatrix} 1.4 \\ 4.2 \end{bmatrix}.$$

Equation (17.10), namely

$$(\nabla h(x^{(1)}))^\top \nabla h(x^{(1)}) d = -(\nabla h(x^{(1)}))^\top h(x^{(1)}),$$

then becomes

$$\begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix} d = - \begin{bmatrix} 1.4 \\ 4.2 \end{bmatrix},$$

which has the solution

$$d^{(1)} = \begin{bmatrix} -0.7 \\ -2.1 \end{bmatrix}.$$

We set

$$x^{(2)} = x^{(1)} + d^{(1)} = \begin{bmatrix} -0.7 \\ -2.1 \end{bmatrix}.$$

Thus we have completed one whole iteration. The subsequent calculations are cumbersome to carry out by hand, and so we stop here. ♢

We should bear in mind that the problem (17.1) is in general not a convex problem. Thus one cannot be sure of finding a *global* optimal solution to (17.1).

Exercise 17.3. Verify (17.6) and (17.7).

Exercise 17.4. Let $\delta_1, \delta_2, \delta_3, \delta_4$ be four given numbers which typically are quite "small". Consider the nonlinear least squares problem in the variable $x \in \mathbb{R}^2$:

$$\text{minimize } f(x) = \frac{1}{2}\Big((h_1(x))^2 + (h_2(x))^2 + (h_3(x))^2 + (h_4(x))^2\Big),$$

where the functions h_i, $i = 1, 2, 3, 4$, are given as follows:

$$\begin{aligned} h_1(x) &= x_1^2 - x_2 - \delta_1, \\ h_2(x) &= x_1^2 + x_2 - \delta_2, \\ h_1(x) &= x_2^2 - x_1 - \delta_3, \\ h_1(x) &= x_2^2 + x_1 - \delta_4. \end{aligned}$$

(1) First assume that $\delta_1 = \delta_2 = \delta_3 = \delta_4 = 0$. Show that $\widehat{x} := 0 \in \mathbb{R}^2$ is a global minimizer of f. (This motivates the choice of $x^{(0)}$ as the starting point below.)

(2) Now suppose that $\delta_1 = -0.1$, $\delta_2 = 0.1$, $\delta_3 = -0.2$, $\delta_4 = 0.2$. Perform one iteration of the Gauss-Newton algorithm starting with $x^{(0)} = 0 \in \mathbb{R}^2$. Determine if the $x^{(1)}$ you find is a local minimizer of f.

Chapter 18

Optimization with constraints: Introduction

So far we have considered optimization problems where all the variables x_j were free, and could take arbitrarily large or small values. In many (most!) applications one doesn't have this freedom. So we now consider the problems where the variables satisfy *constraints*, that is, we consider optimization problems that have the following general form:

$$\left\{\begin{array}{ll}\text{minimize} & f(x),\\ \text{subject to} & x \in \mathcal{F},\end{array}\right. \tag{18.1}$$

where $\mathcal{F}$ is a subset of $\mathbb{R}^n$ and f is a given real-valued function defined (at least) on the set $\mathcal{F}$. The set $\mathcal{F}$ is called the *feasible set* for the problem (18.1). Soon we will consider $\mathcal{F}$'s having a more explicit form, for example a problem with equality constraints:

$$\mathcal{F} = \{x \in \mathbb{R}^n : h_i(x) = 0,\ i = 1, \dots, m\},$$

where $h_1, \dots, h_m$ are given functions from $\mathbb{R}^n$ to $\mathbb{R}$, or a problem with inequality constraints:

$$\mathcal{F} = \{x \in \mathbb{R}^n : g_i(x) \leq 0,\ i = 1, \dots, m\},$$

where $g_1, \dots, g_m$ are given functions from $\mathbb{R}^n$ to $\mathbb{R}$.

Definition 18.1. A point $x \in \mathbb{R}^n$ is called a *feasible solution* to the problem (18.1) if $x \in \mathcal{F}$.

A point $\widehat{x} \in \mathcal{F}$ is called a *local optimal solution* to the problem (18.1) if there exists a $\delta > 0$ such that for all $x \in \mathcal{F}$ that satisfy $\|x - \widehat{x}\| < \delta$, we have $f(\widehat{x}) \leq f(x)$.

A point $\widehat{x} \in \mathcal{F}$ is called a *global optimal solution* to the problem (18.1) if for all $x \in \mathcal{F}$, $f(\widehat{x}) \leq f(x)$.

It is obvious that every global optimal solution is also a local optimal solution, but it can happen for some problems that that there exist local optimal solutions which are not global optimal solutions.

Definition 18.2. A vector $d \in \mathbb{R}^n$ is called a *feasible direction at* $x \in \mathcal{F}$ for the problem (18.1) if there exists an $\epsilon > 0$ such that $x + td \in \mathcal{F}$ for all $t \in (0, \epsilon)$.

A vector $d \in \mathbb{R}^n$ is called a *feasible descent direction at* $x \in \mathcal{F}$ for the problem (18.1) if there exists an $\epsilon > 0$ such that $x + td \in \mathcal{F}$ and $f(x + td) < f(x)$ for all $t \in (0, \epsilon)$.

Lemma 18.3. *Suppose that $\widehat{x} \in \mathcal{F}$ is a local optimal solution to the problem* (18.1). *Then there does* not *exist a feasible descent direction at $\widehat{x} \in \mathcal{F}$ for the problem* (18.1).

Proof. Suppose, on the contrary, that there exists a feasible descent direction $d \in \mathbb{R}^n$ at $\widehat{x} \in \mathcal{F}$ for the problem (18.1). Then exists an $\epsilon > 0$ such that

$$\widehat{x} + td \in \mathcal{F} \text{ and } f(\widehat{x} + td) < f(\widehat{x}) \text{ for all } t \in (0, \epsilon). \tag{18.2}$$

On the other hand, since $\widehat{x}$ is a local optimal solution to the problem (18.1), exists a $\delta > 0$ such that

$$\text{for all } x \in \mathcal{F} \text{ such that } \|x - \widehat{x}\| < \delta, \;\; f(\widehat{x}) \le f(x). \tag{18.3}$$

Now take $x = \widehat{x} + td$, where $t = \frac{1}{2} \min \left\{ \frac{\delta}{\|d\|}, \epsilon \right\}$. Then $x \in \mathcal{F}$ and

$$\|x - \widehat{x}\| = t\|d\| \le \frac{1}{2} \frac{\delta}{\|d\|} \|d\| = \frac{1}{2} \delta < \delta.$$

Thus we arrive at the conclusion that both $f(x) < f(\widehat{x})$ (from (18.2)) and $f(\widehat{x}) \le f(x)$ (from (18.3)) must hold, a contradiction. □

Chapter 19

Optimality conditions: equality constraints

In this chapter we suppose that the set $\mathcal{F}$ is defined via a bunch of equality constraints, that is,

$$\mathcal{F} = \{x \in \mathbb{R}^n : h_i(x) = 0,\ i = 1, \dots, m\},$$

where $h_1, \dots, h_m$ are given functions from $\mathbb{R}^n$ to $\mathbb{R}$. Thus the problem (18.1) now takes the following form:

$$\begin{cases} \text{minimize} & f(x), \\ \text{subject to} & h_i(x) = 0,\ i = 1, \dots, m. \end{cases} \tag{19.1}$$

We will assume that

$$\boxed{f \text{ and the } h_i\text{'s are continuously differentiable.}}$$

Usually, $m < n$, which means that the constraints are given by a nonlinear equation system with more unknowns (n) than the number of equations (m). Most often, this equation system has infinitely many solutions, and the optimization problem consists of determining that solution x for which the objective function f takes the least possible value.

Let $h(x) \in \mathbb{R}^m$ denote the column vector having the components $h_1(x), \dots, h_m(x)$, and let $\nabla h(x)$ be the $m \times n$ matrix with the rows $\nabla h_1(x), \dots, \nabla h_m(x)$, that is,

$$h(x) = \begin{bmatrix} h_1(x) \\ \vdots \\ h_m(x) \end{bmatrix} \quad \text{and} \quad \nabla h(x) = \begin{bmatrix} \dfrac{\partial h_1}{\partial x_1}(x) & \dots & \dfrac{\partial h_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial h_m}{\partial x_1}(x) & \dots & \dfrac{\partial h_m}{\partial x_n}(x) \end{bmatrix}.$$

Definition 19.1. A feasible solution $x \in \mathcal{F}$ is called a *regular point* for the problem (19.1) if the rows of the matrix $\nabla h(x)$ above are linearly independent, that is, the gradient vectors $\nabla h_1(x), \dots, \nabla h_m(x)$ are linearly independent, that is, there do *not* exist scalars $u_1, \dots, u_m$, such that $(u_1, \dots, u_m) \neq (0, \dots, 0)$ and

$$\sum_{i=1}^{m} u_i \nabla h_i(x) = 0.$$

Since the matrix $\nabla h(x)$ usually has fewer rows than columns ($m < n$), the rows are "almost always" linearly independent. So in a certain sense, one would have to be particularly "unlucky" to encounter a nonregular point. But sometimes one has such bad luck!

Lemma 19.2. *Suppose that $\widehat{x} \in \mathcal{F}$ is both a regular point and a local optimal solution to the problem* (19.1)*. Then there cannot exist a vector $d \in \mathbb{R}^n$ which satisfies*

$$\begin{aligned} \nabla f(\widehat{x})d &< 0 \quad \textit{and} \\ \nabla h(\widehat{x})d &= 0. \end{aligned}$$

The proof of the above result is somewhat technical, relying on the Implicit Function Theorem. So we relegate the proof to the appendix.

Using Lemma 19.2, it is rather easy to prove the following important result:

Theorem 19.3. *Suppose that $\widehat{x} \in \mathcal{F}$ is both a regular point and a local optimal solution to the problem* (19.1)*. Then there exists a vector $\widehat{u} \in \mathbb{R}^m$ such that*

$$\nabla f(\widehat{x}) + \widehat{u}^\top \nabla h(\widehat{x}) = 0. \tag{19.2}$$

Proof. Suppose that $\widehat{x} \in \mathcal{F}$ is both a regular point and a local optimal solution to the problem (19.1). Then by Lemma 19.2, there cannot exist a vector d such that $\nabla f(\widehat{x})d < 0$ and $\nabla h(\widehat{x})d = 0$. But then there cannot exist a vector d such that $\nabla f(\widehat{x})d > 0$ and $\nabla h(\widehat{x})d = 0$ either. (Since otherwise, we would have $\nabla f(\widehat{x})(-d) < 0$ and $\nabla h(\widehat{x})(-d) = 0$, which contradicts Lemma 19.2!) Hence

$$\nabla f(\widehat{x})d = 0 \text{ for all } d \text{ satisfying } \nabla h(\widehat{x})d = 0.$$

This means that $(\nabla f(\widehat{x}))^\top \in (\ker \nabla h(\widehat{x}))^\perp = \operatorname{ran}((\nabla h(\widehat{x})^\top)$, that is, there exists a vector $\widehat{u}$ such that $(\nabla f(\widehat{x}))^\top = (\nabla h(\widehat{x}))^\top(-\widehat{u})$, that is, $\nabla f(\widehat{x}) + \widehat{u}^\top \nabla h(\widehat{x}) = 0$. □

Note that (19.2) is a system of n equations. Together with the m constraint equations

$$h_i(\widehat{x}) = 0, \quad i = 1, \dots, m,$$

we have totally a system of $m + n$ (nonlinear) equations in the unknowns $\widehat{x}_1, \dots, \widehat{x}_n$, $\widehat{u}_1, \dots, \widehat{u}_m$. These can be written in a compact form as follows:

$$\left\{ \begin{array}{ll} \nabla f(\widehat{x})^\top + (\nabla h(\widehat{x}))^\top \widehat{u} & = 0, \\ h(\widehat{x}) & = 0. \end{array} \right. \tag{19.3}$$

Example 19.4. Let us revisit the quadratic optimization problem with linear equality constraints:

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}x^\top Hx + c^\top x + c_0, \\ \text{subject to} & Ax = b, \end{array}$$

where $A \in \mathbb{R}^{m\times n}$, $c \in \mathbb{R}^n$, $c_0 \in \mathbb{R}$ and the matrix $H \in \mathbb{R}^{n\times n}$ is symmetric.

This is a special case of the problem (19.1), where

$$\begin{aligned} f(x) &= \frac{1}{2}x^\top Hx + c^\top x + c_0 \text{ and} \\ h(x) &= b - Ax. \end{aligned}$$

Thus we have $(\nabla f(x))^\top = Hx + c$ and $\nabla h(x) = -A$, so that the system (19.3) reduces to the following *linear* system of equations in $\widehat{x}$ and $\widehat{u}$:

$$\begin{bmatrix} H & -A^\top \\ A & 0 \end{bmatrix} \begin{bmatrix} \widehat{x} \\ \widehat{u} \end{bmatrix} = \begin{bmatrix} -c \\ b \end{bmatrix},$$

which is precisely the system (10.3) which we had obtained in Chapter 10 earlier. ♢

Exercise 19.5. Let $Q \in \mathbb{R}^{n\times n}$ be positive semidefinite and $P \in \mathbb{R}^{n\times n}$ be positive definite. Consider the problem in the variable $x \in \mathbb{R}^n$:

$$\begin{array}{ll} \text{maximize} & x^\top Qx, \\ \text{subject to} & x^\top Px = 1. \end{array}$$

(1) Show that every feasible point is a regular point.

(2) Prove that if an optimal solution $\widehat{x}$ exists, then it is an eigenvector of $P^{-1}Q$, and the maximum value of the objective function is then the corresponding eigenvalue.

Exercise 19.6. Consider the following optimization problem:

$$\begin{array}{rl} \text{maximize} & x+y, \\ \text{subject to} & \left(\frac{a}{x}\right)^2 + \left(\frac{b}{y}\right)^2 = 1. \end{array}$$

Find candidate solutions. What does this say about the possibility of carrying a ladder round a corner of two corridors, with widths a and b? (See Figure 1.)

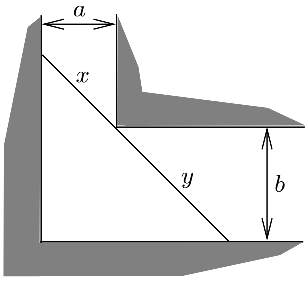


Figure 1. The ladder problem.

Exercise 19.7. Solve the following optimization problem:

$$\begin{array}{rl} \text{maximize} & x^4+y^4+z^4, \\ \text{subject to} & x^2+y^2+z^2 = 1, \\ & x+y+z=1. \end{array}$$

Exercise 19.8. Consider the following optimization problem:

$$\begin{array}{rl} \text{minimize} & x, \\ \text{subject to} & h(x) = 0, \end{array}$$

where h is given by

$$h(x) = \begin{cases} x^2 & \text{if } x < 0, \\ 0 & \text{if } 0 \le x \le 1, \\ (x-1)^2 & \text{if } x > 1. \end{cases}$$

Check that:

(1) the feasible set is $[0,1]$,

(2) that $\widehat{x} := 0$ is a local minimizer,

(3) $f'(\widehat{x}) = 1$ and $h'(\widehat{x}) = 0$.

So it is *not* true that there exists a $\widehat{u}$ such that $\nabla f(\widehat{x}) + \widehat{u}^\top \nabla h(\widehat{x}) = 0$. Does this mean that the statement of Theorem 19.3 is wrong?

Exercise 19.9. A cylindrical tin can with a bottom and a lid is required to have a volume of 1000 cubic centimeters. Find the dimensions of the can that will require the least amount of metal.

Exercise 19.10. Prove that among all triangles with a given perimeter P, the equilateral triangle has the largest area. (The area of a triangle with sides a, b, c is given by $\sqrt{s(s-a)(s-b)(s-c)}$, where s is the semiperimeter, that is, $s = P/2$.)

Exercise 19.11. The output of a manufacturing operation is a quantity $Q = Q(x,y)$ which is a function of the capital equipment x and the hours of labour y. Suppose that the price of labour is p (per hour) and price of investment in equipment is q (per unit). The plan is to spend the total amount b on the manufacturing operation. For optimal production, we want to minimize $Q(x,y)$ subject to $qx+py=b$. Show that at the optimum $(\widehat{x},\widehat{y})$, there holds that

$$\frac{1}{q}\frac{\partial Q}{\partial x} = \frac{1}{p}\frac{\partial Q}{\partial y}.$$

(In other words, at the optimum, the marginal change in the output per "dollar's worth" of additional capital equipment is equal to the marginal change in the output per dollar's worth of additional labour.)

Exercise 19.12. Suppose that x_1, x_2, x_3, x_4, x_5 are real numbers such that

$$\begin{array}{rcl} x_1 + x_2 + x_3 + x_4 + x_5 &=& 8, \quad \text{and} \\ x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 &=& 16. \end{array}$$

We want to determine the largest possible value of x_5. Pose this as an optimization problem, and solve it using the method of Lagrange multipliers. Explain why your solution is a global maximizer.

19.1. Appendix: proof of Lemma 19.2

Proof of Lemma 19.2. Let d be such that $\nabla h(\widehat{x})d = 0$.

Step 1. We will show that for a suitable choice of $u : [-\epsilon, \epsilon] \to \mathbb{R}^m$, the map $x : [-\epsilon, \epsilon] \to \mathbb{R}^n$ defined by

$$x(t) = \widehat{x} + td + (\nabla h(\widehat{x}))^\top u(t) \quad (t \in [-\epsilon, \epsilon]), \tag{19.4}$$

has the following properties:

(1) $x(t) \in \mathcal{F}$ for all $t \in [-\epsilon, \epsilon]$,

(2) $t \mapsto x(t)$ is continuously differentiable,

(3) $x(0) = \widehat{x}$.

To construct this curve $t \mapsto u(t)$, we consider the system of equations $h(\widehat{x} + td + (\nabla h(\widehat{x}))^\top u) = 0$, where for a fixed $t \in \mathbb{R}$, we consider $u \in \mathbb{R}^m$ to be the unknown. This is a nonlinear system of m equations in m unknowns, parameterized continuously by t. When $t = 0$, there is a solution, namely $u(0) := 0$. The Jacobian matrix of the system with respect to u at $t = 0$ is the $m \times m$ matrix $\nabla h(\widehat{x})(\nabla h(\widehat{x}))^\top$, and since $\nabla h(\widehat{x})$ has full row rank ($\widehat{x}$ is regular!), it follows that the matrix $\nabla h(\widehat{x})(\nabla h(\widehat{x}))^\top$ is invertible. So by the Implicit Function Theorem (recalled in Subsection 19.1.1 below), it follows that for each t in a small enough interval $[-\epsilon, \epsilon]$ with $\epsilon > 0$, there is a solution $u(t) \in \mathbb{R}^m$ that satisfies

(1′) $h(\widehat{x} + td + (\nabla h(\widehat{x}))^\top u(t)) = 0$,

(2′) the map $t \mapsto u(t)$ is continuously differentiable, and

(3′) $u(0) = 0$.

It is now immediate that x defined by (19.4) has the desired properties (1), (2), (3) listed above.

Step 2. Since $\widehat{x}$ is a local minimum, we know that for all ξ's in $\mathcal{F}$ close enough to $\widehat{x}$, $f(\widehat{x}) \le f(\xi)$. Now we will take the ξ's to be $x(t)$ with $t > 0$ small enough. Then

$$0 \le f(x(t)) - f(\widehat{x}) = f(x(t)) - f(x(0)),$$

for all $t \in [0, \delta]$ with a small enough $\delta > 0$. But by the Mean Value Theorem applied to the function $t \mapsto f(x(t))$, it follows that there is a $\tau \in (0, t)$ such that

$$0 \le f(x(t)) - f(x(0)) = t\nabla f(x(\tau))d,$$

and so $\nabla f(x(\tau))d \ge 0$. But as $t \to 0$, $\tau \to 0$ as well. Hence $\nabla f(x(0))d = \nabla f(\widehat{x})d \ge 0$. This completes the proof. □

Finally, we remind the reader of the precise statement of the Implicit Function Theorem that we used in the proof above.

19.1.1. Implicit Function Theorem. Suppose we have a set of m equations in $n > m$ variables

$$h_i(u,t) = 0, \quad i = 1, \dots, m.$$

Here $u \in \mathbb{R}^m$ and $t \in \mathbb{R}^{n-m}$. The implicit function theorem addresses the following question:

If $n - m$ of the variables are fixed, then can the equations be solved for the remaining m variables?

Thus, if we select the m variables $u_1, \dots, u_m$, then we are interested in finding out if these can be expressed in terms of the remaining variables, that is, if there are 'nice' functions φ_i such that

$$u_i = \varphi_i(t_1, \dots, t_{n-m}), \quad i = 1, \dots, m.$$

The functions φ_i, if they exist, are called *implicit functions.* As usual, we denote by $h(u,t) \in \mathbb{R}^n$ the column vector having the components $h_1(u,t), \dots, h_m(u,t)$.

Theorem 19.13 (Implicit Function Theorem)**.** *Let* $(u_0, t_0) \in \mathbb{R}^n$ *be such that:*

(1) *h is continuously differentiable in a neighbourhood of* (u_0, t_0),

(2) $h(u_0, t_0) = 0$, *and*

(3) *the Jacobian matrix with respect to u at* (u_0, t_0), *namely*

$$\begin{bmatrix} \frac{\partial h_1}{\partial u_1}(u_0,t_0) & \dots & \frac{\partial h_1}{\partial u_m}(u_0,t_0) \\ \vdots & \ddots & \vdots \\ \frac{\partial h_m}{\partial u_1}(u_0,t_0) & \dots & \frac{\partial h_m}{\partial u_m}(u_0,t_0) \end{bmatrix}, \textit{ is invertible.}$$

Then there is a neighbourhood N of t_0 in $\mathbb{R}^{n-m}$ such that for every t in this neighbourhood N, there is a corresponding vector $u(t) \in \mathbb{R}^m$ such that

(1) $h(u(t), t) = 0$,

(2) $t \mapsto u(t) : N \to \mathbb{R}^m$ *is continuously differentiable, and*

(3) $u(t_0) = u_0$.

Proof. See for example [**R**, Theorem 9.28, page 224-227]. □

Chapter 20

Optimality conditions: inequality constraints

In this chapter we suppose that the set $\mathcal{F}$ is defined via a bunch of inequality constraints, that is,

$$\mathcal{F} = \{x \in \mathbb{R}^n : g_i(x) \leq 0,\ i = 1, \dots, m\},$$

where $g_1, \dots, g_m$ are given functions from $\mathbb{R}^n$ to $\mathbb{R}$. Thus the problem (18.1) now takes the following form:

$$\left\{ \begin{array}{rl} \text{minimize} & f(x), \\ \text{subject to} & g_i(x) \leq 0,\ i = 1, \dots, m. \end{array} \right. \tag{20.1}$$

We will assume that

$$\boxed{f \text{ and the } g_i\text{'s are continuously differentiable.}}$$

In the previous chapter, when we considered equality constraints, we assumed that usually we have that $m < n$. In this chapter when we consider inequality constraints, we cannot make a corresponding similar assumption. This is because it can very well happen that $m > n$ without it being the case that the problem is concocted in a contrived or artificial way. It can also equally well happen that $m < n$, or that $m = n$. Thus we will not distinguish between these three cases in the following treatment, and make no assumption concerning the relative magnitude of m and n.

Definition 20.1. If $x \in \mathcal{F}$, then we denote by $I_a(x)$ ($\subset \{1, \dots, m\}$) the *active* index set, that is,

$$I_a(x) = \{i : g_i(x) = 0\}.$$

In particular if $x \in \mathcal{F}$ and $I_a(x) = \emptyset$, then $g_i(x) < 0$ for all i, that is, all the inequalities are satisfied with a strict inequality. Such points lie in the *interior* of the feasible set $\mathcal{F}$, and it is often the easiest situation for analysis from the optimization point of view, than the case of boundary points x of $\mathcal{F}$ for which $I_a(x) \neq \emptyset$.

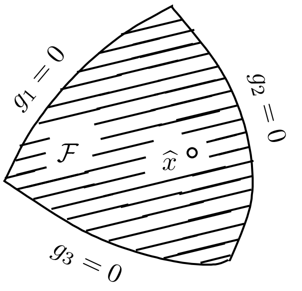


Figure 1. Local extremizer in the interior of the feasible region.

Theorem 20.2. *Suppose that* $\widehat{x} \in \mathcal{F}$ *with* $I_a(\widehat{x}) = \emptyset$ *is a local optimal solution to* (20.1)*. Then* $\nabla f(\widehat{x}) = 0$.

Proof. When $I_a(\widehat{x}) = \emptyset$, then every vector $d \in \mathbb{R}^n$ is a feasible direction for (20.1) at $\widehat{x}$. Indeed for every i, we have $g_i(\widehat{x}) < 0$, and owing to the continuity of g_i, it follows that for all sufficiently small $t > 0$, $g(\widehat{x} + td) < 0$.

Now suppose that $\nabla f(\widehat{x}) \neq 0$. Let $d := -(\nabla f(\widehat{x}))^\top \neq 0$. From the above, we know that this special d is also a feasible direction for (20.1) at $\widehat{x}$. The directional derivative of f at $\widehat{x}$ in the direction d is then given by

$$\nabla f(\widehat{x})d = -\nabla f(\widehat{x})(\nabla f(\widehat{x}))^\top = -\|(\nabla f(\widehat{x}))^\top\|^2 < 0,$$

which implies that (see Lemma 14.2) for sufficiently small $t > 0$, $f(\widehat{x} + td) < f(\widehat{x})$. Thus this d is a feasible descent direction for (20.1) at $\widehat{x}$, and by Lemma 18.3, $\widehat{x}$ cannot be a local optimal solution. □

In the sequel we analyze points $\widehat{x}$ for which $I_a(\widehat{x}) \neq \emptyset$.

Lemma 20.3. *Suppose that* $\widehat{x} \in \mathcal{F}$ *with* $I_a(\widehat{x}) \neq \emptyset$ *is a local optimal solution to* (20.1)*. Then there does* not *exist a* $d \in \mathbb{R}^n$ *such that*

$$\begin{aligned} \nabla f(\widehat{x})d &< 0, \text{ and} && (20.2)\\ \nabla g_i(\widehat{x})d &< 0 \text{ for all } i \in I_a(\widehat{x}). && (20.3) \end{aligned}$$

Proof. Suppose that the vector $d \in \mathbb{R}^n$ satisfies (20.2) and (20.3). Then for each $i \in I_a(\widehat{x})$, we have $g_i(\widehat{x} + td) < g_i(\widehat{x}) = 0$ for all sufficiently small $t > 0$, since the directional derivative $\nabla g_i(\widehat{x})d < 0$ for all $i \in I_a(\widehat{x})$; see Lemma 14.2.

But also for each $i \notin I_a(\widehat{x})$, we have $g_i(\widehat{x} + td) < 0$ for all sufficiently small $t > 0$. This is because g_i is continuous and $g_i(\widehat{x}) < 0$ for $i \notin I_a(\widehat{x})$.

From the above, we conclude that d is a feasible direction for (20.1) at $\widehat{x}$.

Furthermore, $f(\widehat{x} + td) < f(\widehat{x})$ for all sufficiently small $t > 0$, since the directional derivative $\nabla f(\widehat{x})d < 0$ (again using Lemma 14.2).

Consequently, d is a feasible descent direction for (20.1) at $\widehat{x}$. Finally, by Lemma 18.3, $\widehat{x}$ cannot be a local optimal solution. □

Recall Farkas' Lemma (Lemma 6.17):

Lemma 20.4. *Suppose that the $m+1$ vectors q, $p_1, \dots, p_m$ in $\mathbb{R}^n$ are given. Then* exactly one *of the following two systems in $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$ have at least one solution:*

$$(L): \begin{cases} q^\top x < 0, \\ p_1^\top x \geq 0, \\ \vdots \\ p_m^\top x \geq 0. \end{cases} \qquad (R): \begin{cases} q = y_1 p_1 + \dots + y_m p_m, \\ y_1 \geq 0, \\ \vdots \\ y_m \geq 0. \end{cases}$$

We will now prove a consequence of this, which we will use.

Lemma 20.5. *Suppose that vectors $a_1, \dots, a_m \in \mathbb{R}^n$ are given. Then exactly one of the following systems in $d \in \mathbb{R}^n$ and $v \in \mathbb{R}^m$ has at least one solution:*

$$(L'): a_i^\top d < 0, \; i = 1, \dots, m \qquad (R'): \begin{cases} \displaystyle\sum_{i=1}^m v_i a_i = 0, \\ \displaystyle\sum_{i=1}^m v_i > 0, \\ v_i \geq 0, \; i = 1, \dots, m. \end{cases}$$

Proof. Introduce an extra variable t, taking values in $\mathbb{R}$. To say that (L') has a solution is the same as saying that the following system in d and t has a solution:

$$\begin{cases} t < 0, \\ -a_i^\top d + t \geq 0, \; i = 1, \dots m. \end{cases}$$

We see that this is the same as (L) in Farkas' Lemma if

$$x := \begin{bmatrix} d \\ t \end{bmatrix}, \quad q := \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \text{and} \quad \begin{bmatrix} p_1^\top \\ \vdots \\ p_m^\top \end{bmatrix} = \begin{bmatrix} -a_1^\top & 1 \\ \vdots & \vdots \\ -a_m^\top & 1 \end{bmatrix}.$$

With this in mind, (R) in Farkas' Lemma takes the form in $v \in \mathbb{R}^m$:

$$\begin{cases} \displaystyle\sum_{i=1}^m v_i a_i = 0, \\ \displaystyle\sum_{i=1}^m v_i = 1, \\ v_i \geq 0, \; i = 1, \dots, m. \end{cases}$$

But it is clear that this system has a solution iff (R') has a solution. (Why?) So the claim follows from Farkas' Lemma. □

Lemma 20.6. *Let $\widehat{x} \in \mathcal{F}$ with $I_a(\widehat{x}) \neq \emptyset$ be a local optimal solution to* (20.1). *Then there exist scalars $v_0 \geq 0$ and $v_i \geq 0$ ($i \in I_a(\widehat{x})$) such that*

$$v_0 + \sum_{i \in I_a(\widehat{x})} v_i > 0, \quad \textit{and}$$

$$v_0 \nabla f(\widehat{x}) + \sum_{i \in I_a(\widehat{x})} v_i \nabla g_i(\widehat{x}) = 0.$$

Proof. This follows from Lemmas 20.3 and 20.5. □

Definition 20.7. A feasible solution $x \in \mathcal{F}$ with $I_a(x) \neq \emptyset$ is called a *regular point* for the problem (20.1) if there do *not* exist scalars $v_i \geq 0$, $i \in I_a(x)$, such that

$$\sum_{i \in I_a(x)} v_i > 0, \quad \text{and}$$
$$\sum_{i \in I_a(x)} v_i \nabla g_i(x) = 0.$$

A feasible solution $x \in \mathcal{F}$ with $I_a(x) = \emptyset$ is always a regular point for the problem (20.1).

In light of Lemma 20.5, an equivalent way of expressing this definition is the following:

A feasible solution $x \in \mathcal{F}$ with $I_a(x) \neq \emptyset$ is a regular point for the problem (20.1) if there exists at least one vector $d \in \mathbb{R}^n$ such that $\nabla g_i(x) d < 0$ for all $i \in I_a(x)$.

Remark 20.8. From the Definition 20.7, we observe that for $x \in \mathcal{F}$ to be a regular point for the problem (20.1), it is sufficient (but not necessary) that the gradients $\nabla g_i(x)$, $i \in I_a(x)$, are linearly independent.

Lemma 20.9. *Suppose that $\widehat{x} \in \mathcal{F}$ with $I_a(\widehat{x}) \neq \emptyset$ is both a regular point and a local optimal solution to* (20.1). *Then there exist scalars $y_i \geq 0$, $i \in I_a(\widehat{x})$ such that*

$$\nabla f(\widehat{x}) + \sum_{i \in I_a(\widehat{x})} y_i \nabla g_i(\widehat{x}) = 0.$$

Proof. From the Definition 20.7, it follows that the v_0 in Lemma 20.6 cannot be 0 since $\widehat{x}$ is a regular point. Since v_0 must then be strictly positive, one can divide the equation

$$v_0 \nabla f(\widehat{x}) + \sum_{i \in I_a(\widehat{x})} v_i \nabla g_i(\widehat{x}) = 0$$

throughout by v_0, and set $y_i := \dfrac{v_i}{v_0}$ (≥ 0) to obtain the desired conclusion. □

Lemma 20.10. *Suppose that $\widehat{x} \in \mathcal{F}$ with $I_a(\widehat{x}) \neq \emptyset$ is both a regular point and a local optimal solution to* (20.1). *Then there does* not *exist a vector $d \in \mathbb{R}^n$ such that*

$$\nabla f(\widehat{x}) d < 0, \quad \textit{and}$$
$$\nabla g_i(\widehat{x}) d \leq 0 \quad \textit{for all } i \in I_a(\widehat{x}).$$

Proof. This follows from Lemma 20.9 and Farkas' Lemma (with $q = (\nabla f(\widehat{x}))^\top$, $p_i = -(\nabla g_i(\widehat{x}))^\top$). □

The following result is the main result in this chapter. The optimality conditions (1)-(4) given below are called the *Karush-Kuhn-Tucker conditions*, and are abbreviated as "KKT-conditions".

Theorem 20.11. *Suppose that $\widehat{x} \in \mathcal{F}$ is both a regular point and a local optimal solution to* (20.1). *Then there exists a vector $\widehat{y} \in \mathbb{R}^m$ such that:*

(1) $\nabla f(\widehat{x}) + \sum_{i=1}^{m} \widehat{y}_i \nabla g_i(\widehat{x}) = 0,$

(2) $g_i(\widehat{x}) \leq 0,\ i = 1, \dots, m,$

(3) $\widehat{y}_i \geq 0,\ i = 1, \dots, m,$

(4) $\widehat{y}_i g_i(\widehat{x}) = 0,\ i = 1, \dots, m.$

Proof. Suppose first that $I_a(\widehat{x}) \neq \emptyset$. Then by Lemma 20.9, there exist scalars $y_i \geq 0$, $i \in I_a(\widehat{x})$, such that

$$\nabla f(\widehat{x}) + \sum_{i \in I_a(\widehat{x})} y_i \nabla g_i(\widehat{x}) = 0.$$

Define $\widehat{y} \in \mathbb{R}^m$, having components $\widehat{y}_1, \dots, \widehat{y}_m$, by setting

$$\widehat{y}_i = \begin{cases} y_i & \text{if } i \in I_a(\widehat{x}), \\ 0 & \text{if } i \notin I_a(\widehat{x}). \end{cases}$$

Then (1)-(4) are satisfied.

Now suppose that $I_a(\widehat{x}) = \emptyset$. Then by Theorem 20.2, $\nabla f(\widehat{x}) = 0$. By setting $\widehat{y} = 0 \in \mathbb{R}^m$, clearly (1)-(4) are satisfied. □

Lemma 20.12. *In the statement of Theorem 20.11, the condition (4) that $\widehat{y}_i g_i(\widehat{x}) = 0$, $i = 1, \dots, m$, can be replaced by the condition*

$$(4') \quad \sum_{i=1}^{m} \widehat{y}_i g_i(\widehat{x}) = 0.$$

Proof. Suppose first that $\widehat{y}_i g_i(\widehat{x}) = 0$, $i = 1, \dots, m$. Then obviously also (4′) holds.

On the other hand, if (1), (2), (3) and (4′) hold, then

$$\sum_{i=1}^{m} \underbrace{(-\widehat{y}_i g_i(\widehat{x}))}_{\geq 0} = 0,$$

and so $-\widehat{y}_i g_i(\widehat{x}) = 0$ for each i, that is, (4) holds. □

Let $g(x) \in \mathbb{R}^m$ denote the column vector having the components $g_1(x), \dots, g_m(x)$, and let $\nabla g(x)$ be the $m \times n$ matrix with the rows $\nabla g_1(x), \dots, \nabla g_m(x)$, that is,

$$g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{bmatrix} \quad \text{and} \quad \nabla g(x) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(x) & \dots & \frac{\partial g_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_m}{\partial x_1}(x) & \dots & \frac{\partial g_m}{\partial x_n}(x) \end{bmatrix}.$$

Then using Lemma 20.12, we can write Theorem 20.11 in a compact manner:

Theorem 20.13. *Suppose that $\widehat{x} \in \mathcal{F}$ is both a regular point and a local optimal solution to* (20.1). *Then there exists a vector $\widehat{y} \in \mathbb{R}^m$ such that:*

(1) $\nabla f(\widehat{x}) + \widehat{y}^\top \nabla g(\widehat{x}) = 0$,

(2) $g(\widehat{x}) \leq 0$,

(3) $\widehat{y} \geq 0$,

(4) $\widehat{y}^\top g(\widehat{x}) = 0$.

Example 20.14. Let us now consider a quadratic optimization problem with linear *inequality* constraints:

$$\begin{array}{rl} \text{minimize} & \frac{1}{2} x^\top H x + c^\top x + c_0, \\ \text{subject to} & Ax \geq b, \end{array}$$

where $A \in \mathbb{R}^{m \times n}$, $c \in \mathbb{R}^n$, $c_0 \in \mathbb{R}$ and the matrix $H \in \mathbb{R}^{n \times n}$ is symmetric. This is a special case of the problem (20.1), where

$$f(x) = \frac{1}{2} x^\top H x + c^\top x + c_0 \text{ and } g(x) = b - Ax.$$

Thus we have $(\nabla f(x))^\top = Hx+c$ and $\nabla g(x) = -A$, so that the KKT-conditions in Theorem 20.13 reduces to the following system:

(1) $Hx + c = A^\top \widehat{y}$,

(2) $A\widehat{x} \geq b$,

(3) $\widehat{y} \geq 0$,

(4) $\widehat{y}^\top(A\widehat{x} - b) = 0$.

$\Diamond$

Exercise 20.15. In $\mathbb{R}^2$ consider the constraints

$$\begin{cases} x_1 \geq 0, \\ x_2 \geq 0, \\ x_2 - (x_1 - 1)^2 \leq 0. \end{cases}$$

Find the active index set at the point $x = (1, 0) \in \mathbb{R}^2$. Show that the point $x = (1, 0)$ is feasible but not a regular point.

Exercise 20.16. Sketch the region in $\mathbb{R}^2$ determined by the following constraints:

$$\begin{cases} 1 - x - y \geq 0, \\ 5 - x^2 - y^2 \geq 0, \\ x \geq 0. \end{cases}$$

Use the Karush-Kuhn-Tucker conditions to obtain the condition satisfied by ∇f at the point $\widehat{x} = (2, -1)$ in $\mathbb{R}^2$ if the function f has a maximum at $\widehat{x}$ subject to the given constraints. Explain the geometric significance of your condition and sketch ∇f in your diagram.

Exercise 20.17. Consider the following optimization problem:

$$(NP): \begin{cases} \text{minimize} & x_1 \\ \text{subject to} & x_2 - x_1^3 \leq 0, \\ & -x_2 - x_1^3 \leq 0. \end{cases}$$

(1) Depict the feasible region graphically. In the same figure, show the level curves of the objective function using dotted lines. Hence conclude that the origin $(0, 0) \in \mathbb{R}^2$ is a global minimizer.

(2) Write the KKT optimality conditions corresponding to the problem (NP), and show that they are not satisfied at the point $(0, 0)$. Explain this by considering the regularity of the point $(0, 0)$.

Chapter 21

Optimality conditions for convex optimization

In this chapter we shall consider an extremely well-posed class of optimization problems, namely the so-called *convex* problems. For these one can derive much stronger optimization conditions than for general nonlinear problems.

First consider the general formulation of the optimization problem from Chapter 18:

$$\begin{cases} \text{minimize} & f(x), \\ \text{subject to} & x \in \mathcal{F}, \end{cases} \tag{21.1}$$

where the feasible set $\mathcal{F}$ is a given subset of $\mathbb{R}^n$ and the objective function f is a given real-valued function on $\mathcal{F}$.

Definition 21.1. The problem (21.1) is called a *convex optimization problem* if $\mathcal{F}$ is a convex set and f is a convex function on $\mathcal{F}$.

For convex functions, one has the following nice equivalences:

Lemma 21.2. *Suppose that* (21.1) *is a convex optimization problem, and that* $\widehat{x} \in \mathcal{F}$. *Then the following are equivalent:*

(1) $\widehat{x}$ *is a global optimal solution to* (21.1).

(2) $\widehat{x}$ *is a local optimal solution to* (21.1).

(3) *There is no feasible descent direction* d *for* (21.1) *at* $\widehat{x}$.

Proof. That (1)⇒(2) follows from the definitions. Lemma 18.3 gives (2)⇒(3). So the lemma follows once we show that (3)⇒(1), which we prove below.

We will show that if $\widehat{x}$ is *not* a global optimal solution, then there *exists* a feasible descent direction d for (21.1) at $\widehat{x}$.

Suppose that $\widehat{x} \in \mathcal{F}$ is not a global optimal solution. Then there exists (at least one) $x \in \mathcal{F}$ such that $f(x) < f(\widehat{x})$. Let $x(t) = \widehat{x} + td$ for $t \in (0,1)$, where $d := x - \widehat{x}$. So $x(t) = \widehat{x} + t(x - \widehat{x})$. Since $\mathcal{F}$ is a convex set, it follows that $x(t) \in \mathcal{F}$ for all $t \in (0,1)$, which implies in turn that d is a feasible direction for (21.1) at $\widehat{x}$. Also, the convexity of f implies that

$$f(x(t)) = f(\widehat{x} + t(x - \widehat{x})) \leq f(\widehat{x}) + t(f(x) - f(\widehat{x}))$$

for all $t \in (0,1)$. Thus d is a feasible descent direction for (21.1) at $\widehat{x}$. □

Since local and global optimal solutions are the same thing for convex problems, one can simply say "optimal solution" without the adjective "local" or "global". This is reminiscent of the terminology used in the first part of this course, when we dealt with linear optimization (where all problems were convex, although we did not always stress this) and we also encountered this in second part of the course when we dealt with convex quadratic optimization.

In the remainder of this chapter, we will consider problems having the form

$$\left\{\begin{array}{rl} \text{minimize} & f(x), \\ \text{subject to} & g_i(x) \leq 0, \quad i = 1, \dots, m, \end{array}\right. \tag{21.2}$$

where the objective function f and all the constraint functions g_i are *convex* functions. Also, we assume that f and all g_i's are continuously differentiable.

$$\boxed{f \text{ and the } g_i\text{'s are convex and continuously differentiable.}}$$

The feasible set $\mathcal{F}$ for the problem (21.2) is given by

$$\mathcal{F} = \{x \in \mathbb{R}^n : g_i(x) \leq 0, \ i = 1, \dots, m\}. \tag{21.3}$$

The verification that (21.2) is a convex optimization problem is left as an exercise.

Exercise 21.3. $\mathcal{F}$ defined by (21.3) is a convex set if the g_i's are convex.

The following important result shows that for convex problems of the form (21.2), the KKT-conditions are *sufficient* for a global optimal solution.

Theorem 21.4. *Consider the problem* (21.2), *where f and the g_i's are convex and continuously differentiable. Suppose that $\widehat{x} \in \mathbb{R}^n$ and $\widehat{y} \in \mathbb{R}^m$ satisfy the following KKT-conditions:*

(1) $\nabla f(\widehat{x}) + \sum_{i=1}^{m} \widehat{y}_i \nabla g_i(\widehat{x}) = 0,$

(2) $g_i(\widehat{x}) \leq 0,\ i = 1, \dots, m,$

(3) $\widehat{y}_i \geq 0,\ i = 1, \dots, m,$

(4) $\widehat{y}_i g_i(\widehat{x}) = 0,\ i = 1, \dots, m.$

Then $\widehat{x}$ is a (global) optimal solution to the problem (21.2).

Proof. Consider the function $\ell : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\ell(x) = f(x) + \sum_{i=1}^{m} \widehat{y}_i g_i(x) \quad (x \in \mathbb{R}^n).$$

Observe that here $\widehat{y}$ is fixed, while x is the variable vector. By (3), $\widehat{y}_i \geq 0$, which, together with the convexity of f and the g_i's, implies that ℓ is a convex function. Also ℓ is continuously differentiable since f and the g_i's are continuously differentiable.

The condition (1) above gives $\nabla \ell(\widehat{x}) = 0$, and by Theorem 15.10, it follows that $\widehat{x}$ is a global optimal solution for ℓ, that is, $\ell(\widehat{x}) \leq \ell(x)$ for all $x \in \mathbb{R}^n$.

The condition (2) implies that $\widehat{x} \in \mathcal{F}$. Now let $x \in \mathcal{F}$, that is, $g_i(x) \leq 0$ for all $i = 1, \dots, m$. We will now show that $f(\widehat{x}) \leq f(x)$. Indeed, we have

$$f(\widehat{x}) \overset{(4)}{=} \ell(\widehat{x}) \leq \ell(x) = f(x) + \sum_{i=1}^{m} \widehat{y}_i g_i(\widehat{x}) \leq f(x),$$

where the last inequality follows using (3) and the fact that $g_i(x) \leq 0$ for all $i = 1, \dots, m$ ($x \in \mathcal{F}$). This completes the proof. □

Thus the KKT-conditions are *sufficient* for optimality in the case of convex problems. We will soon show that if the convex problem is also "regular", then the KKT-conditions are also in fact necessary.

Definition 21.5. The convex problem (21.2) is said to *regular* if there exists an $x_0 \in \mathbb{R}^n$ such that $g_i(x_0) < 0$ for all $i = 1, \dots, m$.

Lemma 21.6. *If the problem* (21.2) *is regular and convex, then every feasible solution $x \in \mathcal{F}$ is a regular point (in the sense of Definition 20.7) for the problem* (21.2).

Proof. Let $x \in \mathcal{F}$ be a feasible solution for the problem (21.2).

If $I_a(x) = \emptyset$, the x is a regular point by definition; see Definition 20.7.

Now suppose that $I_a(x) \neq \emptyset$. We will show that there exists a $d \in \mathbb{R}^n$ such that $\nabla g_i(x)d < 0$ for all $i \in I_a(x)$. Let $d := x_0 - x$, where x_0 is a vector as in the Definition 21.5. Then we know that $g_i(x_0) < 0$ for each i. Now by the characterization of convexity given by Theorem 15.1, we obtain for *all* i's that

$$0 > g_i(x_0) \geq g_i(x) + \nabla g_i(x)(x_0 - x) = g_i(x) + \nabla g_i(x)d.$$

In particular, for $i \in I_a(x)$, we have $g_i(x) = 0$, and so the above yields that $\nabla g_i(x)d < 0$ for all $i \in I_a(x)$. But this implies that x is a regular point for the problem (21.2) in the sense of Definition 20.7. (Why?) □

In light of the previous lemma, Theorem 20.11, and also 21.4, we have the following:

Theorem 21.7. *Suppose that the problem* (21.2) *is regular and convex. Then $\widehat{x}$ is an (global) optimal solution to* (21.2) *iff there exists a $\widehat{y} \in \mathbb{R}^m$ such that the KKT-conditions are satisfied:*

(1) $\nabla f(\widehat{x}) + \sum_{i=1}^{m} \widehat{y}_i \nabla g_i(\widehat{x}) = 0$,

(2) $g_i(\widehat{x}) \leq 0$, $i = 1, \dots, m$,

(3) $\widehat{y}_i \geq 0$, $i = 1, \dots, m$,

(4) $\widehat{y}_i g_i(\widehat{x}) = 0$, $i = 1, \dots, m$.

Example 21.8. Consider the following problem in the variable vector $x \in \mathbb{R}^n$:

$$\begin{array}{rl} \text{minimize} & \|x - p\|^2, \\ \text{subject to} & \|x - q\|^2 \leq 1, \end{array}$$

where $p, q \in \mathbb{R}^n$ are vectors satisfying

$$\|p\|^2 = 1, \;\; \|q\|^2 = 1, \;\; p^\top q = 0.$$

See Figure 1 for a geometric interpretation when $n = 2$.

Let

$$\begin{array}{rcl} f(x) & := & \|x - p\|^2 = (x - p)^\top (x - p) = x^\top x - 2p^\top x + 1, \text{ and} \\ g(x) & := & \|x - q\|^2 - 1 = (x - q)^\top (x - q) - 1 = x^\top x - 2q^\top x. \end{array}$$

Then the given problem can be rewritten as:

$$\begin{array}{rl} \text{minimize} & f(x), \\ \text{subject to} & g(x) \leq 0. \end{array}$$

We have

$$\begin{array}{rcl} \nabla f(x) & = & 2x^\top - 2p^\top, \text{ and} \\ \nabla g(x) & = & 2x^\top - 2q^\top. \end{array}$$

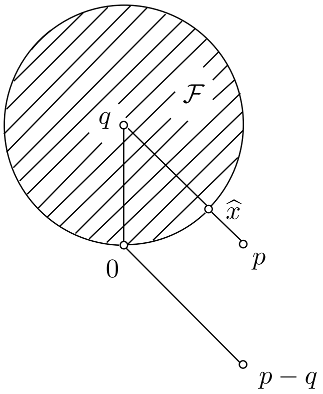


Figure 1. The case when $n = 2$.

Moreover, the Hessians of f and g at x, respectively, are given by

$$\begin{aligned} F(x) &= 2I, \text{ and} \\ G(x) &= 2I. \end{aligned}$$

Since these Hessians are positive semi-definite for all $x \in \mathbb{R}^n$, it follows from Corollary 15.9 that f and g are convex. So the problem is convex.

Moreover, if we take $x_0 := q$, then $g(x_0) = \|q - q\|^2 - 1 = -1 < 0$, and so the problem is also regular.

By Theorem 21.7, the KKT-conditions are necessary and sufficient for a global optimizer!

Since the problem has just one constraint, the KKT-conditions are the following, where $\widehat{x}$ is the sought-after optimal solution and $\widehat{y} \in \mathbb{R}$:

KKT-1: $\nabla f(\widehat{x}) + \widehat{y}\nabla g(\widehat{x}) = 0$,

KKT-2: $g(\widehat{x}) \leq 0$,

KKT-3: $\widehat{y} \geq 0$,

KKT-4: $\widehat{y}g(\widehat{x}) = 0$.

Thus:

KKT-1: $\widehat{x} - p + \widehat{y}(\widehat{x} - q) = 0$,

KKT-2: $\widehat{x}^\top \widehat{x} - 2q^\top \widehat{x} \leq 0$,

KKT-3: $\widehat{y} \geq 0$,

KKT-4: $\widehat{y}(\widehat{x}^\top \widehat{x} - 2q^\top \widehat{x}) = 0$.

We will consider the two possible cases: $\widehat{y} = 0$ and $\widehat{y} > 0$. (The case $\widehat{y} < 0$ is prohibited by KKT-3.)

$\underline{1}^\circ$ $\widehat{y} = 0$. Then KKT-1 implies that $\widehat{x} = p$. Substituting this in KKT-2, we obtain $\widehat{x}^\top \widehat{x} - 2q^\top \widehat{x} = p^\top p - 2q^\top p = 1 - 0 = 1$, which is not ≤ 0. Hence there cannot exist $\widehat{x} \in \mathbb{R}^n$ and $\widehat{y} \in \mathbb{R}$ satisfying the KKT-conditions with $\widehat{y} = 0$.

$\underline{2}^\circ$ $\widehat{y} > 0$. From KKT-1 we have $\widehat{x} = \dfrac{p + \widehat{y}q}{1 + \widehat{y}}$.

Since $\widehat{y} > 0$, KKT-4 gives $\widehat{x}^\top \widehat{x} - 2q^\top \widehat{x} = 0$, that is,

$$\begin{aligned} 0 &= \widehat{x}^\top \widehat{x} - 2q^\top \widehat{x} = \frac{(p+\widehat{y}q)^\top (p+\widehat{y}q)}{(1+\widehat{y})^2} - \frac{2q^\top (p+\widehat{y}q)}{1+\widehat{y}} \\ &= \frac{1+\widehat{y}^2}{(1+\widehat{y})^2} - \frac{2\widehat{y}}{1+\widehat{y}} = \frac{1-2\widehat{y}-\widehat{y}^2}{(1+\widehat{y})^2}. \end{aligned}$$

Thus $1 - 2\widehat{y} - \widehat{y}^2 = 0$, and so $\widehat{y} \in \{-1+\sqrt{2}, -1-\sqrt{2}\}$. But since $\widehat{y} > 0$, we have $\widehat{y} = -1+\sqrt{2}$. The corresponding optimal $\widehat{x}$ is given by

$$\widehat{x} = \frac{p+\widehat{y}q}{1+\widehat{y}} = \frac{p+q(\sqrt{2}-1)}{\sqrt{2}} = q + \frac{p-q}{\sqrt{2}}.$$

These $\widehat{x}$ and $\widehat{y}$ satisfy the KKT-conditions above. Consequently, this $\widehat{x}$ is a global optimal solution to the problem. ♢

Exercise 21.9. Suppose that $G = \{M_1, \dots, M_k\} \subset \mathbb{R}^n$ is a *group* (that is, it is closed under multiplication and taking inverses). A function $f : \mathbb{R}^n \to \mathbb{R}$ is called *G-invariant* if $f(M_i x) = f(x)$ for all $x \in \mathbb{R}^n$ and all $i = 1, \dots, k$. If $x \in \mathbb{R}^n$, then define

$$\overline{x} = \frac{1}{k}\sum_{i=1}^{k} M_i x.$$

Set $\mathcal{S} = \{x \in \mathbb{R}^n : M_i x = x \text{ for } i = 1, \dots, k\}$.

(1) Show that $\mathcal{S}$ is a subspace of $\mathbb{R}^n$.

(2) If $x \in \mathbb{R}^n$, then prove that $\overline{x} \in \mathcal{S}$.

(3) If f is convex and G-invariant, then show that for all $x \in \mathbb{R}^n$, $f(\overline{x}) \le f(x)$.

Now consider the following optimization problem:

$$(P) : \left\{ \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \le 0, \ i = 1, \dots, m. \end{array} \right.$$

Here the g_i are given functions from $\mathbb{R}^n$ to $\mathbb{R}$. The problem (P) is called *G-invariant* if f is G-invariant and the feasible set $\mathcal{F} := \{x \in \mathbb{R}^n : g_i(x) \le 0 \text{ for } i = 1, \dots, m\}$ is *G-invariant*, that is, $x \in \mathcal{F}$ implies that $M_j x \in \mathcal{F}$ for all j.

(4) Show that if (P) is convex and G-invariant, and if there exists a local optimal solution to (P), then there exists a global optimal solution to (P) which belongs to $\mathcal{S}$.

(5) As an example, suppose that the f and g_i are convex and *symmetric*. A symmetric function $h : \mathbb{R}^n \to \mathbb{R}$ satisfies $h(Px) = h(x)$ for all x and all permutation matrices P. Given a permutation π of n elements $\pi : \{1, \dots n\} \to \{1, \dots, n\}$, the corresponding permutation matrix

$$P_\pi = \begin{bmatrix} e_{\pi(1)} \\ \vdots \\ e_{\pi(n)} \end{bmatrix},$$

where e_i is the row vector in $\mathbb{R}^n$ with 1 in the ith position and zeros elsewhere. Then it can be checked that

$$P_\pi \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} x_{\pi(1)} \\ \vdots \\ x_{\pi(n)} \end{bmatrix}.$$

Show that if (P) has a local optimal solution, then it has a global optimal solution of the form $(a, \dots, a) \in \mathbb{R}^n$.

(6) Solve the following optimization problem:

$$\begin{array}{rl} \text{maximize} & x^4 + y^4 + z^4, \\ \text{subject to} & x^2 + y^2 + z^2 \le 1, \\ & x + y + z \le 1. \end{array}$$

Compare your solution with that obtained in Exercise 19.7.

Exercise 21.10. Let $f(x) = x_1^2 x_2^4 x_3^6$, where $x = (x_1, x_2, x_3) \in \mathbb{R}^3$.

(1) Is the function f convex?

(2) Determine if $\widehat{x} = 0 \in \mathbb{R}^3$ is a global optimal solution to the following problem:

$$\begin{array}{rl} \text{minimize} & f(x), \\ \text{subject to} & x^2 + y^2 + z^2 \le 1. \end{array}$$

(3) Find all global optimal solutions to the following problem:

$$\begin{array}{rl} \text{maximize} & f(x), \\ \text{subject to} & x^2 + y^2 + z^2 \le 1. \end{array}$$

Exercise 21.11. Let

$$f(x) = (x_1 + x_2)^2 + (x_2 + x_3)^2 + (x_3 + x_1)^2 - 12x_1 - 8x_2 - 4x_3,$$

where $x = (x_1, x_2, x_3) \in \mathbb{R}^3$.

(1) Determine if $\widehat{x} = (2, 1, 0) \in \mathbb{R}^3$ is a global optimal solution to the following problem:

$$\begin{array}{rl} \text{minimize} & f(x), \\ \text{subject to} & 0 \le x_j \le 2, \ j = 1, 2, 3. \end{array}$$

(2) Determine the values of the three constants c_1, c_2, c_3 such that $\widehat{x} = (2, 1, 0) \in \mathbb{R}^3$ is a global optimal solution to the following problem:

$$\begin{array}{rl} \text{minimize} & f(x), \\ \text{subject to} & (x_1 - c_1)^2 + (x_2 - c_2)^2 + (x_3 - c_3)^2 \le 1. \end{array}$$

Exercise 21.12. Let c be a constant. Consider the following optimization problem:

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}x_1^2 + \frac{1}{2}x_2^2 + \frac{1}{2}x_3^2 - x_1 - x_2 + cx_3 \\ \text{subject to} & x_1 + x_2 \ge 4, \\ & x_2 + x_3 \ge 4, \\ & x_3 + x_1 \ge 4. \end{array}$$

(1) Are there any values of c which make the point $\widehat{x} = (2, 2, 2) \in \mathbb{R}^3$ a global optimal solution to the problem? If yes, find all such values of c.

(2) Also answer the same question as above when $\widehat{x} = (2, 2, 4)$.

Exercise 21.13. Consider the following optimization problem:

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}x^\top x \\ \text{subject to} & Ax \ge b, \\ & x \ge 0, \end{array}$$

where $A = \begin{bmatrix} 2 & -2 & 1 & 1 \\ 1 & 1 & 2 & -2 \end{bmatrix}$ and $b = \begin{bmatrix} 20 \\ 30 \end{bmatrix}$.

(1) Is there a global optimal solution $\widehat{x}$ satisfying $A\widehat{x} = b$ and $\widehat{x} > 0$?

(2) Is there a global optimal solution $\widehat{x}$ satisfying $A\widehat{x} = b$ and moreover $\widehat{x}_3 = \widehat{x}_4 = 0$?

Exercise 21.14. The three points A, B, C in $\mathbb{R}^2$ have coordinates $(1, 0)$, $(-1, \sqrt{3})$, $(-\sqrt{3}, -3)$ respectively.

(1) (*Steiner's problem*) We want to find a point P in $\mathbb{R}^2$ such that *sum* of the distances of P to A, B, C, is minimized. Formulate this as an optimization problem. Verify that $(0, 0)$ is the required point. What is the measure of the angles APB, BPC, CPA? (This can be explained: Imagine that the triangle ABC lies in a horizontal plane, and there are three pulleys at A,B,C. Suppose three threads of long enough but equal lengths are tied at together at one end, and the three free ends have a weight of 1kg attached. The three free ends with weights are passed over the three pulleys, so that they hang under gravity's influence and eventually come to rest. Now nature chooses to minimize the potential energy. But this means that point where the ends of the strings are tied together will precisely be at that point P that minimizes the sum of the distances to the three points A, B, C. We know that the three equal vector forces (tensions in the string) must add up to zero, and so the sin of the angles APB, BPC, CPA must be the same. Hence these angles must be equal. But they add up to 360°, and so each must be 120°. The point P is referred to as the *Toricelli point* of the triangle ABC. Alternately, it is also possible to prove that the Toricelli point is the minimizer using elementary Euclidean geometry.)

(2) Now we want to find a point P in $\mathbb{R}^2$ such that maximum of the distances of P to A, B, C, is minimized. Formulate this as an optimization problem. Find out if the solution to the previous part, namely $\widehat{x} = (0, 0)$, is an optimal solution to your optimization problem.

Exercise 21.15. Let f and g be two real-valued convex and continuously differentiable functions on $\mathbb{R}^n$. Consider the following optimization problem:

$$(P): \left\{ \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g(x) \leq 0. \end{array} \right.$$

Given K points $x^{(1)}, \dots, x^{(K)}$ in $\mathbb{R}^n$, consider the following linear programming problem in the variables $x \in \mathbb{R}^n$ and $z \in \mathbb{R}$:

$$(LP): \left\{ \begin{array}{ll} \text{minimize} & z \\ \text{subject to} & z - \nabla f(x^{(k)})x \geq f(x^{(k)}) - \nabla f(x^{(k)})x^{(k)}, \\ & \quad - \nabla g(x^{(k)})x \geq g(x^{(k)}) - \nabla g(x^{(k)})x^{(k)}, \\ & \text{for } k = 1, \dots, K. \end{array} \right.$$

Suppose that (LP) has an optimal solution $(\widehat{x}, \widehat{z})$.

(1) Show that $\widehat{z}$ is a lower bound for the optimal value of (P), that is, $f(x) \geq \widehat{z}$ for every feasible solution x to (P).

(2) Suppose that the optimal solution $(\widehat{x}, \widehat{z})$ to (LP) is such that $\widehat{x} = x^{(k)}$ for a particular $k \in \{1, \dots, K\}$. Prove that $\widehat{x}$ is then an optimal solution to (P). Moreover, show that the optimal values of (LP) and (P) are the same.

Exercise 21.16. Let $g_1, \dots, g_m$ and f be real-valued convex and differentiable functions on $\mathbb{R}^n$. Consider the following optimization problem:

$$(P): \left\{ \begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad i = 1, \dots, m. \end{array} \right.$$

Suppose that the problem is regular. Suppose also that we have a *guess* $\widehat{x} \in \mathbb{R}^n$ for an optimal solution to (P). One way of finding out whether $\widehat{x}$ is indeed optimal or not is the following.

First we linearize the functions f and $g_1, \dots, g_m$ at the point $\widehat{x}$ (that is compute the first order Taylor polynomials at $\widehat{x}$ for each of these functions). Now we consider a new optimization problem (LP) obtained by replacing the functions in (P) by their respective linearizations. Observe that (LP) is a linear programming problem.

Show that $\widehat{x}$ is am optimal solution to (P) if and only if $\widehat{x}$ is an optimal solution to (LP).

Hint: First write the *dual* to (LP), and then use the duality theory of linear programming and the KKT-conditions.

Exercise 21.17. Consider the optimization problem in the variable $x \in \mathbb{R}^2$:

$$\begin{array}{rl} \text{minimize} & \frac{1}{2}(Ax - b)^\top (Ax - b), \\ \text{subject to} & Ax \geq b, \end{array}$$

where $A = \begin{bmatrix} 2 & -1 \\ -1 & 2 \\ 1 & 1 \end{bmatrix}$ and $b = \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}$.

(So this problem is similar to the one considered in Exercise 11.4, but now we demand all the components in the error vector $Ax - b$ must be nonnegative.)

Show that $(13/6, 11/6)$ is an optimal solution.

Exercise 21.18. Consider the problem

$$(P): \left\{ \begin{array}{ll} \text{minimize} & cx_1 - 4x_2 - 2x_3, \\ \text{subject to} & x_1^2 + x_2^2 \leq 2, \\ & x_2^2 + x_3^2 \leq 2, \\ & x_3^2 + x_1^2 \leq 2, \end{array} \right.$$

where c is a constant.

(1) Verify that (P) is a convex optimization problem.

(2) Write the KKT-conditions for the problem (P).

(3) Are there any values of c which make the point $x = (\frac{7}{5}, \frac{1}{5}, \frac{1}{5})$ an optimal solution to the problem? If so, find all of these possible values of c.

(4) Are there any values of c which make the point $x = (1, 1, 1)$ an optimal solution to the problem? If so, find all of these possible values of c.

Exercise 21.19. Let

$$\begin{aligned} f(x) &= x_1^2 - x_1 x_2 + x_2^2 + x_3^2 - 2x_1 + 4x_2, \\ g_1(x) &= -x_1 - x_2, \\ g_2(x) &= 1 - x_3. \end{aligned}$$

(1) Consider the problem

$$(P_d) : \begin{cases} \text{minimize} & f(x), \\ \text{subject to} & x \in \mathbb{R}^3. \end{cases}$$

Determine a global minimizer to (P_d). Justify your answer.

(2) Consider the problem

$$(P_c) : \begin{cases} \text{minimize} & f(x), \\ \text{subject to} & g_1(x) \leq 0, \\ & g_2(x) \leq 0. \end{cases}$$

Write the KKT-conditions for the problem (P_c). Show that the vector $\widehat{x} = (1, -1, 1)$ satisfies the KKT-conditions for (P_c). Conclude that $\widehat{x}$ is a global optimal solution to (P_c).

Exercise 21.20. Consider the following nonlinear optimization problem:

$$(NP) : \begin{cases} \text{minimize} & e^{-(x_1+x_2)} \\ \text{subject to} & e^{x_1} + e^{x_2} \leq 20, \\ & x_1 \geq 0. \end{cases}$$

(1) Show that (NP) is a regular convex optimization problem.

(2) Write the KKT optimality conditions and solve them. Is there a globally optimal solution to the problem (NP)? Justify your answer.

Chapter 22

Lagrange relaxation

An important tool for handling certain types of optimization problems under constraints is the so-called *Lagrange relaxation* method. This chapter presents some underlying basic concepts behind this method.

We consider the following optimization problem, labelled (P):

$$(P) : \begin{cases} \text{minimize} & f(x), \\ \text{subject to} & g_i(x) \le 0,\ i = 1, \dots, m, \\ & x \in X. \end{cases} \tag{22.1}$$

Here X is a given subset of $\mathbb{R}^n$, while $g_1, \dots, g_m$ and f are given real-valued functions defined (at least) on X. The constraints $g_i(x) \le 0$, $i = 1, \dots, m$, can be written compactly as $g(x) \le 0$, where $g(x)$ denotes the column vector having the components $g_1(x), \dots, g_m(x)$. The constraints in (P) are of two types:

(1) the *explicit* constraints $g(x) \le 0$, and

(2) the *implicit* constraints $x \in X$, which can be of the same type as the explicit ones, namely

$$X = \{x \in \mathbb{R}^n : h_i(x) \le 0,\ i = 1, \dots, k\},$$

where h_i are given real-valued functions.

There exists a certain freedom in the choosing which constraints are regarded as explicit or implicit. We shall comment below how the division of the constraints will be done in order to get the greatest possible benefit of the results. However, first we will define the concept of Lagrange relaxation.

22.1. Definition of the relaxed Lagrange problem

Let $y \in \mathbb{R}^m$ be a given vector with nonnegative components, that is, $y \ge 0$.

The following problem (PR_y) constitutes a *relaxed Lagrange problem* with respect to the explicit constraints $g(x) \le 0$ of the original problem (P) above given by (22.1):

$$(PR_y) : \begin{cases} \text{minimize} & f(x) + y^\top g(x), \\ \text{subject to} & x \in X, \end{cases} \tag{22.2}$$

where $y^\top g(x)$ naturally means the sum $\displaystyle\sum_{i=1}^m y_i g_i(x)$.

On can interpret (PR_y) as if one has put nonnegative "prices" y_i on the explicit constraints in (P) and transferred them to the objective function. The y_i's are usually called *Lagrange multipliers*.

The Lagrange relaxation method is usable in practice if the (PR_y) is substantially simpler to solve than (P).

Thus the basic rule, somewhat simplified, for determining which constraints should be explicit and which ones implicit is that the "hard" constraints must be explicit, while the "easy" constraints must be implicit. Then (PR_y) is a problem with only "easy" constraints.

Example 22.1. Consider that the original problem (P) is of the following form:

$$P : \begin{cases} \text{minimize} & \displaystyle\sum_{j=1}^{n} f_j(x_j), \\ \text{subject to} & \displaystyle\sum_{j=1}^{n} g_{ij}(x_j) \le b_i,\ i = 1, \ldots, m, \\ & x_j^{\min} \le x_j \le x_j^{\max},\ j = 1, \ldots, n, \end{cases}$$

where the f_j and g_{ij} are given convex functions and the $x_j^{\min}, x_j^{\max}, b_i$ are given constants. If we define

$$X = \{x \in \mathbb{R}^n : x_j^{\min} \le x_j \le x_j^{\max},\ j = 1, \ldots, n\},$$

then the relaxed Lagrange problem is:

$$(PR_y) : \begin{cases} \text{minimize} & \displaystyle\sum_{j=1}^{n} f_j(x_j) + \sum_{i=1}^{m} y_i \left(\sum_{j=1}^{n} g_{ij}(x_j) - b_i \right), \\ \text{subject to} & x \in X, \end{cases}$$

which can equivalently be also written as

$$(PR_y) : \begin{cases} \text{minimize} & \displaystyle\sum_{j=1}^{n} \left(f_j(x_j) + \sum_{i=1}^{m} y_i g_{ij}(x_j) \right) - y^\top b, \\ \text{subject to} & x \in X. \end{cases}$$

It can be seen that (PR_y) can be transformed to the following n one variable problems for $j = 1, \ldots, n$:

$$\begin{aligned} \text{minimize} \quad & f_j(x_j) + \sum_{i=1}^{m} y_i g_{ij}(x_j), \\ \text{subject to} \quad & x \in [x_j^{\min}, x_j^{\max}]. \end{aligned}$$

Minimizing n convex one variable functions over a given interval is usually much easier than solving the original problem (P). $\Diamond$

22.2. Global optimality conditions

What is the use of Lagrange relaxation? The underlying motivation is the hope that by choosing a "right" price vector y, one can find an optimal solution to the (easy) problem (PR_y), which is also an optimal solution to the original (hard) problem (P). We shall see later on that this is actually possible in certain cases. However, first we need some additional definitions.

Definition 22.2. A function $L : X \times \mathbb{R}^m \to \mathbb{R}$, defined by

$$L(x, y) = f(x) + y^\top g(x) \quad (x \in X,\ y \in \mathbb{R}^m),$$

is called the *Lagrangian* associated with the problem (P).

Definition 22.3. A pair $(\widehat{x}, \widehat{y}) \in X \times \mathbb{R}^m$ is said to satisfy the *global optimality conditions* associated with (P) if

(1) $L(\widehat{x}, \widehat{y}) = \min\limits_{x \in X} L(x, \widehat{y})$,

(2) $g(\widehat{x}) \leq 0$,

(3) $\widehat{y} \geq 0$,

(4) $\widehat{y}^\top g(\widehat{x}) = 0$.

The condition (1) means that $\widehat{x}$ is an optimal solution to the Lagrange relation problem (PR_y). The condition (4) (in combination with (2) and (3)) means that for i, $\widehat{y}_i = 0$ or $g_i(\widehat{x}) = 0$. The terminology "global optimality conditions", is motivated by the following result.

Theorem 22.4. *If $(\widehat{x}, \widehat{y}) \in X \times \mathbb{R}^m$ satisfy the global optimality conditions associated with (P), then $\widehat{x}$ is an optimal solution to (P).*

Proof. The condition (2) guarantees that $\widehat{x}$ is a feasible solution to (P). Let x be another feasible solution to (P), that is, $x \in X$ and $g(x) \leq 0$. Then we have

$$\begin{array}{rcl} f(x) & \geq & f(x) + \widehat{y}^\top g(x) \quad (g(x) \leq 0 \text{ and using (3)}) \\ & \geq & f(\widehat{x}) + \widehat{y}^\top g(\widehat{x}) \quad (\text{from (1)}) \\ & = & f(\widehat{x}) \quad (\text{from (4)}). \end{array}$$

Thus for all feasible solutions x to (P), we have $f(x) \geq f(\widehat{x})$, and so $\widehat{x}$ is an optimal solution to the problem (P). □

Thus the global optimality conditions are *sufficient* for $\widehat{x}$ to be an optimal solution to (P). However, they are not *necessary*, that is, if $\widehat{x}$ is an optimal solution to (P), then it is not certain that there exists some $\widehat{y} \in \mathbb{R}^m$ such that $(\widehat{x}, \widehat{y})$ satisfy the global optimality conditions associated with (P). For certain classes of problems, the global optimality conditions are also necessary, that is $\widehat{x}$ is an optimal solution to (P) iff there exists a vector $\widehat{y}$ such that $(\widehat{x}, \widehat{y})$ satisfy the global optimality conditions associated with (P).

Example 22.5 (Convex, continuously differentiable functions)**.** Suppose that $X = \mathbb{R}^n$ and the functions $g_1, \dots, g_m$ and f are convex and continuously differentiable. For a given $\widehat{y} \geq 0$, the function

$$x \mapsto L(x, \widehat{y}) = f(x) + \widehat{y}^\top g(x) \quad (x \in \mathbb{R}^n)$$

is also convex and continuously differentiable in x. Thus, by Theorem 15.10, the condition (1) in the global optimality conditions is equivalent with

$$\nabla f(\widehat{x}) + \sum_{i=1}^{m} \widehat{y}_i \nabla g_i(\widehat{x}) = 0.$$

Consequently, in this case, the global optimality conditions are the same as the KKT-conditions.

So in addition to the assumptions above, if we assume also that there exists a point $x_0 \in \mathbb{R}^n$ such that $g_i(x_0) < 0$ for all $i = 1, \dots, m$, then the global optimality conditions are also necessary for $\widehat{x}$ to be an optimal solution to (P); see Definition 21.5 and Theorem 21.7. ◊

22.3. The dual problem

According to the Theorem 22.4 one way to solve the problem (P) is to find $(\widehat{x}, \widehat{y})$ satisfying the global optimality conditions associated with (P). Despite the fact that one can't always be sure that there actually exists a pair $(\widehat{x}, \widehat{y})$ satisfying the global optimality conditions, it can nevertheless be fruitful to seek such a pair. The search itself can give useful information about the original problem (P), for example it can yield lower bounds on the optimal value of (P). The question then arises: is there a systematic method for seeking a pair satisfying the global optimality conditions? A partial answer is given by a result which we shall soon see, namely that the only y which can appear as the "y-part" of a pair satisfying the global optimality conditions is the optimal solution to a certain *dual* problem to (P).

Let $\mathbb{R}^m_+ = \{y \in \mathbb{R}^m : y \geq 0\}$. For a somewhat simplified account, so that the most important ideas are highlighted, we shall assume in the remainder of this chapter that for each $y \in \mathbb{R}^m_+$, there always exists at least one optimal solution to the relaxed Lagrange problem.

The *dual objective function* $\varphi : \mathbb{R}^m_+ \to \mathbb{R}$ is defined as follows:

$$\varphi(y) = \min_{x \in X} \big(f(x) + y^\top g(x)\big) = \min_{x \in X} L(x, y).$$

Lemma 22.6. *If x is a feasible solution to (P) and $y \geq 0$ then $\varphi(y) \leq f(x)$.*

Proof. Let x be a feasible solution to (P), that is, $g(x) \leq 0$ and $x \in X$. Then we have

$$\varphi(y) \leq f(x) + y^\top g(x) \leq f(x),$$

where the first inequality follows from the definition of φ, while the second one is a result of the facts that $y \geq 0$ and $g(x) \leq 0$. □

Thus for each $y \geq 0$, $\varphi(y)$ gives a lower bound for the optimal value of the problem (P). The *dual problem* to (P) is the problem of finding the best (greatest) lower bound, that is, the following maximization problem:

$$(D): \begin{cases} \text{maximize} & \varphi(y), \\ \text{subject to} & y \geq 0. \end{cases} \tag{22.3}$$

Lemma 22.7. *If*

(1) *$\widehat{x}$ is a feasible solution to (P),*

(2) *$\widehat{y} \geq 0$ and*

(3) *$\varphi(\widehat{y}) = f(\widehat{x})$,*

then $\widehat{x}$ is an optimal solution to (P) and $\widehat{y}$ is an optimal solution to (D).

Proof. Let x be a feasible solution to (P) and y be a feasible solution to (D). From Lemma 22.6, we obtain $\varphi(y) \leq f(\widehat{x})$ and $\varphi(\widehat{y}) \leq f(x)$. If these are combined with $\varphi(\widehat{y}) = f(\widehat{x})$, we obtain that

$$\begin{aligned} &\varphi(y) \leq f(\widehat{x}) = \varphi(\widehat{y}) \text{ and} \\ &f(\widehat{x}) = \varphi(\widehat{y}) \leq f(x). \end{aligned}$$

But this means that $\widehat{y}$ is an optimal solution to (D) and $\widehat{x}$ is an optimal solution to (P). □

Now we can strengthen Theorem 22.4.

Theorem 22.8. $(\widehat{x}, \widehat{y}) \in X \times \mathbb{R}^m$ *satisfy the global optimality conditions associated with (P) iff*

(1) *$\widehat{x}$ is an optimal solution to (P),*

(2) *$\widehat{y}$ is an optimal solution to (D), and*

(3) *$\varphi(\widehat{y}) = f(\widehat{x})$.*

Proof. Suppose first that $(\widehat{x}, \widehat{y}) \in X \times \mathbb{R}^m$ satisfy the global optimality conditions associated with (P). Then we have already seen in Theorem 22.4 that $\widehat{x}$ is an optimal solution to (P). We also have that

$$\varphi(\widehat{y}) = \min_{x \in X} L(x, \widehat{y}) = L(\widehat{x}, \widehat{y}) = f(\widehat{x}) + \widehat{y}^\top g(\widehat{x}) = f(\widehat{x}) + 0 = f(\widehat{x}),$$

which proves (3). Also, by Lemma 22.7, $\widehat{y}$ is an optimal solution to (D).

Now suppose that $\widehat{x}$ is an optimal solution to (P), $\widehat{y}$ is an optimal solution to (D), and $\varphi(\widehat{y}) = f(\widehat{x})$. Now $\widehat{x}$ and $\widehat{y}$, being feasible solutions for (P) and (D), respectively, the conditions (2) and (3) of the global optimality conditions are satisfied. Also,

$$\varphi(\widehat{y}) = \min_{x\in X} \big(f(x) + \widehat{y}^\top g(x)\big) \leq f(\widehat{x}) + \widehat{y}^\top g(\widehat{x}) \leq f(\widehat{x}),$$

where the last inequality is a result of the facts that $\widehat{y} \geq 0$ and $g(\widehat{x}) \leq 0$. But we know that $\varphi(\widehat{y}) = f(\widehat{x})$, and so

$$\varphi(\widehat{y}) = f(\widehat{x}) + \widehat{y}^\top g(\widehat{x}) = f(\widehat{x}).$$

Consequently, $\widehat{y}^\top g(\widehat{x}) = 0$, that is, condition (4) of the global optimality conditions is satisfied, and also

$$\min_{x\in X} L(x,\widehat{y}) = \varphi(\widehat{y}) = f(\widehat{x}) + \widehat{y}^\top g(\widehat{x}) = L(\widehat{x},\widehat{y}),$$

that is, condition (1) of the global optimality conditions is satisfied. □

The next result says that the dual objective function is well-suited for maximization.

Theorem 22.9. *φ is a concave function (that is, $-\varphi$ is a convex function) on $\mathbb{R}^m_+$.*

Proof. Let $y, z \in \mathbb{R}^m_+$ and $t \in (0,1)$. Then

$$\begin{aligned}
&\varphi((1-t)y + tz)\\
&= \min_{x\in X}\left(f(x) + ((1-t)y + tz)^\top g(x)\right)\\
&= \min_{x\in X}\left((1-t)\big(f(x) + y^\top g(x)\big) + t\big(f(x) + z^\top g(x)\big)\right)\\
&\geq \min_{x\in X}\left((1-t)\big(f(x) + y^\top g(x)\big)\right) + \min_{x\in X}\left(t\big(f(x) + z^\top g(x)\big)\right)\\
&= (1-t)\min_{x\in X}\left(f(x) + y^\top g(x)\right) + t\min_{x\in X}\left(f(x) + z^\top g(x)\right)\\
&= (1-t)\varphi(y) + t\varphi(z).
\end{aligned}$$

Thus the result follows. □

Example 22.10. Consider the following problem in the variable vector $x \in \mathbb{R}^n$:

$$\begin{array}{rl}\text{minimize} & \frac{1}{2}x^\top x - c^\top x,\\ \text{subject to} & \frac{1}{2}x^\top x - a^\top x \leq 0,\end{array}$$

where $c, a \in \mathbb{R}^n$ are given, and satisfy $a^\top a = 1$, $c^\top c = 1$, $a^\top c = 0$. The Lagrangian for this problem is given by

$$\begin{aligned}
L(x,y) &= \frac{1}{2}x^\top x - c^\top x + y\left(\frac{1}{2}x^\top x - a^\top x\right)\\
&= \frac{1+y}{2}x^\top x - (c + ya)^\top x \quad (x \in \mathbb{R}^n,\ y \in \mathbb{R}).
\end{aligned}$$

The relaxed Lagrange problem (PR_y) is: given $y \geq 0$, minimize the function $x \mapsto L(x,y)$ on $\mathbb{R}^n$. In our case, the optimal solution to (PR_y) is given by

$$\widehat{x}(y) = \frac{1}{1+y}(c + ya).$$

(See Theorem 9.13.) The dual objective function is given by

$$\varphi(y) = L(\widehat{x}(y), y) = -\frac{(c+ya)^\top (c+ya)}{2(1+y)} = \frac{1+y^2}{2(1+y)},$$

where we have used the relations $a^\top a = 1$, $c^\top c = 1$, and $a^\top c = 0$. The dual problem is that of maximizing the map $y \mapsto \varphi(y)$ subject to $y \geq 0$. But

$$\varphi'(y) = \frac{1-2y-y^2}{2(1+y)^2}.$$

If $y \geq 0$, then $\varphi'(y) = 0$ iff $y = \widehat{y} := \sqrt{2} - 1$. Now let

$$\widehat{x} := \widehat{x}(\widehat{y}) = \frac{1}{\sqrt{2}}(c + (\sqrt{2}-1)a).$$

Then it can be verified that the following hold:

(1) $\widehat{x}$ minimizes the map $x \mapsto L(x, \widehat{y})$, since $\widehat{x} = \widehat{x}(\widehat{y})$.

(2) $\widehat{x}$ is a feasible solution to the primal problem.

(3) $\widehat{y} \geq 0$.

(4) $\widehat{y}\left(\frac{1}{2}\widehat{x}^\top \widehat{x} - a^\top \widehat{x}\right) = 0$.

Thus $(\widehat{x}, \widehat{y})$ satisfy the global optimality conditions, which implies that $\widehat{x}$ is an optimal solution to the original (primal) problem. ♢

Exercise 22.11. Suppose that a_j, $j = 1, \dots, n$, and b are given positive constants. Solve the following optimization problem and explain why your solution is globally optimal.

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^n x_j^2, \\ \text{subject to} \quad & \sum_{j=1}^n a_j x_j \geq b, \\ & x_j \geq 0,\ j = 1, \dots, n. \end{aligned}$$

Exercise 22.12. Suppose that a_j, $j = 1, \dots, n$, and b are given positive constants. Solve the following optimization problem and explain why your solution is globally optimal.

$$\begin{aligned} \text{maximize} \quad & \sum_{j=1}^n \log x_j, \\ \text{subject to} \quad & \sum_{j=1}^n a_j x_j \leq b, \\ & x_j > 0,\ j = 1, \dots, n. \end{aligned}$$

Exercise 22.13. Suppose that a_j, b_j, for $j = 1, \dots, n$, and b_0 are given positive constants. Solve the following optimization problem and explain why your solution is globally optimal.

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^n a_j x_j, \\ \text{subject to} \quad & \sum_{j=1}^n \frac{b_j}{x_j} \leq b_0, \\ & x_j > 0,\ j = 1, \dots, n. \end{aligned}$$

Exercise 22.14. Suppose that a_j, c_j, for $j = 1, \dots, n$, and b are given positive constants. Solve the following optimization problem and explain why your solution is globally optimal.

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{n} e^{c_j x_j}, \\ \text{subject to} \quad & \sum_{j=1}^{n} a_j x_j \geq b, \\ & x_j > 0,\ j = 1, \dots, n. \end{aligned}$$

Exercise 22.15. Suppose that a_j, $j = 1, \dots, n$, and b are given positive constants. Consider the following optimization problem (P):

$$(P) : \left\{ \begin{aligned} \text{minimize} \quad & \sum_{j=1}^{n} x_j^3, \\ \text{subject to} \quad & \sum_{j=1}^{n} a_j x_j \geq b, \\ & x_j \geq 0,\ j = 1, \dots, n. \end{aligned} \right.$$

Find the Lagrange dual problem (D) to (P) when the "sum constraint" is relaxed. Find an optimal solution to (D). Hence find an optimal solution to (P).

Exercise 22.16. Consider the following optimization problem (P):

$$(P) : \left\{ \begin{array}{rl} \text{minimize} & x_1^4 + 2x_1x_2 + x_2^2 + x_3^8, \\ \text{subject to} & (x_1 - 2)^2 + (x_2 - 2)^2 + (x_3 - 3)^2 \leq 6, \\ & x_1x_2x_3 \leq 10, \\ & x_1 \geq 1, \\ & x_2 \geq 0, \\ & x_3 \geq 0. \end{array} \right.$$

Use the Lagrange relaxation method to show that $\widehat{x} = (1, 1, 1) \in \mathbb{R}^3$ is a global optimal solution to (P).

Hint: Take $X = \mathbb{R}^3$. Find a $\widehat{y}$ such that the global optimality conditions are satisfied by the pair $(\widehat{x}, \widehat{y})$.

Exercise 22.17. Consider the following optimization problem:

$$(P) : \left\{ \begin{array}{rl} \text{minimize} & f(x), \\ \text{subject to} & g(x) \leq 0, \\ & x \in X, \end{array} \right.$$

where g has m components $g_1, \dots, g_m$. Let $\varphi : \mathbb{R}^m_+ \to (\mathbb{R} \cup \{-\infty\})$ be defined as follows:

$$\varphi(y) = \inf_{x \in X} (f(x) + y^\top g(x)) = \inf_{x \in X} L(x, y).$$

Check that Lemma 22.6 continues to hold with this φ.

Now consider the following dual optimization problem:

$$(D) : \left\{ \begin{array}{rl} \text{maximize} & \varphi(y), \\ \text{subject to} & y \geq 0. \end{array} \right.$$

Verify that Lemma 22.7 and Theorem 22.8 still hold.

Consider the linear programming problem (LP)

$$(LP) : \left\{ \begin{array}{rl} \text{minimize} & c^\top x, \\ \text{subject to} & Ax \geq b, \\ & x \geq 0, \end{array} \right.$$

(where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$), as a special instance of the optimization problem (P). Write the dual optimization problem to (LP) and check that you get the dual program (D) as in (6.2) of Chapter 6.

Exercise 22.18. Consider the problem given in Exercise 21.18 with $c = -6$. Find the value $\varphi(y)$ of the objective function φ of the dual problem (D) to the given problem (P) when $y = (1, 1, 1)$. Is this y optimal for the dual problem?

Hint: When $c = -6$ in the problem (P) and $x = (1, 1, 1)$ is optimal for (P), what is the corresponding value of y obtained from the KKT-conditions in Exercise 21.18?

Exercise 22.19. Consider the problem (P_c) given in Exercise 21.19. Show that the dual problem (D_c) to the problem (P_c) is:

$$(D_c) : \begin{cases} \text{maximize} & -y_1^2 + 2y_1 - \frac{y_2^2}{4} + y_2 - 4, \\ \text{subject to} & y_1 \geq 0, \\ & y_2 \geq 0. \end{cases}$$

Find a global optimal solution to (D_c). Justify your answer.

Hint: Use the solution to Exercise 21.19.

Part 4

Some Linear Algebra

Chapter 23

Subspaces

Recall that a vector space is, roughly speaking, a set of elements (called "vectors"), such that any two vectors can be "added", resulting in a new vector, and any vector can be multiplied by an element from $\mathbb{R}$ so as to give a new vector. The precise definition is recalled below.

Definition 23.1. A *vector space* V is a set together with two functions, $+ : V \times V \to V$, called *vector addition*, and $\cdot : \mathbb{R} \times V \to V$, called *scalar multiplication*, such that the following hold:

(V1) For all $v_1, v_2, v_3 \in V$, $v_1 + (v_2 + v_3) = (v_1 + v_2) + v_3$.

(V2) There exists an element $0 \in V$ such that for all $v \in V$, $v + 0 = v = 0 + v$. (The element 0 is called the *zero vector*.)

(V3) For each element $v \in V$, there exists a unique element in V, denoted by $-v$, such that $v + (-v) = 0 = -v + v$.

(V4) For all $v_1, v_2 \in V$, $v_1 + v_2 = v_2 + v_1$.

(V5) For all $v \in V$, $1 \cdot v = v$.

(V6) For all $\alpha, \beta \in \mathbb{R}$ and all $v \in V$, $\alpha \cdot (\beta \cdot v) = (\alpha\beta) \cdot v$.

(V7) For all $\alpha, \beta \in \mathbb{R}$ and all $v \in V$, $(\alpha + \beta) \cdot v = \alpha \cdot v + \beta \cdot v$.

(V8) For all $\alpha \in \mathbb{R}$ and all $v_1, v_2 \in V$, $\alpha \cdot (v_1 + v_2) = \alpha \cdot v_1 + \alpha \cdot v_2$.

23.1. Definition of a subspace

A subset S of a vector space V is called a *subspace* if

S1. $0 \in S$.

S2. If $v_1, v_2 \in S$, then $v_1 + v_2 \in S$.

S3. If $v \in S$ and $\alpha \in \mathbb{R}$, then $\alpha \cdot v \in S$.

From this definition, it follows that if the vectors $v_1, \dots, v_k$ belong to a subspace S, then every linear combination of these vectors also belongs to S.

Exercise 23.2. Show that if $X \subset \mathbb{R}^n$, then the intersection of all subspaces containing X is a subspace of $\mathbb{R}^n$, and it is the smallest subspace of $\mathbb{R}^n$ that contains X. This subspace is called the *span of* X, and is denoted by $\operatorname{span} X$. What is $\operatorname{span} \emptyset$? If the set $X = \{v_1, \dots, v_k\}$, then show that

$$\operatorname{span} X = \{\alpha_1 v_1 + \cdots + \alpha_k v_k : \alpha_1, \dots, \alpha_k \in \mathbb{R}\}.$$

Exercise 23.3. Show that the set $S = \{A \in \mathbb{R}^{n \times n} : A^\top = A\}$ of all symmetric $n \times n$ matrices is a subspace of the vector space $\mathbb{R}^{n \times n}$.

23.2. Basis for a subspace

A basis of a subspace S is a subset B of S such that $\operatorname{span} B = S$ and B is linearly independent.

If $B = \{v_1, \dots, v_k\}$ is a basis of S, then every vector $v \in S$ can be written in a *unique* way as a linear combination of the basis vectors $v_1, \dots, v_k$. The fact that every vector *can* be written as a linear combination of the basis vectors follows from the spanning property of B, and the *uniqueness* of the associated scalars follows from the linear independence of B.

Exercise 23.4. Show that $B = \emptyset$ is a basis for the subspace $S = \{0\}$.

Exercise 23.5. Find a basis for the subspace S of $\mathbb{R}^{n\times n}$ consisting of all symmetric $n \times n$ matrices.

23.3. Dimension of a subspace

Although there can be many different bases for the same subspace, it turns out that the number of vectors in each basis for the same subspace cannot change. Thus, if $B = \{v_1, \dots, v_k\}$ and $B' = \{u_1, \dots, u_p\}$ are two different bases for the same subspace, then $k = p$. The number of vectors in the basis (which is now a well-defined notion) of a subspace S is called the *dimension* of the subspace, denoted by $\dim S$.

Recall the following important result from linear algebra:

Theorem 23.6. *Suppose that S is a subspace of dimension k. Let $v_1, \dots, v_k$ be vectors from S. Then the following are equivalent:*

(1) *$v_1, \dots, v_k$ are linearly independent.*

(2) *The span of $v_1, \dots, v_k$ is S.*

(3) *$\{v_1, \dots, v_k\}$ is a basis for S.*

Proof. See for example [**T**]. □

A consequence of this result is the following: if one wants to find a basis for a subspace S which we know has dimension k, it suffices to find k linearly independent vectors in S.

Exercise 23.7. What is the dimension of the subspace $S = \{0\}$?

Exercise 23.8. What is the dimension of the subspace S of $\mathbb{R}^{n\times n}$ consisting of all symmetric $n \times n$ matrices.

Exercise 23.9. Suppose that S_1, S_2 are finite-dimensional subspaces of a vector space V. Show that $\dim(S_1 + S_2) + \dim(S_1 \bigcap S_2) = \dim S_1 + \dim S_2$.

23.4. Orthogonal complement

If $X \subset \mathbb{R}^n$, then we define the *orthogonal complement of* X to be the set

$$X^\perp = \{y \in \mathbb{R}^n : y^\top x = 0 \text{ for all } x \in X\}.$$

It is easy to check that $X^\perp$ is a subspace of $\mathbb{R}^n$.

One can show the following important result:

Theorem 23.10. *Let S be a subspace of $\mathbb{R}^n$. Then:*

(1) *For each $x \in \mathbb{R}^n$, there exists a unique $z \in S$ and a unique $y \in S^\perp$ such that $x = z + y$.*

(2) *$(S^\perp)^\perp = S$.*

(3) *If $\dim S = k$, then $\dim(S^\perp) = n - k$.*

Proof. See for example [**T**]. □

Chapter 24

Four fundamental subspaces

Let $A \in \mathbb{R}^{m\times n}$ be a given matrix, and let c_j denote the jth column of A and let r_i denote the ith row of A, that is,

$$A = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{m1} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} c_1 & \dots & c_n \end{bmatrix} = \begin{bmatrix} r_1 \\ \vdots \\ r_m \end{bmatrix}.$$

The transposed matrix $A^\top$ is then given by

$$A = \begin{bmatrix} a_{11} & \dots & a_{m1} \\ \vdots & & \vdots \\ a_{1n} & \dots & a_{mn} \end{bmatrix} = \begin{bmatrix} c_1^\top \\ \vdots \\ c_n^\top \end{bmatrix} = \begin{bmatrix} r_1^\top & \dots & r_m^\top \end{bmatrix}.$$

There are four natural subspaces associated with A:

the column space of A (the range of A),
the row space of A (the range of $A^\top$),
the kernel of A,
the left kernel of A (the kernel of $A^\top$).

The row space and the kernel are subspaces of $\mathbb{R}^n$, while the range space and the left kernel are subspaces of $\mathbb{R}^m$.

24.1. Column space of A

The *column space* or the *range* of A is the following subspace of $\mathbb{R}^m$:

$$\operatorname{ran} A = \{Ax : x \in \mathbb{R}^n\}.$$

In other words, it is the range of the linear transformation given by matrix multiplication by A, that is, the range of the map $x \mapsto Ax$ from $\mathbb{R}^n$ to $\mathbb{R}^m$. An equivalent way of expressing this is that $y \in \operatorname{ran} A$ iff the equation $Ax = y$ has a solution $x \in \mathbb{R}^n$.

Since Ax is the linear combination of the columns of the matrix A by scalars which are the components of x, we have

$$\operatorname{ran} A = \left\{ \sum_{j=1}^{n} x_j c_j : x_j \in \mathbb{R},\ j = 1, \dots, n \right\}.$$

This is the reason behind calling $\operatorname{ran} A$ as the column space of A.

If the columns of A are linearly independent, then they form a basis for $\operatorname{ran} A$. (Why?) In this special case, $\dim(\operatorname{ran} A) = n$.

24.2. Row space of A

The *row space* of A is the column space of $A^\top$, that is, it is the range of $A^\top$. Hence the row space of A is the following subspace of $\mathbb{R}^n$:

$$\operatorname{ran} A^\top = \{A^\top y : y \in \mathbb{R}^m\}.$$

An equivalent way of expressing this is that $x \in \operatorname{ran} A^\top$ iff the equation $A^\top y = x$ has a solution $y \in \mathbb{R}^m$.

Since $A^\top y$ is the linear combination of the columns of the matrix $A^\top$ by scalars which are the components of y, we have

$$\operatorname{ran} A^\top = \left\{ \sum_{i=1}^{m} y_i r_i^\top : y_i \in \mathbb{R},\ i = 1, \dots, m \right\}.$$

Note that in the above, the $r_i^\top$ are the columns of $A^\top$ which are the same as the transposed rows of A. This explains why we call $\operatorname{ran} A^\top$ as the row space of A.

If the rows of A are linearly independent, then the columns of $A^\top$ are linearly independent, and form a basis for $\operatorname{ran} A^\top$. In this special case, $\dim(\operatorname{ran} A^\top) = m$.

24.3. Kernel of A

The *kernel* of A is the subspace of $\mathbb{R}^n$ given by

$$\ker A = \{x \in \mathbb{R}^n : Ax = 0\}.$$

Since the ith entry of Ax is $(Ax)_i = r_i x$, it follows that $\ker A$ is the the set of vectors x which are orthogonal to every row of A:

$$\ker A = \{x \in \mathbb{R}^n : r_i x = 0,\ i = 1, \dots, m\}.$$

If the columns of A are linearly independent, then $Ax = 0$ iff $x = 0$. In this special case, $\ker A = \{0\}$ and $\dim(\ker A) = 0$.

24.4. Left kernel of A

The *left kernel* of A is the kernel of $A^\top$, that is, it is the subspace of $\mathbb{R}^m$ given by

$$\ker A^\top = \{y \in \mathbb{R}^m : A^\top y = 0\}.$$

By taking transposes, $A^\top y = 0$ iff $y^\top A = 0$. Thus

$$\ker A^\top = \{y \in \mathbb{R}^m : y^\top A = 0\}.$$

So we see that $y \in \ker A^\top$ iff the transpose of y when multiplied from the *left* with A, gives the zero vector. This explains why $\ker A^\top$ is called the left kernel of A.

Since the jth entry of $A^\top y$ is $(A^\top y)_j = c_j^\top y$, it follows that $\ker A^\top$ is the the set of vectors y which are orthogonal to every column of A:

$$\ker A^\top = \{y \in \mathbb{R}^m : c_j^\top y = 0,\ j = 1, \dots, n\}.$$

If the rows of A are linearly independent, then $A^\top y = 0$ iff $y = 0$. In this special case, $\ker A^\top = \{0\}$ and $\dim(\ker A^\top) = 0$.

24.5. Orthogonality relations

Theorem 24.1. $(\operatorname{ran} A)^\perp = \ker A^\top$.

Proof. Suppose that $z \in \ker A^\top$. Then $A^\top z = 0$, or equivalently $z^\top A = 0$. So for all $x \in \mathbb{R}^n$, $z^\top Ax = 0x = 0$. But this implies that for all $y \in \operatorname{ran} A$, $z^\top y = 0$. In other words, $z \in (\operatorname{ran} A)^\perp$. So we have proved that $\ker A^\top \subset (\operatorname{ran} A)^\perp$.

We now prove the reverse inclusion. So let $z \in (\operatorname{ran} A)^\perp$. Then for all $y \in \operatorname{ran} A$, $z^\top y = 0$. In other words, for all $x \in \mathbb{R}^n$, $z^\top Ax = 0$. Taking successively $x = e_i$, $i = 1, \dots, n$, where the e_i's denote the standard basis vectors, we obtain that all the components $(z^\top A)_i = 0$, for $i = 1, \dots, n$, of $z^\top A$ are zeros. Hence $z^\top A = 0$, or equivalently $A^\top z = 0$. So $z \in \ker A^\top$. Consequently, also $(\operatorname{ran} A)^\perp \subset \ker A^\top$. □

Taking orthogonal complements, we also have

$$\operatorname{ran} A = ((\operatorname{ran} A)^\perp)^\perp = (\ker A^\top)^\perp.$$

Applying these results to $A^\top$, we obtain

$$(\operatorname{ran} A^\top)^\perp = \ker(A^\top)^\top = \ker A,$$

and

$$\operatorname{ran} A^\top = (\ker(A^\top)^\top)^\perp = (\ker A)^\perp.$$

Summarizing, we have the following four relations:

$$\begin{aligned} (\operatorname{ran} A)^\perp &= \ker A^\top, \\ (\operatorname{ran} A^\top)^\perp &= \ker A, \\ \operatorname{ran} A &= (\ker A^\top)^\perp, \\ \operatorname{ran} A^\top &= (\ker A)^\perp. \end{aligned}$$

Exercise 24.2.

(1) If $A \in \mathbb{R}^{m\times n}$, then prove that ran $A =$ ran $AA^\top$.

(2) If $H \in \mathbb{R}^{n\times n}$ is symmetric, then show that $\ker H$ and ran H are orthogonal, that is, there holds that $(\ker H)^\perp =$ ran H.

24.6. Dimension relations

There is an interesting connection between the dimensions of the four fundamental subspaces. First of all, we recall the so-called *rank-nullity theorem*, saying that if $A \in \mathbb{R}^{m\times n}$, then

$$\dim(\operatorname{ran} A) + \dim(\ker A) = n.$$

See for example [**T**]. Applying this to $A^\top$, we obtain $\dim(\operatorname{ran} A^\top) + \dim(\ker A^\top) = m$. Using the orthogonality relations established in the previous section, and the fact $\dim S^\perp = d - \dim S$ for any subspace S of $\mathbb{R}^d$, we obtain the following: if r denotes the dimension of $\operatorname{ran} A$ (this is also called the *rank* of A), then

$$\begin{aligned} &\dim(\operatorname{ran} A) = r, \\ &\dim(\operatorname{ran} A^\top) = r, \\ &\dim(\ker A) = n - r, \\ &\dim(\ker A^\top) = m - r. \end{aligned}$$

A consequence of these relations are the following equivalences:

(1) The columns of A span $\mathbb{R}^m$ iff the columns of $A^\top$ are linearly independent (iff $r = m$).

(2) The columns of $A^\top$ span $\mathbb{R}^n$ iff the columns of A are linearly independent (iff $r = n$).

Exercise 24.3. Prove the equivalences in (1) and (2) above.

Chapter 25

Bases for fundamental subspaces

In this chapter, we will show how we can determine bases for the four fundamental subspaces. En route we will see a factorization method based on the Gauss-Jordan method. However, we will begin with a theoretical result.

25.1. Result on ranges in a factorization

Theorem 25.1. *Let $A \in \mathbb{R}^{m\times n}$ and $A = BC$, where $B \in \mathbb{R}^{m\times r}$ has linearly independent columns and $C \in \mathbb{R}^{r\times n}$ has linearly independent rows. Then A and $A^\top$ both have ranges of dimension r. Furthermore,*

$$\begin{aligned} \ker A &= \ker C, && (25.1)\\ \ker A^\top &= \ker B^\top, && (25.2)\\ \operatorname{ran} A &= \operatorname{ran} B, && (25.3)\\ \operatorname{ran} A^\top &= \operatorname{ran} C^\top. && (25.4) \end{aligned}$$

Proof. Let $x \in \ker A$, that is, $Ax = 0$. Then $BCx = 0$, that is, $B(Cx) = 0$, and since B has linearly independent columns, it follows that $Cx = 0$, that is, $x \in \ker C$. Hence $\ker A \subset \ker C$. On the other hand, if $x \in \ker C$, then $Cx = 0$, so $Ax = BCx = B0 = 0$. Thus $x \in \ker A$. Hence $\ker C \subset \ker A$. This proves (25.1).

Since $A = BC$, we have $A^\top = C^\top B^\top$, and (25.2) now follows from an application of (25.1), with $A^\top$ replacing A, $C^\top$ replacing B, and $B^\top$ replacing C.

We have $\operatorname{ran} A = (\ker A^\top)^\perp = (\ker B^\top)^\perp = \operatorname{ran} B$, and so we obtain (25.3).

Similarly, $\operatorname{ran} A^\top = (\ker A)^\perp = (\ker C)^\perp = \operatorname{ran} C^\top$, and so we obtain (25.4).

It remains to show that A and $A^\top$ have rank r. Since B has linearly independent columns, and since these span $\operatorname{ran} B$, these columns form a basis for $\operatorname{ran} B$. Consequently, we have that $\dim(\operatorname{ran} A) = \dim(\operatorname{ran} B) = r$, that is, A has rank r. But now apply this result to the factorization $A^\top = C^\top B^\top$, we also obtain that $\dim(\operatorname{ran} A^\top) = r$. □

Also the converse to the above theorem is true: If $A \in \mathbb{R}^{m\times n}$ has rank r, then there is a factorization $A = BC$, where $B \in \mathbb{R}^{m\times r}$ has linearly independent columns and $C \in \mathbb{R}^{r\times n}$ has linearly independent rows. We will prove this in the next section, and we will also learn a method for constructing such B and C starting from the matrix A.

25.2. Factorization using Gauss-Jordan

In the Gauss-Jordan method, a sequence of elementary row operations is carried out, which transform a matrix $A \in \mathbb{R}^{m\times n}$ to a matrix $T \in \mathbb{R}^{m\times n}$ which has a "staircase form". We recall this method quickly by carrying out an example.

Example 25.2 (The Gauss-Jordan method)**.** Suppose that

$$A = \begin{bmatrix} 1 & 2 & 4 & 5 & -3 \\ 2 & 4 & 3 & 5 & -1 \\ 3 & 6 & 2 & 5 & 1 \\ 4 & 8 & 1 & 5 & 3 \end{bmatrix}.$$

We shall use the Gauss-Jordan method for constructing T.

We do the following:

(1) add -2 times the first row to the second row,

(2) add -3 times the first row to the third row, and

(3) add -4 times the first row to the fourth row.

We then obtain

$$\begin{bmatrix} 1 & 2 & 4 & 5 & -3 \\ 0 & 0 & -5 & -5 & 5 \\ 0 & 0 & -10 & -10 & 10 \\ 0 & 0 & -15 & -15 & 15 \end{bmatrix}.$$

Multiplying the second row by $-\frac{1}{5}$, we obtain

$$\begin{bmatrix} 1 & 2 & 4 & 5 & -3 \\ 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & -10 & -10 & 10 \\ 0 & 0 & -15 & -15 & 15 \end{bmatrix}.$$

Now we do the following:

(1) add -4 times the second row to the first row,

(2) add 10 times the second row to the third row, and

(3) add 15 times the second row to the fourth row.

We then obtain

$$T = \begin{bmatrix} 1 & 2 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & -1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Now we see that A has been transformed by elementary row operations into a "staircase matrix form" with two steps, namely the 1s in the first column and the third column. $\Diamond$

The sequence of allowed operations to take A into T corresponds to left multiplication by an invertible matrix P: thus $PA = T$ or $A = P^{-1}T$. Let R be the number of "steps" in the matrix T, and let $\ell := m - r$ and $k := n - r$. Each of the r special columns of T corresponding to these steps have exactly one entry equal to 1 and the others are all 0's. The other $\ell = m - r$ rows of T consist of only zeros. Let $U \in \mathbb{R}^{r\times n}$ be the matrix which is obtained by deleting these ℓ rows from T, and let $0_{\ell\times n}$ denote the zero matrix consisting of precisely these deleted rows of zeros. Then

T can be written as a block matrix with the two blocks U and $0_{\ell\times n}$, that is:

$$T = \left[\begin{array}{c} U \\ 0_{\ell\times n} \end{array}\right].$$

If A has the columns $a_1, \dots, a_n$, then let $\beta_1 < \dots < \beta_r$ be the indices of the columns corresponding to the steps of T. Let $\nu_1 < \dots < \nu_k$ be the indices corresponding to the remaining columns. Let $u_1, \dots, u_n$ be the columns of U. Define $A_\beta \in \mathbb{R}^{m\times r}$ to be the matrix with the columns $a_{\beta_1}, \dots, a_{\beta_r}$. Let $U_\beta \in \mathbb{R}^{r\times r}$ be the matrix with the columns $u_{\beta_1}, \dots, u_{\beta_r}$ and $U_\nu \in \mathbb{R}^{r\times r}$ be the matrix with the columns $u_{\nu_1}, \dots, u_{\nu_k}$. That is,

$$\begin{array}{rcl} A_\beta &=& \left[\begin{array}{ccc} a_{\beta_1} & \dots & a_{\beta_r} \end{array}\right], \quad \text{and} \\ U_\beta &=& \left[\begin{array}{ccc} u_{\beta_1} & \dots & u_{\beta_r} \end{array}\right] = \left[\begin{array}{ccc} e_1 & \dots & e_r \end{array}\right] = I_{r\times r}, \end{array}$$

where $e_1, \dots, e_r$ denote the standard basis vectors of $\mathbb{R}^r$. Furthermore, we have

$$PA = T = \left[\begin{array}{c} U \\ 0_{\ell\times n} \end{array}\right],$$

and so

$$PA_\beta = \left[\begin{array}{c} U_\beta \\ 0_{\ell\times n} \end{array}\right] = \left[\begin{array}{c} I_{r\times r} \\ 0_{\ell\times n} \end{array}\right].$$

From this, it follows that the columns of A_β are linearly independent: indeed, if y is such that $A_\beta y = 0$, then

$$0 = P0 = PA_\beta y = \left[\begin{array}{c} I_{r\times r} \\ 0_{\ell\times n} \end{array}\right] y = \left[\begin{array}{c} y \\ 0 \end{array}\right],$$

which implies that $y = 0$.

Moreover, the rows of U are linearly independent, since if y is such that $y^\top U = 0$, then in particular, $y^\top U_\beta = 0$, that is, $y^\top I_{r\times r} = 0$, and so $y = 0$.

We have

$$A = P^{-1}T = P^{-1}\left[\begin{array}{c} U \\ 0_{\ell\times n} \end{array}\right].$$

Partition $P^{-1} \in \mathbb{R}^{m\times m}$ into two blocks $S_1 \in \mathbb{R}^{m\times r}$ and $S_2 \in \mathbb{R}^{m\times \ell}$, consisting respectively of the first r and the last ℓ columns of P^{-1}. Then we obtain that

$$A = P^{-1}\left[\begin{array}{c} U \\ 0_{\ell\times n} \end{array}\right] = \left[\begin{array}{cc} S_1 & S_2 \end{array}\right]\left[\begin{array}{c} U \\ 0_{\ell\times n} \end{array}\right] = S_1U + S_2 0_{\ell\times n} = S_1U.$$

Since $A = S_1U$, it follows in particular that

$$A_\beta = S_1U_\beta = S_1 I_{r\times r} = S_1.$$

Hence

$$A = A_\beta U. \tag{25.5}$$

By the Theorem 25.1, this means that the matrix A and $A^\top$ have rank r. The rank of A thus coincides with the number of steps in the staircase matrix T. The factorization (25.5) implies that the converse of Theorem 25.1 is true:

Theorem 25.3. *If $A \in \mathbb{R}^{m\times n}$ has rank r, then A can be factorized as $A = BC$, where $B \in \mathbb{R}^{m\times r}$ has linearly independent columns and $C \in \mathbb{R}^{r\times n}$ has linearly independent rows.*

Proof. Just take $B = A_\beta$ and $C = U$ as above. □

Exercise 25.4. Find a factorization $A = BC$ as in the statement of Theorem 25.3 of A, when A is given by:

$$A = \left[\begin{array}{ccc} 2 & 1 & -1 \\ 3 & 1 & 2 \end{array}\right], \quad A = \left[\begin{array}{cc} 1 & 1 \\ 2 & 0 \\ 3 & -1 \end{array}\right].$$

25.3. Basis for ran A and ran $A^\top$

From the factorization given by (25.5), and the Theorem 25.1, it follows immediately that

$$\begin{aligned} \operatorname{ran} A &= \operatorname{ran} A_\beta, \\ \operatorname{ran} A^\top &= \operatorname{ran} U^\top. \end{aligned}$$

This implies that the columns $a_{\beta_1}, \dots, a_{\beta_r}$ of the matrix A_β is a basis for ran A, and the columns of $U^\top$ form a basis for ran $A^\top$.

Exercise 25.5. Find a basis for ran A and a basis for ran $A^\top$ when A is given by:

$$A = \begin{bmatrix} 2 & 1 & -1 \\ 3 & 1 & 2 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \\ 3 & -1 \end{bmatrix}.$$

25.4. Basis for $\ker A$

We have that $\ker A = \ker U$ by Theorem 25.1. Also, we know from the rank-nullity theorem that $\dim(\ker A) = n - r = k$. So for determining a basis for $\ker A$, we should find k linearly independent vectors from $(\ker A =) \ker U$.

If $x \in \ker U$, then $Ux = 0$, that is, $\displaystyle\sum_{i=1}^{r} x_{\beta_i} u_{\beta_i} + \sum_{i=1}^{k} x_{\nu_i} u_{\nu_i} = 0$. But since $U_\beta = I_{r\times r}$, this means that $x_\beta + U_\nu x_\nu = 0$, where $x_\beta \in \mathbb{R}^r$ denotes the column vector with the components $x_{\beta_1}, \dots, x_{\beta_r}$, and x_ν denotes the column vector with the components $x_{\nu_1}, \dots, x_{\nu_k}$. So $x_\beta = -U_\nu x_\nu$.

Now suppose that $x \in \mathbb{R}^n$ is such that $x_\beta = -U_\nu x_\nu$. Then reversing the steps in the calculation of the previous paragraph, we conclude that $x \in \ker U$.

So we have shown that $x \in \ker A$ iff $x_\beta = -U_\nu x_\nu$.

Now we determine for $j = 1, \dots, k$, a $z_j \in \ker A$ as follows. We set $x_\nu := e_j$ and

$$x_\beta := -U_\nu x_\nu = -U_\nu e_j = -u_{\nu_j}.$$

And then, having determined x_ν and x_β, we write the corresponding x, which we define to be our sought after z_j. This z_j now belongs to $\ker A$, by construction.

Now we show that the vectors $z_1, \dots, z_k$ are linearly independent. Let $t_1, \dots, t_k$ be given scalars and let $w := t_1 z_1 + \cdots + t_k z_k$. If w_{ν_j} denotes the ν_jth component of the vector $w \in \mathbb{R}^n$, then we have for $j = 1, \dots, k$, that $w_{\nu_j} = t_j$. (Why?) Hence it follows that if $w = 0$, then $t_j = 0$ for $j = 1, \dots, k$.

Consequently, the vectors $z_1, \dots, z_k$ form a basis for $\ker A$.

Exercise 25.6. Find a basis for $\ker A$ when A is given by:

$$A = \begin{bmatrix} 2 & 1 & -1 \\ 3 & 1 & 2 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \\ 3 & -1 \end{bmatrix}.$$

25.5. Basis for $\ker A^\top$

By now we know how we can find a basis for $\ker A$ by first transforming A into a matrix having a staircase form with the Gauss-Jordan method, and then proceeding as described in the previous section. But we can equally well begin with the matrix $A^\top$ instead of A, transform $A^\top$ into a matrix having a staircase form with the Gauss-Jordan method, and then proceed as described in the previous section to obtain a basis for $\ker A^\top$. En route we get a new basis for ran $A^\top$ and also a new basis for $\operatorname{ran}(A^\top)^\top = \operatorname{ran} A$.

Exercise 25.7. Find a basis for $\ker A^\top$ when A is given by:

$$A = \begin{bmatrix} 2 & 1 & -1 \\ 3 & 1 & 2 \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 1 \\ 2 & 0 \\ 3 & -1 \end{bmatrix}.$$

25.6. An example

We revisit Example 25.2, and determine bases for the four fundamental subspaces.

Example 25.8 (Example 25.2 continued)**.** Recall that in Example 25.2,

$$A = \begin{bmatrix} 1 & 2 & 4 & 5 & -3 \\ 2 & 4 & 3 & 5 & -1 \\ 3 & 6 & 2 & 5 & 1 \\ 4 & 8 & 1 & 5 & 3 \end{bmatrix}.$$

By elementary row transformations, we had arrived at the following matrix T having a staircase form:

$$T = \left[\begin{array}{ccccc} 1 & 2 & 0 & 1 & 1 \\ \cline{1-2} 0 & 0 & \multicolumn{1}{|c}{1} & 1 & -1 \\ \cline{3-5} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{array}\right].$$

Thus $U = \begin{bmatrix} 1 & 2 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 & -1 \end{bmatrix}$.

Basis for $\operatorname{ran} A$**.** A basis for $\operatorname{ran} A$ can be obtained by taking as basis vectors those columns of A which correspond to steps in T. In our case, these are the first and third columns, and so the set

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 4 \\ 3 \\ 2 \\ 1 \end{bmatrix} \right\}$$

constitutes a basis for $\operatorname{ran} A$.

Basis for $\operatorname{ran} A^\top$**.** A basis for $\operatorname{ran} A^\top$ can be obtained by taking the columns of $U^\top$, namely it is the set

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ -1 \end{bmatrix} \right\}.$$

Basis for $\ker A$**.** A basis for $\ker A$ can be determined as follows.

First note that $k = n - r = 5 - 2 = 3$. Set $x_\nu = e_1$, that is, $x_2 = 1$, $x_4 = x_5 = 0$, and calculate x_β, that is, x_1 and x_3, by using $x_\beta = -U_\nu x_\nu$, or equivalently $Ux = 0$. This yields the vector $z_1^\top = \begin{bmatrix} -2 & 1 & 0 & 0 & 0 \end{bmatrix}^\top$.

Then set $x_\nu = e_2$, that is, $x_4 = 1$, $x_2 = x_5 = 0$, and calculate x_1 and x_3, by using $Ux = 0$. Thus we obtain the vector $z_2^\top = \begin{bmatrix} -1 & 0 & -1 & 1 & 0 \end{bmatrix}^\top$.

Finally set $x_\nu = e_3$, that is, $x_5 = 1$, $x_2 = x_4 = 0$, and calculate x_1 and x_3, by using $Ux = 0$. Hence $z_3^\top = \begin{bmatrix} -1 & 0 & 1 & 0 & 1 \end{bmatrix}^\top$.

Thus a basis for $\ker A$ is given by

$$\left\{ \begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

Basis for $\ker A^\top$**.** For determining a basis for $\ker A^\top$ we should first use the Gauss-Jordan method on the matrix

$$A^\top = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 2 & 4 & 6 & 8 \\ 4 & 3 & 2 & 1 \\ 5 & 5 & 5 & 5 \\ -3 & -1 & 1 & 3 \end{bmatrix}$$

in order to bring it to a staircase form using elementary row transformations.

We do the following:

(1) add -2 times the first row to the second row,

(2) add -4 times the first row to the third row,

(3) add -5 times the first row to the fourth row, and

(4) add 3 times the first row to the fifth row.

We then obtain

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 0 & 0 & 0 \\ 0 & -5 & -10 & -15 \\ 0 & -5 & -10 & -15 \\ 0 & 5 & 10 & 15 \end{bmatrix}.$$

We then interchange the second and the third row, giving us

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & -5 & -10 & -15 \\ 0 & 0 & 0 & 0 \\ 0 & -5 & -10 & -15 \\ 0 & 5 & 10 & 15 \end{bmatrix}.$$

Multiplying the second row by $-\frac{1}{5}$, we obtain

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 1 & 2 & 3 \\ 0 & 0 & 0 & 0 \\ 0 & -5 & -10 & -15 \\ 0 & 5 & 10 & 15 \end{bmatrix}.$$

Now we do the following:

(1) add -2 times the second row to the first row,

(2) add 5 times the second row to the fourth row, and

(3) add -5 times the second row to the fifth row.

We then obtain a matrix which has a staircase form, with two steps, namely the 1's in the first and the second columns:

$$\widetilde{T} = \left[\begin{array}{cccc} 1 & 0 & -1 & -2 \\ \cline{1-1} 0 & \multicolumn{1}{|c}{1} & 2 & 3 \\ \cline{2-4} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right] = \left[\begin{array}{c} \widetilde{U} \\ 0_{3\times 4} \end{array}\right],$$

where $\widetilde{U} = \begin{bmatrix} 1 & 0 & -1 & -2 \\ 0 & 1 & 2 & 3 \end{bmatrix}$.

A basis for $\ker A^\top$ can be determined as follows. First note that now $\widetilde{k} = \widetilde{n} - r = 4 - 2 = 2$.

Set $x_\nu = e_1$, that is, $x_3 = 1$, $x_4 = 0$, and calculate x_β, that is, x_1 and x_2, by using $x_\beta = -\widetilde{U}_\nu x_\nu$, or equivalently $\widetilde{U}x = 0$. This yields the vector $\widetilde{z}_1^\top = \begin{bmatrix} 1 & -2 & 1 & 0 \end{bmatrix}^\top$.

Finally set $x_\nu = e_2$, that is, $x_4 = 1$, $x_3 = 0$, and calculate x_1 and x_2, by using $Ux = 0$. This yields the vector $\widetilde{z}_2^\top = \begin{bmatrix} 2 & -3 & 0 & 1 \end{bmatrix}^\top$.

Thus a basis for $\ker A^\top$ is given by

$$\left\{ \begin{bmatrix} -1 \\ 2 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ -3 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

We also remark that a new basis for $\operatorname{ran} A^\top$ can be obtained by taking as basis vectors those columns of $A^\top$ which correspond to steps in $\widetilde{T}$. In our case, these are the first and second columns, and so the set

$$\left\{ \begin{bmatrix} 1 \\ 2 \\ 4 \\ 5 \\ -3 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \\ 3 \\ 5 \\ -1 \end{bmatrix} \right\}$$

constitutes a basis for $\operatorname{ran} A^\top$. A new basis for $\operatorname{ran}(A^\top)^\top = \operatorname{ran} A$ can be obtained by taking the columns of $\widetilde{U}^\top$, namely it is the set

$$\left\{ \begin{bmatrix} 1 \\ 0 \\ -1 \\ -2 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 2 \\ 3 \end{bmatrix} \right\}. \qquad \Diamond$$

Exercise 25.9. Find bases for the four fundamental subspaces of the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 2 & 4 & 3 \end{bmatrix}.$$

Chapter 26

Positive definite and semidefinite matrices

Within nonlinear optimization, and particularly in quadratic optimization, it is important that one can determine, in an efficient manner, whether or not a given matrix is positive definite. In this chapter, we will deal with this question and some related things. We begin with some facts about special types of matrices.

26.1. Diagonal and triangular matrices

(1) A matrix is referred to as a *square matrix* if its number of rows is the same as its number of columns.

(2) A square matrix H is called *symmetric* if $H^\top = H$, that is, its elements satisfy $h_{ij} = h_{ji}$ for all i's and j's.

(3) A square matrix D is called *diagonal* if all its entries which aren't on the diagonal are all zeros, that is, $d_{ij} = 0$ if $i \neq j$. The diagonal elements of a diagonal matrix D are often denoted by d_i (instead of the more precise notation d_{ii}). A diagonal matrix D is always symmetric, and is non-singular (or invertible) iff $d_i \neq 0$ for all i's.

(4) A square matrix L is called *lower triangular* if the elements of the matrix satisfy $l_{ij} = 0$ whenever $i < j$. This means that the elements above the diagonal entries are all zeros. A lower triangular matrix L is non-singular iff all the diagonal entries l_{ii}'s are nonzero.

(5) A square matrix U is called *upper triangular* if the elements of the matrix satisfy $u_{ij} = 0$ whenever $i > j$. This means that the elements below the diagonal entries are all zeros. An upper triangular matrix U is non-singular iff all the diagonal entries u_{ii}'s are nonzero.

Thus a 4 × 4 diagonal matrix D has the following form:

$$D = \begin{bmatrix} d_1 & 0 & 0 & 0 \\ 0 & d_2 & 0 & 0 \\ 0 & 0 & d_3 & 0 \\ 0 & 0 & 0 & d_4 \end{bmatrix}$$

while a 4 × 4 lower triangular matrix L, and a 4 × 4 upper triangular matrix U, have the following form:

$$L = \begin{bmatrix} l_{11} & 0 & 0 & 0 \\ l_{21} & l_{22} & 0 & 0 \\ l_{31} & l_{32} & l_{33} & 0 \\ l_{41} & l_{42} & l_{43} & l_{44} \end{bmatrix}, \quad U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & u_{14} \\ 0 & u_{22} & u_{23} & u_{24} \\ 0 & 0 & u_{33} & u_{34} \\ 0 & 0 & 0 & u_{44} \end{bmatrix}.$$

An important property of diagonal and triangular matrices is that it is easy to solve systems of equations involving these. For example, if one wants to solve the system $Lx = b$, where L is a 4×4 lower triangular matrix as above, then one obtains x as follows:

$$\begin{aligned} x_1 &= \frac{b_1}{l_{11}}, \\ x_2 &= \frac{b_2 - l_{21}x_1}{l_{22}}, \\ x_3 &= \frac{b_3 - l_{31}x_1 - l_{32}x_2}{l_{33}}, \\ x_4 &= \frac{b_4 - l_{41}x_1 - l_{42}x_2 - l_{43}x_3}{l_{44}}. \end{aligned}$$

Similarly, if we want to solve $Ux = b$, where U is a 4×4 upper triangular matrix as above then one obtains x as follows:

$$\begin{aligned} x_1 &= \frac{b_4}{u_{44}}, \\ x_2 &= \frac{b_3 - u_{34}x_4}{u_{33}}, \\ x_3 &= \frac{b_2 - u_{24}x_4 - u_{23}x_3}{u_{22}}, \\ x_4 &= \frac{b_1 - u_{14}x_4 - u_{13}x_3 - u_{12}x_2}{u_{11}}. \end{aligned}$$

The easiest case is the equation system $Dx = b$, where D is a 4×4 diagonal matrix. Then the solution x is given simply by

$$x_1 = \frac{b_1}{d_1}, \quad x_2 = \frac{b_2}{d_2}, \quad x_3 = \frac{b_3}{d_3}, \quad x_4 = \frac{b_4}{d_4}.$$

26.2. Positive definite and semidefinite matrices

Let $H \in \mathbb{R}^{n \times n}$ be a given symmetric matrix and let $x \in \mathbb{R}^n$. Then

$$x^\top H x = \sum_{i=1}^{n} \sum_{j=1}^{n} h_{ij} x_i x_j. \tag{26.1}$$

If $x_1, \dots, x_n$ are thought of as variables, and the matrix entries h_{ij} are constants, then (26.1) is called a *quadratic form* in $x \in \mathbb{R}^n$.

(1) A symmetric matrix $H \in \mathbb{R}^{n \times n}$ is called *positive semidefinite* if $x^\top H x \geq 0$ for all $x \in \mathbb{R}^n$.

(2) A symmetric matrix $H \in \mathbb{R}^{n \times n}$ is called *positive definite* if $x^\top H x > 0$ for all nonzero $x \in \mathbb{R}^n$.

Clearly every positive definite matrix is positive semidefinite.

Example 26.1. Let $H = \begin{bmatrix} 1 & 2 \\ 2 & 5 \end{bmatrix}$.

Then $x^\top H x = x_1^2 + 2x_1x_2 + 2x_2x_1 + 5x_2^2 = (x_1 + 2x_2)^2 + x_2^2 \geq 0$, with equality iff $x_1 + 2x_2 = 0$ and $x_2 = 0$, that is, iff $x_1 = x_2 = 0$. So the matrix is positive definite. $\Diamond$

Example 26.2. Let $H = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$.

Then $x^\top H x = x_1^2 + 2x_1x_2 + 2x_2x_1 + 4x_2^2 = (x_1 + 2x_2)^2 \geq 0$. So the matrix is positive semidefinite. However, the matrix is *not* positive definite, since with $x_1 = -2$ and $x_2 = 1$, we have that $x^\top H x = 0$, but $x \neq 0$. $\Diamond$

Example 26.3. Let $H = \begin{bmatrix} 1 & 2 \\ 2 & 3 \end{bmatrix}$.

Then $x^\top Hx = x_1^2 + 2x_1x_2 + 2x_2x_1 + 3x_2^2 = (x_1 + 2x_2)^2 - x_2^2$. Since with $x_1 = -2$ and $x_2 = 1$, we have that $x^\top Hx = -1 < 0$, it follows that H is not positive semidefinite (even though all entries of H are positive). ♢

Exercise 26.4. Show that $H = \begin{bmatrix} 1 & -1 \\ -1 & 4 \end{bmatrix}$ is positive definite.

26.3. Properties of positive definite matrices

Property 26.5. *If H is symmetric and positive definite, then H is invertible.*

Proof. If H is not invertible, then there exists a nonzero vector x such that $Hx = 0$. Thus $x^\top Hx = x^\top 0 = 0$, contradicting the positive definiteness of H. □

Property 26.6. *If H is symmetric and positive definite, then every diagonal entry $h_{ii} > 0$.*

Proof. We have $e_i^\top He_i = h_{ii}$. □

Property 26.7. *A diagonal matrix D is positive definite iff all the d_i's are positive.*

Proof. This follows from the fact that $x^\top Dx = \sum_{i=1}^{n} d_i x_i^2$. □

Property 26.8. *If $H \in \mathbb{R}^{n\times n}$ is symmetric and positive definite, and if $B \in \mathbb{R}^{n\times k}$ has linearly independent columns, then the matrix $G := B^\top HB \in \mathbb{R}^{k\times k}$ is symmetric and positive definite as well.*

Proof. Since $G^\top = (B^\top HB)^\top = B^\top H^\top B = B^\top HB = G$, we see that G is symmetric. Also, $x^\top Gx = x^\top B^\top HBx = (Bx)^\top H(Bx) \geq 0$ for all $x \in \mathbb{R}^k$, with equality only iff $Bx = 0$, which in turn is equivalent to $x = 0$, since B has linearly independent columns. □

Property 26.9. *Let $H \in \mathbb{R}^{n\times n}$ be the symmetric block matrix*

$$H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix},$$

where $H_1 \in \mathbb{R}^{n_1\times n_1}$, $H_2 \in \mathbb{R}^{n_2\times n_2}$ are symmetric, and $n_1 + n_2 = n$. Then H is positive definite iff H_1 and H_2 are positive definite.

Proof. If: Let $x \in \mathbb{R}^n$ be nonzero. If we partition x into x_1 and x_2, where x_1 is made up of the first n_1 components and x_2 the last n_2 components of x, then either $x_1 \neq 0$ or $x_2 \neq 0$. Then $x_1^\top H_1x_1$ and $x_2^\top H_2x_2$ are both nonnegative, and at least one of them is nonzero. Hence $x^\top Hx = x_1^\top H_1x_1 + x_2^\top H_2x_2 > 0$.

Only if: We have $x^\top Hx = x_1^\top H_1x_1 + x_2^\top H_2x_2$. Suppose that H is positive definite. By taking $x_1 \neq 0$ and $x_2 = 0$ here, we see that H_1 must be positive definite. Next taking x_2 nonzero and $x_1 = 0$, we see that H_2 must also be positive definite. □

Property 26.10. *A symmetric matrix H is positive definite iff all its eigenvalues are positive.*

Proof. If: By the spectral theorem[1], $H = P^\top \Lambda P$, where P is invertible, and Λ is a diagonal matrix with the eigenvalues of H as its diagonal entries. Since all of these eigenvalues are positive, it follows that Λ is positive definite. Moreover, P has linearly independent column vectors, and so we conclude that $P^\top \Lambda P = H$ is positive definite as well.

[1] see for example [**T**]

Only if: Suppose that H is positive definite. By the spectral theorem we know that there is a basis of eigenvectors of H for $\mathbb{R}^n$. Let λ be an eigenvalue of H with an eigenvector x. Then $Hx = \lambda x$ and $x \neq 0$. But then $0 < x^\top Hx = x^\top(\lambda x) = \lambda(x^\top x)$ and since $x^\top x > 0$, it follows that $\lambda > 0$ as well. $\square$

Exercise 26.11 (Sylvester's criterion for positivity)**.** Let $H \in \mathbb{R}^{n\times n}$. For $1 \le k \le n$, the *kth principal submatrix of H* is the $k \times k$ submatrix of H formed by taking just the first k rows and first k columns of H. Its determinant is called the *kth principal minor.* In this exercise we want to prove Sylvester's criterion for positivity, namely that a symmetric matrix $H \in \mathbb{R}^{n\times n}$ is positive definite iff all its principal minors are positive.

(1) Show the 'only if' part by showing that each kth principal submatrix is positive definite, and hence it has a positive determinant.

(2) Let $v_1, \dots, v_n$ be a basis of a vector space V. Suppose that W is a k-dimensional subspace of V. If $m < k$, then show that there exists a nonzero vector in W which is a linear combination of $v_{m+1}, \dots, v_n$. *Hint:* Use Exercise 23.9 with $S_1 := W$ and $S_2 := \operatorname{span}\{v_{m+1}, \dots, v_n\}$.

(3) Let $H \in \mathbb{R}^{n\times n}$ be symmetric. If $w^\top Hw > 0$ for all nonzero vectors w in a k-dimensional subspace W of $\mathbb{R}^n$, then H has at least k positive eigenvalues (counting multiplicity). *Hint:* By the spectral theorem, we know that H has an orthonormal basis of eigenvectors $v_1, \dots, v_n$. Suppose that the first m of these eigenvectors are the ones corresponding to positive eigenvalues, while the others correspond to nonpositive eigenvalues.

(4) Prove Sylvester's criterion for positivity using induction on n. *Hint:* To complete the induction step, note that the nth principal submatrix of $H \in \mathbb{R}^{(n+1)\times(n+1)}$ is positive definite by the induction hypothesis. Thus $w^\top Hw > 0$ for all nonzero vectors w in the n-dimensional subspace $W = \operatorname{span}\{e_1, \dots, e_n\}$ ($\subset \mathbb{R}^{n+1}$). Conclude that at least n eigenvalues of H must be positive. Since $\det H$ is positive as well, argue that all the eigenvalues of H must be positive.

Exercise 26.12. Using Sylvester's criterion of positivity check if the matrices

$$A = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 3 & -1 \\ 1 & -1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 3 & -1 & 2 \\ -1 & 4 & -2 \\ 2 & -2 & 1 \end{bmatrix}$$

are positive definite or not. Are the matrices $-A$, A^3, A^{-1} also positive definite?

Exercise 26.13. True or false:

(1) If A is positive definite, then A^5 is positive definite.

(2) If A is negative definite (that is $-A$ is positive definite), then A^8 is negative definite.

(3) If A is negative definite, then A^{12} is positive definite.

(4) If A is positive definite and B is negative semidefinite, then $A - B$ is positive definite.

26.4. Properties of positive semidefinite matrices

Property 26.14. *If H is symmetric and positive semidefinite, then every diagonal entry $h_{ii} \ge 0$.*

Proof. We have $e_i^\top He_i = h_{ii}$. $\square$

Property 26.15. *If H is symmetric and positive semidefinite and a diagonal entry $h_{ii} = 0$, then $h_{ij} = h_{ji} = 0$ for all j.*

Proof. For $r \in \mathbb{R}$, let $x := re_i + e_j$. Then $x^\top Hx = 2rh_{ij} + h_{jj} \ge 0$. But the choice of r was arbitrary. Hence $h_{ij} = 0$. $\square$

Property 26.16. *A diagonal matrix D is positive semidefinite iff all the d_i's are nonnegative.*

Proof. This follows from the fact that $x^\top Dx = \sum_{i=1}^{n} d_i x_i^2$. $\square$

Property 26.17. *If $H \in \mathbb{R}^{n\times n}$ is symmetric and positive semidefinite and $B \in \mathbb{R}^{n\times k}$, then $G := B^\top HB \in \mathbb{R}^{k\times k}$ is symmetric and positive semidefinite as well.*

Proof. Since $G^\top = (B^\top HB)^\top = B^\top H^\top B = B^\top HB = G$, it follows that G is symmetric. Also, $x^\top Gx = x^\top B^\top HBx = (Bx)^\top H(Bx) \geq 0$ for all $x \in \mathbb{R}^k$. □

Property 26.18. *Let $H \in \mathbb{R}^{n\times n}$ be a symmetric block matrix*

$$H = \begin{bmatrix} H_1 & 0 \\ 0 & H_2 \end{bmatrix},$$

where $H_1 \in \mathbb{R}^{n_1\times n_1}$, $H_2 \in \mathbb{R}^{n_2\times n_2}$ are symmetric, and $n_1+n_2 = n$. Then H is positive semidefinite iff H_1 and H_2 are positive semidefinite.

Proof. If: Let $x \in \mathbb{R}^n$. Partition x into x_1 and x_2, where x_1 is made up of the first n_1 components and x_2 the last n_2 components of x. Then $x_1^\top H_1x_1$ and $x_2^\top H_2x_2$ are both nonnegative. Consequently, $x^\top Hx = x_1^\top H_1x_1 + x_2^\top H_2x_2 \geq 0$.

Only if: We have $x^\top Hx = x_1^\top H_1x_1 + x_2^\top H_2x_2$. Let H be positive semidefinite. By taking $x_1 \in \mathbb{R}^{n_1}$ arbitrary and $x_2 = 0$ here, we see that H_1 must be positive semidefinite. Next taking $x_2 \in \mathbb{R}^{n_2}$ arbitrary and $x_1 = 0$, we see that H_2 must also be positive semidefinite. □

Property 26.19. *A symmetric matrix H is positive semidefinite iff all its eigenvalues are nonnegative.*

Proof. If: By the spectral theorem, $H = P^\top \Lambda P$, where P is invertible, and Λ is a diagonal matrix with the eigenvalues of H as its diagonal entries. Since all of these eigenvalues are nonnegative, it follows that Λ is positive semidefinite. So we conclude that $P^\top \Lambda P = H$ is positive semidefinite as well.

Only if: Suppose that H is positive semidefinite. By the spectral theorem we know that there is a basis of eigenvectors of H for $\mathbb{R}^n$. Let λ be an eigenvalue of H with an eigenvector x. Then $Hx = \lambda x$ and $x \neq 0$. But then $0 \leq x^\top Hx = x^\top(\lambda x) = \lambda(x^\top x)$ and since $x^\top x > 0$, it follows that $\lambda \geq 0$. □

26.5. The matrices $A^\top A$ and $AA^\top$

Let $A \in \mathbb{R}^{m\times n}$ be a given matrix. Set $H = A^\top A \in \mathbb{R}^{n\times n}$ and $G = AA^\top \in \mathbb{R}^{m\times m}$. Then H and G are both symmetric, since

$$\begin{aligned} H^\top &= (A^\top A)^\top = A^\top (A^\top)^\top = A^\top A = H, \\ G^\top &= (AA^\top)^\top = (A^\top)^\top A^\top = AA^\top = G. \end{aligned}$$

Furthermore, H and G are both positive semidefinite, since for every $x \in \mathbb{R}^n$ and every $y \in \mathbb{R}^m$ we have

$$\begin{aligned} x^\top Hx &= x^\top A^\top Ax = (Ax)^\top (Ax) = \|Ax\|^2 \geq 0, \\ y^\top Gy &= y^\top AA^\top y = (A^\top y)^\top (A^\top y) = \|A^\top y\|^2 \geq 0. \end{aligned}$$

Now suppose that A has linearly independent columns. Since I is positive definite, it then follows that $H = A^\top A = A^\top IA$ is positive definite as well.

Similarly, if A has linearly independent rows, then $A^\top$ has linearly independent columns, and applying what we have just proved to $A^\top$, it follows that $G = (A^\top)^\top A^\top = AA^\top$ is positive definite.

The following relations are sometimes useful:

$$\begin{aligned}\ker(A^\top A) &= \ker A,\\ \ker(AA^\top &= \ker A^\top,\\ \operatorname{ran}(A^\top A) &= \operatorname{ran} A^\top,\\ \operatorname{ran}(AA^\top) &= \operatorname{ran} A.\end{aligned}$$

If $A^\top Ax = 0$, then $x^\top A^\top Ax = x^\top 0 = 0$. So $\|Ax\|^2 = x^\top A^\top Ax = 0$, that is, $Ax = 0$. Hence we conclude that $\ker(A^\top A) \subset \ker A$. On the other hand, if $A\xi = 0$, then $A^\top A\xi = A^\top 0 = 0$. Thus also $\ker A \subset \ker(A^\top A)$.

That $\ker(AA^\top) = \ker A^\top$ follows from the previous case by replacing A by $A^\top$. Also,

$$\operatorname{ran}(A^\top A) = (\ker(A^\top A)^\top)^\perp = (\ker(A^\top A))^\perp = (\ker A)^\perp = \operatorname{ran} A^\top.$$

Finally, $\operatorname{ran}(AA^\top) = \operatorname{ran} A$ follows from the previous case by replacing A by $A^\top$.

26.6. $LDL^\top$-factorization of positive definite matrices

In order to determine whether or not a given symmetric matrix is positive definite, one can use the so-called $LDL^\top$*-factorization*, which is based on the following result.

Theorem 26.20. *A symmetric $H \in \mathbb{R}^{n\times n}$ is positive definite iff*

(1) *there exists a lower triangular matrix $L \in \mathbb{R}^{n\times n}$ with 1's on the diagonal (all $l_{ii} = 1$), and*

(2) *there exists a diagonal matrix $D \in \mathbb{R}^{n\times n}$ with all diagonal entries positive (all $d_i > 0$),*

(3) *such that $H = LDL^\top$.*

In this section we will give a constructive proof of this result by describing an algorithm for the $LDL^\top$-factorization of positive definite matrices.

First assume that L and D are as above. Then D is positive definite. Also $L^\top$ has linearly independent columns. Thus it follows that $H = LDL^\top = (L^\top)^\top DL^\top$ is also positive definite, by the properties of positive definite matrices we had studied earlier.

In the remainder of this chapter, we will assume that we have been given a H which is positive definite, and we will show how one can determine L and D with the properties above, such that H has the factorization $H = LDL^\top$.

For making the exposition less technical (and hopefully easier to follow), we limit the description of our method in the special case that H is a 4×4 matrix. The generalization to an $n \times n$ matrix is obvious.

Our goal is to carry out the factorization $H = LDL^\top$, where

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} & h_{14}\\ h_{21} & h_{22} & h_{23} & h_{24}\\ h_{31} & h_{32} & h_{33} & h_{34}\\ h_{41} & h_{42} & h_{43} & h_{44}\end{bmatrix}$$

is given, and we want to find L and D having the form:

$$D = \begin{bmatrix} d_1 & 0 & 0 & 0\\ 0 & d_2 & 0 & 0\\ 0 & 0 & d_3 & 0\\ 0 & 0 & 0 & d_4\end{bmatrix} \quad\text{and}\quad L = \begin{bmatrix} 1 & 0 & 0 & 0\\ l_{21} & 1 & 0 & 0\\ l_{31} & l_{32} & 1 & 0\\ l_{41} & l_{42} & l_{43} & 1\end{bmatrix}.$$

For reasons that will soon become evident, we denote the elements of H by $h_{ij}^{(1)}$. Thus we have

$$H = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} & h_{13}^{(1)} & h_{14}^{(1)} \\ h_{21}^{(1)} & h_{22}^{(1)} & h_{23}^{(1)} & h_{24}^{(1)} \\ h_{31}^{(1)} & h_{32}^{(1)} & h_{33}^{(1)} & h_{34}^{(1)} \\ h_{41}^{(1)} & h_{42}^{(1)} & h_{43}^{(1)} & h_{44}^{(1)} \end{bmatrix},$$

where $h_{ij}^{(1)} = h_{ji}^{(1)}$ for all i and j. Since H is positive definite, it follows in particular that $h_{11}^{(1)} > 0$. Thus we can subtract multiples of the first row from the remaining rows, so that all the elements below the diagonal element $h_{11}^{(1)}$ in the first column become zeros.

This corresponds to premultiplying the matrix H with the matrix E_1 given below. The first row of H is unaffected by these row operations, and so the first row of E_1H is the same as the first row of H. The other rows have been possibly changed, and so these elements will be denoted now by $h_{ij}^{(2)}$. Hence we have that

$$E_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -l_{21} & 1 & 0 & 0 \\ -l_{31} & 0 & 1 & 0 \\ -l_{41} & 0 & 0 & 1 \end{bmatrix} \text{ and } E_1H = \begin{bmatrix} h_{11}^{(1)} & h_{12}^{(1)} & h_{13}^{(1)} & h_{14}^{(1)} \\ 0 & h_{22}^{(2)} & h_{23}^{(2)} & h_{24}^{(2)} \\ 0 & h_{32}^{(2)} & h_{33}^{(2)} & h_{34}^{(2)} \\ 0 & h_{42}^{(2)} & h_{43}^{(2)} & h_{44}^{(2)} \end{bmatrix},$$

where $l_{i1} = \dfrac{h_{i1}^{(1)}}{h_{11}^{(1)}}$ and

$$h_{ij}^{(2)} = h_{ij}^{(1)} - l_{i1}h_{1j}^{(1)} = h_{ij}^{(1)} - \frac{h_{i1}^{(1)}h_{1j}^{(1)}}{h_{11}^{(1)}} \text{ for } i = 2,3,4 \text{ and } j = 2,3,4.$$

Now we can subtract multiples of the first column of E_1H from the other columns, so that all the elements to the right of the diagonal element $h_{11}^{(1)}$ in the first row become zeros. This correspond to postmultiplying the matrix E_1H by the matrix $E_1^\top$. (That the same matrix E_1 appears again, except now transposed, is because of the fact that H is symmetric, so that $h_{1j}^{(1)} = h_{j1}^{(1)}$ for $j = 2,3,4$. Hence the same multipliers which were used earlier in the row operations are now used in the column operations.) The elements $h_{ij}^{(2)}$ are not effected by these column operations since the first column of E_1H has zeros below $h_{11}^{(1)}$. Thus

$$E_1HE_1^\top = \begin{bmatrix} h_{11}^{(1)} & 0 & 0 & 0 \\ 0 & h_{22}^{(2)} & h_{23}^{(2)} & h_{24}^{(2)} \\ 0 & h_{32}^{(2)} & h_{33}^{(2)} & h_{34}^{(2)} \\ 0 & h_{42}^{(2)} & h_{43}^{(2)} & h_{44}^{(2)} \end{bmatrix}.$$

Since H is positive definite, so is $E_1HE_1^\top$, which in turn implies that the matrix

$$\begin{bmatrix} h_{22}^{(2)} & h_{23}^{(2)} & h_{24}^{(2)} \\ h_{32}^{(2)} & h_{33}^{(2)} & h_{34}^{(2)} \\ h_{42}^{(2)} & h_{43}^{(2)} & h_{44}^{(2)} \end{bmatrix}$$

is positive definite as well. In particular, $h_{22}^{(2)} > 0$. Now we are going to repeat the above steps for this smaller positive definite matrix.

One can subtract multiples of the second row of $E_1HE_1^\top$ from the last two rows, so that all the elements below $h_{22}^{(2)}$ in the second column become zeros. This corresponds to premultiplying the matrix $E_1HE_1^\top$ by E_2 given below. The first two rows of $E_1HE_1^\top$ are not affected by these row operations. The last two rows are affected, and we denote their new entries by $h_{ij}^{(3)}$. Thus we have that

$$E_2=\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -l_{32} & 1 & 0 \\ 0 & -l_{42} & 0 & 1 \end{bmatrix} \text{ and } E_2E_1HE_1^\top=\begin{bmatrix} h_{11}^{(1)} & 0 & 0 & 0 \\ 0 & h_{22}^{(2)} & h_{23}^{(2)} & h_{24}^{(2)} \\ 0 & 0 & h_{33}^{(3)} & h_{34}^{(3)} \\ 0 & 0 & h_{43}^{(3)} & h_{44}^{(3)} \end{bmatrix},$$

where $l_{i2} = \dfrac{h_{i2}^{(1)}}{h_{22}^{(2)}}$ and

$$h_{ij}^{(3)} = h_{ij}^{(2)} - l_{i2}h_{2j}^{(2)} = h_{ij}^{(2)} - \frac{h_{i2}^{(2)}h_{2j}^{(2)}}{h_{22}^{(2)}} \text{ for } i = 3,4 \text{ and } j = 3,4.$$

Now we can subtract multiples of the second column of $E_2E_1HE_1^\top$ from the last two columns, so that all the elements to the right of the diagonal element $h_{22}^{(2)}$ in the second row become zeros. This correspond to postmultiplying the matrix $E_2E_1HE_1^\top$ by the matrix $E_2^\top$. The elements $h_{ij}^{(3)}$ are not effected by these column operations. Thus

$$E_2E_1HE_1^\top E_2^\top = \begin{bmatrix} h_{11}^{(1)} & 0 & 0 & 0 \\ 0 & h_{22}^{(2)} & 0 & 0 \\ 0 & 0 & h_{33}^{(3)} & h_{34}^{(3)} \\ 0 & 0 & h_{43}^{(3)} & h_{44}^{(3)} \end{bmatrix}.$$

Since H is positive definite, so is $E_2E_1HE_1^\top E_2^\top$, which in turn implies that the matrix

$$\begin{bmatrix} h_{33}^{(3)} & h_{34}^{(3)} \\ h_{43}^{(3)} & h_{44}^{(3)} \end{bmatrix}$$

is positive definite as well. In particular, $h_{33}^{(3)} > 0$. Now we are going to repeat the above steps for this yet smaller positive definite matrix.

One can subtract a multiple of the third row of $E_2E_1HE_1^\top E_2^\top$ from the last row, so that the elements below $h_{33}^{(3)}$ in the third column becomes zero. This corresponds to premultiplying the matrix $E_2E_1HE_1^\top E_2^\top$ by E_3 given below. The first three rows of $E_2E_1HE_1^\top E_2^\top$ are not affected by these row operations. The last row is affected, and we denote its new entries by $h_{4j}^{(4)}$. Thus we have that

$$\begin{aligned} E_3 &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -l_{43} & 1 \end{bmatrix} \text{ and} \\ E_3E_2E_1HE_1^\top E_2^\top &= \begin{bmatrix} h_{11}^{(1)} & 0 & 0 & 0 \\ 0 & h_{22}^{(2)} & 0 & 0 \\ 0 & 0 & h_{33}^{(3)} & h_{34}^{(3)} \\ 0 & 0 & 0 & h_{44}^{(4)} \end{bmatrix}, \end{aligned}$$

where $l_{43} = \dfrac{h_{43}^{(1)}}{h_{33}^{(3)}}$ and $h_{44}^{(4)} = h_{44}^{(3)} - l_{43}h_{34}^{(3)} = h_{44}^{(3)} - \dfrac{h_{43}^{(3)}h_{34}^{(3)}}{h_{33}^{(3)}}$.

Now we subtract a multiple of the third column of $E_3E_2E_1HE_1^\top E_2^\top$ from the last column, so that the element to the right of the diagonal element $h_{33}^{(3)}$ in the third row becomes a zero. This correspond to postmultiplying the matrix $E_3E_2E_1HE_1^\top E_2^\top$ by the matrix $E_3^\top$. The element $h_{44}^{(4)}$

is not effected by this column operation. Thus

$$E_3E_2E_1HE_1^\top E_2^\top E_3^\top = \begin{bmatrix} h_{11}^{(1)} & 0 & 0 & 0 \\ 0 & h_{22}^{(2)} & 0 & 0 \\ 0 & 0 & h_{33}^{(3)} & 0 \\ 0 & 0 & 0 & h_{44}^{(4)} \end{bmatrix}.$$

Since H is positive definite, so is $E_3E_2E_1HE_1^\top E_2^\top E_3^\top$, which in turn implies that $h_{44}^{(4)} > 0$.

Let D be the diagonal matrix $E_3E_2E_1HE_1^\top E_2^\top E_3^\top$ given above. Since each row operation matrix E_k is invertible, we have that

$$H = E_1^{-1}E_2^{-1}E_3^{-1}D(E_3^\top)^{-1}(E_2^\top)^{-1}(E_1^\top)^{-1}. \tag{26.2}$$

It turns out that it is very easy to calculate the inverses E_k^{-1} and the product $E_1^{-1}E_2^{-1}E_3^{-1}$. Indeed, one has that

$$\begin{aligned} E_1^{-1} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & 0 & 1 & 0 \\ l_{41} & 0 & 0 & 1 \end{bmatrix}, \\ E_2^{-1} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & l_{32} & 1 & 0 \\ 0 & l_{42} & 0 & 1 \end{bmatrix}, \\ E_3^{-1} &= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & l_{43} & 1 \end{bmatrix}, \end{aligned}$$

and

$$E_1^{-1}E_2^{-1}E_3^{-1} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix}.$$

Thus the matrix $L := E_1^{-1}E_2^{-1}E_3^{-1}$ is lower triangular with 1's on the diagonal. Also,

$$\begin{aligned} L^\top &= (E_1^{-1}E_2^{-1}E_3^{-1})^\top = (E_3^{-1})^\top(E_2^{-1})^\top(E_1^{-1})^\top \\ &= (E_3^\top)^{-1}(E_2^\top)^{-1}(E_1^\top)^{-1}. \end{aligned}$$

So (26.2) becomes $H = LDL^\top$, and we have obtained the desired factorization.

26.7. An example of $LDL^\top$-factorization

We will find out if the following matrix H is positive definite or not by carrying out the $LDL^\top$-factorization:

$$H = \begin{bmatrix} 2 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}$$

First add $\frac{1}{2}$ times the first row to the second row and then add $\frac{1}{2}$ times the first column to the second column. Thus

$$E_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad E_1 H E_1^\top = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}.$$

Now add $\frac{2}{3}$ times the second row to the third row and then add $\frac{2}{3}$ times the second column to the third column. Thus

$$E_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & \frac{2}{3} & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } E_2 E_1 H E_1^\top E_2^\top = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & 0 & 0 \\ 0 & 0 & \frac{4}{3} & -1 \\ 0 & 0 & -1 & 2 \end{bmatrix}.$$

Now add $\frac{3}{4}$ times the third row to the fourth row and then add $\frac{3}{4}$ times the third column to the fourth column. Thus

$$E_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & \frac{3}{4} & 1 \end{bmatrix} \text{ and } E_3 E_2 E_1 H E_1^\top E_2^\top E_3^\top = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & 0 & 0 \\ 0 & 0 & \frac{4}{3} & 0 \\ 0 & 0 & 0 & \frac{5}{4} \end{bmatrix}.$$

Thus the $LDL^\top$-factorization is completed, and we have $H = LDL^\top$, where

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 & 0 \\ 0 & -\frac{2}{3} & 1 & 0 \\ 0 & 0 & -\frac{3}{4} & 1 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & 0 & 0 \\ 0 & 0 & \frac{4}{3} & 0 \\ 0 & 0 & 0 & \frac{5}{4} \end{bmatrix}.$$

Since all the diagonal elements of D are > 0, we conclude that H is positive definite.

26.8. Completing squares and $LDL^\top$-factorization

There is a close connection between $LDL^\top$-factorization and good old completion of squares. We will illustrate this connection with the help of the example from the previous section.

Let $x \in \mathbb{R}^4$ and consider the quadratic form

$$x^\top H x = 2x_1^2 - 2x_1x_2 + 2x_2^2 - 2x_2x_3 + 2x_3^2 - 2x_3x_4 + 2x_4^2. \tag{26.3}$$

Using the $LDL^\top$-factorization, we have

$$x^\top H x = x^\top LDL^\top x = (L^\top x)^\top D (L^\top x),$$

where

$$L^\top x = \begin{bmatrix} 1 & -\frac{1}{2} & 0 & 0 \\ 0 & 1 & -\frac{2}{3} & 0 \\ 0 & 0 & 1 & -\frac{3}{4} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} x_1 - \frac{1}{2}x_2 \\ x_2 - \frac{2}{3}x_3 \\ x_3 - \frac{3}{4}x_4 \\ x_4 \end{bmatrix}.$$

Consequently

$$x^\top H x = \begin{bmatrix} x_1 - \frac{1}{2}x_2 \\ x_2 - \frac{2}{3}x_3 \\ x_3 - \frac{3}{4}x_4 \\ x_4 \end{bmatrix}^\top \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & \frac{3}{2} & 0 & 0 \\ 0 & 0 & \frac{4}{3} & 0 \\ 0 & 0 & 0 & \frac{5}{4} \end{bmatrix} \begin{bmatrix} x_1 - \frac{1}{2}x_2 \\ x_2 - \frac{2}{3}x_3 \\ x_3 - \frac{3}{4}x_4 \\ x_4 \end{bmatrix},$$

that is,

$$x^\top Hx = 2\left(x_1 - \frac{1}{2}x_2\right)^2 + \frac{3}{2}\left(x_2 - \frac{1}{2}x_3\right)^2 + \frac{4}{3}\left(x_3 - \frac{1}{2}x_4\right)^2 + \frac{5}{4}x_4^2. \qquad (26.4)$$

Thus using the $LDL^\top$-factorization, the quadratic form given by (26.3) can be written as a sum of squares. An alternative way of doing this is by the completion of squares. We describe this below by means of the same example.

First we eliminate mixed terms which contain the factor x_1 as follows:

$$2x_1^2 - 2x_1x_2 = 2(x_1^2 - x_1x_2) = 2\left(\left(x_1 - \frac{1}{2}x_2\right)^2 - \frac{1}{4}x_2^2\right).$$

This gives

$$x^\top Hx = 2\left(x_1 - \frac{1}{2}x_2\right)^2 + \frac{3}{2}x_2^2 - 2x_2x_3 + 2x_3^2 - 2x_3x_4 + 2x_4^2.$$

Next we eliminate the mixed terms that contain the factor x_2 as follows:

$$\frac{3}{2}x_2^2 - 2x_2x_3 = \frac{3}{2}\left(\left(x_2 - \frac{2}{3}x_3\right)^2 - \frac{4}{9}x_3^2\right).$$

This gives

$$x^\top Hx = 2\left(x_1 - \frac{1}{2}x_2\right)^2 + \frac{3}{2}\left(x_2 - \frac{2}{3}x_3\right)^2 + \frac{4}{3}x_3^2 - 2x_3x_4 + 2x_4^2.$$

Finally, we eliminate the mixed terms that contain the factor x_3 as follows:

$$\frac{4}{3}x_3^2 - 2x_3x_4 + 2x_4^2 = \frac{4}{3}\left(\left(x_3 - \frac{3}{4}x_4\right)^2 - \frac{9}{16}x_4^2\right).$$

This gives

$$x^\top Hx = 2\left(x_1 - \frac{1}{2}x_2\right)^2 + \frac{3}{2}\left(x_2 - \frac{2}{3}x_3\right)^2 + \frac{4}{3}\left(x_3 - \frac{3}{4}x_4\right)^2 + \frac{5}{4}x_4^2,$$

which is the same as (26.4).

26.9. $LDL^\top$-factorization: semidefinite case

We have seen how one can determine whether or not a given symmetric matrix is positive definite. A natural question which then arises is whether there is a similar procedure also for determining if it is positive *semi*definite. The answer is yes, and one can do so with a modified $LDL^\top$-factorization, which allows the diagonal element $h_{ii}^{(i)}$ to be equal to 0.

Theorem 26.21. *A symmetric $H \in \mathbb{R}^{n\times n}$ is positive semidefinite iff*

(1) *there exists a lower triangular matrix $L \in \mathbb{R}^{n\times n}$ with 1's on the diagonal (all the $l_{ii} = 1$), and*

(2) *there exists a diagonal matrix $D \in \mathbb{R}^{n\times n}$ with all diagonal entries nonnegative (all the $d_i \geq 0$),*

(3) *such that $H = LDL^\top$.*

In this section we will give a constructive proof of this result by describing an algorithm for the $LDL^\top$-factorization of positive semidefinite matrices.

First assume that L and D are as above. Then D is positive semidefinite. Thus it follows that $H = LDL^\top = (L^\top)^\top DL^\top$ is also positive semidefinite, by the properties of positive semidefinite matrices we had studied earlier.

In the remainder of this chapter, we will assume that we have been given a H which is positive semidefinite, and we will show how one can determine L and D with the properties above, such that H has the factorization $H = LDL^\top$.

Again for the simplicity of exposition, assume that $n = 4$. Then one can try to use the method we learnt in Section 26.6. But now when H is not necessarily positive definite, it can very well happen that some diagonal element $h_{ii}^{(i)} \leq 0$. But since H is positive semidefinite, it cannot[2] be the case that $h_{ii}^{(i)} < 0$. Thus the "worst" that can happen is that for an i or a few i's $h_{ii}^{(i)} = 0$. Take for example the case that $h_{22}^{(2)} = 0$. But since H is positive semidefinite, it follows then that $h_{23}^{(2)} = h_{32}^{(2)} = h_{24}^{(2)} = h_{42}^{(2)} = 0$, by the property of positive semidefinite matrices we had seen earlier.

But then we arrive at the following scenario:

$$E_1 H E_1^\top = \begin{bmatrix} h_{11}^{(1)} & 0 & 0 & 0 \\ 0 & h_{22}^{(2)} & h_{23}^{(2)} & h_{24}^{(2)} \\ 0 & h_{32}^{(2)} & h_{33}^{(2)} & h_{34}^{(2)} \\ 0 & h_{42}^{(2)} & h_{43}^{(2)} & h_{44}^{(2)} \end{bmatrix} = \begin{bmatrix} h_{11}^{(1)} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & h_{33}^{(2)} & h_{34}^{(2)} \\ 0 & 0 & h_{43}^{(2)} & h_{44}^{(2)} \end{bmatrix}.$$

But then we can simply put $E_2 = I$, so that

$$E_2 E_1 H E_1^\top E_2^\top = \begin{bmatrix} h_{11}^{(1)} & 0 & 0 & 0 \\ 0 & h_{22}^{(2)} & 0 & 0 \\ 0 & 0 & h_{33}^{(3)} & h_{34}^{(3)} \\ 0 & 0 & h_{43}^{(3)} & h_{44}^{(3)} \end{bmatrix},$$

with $h_{22}^{(2)} = 0$ and $h_{ij}^{(3)} = h_{ij}^{(2)}$ for $i = 3, 4$ and $j = 3, 4$.

When we compare this with the corresponding expression from Section 26.6, we note that the difference here is the we now have $h_{22}^{(2)} = 0$ as opposed to having it > 0 before. (Also, E_2 is now the identity matrix, which was typically not the case in Section 26.6, unless both l_{32} and l_{42} happened to be equal to 0.)

Next we continue with the same procedure as in Section 26.6. When we have completed it, we have

$$E_3 E_2 E_1 H E_1^\top E_2^\top E_3^\top = \begin{bmatrix} h_{11}^{(1)} & 0 & 0 & 0 \\ 0 & h_{22}^{(2)} & 0 & 0 \\ 0 & 0 & h_{33}^{(3)} & 0 \\ 0 & 0 & 0 & h_{44}^{(4)} \end{bmatrix} =: D,$$

where all the $h_{ii}^{(i)} \geq 0$. By setting $L := E_1^{-1} E_2^{-1} E_3^{-1}$, we have that $H = LDL^\top$, where L has the form

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ l_{21} & 1 & 0 & 0 \\ l_{31} & l_{32} & 1 & 0 \\ l_{41} & l_{42} & l_{43} & 1 \end{bmatrix}.$$

[2] Here we also use the fact that the E_k's are all invertible.

26.10. A new example of $LDL^\top$-factorization

We will find out if the following matrix H is positive definite, or positive semidefinite or neither, by carrying out the $LDL^\top$-factorization:

$$H = \begin{bmatrix} 1 & -1 & 0 & 0 \\ -1 & 2 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

First add 1 times the first row to the second row and then add 1 times the first column to the second column. Thus

$$E_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad E_1 H E_1^\top = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & -1 & 2 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

Now add 1 times the second row to the third row and then add 1 times the second column to the third column. Thus

$$E_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \text{ and } E_2 E_1 H E_1^\top E_2^\top = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & -1 & 1 \end{bmatrix}.$$

Now add 1 times the third row to the fourth row and then add 1 times the third column to the fourth column. Thus

$$E_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \text{ and } E_3 E_2 E_1 H E_1^\top E_2^\top E_3^\top = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Thus the $LDL^\top$-factorization is completed, and we have $H = LDL^\top$, where

$$L = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix} \quad \text{and} \quad D = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

Since all the diagonal elements of D are ≥ 0, we conclude that H is positive semidefinite. However, it is not positive definite, since there is a diagonal element of D which is equal to 0.

Exercise 26.22. Determine whether H is positive definite or positive semidefinite or neither.

$$H = \begin{bmatrix} 1 & 2 & -1 & 0 \\ 2 & 5 & -1 & -1 \\ -1 & -1 & 0 & -1 \\ 0 & -1 & -1 & 0 \end{bmatrix}, \quad H' = \begin{bmatrix} 1 & -2 & 1 & 0 \\ -2 & 5 & -3 & 1 \\ 1 & -3 & 2 & -1 \\ 0 & 1 & -1 & 3 \end{bmatrix}.$$

Bibliography

[AEP] N. Andréasson, A. Evgrafov and M. Patriksson. *An Introduction to Continuous Optimization.* Studentlitteratur, Lund, 2005.

[GNS] I. Griva, S. Nash and A. Sofer. *Linear and Nonlinear Optimization.* Second edition. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2009.

[LY] D.G. Luenberger and Y. Ye. *Linear and Nonlinear Programming.* Third edition. International Series in Operations Research and Management Science, 116, Springer, New York, 2008.

[R] W. Rudin. *Principles of Mathematical Analysis.* Third edition. McGraw-Hill, Singapore, 1976.

[KS1] K. Svanberg. *Linjär Optimering.* Optimeringslära och systemteori, Institutionen för matematik, Kungliga Tekniska Högskolan, Stockholm, 2007.

[KS2] K. Svanberg. *Kvadratisk Optimering.* Optimeringslära och systemteori, Institutionen för matematik, Kungliga Tekniska Högskolan, Stockholm, 2007.

[KS3] K. Svanberg. *Ickelinjär Optimering.* Optimeringslära och systemteori, Institutionen för matematik, Kungliga Tekniska Högskolan, Stockholm, 2007.

[KS4] K. Svanberg. *Lite Blandad Optimeringsteori.* Optimeringslära och systemteori, Institutionen för matematik, Kungliga Tekniska Högskolan, Stockholm, 2007.

[KS5] K. Svanberg. *Linjär Algebra för Optimerare.* Optimeringslära och systemteori, Institutionen för matematik, Kungliga Tekniska Högskolan, Stockholm, 2007.

[T] S. Treil. *Linear Algebra Done Wrong.* Course notes, Brown University, Providence, R.I., August 2009.

Index