UNIVERSITY OF WATERLOO

Nonlinear Optimization (CO 367)

E. de Klerk, C. Roos, and T. Terlaky

Waterloo, February 23, 2006

Contents

Chapter 0

Introduction

Nothing takes place in the world whose meaning is not that of some maximum or minimum.
L. Euler (1707 – 1783)

0.1 The general nonlinear optimization problem

The general nonlinear optimization (NLO) problem can be written as follows:

$$\begin{array}{lll} (NLO) & \inf & f(x) \\ & \text{s.t.} & h_i(x) = 0, \quad i \in I = \{1, \cdots, p\} \\ & & g_j(x) \leq 0, \quad j \in J = \{1, \cdots, m\} \\ & & x \in \mathcal{C}, \end{array} \tag{1}$$

where $x \in \mathbb{R}^n$, $\mathcal{C} \subseteq \mathbb{R}^n$ is a certain set and $f, h_1, \cdots, h_p, g_1, \cdots, g_m$ are functions defined on $\mathcal{C}$ (or on an open set that contains the set $\mathcal{C}$). The set of feasible solutions will be denoted by $\mathcal{F}$, hence

$$\mathcal{F} = \{x \in \mathcal{C} \mid h_i(x) = 0, \;\; i = 1, \cdots, p \text{ and } g_j(x) \leq 0, \;\; j = 1, \cdots, m\}.$$

We will use the following standard terminology:

- The function f is called the *objective function* of (NLO) and $\mathcal{F}$ is called the *feasible set* (or feasible region);
- If $\mathcal{F} = \emptyset$ then we say that problem (NLO) is *infeasible*;
- If f is not below bounded on $\mathcal{F}$ then we say that problem (NLO) is *unbounded*;
- If the infimum of f over $\mathcal{F}$ is attained at $\bar{x} \in \mathcal{F}$ then we call $\bar{x}$ an *optimal solution* (or minimum or minimizer) of (NLO) and $f(\bar{x})$ the *optimal (objective) value* of (NLO).

Example 0.1 As an example, consider minimization of the 'humpback function'(see Figure 1):

$$\min \; x_1^2(4 - 2.1x_1^2 + \frac{1}{3}x_1^4) + x_1x_2 + x_2^2(-4 + 4x_2^2),$$

subject to the constraints $-2 \leq x_1 \leq 2$ and $-1 \leq x_2 \leq 1$. Note that the feasible set here is simply the rectangle:

$$\mathcal{F} = \{(x_1, x_2) \; : \; -2 \leq x_1 \leq 2, \; -1 \leq x_2 \leq 1\}.$$

This NLO problem has two optimal solutions, namely $(0.0898, \; -0.717)$ and $(-0.0898, \; 0.717)$, as one can (more or less) verify by looking at the contours of the objective function in Figure 1. ∗

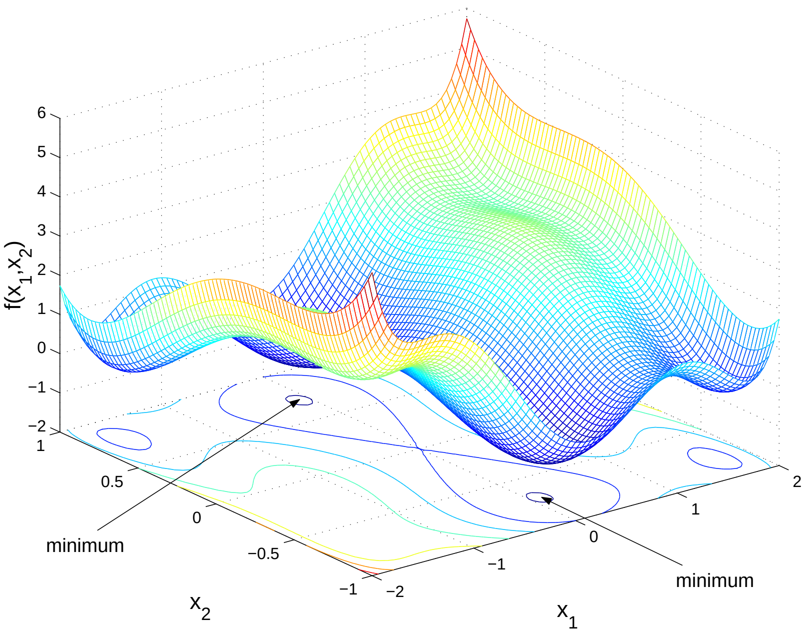


Figure 1: Example of a nonlinear optimization problem with the 'humpback' function as objective function. The contours of the objective function are shown.

Notation

Matrices will be denoted by capital letters $(A, B, P, \ldots)$, vectors by small Latin letters and components of vectors and matrices by the indexed letters [e.g. $z = (z_1, \ldots, z_n)$, $A = (a_{ij})_{i=1}^{m}{}_{j=1}^{n}$]. For the purpose of matrix-vector multiplication, vectors in $\mathbb{R}^n$ will always be viewed as $n \times 1$ matrices (column vectors). Index sets will be denoted by I, J and K.

We now define some classes of NLO problems. Recall that $f : \mathbb{R}^n \to \mathbb{R}$ is called a *quadratic* function if there is a square matrix $Q \in \mathbb{R}^{n\times n}$, a vector $c \in \mathbb{R}^n$ and a number $\gamma \in \mathbb{R}$ such that

$$f(x) = \frac{1}{2}x^T Qx + c^T x + \gamma \text{ for all } x \in \mathbb{R}^n;$$

If $Q = 0$ then f is called *affine* and if $Q = 0$ and $\gamma = 0$ then f is called *linear*. We will abuse this terminology a bit by sometimes referring to affine functions as linear.

Classification of nonlinear optimization problems

We now list a few important classes of optimization problems, with reference to the general problem (1):

Linear Optimization (LO): The functions $f, h_1, \cdots, h_p, g_1, \cdots, g_m$ are affine and the set $\mathcal{C}$ either equals to $\mathbb{R}^n$ or to the nonnegative orthant $\mathbb{R}^n_+$ of $\mathbb{R}^n$. Linear optimization is often called *linear programming*. The reader is probably familiar with the simplex algorithm for solving LO problems.

Unconstrained Optimization: The index sets I and J are empty and $\mathcal{C} = \mathbb{R}^n$.

Quadratic Optimization (QO): The objective function f is quadratic, and all the constraint functions $h_1, \cdots, h_p, g_1, \cdots, g_m$ are affine and the set $\mathcal{C}$ is $\mathbb{R}^n$ or the positive orthant $\mathbb{R}^n_+$ of $\mathbb{R}^n$.

Quadratically Constrained Quadratic Optimization: Same as QO, except that the functions $g_1, \cdots, g_m$ are quadratic.

0.2 A short history of nonlinear optimization

As far as we know, Euclid's book *Elements* was the first mathematical textbook in the history of mankind (4th century B.C.). It contains the following optimization problem.

In a given triangle ABC inscribe a parallelogram $ADEF$ such that $EF \| AB$ and $DE \| AC$ and such that the area of this parallelogram is maximal.

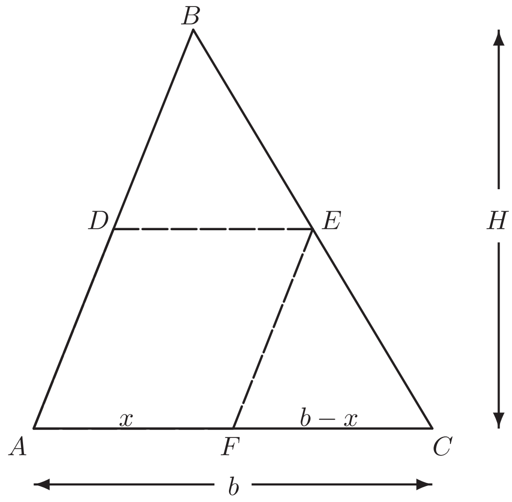


Figure 2: Illustration of Euclid's problem.

Let H denote the height of the triangle, and let b indicate the length of the edge AC. Every inscribed parallelogram of the required form is uniquely determined by choosing a vertex F at a distance $x < b$ from A on the edge AC (see Figure 2).

Exercise 0.1 *Show that Euclid's problem can be formulated as the quadratic optimization problem (QO):*

$$\max_{0<x<b} \frac{Hx(b-x)}{b}. \tag{2}$$

◁

Euclid could show that the maximum is obtained when $x = \frac{1}{2}b$, by using geometric reasoning. A unified methodology for solving nonlinear optimization problems would have to wait until the development of calculus in the 17*th* century. Indeed, in any modern text on calculus we learn to solve problems like (2) by setting the derivative of the *objective function* $f(x) := \frac{Hx(b-x)}{b}$ to zero, and solving the resulting equation to obtain $x = \frac{1}{2}b$.

This modern methodology is due to Fermat (1601 – 1665). Because of this work, Lagrange (1736 – 1813) stated clearly that he considered Fermat to be the inventor of calculus (as opposed to Newton (1643 – 1727) and Liebnitz (1646 – 1716) who were later locked in a bitter struggle for this honour). Lagrange himself is famous for extending the method of Fermat to solve (equality) constrained optimization problems by forming a function now known as the Lagrangian, and applying Fermat's method to the Lagrangian. In the words of Lagrange himself:

> *One can state the following general principle. If one is looking for the maximum or minimum of some function of many variables subject to the condition that these variables are related by a constraint given by one or more equations, then one should add to the function*

whose extremum is sought the functions that yield the constraint equations each multiplied by undetermined multipliers and seek the maximum or minimum of the resulting sum as if the variables were independent. The resulting equations, combined with the constraint equations, will serve to determine all unknowns.

To better understand what Lagrange meant, consider the general NLO problem with only equality constraints ($J = \emptyset$ and $\mathcal{C} = \mathbb{R}^n$ in (1)):

$$\begin{array}{ll} \inf & f(x) \\ \text{s.t.} & h_i(x) = 0, \quad i \in I = \{1, \cdots, p\} \\ & x \in \mathbb{R}^n. \end{array}$$

Now define the Lagrangian function

$$L(x, y) := f(x) + \sum_{i=1}^{p} y_i h_i(x).$$

The new variables y_i are called (Lagrange) multipliers, and are the *undetermined multipliers* Lagrange referred to. Now apply Fermat's method to find the minimum of the function $L(x, y)$, 'as if the variables x and y were independent'. In other words, solve the system of nonlinear equations defined by setting the gradient of L to zero, and retaining the feasibility conditions $h_i(x) = 0$ $(i \in I)$:

$$\nabla L(x, y) = 0, \; h_i(x) = 0 \; (i \in I). \tag{3}$$

If x^* is an optimal solution of NLO then there now exists a vector $y^* \in \mathbb{R}^p$ such that (x^*, y^*) is a solution of the nonlinear equations (3). We can therefore solve the nonlinear system (3) and the x-part of one of the solutions of (3) will yield the optimal solution of (NLO) (provided it exists). Note that it is difficult to know beforehand whether an optimal solution of (NLO) exists.

This brings us to the problem of solving a system of nonlinear equations. Isaac Newton lent his name to perhaps the most widely known algorithm for this problem. In conjunction with Fermat and Lagrange's methods, this yielded one of the first optimization algorithms. It is interesting to note that even today, Newton's optimization algorithm is the most widely used and studied algorithm for nonlinear optimization. The most recent optimization algorithms, namely interior point algorithms, use this method as their 'engine'.

The study of nonlinear optimization in the time of Fermat, Lagrange, Euler and Newton was driven by the realization that many physical principles in nature can be explained via optimization (extremum) principles. For example, the well known principle of Fermat for the refraction of light may be stated as:

in an inhomogeneous medium, light travels from one point to another along the path requiring the shortest time.

Similarly, it was known that many problems in (celestial) mechanics could be formulated as extremum problems.

We return to the problem of deciding whether (NLO) has an optimal solution at all. In the 19th century, Karl Weierstrass (1815 – 1897) proved the famous result — known to any student of analysis — that a continuous function attains its infimum and supremum on a compact set. This gave a practical sufficient condition for the existence of optimal solutions.

In modern times, nonlinear optimization is used in optimal engineering design, finance, statistics and many other fields. It has been said that we live in *the age of optimization*, where everything has to be better and faster than before. Think of designing a car with minimal air resistance, a bridge of minimal weight that still meets essential specifications, a stock portfolio where the risk is minimal and the expected return high; the list is endless. If you can make a mathematical model of your decision problem, then you can optimize it!

Outline of this course

This short history of nonlinear optimization is of course far from complete and has served only to introduce some of the most important topics that will be studied in this course. In Chapters 1 and 2 we will study the methods of Fermat and Lagrange. So-called duality theory based on the methodology of Lagrange will follow in Chapter 3. Then we will turn our attention to optimization algorithms in the remaining chapters. First we will study classical methods like the (reduced) gradient method and Newton's method (Chapter 4), before turning to the modern application of Newton's method in interior point algorithms (Chapter 6). Finally, we conclude with a chapter on special classes of structured nonlinear optimization problems that can be solved very efficiently by interior point algorithms.

In the rest of this chapter we give a few more examples of historical and practical problems to give some idea of the field of nonlinear optimization.

0.3 Some historical examples

These examples of historical nonlinear optimization problems are taken from the wonderful book *Stories about Maxima and Minima* by V.M. Tikhomirov (AMS, 1990). For more background and details the reader is referred to that book. We only state the problems here — the solutions will be presented later in this book in the form of exercises, once the reader has studied optimality conditions. Of course, the mathematical tools we will employ were not available to the originators of the problems. For the (ingenious) original historical solutions the reader is again referred to the book by Tikhomirov.

0.3.1 Tartaglia's problem

Niccolo Tartaglia (1500–1557) posed the following problem:

How do you divide the number 8 into two parts such that the result of multiplying the product of the parts by their difference will be maximal?

If we denote the unknown parts by x_1 and x_2, we can restate the problem as the nonlinear optimization problem:

$$\max\ x_1 x_2 (x_1 - x_2) \tag{4}$$

subject to the constraints

$$x_1 + x_2 = 8,\ x_1 \geq 0,\ x_2 \geq 0.$$

Tartaglia knew that the correct answer is $x_1 = 4 + (4/\sqrt{3})$, $x_2 = 4 - (4/\sqrt{3})$. How can one prove that this is correct? (Solution via Exercise 2.4.)

0.3.2 Kepler's problem

The famous astronomer Johannes Kepler was so intrigued by the geometry of wine barrels that he wrote a book about it in 1615: *New solid geometry of wine barrels*. In this work he considers the following problem (among others).

Given a sphere, inscribe a cylinder of maximal volume.

Kepler knew the cylinder of maximal volume is such that the ratio of its base diameter to the height is $\sqrt{2}$. (And of course the diagonal of the cylinder has the same length as the diameter of the sphere.) How can one show that this is correct? (Solution via Exercises 0.2 and 2.2.)

Exercise 0.2 *Formulate Kepler's problem as a nonlinear optimization problem (NLO).* ◁

0.3.3 Steiner's problem

In the plane of a given triangle, find a point such that the sum of its distances from the vertices of the triangle is minimal.

This problem was included in the first book on optimization, namely *On maximal and minimal values* by Viviani in 1659.

If we denote the vertices of the triangle by the three given vectors $a \in \mathbb{R}^2$, $b \in \mathbb{R}^2$ and $c \in \mathbb{R}^2$, and let $x = [x_1 \ x_2]^T$ denote the vector with the (unknown) coordinates of the point we are looking for, then we can formulate Steiner's problem as the following nonlinear optimization problem.

$$\min_{x \in \mathbb{R}^2} \|x - a\| + \|x - b\| + \|x - c\|$$

The solution is known as the *Torricelli point.* How can one find the Torricelli point for any given triangle? (Solution via Exercise 2.3.)

0.4 Quadratic optimization examples

In countless applications we wish to solve a linear system of the form $Ax = b$. If this system is *over-determined* (more equations than variables), then we can still obtain the so-called *least squares solution* by solving the NLO problem:

$$\min_x \|Ax - b\|^2. \tag{5}$$

Since

$$\|Ax - b\|^2 = (Ax - b)^T (Ax - b) = x^T A^T Ax - 2b^T Ax + b^T b,$$

it follows that problem (5) is a quadratic optimization (QO) problem.

Below we give examples of the least squares and other quadratic optimization problems.

0.4.1 The concrete mixing problem: least square estimation

In civil engineering, different sorts of concrete are needed for different purposes. One of the important characteristics of the concrete are its sand-and-gravel composition, i.e. what percentages of the stones in the sand-and-gravel mixture belong to a certain stone size categories. For each sort concrete the civil engineers can give an ideal sand-and-gravel composition that ensures the desired strength with minimal cement content.

Unfortunately, in the sand-and-gravel mines, such ideal composition can not be found in general. The solution is to mix different sand-and-gravel mixtures in order to approximate the desired quality as closely as possible.

Mathematical model

Let us assume that we have n different stone size categories. The ideal mixture for our actual purpose is given by the vector $c = (c_1, c_2, \cdots, c_n)^T$, where $0 \le c_i \le 1$ for all $i = 1, \cdots, n$ and $\sum_{i=1}^n c_i = 1$. The components c_i indicate what fraction of the sand-and-gravel mixture belongs to the i-th stone size category. Further, let assume that we can get sand-and-gravel mixtures from m different mines, and the stone composition at each mine $j = 1, \cdots, m$ is given by the vectors $A_j = (a_{1j}, \cdots, a_{nj})^T$, where $0 \le a_{ij} \le 1$ for all $i = 1, \cdots, n$ and $\sum_{i=1}^n a_{ij} = 1$. The goal is to find the best possible approximation of the ideal mixture by using the material offered by the m mines.

Let $x = (x_1, \cdots, x_m)$ be a the vector of unknown percentages in the mixture, i.e.

$$\sum_{j=1}^m x_j = 1, \qquad x_j \ge 0.$$

The resulting mixture composition

$$z = \sum_{i=1}^{m} A_j x_j$$

should be as close as possible to the ideal one, i.e. we have to minimize

$$\|z - c\|^2 = (z - c)^T (z - c) = \sum_{j=1}^{n} (z_i - c_i)^2.$$

This optimization problem is a linearly constrained QO problem. We can further simplify this problem by eliminating variable z. Then, introducing matrix $A = (A_1, \cdots, A_m)$ composed from the vectors A_j as its columns, we have the following simple QO problem:

$$\begin{aligned} \min\ (Ax - c)^T (Ax - c) & \\ e^T x = & \ \ 1 \\ x \geq & \ \ 0. \end{aligned}$$

Exercise 0.3 *In the above concrete mixing problem the deviation of the mixture from the targeted ideal composition is measured by the Euclidean distance of the vectors $z = Ax$ and c. The distance of two vectors can be measured alternatively by e.g. the $\| \cdot \|_1$ or by the $\| \cdot \|_\infty$ norms. Restate the mixing problem by using these norms and show that this way, in both cases, pure linear optimization problems can be obtained.* $\triangleleft$

0.4.2 Portfolio analysis (mean–variance models)

An important application of the QO problem is the computation of an efficient frontier for mean–variance models, introduced by Markowitz [31]. Given assets with expected return r_i and covariances v_{ij}, the problem is to find portfolios of the assets that have minimal variance given a level of total return, and maximal return given a level of total variance. Mathematically, if x_i is the proportion invested in asset i then the basic mean–variance problem is

$$\min_x \left\{ \frac{1}{2} x^T V x \ : \ e^T x = 1, \ r^T x = \lambda, \ Dx = d, \ x \geq 0 \right\},$$

where e is an all–one vector, and $Dx = d$ may represent additional constraints on the portfolios to be chosen (for instance those related to volatility of the portfolio). This problem can be viewed as a parametric QO problem, with parameter λ representing the total return of investment. The so-called efficient frontier is then just the optimal value function.

Example 0.2 Consider the following MV-model,

$$\min_x \left\{ x^T V x \ : \ e^T x = 1, \ r^T x = \lambda, \ x \geq 0 \right\}$$

where

$$V = \begin{bmatrix} 0.82 & -0.23 & 0.155 & -0.013 & -0.314 \\ -0.23 & 0.484 & 0.346 & 0.197 & 0.592 \\ 0.155 & 0.346 & 0.298 & 0.143 & 0.419 \\ -0.013 & 0.197 & 0.143 & 0.172 & 0.362 \\ -0.314 & 0.592 & 0.419 & 0.362 & 0.916 \end{bmatrix},$$

$$r = \begin{pmatrix} 1.78 & 0.37 & 0.237 & 0.315 & 0.49 \end{pmatrix}^T.$$

One can check (e.g. by using MATLAB) that for $\lambda > 1.780$ or $\lambda < 0.237$ the QO problem is infeasible. For the values $\lambda \in [0.237, 1.780]$ the QO problem has an optimal solution. $*$

Exercise 0.4 *A mathematical description of this and related portfolio problems is given at:*

http://www-fp.mcs.anl.gov/otc/Guide/CaseStudies/port/formulations.html

Choose your own stock portfolio at the website:

http://www-fp.mcs.anl.gov/otc/Guide/CaseStudies/port/demo.html

and solve this problem remotely via internet to obtain the optimal way of dividing your capital between the stocks you have chosen. In doing this you are free to set the level of risk you are prepared to take. Give the mathematical description of the problem you have solved and report on your results. ◁

Chapter 1

Convex Analysis

If the nonlinear optimization problem (NLO) has a convex objective function and the feasible set is a convex set, then the underlying mathematical structure is much richer than in the general case. For example, one can formulate necessary and sufficient conditions for the existence of optimal solutions in this case. It is therefore important to study convexity in some detail.

1.1 Convex sets and convex functions

Given two points x^1 and x^2 in $\mathbb{R}^n$, any point on the line connecting them is called a *convex combinations of x^1 and x^2*. Formally we have the following definition.

Definition 1.1 *Let two points $x^1, x^2 \in \mathbb{R}^n$ and $0 \leq \lambda \leq 1$ be given. Then the point*

$$x = \lambda x^1 + (1 - \lambda)x^2$$

is a convex combination *of the two points x^1, x^2.*

The set $\mathcal{C} \subset \mathbb{R}^n$ is called convex, *if all convex combinations of any two points $x^1, x^2 \in \mathcal{C}$ are again in $\mathcal{C}$.*

In other words, the line segment connecting two arbitrary points of a convex set is contained in the set.

Figure 1.1 and Figure 1.2 show examples of convex and nonconvex sets in the plane.

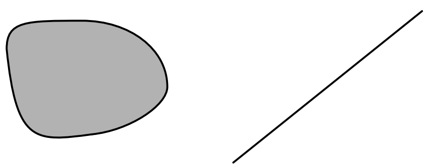


Figure 1.1: Convex sets

Exercise 1.1 *We can define the convex combination of k points as follows. Let the points $x^1, \cdots, x^k \in \mathbb{R}^n$ and $0 \leq \lambda_1, \cdots, \lambda_k$ with $\sum_{i=1}^k \lambda_i = 1$ be given. Then the vector*

$$x = \sum_{i=1}^{k} \lambda_i x^i$$

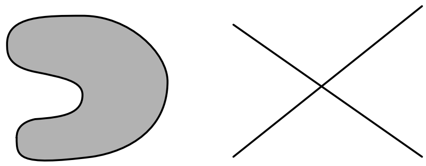


Figure 1.2: Non convex sets

is a convex combination of the given points.

Prove that the set $\mathcal{C}$ is convex if and only if for any $k \geq 2$ all convex combinations of any k points from $\mathcal{C}$ are also in $\mathcal{C}$. ◁

The intersection of (possibly infinitely many) convex sets is again a convex set.

Theorem 1.2 *Let C_i $(i = 1, \ldots)$ be a collection of convex sets. The set*

$$C := \cap_{i=1}^{\infty} C_i$$

ic convex.

Exercise 1.2 *Prove Theorem 1.2.* ◁

Another fundamental property of a convex set is that its closure is again a convex set.

Theorem 1.3 *Let $C \subset \mathbb{R}^n$ be a convex set and let $cl(C)$ denote its closure. Then $cl(C)$ is a convex set.*

We now turn our attention to so-called *convex functions*. A parabola $f(x) = ax^2 + bx + c$ with $a > 0$ is a familiar example of a convex function. Intuitively, it is easier to characterize minima of convex functions than minima of more general functions, and for this reason we will study convex functions in some detail.

Definition 1.4 *A function $f : \mathcal{C} \to R$ defined on a convex set $\mathcal{C}$ is called* convex *if for all $x^1, x^2 \in \mathcal{C}$ and $0 \leq \lambda \leq 1$ one has*

$$f(\lambda x^1 + (1 - \lambda)x^2) \leq \lambda f(x^1) + (1 - \lambda) f(x^2).$$

Exercise 1.3 *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be defined by $f(x) = \|x\|$ for some norm $\|\cdot\|$. Prove that f is a convex function.* ◁

Exercise 1.4 *Show that the following univariate functions are* not *convex:*

$$f(x) = \sin x, \quad f(x) = e^{-x^2}, \quad f(x) = x^3.$$

◁

Definition 1.5 *The epigraph of a function $f : \mathcal{C} \to R$, where $\mathcal{C} \subset \mathbb{R}^n$, is the $(n + 1)$-dimensional set*

$$\{(x, \tau) \; : \; f(x) \leq \tau, x \in \mathcal{C}, \; \tau \in \mathbb{R}\}.$$

Figure 1.3 illustrates the above definition.

Exercise 1.5 *Prove that the function $f : \mathcal{C} \to R$ defined on the convex set $\mathcal{C}$ is convex if and only if the* epigraph *of f is a convex set.* ◁

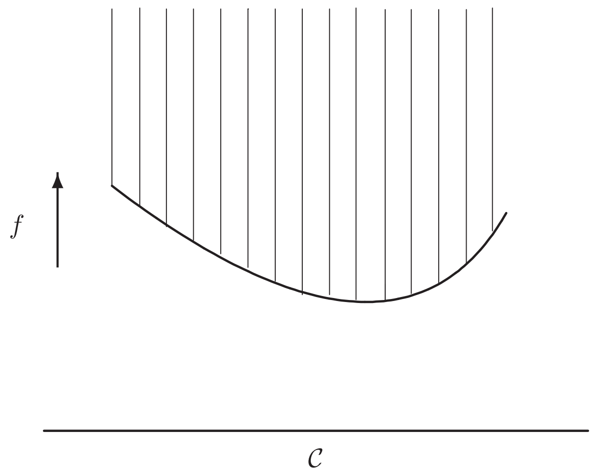


Figure 1.3: The epigraph of a convex function f.

We also will need the concept of a *strictly* convex function. These are convex functions with the nice property that — if a minimum of the function exists — then this minimum is unique.

Definition 1.6 *A (convex) function $f : \mathcal{C} \to \mathbb{R}$, defined on a convex set $\mathcal{C}$, is called* strictly convex *if for all $x^1, x^2 \in \mathcal{C}$ and $0 < \lambda < 1$ one has*

$$f(\lambda x^1 + (1 - \lambda)x^2) < \lambda f(x^1) + (1 - \lambda)f(x^2).$$

We have seen in Exercise 1.5 that a function is convex if and only if its epigraph is convex.

Also, the next exercise shows that a quadratic function is convex if and only if the matrix Q in its definition is positive-semidefinite (PSD).

Exercise 1.6 *Let a symmetric matrix $Q \in \mathbb{R}^{n\times n}$, a vector $c \in \mathbb{R}^n$ and a number $\gamma \in \mathbb{R}$ be given. Prove that the quadratic function*

$$\frac{1}{2}x^T Qx + c^T x + \gamma$$

is convex on $\mathbb{R}^n$ if and only if the matrix Q is PSD, and strictly convex if and only if Q is positive definite. ◁

Exercise 1.7 *Decide whether the following quadratic functions are convex or not. (Hint: use the result from the previous exercise.)*

$$f(x) = x_1^2 + 2x_1x_2 + x_2^2 + 5x_1 - x_2 + \frac{1}{2}, \quad f(x) = x_1^2 + x_2^2 + x_3^2 - 2x_1x_2 - 2x_1x_3 - 2x_2x_3.$$

◁

If we multiply a convex function by -1, then we get a so-called concave function.

Definition 1.7 *A function $f : \mathcal{C} \to \mathbb{R}$, defined on a convex set $\mathcal{C}$, is called (strictly)* concave *if the function $-f$ is (strictly) convex.*

Note that we can change the problem of maximizing a concave function into the problem of minimizing a convex function.

Now we review some further properties of convex sets and convex functions that are necessary to understand and analyze convex optimization problems. First we review some elementary properties of convex sets.

1.2 More on convex sets

1.2.1 Convex hulls and extremal sets

For any set $\mathcal{S} \subset \mathbb{R}^n$ we define the smallest convex set that contains it, the so-called *convex hull* of $\mathcal{S}$, as follows.

Definition 1.8 *Let $\mathcal{S} \subset \mathbb{R}^n$ be an arbitrary set. The set*

$$\mathrm{conv}(\mathcal{S}) := \left\{ x \;\middle|\; x = \sum_{i=1}^{k} \lambda_i x^i,\ x^i \in \mathcal{S},\ i = 1, \cdots, k;\ \lambda_i \in [0,1],\ \sum_{i=1}^{k} \lambda_i = 1,\ k \geq 1 \right\}$$

is called the convex hull *of the set $\mathcal{S}$.*

Observe that $\mathrm{conv}(\mathcal{S})$ is generated by taking all possible convex combinations of points from $\mathcal{S}$.

We now define some important convex subsets of a given convex set $\mathcal{C}$, namely the so-called *extremal sets*, that play an important role in convex analysis. Informally, an extremal set $\mathcal{E} \subset \mathcal{C}$ is a convex subset of $\mathcal{C}$ with the following property: if any point on the line segment connecting two points $x^1 \in \mathcal{C}$ and $x^2 \in \mathcal{C}$ lies in $\mathcal{E}$, then the two points x^1 and x^2 must also lie in $\mathcal{E}$. The faces of a polytope are familiar examples of extreme sets of the polytope.

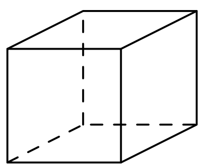
Definition 1.9 *The convex set $\mathcal{E} \subseteq \mathcal{C}$ is an* extremal *set of the convex set $\mathcal{C}$ if, for all $x^1, x^2 \in \mathcal{C}$ and $0 < \lambda < 1$, one has $\lambda x^1 + (1-\lambda)x^2 \in \mathcal{E}$ only if $x^1, x^2 \in \mathcal{E}$.*

An extremal set consisting of only one point is called an *extremal point*. Observe that extremal sets are convex by definition, and the convex set $\mathcal{C}$ itself is always an extremal set of $\mathcal{C}$. It is easy to verify the following result.

Lemma 1.10 *If $\mathcal{E}^1 \subseteq \mathcal{C}$ is an extremal set of the convex set $\mathcal{C}$ and $\mathcal{E}^2 \subseteq \mathcal{E}^1$ is an extremal set of $\mathcal{E}^1$ then $\mathcal{E}^2$ is an extremal set of $\mathcal{C}$.*

***Proof:** Let $x, y \in \mathcal{C}$, $0 < \lambda < 1$ and $z_\lambda = \lambda x + (1-\lambda)y \in \mathcal{E}^2$. Due to $\mathcal{E}^2 \subseteq \mathcal{E}^1$ we have $z_\lambda \in \mathcal{E}^1$, moreover $x, y \in \mathcal{E}^1$ because $\mathcal{E}^1$ is an extremal set of $\mathcal{C}$. Finally, because $\mathcal{E}^2$ is an extremal set of $\mathcal{E}^1$, $x, y \in \mathcal{E}^1$ and $z_\lambda \in \mathcal{E}^2$ we conclude that $x, y \in \mathcal{E}^2$ and thus $\mathcal{E}^2$ is an extremal set of $\mathcal{C}$. □

Example 1.11 Let $\mathcal{C}$ be the cube $\{x \in \mathbb{R}^3 |\ 0 \leq x \leq 1\}$, then the vertices are extremal points, the edges are 1-dimensional extremal sets, the faces are 2-dimensional extremal sets, and the whole cube is a 3-dimensional extremal set of itself.

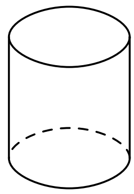


*

Example 1.12 Let $\mathcal{C}$ be the cylinder $\{x \in \mathbb{R}^3 |\ x_1^2 + x_2^2 \leq 1, 0 \leq x_3 \leq 1\}$, then

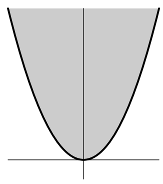
- the points on the circles $\{x \in \mathbb{R}^3 |\ x_1^2 + x_2^2 = 1, x_3 = 1\}$ and $\{x \in \mathbb{R}^3 |\ x_1^2 + x_2^2 = 1, x_3 = 0\}$ are the extremal points,
- the lines $\{x \in \mathbb{R}^3 |\ x_1 = a, x_2 = b, 0 \leq x_3 \leq 1\}$, with $a \in [-1, 1]$ and $b = \sqrt{1-a^2}$ or $b = -\sqrt{1-a^2}$, are the 1-dimensional extremal sets,
- the faces $\{x \in \mathbb{R}^3 |\ x_1^2 + x_2^2 \leq 1, x_3 = 1\}$ and $\{x \in \mathbb{R}^3 |\ x_1^2 + x_2^2 \leq 1, x_3 = 0\}$ are the 2-dimensional extremal sets, and

- the cylinder itself is the only 3-dimensional extremal set.



*

Example 1.13 Let $f(x) = x^2$ and let $\mathcal{C}$ be the epigraph of f, then all points (x_1, x_2) such that $x_2 = x_1^2$ are extremal points. The epigraph itself is the only two dimensional extremal set.



*

Lemma 1.14 *Let $\mathcal{C}$ be a closed convex set. Then all extremal sets of $\mathcal{C}$ are closed.*

In the above examples we have pointed out extremal sets of different dimension without giving a formal definition of what the dimension of a convex set is. To this end, recall from linear algebra that if $\mathcal{L}$ is a (linear) subspace of $\mathbb{R}^n$ and $a \in \mathbb{R}^n$ then $a + \mathcal{L}$ is called an *affine subspace* of $\mathbb{R}^n$. By definition, the dimension of $a + \mathcal{L}$ is the dimension of $\mathcal{L}$.

Definition 1.15 *The smallest affine space $a+\mathcal{L}$ containing a convex set $\mathcal{C} \subseteq \mathbb{R}^n$ is the so-called* affine hull *of $\mathcal{C}$ and denoted by* $\text{aff}(\mathcal{C})$*. The* dimension *of $\mathcal{C}$ is defined as the dimension of* $\text{aff}(\mathcal{C})$.

Given two points $x^1 \in \mathcal{C}$ and $x^2 \in \mathcal{C}$, we call any point that lies on the (infinite) line that passes through x^1 and x^2 an *affine combination* of x^1 and x^2. Formally we have the following definition.

Definition 1.16 *Let two points $x^1, x^2 \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}$ be given. Then the point*

$$x = \lambda x^1 + (1 - \lambda)x^2$$

is an affine combination *of the two points x^1, x^2.*

Observe that in defining the affine combination we do not require that the coefficients λ and $1 - \lambda$ are from the interval $[0, 1]$, while this was required in the definition of the convex combination of points.

Exercise 1.8 *Let $\mathcal{C} \subset \mathbb{R}^n$ be defined by*

$$\mathcal{C} = \left\{ x \;\middle|\; \sum_{i=1}^{n} x_i = 1,\ x \geq 0 \right\}.$$

The set $\mathcal{C}$ is usually called the standard simplex *in $\mathbb{R}^n$.*

1. *Give the extreme points of $\mathcal{C}$; Motivate your answer.*
2. *Show that $\mathcal{C} = \text{conv}\left\{e^1, \ldots, e^n\right\}$, where e^i is the ith standard unit vector.*
3. *What is $\text{aff}(\mathcal{C})$ in this case?*
4. *What is the dimension of $\mathcal{C}$? Motivate your answer.*

◁

Exercise 1.9 *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a given convex set and $k \geq 2$ a given integer. Prove that*

$$\mathrm{aff}(\mathcal{C}) = \left\{ z \;\middle|\; z = \sum_{i=1}^{k} \lambda_i x^i, \; \sum_{i=1}^{k} \lambda_i = 1, \; \lambda_i \in \mathbb{R}, \; x^i \in \mathcal{C}, \; \forall\, i \right\}.$$

◁

Exercise 1.10 *Let $\mathcal{E}$ be an extremal set of the convex set $\mathcal{C}$. Prove that $\mathcal{E} = \mathrm{aff}(\mathcal{E}) \cap \mathcal{C}$. (Hint: Use Exercise 1.9 with $k = 2$.)* ◁

Lemma 1.17 *Let $\mathcal{E}^2 \subset \mathcal{E}^1 \subseteq \mathcal{C}$ be two extremal sets of the convex set $\mathcal{C}$. Then $\dim(\mathcal{E}^2) < \dim(\mathcal{E}^1)$.*

***Proof:** Because $\mathcal{E}^2 \subset \mathcal{E}^1$ we have $\mathrm{aff}(\mathcal{E}^2) \subseteq \mathrm{aff}(\mathcal{E}^1)$. Further, by Exercise 1.10,

$$\mathcal{E}^2 = \mathrm{aff}(\mathcal{E}^2) \cap \mathcal{E}^1.$$

If we assume to the contrary that $\dim(\mathcal{E}^2) = \dim(\mathcal{E}^1)$ then we have $\mathrm{aff}(\mathcal{E}^2) = \mathrm{aff}(\mathcal{E}^1)$ and so

$$\mathcal{E}^2 = \mathrm{aff}(\mathcal{E}^2) \cap \mathcal{E}^1 = \mathrm{aff}(\mathcal{E}^1) \cap \mathcal{E}^1 = \mathcal{E}^1$$

contradicting the assumption $\mathcal{E}^2 \subset \mathcal{E}^1$. □

Lemma 1.18 *Let $\mathcal{C}$ be a nonempty compact (closed and bounded) convex set. Then $\mathcal{C}$ has at least one extremal point.*

***Proof:** Let $\mathcal{F} \subseteq \mathcal{C}$ be the set of points in $\mathcal{C}$ which are furthest from the origin. The set of such points is not empty, because $\mathcal{C}$ is bounded and closed and the norm function is continuous. We claim that any point $z \in \mathcal{F}$ is an extremal point of $\mathcal{C}$.

Let us assume to the contrary that $z \in \mathcal{F}$ is not an extremal point. Then there exist points $x, y \in \mathcal{C}$, both different from z, and a $\lambda \in (0, 1)$ such that

$$z = \lambda x + (1 - \lambda) y.$$

Further, we have $\|x\| \leq \|z\|$ and $\|y\| \leq \|z\|$ because z maximizes the norm of the points over $\mathcal{C}$. Thus by the triangle inequality

$$\|z\| \leq \lambda \|x\| + (1 - \lambda) \|y\| \leq \|z\|$$

which implies that $\|z\| = \|x\| = \|y\|$, i.e. all the three point x, y, z are on the surface of the n-dimensional sphere with radius $\|z\|$ and centered at the origin. This is a contradiction, because these three different points lie on the same line as well. The lemma is proved. □

Observe, that the above proof does not require the use of the origin as reference point. We could choose any point $u \in \mathbb{R}^n$ and prove that the furthest point $z \in \mathcal{C}$ from u is an extremal point of $\mathcal{C}$.

The following theorem shows that a compact convex set is completely determined by its extremal points.

Theorem 1.19 (Krein–Milman Theorem) *Let $\mathcal{C}$ be a compact convex set. Then $\mathcal{C}$ is the convex hull of its extreme points.*

Exercise 1.11 *Let f be a continuous, concave function defined on a compact convex set $\mathcal{C}$. Show that the minimum of f is attained at an extreme point of $\mathcal{C}$. (Hint: Use the Krein-Milman Theorem.)* ◁

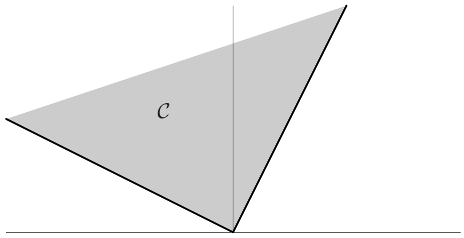
1.2.2 Convex cones

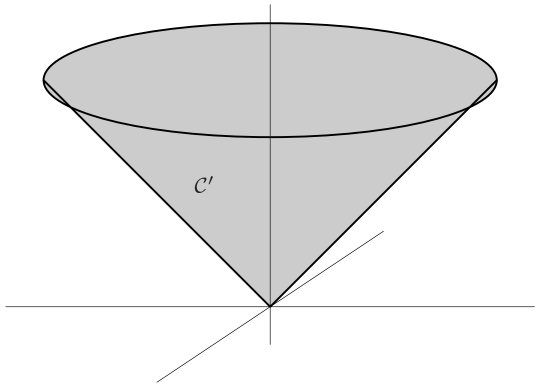
In what follows we define and prove some elementary properties of convex cones.

Definition 1.20 *The set $\mathcal{C} \subset \mathbb{R}^n$ is a* convex cone *if it is a convex set and for all $x \in \mathcal{C}$ and $0 \leq \lambda$ one has $\lambda x \in \mathcal{C}$.*

Example 1.21

- The set $\mathcal{C} = \{(x_1, x_2) \in \mathbb{R}^2 |\ x_2 \geq 2x_1,\ x_2 \geq -\frac{1}{2}x_1\}$ is a convex cone in $\mathbb{R}^2$.
- The set $\mathcal{C}' = \{(x_1, x_2, x_3) \in \mathbb{R}^3 |\ x_1^2 + x_2^2 \leq x_3^2\}$ is a convex cone in $\mathbb{R}^3$.





∗

Definition 1.22 *A convex cone is called* pointed *if it does not contain any subspace except the origin.*

A pointed convex cone could be defined equivalently as a convex cone that does not contain any line.

Lemma 1.23 *A convex cone $\mathcal{C}$ is pointed if and only if the origin 0 is an extremal point of $\mathcal{C}$.*

***Proof:** If the convex cone $\mathcal{C}$ is not pointed, then it contains a nontrivial subspace, in particular, it contains a one dimensional subspace, i.e. a line $\mathcal{L}$ going through the origin. Let $0 \neq x \in \mathcal{L}$, and then we have $-x \in \mathcal{L}$ as well. From here we have $0 = \frac{1}{2}x + \frac{1}{2}(-x) \in \mathcal{C}$, i.e. 0 is not an extremal point.

If the convex cone $\mathcal{C}$ is pointed, then it does not contain any subspace, except the origin 0. In that case we show that 0 is an extremal point of $\mathcal{C}$. If we assume to the contrary that there exists $0 \neq x^1, x^2 \in \mathcal{C}$ and a $\lambda \in (0, 1)$ such that $0 = \lambda x^1 + (1 - \lambda)x^2$, then we derive $x^1 = -\frac{1-\lambda}{\lambda}x^2$. This implies that the line through x^1, the origin 0 and x^2 is in $\mathcal{C}$, contradicting the assumption that $\mathcal{C}$ is pointed. □

Example 1.24 If a convex cone $\mathcal{C} \in \mathbb{R}^2$ is not pointed, then it is either

- a line through the origin,
- a halfspace, or
- $\mathbb{R}^2$.

∗

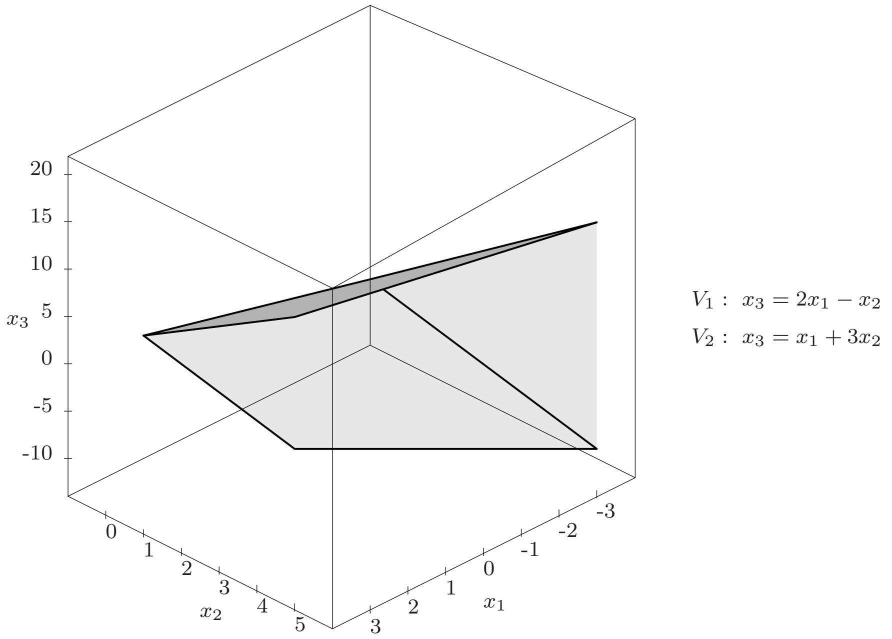
Example 1.25 Let V_1, V_2 be two planes through the origin in $\mathbb{R}^3$, given by the following equations,

$$\begin{aligned} V_1 : &= \{x \in \mathbb{R}^3 \,|\, x_3 = a_1x_1 + a_2x_2\,\}, \\ V_2 : &= \{x \in \mathbb{R}^3 \,|\, x_3 = b_1x_1 + b_2x_2\,\}, \end{aligned}$$

then the convex set

$$\mathcal{C} = \{x \in \mathbb{R}^3 |\ x_3 \geq a_1x_1 + a_2x_2,\ x_3 \leq b_1x_1 + b_2x_2\}$$

is a non-pointed cone.



$*$

Every convex cone $\mathcal{C}$ has an associated *dual cone*. By definition, every vector in the dual cone has a nonnegative inner product with every vector in $\mathcal{C}$.

Definition 1.26 *Let $\mathcal{C} \subseteq \mathbb{R}^n$ be a convex cone. The* dual cone $\mathcal{C}^*$ *is defined by*

$$\mathcal{C}^* := \{\, z \in \mathbb{R}^n \,|\, x^T z \geq 0 \;\; \textit{for all} \;\; x \in \mathcal{C}\}.$$

In the literature the dual cone $\mathcal{C}^*$ is frequently referred to as the *polar* or *positive polar* of the cone $\mathcal{C}$.

Exercise 1.12 *Prove that $(\mathbb{R}^n_+)^* = \mathbb{R}^n_+$, i.e. the nonnegative orthant is a self-dual cone.* ◁

Exercise 1.13 *Let $\mathcal{S}_n$ denote the set of $n \times n$, symmetric positive semidefinite matrices.*
(i) *Prove that $\mathcal{S}_n$ is a convex cone.*
(ii) *Prove that $(\mathcal{S}_n)^* = \mathcal{S}_n$, i.e. $\mathcal{S}_n$ is a self-dual cone.* ◁

Exercise 1.14 *Prove that the dual cone $\mathcal{C}^*$ is a closed convex cone.* ◁

An important, deep theorem [38, 42] says that the dual of the dual cone $(\mathcal{C}^*)^*$ is the closure $\overline{\mathcal{C}}$ of the cone $\mathcal{C}$.

An important cone in the study of convex optimization is the so-called *recession cone*. Given an (unbounded) convex feasible set $\mathcal{F}$ and some $x \in \mathcal{F}$, the recession cone of $\mathcal{F}$ consists of all the directions one can travel in without ever leaving $\mathcal{F}$, when starting from x. Surprisingly, the recession cone does not depend on the choice of x.

Lemma 1.27 *Let us assume that the convex set $\mathcal{C}$ is closed and not bounded. Then*

(*i*) *for each $x \in \mathcal{C}$ there is a non-zero vector $z \in \mathbb{R}^n$ such that $x + \lambda z \in \mathcal{C}$ for all $\lambda \geq 0$, i.e. the set $R(x) = \{z \,|\, x + \lambda z \in \mathcal{C},\; \lambda \geq 0\}$ is not empty;*
(*ii*) *the set $R(x)$ is a closed convex cone (the so-called* recession cone *at x);*
(*iii*) *the cone $R(x) = \mathcal{R}$ is independent of x, thus it is* 'the' *recession cone of the convex set $\mathcal{C}$;*[1]

[1] In the literature the recession cone is frequently referred to as the *characteristic cone* of the convex set $\mathcal{C}$.

(iv) *$\mathcal{R}$ is a pointed cone if and only if $\mathcal{C}$ has at least one extremal point.*

***Proof:** (i) Let $x \in \mathcal{C}$ be given. Because $\mathcal{C}$ is unbounded, then there is a sequence of points $x^1, \cdots, x^k, \cdots$ such that $\|x^k - x\| \to \infty$. Then the vectors

$$y^k = \frac{x^k - x}{\|x^k - x\|}$$

are in the unit sphere. The unit sphere is a closed convex, i.e. compact set, hence there exists an accumulation point $\bar{y}$ of the sequence y^k. We claim that $\bar{y} \in R(x)$. To prove this we take any $\lambda > 0$ and prove that $x + \lambda\bar{y} \in \mathcal{C}$. This claim follows from the following three observations: *1.* If we omit all the points from the sequence y^k for which $\|x - x^k\| < \lambda$ then $\bar{y}$ is still an accumulation point of the remaining sequence y^{k_i}. *2.* Due to the convexity of $\mathcal{C}$ the points

$$x + \lambda y^{k_i} = x + \frac{\lambda}{\|x^{k_i} - x\|}(x^{k_i} - x) = \left(1 - \frac{\lambda}{\|x^{k_i} - x\|}\right)x + \frac{\lambda}{\|x^{k_i} - x\|}x^{k_i}$$

are in $\mathcal{C}$. *3.* Because $\mathcal{C}$ is closed, the accumulation point $x + \lambda\bar{y}$ of the sequence $x + \lambda y^{k_i} \in \mathcal{C}$ is also in $\mathcal{C}$. The proof of the first statement is complete.

(ii)The set $R(x)$ is a cone, because $z \in R(x)$ imply $\mu z \in R(x)$. The convexity of $R(x)$ easily follows from the convexity of $\mathcal{C}$. Finally, if $z^i \in R(x))$ for all $i = 1, 2, \cdots$ and $\bar{z} = \lim_{i\to\infty} z^i$, then for each $\lambda \geq 0$ the closedness of $\mathcal{C}$ and $x + \lambda z^i \in \mathcal{C}$ imply that

$$\lim_{i\to\infty}(x + \lambda z^i) = x + \lambda\bar{z} \in \mathcal{C},$$

hence $\bar{z} \in R(x)$ proving that $R(x)$ is closed.

(iii) Let $x^1, x^2 \in \mathcal{C}$. We have to show that $z \in R(x^1)$ imply $z \in R(x^2)$. Let us assume to the contrary that $z \notin R(x^2)$, i.e. there is an $\alpha > 0$ such that $x^2 + \alpha z \notin \mathcal{C}$. Due to $z \in R(x^1)$ we have $x^1 + (\lambda + \alpha)z \in \mathcal{C}$ for all $\alpha, \lambda \geq 0$. Using the convexity of $\mathcal{C}$ we have that the point

$$x^2_\lambda = x^2 + \frac{\alpha}{\lambda + \alpha}\left(x^1 - x^2 + (\lambda + \alpha)z\right) = x^2\left(1 - \frac{\alpha}{\lambda + \alpha}\right) + \frac{\alpha}{\lambda + \alpha}\left(x^1 + (\lambda + \alpha)z\right)$$

is in $\mathcal{C}$. Further the limit point

$$\lim_{\lambda\to\infty} x^2_\lambda \; = \; x^2 + \alpha z,$$

due to the closedness of $\mathcal{C}$, is also in $\mathcal{C}$.

(iv) We leave the proof of this part as an exercise. □

Exercise 1.15 *Prove part (iv) of Lemma 1.27.* ◁

Corollary 1.28 *The nonempty closed convex set $\mathcal{C}$ is bounded if and only if its recession cone $\mathcal{R}$ consists of the zero vector alone.*

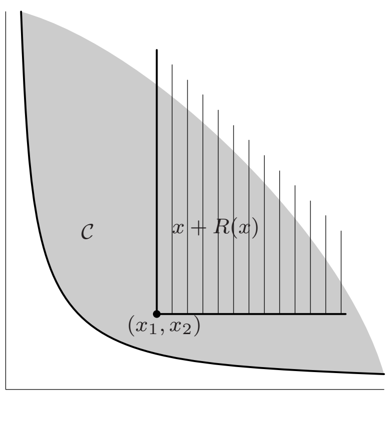
***Proof:** If $\mathcal{C}$ is bounded, then it contains no half line, thus for each $x \in \mathcal{C}$ the set $R(x) = \{0\}$, i.e. $\mathcal{R} = \{0\}$.

The other part of the proof follows form item (i) of Lemma 1.27. □

Example 1.29 Let $\mathcal{C}$ be the epigraph of $f(x) = \frac{1}{x}$. Then every point on the curve $x_2 = \frac{1}{x_1}$ is an extreme point of $\mathcal{C}$. For an arbitrary point $x = (x_1, x_2)$ the recession cone is given by

$$R(x) = \{z \in \mathbb{R}^2 \mid z_1, z_2 \geq 0\}.$$

Hence, $R = R(x)$ is independent of x, and R is a pointed cone of $\mathcal{C}$.



*

Lemma 1.30 *If the convex set $\mathcal{C}$ is closed and has an extremal point, then each extremal set of $\mathcal{C}$ has at least one extremal point as well.*

Proof: Let us assume to the contrary that an extremal set $\mathcal{E} \subset \mathcal{C}$ has no extremal point. Then by item (*iv*) of Lemma 1.27 the recession cone of $\mathcal{E}$ is not pointed, i.e. it contains a line. By statement (*iii*) of the same lemma, this line is contained in the recession cone of $\mathcal{C}$ as well. Applying (*iv*) of Lemma 1.27 again we conclude that $\mathcal{C}$ cannot have an extremal point. This is a contradiction, the lemma is proved. $\square$

Lemma 1.31 *Let $\mathcal{C}$ be a convex set and $\mathcal{R}$ be its recession cone. If $\mathcal{E}$ is an extremal set of $\mathcal{C}$ the recession cone $\mathcal{R}_{\mathcal{E}}$ of $\mathcal{E}$ is an extremal set of $\mathcal{R}$.*

Proof: Clearly $\mathcal{R}_{\mathcal{E}} \subseteq \mathcal{R}$. Let us assume that $\mathcal{R}_{\mathcal{E}}$ is not an extremal set of $\mathcal{R}$. Then there are vectors $z^1, z^2 \in \mathcal{R}$, $z^1 \notin \mathcal{R}_{\mathcal{E}}$ and a $\lambda \in (0,1)$ such that $z = \lambda z^1 + (1-\lambda)z^2 \in \mathcal{R}_{\mathcal{E}}$. Finally, for a certain $\alpha > 0$ and $x \in \mathcal{E}$ we have

$$x^1 = x + \alpha z^1 \in \mathcal{C} \setminus \mathcal{E}, \qquad x^2 = x + \alpha z^2 \in \mathcal{C}$$

and

$$\lambda x^1 + (1-\lambda)x^2 = x + \alpha z \in \mathcal{E}$$

contradicting the extremality of $\mathcal{E}$. $\square$

1.2.3 The relative interior of a convex set

The standard simplex in $\mathbb{R}^3$ was defined as the set

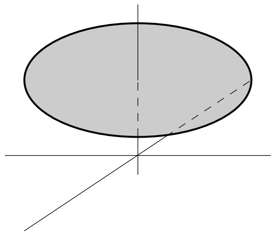
$$\left\{ x \in \mathbb{R}^3 \;\middle|\; \sum_{i=1}^{3} x_i = 1,\ x \geq 0 \right\}.$$

The interior of this convex set is empty, but intuitively the points that do not lie on the 'boundary' of the simplex do form a 'sort of interior'. This leads us to a generalized concept of the interior of a convex set, namely the *relative interior*. If the convex set is full-dimension (i.e. $\mathcal{C} \in \mathbb{R}^n$ has dimension n), then the concepts of interior and relative interior coincide.

Definition 1.32 *Let a convex set $\mathcal{C}$ be given. The point $x \in \mathcal{C}$ is in the* relative interior *of $\mathcal{C}$ if for all $\overline{x} \in \mathcal{C}$ there exists $\tilde{x} \in \mathcal{C}$ and $0 < \lambda < 1$ such that $x = \lambda\overline{x} + (1-\lambda)\tilde{x}$. The set of relative interior points of the set $\mathcal{C}$ will be denoted by $\mathcal{C}^0$.*

The relative interior $\mathcal{C}^0$ of a convex set $\mathcal{C}$ is obviously a subset of the convex set. We will show that the relative interior $\mathcal{C}^0$ is a relatively open (i.e. it coincides with its relative interior) convex set.

Example 1.33 Let $\mathcal{C} = \{x \in \mathbb{R}^3 |\ x_1^2 + x_2^2 \leq 1, x_3 = 1\}$ and $\mathcal{L} = \{x \in \mathbb{R}^3 |\ x_3 = 0\}$, then $\mathcal{C} \subset \text{aff}(\mathcal{C}) = (0,0,1) + \mathcal{L}$. Hence, $\dim \mathcal{C} = 2$ and $\mathcal{C}^0 = \{x \in \mathbb{R}^3 |\ x_1^2 + x_2^2 < 1, x_3 = 1\}$.



*

Lemma 1.34 *Let $\mathcal{C} \subset \mathbb{R}^n$ be a convex set. Then for each $x \in \mathcal{C}^0$, $y \in \mathcal{C}$ and $0 < \lambda \leq 1$ we have*

$$z = \lambda x + (1-\lambda)y \in \mathcal{C}^0 \subseteq \mathcal{C}.$$

***Proof:** Let $u \in \mathcal{C}$ be an arbitrary point. Then we have to show that for each $u \in \mathcal{C}$ there is an $\bar{u} \in \mathcal{C}$ and a $0 < \rho < 1$ such that $z = \rho \bar{u} + (1-\rho)u$. The proof is constructive.

Because $x \in \mathcal{C}^0$, by Definition 1.32 there is an $0 < \alpha < 1$ such that the point

$$v := \frac{1}{\alpha}x + (1 - \frac{1}{\alpha})u$$

is in $\mathcal{C}$. Let

$$\bar{u} = \vartheta v + (1-\vartheta)y, \qquad \text{where} \qquad \vartheta = \frac{\lambda\alpha}{\lambda\alpha + 1 - \lambda}.$$

Due to the convexity of $\mathcal{C}$ we have $\bar{u} \in \mathcal{C}$. Finally, let us define $\rho = \lambda\alpha + 1 - \lambda$. Then one can easily verify that $0 < \rho < 1$ and

$$z = \lambda x + (1-\lambda)y = \rho\bar{u} + (1-\rho)u.$$

Figure 1.4 illustrates the above construction. □

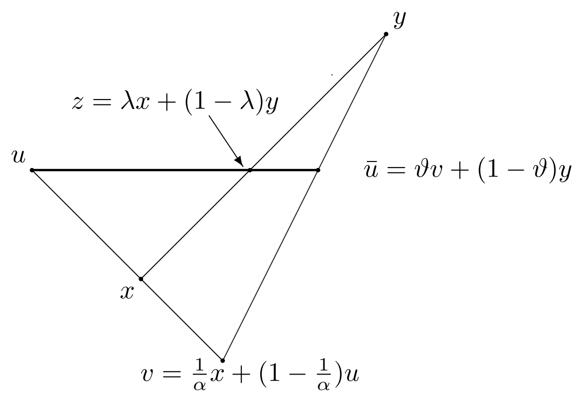


Figure 1.4: If $x \in \mathcal{C}^0$ and $y \in \mathcal{C}$, then the point z is in $\mathcal{C}^0$.

A direct consequence of the above lemma is the following.

Corollary 1.35 *The relative interior $\mathcal{C}^0$ of a convex set $\mathcal{C} \subset \mathbb{R}^n$ is convex.*

Lemma 1.36 *Let $\mathcal{C}$ be a convex set. Then $(\mathcal{C}^0)^0 = \mathcal{C}^0$. Moreover, if $\mathcal{C}$ is nonempty then its relative interior $\mathcal{C}^0$ is nonempty as well.*

***Proof:** The proof of this lemma is quite technical. For a proof the reader is referred to the excellent books of Rockafellar [38] and Stoer and Witzgall [42]. □

1.3 More on convex functions

Now we turn our attention to convex functions.

1.3.1 Basic properties of convex functions

In this section we have collected some useful facts about convex functions.

Lemma 1.37 *Let f be a convex function defined on the convex set $\mathcal{C}$. Then f is continuous on the relative interior $\mathcal{C}^0$ of $\mathcal{C}$.*

***Proof:** Let $p \in \mathcal{C}^0$ be an arbitrary point. Without loss of generality we may assume that $\mathcal{C}$ is full dimensional, p is the origin and $f(p) = 0$.

Let us first consider the one dimensional case. Because 0 is in the interior of the domain $\mathcal{C}$ of f we have a $v > 0$ such that $v \in \mathcal{C}$ and $-v \in \mathcal{C}$ as well. Let us consider the two linear functions

$$\ell_1(x) := x\frac{f(v)}{v} \qquad \text{and} \qquad \ell_2(x) := x\frac{f(-v)}{-v}.$$

One easily checks that the convexity of f implies the following relations:

- $\ell_1(x) \geq f(x)$ if $x \in [0, v]$;
- $\ell_1(x) \leq f(x)$ if $x \in [-v, 0]$;
- $\ell_2(x) \geq f(x)$ if $x \in [-v, 0]$;
- $\ell_2(x) \leq f(x)$ if $x \in [0, v]$.

Then by defining

$$h(x) := \max\{\ell_1(x), \ell_2(x)\} \qquad \text{and} \qquad g(x) := \min\{\ell_1(x), \ell_2(x)\}$$

on the interval $[-v, v]$ we have

$$g(x) \leq f(x) \leq h(x).$$

The linear functions $\ell_1(x)$ and $\ell_2(x)$ are obviously continuous, thus the functions $h(x)$ and $g(x)$ are continuous as well. By observing the relations $f(0) = h(0) = g(0) = 0$ it follows that the function $f(x)$ is continuous at the point 0.

We use an analogous construction for the n-dimensional case. Let us assume again that 0 is an interior point of $\mathcal{C}$ and $f(0) = 0$. Let $v^1, \cdots, v^n, v^{n+1}$ be vectors such that the convex set

$$\left\{ x \mid x = \sum_{i=1}^{n+1} \lambda_i v^i, \;\; \lambda_i \in [0,1], \;\; \sum_{i=1}^{n+1} \lambda_i = 1 \right\}$$

equals the space $\mathbb{R}^n$. For all $i = 1, \cdots, n+1$ let the linear functions (hyperplanes) $L_i(x)$ be defined by $n+1$ of their values: $L_i(0) = 0$ and $L_i(v^j) = f(v^j)$ for all $j \neq i$. Let us further define

$$h(x) := \max\{L_1(x), \cdots, L_{n+1}(x)\} \qquad \text{and} \qquad g(x) := \min\{L_1(x), \cdots, L_{n+1}(x)\}.$$

Then one easily proves that the functions $g(x)$ and $h(x)$ are continuous, $f(0) = h(0) = g(0) = 0$ and

$$g(x) \leq f(x) \leq h(x).$$

Thus the function $f(x)$ is continuous at the point 0. □

Exercise 1.16 *Prove that the functions $g(x)$ and $h(x)$, defined in the proof above, are continuous, $f(0) = h(0) = g(0) = 0$ and*

$$g(x) \leq f(x) \leq h(x).$$

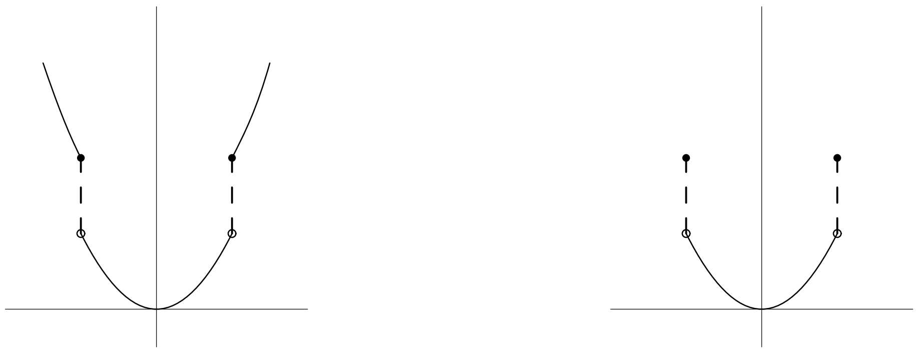
◁

Note that f can be discontinuous on the relative boundary $\mathcal{C} \setminus \mathcal{C}^0$.

Example 1.38 The function

$$f(x) = \begin{cases} x^2 & \text{for } -1 < x < 1 \\ x^2 + 1 & \text{otherwise} \end{cases}$$

is not continuous on $\mathbb{R}$, and it is also not convex. If f is only defined on $\mathcal{C} = \{x \in \mathbb{R} \mid -1 \leq x \leq 1\}$ then f is still not continuous, but it is continuous on $\mathcal{C}^0$ and convex on $\mathcal{C}$.



*

The following result, called Jensen's inequality, is simply a generalization of the inequality $f(\lambda x^1 + (1-\lambda)x^2) \le \lambda f(x^1) + (1-\lambda)f(x^2)$.

Lemma 1.39 (Jensen inequality) *Let f be a convex function defined on a convex set $\mathcal{C} \subseteq \mathbb{R}^n$. Let the points $x^1, \cdots, x^k \in \mathcal{C}$ and $\lambda_1, \cdots, \lambda_k \ge 0$ with $\sum_{i=1}^k \lambda_i = 1$ be given. Then*

$$f\left(\sum_{i=1}^k \lambda_i x^i\right) \le \sum_{i=1}^k \lambda_i f(x^i).$$

***Proof:** The proof is by induction on k. If $k = 2$ then the statement is true by Definition 1.4. Let us assume that the statement holds for a given $k \ge 2$, then we prove that it also holds for $k+1$.

Let the points $x^1, \cdots, x^k, x^{k+1} \in \mathcal{C}$ and $\lambda_1, \cdots, \lambda_k, \lambda_{k+1} \ge 0$ with $\sum_{i=1}^{k+1} \lambda_i = 1$ be given. If at most k of the λ_i, $1 \le i \le k+1$ coefficients are nonzero then, by leaving out the points x^i with zero coefficients, the inequality directly follows from the inductive assumption. Now let us consider the case when all the coefficients λ_i are nonzero. Then by convexity of the set $\mathcal{C}$ we have that

$$\tilde{x} = \sum_{i=1}^k \frac{\lambda_i}{\sum_{j=1}^k \lambda_j} x^i \in \mathcal{C}.$$

Further

$$\begin{aligned}
f\left(\sum_{i=1}^{k+1} \lambda_i x^i\right) &= f\left(\sum_{j=1}^k \lambda_j \sum_{i=1}^k \frac{\lambda_i}{\sum_{j=1}^k \lambda_j} x^i + \lambda_{k+1} x^{k+1}\right) \\
&= f\left(\left[\sum_{j=1}^k \lambda_j\right] \tilde{x} + \lambda_{k+1} x^{k+1}\right) \\
&\le \left[\sum_{j=1}^k \lambda_j\right] f(\tilde{x}) + \lambda_{k+1} f\left(x^{k+1}\right) \\
&\le \left[\sum_{j=1}^k \lambda_j\right] \left(\sum_{i=1}^k \frac{\lambda_i}{\sum_{j=1}^k \lambda_j} f(x^i)\right) + \lambda_{k+1} f(x^{k+1}) \\
&= \sum_{i=1}^{k+1} \lambda_i f(x^i),
\end{aligned}$$

where the first inequality follows from the convexity of the function f (Definition 1.4) and, at the second inequality, the inductive assumption was used. The proof is complete. □

The reader can easily prove the following two lemmas by applying the definitions. We leave the proofs as exercises.

Lemma 1.40 *Let $f^1, \cdots, f^k$ be convex functions defined on a convex set $\mathcal{C} \subseteq \mathbb{R}^n$. Then*

- *for all* $\lambda_1, \cdots, \lambda_k \geq 0$ *the function*

$$f(x) = \sum_{i=1}^{k} \lambda_i f^i(x)$$

  is convex;
- *the function*

$$f(x) = \max_{1 \leq i \leq k} f^i(x)$$

  is convex.

Definition 1.41 *The function* $h : \mathbb{R} \to \mathbb{R}$ *is called*

- *monotonically non-decreasing if for all* $t_1 < t_2 \in \mathbb{R}$ *one has* $h(t_1) \leq h(t_2)$*;*
- *strictly monotonically increasing if for all* $t_1 < t_2 \in \mathbb{R}$ *one has* $h(t_1) < h(t_2)$*.*

Lemma 1.42 *Let* f *be a convex function on the convex set* $\mathcal{C} \subseteq \mathbb{R}^n$ *and* $h : \mathbb{R} \to \mathbb{R}$ *be a convex monotonically non-decreasing function. Then the composite function* $h(f(x)) : \mathcal{C} \to \mathbb{R}$ *is convex.*

Exercise 1.17 *Prove Lemma 1.40 and Lemma 1.42.* ◁

Exercise 1.18 *Assume that the function* h *in Lemma 1.42 is* **not** *monotonically non-decreasing. Give a concrete example that in this case the statement of the lemma fails.* ◁

Definition 1.43 *Let a convex function* $f : \mathcal{C} \to \mathbb{R}$ *defined on the convex set* $\mathcal{C}$ *be given. Let* $\alpha \in \mathbb{R}$ *be an arbitrary number. The set* $\mathcal{D}_\alpha = \{x \in \mathcal{C} \mid f(x) \leq \alpha\}$ *is called a* level set *of the function* f*.*

Lemma 1.44 *If* f *is a convex function on the convex set* $\mathcal{C}$ *then for all* $\alpha \in \mathbb{R}$ *the level set* $\mathcal{D}_\alpha$ *is a (possibly empty) convex set.*

***Proof:** Let $x, y \in \mathcal{D}_\alpha$ and $0 \leq \lambda \leq 1$. Then we have $f(x) \leq \alpha$, $f(y) \leq \alpha$ and we may write

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

Here the first inequality followed from the convexity of the function f. The lemma is proved. □

1.3.2 On the derivatives of convex functions

The first and second order derivatives of (sufficiently differentiable) convex functions have some interesting and useful properties which we review in this section. The reader may recall from elementary calculus that a univariate function f on $\mathbb{R}$ is convex if $f''(x) \geq 0$ for all $x \in \mathbb{R}$ (assuming sufficient differentiability), and that such a function attains its minimum at some $\overline{x} \in \mathbb{R}$ if and only if $f'(\overline{x}) = 0$. We will work towards generalizing these results to multivariate convex functions.

Recall that the gradient ∇f of the function f is defined as the vector formed by the partial derivatives $\frac{\partial f}{\partial x_i}$ of f. Further we introduce the concept of directional derivative.

Definition 1.45 *Let* $x \in \mathbb{R}^n$ *and a direction (vector)* $s \in \mathbb{R}^n$ *be given. The directional derivative* $\delta f(x, s)$ *of the function* f*, at point* x*, in the direction* s*, is defined as*

$$\delta f(x, s) = \lim_{\lambda \to 0} \frac{f(x + \lambda s) - f(x)}{\lambda}$$

if the above limit exists.

If the function f is continuously differentiable then $\frac{\partial f}{\partial x_i} = \delta f(x, e^i)$ where e^i is the $i-$th unit vector. This implies the following result.

Lemma 1.46 *If the function f is continuously differentiable then for all $s \in \mathbb{R}^n$ we have*

$$\delta f(x, s) = \nabla f(x)^T s.$$

The Hesse matrix (or Hessian) $\nabla^2 f(x)$ of the function f at a point $x \in \mathcal{C}$ is composed of the second order partial derivatives of f as

$$(\nabla^2 f(x))_{ij} = \frac{\partial^2 f(x)}{\partial x_i \partial x_j} \quad \text{for all} \quad i, j = 1, \cdots, n.$$

Lemma 1.47 *Let f be a function defined on a convex set $\mathcal{C} \subseteq \mathbb{R}^n$. The function f is convex if and only if the function $\phi(\lambda) = f(x + \lambda s)$ is convex on the interval $[0, 1]$ for all $x \in \mathcal{C}$ and $x + s \in \mathcal{C}$.*

***Proof:** Let us assume that f is a convex function. Then we prove that $\phi(\lambda)$ is convex on the interval $[0, 1]$. Let $\lambda_1, \lambda_2 \in [0, 1]$ and $0 \le \alpha \le 1$. Then one has

$$\begin{aligned} \phi(\alpha\lambda_1 + (1-\alpha)\lambda_2) &= f(x + [\alpha\lambda_1 + (1-\alpha)\lambda_2]s) \\ &= f(\alpha[x + \lambda_1 s] + (1-\alpha)[x + \lambda_2 s]) \\ &\le \alpha f(x + \lambda_1 s) + (1-\alpha) f(x + \lambda_2 s) \\ &= \alpha\phi(\lambda_1) + (1-\alpha)\phi(\lambda_2), \end{aligned}$$

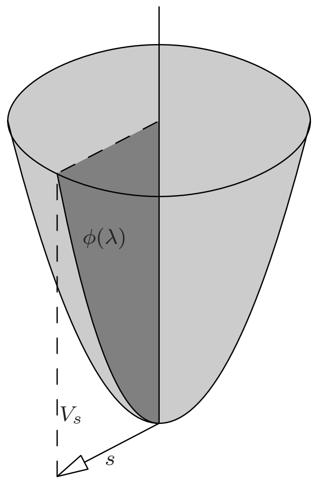
proving the first part of the lemma.

On the other hand, if $\phi(\lambda)$ is convex on the interval $[0, 1]$ for each x, $x + s \in \mathcal{C}$ then the convexity of the function f can be proved as follows. For given $y, x \in \mathcal{C}$ let us define $s := y - x$. Then we write

$$\begin{aligned} f(\alpha y + (1-\alpha)x) &= f(x + \alpha(y - x)) = \phi(\alpha) = \phi(\alpha 1 + (1-\alpha)0) \\ &\le \alpha\phi(1) + (1-\alpha)\phi(0) = \alpha f(y) + (1-\alpha) f(x). \end{aligned}$$

The proof of the lemma is complete. □

Example 1.48 Let $f(x) = x_1^2 + x_2^2$ and let E_f be the epigraph of f. For every $s \in \mathbb{R}^2$, we can define the half-plane $V_s \subset \mathbb{R}^3$ as $\{(x, y) \in \mathbb{R}^2 \times \mathbb{R} |\ x = \mu s, \mu > 0\}$. Now, for $x = (0, 0)$ the epigraph of $\phi(\lambda) = f(x + \lambda s) = f(\lambda s)$ equals $V_s \cap E_f$, which is a convex set. Hence, $\phi(\lambda)$ is convex.



*

Lemma 1.49 *Let f be a continuously differentiable function on the open convex set $\mathcal{C} \subseteq \mathbb{R}^n$. Then the following statements are equivalent.*

1. *The function f is convex on $\mathcal{C}$.*
2. *For any two vectors $x, \overline{x} \in \mathcal{C}$ one has*

$$\nabla f(x)^T(\overline{x} - x) \le f(\overline{x}) - f(x) \le \nabla f(\overline{x})^T(\overline{x} - x).$$

3. *For any $x \in \mathcal{C}$, and any $s \in \mathbb{R}^n$ such that $x+s \in \mathcal{C}$, the function $\phi(\lambda) = f(x+\lambda s)$ is continuously differentiable on the open interval $(0,1)$ and $\phi'(\lambda) = s^T \nabla f(x+\lambda s)$, which is a monotonically non-decreasing function.*

***Proof:** First we prove that *1* implies *2*. Let $0 \le \lambda \le 1$ and $x, \overline{x} \in \mathcal{C}$. Then the convexity of f implies

$$f(\lambda \overline{x} + (1-\lambda)x) \le \lambda f(\overline{x}) + (1-\lambda) f(x).$$

This can be rewritten as

$$\frac{f(x + \lambda(\overline{x} - x)) - f(x)}{\lambda} \le f(\overline{x}) - f(x).$$

Taking the limit as $\lambda \to 0$ and applying Lemma 1.46 the left-hand-side inequality of *2* follows. As one interchanges the role x and $\overline{x}$, the right-hand-side inequality is obtained analogously.

Now we prove that *2* implies *3*. Let $x, x+s \in \mathcal{C}$ and $0 \le \lambda_1, \lambda_2 \le 1$. When we apply the inequalities of *2* with the points $x + \lambda_1 s$ and $x + \lambda_2 s$ the following relations are obtained.

$$(\lambda_2 - \lambda_1) \nabla f(x + \lambda_1 s)^T s \le f(x + \lambda_2 s) - f(x + \lambda_1 s) \le (\lambda_2 - \lambda_1) \nabla f(x + \lambda_2 s)^T s,$$

hence

$$(\lambda_2 - \lambda_1)\phi'(\lambda_1) \le \phi(\lambda_2) - \phi(\lambda_1) \le (\lambda_2 - \lambda_1)\phi'(\lambda_2).$$

Assuming $\lambda_1 < \lambda_2$ we have

$$\phi'(\lambda_1) \le \frac{\phi(\lambda_2) - \phi(\lambda_1)}{\lambda_2 - \lambda_1} \le \phi'(\lambda_2),$$

proving that the function $\phi'(\lambda)$ is monotonically non-decreasing.

Finally we prove that *3* implies *1*. We only have to prove that $\phi(\lambda)$ is convex if $\phi'(\lambda)$ is monotonically non-decreasing. Let us take $0 < \lambda_1 < \lambda_2 < 1$ where $\phi(\lambda_1) \le \phi(\lambda_2)$. Then for $0 \le \alpha \le 1$ we may write

$$\begin{aligned} (1-\alpha)\phi(\lambda_1) \quad &+ \quad \alpha\phi(\lambda_2) - \phi((1-\alpha)\lambda_1 + \alpha\lambda_2) \\ &= \quad \alpha[\phi(\lambda_2) - \phi(\lambda_1)] - [\phi((1-\alpha)\lambda_1 + \alpha\lambda_2) - \phi(\lambda_1)] \\ &= \quad \alpha(\lambda_2 - \lambda_1)\left(\int_0^1 \phi'(\lambda_1 + t(\lambda_2 - \lambda_1))\mathrm{d}t - \int_0^1 \phi'(\lambda_1 + t\alpha(\lambda_2 - \lambda_1))\mathrm{d}t\right) \\ &\ge \quad 0. \end{aligned}$$

The expression for the derivative of ϕ is left as an exercise in calculus (Exercise 1.19). The proof of the Lemma is complete. □

Figure 1.5 illustrates the inequalities at statement *2* of the lemma.

Exercise 1.19 *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be twice continuously differentiable and let $x \in \mathbb{R}^n$ and $s \in \mathbb{R}^n$ be given. Define $\phi : \mathbb{R} \mapsto \mathbb{R}$ via $\phi(\lambda) = f(x + \lambda s)$. Prove that*

$$\phi'(\lambda) = s^T \nabla f(x + \lambda s)$$

and

$$\phi''(\lambda) = s^T \nabla^2 f(x + \lambda s) s.$$

◁

Lemma 1.50 *Let f be a twice continuously differentiable function on the open convex set $\mathcal{C} \subseteq \mathbb{R}^n$. The function f is convex if and only if its Hesse matrix $\nabla^2 f(x)$ is PSD for all $x \in \mathcal{C}$.*

Proof: Let us take an arbitrary $x \in \mathcal{C}$ and $s \in \mathbb{R}^n$, and define $\phi(\lambda) = f(x + \lambda s)$. If f is convex, then $\phi'(\lambda)$ is monotonically non-decreasing. This implies that $\phi''(\lambda)$ is nonnegative for each $x \in \mathcal{C}$ and $0 \le \lambda \le 1$. Thus

$$s^T \nabla^2 f(x) s = \phi''(0) \ge 0$$

proving the positive semidefiniteness of the Hessian $\nabla^2 f(x)$.

On the other hand, if the Hessian $\nabla^2 f(x)$ is positive semidefinite for each $x \in \mathcal{C}$, then

$$s^T \nabla^2 f(x + \lambda s) s = \phi''(\lambda) \ge 0,$$

i.e. $\phi'(\lambda)$ is monotonically non-decreasing proving the convexity of f by Lemma 1.49. The theorem is proved.

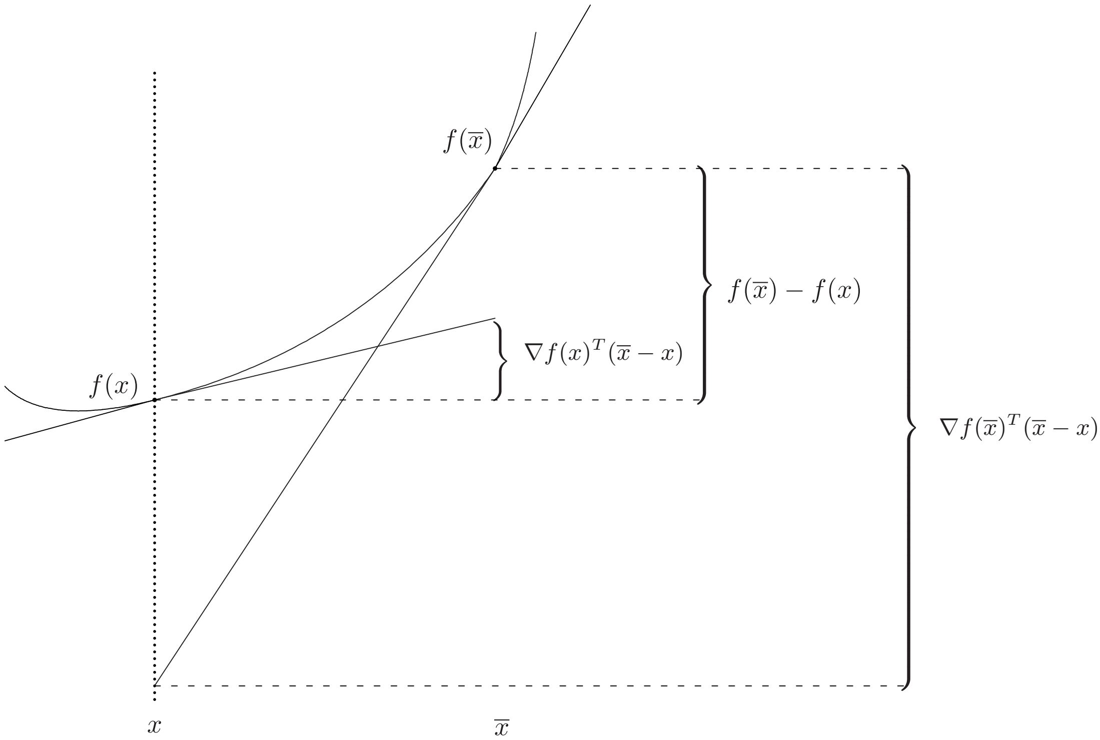


Figure 1.5: Inequalities derived for the gradient of a convex function f.

Exercise 1.20 Prove the following statement analogously as Lemma 1.50 was proved.

Let f be a twice continuously differentiable function on the open convex set $\mathcal{C}$. Then f is strictly convex if its Hesse matrix $\nabla^2 f(x)$ is positive definite (PD) for all $x \in \mathcal{C}$. ◁

Exercise 1.21 *Give an example of a twice continuously differentiable strictly convex function f where $\nabla^2 f(x)$ is* not *positive definite (PD) for all x in the domain of f.* ◁

Chapter 2

Optimality conditions

We consider two cases separately. First optimality conditions for unconstrained optimization are considered. Then optimality conditions for some special constrained optimization problems are derived.

2.1 Optimality conditions for unconstrained optimization

Consider the problem

$$\text{minimize} \quad f(x), \tag{2.1}$$

where $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function. First we define local and global minima of the above problem.

Definition 2.1 *Let a function $f : \mathbb{R}^n \to \mathbb{R}$ be given.*

A point $\overline{x} \in \mathbb{R}^n$ is a local minimum *of the function f if there is an $\epsilon > 0$ such that $f(\overline{x}) \leq f(x)$ for all $x \in \mathbb{R}^n$ when $\|\overline{x} - x\| \leq \epsilon$.*

A point $\overline{x} \in \mathbb{R}^n$ is a strict local minimum *of the function f if there is an $\epsilon > 0$ such that $f(\overline{x}) < f(x)$ for all $x \in \mathbb{R}^n$ when $\|\overline{x} - x\| \leq \epsilon$.*

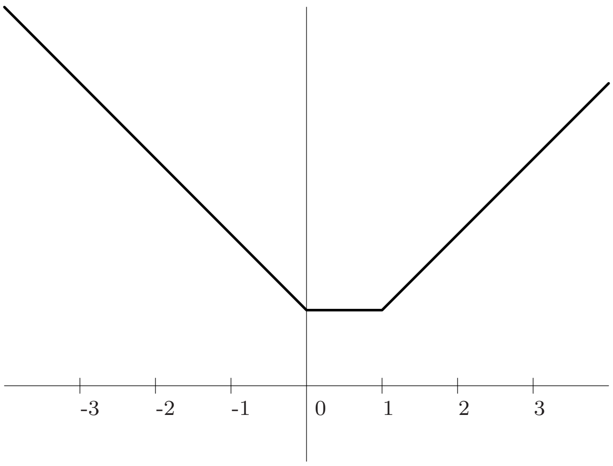
A point $\overline{x} \in \mathbb{R}^n$ is a global minimum *of the function f if $f(\overline{x}) \leq f(x)$ for all $x \in \mathbb{R}^n$.*

A point $\overline{x} \in \mathbb{R}^n$ is a strict global minimum *of the function f if $f(\overline{x}) < f(x)$ for all $x \in \mathbb{R}^n$.*

Convex functions possess the appealing property that a local minimum is global.

Example 2.2 Consider the convex function $f_1 : \mathbb{R} \to \mathbb{R}$ defined as follows.

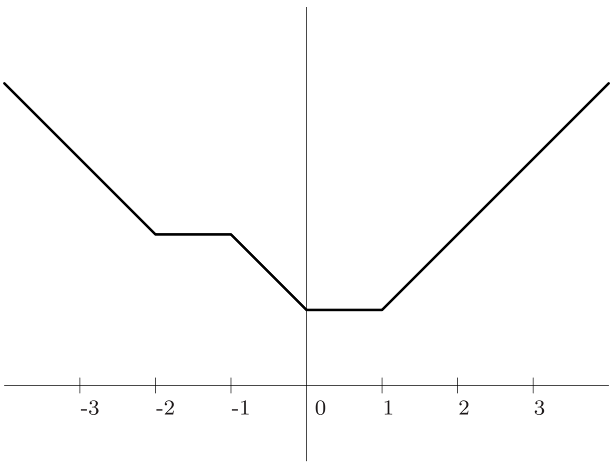
$$f_1(x) = \begin{cases} -x+1 & \text{if } x < 0, \\ 1 & \text{if } 0 \leq x \leq 1, \\ x & \text{if } x > 1. \end{cases}$$



The point $\bar{x} = 0$ is a global minimum of the function f_1 because $f_1(\bar{x}) \leq f_1(x)$ for all $x \in \mathbb{R}$. Because $\bar{x} = 0$ is a global minimum, it follows immediately that it is also a local minimum of the function f_1. The point $\bar{x} = 0$ is neither a strict local nor a strict global minimum point of f_1 because for any $\epsilon > 0$ we can find an x for which $f_1(\bar{x}) = f_1(x)$ applies with $||\bar{x} - x|| \leq \epsilon$.

Now let us consider the non-convex function $f_1 : \mathbb{R} \to \mathbb{R}$ defined as

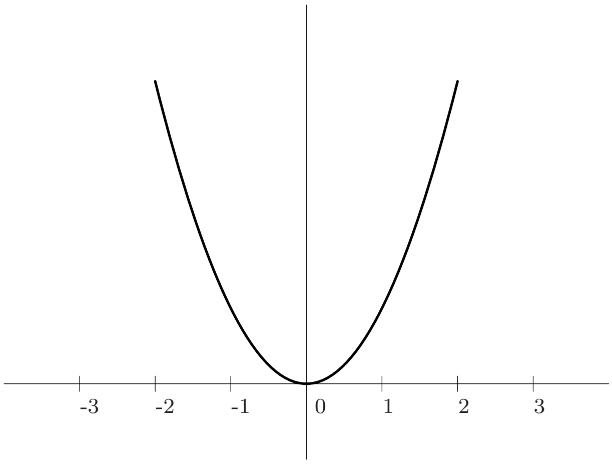
$$f_2(x) = \begin{cases} -x & \text{if } x < 2, \\ 2 & \text{if } -2 \leq x \leq -1, \\ -x+1 & \text{if } -1 < x < 0, \\ 1 & \text{if } 0 \leq x \leq 1, \\ x & \text{if } x > 1. \end{cases}$$



The point $\bar{x} = 0$ is a global minimum of the function f_2 because $f_2(\bar{x}) \leq f_2(x)$ for all $x \in \mathbb{R}$. Because it is a global minimum it is at the same a local minimum as well. The point $\bar{x} = 0$ is neither a strict local, nor a strict global minimum of the function f_2 because for any $\epsilon > 0$ we can find an x for which $f_2(\bar{x}) = f_2(x)$ applies with $||\bar{x} - x|| \leq \epsilon$.

The point $x^* = -2$ is also a local minimum of the function f_2 because $f_2(x^*) \leq f_2(x)$ for all $x \in \mathbb{R}$ when $||x^* - x|| \leq \epsilon$, with $0 < \epsilon < 1$. It is not a strict local minimum because $f_2(x^*) \not< f_2(x)$ for all $x \in \mathbb{R}$ when $||x^* - x|| < \epsilon$, with $\epsilon > 0$. The point $x^* = -2$ is not a global minimum of f_2 because $f_2(-2) > f_2(0)$. $*$

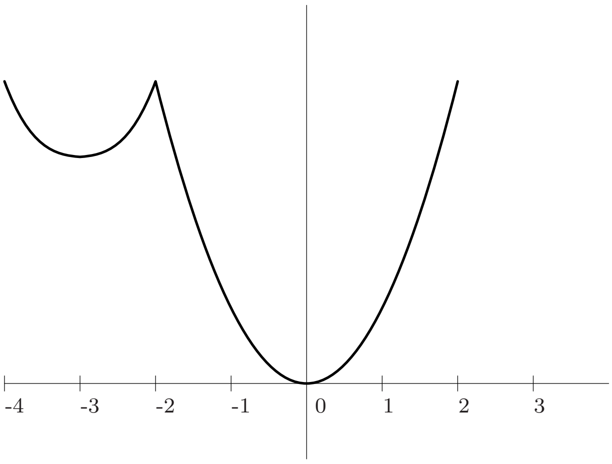
Example 2.3 Consider the convex function $f_1(x) = x^2$ where $x \in \mathbb{R}$.



The point $\bar{x} = 0$ is a strict local minimum of the function f_1 because $f_1(\bar{x}) < f_1(x)$ for all $x \in \mathbb{R}$ when $||\bar{x} - x|| < \epsilon$, with $\epsilon > 0$. The point $\bar{x} = 0$ is also a strict global minimum of the function f_1 because $f_1(\bar{x}) < f_1(x)$ for all $x \in \mathbb{R}$.

Consider the non-convex function $f_2 : \mathbb{R} \to \mathbb{R}$ defined as

$$f_2(x) = \begin{cases} (x+3)^2 + 3 & \text{if } x < -2, \\ x^2 & \text{if } x \geq -2. \end{cases}$$



The point $\bar{x} = 0$ is a strict local minimum as well as an strict global minimum for the function f_2, because $f_2(\bar{x}) < f_2(x)$ for all $x \in \mathbb{R}$ when $||\bar{x} - x|| < \epsilon$, with $\epsilon > 0$, and for all $x \in \mathbb{R}$. The point $x^* = -3$ is a strict local minimum because $f_2(x^*) < f_2(x)$ for all $x \in \mathbb{R}$ when $||x^* - x|| < \epsilon$, with $0 < \epsilon < 1$. The point $x^* = -3$ is not a strict global minimum, because $f_2(-3) > f_2(0)$.

*

Lemma 2.4 *Any (strict) local minimum of a convex function f is a (strict) global minimum of f as well.*

Exercise 2.1 *Prove Lemma 2.4.* ◁

Now we arrive at the famous result of Fermat that says that a necessary condition for $\overline{x}$ to be a minimum of a continuously differentiable function f is that $\nabla f(\overline{x}) = 0$.

Theorem 2.5 (Fermat) *Let f be continuously differentiable. If the point $\overline{x} \in \mathbb{R}^n$ is a minimum of the function f then $\nabla f(\overline{x}) = 0$.*

Proof: As $\overline{x}$ is a minimum, one has

$$f(\overline{x}) \leq f(\overline{x} + \lambda s) \text{ for all } s \in \mathbb{R}^n \text{ and } \lambda \in R.$$

By bringing $f(\overline{x})$ to the right hand side and dividing by $\lambda > 0$ we have

$$0 \leq \frac{f(\overline{x} + \lambda s) - f(\overline{x})}{\lambda}.$$

Taking the limit as $\lambda \to 0$ results in

$$0 \leq \delta f(\overline{x}, s) = \nabla f(\overline{x})^T s \quad \text{for all} \quad s \in \mathbb{R}^n.$$

As $s \in \mathbb{R}^n$ is arbitrary we conclude that $\nabla f(\overline{x}) = 0$. □

Remark: In the above theorem it is enough to assume that the partial derivatives of f exist. The same proof applies if we choose e^i, the standard unit vectors instead of the arbitrary direction s.

Exercise 2.2 *Consider Kepler's problem as formulated in Exercise 0.2.*

1. *Show that Kepler's problem can be written as the problem of minimizing a nonlinear univariate function on an open interval.*
2. *Show that the solution given by Kepler is indeed optimal by using Theorem 2.5.*

◁

Observe that the above theorem contains only a one sided implication. It does not say anything about a minimum of f if $\nabla f(\overline{x}) = 0$. Such points are not necessarily minimum points. These points are called *stationary* points. Think of the stationary (inflection) point $x = 0$ of the univariate function $f(x) = x^3$. In other words, Fermat's result only gave a *necessary* condition for a minimum, namely $\nabla f(\overline{x}) = 0$. We will now see that this is also a sufficient condition if f is convex.

Theorem 2.6 *Let f be a continuously differentiable convex function. The point $\overline{x} \in \mathbb{R}^n$ is a minimum of the function f if and only if $\nabla f(\overline{x}) = 0$.*

Proof: As $\overline{x}$ is a minimum of f then by Theorem 2.5 we have $\nabla f(\overline{x}) = 0$. On the other hand, if f is a convex function and $\nabla f(\overline{x}) = 0$ then

$$f(x) - f(\overline{x}) \geq \nabla f(\overline{x})^T (x - \overline{x}) = 0 \quad \text{for all} \quad x \in \mathbb{R}^n,$$

hence the theorem is proved. □

Exercise 2.3 *We return to Steiner's problem (see Section 0.3.3) of finding the Torricelli point of a given triangle, that was defined as the solution of the optimization problem*

$$\min_{x \in \mathbb{R}^2} \|x - a\| + \|x - b\| + \|x - c\|, \tag{2.2}$$

where a, b, and c are given vectors in $\mathbb{R}^2$ that form the vertices of the given triangle.

1. *Show that the objective function is convex.*
2. *Give necessary and sufficient conditions for a minimum of (2.2). (In other words, give the equations that determine the Torricelli point. You may assume that all three angles of the triangle are smaller than $\frac{\pi}{2}$.)*
3. *Find the Torricelli point of the triangle with vertices $(0, 0)$, $(3, 0)$ and $(1, 2)$.*

◁

If f is a twice continuously differentiable (not necessarily convex) function then second order sufficient conditions for local minima are derived as follows. Let $\nabla^2 f(x)$ denote the Hesse matrix of f at the point x.

Theorem 2.7 *Let f be a twice continuously differentiable function. If at a point $\overline{x} \in \mathbb{R}^n$ it holds that $\nabla f(\overline{x}) = 0$ and $\nabla^2 f(x)$ is positive semidefinite in an $\epsilon-$neighborhood ($\epsilon > 0$) of $\overline{x}$ then the point $\overline{x}$ is a local minimum of the function f. (If we assume positive* definiteness *then we get a* strict *local minimum.)*

Proof: Taking the Taylor series expansion of f at $\overline{x}$ we have

$$f(x) = f(\overline{x} + (x - \overline{x})) = f(\overline{x}) + \nabla f(\overline{x})^T (x - \overline{x}) + \frac{1}{2}(x - \overline{x})^T \nabla^2 f(\overline{x} + \alpha(x - \overline{x}))(x - \overline{x})$$

for some $0 \leq \alpha \leq 1$. Using the assumptions we have the result $f(x) \geq f(\overline{x})$ as x is in the neighborhood of $\overline{x}$ where the Hesse matrix is positive semidefinite. □

Corollary 2.8 *Let f be a twice continuously differentiable function. If at $\overline{x} \in \mathbb{R}^n$ the gradient $\nabla f(\overline{x}) = 0$ and the Hessian $\nabla^2 f(\overline{x})$ is positive definite then the point $\overline{x}$ is a strict local minimum of the function f.*

Proof: Since f is twice continuously differentiable, it follows from the positive definiteness of the Hesse matrix at $\overline{x}$ that it is positive definite in a neighborhood of $\overline{x}$. Hence the claim follows from theorem 2.7. □

2.2 Optimality conditions for constrained optimization

The following theorem generalizes the optimality conditions for a convex function on $\mathbb{R}^n$ (Theorem 2.6), by replacing $\mathbb{R}^n$ by any relatively open convex set $\mathcal{C} \subseteq \mathbb{R}^n$.

Theorem 2.9 *Let us consider the optimization problem* $\min\{ f(x) : x \in \mathcal{C}\}$ *where $\mathcal{C}$ is a relatively open convex set and f is a convex differentiable function. The point $\overline{x}$ is an optimal solution of this problem if and only if $\nabla f(\overline{x})^T s = 0$ for all $s \in \mathcal{L}$, where $\mathcal{L}$ denotes the linear subspace with* $\mathrm{aff}(\mathcal{C}) = x + \mathcal{L}$ *for any $x \in \mathcal{C}$. Here* $\mathrm{aff}(\mathcal{C})$ *denotes the affine hull of $\mathcal{C}$.*

Proof: Let $s \in \mathcal{L}$ and $\lambda \in \mathbb{R}$. If $\overline{x}$ is a minimum, one has

$$f(\overline{x}) \leq f(\overline{x} + \lambda s) \text{ if } \overline{x} + \lambda s \in \mathcal{C}.$$

Note that $\overline{x} + \lambda s \in \mathrm{aff}(\mathcal{C})$ since $s \in \mathcal{L}$, and $\overline{x} + \lambda s \in \mathcal{C}$ if λ is sufficiently small, since $\mathcal{C}$ is a relatively open set.

By bringing $f(\overline{x})$ to the right hand side and dividing by $\lambda > 0$ we have

$$0 \leq \frac{f(\overline{x} + \lambda s) - f(\overline{x})}{\lambda} \quad \text{for all} \quad s \in \mathcal{L},$$

if $\lambda > 0$ is sufficiently small. Taking the limit as $\lambda \downarrow 0$ results in

$$0 \leq \delta f(\overline{x}, s) = \nabla f(\overline{x})^T s \quad \text{for all} \quad s \in \mathcal{L}.$$

We conclude that $\nabla f(\overline{x})^T s = 0$ for all $s \in \mathcal{L}$.

Conversely, if f is a convex function and $\nabla f(\overline{x})^T s = 0$ for all $s \in \mathcal{L}$ then, for any $x \in \mathcal{C}$,

$$f(x) - f(\overline{x}) \geq \nabla f(\overline{x})^T (x - \overline{x}) = 0,$$

since $s = (x - \overline{x}) \in \mathcal{L}$, hence the theorem is proved. $\square$

A crucial assumption of the above lemma is that the set $\mathcal{C}$ is a relatively open set. In general this is not the case because the level sets of convex optimization problems are closed. However as we will see later the barrier function approach will result in such relatively open feasible sets. This is an important feature of interior point methods that will be discussed later on. If the set of feasible solutions is not relatively open, similar results by using similar techniques can be derived (see Theorem 2.14).

Exercise 2.4 *We return to Tartaglia's problem (4) in Section 0.3.1.*

1. *Eliminate one of the variables and show that the resulting problem can be written as the problem of minimizing a univariate convex function on an open interval.*
2. *Show that the answer given by Tartaglia is indeed the optimal solution, by applying Theorem 2.9.*

$\triangleleft$

Now let us consider the general convex optimization problem, as given earlier in (1), but without equality constraints.

$$\begin{aligned} (CO) \quad \min \quad & f(x) \\ \text{s.t.} \quad & g_j(x) \leq 0, \quad j = 1, \cdots, m \\ & x \in \mathcal{C}, \end{aligned} \tag{2.3}$$

where $\mathcal{C} \subseteq \mathbb{R}^n$ is a convex set and $f, g_1, \cdots, g_m$ are convex functions on $\mathcal{C}$ (or on an open set that contains the set $\mathcal{C}$). Almost always we will assume that the functions are differentiable. The set of feasible solutions will be denoted by $\mathcal{F}$, hence

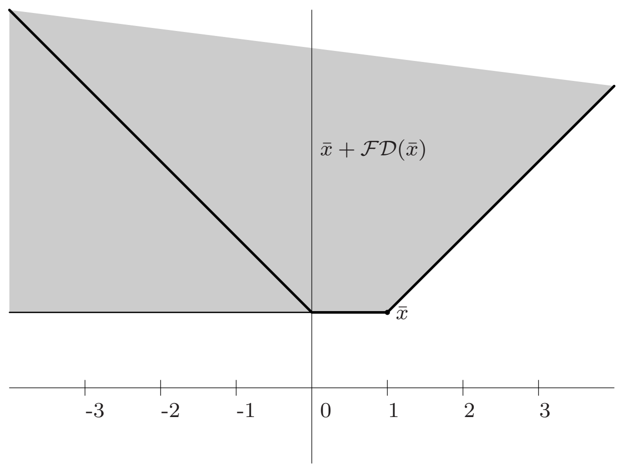
$$\mathcal{F} = \{x \in \mathcal{C} \mid g_j(x) \leq 0, \quad j = 1, \cdots, m\}.$$

Definition 2.10 *The vector $s \in \mathbb{R}^n$ is called a* feasible direction *at a point $x \in \mathcal{F}$ if there is a $\lambda_0 > 0$ such that $x + \lambda s \in \mathcal{F}$ for all $0 \leq \lambda \leq \lambda_0$. The set of feasible directions at the feasible point $x \in \mathcal{F}$ is denoted by $\mathcal{FD}(x)$.*

Example 2.11 Assume that the feasible set $\mathcal{F} \subset \mathbb{R}^2$ is defined by the three constraints

$$-x_1 - x_2 + 1 \leq 0, \; 1 - x_2 \leq 0, \; x_1 - x_2 \leq 0.$$

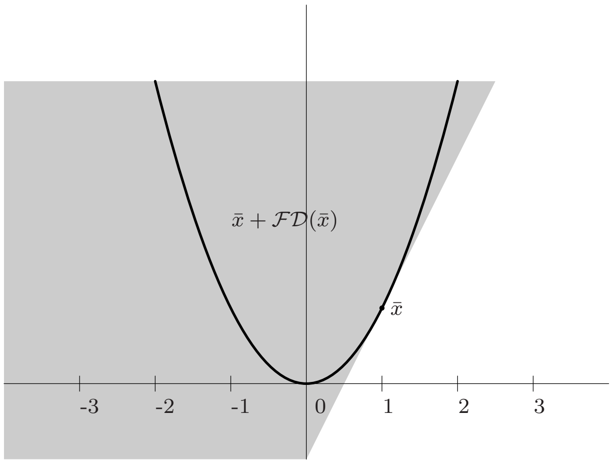
If $\bar{x} = (1, 1)$, then the set of feasible directions at $\bar{x}$ is $\mathcal{FD}(\bar{x}) = \{s \in \mathbb{R}^2 | \; s_2 \geq s_1, s_2 \geq 0\}$. Note that in this case $\mathcal{FD}(\bar{x})$ is a closed convex set.



*

Example 2.12 Assume that the feasible set $\mathcal{F} \subset \mathbb{R}^2$ is defined by the single constraint $x_1^2 - x_2 \leq 0$.

If $\bar{x} = (1,1)$, then the set of feasible directions at $\bar{x}$ is $\mathcal{FD}(\bar{x}) = \{s \in \mathbb{R}^2 | \; s_2 > 2s_1\}$. Observe that now $\mathcal{FD}(\bar{x})$ is an open set.



*

Lemma 2.13 *For any $x \in \mathcal{F}$ the set of feasible directions $\mathcal{FD}(x)$ is a convex cone.*

Proof: Let $\vartheta > 0$. Obviously, $s \in \mathcal{FD}(x)$ implies $(\vartheta s) \in \mathcal{FD}(x)$ since $x + \frac{\lambda}{\vartheta}(\vartheta s) = x + \lambda s \in \mathcal{F}$, hence $\mathcal{FD}(x)$ is a cone. To prove the convexity of $\mathcal{FD}(x)$ let us take $s, \overline{s} \in \mathcal{FD}(x)$. Then by definition we have $x + \lambda s \in \mathcal{F}$ and $x + \lambda \overline{s} \in \mathcal{F}$ for some $\lambda > 0$ (observe that a common λ can be taken). Further, for $0 \leq \alpha \leq 1$ we write

$$x + \lambda(\alpha s + (1-\alpha)\overline{s}) = \alpha(x + \lambda s) + (1-\alpha)(x + \lambda \overline{s}).$$

Due to the convexity of $\mathcal{F}$ the right hand side of the above equation is in $\mathcal{F}$, hence the convexity of $\mathcal{FD}(x)$ follows. □

In view of the above lemma we may speak about the cone of feasible directions $\mathcal{FD}(x)$ for any $x \in \mathcal{F}$. Note that the cone of feasible directions is not necessarily closed even if the set of feasible solutions $\mathcal{F}$ is closed. Figure 2.1 illustrates the cone $\mathcal{FD}(x)$ for three different choices of $\mathcal{F}$ and x.

We will now formulate an optimality condition in terms of the cone of feasible directions. It states that a feasible solution is optimal if and only if the gradient of the objective in that point has an acute angle with all feasible directions at that point (no feasible descent direction exists).

Theorem 2.14 *The feasible point $\overline{x} \in \mathcal{F}$ is an optimal solution of the convex optimization problem (CO) if and only if for all $s \in \mathcal{FD}(\overline{x})$ one has $\delta f(\overline{x}, s) \geq 0$.*

Proof: Observing that $s \in \mathcal{FD}(\overline{x})$ if and only if $s = \lambda(x - \overline{x})$ for some $x \in \mathcal{F}$ and some $\lambda > 0$, the result follows in the same way as in the proof of Theorem 2.9. □

2.2.1 A geometric interpretation

The purpose of this section is to give a geometric interpretation of the result of Theorem 2.14. In doing so, we will look at where we are going in the rest of this chapter, and what we would like to

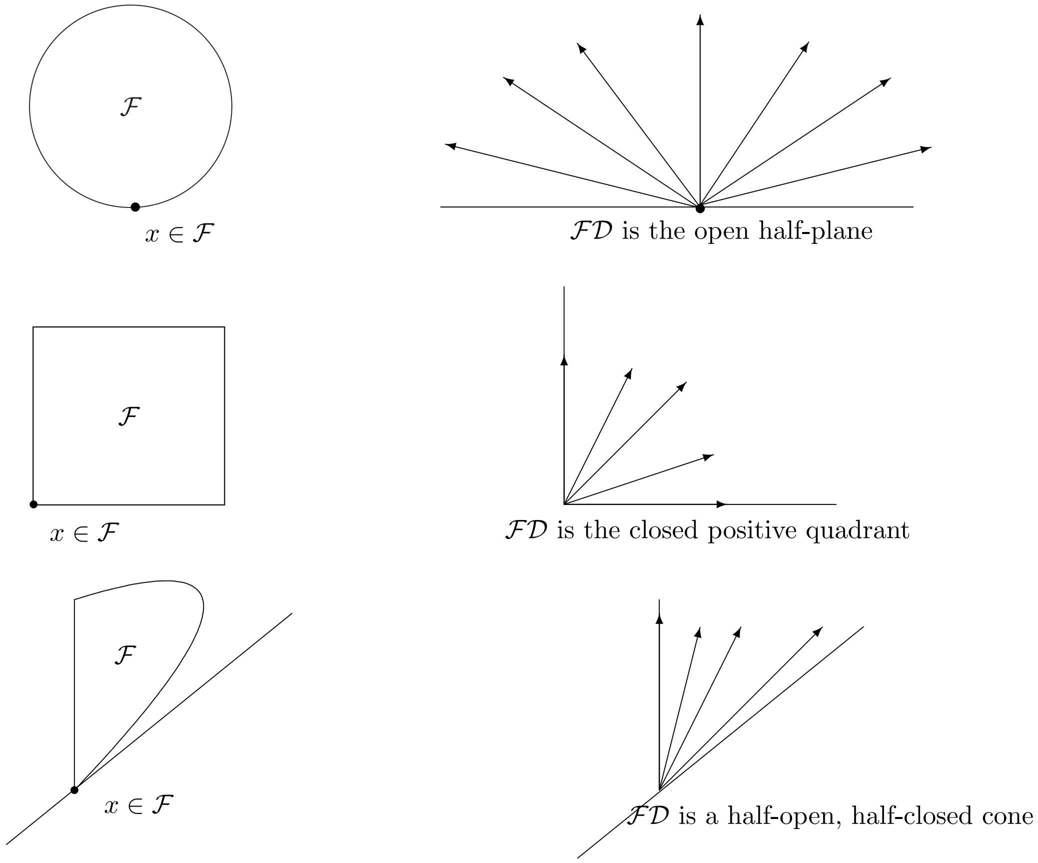


Figure 2.1: Convex feasible sets and cones of feasible directions.

prove. The results in this section are not essential or even necessary in developing the theory further, but should provide some geometric insight and a taste of things to come.

Theorem 2.14 gives us necessary and sufficient optimality conditions, but is not a practical test because we do not have a description of the cone $\mathcal{FD}(\overline{x})$ and therefore cannot perform the test:

$$\text{`is } \delta f(\overline{x}, s) \equiv \nabla f(\overline{x})^T s \geq 0 \text{ for all } s \in \mathcal{FD}(\overline{x})\text{?'} \tag{2.4}$$

It is easy to give a *sufficient* condition for (2.4) to hold, which we will now do. This condition will depend only on the constraint functions that are zero (active) at $\overline{x}$.

Definition 2.15 *A constraint* $g_i(x) \leq 0$ *is called* active *at* $\overline{x} \in \mathcal{F}$ *if* $g_i(\overline{x}) = 0$.

Now let $I_{\overline{x}}$ denote the index set of the active constraints at $\overline{x}$, and assume that $\mathcal{C} = \mathbb{R}^n$.

We now give a sufficient condition for (2.4) to hold (i.e. for $\overline{x}$ to be an optimal solution).

Theorem 2.16 *A point* $\overline{x} \in \mathcal{F}$ *is an optimal solution of problem (CO) (if* $\mathcal{C} = \mathbb{R}^n$*) if*

$$\nabla f(\overline{x}) = -\sum_{i \in I_{\overline{x}}} \overline{y}_i \nabla g_i(\overline{x}), \tag{2.5}$$

for some nonnegative vector $\overline{y}$*, where* $I_{\overline{x}}$ *denotes the index set of the active constraints at* $\overline{x}$*, as before.*

The condition (2.5) is called the Karush-Kuhn-Tucker (KKT) optimality condition,. One can check whether it holds for a given $\overline{x} \in \mathcal{F}$ by using techniques from linear optimization.

The proof that (2.5) is indeed a sufficient condition for optimality follows from the next two exercises.

Exercise 2.5 *Let $s \in \mathcal{FD}(\overline{x})$ be a given feasible direction at $\overline{x} \in \mathcal{F}$ for (CO) and let $\mathcal{C} = \mathbb{R}^n$. One has*

$$\nabla g_i(\overline{x})^T s \leq 0 \text{ for all } i \in I_{\overline{x}}.$$

(Hint: Use Lemma 1.49.) ◁

Exercise 2.6 *Let $\overline{x} \in \mathcal{F}$ be a feasible solution of (CO) where $\mathcal{C} = \mathbb{R}^n$. Use the previous exercise and Theorem 2.14 to show that, if there exists a $\overline{y} \geq 0$ such that*

$$\nabla f(\overline{x}) = -\sum_{i \in I_{\overline{x}}} \overline{y}_i \nabla g_i(\overline{x}),$$

then $\overline{x}$ is an optimal solution of (CO). ◁

Exercise 2.7 *We wish to design a cylindrical can with height h and radius r such that the volume is at least V units and the total surface area is minimal.*

We can formulate this as the following optimization problem:

$$p^* := \min 2\pi r^2 + 2\pi rh$$

subject to

$$\pi r^2 h \geq V, \; r > 0, \; h > 0.$$

1. *Show that we can rewrite the above problem as the following optimization problem:*

$$p^* = \min 2\pi \left(e^{2x_1} + e^{x_1+x_2}\right),$$

subject to

$$\ln\left(\frac{V}{\pi}\right) - 2x_1 - x_2 \leq 0, \; x_1 \in \mathbb{R}, \; x_2 \in \mathbb{R}.$$

2. *Prove that the new problem is a convex optimization problem (CO).*

3. *Prove that the optimal design is where $r = \frac{1}{2}h = \left(\frac{V}{2\pi}\right)^{\frac{1}{3}}$ by using the result of Exercise 2.6.*

◁

The KKT condition (2.5) is sufficient for optimality, but is not a necessary condition for optimality in general, as the next example shows.

Example 2.17 Consider the problem of the form (CO):

$$\min x \text{ subject to } x^2 \leq 0, \; x \in \mathbb{R}.$$

Obviously, the unique optimal solution is $\bar{x} = 0$, and the constraint $g(x) := x^2 \leq 0$ is active at $\bar{x}$.

If we write out condition (2.5), we get

$$1 \equiv \nabla f(\overline{x}) = -\overline{y}\nabla g(\overline{x}) \equiv -\overline{y}(2(0)) = 0,$$

which is obviously not satisfied for any choice of $\overline{y} \geq 0$. In other words, we cannot prove that $\bar{x} = 0$ is an optimal solution by using the KKT condition. ∗

In the rest of the chapter we will show that the KKT conditions are also *necessary* optimality conditions for (CO), if the feasible set $\mathcal{F}$ satisfies an additional assumption called the *Slater condition.*

2.2.2 The Slater condition

We still consider the convex optimization (CO) problem in the form:

$$
\begin{array}{lll}
(CO) & \min & f(x) \\
 & \text{s.t.} & g_j(x) \le 0, \quad j = 1, \cdots, m \\
 & & x \in \mathcal{C},
\end{array}
$$

where $\mathcal{C} \subseteq \mathbb{R}^n$ is a convex set and $f, g_1, \cdots, g_m$ are convex functions on $\mathcal{C}$ (or on an open set that contains the set $\mathcal{C}$). Almost always we will assume that the functions f and g_j are differentiable. The set of indices $\{1, \cdots, m\}$ is denoted by J, and the set of feasible solutions by $\mathcal{F}$, hence

$$\mathcal{F} = \{x \in \mathcal{C} \mid g_j(x) \le 0, \quad j \in J\}.$$

We now introduce the assumption on $\mathcal{F}$ that we referred to in the previous section, namely the *Slater condition.*

Definition 2.18 *A vector (point) $x^0 \in \mathcal{C}^0$ is called a* Slater point *of (CO) if*

$$
\begin{array}{lll}
g_j(x^0) < 0, & \textit{for all } j \textit{ where} & g_j \textit{ is nonlinear,} \\
g_j(x^0) \le 0, & \textit{for all } j \textit{ where} & g_j \textit{ is linear.}
\end{array}
$$

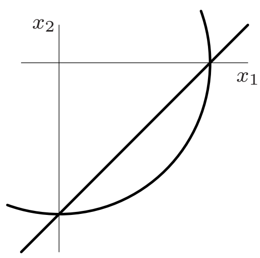
If a Slater point exists we say that (CO) is Slater regular *or (CO) satisfies the Slater condition, or (CO) satisfies the Slater constraint qualification.*

Example 2.19

1. Let us consider the optimization problem

$$
\begin{array}{ll}
\min & f(x) \\
\text{s.t.} & x_1^2 + x_2^2 \le 4 \\
 & x_1 - x_2 \ge 2 \\
 & x_2 \ge 0 \\
 & \mathcal{C} = \mathbb{R}^2.
\end{array}
$$

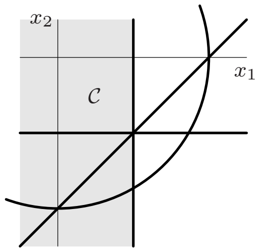
The feasible region $\mathcal{F}$ contains only one point, $(2, 0)$, for which the non-linear constraint becomes an equality. Hence, the problem is not Slater regular.



2. Let us consider the optimization problem

$$
\begin{array}{ll}
\min & f(x) \\
\text{s.t.} & x_1^2 + x_2^2 \le 4 \\
 & x_1 - x_2 \ge 2 \\
 & x_2 \ge -1 \\
 & \mathcal{C} = \{x \mid x_1 \le 1\}.
\end{array}
$$

Again the feasible region contains only one point, $(1, -1)$. For this point the non-linear constraint holds with strict inequality. However, this point does not lie in the relative interior of $\mathcal{C}$. Hence, the problem is not Slater regular.



*

Exercise 2.8 *Assume that (CO) satisfies the Slater condition. Prove that any $x \in \mathcal{F}^0$ is a Slater point of (CO).* ◁

Exercise 2.9 *By solving a so-called first-phase problem one can check whether a given problem of the form (CO) satisfies the Slater condition. Let us assume that $\mathcal{C} = \mathbb{R}^n$ and consider the first-phase problem*

$$\begin{array}{ll} \min & \tau \\ \text{s.t.} & g_j(x) - \tau \leq 0, \quad j = 1, \cdots, m \\ & x \in \mathbb{R}^n, \ \tau \in \mathbb{R}, \end{array}$$

where τ is an auxiliary variable.

(a) Show that the first-phase problem is Slater regular.

(b) What information can you gain about problem (CO) by looking at the optimal objective value τ^* of the first–phase problem? (Consider the cases: $\tau^* > 0$, $\tau^* = 0$ and $\tau^* < 0$.) ◁

We can further refine our definition. Some constraint functions $g_j(x)$ might take the value zero for all feasible points. Such constraints are called *singular* while the others are called *regular*. Hence the index set of singular constraints is defined as

$$J_s = \{j \in J \mid g_j(x) = 0 \text{ for all } x \in \mathcal{F}\},$$

while the index set of regular (qualified) constraints is defined as the complement of the singular set

$$J_r = J \setminus J_s = \{j \in J \mid g_j(x) < 0 \text{ for some } x \in \mathcal{F}\}.$$

Remark: Note, that if (CO) is Slater regular, then all singular functions must be linear.

Definition 2.20 *A Slater point $x^* \in \mathcal{C}^0$ is called an* Ideal Slater point *of the convex optimization problem (CO) if*

$$\begin{array}{ll} g_j(x^*) < 0 & \text{for all } j \in J_r, \\ g_j(x^*) = 0 & \text{for all } j \in J_s. \end{array}$$

First we show an elementary property.

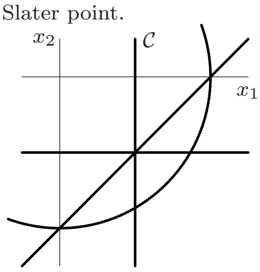
Lemma 2.21 *If the convex optimization problem (CO) is Slater regular then there exists an ideal Slater point $x^* \in \mathcal{F}$.*

Proof: According to the assumption, there exists a Slater point $x^0 \in \mathcal{C}^0$ and there exist points $x^k \in \mathcal{F}$ for all $k \in J_r$ such that $g_k(x^k) < 0$. Let $\lambda_0 > 0$, $\lambda_k > 0$ for all $k \in J_r$ such that $\lambda_0 + \sum_{k \in J_r} \lambda_k = 1$, then $x^* = \lambda_0 x^0 + \sum_{j \in J_r} \lambda_k x^k$ is an ideal Slater point. This last statement follows from the convexity of the functions g_j. □

Example 2.22
1. Let us consider the optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x_1^2 + x_2^2 \leq 4 \\ & x_1 - x_2 \geq 2 \\ & x_2 \geq -1 \\ & \mathcal{C} = \{x \mid x_1 = 1\}. \end{array}$$

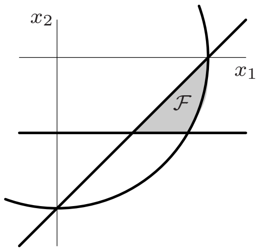
The feasible region contains only one point, $(1, -1)$, but now this point does lie in the relative interior of the convex set $\mathcal{C}$. Hence, this point is an ideal Slater point.



2. Let us consider the optimization problem

$$\begin{array}{ll} \min & f(x) \\ \text{s.t.} & x_1^2 + x_2^2 \leq 4 \\ & x_1 - x_2 \geq 2 \\ & x_2 \geq -1 \\ & \mathcal{C} = \mathbb{R}^2. \end{array}$$

Now, the point $(1, -1)$ is again a Slater point, but not an ideal Slater point. The point $(\frac{3}{2}, -\frac{3}{4})$ is an ideal Slater point.



*

Exercise 2.10 *Prove that any ideal Slater point of (CO) is in the relative interior of $\mathcal{F}$.* ◁

2.2.3 Convex Farkas lemma

The convex Farkas lemma is an example of a *theorem of alternatives*, which means that it is a statement of the type: for two specific systems of inequalities (I) and (II), (I) has a solution if and only if (II) has no solution. It will play an essential role in developing the KKT theory.

Before stating the convex Farkas lemma, we present a simple separation theorem. It essentially states that disjoint convex sets can be separated by a (hyper)plane (geometrically, in $\mathbb{R}^2$ or $\mathbb{R}^3$, the convex sets lie on different sides of the separating plane). Its proof can be found in most textbooks (see e.g. [2]).

Theorem 2.23 *Let $\mathcal{U} \subseteq \mathbb{R}^n$ be a convex set and a point $w \in \mathbb{R}^n$ with $w \notin \mathcal{U}$ be given. Then there is a separating hyperplane $\{x \mid a^T x = \alpha\}$, with $a \in \mathbb{R}^n$, $\alpha \in \mathbb{R}$ such that*

1. *$a^T w \leq \alpha$;*
2. *$a^T u \geq \alpha$ for all $u \in \mathcal{U}$ but $\mathcal{U}$ is not a subset of the hyperplane.*

Note that the last property says that there is a $\overline{u} \in \mathcal{U}$ such that $a^T\overline{u} > \alpha$.

Now we are ready to prove the convex Farkas Lemma. The proof here is a simplified version of the proofs in the books [38, 42].

Lemma 2.24 (Farkas) *The convex optimization problem (CO) is given and we assume that the Slater regularity condition is satisfied. The inequality system*

$$\begin{aligned} &f(x) < 0 \\ &g_j(x) \leq 0, \qquad j = 1, \cdots, m \\ &x \in \mathcal{C}, \end{aligned} \tag{2.6}$$

has no solution if and only if there exists a vector $y = (y_1, \cdots, y_m) \geq 0$ such that

$$f(x) + \sum_{j=1}^{m} y_j g_j(x) \geq 0 \quad \textit{for all } x \in \mathcal{C}. \tag{2.7}$$

Before proving this important result we make remark. The systems (2.6) and (2.7) are called *alternative systems*, i.e. exactly one of them has a solution.

Proof: If the system (2.6) has a solution then clearly (2.7) cannot be true for that solution. This is the trivial part of the lemma. Note that this part is true without any regularity condition.

To prove the other side let us assume that (2.6) has no solution. With $u = (u_0, \cdots, u_m)$, we define the set $\mathcal{U} \in \mathbb{R}^{m+1}$ as follows.

$$\mathcal{U} = \{u \mid \exists x \in \mathcal{C} \text{ with } u_0 > f(x),\ u_j \geq g_j(x) \text{ if } j \in J_r,\ \ u_j = g_j(x) \text{ if } j \in J_s\}.$$

Clearly the set $\mathcal{U}$ is convex (note that due to the Slater condition singular functions are linear) and due to the infeasibility of (2.6) it does not contain the origin. Hence according to Theorem 2.23 there exists a separating hyperplane defined by an appropriate vector $(y_0, y_1, \cdots, y_m)$ and $\alpha = 0$ such that

$$\sum_{j=0}^{m} y_j u_j \geq 0 \quad \text{for all } u \in \mathcal{U} \tag{2.8}$$

and for some $\overline{u} \in \mathcal{U}$ one has

$$\sum_{j=0}^{m} y_j \overline{u}_j > 0. \tag{2.9}$$

The rest of the proof is divided into four parts.

I. First we prove that $y_0 \geq 0$ and $y_j \geq 0$ for all $j \in J_r$.

II. Secondly we establish that (2.8) holds for $u = (f(x), g_1(x), \cdots, g_m(x))$ if $x \in \mathcal{C}$.

III. Then we prove that y_0 must be positive.

IV. Finally, it is shown by using induction that we can assume $y_j > 0$ for all $j \in J_s$.

I. First we show that $y_0 \geq 0$ and $y_j \geq 0$ for all $j \in J_r$. Let us assume that $y_0 < 0$. Let us take an arbitrary $(u_0, u_1, \cdots, u_m) \in \mathcal{U}$. By definition $(u_0 + \lambda, u_1, \cdots, u_m) \in \mathcal{U}$ for all $\lambda \geq 0$. Hence by (2.8) one has

$$\lambda y_0 + \sum_{j=0}^{m} y_j u_j \geq 0 \quad \text{for all } \lambda \geq 0.$$

For sufficiently large λ the left hand side is negative, which is a contradiction, i.e. y_0 must be nonnegative. The proof of the nonnegativity of all y_j as $j \in J_r$ goes analogously.

II. Secondly we establish that

$$y_0 f(x) + \sum_{j=1}^{m} y_j g_j(x) \geq 0 \quad \text{for all } x \in \mathcal{C}. \tag{2.10}$$

This follows from the observation that for all $x \in \mathcal{C}$ and for all $\lambda > 0$ one has $u = (f(x) + \lambda, g_1(x), \cdots, g_m(x)) \in \mathcal{U}$, thus

$$y_0(f(x) + \lambda) + \sum_{j=1}^{m} y_j g_j(x) \geq 0 \quad \text{for all } x \in \mathcal{C}.$$

Taking the limit as $\lambda \longrightarrow 0$ the claim follows.

III. Thirdly we show that $y_0 > 0$. The proof is by contradiction. We already know that $y_0 \geq 0$. Let us assume to the contrary that $y_0 = 0$. Hence from (2.10) we have

$$\sum_{j \in J_r} y_j g_j(x) + \sum_{j \in J_s} y_j g_j(x) = \sum_{j=1}^{m} y_j g_j(x) \geq 0 \quad \text{for all } x \in \mathcal{C}.$$

Taking an ideal Slater point $x^* \in \mathcal{C}^0$ one has

$$g_j(x^*) = 0 \ \text{ if } \ j \in J_s,$$

whence

$$\sum_{j \in J_r} y_j g_j(x^*) \geq 0.$$

Since $y_j \geq 0$ and $g_j(x^*) < 0$ for all $j \in J_r$, this implies $y_j = 0$ for all $j \in J_r$. This results in

$$\sum_{j \in J_s} y_j g_j(x) \geq 0 \quad \text{for all } x \in \mathcal{C}. \tag{2.11}$$

Now, from (2.9), with $\overline{x} \in \mathcal{C}$ such that $\overline{u}_j = g_j(\overline{x})$ if $i \in J_s$ we have

$$\sum_{j \in J_s} y_j g_j(\overline{x}) > 0. \tag{2.12}$$

Because the ideal Slater point x^* is in the relative interior of $\mathcal{C}$ there exist a vector $\tilde{x} \in \mathcal{C}$ and $0 < \lambda < 1$ such that $x^* = \lambda\overline{x} + (1-\lambda)\tilde{x}$. Using that $g_j(x^*) = 0$ for $j \in J_s$ and that the singular functions are linear one gets

$$\begin{aligned}
0 = &\ \textstyle\sum_{j \in J_s} y_j g_j(x^*) \\
= &\ \textstyle\sum_{j \in J_s} y_j g_j(\lambda\overline{x} + (1-\lambda)\tilde{x}) \\
= &\ \textstyle\lambda \sum_{j \in J_s} y_j g_j(\overline{x}) + (1-\lambda)\sum_{j \in J_s} y_j g_j(\tilde{x}) \\
> &\ \textstyle(1-\lambda)\sum_{j \in J_s} y_j g_j(\tilde{x}).
\end{aligned}$$

Here the last inequality follows from (2.12). The inequality

$$(1-\lambda)\sum_{j \in J_s} y_j g_j(\tilde{x}) < 0$$

contradicts (2.11). Hence we have proved that $y_0 > 0$.

At this point we have (2.10) with $y_0 > 0$ and $y_j \geq 0$ for all $j \in J_r$. Dividing by $y_0 > 0$ in (2.10) and by defining $y_j := \frac{y_j}{y_0}$ for all $j \in J$ we obtain

$$f(x) + \sum_{j=1}^{m} y_j g_j(x) \geq 0 \quad \text{for all } x \in \mathcal{C}. \tag{2.13}$$

We finally show that y may be taken such that $y_j > 0$ for all $j \in J_s$.

IV. To complete the proof we show by induction on the cardinality of J_s that one can make y_j positive for all $j \in J_s$. Observe that if $J_s = \emptyset$ then we are done. If $|J_s| = 1$ then we apply the results proved till this point to the inequality system

$$\begin{aligned} &g_s(x) < 0, \\ &g_j(x) \leq 0, \qquad j \in J_r, \\ &x \in \mathcal{C} \end{aligned} \tag{2.14}$$

where $\{s\} = J_s$. The system (2.14) has no solution, it satisfies the Slater condition, and therefore there exists a $\hat{y} \in \mathbb{R}^{m-1}$ such that

$$g_s(x) + \sum_{j \in J_r} \hat{y}_j g_j(x) \geq 0 \quad \text{for all } x \in \mathcal{C}, \tag{2.15}$$

where $\hat{y}_j \geq 0$ for all $j \in J_r$. Adding a sufficiently large positive multiple of (2.15) to (2.13) one obtains a positive coefficient $\hat{y}_s > 0$ for $g_s(x)$.

The general inductive step goes analogously. Assuming that the result is proved if $|J_s| = k$ then the result is proved for the case $|J_s| = k + 1$. Let $s \in J_s$ then $|J_s \setminus \{s\}| = k$, and hence the inductive assumption applies to the system

$$\begin{aligned} &g_s(x) < 0 \\ &g_j(x) \leq 0, \qquad j \in J_s \setminus \{s\}, \\ &g_j(x) \leq 0, \qquad j \in J_r, \\ &x \in \mathcal{C}. \end{aligned} \tag{2.16}$$

By construction the system (2.16) has no solution, it satisfies the Slater condition, and by the inductive assumption we have a $\hat{y} \in \mathbb{R}^{m-1}$ such that

$$g_s(x) + \sum_{j \in J_r \cup J_s \setminus \{s\}} \hat{y}_j g_j(x) \geq 0 \quad \text{for all } x \in \mathcal{C}. \tag{2.17}$$

where $\hat{y}_j > 0$ for all $j \in J_s \setminus \{s\}$ and $\hat{y}_j \geq 0$ for all $j \in J_r$. Adding a sufficiently large multiple of (2.17) to (2.13), one obtains the desired nonnegative multipliers. □

Remark: Note, that finally we proved slightly more than was stated. We have proved that the multipliers of all the singular constraints can be made strictly positive.

Example 2.25 [Farkas Lemma]

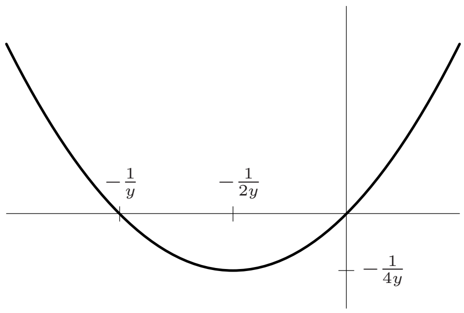
1. Let us consider the convex optimization problem

$$\begin{array}{lll} \text{(CO)} & \min & x \\ & \text{s.t.} & x^2 \leq 0 \\ & & x \in \mathbb{R}. \end{array}$$

Then (CO) is *not Slater regular*.
The system

$$\begin{array}{rcl} x & < & 0 \\ x^2 & \leq & 0 \end{array}$$

has no solution, but for every $y > 0$ the quadratic function $f(x) = x + yx^2$ has two zeroes.



So, there is no $y \geq 0$ such that

$$x + yx^2 \geq 0 \text{ for all } x \in \mathbb{R}.$$

Hence, *the Farkas Lemma does not hold* for (CO).

2. Let us consider the convex optimization problem

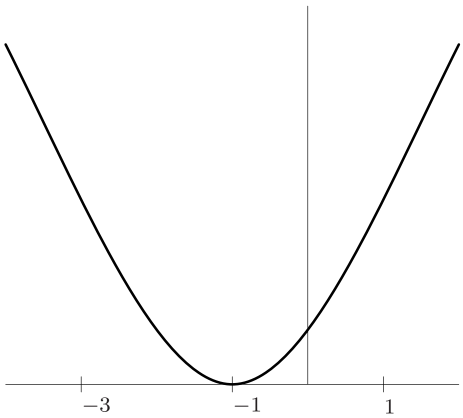
$$\begin{array}{lll} \text{(CO)} & \min & 1 + x \\ & \text{s.t.} & x^2 - 1 \leq 0 \\ & & x \in \mathbb{R}. \end{array}$$

Then (CO) is Slater regular (0 is an ideal Slater point). The system

$$\begin{array}{rcl} 1 + x & < & 0 \\ x^2 - 1 & \leq & 0 \end{array}$$

has no solution. If we let $y = \frac{1}{2}$ the quadratic function

$$g(x) = x + 1 + y(x^2 - 1) = \frac{1}{2}x^2 + x + \frac{1}{2}$$



has only one zero, thus one has

$$\frac{1}{2}x^2 + x + \frac{1}{2} \geq 0 \text{ for all } x \in \mathbb{R}.$$

*

Exercise 2.11 *Let the matrices $A : m \times n$ and the vector $b \in \mathbb{R}^m$ be given. Apply the convex Farkas Lemma 2.24 to prove that exactly one of the following alternative systems (I) or (II) is solvable:*

$$(I) \qquad Ax \leq b, \qquad x \geq 0,$$

or

$$(II) \qquad A^T y \geq 0, \qquad y \geq 0, \qquad b^T y < 0.$$

◁

Exercise 2.12 *Let the matrices $A : m \times n$, $B : k \times n$ and the vectors $a \in \mathbb{R}^m$, $b \in \mathbb{R}^k$ be given. With a proper reformulation, apply the convex Farkas Lemma 2.24 to the inequality system*

$$Ax \leq a, \qquad Bx < b, \qquad x \geq 0$$

to derive its alternative system. ◁

Exercise 2.13 *Let the matrix $A : m \times n$ and the vectors $c \in \mathbb{R}^n$ and $b \in R^m$ be given. Apply the convex Farkas Lemma 2.24 to prove the so-called Goldman–Tucker theorem for the LO problem:*

$$\min \{c^T x \ : \ Ax = b, \qquad x \geq 0\}$$

when it admits an optimal solution. In other words, prove that there exists an optimal solution x^* and an optimal solution (y^*, s^*) of the dual LO problem

$$\max \{b^T y \ : \ A^T y + s = c, \qquad s \geq 0\}$$

such that

$$x^* + s^* > 0.$$

◁

2.2.4 Karush–Kuhn–Tucker theory

For the convex optimization problem (CO) one defines the Lagrange function (or *Lagrangian*)

$$L(x, y) := f(x) + \sum_{j=1}^{m} y_j g_j(x) \tag{2.18}$$

where $x \in \mathcal{C}$ and $y \geq 0$. Note that for fixed y the Lagrange function is convex in x.

Definition 2.26 *A vector pair $(\overline{x}, \overline{y}) \in \mathbb{R}^{n+m}$, $\overline{x} \in \mathcal{C}$ and $\overline{y} \geq 0$ is called a saddle point of the Lagrange function L if*

$$L(\overline{x}, y) \leq L(\overline{x}, \overline{y}) \leq L(x, \overline{y}) \tag{2.19}$$

for all $x \in \mathcal{C}$ and $y \geq 0$.

One easily sees that (2.19) is equivalent with

$$L(\overline{x}, y) \leq L(x, \overline{y}) \quad \text{for all} \quad x \in \mathcal{C}, \quad y \geq 0.$$

We will see (in the proof of Theorem 2.30) that the $\overline{x}$ part of a saddle point is always an optimal solution of (CO).

Example 2.27 **[Saddle point]** Let us consider the convex optimization problem

$$\begin{array}{lll} \text{(CO)} & \min & -x + 2 \\ & \text{s.t.} & e^x - 4 \leq 0 \\ & & x \in \mathbb{R} \end{array}$$

Then the Lagrange function of (CO) is given by

$$L(x, y) = -x + 2 + y(e^x - 4),$$

where the Lagrange multiplier y is non-negative. For fixed $y > 0$ we have

$$\frac{\partial}{\partial x} L(x, y) = -1 + y e^x = 0$$

for $x = -\log y$, thus $L(-\log y, y) = \log y - 4y + 3$ is a minimum.
On the other hand, for feasible x, i.e. if $x \leq \log 4$, we have

$$\sup_{y \geq 0} \ y(e^x - 4) = 0.$$

Hence, defining $\psi(y) = \inf_{x \in \mathbb{R}} L(x, y)$ and $\phi(x) = \sup_{y \geq 0} L(x, y)$ we have

$$\psi(y) = \begin{cases} \log y - 4y + 3 & \text{for } y > 0, \\ -\infty & \text{for } y = 0; \end{cases}$$

$$\phi(x) = \begin{cases} -x + 2 & \text{for } x \leq \log 4, \\ \infty & \text{for } x > \log 4. \end{cases}$$

Now, we have

$$\frac{d}{dy} \psi(y) = \frac{1}{y} - 4 = 0$$

for $y = \frac{1}{4}$, i.e. this value gives the maximum of $\psi(y)$. Hence, $\sup_{y \geq 0} \psi(y) = -\log 4 + 2$. The function $\phi(x)$ is minimal for $x = \log 4$, thus $\inf_{x \in \mathbb{R}} \phi(x) = -\log 4 + 2$ and we conclude that $(\log 4, \frac{1}{4})$ is a saddle point of the Lagrange function $L(x, y)$. Note that $x = \log 4$ is the optimal solution of (CO). *

Lemma 2.28 *A saddle point* $(\overline{x},\overline{y}) \in \mathbb{R}^{n+m}$, $\overline{x} \in \mathcal{C}$ *and* $\overline{y} \geq 0$ *of* $L(x,y)$ *satisfies the relation*

$$\inf_{x\in\mathcal{C}} \sup_{y\geq 0} L(x,y) = L(\overline{x},\overline{y}) = \sup_{y\geq 0} \inf_{x\in\mathcal{C}} L(x,y). \tag{2.20}$$

Proof: For any $(\hat{x},\hat{y})$ one has

$$\inf_{x\in\mathcal{C}} L(x,\hat{y}) \leq L(\hat{x},\hat{y}) \leq \sup_{y\geq 0} L(\hat{x},y),$$

hence one can take the supremum of the left hand side and the infimum of the right hand side resulting in

$$\sup_{y\geq 0} \inf_{x\in\mathcal{C}} L(x,y) \leq \inf_{x\in\mathcal{C}} \sup_{y\geq 0} L(x,y). \tag{2.21}$$

Further using the saddle point inequality (2.19) one obtains

$$\inf_{x\in\mathcal{C}} \sup_{y\geq 0} L(x,y) \leq \sup_{y\geq 0} L(\overline{x},y) \leq L(\overline{x},\overline{y}) \leq \inf_{x\in\mathcal{C}} L(x,\overline{y}) \leq \sup_{y\geq 0} \inf_{x\in\mathcal{C}} L(x,y). \tag{2.22}$$

Combining (2.22) and (2.21) the equality (2.20) follows. □

Condition (2.20) is a property of saddle points. If some $\overline{x} \in \mathcal{C}$ and $\overline{y} \geq 0$ satisfy (2.20), it does not imply that $(\overline{x},\overline{y})$ is a saddle point though, as the following example shows.

Example 2.29 Let (CO) be given by

$$\min e^x \text{ subject to } x \leq 0.$$

Here

$$L(x,y) = e^x + yx.$$

It is easy to verify that $\overline{x} = -1$ and $\overline{y} = e^{-1}$ satisfy (2.20). Indeed, $L(\overline{x},\overline{y}) = 0$ and

$$\inf_{x\in\mathcal{C}} \sup_{y\geq 0} L(x,y) = 0,$$

by letting x tend to $-\infty$. Likewise

$$\sup_{y\geq 0} \inf_{x\in\mathcal{C}} L(x,y) = 0.$$

However, $(\overline{x},\overline{y})$ is *not* a saddle point of L. (This example does not have an optimal solution, and, as we have mentioned, the $\overline{x}$ part of a saddle point is always an optimal solution of (CO).) ∗

We still do not know if a saddle point exists or not. Assuming Slater regularity, the next result states that $L(x,y)$ has a saddle point if and only if (CO) has an optimal solution.

Theorem 2.30 (Karush–Kuhn–Tucker) *The convex optimization problem (CO) is given. Assume that the Slater regularity condition is satisfied. The vector* $\overline{x}$ *is an optimal solution of (CO) if and only if there is a vector* $\overline{y}$ *such that* $(\overline{x},\overline{y})$ *is a saddle point of the Lagrange function* L.

Proof: The easy part of the theorem is to prove that if $(\overline{x},\overline{y})$ is a saddle point of $L(x,y)$ then $\overline{x}$ is optimal for (CO). The proof of this part does not need any regularity condition. From the saddle point inequality (2.19) one has

$$f(\overline{x}) + \sum_{j=1}^{m} y_j g_j(\overline{x}) \leq f(\overline{x}) + \sum_{j=1}^{m} \overline{y}_j g_j(\overline{x}) \leq f(x) + \sum_{j=1}^{m} \overline{y}_j g_j(x)$$

for all $y \geq 0$ and for all $x \in \mathcal{C}$. From the first inequality one easily derives $g_j(\overline{x}) \leq 0$ for all $j = 1, \cdots, m$ hence $\overline{x} \in \mathcal{F}$ is a feasible solution of (CO). Taking the two extreme sides of the above inequality and substituting $y = 0$ we have

$$f(\overline{x}) \leq f(x) + \sum_{j=1}^{m} \overline{y}_j g_j(x) \leq f(x)$$

for all $x \in \mathcal{F}$, i.e. $\overline{x}$ is optimal.

To prove the other direction we need Slater regularity and the Convex Farkas Lemma 2.24. Let us take an optimal solution $\overline{x}$ of the convex optimization problem (CO). Then the inequality system

$$\begin{aligned} &f(x) - f(\overline{x}) < 0 \\ &g_j(x) \le 0, \qquad\qquad j = 1, \cdots, m \\ &x \in \mathcal{C} \end{aligned}$$

is infeasible. By the Convex Farkas Lemma 2.24 there exists a $\overline{y} \ge 0$ such that

$$f(x) - f(\overline{x}) + \sum_{j=1}^{m} \overline{y}_j g_j(x) \ge 0$$

for all $x \in \mathcal{C}$. Using that $\overline{x}$ is feasible one easily derive the saddle point inequality

$$f(\overline{x}) + \sum_{j=1}^{m} y_j g_j(\overline{x}) \le f(\overline{x}) \le f(x) + \sum_{j=1}^{m} \overline{y}_j g_j(x)$$

for all $y \ge 0$ and $x \in \mathcal{C}$, which completes the proof. □

The following corollaries lead us to the Karush–Kuhn-Tucker (KKT) optimality conditions.

Corollary 2.31 *Under the assumptions of Theorem 2.30 the vector $\overline{x} \in \mathcal{C}$ is an optimal solution of (CO) if and only if there exists a $\overline{y} \ge 0$ such that*

$$\begin{aligned} &(i) \quad f(\overline{x}) = \min_{x \in \mathcal{C}} \{ f(x) + \sum_{j=1}^{m} \overline{y}_j g_j(x) \} \qquad \text{and} \\ &(ii) \quad \sum_{j=1}^{m} \overline{y}_j g_j(\overline{x}) = \max_{y \ge 0} \{ \sum_{j=1}^{m} y_j g_j(\overline{x}) \}. \end{aligned}$$

Proof: Easily follows from the theorem. □

Corollary 2.32 *Under the assumptions of Theorem 2.30 the vector $\overline{x} \in \mathcal{F}$ is an optimal solution of (CO) if and only if there exists a $\overline{y} \ge 0$ such that*

$$\begin{aligned} &(i) \quad f(\overline{x}) = \min_{x \in \mathcal{C}} \{ f(x) + \sum_{j=1}^{m} \overline{y}_j g_j(x) \} \qquad \text{and} \\ &(ii) \quad \sum_{j=1}^{m} \overline{y}_j g_j(\overline{x}) = 0. \end{aligned}$$

Proof: Easily follows from the Corollary 2.31. □

Corollary 2.33 *Let us assume that $\mathcal{C} = \mathbb{R}^n$ and the functions $f, g_1, \cdots, g_m$ are continuously differentiable functions. Under the assumptions of Theorem 2.30 the vector $\overline{x} \in \mathcal{F}$ is an optimal solution of (CO) if and only if there exists a $\overline{y} \ge 0$ such that*

$$\begin{aligned} &(i) \quad 0 = \nabla f(\overline{x}) + \sum_{j=1}^{m} \overline{y}_j \nabla g_j(\overline{x}) \qquad \text{and} \\ &(ii) \quad \sum_{j=1}^{m} \overline{y}_j g_j(\overline{x}) = 0. \end{aligned}$$

Proof: Follows directly from Corollary 2.32 and the convexity of the function $f(x)+\sum_{j=1}^{m} \overline{y}_j g_j(x),\ x \in \mathcal{C}$. $\square$

Exercise 2.14 *Prove the above three Corollaries.* $\triangleleft$

Note that the last corollary stays valid if $\mathcal{C}$ is a full dimensional open subset of $\mathbb{R}^n$. If the set $\mathcal{C}$ is not full dimensional, then the right hand side vector, the $x-$gradient of the Lagrange function has to be orthogonal to any direction in the affine hull of $\mathcal{C}$ (*cf.* Theorem 2.9). To check the validity of these statements is left to the reader.

Now the notion of Karush–Kuhn–Tucker (KKT) point is defined.

Definition 2.34 (KKT point) *Let us assume that $\mathcal{C} = \mathbb{R}^n$ and the functions $f, g_1, \cdots, g_m$ are continuously differentiable functions. The vector $(\overline{x}, \overline{y}) \in \mathbb{R}^{n+m}$ is called a Karush–Kuhn–Tucker (KKT) point of (CO) if*

$$\begin{aligned}
&(i) && g_j(\overline{x}) \leq 0, \text{ for all } j \in J,\\
&(ii) && 0 = \nabla f(\overline{x}) + \sum_{j=1}^{m} \overline{y}_j \nabla g_j(\overline{x})\\
&(iii) && \sum_{j=1}^{m} \overline{y}_j g_j(\overline{x}) = 0,\\
&(iv) && \overline{y} \geq 0.
\end{aligned}$$

It is important to understand that — under the assumptions of Corollary 2.33 — $(\overline{x}, \overline{y})$ is a saddle point of the Lagrangian of (CO) if and only if it is a KKT point of (CO). The proof is left as an exercise.

Exercise 2.15 *Let us assume that $\mathcal{C} = \mathbb{R}^n$ and the functions $f, g_1, \cdots, g_m$ are continuously differentiable convex functions and the assumptions of Theorem 2.30 hold. Show that $(\overline{x}, \overline{y})$ is a saddle point of the Lagrangian of (CO) if and only if it is a KKT point of (CO).* $\triangleleft$

The so-called Karush–Kuhn–Tucker sufficient optimality conditions now follow from Corollary 2.33.

Corollary 2.35 *Let us assume that $\mathcal{C} = \mathbb{R}^n$ and the functions $f, g_1, \cdots, g_m$ are continuously differentiable convex functions and the assumptions of Theorem 2.30 hold. Let the vector $(\overline{x}, \overline{y})$ be a KKT point, then $\overline{x}$ is an optimal solution of (CO).*

Thus we have derived necessary and sufficient optimality conditions for the convex optimization problem (CO) under the Slater regularity assumption. Note that if an optimization problem is not convex, or does not satisfy any regularity condition, then only weaker results can be proven.

Chapter 3

Duality in convex optimization

Every optimization problem has an associated dual optimization problem. Under some assumptions, a convex optimization problem (CO) and its dual have the same optimal objective values. We can therefore use the dual problem to show that a certain solution of (CO) is in fact optimal. Moreover, some optimization algorithms solve (CO) and its dual problem at the same time, and when the objective values are the same then optimality has been proved. One can easily derive dual problems and duality results from the KKT theory or from the Convex Farkas Lemma. First we define the more general Lagrange dual and then we specialize it to get the so-called Wolfe dual for convex problems.

3.1 Lagrange dual

Definition 3.1 *Denote* $\psi(y) = \inf_{x\in\mathcal{C}}\{f(x) + \sum_{j=1}^m y_j g_j(x)\}$*. The problem*

$$\begin{array}{ll} (LD) & \sup \psi(y) \\ & y \geq 0 \end{array}$$

is called the Lagrange dual of the convex optimization problem (CO).

Lemma 3.2 *The Lagrange Dual (LD) of (CO) is a convex optimization problem, even if the functions* $f, g_1, \cdots, g_m$ *are not convex.*

Proof: Because the maximization of $\psi(y)$ is equivalent to the minimization of $-\psi(y)$, we have only to prove that the function $-\psi(y)$ is convex, i.e. $\psi(y)$ is concave. Let $\overline{y}, \hat{y} \geq 0$ and $0 \leq \lambda \leq 1$. Using that the infimum of the sum of two functions is larger than the sum of the two separate infimums one has:

$$\begin{aligned}
\psi(\lambda\overline{y} + (1-\lambda)\hat{y}) &= \inf_{x\in\mathcal{C}}\left\{f(x) + \sum_{j=1}^m (\lambda\overline{y}_j + (1-\lambda)\hat{y}_j)g_j(x)\right\} \\
&= \inf_{x\in\mathcal{C}}\left\{\lambda\left[f(x) + \sum_{j=1}^m \overline{y}_j g_j(x)\right] + (1-\lambda)\left[f(x) + \sum_{j=1}^m \hat{y}_j g_j(x)\right]\right\} \\
&\geq \inf_{x\in\mathcal{C}}\left\{\lambda\left[f(x) + \sum_{j=1}^m \overline{y}_j g_j(x)\right]\right\} + \inf_{x\in\mathcal{C}}\left\{(1-\lambda)\left[f(x) + \sum_{j=1}^m \hat{y}_j g_j(x)\right]\right\} \\
&= \lambda\psi(\overline{y}) + (1-\lambda)\psi(\hat{y}).
\end{aligned}$$

□

Definition 3.3 *If $\overline{x}$ is a feasible solution of (CO) and $\overline{y} \geq 0$ then we call the quantity*

$$f(\overline{x}) - \psi(\overline{y})$$

the duality gap at $\overline{x}$ and $\overline{y}$.

It is easy to prove the so-called weak duality theorem, which states that the duality gap is always nonnegative.

Theorem 3.4 *If $\overline{x}$ is a feasible solution of (CO) and $\overline{y} \geq 0$ then*

$$\psi(\overline{y}) \leq f(\overline{x})$$

and equality holds if and only if $\inf_{x \in \mathcal{C}} \{f(x) + \sum_{j=1}^{m} \overline{y}_j g_j(x)\} = f(\overline{x})$.

Proof: By straightforward calculations one has

$$\psi(\overline{y}) = \inf_{x \in \mathcal{C}} \{f(x) + \sum_{j=1}^{m} \overline{y}_j g_j(x)\} \leq f(\overline{x}) + \sum_{j=1}^{m} \overline{y}_j g_j(\overline{x}) \leq f(\overline{x}).$$

Equality holds if and only if $\inf_{x \in \mathcal{C}} \{f(x) + \sum_{j=1}^{m} \overline{y}_j g_j(x)\} = f(\overline{x})$ and hence $\overline{y}_j g_j(x) = 0$ for all $j \in J$. □

One easily derives the following corollary.

Corollary 3.5 *If $\overline{x}$ is a feasible solution of (CO), $\overline{y} \geq 0$ and $\psi(\overline{y}) = f(\overline{x})$ then the vector $\overline{x}$ is an optimal solution of (CO) and $\overline{y}$ is optimal for (LD). Further if the functions $f, g_1, \cdots, g_m$ are continuously differentiable then $(\overline{x}, \overline{y})$ is a KKT-point.*

To prove the so-called strong duality theorem one needs a regularity condition.

Theorem 3.6 *Let us assume that (CO) satisfies the Slater regularity condition. Let $\overline{x}$ be a feasible solution of (CO). The vector $\overline{x}$ is an optimal solution of (CO) if and only if there exists a $\overline{y} \geq 0$ such that $\overline{y}$ is an optimal solution of (LD) and*

$$\psi(\overline{y}) = f(\overline{x}).$$

Proof: Directly follows from Corollary 2.31. □

Exercise 3.1 *Prove Theorem 3.6.* ◁

Remark: If the convex optimization problem does not satisfy a regularity condition, then it is not true in general that the duality gap is zero. It is also not always true (even not under regularity condition) that the convex optimization problem has an optimal solution. Frequently only the supremum or the infimum of the objective function exists.

Example 3.7 [Lagrange dual] Let us consider again the problem (see Example 2.25)

$$\begin{array}{lll} \text{(CO)} & \min & x \\ & \text{s.t.} & x^2 \leq 0 \\ & & x \in \mathbb{R}. \end{array}$$

As we have seen this (CO) problem is *not Slater regular* and the Convex Farkas Lemma 2.24 does not apply to the system

$$\begin{aligned} x &< 0 \\ x^2 &\leq 0. \end{aligned}$$

On the other hand, we have

$$\psi(y) = \inf_{x \in \mathbb{R}} (x + yx^2) = \begin{cases} -\frac{1}{4y} & \text{for } y > 0 \\ -\infty & \text{for } y = 0. \end{cases}$$

The Lagrange dual reads

$$\sup_{y \geq 0} \psi(y).$$

The optimal value of the Lagrange dual is zero, i.e. in spite of the lack of Slater regularity there is *no duality gap*. ∗

3.2 Wolfe dual

Observing the similarity of the formulation of the Lagrange dual (LD) and the conditions occurring in the corollaries of the KKT-Theorem 2.30 the so-called Wolfe dual is obtained.

Definition 3.8 *Assume that $\mathcal{C} = \mathbb{R}^n$ and the functions $f, g_1, \cdots, g_m$ are continuously differentiable and convex. The problem*

$$(WD) \quad \sup_{x,y} \left\{ f(x) + \sum_{j=1}^m y_j g_j(x) \right\}$$
$$\nabla f(x) + \sum_{j=1}^m y_j \nabla g_j(x) = 0,$$
$$y \geq 0, \; x \in \mathbb{R}^n,$$

is called the Wolfe Dual of the convex optimization problem (CO).

Note that the variables in (WD) are both $y \geq 0$ and $x \in \mathbb{R}^n$, and that the Lagrangian $L(x, y)$ is the objective function of (WD). For this reason, the Wolfe dual does not have a concave objective function in general, but it is still very useful tool, as we will see. In particular, if the Lagrange function has a saddle point, $\mathcal{C} = \mathbb{R}^n$ and the functions $f, g_1, \cdots, g_m$ are continuously differentiable and convex, then the two dual problems are equivalent. Using the results of the previous section one easily proves weak and strong duality results, as we will now show. A more detailed discussion of duality theory can be found in [2, 28].

Theorem 3.9 (Weak duality for the Wolfe dual) *Assume that $\mathcal{C} = \mathbb{R}^n$ and the functions $f, g_1, \cdots, g_m$ are continuously differentiable and convex. If $\hat{x}$ is a feasible solution of (CO) and $(\overline{x}, \overline{y})$ is a feasible solution for (WD) then*

$$L(\overline{x}, \overline{y}) \leq f(\hat{x}).$$

In other words, weak duality holds for (CO) and (WD).

Proof: Let $(\overline{x}, \overline{y})$ be a feasible solution for (WD). Since the functions f and $g_1, \ldots, g_m$ are convex and continuously differentiable, and $\overline{y} \geq 0$, the function

$$h(x) := f(x) + \sum_{j=1}^m \overline{y}_j g_j(x)$$

must also be convex and continuously differentiable (see Lemma 1.40). Since $(\overline{x}, \overline{y})$ is feasible for (WD), one has

$$\nabla h(\overline{x}) = \nabla f(\overline{x}) + \sum_{j=1}^m \overline{y}_j \nabla g_j(\overline{x}) = 0.$$

This means that $\overline{x}$ is a minimizer of the function h, by Lemma 2.6. In other words

$$f(\overline{x}) + \sum_{j=1}^{m} \overline{y}_j g_j(\overline{x}) \le f(x) + \sum_{j=1}^{m} \overline{y}_j g_j(x) \quad \forall x \in \mathbb{R}^n. \tag{3.1}$$

Let $\hat{x}$ be an arbitrary feasible solution of (CO). Setting $x = \hat{x}$ in (3.1) one gets

$$f(\overline{x}) + \sum_{j=1}^{m} \overline{y}_j g_j(\overline{x}) \le f(\hat{x}) + \sum_{j=1}^{m} \overline{y}_j g_j(\hat{x}) \le f(\hat{x}),$$

where the last inequality follows from $\overline{y} \ge 0$ and $g_j(\hat{x}) \le 0$ $(j = 1, \ldots, m)$. This completes the proof. □

Theorem 3.10 (Strong duality for the Wolfe dual) *Assume that $\mathcal{C} = \mathbb{R}^n$ and the functions $f, g_1, \cdots, g_m$ are continuously differentiable and convex. Also assume that (CO) satisfies the Slater regularity condition. Let $\overline{x}$ be a feasible solution of (CO). Then $\overline{x}$ is an optimal solution of (CO) if and only if there exists a $\overline{y} \ge 0$ such that $(\overline{x}, \overline{y})$ is an optimal solution of (WD).*

Proof: Follows directly from Corollary 2.33. □

Warning! Remember, we are only allowed to form the Wolfe dual of a nonlinear optimization problem if it is a *convex* optimization problem. We may replace the infimum in the definition of $\psi(y)$ by the condition that the x-gradient is zero only if all the functions f and g_j, $\forall j$ are convex and if we know that the infimum is attained. Else, the condition

$$\nabla f(x) + \sum_{j=1}^{m} y_j \nabla g_j(x) = 0$$

allows solutions which are possibly maxima, saddle points or inflection points, or it may not have any solution. In such cases no duality relation holds in general. For nonconvex problems one has to work with the Lagrange dual.

Example 3.11 [Wolfe dual] Let us consider the convex optimization problem

$$\begin{array}{llrcl} \text{(CO)} & \min & x_1 + e^{x_2} & & \\ & \text{s.t.} & 3x_1 - 2e^{x_2} & \ge & 10 \\ & & x_2 & \ge & 0 \\ & & x & \in & \mathbb{R}^2. \end{array}$$

Then the optimal value is 5 with $x = (4, 0)$. Note that the Slater condition holds for this example.

Wolfe dual The Wolfe dual of (CO) is given by

$$\begin{array}{llr} \text{(WD)} & \sup & x_1 + e^{x_2} + y_1(10 - 3x_1 + 2e^{x_2}) - y_2 x_2 \\ & \text{s.t.} & 1 - 3y_1 = 0 \\ & & e^{x_2} + 2e^{x_2} y_1 - y_2 = 0 \\ & & x \in \mathbb{R}^2, y \ge 0, \end{array}$$

which is a non-convex problem. The first constraint gives $y_1 = \frac{1}{3}$, and thus the second constraint becomes

$$\frac{5}{3} e^{x_2} - y_2 = 0.$$

Now we can eliminate y_1 and y_2 from the object function. We get the function

$$f(x_2) = \frac{5}{3} e^{x_2} - \frac{5}{3} x_2 e^{x_2} + \frac{10}{3}.$$

This function has a maximum when

$$f'(x_2) = -\frac{5}{3} x_2 e^{x_2} = 0,$$

which is only true when $x_2 = 0$ and $f(0) = 5$. Hence the optimal value of (WD) is 5 and then $(x, y) = (4, 0, \frac{1}{3}, \frac{5}{3})$.

Lagrange dual We can double check this answer by using the Lagrange dual. Let

$$\begin{aligned}\psi(y) &= \inf_{x\in\mathbb{R}^2}\{x_1 + e^{x_2} + y_1(10 - 3x_1 + 2e^{x2}) - y_2x_2\} \\ &= \inf_{x_1\in\mathbb{R}}\{x_1 - 3y_1x_1\} + \inf_{x_2\in\mathbb{R}}\{(1+2y_1)e^{x_2} - y_2x_2\} + 10y_1.\end{aligned}$$

We have

$$\inf_{x_1\in\mathbb{R}}\{x_1 - 3y_1x_1\} = \begin{cases} 0 & \text{for } y_1 = \frac{1}{3} \\ -\infty & \text{otherwise.}\end{cases}$$

Now, for fixed y_1, y_2, with $y_2 > 0$ let

$$g(x_2) = (1+2y_1)e^{x_2} - y_2x_2.$$

Then g has a minimum when

$$g'(x_2) = (1+2y_1)e^{x_2} - y_2 = 0,$$

i.e., when $x_2 = \log\left(\frac{y_2}{1+2y_1}\right)$. Further, $g(\log\left(\frac{y_2}{1+2y_1}\right)) = y_2 - y_2\log\left(\frac{y_2}{1+2y_1}\right)$. Hence, we have

$$\inf_{x_2\in\mathbb{R}}\{(1+2y_1)e^{x_2} - y_2x_2\} = \begin{cases} y_2 - y_2\log\left(\frac{y_2}{1+2y_1}\right) & \text{for } y_2 > 0 \\ 0 & \text{for } y_2 = 0.\end{cases}$$

Thus the Lagrange dual becomes

$$\begin{aligned}\text{(LD)} \qquad \sup\psi(y) &= 10y_1 + y_2 - y_2\log\left(\frac{y_2}{1+2y_1}\right) \\ \text{s.t.} \qquad y_1 &= \frac{1}{3} \\ y_2 &\geq 0.\end{aligned}$$

Now we have

$$\frac{d}{dy_2}\psi(\frac{1}{3}, y_2) = \log(\frac{3y_2}{5}) = 0$$

when $y_2 = \frac{5}{3}$, and $\psi(\frac{1}{3}, \frac{5}{3}) = 5$.

*

Exercise 3.2 *Prove that — under the assumptions of Theorem 3.10 — the Lagrange and Wolfe duals of the optimization problem (CO) are equivalent.* ◁

Exercise 3.3 *We wish to design a rectangular prism (box) with length l, width b, and height h such that the volume of the box is at least V units, and the total surface area is minimal. This problem has the following (nonconvex) formulation:*

$$\min_{l,b,h} 2(lb + bh + lh), \quad lbh \geq V, \quad l, b, h > 0. \tag{3.2}$$

i) Transform the problem (3.2) by introducing new variables to obtain:

$$\min_{x_1,x_2,x_3} 2(e^{x_1+x_2} + e^{x_2+x_3} + e^{x_1+x_3}), \quad x_1 + x_2 + x_3 \geq \ln(V), \quad x_1, x_2, x_3 \in \mathbb{R}. \tag{3.3}$$

ii) Show that the transformed problem is convex and satisfies Slater's regularity condition.

iii) Show that the Lagrange dual of problem (3.3) is:

$$\max_{\lambda\geq 0}\left(\frac{3}{2} + \ln(V)\right)\lambda - \frac{3}{2}\lambda\ln\left(\frac{\lambda}{4}\right). \tag{3.4}$$

iv) Show that the Wolfe dual of problem (3.3) is the same as the Lagrange dual.

v) Use the KKT conditions of problem (3.3) to show that the cube ($l = b = h = V^{1/3}$) is the optimal solution of problem (3.2).

vi) Use the dual problem (3.4) to derive the same result as in part v).

◁

3.3 Examples for dual problems

In this section we derive the Lagrange and/or the Wolfe dual of some specific convex optimization problems.

Linear optimization

Let $A : m \times n$ be a matrix, $b \in \mathbb{R}^m$ and $c, x \in \mathbb{R}^n$. The primal Linear Optimization (LO) problem is given as

$$(\text{LO}) \qquad \min\{c^T x \mid Ax = b, \;\; x \geq 0\}.$$

Here we can say that $\mathcal{C} = \mathbb{R}^n$. Obviously all the constraints are continuously differentiable. The inequality constraints can be given as $g_j(x) = (a^j)^T x - b_j$ if $j = 1, \cdots, m$ and $g_j(x) = (-a^{j-m})^T x + b_{j-m}$ if $j = m+1, \cdots, 2m$ and finally $g_j(x) = -x_{j-2m}$ if $j = 2m+1, \cdots, 2m+n$. Here a^j denotes the jth row of matrix A. Denoting the Lagrange multipliers by y^-, y^+ and s respectively the Wolfe dual (WD) of (LO) has the following form:

$$\begin{aligned} \max \quad & c^T x + (y^-)^T (Ax - b) + (y^+)^T (-Ax + b) + s^T(-x) \\ & c + A^T y^- - A^T y^+ - s = 0, \\ & y^- \geq 0, \;\; y^+ \geq 0, \;\; s \geq 0. \end{aligned}$$

As we substitute $c = -A^T y^- + A^T y^+ + s$ in the objective and introduce the notation $y = y^+ - y^-$ the standard dual linear optimization problem follows.

$$\begin{aligned} \max \quad & b^T y \\ & A^T y + s = c, \\ & s \geq 0. \end{aligned}$$

Quadratic optimization

The quadratic optimization problem is considered in the symmetric form. Let $A : m \times n$ be a matrix, $Q : n \times n$ be a positive semi-definite symmetric matrix, $b \in \mathbb{R}^m$ and $c, x \in \mathbb{R}^n$. The primal Quadratic Optimization (QO) problem is given as

$$(\text{QO}) \qquad \min\{c^T x + \frac{1}{2} x^T Q x \mid Ax \geq b, \;\; x \geq 0\}.$$

Here we can say that $\mathcal{C} = \mathbb{R}^n$. Obviously all the constraints are continuously differentiable. The inequality constraints can be given as $g_j(x) = (-a^j)^T x + b_j$ if $j = 1, \cdots, m$ and $g_j(x) = -x_{j-m}$ if $j = m+1, \cdots, m+n$. Denoting the Lagrange multipliers by y and s respectively the Wolfe dual (WD) of (QO) has the following form:

$$\begin{aligned} \max \quad & c^T x + \tfrac{1}{2} x^T Q x + y^T(-Ax + b) + s^T(-x) \\ & c + Qx - A^T y - s = 0, \\ & y \geq 0, \;\; s \geq 0. \end{aligned}$$

As we substitute $c = -Qx + A^T y + s$ in the objective the dual quadratic optimization problem follows.

$$\begin{aligned} \max \quad & b^T y - \tfrac{1}{2} x^T Q x \\ & -Qx + A^T y + s = c, \\ & y \geq 0, \;\; s \geq 0. \end{aligned}$$

Observe that the vector x occurring in this dual is not necessarily feasible for (QO)! To eliminate the x variables another form of the dual can be presented.

Since Q is a positive semidefinite symmetric matrix, it can be represented as the product of two matrices $Q = D^T D$ (use e.g. Cholesky factorization), one can introduce the vector $z = Dx$. Hence the following (QD) dual problem is obtained:

$$\begin{aligned} \max \qquad & b^T y - \tfrac{1}{2} z^T z \\ & -D^T z + A^T y + s = c, \\ & y \ge 0, \;\; s \ge 0. \end{aligned}$$

Note that the optimality conditions are $x^T s = 0,\;\; y^T(Ax - b) = 0$ and $z = Dx$.

Constrained maximum likelihood estimation

Maximum Likelihood Estimation frequently occurs in statistics. This problem can also be used to illustrate duality in convex optimization. In this problem we are given a finite set of sample points $x_i, \quad (1 \le i \le n)$. The most probable density values at the sample points are to be determined that satisfy some linear (e.g. convexity) constraints. Formally, the problem is defined as one has to determine the maximum of the Likelihood function $\Pi_{i=1}^n x_i$ under the conditions

$$Ax \ge 0, \;\; d^T x = 1, \;\; x \ge 0.$$

Here $Ax \ge 0$ represents the linear constraints, the density values x_i are nonnegative and the condition $d^T x = 1$ ensures that the (approximate) integral of the density function is one. Since the logarithm function is monotone the objective can equivalently replaced by

$$\min \quad -\sum_{i=1}^n \ln x_i.$$

It is easy to check that the so defined problem is a convex optimization problem. Again we can take $\mathcal{C} = \mathbb{R}^n$ and all the constraints are linear, hence continuously differentiable. Denoting the Lagrange multipliers by $y \in \mathbb{R}^m$, $t \in \mathbb{R}$ and $s \in \mathbb{R}^n$ respectively the Wolfe dual (WD) of this problem has the following form:

$$\begin{aligned} \max \quad & -\sum_{i=1}^n \ln x_i + y^T(-Ax) + t(d^T x - 1) + s^T(-x) \\ & -X^{-1}e - A^T y + td - s = 0, \\ & y \ge 0, \;\; s \ge 0. \end{aligned}$$

Here the notation $e = (1, \cdots, 1) \in \mathbb{R}^n$ and $X = \text{diag}(x)$ is used. Also note that for simplicity we did not split the equality constraint into two inequalities but we used immediately that its multiplier is a free variable. Multiplying the first constraint by x^T one has

$$-x^T X^{-1} e - x^T A^T y + t x^T d - x^T s = 0.$$

Using $d^T x = 1$, $x^T X^{-1} e = n$ and the optimality conditions $y^T Ax = 0$, $x^T s = 0$ we have

$$t = n.$$

Observe further that due to the logarithm in the primal objective, the primal optimal solution is necessarily strictly positive, hence the dual variable s must be zero at the optimum. Combining these results the dual problem is

$$\begin{aligned} \max \qquad & -\sum_{i=1}^n \ln x_i \\ & X^{-1}e + A^T y = nd, \\ & y \ge 0. \end{aligned}$$

Eliminating the variables $x_i > 0$ from the constraints one has $x_i = \frac{1}{nd_i - a_i^T y}$ and $-\ln x_i = \ln(nd_i - a_i^T y)$ for all $i = 1, \cdots, n$. Now we have the final form of our dual problem:

$$\max \quad \sum_{i=1}^{n} \ln(nd_i - a_i^T y)$$
$$A^T y \leq nd,$$
$$y \geq 0.$$

3.4 Some examples with positive duality gap

Example 3.12 This example is due to Duffin. It shows that positive duality gap might occur for convex problems when the problem does not satisfy the Slater regularity condition. Moreover, it makes clear that the Wolfe dual might be significantly weaker than the Lagrange dual.

Let us consider the convex optimization problem

$$\begin{array}{lll} \text{(CO)} & \min & e^{-x_2} \\ & \text{s.t.} & \sqrt{x_1^2 + x_2^2} - x_1 \leq 0 \\ & & x \in \mathbb{R}^2. \end{array}$$

The feasible region is $\mathcal{F} = \{x \in \mathbb{R}^2 | \; x_1 \geq 0, x_2 = 0\}$. The only constraint is non-linear and singular, thus (CO) is not Slater regular. The optimal value of the object function is 1.

The Lagrange function is given by

$$L(x, y) = e^{-x_2} + y(\sqrt{x_1^2 + x_2^2} - x_1).$$

Let us first consider the Wolfe dual (WD):

$$\begin{array}{rcl} \sup e^{-x_2} + y(\sqrt{x_1^2 + x_2^2} - x_1) & & \\ -y + y\frac{x_1}{\sqrt{x_1^2+x_2^2}} & = & 0 \\ -e^{-x_2} + y\frac{x_2}{\sqrt{x_1^2+x_2^2}} & = & 0 \\ y & \geq & 0. \end{array}$$

The first constraint imply that $x_2 = 0$ and $x_1 \geq 0$, but these values do not satisfy the second constraint. Thus the Wolfe dual is infeasible, yielding an infinitely large duality gap.

Let us see if we can do better by using the Lagrange dual. Now, let $\epsilon = \sqrt{x_1^2 + x_2^2} - x_1$, then

$$x_2^2 - 2\epsilon x_1 - \epsilon^2 = 0.$$

Hence, for any $\epsilon > 0$ we can find $x_1 > 0$ such that $\epsilon = \sqrt{x_1^2 + x_2^2} - x_1$ even if x_2 goes to infinity. However, when x_2 goes to infinity e^{-x_2} goes to 0. So,

$$\psi(y) = \inf_{x \in \mathbb{R}^2} e^{-x_2} + y\left(\sqrt{x_1^2 + x_2^2} - x_1\right) = 0,$$

thus the optimal value of the Lagrange dual

$$\begin{array}{lll} \text{(LD)} & \max & \psi(y) \\ & \text{s.t.} & y \geq 0 \end{array}$$

is 0. This gives a nonzero duality gap that equals to 1.

Observe that the Wolfe dual becomes infeasible because the infimum in the definition of $\psi(y)$ exists, but it is not attained. $*$

Example 3.13 [Basic model with zero duality gap] Let us first consider the following simple convex optimization problem.

$$\begin{array}{rcl} \min & x_1 & \\ \text{s.t.} & x_1^2 & \leq 0 \\ & -x_2 & \leq 0 \\ & -1 - x_1 & \leq 0. \end{array} \tag{3.5}$$

Here the convex set $\mathcal{C}$ where the above functions are defined is $\mathbb{R}^2$. It is clear that the set of feasible solutions is given by

$$\mathcal{F} = \{(x_1, x_2) \,|\, x_1 = 0, \;\; x_2 \geq 0\},$$

thus any feasible vector $(x_1, x_2) \in \mathcal{F}$ is optimal and the optimal value of this problem is 0. Because $x_1 = 0$ for all feasible solutions the Slater regularity condition does not hold for (3.5).

Let us make the Lagrange dual of (3.5). The Lagrange multipliers (y_1, y_2, y_3) are nonnegative and the Lagrange function

$$L(x, y) = x_1 + y_1 x_1^2 - y_2 x_2 - y_3(1 + x_1)$$

is defined on $x \in \mathbb{R}^2$ and $y \in \mathbb{R}^3$, $y \geq 0$.

The Lagrange dual is defined as

$$\begin{array}{rl} \max & \psi(y) \\ \text{s.t.} & y \;\geq\; 0. \end{array} \tag{3.6}$$

where

$$\begin{aligned} \psi(y) &= \inf_{x \in \mathbb{R}^2} \{x_1 + y_1 x_1^2 - y_2 x_2 - y_3(1 + x_1)\} \\ &= \inf_{x \in \mathbb{R}^2} \{x_1(1 - y_3) + y_1 x_1^2 - y_2 x_2 - y_3\} \\ &= \begin{cases} -\infty & \text{if} \quad y_2 \neq 0 \text{ or } y_1 = 0 \text{ but } y_3 \neq 1; \\ 0 & \text{if} \quad y_2 = 0,\;\; y_1 = 0 \text{ and } y_3 = 1; \\ -y_3 - \dfrac{(1 - y_3)^2}{4y_1} & \quad\;\; y_2 = 0 \text{ and } y_1 \neq 0. \end{cases} \end{aligned}$$

The last expression in the formula above is obtained by minimizing the convex quadratic function $x_1(1-y_3)+y_1x_1^2-y_3$ where y_1 and y_3 are fixed. Because this last expression is nonpositive, the maximum of $\psi(y)$ is zero. Thus for this problem both the primal and the dual problems have optimal solutions with equal (zero) optimal objective values. ∗

Example 3.14 [A variant with positive duality gap] Let us consider the same problem as in the previous example (see problem (3.5)) with a different representation of the feasible set. As we will see the new formulation results in a quite different dual. The new dual has also an optimal solution but now the duality gap is positive.

$$\begin{array}{rrcl} & \min \;\; x_1 & & \\ \text{s.t.} & x_0 - s_0 & = & 0 \\ & x_1 - s_1 & = & 0 \\ & x_2 - s_2 & = & 0 \\ & 1 + x_1 - s_3 & = & 0 \\ & x_0 & = & 0 \\ & x \in \mathbb{R}^3, s \in \mathcal{C}. & & \end{array} \tag{3.7}$$

Note that (3.7) has the correct form: the constraints are linear, hence convex, and the vector (x, s) of the variables belong to the convex set $\mathbb{R}^3 \times \mathcal{C}$. Here the convex set $\mathcal{C}$ is defined as follows:

$$\mathcal{C} = \{s = (s_0, s_1, s_2, s_3) \,|\, s_0 \geq 0, \;\; s_2 \geq 0, \;\; s_3 \geq 0, \;\; s_0 s_2 \geq s_1^2\}.$$

It is clear that the set of feasible solutions is

$$\mathcal{F} = \{(x, s) \,|\, x_0 = 0, \;\; x_1 = 0, \;\; x_2 \geq 0, \;\; s_0 = 0, \;\; s_1 = 0, \;\; s_2 \geq 0, \;\; s_3 = 1\},$$

thus any feasible vector $(x, s) \in \mathcal{F}$ is optimal and the optimal value of this problem is 0.

Exercise 3.4 *1. Prove that the function $s_1^2 - s_0 s_2$ is not convex.*

2. Prove that the set $\mathcal{C}$ is convex.

3. Prove that problem (3.7) does not satisfy the Slater regularity condition.

◁

Due to the equality constraints the Lagrange multipliers $(y_0, y_1, y_2, y_3, y_4)$ are free and the Lagrange function

$$L(x, s, y) = x_1 + y_0(x_0 - s_0) + y_1(x_1 - s_1) + y_2(x_2 - s_2) + y_3(1 + x_1 - s_3) + +y_4 x_0)$$

is defined for $x \in \mathbb{R}^3$, $s \in \mathcal{C}$ and $y \in \mathbb{R}^5$.

The Lagrange dual is defined as

$$\begin{array}{rl} \max & \psi(y) \\ \text{s.t.} & y \in \mathbb{R}^5 \end{array} \tag{3.8}$$

where

$$\begin{aligned}
\psi(y) &= \inf_{x\in \mathbb{R}^3,\ s\in\mathcal{C}} L(x,s,y) \\
&= \inf_{x\in \mathbb{R}^3,\ s\in\mathcal{C}} \{x_1(1+y_1+y_3)+x_0(y_4+y_0)+x_2y_2-s_0y_0-s_1y_1-s_2y_2-s_3y_3+y_3\} \\
&= \begin{cases} y_3 & \text{if } \ 1+y_1+y_3=0,\ y_4+y_0=0,\ y_2=0,\ y_3\le 0,\ y_0\le 0,\ y_1=0; \\ -\infty & \text{otherwise.}\end{cases}
\end{aligned}$$

The last equality requires some explanation.

- If $1+y_1+y_3\neq 0$ then $L(x,s,y)=x_1(1+y_1+y_3)+y_3$ for $x_0=x_2=0,\ s=0$. So $\inf L(x,s,y)=-\infty$ in this case.
- If $y_4+y_0\neq 0$ then $L(x,s,y)=x_0(y_4+y_0)+y_3$ for $x_1=x_2=0,\ s=0$. So $\inf L(x,s,y)=-\infty$ in this case.
- If $y_2\neq 0$ then $L(x,s,y)=x_2y_2+y_3$ for $x_0=x_1=0,\ s=0$. So $\inf L(x,s,y)=-\infty$ in this case.
- If $y_0>0$ then $L(x,s,y)=-s_0y_0+y_3$ for $x=0,\ s_0\ge 0,\ s_1=0,\ s_2=0,\ s_3=0$. So $\inf L(x,s,y)=-\infty$ in this case.
- If $y_3>0$ then $L(x,s,y)=-s_3y_3+y_3$ for $x=0,\ s_0=0,\ s_1=0,\ s_2=0,\ s_3\ge 0$. So $\inf L(x,s,y)=-\infty$ in this case.
- If $y_2=0$ but $y_1\neq 0$ then $L(x,s,y)=-\frac{1}{\tau}y_0-\frac{y_1}{|y_1|}\tau y_1+y_3$ for $x=0,\ s_3=0$ and $(s_0,s_1,s_2,s_3)=(\frac{1}{\tau},\frac{y_1}{|y_1|}\tau,\tau^2,0)\in\mathcal{C}$. So $\inf L(x,s,y)=-\infty$ (one obtains this a let $\tau\to\infty$) in this case.

Summarizing the above results we conclude that the Lagrange dual reduces to

$$\begin{array}{c} \max \ \ y_3 \\ y_0\le 0,\ y_1=0,\ y_2=0,\ y_3=-1,\ y_4=-y_0. \end{array}$$

Here for any feasible solution $y_3=-1$, thus the optimal value of the Lagrange dual is -1, i.e. both the primal problem (3.7) and its dual (3.8) have optimal solutions, but their optimal values are not equal. $*$

Exercise 3.5 *Modify the above problem so that for a given $\gamma>0$ the nonzero duality gap at optimum will be equal to γ.* $\triangleleft$

Example 3.15 [Duality for non convex problems 1] Let us consider the non-convex optimization problem

$$\begin{array}{lll} \text{(NCO)} & \min & x_1^2-2x_2 \\ & \text{s.t.} & x_1^2+x_2^2=4 \\ & & x\in\mathbb{R}^2. \end{array}$$

Then the optimal value is -4, with $x=(0,2)$.

Lagrange dual The Lagrange function of (NCO) is given by

$$L(x,y)=x_1^2-2x_2+y(x_1^2+x_2^2-4),\quad \text{where } y\in\mathbb{R},$$

and then

$$\begin{aligned}
\psi(y) &= \inf_{x\in\mathbb{R}^2} L(x,y) \\
&= \inf_{x_1}\{(1+y)x_1^2\}+\inf_{x_2}\{yx_2^2-2x_2\}-4y.
\end{aligned}$$

We have

$$\begin{aligned}
\inf_{x_1}\{(1+y)x_1^2\} &= \begin{cases} 0 & \text{for } y\ge -1 \\ -\infty & \text{for } y<-1 \end{cases} \\
\inf_{x_2}\{yx_2^2-2x_2\} &= \begin{cases} -\frac{1}{y} & \text{for } y>0 \\ -\infty & \text{for } y\le 0. \end{cases}
\end{aligned}$$

Hence, the Lagrange dual is

$$\begin{array}{lll} \text{(LD)} & \sup & -\dfrac{1}{y}-4y \\ & & y>0, \end{array}$$

which is a convex problem, and the optimal value is -4, with $y=\frac{1}{2}$. Note that although the problem is not convex, and does not satisfy the Slater regularity condition, the duality gap is zero.

*

Example 3.16 [Duality for non convex problems 2] Let us consider the non-convex optimization problem

$$\begin{array}{lll} \text{(CLO)} & \min & x_1^2 - x_2^2 \\ & \text{s.t.} & x_1 + x_2 \le 2 \\ & & x \in \mathcal{C} = \{x \in \mathbb{R}^2 |\ -2 \le x_1, x_2 \le 4\}. \end{array}$$

Then we have the optimal value -12 with $x = (-2, 4)$. The Lagrange function of (CLO) is given by

$$L(x, y) = x_1^2 - x_2^2 + y(x_1 + x_2 - 2), \qquad \text{where} \quad y \ge 0.$$

Thus for $y \ge 0$ we have

$$\begin{aligned} \psi(y) &= \inf_{x \in \mathcal{C}} L(x, y) \\ &= \inf_{-2 \le x_1 \le 4} \{x_1^2 + y x_1\} + \inf_{-2 \le x_2 \le 4} \{-x_2^2 + y x_2\} - 2y, \end{aligned}$$

Now, $x_1^2 + yx_1$ is a parabola which has its minimum at $x_1 = -\frac{y}{2}$. So, this minimum lies within $\mathcal{C}$ when $y \le 4$. When $y \ge 4$ the minimum is reached at the boundary of $\mathcal{C}$. The minimum of the parabola $-x_2^2 + yx_2$ is always reached at the boundaries of $\mathcal{C}$, at $x_2 = -2$ when $y \ge 2$, and at $x_2 = 4$ when $y \le 2$. Hence, we have

$$\psi(y) = \begin{cases} -\frac{y^2}{4} + 2y - 16 & \text{for } y \le 2 \\ -\frac{y^2}{4} - 4y - 4 & \text{for } 2 \le y \le 4 \\ -6y & \text{for } y \ge 4. \end{cases}$$

Maximizing $\psi(y)$ for $y \ge 0$ gives

$$\begin{aligned} \sup_{0 \le y \le 2} \psi(y) &= -13, \\ \sup_{2 \le y \le 4} \psi(y) &= -13, \\ \sup_{y \ge 4} \psi(y) &= -24. \end{aligned}$$

Hence, the optimal value of the Lagrange dual is -13, and we have a nonzero duality gap that equals to 1. *

3.5 Semidefinite optimization

The Primal and the Dual Problem

Let $A_0, A_1, \cdots, A_n \in \mathbb{R}^{m \times m}$ be symmetric matrices. Further let $c \in \mathbb{R}^n$ be a given vector and $x \in \mathbb{R}^n$ be the vector of unknowns in which the optimization is done. The *primal semidefinite optimization problem* is defined as

$$\begin{array}{lll} (PSO) & \min & c^T x \\ & \text{s.t.} & -A_0 + \sum_{k=1}^{n} A_k x_k \succeq 0, \end{array} \tag{3.9}$$

where $\succeq 0$ indicates that the left hand side matrix has to be positive semidefinite. It is clear that the primal problem (PSO) is a convex optimization problem since the convex combination of positive semidefinite matrices is also positive semidefinite. For convenience the notation

$$F(x) = -A_0 + \sum_{k=1}^{n} A_k x_k$$

will be used.

The *dual problem of the semidefinite optimization problem*, as given e.g. in [44], is as follows:

$$\begin{array}{llll} (DSP) & \max & \text{Tr}(A_0 Z) & \\ & \text{s.t.} & \text{Tr}(A_k Z) = c_k, & \text{for all} \quad k = 1, \cdots, n, \\ & & Z \succeq 0, & \end{array} \tag{3.10}$$

where $Z \in \mathbb{R}^{m\times m}$ is the matrix of variables. Again, the dual of the semidefinite optimization problem is convex. The trace of a matrix is a linear function of the matrix and the convex combination of positive semidefinite matrices is also positive semidefinite.

Theorem 3.17 *(Weak duality) If $x \in \mathbb{R}^n$ is primal feasible and $Z \in \mathbb{R}^{m\times m}$ is dual feasible, then*

$$c^T x \geq Tr(A_0 Z)$$

with equality if and only if

$$F(x)Z = 0.$$

Proof: Using the dual constraints and some elementary properties of the trace of matrices one may write

$$c^T x - \text{Tr}(A_0 Z) = \sum_{k=1}^{n} \text{Tr}(A_k Z)x_k - \text{Tr}(A_0 Z) = \text{Tr}((\sum_{k=1}^{n} A_k x_k - A_0)Z) = \text{Tr}(F(x)Z) \geq 0.$$

Here the last inequality holds because both matrices $F(x)$ and Z are positive semidefinite. Equality holds if and only if $F(x)Z = 0$, which completes the proof. □

The Dual as Lagrange–Wolfe Dual

First we give another equivalent form of the (PSO) problem in order to be able to derive the dual problem more easily. Clearly problem (PSO) can equivalently be given as

$$(PSO') \qquad \begin{aligned} \min \quad & c^T x \\ \text{s.t.} \quad & -F(x) + S = 0 \\ & S \succeq 0, \end{aligned} \tag{3.11}$$

where $S \in \mathbb{R}^{m\times m}$ is a symmetric matrix. It plays the role of the usual "slack variables". The Lagrange function $L(x, S, Z)$ of problem (PSO') is defined on the set $\{ (x, S, Z) \,|\, x \in \mathbb{R}^n,\ S \in \mathbb{R}^{m\times m},\ S \succeq 0,\ Z \in \mathbb{R}^{m\times m}, \}$ and is given by

$$L(x, S, Z) = c^T x - e^T (F(x) \circ Z)e + e^T (S \circ Z)e,$$

where $e^T = (1, \cdots, 1) \in \mathbb{R}^n$ and $X \circ Z$ denotes the Minkowski (coordinatewise) product of matrices. Before going on we observe that $e^T (S\circ Z)e = \text{Tr}(SZ)$, hence the Lagrange function can be reformulated as

$$L(x, S, Z) = c^T x - \sum_{k=1}^{n} x_k \text{Tr}(A_k Z) + \text{Tr}(A_0 Z) + \text{Tr}(SZ). \tag{3.12}$$

Before formulating the Lagrange dual of (PSO') note that we can assume that the matrix Z is symmetric, since $F(x)$ is symmetric. The Lagrange dual of problem (PSO') is

$$(DSDL) \qquad \max \{\psi(Z) \mid Z \in \mathbb{R}^{m\times m}\} \tag{3.13}$$

where

$$\psi(Z) = \min\{ L(x, S, Z) \,|\, x \in \mathbb{R}^n,\ S \in \mathbb{R}^{m\times m},\ S \succeq 0\}. \tag{3.14}$$

As we did in deriving the Wolfe dual, one easily derives optimality conditions to calculate $\psi(Z)$. Since the minimization in (3.14) is done in the free variable x, the positive semidefinite matrix of variables S and, further the function $L(x, S, Z)$ is separable w.r.t. x and S we can take these minimums separately.

If we minimize in S all the terms in (3.12) but $\text{Tr}(SZ)$ are constant. The matrix S is positive semidefinite, hence

$$\min_S \text{Tr}(SZ) = \begin{cases} 0 & \text{if } Z \succeq 0, \\ -\infty & \text{otherwise.} \end{cases} \tag{3.15}$$

If we minimize (3.14) in x, we need to equate the $x-$gradient of $L(x, S, Z)$ to zero (remember to the Wolfe dual). This requirement leads to

$$c_k - \mathrm{Tr}(A_k Z) = 0 \quad \text{for all} \quad k = 1, \cdots, n. \tag{3.16}$$

Multiplying the equations of (3.16) by x_k and summing up one obtains

$$c^T x - \sum_{k=1}^{n} x_k \mathrm{Tr}(A_k Z) = 0.$$

By combining the last formula and the results presented in (3.15) and in (3.16) the simplified form of the Lagrange dual (3.13), the Lagrange–Wolfe dual

$$\begin{array}{lll} (DSO) & \max & \mathrm{Tr}(A_0 Z) \\ & \text{s.t.} & \mathrm{Tr}(A_k Z) = c_k, \quad \text{for all} \quad k = 1, \cdots, n, \\ & & Z \succeq 0, \end{array}$$

follows. The reader readily verifies that this is identical to (3.10).

Exercise 3.6 *Consider the problem given in Example 3.13.*

1. *Prove that problem (3.7) is a semidefinite optimization problem.*
2. *Give the dual semidefinite problem.*
3. *Prove that there is a positive duality gap for this primal–dual semidefinite optimization pair.*

◁

3.6 Duality in cone-linear optimization

In this section we deal with *cone-linear optimization* problems. A cone-linear optimization problem is a natural generalization of the well known standard linear optimization problem

$$\min \{ c^T x \mid Ax \geq b, \; x \geq 0 \},$$

where $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and $c \in \mathbb{R}^n$. The inequality conditions can be reformulated by observing that the conditions $Ax \geq b$ and $x \geq 0$ mean that the vector $Ax - b$ and x should be in the positive orthant

$$\mathbb{R}^m_+ := \{ x \in \mathbb{R}^m \mid x \geq 0 \}$$

and $\mathbb{R}^n_+$, respectively. One observes, that the positive orthants $\mathbb{R}^m_+$ and $\mathbb{R}^n_+$ are convex cones, i.e. the linear optimization problem can be restated as the following cone-linear optimization problem

$$\begin{array}{rcl} \min c^T x & & \\ Ax - b & \in & \mathbb{R}^m_+ \\ x & \in & \mathbb{R}^n_+. \end{array}$$

The dual problem

$$\max \{ b^T y \mid A^T y \leq c, \; y \geq 0 \},$$

can similarly be reformulated in the conic form:

$$\begin{array}{rcl} \max b^T y & & \\ c - A^T y & \in & \mathbb{R}^n_+ \\ y & \in & \mathbb{R}^m_+. \end{array}$$

The natural question arises: how one can derive dual problems for general cone-linear optimization problems where, in the above given formulation the simple polyhedral convex cones $\mathbb{R}^m_+$ and $\mathbb{R}^n_+$ are replaced by arbitrary convex cones $\mathcal{C}_1 \subseteq \mathbb{R}^m$ and $\mathcal{C}_2 \subseteq \mathbb{R}^n$. The *cone-linear optimization* problem is defined as follows:

$$\begin{array}{rcl} \min\ c^T x & & \\ Ax - b & \in & \mathcal{C}_1 \\ x & \in & \mathcal{C}_2. \end{array} \tag{3.17}$$

The Dual of a Cone-linear Problem

First, by introducing slack variables s, we give another equivalent form of the cone-linear problem (3.17)

$$\begin{array}{rcl} \min\ c^T x & & \\ s - Ax + b & = & 0 \\ s & \in & \mathcal{C}_1 \\ x & \in & \mathcal{C}_2. \end{array}$$

In this optimization problem we have linear equality constraints $s - Ax + b = 0$ and the vector (s, x) must be in the convex cone

$$\mathcal{C}_1 \times \mathcal{C}_2 := \{\, (s, x) \mid s \in \mathcal{C}_1,\ x \in \mathcal{C}_2 \,\}.$$

The Lagrange function $L(s, x, y)$ of the above problem is defined on the set

$$\{\, (s, x, y) \mid s \in \mathcal{C}_1,\ x \in \mathcal{C}_2,\ y \in \mathbb{R}^m \,\}$$

and is given by

$$L(s, x, y) = c^T x + y^T(s - Ax + b) = b^T y + s^T y + x^T(c - A^T y). \tag{3.18}$$

Hence, the Lagrange dual of the cone-linear problem is given by

$$\max_{y \in \mathbb{R}^m} \ \psi(y)$$

where

$$\psi(y) \ = \ \min\{\, L(s, x, y) \mid s \in \mathcal{C}_1,\ x \in \mathcal{C}_2 \}. \tag{3.19}$$

As we did in deriving the Wolfe dual, one easily derives optimality conditions to calculate $\psi(y)$. Since the minimization in (3.19) is done in the variables $s \in \mathcal{C}_1$ and $x \in \mathcal{C}_2$, and the function $L(s, x, y)$ is separable w.r.t. x and s, we can take these minimums separately.

If we minimize in s all the terms in (3.18) but $s^T y$ are constant. The vector s is in the cone $\mathcal{C}_1$, hence

$$\min_{s \in \mathcal{C}_1} \ s^T y = \begin{cases} 0 & \text{if } y \in \mathcal{C}_1^*, \\ -\infty & \text{otherwise.} \end{cases} \tag{3.20}$$

If we minimize (3.19) in x then all the terms in (3.18) but $x^T(c - A^T y)$ are constant. The vector x is in the cone $\mathcal{C}_2$, hence

$$\min_{x \in \mathcal{C}_2} \ x^T(c - A^T y) = \begin{cases} 0 & \text{if } c - A^T y \in \mathcal{C}_2^*, \\ -\infty & \text{otherwise.} \end{cases} \tag{3.21}$$

By combining (3.20) and (3.21) we have

$$\psi(y) = \begin{cases} b^T y & \text{if } y \in \mathcal{C}_1^* \ \text{ and } c - A^T y \in \mathcal{C}_2^*, \\ -\infty & \text{otherwise.} \end{cases} \tag{3.22}$$

Thus the *dual* of the cone-linear optimization problem (3.17) is the following cone-linear problem:

$$\begin{array}{rcl} \max\ b^T y & & \\ c - A^T y & \in & \mathcal{C}_2^* \\ y & \in & \mathcal{C}_1^*. \end{array} \tag{3.23}$$

Exercise 3.7 *Derive the dual semidefinite optimization problem (DSO) by using the general cone-dual problem (3.23).* ◁

To illustrate the duality relation between (3.17) and (3.23) we prove the following weak duality theorem.

Theorem 3.18 *(Weak duality) If $x \in \mathbb{R}^n$ is a feasible solution of the primal problem (3.17) and $y \in \mathbb{R}^m$ is a feasible solution of the dual problem (3.23) then*

$$c^T x \geq b^T y$$

with equality if and only if

$$x^T(c - A^T y) = 0 \qquad \textit{and} \qquad y^T(Ax - b) = 0.$$

Proof: Using the definition of the dual cone one may write

$$c^T x - b^T y = x^T(c - A^T y) + y^T(Ax - b) \geq 0.$$

Due to the nonnegativity of the vectors x, $c - A^T y$, y and $Ax - b$, equality holds if and only if $x^T(c - A^T y) = 0$ and $y^T(Ax - b) = 0$, which completes the proof. □

Chapter 4

Algorithms for unconstrained optimization

4.1 A generic algorithm

The problem considered in this section is

$$\begin{aligned} \min \quad & f(x) \\ \text{s.t.} \quad & x \in \mathcal{C}, \end{aligned} \tag{4.1}$$

where $\mathcal{C}$ is a relatively open convex set. For typical unconstrained optimization problems one has $\mathcal{C} = \mathbb{R}^n$, the trivial full dimensional open set, but for other applications (like in interior point methods) one frequently has lower dimensional relatively open convex sets. A generic algorithm for minimizing the function $f(x)$ can be presented as follows.

Generic Algorithm

Input:
x^0 is a given (relative interior) feasible point;

For $k = 0, 1, \ldots$ do
Step 1: Find a **search direction** s^k with $\delta f(x^k, s^k) < 0$;
(This should be a descending feasible direction in the constrained case.)
Step 1a: If no such direction exists STOP, optimum found.
Step 2: Line search : find $\lambda_k = \arg\min_{\lambda} f(x^k + \lambda s^k)$;
Step 3: $x^{k+1} = x^k + \lambda_k s^k, \quad k = k + 1$;
Step 4: If **stopping criteria** are satisfied STOP.

The crucial elements of all algorithms, besides the selection of a starting point are printed boldface in the scheme, given above.

To generate a search direction is the crucial element of all minimization algorithms. Once a search direction is obtained, then one performs the *line search procedure*. Before we discuss these aspects in detail we turn to the question of the convergence rate of an algorithm.

4.2 Rate of convergence

Assume that an algorithm generates an n dimensional convergent sequence of iterates $x^1, x^2, \ldots, x^k, \ldots \to \overline{x}$, as a minimizing sequence of the continuous function $f(x) : \mathbb{R}^n \to \mathbb{R}$.

One can define a scalar sequence $\alpha_k = ||x^k - \overline{x}||$ with limit $\alpha = 0$, or a sequence $\alpha_k = f(x^k)$ with limit $\alpha = f(\overline{x})$. The rate of convergence of these sequences gives an indication of 'how fast' the iterates converge. In order to quantify the concept of rate (or order) of convergence, we need the following definition.

Definition 4.1 *Let $\alpha_1, \alpha_2, \ldots, \alpha_k, \ldots \to \alpha$ be a convergent sequence with $\alpha_k \neq \alpha$ for all k. We say that the order of convergence of this sequence is p^* if*

$$p^* = \sup \left\{ p \ : \limsup_{k \to \infty} \frac{|\alpha_{k+1} - \alpha|}{|\alpha_k - \alpha|^p} < \infty \right\}.$$

The larger p^* is, the faster the convergence. Let $\beta = \limsup_{k \to \infty} \frac{|\alpha_{k+1} - \alpha|}{|\alpha_k - \alpha|^{p^*}}$. If $p^* = 1$ and $0 < \beta < 1$ we are speaking about *linear (or geometric rate of) convergence*. If $p^* = 1$ and $\beta = 0$ the convergence rate is *super-linear*, while if $\beta = 1$ the convergence rate is *sub-linear*. If $p^* = 2$ then the convergence is *quadratic*.

Exercise 4.1 *Show that the sequence $\alpha_k = a^k$, where $0 < a < 1$ converges linearly to zero while $\beta = a$.* ◁

Exercise 4.2 *Show that the sequence $\alpha_k = a^{(2^k)}$, where $0 < a < 1$, converges quadratically to zero.* ◁

Exercise 4.3 *Show that the sequence $\alpha_k = \frac{1}{k}$ converges sub-linearly to zero.* ◁

Exercise 4.4 *Show that the sequence $\alpha_k = (\frac{1}{k})^k$ converges super-linearly to zero.* ◁

Exercise 4.5 *Construct a sequence that converges to zero with the order of four.* ◁

4.3 Line search

Line search in fact means one dimensional optimization, since the function $f(x^k + \lambda s^k)$ is the function of the single variable λ. Hence our problem in this part is to find the minimum of a one dimensional function $\phi(\lambda) := f(x^k + \lambda s^k)$, or if it is differentiable one has to find a zero of its derivative $\phi'(\lambda)$.

Exercise 4.6 *Assume that f is continuously differentiable, x^k and s^k are given, and λ_k is obtained via exact line search:*

$$\lambda_k = \arg\min_{\lambda} f(x^k + \lambda s^k).$$

Show that $\nabla f(x^k + \lambda_k s^k)^T s^k = 0$. ◁

Below we present four line search methods, that require different levels of information about $\phi(\lambda)$:

- The Dichotomous search and Golden section methods, that use only function evaluations of ϕ;
- bisection, that evaluates $\phi'(\lambda)$ (ϕ has to be continuously differentiable);
- Newton's method, that evaluates both $\phi'(\lambda)$ and $\phi''(\lambda)$.

4.3.1 Dichotomous and Golden section search

Assume that ϕ is convex and has a minimizer, and that we know an interval $[a, b]$ that contains this minimizer. We wish to reduce the size of this 'interval of uncertainty' by evaluating ϕ at points in $[a, b]$.

Say we evaluate $\phi(\lambda)$ at two points $\bar{a} \in (a, b)$ and $\bar{b} \in (a, b)$, where $\bar{a} < \bar{b}$.

Lemma 4.2 *If $\phi(\bar{a}) < \phi(\bar{b})$ then the minimum of ϕ is contained in the interval $[a, \bar{b}]$. If $\phi(\bar{a}) \geq \phi(\bar{b})$ then the minimum of ϕ is contained in the interval $[\bar{a}, b]$.*

Exercise 4.7 *Prove Lemma 4.2.* ◁

The lemma suggest a simple algorithm to reduce the interval of uncertainty.

Input:

$\epsilon > 0$ is the accuracy parameter;

$a_0,\ b_0$ are given such that $[a_0, b_0]$ contains the minimizer of $\phi(\lambda)$;

For $k = 0, 1, \ldots$, do:

Step 1: If $|a_k - b_k| < \epsilon$ STOP.

Step 2: Choose $\bar{a}_k \in (a_k, b_k)$ and $\bar{b}_k \in (a_k, b_k)$, such that $\bar{a}_k < \bar{b}_k$;

Step 3a: If $\phi(\bar{a}_k) < \phi(\bar{b}_k)$ then the minimum of ϕ is contained in the interval $[a_k, \bar{b}_k]$; set $b_{k+1} = \bar{b}_k$ and $a_{k+1} = a_k$;

Step 3b: If $\phi(\bar{a}_k) \geq \phi(\bar{b}_k)$ then the minimum of ϕ is contained in the interval $[\bar{a}_k, b_k]$; set $a_{k+1} = \bar{a}_k$ and $b_{k+1} = b_k$;

We have not specified yet how we should choose the values $\bar{a}_k$ and $\bar{b}_k$ in iteration k (Step 2 of the algorithm). There are many ways to do this. One is to choose $\bar{a}_k = \frac{1}{2}(a_k + b_k) - \delta$ and $\bar{b}_k = \frac{1}{2}(a_k + b_k) + \delta$ where $\delta > 0$ is a (very) small fixed constant. This is called *Dichotomous Search*.

Exercise 4.8 *Prove that — when using Dichotomous Search — the interval of uncertainty is reduced by a factor $(\frac{1}{2} + \delta)^{t/2}$ after t function evaluations.* ◁

There is a more clever way to choose $\bar{a}_k$ and $\bar{b}_k$, which reduces the number of function evaluations per iteration from two to one, while still shrinking the interval of uncertainty by a constant factor. It is based on a geometric concept called the *Golden section*.

The golden section of a line segment is its division into two unequal segments, such that the ratio of the longer of the two segments to the whole segment is equal to the ratio of the shorter segment to the longer segment.

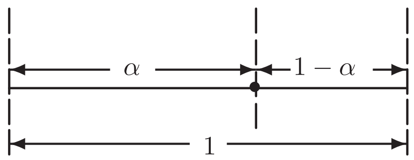


Figure 4.1: The golden section: $\alpha \approx 0.618$.

With reference to Figure 4.1, we require that the value α is chosen such that the following ratios are equal:

$$\frac{1-\alpha}{\alpha} = \frac{\alpha}{1}.$$

This is the same as $\alpha^2 + \alpha - 1 = 0$ which has only one root in the interval $[0, 1]$, namely $\alpha \approx 0.618$.

Returning to the line search procedure, we simply choose $\bar{a}_k$ and $\bar{b}_k$ as the points that correspond to the golden section (see Figure 4.2).

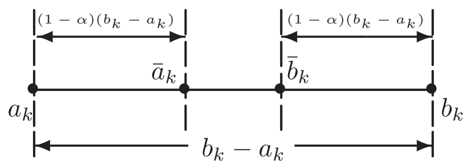


Figure 4.2: Choosing $\bar{a}_k$ and $\bar{b}_k$ via the Golden section rule.

The reasoning behind this is as follows. Assume that we know the values $\phi(\bar{a}_k)$ and $\phi(\bar{b}_k)$ during iteration k. Assume that $\phi(\bar{a}_k) < \phi(\bar{b}_k)$, so that we set $b_{k+1} = \bar{b}_k$ and $a_{k+1} = a_k$. Now, by the definition of the golden section, $\bar{b}_{k+1}$ is equal to $\bar{a}_k$ (see Figure 4.3).

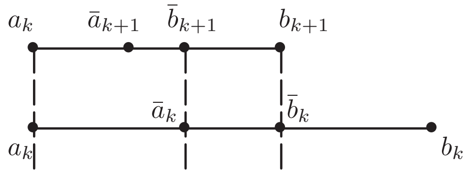


Figure 4.3: Illustration of consecutive iterations of the Golden section rule when $\phi(\bar{a}_k) < \phi(\bar{b}_k)$.

In other words, we do not have to evaluate ϕ at $\bar{b}_{k+1}$, because we already know this value. In iteration $k+1$ we therefore only have to evaluate $\phi(\bar{a}_{k+1})$ in this case. The analysis for the case where $\phi(\bar{a}_k) \geq \phi(\bar{b}_k)$ is perfectly analogous.

Exercise 4.9 *Prove that — when using Golden section search — the interval of uncertainty is reduced by a factor* 0.618^{t-1} *after t function evaluations.* ◁

The Golden section search requires fewer function evaluations than the Dichotomous search method to reduce the length interval of uncertainty to a given $\epsilon > 0$; see Exercise 4.10. If one assumes that the time it takes to evaluate ϕ dominates the work per iteration, then it is more important to count the total number of function evaluations than the number of iterations.

Exercise 4.10 *Show that the Dichotomous search algorithm terminates after at most*

$$2\left(\frac{\log\left(\frac{b_0-a_0}{\epsilon}\right)}{\log\left(\frac{2}{1+2\delta}\right)}\right)$$

function evaluations, and that the Golden section search terminates after at most

$$1+\left(\frac{\log\left(\frac{b_0-a_0}{\epsilon}\right)}{\log\left(\frac{1}{0.618}\right)}\right)$$

function evaluations. Which of the two bounds is better? ◁

4.3.2 Bisection

The Bisection method (also called Bolzano's method) is used to find a root of $\phi'(\lambda)$ (here we assume ϕ to be continuously differentiable). Recall that such a root corresponds to a minimum of ϕ if ϕ is convex.

The algorithm is similar to the Dichotomous and Golden section search ones, in the sense that it too uses an interval of uncertainty that is reduced at each iteration. In the case of the bisection method the interval of uncertainty contains a root of $\phi'(\lambda)$.

The algorithm proceeds as follows.

Input:

$\epsilon > 0$ is the accuracy parameter;

$a_0,\ b_0$ are given such that $\phi'(a_0) < 0$ and $\phi'(b_0) > 0$;

For $k = 0, 1, \ldots,$ do:

Step 1: If $|b_k - a_k| < \epsilon$ STOP.

Step 2: Let $\lambda = \frac{1}{2}(a_k + b_k)$;

Step 3: If $\phi'(\lambda) < 0$ then $a_{k+1} := \lambda$ and $b_{k+1} = b_k$;

Step 4: If $\phi'(\lambda) > 0$ then $b_{k+1} := \lambda$ and $a_{k+1} = a_k$.

Exercise 4.11 *Prove that that the bisection algorithm uses at most* $\log_2 \frac{|b_0 - a_0|}{\epsilon}$ *function evaluations before terminating.* ◁

Nota that the function $\phi'(\lambda)$ does not have to be differentiable in order to perform the bisection procedure.

4.3.3 Newton's method

Newton's method is another algorithm for finding a root of ϕ'. Once again, such a root corresponds to a minimum of ϕ if ϕ is convex. Newton's method requires that ϕ be twice continuously differentiable and strictly convex, and works as follows: we construct the linear Taylor approximation to ϕ' at the current iterate λ_k, namely

$$l(\lambda) := \phi'(\lambda_k) + \phi''(\lambda_k)(\lambda - \lambda_k).$$

Next we find the root of $l(\lambda)$ and set λ_{k+1} to be equal to this root. This means that λ_{k+1} is given by

$$\lambda_{k+1} = \lambda_k - \frac{\phi'(\lambda_k)}{\phi''(\lambda_k)}.$$

Now we repeat the process with λ_{k+1} as the current iterate.

There is an equivalent interpretation of this procedure: take the quadratic Taylor approximation of ϕ at the current iterate λ_k, namely

$$q(\lambda) = \phi(\lambda_k) + \phi'(\lambda_k)(\lambda - \lambda_k) + \frac{1}{2}\phi''(\lambda_k)(\lambda - \lambda_k)^2,$$

and set λ_{k+1} to be the minimum of q. The minimum of q is attained at

$$\lambda_{k+1} = \lambda_k - \frac{\phi'(\lambda_k)}{\phi''(\lambda_k)},$$

and λ_{k+1} becomes the new iterate (new approximation to the minimum). Note that the two interpretations are indeed equivalent.

Newton's algorithm can be summarized as follows.

Input:

$\epsilon > 0$ is the accuracy parameter;

λ_0 is the given initial point; $k = 0$;

For $k = 0, 1, \ldots,$ do:

Step 1: Let $\lambda_{k+1} = \lambda_k - \frac{\phi'(\lambda_k)}{\phi''(\lambda_k)}$;

Step 2: If $|\lambda_{k+1} - \lambda_k| < \epsilon$ STOP.

Newton's method as presented above may not converge to the global minimum of ϕ. On the other hand, Newton's method has some spectacular properties. It converges quadratically if the following conditions are met:

1. the starting point is sufficiently close to the minimum point;
2. in addition to being convex, the function ϕ has a property called *self-concordance*, which we will discuss later.

The next two examples illustrate the possible scenarios.

Example 4.3 Let us apply Newton's method to $\phi(\lambda) = \lambda - \log(1+\lambda)$. Note that the domain of ϕ is $(-1, \infty)$. The first and second derivatives of ϕ are given by

$$\phi'(\lambda) = \frac{\lambda}{1+\lambda}, \ \phi''(\lambda) = \frac{1}{(1+\lambda)^2},$$

and it is therefore clear that ϕ is strictly convex on its domain, and that $\lambda = 0$ is the minimizer of ϕ.

The iterates from Newton's method satisfy the recursive relation

$$\lambda_{k+1} = \lambda_k - [\phi''(\lambda_k)]^{-1}\phi'(\lambda_k) = \lambda_k - \lambda_k(1+\lambda_k) = -\lambda_k^2.$$

This implies quadratic convergence if $|\lambda_0| < 1$ (see Exercise 4.12).

On the other hand, note that Newton's method *fails* if $\lambda_0 \geq 1$. For example, if $\lambda_0 = 1$ then $\lambda_1 = -1$, which is not in the domain of ϕ!

We mention that the convex function ϕ has the self-concordance property mentioned above. This will be shown in Exercise 6.16. $*$

Exercise 4.12 *This exercise refers to Example 4.3. Prove that, if the sequence $\{\lambda_k\}$ satisfies*

$$\lambda_{k+1} = -\left(\lambda_k\right)^2,$$

then $\lambda_k \to 0$ with a quadratic rate of convergence if $|\lambda_0| < 1$. $\triangleleft$

In the following example, Newton's method converges to the minimum, but the rate of convergence is only linear.

Example 4.4 Let $m \geq 2$ be even and define

$$\phi(\lambda) = \lambda^m.$$

Clearly, ϕ has a unique minimizer, namely $\lambda = 0$. Suppose we start Newton's method at some nonzero $\lambda_0 \in \mathbb{R}$.

The derivatives of ϕ are

$$\begin{aligned} \phi'(\lambda) &= m\lambda^{m-1} \\ \phi''(\lambda) &= m(m-1)\lambda^{m-2}. \end{aligned}$$

Hence, the iterates from Newton's method satisfy the recursive relation

$$\lambda_{k+1} = \lambda_k - \left(\phi''(\lambda_k)\right)^{-1}\phi'(\lambda_k) = \lambda_k + \frac{-1}{m-1}\lambda_k = \frac{m-2}{m-1}\lambda_k.$$

This shows that Newton's method is exact if ϕ is quadratic (if $m = 2$), whereas for $m > 2$ the Newton process converges to 0 with a *linear convergence rate* (see Exercise 4.13). $*$

Exercise 4.13 *This exercise refers to Example 4.4. Prove that, if the sequence $\{\lambda_k\}$ satisfies*

$$\lambda_{k+1} = \frac{m-2}{m-1}\lambda_k,$$

where $m > 2$ is even, then $\lambda_k \to 0$ with a linear rate of convergence, if $\lambda_0 \neq 0$. $\triangleleft$

4.4 Gradient method

We now return to the generic algorithm on page 63, and look at some different choices for the search direction. The gradient method uses the negative gradient $(-\nabla f(x^k))$ of the function f as the search direction.[1] This direction is frequently referred to as the *steepest descent direction*. This name is justified by observing that the normalized directional derivative is minimized by the negative gradient

$$\delta f(x, -\nabla f(x)) = -\nabla f(x)^T \nabla f(x) = \min_{||s||=||\nabla f(x)||} \{\nabla f(x)^T s\}.$$

Exercise 4.14 *Let $f : \mathbb{R}^n \mapsto \mathbb{R}$ be continuously differentiable and let $\bar{x} \in \mathbb{R}^n$ be given. Assume that the level set $\{x \in \mathbb{R}^n \mid f(x) = f(\bar{x})\}$, is in fact a curve (contour). Show that $\nabla f(\bar{x})$ is orthogonal to the tangent line to the curve at $\bar{x}$.* ◁

To calculate the gradient is relatively cheap which indicates that the gradient method can be quite efficient. Although it works fine in many applications, several theoretical and practical disadvantages can be mentioned. First, the minimization of a convex quadratic function by the gradient method is not a finite process in general. Slow convergence, due to a sort of "zigg–zagging" sometimes takes place. Secondly, the order of convergence is no better than linear in general.

Figure 4.4 illustrates the zig-zag behavior that may occur when using the gradient method.

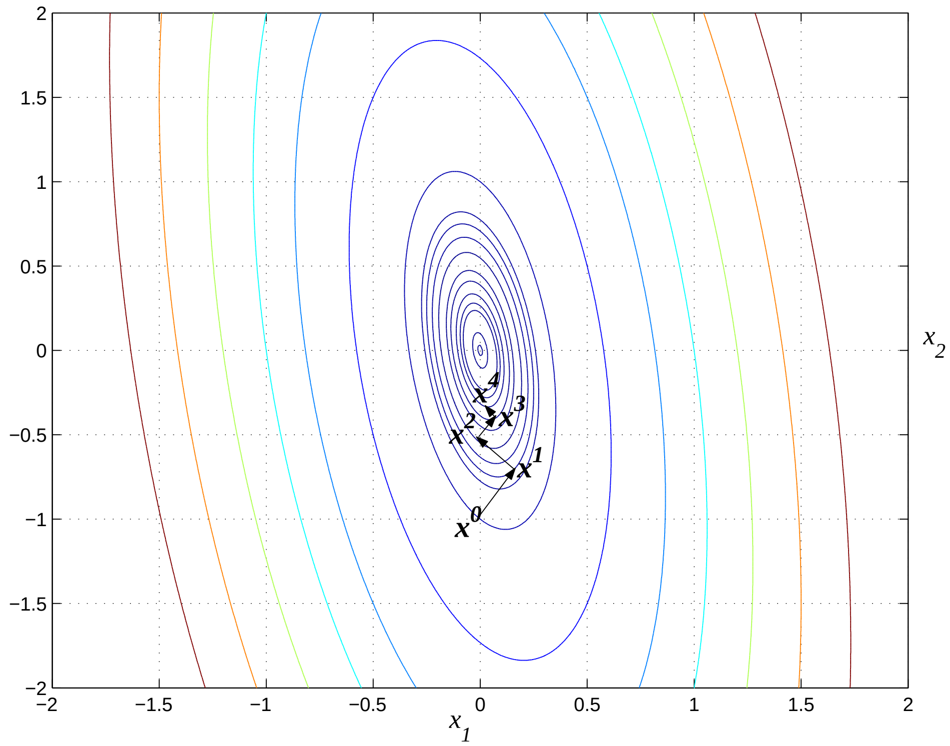


Figure 4.4: Iterates of the gradient method for the function $f(x) = 9x_1^2 + 2x_1x_2 + x_2^2$.

Exercise 4.15 *Calculate the steepest descent direction for the quadratic function*

$$f(x) = \frac{1}{2}x^T Qx + q^T x - \beta,$$

where the matrix Q is positive definite. Calculate the exact step length in the line search as well. ◁

[1] Here, for the sake of simplicity, it is assumed that $\mathcal{C} = \mathbb{R}^n$. In other cases the negative gradient might point out of the feasible set $\mathcal{C}$.

Exercise 4.16 *Prove that subsequent search directions of the gradient method are always orthogonal (i.e.* $s^k \perp s^{k+1}$; $k = 0, 1, 2, \ldots$*) if exact line search is used.* ◁

The following theorem gives a convergence result for the gradient method.

Theorem 4.5 *Let* f *be continuously differentiable. Starting from the initial point* x^0 *using exact line search the gradient method produces a decreasing sequence* $x^0, x^1, x^2, \cdots$ *such that* $f(x^k) > f(x^{k+1})$ *for* $k = 0, 1, 2, \cdots$*. Assume that the level set* $D = \{x : f(x) \leq f(x^0)\}$ *is compact, then any accumulation point* $\overline{x}$ *of the generated sequence* $x^0, x^1, x^2, \cdots, x^k, \cdots$ *is a stationary point (i.e.* $\nabla f(\overline{x}) = 0$*) of* f*. Further if the function* f *is a convex function, then* $\overline{x}$ *is a global minimizer of* f*.*

Proof: Since D is compact and f is continuous we have that f is bounded on D, hence we have a convergent subsequence $x^{k_j} \to \overline{x}$ with $f(x^{k_j}) \to f^*$ as $k_j \to \infty$. By continuity of f we have $f(\overline{x}) = f^*$. Since the search direction is the gradient of f we have

$$\overline{s} = \lim_{k_j \to \infty} s^{k_j} = -\lim_{k_j \to \infty} \nabla f(x^{k_j}) = -\nabla f(\overline{x}).$$

Multiplying by $\nabla f(\overline{x})$ we have

$$\overline{s}^T \nabla f(\overline{x}) = -\nabla f(\overline{x})^T \nabla f(\overline{x}) \leq 0. \tag{4.2}$$

On the other hand using the construction of the iteration sequence and the convergent subsequence we write

$$f(x^{k_{j+1}}) \leq f(x^{k_j+1}) \leq f(x^{k_j} + \lambda s^{k_j}).$$

Taking the limit in the last inequality we have

$$f(\overline{x}) \leq f(\overline{x} + \lambda \overline{s})$$

which leads to $\delta f(\overline{x}, \overline{s}) = \overline{s}^T \nabla f(\overline{x}) \geq 0$. Combining this result with (4.2) we have $\nabla f(\overline{x}) = 0$, and the theorem is proved. □

4.5 Newton's method

We now extend Newton's method to the multivariate case. To apply Newton's method we have to assume that the function f is a twice continuously differentiable function with positive definite Hessian on its domain. Newton's search direction in multidimensional optimization is again based on minimizing the second order approximation of the function f. The quadratic Taylor approximation at the current iterate x^k is given by:

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

Since the Hessian $\nabla^2 f(x^k)$ is positive definite, the function $q(x)$ is strictly convex (see Exercise 1.20). Hence the minimum of $q(x)$ is attained when its gradient

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

equals to the zero vector, i.e. at the point

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

The classical Newton method does not apply line search, one takes the full Newton step. If line search is applied then typically we are far from the solution, the step length is usually less than one. We refer to this as the *damped Newton* method.

In addition we have to mention that to compute and invert the Hesse matrix is more expensive than to compute only the gradient. Several methods are developed to reduce this cost while preserving the advantages of Newton's method. These are the so-called *quasi-Newton* methods of which the most popular are the methods which use *conjugate directions*, to be discussed later.

Anyway, the compensation for the extra cost in Newton's method is a better search direction. Just note that the minimization of a convex quadratic function happens in one step.

Exercise 4.17 *Let* $f(x) = \frac{1}{2}x^T Ax - b^T x$ *where* A *is positive definite and* $b \in \mathbb{R}^n$*. Assume that we apply Newton's method to minimize* f*. Show that* $x^1 = A^{-1}b$, i.e. x^1 *is the minimum of* f*, regardless of the starting point* x^0*.* ◁

If the Hessian $\nabla^2 f(x)$ is not positive definite, or is ill-conditioned (the ratio of the largest and smallest eigenvalue is large) then it is not (or hardly) invertible. In this case additional techniques are needed to circumvent these difficulties. In the *trust region* method, $\nabla^2 f(x)$ is replaced by $(\nabla^2 f(x)+\alpha I)$ where I is the identity matrix and α is changed dynamically. Observe that if $\alpha = 0$ then we have the Hessian, hence we have the Newton step, while as $\alpha \to \infty$ this matrix approaches a multiple of the identity matrix and so the search direction is asymptotically getting parallel to the negative gradient.

The interested reader can consult the following books for more details on trust region methods [2, 3, 16, 9].

Exercise 4.18 *Let* $x \in \mathbb{R}^n$ *and* f *be twice continuously differentiable. Show that* $s = -H\nabla f(x)$ *is a descent direction of* f *at* x *for any positive definite matrix* H*, if* $\nabla f(x) \neq 0$*. Which choice of* H *gives:*

- *the steepest descent direction?*
- *Newton's direction (for convex* f*)?*

◁

Exercise 4.19 *Consider the unconstrained optimization problem:*

$$\min\,(x_1 - 2)^4 + (x_1 - 2x_2)^2\,.$$

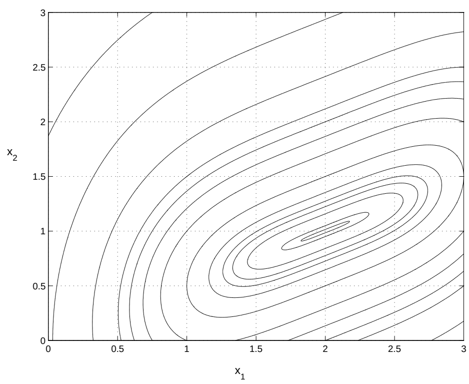


Figure 4.5: Contours of the function. Note that the minimum is at $[2, 1]^T$.

1. *Perform two iterations of the gradient method, starting from* $x^0 = [0, 3]^T$*.*
2. *Perform four iterations of Newton's method (without line search), with the same starting point* x^0*.*

◁

Relation with Newton's method for solving nonlinear equations

The reader may be familiar with Newton's method to solve nonlinear systems of equations. Here we show that Newton's optimization method is obtained by setting the gradient of f to zero and using Newton's method for nonlinear equations to solve the resulting equations.

Assume we have a nonlinear system of equations

$$F(x) = 0$$

to solve, where $F(x)$ is a differentiable mapping from $\mathbb{R}^n \to \mathbb{R}^m$. Given any point $x^k \in \mathbb{R}^n$, Newton's method proceeds as follows. Let us first linearize the nonlinear equation at x^k by approximating $F(x)$ by $F(x^k) + JF(x^k)(x - x^k)$ where $JF(x)$ denotes the Jacobian of F, an $m \times n$ matrix defined as

$$JF(x)_{ij} = \frac{\partial F_i(x)}{\partial x_j} \quad \text{where} \quad i = 1, \cdots, m; \;\; j = 1, \cdots, n.$$

Now we take a step so that the iterate after the step satisfies the linearized equation

$$JF(x^k)(x^{k+1} - x^k) = -F(x^k). \tag{4.3}$$

This is a linear system of equations, hence a solution (if it exists) can be found by standard linear algebra.

Observe, that if we want to minimize a strictly convex function $f(x)$ one can interpret this problem as solving the nonlinear system of equations $\nabla f(x) = 0$. The solution of this system by Newton's method, as we have a point x^k, leads to (apply (4.3))

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

The Jacobian of the gradient is exactly the Hessian of the function $f(x)$ hence it is positive definite and we have

$$x^{k+1} = x^k - (\nabla^2 f(x^k))^{-1} \nabla f(x^k)$$

as we have seen above.

4.6 Methods of Conjugate directions

Let A be an $n \times n$ symmetric positive definite matrix and $b \in \mathbb{R}^n$. We consider the problem of minimizing the strictly convex quadratic function

$$q(x) = \frac{1}{2} x^T A x - b^T x.$$

We will study a class of algorithms that use so-called conjugate search directions to minimize q.

Definition 4.6 *The directions (vectors)* $s^1, \cdots, s^k \in \mathbb{R}^n$ *are called* conjugate (or $A-$conjugate) directions *if* $(s^i)^T A s^j = 0$ *for all* $1 \le i \ne j \le k$.

Note that conjugate directions are mutually orthogonal if $A = I$.

Exercise 4.20 *Let* A *be* $n \times n$ *symmetric positive definite and* $s^1, \ldots, s^k$ $(k \le n)$ *be A-conjugate. Prove that* $s^1, \ldots, s^k$ *are linearly independent.* ◁

If one uses A-conjugate directions in the generic algorithm to minimize q, then the minimum is found in at most n iterations. The next theorem establishes this important fact.

Theorem 4.7 *Let $s^0, \cdots, s^k \in \mathbb{R}^n$ be conjugate directions with respect to A. Let x^0 be given and let*

$$x^{i+1} := arg\min \; q(x^i + \lambda s^i) \qquad i = 0, \cdots, k.$$

Then x^{k+1} minimizes $q(x)$ on the affine space $H = x^0 + span(s^0, \cdots, s^k)$.

Proof: One has to show (see Theorem 2.9) that $\nabla q(x^{k+1}) \perp s^1, \cdots, s^k$. Recall that

$$x^{i+1} := x^i + \lambda^i s^i \qquad i = 0, \cdots, k$$

where λ^i indicates the line-search minimum, thus

$$x^{k+1} := x^1 + \lambda^0 s^0 + \cdots + \lambda^k s^k = x^i + \lambda^i s^i + \cdots + \lambda^k s^k,$$

for any fixed $i \leq k$. Due to exact line-search we have $\nabla q(x^{i+1})^T s^i = 0$ (see Exercise 4.6). Using $\nabla q(x) = Ax - b$, we get

$$\nabla q(x^{k+1}) := \nabla q(x^i + \lambda^i s^i) + \sum_{j=i+1}^{k} \lambda^j A s^j.$$

Taking the inner product on both sides with s^i yields

$$(s^i)^T \nabla q(x^{k+1}) := (s^i)^T \nabla q(x^{i+1}) + \sum_{j=i+1}^{k} \lambda^j (s^i)^T A s^j.$$

Hence $(s^i)^T \nabla q(x^{k+1}) = 0$. □

Corollary 4.8 *Let x^k be defined as in Theorem 4.7. Then $x^n = A^{-1}b$,* i.e. *x^n is the minimizer of $q(x) = \frac{1}{2}x^T Ax - b^T x$.*

Exercise 4.21 *Show that the result in Corollary 4.8 follows from Theorem 4.7.* ◁

4.6.1 The method of Powell

To formulate algorithms that use conjugate directions, we need tools to construct conjugate directions. The next theorem may seem a bit technical, but it gives us such a tool.

Theorem 4.9 *Let $\mathcal{L}$ be a linear subspace, $H_1 := x^1 + \mathcal{L}$ and $H_2 := x^2 + \mathcal{L}$ be two parallel affine spaces where x^1 and x^2 are the minimizers of $q(x)$ over H_1 and H_2, respectively.*

Then for every $s \in \mathcal{L}$, $(x^2 - x^1)$ and s are conjugate with respect to A.

Proof: Assume x^1 minimizes $q(x)$ over $H_1 = x^1 + \mathcal{L}$ and x^2 minimizes $q(x)$ over $H_2 = x^2 + \mathcal{L}$. Let $s \in \mathcal{L}$. Now

$$\begin{aligned} q(x^1 + \lambda s) \geq q(x^1) &\quad \Rightarrow \quad s^T \nabla q(x^1) = 0 \\ q(x^2 + \lambda s) \geq q(x^2) &\quad \Rightarrow \quad s^T \nabla q(x^2) = 0 \end{aligned}$$

This implies that

$$s^T \left(\nabla q(x^2) - \nabla q(x^1) \right) = s^T A(x^2 - x^1) = 0.$$

In other words, for any $s \in \mathcal{L}$, s and $x^2 - x^1$ are A-conjugate directions. □

The basic ingredients of the method of Powell are as follows:

- The algorithm constructs conjugate directions t^1, ..., t^n by using the result of Theorem 4.9. The method requires one *cycle* of $n+1$ line searches to construct each conjugate direction t^i. Thus the first conjugate direction t^1 is constructed at the end of cycle 1, *etc.*
- It starts with a fixed set of linearly independent directions s^1, ..., s^n to achieve this. (Usually the standard unit vectors.)
- In the first cycle, the method performs successive exact line searches using the directions s^1, ..., s^n (in that order). In the second cycle the directions s^2, ..., s^n, t^1 are used (in that order). In the third cycle the directions s^3, ..., s^n, t^1, t^2 are used, *etc.*
- The method terminates after n cycles due to the result in Theorem 4.7.

We will now state the algorithm, but first a word about notation. As mentioned before, the second cycle uses the directions s^2, ..., s^n, t^1. In order to state the algorithm in a compact way, the search directions used during cycle k are called $s^{(k,1)}$, ..., $s^(k,n)$.

The iterates generated during cycle k via successive line searches will be called $z^{(k,1)}, \ldots, z^{(k,n)}$, and x^k will denote the iterate at the end of *cycle* k.

Powell's algorithm

Input A starting point x^0, a set of linearly independent vectors s^1, ..., s^n.
Initialization Set $s^{(1,i)} = s^i, \;\; i = 1, \cdots, n$.

For $k = 1, 2, \ldots, n$ **do:**
(Cycle k:)
Let $z^{(k,1)} = x^{k-1}$ and $z^{(k,i+1)} := \arg\min q\left(z^{(k,i)} + \lambda s^{(k,i)}\right) \qquad i = 1, \cdots, n.$
Let $x^k := \text{argmin } q(z^{(k,n+1)} + \lambda t^k)$ where $t^k := z^{(k,n+1)} - x^{k-1}$.
Let $s^{(k+1,i)} = s^{(k,i+1)}, \;\; i = 1, \cdots, n-1$ and $s^{(k+1,n)} := t^k$.

It may not be clear to the reader why the directions $t^1, t^2, \ldots$ are indeed conjugate directions. As mentioned before, we will invoke Theorem 4.9 to prove this.

Lemma 4.10 *The vectors $t^1, \ldots, t^n$ generated by Powell's algorithm are A-conjugate.*

Proof: The proof is by induction. Assume that $t^1, \ldots, t^k$ are conjugate at the end of cycle k of the algorithm. By the definition of x^k in the statement of the algorithm, and by Theorem 4.7, x^k minimizes q over the affine space $x^k + \text{span}\{t^1, \ldots, t^k\}$.

In cycle $k+1$, $z^{(k+1,n+1)}$ is obtained after successive line searches along the directions $\{s^1, \ldots, s^{n-k}, t^1, \ldots, t^k\}$. By Theorem 4.7, $z^{(k+1,n+1)}$ minimizes q over the affine space $z^{(k+1,n)} + \text{span}\{t^1, \ldots, t^k\}$.

Now define $t^{k+1} = z^{(k,n+1)} - x^{k-1}$. By Theorem 4.9, t^{k+1} is A-conjugate to every vector in $\text{span}\{t^1, \ldots, t^k\}$, and in particular to $\{t^1, \ldots, t^k\}$. □

Example 4.11 We consider the problem

$$\min \; f(x) = 5x_1^2 + 2x_1x_2 + x_2^2 + 7.$$

The minimum is attained at $x_1 = x_2 = 0$.

We choose s^1 and s^2 as the standard unit vectors in $\mathbb{R}^2$, and the starting point is: $x^0 = [1, 2]^T$. The progress of Powell's method for this example is illustrated in Figure 4.6. We will describe the progress with giving the actual numerical values, in order to keep things simple.

Note that, at the start of cycle 1, successive line searches are done using $s^1 = [0\ 1]^T$ and $s^2 = [1\ 0]^T$. Then the first conjugate direction t^1 is generated by connecting x^0 with the last point obtained, and a line search is performed along t^1 to obtain the point x^1.

In cycle 2, successive line searches are done using s^2 and t^1. Then the second conjugate direction t^2 is generated by connecting x^1 with the last point obtained, and a line search is performed along t^2 to obtain the point x^2.

Note that x^2 is optimal, as it should be. *

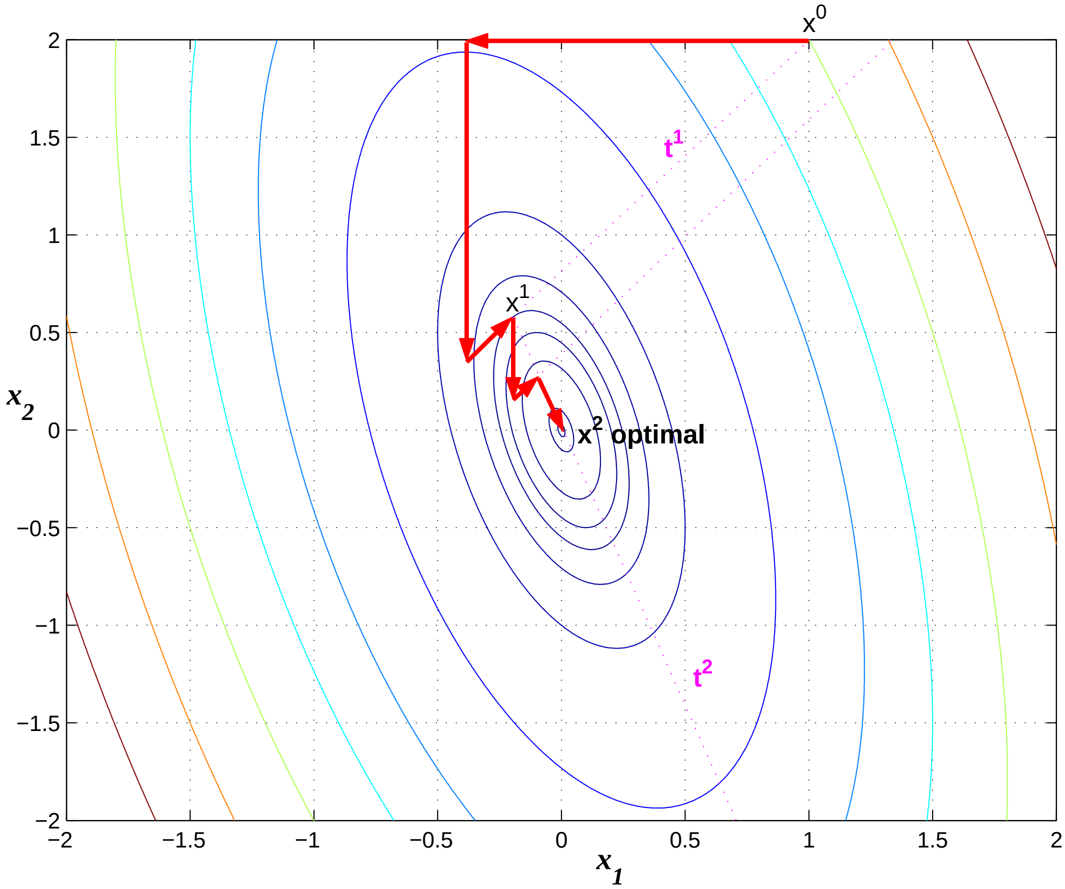


Figure 4.6: Iterates generated by Powell's algorithm for the function $f(x) = 5x_1^2 + 2x_1x_2 + x_2^2 + 7$, starting from $x^0 = [1, 2]^T$.

Discussion of Powell's method

- We may apply Powell's algorithm to any function (not necessarily quadratic); The only change to the Algorithm on page 74 is that the quadratic function $q(x)$ is replaced by a general $f(x)$. Of course, in this case it does not make sense to speak of conjugate directions, and there is no guarantee that x^n will be the optimal solution. For this reason it is customary to restart the algorithm from x^n.

- Powell's algorithm uses *only line searches*, and finds the exact minimum of a strictly convex quadratic function after at most $n(n+1)$ line-searches. For a general (convex) function f, Powell's method can be combined with the Golden section line search procedure to obtain an algorithm for minimizing f that does *not* require gradient information.

- Storage requirements: The algorithm stores n n-vectors (the current set of search directions) at any given time.

Let us compare Powell's method to Newton's method and the gradient method. Newton's method requires only one step to minimize a strictly convex quadratic function, but requires both gradient and Hessian information for general functions. The gradient method requires only gradient information, but does not always converge in a finite number of steps (not even for strictly convex quadratic functions).

In conclusion, Powell method is an attractive algorithm for minimizing 'black box' functions where gradient and Hessian information is not available (or too expensive to compute).

4.6.2 The Fletcher-Reeves method

The method of Fletcher and Reeves is also a conjugate gradient method to

$$\text{minimize } q(x) = \frac{1}{2}x^T Ax - b^T x,$$

but is simpler to state than the method of Powell.

Before giving a formal statement of the algorithm, we list the key ingredients:

- The first search direction is the steepest descent direction: $s^0 = -\nabla q(x^0)$.
- The search direction at iteration k, namely s^k, is constructed so that it is conjugate with respect to the preceding directions $s^0, \ldots, s^{k-1}$, as well as a linear combination of $-\nabla q(x^k)$ and $s^0, \ldots, s^{k-1}$.
- We will show that these requirements imply that

$$s^k = -\nabla q(x^k) + \left(\frac{\|\nabla q(x^k)\|^2}{\|\nabla q(x^{k-1})\|^2}\right) s^{k-1}.$$

- Note that, unlike Powell's method, this method requires *gradient information.* The advantage over Powell's method is that we only have to store two n-vectors and do $n+1$ line searches.
- We may again use the method to minimize non-quadratic functions, but then convergence is not assured.

Let us consider the situation during iteration k, *i.e.* assume that x^k, $\nabla q(x^k)$ and $s^1, \cdots, s^{k-1}$ conjugate directions be given.

We want to find values $\beta_{k,0}$ .... $\beta_{k,k-1}$ such that

$$s^k := -\nabla q(x^k) + \beta_{k,0}s^0 + \cdots + \beta_{k,k-1}s^{k-1},$$

and s^k is conjugate with respect to $s^0, \cdots, s^{k-1}$.

We require A-conjugacy, *i.e.* $s_i^T As_k = 0$, which implies:

$$\beta_{k,i} = \frac{\nabla q(x^k)^T As^i}{(s^i)^T As^i} \qquad (i = 0, \ldots, k-1).$$

We will now show that $\beta_{k,i} = 0$ if $i < k-1$. To this end, note that

$$\nabla q(x^{i+1}) - \nabla q(x^i) = A(x_{i+1} - x_i) = \lambda_i As^i.$$

Therefore

$$\beta_{k,i} = \frac{\nabla q(x^k)^T(\nabla q(x^{i+1}) - \nabla q(x^i))}{(s^i)^T(\nabla q(x^{i+1}) - \nabla q(x^i))} \quad (i < k).$$

For any $i < k$ we have

$$s^i = -\nabla q(x^i) + \beta_{i,1}s^1 + \cdots + \beta_{i,i-1}s^{i-1}.$$

By Theorem 4.7 we have

$$\nabla q(x^k) \perp s^i \quad (i = 0, \ldots, k-1).$$

Therefore

$$\nabla q(x^i)^T \nabla q(x^k) = 0 \quad (i < k),$$

and

$$\nabla q(x^i)^T s^i = -\|\nabla q(x^i)\|^2 \quad (i < k).$$

Therefore $\beta_{k,i} = 0$ if $i < k-1$. Also, due to exact line-search, we have $(s^i)^T(\nabla q(x^{i+1})) = 0$ (see Exercise 4.6). Therefore

$$\beta_{k,k-1} = \frac{\|\nabla q(x^k)\|^2}{\|\nabla q(x^{k-1})\|^2}.$$

Fletcher-Reeves algorithm

Let x^0 be an initial point.

Step 0**.** Let $s^0 = -\nabla q(x^0)$ and
$x^1 := \arg\min q(x^0 + \lambda s^0)$.

Step k**.** Let x^k, $\nabla q(x^k)$ and $s^0, \cdots, s^{k-1}$ conjugate directions be given. Set

$$s^k = -\nabla q(x^k) + \left(\frac{\|\nabla q(x^k)\|^2}{\|\nabla q(x^{k-1})\|^2} \right) s^{k-1}.$$

Set $x^{k+1} := \text{argmin} q(x^k + \lambda s^k)$.

Exercise 4.22

$$\min x_1^2 + 2x_2^2 + 2x_3^2 + 2x_1x_2 + 2x_2x_3.$$

1. *Solve this problem using the conjugate gradient method of Powell. Use exact line search and the starting point* $[2, 4, 10]^T$*. Use the standard unit vectors as* s^1*,* s^2 *and* s^3*.*
2. *Solve this problem using the Fletcher-Reeves conjugate gradient method. Use exact line search and the starting point* $[2, 4, 10]^T$*.*

◁

4.7 Quasi-Newton methods

Recall that the Newton direction at iteration k is given by:

$$s^k = -\left[\nabla^2 f(x^k)\right]^{-1} \nabla f(x^k).$$

Quasi-Newton methods use a positive definite approximation H_k to $\left[\nabla^2 f(x^k)\right]^{-1}$. The approximation H_k is updated at each iteration, say

$$H_{k+1} = H_k + D_k,$$

where D_k denotes the update.

Let A be an $n \times n$ symmetric PD matrix, and consider once more the strictly convex quadratic function

$$q(x) = \frac{1}{2} x^T A x - b^T x.$$

The Newton direction for q at x^k is:

$$s^k = -\left[\nabla^2 q(x^k)\right]^{-1} \nabla q(x^k) = -A^{-1} \nabla q(x^k).$$

Note that

$$\nabla q(x^{k+1}) - \nabla q(x^k) = A\left(x^{k+1} - x^k\right).$$

We introduce the notation

$$y^k := \nabla q(x^{k+1}) - \nabla q(x^k), \quad \sigma^k = x^{k+1} - x^k.$$

Notice that $y^k = A\sigma^k$ *i.e.* $\sigma^k = A^{-1} y^k$.

The secant condition

We will use a search direction of the form

$$s^k = -H_k \nabla q(x^k)$$

where H_k is an approximation of $[\nabla^2 q(x^k)]^{-1} = A^{-1}$, and subsequently perform the usual *line search*:

$$x^{k+1} = \arg\min q(x^k + \lambda s^k).$$

Since

$$y^k := \nabla q(x^{k+1}) - \nabla q(x^k), \;\; \sigma^k = x^{k+1} - x^k,$$

and $\sigma^k = A^{-1}y^k$, we require that $\sigma^k = H_{k+1}y^k$. This is called the *secant condition* (*quasi-Newton property*).

The hereditary property

Since

$$y^k := \nabla q(x^{k+1}) - \nabla q(x^k), \;\; \sigma^k = x^{k+1} - x^k,$$

and $\nabla q(x) = Ax - b$, it holds that

$$\sigma^i = A^{-1}y^i \qquad (i = 0, \ldots, k-1).$$

We therefore require that our approximation H_k also satisfies

$$\sigma^i = H_k y^i \qquad (i = 0, \ldots, k-1).$$

This is called the *hereditary property*.

Since

$$\sigma^i = A^{-1}y^i \text{ and } \sigma^i = H_n y^i \quad (i = 0, \ldots, n-1),$$

it follows that $H_n A\sigma^i = \sigma^i$ $(i = 0, \ldots, n-1)$. If the σ^i $(i = 0, \ldots, n-1)$ are linearly independent, this implies $H_n = A^{-1}$.

Discussion

We showed that — if the σ^i $(i = 0, \ldots, n-1)$ are linearly independent — we have $H_n = A^{-1} = \left[\nabla^2 q(x^n)\right]^{-1}$ (the approximation has become exact!) In iteration n, we therefore use the search direction

$$s^n = -H_n \nabla q(x^n) = -A^{-1}\nabla q(x^n).$$

But this is simply the *Newton direction* at x^n! In other words, we find the minimum of q no later than in iteration n.

Generic Quasi-Newton algorithm

Step 0: Let x^0 be given and set $H_0 = I$.

Step k**:** Calculate the search direction $s^k = -H_k \nabla q(x^k)$ and perform the usual line search $x^{k+1} = \arg\min q(x^k + \lambda s^k)$.

We choose D_k in such a way that:

i $H_{k+1} = H_k + D_k$ is symmetric positive definite;

ii $\sigma^k = H_{k+1}y^k$ (secant condition);

iii $\sigma^i = H_{k+1}y^i$ $(i = 0, \ldots, k-1)$ (hereditary property).

4.7.1 The DFP update

The *Davidon-Fletcher-Powell (DFP)* rank-2 update is defined by

$$D_k = \frac{\sigma^k {\sigma^k}^T}{{\sigma^k}^T y^k} - \frac{H_k y^k {y^k}^T H_k}{{y^k}^T H_k y^k}.$$

We will show that:

i If $y_k^T \sigma_k > 0$, then H_{k+1} is positive definite.

ii $H_{k+1} = H_k + D_k$ satisfies the secant condition: $\sigma^k = H_{k+1} y^k$.

iii The hereditary property holds: $\sigma^i = H_{k+1} y^i$ $(i = 0, \ldots, k-1)$.

Exercise 4.23 *Show that, if H_k is positive definite, then*

$$H_{k+1} = H_k + D_k = H_k + \frac{\sigma^k {\sigma^k}^T}{{\sigma^k}^T y^k} - \frac{H_k y^k {y^k}^T H_k}{{y^k}^T H_k y^k},$$

is also positive definite if $(\sigma^k)^T y^k > 0$.

Hint 1: For ease of notation, show that

$$H + \frac{\sigma \sigma^T}{\sigma^T y} - \frac{H y y^T H}{y^T H y},$$

is positive definite if the matrix H is P.D. and the vectors y, σ satisfy $y^T \sigma > 0$.

Hint 2: Set $H = LL^T$ and show that

$$v^T \left(H + \frac{\sigma \sigma^T}{\sigma^T y} - \frac{H y y^T H}{y^T H y} \right) v > 0 \qquad \forall v \in \mathbb{R}^n \setminus \{0\}.$$

Hint 3: Use the Cauchy-Schwartz inequality

$$(a^T a)(b^T b) - (a^T b)^2 > 0 \text{ if } a \neq kb \text{ for all } k \in \mathbb{R},$$

to obtain the required inequality. ◁

Exercise 4.24 *Prove that $H_{k+1} = H_k + D_k$ satisfies the secant condition: $\sigma^k = H_{k+1} y^k$.* ◁

We now prove that the DFP update satisfies the hereditary property. At the same time, we will show that the search directions of the DFP method are conjugate.

Lemma 4.12 *Let $H_0 = I$. One has*

$$\sigma^i = H_{k+1} y^i \quad (i = 0, \ldots, k), \; k \geq 0, \tag{4.4}$$

and σ^0, ..., σ^k are mutually conjugate.

Proof: We will use *induction on k*. The reader may verify that (4.4) holds for $k = 0$. Induction assumption:

$$\sigma^i = H_k y^i \quad (i = 0, \ldots, k-1),$$

and σ^0, ..., σ^{k-1} are *mutually conjugate.*

We now use

$$\sigma^k = \lambda_k s^k = -\lambda_k H_k \nabla q(x^k),$$

to get

$$
\begin{aligned}
(\sigma^k)^T A\sigma^i &= -\lambda_k (H_k \nabla q(x^k))^T A\sigma^i \\
&= -\lambda_k \nabla q(x^k)^T H_k A\sigma^i.
\end{aligned}
$$

Now use the induction assumption $\sigma^i = H_k y^i \equiv H_k A\sigma^i$ $(i = 0, \ldots, k-1)$, to get:

$$(\sigma^k)^T A\sigma^i = \nabla q(x^k)^T \sigma^i \quad (i = 0, \ldots, k-1).$$

Since σ^0, ..., σ^{k-1} *mutually conjugate*, Theorem 4.7 implies that:

$$\nabla q(x^k)^T \sigma^i = 0 \quad (i = 0, \ldots, k-1).$$

Substituting we get

$$(\sigma^k)^T A\sigma^i = 0 \quad (i = 0, \ldots, k-1),$$

i.e. σ^0, ..., σ^k are *mutually conjugate*. We use this to prove the hereditary property. Note that

$$H_{k+1} y^i = H_k y^i + \frac{\sigma^k {\sigma^k}^T y^i}{{\sigma^k}^T y^k} - \frac{H_k y^k {y^k}^T H_k y^i}{{y^k}^T H_k y^k}.$$

We can simplify this, using:

$${\sigma^k}^T y^i = {\sigma^k}^T A\sigma^i = 0 \quad (i = 0, \ldots, k-1).$$

We get

$$H_{k+1} y^i = H_k y^i - \frac{H_k y^k {y^k}^T H_k y^i}{{y^k}^T H_k y^k}. \tag{4.5}$$

By the induction assumption $\sigma^i = H_k y^i$ $(i = 0, \ldots, k-1)$, and therefore

$${y^k}^T H_k y^i = {y^k}^T \sigma^i = {\sigma^k}^T A\sigma^i = 0 \quad (i = 0, \ldots, k-1).$$

Substituting in (4.5) we get the required

$$H_{k+1} y^i = H_k y^i = \sigma^i \quad (i = 0, \ldots, k-1).$$

□

DFP updates: discussion

- We have shown that the DFP updates preserve the required properties: positive definiteness, the secant condition, and the hereditary property.
- We have also shown that the DFP directions are mutually conjugate for quadratic functions.
- The DFP method can be applied to non-quadratic functions, but then the convergence of the DFP method is an open problem, even if the function is convex.
- In practice DFP performs quite well, but the method of choice today is the so-called BFGS update.

4.7.2 The BFGS update

The *Broyden-Fletcher-Goldfarb-Shanno (BFGS) update* is defined via

$$D_k = \frac{\tau_k \sigma^k {\sigma^k}^T - \sigma^k {y^k}^T H_k - H_k y^k {\sigma^k}^T}{{\sigma^k}^T y^k},$$

where

$$\tau_k = 1 + \frac{{y^k}^T H_k y^k}{{\sigma^k}^T y^k}.$$

i If $y_k^T \sigma_k > 0$, then $H_{k+1} = H_k + D_k$ is positive definite.

ii H_{k+1} satisfies the secant and hereditary conditions.

Exercise 4.25 *Consider the BFGS update:*

$$D_k = \frac{\tau_k \sigma^k {\sigma^k}^T - \sigma^k {y^k}^T H_k - H_k y^k {\sigma^k}^T}{{\sigma^k}^T y^k},$$

where $y^k := \nabla q(x^{k+1}) - \nabla q(x^k)$, $\sigma^k := x^{k+1} - x^k$, *and*

$$\tau_k = 1 + \frac{{y^k}^T H_k y^k}{{\sigma^k}^T y^k}.$$

(a) Show that if $y_k^T \sigma_k > 0$, *and* H_k *is positive definite, then* $H_{k+1} = H_k + D_k$ *is positive definite.*

(b) Show that the BFGS update satisfies the secant condition: $\sigma^k = H_{k+1} y^k$.

◁

How do we guarantee ${\sigma^k}^T y^k > 0$? Note that $\sigma^k = \lambda_k s^k$ and $y^k = \nabla f(x^{k+1}) - \nabla f(x^k)$. Thus we need to maintain

$$\nabla f(x^{k+1})^T s^k > \nabla f(x^k)^T s^k.$$

This can be guaranteed by using a special line-search.

The convergence of the BFGS method for convex functions was proved in 1976 by Powell. In practice, BFGS outperforms DFP and is currently the Quasi-Newton method of choice.

Exercise 4.26 *Consider the unconstrained optimization problem:*

$$\min 5x_1^2 + 2x_1x_2 + x_2^2 + 7.$$

See Figure 4.7 for a contour plot.

1. *Perform two iterations using the DFP Quasi-Newton method. Use exact line search and the starting point* $[1, 2]^T$. *Plot the iterates.*
2. *Perform two iterations using the BFGS Quasi-Newton method. Use exact line search and the starting point* $[1, 2]^T$. *Plot the iterates.*

◁

4.8 Stopping criteria

The stopping criteria is a relatively simple but essential part of the algorithms. If the algorithm generates both primal and dual solutions then the algorithm stops to iterate as the (relative) duality gap is less than a predefined threshold value $\epsilon > 0$. The duality gap is defined as

$$\text{primal obj. value} - \text{dual obj. value}$$

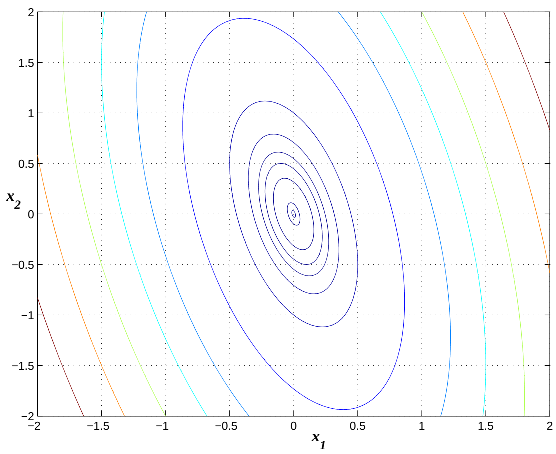


Figure 4.7: Contours of the objective function. Note that the minimum is at $[0,0]^T$.

while the relative duality gap is usually defined as

$$\frac{\text{primal obj. value} - \text{dual obj. value}}{1 + |\text{primal obj. value}|}.$$

In unconstrained optimization it happens often that one uses a primal algorithm and then there is no such absolute measure to show how close we are to the optimum. Usually the algorithm is then stopped as there is no sufficient improvement in the objective, or if the iterates are too close to each other or if the length of the gradient or the length of the Newton step in an appropriate norm is small. All these criteria can be scaled (relative to) some characteristic number describing the dimensions of the problem. We give just two examples here. The relative improvement of the objective is not sufficient and the algorithm is stopped if at two subsequent iterate x^k, x^{k+1}

$$\frac{|f(x^k) - f(x^{k+1})|}{1 + |f(x^k)|} \le \epsilon.$$

In Newton's method we conclude that we are close to the minimum of the function if the length of the full Newton step in the norm induced by the Hessian is small, i.e.

$$\begin{aligned} ||(\nabla^2 f(x^k))^{-1}\nabla f(x^k)||_{\nabla^2 f(x^k)} &= (\nabla f(x^k))^T(\nabla^2 f(x^k))^{-1}\nabla^2 f(x^k)(\nabla^2 f(x^k))^{-1}\nabla f(x^k) \\ &= (\nabla f(x^k))^T(\nabla^2 f(x^k))^{-1}\nabla f(x^k) \\ &\le \epsilon. \end{aligned}$$

This criteria can also be interpreted as the length of the gradient measured in the norm induced by the inverse Hessian. This last measure is used in interior point methods to control the Newton process efficiently.

Chapter 5

Algorithms for constrained optimization

5.1 The reduced gradient method

The reduced gradient method can be viewed as the logical extension of the gradient method to constrained optimization problems. We start with linearly constrained optimization problems.

To this end, consider the following linearly constrained convex problem

$$(LC) \quad \begin{array}{ll} \min & f(x) \\ \text{s.t.} & Ax = b, \\ & x \geq 0. \end{array} \tag{5.1}$$

Assumptions:

- f is continuously differentiable;
- Every subset of m columns of the $m \times n$ matrix A is linearly independent;
- each extreme point of the feasible set has at least m positive components (non-degeneracy assumption).

Exercise 5.1 *Prove that under the non-degeneracy assumption, every $x \in \mathcal{F}$ has at least m positive components.* ◁

If $x \in \mathcal{F}$, we call a set of m columns B of A a *basis* if $x_i > 0$ when column i is a column of B. We partition x into *basic* x_B and *non-basic variables* x_N such that the basic variables $x_B > 0$ correspond to the columns of B. Note that x_N does not have to be zero.

For simplicity of notation we assume that we can partition the matrix A as $A = [B, N]$. We partition x accordingly: $x^T = [x_B, x_N]^T$. Thus we can rewrite $Ax = b$ as

$$Bx_B + Nx_N = b,$$

such that

$$x_B = B^{-1}b - B^{-1}Nx_N.$$

(Recall that B^{-1} exists by assumption.)

Given $x \in \mathcal{F}$, we will choose B as the columns corresponding to the m *largest components* of x.

The basic variables x_B can now be eliminated from problem (5.1) to obtain the *reduced problem*

$$\begin{aligned} \min \quad & f_N(x_N) \\ \text{s.t.} \quad & B^{-1}b - B^{-1}Nx_N \geq 0, \\ & x_N \geq 0, \end{aligned}$$

where $f_N(x_N) = f(x) = f(B^{-1}b - B^{-1}Nx_N, x_N)$.

Note that any *feasible direction* s for problem (LC) in (5.1) must satisfy $As = 0$. If we write $s^T = [s_B^T, s_N^T]$ for a given basis B, the condition $As = 0$ can be rewritten as

$$Bs_B + Ns_N = 0.$$

We can solve this equation to obtain:

$$s_B = -(B)^{-1}Ns_N. \tag{5.2}$$

The choice of search direction

Recall that s is a *descent direction* of f at $x \in \mathcal{F}$ if and only if $\nabla f(x)^T s < 0$, which translates to

$$\nabla_B f(x)^T s_B + \nabla_N f(x)^T s_N < 0.$$

Here $\nabla_B f(x)$ is the gradient with respect to the basic variables, etc.

Substitute s_B from (5.4) to get:

$$\nabla f(x)^T s = \left(-\nabla_B f(x)^T (B)^{-1}N + \nabla_N f(x)^T\right) s_N.$$

Definition 5.1 *We call*

$$r := \left(-\nabla_B f(x)^T (B)^{-1}N + \nabla_N f(x)^T\right)^T \tag{5.3}$$

the reduced gradient *of f at x for the given basis B.*

Note that

$$\nabla f(x)^T s = r^T s_N.$$

In other words, the reduced gradient r plays the same role in the reduced problem as the gradient ∇f did in the original problem (LC). In fact, the reduced gradient is exactly the gradient of the function f_N with respect to x_N in the reduced problem.

Exercise 5.2 *Prove that $r = \nabla_N f_N(x_N)$, where $f_N(x_N) = f(x) = f(B^{-1}b - B^{-1}Nx_N, x_N)$.* ◃

Recall that the gradient method uses the search direction $s = -\nabla f(x)$. Analogously, the basic idea for the reduced gradient method is to use the negative reduced gradient $s_N = -r$ as search direction for the variables x_N, and then calculating the search direction for the variables x_B from

$$s_B = -(B)^{-1}Ns_N. \tag{5.4}$$

At iteration k of the algorithm we then perform a line search: find $0 \leq \lambda \leq \lambda_{\max}$ such that

$$x^{k+1} = x^k + \lambda s^k \geq 0,$$

where $\lambda_{\max}$ is an upper bound on the maximal feasible step length and is given by

$$\lambda_{\max} = \begin{cases} \min_{j:(s^k)_j<0} \frac{-(x^k)_j}{(s^k)_j} & \text{if } s^k \not\geq 0 \\ \infty & \text{if } s^k \geq 0 \end{cases} \tag{5.5}$$

This choice for $\lambda_{\max}$ guarantees that $x^{k+1} \geq 0$ and $Ax^{k+1} = b$.

Necessary modifications to the search direction

If we choose $s_N = -r$, then it may happen that

$$(s_N)_i < 0 \text{ and } (x_N)_i = 0$$

for some index i.

In this case $\lambda_{\max} = 0$ and we cannot make a step. One possible solution is as follows: for the nonbasic components set

$$(s_N)_i = \begin{cases} -(x_N)_i r_i & \text{if } r_i > 0 \\ -r_i & \text{if } r_i \leq 0 \end{cases} \tag{5.6}$$

Note that this prevents zero and 'very small' steps.

Convergence results

Since the reduced gradient method may be viewed as an extension of the gradient method, it may come as no surprise that analogous converge results hold for the reduced gradient method as for the gradient method. In this section we state some convergence results and emphasize the analogy with the results we have already derived for the gradient method (see Theorem 4.5).

Assume that the reduced gradient method generates iterates $\{x^k\}$, $k = 0, 1, 2, \ldots$

Theorem 5.2 *The search direction s^k at x^k is always a feasible descent direction unless $s^k = 0$. If $s^k = 0$, then x^k is a KKT point of problem (LC).*

Compare this to the gradient method where, by definition, $s^k = 0$ if and only if x^k is a stationary point ($\nabla f(x^k) = 0$).

Exercise 5.3 *Prove Theorem 5.2.* ◁

Theorem 5.3 *Any accumulation point of $\{x^k\}$ is a KKT point.*

Compare this to the gradient method where any accumulation point of $\{x^k\}$ is a *stationary point* under some assumptions (see Theorem 4.5).

The proof of Theorem 5.3 is beyond the scope of this course. A detailed proof is given in [2], Theorem 10.6.3.

The reduced gradient algorithm: a summary

To summarize, we give a statement of the complete algorithm.

1. **Initialization**

   Choose a starting point $x^0 \geq 0$ such that $Ax = b$. Let $k = 0$.

2. **Main step**

   [1.1] Form B from those columns of A that correspond to the m largest components of x^k. Define N as the remaining columns of A. Define x_B as the elements of x^k that correspond to B, and define x_N similarly.

   [1.2] Compute the reduced gradient r from (5.3).

   [1.3] Compute s_N from (5.6) and s_B from (5.4). Form s^k from s_B and s_N.

   [1.4] If $s^k = 0$, STOP (x^k is a KKT point).

3. **Line search**

   [2.1] Compute $\lambda_{\max}$ from (5.5).

   [2.2] Perform the line search

$$\lambda_k := \arg \min_{0 \le \lambda \le \lambda_{\max}} f\left(x^k + \lambda s^k\right).$$

   [2.3] Set $x^{k+1} = x^k + \lambda_k s^k$ and replace k by $k+1$.

   [2.4] Repeat the main step.

Remarks:

- During the algorithm the solution x^k is not necessarily a basic solution, hence positive coordinates in x_N^k may appear. These variables are usually referred to as *superbasic* variables.
- Recall that we have made a non-degeneracy assumption that is difficult to check in practice. If degeneracy occurs in practice, similar techniques as in the linear optimization case are applied to resolve degeneracy and prevent cycling.
- The convex simplex method is obtained as the specialization of the above reduced gradient scheme if the definition of the search direction s_N is modified. We allow only one coordinate j of s_N to be nonzero and defined as $s_j = -\frac{\partial f_N(x_N^k)}{\partial x_j} > 0$. The rest of the s_N coordinates is defined to be zero and $s_B = -B^{-1}Ns_N = -B^{-1}a_j s_j$, where a_j is the j-th column of the matrix A.
- The simplex method of LO is obtained as a further specialization of the convex simplex method. One assumes that the objective function is linear and the initial solution is a basic solution.

Example 5.4 [Reduced gradient method 1] Consider the following problem:

$$\begin{array}{lrcl} \min & x^2 & & \\ \text{s.t.} & x & \geq & 2 \\ & x & \geq & 0. \end{array}$$

We solve this problem by using the Reduced Gradient Method starting from the starting point $x^0 = 5$ with objective value 25. We start with converting the constraint in an equality-constraint:

$$\begin{array}{lrcl} \min & x^2 & & \\ \text{s.t.} & x - y & = & 2 \\ & x, y & \geq & 0. \end{array}$$

The value of the slack variable y^0 is 3. We therefore choose variable x as the basic variable. This results in $B = 1$ and $N = -1$. We eliminate the basic variable:

$$f_N(x_N) = f(B^{-1}b - B^{-1}Nx_N, x_N) = f(2 + y, y).$$

This gives us the following problem:

$$\begin{array}{lrcl} \min & (2+y)^2 & & \\ \text{s.t.} & 2 + y & \geq & 0 \\ & y & \geq & 0. \end{array}$$

Iteration 1

The search directions are:

$$\begin{aligned} s_N^0 = s_y^0 &= -\frac{\delta f_N(y^0)}{\delta y} = -(2(2 + y^0)) = -10, \\ s_B^0 = s_x^0 &= -B^{-1}Ns_N^0 = (-1) \cdot (-1) \cdot (-10) = -10. \end{aligned}$$

The new values of the variables are, depending on the step-length λ:

$$\begin{aligned} x^1 = x^0 + \lambda s_x^0 &= 5 - 10\lambda \\ y^1 = y^0 + \lambda s_y^0 &= 3 - 10\lambda \end{aligned}$$

which stay non-negative if $\lambda \leq \bar{\lambda} = \frac{3}{10}$.
We now have to solve the one-dimensional problem:

$$\min \quad (5 - 10\lambda)^2.$$

The minimum is attained when

$$-20(5 - 10\lambda) = 0,$$

i.e. when $\lambda = \frac{1}{2}$. Since the $\lambda = \frac{1}{2}$ is larger than $\bar{\lambda} = \frac{3}{10}$ that preserves the non-negativity of the variables, we have to take $\lambda = \frac{3}{10}$ as the step-length. This results in $x^1 = 2$ and $y^1 = 0$ with 4 as the objective value.

Iteration 2

Since $x > y$, we use the variable x as basic variable again. First we compute the search direction of y. Because $y^0 = 0$ the search direction has to be non-negative else it will get the value 0:

$$s_N^3 = s_y^3 = -2(2 + y^3) = -4.$$

This means:

$$\begin{aligned} s_N^3 &= s_y^3 = 0, \\ s_B^3 &= s_x^3 = 0. \end{aligned}$$

Thus the optimum point is $x^{opt} = 2$. *

Example 5.5 [Reduced gradient method 2] Consider the following problem:

$$\begin{aligned} \min \quad & x_1^2 + x_2^2 + x_3^2 + x_4^2 - 2x_1 - 3x_4 \\ \text{s.t.} \quad & 2x_1 + x_2 + x_3 + 4x_4 = 7 \\ & x_1 + x_2 + 2x_3 + x_4 = 6 \\ & x_1, x_2, x_3, x_4 \geq 0. \end{aligned}$$

We perform one iteration of the Reduced Gradient Method starting from the point $x^0 = (x_1^0, x_2^0, x_3^0, x_4^0)^T = (2, 2, 1, 0)^T$ with an objective value 5. At this point x^0 we consider x_1 and x_2 as basic variables. This results in

$$B = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix} \text{ and } N = \begin{pmatrix} 1 & 4 \\ 2 & 1 \end{pmatrix}.$$

We eliminate the basic variables to obtain the reduced problem:

$$\begin{aligned} f_N(x_N) &= f(B^{-1}b - B^{-1}Nx_N, x_N) \\ &= f(\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 7 \\ 6 \end{pmatrix} - \begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 4 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} x_3 \\ x_4 \end{pmatrix}, x_3, x_4) \\ &= f(1 + x_3 - 3x_4, 5 - 3x_3 + 2x_4, x_3, x_4). \end{aligned}$$

This results in the following problem:

$$\begin{aligned} \min \quad & (1 + x_3 - 3x_4)^2 + (5 - 3x_3 + 2x_4)^2 + x_3^2 + x_4^2 - 2(1 + x_3 - 3x_4) - 3x_4 \\ & 1 + x_3 - 3x_4 \geq 0 \\ & 5 - 3x_3 + 2x_4 \geq 0 \\ & x_3, x_4 \geq 0. \end{aligned}$$

Iteration 1

The search directions are:

$$\begin{aligned} s_N^0 = \begin{pmatrix} s_3^0 \\ s_4^0 \end{pmatrix} &= \begin{pmatrix} -\frac{\delta f_N(x_3^0)}{\delta x_3} \\ -\frac{\delta f_N(x_4^0)}{\delta x_4} \end{pmatrix} \\ &= \begin{pmatrix} -(2(1 + x_3^0 - 3x_4^0) - 6(5 - 3x_3^0 + 2x_4^0) + 2x_3^0 - 2) \\ -(-6(1 + x_3^0 - 3x_4^0) + 4(5 - 3x_3^0 + 2x_4^0) + 2x_4^0 + 3) \end{pmatrix} \\ &= \begin{pmatrix} 8 \\ 1 \end{pmatrix}. \end{aligned}$$

Because $x_4^0 = 0$ the search direction s_4^0 has to be non-negative. We see that this is true.

$$s_B^0 = \begin{pmatrix} s_1^0 \\ s_2^0 \end{pmatrix} \quad = \quad -B^{-1}Ns_N^0 = -\begin{pmatrix} 1 & -1 \\ -1 & 2 \end{pmatrix}\begin{pmatrix} 1 & 4 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 8 \\ 1 \end{pmatrix} = \begin{pmatrix} 5 \\ -22 \end{pmatrix}.$$

We now have to make a line search to obtain the new variables. These new variables as a function of the step length λ are:

$$\begin{aligned} x_1^1 = x_1^0 + \lambda s_1^0 &= 2 + 5\lambda \\ x_2^1 = x_2^0 + \lambda s_2^0 &= 2 - 22\lambda \\ x_3^1 = x_3^0 + \lambda s_3^0 &= 1 + 8\lambda \\ x_4^1 = x_4^0 + \lambda s_4^0 &= \lambda \end{aligned}$$

which stay non-negative if $\lambda \leq \bar{\lambda} = \frac{2}{22} \approx 0.09$.
We proceed by solving

$$\min \quad (2 + 5\lambda)^2 + (2 - 22\lambda)^2 + (1 + 8\lambda)^2 + \lambda_2 - 2(2 + 5\lambda) - 3\lambda.$$

This means

$$\begin{aligned} 10(2 + 5\lambda) - 44(2 - 22\lambda) + 16(1 + 8\lambda) + 2\lambda - 13 &= 0 \\ \lambda &= \frac{65}{1148} \approx 0.06 \qquad (\lambda < \bar{\lambda} = \frac{2}{22}). \end{aligned}$$

The minimizer $\lambda = \frac{65}{1148}$ is smaller than $\bar{\lambda} = \frac{2}{22}$, so non-negativity of the variables is preserved. Thus the new iterate is $x^1 = (x_1^1, x_2^1, x_3^1, x_4^1)^T = (2.28, 0.75, 1.45, 0.06)^T$ with an objective value of 3.13.

*

Exercise 5.4 *Perform two iterations of the reduced gradient method for the following linearly constrained convex optimization problem:*

$$\begin{aligned} \min \ & x_1^2 + x_2^4 + (x_3 - x_4)^2 \\ \text{s.t.} \ & x_1 + 2x_2 + 3x_3 + 4x_4 = 10 \\ & x \geq 0. \end{aligned}$$

Let the initial point be given as $x^0 = (1, 1, 1, 1)$ *and use* x_1 *as the initial basic variable.* ◁

5.2 Generalized reduced gradient (GRG) method

The reduced gradient method can be generalized to nonlinearly constrained optimization problems. Similarly to the linearly constrained case we consider the problem with equality constraints and non-negative variables as follows.

$$\begin{aligned} (NC) \quad \min \quad & f(x) \\ \text{s.t.} \quad & h_j(x) = 0, \ \ j = 1, \cdots, m \\ & x \geq 0, \end{aligned} \tag{5.7}$$

where the functions $f, h_1, \cdots, h_m$ supposed to be continuously differentiable. [1]

The basis idea is to replace the nonlinear equations by their linear Taylor approximation at the current value of x, and then apply the reduced gradient algorithm to the resulting problem.

We assume that the gradients of the constraint functions h_j are linearly independent at every point $x \geq 0$, and that each feasible x has at least m positive components. These assumptions ensure that we can always apply the reduced gradient algorithm to the linearized problem. The extra difficulty here is that — since the feasible region $\mathcal{F}$ is not convex — this procedure may produce iterates that lie outside $\mathcal{F}$, and then some extra effort is needed to restore feasibility.

[1] The problem (NC) is in general not convex. It is a (CO) problem if and only if the functions h_j are affine.

Let a feasible solution $x^k \geq 0$ with $h_j(x^k) = 0$ for all j be given. By assumption the Jacobian matrix of the constraints $H(x) = (h_1(x), \cdots, h_m(x))^T$ at each $x \geq 0$ has full rank and, for simplicity at the point x^k will be denoted by

$$A = JH(x^k).$$

Let us assume that a basis B, where $x_B^k > 0$ is given. Then a similar construction as in the linear case apply. We generate a reduced gradient search direction by virtually keeping the linearized constraints valid. This direction by construction will be in the null space of A. More specifically for the linearized constraints we have

$$H(x^k) + JH(x^k)(x - x^k) = 0 + A(x - x^k) = 0.$$

From this one has

$$Bx_B + Nx_N = Ax^k$$

and by introducing the notation $b = Ax^k$ we have

$$x_B = B^{-1}b - B^{-1}Nx_N$$

hence the basic variables x_B can be eliminated from the linearization of the problem (5.7) to result

$$\begin{aligned} \min \quad & f_N(x_N) \\ \text{s.t.} \quad & B^{-1}b - B^{-1}Nx_N \geq 0, \\ & x_N \geq 0. \end{aligned}$$

where $f_N(x_N) = f(x) = f(B^{-1}b - B^{-1}Nx_N, x_N)$. Using the notation

$$\nabla f(x)^T = ((\nabla_B f(x))^T, (\nabla_N f(x))^T),$$

the gradient of f_N, namely the *reduced gradient* can be expressed as

$$\nabla_N f(x)^T = -(\nabla_B f(x))^T B^{-1}N + (\nabla_N f(x))^T.$$

From this point on the generation of the search direction s proceeds in exactly the same way as in the linearly constrained case. Due to the nonlinearity of the constraints $H(x^{k+1}) = H(x^k + \lambda s) = 0$ will not hold in general. Hence something more has to be done to restore feasibility.

Special care has to be taken to control the step size. A larger step size might allow larger improvement of the objective but, on the other hand results in larger infeasibility of the constraints. A good compromise must be made.

In old versions of the GRG method Newton's method is applied to the nonlinear equality system $H(x) = 0$ from the initial point x^{k+1} to produce a next feasible iterate. In more recent implementations the reduced gradient direction is combined by a direction from the orthogonal subspace (the range space of A^T) and then a modified (nonlinear, discrete) line search is performed. These schemes are quite complicated and not discussed here in more detail.

Example 5.6 [Generalized reduced gradient method 1] We consider the following problem:

$$\begin{aligned} \min \quad & x_1^2 + x_2^2 + 12x_1 - 4x_2 \\ \text{s.t.} \quad & x_1^2 - 2x_2 = 0 \\ & x_1, x_2 \geq 0. \end{aligned}$$

We perform two steps of the Generalized Reduced Gradient Method starting from the point $x^0 = (x_1^0, x_2^0)^T = (4, 8)^T$ with an objective value of 96. We will plot the progress of the algorithm in Figure 5.1. At the point x^0 we consider x_2 as the basic variable. First we have to linearize the nonlinearly constraint:

$$A = (N, B) = JH(x^0) = (2x_1^0 \quad -2) = (-8 \quad -2). \qquad b = Ax^0 = (-8 \quad -2)\begin{pmatrix} 4 \\ 8 \end{pmatrix} = 16.$$

Now we eliminate the basic variable:

$$f_N(x_N) = f(B^{-1}b - B^{-1}Nx_N, x_N) = f(x_1, -\frac{1}{2} \cdot 16 + \frac{1}{2}x_1).$$

This leads us to the following problem:

$$\begin{array}{rrcl} \min & x_1^2 + (4x_1 - 8)^2 + 12x_1 - 4(4x_1 - 8) & & \\ & \text{s.t.} \quad 4x_1 - 8 & \geq & 0 \\ & x_1 & \geq & 0. \end{array}$$

Iteration 1

The search direction is:

$$\begin{aligned} s_N^0 = s_1^0 &= -\frac{\delta f_N(x_1^0)}{\delta x_1} = -(2x_1^0 + 8(4x_1 - 8) + 12 - 16)) = -68 \\ s_B^0 = s_1^0 &= -B^{-1}Ns_N^0 = \frac{1}{2} \cdot 8 \cdot -68 = -272. \end{aligned}$$

The new variables as a function of the step length λ are:

$$\begin{aligned} x_1^1 = x_1^0 + \lambda s_1^0 &= 4 - 68\lambda \\ x_2^1 = x_2^0 + \lambda s_2^0 &= 8 - 272\lambda \end{aligned}$$

which stay non-negative if $\lambda \leq \bar{\lambda} = \frac{1}{34}$.

We do this by solving

$$\min \quad (4 - 68\lambda)^2 + (8 - 272)^2 + 12(4 - 68\lambda) - 4(8 - 272\lambda)$$

This means

$$\begin{aligned} -136(4 - 68\lambda) - 544(8 - 272\lambda) - 816 + 1088 &= 0 \\ \lambda &= \frac{1}{34} \qquad (\lambda = \bar{\lambda}). \end{aligned}$$

This results in $x^1 = (x_1^1, x_2^1)^T = (2, 0)^T$. But due to the nonlinearity of the constraint, the constraint will not hold with these values. To find a solution for which the constraint will hold, we consider the x_N as a fixed variable. The x_B will change in a value for which the constraint holds, this means $x_B = 2$. The objective value is 24.

Iteration 2

Because x_2^1 stayed positive we now use x_1^1 as basic variable again. But first we have to linearize the nonlinearly constraint with the values of iteration 1:

$$A = JH(x^1) = (2x_1^1 \quad -2) = (4 \quad -2). \qquad b = Ax^1 = (4 \quad -2)\begin{pmatrix} 2 \\ 2 \end{pmatrix} = 4.$$

We eliminate the basic variable:

$$f_N(x_N) = f(B^{-1}b - B^{-1}Nx_N, x_N) = f(x_1, -\frac{1}{2} \cdot 4 + \frac{1}{2} \cdot 4 \cdot x_1) = f(x_1, 2x_1 - 2).$$

This gives us the following problem:

$$\begin{array}{rrcl} \min & x_1^2 + (2x_1 - 2)^2 + 12x_1 - 4(2x_1 - 2) & & \\ & \text{s.t.} \quad 2x_1 - 2 & \geq & 0 \\ & x_1 & \geq & 0. \end{array}$$

The search direction is:

$$\begin{aligned} s_N^1 = s_1^1 &= -\frac{\delta f_N(x_1^1}{\delta x_1} = -(2x_1^1 + 4(2x_1^1 - 2) + 12 - 8) = -16 \\ s_B^1 = s_2^1 &= -B^{-1}Ns_N^1 = -32. \end{aligned}$$

The new variables as a function of the step length λ are:

$$\begin{aligned} x_1^2 = x_1^1 + \lambda s_1^1 = 2 - 16\lambda \\ x_2^2 = x_2^1 + \lambda s_2^1 = 2 - 32\lambda \end{aligned}$$

which stay non-negative if $\lambda \leq \bar{\lambda} = \frac{2}{32}$.

Now we have to solve

$$\min \quad 2(2 - 16\lambda)^2 + (2 - 32\lambda)^2 + 12(2 - 16\lambda) - 4(2 - 32\lambda).$$

This means:

$$\begin{aligned} -32(2-16\lambda) - 64(2-32\lambda) - 192 + 128 &= 0 \\ \lambda &= \frac{1}{10} \qquad (\lambda > \bar{\lambda} = \frac{1}{16}). \end{aligned}$$

As we can see has $\lambda = \frac{1}{10}$ a larger value than $\bar{\lambda} = \frac{1}{16}$. In order to get non-negative values for the variables we have to use the value $\frac{1}{16}$ as step length. This gives us $x^2 = (x_1^2, x_2^2)^T = (1,0)^T$. To get variables for which the constraint holds, we take the x_N as fixed variable. This leads to $x^2 = (1, \frac{1}{2})^T$ with an objective value of $11\frac{1}{4}$.

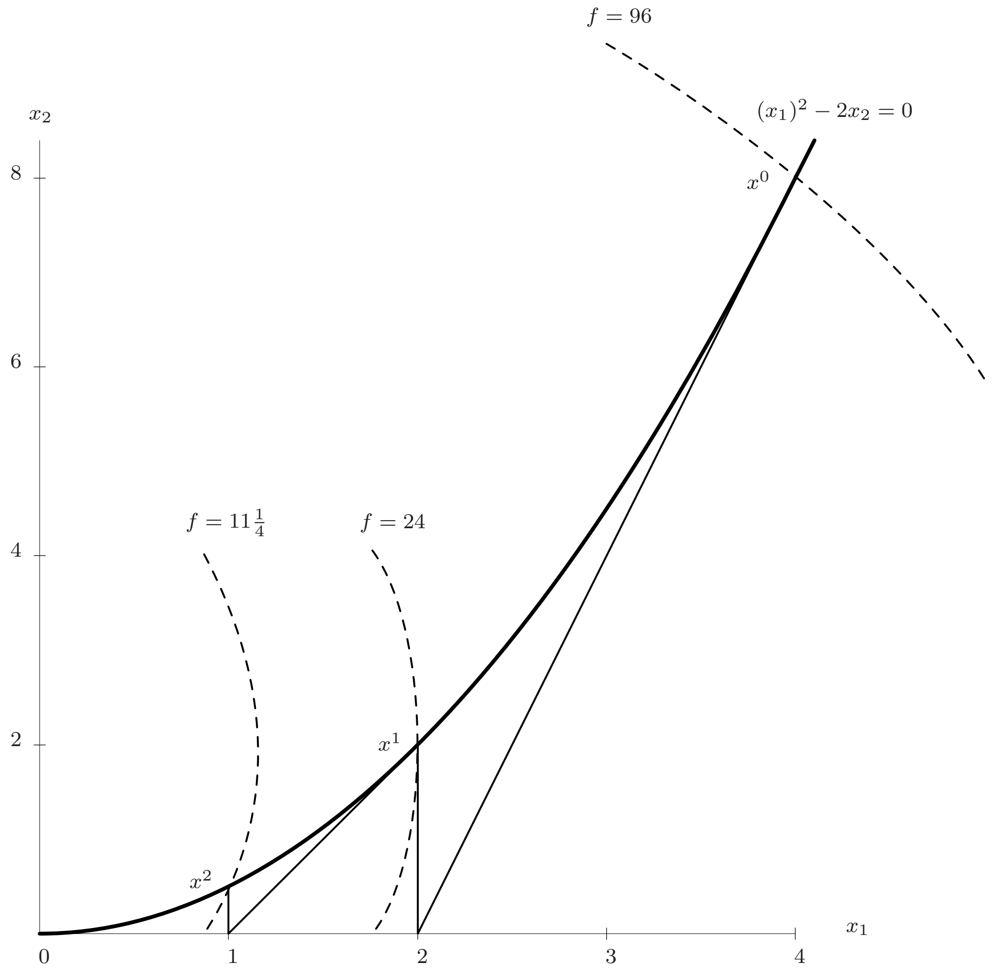


Figure 5.1: Illustration of Example 5.6.

*

Example 5.7 [Generalized reduced gradient method 2] We consider the following problem:

$$\begin{aligned} \min \quad & 2x_1^2 + 3x_2^2 & & \\ \text{s.t.} \quad & 3x_1^2 + 2x_2^2 &=& \; 20 \\ & x_1, x_2 &\geq& \; 0. \end{aligned}$$

We solve this problem by using three steps of the Generalized Reduced Gradient Method starting from the point $x^0 = (x_1^0, x_2^0)^T = (2,2)^T$ with an objective value of 20. At this point x^0 we consider x_1 as basic variable. First we have to linearize the nonlinearly constraint:

$$A = (B, N) = JH(x^0) = (6x_1^0 \;\; 4x_2^0) = (12 \;\; 8). \qquad b = Ax^0 = (12 \;\; 8)\begin{pmatrix} 2 \\ 2 \end{pmatrix}.$$

Now we eliminate the basic variables:

$$f_N(x_N) \quad = \quad f(B^{-1}b - B^{-1}Nx_N, x_N) = f(\frac{40}{12} - \frac{8}{12}x_2, x_2).$$

This leads us to the following problem:

$$\begin{array}{lrcl} \min & 2(\frac{40}{12} - \frac{8}{12}x_2)^2 + 3x_2^2 & & \\ \text{s.t.} & \frac{40}{12} - \frac{8}{12}x_2 & \geq & 0 \\ & x_2 & \geq & 0. \end{array}$$

Iteration 1 The search direction is:

$$\begin{aligned} s_N^0 = s_2^0 &= -\frac{\delta f_N(x_2^0)}{\delta x_2} = -(-\frac{32}{12}(\frac{40}{12} - \frac{8}{12}x_2^0) + 6x_2^0) = -\frac{20}{3} \\ s_B^0 = s_1^0 &= -B^{-1}Ns_N^0 = -\frac{1}{12} \cdot 8 \cdot -\frac{20}{3} = \frac{40}{9}. \end{aligned}$$

The new variables as a function of λ are:

$$x_1^1 = x_1^0 + \lambda s_1^0 = 2 + \frac{40}{9}\lambda$$
$$x_2^1 = x_2^0 + \lambda s_2^0 = 2 - \frac{20}{3}\lambda$$

which are non-negative as long as $\lambda \leq \bar{\lambda} = \frac{2}{\frac{20}{3}} = \frac{3}{10}$.

We do this by solving

$$\min \quad 2(2 + \frac{40}{9}\lambda)^2 + 3(2 - \frac{20}{3}\lambda)^2.$$

This means

$$\begin{aligned} \frac{160}{9}(2 + \frac{40}{9}\lambda) - \frac{120}{3}(2 - \frac{20}{3}\lambda) &= 0 \\ \lambda &= \frac{9}{70} \qquad (\lambda < \bar{\lambda} = \frac{3}{10}). \end{aligned}$$

This results in $x^1 = (x_1^1, x_2^1)^T = (\frac{18}{7}, \frac{8}{7})^T$. But due to the nonlinearity of the constraint, the constraint will not hold with these values. To find a solution for which the constraint will hold, we consider the x_N as a fixed variable. The x_B will change in a value for which the constraint holds, this means $x_B = 2.41$. The objective value is 15.52.

Iteration 2

Because x_1^1 stayed positive we use x_1 as basic variable again. First with the values of iteration 1 we linearize the nonlinearly constraint again:

$$A = JH(x^1) = (6x_1^1 \;\; 4x_2^1) = (14.45 \;\; 4.57). \quad b = Ax^1 = (14.45 \;\; 4.57)\begin{pmatrix} 2.41 \\ 1.14 \end{pmatrix} = 40.$$

We eliminate the basic variable:

$$f_N(x_N) = f(B^{-1}b - B^{-1}Nx_N, x_N) = f(2.77 - 0.32x_2, x_2).$$

This gives us the following problem:

$$\begin{array}{lrcl} \min & 2(2.77 - 0.32x_2)^2 + 3x_2^2 & & \\ \text{s.t.} & 2.77 - 0.32x_2 & \geq & 0 \\ & x_2 & \geq & 0. \end{array}$$

The search direction is:

$$\begin{aligned} s_N^1 = s_2^1 &= -\frac{\delta f_N(x_2^1}{\delta x_2} = -(-4 \cdot 0.32(2.77 - 0.32x_2^1) + 6x_2^1) = -3.78 \\ s_B^1 = s_1^1 &= -B^{-1}Ns_N^1 = 1.2. \end{aligned}$$

The new variables, depending on the step length λ, are:

$$x_1^2 = x_1^1 + \lambda s_1^1 = 2.41 + 1.20\lambda$$
$$x_2^2 = x_2^1 + \lambda s_2^1 = 1.14 - 3.78\lambda$$

which stay non-negative if $\lambda \leq \bar{\lambda} = \frac{1.14}{3.78} \approx 0.30$.

Now we have to solve

$$\min \quad 2(2.41 + 1.20\lambda)^2 + 3(1.14 - 3.78\lambda)^2.$$

This means:

$$\begin{aligned} 4.80(2.41 + 1.20\lambda) - 22.68(1.14 - 3.78\lambda) &= 0 \\ \lambda &= 0.156 \qquad (\lambda < \bar{\lambda} \approx 0.3). \end{aligned}$$

This gives us $x^2 = (x_1^2, x_2^2)^T = (2.6, 0.55)^T$. To get variables for which the constraint holds, we take the x_N as fixed variable. This leads to $x^2 = (2.52, 0.55)^T$ with an objective value of 13.81.

Iteration 3

Again we can use x_1 as basic variable. We start this iteration with linearization of the constraint:

$$A = JH(x^2) = (6x_1^2 \;\; 4x_2^2) = (15.24 \;\; 2.2). \qquad b = Ax^2 = (15.24 \;\; 2.2) \begin{pmatrix} 2.54 \\ 0.55 \end{pmatrix} = 39.9.$$

Eliminating the basic variable:

$$f_N(x_N) = f(2.62 - 0.14x_2, x_2).$$

This gives us the following problem:

$$\begin{aligned} \min \quad & 2(2.62 - 0.14x_2)^2 + 3x_2^2 \\ \text{s.t.} \quad & 2.62 - 0.14x_2 \geq 0 \\ & x_2 \geq 0. \end{aligned}$$

Search directions:

$$\begin{aligned} s_N^2 = s_2^2 &= -\frac{\delta f_N(x_2^2}{\delta x_2} = -(-0.56(2.62 - 0.14x_2^2) + 6x_2^2) = -1.88 \\ s_B^2 = s_1^2 &= -B^{-1}Ns_N^2 = 0.27. \end{aligned}$$

New variables as a function of λ:

$$\begin{aligned} x_1^3 = x_1^2 + \lambda s_1^2 = 2.52 + 0.27\lambda \\ x_2^3 = x_2^2 + \lambda s_2^2 = 0.55 - 1.88\lambda \end{aligned}$$

which stay non-negative if $\lambda \leq \bar{\lambda} = \frac{0.55}{1.88} \approx 0.293$.

Now we solve

$$\min \quad 2(2.54 + 0.27\lambda)^2 + 3(0.55 - 1.88\lambda)^2.$$

This means:

$$\begin{aligned} 1.08(2.52 + 0.27\lambda) - 5.64(0.55 - 1.88\lambda) &= 0 \\ \lambda &= 0.161. \end{aligned}$$

This gives us the variables $x_1^3 = 2.58$ and $x_2^3 = 0.25$. Correcting the x_B results in $x^3 = (2.57, 0.25)^T$ with objective value 13.39. *

Exercise 5.5 *Perform one iteration of the generalized reduced gradient method to solve the following nonlinearly constrained convex optimization problem:*

$$\begin{aligned} \min \quad & x_1^2 + x_2^4 + (x_3 - x_4)^2 \\ \text{s.t.} \quad & x_1^2 + x_2^2 + x_3^2 + x_4^2 \leq 4 \\ & x \geq 0. \end{aligned}$$

Let the initial point be given as $x^0 = (1, 1, 1, 1)$. ◁

(You might need MAPLE or MATLAB to make the necessary calculations.)

Bibliography

[1] Anstreicher, K.M. (1990), A Standard Form Variant, and Safeguarded Linesearch, for the Modified Karmarkar Algorithm, *Mathematical Programming* 47, 337–351.

[2] M.S. Bazarraa, H.D. Sherali and C.M. Shetty, *Nonlinear Programming: Theory and Algorithms*, John Wiley and Sons, New York (1993).

[3] D.P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Belmont, MA 02178-9998, (1995).

[4] V. Chvátal, *Linear Programming*, (W.H. Freeman and Company, 1983).

[5] R. W. Cottle and S-M. Guu. Two Characterizations of Sufficient Matrices. *Linear Algebra and Its Applications*, ??–??, 199?.

[6] R. W. Cottle, J.-S. Pang and V. Venkateswaran. Sufficient Matrices and the Linear Complementarity Problem. *Linear Algebra and Its Applications*, 114/115:231-249, 1989.

[7] R. W. Cottle, J.-S. Pang and R. E. Stone. *The Linear Complementarity Problem.* Academic, Boston, 1992.

[8] Duffin, R.J., Peterson, E.L. and Zener, C. (1967), *Geometric Programming*, John Wiley & Sons, New York.

[9] P.E. Gill, W. Murray and M.H. Wright, *Practical Optimization* Academic Press, London, (1981).

[10] P.E. Gill, W. Murray and M.H. Wright, *Numerical Linear Algebra and Optimization, Vol.1.* Addison Wiley P.C. New York, (1991).

[11] R.A. Horn and C.R. Jonson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK (1985).

[12] Han, C.–G., Pardalos, P.M. and Ye, Y. (1991), On Interior–Point Algorithms for Some Entropy Optimization Problems, Working Paper, Computer Science Department, The Pennsylvania State University, University Park, Pennsylvania.

[13] D. den Hertog, (1994), *Interior Point Approach to Linear, Quadratic and Convex Programming: Algorithms and Complexity*, Kluwer A.P.C., Dordrecht.

[14] D. den Hertog, C. Roos and T. Terlaky, The Linear Comlementarity Problem, Sufficient Matrices and the Criss–Cross Method, (1993) *Linear Algebra and Its Applications*, 187. 1–14.

[15] Hertog, D. den, Jarre, F., Roos, C. and Terlaky, T. (1995), A Sufficient Condition for Self-Concordance with Application to Some Classes of Structured Convex Programming Problems, *Mathematical Programming* 69, 75–88.

[16] J.-B. Hirriart–Urruty and C. Lemarèchal, *Convex Analysis and Minimization Algorithms I and II*, Springer Verlag, Berlin, Heidelberg (1993).

[17] R. Horst, P.M. Pardalos and N.V. Thoai, (1995) *Introduction to Global Optimization*, Kluwer A.P.C. Dordrecht.

[18] P. Kas and E. Klafszky, (1993) On the Dduality of the Mixed Entropy Programming, *Optimization* 27. 253–258.

[19] E. Klafszky and T. Terlaky, (1992) Some Generalizations of the Criss–Cross Method for Quadratic Programming, *Mathemathische Operationsforschung und Statistics ser. Optimization* 24. 127–139.

[20] Jarre, F. (1990), Interior-point Methods for Classes of Convex Programs. Technical Report SOL 90–16, Systems Optimization Laboratory, Department of Operations Research, Stanford University, Stanford, California.

[21] Jarre, F. (1996), Interior-point Methods for Convex Programming. In T. Terlaky (ed.), Interior-point Methods for Mathematical Programming, Kluwer A.P.C., Dordrecht, pp. 255–296.

[22] Jarre, F. (1994), Interior-point Methods via Self-concordance or Relative Lipschitz Condition. Habilitationsschrift, Würzburg, Germany.

[23] Karmarkar, N.K. (1984), A New Polynomial–Time Algorithm for Linear Programming, *Combinatorica* 4, 373–395.

[24] Klafszky, E. (1976), Geometric Programming, Seminar Notes 11.1976, Hungarian Committee for Systems Analysis, Budapest.

[25] Kortanek, K.O. and No, H. (1992), A Second Order Affine Scaling Algorithm for the Geometric Programming Dual with Logarithmic Barrier, *Optimization* 23, 501–507.

[26] D.C. Lay *Linear Algebra and Its Applications* Addision Wiley (1994).

[27] F. Lootsma, *Algorithms for Unconstrained Optimization* Dictaat Nr. a85A, en WI-385, Fac. TWI, TU Delft.

[28] F. Lootsma, *Duality in Non–Linear Programming* Dictaat Nr. a85D, Fac. TWI, TU Delft.

[29] F. Lootsma, *Algorithms fo Constrained Optimization* Dictaat Nr. a85B, Fac. TWI, TU Delft.

[30] H.M. Markowitz, (1956), The Optimization of a Quadratic Function Subject to Linear Constraints, *Naval Research Logistics Quarterly* **3**, 111-133.

[31] H.M. Markowitz, (1959), *Portfolio Selection, Efficient Diversification of Investments*, Cowles Foundation for Research in Economics at Yale University, Monograph 16, John Wiley & Sons, New York.

[32] Nemirovsky, A.S. (1999), *Convex Optimization in Engineering*, Dictaat Nr. WI-485, Fac. ITS/TWI, TU Delft.

[33] Nesterov, Y.E. and Nemirovsky, A.S. (1989), Self–Concordant Functions and Polynomial Time Methods in Convex Programming, Report, Central Economical and Mathematical Institute, USSR Academy of Science, Moscow, USSR.

[34] Nesterov, Y.E. and Nemirovsky, A.S. (1994), *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia.

[35] Peterson, E.L., and Ecker, J.G. (1970), Geometric Programming: Duality in Quadratic Programming and l_p Approximation I, in: H.W. Kuhn and A.W. Tucker (eds.), *Proceedings of the International Symposium of Mathematical Programming*, Princeton University Press, New Jersey.

[36] Peterson, E.L., and Ecker, J.G. (1967), Geometric Programming: Duality in Quadratic Programming and l_p Approximation II, *SIAM Journal on Applied Mathematics* 13, 317–340.

[37] Peterson, E.L., and Ecker, J.G. (1970), Geometric Programming: Duality in Quadratic Programming and l_p Approximation III, *Journal of Mathematical Analysis and Applications* 29, 365–383.

[38] R.T. Rockafellar, *Convex Analysis*, Princeton, New Jersey, Princeton University Press (1970).

[39] C. Roos and T. Terlaky, *Introduction to Linear Optimization* Dictaat WI187, Fac. ITS/TWI, TU Delft (1997).

[40] C. Roos, T. Terlaky and J.-Ph. Vial (1997), *Theory and Algorithms for Linear Optimization: An Interior Point Approach.* John Wiley and Sons.

[41] S.S. Sapatnekar (1992), A convex programming approach to problems in VSLI designs. Ph. D. Thesis. University of Illinois at Urbana Campaign. 70–100.

[42] J. Stoer and C. Witzgall, *Convexity and Optimization in Finite Dimensions I*, Springer Verlag, Berlin, Heidelberg (1970).

[43] Terlaky, T. (1985), On l_p programming, *European Journal of Operational Research* 22, 70–100.

[44] Vandenberghe, L. and Boyd, S. (1994) Semidefinite Programming, Report December 7, 1994, Information Systems Laboratory, Electrical Engineering Department, Stanford University, Stanford CA 94305.

[45] Ye, Y. and Potra, F. (1990), An Interior–Point Algorithm for Solving Entropy Optimization Problems with Globally Linear and Locally Quadratic Convergence Rate, Working Paper Series No. 90–22, Department of Management Sciences, The University of Iowa, Iowa City, Iowa. To Appear in *SIAM Journal on Optimization*.

Index