Optimization

Lecture Notes

Athanasios P. Liavas
School of Electrical and Computer Engineering
Technical University of Crete
Email: liavas@telecom.tuc.gr

European Union
European Social Fund

Operational Programme
Human Resources Development,
Education and Lifelong Learning

Co-financed by Greece and the European Union

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη

"Support for the Internationalization of Higher Education, School of Electrical and Computer Engineering, Technical University of Crete" (MIS 5150766), under the call for proposals "Support of the Internationalization of Higher Education- Technical University of Crete" (EDULLL 153). The project is co-financed by Greece and the European Union (European Social Fund-ESF) through the Operational Programme "Human Resources Development, Education and Lifelong Learning 2014-2020."



Operational Programme
Human Resources Development,
Education and Lifelong Learning

Co-financed by Greece and the European Union



Introduction

These notes are material that is developed during the teaching of the course "Optimization." Optimization is an area of Applied Mathematics with an extremely large impact on Engineering Sciences, Economics, Operational Research etc.

The area of (Convex) Optimization is extremely active, with a wealth of contributions by members of various scientific communities, such as Mathematicians, Engineers, Economists.

In this course, we will cover a relatively small subset of the region, which could be characterized as introductory to the area of convex optimization.

Coverage will include

1. brief review of real function theory concepts,
2. elements of the theory of convex sets and convex functions,
3. definition of basic concepts of (convex) optimization problems,
4. extraction of optimality conditions,
5. definition of unconstrained convex optimization problems and the steepest descent and Newton algorithms,
6. definition of convex optimization problems with linear constraints and solution by a generalization of the Newton algorithm,
7. definition of convex optimization problems and solution by interior point method.

The following excellent books have significantly influenced the formatting of this material:

1. S. Boyd and L. Vandenberghe. Convex Optimization. Cambridge University Press, 2004.
2. M. Bazaraa, H. Sherali, C. Shetty. Nonlinear Programming. Theory and Algorithms. Wiley, 2nd Edition, 1993.
3. A. Beck. Introduction to Nonlinear Optimization. SIAM, 2014.
4. D. Bertsekas. Nonlinear Programming. Athena Scientific, 2nd Edition, 1999.

A very useful book covering vector calculus is the following:

1. J. Marsden and A. Tromba, Vector Calculus, W. H. Freeman; Sixth edition (December 16, 2011).

Your comments and suggestions are most welcome.

The instructor,

Athanasios P. Liavas

Chapter 1

Euclidean spaces

In this chapter, we will briefly present some basic elements of the theory of Euclidean spaces. The book by Marsden and Tromba contains an extensive description and is an excellent source.

We assume that the reader is familiar with the n-dimensional Euclidean space $\mathbb{R}^n$, with $n \in \mathbb{N}$. If $n = 1$, then the corresponding space will be denoted as $\mathbb{R}$. In the remainder of the chapter, we generally assume that we work in $\mathbb{R}^n$.

1.1 Points - Vectors

Definition 1.1.1. A **vector** is an ordered n-tuple $(x_1, x_2, \ldots, x_n)$, with $x_i \in \mathbb{R}$, for $i = 1, \ldots, n$.

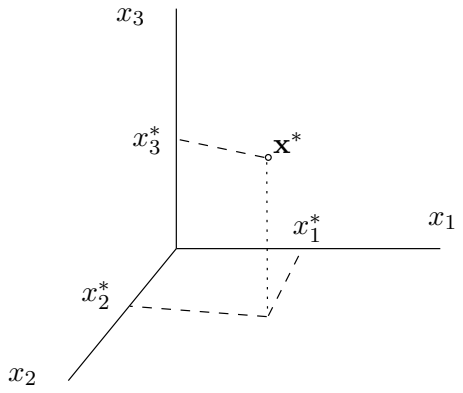


Figure 1.1: Point in three-dimensional space.

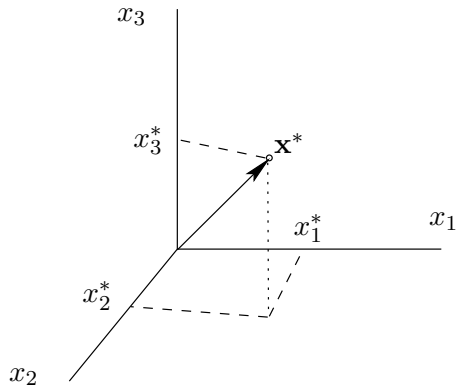


Figure 1.2: Vector in three-dimensional space.

Each vector will be denoted in three equivalent ways. More specifically, the three-dimensional vector $\mathbf{v} = (x, y, z)$ will be denoted as

$$\mathbf{v} = (x, y, z) = \begin{bmatrix} x \\ y \\ z \end{bmatrix} = [\, x \;\; y \;\; z]^T \,, \tag{1.1}$$

where the symbol $[\,\cdot\,]^T$ means **transpose**.

Each vector corresponds to a point in n-dimensional space. For example, in Figure 1.1, we draw the point $\mathbf{x}^* = (x_1^*, x_2^*, x_3^*)$.

A more "geometric" representation of a vector $\mathbf{x}^* = (x_1^*, x_2^*, x_3^*)$ is given by the directed line segment, starting at the origin of the axes, i.e., the point $\mathbf{0} = (0, 0, 0)$, and ending at the point $\mathbf{x}^*$ (see Figure 1.2).

1.1.1 Vector Operations

Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ be vectors. Their sum $\mathbf{z} = \mathbf{x} + \mathbf{y}$ is defined as the vector $\mathbf{z} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n)$.

Schematically, the sum of the vectors $\mathbf{x}$ and $\mathbf{y}$ is represented by the diagonal of the parallelogram whose sides are the vectors $\mathbf{x}$ and $\mathbf{y}$ (see Fig. 1.3, for $n = 2$).

The vector $\mathbf{y} - \mathbf{x} = (y_1 - x_1, y_2 - x_2, \ldots, y_n - x_n)$ is the vector we should add to $\mathbf{x}$ to get $\mathbf{y}$ (see Figure 1.3, for $n = 2$).

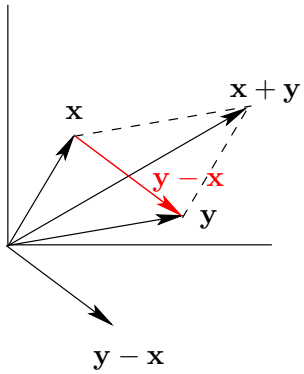


Figure 1.3: Vector addition and subtraction.

Let $\mathbf{x} = (x_1, \ldots, x_n)$ be a vector. Then, the vector $a\mathbf{x} = (ax_1, \ldots, ax_n)$, with $a \in \mathbb{R}$, is called a **scalar multiple** of $\mathbf{x}$. If $a > 0$, then $a\mathbf{x}$ has the same direction as $\mathbf{x}$ while, if $a < 0$, then $a\mathbf{x}$ has direction opposite to that of $\mathbf{x}$.

1.1.2 Inner product and Norm

Definition 1.1.2. Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be vectors. Their inner product is defined as follows

$$\langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y} = \mathbf{y}^T \mathbf{x} = \sum_{i=1}^{n} x_i y_i. \tag{1.2}$$

Definition 1.1.3. The Euclidean norm of the vector $\mathbf{x} = (x_1, \ldots, x_n)$ is the non-negative real number defined as

$$\|\mathbf{x}\|_2 = \langle \mathbf{x}, \mathbf{x} \rangle^{\frac{1}{2}} = \left(x_1^2 + \cdots + x_n^2 \right)^{\frac{1}{2}}. \tag{1.3}$$

The following important properties can be proved.

1. If $\mathbf{x} \in \mathbb{R}^n$, then $\|\mathbf{x}\|_2 \geq 0$, with equality if, and only if, $\mathbf{x} = \mathbf{0}$.
2. If $\mathbf{x} \in \mathbb{R}^n$ and $a \in \mathbb{R}$, then $\|a\mathbf{x}\|_2 = |a| \, \|\mathbf{x}\|_2$.
3. If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then $\|\mathbf{x} + \mathbf{y}\|_2 \leq \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2$, with equality if, and only if, $\mathbf{x}$ and $\mathbf{y}$ are nonnegative scalar multiples of each other.

Any function $\| \cdot \| : \mathbb{R}^n \to \mathbb{R}_+$ satisfying the above conditions is called a norm function on $\mathbb{R}^n$.

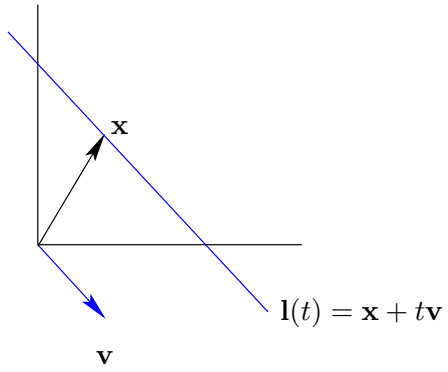


Figure 1.4: Line parallel to vector $\mathbf{v}$ passing through point $\mathbf{x}$.

Theorem 1.1.1. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be nonzero vectors. Then,

$$\langle \mathbf{x}, \mathbf{y} \rangle = \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \cos \angle(\mathbf{x}, \mathbf{y}), \tag{1.4}$$

where $\angle(\mathbf{x}, \mathbf{y})$ is the angle between the vectors $\mathbf{x}$ and $\mathbf{y}$.

Proof. See a good book on Calculus of Several Variables. □

Corollary 1.1.1. (*Cauchy–Schwarz inequality*) Let $\mathbf{x}$ and $\mathbf{y}$ be vectors in $\mathbb{R}^n$. Then,

$$|\langle \mathbf{x}, \mathbf{y} \rangle| \leq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2 \tag{1.5}$$

with equality if, and only if, $\mathbf{x}$ and $\mathbf{y}$ are linearly dependent.

Proof. For nonzero vectors, the result follows from Theorem 1.1.1 and the fact that $|\cos(\theta)| \leq 1$, for every angle θ. If either vector is zero, the inequality and the equality statement follow immediately. □

Definition 1.1.4. The vectors $\mathbf{x}$ and $\mathbf{y}$ in $\mathbb{R}^n$ are called orthogonal if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$.

If the vectors $\mathbf{x}$ and $\mathbf{y}$ are orthogonal, then we write $\mathbf{x} \perp \mathbf{y}$.

1.1.3 Lines

Let $\mathbf{x}, \mathbf{v} \in \mathbb{R}^n$ be vectors. The points described by the relationship

$$\mathbf{l}(t) = \mathbf{x} + t\mathbf{v},$$

for $t \in \mathbb{R}$, define the **line** which passes through point $\mathbf{x}$ and is parallel to the vector $\mathbf{v}$ (see Figure 1.4).

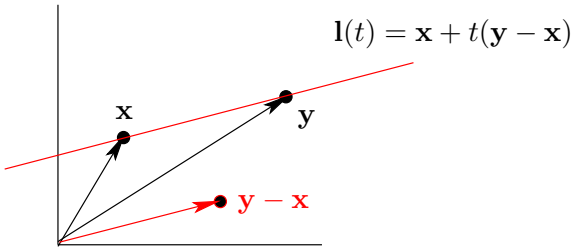


Figure 1.5: A straight line that passes through the points $\mathbf{x}$ and $\mathbf{y}$.

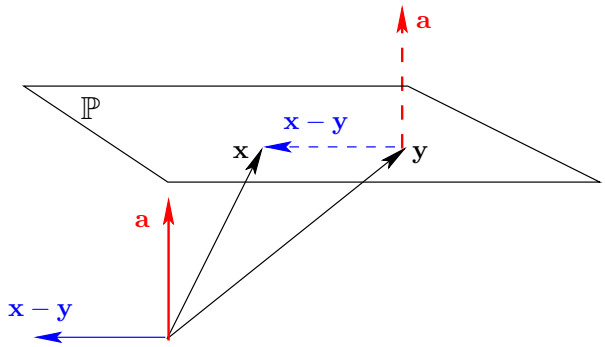


Figure 1.6: Hyperplane passing through $\mathbf{y}$ perpendicular to $\mathbf{a}$.

1.1.4 Straight line passing through the points x and y

Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ be vectors. The points defined by the relation

$$\mathbf{l}(t) = \mathbf{x} + t(\mathbf{y} - \mathbf{x}), \quad \text{for } t \in \mathbb{R}, \tag{1.6}$$

define the line that passes through the points $\mathbf{x}$ and $\mathbf{y}$ (see Figure 1.5).

We observe that, for $t = 0$, $\mathbf{l}(t) = \mathbf{x}$ while, for $t = 1$, $\mathbf{l}(t) = \mathbf{y}$. An alternative description of $\mathbf{l}(t)$ is as follows:

$$\mathbf{l}(t) = (1 - t)\mathbf{x} + t\mathbf{y}. \tag{1.7}$$

For $t \in [0, 1]$, relation (1.7) defines the line segment connecting the points $\mathbf{x}$ and $\mathbf{y}$.

1.1.5 Hyperplane perpendicular to a vector

An **affine set** $\mathbb{P}$ is a set of points such that the entire straight line through any two points of $\mathbb{P}$ belongs to $\mathbb{P}$. That is, if $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{P}$ and $\theta \in \mathbb{R}$, then $\theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2 \in \mathbb{P}$.

Let $\mathbb{P}$ be a hyperplane in $\mathbb{R}^n$, $\mathbf{y}$ a point of $\mathbb{P}$, and $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$ a vector perpendicular to $\mathbb{P}$ (see Fig. 1.6). Then, for every $\mathbf{x} \in \mathbb{P}$, the vector $\mathbf{x} - \mathbf{y}$ is parallel to $\mathbb{P}$. Therefore, the vector $\mathbf{a}$ is perpendicular to $\mathbf{x} - \mathbf{y}$.

The mathematical description of the points of $\mathbb{P}$ is as follows:

$$\begin{aligned}\mathbb{P} &= \{\mathbf{x} \in \mathbb{R}^n \,|\, \langle \mathbf{x} - \mathbf{y}, \mathbf{a} \rangle = 0\} \\ &= \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{x}^T\mathbf{a} = \mathbf{y}^T\mathbf{a}\} \\ &= \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{x}^T\mathbf{a} = c\},\end{aligned} \tag{1.8}$$

where $c := \mathbf{y}^T\mathbf{a}$.

In general, for $\mathbf{a} \neq \mathbf{0}$, we say that the equation (1.8) defines the **hyperplane** in $\mathbb{R}^n$ which passes through the point $\mathbf{y}$ and is perpendicular to the vector $\mathbf{a}$.

If $n = 2$, then the hyperplane is simply a straight line, while, if $n = 3$, then the hyperplane is a plane. In general, a hyperplane in $\mathbb{R}^n$ is an affine set of dimension $(n-1)$ (equivalently, with $(n-1)$ degrees of freedom). As we shall see in the sequel, expression (1.8) is very important.

Exercise: Prove that the set $\mathbb{P}$, defined by the relation (1.8), is an affine set.

1.1.6 Open and Closed Sets

Definition 1.1.5. Let $\mathbf{x} \in \mathbb{R}^n$ and $r \in \mathbb{R}_{++}$. The set

$$\mathbb{B}(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{x}\|_2 < r\} \tag{1.9}$$

is called the (Euclidean) **ball** with center $\mathbf{x}$ and radius r.

The set $\mathbb{B}(\mathbf{x}, r)$ is also called a **neighborhood** of $\mathbf{x}$.

Definition 1.1.6. The point $\mathbf{x} \in \mathbb{S}$ is called **interior point** of $\mathbb{S}$ if $\mathbb{S}$ contains a neighborhood of $\mathbf{x}$, that is, if there exists $r_{\mathbf{x}} \in \mathbb{R}_{++}$ such that $\mathbb{B}(\mathbf{x}, r_{\mathbf{x}}) \subset \mathbb{S}$.

Definition 1.1.7. The set of interior points of the set $\mathbb{S}$ is called the **interior** of $\mathbb{S}$.

Definition 1.1.8. A point $\mathbf{x}$ is called a **boundary point** of the set $\mathbb{S}$ if *every* neighborhood of $\mathbf{x}$ contains at least one point which belongs to $\mathbb{S}$ and at least one point which does not belong to $\mathbb{S}$. The set of all the boundary points of $\mathbb{S}$ is called the **boundary** of $\mathbb{S}$.

A boundary point of the set $\mathbb{S}$ may not belong to $\mathbb{S}$.

Definition 1.1.9. A set is called **open** if it contains only its interior points or, equivalently, if it contains none of its boundary points.

Definition 1.1.10. A set is called **closed** if it contains its boundary.

It can be shown that a set is open if, and only if, its complement is closed.

A set may be neither open nor closed (give examples).

Definition 1.1.11. A set is called **bounded** if it is contained in a ball of finite radius.

Definition 1.1.12. A set is called **compact** if it is closed and bounded.

Theorem 1.1.2. (*Theorem Weierstrass*) Let $\mathbb{U} \subseteq \mathbb{R}^n$ be a compact set and let $f : \mathbb{U} \to \mathbb{R}$ be a continuous function. Then, there exists a point $\mathbf{x}_0 \in \mathbb{U}$ such that $f(\mathbf{x}_0) \leq f(\mathbf{x})$, for every $\mathbf{x} \in \mathbb{U}$.

Proof. The proof requires the use of some advanced Real Analysis concepts and is beyond the scope of this course. □

In words, this important theorem says that, if $\mathbb{U}$ is compact and f is continuous, then there is a point of $\mathbb{U}$ which attains the minimum value of f in $\mathbb{U}$. If $\mathbb{U}$ is not closed or bounded, then we can easily construct examples in which the minimum value of f in $\mathbb{U}$ is not attained.

1.2 Functions

Definition 1.2.1. Let $\mathbb{U} \subseteq \mathbb{R}^n$. The transformation $f : \mathbb{U} \to \mathbb{R}^m$, which maps each element $\mathbf{x} \in \mathbb{U}$ to one element of $\mathbb{R}^m$, is called a **function** from $\mathbb{U}$ into $\mathbb{R}^m$.

If in the definition 1.2.1 we have $m = 1$, then f is called a real function of n variables and is denoted as $f(\mathbf{x})$ or as $f(x_1, \ldots, x_n)$.

For example, the function

$$f(x_1, x_2) = \sqrt{x_1^2 + x_2^2}, \text{ with } x_1, x_2 \in \mathbb{R},$$

is a real function of two variables, while

$$f(x_1, x_2, x_3) = x_1(x_2^2 + \log_2 x_3), \text{ with } x_1, x_2 \in \mathbb{R}, x_3 \in \mathbb{R}_{++},$$

is a real function of three variables.

Definition 1.2.2. Let $f : \mathbb{U} \subseteq \mathbb{R}^n \to \mathbb{R}$. Then, the set

$$\mathbb{G}^f = \{(\mathbf{x}, f(\mathbf{x})) \in \mathbb{R}^{n+1} \,|\, \mathbf{x} \in \mathbb{U}\} \tag{1.10}$$

is called the **graph** of f.

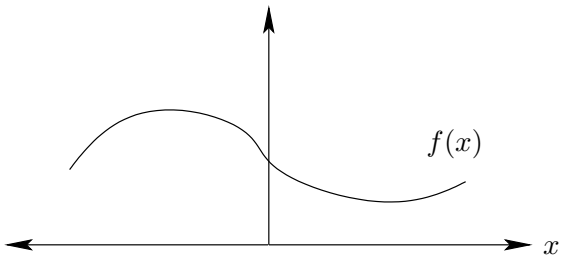


Figure 1.7: Graph of a real function of one variable.

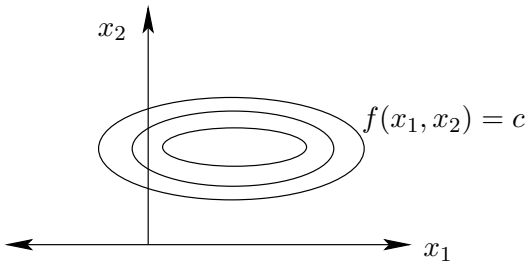


Figure 1.8: Level sets of a real function of two variables.

If $n = 1$, then the graph is a curve in $\mathbb{R}^2$ (see Figure 1.7), while, if $n = 2$, then the graph is a surface in $\mathbb{R}^3$.

If $n > 2$, then the graph is a hypersurface in $\mathbb{R}^{n+1}$, which cannot be visualized.

1.2.1 Curves

Definition 1.2.3. A function $\mathbf{f} : \mathbb{A} \subseteq \mathbb{R} \to \mathbb{R}^n$ is called a **curve** in $\mathbb{R}^n$.

Example 1.2.1. The function $\mathbf{f} : [0, 2\pi] \to \mathbb{R}^2$, with $\mathbf{f}(t) = (\cos(t), \sin(t))$ defines the circle centered at point $(0, 0)$ with radius 1.

1.2.2 Level Sets

Definition 1.2.4. Let $f : \mathbb{U} \subseteq \mathbb{R}^n \to \mathbb{R}$ and $c \in \mathbb{R}$. The set

$$\mathbb{S}_c := \{\mathbf{x} \in \mathbb{U} \,|\, f(\mathbf{x}) = c\} \tag{1.11}$$

is called the **level set** of f, for value c.

If $n = 2$, then we are talking about a level curve (see Figure 1.8), while if $n = 3$ then we are talking about a level surface.

1.2.3 Limits of Sequences - Continuity of Functions

We assume that the reader is familiar with the concepts of vector sequences and convergence of sequences of vectors as well as the concept of continuity of functions. The book by Marsden and Tromba contains extensive coverage of these basic concepts.

1.2.4 Derivative

The concept of differentiation is about approximating functions by linear (actually, affine) functions. Next, we list basic definitions and properties of the derivative function.

Definition 1.2.5. Let $\mathbb{U}$ be an open subset of $\mathbb{R}^n$ and $f : \mathbb{U} \to \mathbb{R}^m$. f is called **differentiable** at the point $\mathbf{x} \in \mathbb{U}$ if there exists a $(m \times n)$ matrix $\mathbf{A}^f_{\mathbf{x}}$ such that

$$\lim_{\mathbf{y}\to\mathbf{x}} \left(\frac{\|f(\mathbf{y}) - f(\mathbf{x}) - \mathbf{A}^f_{\mathbf{x}}(\mathbf{y} - \mathbf{x})\|}{\|\mathbf{y} - \mathbf{x}\|} \right) = 0. \tag{1.12}$$

The matrix $\mathbf{A}^f_{\mathbf{x}}$ is called the **derivative** (or Jacobian) of f at the point $\mathbf{x}$ and is denoted as $Df(\mathbf{x})$. The function f is called differentiable if it is differentiable at every point $\mathbf{x} \in \mathbb{U}$.

Next, we state three important properties of the derivative as theorems.

Theorem 1.2.1. If $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^m$ are differentiable at the point $\mathbf{x} \in \mathbb{R}^n$, then $(f + g)(\mathbf{x}) = f(\mathbf{x}) + g(\mathbf{x})$ is also differentiable at $\mathbf{x}$ and

$$D(f + g)(\mathbf{x}) = Df(\mathbf{x}) + Dg(\mathbf{x}). \tag{1.13}$$

Proof. See a good book on Calculus of Several Variables. □

Theorem 1.2.2. If $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}$ are differentiable at the point $\mathbf{x} \in \mathbb{R}^n$, then $(f \cdot g)(\mathbf{x}) = f(\mathbf{x}) \cdot g(\mathbf{x})$ is also differentiable at $\mathbf{x}$ and

$$D(f \cdot g)(\mathbf{x}) = g(\mathbf{x})Df(\mathbf{x}) + f(\mathbf{x})Dg(\mathbf{x}). \tag{1.14}$$

Proof. See a good book on Calculus of Several Variables. □

Theorem 1.2.3. Let $f : \mathbb{R}^n \to \mathbb{R}^m$ and $h : \mathbb{R}^k \to \mathbb{R}^n$. If h is differentiable at $\mathbf{x} \in \mathbb{R}^k$ and f is differentiable at $h(\mathbf{x}) \in \mathbb{R}^n$, then $f \circ h$ is differentiable at $\mathbf{x} \in \mathbb{R}^k$ and

$$D(f \circ h)(\mathbf{x}) = Df(h(\mathbf{x}))Dh(\mathbf{x}), \tag{1.15}$$

where in the right side of (1.15) we have matrix multiplication.

Proof. See a good book on Calculus of Several Variables. □

In the Appendix, at the end of the chapter, we prove in detail a special case of Theorem 1.2.3.

If a function f is differentiable and its derivative, Df, is a continuous function, then we say that f belongs to the family of functions C^1.

Definition 1.2.6. Let $\mathbb{U}$ be an open subset of $\mathbb{R}^n$, $f : \mathbb{U} \to \mathbb{R}$, $\mathbf{x} \in \mathbb{U}$, and $\mathbf{e}_j$, for $j = 1, \ldots, n$, the n-dimensional vector with elements 0 everywhere except from the j-th position in which it has element 1. If the following limit exists

$$\lim_{h \to 0} \frac{f(\mathbf{x} + h\mathbf{e}_j) - f(\mathbf{x})}{h}, \tag{1.16}$$

it is called the **partial derivative** of f with respect to the j-th coordinate, x_j, at the point $\mathbf{x}$, and is denoted as $\frac{\partial f(\mathbf{x})}{\partial x_j}$ or as $\frac{\partial f(x_1,\ldots,x_n)}{\partial x_j}$.

Next, we state an important theorem that connects the derivative of a function f with its partial derivatives.

Theorem 1.2.4. Let $\mathbb{U} \subseteq \mathbb{R}^n$ be an open set and $f : \mathbb{U} \to \mathbb{R}$.

1. If f is differentiable at $\mathbf{x} \in \mathbb{U}$, then the partial derivatives $\frac{\partial f(\mathbf{x})}{\partial x_i}$ exist and the derivative is equal to $Df(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1} \ \cdots \ \frac{\partial f(\mathbf{x})}{\partial x_n}\right]$.
2. If the partial derivatives of f exist in a neighborhood of $\mathbf{x}$ and are continuous at $\mathbf{x}$, then f is differentiable at $\mathbf{x}$ and $Df(\mathbf{x}) = \left[\frac{\partial f(\mathbf{x})}{\partial x_1} \ \cdots \ \frac{\partial f(\mathbf{x})}{\partial x_n}\right]$.
3. f is C^1 in $\mathbb{U}$ if, and only if, the partial derivatives of f exist and are continuous in $\mathbb{U}$.

Proof. See a standard text on Calculus of Several Variables. □

The existence of the partial derivatives of f does not automatically imply the differentiability of f. The continuity of the partial derivatives is very important.

Definition 1.2.7. If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then the function $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, which is defined as

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_n} \end{bmatrix} = Df(\mathbf{x})^T, \tag{1.17}$$

is called the **gradient** of f at the point $\mathbf{x}$.

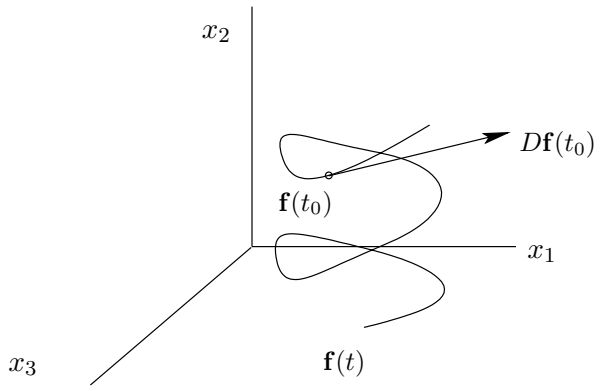


Figure 1.9: Curve derivative.

Later, we will prove some very important properties of $\nabla f(\mathbf{x})$.

Definition 1.2.8. Let $f : \mathbb{R}^n \to \mathbb{R}$. If ∇f is differentiable at a point $\mathbf{x} \in \mathbb{R}^n$, then we say that f is twice differentiable at the point $\mathbf{x}$. The second derivative of f is the derivative of ∇f and is denoted as follows:

$$\nabla^2 f(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 f(\mathbf{x})}{\partial x_1^2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_1} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_1} \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_2} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2^2} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n \partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(\mathbf{x})}{\partial x_1 \partial x_n} & \frac{\partial^2 f(\mathbf{x})}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f(\mathbf{x})}{\partial x_n^2} \end{bmatrix}. \tag{1.18}$$

The $(n \times n)$ matrix $\nabla^2 f(\mathbf{x})$ is called the **Hessian** of f at the point $\mathbf{x}$. If the second-order partial derivatives exist in a neighborhood of $\mathbf{x}$ and are continuous at $\mathbf{x}$, then the order of differentiation does not matter (i.e., $\nabla^2 f(\mathbf{x})$ is a symmetric matrix).

1.2.5 Derivative of a curve

Definition 1.2.9. If $\mathbf{f} : \mathbb{A} \subseteq \mathbb{R} \to \mathbb{R}^n$, with $\mathbf{f}(t) = (f_1(t), \ldots, f_n(t))$, then the $(n \times 1)$ vector $D\mathbf{f}(t_0) = (Df_1(t_0), \ldots, Df_n(t_0))$ is called the derivative of the curve $\mathbf{f}$ at the point $(f_1(t_0), \ldots, f_n(t_0))$.

The derivative $D\mathbf{f}$ can be expressed as

$$D\mathbf{f}(t_0) = \lim_{h \to 0} \frac{\mathbf{f}(t_0 + h) - \mathbf{f}(t_0)}{h}. \tag{1.19}$$

The vector $D\mathbf{f}(t_0)$ is parallel to the line which is tangent to the curve $\mathbf{f}$ at the point $\mathbf{f}(t_0)$ (see Figure 1.9) and is expressed as follows:

$$\mathbf{l}(t) = \mathbf{f}(t_0) + tD\mathbf{f}(t_0). \tag{1.20}$$

1.2.6 Directional Derivative

Definition 1.2.10. Let $\mathbb{U} \subseteq \mathbb{R}^n$ be an open set and $f : \mathbb{U} \to \mathbb{R}$. Let $\mathbf{x} \in \mathbb{U}$ and $\mathbf{0} \neq \mathbf{h} \in \mathbb{R}^n$. If the limit

$$\lim_{t \to 0^+} \left(\frac{f(\mathbf{x} + t\mathbf{h}) - f(\mathbf{x})}{t} \right) \tag{1.21}$$

exists, it is called the **directional derivative** of f, at the point $\mathbf{x}$ along the direction $\mathbf{h}$, and is denoted as $Df(\mathbf{x}; \mathbf{h})$.

If the function f is differentiable, then an equivalent expression for the directional derivative is as follows:

$$Df(\mathbf{x}; \mathbf{h}) = \frac{d}{dt} f(\mathbf{x} + t\mathbf{h}) \bigg|_{t=0}. \tag{1.22}$$

This holds because

$$\begin{aligned} \frac{d}{dt} f(\mathbf{x} + t\mathbf{h}) \bigg|_{t=0} &= \lim_{\Delta t \to 0} \frac{f(\mathbf{x} + (t + \Delta t)\mathbf{h}) - f(\mathbf{x} + t\mathbf{h})}{\Delta t} \bigg|_{t=0} \\ &= \lim_{\Delta t \to 0} \frac{f(\mathbf{x} + \Delta t\, \mathbf{h}) - f(\mathbf{x})}{\Delta t}. \end{aligned} \tag{1.23}$$

Therefore, in this case, the two-sided limit, $\lim_{t \to 0}$ exists, and is equal to the limit from the right, $\lim_{t \to 0^+}$.

Theorem 1.2.5. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and $\mathbf{x} \in \mathbb{R}^n$. Then, for every $\mathbf{h} \in \mathbb{R}^n$, the directional derivative $Df(\mathbf{x}; \mathbf{h})$ exists and is given by

$$Df(\mathbf{x}; \mathbf{h}) = Df(\mathbf{x})\, \mathbf{h} = \nabla f(\mathbf{x})^T \mathbf{h}. \tag{1.24}$$

Proof. Let $\mathbf{c}(t) = \mathbf{x} + t\mathbf{h}$. Then, $f(\mathbf{x} + t\mathbf{h}) = f(\mathbf{c}(t))$. Using the chain rule, we have that

$$\frac{d}{dt} f(\mathbf{c}(t)) = Df(\mathbf{c}(t)) \cdot D\mathbf{c}(t). \tag{1.25}$$

Furthermore, $\mathbf{c}(0) = \mathbf{x}$ and $D\mathbf{c}(t) = \mathbf{h}$. Therefore,

$$\begin{aligned} Df(\mathbf{x}; \mathbf{h}) = \frac{d}{dt} f(\mathbf{x} + t\mathbf{h}) \bigg|_{t=0} = \frac{d}{dt} f(\mathbf{c}(t)) \bigg|_{t=0} &= Df(\mathbf{c}(t))\, D\mathbf{c}(t)|_{t=0} \\ = Df(\mathbf{c}(0))\, D\mathbf{c}(0) = Df(\mathbf{x})\, \mathbf{h} &= \nabla f(\mathbf{x})^T \mathbf{h}, \end{aligned} \tag{1.26}$$

completing the proof. □

Choosing $\mathbf{h} \in \mathbb{R}^n$ with $\|\mathbf{h}\|_2 = 1$, we can interpret $Df(\mathbf{x}; \mathbf{h})$ as the rate of change of f at the point $\mathbf{x}$ along the direction $\mathbf{h}$.

A direct consequence of the Theorem 1.2.5 is the following.

Corollary 1.2.1. If f is differentiable at $\mathbf{x}$, then, for every $\mathbf{h} \in \mathbb{R}^n$, $Df(\mathbf{x}; -\mathbf{h})$ exists and

$$Df(\mathbf{x}; -\mathbf{h}) = -Df(\mathbf{x}; \mathbf{h}). \tag{1.27}$$

A very important theorem related to the rate of change of a function f is as follows.

Theorem 1.2.6. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at $\mathbf{x}_0 \in \mathbb{R}^n$, with $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$. Then, the vector $\nabla f(\mathbf{x}_0)$ "points" to the direction along which the rate of increase of f at $\mathbf{x}_0$ is maximum. On the other hand, $-\nabla f(\mathbf{x}_0)$ "points" to the direction along which the rate of decrease of f at $\mathbf{x}_0$ is maximum.

Proof. Let $\mathbf{h}$ be a vector with unit Euclidean norm in $\mathbb{R}^n$. The rate of change of f at the point $\mathbf{x}_0$, along the direction $\mathbf{h}$, equals $\nabla f(\mathbf{x}_0)^T \mathbf{h}$. Then,

$$\nabla f(\mathbf{x}_0)^T \mathbf{h} = \|\nabla f(\mathbf{x}_0)\|_2 \cos \angle (\nabla f(\mathbf{x}_0), \mathbf{h}) \tag{1.28}$$

and

$$\left|\nabla f(\mathbf{x}_0)^T \mathbf{h}\right| = \|\nabla f(\mathbf{x}_0)\|_2 \left|\cos \angle (\nabla f(\mathbf{x}_0), \mathbf{h})\right| \leq \|\nabla f(\mathbf{x}_0)\|_2 , \tag{1.29}$$

from which we obtain that

$$-\|\nabla f(\mathbf{x}_0)\|_2 \leq \nabla f(\mathbf{x}_0)^T \mathbf{h} \leq \|\nabla f(\mathbf{x}_0)\|_2 . \tag{1.30}$$

The left inequality holds as equality for $\mathbf{h} = -\frac{1}{\|\nabla f(\mathbf{x}_0)\|_2} \nabla f(\mathbf{x}_0)$, while the right inequality holds as equality for $\mathbf{h} = \frac{1}{\|\nabla f(\mathbf{x}_0)\|_2} \nabla f(\mathbf{x}_0)$. □

1.2.7 Gradient and function level sets

In this subsection, we prove another important result for the gradient.

We repeat that a function $\mathbf{c} : [a, b] \subseteq \mathbb{R} \to \mathbb{R}^m$ is a curve. The curve $\mathbf{c}(t)$ can be expressed as

$$\mathbf{c}(t) = \begin{bmatrix} c_1(t) \\ \vdots \\ c_m(t) \end{bmatrix}, \tag{1.31}$$

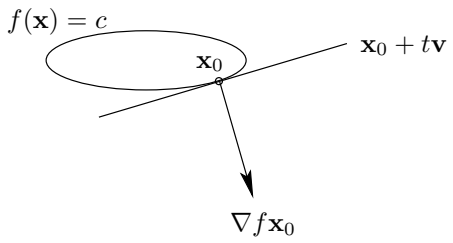


Figure 1.10: Relationship of the gradient vector ∇f and the level set of f at the point $\mathbf{x}_0$.

with $c_i : [a, b] \to \mathbb{R}$, for $i = 1, \ldots, m$. As we have seen, if $\mathbf{c}$ is differentiable at $t \in [a, b]$, then its derivative is the m-dimensional vector

$$D\mathbf{c}(t) = \lim_{h \to 0} \frac{\mathbf{c}(t+h) - \mathbf{c}(t)}{h}. \tag{1.32}$$

The i-th element of $D\mathbf{c}(t)$ is equal to the derivative of $c_i(t)$, for $i = 1, \ldots, m$, and $D\mathbf{c}(t)$ is parallel to the straight line which is tangent to the curve $\mathbf{c}$, at the point $\mathbf{c}(t)$.

Theorem 1.2.7. Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable real function and $\mathbf{x}_0$ a point of the level set $\mathbb{S}_c = \{\mathbf{x} \in \mathbb{R}^n \,|\, f(\mathbf{x}) = c\}$, with $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$. Then, the gradient vector $\nabla f(\mathbf{x}_0)$ is orthogonal to the level set $\mathbb{S}_c$ in the following sense (see Figure 1.10). If $\mathbf{c}(t)$ is a smooth curve contained in the set $\mathbb{S}_c$, with $\mathbf{c}(0) = \mathbf{x}_0$, and $\mathbf{v}$ be a vector parallel to the line tangent to $\mathbf{c}(t)$ at the point $\mathbf{x}_0$, for example, $\mathbf{v} = D\mathbf{c}(0)$, then

$$\nabla f(\mathbf{x}_0)^T \mathbf{v} = 0. \tag{1.33}$$

Proof. Since the curve $\mathbf{c}(t)$ is contained in the set $\mathbb{S}_c$, we have that $f(\mathbf{c}(t)) = c$. Moreover, $\mathbf{c}(0) = \mathbf{x}_0$. From the chain rule, we have that

$$\begin{aligned} 0 = \frac{d}{dt} f(\mathbf{c}(t))\bigg|_{t=0} &= Df(\mathbf{c}(t)) D\mathbf{c}(t)|_{t=0} \\ &= Df(\mathbf{c}(0)) D\mathbf{c}(0) \\ &= Df(\mathbf{x}_0)\mathbf{v} \\ &= \nabla f(\mathbf{x}_0)^T \mathbf{v}, \end{aligned} \tag{1.34}$$

completing the proof. □

As a consequence of the Theorem 1.2.7, it is reasonable to define the plane which is tangent to the level set of f, $\mathbb{S}_c$, at point $\mathbf{x}_0 \in \mathbb{S}_c$, as follows.

Definition 1.2.11. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable and $\mathbb{S}_c = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) = c\}$. Then, the plane tangent to $\mathbb{S}_c$ at the point $\mathbf{x}_0 \in \mathbb{S}_c$, with $\nabla f(\mathbf{x}_0) \neq \mathbf{0}$, is defined as

$$\mathbb{P}^f_{\mathbf{x}_0} = \{\mathbf{x} \in \mathbb{R}^n \mid \nabla f(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0) = 0\}. \tag{1.35}$$

Obviously, $\mathbb{P}^f_{\mathbf{x}_0}$ contains the point $\mathbf{x}_0$. If $\mathbf{x} \neq \mathbf{x}_0$ belongs to the tangent plane, then the vector $\mathbf{x} - \mathbf{x}_0$ is parallel to the tangent plane. But, we have proved that $\nabla f(\mathbf{x}_0)$ is perpendicular to every tangent vector of any curve which lies on the surface $\mathbb{S}_c$ and passes through the point $\mathbf{x}_0$. Therefore, the definition (1.35) is satisfactory.

1.2.8 Taylor expansions

Let $f : \mathbb{R}^n \to \mathbb{R}$ have continuous derivatives of sufficiently high order in a neighborhood of $\mathbf{x}_0$. Then, as $\mathbf{x} \to \mathbf{x}_0$, the following very important relationships hold true:

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0) + O(\|\mathbf{x} - \mathbf{x}_0\|^2) \\ &= f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0) \\ &\quad + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \nabla^2 f(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + O(\|\mathbf{x} - \mathbf{x}_0\|^3), \end{aligned} \tag{1.36}$$

which we call first- and second-order Taylor expansions, respectively.

If, in the above expressions, we ignore the terms denoted by $O(\cdot)$, then we have the first- and second-order Taylor approximations, respectively.

Moreover, assuming that the required derivatives are continuous on an open set containing the line segment joining $\mathbf{x}_0$ and $\mathbf{x}$, it can be shown that the following *exact* expressions hold true:

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{z})^T(\mathbf{x} - \mathbf{x}_0) \tag{1.37}$$

for some $\mathbf{z} = \theta\mathbf{x} + (1 - \theta)\mathbf{x}_0$, with $0 \leq \theta \leq 1$, and

$$f(\mathbf{x}) = f(\mathbf{x}_0) + \nabla f(\mathbf{x}_0)^T(\mathbf{x} - \mathbf{x}_0) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_0)^T \nabla^2 f(\mathbf{w})(\mathbf{x} - \mathbf{x}_0), \tag{1.38}$$

for some $\mathbf{w} = \theta\mathbf{x} + (1 - \theta)\mathbf{x}_0$, with $0 \leq \theta \leq 1$ (for the proofs see any good book on Calculus of Several Variables).

1.3 Appendix

In this appendix, we use first-order Taylor expansions and give a proof of a special form of the chain rule.

Let $\mathbf{c} : \mathbb{R} \to \mathbb{R}^n$ be a smooth curve and $f : \mathbb{R}^n \to \mathbb{R}$ be a smooth real function. We define $g : \mathbb{R} \to \mathbb{R}$, with $g(t) = (f \circ \mathbf{c})(t) = f(\mathbf{c}(t))$. Suppose we want to calculate the derivative $Dg(t) = D(f \circ \mathbf{c})(t)$. We remind you that

$$\mathbf{c}(t + \Delta t) = \mathbf{c}(t) + D\mathbf{c}(t)\Delta t + O(\Delta t^2), \tag{1.39}$$

$$f(\mathbf{x} + \Delta\mathbf{x}) = f(\mathbf{x}) + Df(\mathbf{x})\Delta\mathbf{x} + O(\|\Delta\mathbf{x}\|_2^2) \tag{1.40}$$

Then

$$\begin{aligned} Dg(t) &= \lim_{\Delta t \to 0} \frac{g(t + \Delta t) - g(t)}{\Delta t} \\ &= \lim_{\Delta t \to 0} \frac{f(\mathbf{c}(t + \Delta t)) - f(\mathbf{c}(t))}{\Delta t} \\ &\overset{(1.39)}{=} \lim_{\Delta t \to 0} \frac{f\left(\mathbf{c}(t) + D\mathbf{c}(t)\Delta t + O(\Delta t^2)\right) - f(\mathbf{c}(t))}{\Delta t} \\ &\overset{(1.40)}{=} \lim_{\Delta t \to 0} \frac{f\left(\mathbf{c}(t)\right) + Df(\mathbf{c}(t))D\mathbf{c}(t)\Delta t + O(\Delta t^2) - f(\mathbf{c}(t))}{\Delta t} \\ &= \lim_{\Delta t \to 0} [Df(\mathbf{c}(t))D\mathbf{c}(t) + O(\Delta t)] \\ &= Df(\mathbf{c}(t))D\mathbf{c}(t). \end{aligned} \tag{1.41}$$

1.4 Statements - Proofs

Statements are denoted by capital letters, e.g., P, Q, and will be true or false. For example, the sentence

$$P = \text{“2 is an even number”} \tag{1.42}$$

is a true statement.

If P is a statement, then the **negation** of P is a statement, denoted by not P, and is true if P is false and false if P is true (see the truth table below).

P	not P
T	F
F	T

An important way of constructing statements is by implication. Symbolically, we have

$$(P \Rightarrow Q) \qquad (\text{If } P, \text{ then } Q) \qquad (P \text{ implies } Q). \tag{1.43}$$

Here, P is called the **hypothesis** and Q is called the **conclusion** of the implication.

If the implication holds, we say that P is sufficient for Q and Q is necessary for P. The truth table of the statement $P \Rightarrow Q$ is as follows:

P	Q	$P \Rightarrow Q$
F	F	T
F	T	T
T	F	F
T	T	T

That is, if the hypothesis P is false, then the implication is true regardless of whether the conclusion Q is true or false (this, at first, might seem a bit strange). If the hypothesis P is true and the conclusion Q false, then the implication is false. If both the hypothesis and the conclusion are true, then the implication is true.

The implication $Q \Rightarrow P$ is called the **converse** of $P \Rightarrow Q$ while the implication $(\text{not}\, Q) \Rightarrow (\text{not}\, P)$ is called the **contrapositive** of $P \Rightarrow Q$. The corresponding truth tables are as follows:

P	Q	$P \Rightarrow Q$	$Q \Rightarrow P$	$(\text{not}\, Q) \Rightarrow (\text{not}\, P)$	$\text{not}(P \wedge \text{not}(Q))$
F	F	T	T	T	T
F	T	T	F	T	T
T	F	F	T	F	F
T	T	T	T	T	T

We observe that the statement $P \Rightarrow Q$ is logically equivalent (has the same truth table) with the statement $\text{not}\, Q \Rightarrow \text{not}\, P$. Another statement equivalent to $P \Rightarrow Q$ is $\text{not}\, (P \,\text{and}\, (\text{not}\, Q))$.

According to the above, in order to prove a statement of the form $P \Rightarrow Q$, we have the following basic options:

1. **Direct proof:** We start from the assumption P and, using "correct reasoning," we arrive at conclusion Q.
2. **Proof via contraposition:** We start from the hypothesis $(\text{not}\, Q)$ and, using "correct reasoning," we arrive at conclusion $(\text{not}\, P)$.
3. **Proof via reductio ad absurdum:** We start from the hypothesis $P \,\text{and}\, (\text{not}\, Q)$ and, using "correct reasoning," we arrive at an assertion which is false.

1.4.1 Quantifiers

Important symbols that we use to construct statements are the following: $\forall$ and $\exists$. The first one denotes the universal quantifier while the second denotes the existential quantifier. For example, consider the statements

$$(\forall x)(\exists y)(x + y = 0) \tag{1.44}$$

and

$$(\exists y)(\forall x)(x + y = 0). \tag{1.45}$$

The first statement is true, while the second is false. Therefore, the "order" of the quantifiers matters.

The negation of simple statements with quantifiers is constructed as follows:

$$\text{not}\,((\forall x)P) \quad \equiv \quad (\exists x)\,\text{not}(P), \tag{1.46}$$

$$\text{not}\,((\exists x)P) \quad \equiv \quad (\forall x)\,\text{not}(P). \tag{1.47}$$

If the sentences are more complex, we work as follows:

$$\begin{aligned} & \text{not}\,[(\forall x)(\exists y)(x + y = 0)] \\ \equiv\ & (\exists x)\,\text{not}\,[(\exists y)(x + y = 0)] \\ \equiv\ & (\exists x)\,(\forall y)\,\text{not}(x + y = 0) \\ \equiv\ & (\exists x)\,(\forall y)\,(x + y \neq 0). \end{aligned}$$

The last statement asserts that there exists x such that for every y the sum $x+y$ is nonzero. This statement is false.

Chapter 2

Convex sets

2.1 Affine sets

Definition 2.1.1. A set $\mathbb{C} \subseteq \mathbb{R}^n$ is called **affine** if the points of the lines joining any two points of $\mathbb{C}$ belong to $\mathbb{C}$. That is, if $\mathbf{x}, \mathbf{y} \in \mathbb{C}$ and $\theta \in \mathbb{R}$, then $\theta\mathbf{x} + (1-\theta)\mathbf{y} \in \mathbb{C}$.

In other words, the set $\mathbb{C}$ is called affine if it contains every linear combination of any *two* of its points, under the condition that the coefficients of the linear combination sum to one.

Example 2.1.1. Prove that the solution set of the equation $\mathbf{Ax} = \mathbf{b}$ is an affine set.

The proof is as follows. If the system has no solutions, then the solution set is $\emptyset$, which is an affine set.

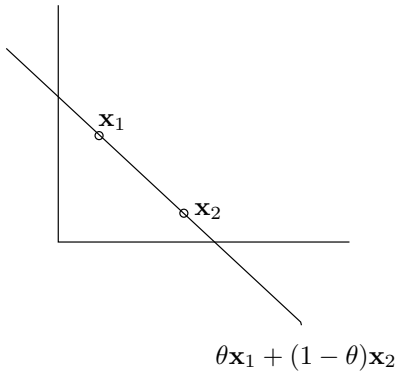


Figure 2.1: The affine set $\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2$.

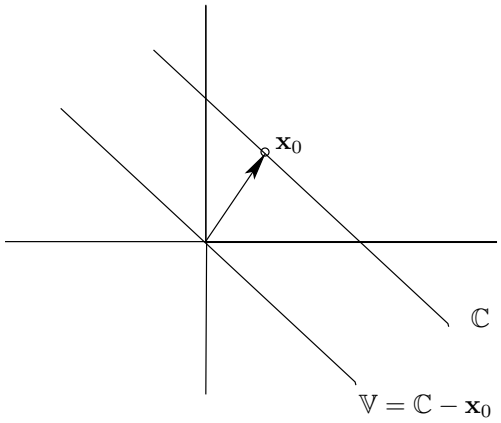


Figure 2.2: Affine set $\mathbb{C}$ and subspace $\mathbb{V} - \mathbf{x}_0$.

Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be solutions of the equation $\mathbf{Ax} = \mathbf{b}$ and $\theta \in \mathbb{R}$. Then

$$\begin{aligned} & \mathbf{A}\mathbf{x}_1 = \mathbf{b}, \quad \mathbf{A}\mathbf{x}_2 = \mathbf{b} \\ \Rightarrow \quad & \mathbf{A}(\theta\,\mathbf{x}_1) = \theta\,\mathbf{b}, \quad \mathbf{A}((1-\theta)\,\mathbf{x}_2) = (1-\theta)\,\mathbf{b} \\ \Rightarrow \quad & \mathbf{A}\left(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2\right) = \mathbf{b}. \end{aligned} \tag{2.1}$$

Therefore, the point $\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2$ is a solution of the equation $\mathbf{Ax} = \mathbf{b}$. So, the solution set is affine. □

Definition 2.1.2. If $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$, $\theta_1, \ldots, \theta_k \in \mathbb{R}$, with

$$\theta_1 + \cdots + \theta_k = 1, \tag{2.2}$$

then every expression of the form $\theta_1\mathbf{x}_1 + \cdots + \theta_k\mathbf{x}_k$ is called an **affine combination** of $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

Using induction and the Definition 2.1.1, it can be shown that an affine set contains every affine combination of its elements. That is, if $\mathbb{C}$ is an affine set, $k \in \mathbb{N}$, $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{C}$, and $\theta_1 + \cdots + \theta_k = 1$, then $\theta_1\mathbf{x}_1 + \cdots + \theta_k\mathbf{x}_k \in \mathbb{C}$ (prove it).

Theorem 2.1.1. If $\mathbb{C}$ is an affine set and $\mathbf{x}_0 \in \mathbb{C}$, then the set

$$\mathbb{V} = \mathbb{C} - \mathbf{x}_0 = \{\mathbf{x} - \mathbf{x}_0 \,|\, \mathbf{x} \in \mathbb{C}\} \tag{2.3}$$

is a linear subspace.

Proof. We must prove that if $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$ and $a_1, a_2 \in \mathbb{R}$, then $a_1\mathbf{v}_1 + a_2\mathbf{v}_2 \in \mathbb{V}$.

Suppose that $\mathbf{v}_1, \mathbf{v}_2 \in \mathbb{V}$. Then, there exist $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{C}$ such that $\mathbf{v}_1 = \mathbf{x}_1 - \mathbf{x}_0$ and $\mathbf{v}_2 = \mathbf{x}_2 - \mathbf{x}_0$. Let $a_1, a_2 \in \mathbb{R}$. Then

$$\begin{aligned} a_1\mathbf{v}_1 + a_2\mathbf{v}_2 &= a_1(\mathbf{x}_1 - \mathbf{x}_0) + a_2(\mathbf{x}_2 - \mathbf{x}_0) \\ &= a_1\mathbf{x}_1 + a_2\mathbf{x}_2 - (a_1 + a_2)\mathbf{x}_0 \\ &= a_1\mathbf{x}_1 + a_2\mathbf{x}_2 - (a_1 + a_2)\mathbf{x}_0 + \mathbf{x}_0 - \mathbf{x}_0 \\ &= [a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + (1 - a_1 - a_2)\mathbf{x}_0] - \mathbf{x}_0 \\ &= \mathbf{x}_* - \mathbf{x}_0, \end{aligned} \tag{2.4}$$

where $\mathbf{x}_* := a_1\mathbf{x}_1 + a_2\mathbf{x}_2 + (1 - a_1 - a_2)\mathbf{x}_0 \in \mathbb{C}$ (why?). So, any linear combination of $\mathbf{v}_1$ and $\mathbf{v}_2$ can be expressed as $\mathbf{x}_* - \mathbf{x}_0$, with $\mathbf{x}_* \in \mathbb{C}$ and, thus, belongs to $\mathbb{V}$. □

Similarly, it can be shown that every affine set $\mathbb{C}$ can be expressed as a translation of a linear subspace $\mathbb{V}$, that is,

$$\mathbb{C} = \mathbb{V} + \mathbf{x}_0 = \{\mathbf{v} + \mathbf{x}_0 \,|\, \mathbf{v} \in \mathbb{V}\}, \tag{2.5}$$

with $\mathbf{x}_0 \in \mathbb{C}$. Vector $\mathbf{x}_0$ in the above definitions is not unique. Actually, it might be any point of $\mathbb{C}$.

The dimension of $\mathbb{C}$ is defined as the dimension of $\mathbb{V}$.

Definition 2.1.3. The set of the affine combinations of the elements of a set $\mathbb{C} \subseteq \mathbb{R}^n$ is called the **affine hull** of $\mathbb{C}$ and is denoted as **aff** $\mathbb{C}$.

That is,

$$\textbf{aff}\,\mathbb{C} = \{\theta_1\mathbf{x}_1 + \cdots + \theta_k\mathbf{x}_k \,|\, k \in \mathbb{N},\ \mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{C}, \theta_1 + \cdots + \theta_k = 1\}. \tag{2.6}$$

The affine hull of a set $\mathbb{C}$ is the smallest affine set that contains $\mathbb{C}$, in the following sense: if $\mathbb{S}$ is an affine set with $\mathbb{C} \subseteq \mathbb{S}$, then **aff** $\mathbb{C} \subseteq \mathbb{S}$.

Exercise: Let $\mathbb{C} \subseteq \mathbb{R}^n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{C}$. Let each of the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_l$ be an affine combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$, and each of the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ be an affine combination of $\mathbf{y}_1, \ldots, \mathbf{y}_l$. Prove that each of the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ is an affine combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

2.1.1 Affine Dimension and Relative Interior

The **affine dimension** of a set $\mathbb{C}$ is the dimension of its affine hull.

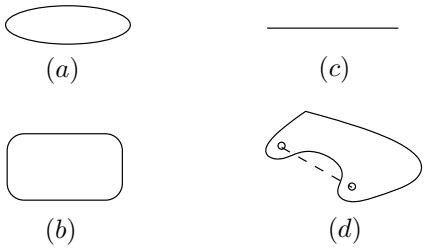


Figure 2.3: Examples of convex sets $(a), (b), (c)$ and non-convex set (d).

If the affine dimension of a set $\mathbb{C} \subseteq \mathbb{R}^n$ is less than n, then $\mathbb{C}$ lies in the affine set $\mathbf{aff}\,\mathbb{C} \neq \mathbb{R}^n$.

We define as **relative interior** of the set $\mathbb{C}$ the set

$$\mathbf{relint}\,\mathbb{C} = \{\mathbf{x} \in \mathbb{C} \,|\, \mathbb{B}(\mathbf{x}, r) \cap \mathbf{aff}\,\mathbb{C} \subseteq \mathbb{C}, \text{ for some } r > 0\}, \tag{2.7}$$

where

$$\mathbb{B}(\mathbf{x}, r) = \{\mathbf{y} \in \mathbb{R}^n \,|\, \|\mathbf{y} - \mathbf{x}\|_2 \leq r\}, \tag{2.8}$$

and as **relative boundary** the set $\mathbf{cl}\,\mathbb{C} \setminus \mathbf{relint}\,\mathbb{C}$.

Example 2.1.2. Consider the square

$$\mathbb{C} = \{\mathbf{x} \in \mathbb{R}^3 \,|\, -1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1, x_3 = 0\}. \tag{2.9}$$

The affine hull of $\mathbb{C}$ is the set $\mathbf{aff}\,\mathbb{C} = \{\mathbf{x} \in \mathbb{R}^3 \,|\, x_3 = 0\}$. The interior of $\mathbb{C}$ is empty (why?) while its boundary is set $\mathbb{C}$ itself. But the relative interior of $\mathbb{C}$ is the set

$$\mathbf{relint}\,\mathbb{C} = \{\mathbf{x} \in \mathbb{R}^3 \,|\, -1 < x_1 < 1, -1 < x_2 < 1, x_3 = 0\}, \tag{2.10}$$

and its relative boundary is the set

$$\{\mathbf{x} \in \mathbb{R}^3 \,|\, \max\{|x_1|, |x_2|\} = 1, x_3 = 0\}. \tag{2.11}$$

2.2 Convex sets

Definition 2.2.1. A set $\mathbb{C} \subseteq \mathbb{R}^n$ is called **convex** if the points of the line segments connecting any two points of $\mathbb{C}$ belong to $\mathbb{C}$. That is, for every $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{C}$ and $0 \leq \theta \leq 1$,

$$\theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2 \in \mathbb{C}. \tag{2.12}$$

We can easily see that every affine set is convex.

Definition 2.2.2. Let $k \in \mathbb{N}$, $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$, and $\theta_1, \ldots, \theta_k \in \mathbb{R}$, with $\theta_i \geq 0$, for $i = 1, \ldots, k$, and $\theta_1 + \cdots + \theta_k = 1$. The point $\theta_1\mathbf{x}_1 + \cdots + \theta_k\mathbf{x}_k$ is called a **convex combination** of $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

Theorem 2.2.1. A set $\mathbb{C}$ is convex if, and only if, it contains all convex combinations of its points, that is, if $k \in \mathbb{N}$, $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{C}$, $\theta_1, \ldots, \theta_k \in \mathbb{R}$, with $\theta_i \geq 0$, for $i = 1, \ldots, k$, and $\theta_1 + \cdots + \theta_k = 1$, then $\theta_1\mathbf{x}_1 + \cdots + \theta_k\mathbf{x}_k \in \mathbb{C}$.

Proof. (Converse) The proof of the converse is simple and we will start with it. Since $\mathbb{C}$ contains the convex combinations of its elements for any $k \in \mathbb{N}$, it also contains the convex combinations of its elements for $k = 2$. Therefore, it is a convex set (recall the definition of a convex set).

(Direct) The proof will be done by induction. For $k = 2$, the theorem holds by the definition of convex set. We assume that the theorem holds for $k = n$.

Let $k = n + 1$, $\mathbf{x}_i \in \mathbb{C}$, $\theta_i \in \mathbb{R}$, $\theta_i \geq 0$, for $i = 1, \ldots, n + 1$, with $\sum_{i=1}^{n+1} \theta_i = 1$ and $\mathbf{x} = \sum_{i=1}^{n+1} \theta_i\mathbf{x}_i$. Define $\theta := \sum_{i=1}^{n} \theta_i$. If $\theta = 0$, then $\theta_{n+1} = 1$ and $\mathbf{x} = \mathbf{x}_{n+1} \in \mathbb{C}$. Thus, suppose that $\theta > 0$. Then $\theta_{n+1} = 1 - \theta$ and

$$\mathbf{x} = \sum_{i=1}^{n+1} \theta_i\mathbf{x}_i = \sum_{i=1}^{n} \theta_i\mathbf{x}_i + \theta_{n+1}\mathbf{x}_{n+1} = \theta \left(\sum_{i=1}^{n} \frac{\theta_i}{\theta}\mathbf{x}_i \right) + \theta_{n+1}\mathbf{x}_{n+1}. \tag{2.13}$$

If we define $\theta'_i := \frac{\theta_i}{\theta}$, for $i = 1, \ldots, n$, we observe that the expression $\sum_{i=1}^{n} \theta'_i\mathbf{x}_i$ is a convex combination of n elements of $\mathbb{C}$ (observe that $\theta'_i \geq 0$, for $i = 1, \ldots, n$, and $\sum_{i=1}^{n} \theta'_i = 1$). We define $\mathbf{x}' = \sum_{i=1}^{n} \theta'_i\mathbf{x}_i$. From the induction hypothesis, we conclude that $\mathbf{x}' \in \mathbb{C}$. Finally, we find that $\mathbf{x} \in \mathbb{C}$ because it can be expressed as a convex combination of two points of $\mathbb{C}$, as follows:

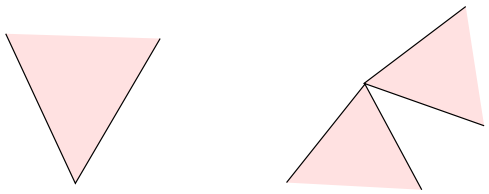
$$\mathbf{x} = \theta\mathbf{x}' + (1 - \theta)\mathbf{x}_{n+1}. \tag{2.14}$$

Therefore, the proposition holds for $k = n + 1$ and the proof is complete. □

Definition 2.2.3. The **convex hull** of a set $\mathbb{C}$ is the set of all convex combinations of its elements and is denoted as $\mathbf{conv}\,\mathbb{C}$. That is,

$$\mathbf{conv}\,\mathbb{C} = \{\theta_1\mathbf{x}_1 + \cdots + \theta_k\mathbf{x}_k \,|\, k \in \mathbb{N},\ \mathbf{x}_i \in \mathbb{C}, \theta_i \geq 0, i = 1, \ldots, k, \theta_1 + \cdots + \theta_k = 1\}. \tag{2.15}$$

It can be shown that the set $\mathbf{conv}\,\mathbb{C}$ is the smallest convex set containing $\mathbb{C}$ (prove it).



(a) (b)

Figure 2.4: Convex cone (a), non-convex cone (b).

Exercise: Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$. Find the set $\mathbf{conv}\{\mathbf{x}_1, \mathbf{x}_2\}$.

Exercise: Let $\mathbb{C} \subseteq \mathbb{R}^n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{C}$. Let each of the vectors $\mathbf{y}_1, \ldots, \mathbf{y}_l$ be a convex combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$, and each of the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ be a convex combination of $\mathbf{y}_1, \ldots, \mathbf{y}_l$. Prove that each of the vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$ is a convex combination of $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

2.3 Cones

Definition 2.3.1. A set $\mathbb{C} \subseteq \mathbb{R}^n$ is called a **cone** if for every $\mathbf{x} \in \mathbb{C}$ and $\theta \geq 0$, we have that $\theta\mathbf{x} \in \mathbb{C}$.

Notice that if $\mathbb{C} \neq \emptyset$, then $\mathbf{0} \in \mathbb{C}$ (why?).

A set $\mathbb{C}$ is a convex cone if it is, at the same time, a convex set and a cone (see Figure 2.4).

Exercise: Let $\mathbb{C}$ be a convex cone. Prove that, if $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{C}$ and $\theta_1, \theta_2 \geq 0$, then

$$\theta_1\mathbf{x}_1 + \theta_2\mathbf{x}_2 \in \mathbb{C}. \tag{2.16}$$

Proof: Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{C}$ and $\theta_1, \theta_2 \geq 0$. If $\theta_1 = \theta_2 = 0$, then $\theta_1\mathbf{x}_1 + \theta_2\mathbf{x}_2 = \mathbf{0} \in \mathbb{C}$.

Let $\theta_1 \neq 0$ or $\theta_2 \neq 0$. Then, due to the convexity of $\mathbb{C}$, we have that

$$\frac{\theta_1}{\theta_1 + \theta_2}\mathbf{x}_1 + \frac{\theta_2}{\theta_1 + \theta_2}\mathbf{x}_2 \in \mathbb{C}. \tag{2.17}$$

Moreover, $\mathbb{C}$ is a cone, therefore,

$$(\theta_1 + \theta_2)\left(\frac{\theta_1}{\theta_1 + \theta_2}\mathbf{x}_1 + \frac{\theta_2}{\theta_1 + \theta_2}\mathbf{x}_2\right) \in \mathbb{C}. \tag{2.18}$$

That is, $\theta_1\mathbf{x}_1 + \theta_2\mathbf{x}_2 \in \mathbb{C}$. □

Definition 2.3.2. Let $k \in \mathbb{N}$, $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$, and $\theta_1, \ldots, \theta_k \in \mathbb{R}$, with $\theta_i \geq 0$, for $i = 1, \ldots, k$. The vectors of the form

$$\theta_1 \mathbf{x}_1 + \cdots + \theta_k \mathbf{x}_k \tag{2.19}$$

are called **conic combinations** of $\mathbf{x}_1, \ldots, \mathbf{x}_k$.

A set $\mathbb{C}$ is a convex cone if, and only if, it contains all conic combinations of its elements.

Definition 2.3.3. The **conic hull** of a set $\mathbb{C}$ is the set of the conic combinations of its elements, that is, the set

$$\{\theta_1 \mathbf{x}_1 + \cdots + \theta_k \mathbf{x}_k \,|\, k \in \mathbb{N},\ \mathbf{x}_i \in \mathbb{C}, \theta_i \geq 0, i = 1, \ldots, k\}. \tag{2.20}$$

It can be shown that the conic hull of a set $\mathbb{C}$ is the smallest convex cone which contains $\mathbb{C}$.

2.4 Examples of Convex Sets

$\emptyset$, the singleton $\{x_0\}$, and $\mathbb{R}^n$ are affine and, thus, convex sets.

Straight lines are affine and, therefore, convex sets, while line segments are convex but not affine sets.

2.5 Hyperplanes - Halfspaces

Let $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$. A **hyperplane** in $\mathbb{R}^n$ is a set of the form

$$\mathbb{P} = \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{a}^T \mathbf{x} - b = 0\}. \tag{2.21}$$

A hyperplane divides $\mathbb{R}^n$ into two **half-spaces**, as follows:

$$\mathbb{P}^+ = \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{a}^T \mathbf{x} - b \geq 0\}, \quad \mathbb{P}^- = \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{a}^T \mathbf{x} - b \leq 0\}. \tag{2.22}$$

The set $\mathbb{P}$ is affine and, therefore, convex. The sets $\mathbb{P}^+$ and $\mathbb{P}^-$ are convex but not affine (prove it).

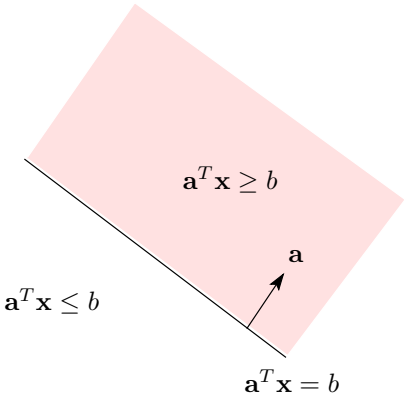


Figure 2.5: Hyperplane - Halfspaces.

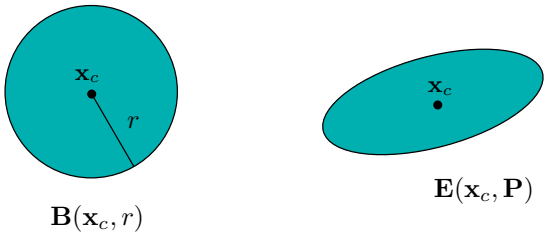


Figure 2.6: Euclidean ball and ellipsoid.

2.5.1 Euclidean balls

The Euclidean ball in $\mathbb{R}^n$, with center $\mathbf{x}_c$ and radius $r \geq 0$, is defined as

$$\mathbb{B}(\mathbf{x}_c, r) = \{\mathbf{x} \in \mathbb{R}^n \,|\, \|\mathbf{x} - \mathbf{x}_c\|_2 \leq r\}. \tag{2.23}$$

Another representation of the Euclidean ball is as follows:

$$\mathbb{B}(\mathbf{x}_c, r) = \{\mathbf{x}_c + r\mathbf{x} \,|\, \|\mathbf{x}\|_2 \leq 1\}. \tag{2.24}$$

Euclidean balls are convex sets. The proof is as follows. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{B}(\mathbf{x}_c, r)$. This means that $\|\mathbf{x}_i - \mathbf{x}_c\|_2 \leq r$, for $i = 1, 2$. We must prove that, for $0 \leq \theta \leq 1$, we have that

$$\mathbf{x} := \theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2 \in \mathbb{B}(\mathbf{x}_c, r),$$

that is, $\|\mathbf{x} - \mathbf{x}_c\|_2 \leq r$.

We proceed as follows:

$$\begin{aligned}\|\mathbf{x}-\mathbf{x}_c\|_2 &= \|\theta\mathbf{x}_1+(1-\theta)\mathbf{x}_2-\mathbf{x}_c\|_2\\ &= \|\theta(\mathbf{x}_1-\mathbf{x}_c)+(1-\theta)(\mathbf{x}_2-\mathbf{x}_c)\|_2\\ &\leq \|\theta(\mathbf{x}_1-\mathbf{x}_c)\|_2+\|(1-\theta)(\mathbf{x}_2-\mathbf{x}_c)\|_2\\ &= \theta\|\mathbf{x}_1-\mathbf{x}_c\|_2+(1-\theta)\|\mathbf{x}_2-\mathbf{x}_c\|_2\\ &\leq \theta\, r+(1-\theta)\, r\\ &= r,\end{aligned} \tag{2.25}$$

therefore, $\mathbf{x}\in\mathbb{B}(\mathbf{x}_c,r)$.

2.5.2 Ellipsoids

If $\mathbf{P}=\mathbf{P}^T\succ\mathbf{O}$ and $\mathbf{x}_c\in\mathbb{R}^n$, then the set

$$\mathbb{E}(\mathbf{x}_c,\mathbf{P})=\left\{\mathbf{x}\in\mathbb{R}^n\,|\,(\mathbf{x}-\mathbf{x}_c)^T\mathbf{P}^{-1}(\mathbf{x}-\mathbf{x}_c)\leq 1\right\} \tag{2.26}$$

is called an **ellipsoid** with center $\mathbf{x}_c$.

An alternative description of the same ellipsoid is as follows:

$$\mathbb{E}(\mathbf{x}_c,\mathbf{P})=\{\mathbf{x}_c+\mathbf{A}\mathbf{u}\,|\,\|\mathbf{u}\|_2\leq 1\}, \tag{2.27}$$

with $\mathbf{A}=\mathbf{P}^{\frac{1}{2}}$.

Ellipsoids are convex sets. Next, we will give a proof for the case where $\mathbf{x}_c=\mathbf{0}$ (the proof generalizes to every $\mathbf{x}_c$).

Let $\mathbf{x}_1,\mathbf{x}_2\in\mathbb{E}(\mathbf{0},\mathbf{P})$, that is, $\mathbf{x}_i^T\mathbf{P}^{-1}\mathbf{x}_i\leq 1$, for $i=1,2$. We must prove that, if $0\leq\theta\leq 1$ and $\mathbf{x}:=\theta\mathbf{x}_1+(1-\theta)\mathbf{x}_2$, then $\mathbf{x}^T\mathbf{P}^{-1}\mathbf{x}\leq 1$. First, we have

$$\begin{aligned}\mathbf{x}^T\mathbf{P}^{-1}\mathbf{x} &= \left(\theta\mathbf{x}_1+(1-\theta)\mathbf{x}_2\right)^T\mathbf{P}^{-1}\left(\theta\mathbf{x}_1+(1-\theta)\mathbf{x}_2\right)\\ &= \theta^2\mathbf{x}_1^T\mathbf{P}^{-1}\mathbf{x}_1+(1-\theta)^2\mathbf{x}_2^T\mathbf{P}^{-1}\mathbf{x}_2+2\theta(1-\theta)\mathbf{x}_1^T\mathbf{P}^{-1}\mathbf{x}_2\\ &\leq \theta^2+(1-\theta)^2+2\theta(1-\theta)\mathbf{x}_1^T\mathbf{P}^{-1}\mathbf{x}_2.\end{aligned} \tag{2.28}$$

If we define $\mathbf{y}_i:=\mathbf{P}^{-\frac{1}{2}}\mathbf{x}_i$, for $i=1,2$, then $\|\mathbf{y}_i\|_2^2=\mathbf{x}_i^T\mathbf{P}^{-1}\mathbf{x}_i\leq 1$, for $i=1,2$, and

$$\begin{aligned}\mathbf{x}_1^T\mathbf{P}^{-1}\mathbf{x}_2 &= \left(\mathbf{P}^{-\frac{1}{2}}\mathbf{x}_1\right)^T\left(\mathbf{P}^{-\frac{1}{2}}\mathbf{x}_2\right)\\ &= \mathbf{y}_1^T\mathbf{y}_2\\ &\overset{(a)}{\leq} \|\mathbf{y}_1\|_2\|\mathbf{y}_2\|_2\\ &\leq 1,\end{aligned} \tag{2.29}$$

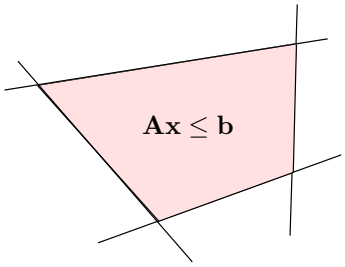


Figure 2.7: Polyhedron $\mathbf{Ax} \leq \mathbf{b}$.

where in step (a) we used the Cauchy-Schwarz inequality. Since $0 \leq \theta \leq 1$, we have that $\theta(1-\theta) \geq 0$, and

$$\theta(1-\theta)\mathbf{x}_1^T\mathbf{P}^{-1}\mathbf{x}_2 \leq \theta(1-\theta). \tag{2.30}$$

Therefore,

$$\mathbf{x}^T\mathbf{P}^{-1}\mathbf{x} \leq \theta^2 + (1-\theta)^2 + 2\theta(1-\theta) = 1. \tag{2.31}$$

Thus, the point $\mathbf{x}$ belongs to the ellipsoid $\mathbb{E}(\mathbf{0}, \mathbf{P})$ and the set $\mathbb{E}(\mathbf{0}, \mathbf{P})$ is convex.

2.5.3 Polyhedra

We define a **polyhedron** as the solution of a finite set of linear inequalities and linear equalities (see Figure 2.7). For example, the set

$$\mathbb{D} = \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{a}_j^T\mathbf{x} - b_j \leq 0, j = 1, \ldots, m, \mathbf{c}_j^T\mathbf{x} - d_j = 0, j = 1, \ldots, p\} \tag{2.32}$$

is a polyhedron.

Thus, a polyhedron is the intersection of a finite number of half-spaces and hyperplanes. Affine sets, line segments, and half-spaces are polyhedra.

Exercise: Prove that polyhedra are convex sets.

2.6 Set operations that preserve convexity

2.6.1 Intersection

If $\mathbb{S}_1, \mathbb{S}_2 \subseteq \mathbb{R}^n$ are convex sets, then their intersection $\mathbb{S}_1 \cap \mathbb{S}_2$ is a convex set. The above proposition generalizes to any number (finite or infinite) of convex sets.

More specifically, if $i \in \mathcal{I}$, where $\mathcal{I}$ is any set of indices, and $\mathbb{S}_i$ is a convex set, for each $i \in \mathcal{I}$, then the set

$$\mathbb{S} = \cap_{i \in \mathcal{I}} \mathbb{S}_i \tag{2.33}$$

is convex. The proof is as follows.

Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{S}$. This means that $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{S}_i$, for $i \in \mathcal{I}$. If $0 \leq \theta \leq 1$, then, due to the convexity of $\mathbb{S}_i$, for $i \in \mathcal{I}$, we have that $\mathbf{x} := \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathbb{S}_i$, for $i \in \mathcal{I}$. Therefore, $\mathbf{x} \in \mathbb{S}$, proving that $\mathbb{S}$ is a convex set.

2.6.2 Cartesian product

Theorem 2.6.1. If $\mathbb{S}_1 \subseteq \mathbb{R}^n$ and $\mathbb{S}_2 \subseteq \mathbb{R}^m$ are convex sets, then the Cartesian product $\mathbb{S} = \mathbb{S}_1 \times \mathbb{S}_2 \subseteq \mathbb{R}^{n+m}$ is a convex set.

Proof. We must prove that, if $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{S}$ and $0 \leq \theta \leq 1$, then $\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathbb{S}$.

Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{S}$. Then, there exist $\mathbf{x}_{11}, \mathbf{x}_{12} \in \mathbb{S}_1$ and $\mathbf{x}_{21}, \mathbf{x}_{22} \in \mathbb{S}_2$ such that

$$\mathbf{x}_1 = \begin{bmatrix} \mathbf{x}_{11} \\ \mathbf{x}_{21} \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} \mathbf{x}_{12} \\ \mathbf{x}_{22} \end{bmatrix}.$$

Since the sets $\mathbb{S}_1$ and $\mathbb{S}_2$ are convex, for each $0 \leq \theta \leq 1$, $\theta\mathbf{x}_{11} + (1-\theta)\mathbf{x}_{12} \in \mathbb{S}_1$ and $\theta\mathbf{x}_{21} + (1-\theta)\mathbf{x}_{22} \in \mathbb{S}_2$. Therefore,

$$\begin{aligned} & \begin{bmatrix} \theta\mathbf{x}_{11} + (1-\theta)\mathbf{x}_{12} \\ \theta\mathbf{x}_{21} + (1-\theta)\mathbf{x}_{22} \end{bmatrix} \in \mathbb{S}_1 \times \mathbb{S}_2 \\ \implies & \theta \begin{bmatrix} \mathbf{x}_{11} \\ \mathbf{x}_{21} \end{bmatrix} + (1-\theta) \begin{bmatrix} \mathbf{x}_{12} \\ \mathbf{x}_{22} \end{bmatrix} \in \mathbb{S}_1 \times \mathbb{S}_2 \\ \implies & \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathbb{S}, \end{aligned} \tag{2.34}$$

completing the proof. □

2.6.3 Image of affine function

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is called affine if it is the sum of a linear function and a constant, that is, if it is of the form $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$, with $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$.

Exercise: Let $f : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$. A function f is linear if, for any $k \in \mathbb{N}$, $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$ and $a_1, \ldots, a_k \in \mathbb{R}$, it holds true that

$$f\left(\sum_{i=1}^{k} a_i \mathbf{x}_i\right) = \sum_{i=1}^{k} a_i f(\mathbf{x}_i). \tag{2.35}$$

Prove that the function $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ is linear.

Exercise: Let $f : \mathbb{R}^n \to \mathbb{R}^m$, $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$. A function f is affine if, for any $k \in \mathbb{N}$, $\mathbf{x}_1, \ldots, \mathbf{x}_k \in \mathbb{R}^n$ and $a_1, \ldots, a_k \in \mathbb{R}$, with $\sum_{i=1}^{k} a_i = 1$, it holds true that

$$f\left(\sum_{i=1}^{k} a_i \mathbf{x}_i\right) = \sum_{i=1}^{k} a_i f(\mathbf{x}_i). \tag{2.36}$$

Prove that the function $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ is affine.

Theorem 2.6.2. Let $\mathbb{S} \subseteq \mathbb{R}^n$ be a convex set and $f : \mathbb{R}^n \to \mathbb{R}^m$ an affine function. Then, the image of $\mathbb{S}$ under f,

$$f(\mathbb{S}) = \{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathbb{S}\}, \tag{2.37}$$

is a convex set.

Proof. We must prove that, if $\mathbf{y}_1, \mathbf{y}_2 \in f(\mathbb{S})$ and $0 \leq \theta \leq 1$, then $\mathbf{y} = \theta\mathbf{y}_1 + (1-\theta)\mathbf{y}_2 \in f(\mathbb{S})$.

Let $\mathbf{y}_1, \mathbf{y}_2 \in f(\mathbb{S})$. This means that there exist $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{S}$ such that $\mathbf{y}_i = \mathbf{A}\mathbf{x}_i + \mathbf{b}$, for $i = 1, 2$.

Due to the convexity of $\mathbb{S}$, we have that, for $0 \leq \theta \leq 1$, $\mathbf{x} = \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathbb{S}$. The image of $\mathbf{x}$ is given by the relation

$$\begin{aligned} f(\mathbf{x}) &= \mathbf{A}\mathbf{x} + \mathbf{b} \\ &= \mathbf{A}\left(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2\right) + \mathbf{b} \\ &= \theta\left(\mathbf{A}\mathbf{x}_1 + \mathbf{b}\right) + (1-\theta)\left(\mathbf{A}\mathbf{x}_2 + \mathbf{b}\right) \\ &= \theta\mathbf{y}_1 + (1-\theta)\mathbf{y}_2 \\ &= \mathbf{y}. \end{aligned} \tag{2.38}$$

Therefore, $\mathbf{y} = f(\mathbf{x}) \in f(\mathbb{S})$ and the proof is complete. □

Corollary 2.6.1. If $f : \mathbb{R}^k \to \mathbb{R}^n$ is an affine function and $\mathbb{S} \subseteq \mathbb{R}^n$ is a convex set, then the inverse image of $\mathbb{S}$ under f, i.e., the set

$$f^{-1}(\mathbb{S}) = \{\mathbf{x} \,|\, f(\mathbf{x}) \in \mathbb{S}\}, \tag{2.39}$$

is a convex set.

Proof. Let $\mathbf{x}_1, \mathbf{x}_2 \in f^{-1}(\mathbb{S})$ and $0 \leq \theta \leq 1$. Then $f(\mathbf{x}_1), f(\mathbf{x}_2) \in \mathbb{S}$. Since $\mathbb{S}$ is convex,

$$\theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2) \in \mathbb{S}. \tag{2.40}$$

Because f is affine,

$$f\left(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2\right) = \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2) \in \mathbb{S}. \tag{2.41}$$

Therefore, $\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in f^{-1}(\mathbb{S})$, and hence $f^{-1}(\mathbb{S})$ is convex. □

Next, we will present some important applications of Theorem 2.6.2.

2.6.4 Set Scaling

Corollary 2.6.2. If $\mathbb{S} \subseteq \mathbb{R}^n$ is a convex set and $a \in \mathbb{R}$, then the set $a\,\mathbb{S} = \{a\mathbf{x} \,|\, \mathbf{x} \in \mathbb{S}\}$ is convex.

Proof. This proposition is a consequence of Theorem 2.6.2, applied to the affine (actually, linear) function $f : \mathbb{R}^n \to \mathbb{R}^n$ with $f(\mathbf{x}) = a\,\mathbf{x}$, since $a\,\mathbb{S} = f(\mathbb{S})$. □

2.6.5 Set translation

Corollary 2.6.3. If $\mathbb{S} \subseteq \mathbb{R}^n$ is a convex set and $\mathbf{a} \in \mathbb{R}^n$, then the set $\mathbb{S}+\mathbf{a} = \{\mathbf{x}+\mathbf{a} \,|\, \mathbf{x} \in \mathbb{S}\}$ is convex.

Proof. This proposition is a consequence of Theorem 2.6.2, applied to the affine function $f : \mathbb{R}^n \to \mathbb{R}^n$ with $f(\mathbf{x}) = \mathbf{x} + \mathbf{a}$, since $\mathbb{S} + \mathbf{a} = f(\mathbb{S})$. □

2.6.6 Projection

Corollary 2.6.4. The projection of a convex set onto some of its coordinates results in a convex set. More specifically, if $\mathbb{S} \subseteq \mathbb{R}^{n+m}$ is a convex set, then the set

$$\mathbb{T} = \{\mathbf{x}_1 \in \mathbb{R}^n \,|\, \text{there exists } \mathbf{x}_2 \in \mathbb{R}^m \text{ such that } (\mathbf{x}_1, \mathbf{x}_2) \in \mathbb{S}\} \tag{2.42}$$

is convex.

Proof. The proof is left to the reader. □

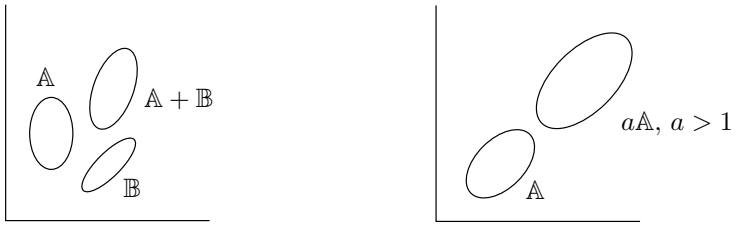


Figure 2.8: Sum of convex sets and scaling of convex sets.

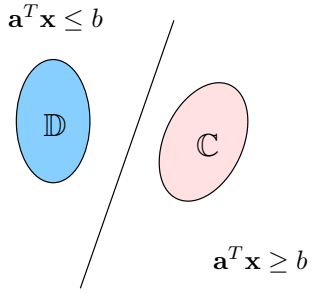


Figure 2.9: Hyperplane separating convex sets.

2.6.7 Sum of sets

If $\mathbb{S}_1, \mathbb{S}_2 \subseteq \mathbb{R}^n$, then their sum is defined as

$$\mathbb{S}_1 + \mathbb{S}_2 = \{\mathbf{x}_1 + \mathbf{x}_2 \,|\, \mathbf{x}_1 \in \mathbb{S}_1, \mathbf{x}_2 \in \mathbb{S}_2\}. \tag{2.43}$$

Corollary 2.6.5. If $\mathbb{S}_1$ and $\mathbb{S}_2$ are convex, then $\mathbb{S}_1 + \mathbb{S}_2$ is convex.

Proof. One way to prove this corollary is to observe the following:

1. $\mathbb{S}_1 \times \mathbb{S}_2$ is convex (Cartesian product of convex sets),
2. the image of the convex set $\mathbb{S}_1 \times \mathbb{S}_2$ under the affine transformation $f : \mathbb{S}_1 \times \mathbb{S}_2 \to \mathbb{R}^n$ defined as
$$f(\mathbf{x}_1, \mathbf{x}_2) = [\,\mathbf{I}_n \;\mathbf{I}_n\,] \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \mathbf{x}_1 + \mathbf{x}_2 \tag{2.44}$$
is the sum $\mathbb{S}_1 + \mathbb{S}_2$.

2.7 Separating hyperplanes

Theorem 2.7.1. Let $\mathbb{C}, \mathbb{D} \subseteq \mathbb{R}^n$ be convex sets that do not intersect, that is, $\mathbb{C} \cap \mathbb{D} = \emptyset$. Then, there exist $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$ and $b \in \mathbb{R}$, such that $\mathbf{a}^T\mathbf{x} \geq b$ for every $\mathbf{x} \in \mathbb{C}$ and $\mathbf{a}^T\mathbf{x} \leq b$ for every $\mathbf{x} \in \mathbb{D}$ (see Figure 2.9).

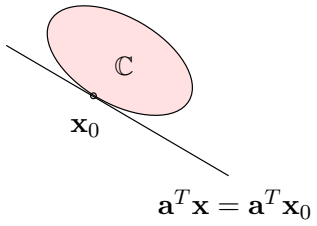


Figure 2.10: Supporting hyperplane of set $\mathbb{C}$ at point $\mathbf{x}_0$.

The hyperplane $\{\mathbf{x} \,|\, \mathbf{a}^T\mathbf{x} = b\}$ is called a **separating hyperplane** of $\mathbb{C}$ and $\mathbb{D}$.

Proof. See pages 46–48 of the book by Boyd and Vandenberghe. □

2.8 Supporting hyperplanes

Theorem 2.8.1. Let $\mathbb{C} \subseteq \mathbb{R}^n$ be a nonempty convex set and $\mathbf{x}_0 \in \mathbf{bd}\,\mathbb{C}$. Then, there exists $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$ such that $\mathbf{a}^T\mathbf{x} \leq \mathbf{a}^T\mathbf{x}_0$, for every $\mathbf{x} \in \mathbb{C}$.

Proof. See page 51 of Boyd and Vandenberghe. □

Theorem 2.8.1 says that there exists a hyperplane $\mathbb{P} = \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{x}_0\}$ that passes through $\mathbf{x}_0$ and such that $\mathbb{C}$ lies entirely in one of the two closed half-spaces determined by $\mathbb{P}$ (see Fig. 2.10).

Chapter 3

Convex functions

In this Chapter, we shall briefly consider significant properties of convex functions.

3.1 Convex Functions

Definition 3.1.1. A function $f : \mathbf{dom}f \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **convex** if the set $\mathbf{dom}f$ is convex and if, for each $\mathbf{x}, \mathbf{y} \in \mathbf{dom}f$ and $0 \leq \theta \leq 1$, it holds that

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) \leq \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}). \tag{3.1}$$

Geometrically, the inequality (3.1) implies that no point of the line segment connecting the points $(\mathbf{x}, f(\mathbf{x}))$ and $(\mathbf{y}, f(\mathbf{y}))$ lies under the graph of f (see Fig. 3.1).

Definition 3.1.2. A function $f : \mathbf{dom}f \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **strictly convex** if the set $\mathbf{dom}f$ is convex and if, for every $\mathbf{x}, \mathbf{y} \in \mathbf{dom}f$, with $\mathbf{x} \neq \mathbf{y}$, and $0 < \theta < 1$, it holds that

$$f(\theta\mathbf{x} + (1-\theta)\mathbf{y}) < \theta f(\mathbf{x}) + (1-\theta)f(\mathbf{y}). \tag{3.2}$$

Geometrically, the inequality (3.2) means that, with the exception of the points $(\mathbf{x}, f(\mathbf{x}))$ and $(\mathbf{y}, f(\mathbf{y}))$, the line segment connecting the points $(\mathbf{x}, f(\mathbf{x}))$ and $(\mathbf{y}, f(\mathbf{y}))$ lies above the graph of f (see Fig. 3.1).

The function f is called **concave** if $-f$ is convex.

An affine function is, at the same time, convex and concave.

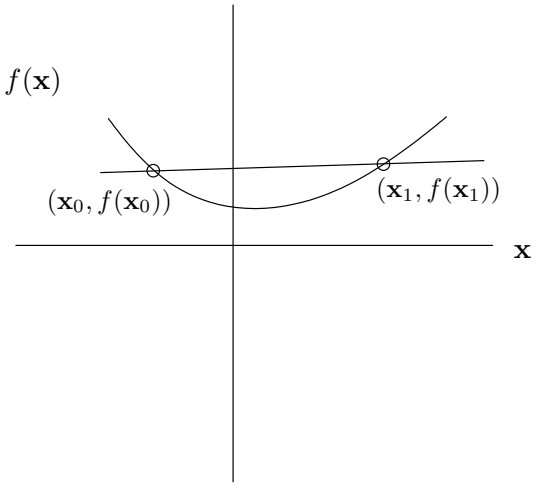


Figure 3.1: Convex function.

The relation (3.1) can be generalized as follows. Let $f : \mathbf{dom}f \subseteq \mathbb{R}^n \to \mathbb{R}$ be a convex function, $\mathbf{x}_i \in \mathbf{dom}f$, for $i = 1, \ldots, n$, and $a_i \geq 0$, for $i = 1, \ldots, n$, with $a_1 + \cdots + a_n = 1$. Then,

$$f\left(\sum_{i=1}^{n} a_i \mathbf{x}_i\right) \leq \sum_{i=1}^{n} a_i f(\mathbf{x}_i). \tag{3.3}$$

This inequality is sometimes called Jensen's inequality.

It can be shown that a function f is convex if, and only if, its restriction on the intersection of any straight line in $\mathbb{R}^n$ and its domain $\mathbf{dom}f$ is a convex function.

Theorem 3.1.1. The function $f : \mathbf{dom}f \subseteq \mathbb{R}^n \to \mathbb{R}$ is convex if, and only if, for each $\mathbf{x} \in \mathbf{dom}f$ and $\mathbf{v} \in \mathbb{R}^n$, the function $g(t) = f(\mathbf{x}+t\mathbf{v})$, with domain the set $\{t \in \mathbb{R} \,|\, \mathbf{x}+t\mathbf{v} \in \mathbf{dom}f\}$, is convex.

Proof. A proof appears in the Appendix at the end of this chapter (for a two-dimensional example, see Fig. 3.2, where $\mathbf{x} = (1, 0)$ and $\mathbf{v} = (0, 1)$). □

Convex functions have some very important properties. For example, they are continuous on the interior of their domain. Discontinuities can only exist on the boundary of their domain.

The theory of convex functions is extensive. In this chapter, we will cover only the topics which are necessary for our developments.

In the sequel, we present equivalent definitions of convexity when the function is sufficiently

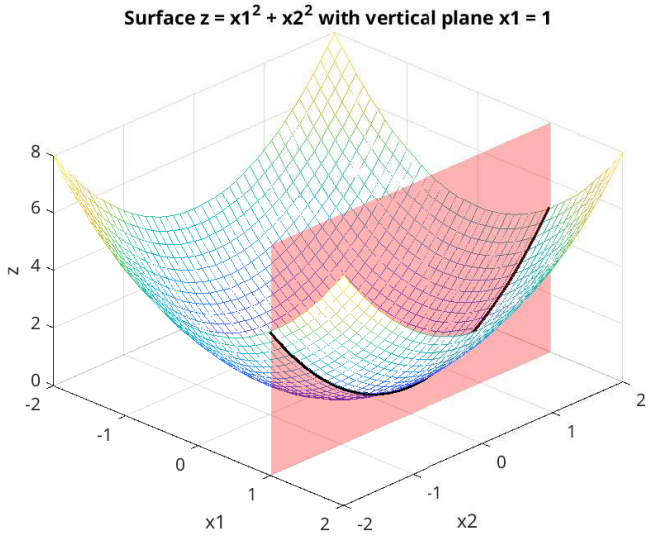


Figure 3.2: The intersection of the graph of a convex function $f(\mathbf{x})$ and the vertical plane is the graph of the function $g(t) = f(\mathbf{x} + t\mathbf{v})$, for $\mathbf{x} = (1, 0)$ and $\mathbf{v} = (0, 1)$.

smooth.

3.2 First Order Condition

Theorem 3.2.1. Let $\mathbf{dom} f \subseteq \mathbb{R}^n$ be an open set and $f : \mathbf{dom} f \to \mathbb{R}$ be a differentiable function. Function f is convex if, and only if, $\mathbf{dom} f$ is a convex set and

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}), \tag{3.4}$$

for any $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f$.

The inequality (3.4) is very important. It states that the first-order Taylor approximation at any point of the domain of a convex function is a global underestimator of the function. That is, from local information (the values of the function and its derivative at a point), we get global information (a global underestimator of the function).

Proof. (*Direct*) Let f be a convex function and $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f$. Then, for $0 < \theta \leq 1$ we have that $(1 - \theta)\mathbf{x} + \theta\mathbf{y} = \mathbf{x} + \theta(\mathbf{y} - \mathbf{x}) \in \mathbf{dom} f$ and

$$f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})) \leq (1 - \theta)f(\mathbf{x}) + \theta f(\mathbf{y}).$$

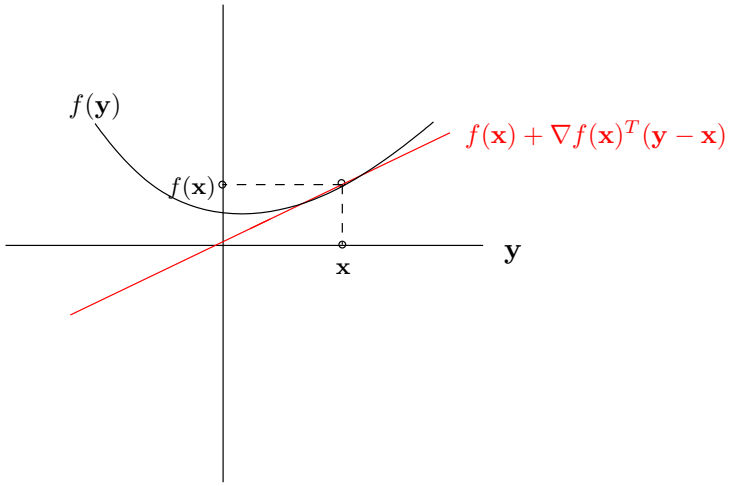


Figure 3.3: First order convexity condition.

Dividing both sides of the above inequality by θ, and after simple algebraic manipulations, we obtain

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \frac{f(\mathbf{x} + \theta(\mathbf{y} - \mathbf{x})) - f(\mathbf{x})}{\theta}.$$

Taking the limit for $\theta \to 0^+$, we get (why?)

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}),$$

which is the relation to be proved.

(*Converse*) Assume that (3.4) holds for every $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f$. Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{dom} f$, with $\mathbf{x}_1 \neq \mathbf{x}_2$, and $0 \leq \theta \leq 1$ and set $\mathbf{x} = \theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2$.

Using (3.4) twice, we get

$$f(\mathbf{x}_1) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{x}_1 - \mathbf{x}), \tag{3.5}$$

$$f(\mathbf{x}_2) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{x}_2 - \mathbf{x}). \tag{3.6}$$

Multiplying (3.5) by θ and (3.6) with $(1 - \theta)$ and adding the inequalities, we obtain

$$\begin{aligned} \theta f(\mathbf{x}_1) + (1 - \theta) f(\mathbf{x}_2) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T \left(\theta\mathbf{x}_1 + (1 - \theta)\mathbf{x}_2 - \mathbf{x}\right) \\ &= f(\mathbf{x}), \end{aligned}$$

proving that f is convex. □

3.3 Second Order Condition

Theorem 3.3.1. Let $\mathbf{dom}f \subseteq \mathbb{R}^n$ be an open set and $f : \mathbf{dom}f \to \mathbb{R}$ be continuously twice differentiable function. Function f is convex if, and only if, $\mathbf{dom}f$ is a convex set and

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{O} \tag{3.7}$$

for any $\mathbf{x} \in \mathbf{dom}f$.

Proof. (Direct) Let f be a convex function, $\mathbf{x} \in \mathbf{dom}f$, and $\mathbf{h} \in \mathbb{R}^n$. Since $\mathbf{dom}f$ is an open set, we have that $\mathbf{x} + \lambda\mathbf{h} \in \mathbf{dom}f$, for $|\lambda|$ sufficiently small. From (3.4), we get

$$f(\mathbf{x} + \lambda\mathbf{h}) \geq f(\mathbf{x}) + \lambda\nabla f(\mathbf{x})^T\mathbf{h}.$$

From the second-order Taylor expansion, we have

$$f(\mathbf{x} + \lambda\mathbf{h}) = f(\mathbf{x}) + \lambda\nabla f(\mathbf{x})^T\mathbf{h} + \frac{\lambda^2}{2}\,\mathbf{h}^T\nabla^2 f(\mathbf{x})\mathbf{h} + o(\lambda^2).$$

Combining the above two relations, we obtain that

$$\frac{\lambda^2}{2}\,\mathbf{h}^T\nabla^2 f(\mathbf{x})\mathbf{h} + o(\lambda^2) \geq 0.$$

Dividing by λ^2, we obtain

$$\frac{1}{2}\,\mathbf{h}^T\nabla^2 f(\mathbf{x})\mathbf{h} + o(1) \geq 0.$$

Taking the limit as $\lambda \to 0$, we obtain

$$\mathbf{h}^T\nabla^2 f(\mathbf{x})\mathbf{h} \geq 0,$$

for every $\mathbf{h} \in \mathbb{R}^n$, which gives that, for every $\mathbf{x} \in \mathbf{dom}f$,

$$\nabla^2 f(\mathbf{x}) \succeq \mathbf{O}.$$

(*Converse*) We must prove that, if $\nabla^2 f(\mathbf{x}) \succeq \mathbf{O}$ for every $\mathbf{x} \in \mathbf{dom}f$, then f is convex.

From the second-order Taylor expansion, we get

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}\,(\mathbf{y} - \mathbf{x})^T\nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}),$$

with $\mathbf{z} = \lambda\mathbf{x} + (1 - \lambda)\mathbf{y}$, for some $\lambda \in (0, 1)$.

Since we have assumed that $\nabla^2 f(\mathbf{x}) \succeq \mathbf{O}$ for every $\mathbf{x} \in \mathbf{dom} f$, we have that $\nabla^2 f(\mathbf{z}) \succeq \mathbf{O}$, which implies that

$$(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}) \geq 0.$$

Therefore, for each $\mathbf{x}, \mathbf{y} \in \mathbf{dom} f$,

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T (\mathbf{y} - \mathbf{x}),$$

and, by Theorem 3.4, we conclude that f is convex. □

3.4 Examples of Convex Functions

Example 3.4.1. In the sequel, we provide examples of scalar convex and concave functions.

1. $f(x) = e^{ax}$ is convex in $\mathbb{R}$, for every $a \in \mathbb{R}$.
2. $f(x) = x^a$ is convex in $\mathbb{R}_{++}$ if $a \geq 1$ or $a \leq 0$, and concave if $0 \leq a \leq 1$.
3. $f(x) = |x|^p$ is convex in $\mathbb{R}$, for $p \geq 1$.
4. $f(x) = \log(x)$ is concave in $\mathbb{R}_{++}$.

Example 3.4.2. Quadratic functions. Let $f : \mathbb{R}^n \to \mathbb{R}$, with

$$f(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T \mathbf{P} \mathbf{x} + \mathbf{q}^T \mathbf{x} + r,$$

with $\mathbf{P} \in \mathbb{R}^{n \times n}$ and $\mathbf{P} = \mathbf{P}^T$, $\mathbf{q} \in \mathbb{R}^n$ and $r \in \mathbb{R}$.

It can be proved that

$$\nabla f(\mathbf{x}) = \mathbf{P}\mathbf{x} + \mathbf{q} \tag{3.8}$$

and

$$\nabla^2 f(\mathbf{x}) = \mathbf{P}. \tag{3.9}$$

The proof follows from a more general approach which is based on Taylor expansions and proceeds as follows. Let $g : \mathbb{R}^n \to \mathbb{R}$. If we can express $g(\mathbf{x} + \mathbf{d})$ as

$$g(\mathbf{x} + \mathbf{d}) = g(\mathbf{x}) + \mathbf{a}^T \mathbf{d} + \frac{1}{2} \mathbf{d}^T \mathbf{A} \mathbf{d} + o(\|\mathbf{d}\|^2), \tag{3.10}$$

where $\mathbf{A} = \mathbf{A}^T$, then

$$\nabla g(\mathbf{x}) = \mathbf{a} \tag{3.11}$$

and

$$\nabla^2 g(\mathbf{x}) = \mathbf{A}. \tag{3.12}$$

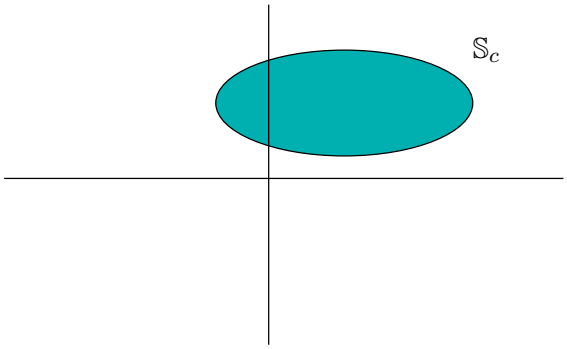


Figure 3.4: Sublevel set of a convex function.

We proceed as follows:

$$\begin{aligned} f(\mathbf{x}+\mathbf{d}) &= \frac{1}{2}(\mathbf{x}+\mathbf{d})^T\mathbf{P}(\mathbf{x}+\mathbf{d}) + \mathbf{q}^T(\mathbf{x}+\mathbf{d}) + r \\ &= \left(\frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + r\right) + \frac{1}{2}\mathbf{d}^T\mathbf{P}\mathbf{x} + \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{d} + \mathbf{q}^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T\mathbf{P}\mathbf{d} \qquad (3.13) \\ &\overset{(a)}{=} f(\mathbf{x}) + (\mathbf{P}\mathbf{x}+\mathbf{q})^T\mathbf{d} + \frac{1}{2}\mathbf{d}^T\mathbf{P}\mathbf{d}, \end{aligned}$$

where at point (a) we used the fact that $\mathbf{d}^T\mathbf{P}\mathbf{x} = \mathbf{x}^T\mathbf{P}\mathbf{d}$ (why?).

From the second-order conditions, we conclude that f is convex (strictly convex) if, and only if, $\mathbf{P} \succeq \mathbf{O}$ ($\mathbf{P} \succ \mathbf{O}$).

Example 3.4.3. In the sequel, we provide examples of multivariable convex and concave functions.

1. $f(\mathbf{x}) = \|\mathbf{x}\|$ is convex in $\mathbb{R}^n$, where $\|\cdot\|$ is any norm in $\mathbb{R}^n$.
2. $f(\mathbf{x}) = \max\{x_1, \ldots, x_n\}$ is convex in $\mathbb{R}^n$.
3. The function $f(\mathbf{x}) = \log\left(e^{x_1} + \cdots + e^{x_n}\right)$ is convex in $\mathbb{R}^n$.
4. The function $f(\mathbf{x}) = \left(\prod_{i=1}^n x_i\right)^{\frac{1}{n}}$ is concave in $\mathbb{R}^n_{++}$.

3.5 Sublevel Sets

Definition 3.5.1. Let $f : \mathbf{dom}f \subseteq \mathbb{R}^n \to \mathbb{R}$. The set

$$\mathbb{S}_c = \{\mathbf{x} \in \mathbf{dom}f \mid f(\mathbf{x}) \leq c\} \qquad (3.14)$$

is called the c-**sublevel set** of f.

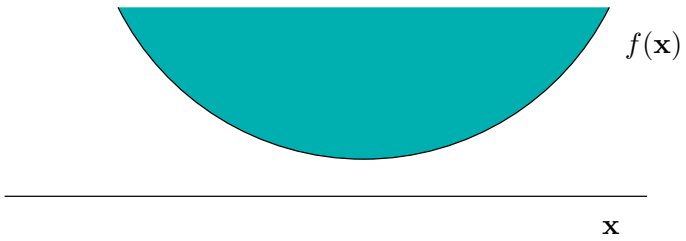


Figure 3.5: Epigraph of a convex function.

Theorem 3.5.1. If $f : \mathbf{dom}f \subseteq \mathbb{R}^n \to \mathbb{R}$ is a convex function, then the c-sublevel set $\mathbb{S}_c$ is convex, for every $c \in \mathbb{R}$.

Proof. [We must prove that if $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{S}_c$ *and* $0 \leq \theta \leq 1$*, then* $\mathbf{x} = \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathbb{S}_c$*].*

Let $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{S}_c$ and $0 \leq \theta \leq 1$. Then, $f(\mathbf{x}_1) \leq c$ and $f(\mathbf{x}_2) \leq c$, and $\mathbf{x} = \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathbf{dom}f$. Furthermore,

$$\begin{aligned} f(\mathbf{x}) = f(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2) &\leq \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2) \\ &\leq \theta\, c + (1-\theta)\, c \\ &= c. \end{aligned}$$

Therefore, $\mathbf{x} = \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathbb{S}_c$. □

We must point out that the converse proposition does *not* hold. For example, the sublevel sets of the function $f : \mathbb{R}_{++} \to \mathbb{R}$, with $f(x) = \log(x)$, are convex. But, the function f is concave.

3.6 Epigraph

Definition 3.6.1. Let $f : \mathbf{dom}f \subseteq \mathbb{R}^n \to \mathbb{R}$. The set

$$\mathbf{epi}f = \{(\mathbf{x}, t) \in \mathbb{R}^{n+1} \,|\, \mathbf{x} \in \mathbf{dom}f, f(\mathbf{x}) \leq t\}$$

is called the **epigraph** of f.

Theorem 3.6.1. Let $f : \mathbf{dom}f \subseteq \mathbb{R}^n \to \mathbb{R}$. Function f is convex if, and only if, $\mathbf{epi}f$ is a convex set.

Proof. (Direct). [We must prove that if f is a convex function, then $\mathbf{epi}f$ *is a convex set.]*

We assume that f is convex. To prove that $\mathbf{epi}f$ is a convex set, we need to prove that if

$$(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2) \in \mathbf{epi}f, \quad 0 \leq \theta \leq 1,$$

then

$$\theta(\mathbf{x}_1, t_1) + (1-\theta)(\mathbf{x}_2, t_2) = (\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2, \theta t_1 + (1-\theta)t_2) \in \mathbf{epi}f.$$

Assume that $(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2) \in \mathbf{epi}f$, which implies that

$$\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{dom}f, \ f(\mathbf{x}_1) \leq t_1, \ f(\mathbf{x}_2) \leq t_2. \tag{3.15}$$

Since f is convex, we have, for $0 \leq \theta \leq 1$, $\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathbf{dom}f$ and

$$\begin{aligned} f(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2) &\leq \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2) \\ &\overset{(3.15)}{\leq} \theta t_1 + (1-\theta) t_2, \end{aligned}$$

that is, $\theta(\mathbf{x}_1, t_1) + (1-\theta)(\mathbf{x}_2, t_2) \in \mathbf{epi}f$.

(Converse). [We must prove that if $\mathbf{epi}f$ *is a convex set, then f is a convex function.]*

We assume that the set $\mathbf{epi}f$ is convex. We know that the projection of $\mathbf{epi}f$ onto its first n coordinates is a convex set. Therefore, the set $\mathbf{dom}f$ is convex.

In addition, if $(\mathbf{x}_1, t_1), (\mathbf{x}_2, t_2) \in \mathbf{epi}f$ and $0 \leq \theta \leq 1$, then

$$\theta(\mathbf{x}_1, t_1) + (1-\theta)(\mathbf{x}_2, t_2) \in \mathbf{epi}f. \tag{3.16}$$

Note that, in the above expression, we can take $t_1 = f(\mathbf{x}_1)$ and $t_2 = f(\mathbf{x}_2)$ (why?). In this case, we have that

$$\theta(\mathbf{x}_1, f(\mathbf{x}_1)) + (1-\theta)(\mathbf{x}_2, f(\mathbf{x}_2)) = (\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2, \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2)) \in \mathbf{epi}f, \tag{3.17}$$

that is,

$$f(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2) \leq \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2).$$

Thus, f is a convex function. □

3.7 Function operations that preserve convexity

3.7.1 Non-negative weighted sums

Theorem 3.7.1. Let $f_k : \mathbb{R}^n \to \mathbb{R}$, for $k = 1, \ldots, K$, be convex functions and $w_k \geq 0$, for $k = 1, \ldots, K$. Then, the function

$$f(\mathbf{x}) = \sum_{k=1}^{K} w_k f_k(\mathbf{x})$$

is convex.

Proof. From the convexity of f_k, for $k = 1, \ldots, K$, we have that, if $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and $0 \leq \theta \leq 1$, then

$$f_k(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2) \leq \theta f_k(\mathbf{x}_1) + (1-\theta) f_k(\mathbf{x}_2), \text{ for } k = 1, \ldots, K.$$

Multiplying both sides of each inequality by w_k, for $k = 1, \ldots, K$, and summing, we get

$$\begin{aligned} \sum_{k=1}^{K} w_k f_k(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2) &\leq \sum_{k=1}^{K} w_k \left(\theta f_k(\mathbf{x}_1) + (1-\theta) f_k(\mathbf{x}_2)\right) \\ &= \theta \sum_{k=1}^{K} w_k f_k(\mathbf{x}_1) + (1-\theta) \sum_{k=1}^{K} w_k f_k(\mathbf{x}_2), \end{aligned}$$

which gives that

$$f(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2) \leq \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2).$$

□

3.7.2 Composition with an affine function

Theorem 3.7.2. Let $f : \mathbf{dom} f \subseteq \mathbb{R}^n \to \mathbb{R}$, $\mathbf{A} \in \mathbb{R}^{n \times m}$, and $\mathbf{b} \in \mathbb{R}^n$. Let $g : \mathbb{R}^m \to \mathbb{R}$, with $g(\mathbf{x}) = f(\mathbf{A}\mathbf{x} + \mathbf{b})$ and domain $\mathbf{dom}\, g = \{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{A}\mathbf{x} + \mathbf{b} \in \mathbf{dom}\, f\}$. If f is a convex function, then g is also convex. If f is a concave function, then g is also concave.

Proof. It is left to the reader. □

3.7.3 Pointwise maxima

Theorem 3.7.3. Let $f_k : \mathbb{R}^n \to \mathbb{R}$, for $k = 1, \ldots, K$, be convex functions. Then, the function

$$f(\mathbf{x}) = \max\{f_1(\mathbf{x}), \ldots, f_K(\mathbf{x})\}$$

is convex.

Proof. From the convexity of f_k, for $k = 1, \ldots, K$, we have that, if $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^n$ and $0 \leq \theta \leq 1$, then

$$f_k(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2) \leq \theta f_k(\mathbf{x}_1) + (1-\theta) f_k(\mathbf{x}_2), \text{ for } k = 1, \ldots, K.$$

Therefore,

$$\begin{aligned}
\max_{k=1,\ldots,K} \{f_k(\theta \mathbf{x}_1 + (1-\theta)\mathbf{x}_2)\} &\leq \max_{k=1,\ldots,K} \{\theta f_k(\mathbf{x}_1) + (1-\theta) f_k(\mathbf{x}_2)\} \\
&\overset{(!)}{\leq} \max_{k=1,\ldots,K} \{\theta f_k(\mathbf{x}_1)\} + \max_{k=1,\ldots,K} \{(1-\theta) f_k(\mathbf{x}_2)\} \\
&= \theta \max_{k=1,\ldots,K} \{f_k(\mathbf{x}_1)\} + (1-\theta) \max_{k=1,\ldots,K} \{f_k(\mathbf{x}_2)\} \\
&= \theta f(\mathbf{x}_1) + (1-\theta) f(\mathbf{x}_2).
\end{aligned}$$

The only non-trivial inequality is (!) , which, essentially, is equivalent to the inequality

$$\max_{k=1,\ldots,K} \{a_k + b_k\} \leq \max_{k=1,\ldots,K} \{a_k\} + \max_{k=1,\ldots,K} \{b_k\},$$

for any scalars a_k, b_k, for $k = 1, \ldots, K$ (try to prove this inequality). □

3.7.4 Composition of functions

Let $f, h : \mathbb{R} \to \mathbb{R}$ be twice differentiable and let $g(t) = (f \circ h)(t)$. Assume that $\mathbf{dom} g$ is an interval. Then g is convex if, and only if, $D^2 g(t) \geq 0$ for every t in the interior of $\mathbf{dom} g$.

We know that $Dg(t) = Df(h(t))\, Dh(t)$ and

$$D^2 g(t) = D^2 f(h(t)) \left(Dh(t)\right)^2 + Df(h(t)) D^2 h(t).$$

With appropriate choice of f and h, we can have $D^2 g(t) \geq 0$ for each $t \in \mathbf{dom} g$.

For example, if f is convex and non-decreasing, which implies that $Df(t) \geq 0$ and $D^2 f(t) \geq 0$, and h is convex, which implies that $D^2 h(t) \geq 0$, then g is convex (why?).

For more examples, see pages 83–87 of Boyd–Vandenberghe.

3.7.5 Partial minimization

Theorem 3.7.4. Let $f : \mathbb{R}^{n+m} \to \mathbb{R}$, $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Let $f(\mathbf{x}, \mathbf{y})$ be convex, with respect to $(\mathbf{x}, \mathbf{y})$, and let $\mathbb{C} \subseteq \mathbb{R}^m$ be a nonempty convex set. Then, $g(\mathbf{x}) = \inf_{\mathbf{y} \in \mathbb{C}} f(\mathbf{x}, \mathbf{y})$ is convex, with respect to $\mathbf{x}$, if $g(\mathbf{x}) > -\infty$ for every $\mathbf{x}$ in

$$\mathbf{dom}\, g = \{\mathbf{x} \,|\, (\mathbf{x}, \mathbf{y}) \in \mathbf{dom} f \text{ for some } \mathbf{y} \in \mathbb{C}\},$$

and $g(\mathbf{x}) < \infty$, for some $\mathbf{x} \in \mathbf{dom} g$.

Proof. We must prove that, if $\mathbf{x}_1, \mathbf{x}_2 \in \mathbf{dom} g$ and $0 \leq \theta \leq 1$, then $\mathbf{x} = \theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2 \in \mathbf{dom} g$ and $g(\mathbf{x}) \leq \theta g(\mathbf{x}_1) + (1-\theta) g(\mathbf{x}_2)$.

For each $\epsilon > 0$, we can find $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{C}$ such that

$$(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2) \in \mathbf{dom} f$$

$$g(\mathbf{x}_1) + \epsilon \geq f(\mathbf{x}_1, \mathbf{y}_1), \quad g(\mathbf{x}_2) + \epsilon \geq f(\mathbf{x}_2, \mathbf{y}_2).$$

Since $\mathbb{C}$ and $\mathbf{dom} f$ are convex, $\theta\mathbf{y}_1 + (1-\theta)\mathbf{y}_2 \in \mathbb{C}$ and $(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2, \theta\mathbf{y}_1 + (1-\theta)\mathbf{y}_2) \in \mathbf{dom} f$. Multiplying the first inequality by θ and the second with $(1-\theta)$ and adding, we obtain

$$\begin{aligned} \theta g(\mathbf{x}_1) + (1-\theta) g(\mathbf{x}_2) + \epsilon &\geq \theta f(\mathbf{x}_1, \mathbf{y}_1) + (1-\theta) f(\mathbf{x}_2, \mathbf{y}_2) \\ &\geq f(\theta\mathbf{x}_1 + (1-\theta)\mathbf{x}_2, \theta\mathbf{y}_1 + (1-\theta)\mathbf{y}_2) \\ &= f(\mathbf{x}, \theta\mathbf{y}_1 + (1-\theta)\mathbf{y}_2) \\ &\geq g(\mathbf{x}). \end{aligned} \tag{3.18}$$

Therefore, $\mathbf{x} \in \mathbf{dom} g$ and, since the inequality holds for every $\epsilon > 0$, we have that $g(\mathbf{x}) \leq \theta g(\mathbf{x}_1) + (1-\theta) g(\mathbf{x}_2)$. Therefore, g is a convex function (see, OPTIII, p. 68). □

Appendix

Sketch of proof of Theorem 3.1.1, assuming $f : \mathbb{R}^n \to \mathbb{R}$. (Work out the details for $f : \mathbf{dom} f \subseteq \mathbb{R}^n \to \mathbb{R}$, with $\mathbf{dom} f$ a convex set).

Let $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, and $g_{\mathbf{x},\mathbf{y}} : \mathbb{R} \to \mathbb{R}$, with $g_{\mathbf{x},\mathbf{y}}(t) = f(t\mathbf{x} + (1-t)\mathbf{y})$. Hereafter, to simplify the notation, $g_{\mathbf{x},\mathbf{y}}$ will be denoted as g.

Direct: If f is a convex function, then g is a convex function.

If $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function and $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then $g_{\mathbf{x},\mathbf{y}} : \mathbb{R} \to \mathbb{R}$, with $g_{\mathbf{x},\mathbf{y}}(t) = f(t\mathbf{x} + (1-t)\mathbf{y})$, is a convex function.

To prove that g is a convex function, we must prove that, for every $t_1, t_2 \in \mathbb{R}$ and $0 \leq \lambda \leq 1$, it must be true that

$$g(\lambda t_1 + (1-\lambda)t_2) \leq \lambda g(t_1) + (1-\lambda)g(t_2). \tag{3.19}$$

Equivalently, we must prove that

$$\begin{aligned} f\left((\lambda t_1 + (1-\lambda)t_2)\mathbf{x} + (1-(\lambda t_1 + (1-\lambda)t_2))\,\mathbf{y}\right) \leq \\ \lambda f\left(t_1\mathbf{x} + (1-t_1)\mathbf{y}\right) + (1-\lambda) f\left(t_2\mathbf{x} + (1-t_2)\mathbf{y}\right). \end{aligned} \tag{3.20}$$

After simple algebraic operations, we get

$$\begin{aligned} (\lambda t_1 + (1-\lambda)t_2)\mathbf{x} + (1-(\lambda t_1 + (1-\lambda)t_2))\,\mathbf{y} = \\ \lambda\left(t_1\mathbf{x} + (1-t_1)\mathbf{y}\right) + (1-\lambda)\left(t_2\mathbf{x} + (1-t_2)\mathbf{y}\right). \end{aligned} \tag{3.21}$$

Therefore, (3.20) is equivalently written as

$$\begin{aligned} f\left(\lambda\left(t_1\mathbf{x} + (1-t_1)\mathbf{y}\right) + (1-\lambda)\left(t_2\mathbf{x} + (1-t_2)\mathbf{y}\right)\right) \leq \\ \lambda f\left(t_1\mathbf{x} + (1-t_1)\mathbf{y}\right) + (1-\lambda) f\left(t_2\mathbf{x} + (1-t_2)\mathbf{y}\right), \end{aligned} \tag{3.22}$$

which is true, because we have assumed that f is a convex function.

Converse: If the restriction of f to every line is convex, then f is a convex function.

We must prove that, if $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and $0 \leq \lambda \leq 1$, then

$$f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}) \leq \lambda f(\mathbf{x}) + (1-\lambda) f(\mathbf{y}). \tag{3.23}$$

From the definition of g, we have

$$g(\lambda) = f(\lambda\mathbf{x} + (1-\lambda)\mathbf{y}),\ g(1) = f(\mathbf{x}),\ g(0) = f(\mathbf{y}). \tag{3.24}$$

From the convexity of g, we get

$$g(\lambda) = g(\lambda \cdot 1 + (1-\lambda)\cdot 0) \leq \lambda g(1) + (1-\lambda) g(0), \tag{3.25}$$

which is equivalent to (3.23), completing the proof.

Chapter 4

Optimization Problems

4.1 Optimization Problems

Definition 4.1.1. An optimization problem is defined as

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m, \\ & h_i(\mathbf{x}) = 0, \quad i = 1, \ldots, p, \end{array} \tag{4.1}$$

where

1. the vector $\mathbf{x} \in \mathbb{R}^n$ is called the **optimization variable**,
2. the function $f_0 : \mathbf{dom}\, f_0 \subseteq \mathbb{R}^n \to \mathbb{R}$ is called **cost function** or **objective function**,
3. the inequalities $f_i(\mathbf{x}) \leq 0$, with $f_i : \mathbf{dom}\, f_i \subseteq \mathbb{R}^n \to \mathbb{R}$, for $i = 1, \ldots, m$, are called **inequality constraints**, and
4. the equalities $h_i(\mathbf{x}) = 0$, with $h_i : \mathbf{dom}\, h_i \subseteq \mathbb{R}^n \to \mathbb{R}$, for $i = 1, \ldots, p$, are called **equality constraints**.

The set of points for which all functions are defined is

$$\mathbb{D} := \bigcap_{i=0}^{m} \mathbf{dom}\, f_i \cap \bigcap_{i=1}^{p} \mathbf{dom}\, h_i. \tag{4.2}$$

Definition 4.1.2. A point $\mathbf{x} \in \mathbb{D}$ is called **feasible** if it satisfies all constraints.

An optimization problem is called **feasible** if there exists a feasible point for that problem. Otherwise, it is called **infeasible**.

The set of feasible points of an optimization problem is called **feasible set** of the problem. The feasible set of the optimization problem (4.1) is the set

$$\mathbb{X} := \{\mathbf{x} \in \mathbb{D} \,|\, f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m,\ h_i(\mathbf{x}) = 0,\ i = 1, \ldots, p\}. \tag{4.3}$$

The **optimal value** of the problem (4.1) is defined as

$$p_* := \inf \{f_0(\mathbf{x}) \,|\, \mathbf{x} \in \mathbb{X}\}. \tag{4.4}$$

p_* can take the values $\pm\infty$. If the problem is infeasible, then $p_* = \infty$. If there are feasible points $\mathbf{x}_k \in \mathbb{D}$, with $f_0(\mathbf{x}_k) \to -\infty$, when $k \to \infty$, then $p_* = -\infty$ and the problem is called **unbounded below**.

4.1.1 Local and Global Minima

Definition 4.1.3. A point $\mathbf{x}_* \in \mathbb{X}$ is called **optimal point** if $f_0(\mathbf{x}_*) = p_*$. The set

$$\mathbb{X}_{\text{opt}} := \{\mathbf{x} \in \mathbb{X} \,|\, f_0(\mathbf{x}) = p_*\} \tag{4.5}$$

is called **optimal set**.

Alternatively, we say that $\mathbf{x} \in \mathbb{X}$ is an optimal point if $f_0(\mathbf{x}) \leq f_0(\mathbf{y})$ for every $\mathbf{y} \in \mathbb{X}$.

If there is an optimal point for the problem (4.1), then we say that the problem has a solution and the optimal value is attained.

If the set $\mathbb{X}_{\text{opt}}$ is empty, then we say that the optimal value is not attained (this is always the case when the problem is unbounded from below).

Definition 4.1.4. The point $\mathbf{x} \in \mathbb{X}$ is called **locally optimal** if there exists $\epsilon > 0$ such that $f_0(\mathbf{x}) \leq f_0(\mathbf{y})$ for every $\mathbf{y} \in \mathbb{X}$, with $\|\mathbf{y} - \mathbf{x}\|_2 < \epsilon$.

If $\mathbf{x} \in \mathbb{X}$ and $f_i(\mathbf{x}) = 0$, for some i in the set $\{1, \ldots, m\}$, then we say that the i-th inequality constraint is **active** at $\mathbf{x}$. On the other hand, if $f_i(\mathbf{x}) < 0$, for some i in the set $\{1, \ldots, m\}$, then we say that the i-th inequality constraint is **inactive** at $\mathbf{x}$.

A constraint is called **redundant** if its deletion does not change the feasible set.

4.2 Convex Optimization Problems

Definition 4.2.1. A problem of the form

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ & \mathbf{a}_i^T \mathbf{x} = b_i, \quad i = 1, \ldots, p, \end{array} \tag{4.6}$$

is called a **convex optimization problem** if the functions f_i, for $i = 0, \ldots, m$, are convex. Here, $\mathbb{D} := \bigcap_{i=0}^m \mathbf{dom}\, f_i$.

The feasible set of the optimization problem (4.6) is

$$\mathbb{X} := \{\mathbf{x} \in \mathbb{D} \,|\, f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m,\ \mathbf{a}_i^T \mathbf{x} = b_i,\ i = 1, \ldots, p\}. \tag{4.7}$$

Theorem 4.2.1. The feasible set of a convex optimization problem is convex.

Proof. The feasible set of a convex optimization problem is the intersection of the (convex) domains of the functions f_i, for $i = 0, \ldots, m$, the (convex) 0-sublevel sets $\{\mathbf{x} \in \mathbf{dom} f_i \,|\, f_i(\mathbf{x}) \leq 0\}$, for $i = 1, \ldots, m$, and the hyperplanes $\{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{a}_i^T \mathbf{x} = b_i\}$, for $i = 1, \ldots, p$. Therefore, it is convex. □

Theorem 4.2.2. Consider the convex optimization problem (4.6). Then, the set of points $\mathbf{x} \in \mathbb{X}$ which minimize f_0 is convex. Moreover, every local minimum of f_0 is a global minimum.

Proof. Assume that the problem has a solution and optimal value p_*. Then, the set of optimal points is the p_*-sublevel set of f_0 restricted to the convex feasible set $\mathbb{X}$,

$$\mathbb{X}_{\text{opt}} = \{\mathbf{x} \in \mathbb{X} \,|\, f_0(\mathbf{x}) \leq p_*\}, \tag{4.8}$$

and is therefore convex.

(We shall prove the second statement by Reductio ad Absurdum). Let $\mathbf{x} \in \mathbb{X}$ be a local but not global minimum of f_0. Then, there exists $\mathbf{y} \in \mathbb{X}$ such that $f_0(\mathbf{y}) < f_0(\mathbf{x})$. If we consider f_0 on the line segment $\theta\mathbf{y} + (1-\theta)\mathbf{x}$, with $0 \leq \theta \leq 1$, we have

$$f_0(\theta\mathbf{y} + (1-\theta)\mathbf{x}) \leq \theta f_0(\mathbf{y}) + (1-\theta) f_0(\mathbf{x}) < f_0(\mathbf{x}), \tag{4.9}$$

For every $\theta > 0$, the right-hand side is strictly smaller than $f_0(\mathbf{x})$. Moreover, by choosing $\theta > 0$ sufficiently small, we can make $\|\theta(\mathbf{y} - \mathbf{x})\|_2$ arbitrarily small. Hence there are feasible points arbitrarily close to $\mathbf{x}$ with strictly smaller objective value, contradicting the assumption that $\mathbf{x}$ is a local minimum. The proof is complete. □

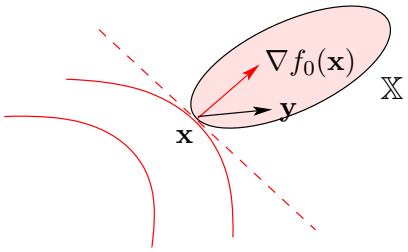


Figure 4.1: Geometric interpretation of the optimality condition (4.12).

4.3 An Optimality Criterion for Differentiable Cost Functions

Let $f_0 : \mathbf{dom}\, f_0 \subseteq \mathbb{R}^n \to \mathbb{R}$ be a convex and differentiable function. We know that, for every $\mathbf{x}, \mathbf{y} \in \mathbf{dom}\, f_0$, the following inequality holds:

$$f_0(\mathbf{y}) \geq f_0(\mathbf{x}) + \nabla f_0(\mathbf{x})^T(\mathbf{y} - \mathbf{x}). \tag{4.10}$$

Theorem 4.3.1. Let

$$\mathbb{X} := \{\mathbf{x} \in \mathbb{D} \,|\, f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m,\ \mathbf{a}_i^T\mathbf{x} = b_i,\ i = 1, \ldots, p\}. \tag{4.11}$$

The point $\mathbf{x} \in \mathbb{X}$ is optimal for the convex optimization problem (4.6) if, and only if,

$$\nabla f_0(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) \geq 0, \text{ for every } \mathbf{y} \in \mathbb{X}. \tag{4.12}$$

Proof. (*Direct, through reductio ad absurdum*) Assume that $\mathbf{x} \in \mathbb{X}$ is optimal but there exists a point $\mathbf{y} \in \mathbb{X}$ such that $\nabla f_0(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) < 0$. If we define, for $0 \leq \theta \leq 1$,

$$\mathbf{z}(\theta) := \theta\mathbf{y} + (1 - \theta)\mathbf{x} = \mathbf{x} + \theta(\mathbf{y} - \mathbf{x}), \tag{4.13}$$

we have that $\mathbf{z}(\theta) \in \mathbb{X}$, for every $0 \leq \theta \leq 1$, because $\mathbb{X}$ is convex. For small and positive θ, we have that $f_0(\mathbf{z}(\theta)) < f_0(\mathbf{x})$ because[1]

$$\frac{d}{d\theta} f_0(\mathbf{z}(\theta))\bigg|_{\theta=0} = \nabla f_0(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) < 0. \tag{4.14}$$

Therefore, $\mathbf{x}$ is not optimal, which is false.

[1]For a detailed proof, see the Appendix at the end of this chapter.

(*Converse*) If (4.12) holds, then, due to (4.10), we have that

$$f_0(\mathbf{y}) \geq f_0(\mathbf{x}), \quad \text{for each } \mathbf{y} \in \mathbb{X}, \tag{4.15}$$

that is, $\mathbf{x}$ is an optimal point for the problem (4.6).

The proof is complete. □

4.3.1 Unconstrained Problems

Theorem 4.3.2. Let $f_0 : \mathbb{R}^n \to \mathbb{R}$ be a convex differentiable function. The point $\mathbf{x} \in \mathbb{R}^n$ is an optimal point for the unconstrained minimization problem if, and only if,

$$\nabla f_0(\mathbf{x}) = \mathbf{0}. \tag{4.16}$$

Proof. By Theorem 4.3.1, we should have

$$\nabla f_0(\mathbf{x})^T \mathbf{h} \geq 0, \quad \text{for every } \mathbf{h} \in \mathbb{R}^n. \tag{4.17}$$

This means that $\nabla f_0(\mathbf{x}) = \mathbf{0}$. Because, if $\nabla f_0(\mathbf{x}) \neq \mathbf{0}$ and we choose $\mathbf{h} = -\nabla f_0(\mathbf{x})$, we get that

$$-\|\nabla f_0(\mathbf{x})\|_2^2 \geq 0, \tag{4.18}$$

which is false. □

4.3.2 Problems with linear constraints

Consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}. \end{array} \tag{4.19}$$

The point $\mathbf{x}_* \in \mathbb{X}$ is optimal if, and only if,

$$\nabla f_0(\mathbf{x}_*)^T (\mathbf{y} - \mathbf{x}_*) \geq 0, \tag{4.20}$$

for every $\mathbf{y}$, with $\mathbf{A}\mathbf{y} = \mathbf{b}$.

4.3.3 Optimization on the non-negative orthant

Let the problem

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) \\ \text{subject to} & \mathbf{x} \succeq \mathbf{0}. \end{array} \tag{4.21}$$

The point $\mathbf{x}_* \succeq \mathbf{0}$ is optimal if, and only if,

$$\nabla f_0(\mathbf{x}_*)^T (\mathbf{y} - \mathbf{x}_*) \geq 0, \text{ for every } \mathbf{y} \succeq \mathbf{0}. \tag{4.22}$$

4.4 Equivalent problems

In rough terms, we say that two optimization problems are equivalent when "from the solution of one we can compute the solution of the other."

Example 4.4.1. Let $r \geq 0$. The problem

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) \\ \text{subject to} & \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2 \leq r, \end{array} \tag{4.23}$$

is equivalent to the problem

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) \\ \text{subject to} & \|\mathbf{A}\mathbf{x} - \mathbf{b}\|_2^2 \leq r^2. \end{array} \tag{4.24}$$

The two problems have the same feasible set and objective function, and therefore the same optimal points and optimal value. But the constraint of the first problem refers to a nondifferentiable function, while the corresponding function for the second problem is differentiable. □

More complex (and more interesting) examples of equivalent problems will be mentioned later in the course.

4.4.1 Epigraph Form

The problem

$$\begin{array}{ll} \underset{\mathbf{x},t}{\text{minimize}} & t \\ \text{subject to} & f_0(\mathbf{x}) - t \leq 0 \\ & f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m, \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \end{array} \tag{4.25}$$

is equivalent to the problem (4.6) and is called optimization problem in **epigraph form**.

4.5 Linear optimization problems

The optimization problem with linear cost function and affine equality and inequality constraints

$$\begin{array}{ll} \text{minimize} & \mathbf{c}^T\mathbf{x} \\ \text{subject to} & \mathbf{G}\mathbf{x} \preceq \mathbf{h} \\ & \mathbf{A}\mathbf{x} = \mathbf{b} \end{array} \tag{4.26}$$

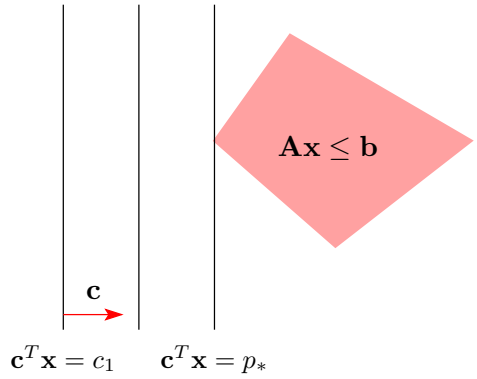


Figure 4.2: Linear optimization problem in inequality form.

is called a **linear programming problem** or **linear program** (see Fig. 4.2). Two forms of linear programs are very common.

The **standard form**, in which the problem is expressed as

$$\begin{array}{ll} \text{minimize} & \mathbf{c}^T\mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} \succeq \mathbf{0} \end{array} \tag{4.27}$$

and the **inequality form**, in which the problem is expressed as

$$\begin{array}{ll} \text{minimize} & \mathbf{c}^T\mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} \preceq \mathbf{b}. \end{array} \tag{4.28}$$

It can be proved that any linear program can be expressed in standard or inequality form (via slack variables).

4.6 Quadratic optimization problems

The optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + r \\ \text{subject to} & \mathbf{G}\mathbf{x} \preceq \mathbf{h} \\ & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{array} \tag{4.29}$$

is called a **quadratic optimization problem**. Assuming $\mathbf{P}$ is symmetric, the problem is convex if $\mathbf{P} \succeq \mathbf{O}$.

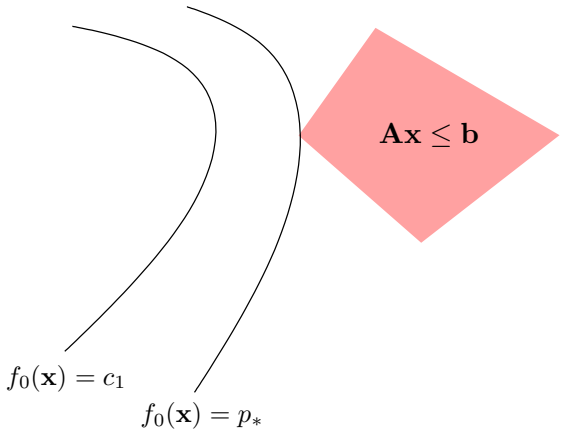


Figure 4.3: Convex quadratic optimization problem.

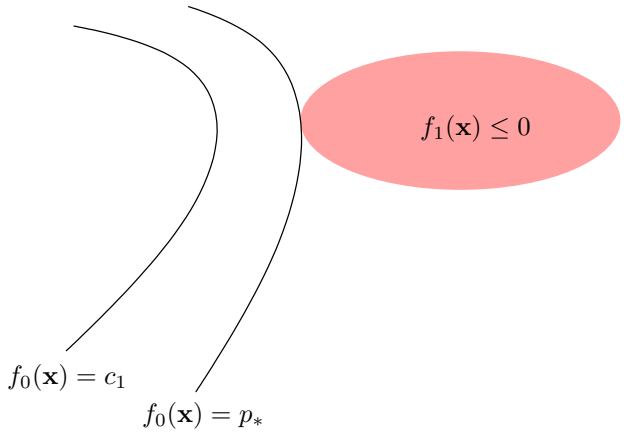


Figure 4.4: Convex quadratic optimization problem with a convex quadratic constraint.

The optimization problem

$$\begin{array}{ll} \text{minimize} & \frac{1}{2}\,\mathbf{x}^T\mathbf{P}_0\mathbf{x} + \mathbf{q}_0^T\mathbf{x} + r_0 \\ \text{subject to} & \frac{1}{2}\,\mathbf{x}^T\mathbf{P}_i\mathbf{x} + \mathbf{q}_i^T\mathbf{x} + r_i \leq 0, \quad i = 1, \ldots, m \\ & \mathbf{Ax} = \mathbf{b}, \end{array} \tag{4.30}$$

is called a **quadratically constrained quadratic program** (QCQP). Assuming the matrices $\mathbf{P}_i$ are symmetric, the problem is convex if $\mathbf{P}_i \succeq \mathbf{O}$, for $i = 0, \ldots, m$.

4.6.1 Least-Squares

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{b} \in \mathbb{R}^m$. The optimization problem

$$\min_{\mathbf{x}} \; f_0(\mathbf{x}) = \frac{1}{2}\,\|\mathbf{Ax} - \mathbf{b}\|_2^2 = \frac{1}{2}\,(\mathbf{Ax} - \mathbf{b})^T(\mathbf{Ax} - \mathbf{b}) \tag{4.31}$$

is called a linear least-squares problem.

Problem: Write the function f_0 as a quadratic function, compute its first and second derivatives, and solve the least-squares problem.

The optimization problem

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & \|\mathbf{A}\mathbf{x}-\mathbf{b}\|_2^2 \\ \text{subject to} & l_i \le x_i \le u_i, \quad i=1,\dots,n, \end{array} \tag{4.32}$$

is called a linear least-squares problem with box constraints. In general, this problem does not admit a simple closed-form solution.

Appendix

Let $a \in \mathbb{R}$. We recall that $f(x) = O(g(x))$, for $x \to a$, if there are constants $\delta > 0$ and $K := K(a,\delta) > 0$ or $K := K(\delta) > 0$ such that

$$|f(x)| \le K|g(x)|, \quad \text{for each } x \in \mathbb{R}, \text{ with } 0 < |x-a| < \delta. \tag{4.33}$$

Theorem 4.6.1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable at $\mathbf{x}$ and let $\Delta\mathbf{x} \in \mathbb{R}^n$ satisfy $\nabla f(\mathbf{x})^T \Delta\mathbf{x} < 0$. Then, for every sufficiently small $t > 0$, we have

$$f(\mathbf{x} + t\Delta\mathbf{x}) < f(\mathbf{x}). \tag{4.34}$$

Proof. Since f is differentiable at $\mathbf{x}$,

$$f(\mathbf{x} + t\Delta\mathbf{x}) = f(\mathbf{x}) + t\nabla f(\mathbf{x})^T \Delta\mathbf{x} + o(t), \qquad t \to 0^+. \tag{4.35}$$

Equivalently,

$$\frac{f(\mathbf{x} + t\Delta\mathbf{x}) - f(\mathbf{x})}{t} = \nabla f(\mathbf{x})^T \Delta\mathbf{x} + o(1). \tag{4.36}$$

Because $\nabla f(\mathbf{x})^T \Delta\mathbf{x} < 0$, the right-hand side is strictly negative for all sufficiently small $t > 0$. Therefore,

$$f(\mathbf{x} + t\Delta\mathbf{x}) - f(\mathbf{x}) < 0, \tag{4.37}$$

and hence $f(\mathbf{x} + t\Delta\mathbf{x}) < f(\mathbf{x})$. □

Chapter 5

Unconstrained convex optimization

5.1 Unconstrained convex optimization problems

The problem we shall study in this chapter is the unconstrained convex optimization problem

$$\min_{\mathbf{x}} f(\mathbf{x}), \tag{5.1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable convex real-valued function.

We assume that the problem (5.1) has a solution and that the minimum value of f equals p_*. Therefore, there exists $\mathbf{x}_* \in \mathbb{R}^n$ such that $f(\mathbf{x}_*) = p_*$.

As we have seen, a point $\mathbf{x}_* \in \mathbb{R}^n$ is a solution of the problem (5.1) if, and only if,

$$\nabla f(\mathbf{x}_*) = \mathbf{0}. \tag{5.2}$$

The relation (5.2) is a system of (usually, non-linear) equations, which rarely has a closed-form solution, and which is usually solved through iterative procedures.

An iterative procedure for solving (5.2) is a procedure which generates a sequence of points $\mathbf{x}_k \in \mathbb{R}^n$ such that $\mathbf{x}_k \to \mathbf{x}_*$, when $k \to \infty$.

The system (5.2) can be solved through direct or indirect iterative procedures.

A direct iterative process tries to compute $\mathbf{x}_*$ by solving the system (5.2), while an indirect one tries to take advantage of properties of f and compute $\mathbf{x}_*$ indirectly.

5.2 Starting point and sublevel sets

The methods we will present start from an initial point $\mathbf{x}_0$ which belongs to $\mathbf{dom}\, f$ and for which it should hold that the set

$$\mathbb{S} = \{\mathbf{x} \,|\, f(\mathbf{x}) \leq f(\mathbf{x}_0)\} \tag{5.3}$$

is closed.[1]

This condition is satisfied if $\mathbf{x}_0 \in \mathbf{dom}\, f$ and f is a closed function, that is, when all sublevel sets of f are closed sets.

Two important cases of closed functions are as follows:

1. continuous functions f with domain $\mathbf{dom}\, f = \mathbb{R}^n$ are closed. For example, the function $f : \mathbb{R}^n \to \mathbb{R}$ with formula

$$f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + c, \tag{5.4}$$

   with $\mathbf{P} = \mathbf{P}^T \succ \mathbf{0}$, is closed.

2. continuous functions with domain $\mathbf{dom}\, f$ open set, for which function $f(\mathbf{x})$ tends to infinity when $\mathbf{x}$ approaches $\mathbf{bd}\,\mathbf{dom}\, f$, are closed. For example, the function $f : \mathbf{dom}\, f \subseteq \mathbb{R}^n \to \mathbb{R}$ defined as

$$f(\mathbf{x}) = -\sum_{i=1}^{m} \log(b_i - \mathbf{a}_i^T\mathbf{x}). \tag{5.5}$$

   The domain of f is the set

$$\mathbf{dom}\, f = \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{a}_i^T\mathbf{x} < b_i,\ i = 1, \ldots, m\}. \tag{5.6}$$

   Notice that, as $\mathbf{x}$ approaches the boundary of the domain, the function f tends to infinity.

5.3 Descent Methods

The first method we will study is an indirect iterative process to solve (5.1) or, equivalently, (5.2) and is described as follows.

[1] This guarantees that limits of sequences of points $\mathbf{x}_k \in \mathbb{S}$, for $k = 1, 2, \ldots$, will be points of $\mathbb{S}$.

Suppose that, at the k-th step of the process, our estimate of the solution of (5.2) is the point $\mathbf{x}_k$, with $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$.

The next point is given by the relation

$$\mathbf{x}_{k+1} = \mathbf{x}_k + t_k \Delta \mathbf{x}_k, \tag{5.7}$$

for suitably chosen $t_k > 0$ and $\Delta \mathbf{x}_k$.

It seems reasonable to choose t_k and $\Delta \mathbf{x}_k$ such that

$$f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k), \tag{5.8}$$

with equality only if $\mathbf{x}_k = \mathbf{x}_*$. Such iterative methods are called **descent methods**.

Due to the convexity of f, we know that if $\nabla f(\mathbf{x}_k)^T(\mathbf{x}_{k+1} - \mathbf{x}_k) = t_k \nabla f(\mathbf{x}_k)^T \Delta \mathbf{x}_k \geq 0$, then $f(\mathbf{x}_{k+1}) \geq f(\mathbf{x}_k)$ (why?). Therefore, in order to develop a descent method, we must choose $\Delta \mathbf{x}_k$ such that

$$\nabla f(\mathbf{x}_k)^T \Delta \mathbf{x}_k < 0, \tag{5.9}$$

that is, $\cos \angle (\Delta \mathbf{x}_k, \nabla f(\mathbf{x}_k)) < 0$. Such a direction of movement is called a **descent direction**.

In this case, t_k is chosen as

$$t_k = \underset{t>0}{\operatorname{argmin}} f(\mathbf{x}_k + t \Delta \mathbf{x}_k). \tag{5.10}$$

5.4 Gradient Descent Method

Among the most common choices for $\Delta \mathbf{x}_k$ is to set

$$\Delta \mathbf{x}_k = -\nabla f(\mathbf{x}_k). \tag{5.11}$$

This direction is called **negative gradient** direction. If $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$, then the choice (5.11) satisfies relation (5.9). Geometrically, we move from the point $\mathbf{x}_k$ along the direction which gives the maximum rate of decrease of f among all directions of the same Euclidean norm.

In Table 5.1, we present the gradient descent method. Usually, for reasons we shall explain later, as termination criterion we use the inequality $\|\nabla f(\mathbf{x}_k)\|_2 < \epsilon$, for some "small" $\epsilon > 0$.

$\mathbf{x}_0 \in \mathbb{R}^n,\ k = 0.$
While (stopping criterion is FALSE)
1. $\Delta\mathbf{x}_k := -\nabla f(\mathbf{x}_k)$.
2. Line search and choose t_k.
3. $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\Delta\mathbf{x}_k$.
4. $k := k + 1$.

Table 5.1: Gradient Descent Algorithm.

5.4.1 Line-Search

As we have seen, for a given descent direction $\Delta\mathbf{x}_k$, the line-search problem is expressed as

$$t_k = \underset{t>0}{\operatorname{argmin}}\ f(\mathbf{x}_k + t\Delta\mathbf{x}_k). \tag{5.12}$$

5.4.2 Exact line-search

When the solution of the problem (5.12) can be expressed in closed-form or is relatively easy to compute accurately, then we adopt the exact line-search method to compute t_k.

Example 5.4.1. Exact line-search for convex quadratic cost function and negative gradient direction.

Let the quadratic cost function $f : \mathbb{R}^n \to \mathbb{R}$, with

$$f(\mathbf{x}) = \frac{1}{2}\,\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x}, \tag{5.13}$$

with $\mathbf{P} \in \mathbb{R}^{n\times n}$, $\mathbf{P} = \mathbf{P}^T \succ \mathbf{O}$, and $\mathbf{q} \in \mathbb{R}^n$.

As we have shown, the gradient of the function f is equal to

$$\nabla f(\mathbf{x}) = \mathbf{P}\mathbf{x} + \mathbf{q}. \tag{5.14}$$

To solve the exact line-search problem, we assume that $\nabla f(\mathbf{x}) \neq \mathbf{0}$, we define the function

$$g(t) := f\left(\mathbf{x} - t\nabla f(\mathbf{x})\right), \tag{5.15}$$

and we look for

$$t_* = \underset{t\geq 0}{\operatorname{argmin}}\, g(t). \tag{5.16}$$

The function $g(t)$ is expressed as follows:

$$\begin{aligned} g(t) &= \frac{1}{2}\left(\mathbf{x} - t\nabla f(\mathbf{x})\right)^T \mathbf{P}\left(\mathbf{x} - t\nabla f(\mathbf{x})\right) + \mathbf{q}^T\left(\mathbf{x} - t\nabla f(\mathbf{x})\right) \\ &= \frac{1}{2}t^2\,\nabla f(\mathbf{x})^T\mathbf{P}\nabla f(\mathbf{x}) - t\left(\mathbf{P}\mathbf{x} + \mathbf{q}\right)^T\nabla f(\mathbf{x}) + \frac{1}{2}\,\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} \\ &= \frac{1}{2}t^2\,\nabla f(\mathbf{x})^T\mathbf{P}\nabla f(\mathbf{x}) - t\,\|\nabla f(\mathbf{x})\|^2 + \frac{1}{2}\,\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} \\ &= \frac{1}{2}\mathtt{a}t^2 - \mathtt{b}t + \mathtt{c}, \end{aligned} \tag{5.17}$$

with $\mathtt{a} := \nabla f(\mathbf{x})^T\mathbf{P}\nabla f(\mathbf{x}) > 0$, $\mathtt{b} := \|\nabla f(\mathbf{x})\|^2$, and $\mathtt{c} := \frac{1}{2}\,\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x}$. The derivative of g is

$$\frac{dg(t)}{dt} = \mathtt{a}t - \mathtt{b}. \tag{5.18}$$

The minimum of $g(t)$ is attained at the solution of the equation $\frac{dg(t)}{dt} = 0$, that is

$$t_* = \frac{\mathtt{b}}{\mathtt{a}} = \frac{\|\nabla f(\mathbf{x})\|^2}{\nabla f(\mathbf{x})^T\mathbf{P}\nabla f(\mathbf{x})} > 0. \tag{5.19}$$

So, in this case, there is a closed-form solution to the exact line-search problem. □

5.4.3 Backtracking line-search

In many cases, solving the problem (5.12) is difficult. In these cases, we can adopt *inexact* line-search techniques (note that there are many such techniques).

A common inexact line-search technique is the **backtracking line-search**. For $\mathbf{x} \in \mathbf{dom} f$, the backtracking line-search takes as input a descent direction, $\Delta\mathbf{x}$, and returns a value $t > 0$ such that the value $f(\mathbf{x} + t\Delta\mathbf{x})$ is "sufficiently" smaller than $f(\mathbf{x})$.

The method uses parameters α and β, with $0 < \alpha < 0.5$ and $0 < \beta < 1$, and is described in Table 5.4.3. It starts by setting $t = 1$, and decreases t by a multiplicative factor β until

$$f(\mathbf{x} + t\Delta\mathbf{x}) \leq f(\mathbf{x}) + \alpha t\nabla f(\mathbf{x})^T\Delta\mathbf{x}. \tag{5.20}$$

1. It can be shown that condition (5.20) is satisfied for all sufficiently small $t > 0$; therefore, the backtracking line-search algorithm always terminates.

   The proof is based on (1) the first-order Taylor expansion and (2) the fact that the vector $\Delta\mathbf{x}$ is a descent direction, that is, $\nabla f(\mathbf{x})^T\Delta\mathbf{x} < 0$.

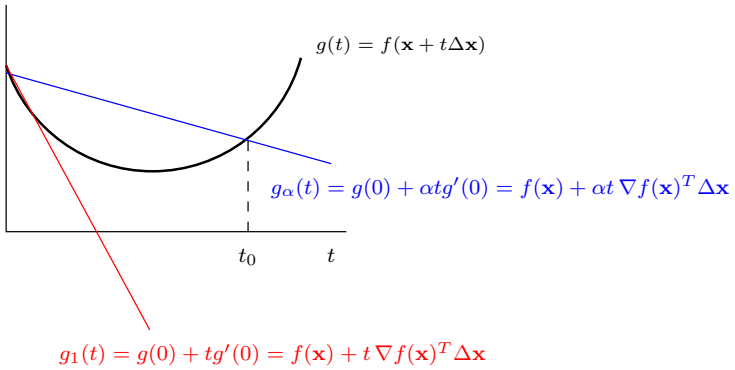


Figure 5.1: Backtracking line-search.

Let $\Delta\mathbf{x}$ be a descent direction at $\mathbf{x}$, $\alpha \in (0, 0.5)$, $\beta \in (0, 1)$.
$t := 1$.
While $\big(f(\mathbf{x} + t\Delta\mathbf{x}) > f(\mathbf{x}) + \alpha t \nabla f(\mathbf{x})^T \Delta\mathbf{x} \big)$
    1. $t := \beta t$.

Table 5.2: Backtracking line-search.

More specifically, from the first-order Taylor expansion, we have that, for sufficiently small t,

$$f(\mathbf{x} + t\Delta\mathbf{x}) = f(\mathbf{x}) + t\nabla f(\mathbf{x})^T \Delta\mathbf{x} + o(t). \tag{5.21}$$

Since $(1 - \alpha)\nabla f(\mathbf{x})^T \Delta\mathbf{x} < 0$, for all sufficiently small $t > 0$ we have

$$f(\mathbf{x} + t\Delta\mathbf{x}) < f(\mathbf{x}) + \alpha t \nabla f(\mathbf{x})^T \Delta\mathbf{x}. \tag{5.22}$$

2. The first-order Taylor estimate of the decrease in the value of f is

$$f(\mathbf{x}) - f(\mathbf{x} + t\Delta\mathbf{x}) \approx -t\nabla f(\mathbf{x})^T \Delta\mathbf{x}. \tag{5.23}$$

Therefore, the backtracking line-search algorithm terminates when the decrease of f is at least $-\alpha t \nabla f(\mathbf{x})^T \Delta\mathbf{x}$, that is, at least α times the first-order Taylor estimate of the decrease (see Figure 5.1).

5.5 Convergence Analysis for Strongly Convex Functions

Convergence analysis of methods for solving optimization problems is extremely important because it provides information about the speed of convergence of the methods.

Additional features of the method, such as, for example, the computational complexity of each step of the method, complete the description of the method.

Convergence analysis is usually accompanied by assumptions about the nature of the optimization problem, for example, assumptions about differentiability, strict or strong convexity of the cost function, etc.

In the sequel, we will study the convergence of the gradient descent method, assuming that the cost function is strongly convex.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a **strongly convex** twice continuously differentiable function. More specifically, we assume that there exist $0 < m \leq M < \infty$ such that

$$m\mathbf{I} \preceq \nabla^2 f(\mathbf{x}) \preceq M\mathbf{I}, \quad \text{for every } \mathbf{x} \in \mathbb{S}, \tag{5.24}$$

where $\mathbf{I}$ is the $(n \times n)$ identity matrix and $\mathbb{S} := \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) \leq f(\mathbf{x}_0)\}$.

In practice, usually, the constants m and M are *unknown*. But, the assumption of their existence allows **convergence analysis** of the gradient descent algorithm and leads to extremely important conclusions. For this reason, in what follows, we will assume the existence of these constants.

Relation (5.24) implies that, for any $\mathbf{x} \in \mathbb{S}$ and $\mathbf{y} \in \mathbb{R}^n$,

$$m\mathbf{y}^T\mathbf{I}\mathbf{y} \leq \mathbf{y}^T\nabla^2 f(\mathbf{x})\mathbf{y} \leq M\mathbf{y}^T\mathbf{I}\mathbf{y}, \tag{5.25}$$

or, equivalently,

$$m\|\mathbf{y}\|_2^2 \leq \mathbf{y}^T\nabla^2 f(\mathbf{x})\mathbf{y} \leq M\|\mathbf{y}\|_2^2. \tag{5.26}$$

Before proceeding, we mention that (5.24) is equivalent to

$$\frac{1}{M}\,\mathbf{I} \preceq \left(\nabla^2 f(\mathbf{x})\right)^{-1} \preceq \frac{1}{m}\,\mathbf{I}, \quad \text{for any } \mathbf{x} \in \mathbb{S}. \tag{5.27}$$

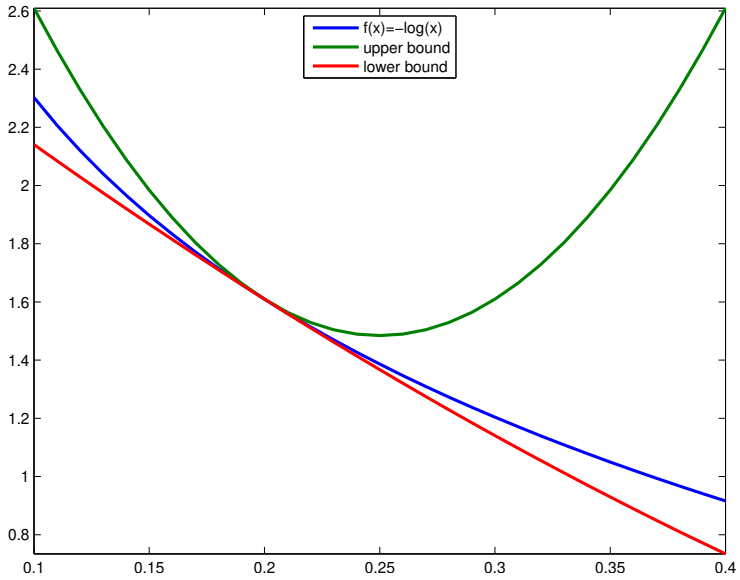


Figure 5.2: Quadratic total overestimator and underestimator of $f(x) = -\log(x)$, in the interval $x \in [0.1, 0.4]$, at the point $\mathbf{x}_0 = 0.2$.

We recall that, for any $\mathbf{x}, \mathbf{y} \in \mathbb{S}$,

$$f(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{z})(\mathbf{y} - \mathbf{x}), \tag{5.28}$$

for some $\mathbf{z}$ on the line segment that connects the points $\mathbf{x}$ and $\mathbf{y}$.

Using (5.26), we can prove the inequalities, for any $\mathbf{x}, \mathbf{y} \in \mathbb{S}$:

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{m}{2}\|\mathbf{y} - \mathbf{x}\|_2^2 \tag{5.29}$$

and

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{M}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \tag{5.30}$$

which will prove extremely useful in the sequel.

Their usefulness basically stems from the fact that, for a given $\mathbf{x}$, the expressions on the right-hand sides of (5.29) and (5.30) are quadratic functions of $\mathbf{y}$ which provide, respectively, a global underestimator and a global overestimator of f, in the set $\mathbb{S}$.

Another inequality which will be useful later is

$$p_* \geq f(\mathbf{x}) - \frac{1}{2m}\|\nabla f(\mathbf{x})\|_2^2, \tag{5.31}$$

which holds for every $\mathbf{x} \in \mathbb{S}$. The proof is based on (5.29) and is as follows. For a given $\mathbf{x}$, the right-hand side of (5.29) is a quadratic function of $\mathbf{y}$, with minimum value $f(\mathbf{x}) - \frac{1}{2m}\|\nabla f(\mathbf{x})\|_2^2$ (prove it). Hence, by (5.29), we have the inequality (why?)

$$\min_{\mathbf{y}\in\mathbb{S}} f(\mathbf{y}) \geq \min_{\mathbf{y}\in\mathbb{S}} \left(f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y}-\mathbf{x}) + \frac{m}{2}\|\mathbf{y}-\mathbf{x}\|_2^2 \right), \tag{5.32}$$

which implies (5.31).

An interesting interpretation of (5.31) is as follows. If, for some $\mathbf{x} \in \mathbb{S}$, we have that the value $\|\nabla f(\mathbf{x})\|_2$ is "small," then the value of $f(\mathbf{x})$ is "close" to the optimal value, p_*. This interpretation offers a terminating condition of the gradient descent algorithm, for example, the condition $\|\nabla f(\mathbf{x})\|_2 < \epsilon$, for a "small" positive ϵ.

Finally, using (5.29), it can be shown that, for every $\mathbf{x} \in \mathbb{S}$,

$$\|\mathbf{x} - \mathbf{x}_*\|_2 \leq \frac{2}{m}\,\|\nabla f(\mathbf{x})\|_2\,. \tag{5.33}$$

The proof is as follows. Setting $\mathbf{y} = \mathbf{x}_*$ in (5.29), we get

$$\begin{aligned} p_* = f(\mathbf{x}_*) &\geq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{x}_*-\mathbf{x}) + \frac{m}{2}\|\mathbf{x}_*-\mathbf{x}\|_2^2 \\ &\geq f(\mathbf{x}) - \|\nabla f(\mathbf{x})\|_2\|\mathbf{x}_*-\mathbf{x}\|_2 + \frac{m}{2}\|\mathbf{x}_*-\mathbf{x}\|_2^2, \end{aligned} \tag{5.34}$$

where, in the second line, we used the Cauchy-Schwarz inequality. Since $p_* \leq f(\mathbf{x})$, we have

$$\frac{m}{2}\|\mathbf{x}_*-\mathbf{x}\|_2^2 - \|\nabla f(\mathbf{x})\|_2\|\mathbf{x}_*-\mathbf{x}\|_2 \leq p_* - f(\mathbf{x}) \leq 0, \tag{5.35}$$

from which we get (5.33). This inequality states that, if for some $\mathbf{x}$ we have that the value $\|\nabla f(\mathbf{x})\|_2$ is "small," then $\mathbf{x}$ is "close" to the optimal point $\mathbf{x}_*$. Therefore, we have another argument for using the quantity $\|\nabla f(\mathbf{x})\|_2$ as a terminating criterion for the gradient descent algorithm.

During the analysis and in order to simplify notation, we will use expression $\mathbf{x}_+ = \mathbf{x} + t\Delta\mathbf{x}$ instead of the expression $\mathbf{x}_{k+1} = \mathbf{x}_k + t_k\Delta\mathbf{x}_k$.

5.5.1 Exact line-search

In this subsection, we will study the convergence speed of the gradient descent method with exact line-search.

We define $\tilde{f} : \mathbb{R} \to \mathbb{R}$, with $\tilde{f}(t) = f(\mathbf{x} - t\nabla f(\mathbf{x}))$. From (5.30), for $\mathbf{y} = \mathbf{x} - t\nabla f(\mathbf{x})$, we get

$$\tilde{f}(t) \leq f(\mathbf{x}) - t\|\nabla f(\mathbf{x})\|_2^2 + \frac{Mt^2}{2}\|\nabla f(\mathbf{x})\|_2^2. \tag{5.36}$$

If we apply exact line-search and minimize, with respect to t, both sides of (5.36), we obtain

$$\min_{t>0} \tilde{f}(t) \leq \min_{t>0} \left(f(\mathbf{x}) - t\|\nabla f(\mathbf{x})\|_2^2 + \frac{Mt^2}{2}\|\nabla f(\mathbf{x})\|_2^2 \right). \tag{5.37}$$

The optimal value of the left-hand side equals $\tilde{f}(t_{\text{exact}})$. The right-hand side is a quadratic function of t, which is minimized for $t = \frac{1}{M}$, and has minimum value equal to $f(\mathbf{x}) - \frac{1}{2M}\|\nabla f(\mathbf{x})\|_2^2$ (prove it). Therefore,

$$f(\mathbf{x}_+) := \tilde{f}(t_{\text{exact}}) \leq f(\mathbf{x}) - \frac{1}{2M}\|\nabla f(\mathbf{x})\|_2^2. \tag{5.38}$$

Subtracting p_* from both sides, we get

$$f(\mathbf{x}_+) - p_* \leq f(\mathbf{x}) - p_* - \frac{1}{2M}\|\nabla f(\mathbf{x})\|_2^2. \tag{5.39}$$

From (5.31), we get

$$\|\nabla f(\mathbf{x})\|_2^2 \geq 2m(f(\mathbf{x}) - p_*) \Longrightarrow -\frac{1}{2M}\|\nabla f(\mathbf{x})\|_2^2 \leq -\frac{m}{M}\left(f(\mathbf{x}) - p_*\right). \tag{5.40}$$

Combining (5.39) and (5.40), we get

$$f(\mathbf{x}_+) - p_* \leq \left(1 - \frac{m}{M}\right)\left(f(\mathbf{x}) - p_*\right). \tag{5.41}$$

Defining

$$c := 1 - \frac{m}{M}, \tag{5.42}$$

and applying the above inequality recursively, we obtain

$$f(\mathbf{x}_k) - p_* \leq c^k \left(f(\mathbf{x}_0) - p_*\right). \tag{5.43}$$

We observe that $0 \leq c < 1$. Therefore, (5.43) implies that $f(\mathbf{x}_k) \to p_*$, when $k \to \infty$.

In particular, in order to derive a sufficient condition ensuring that we have achieved

accuracy $\epsilon > 0$, we solve for k for which the following hold true:

$$\begin{aligned} & c^k(f(\mathbf{x}_0) - p_*) \leq \epsilon \\ \iff \quad & \frac{f(\mathbf{x}_0) - p_*}{\epsilon} \leq \frac{1}{c^k} = \left(\frac{1}{c}\right)^k \\ \iff \quad & \log\left(\frac{f(\mathbf{x}_0) - p_*}{\epsilon}\right) \leq k \log\left(\frac{1}{c}\right) \\ \iff \quad & k \geq \frac{\log\left(\frac{f(\mathbf{x}_0)-p_*}{\epsilon}\right)}{\log\left(\frac{1}{c}\right)}. \end{aligned} \tag{5.44}$$

Considering (5.43), we conclude that if

$$k \geq k_\epsilon := \frac{\log\left(\frac{f(\mathbf{x}_0)-p_*}{\epsilon}\right)}{\log\left(\frac{1}{c}\right)}, \tag{5.45}$$

then $f(\mathbf{x}_k) - p_* \leq \epsilon$.

We define the **condition number** of the problem as the quantity

$$\mathcal{K} := \frac{M}{m}. \tag{5.46}$$

For large $\mathcal{K}$, we have that[2]

$$\log\left(\frac{1}{c}\right) = -\log\left(1 - \frac{m}{M}\right) \approx \frac{m}{M} = \frac{1}{\mathcal{K}}. \tag{5.47}$$

In this case,

$$k_\epsilon \approx \mathcal{K} \log\left(\frac{f(\mathbf{x}_0) - p_*}{\epsilon}\right), \tag{5.48}$$

that is, the maximum required number of iterations to obtain an objective value within ϵ of the optimal value p_* increases *linearly* with the condition number of the problem. Moreover, it depends logarithmically on the ratio $\frac{f(\mathbf{x}_0)-p_*}{\epsilon}$.

Another interpretation of (5.48) is as follows: to decrease the target accuracy tolerance ϵ by a factor of 10, we must perform additional (at most) $O(1)\mathcal{K}$ iterations (why?).

5.5.2 Another interpretation of the gradient descent method

Suppose that we are at the point $\mathbf{x}_k$. The second-order Taylor approximation of f around $\mathbf{x}_k$ is as follows:

$$f(\mathbf{x}) \approx f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_k)^T \nabla^2 f(\mathbf{x}_k)(\mathbf{x} - \mathbf{x}_k). \tag{5.49}$$

[2]Recall that $\log(1 - x) = -x + O(x^2)$.

If, in the above relationship, instead of the Hessian $\nabla^2 f(\mathbf{x}_k)$ we put the matrix $\frac{1}{t}\mathbf{I}$, we get

$$f(\mathbf{x}) \approx f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x} - \mathbf{x}_k) + \frac{1}{2t}\|\mathbf{x} - \mathbf{x}_k\|_2^2 =: g(\mathbf{x}). \tag{5.50}$$

Function $g(\mathbf{x})$ is a quadratic convex function of $\mathbf{x}$, with derivative

$$\nabla g(\mathbf{x}) = \frac{1}{t}(\mathbf{x} - \mathbf{x}_k) + \nabla f(\mathbf{x}_k). \tag{5.51}$$

The point that minimizes the function $g(\mathbf{x})$ is

$$\nabla g(\mathbf{x}_+) = \mathbf{0} \Rightarrow \mathbf{x}_+ = \mathbf{x}_k - t\nabla f(\mathbf{x}_k). \tag{5.52}$$

Therefore, the iteration of the gradient descent method at the point $\mathbf{x}_k$ can be interpreted as the optimization of a simple quadratic approximation of f around the point $\mathbf{x}_k$. The smaller t is, the larger the contribution of the second-order term, which means that the optimal point will be near $\mathbf{x}_k$.

5.6 Newton method

Next, we describe an extremely important method for solving systems of nonlinear equations and optimization problems, the Newton method.

5.6.1 Newton step

The direction of movement

$$\Delta\mathbf{x}_{\text{Nt}} = -\left(\nabla^2 f(\mathbf{x})\right)^{-1}\nabla f(\mathbf{x}) \tag{5.53}$$

is called **Newton step**.

If $\nabla^2 f(\mathbf{x}) \succ \mathbf{O}$, then

$$\nabla f(\mathbf{x})^T \Delta\mathbf{x}_{\text{Nt}} = -\nabla f(\mathbf{x})^T\left(\nabla^2 f(\mathbf{x})\right)^{-1}\nabla f(\mathbf{x}) \leq 0, \tag{5.54}$$

with equality if, and only if, $\nabla f(\mathbf{x}) = \mathbf{0}$. Thus, if $\nabla f(\mathbf{x}) \neq \mathbf{0}$, the Newton step is a descent direction of f at the point $\mathbf{x}$.

The Newton step can be interpreted as follows.

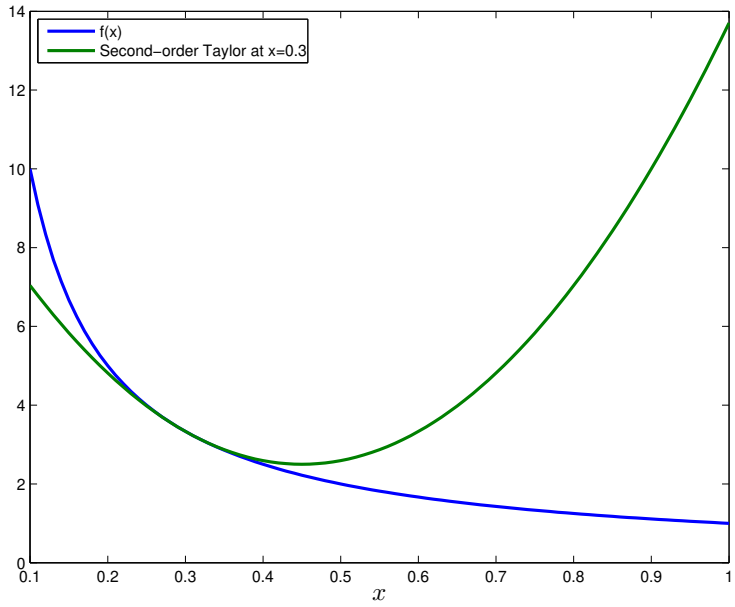


Figure 5.3: Function $f(x) = x^{-1}$ and second-order Taylor approximation at the point $x = 0.3$.

1. *Second-order approximation minimization.* The second-order Taylor approximation of f at the point $\mathbf{x}$ is given by the relation

$$\hat{f}_{\mathbf{x}}(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}). \tag{5.55}$$

   This quadratic function is minimized at $\mathbf{y} = \mathbf{x} + \Delta\mathbf{x}_{\mathrm{Nt}}$ (see Figure 5.3) (prove it).

2. *Method of steepest descent, in terms of the Hessian norm.* The Newton step is proportional to the direction of steepest descent of f at $\mathbf{x}$ for the norm

$$\|\mathbf{u}\|_{\nabla^2 f(\mathbf{x})} = \left(\mathbf{u}^T \nabla^2 f(\mathbf{x})\mathbf{u}\right)^{\frac{1}{2}}. \tag{5.56}$$

3. *Solution of a linearized optimality condition.* The linear approximation of the equation $\nabla f(\mathbf{x}_*) = \mathbf{0}$, around point $\mathbf{x}$, gives

$$\nabla f(\mathbf{x} + \mathbf{v}) \approx \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\mathbf{v} = \mathbf{0}, \tag{5.57}$$

   which gives $\mathbf{v} = -\left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x})$.

We present the Newton method in Table 5.3.

$\mathbf{x} \in \mathbb{R}^n,\ \epsilon > 0.$
While (TRUE)
1. $\Delta \mathbf{x}_{\mathrm{Nt}} := -\left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x})$.
2. $\lambda^2 := \nabla f(\mathbf{x})^T \left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x})$.
3. **quit** if $\frac{\lambda^2}{2} \leq \epsilon$.
4. Perform backtracking line search and compute t.
5. $\mathbf{x} := \mathbf{x} + t\Delta \mathbf{x}_{\mathrm{Nt}}$.

Table 5.3: Newton method.

5.6.2 Newton decrement

The quantity

$$\lambda(\mathbf{x}) := \left(\nabla f(\mathbf{x})^T \left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x})\right)^{\frac{1}{2}} \tag{5.58}$$

is called **Newton decrement**. It plays an important role in the analysis of the Newton method and can be used in the terminating condition of the algorithm.

If $\hat{f}_{\mathbf{x}}(\mathbf{y})$ is the second-order approximation of f, at the point $\mathbf{x}$, that is,

$$\hat{f}_{\mathbf{x}}(\mathbf{y}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{1}{2}(\mathbf{y} - \mathbf{x})^T \nabla^2 f(\mathbf{x})(\mathbf{y} - \mathbf{x}), \tag{5.59}$$

then

$$\begin{aligned} f(\mathbf{x}) - \inf_{\mathbf{y}} \hat{f}_{\mathbf{x}}(\mathbf{y}) &= f(\mathbf{x}) - \hat{f}_{\mathbf{x}}(\mathbf{x} + \Delta \mathbf{x}_{\mathrm{Nt}}) \\ &= -\nabla f(\mathbf{x})^T \Delta \mathbf{x}_{\mathrm{Nt}} - \frac{1}{2}\Delta \mathbf{x}_{\mathrm{Nt}}^T \nabla^2 f(\mathbf{x}) \Delta \mathbf{x}_{\mathrm{Nt}} \\ &= \nabla f(\mathbf{x})^T \left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x}) - \frac{1}{2} \nabla f(\mathbf{x})^T \left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x}) \\ &= \frac{1}{2}\lambda^2(\mathbf{x}). \end{aligned} \tag{5.60}$$

That is, the quantity $\frac{\lambda^2(\mathbf{x})}{2}$ is an estimate of the quantity $f(\mathbf{x}) - p_*$, based on the quadratic approximation of f at the point $\mathbf{x}$.

In addition, the Newton decrement appears in the backtracking line-search because

$$-\lambda^2(\mathbf{x}) = \nabla f(\mathbf{x})^T \Delta \mathbf{x}_{\mathrm{Nt}} = \left.\frac{d}{dt} f(\mathbf{x} + t\Delta \mathbf{x}_{\mathrm{Nt}})\right|_{t=0}. \tag{5.61}$$

5.6.3 Local convergence of the Newton algorithm

In this subsection, we will study the convergence of the Newton algorithm assuming it starts from a point that is close enough to the optimal point.

The result we will prove states that, locally, Newton's method has quadratic convergence.

In practice, this means that, if we start from a point that is sufficiently close to the solution, then the error is approximately squared at each iteration; equivalently, the number of correct digits roughly doubles at each iteration.

A full analysis of the convergence of the Newton algorithm is given in the Appendix of the chapter.

Technical background

Before proceeding to the analysis of the Newton algorithm, we introduce some notation which we will use next.

Let $\mathbf{g} = (g_1, \ldots, g_n) : \mathbb{R} \to \mathbb{R}^n$. We define

$$\int_a^b \mathbf{g}(t)dt := \begin{bmatrix} \int_a^b g_1(t)dt \\ \vdots \\ \int_a^b g_n(t)dt \end{bmatrix}. \tag{5.62}$$

If $\mathbf{A} : \mathbb{R} \to \mathbb{R}^{m\times n}$, then we define

$$\int_a^b \mathbf{A}(t)\mathbf{g}(t)dt := \begin{bmatrix} \int_a^b \mathbf{a}_1^T(t)\mathbf{g}(t)dt \\ \vdots \\ \int_a^b \mathbf{a}_m^T(t)\mathbf{g}(t)dt \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^n \int_a^b a_{1,i}(t)g_i(t)dt \\ \vdots \\ \sum_{i=1}^n \int_a^b a_{m,i}(t)g_i(t)dt \end{bmatrix}. \tag{5.63}$$

Furthermore, notice that we can write

$$\mathbf{c} = \int_0^1 \mathbf{c}\, dt, \quad \mathbf{Ac} = \int_0^1 \mathbf{Ac}\, dt. \tag{5.64}$$

We define the norm

$$\left\| \int_a^b \mathbf{g}(t)dt \right\|_2 = \sqrt{\sum_{i=1}^n \left(\int_a^b g_i(t)dt \right)^2}. \tag{5.65}$$

The following inequalities will be useful later

$$\left\| \int_a^b \mathbf{A}(t)\mathbf{g}(t)dt \right\|_2 \leq \int_a^b \|\mathbf{A}(t)\mathbf{g}(t)\|_2 dt \leq \int_a^b \|\mathbf{A}(t)\|_2 \|\mathbf{g}(t)\|_2 dt. \tag{5.66}$$

The left inequality is a generalization of the triangle inequality, while the right inequality results from an application of the relationship

$$\|\mathbf{A}\mathbf{g}\|_2 \leq \|\mathbf{A}\|_2\|\mathbf{g}\|_2, \quad \text{for } \mathbf{A} \in \mathbb{R}^{m\times n},\ \mathbf{g} \in \mathbb{R}^n. \tag{5.67}$$

At this point, we have all the technical tools at our disposal to proceed with the local convergence analysis of the Newton algorithm.

Theorem 5.6.1. Let $f : \mathbb{R}^n \to \mathbb{R}$ be twice continuously differentiable. Let us assume that

1. $\exists m > 0$ such that $\nabla^2 f(\mathbf{x}) \succeq m\mathbf{I}$, for every $\mathbf{x} \in \mathbb{R}^n$. Notice that this assumption leads to the inequality
$$\left(\nabla^2 f(\mathbf{x})\right)^{-1} \preceq \frac{1}{m}\mathbf{I}, \text{ for any } \mathbf{x} \in \mathbb{R}^n.$$
2. $\exists L > 0$ such that[3] $\|\nabla^2 f(\mathbf{x}) - \nabla^2 f(\mathbf{y})\|_2 \leq L\|\mathbf{x} - \mathbf{y}\|_2$, for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Let $\{\mathbf{x}_k\}$ be the sequence produced by the Newton method and $\mathbf{x}_*$ the optimal point. Then, for $k = 0, 1, \ldots$, the following inequality holds true

$$\|\mathbf{x}_{k+1} - \mathbf{x}_*\|_2 \leq \frac{L}{2m}\|\mathbf{x}_k - \mathbf{x}_*\|_2^2. \tag{5.68}$$

Moreover, if $\|\mathbf{x}_0 - \mathbf{x}_*\|_2 \leq \frac{m}{L}$, then, for $k = 0, 1, \ldots$, we have that

$$\|\mathbf{x}_k - \mathbf{x}_*\|_2 \leq \frac{2m}{L}\left(\frac{1}{2}\right)^{2^k}. \tag{5.69}$$

Proof. We study the algorithm with $t_k = 1$. From the definition of the algorithm, we have

$$\begin{aligned}
\mathbf{x}_{k+1} - \mathbf{x}_* &= \mathbf{x}_k - \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1}\nabla f(\mathbf{x}_k) - \mathbf{x}_* \\
&\overset{\nabla f(\mathbf{x}_*)=\mathbf{0}}{=} \mathbf{x}_k - \mathbf{x}_* + \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1}\left(\nabla f(\mathbf{x}_*) - \nabla f(\mathbf{x}_k)\right) \\
&\overset{(!!)}{=} \mathbf{x}_k - \mathbf{x}_* + \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1}\int_0^1 \nabla^2 f(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k))(\mathbf{x}_* - \mathbf{x}_k)dt \\
&= \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1}\int_0^1 \left[\nabla^2 f(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k)) - \nabla^2 f(\mathbf{x}_k)\right](\mathbf{x}_* - \mathbf{x}_k)dt,
\end{aligned} \tag{5.70}$$

where at point (!!) we worked as follows. We define the function $\mathbf{g} : \mathbb{R} \to \mathbb{R}^n$, with

$$\mathbf{g}(t) := \nabla f(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k)).$$

[3]The coefficient L, which can be interpreted as a bound on the third derivative of f, can be taken to be zero for a quadratic function. More generally, L measures how well f can be approximated by a quadratic model, so we can expect the Lipschitz constant L to play a critical role in the performance of Newton's method. Intuition suggests that Newton's method will work very well for a function whose quadratic model varies slowly (i.e., has small L).

If $D\mathbf{g}$ is the derivative of $\mathbf{g}$ (note that $D\mathbf{g} : \mathbb{R} \to \mathbb{R}^n$), then

$$\mathbf{g}(1) - \mathbf{g}(0) = \int_0^1 D\mathbf{g}(t)dt. \tag{5.71}$$

Using the chain rule, we get

$$\begin{aligned} D\mathbf{g}(t) &= D\nabla f(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k)) \\ &= \nabla^2 f(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k))D(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k)) \\ &= \nabla^2 f(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k))(\mathbf{x}_* - \mathbf{x}_k). \end{aligned} \tag{5.72}$$

Therefore, (5.71) is written as

$$\nabla f(\mathbf{x}_*) - \nabla f(\mathbf{x}_k) = \int_0^1 \nabla^2 f(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k))(\mathbf{x}_* - \mathbf{x}_k)dt. \tag{5.73}$$

This relationship was used at point (!!).

From (5.70), we get

$$\begin{aligned} &\|\mathbf{x}_{k+1} - \mathbf{x}_*\|_2 \\ &\quad\leq \| \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1} \|_2 \\ &\qquad\qquad \left\| \int_0^1 \left[\nabla^2 f(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k)) - \nabla^2 f(\mathbf{x}_k)\right] (\mathbf{x}_* - \mathbf{x}_k)dt \right\|_2 \\ &\quad\leq \| \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1} \|_2 \\ &\qquad\qquad \int_0^1 \left\| \left[\nabla^2 f(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k)) - \nabla^2 f(\mathbf{x}_k)\right] (\mathbf{x}_* - \mathbf{x}_k) \right\|_2 dt \\ &\quad\leq \| \left(\nabla^2 f(\mathbf{x}_k)\right)^{-1} \|_2 \\ &\qquad\qquad \int_0^1 \left\| \nabla^2 f(\mathbf{x}_k + t(\mathbf{x}_* - \mathbf{x}_k)) - \nabla^2 f(\mathbf{x}_k) \right\|_2 \|(\mathbf{x}_* - \mathbf{x}_k)\|_2 \, dt \\ &\quad\overset{(*)}{\leq} \frac{1}{m} \int_0^1 Lt\|\mathbf{x}_k - \mathbf{x}_*\|_2^2 dt \\ &\quad= \frac{L}{2m}\|\mathbf{x}_k - \mathbf{x}_*\|_2^2, \end{aligned}$$

where, at point $(*)$, we used the two assumptions. The first part of the theorem is proved. The second part is proved by induction. For $k = 0$, we assume that

$$\|\mathbf{x}_0 - \mathbf{x}_*\|_2 \leq \frac{m}{L} = \frac{2m}{L}\left(\frac{1}{2}\right)^{2^0}. \tag{5.74}$$

Let the assumption hold for some k, that is,

$$\|\mathbf{x}_k - \mathbf{x}_*\|_2 \leq \frac{2m}{L}\left(\frac{1}{2}\right)^{2^k}. \tag{5.75}$$

We will prove that it holds for $k+1$. From the first part of the theorem, we have that

$$\begin{aligned}\|\mathbf{x}_{k+1} - \mathbf{x}_*\|_2 &\leq \frac{L}{2m}\|\mathbf{x}_k - \mathbf{x}_*\|_2^2 \\ &\leq \frac{L}{2m}\left(\frac{2m}{L}\left(\frac{1}{2}\right)^{2^k}\right)^2 \\ &= \frac{2m}{L}\left(\frac{1}{2}\right)^{2^{k+1}},\end{aligned} \tag{5.76}$$

completing the proof. □

Chapter 6

Optimality Conditions

In this chapter, we derive necessary and sufficient conditions characterizing optimal solutions of convex optimization problems. First, we prove Farkas' Lemma and use it to prove the Fritz John conditions. Then, we make an additional assumption and prove the Karush-Kuhn-Tucker conditions.

Our approach, which is primarily geometric, is different from that of Boyd and Vandenberghe,[1] and is based on material from the book by Bazaraa, Sherali, Shetty and notes by M. Epelman (available online) (see also Chapters 10 and 11 of the book by A. Beck).

The optimality conditions (FJ, KKT) are extremely important because

1. they offer geometric interpretation and deeper understanding of the problem,
2. important constrained optimization algorithms search the optimal point by, essentially, searching for the point which satisfies the optimality conditions.

6.1 Necessary optimality conditions

Consider the convex optimization problem

$$\begin{array}{lll} \text{minimize} & f_0(\mathbf{x}) & \\ \text{subject to} & f_i(\mathbf{x}) \leq 0, & i = 1, \ldots, m \\ & h_i(\mathbf{x}) = 0, & i = 1, \ldots, p, \end{array} \tag{6.1}$$

with $h_i(\mathbf{x}) = \mathbf{a}_i^T \mathbf{x} - b_i$, for $i = 1, \ldots, p$.

[1] Which can be seen as complementary to the geometric approach and deserves independent study.

We assume that the functions f_i, for $i = 0, \ldots, m$, are differentiable, with $f_i : \mathbf{dom} f_i \subseteq \mathbb{R}^n \to \mathbb{R}$, with $\mathbf{dom} f_i$, for $i = 0, \ldots, m$, open convex sets.

Equality constraints can be expressed in the form

$$\mathbf{A}\mathbf{x} = \mathbf{b}, \tag{6.2}$$

with

$$\mathbf{A} := \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_p^T \end{bmatrix}, \quad \mathbf{b} := \begin{bmatrix} b_1 \\ \vdots \\ b_p \end{bmatrix}. \tag{6.3}$$

Let $\mathbb{D} := \bigcap_{i=0}^m \mathbf{dom} f_i$. The feasible set of problem (6.1) is the set

$$\mathbb{X} := \{\, \mathbf{x} \in \mathbb{D} \mid\ f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m,\ \mathbf{A}\mathbf{x} = \mathbf{b} \}\,. \tag{6.4}$$

Definition 6.1.1. Let $\mathbf{x} \in \mathbb{X}$.

1. The set $\mathbb{F}_0 := \{\mathbf{d} \in \mathbb{R}^n \mid \nabla f_0(\mathbf{x})^T \mathbf{d} < 0\}$ is called the "set of strict descent directions" of f_0, at the point $\mathbf{x}$.
2. The set $\mathbb{I} := \{i \in \{1, \ldots, m\} \mid f_i(\mathbf{x}) = 0\}$ is called set of indices of the active inequality constraints of problem (6.1), at the point $\mathbf{x}$.
3. The set $\mathbb{G}_0 := \{\mathbf{d} \in \mathbb{R}^n \mid \nabla f_i(\mathbf{x})^T \mathbf{d} < 0,\ \text{for } i \in \mathbb{I}\}$ is called the "set of strict interior directions of the active inequality constraints" of problem (6.1), at the point $\mathbf{x}$.
4. The set $\mathbb{H}_0 := \{\mathbf{d} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{d} = \mathbf{0}\}$ is called set of allowed directions of the equality constraints of problem (6.1), at the point $\mathbf{x}$.

Note: Sets $\mathbb{F}_0$ and $\mathbb{G}_0$ are not cones in the usual sense, because they do not contain the zero vector. They are, however, closed under multiplication by positive scalars.

The first important result for the characterization of optimal solutions of problem (6.1) is as follows.

Theorem 6.1.1. If $\mathbf{x} \in \mathbb{X}$ is an optimal point of the problem (6.1), then

$$\mathbb{F}_0 \cap \mathbb{G}_0 \cap \mathbb{H}_0 = \emptyset. \tag{6.5}$$

Proof. We will prove this result by reductio ad absurdum.

Let $\mathbf{x} \in \mathbb{X}$ be an optimal point for the problem (6.1) and $\mathbb{F}_0 \cap \mathbb{G}_0 \cap \mathbb{H}_0 \neq \emptyset$. Let $\mathbf{d} \in \mathbb{F}_0 \cap \mathbb{G}_0 \cap \mathbb{H}_0$ and $\theta > 0$.

Then, for all sufficiently small θ and $i \in \mathbb{I}$, we have that

$$f_i(\mathbf{x} + \theta\mathbf{d}) < f_i(\mathbf{x}) = 0 \text{ and } \mathbf{A}(\mathbf{x} + \theta\mathbf{d}) = \mathbf{b}.$$

Furthermore, since for $i \notin \mathbb{I}$ we have that $f_i(\mathbf{x}) < 0$, then, for $i \notin \mathbb{I}$ and sufficiently small θ, due to the continuity of f_i, we have that $f_i(\mathbf{x} + \theta\mathbf{d}) < 0$.

That is, for sufficiently small θ, the points $\mathbf{x} + \theta\mathbf{d}$ are feasible points of the problem (6.1).

At the same time, however, for sufficiently small θ, we have that $f_0(\mathbf{x}+\theta\mathbf{d}) < f_0(\mathbf{x})$, which is false, because we have assumed that the point $\mathbf{x}$ is optimal. □

Theorem 6.1.1 indicates that, if the point $\mathbf{x} \in \mathbb{X}$ is optimal for the problem (6.1), then there is no strictly feasible direction $\mathbf{d}$ which, at the same time, is a descent direction of f_0.

6.2 Farkas' Lemma

Next, we will translate the geometric condition (6.5) into an algebraic relationship between the gradients of the cost and constraint functions.

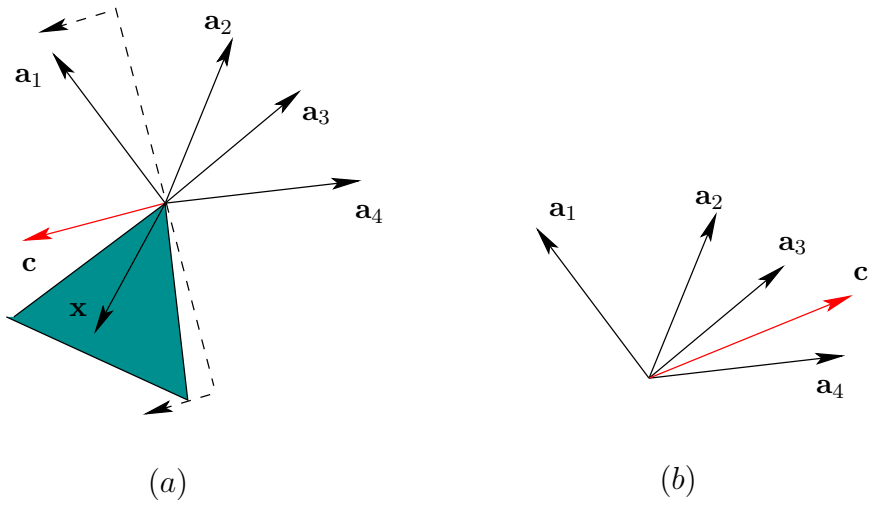
First, we prove the best-known result from a class of results that are known as **Theorems of the Alternatives**.

Lemma 6.2.1. *Farkas' Lemma.* Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{c} \in \mathbb{R}^n$. Then, exactly one of the following systems has a solution:

$$\begin{aligned} &(1) \quad \mathbf{A}\mathbf{x} \leq \mathbf{0},\ \mathbf{c}^T\mathbf{x} > 0, \\ &(2) \quad \mathbf{A}^T\mathbf{y} = \mathbf{c},\ \mathbf{y} \geq \mathbf{0}. \end{aligned} \tag{6.6}$$

Proof. Assume that system (2) has a solution, that is, there exists $\mathbf{y} \geq \mathbf{0}$ such that $\mathbf{A}^T\mathbf{y} = \mathbf{c}$.

Next, we consider the system (1) and more specifically the inequality $\mathbf{A}\mathbf{x} \leq \mathbf{0}$. The inequality is satisfied for $\mathbf{x} = \mathbf{0}$. But this solution does not satisfy the inequality $\mathbf{c}^T\mathbf{x} > 0$.



(a) (b)

Figure 6.1: Farkas' Lemma (a) the system (1) has a solution (b) the system (2) has a solution - in dark green color the cone of possible solutions.

If there is no $\mathbf{x} \neq \mathbf{0}$ satisfying the inequality $\mathbf{Ax} \leq \mathbf{0}$, then the system (1) has no solution. If there exists $\mathbf{x} \neq \mathbf{0}$ which satisfies inequality $\mathbf{Ax} \leq \mathbf{0}$, then, since $\mathbf{y} \geq \mathbf{0}$ and $\mathbf{Ax} \leq \mathbf{0}$, we get that $\mathbf{c}^T\mathbf{x} = \mathbf{y}^T\mathbf{Ax} \leq 0$. That is, in this case too, the system (1) has no solution.

Next, we assume that the system (2) has no solution.

We define the set $\mathbb{M} := \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{x} = \mathbf{A}^T\mathbf{y}, \mathbf{y} \geq \mathbf{0}\}$. We know that the set $\mathbb{M}$ is a closed convex cone, because if

$$\mathbf{A} := \begin{bmatrix} \mathbf{a}_1^T \\ \vdots \\ \mathbf{a}_m^T \end{bmatrix}, \tag{6.7}$$

then $\mathbb{M} = \{\sum_{i=1}^m y_i\mathbf{a}_i \,|\, y_i \geq 0, \text{ for } i = 1, \ldots, m\}$. That is, $\mathbb{M}$ is the conic hull of $\mathbf{a}_1, \ldots, \mathbf{a}_m$.

In addition, from the assumption about the non existence of a solution for the system (2), we know that $\mathbf{c} \notin \mathbb{M}$.

Therefore, there exists a hyperplane which separates the point $\mathbf{c}$ and the convex set $\mathbb{M}$. That is, there exist $\mathbf{p} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$ such that $\mathbf{p}^T\mathbf{c} > \alpha$ and $\mathbf{p}^T\mathbf{x} \leq \alpha$ for every $\mathbf{x} \in \mathbb{M}$. Then

1. Since $\mathbf{0} \in \mathbb{M}$, we have that $0 = \mathbf{p}^T\mathbf{0} \leq \alpha$ and $\mathbf{p}^T\mathbf{c} > \alpha \geq 0$, thus, $\mathbf{p}^T\mathbf{c} > 0$.
2. We recall that for each $\mathbf{x} \in \mathbb{M}$, we have that $\alpha \geq \mathbf{p}^T\mathbf{x}$.

Furthermore, $\mathbf{x} \in \mathbb{M}$ implies that $\mathbf{x} = \mathbf{A}^T\mathbf{y}$, with $\mathbf{y} \geq \mathbf{0}$. Therefore, for every $\mathbf{y} \geq \mathbf{0}$, we have

$$\alpha \geq \mathbf{p}^T\mathbf{x} = \mathbf{p}^T\mathbf{A}^T\mathbf{y} = \mathbf{y}^T\mathbf{A}\mathbf{p}.$$

But, since $\mathbf{y} \geq \mathbf{0}$ can become arbitrarily large, the inequality $\alpha \geq \mathbf{y}^T\mathbf{A}\mathbf{p}$, for every $\mathbf{y} \geq \mathbf{0}$, implies that $\mathbf{A}\mathbf{p} \leq \mathbf{0}$.

This happens because, if some element of $\mathbf{A}\mathbf{p}$ is greater than zero, say $(\mathbf{A}\mathbf{p})_i > 0$, then we can choose a $\mathbf{y}$ which has zeros everywhere except the i-th position where it has arbitrarily large value, y_i, so that the inequality $\alpha \geq \mathbf{y}^T\mathbf{A}\mathbf{p}$ is not satisfied.

Therefore, we constructed a vector $\mathbf{p}$ such that $\mathbf{A}\mathbf{p} \leq \mathbf{0}$ and $\mathbf{c}^T\mathbf{p} > 0$. Thus, we found a solution for the system (1), completing the proof. □

The following result is an important application of Farkas' Lemma.

Lemma 6.2.2. *(Basic Lemma)* Exactly one of the following two systems has a solution.

$$\begin{array}{ll} (1) & \mathbf{A}\mathbf{x} < \mathbf{0},\ \mathbf{B}\mathbf{x} \leq \mathbf{0},\ \mathbf{H}\mathbf{x} = \mathbf{0} \\ (2) & \mathbf{A}^T\mathbf{u} + \mathbf{B}^T\mathbf{w} + \mathbf{H}^T\mathbf{v} = \mathbf{0},\ \mathbf{u} \geq \mathbf{0},\ \mathbf{w} \geq \mathbf{0},\ \mathbf{1}^T\mathbf{u} = 1, \end{array} \tag{6.8}$$

where $\mathbf{1} := (1, \ldots, 1)$.

Proof. The system (1) is equivalently expressed as

$$\begin{array}{c} \mathbf{A}\mathbf{x} + e\mathbf{1} \leq \mathbf{0},\ e > 0 \\ \mathbf{B}\mathbf{x} \leq \mathbf{0} \\ \mathbf{H}\mathbf{x} \leq \mathbf{0} \\ -\mathbf{H}\mathbf{x} \leq \mathbf{0}. \end{array} \tag{6.9}$$

A more condensed form of the above system is as follows:

$$\begin{bmatrix} \mathbf{A} & \mathbf{1} \\ \mathbf{B} & \mathbf{0} \\ \mathbf{H} & \mathbf{0} \\ -\mathbf{H} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ e \end{bmatrix} \leq \mathbf{0}, \quad \begin{bmatrix} 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ e \end{bmatrix} > 0. \tag{6.10}$$

The system (6.10) has exactly the form of system (1) of Lemma 6.2.1.

The system corresponding to the system (2) of Lemma 6.2.1 has the form

$$\begin{bmatrix} \mathbf{A} & \mathbf{1} \\ \mathbf{B} & \mathbf{0} \\ \mathbf{H} & \mathbf{0} \\ -\mathbf{H} & \mathbf{0} \end{bmatrix}^T \begin{bmatrix} \mathbf{u} \\ \mathbf{w} \\ \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}, \qquad (\mathbf{u}, \mathbf{w}, \mathbf{v}_1, \mathbf{v}_2) \geq \mathbf{0}, \tag{6.11}$$

or, equivalently,

$$\mathbf{A}^T\mathbf{u} + \mathbf{B}^T\mathbf{w} + \mathbf{H}^T(\mathbf{v}_1 - \mathbf{v}_2) = \mathbf{0}, \quad \mathbf{1}^T\mathbf{u} = 1, \quad (\mathbf{u}, \mathbf{w}, \mathbf{v}_1, \mathbf{v}_2) \geq \mathbf{0}. \tag{6.12}$$

Setting $\mathbf{v} := \mathbf{v}_1 - \mathbf{v}_2$, completes the proof. □

Note: The relation $\mathbf{1}^T\mathbf{u} = 1$ implies that $\mathbf{u} \neq \mathbf{0}$. Therefore, in the statement of the theorem, the normalization $\mathbf{1}^T\mathbf{u} = 1$ can equivalently be replaced by $\mathbf{u} \neq \mathbf{0}$, together with $\mathbf{u} \geq \mathbf{0}$, because the homogeneous equality in system (2) can then be rescaled so that $\mathbf{1}^T\mathbf{u} = 1$.

6.3 First-order optimality conditions

Next, we use Lemma 6.2.2 and prove important algebraic optimality conditions for the optimization problem (6.1).

Theorem 6.3.1. *(Fritz John necessary conditions)* Let $\mathbf{x} \in \mathbb{X}$ be an optimal point for the problem (6.1). Then, there exists a vector $(u_0, \mathbf{u}, \mathbf{v})$ such that

$$\begin{aligned} u_0\nabla f_0(\mathbf{x}) + \sum_{i=1}^{m} u_i \nabla f_i(\mathbf{x}) + \mathbf{A}^T\mathbf{v} &= \mathbf{0}, \\ (u_0, \mathbf{u}) \geq (0, \mathbf{0}),\ (u_0, \mathbf{u}) &\neq (0, \mathbf{0}), \\ u_i f_i(\mathbf{x}) = 0,\ i &= 1, \ldots, m. \end{aligned} \tag{6.13}$$

Proof. Without loss of generality, we assume $\mathbb{I} = \{1, \ldots, l\}$ and define

$$\mathcal{A} := \begin{bmatrix} \nabla f_0(\mathbf{x})^T \\ \nabla f_1(\mathbf{x})^T \\ \vdots \\ \nabla f_l(\mathbf{x})^T \end{bmatrix}. \tag{6.14}$$

Since $\mathbf{x} \in \mathbb{X}$ is an optimal point for the problem (6.1), from Theorem 6.1.1, we conclude that $\mathbb{F}_0 \cap \mathbb{G}_0 \cap \mathbb{H}_0 = \emptyset$.

That is, there is no $\mathbf{d}$ which satisfies the relations $\mathcal{A}\mathbf{d} < \mathbf{0}$ and $\mathbf{A}\mathbf{d} = \mathbf{0}$. Therefore, from Lemma 6.2.2 for $\mathbf{B} = \mathbf{0}$, we conclude that there exist vectors $(u_0, u_1, \ldots, u_l)$ and $\mathbf{v}$ such that

$$u_0 \nabla f_0(\mathbf{x}) + \sum_{i=1}^{l} u_i \nabla f_i(\mathbf{x}) + \mathbf{A}^T \mathbf{v} = \mathbf{0}, \tag{6.15}$$

with $(u_0, u_1, \ldots, u_l) \geq \mathbf{0}$ and $(u_0, u_1, \ldots, u_l) \neq \mathbf{0}$. We set $u_{l+1} = \cdots = u_m = 0$.

Therefore, we get that $(u_0, \mathbf{u}) \geq (0, \mathbf{0})$, $(u_0, \mathbf{u}) \neq (0, \mathbf{0})$ and $u_i f_i(\mathbf{x}) = 0$, for $i = 1, \ldots, m$. The proof is complete. □

Next, we use an additional assumption and prove the necessary KKT conditions.

Theorem 6.3.2. *(KKT necessary conditions)* Let $\mathbf{x} \in \mathbb{X}$ be an optimal point of the optimization problem (6.1). Moreover, assume that the gradients of all active constraints of problem (6.1)[2] at the point $\mathbf{x}$ are linearly independent. Then, there exist vectors $\mathbf{u}$ and $\mathbf{v}$ such that

$$\begin{aligned} \nabla f_0(\mathbf{x}) + \sum_{i=1}^{m} u_i \nabla f_i(\mathbf{x}) + \mathbf{A}^T \mathbf{v} = \mathbf{0}, \\ \mathbf{u} \geq \mathbf{0}, \quad u_i f_i(\mathbf{x}) = 0, \quad \text{for } i = 1, \ldots, m. \end{aligned} \tag{6.16}$$

Proof. We have assumed that $\mathbf{x}$ is an optimal point of the problem (6.1). Therefore, it must satisfy the Fritz John conditions of (6.13).

If $u_0 > 0$, then we can normalize the Fritz John multipliers by dividing the stationarity equation by u_0: redefine $u_i := u_i/u_0$, for $i = 1, \ldots, m$, and $v_i := v_i/u_0$, for $i = 1, \ldots, p$. This gives (6.16).

If $u_0 = 0$, then there exist $\mathbf{u} \geq \mathbf{0}$, with $\mathbf{u} \neq \mathbf{0}$, and $\mathbf{v}$ such that

$$\sum_{i \in \mathbb{I}} u_i \nabla f_i(\mathbf{x}) + \sum_{i=1}^{p} v_i \mathbf{a}_i = \mathbf{0}. \tag{6.17}$$

This, however, is impossible, because we have assumed that the gradients of the active constraints at point $\mathbf{x}$ are linearly independent. The proof is complete. □

[2]The active constraints consist of the active inequality constraints together with **all** equality constraints.

Next, we list sufficient optimality conditions.

Theorem 6.3.3. *(KKT sufficient conditions)* Let the point $\mathbf{x} \in \mathbb{X}$ satisfy the Karush-Kuhn-Tucker conditions, that is, for some $\mathbf{u} \in \mathbb{R}^m$ and $\mathbf{v} \in \mathbb{R}^p$, the following relations hold:

$$\begin{aligned} \nabla f_0(\mathbf{x}) + \sum_{i=1}^{m} u_i \nabla f_i(\mathbf{x}) + \mathbf{A}^T \mathbf{v} = \mathbf{0}, \\ \mathbf{u} \geq \mathbf{0}, \quad u_i f_i(\mathbf{x}) = 0, \quad \text{for } i = 1, \ldots, m. \end{aligned} \tag{6.18}$$

Then, $\mathbf{x}$ is an optimal point for the problem (6.1).

Proof. We know that the feasible set $\mathbb{X}$ of the problem (6.1) is convex. Let $\widehat{\mathbf{x}} \in \mathbb{X}$ be a point different from $\mathbf{x}$. The points $\theta\widehat{\mathbf{x}} + (1-\theta)\mathbf{x}$, with $0 \leq \theta \leq 1$, are feasible. This means that, for every $i \in \mathbb{I}$ and $0 \leq \theta \leq 1$,

$$f_i(\theta\widehat{\mathbf{x}} + (1-\theta)\mathbf{x}) = f_i(\mathbf{x} + \theta(\widehat{\mathbf{x}} - \mathbf{x})) \leq 0 = f_i(\mathbf{x}). \tag{6.19}$$

Since the value of f_i does not increase when moving from $\mathbf{x}$ along the direction $\widehat{\mathbf{x}} - \mathbf{x}$, we have that, for each $i \in \mathbb{I}$,

$$\nabla f_i(\mathbf{x})^T (\widehat{\mathbf{x}} - \mathbf{x}) \leq 0. \tag{6.20}$$

Also, since $\widehat{\mathbf{x}}, \mathbf{x} \in \mathbb{X}$, we have that

$$\mathbf{A}(\widehat{\mathbf{x}} - \mathbf{x}) = \mathbf{0}. \tag{6.21}$$

Using (6.18), (6.20), and (6.21), we obtain that, for any $\widehat{\mathbf{x}} \in \mathbb{X}$,

$$\begin{aligned} \nabla f_0(\mathbf{x})^T (\widehat{\mathbf{x}} - \mathbf{x}) &= -\left(\sum_{i=1}^{m} u_i \nabla f_i(\mathbf{x})^T + \mathbf{v}^T \mathbf{A} \right) (\widehat{\mathbf{x}} - \mathbf{x}) \\ &= -\sum_{i=1}^{m} u_i \nabla f_i(\mathbf{x})^T (\widehat{\mathbf{x}} - \mathbf{x}) - \mathbf{v}^T \mathbf{A}(\widehat{\mathbf{x}} - \mathbf{x}) \\ &\geq 0. \end{aligned} \tag{6.22}$$

We recall that a necessary and sufficient condition for the point $\mathbf{x}_*$ to be a solution for the problem $\min_{\mathbf{x} \in \mathbb{X}} f_0(\mathbf{x})$ is as follows:

$$\nabla f_0(\mathbf{x}_*)^T (\mathbf{y} - \mathbf{x}_*) \geq 0, \ \forall \mathbf{y} \in \mathbb{X}. \tag{6.23}$$

Therefore, from (6.22), we conclude that, for any $\widehat{\mathbf{x}} \in \mathbb{X}$, $f_0(\widehat{\mathbf{x}}) \geq f_0(\mathbf{x})$. That is, $\mathbf{x}$ is an optimal point of the problem (6.1). □

6.4 Constraint qualification

In Theorem 6.3.2, we assumed that

1. the point $\mathbf{x}$ is optimal for problem (6.1),
2. "a certain requirement" is satisfied by the constraints,

and we proved the necessity of the KKT conditions.

The extra "requirement," which allows us to prove the KKT conditions, is called **constraint qualification**.

We have already proved that if the gradients of the active constraints of the problem (6.1) at the optimal point $\mathbf{x}$ are linearly independent, then the KKT conditions hold true.

However, this requirement is *local*, because it only concerns the optimal point and can be checked only if we have in advance some optimal point in mind.

An extremely useful *global* condition is the following.

Definition 6.4.1. *(Slater Condition)* Consider the optimization problem (6.1). The **Slater condition** is satisfied if there exists a feasible point $\bar{\mathbf{x}} \in \mathbb{X}$ such that $f_i(\bar{\mathbf{x}}) < 0$, for $i = 1, \ldots, m$.[3]

Theorem 6.4.1. *(Slater condition and necessity of the KKT conditions)* Let the Slater condition be satisfied for the optimization problem (6.1). Then, the KKT conditions are necessary to characterize the optimal point of the problem, that is, if $\mathbf{x} \in \mathbb{X}$ is an optimal point of the problem (6.1), then it satisfies the KKT conditions.

Proof. Let $\mathbf{x} \in \mathbb{X}$ be an optimal point of the problem (6.1). Then, the Fritz John conditions hold, i.e., there is a vector $(u_0, \mathbf{u}, \mathbf{v})$, with $(u_0, \mathbf{u}) \geq \mathbf{0}$ and $(u_0, \mathbf{u}) \neq \mathbf{0}$, such that

$$u_0 \nabla f_0(\mathbf{x}) + \sum_{i=1}^{m} u_i \nabla f_i(\mathbf{x}) + \mathbf{A}^T \mathbf{v} = \mathbf{0}, \tag{6.24}$$

with $u_i f_i(\mathbf{x}) = 0$, for $i = 1, \ldots, m$. If $u_0 \neq 0$, then we can divide by u_0 and get the KKT conditions.

Assume that $u_0 = 0$. Since we have assumed that the Slater condition is satisfied, there

[3] Obviously, since $\bar{\mathbf{x}} \in \mathbb{X}$, the relation $\mathbf{A}\bar{\mathbf{x}} = \mathbf{b}$ will also be satisfied.

exists $\bar{\mathbf{x}} \in \mathbb{X}$ such that, for each $i \in \mathbb{I}$,

$$0 = f_i(\mathbf{x}) > f_i(\bar{\mathbf{x}}). \tag{6.25}$$

From the convexity of f_i, we have that

$$f_i(\bar{\mathbf{x}}) \geq f_i(\mathbf{x}) + \nabla f_i(\mathbf{x})^T(\bar{\mathbf{x}} - \mathbf{x}). \tag{6.26}$$

From the relations (6.25) and (6.26), we conclude that, for every $i \in \mathbb{I}$, the following inequality holds

$$\nabla f_i(\mathbf{x})^T(\bar{\mathbf{x}} - \mathbf{x}) < 0. \tag{6.27}$$

Furthermore, since $\bar{\mathbf{x}}$ and $\mathbf{x}$ are feasible points, we have that $\mathbf{A}(\bar{\mathbf{x}} - \mathbf{x}) = \mathbf{0}$. From (6.24) and the assumption $u_0 = 0$, we get that one (or some) of u_i, for $i \in \mathbb{I}$, should be positive, therefore

$$\begin{aligned} 0 = \mathbf{0}^T(\bar{\mathbf{x}} - \mathbf{x}) &= \left(\sum_{i=1}^{m} u_i \nabla f_i(\mathbf{x})^T + \mathbf{v}^T \mathbf{A} \right) (\bar{\mathbf{x}} - \mathbf{x}) \\ &= \sum_{i \in \mathbb{I}} u_i \nabla f_i(\mathbf{x})^T (\bar{\mathbf{x}} - \mathbf{x}) \\ &< 0, \end{aligned} \tag{6.28}$$

which is false. Thus, if the Slater condition holds, then we must have that $u_0 > 0$, completing the proof. □

6.5 Geometric interpretation for problems with inequality constraints

In order to give a geometric interpretation to the KKT conditions, we consider the convex optimization problem with only inequality constraints. Furthermore, we assume that Slater's condition holds. In this case, the KKT conditions are sufficient and necessary optimality conditions. If $\mathbb{I} = \{1, \ldots, l\}$, then the point $\mathbf{x} \in \mathbb{X}$ is optimal if, and only if, there exists a vector $(u_1, \ldots, u_l)$ with $(u_1, \ldots, u_l) \geq \mathbf{0}$ such that

$$\sum_{i=1}^{l} u_i \nabla f_i(\mathbf{x}) = -\nabla f_0(\mathbf{x}) \tag{6.29}$$

That is, the point $\mathbf{x} \in \mathbb{X}$ is optimal for the optimization problem if, and only if, the negative gradient of f_0 at the point $\mathbf{x}$, $-\nabla f_0(\mathbf{x})$, lies in the cone generated by the gradients of the active inequality constraints at the point $\mathbf{x}$ (see Figure 6.2).

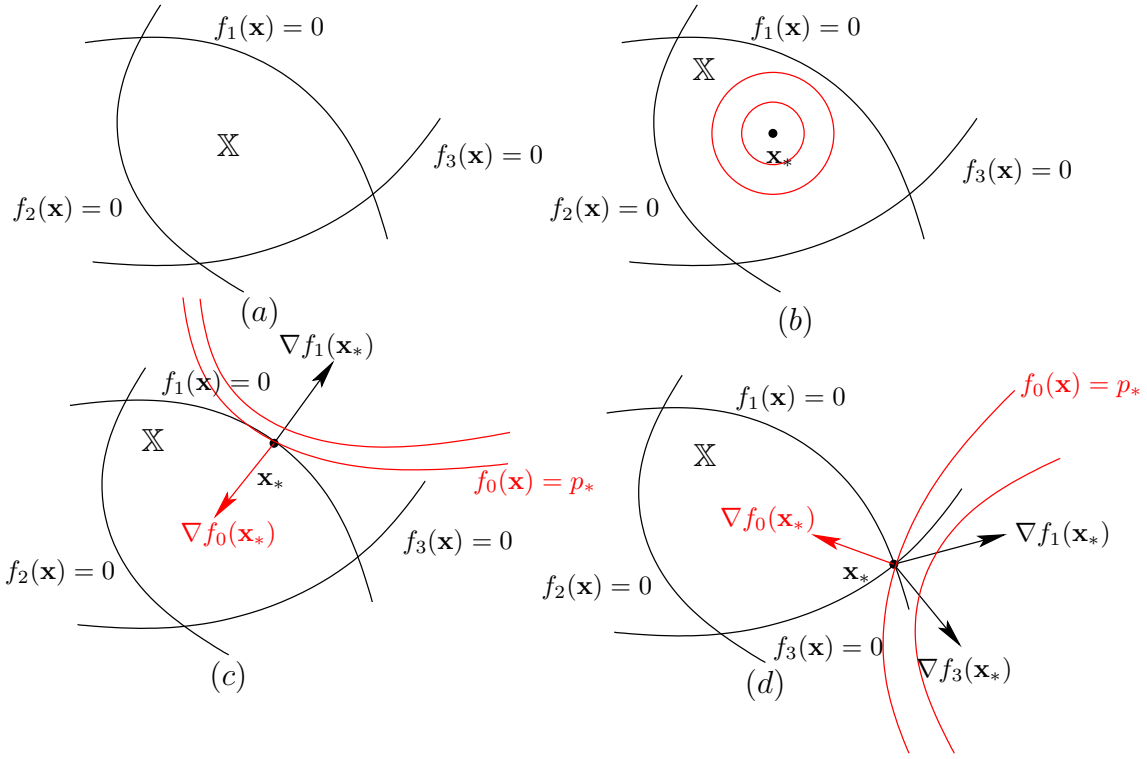


Figure 6.2: KKT conditions. (a) Feasible set of inequality constraints $\mathbb{X} = \{\mathbf{x} \in \mathbb{R}^n : f_i(\mathbf{x}) \leq 0, i = 1, 2, 3\}$. (b) Optimal point $\mathbf{x}_*$ inside the feasible set (c) Optimal point $\mathbf{x}_*$ with one active constraint (d) Optimal point $\mathbf{x}_*$ with two active constraints.

In Figure 6.3, we depict the optimization problem of the convex function $f_0(\mathbf{x})$ under the affine equality constraint $h_1(\mathbf{x}) = \mathbf{a}^T\mathbf{x} - b = 0$. Notice that $\nabla f_0(\mathbf{x}_*)$ is collinear with $\nabla h_1(\mathbf{x}_*)$ but can have the same or opposite direction (i.e., $v \in \mathbb{R}$).

6.6 Examples

First, we give an example where the KKT conditions *do not* hold.

Example 6.6.1. Consider the optimization problem

$$\begin{array}{ll} \text{minimize} & x_1 \\ \text{subject to} & (x_1 - 1)^2 + (x_2 - 1)^2 \leq 1 \\ & (x_1 - 1)^2 + (x_2 + 1)^2 \leq 1. \end{array} \tag{6.30}$$

The only feasible point is $\mathbf{x} = (1, 0)$. Prove that, at this point, the Fritz John conditions hold true but the KKT conditions do not hold true. Try to understand why this happens.

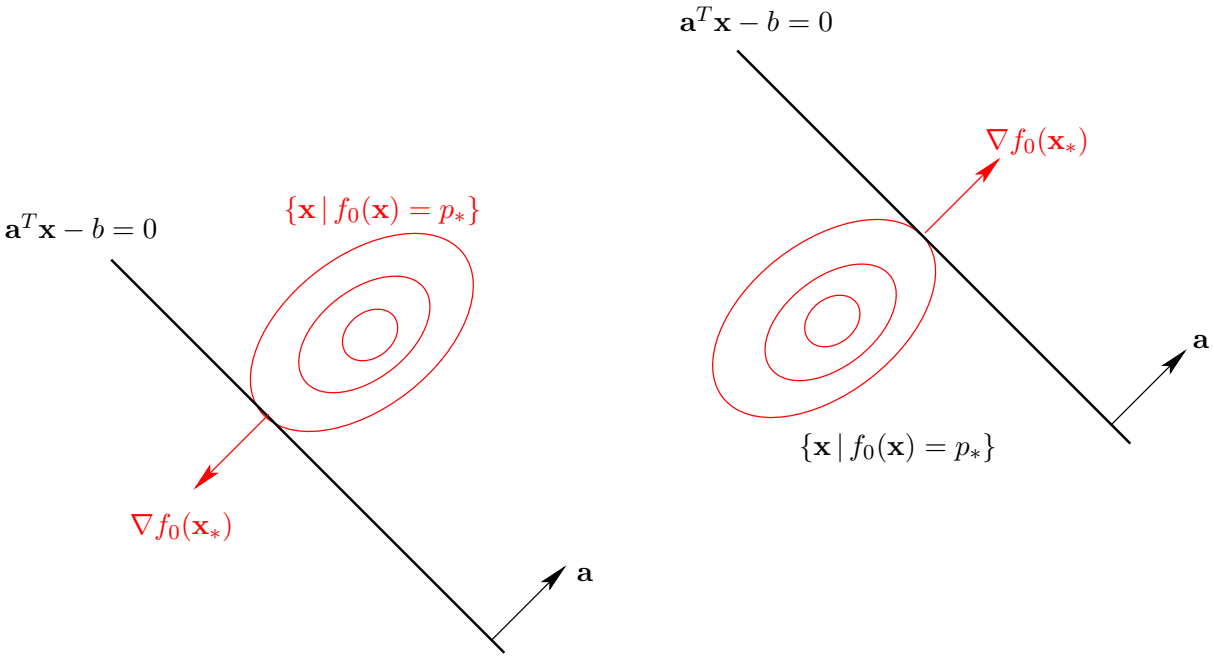


Figure 6.3: KKT conditions with a linear constraint for $v_1 > 0$ and $v_1 < 0$.

Are the gradients of the active constraints independent or does the Slater condition hold true? □

We continue with two very important examples.

Example 6.6.2. Let $f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x}$, where $\mathbf{P} = \mathbf{P}^T \succeq \mathbf{0}$, and consider the quadratic problem with affine equality constraints

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}. \end{array} \tag{6.31}$$

The KKT conditions are

$$\begin{aligned} \nabla f(\mathbf{x}) + \mathbf{A}^T\mathbf{v} &= \mathbf{0} \\ \mathbf{A}\mathbf{x} &= \mathbf{b}. \end{aligned} \tag{6.32}$$

Recall that $\nabla f(\mathbf{x}) = \mathbf{P}\mathbf{x} + \mathbf{q}$. Then, the KKT conditions can be expressed as the system of linear equations

$$\begin{bmatrix} \mathbf{P} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{x} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} -\mathbf{q} \\ \mathbf{b} \end{bmatrix}. \tag{6.33}$$

If there exists a solution of the system (6.33), say, $(\mathbf{x}_*, \mathbf{v}_*)$, then $\mathbf{x}_*$ is a primal optimal point. The multiplier $\mathbf{v}_*$ is an optimal Lagrange multiplier (and, in the duality framework developed later, a dual optimal point). □

We will say more about this very important problem later.

Example 6.6.3. For $\mathbf{x}_k \in \mathbf{dom}\, f$, with $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$, $\mathbf{A} = \mathbf{A}^T \succ 0$, and $t > 0$, compute the solution of the problem

$$\begin{array}{ll} \text{minimize} & \nabla f(\mathbf{x}_k)^T \mathbf{x} \\ \text{subject to} & (\mathbf{x} - \mathbf{x}_k)^T \mathbf{A} (\mathbf{x} - \mathbf{x}_k) \leq t^2. \end{array} \tag{6.34}$$

The problem (6.34) can be written as

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) := \nabla f(\mathbf{x}_k)^T \mathbf{x} \\ \text{subject to} & f_1(\mathbf{x}) \leq 0, \end{array} \tag{6.35}$$

with $f_1(\mathbf{x}) := (\mathbf{x} - \mathbf{x}_k)^T \mathbf{A} (\mathbf{x} - \mathbf{x}_k) - t^2$.

The KKT conditions are

$$\begin{array}{ll} (a) & \nabla f(\mathbf{x}_k) + 2\lambda\, \mathbf{A}(\mathbf{x}_* - \mathbf{x}_k) = \mathbf{0} \\ (b) & f_1(\mathbf{x}_*) = (\mathbf{x}_* - \mathbf{x}_k)^T \mathbf{A} (\mathbf{x}_* - \mathbf{x}_k) - t^2 \leq 0 \\ (c) & \lambda \geq 0,\ \lambda\, f_1(\mathbf{x}_*) = 0. \end{array} \tag{6.36}$$

From the relation (a), we find that we must have $\lambda > 0$ (why?). From this, we conclude that $f_1(\mathbf{x}_*) = 0$ (why?).

From (a), we get

$$(\mathbf{x}_* - \mathbf{x}_k) = -\frac{1}{2\lambda} \mathbf{A}^{-1} \nabla f(\mathbf{x}_k). \tag{6.37}$$

Substituting this value into (b) and noting that $f_1(\mathbf{x}_*) = 0$, we have

$$\begin{array}{ll} & \dfrac{1}{4\lambda^2} \nabla f(\mathbf{x}_k)^T \mathbf{A}^{-1} \nabla f(\mathbf{x}_k) = t^2 \\ \Longrightarrow & \lambda = \dfrac{1}{2t} \left(\nabla f(\mathbf{x}_k)^T \mathbf{A}^{-1} \nabla f(\mathbf{x}_k) \right)^{\frac{1}{2}}. \end{array} \tag{6.38}$$

Substituting this value of λ into (6.37), we obtain

$$\mathbf{x}_* = \mathbf{x}_k - \frac{t}{(\nabla f(\mathbf{x}_k)^T \mathbf{A}^{-1} \nabla f(\mathbf{x}_k))^{\frac{1}{2}}} \mathbf{A}^{-1} \nabla f(\mathbf{x}_k). \tag{6.39}$$

We observe that, indeed, $f_1(\mathbf{x}_*) = 0$.

In addition, we make the following important observations:

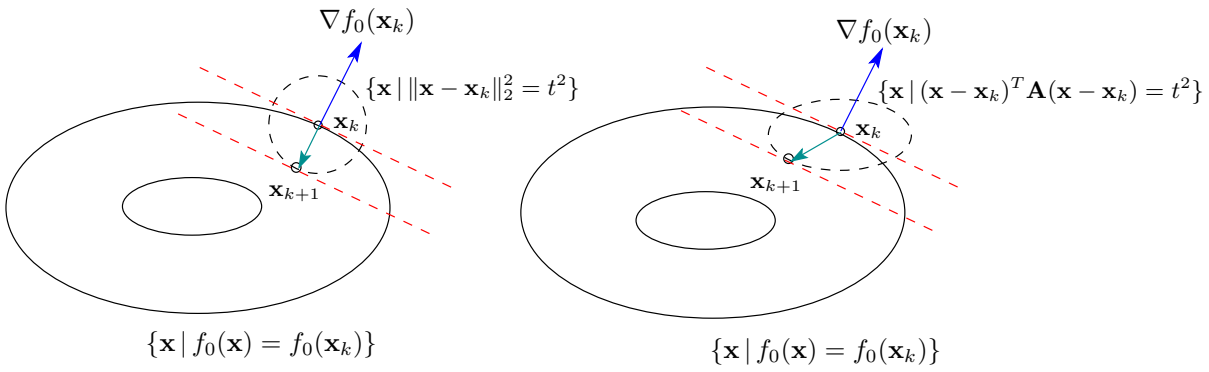


Figure 6.4: Problem solution (6.34) for $\mathbf{A} = \mathbf{I}$ and $\mathbf{A} = \nabla^2 f(\mathbf{x}_k)$.

1. Setting $\mathbf{A} = \mathbf{I}$, we obtain a step in the negative-gradient direction, normalized so that $\|\mathbf{x}_* - \mathbf{x}_k\|_2 = t$.
2. If $\nabla^2 f(\mathbf{x}_k) \succ \mathbf{0}$ and we set $\mathbf{A} = \nabla^2 f(\mathbf{x}_k)$, we obtain a step in the Newton direction, scaled to lie on the boundary of the trust region. It equals the full Newton step only for the corresponding choice of t.
3. The solution would be exactly the same if instead of $f_0(\mathbf{x}) = \nabla f(\mathbf{x}_k)^T\mathbf{x}$ we had set $f_0(\mathbf{x}) = f(\mathbf{x}_k) + \nabla f(\mathbf{x}_k)^T(\mathbf{x} - \mathbf{x}_k)$ (why?), which is the first order Taylor approximation of $f(\mathbf{x})$ around the point $\mathbf{x}_k$.

Therefore, we get a geometric interpretation of the negative-gradient and Newton directions, for unconstrained problems, as solutions of the minimization of the first-order approximation of the cost function at point $\mathbf{x}_k$, subject to a trust-region constraint defined by a suitably chosen norm in $\mathbb{R}^n$ (see Figure 6.4). □

Example 6.6.4. Let $\mathbb{S} = \{\mathbf{x} \in \mathbb{R}^2 \,|\, x_1^2 + x_2^2 \leq 1, x_1 \geq 0, x_2 \geq 0\}$. Minimize $f_0(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$, with (1) $\mathbf{c} = [1\ 1]^T$ and (2) $\mathbf{c} = [-2\ \ -1]^T$. Compute the point

$$\mathbf{x}_* = \arg\min_{\mathbf{x}\in\mathbb{S}} f_0(\mathbf{x}).$$

Point $\mathbf{x}_*$ is the solution of the convex problem

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) = \mathbf{c}^T\mathbf{x}, \\ \text{subject to} & f_1(\mathbf{x}) = x_1^2 + x_2^2 - 1 \leq 0 \\ & f_2(\mathbf{x}) = -x_1 \leq 0 \\ & f_3(\mathbf{x}) = -x_2 \leq 0. \end{array} \tag{6.40}$$

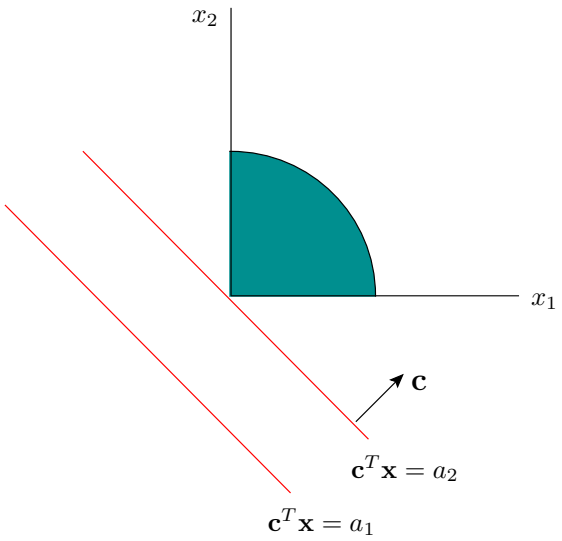


Figure 6.5: Feasible set $\mathbb{S}$ and level sets of $f_0(\mathbf{x})$, with $\mathbf{c} = [1\ 1]^T$, and $a_1 < a_2$.

The KKT conditions for this problem are as follows:

$$\begin{aligned} \nabla f_0(\mathbf{x}_*) + \lambda_1 \nabla f_1(\mathbf{x}_*) + \lambda_2 \nabla f_2(\mathbf{x}_*) + \lambda_3 \nabla f_3(\mathbf{x}_*) &= \mathbf{0} \\ \lambda_i &\geq 0, \quad i = 1, 2, 3, \\ f_i(\mathbf{x}_*) &\leq 0, \quad i = 1, 2, 3, \\ \lambda_i f_i(\mathbf{x}_*) &= 0, \quad i = 1, 2, 3, \end{aligned} \tag{6.41}$$

with

$$\nabla f_0(\mathbf{x}_*) = \mathbf{c},\ \nabla f_1(\mathbf{x}_*) = \begin{bmatrix} 2x_{*,1} \\ 2x_{*,2} \end{bmatrix},\ \nabla f_2(\mathbf{x}_*) = \begin{bmatrix} -1 \\ 0 \end{bmatrix},\ \nabla f_3(\mathbf{x}_*) = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

1. First, we consider the case where $\mathbf{c} = [1\ 1]^T$ (see Figure 6.5). Substituting into (6.41), we get

$$\begin{aligned} \begin{bmatrix} 2x_{*,1}\lambda_1 - \lambda_2 \\ 2x_{*,2}\lambda_1 - \lambda_3 \end{bmatrix} &= \begin{bmatrix} -1 \\ -1 \end{bmatrix} \\ \lambda_i &\geq 0, \quad i = 1, 2, 3, \\ f_i(\mathbf{x}_*) &\leq 0, \quad i = 1, 2, 3, \\ \lambda_i f_i(\mathbf{x}_*) &= 0, \quad i = 1, 2, 3. \end{aligned} \tag{6.42}$$

   If $\lambda_1 > 0$, then

$$x_{*,1} = \frac{-1 + \lambda_2}{2\lambda_1}, \quad x_{*,2} = \frac{-1 + \lambda_3}{2\lambda_1}. \tag{6.43}$$

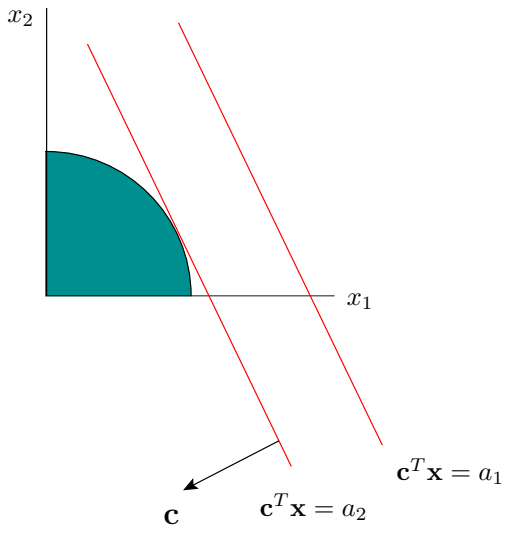


Figure 6.6: Feasible set $\mathbb{S}$ and level sets of $f_0(\mathbf{x})$, with $\mathbf{c} = [-2 \ -1]^T$, and $a_1 < a_2$.

For $\mathbf{x}_*$ to be feasible, $x_{*,1} \geq 0$ and $x_{*,2} \geq 0$ should hold true. For this to happen, we must have $\lambda_2 \geq 1$ and $\lambda_3 \geq 1$.

Since, however, it should be true that $\lambda_i f_i(\mathbf{x}_*) = 0$, for $i = 2, 3$, we should have that $x_{*,1} = x_{*,2} = 0$.

But, in this case, we have that $x_{*,1}^2 + x_{*,2}^2 - 1 < 0$ and the relation $\lambda_1 f_1(\mathbf{x}_*) = 0$ does not hold true.

Thus, we cannot have $\lambda_1 > 0$.

If $\lambda_1 = 0$, then, from the first equation of (6.42), we obtain that $\lambda_2 = 1$ and $\lambda_3 = 1$, which implies that $x_{*,1} = x_{*,2} = 0$. We observe that, if we set these values, we satisfy the KKT conditions.

Thus, the optimal solution is $\mathbf{x}_* = (0, 0)$, $\lambda_1 = 0$, $\lambda_2 = 1$, and $\lambda_3 = 1$.

2. Next, we consider the case where $\mathbf{c} = [-2 \ -1]^T$. From Figure 6.6, we conclude that only the constraint $f_1(\mathbf{x}) \leq 0$ will be active at the optimal point. Therefore, we have

$\lambda_2 = \lambda_3 = 0$, and

$$\begin{aligned} \begin{bmatrix} 2x_{*,1}\lambda_1 \\ 2x_{*,2}\lambda_1 \end{bmatrix} &= \begin{bmatrix} 2 \\ 1 \end{bmatrix} \\ \lambda_i \geq 0, \quad & i = 1,2,3, \\ f_i(\mathbf{x}_*) \leq 0, \quad & i = 1,2,3, \\ \lambda_i f_i(\mathbf{x}_*) = 0, \quad & i = 1,2,3. \end{aligned} \tag{6.44}$$

From the first equation, we get that $\lambda_1 = \frac{1}{x_{*,1}} = \frac{1}{2x_{*,2}}$. Therefore, $x_{*,1} = 2x_{*,2}$. Substituting in the relationship $f_1(\mathbf{x}_*) = 0$, we get

$$x_{*,1}^2 + x_{*,2}^2 = 1 \Rightarrow 4x_{*,2}^2 + x_{*,2}^2 = 1 \Rightarrow x_{*,2} = \frac{1}{\sqrt{5}}. \tag{6.45}$$

Thus, we get that $x_{*,1} = \frac{2}{\sqrt{5}}$, $x_{*,2} = \frac{1}{\sqrt{5}}$, $\lambda_1 = \frac{\sqrt{5}}{2}$, $\lambda_2 = \lambda_3 = 0$. □

6.7 Projection onto a convex set

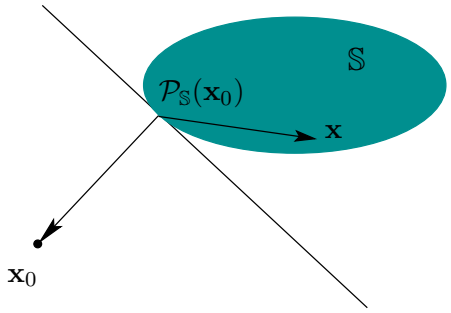


Figure 6.7: Projection of point $\mathbf{x}_0$ onto convex set $\mathbb{S}$.

Next, we turn to the very important problem of finding the projection onto a closed convex set. Let $\mathbb{S} \subset \mathbb{R}^n$ be a (non-empty) closed convex set and point $\mathbf{x}_0 \in \mathbb{R}^n$. The solution of the problem

$$\begin{aligned} \text{minimize} \quad & f_0(\mathbf{x}) := \|\mathbf{x}_0 - \mathbf{x}\|_2^2 \\ \text{subject to} \quad & \mathbf{x} \in \mathbb{S}, \end{aligned} \tag{6.46}$$

is called the **projection** of the point $\mathbf{x}_0$ onto the set $\mathbb{S}$ and is denoted by $\mathcal{P}_{\mathbb{S}}(\mathbf{x}_0)$.

We know that the point $\mathbf{x}_*$ minimizes the convex function $f(\mathbf{x})$ over the set $\mathbb{X}$ if, and only if,

$$\nabla f(\mathbf{x}_*)^T(\mathbf{x} - \mathbf{x}_*) \geq 0, \quad \forall \mathbf{x} \in \mathbb{X}. \tag{6.47}$$

Therefore, the point $\mathcal{P}_{\mathbb{S}}(\mathbf{x}_0)$ is a solution of the problem (6.46) if, and only if,

$$(\mathbf{x}_0 - \mathcal{P}_{\mathbb{S}}(\mathbf{x}_0))^T(\mathbf{x} - \mathcal{P}_{\mathbb{S}}(\mathbf{x}_0)) \leq 0, \quad \forall \mathbf{x} \in \mathbb{S}. \tag{6.48}$$

Next, we study a very important example whose solution is given in closed form.

Example 6.7.1. Find the projection of the point $\mathbf{x}_0 \in \mathbb{R}^n$ onto the convex set $\mathbb{P} := \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{x} \geq 0\}$.

This problem is written as follows:

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) := \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}\|_2^2 \\ \text{subject to} & f_i(\mathbf{x}) = -x_i \leq 0, \quad i = 1, \ldots, n. \end{array} \tag{6.49}$$

The KKT conditions are as follows:

$$\begin{aligned} \nabla f_0(\mathbf{x}_*) + \sum_{i=1}^{n} \lambda_i \nabla f_i(\mathbf{x}_*) &= \mathbf{0} \\ \lambda_i &\geq 0, \quad i = 1, \ldots, n \\ f_i(\mathbf{x}_*) &\leq 0, \quad i = 1, \ldots, n \\ \lambda_i f_i(\mathbf{x}_*) &= 0, \quad i = 1, \ldots, n. \end{aligned} \tag{6.50}$$

Defining $\boldsymbol{\lambda} := (\lambda_1, \ldots, \lambda_n)$, we write the first relation of (6.50) as

$$\begin{aligned} & (\mathbf{x}_* - \mathbf{x}_0) - \boldsymbol{\lambda} = \mathbf{0} \\ \implies \quad & x_{*,i} = x_{0,i} + \lambda_i, \quad i = 1, \ldots, n. \end{aligned} \tag{6.51}$$

Therefore, for $i = 1, \ldots, n$, we have

$$(a)\ \lambda_i \geq 0, \quad (b)\ x_{*,i} = x_{0,i} + \lambda_i \geq 0, \quad (c)\ \lambda_i x_{*,i} = \lambda_i(x_{0,i} + \lambda_i) = 0. \tag{6.52}$$

We will consider the above relations separately for each i.

1. Let $x_{0,i} \geq 0$. Then, due to (a) and (c), we have that $\lambda_i = 0$. Therefore, from (b) we get $x_{*,i} = x_{0,i}$.
2. Let $x_{0,i} < 0$. Then, due to (b), we have that $\lambda_i > 0$. Therefore, due to (c), we have that $x_{*,i} = x_{0,i} + \lambda_i = 0$.

Thus, $\mathcal{P}_{\mathbb{P}}(\mathbf{x}_0) = (\mathbf{x}_0)_+ = \max\{\mathbf{0}, \mathbf{x}_0\}$, where the operator $\max\{\cdot, \cdot\}$ is applied elementwise. $\square$

Exercise: Find the projection of the point $\mathbf{x}_0 \in \mathbb{R}^n$ onto the set

$$\mathbf{B}(\mathbf{0}, r) := \{\mathbf{x} \in \mathbb{R}^n \,|\, \|\mathbf{x}\|_2 \leq r\}.$$

Example 6.7.2. Let $\mathbf{0} \neq \mathbf{a} \in \mathbb{R}^n$ and hyperplane $\mathbb{H} := \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{a}^T\mathbf{x} = b\}$. Find the distance of a point $\mathbf{x}_0 \in \mathbb{R}^n$ from $\mathbb{H}$.

We consider the problem

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) := \frac{1}{2}\|\mathbf{x}_0 - \mathbf{x}\|_2^2 \\ \text{subject to} & f_1(\mathbf{x}) = \mathbf{a}^T\mathbf{x} - b = 0. \end{array} \tag{6.53}$$

The KKT conditions are as follows:

$$\begin{array}{l} \nabla f_0(\mathbf{x}_*) + v_1 \nabla f_1(\mathbf{x}_*) = 0 \Longrightarrow \mathbf{x}_* - \mathbf{x}_0 + v_1\mathbf{a} = 0, \\ \mathbf{a}^T\mathbf{x}_* - b = 0. \end{array} \tag{6.54}$$

Multiplying the first equation from the left by the vector $\mathbf{a}^T$, we get

$$\mathbf{a}^T\mathbf{x}_* - \mathbf{a}^T\mathbf{x}_0 + v_1\|\mathbf{a}\|_2^2 = 0 \Longrightarrow v_1 = \frac{1}{\|\mathbf{a}\|_2^2}\left(\mathbf{a}^T\mathbf{x}_0 - b\right). \tag{6.55}$$

Therefore, the projection of $\mathbf{x}_0$ onto $\mathbb{H}$ is as follows:

$$\mathbf{x}_* = \mathbf{x}_0 - \frac{\mathbf{a}^T\mathbf{x}_0 - b}{\|\mathbf{a}\|_2^2}\,\mathbf{a}, \tag{6.56}$$

and the distance of $\mathbf{x}_0$ from $\mathbb{H}$ is equal to

$$\|\mathbf{x}_* - \mathbf{x}_0\|_2 = \frac{|\mathbf{a}^T\mathbf{x}_0 - b|}{\|\mathbf{a}\|_2}. \tag{6.57}$$

Exercise: Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{b} \in \mathbb{R}^m$. Find the projection of the point $\mathbf{x}_0 \in \mathbb{R}^n$ onto the set

$$\mathbb{S} := \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{A}\mathbf{x} = \mathbf{b}\}.$$

Exercise: Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ with $\mathbf{a} \leq \mathbf{b}$. Find the projection of the point $\mathbf{x}_0 \in \mathbb{R}^n$ onto the set

$$\mathbb{S} := \{\mathbf{x} \in \mathbb{R}^n \,|\, \mathbf{a} \leq \mathbf{x} \leq \mathbf{b}\}.$$

Chapter 7

Duality

For every optimization problem there is another optimization problem which is closely related to the original. The original is called the **primal** problem while the related one is called the **Lagrange dual** problem.

Under certain conditions, for example, convexity together with an appropriate constraint qualification, the primal problem and its dual have the same optimal value and, moreover, in some cases, it is possible to compute the solution of the primal problem easily via the solution of the dual.

7.1 Primal and Dual Problem

Let us consider the minimization problem

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq 0, \quad i = 1, ..., m, \\ & h_i(\mathbf{x}) = 0, \quad i = 1, ..., p, \end{array} \tag{7.1}$$

with $\mathbf{x} \in \mathbb{R}^n$.

Let $\mathbb{D} := \bigcap_{i=0}^{m} \mathbf{dom}\, f_i \cap \bigcap_{i=1}^{p} \mathbf{dom}\, h_i$. The feasible set is given by

$$\mathbb{X} := \{\mathbf{x} \in \mathbb{D} \,|\, f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m,\ h_i(\mathbf{x}) = 0,\ i = 1, \ldots, p\}. \tag{7.2}$$

For the moment, we do not assume that the problem (7.1) is a convex optimization problem. We assume that there exists a solution to the problem, say, $\mathbf{x}_* \in \mathbb{X}$, with $f_0(\mathbf{x}_*) = p_* > -\infty$. We set $\boldsymbol{\lambda} := [\, \lambda_1 \, \cdots \, \lambda_m \,]^T$ and $\mathbf{v} := [\, v_1 \, \cdots \, v_p \,]^T$ and define the Lagrangian $L : \mathbb{D} \times \mathbb{R}^m \times$

$\mathbb{R}^p \to \mathbb{R}$, as

$$L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} v_i h_i(\mathbf{x}). \tag{7.3}$$

The vectors $\boldsymbol{\lambda}$ and $\mathbf{v}$ are called **dual variables** or **Lagrange multiplier vectors**.

First, we note that

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{v}} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) = \begin{cases} f_0(\mathbf{x}), & \text{if } \mathbf{x} \in \mathbb{X}, \\ +\infty, & \text{if } \mathbf{x} \notin \mathbb{X}. \end{cases} \tag{7.4}$$

Therefore, the problem (7.1) is equivalent to the problem

$$\min_{\mathbf{x} \in \mathbb{D}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{v}} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}). \tag{7.5}$$

The dual problem is the following:

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{v}} \min_{\mathbf{x} \in \mathbb{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}). \tag{7.6}$$

In the sequel, we study the dual problem and its relation to the primal.

7.2 The Lagrange dual function

We define the **Lagrange dual function**, or, simply, dual function, $g : \mathbb{R}^m_+ \times \mathbb{R}^p \to \mathbb{R} \cup \{-\infty\}$, as

$$\begin{aligned} g(\boldsymbol{\lambda}, \mathbf{v}) &:= \inf_{\mathbf{x} \in \mathbb{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) \\ &= \inf_{\mathbf{x} \in \mathbb{D}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{p} v_i h_i(\mathbf{x}) \right). \end{aligned} \tag{7.7}$$

When at a point $(\boldsymbol{\lambda}, \mathbf{v})$ the Lagrangian is unbounded below with respect to $\mathbf{x}$, we set $g(\boldsymbol{\lambda}, \mathbf{v}) = -\infty$. We define the domain of g as

$$\mathbf{dom}\, g = \{(\boldsymbol{\lambda}, \mathbf{v}) \in \mathbb{R}^m_+ \times \mathbb{R}^p : g(\boldsymbol{\lambda}, \mathbf{v}) > -\infty\}. \tag{7.8}$$

Theorem 7.2.1. Let g be the Lagrange dual function of (7.7). Then

1. $\mathbf{dom}\, g$ is a convex set.
2. The function g is concave.

Proof. First, we prove that the domain of g is a convex set. Let $(\boldsymbol{\lambda}_1, \mathbf{v}_1), (\boldsymbol{\lambda}_2, \mathbf{v}_2) \in \mathbf{dom} g$, that is,

$$\begin{aligned} g(\boldsymbol{\lambda}_1, \mathbf{v}_1) &= \inf_{\mathbf{x}\in\mathbb{D}} L(\mathbf{x}, \boldsymbol{\lambda}_1, \mathbf{v}_1) > -\infty, \\ g(\boldsymbol{\lambda}_2, \mathbf{v}_2) &= \inf_{\mathbf{x}\in\mathbb{D}} L(\mathbf{x}, \boldsymbol{\lambda}_2, \mathbf{v}_2) > -\infty. \end{aligned} \tag{7.9}$$

Assume that $\alpha \in [0, 1]$. Then,

$$\begin{aligned} &g(\alpha\boldsymbol{\lambda}_1 + (1-\alpha)\boldsymbol{\lambda}_2, \alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2) \\ &\quad= \inf_{\mathbf{x}\in\mathbb{D}} L(\mathbf{x}, \alpha\boldsymbol{\lambda}_1 + (1-\alpha)\boldsymbol{\lambda}_2, \alpha\mathbf{v}_1 + (1-\alpha)\mathbf{v}_2) \\ &\quad\stackrel{!}{=} \inf_{\mathbf{x}\in\mathbb{D}} (\alpha L(\mathbf{x}, \boldsymbol{\lambda}_1, \mathbf{v}_1) + (1-\alpha) L(\mathbf{x}, \boldsymbol{\lambda}_2, \mathbf{v}_2)) \\ &\quad\geq \inf_{\mathbf{x}\in\mathbb{D}} \alpha L(\mathbf{x}, \boldsymbol{\lambda}_1, \mathbf{v}_1) + \inf_{\mathbf{x}\in\mathbb{D}} (1-\alpha)\, L(\mathbf{x}, \boldsymbol{\lambda}_2, \mathbf{v}_2) \\ &\quad= \alpha \inf_{\mathbf{x}\in\mathbb{D}} L(\mathbf{x}, \boldsymbol{\lambda}_1, \mathbf{v}_1) + (1-\alpha) \inf_{\mathbf{x}\in\mathbb{D}} L(\mathbf{x}, \boldsymbol{\lambda}_2, \mathbf{v}_2) \\ &\quad= \alpha g(\boldsymbol{\lambda}_1, \mathbf{v}_1) + (1-\alpha) g(\boldsymbol{\lambda}_2, \mathbf{v}_2) \\ &\quad> -\infty. \end{aligned}$$

Thus, $\mathbf{dom}\, g$ is a convex set. During the proof, we also proved the concavity of g. □

Theorem 7.2.2. Let g be the Lagrange dual function of (7.7). Then, for every $\boldsymbol{\lambda} \geq \mathbf{0}$ and $\mathbf{v}$, we have

$$g(\boldsymbol{\lambda}, \mathbf{v}) \leq p_*. \tag{7.10}$$

Proof. Let $\bar{\mathbf{x}}$ be a feasible point of the problem (7.1), that is, $\bar{\mathbf{x}} \in \mathbb{X}$. Then

$$\sum_{i=1}^{m} \lambda_i f_i(\bar{\mathbf{x}}) + \sum_{i=1}^{p} v_i h_i(\bar{\mathbf{x}}) \leq 0,$$

because every term of the first sum is a non-positive number and every term of the second sum equals zero.

Therefore,

$$L(\bar{\mathbf{x}}, \boldsymbol{\lambda}, \mathbf{v}) = f_0(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \lambda_i f_i(\bar{\mathbf{x}}) + \sum_{i=1}^{p} v_i h_i(\bar{\mathbf{x}}) \leq f_0(\bar{\mathbf{x}}) \tag{7.11}$$

and

$$g(\boldsymbol{\lambda}, \mathbf{v}) = \inf_{\mathbf{x}\in\mathbb{D}} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) \leq L(\bar{\mathbf{x}}, \boldsymbol{\lambda}, \mathbf{v}) \leq f_0(\bar{\mathbf{x}}). \tag{7.12}$$

The above relation holds for every $\bar{\mathbf{x}} \in \mathbb{X}$, so it also holds for the optimal point $\mathbf{x}_*$. Therefore,

$$g(\boldsymbol{\lambda}, \mathbf{v}) \leq f_0(\mathbf{x}_*) = p_*, \tag{7.13}$$

proving the theorem. □

Through the dual function, we derive a non-trivial lower bound for p_* only when $\boldsymbol{\lambda} \geq 0$ and $(\boldsymbol{\lambda}, \mathbf{v}) \in \mathbf{dom}\, g$, that is, $g(\boldsymbol{\lambda}, \mathbf{v}) > -\infty$.

The pairs $(\boldsymbol{\lambda}, \mathbf{v})$ with $\boldsymbol{\lambda} \geq 0$ and $(\boldsymbol{\lambda}, \mathbf{v}) \in \mathbf{dom}\, g$ are called **dual feasible** points.

Example 7.2.1. Consider the optimization problem

$$\text{minimize } \mathbf{x}^T\mathbf{x}, \quad \text{subject to } \mathbf{A}\mathbf{x} = \mathbf{b}. \tag{7.14}$$

with $\mathbf{A} \in \mathbb{R}^{p\times n}$. The Lagrangian is given by

$$L(\mathbf{x}, \mathbf{v}) = \mathbf{x}^T\mathbf{x} + \mathbf{v}^T(\mathbf{A}\mathbf{x} - \mathbf{b}), \tag{7.15}$$

with $\mathbf{dom}\, L = \mathbb{R}^n \times \mathbb{R}^p$. The dual function is defined as

$$g(\mathbf{v}) := \inf_{\mathbf{x}\in\mathbb{R}^n} L(\mathbf{x}, \mathbf{v}). \tag{7.16}$$

$L(\mathbf{x}, \mathbf{v})$ is a *convex* quadratic function of $\mathbf{x}$. Thus, the point $\mathbf{x}$ which minimizes the Lagrangian is given by the solution of the equation

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}) = \mathbf{0} \Rightarrow 2\mathbf{x} + \mathbf{A}^T\mathbf{v} = \mathbf{0} \Rightarrow \mathbf{x} = -\frac{1}{2}\mathbf{A}^T\mathbf{v}. \tag{7.17}$$

Therefore,

$$g(\mathbf{v}) = L\left(-\frac{1}{2}\mathbf{A}^T\mathbf{v}, \mathbf{v}\right) = -\frac{1}{4}\mathbf{v}^T\mathbf{A}\mathbf{A}^T\mathbf{v} - \mathbf{b}^T\mathbf{v}, \tag{7.18}$$

which is a *concave* quadratic function of $\mathbf{v} \in \mathbb{R}^p$.

From weak duality, we have that, for every $\mathbf{v} \in \mathbb{R}^p$,

$$-\frac{1}{4}\mathbf{v}^T\mathbf{A}\mathbf{A}^T\mathbf{v} - \mathbf{b}^T\mathbf{v} \leq \inf_{\mathbf{x}\in\mathbb{R}^n} \{\mathbf{x}^T\mathbf{x} \,|\, \mathbf{A}\mathbf{x} = \mathbf{b}\}. \tag{7.19}$$

□

Example 7.2.2. *(Linear problem in standard form)* Consider the linear programming problem in standard form

$$\begin{aligned} &\text{minimize} && \mathbf{c}^T\mathbf{x} \\ &\text{subject to} && \mathbf{A}\mathbf{x} = \mathbf{b}, \\ & && \mathbf{x} \geq \mathbf{0}. \end{aligned} \tag{7.20}$$

The inequality $\mathbf{x} \geq \mathbf{0}$ is equivalent to the inequality $-\mathbf{x} \leq \mathbf{0}$, that is, $-x_i \leq 0$, for $i = 1, \ldots, n$.

The Lagrangian is as follows:

$$\begin{aligned} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) &= \mathbf{c}^T\mathbf{x} - \sum_{i=1}^{n} \lambda_i x_i + \mathbf{v}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= -\mathbf{b}^T\mathbf{v} + (\mathbf{c} + \mathbf{A}^T\mathbf{v} - \boldsymbol{\lambda})^T\mathbf{x}. \end{aligned} \tag{7.21}$$

The dual function is given by

$$g(\boldsymbol{\lambda}, \mathbf{v}) := \inf_{\mathbf{x}\in\mathbb{R}^n} L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) = -\mathbf{b}^T\mathbf{v} + \inf_{\mathbf{x}\in\mathbb{R}^n} \left\{(\mathbf{c} + \mathbf{A}^T\mathbf{v} - \boldsymbol{\lambda})^T\mathbf{x}\right\}. \tag{7.22}$$

If there is no bound on $\mathbf{x}$, then a linear function of $\mathbf{x}$ is bounded below only if it is identically zero.

Therefore,

$$g(\boldsymbol{\lambda}, \mathbf{v}) = \begin{cases} -\mathbf{b}^T\mathbf{v}, & \mathbf{c} + \mathbf{A}^T\mathbf{v} - \boldsymbol{\lambda} = \mathbf{0}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{7.23}$$

The inequality (7.13) is non-trivial only for pairs $(\boldsymbol{\lambda}, \mathbf{v}) \in \mathbb{R}^n_+ \times \mathbb{R}^p$ such that

$$\mathbf{c} + \mathbf{A}^T\mathbf{v} - \boldsymbol{\lambda} = \mathbf{0}. \tag{7.24}$$

These pairs are the solution of the system of linear equations

$$\begin{bmatrix} -\mathbf{I} & \mathbf{A}^T \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{v} \end{bmatrix} = -\mathbf{c}. \tag{7.25}$$

Therefore, they define an affine set.

For these pairs of points, the function $-\mathbf{b}^T\mathbf{v}$ provides a non-trivial lower bound for the optimal value of the problem (7.20). □

7.3 The Lagrange dual problem

We have seen that, for each pair $(\boldsymbol{\lambda}, \mathbf{v})$, with $\boldsymbol{\lambda} \geq \mathbf{0}$, the function $g(\boldsymbol{\lambda}, \mathbf{v})$ is a lower bound for the optimal value of the cost function p_*.

An extremely important question is the following: *"What is the largest lower bound that we can extract from the dual function $g(\boldsymbol{\lambda}, \mathbf{v})$?"*

To answer this question, we define the problem

$$\begin{aligned} &\max_{\boldsymbol{\lambda},\mathbf{v}} \; g(\boldsymbol{\lambda},\mathbf{v}) \\ &\text{s.t. } \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned} \tag{7.26}$$

The problem (7.1) is called the **primal** problem, while the problem (7.26) is called the **Lagrange dual** problem.

If the problem (7.26) has a solution, then an optimal point $(\boldsymbol{\lambda}_*, \mathbf{v}_*)$ is called a **dual optimal point**, and its components are called **optimal Lagrange multipliers**.

The problem (7.26) is *always* a convex optimization problem because the constraint $\boldsymbol{\lambda} \geq \mathbf{0}$ is convex and the function $g(\boldsymbol{\lambda}, \mathbf{v})$ is concave.

Example 7.3.1. *(Linear problem in standard form)* Consider the linear programming problem in standard form

$$\begin{aligned} \text{minimize} \quad & \mathbf{c}^T\mathbf{x} \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \\ & \mathbf{x} \geq \mathbf{0}. \end{aligned} \tag{7.27}$$

As we have seen, the dual function is

$$g(\boldsymbol{\lambda},\mathbf{v}) = \begin{cases} -\mathbf{b}^T\mathbf{v}, & \mathbf{c} + \mathbf{A}^T\mathbf{v} - \boldsymbol{\lambda} = \mathbf{0}, \\ -\infty, & \text{otherwise.} \end{cases} \tag{7.28}$$

The dual problem is

$$\begin{aligned} \text{maximize} \quad & g(\boldsymbol{\lambda},\mathbf{v}) \\ \text{subject to} \quad & \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned} \tag{7.29}$$

The function g takes values other than $-\infty$ only if $\mathbf{c} + \mathbf{A}^T\mathbf{v} - \boldsymbol{\lambda} = \mathbf{0}$.

Therefore, an equivalent expression for the dual problem is as follows:

$$\begin{aligned} \text{maximize} \quad & -\mathbf{b}^T\mathbf{v} \\ \text{subject to} \quad & \mathbf{c} + \mathbf{A}^T\mathbf{v} - \boldsymbol{\lambda} = \mathbf{0}, \\ & \boldsymbol{\lambda} \geq \mathbf{0}. \end{aligned} \tag{7.30}$$

or, equivalently,

$$\begin{aligned} \text{maximize} \quad & -\mathbf{b}^T\mathbf{v} \\ \text{subject to} \quad & \mathbf{c} + \mathbf{A}^T\mathbf{v} \geq \mathbf{0}, \end{aligned} \tag{7.31}$$

which is a linear programming problem in inequality form. □

You can find more examples in the book by Boyd and Vandenberghe.

7.4 Weak duality

We know that $g(\boldsymbol{\lambda}, \mathbf{v}) \leq p_*$, for every $\boldsymbol{\lambda} \geq \mathbf{0}$ and $\mathbf{v}$. If we define

$$d_* := \sup_{\boldsymbol{\lambda} \geq \mathbf{0}, \mathbf{v}} g(\boldsymbol{\lambda}, \mathbf{v}), \tag{7.32}$$

then we get that

$$d_* \leq p_*. \tag{7.33}$$

The inequality (7.33) is called **weak duality**. The relation (7.33) holds even when the values of d_* and p_* are $\pm\infty$. For example, if $p_* = -\infty$, that is, if the primal problem is unbounded below, then weak duality implies $d_* = -\infty$. The converse need not hold: $d_* = -\infty$ can also occur when the dual problem is infeasible while the primal problem is not unbounded below.

The quantity $p_* - d_*$ is called **optimal duality gap** and is always non-negative.

7.5 Strong Duality

If $d_* = p_*$, then we say that **strong duality** holds. Strong duality does not always hold. But, if the original problem is convex, then, very often, it does. For example, if the primal problem is convex and Slater's condition is satisfied, then it can be proved that strong duality holds (see Section 5.3.2 of B&V's book).

7.6 Optimality Conditions

Next, we assume that strong duality holds and derive optimality conditions for the primal problem.[1]

[1] This approach is complementary to that of the previous chapter.

7.6.1 Complementary slackness

Let $\mathbf{x}_*$ be the primal optimal point and $(\boldsymbol{\lambda}_*, \mathbf{v}_*)$, with $\boldsymbol{\lambda}_* \geq \mathbf{0}$, the dual optimal point. Then

$$\begin{aligned} f_0(\mathbf{x}_*) = g(\boldsymbol{\lambda}_*, \mathbf{v}_*) &= \inf_{\mathbf{x}\in\mathbb{D}} \underbrace{\left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_{*,i} f_i(\mathbf{x}) + \sum_{i=1}^{p} v_{*,i} h_i(\mathbf{x}) \right)}_{L(\mathbf{x}, \boldsymbol{\lambda}_*, \mathbf{v}_*)} \\ &\overset{(a)}{\leq} f_0(\mathbf{x}_*) + \sum_{i=1}^{m} \lambda_{*,i} f_i(\mathbf{x}_*) + \sum_{i=1}^{p} v_{*,i} h_i(\mathbf{x}_*) \\ &\overset{(b)}{\leq} f_0(\mathbf{x}_*), \end{aligned} \tag{7.34}$$

where

1. the inequality (a) holds because the infimum of $L(\mathbf{x}, \boldsymbol{\lambda}_*, \mathbf{v}_*)$, over all $\mathbf{x} \in \mathbb{D}$, is less than or equal to $L(\mathbf{x}_*, \boldsymbol{\lambda}_*, \mathbf{v}_*)$,
2. the inequality (b) holds because $\lambda_{*,i} f_i(\mathbf{x}_*) \leq 0$, for $i = 1, \ldots, m$, and $v_{*,i} h_i(\mathbf{x}_*) = 0$, for $i = 1, \ldots, p$.

With some thought, we realize that the last two inequalities of (7.34) hold as equalities. Therefore, it must be true that

$$\sum_{i=1}^{m} \lambda_{*,i} f_i(\mathbf{x}_*) = 0, \tag{7.35}$$

from which it follows that

$$\lambda_{*,i} f_i(\mathbf{x}_*) = 0, \quad \text{for } i = 1, ..., m. \tag{7.36}$$

The relation (7.36) is called **complementary slackness**.

7.6.2 KKT Conditions

Since the inequality (a) of (7.34) holds as equality, we conclude that $\mathbf{x}_*$ minimizes the function $L(\mathbf{x}, \boldsymbol{\lambda}_*, \mathbf{v}_*)$.

Assume that the functions f_i, for $i = 0, \ldots, m$, and h_i, for $i = 1, \ldots, p$, are differentiable on an open set containing $\mathbb{D}$ (or, more generally, on an open neighborhood of $\mathbf{x}_*$).

Note that the function $L(\mathbf{x}, \boldsymbol{\lambda}_*, \mathbf{v}_*)$ is a differentiable function of $\mathbf{x}$. Therefore, at the point $\mathbf{x}_*$ which minimizes $L(\mathbf{x}, \boldsymbol{\lambda}_*, \mathbf{v}_*)$, the derivative, with respect to $\mathbf{x}$, must be equal to zero,

that is,

$$\begin{aligned} & \nabla L_{\mathbf{x}}(\mathbf{x}_*, \boldsymbol{\lambda}_*, \mathbf{v}_*) = \mathbf{0}, \\ \Longrightarrow \quad & \nabla f_0(\mathbf{x}_*) + \sum_{i=1}^{m} \lambda_{*,i} \nabla f_i(\mathbf{x}_*) + \sum_{i=1}^{p} v_{*,i} \nabla h_i(\mathbf{x}_*) = \mathbf{0}. \end{aligned} \tag{7.37}$$

If we collect all the relations which must be satisfied by $\mathbf{x}_*$, $\boldsymbol{\lambda}_*$, and $\mathbf{v}_*$, we get

$$\begin{aligned} \nabla f_0(\mathbf{x}_*) + \sum_{i=1}^{m} \lambda_{*,i} \nabla f_i(\mathbf{x}_*) + \sum_{i=1}^{p} v_{*,i} \nabla h_i(\mathbf{x}_*) = \mathbf{0}, & \\ f_i(\mathbf{x}_*) \leq 0, \ i = 1, ..., m, & \\ h_i(\mathbf{x}_*) = 0, \ i = 1, ..., p, & \\ \lambda_{*,i} \geq 0, \ i = 1, ..., m, & \\ \lambda_{*,i} f_i(\mathbf{x}_*) = 0, \ i = 1, ..., m. & \end{aligned} \tag{7.38}$$

That is, we get the KKT conditions.

Therefore, for any optimization problem with differentiable functions for which strong duality holds, every pair of primal and dual optimal points satisfies the KKT conditions.

We will not expand further because we have already seen in the previous chapter in which cases the KKT conditions are sufficient and necessary optimality conditions of convex optimization problems (see the section 5.3.3 of the book B&V).

7.7 Usefulness of duality

The dual problem is very useful for many reasons. Below, we briefly present two of them.

7.7.1 Suboptimality Guarantee and Algorithm Termination Criteria

If $(\bar{\boldsymbol{\lambda}}, \bar{\mathbf{v}})$ is a feasible point for the dual problem and $\bar{\mathbf{x}}$ is a feasible point for the primal problem, then

$$g(\bar{\boldsymbol{\lambda}}, \bar{\mathbf{v}}) \leq p_* \leq f_0(\bar{\mathbf{x}}). \tag{7.39}$$

Let us assume that

$$f_0(\bar{\mathbf{x}}) - g(\bar{\boldsymbol{\lambda}}, \bar{\mathbf{v}}) = \bar{\epsilon} > 0. \tag{7.40}$$

Then, from the left inequality of (7.39), we obtain that

$$-p_* \leq -g(\bar{\boldsymbol{\lambda}}, \bar{\mathbf{v}}) \tag{7.41}$$

and, from (7.40), we obtain that

$$f_0(\bar{\mathbf{x}}) - p_* \leq f_0(\bar{\mathbf{x}}) - g(\bar{\boldsymbol{\lambda}}, \bar{\mathbf{v}}) = \bar{\epsilon}. \tag{7.42}$$

That is, if we know a feasible primal point $\bar{\mathbf{x}}$ and a feasible dual point $(\bar{\boldsymbol{\lambda}}, \bar{\mathbf{v}})$ such that the relation (7.40) holds, then we can conclude that $\bar{\mathbf{x}}$ is (at most) $\bar{\epsilon}$–suboptimal. This conclusion can be used as a termination criterion for some optimization algorithms.

Other cases for which this conclusion is extremely useful are cases where the primal problem is extremely difficult to solve; therefore, we may be satisfied with some "good" suboptimal solutions.

7.7.2 Solving a primal problem through the dual

In some cases (1) the solution of the dual problem is much easier than the solution of the primal problem and (2) we can (easily) solve the primal problem using the solution of the dual.

Example 7.7.1. We consider the problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) = \mathbf{x}^T\mathbf{x} \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{array} \tag{7.43}$$

with $\mathbf{A} \in \mathbb{R}^{p\times n}$ and $\mathbf{rank}(\mathbf{A}) = p$.

As we have seen, the dual function is given by the relation

$$g(\mathbf{v}) = -\frac{1}{4}\mathbf{v}^T\mathbf{A}\mathbf{A}^T\mathbf{v} - \mathbf{b}^T\mathbf{v}. \tag{7.44}$$

The dual problem is the *unconstrained convex* quadratic problem

$$\text{minimize} \quad -g(\mathbf{v}) = \tfrac{1}{4}\mathbf{v}^T\mathbf{A}\mathbf{A}^T\mathbf{v} + \mathbf{b}^T\mathbf{v}. \tag{7.45}$$

Since we have assumed that the rows of $\mathbf{A}$ are linearly independent, the $(p \times p)$ matrix $\mathbf{A}\mathbf{A}^T$ is invertible. The solution of the dual problem is equal to

$$\mathbf{v}_* = -2\,(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}. \tag{7.46}$$

The KKT conditions are necessary and sufficient for this convex problem with affine equality constraints (why?). Thus,

$$\begin{array}{ll} & \nabla f(\mathbf{x}_*) + \mathbf{A}^T\mathbf{v}_* = \mathbf{0} \\ \Longrightarrow & 2\mathbf{x}_* - 2\mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b} = \mathbf{0}. \end{array} \tag{7.47}$$

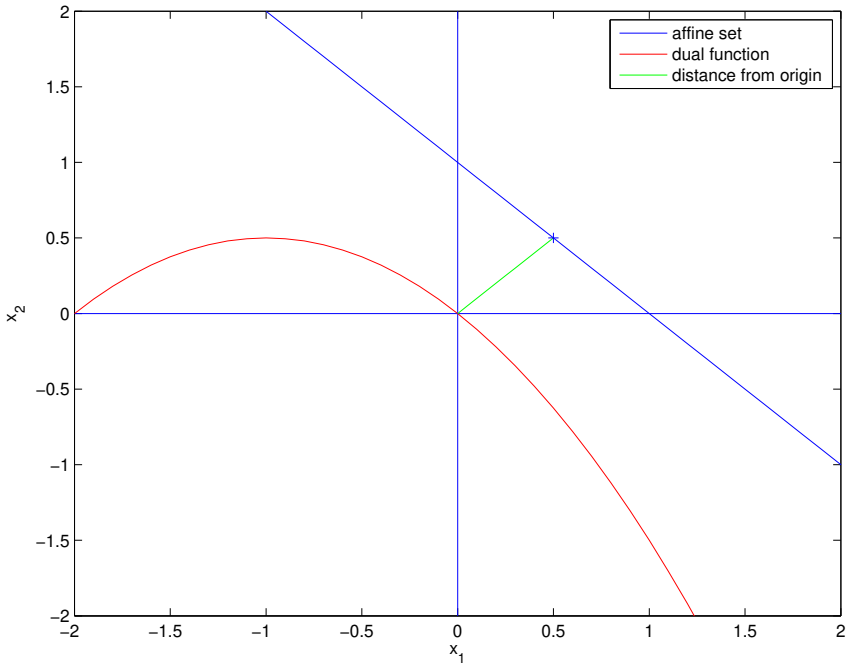


Figure 7.1: Affine set, dual function and affine set point of minimum Euclidean norm.

Therefore, the optimal primal solution is given by the relation

$$\mathbf{x}_* = \mathbf{A}^T(\mathbf{A}\mathbf{A}^T)^{-1}\mathbf{b}. \tag{7.48}$$

Observe that $\mathbf{A}\mathbf{x}_* = \mathbf{b}$. That is, we solved the primal problem *with affine constraints* by using the solution of the *unconstrained* dual problem.

In Figure 7.1, we draw

1. the affine set $\mathbf{a}^T\mathbf{x} = b$ in $\mathbb{R}^2$, for $\mathbf{a}^T = [\,1\ 1\,]$ and $b = 1$,
2. the optimal point $\mathbf{x}_* = (0.5, 0.5)$, with $f(\mathbf{x}_*) = \mathbf{x}_*^T\mathbf{x}_* =0.5$,
3. the dual function $g(v)$, with maximum value $g(v_*) = g(-1) = 0.5$.

□

Chapter 8

Convex optimization with affine equality constraints

In this chapter, we will study problems of the form

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{array} \tag{8.1}$$

with $f : \mathbf{dom}\, f \subseteq \mathbb{R}^n \to \mathbb{R}$ being a convex twice differentiable function, $\mathbf{A} \in \mathbb{R}^{p\times n}$, with $\mathbf{rank}(\mathbf{A}) = p$, and $\mathbf{b} \in \mathbb{R}^p$.

We assume that the problem has a solution, say, $\mathbf{x}_*$, and define $p_* := f(\mathbf{x}_*) = \inf\{f(\mathbf{x}) \mid \mathbf{A}\mathbf{x} = \mathbf{b}\}$. We know that the point $\mathbf{x}_*$ is optimal if, and only if, the KKT conditions are satisfied, which, in this case, are as follows:

$$\nabla f(\mathbf{x}_*) + \mathbf{A}^T\mathbf{v}_* = \mathbf{0}, \quad \mathbf{A}\mathbf{x}_* = \mathbf{b}, \tag{8.2}$$

with $\mathbf{v}_* \in \mathbb{R}^p$. The relations (8.2) constitute a system of $(n + p)$ (usually, nonlinear) equations.

The linear system of equations $\mathbf{A}\mathbf{x}_* = \mathbf{b}$ is called the system of **primal feasibility**. The system of equations $\nabla f(\mathbf{x}_*) + \mathbf{A}^T\mathbf{v}_* = \mathbf{0}$ constitutes the **stationarity** condition and is, in general, a system of nonlinear equations.

There are several approaches towards the solution of problem (8.2). We mention the following.

1. We can remove the equality constraints and then work on problems without constraints, as in the previous chapter.

2. If the dual function is differentiable, then we can solve the dual problem, which is an unconstrained problem, and then, somehow, compute the primal optimal solution.
3. We can extend the Newton method from unconstrained problems to problems with affine equality constraints. This will be the basic approach we will follow in the sequel.

8.1 Solving a primal problem via the dual

We assume that strong duality holds. The dual function for problem (8.1) is defined as:

$$\begin{aligned} g(\mathbf{v}) &= \inf_{\mathbf{x}\in\mathbf{dom}\, f} \left\{ f(\mathbf{x}) + \mathbf{v}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \right\} \\ &= -\mathbf{b}^T\mathbf{v} + \inf_{\mathbf{x}\in\mathbf{dom}\, f} \left\{ f(\mathbf{x}) + \mathbf{v}^T\mathbf{A}\mathbf{x} \right\}. \end{aligned} \tag{8.3}$$

If the function g is twice differentiable on its domain, then we can use an appropriate algorithm for smooth optimization to solve the dual problem

$$\max_{\mathbf{v}} g(\mathbf{v}). \tag{8.4}$$

Then, the optimal dual variable, $\mathbf{v}_*$, can be used to compute the optimal primal variable, $\mathbf{x}_*$.

Example 8.1.1. *(Analytic center with equality constraints)* Consider the problem

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}) = -\sum_{i=1}^n \log x_i \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{array} \tag{8.5}$$

with $\mathbf{dom}\, f = \mathbb{R}^n_{++}$ and $\mathbf{A} \in \mathbb{R}^{p\times n}$ with $\mathbf{rank}(\mathbf{A}) = p$.

The Lagrangian is as follows:

$$L(\mathbf{x}, \mathbf{v}) = f(\mathbf{x}) + \mathbf{v}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = -\sum_{i=1}^n \log x_i + \mathbf{v}^T(\mathbf{A}\mathbf{x} - \mathbf{b}). \tag{8.6}$$

For a given $\mathbf{v}$, $L(\mathbf{x}, \mathbf{v})$ is a convex function of $\mathbf{x}$ (prove it).

The derivative of L, with respect to $\mathbf{x}$, is as follows

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}) = -\begin{bmatrix} x_1^{-1} & \cdots & x_n^{-1} \end{bmatrix}^T + \mathbf{A}^T\mathbf{v}. \tag{8.7}$$

Solving the equation $\nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}) = \mathbf{0}$, we obtain

$$x_i = \frac{1}{(\mathbf{A}^T\mathbf{v})_i}, \quad \text{for} \quad i = 1, \ldots, n. \tag{8.8}$$

The dual function is as follows

$$g(\mathbf{v}) = \inf_{\mathbf{x}} L(\mathbf{x}, \mathbf{v}) = L(1./(\mathbf{A}^T\mathbf{v}), \mathbf{v}) = \sum_{i=1}^{n} \log\left(\mathbf{A}^T\mathbf{v}\right)_i + n - \mathbf{b}^T\mathbf{v}. \tag{8.9}$$

The dual problem is

$$\max_{\mathbf{v}} g(\mathbf{v}) = -\mathbf{b}^T\mathbf{v} + n + \sum_{i=1}^{n} \log\left(\mathbf{A}^T\mathbf{v}\right)_i, \tag{8.10}$$

whose domain imposes the *implicit constraint* $\mathbf{A}^T\mathbf{v} > \mathbf{0}$. Suppose that the dual problem has a solution $\mathbf{v}_*$. Then, from (8.8), we get that

$$x_{*,i} = \frac{1}{(\mathbf{A}^T\mathbf{v}_*)_i}, \quad \text{for} \quad i = 1, \ldots, n. \tag{8.11}$$

Thus, if we solve the dual problem, then we can very easily solve the primal problem. ■

8.2 Convex quadratic problems with affine equality constraints

Before proceeding to the extension of the Newton method for the solution of convex optimization problems with affine equality constraints, we will study a simple but extremely important problem, which has a closed form solution.

The solution of this problem will be used as a building block for the development of algorithms for convex optimization problems with affine equality constraints.

Consider the quadratic problem with affine equality constraints

$$\begin{array}{ll} \min\limits_{\mathbf{x}} & f(\mathbf{x}) = \frac{1}{2}\,\mathbf{x}^T\mathbf{P}\mathbf{x} + \mathbf{q}^T\mathbf{x} + r \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{array} \tag{8.12}$$

with $\mathbf{P} = \mathbf{P}^T \in \mathbb{R}^{n\times n}$, $\mathbf{P} \succeq \mathbf{O}$, $\mathbf{q} \in \mathbb{R}^n$, and $r \in \mathbb{R}$.

The KKT conditions for this problem are

$$\mathbf{P}\mathbf{x}_* + \mathbf{q} + \mathbf{A}^T\mathbf{v}_* = \mathbf{0}, \quad \mathbf{A}\mathbf{x}_* = \mathbf{b}, \tag{8.13}$$

or, equivalently,

$$\begin{bmatrix} \mathbf{P} & \mathbf{A}^T \\ \mathbf{A} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \mathbf{x}_* \\ \mathbf{v}_* \end{bmatrix} = \begin{bmatrix} -\mathbf{q} \\ \mathbf{b} \end{bmatrix}, \tag{8.14}$$

that is, they are a system of $(n + p)$ linear equations.

If the coefficient matrix is invertible, then there exists a *unique* solution for the primal–dual optimal pair $(\mathbf{x}_*, \mathbf{v}_*)$. It can be shown that, if $\mathbf{P} \succ \mathbf{O}$, then the coefficient matrix is invertible. (For more details, see B&V, pages 522, 523).

8.3 Newton's method starting from a feasible point

In the sequel, we extend the Newton method from unconstrained problems to problems with affine equality constraints.

The approaches are quite similar. However, there are the following differences:

1. We must start from a feasible point, say, $\mathbf{x}_0$, that is, we must have $\mathbf{x}_0 \in \mathbf{dom}\, f$ and $\mathbf{A}\mathbf{x}_0 = \mathbf{b}$. Often, finding an initial feasible point is difficult and, perhaps, requires the solution of an optimization problem, in the form of **feasibility check**.
2. The definition of the Newton step should be done so that the equality constraints are taken into account, that is, we should have that $\mathbf{A}\Delta\mathbf{x}_{\mathrm{Nt}} = \mathbf{0}$. As we shall see below, this is relatively simple to attain.

8.4 Newton step

As in the case of unconstrained optimization problems, the Newton step can be derived through various approaches.

8.4.1 Definition via the minimization of the second-order Taylor approximation

The first approach for the computation of the Newton step for the problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{array} \tag{8.15}$$

at a feasible point $\mathbf{x}$, that is, $\mathbf{x} \in \mathbf{dom}\, f$ and $\mathbf{A}\mathbf{x} = \mathbf{b}$, is as follows.

We replace the cost function f by its second-order Taylor approximation, at the point $\mathbf{x}$, constructing the quadratic problem, with variable $\mathbf{z}$,

$$\begin{array}{ll} \min\limits_{\mathbf{z}} & \widehat{f}(\mathbf{x}+\mathbf{z}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^T\mathbf{z} + \frac{1}{2}\,\mathbf{z}^T\nabla^2 f(\mathbf{x})\mathbf{z} \\ \text{subject to} & \mathbf{A}(\mathbf{x}+\mathbf{z}) = \mathbf{b}. \end{array} \tag{8.16}$$

We define as Newton step for the problem (8.15), at the feasible point $\mathbf{x}$, the solution of the problem (8.16), which we call $\Delta\mathbf{x}_{\mathrm{Nt}}$.[1] Essentially, the Newton step, $\Delta\mathbf{x}_{\mathrm{Nt}}$, is the vector that must be added to $\mathbf{x}$ so that the second-order Taylor approximation of f, at $\mathbf{x}$, is minimized at the point $\mathbf{x} + \Delta\mathbf{x}_{\mathrm{Nt}}$.

The KKT conditions for the quadratic problem (8.16) are as follows

$$\left[\begin{array}{cc} \nabla^2 f(\mathbf{x}) & \mathbf{A}^T \\ \mathbf{A} & \mathbf{O} \end{array}\right] \left[\begin{array}{c} \Delta\mathbf{x}_{\mathrm{Nt}} \\ \mathbf{w} \end{array}\right] = \left[\begin{array}{c} -\nabla f(\mathbf{x}) \\ \mathbf{0} \end{array}\right], \tag{8.17}$$

where $\mathbf{w}$ is the optimal dual variable for the given quadratic problem. If $\nabla^2 f(\mathbf{x}) \succ \mathbf{O}$, then it can be proved that the system (8.17) has a unique solution.

As for unconstrained optimization problems, we observe that, if f is quadratic and a full Newton step is taken, then the Newton method reaches the optimal solution in one step (under the stated nonsingularity assumptions). Therefore, we expect that the Newton algorithm will behave (very) well when f is (very) close to a quadratic function.

8.4.2 Definition via the solution of linearized optimality conditions

Another interpretation of the Newton step, at the point $\mathbf{x}$, is as follows. We have seen that the optimality conditions for the problem (8.1) are given by the relations

$$\nabla f(\mathbf{x}_*) + \mathbf{A}^T\mathbf{v}_* = \mathbf{0}, \quad \mathbf{A}\mathbf{x}_* = \mathbf{b}. \tag{8.18}$$

[1] We assume that there is a solution.

Instead of the system (8.18), which, usually, is nonlinear, we solve the following linear approximation.

We replace $\mathbf{x}_*$ with $\mathbf{x} + \Delta\mathbf{x}_{\rm Nt}$ and $\mathbf{v}_*$ with $\mathbf{w}$. If, moreover, we replace the gradient $\nabla f(\mathbf{x} + \Delta\mathbf{x}_{\rm Nt})$ by its first-order Taylor approximation, at the point $\mathbf{x}$, then we get

$$\nabla f(\mathbf{x}_*) = \nabla f(\mathbf{x} + \Delta\mathbf{x}_{\rm Nt}) \approx \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\,\Delta\mathbf{x}_{\rm Nt}, \tag{8.19}$$

and an approximation of (8.18) is as follows:

$$\begin{aligned} \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\,\Delta\mathbf{x}_{\rm Nt} + \mathbf{A}^T\mathbf{w} &= \mathbf{0} \\ \mathbf{A}(\mathbf{x} + \Delta\mathbf{x}_{\rm Nt}) &= \mathbf{b}. \end{aligned} \tag{8.20}$$

Considering that $\mathbf{A}\mathbf{x} = \mathbf{b}$, the above relations are expressed as

$$\nabla^2 f(\mathbf{x})\Delta\mathbf{x}_{\rm Nt} + \mathbf{A}^T\mathbf{w} = -\nabla f(\mathbf{x}), \quad \mathbf{A}\Delta\mathbf{x}_{\rm Nt} = \mathbf{0}, \tag{8.21}$$

which are the same expressions as those in (8.17).

8.4.3 Analytic expression for the Newton step

Assuming $\nabla^2 f(\mathbf{x}) \succ \mathbf{O}$, the analytic expressions for $\Delta\mathbf{x}_{\rm Nt}$ and $\mathbf{w}$ are as follows:

$$\begin{aligned} \Delta\mathbf{x}_{\rm Nt} = &- \left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x}) \\ &+ \left(\nabla^2 f(\mathbf{x})\right)^{-1} \mathbf{A}^T \left(\mathbf{A}\left(\nabla^2 f(\mathbf{x})\right)^{-1} \mathbf{A}^T\right)^{-1} \mathbf{A}\left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x}), \end{aligned} \tag{8.22}$$

$$\mathbf{w} = -\left(\mathbf{A}\left(\nabla^2 f(\mathbf{x})\right)^{-1} \mathbf{A}^T\right)^{-1} \mathbf{A}\left(\nabla^2 f(\mathbf{x})\right)^{-1} \nabla f(\mathbf{x}). \tag{8.23}$$

Notice that these two quantities satisfy (8.21) and, in particular,

$$\Delta\mathbf{x}_{\rm Nt} = -\left(\nabla^2 f(\mathbf{x})\right)^{-1} \left(\nabla f(\mathbf{x}) + \mathbf{A}^T\mathbf{w}\right). \tag{8.24}$$

8.4.4 Newton decrement

As in unconstrained optimization problems, we define the Newton decrement, at the feasible point $\mathbf{x}$, as follows:

$$\lambda(\mathbf{x}) := \left(\Delta\mathbf{x}_{\rm Nt}^T \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}_{\rm Nt}\right)^{\frac{1}{2}}. \tag{8.25}$$

The quantity $\lambda(\mathbf{x})$ provides an approximation of the difference $f(\mathbf{x}) - p_*$, based on the second-order Taylor approximation of f, at the point $\mathbf{x}$. More specifically,

$$f(\mathbf{x}) - \inf_{\mathbf{z}} \left\{ \widehat{f}(\mathbf{x} + \mathbf{z}) \mid \mathbf{A}\mathbf{z} = \mathbf{0} \right\} = \frac{\lambda^2(\mathbf{x})}{2}. \tag{8.26}$$

The proof is as follows:

$$\begin{aligned}\inf_{\mathbf{z}} \left\{ \widehat{f}(\mathbf{x}+\mathbf{z}) \,|\, \mathbf{A}\mathbf{z} = \mathbf{0} \right\} &= \widehat{f}(\mathbf{x}+\Delta\mathbf{x}_{\mathrm{Nt}}) \\ &= f(\mathbf{x}) + \nabla f(\mathbf{x})^T \Delta\mathbf{x}_{\mathrm{Nt}} + \frac{1}{2}\,\Delta\mathbf{x}_{\mathrm{Nt}}^T \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}_{\mathrm{Nt}}.\end{aligned} \tag{8.27}$$

From (8.21), we get that

$$\begin{aligned}\nabla f(\mathbf{x})^T \Delta\mathbf{x}_{\mathrm{Nt}} &= -\left(\Delta\mathbf{x}_{\mathrm{Nt}}^T \nabla^2 f(\mathbf{x}) + \mathbf{w}^T \mathbf{A}\right) \Delta\mathbf{x}_{\mathrm{Nt}} \\ &= -\Delta\mathbf{x}_{\mathrm{Nt}}^T \nabla^2 f(\mathbf{x}) \Delta\mathbf{x}_{\mathrm{Nt}} \\ &= -\lambda^2(\mathbf{x}).\end{aligned} \tag{8.28}$$

Combining (8.27), (8.28), and (8.25), we get (8.26).

Finally, the Newton decrement appears in the line search because

$$\left.\frac{d}{dt} f(\mathbf{x} + t\Delta\mathbf{x}_{\mathrm{Nt}})\right|_{t=0} = \nabla f(\mathbf{x})^T \Delta\mathbf{x}_{\mathrm{Nt}} = -\lambda^2(\mathbf{x}). \tag{8.29}$$

8.4.5 Feasible descent direction

Suppose we are at the feasible point $\mathbf{x}$. Then, $\mathbf{z} \in \mathbb{R}^n$ is a feasible direction of movement if $\mathbf{A}\mathbf{z} = \mathbf{0}$, and descent direction if, for sufficiently small t, $f(\mathbf{x} + t\mathbf{z}) < f(\mathbf{x})$.

The Newton step $\Delta\mathbf{x}_{\mathrm{Nt}}$ is always a feasible direction and, whenever $\lambda(\mathbf{x}) > 0$, it is a descent direction, because

1. from the second system of equations of (8.17), we have that $\mathbf{A}\Delta\mathbf{x}_{\mathrm{Nt}} = \mathbf{0}$,
2. due to (8.29), if $\lambda(\mathbf{x}) > 0$, then the directional derivative is strictly negative and, for sufficiently small $t > 0$, $f(\mathbf{x} + t\Delta\mathbf{x}_{\mathrm{Nt}}) < f(\mathbf{x})$.

8.4.6 Newton's algorithm starting from a feasible point

The Newton algorithm starting from a feasible point is presented in Table 8.1.

As we see in Fig. 8.1, if we start from a feasible point, $\mathbf{x}_0$, then, for $k \geq 0$, the points $\mathbf{x}_k$ remain feasible. With a line search that accepts a positive step along a nonzero Newton descent direction, the objective decreases at each nonterminal iteration.

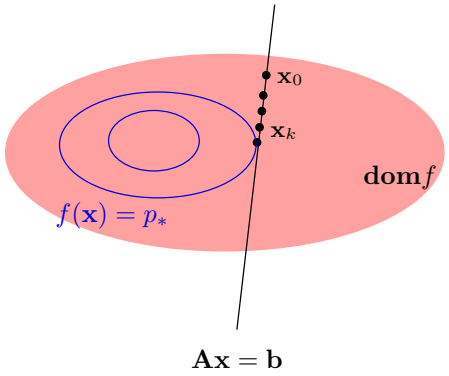


Figure 8.1: Trajectory $\mathbf{x}_0, \ldots, \mathbf{x}_k$ of the Newton algorithm, starting from a feasible point.

$\mathbf{x} \in \mathbf{dom}\, f$, $\mathbf{Ax} = \mathbf{b}$, tolerance ϵ.
while (TRUE)

1. compute Newton step and decrement: $\Delta\mathbf{x}_{\rm Nt}, \lambda(\mathbf{x})$.
2. quit if $\lambda^2(\mathbf{x}) \leq \epsilon$.
3. Perform backtracking line search and choose t.
4. $\mathbf{x} := \mathbf{x} + t\Delta\mathbf{x}_{\rm Nt}$.

Table 8.1: The Newton algorithm for problems with equality constraints starting from a feasible point.

8.5 Convergence Analysis

The Newton method for problems with linear equality constraints, starting from a feasible point, behaves like the Newton method for unconstrained optimization problems.

For details of the convergence analysis of the method see Section 10.2.4 of the book of Boyd and Vandenberghe.

8.6 Newton algorithm starting from an infeasible point

Next, we will study the important case where we start the Newton algorithm from an infeasible $\mathbf{x} \in \mathbf{dom}\, f$, that is, $\mathbf{Ax} \neq \mathbf{b}$.

This may happen either because the computation of a feasible point is difficult, for example, when $\mathbf{dom}\, f \subset \mathbb{R}^n$, or for any other reason.

8.7 Newton step at infeasible points

We repeat that the optimality conditions for the problem (8.1) are as follows:

$$\nabla f(\mathbf{x}_*) + \mathbf{A}^T\mathbf{v}_* = \mathbf{0}, \quad \mathbf{A}\mathbf{x}_* = \mathbf{b}.$$

Let $\mathbf{x} \in \mathbf{dom}\, f$ be a point which is *not* feasible, that is,

$$\mathbf{A}\mathbf{x} \neq \mathbf{b}.$$

Our goal is to find a step $\Delta\mathbf{x}$ such that the point $\mathbf{x} + \Delta\mathbf{x}$ is feasible and approximately optimal, that is,

$$\mathbf{A}(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}, \quad \mathbf{x} + \Delta\mathbf{x} \approx \mathbf{x}_*.$$

Replacing $\mathbf{x}_*$ in the optimality conditions by $\mathbf{x}+\Delta\mathbf{x}$ and $\mathbf{v}_*$ by $\mathbf{w}$, and using the first-order Taylor approximation

$$\nabla f(\mathbf{x} + \Delta\mathbf{x}) \approx \nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\Delta\mathbf{x}, \tag{8.30}$$

we obtain

$$\nabla f(\mathbf{x}) + \nabla^2 f(\mathbf{x})\Delta\mathbf{x} + \mathbf{A}^T\mathbf{w} = \mathbf{0}, \quad \mathbf{A}(\mathbf{x} + \Delta\mathbf{x}) = \mathbf{b}, \tag{8.31}$$

or, equivalently,

$$\begin{bmatrix} \nabla^2 f(\mathbf{x}) & \mathbf{A}^T \\ \mathbf{A} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \mathbf{w} \end{bmatrix} = - \begin{bmatrix} \nabla f(\mathbf{x}) \\ \mathbf{A}\mathbf{x} - \mathbf{b} \end{bmatrix}. \tag{8.32}$$

The only difference between the relation (8.32), which defines the Newton step at an infeasible point, and the relation (8.17), which defines the Newton step at a feasible point, is in the second term of the right-hand side, which is $\mathbf{A}\mathbf{x} - \mathbf{b}$ in (8.32), while it is zero in (8.17). If $\mathbf{x}$ is feasible, then the two systems are identical. If, however, $\mathbf{x}$ is not feasible, then the two terms are different.

We note that if $\mathbf{x}$ is not feasible and we take the full step $\Delta\mathbf{x}$, then the new point $\mathbf{x} + \Delta\mathbf{x}$ will *be* feasible.

An extremely interesting case arises when $\mathbf{x}$ is *not* a feasible point but it is *not* possible to move by $\Delta\mathbf{x}$, because $\mathbf{dom}\, f \subset \mathbb{R}^n$ and $\mathbf{x} + \Delta\mathbf{x} \notin \mathbf{dom}\, f$. We will consider this case in the sequel.

8.7.1 Interpretation as primal–dual Newton step

Next, we give another interpretation of the relation (8.32) in the context of the algorithm *primal–dual*, which updates both the primal variable $\mathbf{x}$ and the dual variable $\mathbf{v}$ with the goal of solving the KKT conditions.

Let the function $r : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n \times \mathbb{R}^p$ be defined as

$$r(\mathbf{x}, \mathbf{v}) := (r_{\text{dual}}(\mathbf{x}, \mathbf{v}), r_{\text{primal}}(\mathbf{x}, \mathbf{v})), \tag{8.33}$$

with

$$r_{\text{dual}}(\mathbf{x}, \mathbf{v}) := \nabla f(\mathbf{x}) + \mathbf{A}^T\mathbf{v}, \quad r_{\text{primal}}(\mathbf{x}, \mathbf{v}) := \mathbf{A}\mathbf{x} - \mathbf{b}. \tag{8.34}$$

The optimality conditions (8.2) of the problem (8.1) can be expressed as

$$r(\mathbf{x}_*, \mathbf{v}_*) = \mathbf{0}. \tag{8.35}$$

The quantities $r_{\text{dual}}(\mathbf{x}, \mathbf{v})$ and $r_{\text{primal}}(\mathbf{x}, \mathbf{v})$ are called, respectively, the **dual residual** and the **primal residual** at the point $(\mathbf{x}, \mathbf{v})$. Their norms measure the violation of the KKT conditions."

The first-order Taylor approximation of the function r, at the point $\mathbf{y} = (\mathbf{x}, \mathbf{v})$, is given by

$$r(\mathbf{y} + \mathbf{z}) \approx r_{\mathbf{y}}(\mathbf{z}) = r(\mathbf{y}) + Dr(\mathbf{y})\mathbf{z}, \tag{8.36}$$

where $Dr(\mathbf{y})$ is the derivative of r at $\mathbf{y}$. Function $r_{\mathbf{y}}(\mathbf{z})$ is an affine function of $\mathbf{z}$.

We define as **primal-dual Newton step**

$$\Delta\mathbf{y}_{\text{pd}} := (\Delta\mathbf{x}_{\text{pd}}, \Delta\mathbf{v}_{\text{pd}}), \tag{8.37}$$

at the point $\mathbf{y}$, as the vector $\mathbf{z}$ for which the right-hand side of (8.36) becomes equal to zero, i.e.,

$$Dr(\mathbf{y})\Delta\mathbf{y}_{\text{pd}} = -r(\mathbf{y}). \tag{8.38}$$

That is, the primal–dual Newton step is defined in terms of $\mathbf{x}$ *and* $\mathbf{v}$.

The derivative $Dr(\mathbf{y}) = Dr(\mathbf{x}, \mathbf{v})$ is as follows

$$Dr(\mathbf{x}, \mathbf{v}) = \begin{bmatrix} Dr_{\text{dual}}(\mathbf{x}, \mathbf{v}) \\ Dr_{\text{primal}}(\mathbf{x}, \mathbf{v}) \end{bmatrix} = \begin{bmatrix} \nabla^2 f(\mathbf{x}) & \mathbf{A}^T \\ \mathbf{A} & \mathbf{O} \end{bmatrix}, \tag{8.39}$$

therefore, (8.38) can be expressed as

$$\begin{bmatrix} \nabla^2 f(\mathbf{x}) & \mathbf{A}^T \\ \mathbf{A} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}_{\text{pd}} \\ \Delta\mathbf{v}_{\text{pd}} \end{bmatrix} = - \begin{bmatrix} r_{\text{dual}}(\mathbf{x}, \mathbf{v}) \\ r_{\text{primal}}(\mathbf{x}, \mathbf{v}) \end{bmatrix} = - \begin{bmatrix} \nabla f(\mathbf{x}) + \mathbf{A}^T\mathbf{v} \\ \mathbf{A}\mathbf{x} - \mathbf{b} \end{bmatrix}. \tag{8.40}$$

Setting $\mathbf{v}_+ = \mathbf{v} + \Delta\mathbf{v}_{\rm pd}$, we obtain the relation

$$\begin{bmatrix} \nabla^2 f(\mathbf{x}) & \mathbf{A}^T \\ \mathbf{A} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x}_{\rm pd} \\ \mathbf{v}_+ \end{bmatrix} = - \begin{bmatrix} \nabla f(\mathbf{x}) \\ \mathbf{A}\mathbf{x} - \mathbf{b} \end{bmatrix}, \tag{8.41}$$

which is identical to (8.32) for

$$\Delta\mathbf{x} = \Delta\mathbf{x}_{\rm pd}, \quad \mathbf{w} = \mathbf{v}_+ = \mathbf{v} + \Delta\mathbf{v}_{\rm pd}. \tag{8.42}$$

8.7.2 Merit function

We recall that the Newton direction, $\Delta\mathbf{x}$, at a **infeasible point x** is given by the relation (8.31). The direction $\Delta\mathbf{x}$ is not necessarily a descent direction for f, at the point $\mathbf{x}$, because quantity

$$\begin{aligned} \frac{d}{dt} f(\mathbf{x} + t\Delta\mathbf{x})\Big|_{t=0} &= \nabla f(\mathbf{x})^T \Delta\mathbf{x} \\ &\overset{(8.31)}{=} -\Delta\mathbf{x}^T \left(\nabla^2 f(\mathbf{x}) \Delta\mathbf{x} + \mathbf{A}^T \mathbf{w} \right) \\ &\overset{(8.41)}{=} -\Delta\mathbf{x}^T \nabla^2 f(\mathbf{x}) \Delta\mathbf{x} + (\mathbf{A}\mathbf{x} - \mathbf{b})^T \mathbf{w} \end{aligned} \tag{8.43}$$

is not necessarily negative (unless $\mathbf{A}\mathbf{x} = \mathbf{b}$). This fact has several consequences. For example, we cannot speak of "descent Newton step," with all its implications.

However, the interpretation of the quantity $\Delta\mathbf{y}_{\rm pd} = (\Delta\mathbf{x}_{\rm pd}, \Delta\mathbf{v}_{\rm pd})$ as primal–dual Newton step allows the step $\Delta\mathbf{y}_{\rm pd}$ to be related to another function.

The directional derivative of the function $\|r\|_2^2$, at the point $\mathbf{y}$ along the direction $\Delta\mathbf{y}_{\rm pd}$, is[2]

$$\begin{aligned} \frac{d}{dt} \left\| r(\mathbf{y} + t\Delta\mathbf{y}_{\rm pd}) \right\|_2^2 \Big|_{t=0} &= 2\, r(\mathbf{y})^T Dr(\mathbf{y}) \Delta\mathbf{y}_{\rm pd} \\ &= -2\, r(\mathbf{y})^T r(\mathbf{y}) \\ &= -2 \left\| r(\mathbf{y}) \right\|_2^2 \\ &\leq 0. \end{aligned} \tag{8.44}$$

[2]Let $g(\mathbf{u}) = \|\mathbf{u}\|_2^2$. From the chain rule, we have

$$D(g \circ r \circ \mathbf{c})(t) = Dg((r \circ \mathbf{c})(t))\, D(r \circ \mathbf{c})(t) = Dg((r \circ \mathbf{c})(t))\, Dr(\mathbf{c}(t))\, D\mathbf{c}(t).$$

Setting $\mathbf{c}(t) = \mathbf{y} + t\Delta\mathbf{y}_{pd}$, $(r \circ \mathbf{c})(t) = r(\mathbf{c}(t))$ and $(g \circ r \circ \mathbf{c})(t) = \|r(\mathbf{c}(t))\|_2^2$, we get (8.44).

Thus, whenever $r(\mathbf{y}) \neq \mathbf{0}$, the direction $\Delta\mathbf{y}_{\text{pd}}$ is a strict descent direction for the merit function $\|r\|_2^2$.

This fact allows the use of the quantity $\|r\|_2$ as a measure of progress of the Newton method, starting from an infeasible point, and can be used, for example, in a line search and/or in algorithm termination criteria.

8.7.3 Full-Step Feasibility Property

As we have seen, if we are at an infeasible point and take the full step $\Delta\mathbf{x}_{\text{pd}}$ (i.e., $t = 1$), then the new point is feasible. Subsequent full Newton steps preserve feasibility.

If, however, we move by $t\Delta\mathbf{x}_{\text{pd}}$, with $t \in (0, 1)$, then the new point will not be feasible. If we define $\mathbf{x}_+ = \mathbf{x} + t\Delta\mathbf{x}_{\text{pd}}$, and $\mathbf{v}_+ = \mathbf{v} + t\Delta\mathbf{v}_{\text{pd}}$, then

$$\begin{aligned} r_{\text{primal}}(\mathbf{x}_+, \mathbf{v}_+) &= \mathbf{A}(\mathbf{x} + t\Delta\mathbf{x}_{\text{pd}}) - \mathbf{b} \\ &\overset{(a)}{=} (1-t)(\mathbf{A}\mathbf{x} - \mathbf{b}) \\ &= (1-t) r_{\text{primal}}(\mathbf{x}, \mathbf{v}), \end{aligned} \tag{8.45}$$

where at point (a) we used the fact that, because of the relation (8.40), we have

$$\mathbf{A}\Delta\mathbf{x}_{\text{pd}} = -(\mathbf{A}\mathbf{x} - \mathbf{b}).$$

Working recursively, we obtain

$$r_{\text{primal}}(\mathbf{x}_k, \mathbf{v}_k) = \left(\prod_{i=0}^{k-1} (1 - t_i) \right) r_{\text{primal}}(\mathbf{x}_0, \mathbf{v}_0). \tag{8.46}$$

Obviously, if, for some k, we have $t_k = 1$, then the points $\mathbf{x}_l$, for $l > k$, will be feasible and the primal residual will be zero.

8.7.4 Newton algorithm starting from an infeasible point

The Newton algorithm starting from an infeasible point is listed in Table 8.2 (see also Fig. 8.2).

An obvious variation is as follows. If, at some step, we choose $t = 1$, then the next point will be feasible and we can continue with the Newton algorithm starting from a feasible point.

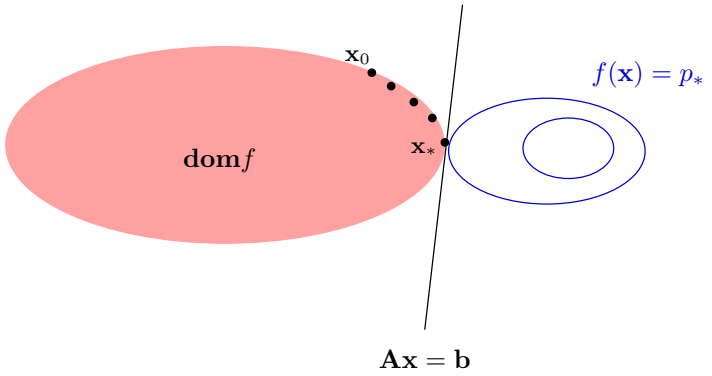


Figure 8.2: Trajectory $\mathbf{x}_0, \ldots, \mathbf{x}_*$ of the Newton algorithm, starting from an infeasible point.

$\mathbf{x} \in \mathbf{dom}\, f, \mathbf{v} \in \mathbb{R}^p$, tolerance ϵ.
Repeat

1. compute primal and dual Newton steps : $\Delta\mathbf{x}_{\rm pd}, \Delta\mathbf{v}_{\rm pd}$.
2. Backtracking line search
   $t := 1$
   while $\|r(\mathbf{x} + t\Delta\mathbf{x}_{\rm pd}, \mathbf{v} + t\Delta\mathbf{v}_{\rm pd})\|_2 > (1 - \alpha t)\|r(\mathbf{x}, \mathbf{v})\|_2$
   $t := \beta\, t$.
3. $\mathbf{x} := \mathbf{x} + t\Delta\mathbf{x}_{\rm pd}, \mathbf{v} := \mathbf{v} + t\Delta\mathbf{v}_{\rm pd}$.

until $\|r(\mathbf{x}, \mathbf{v})\|_2 \leq \epsilon$.

Table 8.2: The Newton algorithm for problems with equality constraints starting from an infeasible point (primal–dual).

The main difference between these two approaches lies in the line search. However, this difference does not seem to significantly differentiate the behavior of the methods.

8.7.5 In practice: Newton method for convex problems with affine equality constraints

As we have seen, if $\mathbf{dom}\, f \subset \mathbb{R}^n$, then finding a feasible point may be difficult. In this case, one approach is to solve the problem in two phases.

In Phase I, we solve a **feasibility problem**. If this problem has a solution, then we can proceed to Phase II, using the Newton algorithm starting from a feasible point.

If $\mathbf{dom}\, f$ is relatively simple and we know that the problem is feasible, then it may be preferable to use the Newton algorithm starting from an infeasible point.

Example 8.7.1. Consider the problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) = -\sum_{i=1}^{n} \log x_i \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{array} \tag{8.47}$$

with $\mathbf{dom}\, f = \mathbb{R}^n_{++}$. We can adopt the following approaches.

1. Find a point $\mathbf{x}_0$, if it exists, with $\mathbf{x}_0 \in \mathbf{dom}\, f$ and $\mathbf{A}\mathbf{x}_0 = \mathbf{b}$, and, then, use the Newton algorithm starting from a feasible point. Finding a feasible starting point requires the solution of the **feasibility problem**

$$\begin{array}{ll} \text{minimize} & f(\mathbf{x}) = 0 \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \\ & \mathbf{x} > \mathbf{0}. \end{array} \tag{8.48}$$

   We note that, for the moment, we do not know how to solve the problem (8.48), because it contains inequalities (we will solve problems with inequality constraints in the next chapter).

2. Start from an infeasible point $\mathbf{x} \in \mathbf{dom}\, f$, for example, $\mathbf{x} = \mathbf{1}$, and use the primal–dual algorithm.

3. Solve the dual problem (if it has a solution)

$$\max_{\mathbf{v}} g(\mathbf{v}) = -\mathbf{b}^T\mathbf{v} + \sum_{i=1}^{n} \log\left(\mathbf{A}^T\mathbf{v}\right)_i + n, \tag{8.49}$$

   and compute the solution of the primal problem via the solution of the dual. More specifically, if $\mathbf{v}_*$ is the solution of the dual problem, then

$$x_{*,i} = 1/\left(\mathbf{A}^T\mathbf{v}_*\right)_i, \quad \text{for } i = 1, \ldots, n. \tag{8.50}$$

■

We understand that, in general, there are many approaches for the solution of an optimization problem. It is extremely important to know the advantages and disadvantages of each approach, so that we can choose the most suitable one.

Chapter 9

Convex Optimization Problems

In this chapter, we study convex optimization problems in their general form, that is, convex minimization problems with convex inequality constraints and affine equality constraints.

There are several approaches for the solution of these important problems. We will focus on the *interior point method*, stating that it is one of the most widely used methods for solving small- to medium-scale convex optimization problems.

9.1 Convex optimization problems

Consider the convex optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m, \\ & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{array} \tag{9.1}$$

where $f_i : \mathbf{dom}\, f_i \subseteq \mathbb{R}^n \to \mathbb{R}$, for $i = 0, \ldots, m$, are convex, twice differentiable functions defined on open, convex domains $\mathbf{dom}\, f_i$, for $i = 0, \ldots, m$, $\mathbf{A} \in \mathbb{R}^{p \times n}$, with $\mathbf{rank}(\mathbf{A}) = p$, and $\mathbf{b} \in \mathbb{R}^p$.

The set for which the problem (9.1) is defined is $\mathbb{D} := \bigcap_{i=0}^{m} \mathbf{dom}\, f_i$, while the feasible set is

$$\mathbb{X} := \{\mathbf{x} \in \mathbb{D} \,|\, f_i(\mathbf{x}) \leq 0,\ i = 1, \ldots, m,\ \mathbf{A}\mathbf{x} = \mathbf{b}\}. \tag{9.2}$$

We assume that the problem (9.1) has a solution, say, $\mathbf{x}_*$, and define $p_* := f_0(\mathbf{x}_*)$. Further, we assume that it is strictly feasible. That is, there exists $\mathbf{x} \in \mathbb{D}$ such that $\mathbf{A}\mathbf{x} = \mathbf{b}$ and

$f_i(\mathbf{x}) < 0$, for $i = 1, \ldots, m$ (i.e., the Slater condition is satisfied). In this case, the KKT conditions are necessary and sufficient optimality conditions.

If we define $\boldsymbol{\lambda}_* := [\lambda_{*,1} \cdots \lambda_{*,m}]^T$, then the KKT conditions are expressed as:

$$\begin{aligned} \nabla f_0(\mathbf{x}_*) + \sum_{i=1}^{m} \lambda_{*,i} \nabla f_i(\mathbf{x}_*) + \mathbf{A}^T \mathbf{v}_* &= \mathbf{0} \\ \mathbf{A}\mathbf{x}_* &= \mathbf{b} \\ f_i(\mathbf{x}_*) &\leq 0, \ i = 1, \ldots, m \\ \boldsymbol{\lambda}_* &\geq \mathbf{0} \\ \lambda_{*,i} f_i(\mathbf{x}_*) &= 0, \ i = 1, \ldots, m. \end{aligned} \tag{9.3}$$

9.2 Interior Point Methods

Interior point methods solve problem (9.1) by solving a *sequence* of problems with affine equality constraints. These problems are "approximations" of problem (9.1) and their solutions are strictly feasible for the problem (9.1). Next, we describe the process.

9.2.1 Logarithmic barrier function

The problem (9.1) can be expressed as

$$\begin{aligned} \text{minimize} \quad & f_0(\mathbf{x}) + \sum_{i=1}^{m} I_-(f_i(\mathbf{x})) \\ \text{subject to} \quad & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{aligned} \tag{9.4}$$

with

$$I_-(u) := \begin{cases} 0, & u \leq 0, \\ \infty, & u > 0. \end{cases} \tag{9.5}$$

The cost function of the problem (9.4) is, in general, non-differentiable (and extended-valued); therefore, standard gradient or Newton methods cannot be applied directly to it.

An approximation of the function $I_-(u)$ is $\hat{I}_-(u)$, which is defined as

$$\hat{I}_-(u) := -\frac{1}{t} \log(-u), \tag{9.6}$$

with $\mathbf{dom}\, \hat{I}_- = -\mathbb{R}_{++}$ and $t > 0$. The parameter t determines the approximation quality (see Figure 9.1). The function $\hat{I}_-$ is convex; when extended to take the value $+\infty$ outside

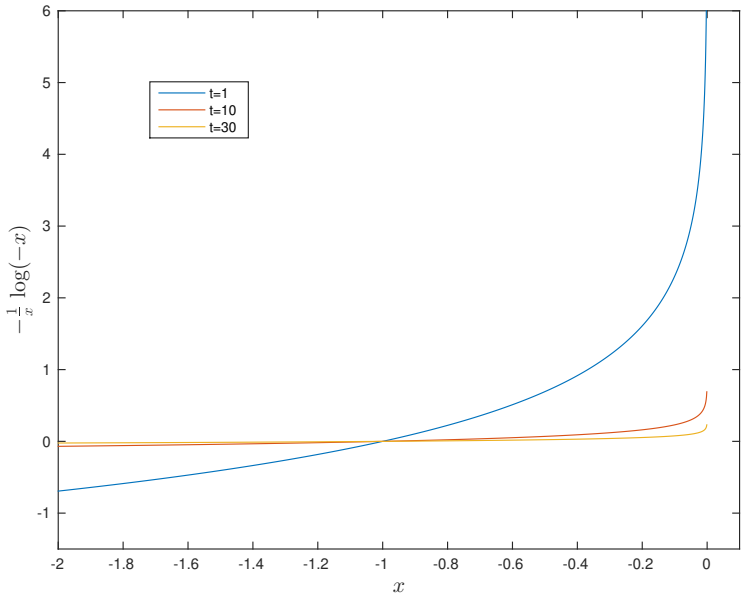


Figure 9.1: Approximations of the function $I_-(u)$, for $t = 1, 10, 30$.

its domain, it is closed (in particular, it tends to infinity as $u \uparrow 0$) and it will be extremely useful in the sequel.

First, we define the **logarithmic barrier function** for the problem (9.1)

$$\phi(\mathbf{x}) := -\sum_{i=1}^{m} \log(-f_i(\mathbf{x})), \tag{9.7}$$

with $\mathbf{dom}\,\phi = \{\mathbf{x} \in \bigcap_{i=1}^{m} \mathbf{dom}\, f_i \mid \ f_i(\mathbf{x}) < 0,\ i = 1, \ldots, m\}$. As we will prove next, the function ϕ is convex. Moreover, it is twice differentiable.

Next, we define the optimization problem

$$\begin{array}{ll} \text{minimize} & f_0(\mathbf{x}) + \frac{1}{t}\,\phi(\mathbf{x}) \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{array} \tag{9.8}$$

for $\mathbf{x} \in \mathbf{dom}\, f_0 \cap \mathbf{dom}\,\phi$ and $t > 0$.

Adding the function $\frac{1}{t}\,\phi(\mathbf{x})$ to the cost function $f_0(\mathbf{x})$, intuitively, "raises a barrier" on the boundary of the feasible set $\mathbb{X}$, essentially "trapping" the solution of the problem (9.8) inside $\mathbb{X}$ (but allowing it to reach "very close" to its boundary, for "large" t).

The problem (9.8) is convex and the Newton algorithm for problems with affine equality

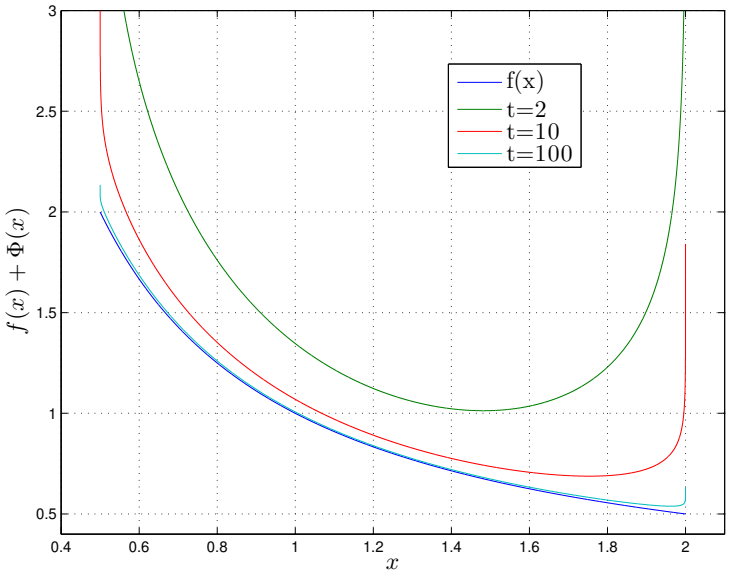


Figure 9.2: Functions $f_0(x) + \frac{1}{t}\,\phi(x)$, for $t = 2, 10, 100$.

constraints can be used for its solution. Obviously, other algorithms can be used as well. We will focus on the Newton algorithm.

Example 9.2.1. Consider the simple optimization problem with inequality constraints

$$\begin{array}{ll} \text{minimize} & f_0(x) = \frac{1}{x} \\ \text{subject to} & 0.5 \leq x \leq 2. \end{array} \tag{9.9}$$

The function $\phi(x)$ in this case is given by

$$\phi(x) = -\log(-(x-2)) - \log(-(0.5-x)). \tag{9.10}$$

In Figure 9.2, we plot the function $f_0(x) + \frac{1}{t}\,\phi(x)$, for $t = 2, 10, 100$. We observe that, as t increases, the function $f_0(x) + \frac{1}{t}\,\phi(x)$ approaches $f_0(x)$ except at the endpoints of the line interval $[0.5, 2]$ where it increases sharply, forming a "barrier." ■

Next, we study the problem

$$\begin{array}{ll} \text{minimize} & t f_0(\mathbf{x}) + \phi(\mathbf{x}) \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b} \end{array} \tag{9.11}$$

which is equivalent to the problem (9.8).

The first and second derivatives of $\phi(\mathbf{x})$ are as follows (prove it!):

$$\begin{aligned}\nabla\phi(\mathbf{x}) &= \sum_{i=1}^{m} \frac{1}{-f_i(\mathbf{x})}\nabla f_i(\mathbf{x}),\\ \nabla^2\phi(\mathbf{x}) &= \sum_{i=1}^{m} \frac{1}{f_i^2(\mathbf{x})}\nabla f_i(\mathbf{x})\nabla f_i(\mathbf{x})^T - \sum_{i=1}^{m} \frac{1}{f_i(\mathbf{x})}\nabla^2 f_i(\mathbf{x}).\end{aligned} \tag{9.12}$$

From the form of $\nabla^2\phi(\mathbf{x})$ and the fact that $f_i(\mathbf{x}) < 0$ for $\mathbf{x} \in \mathbf{dom}\,\phi$, we easily see that $\phi(\mathbf{x})$ is a convex function (because its Hessian is positive semidefinite).

9.2.2 Central Path

We assume that, for $t > 0$, there is a solution to the problem (9.11), say, $\mathbf{x}_*(t)$ (hence, $f_0(\mathbf{x}_*(t)) > -\infty$.) The function $t \mapsto \mathbf{x}_*(t)$ is called the **central path**, while each point of the central path is called **central point**.

If $\mathbf{x}_*(t)$ is a central point, then there exists $\hat{\mathbf{v}} \in \mathbb{R}^p$ that satisfies the KKT conditions for the problem (9.11), that is,

$$t\nabla f_0(\mathbf{x}_*(t)) + \nabla\phi(\mathbf{x}_*(t)) + \mathbf{A}^T\hat{\mathbf{v}} = \mathbf{0}, \quad \mathbf{A}\mathbf{x}_*(t) = \mathbf{b}, \tag{9.13}$$

or, equivalently,

$$t\nabla f_0(\mathbf{x}_*(t)) + \sum_{i=1}^{m} \frac{1}{-f_i(\mathbf{x}_*(t))}\nabla f_i(\mathbf{x}_*(t)) + \mathbf{A}^T\hat{\mathbf{v}} = \mathbf{0}, \quad \mathbf{A}\mathbf{x}_*(t) = \mathbf{b}. \tag{9.14}$$

Moreover, $\mathbf{x}_*(t) \in \mathbf{dom}\, f_0 \cap \mathbf{dom}\,\phi$, therefore,

$$f_i(\mathbf{x}_*(t)) < 0, \quad \text{for } i = 1, \ldots, m. \tag{9.15}$$

9.2.3 Dual feasible points from central path points

From (9.14), we can draw the following important conclusion about the central path. Each central point, $\mathbf{x}_*(t)$, defines a dual feasible point for the problem (9.1) and, thus, a lower bound for p_*. More specifically, define

$$\lambda_{*,i}(t) := -\frac{1}{tf_i(\mathbf{x}_*(t))}, \quad \text{for } i = 1, \ldots, m, \quad \mathbf{v}_*(t) = \frac{\hat{\mathbf{v}}}{t}. \tag{9.16}$$

It can be proved that the point $(\boldsymbol{\lambda}_*(t), \mathbf{v}_*(t))$, with $\boldsymbol{\lambda}_*(t) := [\,\lambda_{*,1}(t) \;\cdots\; \lambda_{*,m}(t)\,]^T$, is a dual feasible point for the problem (9.1).[1]

[1] Recall that a point $(\boldsymbol{\lambda}, \mathbf{v})$ is dual feasible if $\boldsymbol{\lambda} \geq \mathbf{0}$ and $g(\boldsymbol{\lambda}, \mathbf{v}) > -\infty$.

The proof is as follows. By definition, we have that $\boldsymbol{\lambda}_*(t) > \mathbf{0}$, because $t > 0$ and $f_i(\mathbf{x}_*(t)) < 0$, for $i = 1, \ldots, m$.

The relation (9.14) can be expressed as

$$\nabla f_0(\mathbf{x}_*(t)) + \sum_{i=1}^{m} \lambda_{*,i}(t) \nabla f_i(\mathbf{x}_*(t)) + \mathbf{A}^T \mathbf{v}_*(t) = \mathbf{0}. \tag{9.17}$$

The Lagrangian for the problem (9.1) is

$$L(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \mathbf{v}^T(\mathbf{A}\mathbf{x} - \mathbf{b}). \tag{9.18}$$

Let $\boldsymbol{\lambda} = \boldsymbol{\lambda}_*(t)$ and $\mathbf{v} = \mathbf{v}_*(t)$. Then, we know that

$$g(\boldsymbol{\lambda}_*(t), \mathbf{v}_*(t)) = \inf_{\mathbf{x} \in \mathbb{D}} L(\mathbf{x}, \boldsymbol{\lambda}_*(t), \mathbf{v}_*(t)). \tag{9.19}$$

The function $L(\mathbf{x}, \boldsymbol{\lambda}_*(t), \mathbf{v}_*(t))$ is a convex function of $\mathbf{x}$ (why?). Therefore, a necessary and sufficient condition for its minimization over the open set $\mathbb{D}$ is

$$\nabla_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}_*(t), \mathbf{v}_*(t)) = \mathbf{0}. \tag{9.20}$$

We easily see that, due to (9.17), $\mathbf{x}_*(t)$ satisfies (9.20) and, therefore, minimizes function $L(\mathbf{x}, \boldsymbol{\lambda}_*(t), \mathbf{v}_*(t))$.

The dual function at point $(\boldsymbol{\lambda}_*(t), \mathbf{v}_*(t))$ is

$$\begin{aligned} g(\boldsymbol{\lambda}_*(t), \mathbf{v}_*(t)) &= L(\mathbf{x}_*(t), \boldsymbol{\lambda}_*(t), \mathbf{v}_*(t)) \\ &= f_0(\mathbf{x}_*(t)) + \sum_{i=1}^{m} \lambda_{*,i}(t) f_i(\mathbf{x}_*(t)) + \mathbf{v}_*(t)^T(\mathbf{A}\mathbf{x}_*(t) - \mathbf{b}) \\ &\overset{(9.16)}{=} f_0(\mathbf{x}_*(t)) - \frac{m}{t}. \end{aligned} \tag{9.21}$$

Since $f_0(\mathbf{x}_*(t)) > -\infty$, we have that $g(\boldsymbol{\lambda}_*(t), \mathbf{v}_*(t)) > -\infty$. So, the pair $(\boldsymbol{\lambda}_*(t), \mathbf{v}_*(t))$ is a dual feasible point for the problem (9.1). Therefore,

$$g(\boldsymbol{\lambda}_*(t), \mathbf{v}_*(t)) \leq p_*. \tag{9.22}$$

Combining the relations (9.21) and (9.22), we obtain the very important inequality

$$f_0(\mathbf{x}_*(t)) - p_* \leq \frac{m}{t}. \tag{9.23}$$

That is,

1. for a given t, the point $\mathbf{x}_*(t)$ is at most $\frac{m}{t}$–suboptimal for the problem (9.1),
2. as $t \to \infty$, the objective-value gap $f_0(\mathbf{x}_*(t)) - p_*$ converges to zero. In particular, any limit point of $\mathbf{x}_*(t)$ that exists is optimal for the problem (9.1).

9.2.4 Interpretation via KKT

We have seen that relations

$$\begin{aligned} \mathbf{A}\mathbf{x} = \mathbf{b}, \quad f_i(\mathbf{x}) \leq 0, \quad i = 1, \ldots, m \\ \boldsymbol{\lambda} \geq \mathbf{0} \\ \nabla f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \nabla f_i(\mathbf{x}) + \mathbf{A}^T \mathbf{v} = \mathbf{0} \\ \lambda_i f_i(\mathbf{x}) = -\frac{1}{t}, \quad i = 1, \ldots, m \end{aligned} \tag{9.24}$$

are satisfied by $\mathbf{x}_*(t)$, $\boldsymbol{\lambda}_*(t)$, and $\mathbf{v}_*(t)$.

The relations (9.24) are the KKT conditions for the problem (9.1), with the only difference being that, instead of zero in the right-hand side of the fourth line, there is the term $-\frac{1}{t}$.

Therefore, for "large" t, $\mathbf{x}_*(t)$, $\boldsymbol{\lambda}_*(t)$, and $\mathbf{v}_*(t)$ "almost" satisfy the KKT conditions for the problem (9.1).

9.3 The Barrier Method

We have seen that the point $\mathbf{x}_*(t)$ is at most $\frac{m}{t}$–suboptimal for the problem (9.1). An approach for the solution of (9.1), with tolerance ϵ, is to solve the problem

$$\begin{array}{ll} \text{minimize} & \frac{m}{\epsilon} f_0(\mathbf{x}) + \phi(\mathbf{x}) \\ \text{subject to} & \mathbf{A}\mathbf{x} = \mathbf{b}. \end{array} \tag{9.25}$$

In general, this problem is hard to solve for "small" ϵ. An efficient approach for the solution of problem (9.25) is the iterative solution of problems of the form (9.11), with increasing t. The solution of the problem for a particular t is the starting point for the next problem (with a larger t).

This method, which is called the *barrier method*, is described in Table 9.3. It was originally proposed by Fiacco and McCormick in the 1960s. But, during the next decade, it was not widely accepted mainly because there were arguments against the numerical stability of the method. For example, for large t, the Hessian of the function $tf_0 + \phi$ can become ill-conditioned, which can lead to difficulties in the numerical implementation of the method.

The method reappeared during the 1980s, when it was associated with Karmarkar's algorithm for the solution of linear programming problems. Today, interior-point methods

remain an important approach for the solution of small- to medium-scale convex optimization problems.

$\mathbf{x} \in \mathbf{dom}\, f_0 \cap \mathbf{dom}\, \phi,\ \mathbf{A}\mathbf{x} = \mathbf{b},\ t > 0,\ \mu > 1,$ tolerance $\epsilon > 0$.

repeat

1. Compute $\mathbf{x}_*(t)$, by minimizing $tf_0 + \phi$ subject to $\mathbf{A}\mathbf{x} = \mathbf{b}$, starting at $\mathbf{x}$.
2. $\mathbf{x} := \mathbf{x}_*(t)$.
3. quit if $\frac{m}{t} < \epsilon$.
4. $t := \mu\, t$.

Table 9.1: Barrier method for convex optimization problems.

9.4 Feasibility and Phase I

One approach towards finding an initial feasible point for the problem (9.1) is the solution of the feasibility problem

$$\begin{array}{ll} \underset{\mathbf{x},s}{\text{minimize}} & s \\ \text{subject to} & f_i(\mathbf{x}) \leq s,\ i = 1, \ldots, m \\ & \mathbf{A}\mathbf{x} = \mathbf{b}, \end{array} \tag{9.26}$$

with optimization variables $\mathbf{x} \in \mathbb{R}^n$ and $s \in \mathbb{R}$ (we recall that $\mathbf{A} \in \mathbb{R}^{p \times n}$ and $\mathbf{b} \in \mathbb{R}^p$).

If there exists $\bar{\mathbf{x}} \in \mathbb{D}$ such that $\mathbf{A}\bar{\mathbf{x}} = \mathbf{b}$, then the problem (9.26) is strictly feasible, because we can put as initial point the vector $(\bar{\mathbf{x}}, \bar{s})$ with

$$\bar{s} > \max\{f_1(\bar{\mathbf{x}}), \ldots, f_m(\bar{\mathbf{x}})\}. \tag{9.27}$$

If $p < n$, then the system $\mathbf{A}\mathbf{x} = \mathbf{b}$ has infinitely many solutions (recall that we have assumed that the rows of $\mathbf{A}$ are linearly independent). One solution of $\mathbf{A}\mathbf{x} = \mathbf{b}$ is given by $\bar{\mathbf{x}} = \mathbf{A}^\sharp \mathbf{b}$, where $\mathbf{A}^\sharp$ is the pseudoinverse of $\mathbf{A}$. For Phase I, however, we also require $\bar{\mathbf{x}} \in \mathbb{D}$.

Let $(\bar{\mathbf{x}}_*, \bar{s}_*)$ be the optimal point of the problem (9.26) and $\bar{p}_*$ the optimal value. Then

1. if $\bar{p}_* > 0$, then the problem (9.1) is *infeasible*.
2. if $\bar{p}_* < 0$, then the problem (9.1) is strictly feasible and a strictly feasible point for problem (9.1) is $\bar{\mathbf{x}}_*$.

3. if $\bar{p}_* = 0$, then the original problem is feasible but not strictly feasible (assuming the Phase I optimum is attained). In practice, it may happen that $|\bar{p}_*| < \epsilon$, for very small $\epsilon > 0$. Then, the inequalities $f_i(\mathbf{x}) \leq -\epsilon$, for $i = 1, \ldots, m$, are infeasible, while the inequalities $f_i(\mathbf{x}) \leq \epsilon$, for $i = 1, \ldots, m$, are feasible.

We note that it is not necessary to solve the problem (9.26). It suffices to find an $\mathbf{x}$ which leads to $s < 0$.

9.5 Primal–Dual Interior Point Method

9.5.1 Primal–dual direction

We define

$$\mathbf{f}(\mathbf{x}) := \begin{bmatrix} f_1(\mathbf{x}) \\ \vdots \\ f_m(\mathbf{x}) \end{bmatrix}, \quad D\mathbf{f}(\mathbf{x}) := \begin{bmatrix} \nabla f_1(\mathbf{x})^T \\ \vdots \\ \nabla f_m(\mathbf{x})^T \end{bmatrix}. \tag{9.28}$$

We define the vector $r_t(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v})$ as follows:

$$r_t(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) := \begin{bmatrix} \nabla f_0(\mathbf{x}) + D\mathbf{f}(\mathbf{x})^T \boldsymbol{\lambda} + \mathbf{A}^T \mathbf{v} \\ -\mathbf{diag}(\boldsymbol{\lambda})\mathbf{f}(\mathbf{x}) - \frac{1}{t}\,\mathbf{1}_m \\ \mathbf{A}\mathbf{x} - \mathbf{b} \end{bmatrix}. \tag{9.29}$$

If $\mathbf{x} \in \mathbf{dom}\, f_0 \cap \mathbf{dom}\, \phi$, $\boldsymbol{\lambda} > \mathbf{0}$, and $\mathbf{v}$ satisfy the equation $r_t(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v}) = \mathbf{0}$, then $\mathbf{x}$ is a central point for the parameter t, and $\boldsymbol{\lambda}$ and $\mathbf{v}$ are the associated dual variables. If the central point and associated dual variables are unique, they equal $\mathbf{x}_*(t)$, $\boldsymbol{\lambda}_*(t)$, and $\mathbf{v}_*(t)$, respectively.

In particular, $\mathbf{x}_*(t)$ is a primal feasible point and $(\boldsymbol{\lambda}_*(t), \mathbf{v}_*(t))$ is a dual feasible point, with duality gap $\frac{m}{t}$.

Suppose we are at the point $\mathbf{y} = (\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v})$, with $\mathbf{x} \in \mathbf{dom}\, f_0 \cap \mathbf{dom}\, \phi$ and $\boldsymbol{\lambda} > \mathbf{0}$.

The vector

$$r_{t,\mathrm{d}}(\mathbf{y}) := \nabla f_0(\mathbf{x}) + D\mathbf{f}(\mathbf{x})^T \boldsymbol{\lambda} + \mathbf{A}^T \mathbf{v} \tag{9.30}$$

is called **dual residual**, the vector

$$r_{t,\mathrm{p}}(\mathbf{y}) := \mathbf{A}\mathbf{x} - \mathbf{b} \tag{9.31}$$

is called **primal residual**, while the vector

$$r_{t,c}(\mathbf{y}) := -\mathbf{diag}(\boldsymbol{\lambda})\mathbf{f}(\mathbf{x}) - \frac{1}{t}\,\mathbf{1}_m, \tag{9.32}$$

where $\mathbf{1}_m$ is the m-dimensional vector with unit elements, is called **centrality residual**.

The first-order approximation of r_t, at the point $\mathbf{y}$, is as follows:

$$r_t(\mathbf{y}+\mathbf{z}) \approx r_{t,\mathbf{y}}(\mathbf{z}) = r_t(\mathbf{y}) + Dr_t(\mathbf{y})\mathbf{z}. \tag{9.33}$$

We define as the Newton step $\Delta\mathbf{y} = (\Delta\mathbf{x}, \Delta\boldsymbol{\lambda}, \Delta\mathbf{v})$, at the point $\mathbf{y}$, the vector $\mathbf{z}$ for which $r_{t,\mathbf{y}}(\mathbf{z})$ vanishes, i.e.

$$Dr_t(\mathbf{y})\Delta\mathbf{y} = -r_t(\mathbf{y}). \tag{9.34}$$

Calculating the derivative Dr_t, at the point $\mathbf{y}$, we obtain

$$\begin{bmatrix} \nabla^2 f_0(\mathbf{x}) + \sum_{i=1}^m \lambda_i \nabla^2 f_i(\mathbf{x}) & D\mathbf{f}(\mathbf{x})^T & \mathbf{A}^T \\ -\mathbf{diag}(\boldsymbol{\lambda})D\mathbf{f}(\mathbf{x}) & -\mathbf{diag}(\mathbf{f}(\mathbf{x})) & \mathbf{O} \\ \mathbf{A} & \mathbf{O} & \mathbf{O} \end{bmatrix} \begin{bmatrix} \Delta\mathbf{x} \\ \Delta\boldsymbol{\lambda} \\ \Delta\mathbf{v} \end{bmatrix} = - \begin{bmatrix} r_{t,\mathrm{d}}(\mathbf{y}) \\ r_{t,\mathrm{c}}(\mathbf{y}) \\ r_{t,\mathrm{p}}(\mathbf{y}) \end{bmatrix}. \tag{9.35}$$

We define as **primal-dual search direction**, $\Delta\mathbf{y}_{\mathrm{pd}} := (\Delta\mathbf{x}_{\mathrm{pd}}, \Delta\boldsymbol{\lambda}_{\mathrm{pd}}, \Delta\mathbf{v}_{\mathrm{pd}})$, the Newton step $\Delta\mathbf{y}$.

The variables $\mathbf{x}_k$ and $(\boldsymbol{\lambda}_k, \mathbf{v}_k)$ of the primal–dual interior point algorithm are not necessarily feasible points of the problem (9.1), but only asymptotically (i.e., at the convergence of the algorithm). This means that we have no obvious way to compute the duality gap during algorithm execution.

For $\mathbf{x} \in \mathbf{dom}\, f_0 \cap \mathbf{dom}\, \phi$ and $\boldsymbol{\lambda} > \mathbf{0}$, we define the **surrogate duality gap** as

$$\hat{\eta}(\mathbf{x}, \boldsymbol{\lambda}) := -\mathbf{f}(\mathbf{x})^T\boldsymbol{\lambda}. \tag{9.36}$$

If $\mathbf{x}$ is primal feasible and $(\boldsymbol{\lambda}, \mathbf{v})$ is dual feasible, then $\hat{\eta}(\mathbf{x}, \boldsymbol{\lambda})$ equals the primal–dual gap whenever stationarity holds. Along the central path it equals m/t.

Accordingly, in the primal–dual method we associate the surrogate duality gap $\hat{\eta}$ with the parameter $t = \frac{m}{\hat{\eta}}$.

9.5.2 Line search

Let

$$\mathbf{x}_+ := \mathbf{x} + s\Delta\mathbf{x}_{\mathrm{pd}},\ \boldsymbol{\lambda}_+ := \boldsymbol{\lambda} + s\Delta\boldsymbol{\lambda}_{\mathrm{pd}},\ \mathbf{v}_+ := \mathbf{v} + s\Delta\mathbf{v}_{\mathrm{pd}}. \tag{9.37}$$

A line search method should guarantee that $\boldsymbol{\lambda}_+ > \mathbf{0}$, $\mathbf{f}(\mathbf{x}_+) < \mathbf{0}$ and the function $\|r(\cdot)\|_2$ decreases quite a bit (see Boyd & Vandenberghe, page 613). We proceed as follows:

1. Compute the quantity $s^{\max} := \sup\{s \in [0,1] \,|\, \boldsymbol{\lambda} + s\Delta\boldsymbol{\lambda}_{\mathrm{pd}} > \mathbf{0}\}$, and choose an initial $s < s^{\max}$ (typically slightly smaller than $s^{\max}$).
2. Backtrack by $\beta \in (0,1)$ until $\mathbf{f}(\mathbf{x}_+) < \mathbf{0}$.
3. Backtrack by $\beta \in (0,1)$ until

$$\|r_t(\mathbf{x}_+, \boldsymbol{\lambda}_+, \mathbf{v}_+)\|_2 \leq (1 - \alpha s)\|r_t(\mathbf{x}, \boldsymbol{\lambda}, \mathbf{v})\|_2 \tag{9.38}$$

for $\alpha \in (0, 0.5)$.

The relation (9.38) is compatible with the backtracking line search we have encountered in previous algorithms. Under the standard regularity assumptions for the primal–dual Newton method, the backtracking procedure terminates after a finite number of steps.

9.5.3 Primal-dual algorithm for convex optimization problems

The primal-dual algorithm is presented next. We note that we should start from a point that strictly satisfies all inequality constraints.

$\mathbf{x} \in \mathbf{dom}\, f_0 \cap \mathbf{dom}\, \phi,\ \boldsymbol{\lambda} > \mathbf{0},\ \mu > 1,\ \epsilon_{\mathrm{feas}} > 0,\ \epsilon > 0.$

repeat

1. $t := \mu \frac{m}{\hat{\eta}(\mathbf{x}, \boldsymbol{\lambda})}$.
2. Compute $\Delta\mathbf{y}_{\mathrm{pd}}$.
3. perform line search and choose step s.
4. $\mathbf{y} := \mathbf{y} + s\Delta\mathbf{y}_{\mathrm{pd}}$.

until $\left(\|r_{t,\mathrm{p}}(\mathbf{y})\|_2 < \epsilon_{\mathrm{feas}},\ \|r_{t,\mathrm{d}}(\mathbf{y})\|_2 < \epsilon_{\mathrm{feas}},\ \hat{\eta}(\mathbf{x}, \boldsymbol{\lambda}) < \epsilon\right).$
