Contents

Introduction

The goal of *Mathematical Programming* is the design of mathematical solution methods for optimization problems. These methods should be *algorithmic* in the sense that they can be converted into computer programs without much effort. The main concern, however, is not the eventual concrete implementation of an algorithm but the necessary prerequisit thereof: the exhibition of a solution strategy that hopefully makes the ensuing algorithm "efficient" in practice. Mathematical programming thus offers an approach to the theory of mathematical optimization that is very much motivated by the question whether certain parameters (solutions of an optimization problem or eigenvalues of a matrix) not only exist in an abstract way but can actually be computed well enough to satisfy practical needs.

Mathematical optimization traditionally decomposes into three seemingly rather disjoint areas: *Discrete* (or *combinatorial*) *optimization*, *linear optimization* and *nonlinear optimization*. Yet, a closer look reveals a different picture. Efficiently solvable discrete optimization problems are typically those that can be cast into the framework of linear optimization. And, as a rule of thumb, nonlinear problems are solved by repeated linear (or quadratic) approximation.

The dominant role of linearity in optimization is not surprising. It has long been known that much of the structural analysis of mathematical optimization can be achieved taking advantage of the language of vector spaces (see, for example, Luenberger's elegant classic treatment [55]). Moreover, it appears to be an empirical fact that not only computations in linear algebra can be carried out numerically efficiently in practice but that, indeed, efficient numerical computation is tantamount to being able to reduce the computational task as much as possible to linear algebra.

The present book wants to introduce the reader to the fundamental algorithmic techniques in mathematical programming with a strong emphasis on the central position of linear algebra both in the structural analysis and the computational procedures. Although an optimization problem often admits a geometric picture involving sets of points in Euclidean space, which may guide the intuition in the structural analysis, we stress the role of the *presentation* of a problem in terms of explicit functions that encode the set of admissible solutions and the quantity to be optimized. The presentation is crucial for the design of a solution method and its efficiency.

The book attempts to be as much self-contained as possible. Only basic knowledge about (real) vector spaces and differentiable functions is assumed at the outset. Chapter 1 reviews this material, providing proofs of facts that might be not (yet) so familiar to the reader. We really begin in Chapter 2, which introduces the fundamental techniques of numerical linear algebra we will rely on later. Chapter 3 provides the corresponding geometric point of view. Then linear programs are treated.

Having linear programming techniques at our disposal, we investigate discrete optimization problems and discuss theories for analyzing their ”complexity” with respect to their solvability by ”efficient” algorithms. Nonlinear programs proper are presented in the last three chapters. Convex minimization problems occupy here a position between linear and nonlinear structures: while the feasible sets of linear programs are *finite* intersections of half-spaces, convex problems may be formulated with respect to *infinite* intersections of half-spaces. Convex optimization problems mark the border of efficient solvability. For example, quadratic optimization problems turn out to be “efficiently” solvable if and only if they are convex.

The book contains many items marked ”**Ex**”. These items are intended to provide both ”examples” and ”exercises” to which also details of proofs or additional observations are deferred. They are meant to be an integral part of the presentation of the material. We cordially invite the interested reader to test his or her understanding of the text by working them out in detail .

CHAPTER 1

Real Vector Spaces

This chapter is a brief review of the basic notions and facts from linear algebra and analysis that we will use as tools in mathematical programming. The reader is assumed to be already familiar with most of the material in this chapter. The proofs we sketch here (as well as the exercises) are mainly meant as reminders.

1.1. Linear and Affine Spaces

We discuss mathematical programming problems to a very large extent within the model of vector spaces V (or W *etc.*) over the field $\mathbb{R}$ of real numbers. The fundamental operations that define the structure of a vector space are: adding two vectors and multiplying a vector with a *scalar* (here: a real number $\lambda \in \mathbb{R}$). So we can carry out algebraic manipulations taking advantage of the following properties:

- If $\mathbf{v}, \mathbf{w}$ are elements of the vector space V, then the sum $\mathbf{z} = \mathbf{v} + \mathbf{w}$ is an element of V.
- If $\mathbf{v} \in V$ is a vector in V and $\lambda \in \mathbb{R}$ a scalar, then $\mathbf{z} = \lambda\mathbf{v}$ is a vector in V.
- The order in which vectors are added is irrelevant: $\mathbf{v} + \mathbf{w} = \mathbf{w} + \mathbf{v}$.

These properties reflect the *elementary operations* to which *all*(!) computations in linear algebra reduce. We note that the vector $\mathbf{0} = 0 \cdot \mathbf{v} \in V$ is uniquely determined (and independent of the particular choice of $\mathbf{v} \in V$).

There are two especially important examples of vector spaces in mathematical programming:

(1) $V = \mathbb{R}^n$, the vector space of all n-tuples $\mathbf{x} = (x_1, \ldots, x_n)^T$ of real numbers $x_j \in \mathbb{R}$. Addition and scalar multiplication of vectors in $\mathbb{R}^n$ is carried out componentwise.

(2) $V = \mathbb{R}^{m \times n}$, the vector space of all $(m \times n)$-matrices $\mathbf{A} = (a_{ij})$ of real numbers a_{ij}, where $i = 1, \ldots, m,\ \ j = 1, \ldots, n$. Addition and scalar multiplication is again componentwise.

REMARK. Having the full field $\mathbb{R}$ of real numbers available as field of scalars is important mainly when we deal with *limits* (as they are implicit in the notions of *continuous* or *differentiable* functions), or when we want to take square roots. In most other cases, it would suffice to restrict the scalars to the subfield $\mathbb{Q} \subset \mathbb{R}$ of rational numbers. Indeed, the numerical parameters one usually encounters and deals with in the computational practice are rational numbers. For convenience, we will usually assume $\mathbb{R}$ as the underlying field

of scalars.
While theoretical aspects of vector spaces are often fruitfully studied admitting scalars from the field $\mathbb{C}$ of complex numbers, this generality is less "natural" in mathematical programming, where pairs of scalars often must be compared with respect to the ordering $x < y$ of the real numbers (which cannot be extended to the complex numbers in an algebraically "reasonable" fashion).

We think of a vector $\mathbf{x} \in \mathbb{R}^n$ usually as a *column vector*, *i.e.*, as a $(n \times 1)$-matrix:

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

which we often abbreviate as $\mathbf{x} = (x_j)$. The *transpose* of $\mathbf{x}$ is the corresponding *row vector* $\mathbf{x}^T = (x_1, \ldots, x_n)$.

A matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ can be thought of as either an ordered set of n column vectors $\mathbf{A}_{\cdot j} = (a_{1j}, \ldots, a_{mj})^T \in \mathbb{R}^m$ or as an ordered set of m row vectors $\mathbf{A}_{i\cdot} = (a_{i1}, \ldots, a_{in})$ of *length* n. Alternatively, a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ can be viewed as a vector with $m \cdot n$ components a_{ij}, *i.e.*, as an element in $\mathbb{R}^{m \cdot n}$.

Of particular importance is the *identity* matrix $\mathbf{I} = (e_{ij}) \in \mathbb{R}^{n \times n}$, defined by

$$e_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

The n column vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$ of $\mathbf{I}$ are the so-called *unit vectors* of $\mathbb{R}^n$.

A *linear subspace* of the vector space V is a (non-empty) subset $W \subseteq V$ that is a vector space in its own right (relative to the addition and scalar multiplication in V). We then say that W is *closed* with respect to the operations of adding vectors and multiplying them by scalars.

Because the intersection of linear subspaces is again a linear subspace, we observe: For every subset S of the vector space V, there exists a unique smallest linear subspace span S, called the *span* (or *linear hull*) of S, that contains S.

The vector $\mathbf{z} \in V$ is a *linear combination* of the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k$, if there are scalars $\lambda_1, \ldots, \lambda_k \in \mathbb{R}$ such that $\mathbf{z} = \lambda_1 \mathbf{v}_1 + \ldots + \lambda_k \mathbf{v}_k$. Note that every linear combination results from a finite sequence of elementary operations:

$$\mathbf{z}_1 = \lambda_1 \mathbf{v}_1, \ \mathbf{z}_2 = \mathbf{z}_1 + \lambda_2 \mathbf{v}_2, \ \ldots, \ \mathbf{z}_k = \mathbf{z}_{k-1} + \lambda_k \mathbf{v}_k \ .$$

Ex. 1.1. *Show:* span *S consists of all (finite) linear combinations of vectors in S.*

If $\mathbf{v}_1, \ldots, \mathbf{v}_n$ are the column vectors of the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, and $\mathbf{x} = (x_1, \ldots, x_n)^T$ is a vector of coefficients x_j, we describe the resulting linear combination in matrix notation:

$$\mathbf{A}\mathbf{x} = x_1 \mathbf{v}_1 + \ldots + x_n \mathbf{v}_n \qquad \in \mathbb{R}^m \ .$$

In this case, span $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\} = \{\mathbf{A}\mathbf{x} \in \mathbb{R}^m \mid \mathbf{x} \in \mathbb{R}^n\} = \text{col } \mathbf{A}$ is called the *column space* of $\mathbf{A}$. Similarly, the linear hull of the row vectors of $\mathbf{A}$ is the *row space* row $\mathbf{A}$. Interchanging the rows and columns of $\mathbf{A}$ (*i.e*, passing from $\mathbf{A} = (a_{ij})$ to its *transpose* $\mathbf{A}^T = (\overline{a}_{ij})$, where $\overline{a}_{ij} = a_{ji}$), we have

$$\text{row } \mathbf{A} = \text{col } \mathbf{A}^T \ .$$

A non-empty subset $S \subseteq V$ of vectors is called (linearly) *independent* if no vector $\mathbf{s} \in S$ can be expressed as a linear combination of vectors in $S \setminus \{\mathbf{s}\}$.

Ex. 1.2. *Show:* $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ *is independent if and only if* $\lambda_1\mathbf{v}_1 + \ldots + \lambda_k\mathbf{v}_k = \mathbf{0}$ *implies* $\lambda_i = 0$ *for all* $i = 1, \ldots, k$.

A minimal subset $B \subseteq V$ such that $V = \text{span } B$ is a *basis* of V. Note that a non-empty basis is necessarily an independent set. The *dimension* of V is the number of elements in a basis B:

$$\dim V \ = \ |B| \ .$$

From linear algebra, we know that any two finite bases of a vector space have the same cardinality. This fact implies that "dim V" is well-defined and equals the maximal size of an independent subset of V. (There are many interesting infinite-dimensional vector spaces. For the applications in mathematical programming, however, we will always be concerned with finite-dimensional vector spaces).

A finite-dimensional vector space V with basis $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ can be identified with $\mathbb{R}^n$ in the following way. Because B is linearly independent, each vector $\mathbf{v} \in V$ has a unique representation

$$\mathbf{v} = x_1\mathbf{v}_1 + \ldots + x_n\mathbf{v}_n \ , \quad \text{where } \mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n \ .$$

So we can represent the element $\mathbf{v} \in V$ by the vector $\mathbf{x} \in \mathbb{R}^n$ of its coordinates with respect to B. It is not difficult to verify that this representation is compatible with the operations of adding vectors and multiplying vectors by scalars $\lambda \in \mathbb{R}$.

REMARK. The identification of V with $\mathbb{R}^n$ depends, of course, on the basis B we choose for the representation. Different bases lead to different representations.

In the case $V = \mathbb{R}^n$, the *unit vectors* $\mathbf{e}_1, \ldots, \mathbf{e}_n \in \mathbb{R}^n$ form the so-called *standard basis*. With respect to the latter, a vector $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$ has the representation

$$\mathbf{x} = x_1\mathbf{e}_1 + \ldots + x_n\mathbf{e}_n \ .$$

An *affine subspace* (also called a *linear variety*) of the vector space V is a subset $L \subseteq V$ such that either $L = \emptyset$ or there exists a vector $\mathbf{p} \in V$ and a linear subspace $W \subseteq V$ with the property

$$L = \mathbf{p} + W = \{\mathbf{p} + \mathbf{w} \mid \mathbf{w} \in W\} \ .$$

It is easy to verify that the affine subspace $L' = \mathbf{p}' + W'$ equals $L = \mathbf{p} + W$ if and only if $\mathbf{p}' \in L$ and $W' = W$. We define the *affine* dimension of L as

$$\dim L = \begin{cases} \dim W & \text{if } L \neq \emptyset, \\ -1 & \text{if } L = \emptyset. \end{cases}$$

An affine subspace of dimension 0 is called a *point* (and is identified with the element it contains). A *line* is an affine subspace of dimension 1. A *plane* is an affine subspace of dimension 2. An affine subspace H of V of dimension $\dim H = \dim V - 1$ is called a *hyperplane* of V.

Ex. 1.3. *Show:*

(a) *The linear subspaces are exactly the affine subspaces L with $\mathbf{0} \in L$.*
(b) *The intersection of affine subspaces is an affine subspace.*

A linear combination $\mathbf{z} = \lambda_1 \mathbf{v}_1 + \ldots + \lambda_k \mathbf{v}_k$ is said to be an *affine combination* of $\mathbf{v}_1, \ldots, \mathbf{v}_k$ if the scalars $\lambda_i \in \mathbb{R}$ satisfy the condition $\sum_{i=1}^{k} \lambda_i = 1$.

Ex. 1.4. *Show that the subset $L \subseteq V$ is an affine subspace if and only if L contains all (finite) affine combinations of elements of L.*

For an arbitrary subset $S \subseteq V$, we denote by aff S the set of all finite affine combinations of elements in S and call it the *affine hull* (or *affine span*) of S. By Ex. 1.4, aff S is the smallest affine subspace that contains S.

1.2. Maps and Matrices

Assume that V and W are vector spaces with bases $\{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ and $\{\mathbf{w}_1, \ldots, \mathbf{w}_m\}$ respectively. Every $\mathbf{v} \in V$ can be uniquely expressed as a linear combination $\mathbf{v} = x_1 \mathbf{v}_1 + \ldots + x_n \mathbf{v}_n$, with coefficients $x_j \in \mathbb{R}$, and every $\mathbf{w} \in W$ as a linear combination $\mathbf{w} = y_1 \mathbf{w}_1 + \ldots + y_m \mathbf{w}_m$, with coefficients $y_i \in \mathbb{R}$.

The map $f : V \to W$ is said to be *linear* if f is "compatible" with the vector space operations. This means that we have for all $\mathbf{u}, \mathbf{v} \in V$ and $\lambda \in \mathbb{R}$:

$$f(\mathbf{u}) + f(\mathbf{v}) = f(\mathbf{u}) + f(\mathbf{v})$$
$$f(\lambda \mathbf{u}) = \lambda f(\mathbf{u}).$$

It follows that the linear map f is completely determined by the images of the basis vectors $\mathbf{v}_j$ and hence by the coefficients a_{ij} that describe these images in terms of the basis vectors $\mathbf{w}_i$:

$$f(\mathbf{v}_j) = a_{1j} \mathbf{w}_1 + a_{2j} \mathbf{w}_2 + \ldots + a_{mj} \mathbf{w}_m \ .$$

So the map f is "encoded" by the $(m \times n)$-matrix $\mathbf{A} = (a_{ij})$: If $\mathbf{x}$ is the vector of coefficients of the vector $\mathbf{v} \in V$, then $\mathbf{y} = \mathbf{A}\mathbf{x}$ is the vector of coefficients of the image $f(\mathbf{v}) \in W$. In that sense, a linear map corresponds to a matrix (that also depends on our choice of bases for V and W!) Conversely, we can construct a

linear map f from a given $(m \times n)$-matrix $\mathbf{A}$ as follows. For every basis vector $\mathbf{v}_j$, we define

$$f(\mathbf{v}_j) = a_{1j}\mathbf{w}_1 + a_{2j}\mathbf{w}_2 + \ldots + a_{mj}\mathbf{w}_m \ .$$

Because an arbitrary vector $\mathbf{v} \in V$ can be uniquely expressed as a linear combination $\mathbf{v} = x_1\mathbf{v}_1 + \ldots + x_n\mathbf{v}_n$, we obtain the well-defined extension

$$f(\mathbf{v}) = x_1 f(\mathbf{v}_1) + \ldots + x_n f(\mathbf{v}_n) \ .$$

The composition of two (or more) linear maps corresponds to the product of their associated matrices. If $f : V \to W$ and $g : W \to Z$ are linear maps represented by $\mathbf{A} \in \mathbb{R}^{m\times n}$ resp. $\mathbf{B} \in \mathbb{R}^{k\times m}$ (with respect to fixed bases in V, W and Z), then also

$$g \circ f : V \to Z \ , \quad \text{where} \quad g \circ f(\mathbf{u}) = g(f(\mathbf{u})) \ ,$$

is a linear map. The matrix $\mathbf{C} = (c_{ij}) \in \mathbb{R}^{k\times n}$ describing $g \circ f$ is the *product* $\mathbf{C} = \mathbf{BA}$, *i.e.*, the elements c_{ij} are the inner products (*cf.* Section 1.3.1) of the rows of $\mathbf{B}$ with the columns of $\mathbf{A}$:

$$c_{ij} \ = \ \mathbf{B}_{i\cdot}\,\mathbf{A}_{\cdot j} \ = \ \sum_{l=1}^{m} b_{il}a_{lj} \ .$$

REMARK. The product matrix $\mathbf{C} = \mathbf{BA}$ admits two "dual" points of view:

$$\mathbf{C}_{\cdot j} = \mathbf{BA}_{\cdot j} = \sum_{l=1}^{m} a_{lj}\mathbf{B}_{\cdot l} \quad \text{and} \quad \mathbf{C}_{i\cdot} = \mathbf{B}_{i\cdot}\mathbf{A} = \sum_{l=1}^{m} b_{il}\mathbf{A}_{l\cdot} \ .$$

That is: The jth column $\mathbf{C}_{\cdot j}$ is the linear combination of the columns of $\mathbf{B}$ according to the coefficients of the jth column of $\mathbf{A}$. Dually, the ith row $\mathbf{C}_{i\cdot}$ is the linear combination of the rows of $\mathbf{A}$ according to the coefficients of the ith row of $\mathbf{B}$. Hence

$$\text{row } \mathbf{BA} \subseteq \text{row } \mathbf{A} \quad \text{and} \quad \text{col } \mathbf{BA} \subseteq \text{col } \mathbf{B} \ .$$

The Kernel. If $f : V \to W$ is a linear map and $L \subseteq V$ is a linear space, then $f(L) \subseteq W$ is a linear subspace of W. Similarly, if $L \subseteq W$ is a linear space of W, then $f^{-1}(L) = \{\mathbf{v} \in V \mid f(\mathbf{v}) \in L\}$ is a subspace of V. In particular, the *kernel* and the *image* of f, defined by

$$\begin{aligned} \ker f \ &= \ \{\mathbf{v} \in V \mid f(\mathbf{v}) = \mathbf{0}\} \subseteq V \ , \\ \text{im } f \ &= \ f(V) \subseteq W \end{aligned}$$

are subspaces of V resp. W. Their dimensions are related *via*

$$\dim \ker f + \dim \text{im } f = \dim V \ .$$

(To see this, take a basis of $\ker f$, extend it to a basis of V and verify that the extension is mapped to a basis of $\text{im } f$.) Note that we may usually (or "without loss of generality" – henceforth abbreviated: "*w.l.o.g.*") assume $\text{im } f = W$, *i.e.*, f is *surjective* (otherwise we simply replace W by $W' = \text{im } f$ in our reasoning).

Ex. 1.5. *Use the dimension formula to show that a linear map* $f : \mathbb{R}^n \to \mathbb{R}^n$ *is* surjective *(*i.e., $f(\mathbb{R}^n) = \mathbb{R}^n$*) if and only if it is* injective *(*i.e., $f(\mathbf{v}_1) = f(\mathbf{v}_2)$ *implies* $\mathbf{v}_1 = \mathbf{v}_2$*).*

If $f : \mathbb{R}^n \to \mathbb{R}^m$ is represented by $\mathbf{A} \in \mathbb{R}^{m\times n}$ with respect to the standard bases in $\mathbb{R}^n$ and $\mathbb{R}^n$, then

$$f(\mathbf{e}_j) = a_{1j}\mathbf{e}_1 + \ldots + a_{mj}\mathbf{e}_m = \mathbf{A}_{\cdot j} \, .$$

Hence the vector $\mathbf{x} = x_1\mathbf{e}_1 + \ldots + x_n\mathbf{e}_n \in \mathbb{R}^n$ is mapped to

$$f(\mathbf{x}) = x_1 f(\mathbf{e}_1) + \ldots + x_n f(\mathbf{e}_n) = x_1\mathbf{A}_{\cdot 1} + \ldots + x_n\mathbf{A}_{\cdot n} = \mathbf{A}\mathbf{x}$$

and we have $\operatorname{im} f = \operatorname{col} \mathbf{A}$. The kernel of $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ is also referred to as the *kernel* of $\mathbf{A}$, *i.e.*,

$$\ker \mathbf{A} \;=\; \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{0}\} \, .$$

Now suppose that $f(\mathbf{x}) = \mathbf{A}\mathbf{x}$ is surjective, *i.e.*, $\operatorname{col} \mathbf{A} = \mathbb{R}^m$. Then $\mathbf{A}$ must contain m columns, say $\mathbf{A}_{\cdot 1}, \ldots, \mathbf{A}_{\cdot m}$ that form a basis of $\mathbb{R}^m$. Let $g : \mathbb{R}^m \to \mathbb{R}^m$ be the (unique) linear map with the property $g(\mathbf{A}_{\cdot j}) = \mathbf{e}_j, \;\; j = 1, \ldots, m$. (Note that the first $\mathbf{e}_j$ denotes here the jth unit vector in $\mathbb{R}^n$, while the second $\mathbf{e}_j$ is the jth unit vector in $\mathbb{R}^m$.) So $g \circ f(\mathbf{e}_j) = \mathbf{e}_j$ holds for $j = 1, \ldots, m$. Therefore, if the matrix $\mathbf{B} \in \mathbb{R}^{m\times m}$ represents g, *i.e.*, if $g(\mathbf{y}) = \mathbf{B}\mathbf{y}$, we have

$$(\mathbf{B}\mathbf{A})\mathbf{e}_j = \mathbf{e}_j \qquad \text{for } j = 1, \ldots m$$

and conclude that $\mathbf{B}\mathbf{A} \in \mathbb{R}^{m\times n}$ must be of the form $\mathbf{B}\mathbf{A} = [\mathbf{I} \mid \mathbf{N}]$, where $\mathbf{I} \in \mathbb{R}^{m\times m}$ is the identity matrix and $\mathbf{N} \in \mathbb{R}^{m\times(n-m)}$.

Ex. 1.6. *Verify that* $\mathbf{B}\mathbf{A} = [\mathbf{I} \mid \mathbf{N}]$ *is the matrix that represents* f *with respect to the standard basis in* $\mathbb{R}^n$ *and the basis* $\{\mathbf{A}_{\cdot 1}, \ldots, \mathbf{A}_{\cdot m}\}$ *in* $\mathbb{R}^m$*.*

In the special case $m = n$, the composition $g \circ f = \mathrm{id}$ yields the identity map and $\mathbf{B}\mathbf{A} = \mathbf{I} \in \mathbb{R}^{n\times n}$. The matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$ is then the so-called *inverse* of $\mathbf{A} \in \mathbb{R}^{n\times n}$, denoted by $\mathbf{A}^{-1}$. So $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. Moreover, $f \circ g(\mathbf{A}_{\cdot j}) = f(\mathbf{e}_j) = \mathbf{A}_{\cdot j}$ implies that also $f \circ g = id$ and, therefore, $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ holds.

In the general case $m \le n$, we observe that $g : \mathbb{R}^m \to \mathbb{R}^m$ (defined by $g(\mathbf{A}_{\cdot j}) = \mathbf{e}_j$) has an inverse g^{-1} (defined by $g^{-1}(\mathbf{e}_j) = \mathbf{A}_{\cdot j}$). In particular, $\mathbf{B}^{-1} \in \mathbb{R}^{m\times m}$ exists and we have

$$\mathbf{A}\mathbf{x} = \mathbf{0} \;\Rightarrow\; \mathbf{B}\mathbf{A}\mathbf{x} = 0 \quad \text{and} \quad (\mathbf{B}\mathbf{A})\mathbf{x} = \mathbf{0} \;\Rightarrow\; \mathbf{A}\mathbf{x} = \mathbf{B}^{-1}(\mathbf{B}\mathbf{A})\mathbf{x} = \mathbf{0} \, ,$$

i.e., $\ker \mathbf{B}\mathbf{A} = \ker \mathbf{A}$.

REMARK (ELEMENTARY ROW OPERATIONS). By an *elementary row operation* on the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ we understand one of the operations:

- Addition of one row to another row;
- Multiplication of one row by a non-zero scalar;
- Interchange of two rows;

while keeping the other rows fixed. So elementary row operations consist of elementary operations on the row vectors of $\mathbf{A}$. Hence an elementary row operation can be described by multiplying $\mathbf{A}$ from the left with an matrix $\mathbf{B} \in \mathbb{R}^{m\times m}$ (see Ex. 1.7). Because an elementary row operation can be reversed by a similar operation, $\mathbf{B}$ is invertible. Moreover,

$$\text{row } \mathbf{A} = \text{row } \mathbf{B}^{-1}(\mathbf{BA}) \subseteq \text{row } \mathbf{BA} \subseteq \text{row } \mathbf{A}$$

shows row $\mathbf{BA}$ = row $\mathbf{A}$. In other words: the row space of $\mathbf{A}$ is invariant under elementary row operations.

Ex. 1.7. *Design the matrix* $\mathbf{B} \in \mathbb{R}^{5\times 5}$ *with the property that* $\mathbf{BA}$ *arises from* $\mathbf{A}$ *by subtracting* 17 *times the first row of* $\mathbf{A}$ *from the fourth row.*

Affine Maps. Restricting ourselves directly to $V = \mathbb{R}^n$ and $W = \mathbb{R}^m$ and their standard bases, we define an *affine map* $f : \mathbb{R}^n \to \mathbb{R}^m$ as a map of the form

$$f(\mathbf{x}) = \mathbf{Ax} - \mathbf{b} \quad \text{with } \mathbf{A} \in \mathbb{R}^{m\times n},\ \mathbf{b} \in \mathbb{R}^m\ .$$

Clearly, the image im f of an affine map is an affine subspace of $\mathbb{R}^m$ and, conversely, the *inverse image* $f^{-1}(L) = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) \in L\}$ of an affine subspace $L \subseteq \mathbb{R}^m$ is an affine subspace of $\mathbb{R}^n$. In particular

$$f^{-1}(\{\mathbf{0}\}) = \{\mathbf{x} \in \mathbb{R}^n \mid f(\mathbf{x}) = \mathbf{0}\}$$

is an affine subspace of $\mathbb{R}^n$. In case $f^{-1}(\{\mathbf{0}\})$ is non-empty, we may chose an arbitrary element $\mathbf{p} \in f^{-1}(\{\mathbf{0}\})$ and obtain the representation

$$f^{-1}(\{\mathbf{0}\}) = \mathbf{p} + \ker \mathbf{A}\ \ (= \{\mathbf{p} + \mathbf{x} \mid \mathbf{x} \in \ker \mathbf{A}\})\ .$$

If $n = m$ and $f(\mathbf{x}) = \mathbf{Ax} - \mathbf{b}$ is invertible (*i.e.*, $\mathbf{A}^{-1}$ exists), f is called an *affine transformation*.

1.3. Inner Products and Norms

1.3.1. Inner Products. An *inner product* on $\mathbb{R}^n$ is a map $\langle\cdot|\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ such that for all vectors $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and scalars $\lambda \in \mathbb{R}$,

$$\langle \mathbf{x}|\mathbf{y}\rangle = \langle \mathbf{y}|\mathbf{x}\rangle \tag{1.1}$$

$$\langle \lambda\mathbf{x}|\mathbf{y}\rangle = \lambda\langle \mathbf{x}|\mathbf{y}\rangle \tag{1.2}$$

$$\langle \mathbf{x}+\mathbf{y}|\mathbf{z}\rangle = \langle \mathbf{x}|\mathbf{z}\rangle + \langle \mathbf{y}|\mathbf{z}\rangle \tag{1.3}$$

$$\langle \mathbf{x}|\mathbf{x}\rangle > 0 \qquad \text{if } \mathbf{x} \neq \mathbf{0}. \tag{1.4}$$

By (1.1)-(1.3), an inner product is a *symmetric bilinear form* and, by (1.4), *positive definite*. We will usually be concerned with the so-called *standard inner product* for $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$, which is a special case of the matrix product:

$$\langle \mathbf{x}|\mathbf{y}\rangle = \mathbf{x}^T\mathbf{y} = \sum_{j=1}^{n} x_j y_j\ .$$

Ex. 1.8. *Show that* $\langle \mathbf{x}|\mathbf{y}\rangle = \mathbf{x}^T\mathbf{y}$ *defines an inner product on* $\mathbb{R}^n$.

It is also useful to consider the standard inner product in the vector space $\mathbb{R}^{m\times n}$ of $(m \times n)$-matrices. If we think of two matrices $\mathbf{A} = (a_{ij})$ and $\mathbf{B} = (b_{ij})$ in $\mathbb{R}^{m\times n}$ as two vectors of length mn in $\mathbb{R}^{mn}$, their inner product should be the sum of the componentwise products. We will denote this inner product by

$$\mathbf{A} \circ \mathbf{B} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ij} \ .$$

REMARK. Defining the *trace* of a matrix $\mathbf{C} = (c_{ij}) \in \mathbb{R}^{n\times n}$ as the sum of all diagonal elements,

$$\text{tr } \mathbf{C} = \sum_{i} c_{ii} \ ,$$

one obtains $\mathbf{A} \circ \mathbf{B} = \text{tr } (\mathbf{A}^T\mathbf{B})$ (see Ex. 1.9).

EX. 1.9. *Show:* $\mathbf{A} \circ \mathbf{B} = \text{tr } (\mathbf{A}^T\mathbf{B})$, $\mathbf{A} \circ \mathbf{B} = \mathbf{B} \circ \mathbf{A}$, $\mathbf{A} \circ \mathbf{B} = \mathbf{I} \circ (\mathbf{A}^T\mathbf{B})$. *Give an example of matrices* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n\times n}$ *such that* $\mathbf{AB} \neq \mathbf{BA}$.

From the fundamental properties of the inner product, we can derive the inequality of *Cauchy-Schwarz*:

LEMMA 1.1 (Cauchy-Schwarz)**.** *Let* $\langle\cdot|\cdot\rangle$ *be an inner product on* $\mathbb{R}^n$. *Then all vectors* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ *satisfy the inequality*

$$\langle \mathbf{x}|\mathbf{y}\rangle^2 \ \leq \ \langle \mathbf{x}|\mathbf{x}\rangle\langle \mathbf{y}|\mathbf{y}\rangle \ .$$

Equality holds if and only if $\mathbf{x}$ *is a scalar multiple of* $\mathbf{y}$.

Proof. We may assume $\langle \mathbf{x}|\mathbf{x}\rangle = \langle \mathbf{y}|\mathbf{y}\rangle = 1$ *w.l.o.g.* (otherwise we could scale, say, $\mathbf{x}$ with a nonzero scalar λ so that $\langle \lambda\mathbf{x}|\lambda\mathbf{x}\rangle = 1$, which would just multiply both sides of the inequality with λ^2.) We then find

$$\begin{aligned} 0 \ \leq \ \langle \mathbf{x}-\mathbf{y}|\mathbf{x}-\mathbf{y}\rangle \ &= \ \langle \mathbf{x}|\mathbf{x}\rangle - 2\langle \mathbf{x}|\mathbf{y}\rangle + \langle \mathbf{y}|\mathbf{y}\rangle \ = \ 2 - 2\langle \mathbf{x}|\mathbf{y}\rangle \\ 0 \ \leq \ \langle \mathbf{x}+\mathbf{y}|\mathbf{x}+\mathbf{y}\rangle \ &= \ \langle \mathbf{x}|\mathbf{x}\rangle + 2\langle \mathbf{x}|\mathbf{y}\rangle + \langle \mathbf{y}|\mathbf{y}\rangle \ = \ 2 + 2\langle \mathbf{x}|\mathbf{y}\rangle \ . \end{aligned}$$

So $|\langle \mathbf{x}|\mathbf{y}\rangle| \leq 1 = \langle \mathbf{x}|\mathbf{x}\rangle\langle \mathbf{y}|\mathbf{y}\rangle$. By property (1.4) of the inner product, equality can only hold if $\mathbf{x} - \mathbf{y} = \mathbf{0}$ or $\mathbf{x} + \mathbf{y} = \mathbf{0}$, *i.e.*, if $\mathbf{x} = \mathbf{y}$ or $\mathbf{x} = -\mathbf{y}$. The claim follows. ◇

Inner Products and Positive Definite Matrices. Let $\langle\cdot|\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ be an inner product on $\mathbb{R}^n$, and fix a basis $B = \{\mathbf{v}_1, \ldots, \mathbf{v}_n\}$ of $\mathbb{R}^n$. Relative to B, we can describe the inner product via the *Gram matrix* of inner products

$$\mathbf{G} = \mathbf{G}(\mathbf{v}_1, \ldots, \mathbf{v}_n) \ = \ (g_{ij}) \ = \ (\langle \mathbf{v}_i|\mathbf{v}_j\rangle) \ .$$

The symmetry of the inner product means that $\mathbf{G}$ is a *symmetric* matrix, *i.e.*, $\mathbf{G} = \mathbf{G}^T$ (or equivalently, $g_{ij} = g_{ji}$ for all i, j). If $\mathbf{x} = (x_1, \ldots, x_n)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$ are the n-tuples of scalars of the linear combinations $\mathbf{u} = x_1\mathbf{v}_1 +$

$\ldots + x_n\mathbf{v}_n$ and $\mathbf{w} = y_1\mathbf{v}_1 + \ldots + y_n\mathbf{v}_n$, then the bilinearity of the inner product yields

$$\langle \mathbf{u}|\mathbf{w}\rangle = \mathbf{x}^T\mathbf{G}\mathbf{y} = \sum_{i=1}^{n}\sum_{j=1}^{n} g_{ij}x_i y_j \ . \tag{1.5}$$

Moreover, if $\mathbf{x} \neq \mathbf{0}$, then $\mathbf{u} \neq \mathbf{0}$ and $\mathbf{x}^T\mathbf{G}\mathbf{x} = \langle \mathbf{u}|\mathbf{u}\rangle > 0$.

Conversely, let us call the symmetric matrix $\mathbf{Q} = (q_{ij}) \in \mathbb{R}^{n\times n}$ *positive definite* if $\mathbf{x}^T\mathbf{Q}\mathbf{x} > 0$ holds for all $\mathbf{x} \neq \mathbf{0}$. Then $\mathbf{Q}$ gives rise to an inner product $\langle \mathbf{x}|\mathbf{y}\rangle = \mathbf{x}^T\mathbf{Q}\mathbf{y}$ (with $\mathbf{Q}$ as its Gram matrix with respect to the standard basis in $\mathbb{R}^n$).

REMARK. Our discussion exhibits inner products of finite-dimensional vector spaces and positive definite matrices as manifestations of the same mathematical phenomenon: The positive definite matrices are exactly the Gram matrices of inner products. In particular, the standard inner product on $\mathbb{R}^n$ has the $(n \times n)$-identity matrix $\mathbf{I}$ as its Gram matrix relative to the standard basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$.

Inner Products and Orthogonal Matrices. Given an inner product $\langle\cdot|\cdot\rangle : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$, we say that the vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ are *orthogonal* if $\langle \mathbf{x}|\mathbf{y}\rangle = 0$. (See also Ex. 1.13.) A system of vectors $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ in $\mathbb{R}^n$ is called *orthonormal* provided

$$\langle \mathbf{v}_i|\mathbf{v}_j\rangle \ = \ \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Consider now $\mathbb{R}^n$ with respect to the standard inner product $\langle \mathbf{x}|\mathbf{y}\rangle = \mathbf{x}^T\mathbf{y}$. The matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is called *orthogonal* if $\mathbf{A}^T\mathbf{A} = \mathbf{I}$ (*i.e.*, if $\mathbf{A}^{-1} = \mathbf{A}^T$). This property means that the column vectors $\mathbf{A}_{\cdot 1}, \ldots, \mathbf{A}_{\cdot n}$ of $\mathbf{A}$ satisfy

$$\mathbf{A}_{\cdot i}^T\mathbf{A}_{\cdot j} = (\mathbf{A}\mathbf{e}_i)^T\mathbf{A}\mathbf{e}_j = \mathbf{e}_i^T\mathbf{e}_j = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j \ , \end{cases}$$

i.e., the columns $\mathbf{A}_{\cdot 1}, \ldots, \mathbf{A}_{\cdot n}$ (and then also the rows of $\mathbf{A}$) form an orthonormal basis of $\mathbb{R}^n$.

Ex. 1.10. *Show that $\mathbf{A} \in \mathbb{R}^{n\times n}$ is orthogonal if and only if $\mathbf{x}^T\mathbf{y} = (\mathbf{A}\mathbf{x})^T(\mathbf{A}\mathbf{y})$ holds for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Prove that the vectors of an orthonormal system are linearly independent.*

1.3.2. Norms. We can speak reasonably about the "length" of a vector in $\mathbb{R}^n$ only relative to a given *norm* on $\mathbb{R}^n$, that is, a map $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ such that for all vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and scalars $\lambda \in \mathbb{R}$,

$$\|\mathbf{x}\| > \mathbf{0} \qquad \text{for } \mathbf{x} \neq \mathbf{0}; \tag{1.6}$$

$$\|\lambda\mathbf{x}\| = |\lambda| \cdot \|\mathbf{x}\|; \tag{1.7}$$

$$\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|. \tag{1.8}$$

Inequality (1.8) is the *triangle inequality*.

Every inner product $\langle\cdot|\cdot\rangle$ on $\mathbb{R}^n$ gives rise to a norm (see Ex. 1.11) *via*

$$\|\mathbf{x}\| \;=\; \sqrt{\langle\mathbf{x}|\mathbf{x}\rangle}\;.$$

The norm arising from the standard inner product on $\mathbb{R}^n$ is the *Euclidean norm*

$$\|\mathbf{x}\| \;=\; \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{x_1^2+\ldots+x_n^2}\;.$$

REMARK. $\mathbb{R}^n$ admits also norms that do not arise from inner products (see Ex. 1.14). In case of ambiguity, the Euclidean norm is denoted by $\|\mathbf{x}\|_2$.

Ex. 1.11. *Let $\langle\cdot|\cdot\rangle$ be an inner product on $\mathbb{R}^n$.*

(a) *Use the inequality of Cauchy-Schwarz to show that $\|\mathbf{x}\| = \sqrt{\langle\mathbf{x}|\mathbf{x}\rangle}$ defines a norm.*
(b) *Show that the norm defined in* (a) *satisfies the so-called* parallelogram equality*:* $\|\mathbf{x}+\mathbf{y}\|^2+\|\mathbf{x}-\mathbf{y}\|^2 = 2\|\mathbf{x}\|^2+2\|\mathbf{y}\|^2$ *for all* $\mathbf{x},\mathbf{y}\in\mathbb{R}^n$.
(c) *Let $\|\cdot\|$ be a norm that satisfies the parallelogram equality. Show that* $\langle\mathbf{x}|\mathbf{y}\rangle = \frac{1}{4}(\|\mathbf{x}+\mathbf{y}\|^2-\|\mathbf{x}-\mathbf{y}\|^2)$ *defines an inner product.*

Hint for (c): *Verify the claim first for vectors with components in $\mathbb{Z}$ and $\mathbb{Q}$. Deduce then the statement for $\mathbb{R}$ from the continuity of the norm (cf. Section 1.4.2 below).*

Extending the Euclidean norm of vectors to matrices $\mathbf{A}=(a_{ij})\in\mathbb{R}^{m\times n}$, we obtain the *Frobenius norm*:

$$\|\mathbf{A}\|_F = \sqrt{\mathbf{A}\circ\mathbf{A}} = \|(a_{ij})\|_2\;. \tag{1.9}$$

Ex. 1.12. *Show for every $\mathbf{A}\in\mathbb{R}^{m\times n}$ and $\mathbf{x}\in\mathbb{R}^n$: $\|\mathbf{A}\mathbf{x}\|_2 \le \|\mathbf{A}\|_F\|\mathbf{x}\|_2$.*

Ex. 1.13 ("Theorem of Pythagoras")**.** *Let $\langle\cdot|\cdot\rangle$ be an inner product on $\mathbb{R}^n$ with associated norm $\|\mathbf{x}\| = \sqrt{\langle\mathbf{x}|\mathbf{x}\rangle}$. Say that the vector* $\mathbf{a}$ *is* perpendicular *to the vector* $\mathbf{b}$ *if the* distance $\|\mathbf{a}-\mathbf{b}\|$ *from* $\mathbf{a}$ *to* $\mathbf{b}$ *is the same as the distance* $\|\mathbf{a}-(-\mathbf{b})\|$ *from* $\mathbf{a}$ *to* $(-\mathbf{b})$*. Show:*

$$\|\mathbf{a}-\mathbf{b}\| = \|\mathbf{a}-(-\mathbf{b})\| \qu\Leftrightarrow\quad \langle\mathbf{a}|\mathbf{b}\rangle = 0 \qu\Leftrightarrow\quad \|\mathbf{a}\|^2+\|\mathbf{b}\|^2 = \|\mathbf{a}-\mathbf{b}\|^2\;. \tag{1.10}$$

Ex. 1.14. *Define for the vector $\mathbf{x}=(x_1,\ldots,x_n)^T\in\mathbb{R}^n$,*

$\|\mathbf{x}\|_1 \;=\; |x_1|+\ldots+|x_n|$ ("sum norm")
$\|\mathbf{x}\|_\infty \;=\; \max\limits_{1\le i\le n}|x_i|$ ("maximum norm").

Show that both $\|.\|_1$ and $\|.\|_\infty$ are norms but do not *arise from inner products. (Hint: use Ex. 1.11(b))*

1.4. Continuous and Differentiable Functions

1.4.1. Topology of $\mathbb{R}^n$. A norm on the vector space $\mathbb{R}^n$ allows us to measure the distance between vectors and, therefore, to specify "neighborhoods" *etc.* We will investigate these concepts with respect to the Euclidean norm and denote it simply $\|\cdot\| = \|\cdot\|_2$ (unless explicitly specified otherwise). For two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ we define their *(Euclidean) distance* as $\|\mathbf{x} - \mathbf{y}\|$. In the case $n = 1$, of course, we also use the familiar notation of the absolute value $|x - y|$. (For a general in-depth treatment of the analysis in $\mathbb{R}^n$ and further details we refer, *e.g.* to Rudin [69]).

The set of real numbers $\mathbb{R}$ has (by definition) the following so-called *completeness property* : A non-decreasing infinite sequence of real numbers $r_1 \le r_2 \le \dots$ has a (unique) limit

$$r = \lim_{k\to\infty} r_k \in \mathbb{R}$$

if and only if the exists a *bound* $M \in \mathbb{R}$ such that $r_k \le M$ holds for all k. As a consequence of this property (*cf.* Ex.1.15), every subset $S \subseteq \mathbb{R}$ has a unique *infimum*, which is defined as the largest lower bound for all $s \in S$ and denoted by $\inf S$. (If S is not bounded from below, then $\inf S = -\infty$ and if $S = \emptyset$, then $\inf S = +\infty$.) The *supremum* $\sup S$ is defined similarly.

Ex. 1.15. *Let $S \subseteq \mathbb{R}$ be non-empty and $s_1 \in \mathbb{R}$ be such that $s_1 \le s$ for all $s \in S$. Define a sequence $s_1 \le s_2 \le \dots$ of lower bounds for S as follows. Given s_k, we set*

$$s_{k+1} \;=\; \begin{cases} s_k + 1/k & \text{if } s_k + 1/k \;\le\; s \text{ for all } s \in S, \\ s_k & \text{otherwise.} \end{cases}$$

Show: $\bar{s} = \lim_{k\to\infty} s_k = \inf S$. (Hint: Use $\sum_{k=1}^{\infty} 1/k = \infty$.)

More generally, we say that a sequence $\mathbf{s}_1, \mathbf{s}_2, \dots$ of points $\mathbf{s}_k \in \mathbb{R}^n$ *converges* if the exists some $\mathbf{s} \in \mathbb{R}^n$ such that

$$\lim_{k\to\infty} \|\mathbf{s} - \mathbf{s}_k\| \;=\; 0 \;,$$

which is denoted by $\mathbf{s} = \lim_{k\to\infty} \mathbf{s}_k$ or just $\mathbf{s}_k \to \mathbf{s}$. A subset $S \subset \mathbb{R}^n$ is *bounded* if there exists some $M \in \mathbb{R}$ such that $\|\mathbf{s}\| \le M$ holds for all $\mathbf{s} \in S$. The completeness property of $\mathbb{R}$ then implies the following (*cf.* Ex. 1.16):

- $S \subset \mathbb{R}^n$ is bounded if and only if every infinite sequence $\mathbf{s}_1, \mathbf{s}_2, \dots$ of points $\mathbf{s}_k \in S$ admits a convergent subsequence $\mathbf{s}_{k_1}, \mathbf{s}_{k_2}, \dots$.

We refer to the limit $\bar{\mathbf{s}} = \lim_{i\to\infty} \mathbf{s}_{k_i}$ of a convergent subsequence $(\mathbf{s}_{k_i})$ as an *accumulation point* of S (or the sequence $(\mathbf{s}_k)$).

Ex. 1.16. *Show: $S \subset \mathbb{R}^n$ is bounded if and only if every infinite sequence $(\mathbf{s}_k)$ in S admits a convergent subsequence $(\mathbf{s}_{k_i})$.*
(Hint: *Assume that S is contained in an n-*dimensional cube $Q_0 = [-M, M]^n$. *Subdivide Q_0 into 2^n smaller cubes. One of these, say Q_1, contains infinitely many $\mathbf{s}_k$. Let $\mathbf{s}_{k_1}$ be the first $\mathbf{s}_k$ that is contained in Q_1 and proceed by subdividing Q_1.)*

An *open ball* (of radius $r > 0$ and centered at $\mathbf{x}_0 \in \mathbb{R}^n$) is a subset of the form

$$U_r(\mathbf{x}_0) \;=\; \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x} - \mathbf{x}_0\| < r\}\ .$$

A set $U \subseteq \mathbb{R}^n$ is *open* if U contains with any $\mathbf{x}_0$ also some open ball $U_r(\mathbf{x}_0)$. (In other words: An open set is precisely the union of all the open balls it contains.) A subset $C \subseteq \mathbb{R}^n$ is *closed* if $\mathbb{R}^n \backslash C$ is open.

(Arbitrary) unions of open sets are open and, correspondingly, intersections of closed sets are closed. In particular, every $S \subseteq \mathbb{R}^n$ has a unique smallest closed set containing S, namely the intersection cl S of all closed sets containing S. cl S is the so-called *closure* of S. Similarly, every $S \subseteq \mathbb{R}^n$ admits a unique maximal open set int S contained in S, namely the union of all open balls contained in S, the *interior* of S.

The *boundary* of $S \subseteq \mathbb{R}^n$ is defined as $\partial S = \text{cl } S \backslash \text{int } S$. Equivalently (*cf.* Ex. 1.17), ∂S consists of all points $\mathbf{x} \in \mathbb{R}^n$ that are accumulation points of both S and $\mathbb{R}^n \backslash S$.

Ex. 1.17. *Show: If* $\mathbf{x} \in \text{cl } S \backslash \text{int } S$ *then every* $U_{1/k}(\mathbf{x})$ $(k = 1, 2, \ldots)$ *intersects both* S *and* $\mathbb{R}^n \backslash S$. *Let* $\mathbf{s}_k \in U_{1/k}(\mathbf{x}) \cap S$ *and* $\mathbf{s}'_k \in U_{1/k}(\mathbf{x}) \backslash S$ *and observe* $\mathbf{s}_k \to \mathbf{x}$ *and* $\mathbf{s}'_k \to \mathbf{x}$.

Ex. 1.18. *Show:*
(i) Open intervals (a, b) *are open subsets of* $\mathbb{R}$.
(ii) An open line segment $(\mathbf{a}, \mathbf{b}) = \{(1 - \lambda)\mathbf{a} + \lambda\mathbf{b} \mid \lambda \in (0, 1)\}$ *(*$\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$*) is* not *open in* $\mathbb{R}^n$, $n \geq 2$.

REMARK (RELATIVE TOPOLOGY). The topological concepts of $\mathbb{R}^n$ carry over to subsets of $\mathbb{R}^n$. For our purposes, it suffices to consider affine subspaces $L \subseteq \mathbb{R}^n$. If dim $L = k$, it is occasionally more "natural" to think of L as an "isomorphic" copy of $\mathbb{R}^k$ and define the relevant notions accordingly ("relative to L").
A *relatively open ball* is a set of the form $U_r(\mathbf{x}_0) \cap L$, $\mathbf{x}_0 \in L$, and the *relative interior* of a subset $S \subset L$ is the union of all the relatively open balls S contains.

A subset $S \subset \mathbb{R}^n$ is *compact*, if it is bounded and closed. So S is compact if and only if every infinite sequence $(\mathbf{s}_k)$ in S has an accumulation point $\bar{\mathbf{s}} \in S$. (The existence of $\bar{\mathbf{s}}$ is equivalent to the boundedness of S and $\bar{\mathbf{s}} \in S$ is equivalent to the closedness of S.) This observation is sometimes referred to as the *Theorem of Bolzano-Weierstrass*.

Ex. 1.19. *Let* $S \subset \mathbb{R}$ *be compact. Show that inf S and sup S belong to S.*

1.4.2. Continuous Functions. Let S be a given subset of $\mathbb{R}^n$. Then the function $f : S \to \mathbb{R}^m$ is said to be *continuous* at the point $\mathbf{x}_0 \in S$ if for every $\varepsilon > 0$ there exists some $\delta > 0$ so that

$$\|f(\mathbf{x}_0) - f(\mathbf{x})\|_2 < \varepsilon \text{ holds whenever } \mathbf{x} \in S \text{ and } \|\mathbf{x}_0 - \mathbf{x}\|_2 < \delta,$$

or, equivalently, $f(U_\delta(\mathbf{x}_0) \cap S) \subseteq U_\varepsilon(f(\mathbf{x}_0))$. We denote this property by

$$\boxed{f(\mathbf{x}_0) = \lim_{\mathbf{x} \to \mathbf{x}_0} f(\mathbf{x})}$$

We say that f is *continuous on* S if f is continuous at every $\mathbf{x}_0 \in S$.

It is easily verified that sums and products of real-valued continuous functions are again continuous. Furthermore, compositions of continuous functions are continuous.

LEMMA 1.2. *Let $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ be an arbitrary norm on $\mathbb{R}^n$. Then $f(\mathbf{x}) = \|\mathbf{x}\|$ is a continuous function.*

Proof. We first consider $\mathbf{x}_0 = \mathbf{0}$ and let $M = n \cdot \max_j \|\mathbf{e}_j\|$ (where $\mathbf{e}_j$ is the jth unit vector). Then

$$\|\mathbf{x}\| = \|\sum_j x_j \mathbf{e}_j\| \le \sum_i |x_j| \cdot \|\mathbf{e}_j\| \le n(\max_j \|\mathbf{e}_j\|)(\max_j |x_j|) \le M\|\mathbf{x}\|_2 \ .$$

So $\|\mathbf{x}\|_2 \to 0$ implies $f(\mathbf{x}) \to 0 = f(\mathbf{0})$, i.e., f is continuous at $\mathbf{x}_0 = \mathbf{0}$. The continuity of f in general now follows from the observation that $\mathbf{x} \to \mathbf{x}_0$ is equivalent with $(\mathbf{x} - \mathbf{x}_0) \to \mathbf{0}$.

◇

If $C \subseteq \mathbb{R}^n$ is compact and $f : C \to \mathbb{R}^m$ is continuous, then the image $f(C)$ is compact in $\mathbb{R}^m$. For a proof of this fact, it suffices to show that every infinite sequence $f(\mathbf{z}_1), f(\mathbf{z}_2), \ldots$ of images has an accumulation point $f(\bar{\mathbf{z}}) \in f(C)$. Now C is compact. So the sequence $\mathbf{z}_1, \mathbf{z}_2, \ldots$ has a subsequence $\mathbf{z}_{k_1}, \mathbf{z}_{k_2}, \ldots$ and a point $\bar{\mathbf{z}} \in C$ such that $\mathbf{z}_{k_i} \to \bar{\mathbf{z}}$. The continuity of f guarantees $f(\mathbf{z}_{k_i}) \to f(\bar{\mathbf{z}})$, which establishes $f(\bar{\mathbf{z}})$ as an accumulation point in $f(C)$.

As a consequence, we obtain the following existence result which is used repeatedly in the analysis of optimization problems.

THEOREM 1.1. (Weierstrass) *Let $C \subset \mathbb{R}^n$ be a nonempty compact set and $f : C \to \mathbb{R}$ be continuous. Then f attains its maximum and minimum value on C,* i.e. *there exist $\mathbf{x}_0, \mathbf{x}_1 \in C$ with*

$$f(\mathbf{x}_0) \le f(\mathbf{x}) \le f(\mathbf{x}_1) \qquad \textit{for all } \mathbf{x} \in C.$$

Proof. If $f : C \to \mathbb{R}$ is continuous, then $F = f(C) \subset \mathbb{R}$ is compact. Hence $\inf F \in \mathbb{R}$ and $\sup F \in \mathbb{R}$ exist and belong to F. The Theorem follows.

◇

REMARK. The *unit spere* $S := \{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\|_2 = 1\}$ is a compact set. If $\|\cdot\|$ is any norm on $\mathbb{R}^n$, then in view of Lemma 1.2

$$\alpha := \min\{\|\mathbf{x}\| \mid \mathbf{x} \in S\} > 0 \quad \text{and} \quad \beta = \max\{\|\mathbf{x}\| \mid \mathbf{x} \in S\}$$

exist, showing that $\|\cdot\|$ is *equivalent* to the Euclidean norm $\|\cdot\|_2$ in the sense that

$$\alpha\|\mathbf{x}\|_2 \ \le \ \|\mathbf{x}\| \le \ \beta\|\mathbf{x}\|_2 \quad \text{for all } \mathbf{x} \in \mathbb{R}^n \ .$$

As a consequence, the definition of continuity (as well as the notion of differentiability below) does not depend on the particular norm we use.

1.4.3. Differentiable Functions. From a computational point of view, linear and affine functions are the easiest to deal with. Therefore, we are interested in the question when a not necessarily linear function can be, at least locally, approximated by a linear function. Again, we focus right directly on the vector spaces $\mathbb{R}^n$ and $\mathbb{R}^m$, equipped with the standard inner product and the Euclidean norm. In each case, we choose as reference the standard basis of unit vectors $\mathbf{e}_j = (\dots, 0, 1, 0, \dots)^T$.

Let U be an open subset in $\mathbb{R}^n$. The function $f : U \to \mathbb{R}^m$ is said to be *differentiable* at the point $\mathbf{x}_0 \in U$ if there exists a matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ and a function $\varphi : U \to \mathbb{R}^m$ such that $\lim_{\mathbf{h}\to\mathbf{0}} \varphi(\mathbf{h}) = \mathbf{0}$ and

$$f(\mathbf{x}_0 + \mathbf{h}) \;=\; f(\mathbf{x}_0) + \mathbf{A}\mathbf{h} + \|\mathbf{h}\|\varphi(\mathbf{h}) \quad \text{for all } \mathbf{x}_0 + \mathbf{h} \in U \ .$$

A shorter way of expressing these conditions is offered by the notation

$$\boxed{f(\mathbf{x}_0 + \mathbf{h}) = f(\mathbf{x}_0) + \mathbf{A}\mathbf{h} + o(\|\mathbf{h}\|)}$$

(Recall that $o(\|\mathbf{h}\|^k)$ generally denotes a term of the form $\|\mathbf{h}\|^k\eta(\mathbf{h})$, where $\eta(\mathbf{h})$ is a function satisfying $\lim_{\mathbf{h}\to\mathbf{0}} \eta(\mathbf{h}) = \mathbf{0}$.)

The definition says that the differentiable function f can be approximated near $\mathbf{x}_0$ *via* the affine function

$$\tilde{f}(\mathbf{h}) = f(\mathbf{x}_0) + \mathbf{A}\mathbf{h} \ .$$

The matrix $\mathbf{A}$ is called the *derivative* of f at $\mathbf{x}_0$ and is generally denoted by $\nabla f(\mathbf{x}_0)$ (and by $f'(x_0)$ in case $n = 1$). We call f *differentiable on* U if f is differentiable at every $\mathbf{x}_0 \in U$. The derivative $\nabla f(\mathbf{x}_0)$ turns out to be nothing but the *Jacobian* matrix associated with f (see p. 17).

EX. 1.20. *Show that $f : U \to \mathbb{R}^m$ is necessarily continuous at $\mathbf{x}_0 \in U$ provided f is differentiable at $\mathbf{x}_0$.*

Derivatives of Functions in One Variable. The analysis of functions of several variables can often be reduced to the one-dimensional case. Therefore, we briefly review some basic and important facts for functions f in one real variable x (with f' denoting the derivative).

Let f be defined on the open interval $(a, b) \subseteq \mathbb{R}$ and differentiable at the point $x_0 \in (a, b)$. The following is a key observation for many optimization problems.

LEMMA 1.3. *If $f'(x_0) > 0$, then there exists some $\delta > 0$ such that*

$$f(x_0 - h) < f(x_0) < f(x_0 + h) \qquad \textit{whenever } 0 < h < \delta \ .$$

Proof. Because $f'(x_0)$ exists, we can write

$$f(x_0 + h) = f(x_0) + hf'(x_0) + |h|\varphi(h)$$

with $\lim_{h\to 0} \varphi(h) = 0$. Hence, if $f'(x_0) > 0$, there exists some $\delta > 0$ such that

$$|\varphi(h)| \;<\; f'(x_0) \quad \text{whenever } |h| \le \delta \ ,$$

which implies the Lemma.

◇

An immediate consequence of Lemma 1.3 is the *extremum principle*: If $f : (a,b) \to \mathbb{R}$ is differentiable at $x_0 \in (a,b)$ and if either $f(x_0) = \max_{x\in(a,b)} f(x)$ or $f(x_0) = \min_{x\in(a,b)} f(x)$, then

$$\boxed{f'(x_0) \;=\; 0}$$

.

REMARK. The extremum principle exhibits the *critical equation* $f'(x) = 0$ as a necessary (and very useful) *optimality condition* for $f(x)$: When one searches for a maximizer or a minimizer of $f(x)$, one may restrict one's attention to the so-called *critical points* x_0 that satisfy the critical equation.

THEOREM 1.2. (Mean Value Theorem) *Let $f : [a,b] \to \mathbb{R}$ be continuous on the closed interval $[a,b]$ and differentiable on the open interval (a,b). Then there exists some $\xi \in (a,b)$ such that*

$$f(b) - f(a) = (b-a)\, f'(\xi) \ .$$

Proof. Define $g(x) = (b-a)\, f(x) - [f(b) - f(a)]x$ and observe $g(a) = g(b)$. If g is constant on $[a,b]$, then every $\xi \in (a,b)$ has the claimed property

$$0 = g'(\xi) = (b-a)\, f'(\xi) - f(b) + f(a) \ .$$

Just like f, also g is differentiable on (a,b) and continuous on $[a,b]$. Hence, if g is not constant on the compact set $[a,b] \subseteq \mathbb{R}$, there exists some $\xi \in (a,b)$ such that

$$g(\xi) = \max_{x\in(a,b)} g(x) \qquad \text{or} \qquad g(\xi) = \min_{x\in(a,b)} g(x) \ .$$

In either case, the extremum principle yields $g'(\xi) = 0$ and the Theorem follows as before.

◇

EX. 1.21. *Show:* $1 + x \le e^x$ *for* $x \in \mathbb{R}$.

As an application of the mean Value Theorem we obtain the second order *Taylor formula*.

LEMMA 1.4. *Let $U = (-t_0, t_0) \subseteq \mathbb{R}$ and assume $p : U \to \mathbb{R}$ is twice differentiable. Then, for any given $t \in U$ there exists $0 < \theta < 1$, such that*

$$p(t) = p(0) + tp'(0) + \frac{1}{2}t^2 p''(\theta t) \ .$$

Proof. Assume *w.l.o.g.* that $t > 0$ and consider, for $\alpha \in \mathbb{R}$ (to be defined below) the twice differentiable function $g : [0,t] \to \mathbb{R}$ defined by

$$g(\tau) := p(\tau) - [p(0) + \tau p'(0) + \frac{1}{2}\tau^2 \alpha] \ .$$

Then $g(0) = 0,\ g'(0) = 0$ and α can be chosen so that $g(t) = 0$. We are left to show that this choice yields $\alpha = p''(\theta t)$ for some $0 < \theta < 1$.
Since $g(0) = g(t) = 0$, the Mean Value Theorem yields some $t_1 \in (0, t)$ with $g'(t_1) = 0$. Applying the Mean Value Theorem to g' with $g'(0) = g'(t_1) = 0$, we find some $t_2 \in (0, t_1)$ such that $g''(t_2) = 0$. Hence

$$0 = g''(t_2) = p''(t_2) - \alpha$$

and the claim follows with $\theta t = t_2$.

◇

Directional Derivatives and the Gradient. A function $f : U \to \mathbb{R}^m$ assigns to each vector $\mathbf{x} \in U$ a vector

$$f(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))^T \in \mathbb{R}^m .$$

It follows from the definition that f is differentiable at $\mathbf{x}_0$ if and only if each of the real-valued component functions $f_i : U \to \mathbb{R}$ is differentiable at $\mathbf{x}_0$. So we may restrict our attention to the case $m = 1$.

Consider the differentiable real-valued function $f : U \to \mathbb{R}$ at the point $\mathbf{x}_0 \in U$ and fix a "direction" $\mathbf{d} \in \mathbb{R}^n$. Then f induces locally a function

$$p_{\mathbf{d}}(t) = f(\mathbf{x}_0 + t\mathbf{d})$$

of the real parameter t. Moreover, the representation

$$f(\mathbf{x}_0 + t\mathbf{d}) = f(\mathbf{x}_0) + t\nabla f(\mathbf{x}_0)\mathbf{d} + \|t\mathbf{d}\|\varphi(t\mathbf{d})$$

immediately shows that the derivative $p'_{\mathbf{d}}(0)$ exists. In fact, letting $\tilde{\varphi}(t) = \|\mathbf{d}\|\varphi(t\mathbf{d})$, we have

$$p_{\mathbf{d}}(t) = p_{\mathbf{d}}(0) + t\nabla f(\mathbf{x}_0)\mathbf{d} + |t|\tilde{\varphi}(t)$$

and hence $p'_{\mathbf{d}}(0) = \nabla f(\mathbf{x}_0)\mathbf{d}$.

Choosing the direction $\mathbf{d}$ as a unit vector $\mathbf{e}_j$, we recognize what the derivative matrix $\nabla f(\mathbf{x}_0) = (a_1, \ldots, a_n) \in \mathbb{R}^{1\times n}$ actually is and how it can be computed. Recall first that the *partial derivative* $\partial f/\partial x_j$ of f at $\mathbf{x}_0$ is defined to be the derivative $p'_{\mathbf{e}_j}(0)$ of the function $p_{\mathbf{e}_j}(t) = f(\mathbf{x}_0 + t\mathbf{e}_j)$, which implies

$$a_j = \nabla f(\mathbf{x}_0)\mathbf{e}_j = \frac{\partial f(\mathbf{x}_0)}{\partial x_j} .$$

Hence, the derivative of $f : U \to \mathbb{R}$ is given by

$$\nabla f(\mathbf{x}_0) = \left[\frac{\partial f(\mathbf{x}_0)}{\partial x_1}, \ldots, \frac{\partial f(\mathbf{x}_0)}{\partial x_n}\right]$$

and is called the *gradient* of f at $\mathbf{x}_0$.

Ex. 1.22. *Let $f : \mathbb{R}^2 \to \mathbb{R}$ be given by $f(x, y) = x \cdot y$. Show $\nabla f(x, y) = (y, x)$.*

In general, we note for $p_{\mathbf{d}}(t) = f(\mathbf{x}_0 + t\mathbf{d})$ the formula for the *directional derivative of* f *at* $\mathbf{x}_0$ *with respect to* $\mathbf{d}$:

$$\boxed{p'_{\mathbf{d}}(0) = \nabla f(\mathbf{x}_0)\mathbf{d} = \sum_{j=1}^{n} \frac{\partial f(\mathbf{x}_0)}{\partial x_j} d_j} \tag{1.11}$$

Formula (1.11) allows us to determine the direction with respect to which f offers the largest marginal change at $\mathbf{x}_0$. With $p_{\mathbf{d}}(t)$ defined as before, we want to solve the optimization problem

$$\max_{\mathbf{d}\in\mathbb{R}^n} |p'_{\mathbf{d}}(0)| \quad \text{subject to } \|\mathbf{d}\| = 1.$$

We assume $\nabla f(\mathbf{x}_0) \neq \mathbf{0}^T$. Applying the Cauchy-Schwarz inequality to (1.11), we deduce

$$|p'_{\mathbf{d}}(0)| \;\leq\; \|\nabla f(\mathbf{x}_0)\| \cdot \|\mathbf{d}\| = \|\nabla f(\mathbf{x}_0)\|$$

with equality if and only if $\mathbf{d}^T = \lambda \nabla f(\mathbf{x}_0)$ for some $\lambda \in \mathbb{R}$. Hence we find that the gradient $\nabla f(\mathbf{x}_0)$ yields the direction $\mathbf{d}$ into which f exhibits the largest marginal change at $\mathbf{x}_0$. Depending on the sign of λ, we obtain the direction of largest increase or largest decrease.

Moreover, we have the general *extremum principle*: If $\mathbf{x}_0 \in U$ is a *local minimizer* or *maximizer* of f in the sense that for some $\varepsilon > 0$

$$f(\mathbf{x}_0) = \max_{\mathbf{x}\in U_\varepsilon(\mathbf{x}_0)} f(\mathbf{x}) \qquad \text{or} \qquad f(\mathbf{x}_0) = \min_{\mathbf{x}\in U_\varepsilon(\mathbf{x}_0)} f(\mathbf{x}) \,,$$

the one-dimensional extremum principle says that $0 = p'_{\mathbf{d}}(0) = \nabla f(\mathbf{x}_0)\mathbf{d}$ must hold for all directions $\mathbf{d}$, which implies that $\mathbf{x}_0$ must be a *critical point*, *i.e.,* satisfy the *critical equation*

$$\boxed{\nabla f(\mathbf{x}_0) = \mathbf{0}^T} \tag{1.12}$$

For general $f : U \to \mathbb{R}^m$, the same reasoning as in the case $m = 1$ shows that the derivative $\nabla f(\mathbf{x}_0)$ has as rows exactly the gradients of the component functions f_i of f. Hence the derivative $\nabla f(\mathbf{x}_0)$ of f at $\mathbf{x}_0$ is the *Jacobian* matrix

$$\nabla f(\mathbf{x}_0) \;=\; \left(\frac{\partial f_i(\mathbf{x}_0)}{\partial x_j}\right) \;\in \mathbb{R}^{m\times n} \,.$$

Ex. 1.23. *Show that an affine function* $f(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b}$ *has the derivative* $\nabla f(\mathbf{x}) = \mathbf{A}$ *at every* $\mathbf{x}$*. In particular, a linear function* $g(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ *has gradient* $\nabla g(\mathbf{x}) = \mathbf{c}^T$.
Let $\mathbf{Q}$ *be a symmetric matrix and consider the* quadratic function $q(\mathbf{x}) = \mathbf{x}^T Q\mathbf{x}$*. Show that* $\nabla q(\mathbf{x}) = 2\mathbf{x}^T\mathbf{Q}$ *holds for all* $\mathbf{x}$*.*

The existence of the Jacobian $\nabla f(\mathbf{x})$ (*i.e.*, the existence of the partial derivatives $\partial f_i(\mathbf{x})/\partial x_j$) alone does not necessarily guarantee the differentiability of f at $\mathbf{x}$. A sufficient – and in practice quite useful – condition is the continuity of the (generally non-linear) map $\mathbf{x} \mapsto \nabla f(\mathbf{x})$.

LEMMA 1.5. *Let $U \subseteq \mathbb{R}^n$ be open and $f : U \to \mathbb{R}^m$ have continuous partial derivative functions $\mathbf{x} \mapsto \partial f_i(\mathbf{x})/\partial x_j$ for all $i = 1 \ldots, m$ and $j = 1, \ldots, n$ (i.e., the function $\mathbf{x} \mapsto \nabla f(\mathbf{x})$ exists and is continuous). Then f is differentiable on U.*

Proof. As noted above, it suffices to assume $m = 1$. Given $\mathbf{x}_0 \in U$ and the vector $\mathbf{d} = (d_1, \ldots, d_n)^T \in \mathbb{R}^n$, we let

$$\mathbf{x}_k = \mathbf{x}_0 + d_1\mathbf{e}_1 + \ldots + d_k\mathbf{e}_k \quad \text{for } k = 1, \ldots, n \ .$$

For $\|\mathbf{d}\|$ sufficiently small, we have $\mathbf{x}_k \in U$ for all $k = 1, \ldots, n$ and

$$f(\mathbf{x}_0 + \mathbf{d}) - f(\mathbf{x}_0) = \sum_{k=1}^{n} [\, f(\mathbf{x}_k) - f(\mathbf{x}_{k-1})] \, .$$

Applying the Mean Value Theorem to the functions $\xi \mapsto f(\mathbf{x}_{k-1} + \xi\mathbf{e}_k)$, we obtain numbers $\xi_k \in (0, d_k)$ such that

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k-1}) = d_k \frac{\partial f(\mathbf{x}_{k-1} + \xi_k \mathbf{e}_k)}{\partial x_k} \ ,$$

whence we deduce

$$\begin{aligned} \|\mathbf{d}\|\varphi(\mathbf{d}) &= f(\mathbf{x}_0 + \mathbf{d}) - f(\mathbf{x}_0) - \nabla f(\mathbf{x}_0)\mathbf{d} \\ &= \sum_{k=1}^{n} d_k \left[\frac{\partial f(\mathbf{x}_{k-1} + \xi_k \mathbf{e}_k)}{\partial x_k} - \frac{\partial f(\mathbf{x}_0)}{\partial x_k} \right] . \end{aligned}$$

Because $\|\mathbf{d}\|^{-1} \cdot |d_k| \leq 1$ and $\lim_{\mathbf{d}\to\mathbf{0}} (\mathbf{x}_{k-1} + \xi_k \mathbf{e}_k) = \mathbf{x}_0$, continuity of the partial derivatives finally implies

$$\lim_{\mathbf{d}\to\mathbf{0}} \|\varphi(\mathbf{d})\| \ \leq \ \lim_{\mathbf{d}\to\mathbf{0}} \sum_{k=1}^{n} \left| \frac{\partial f(\mathbf{x}_{k-1} + \xi_k \mathbf{e}_k)}{\partial x_k} - \frac{\partial f(\mathbf{x}_0)}{\partial x_k} \right| \ = \ 0 \ .$$

◇

If $f : U \to \mathbb{R}^m$ has continuous partial derivatives, we call f a C^1-*function*. In general, C^k denotes the class of functions with continuous partial derivatives up to order k (*cf.* Section1.4.4 below).

The Chain Rule. Let $U \subseteq \mathbb{R}^n$ and $S \subseteq \mathbb{R}^m$ be open sets and assume that the function $f : U \to S$ is differentiable at $\mathbf{x}_0 \in U$. If $g : S \to \mathbb{R}^k$ is differentiable at $\mathbf{y}_0 = f(\mathbf{x}_0)$, the composite function $h : U \to \mathbb{R}^k$, given by $h(\mathbf{x}) = g(f(\mathbf{x}))$, can be linearly approximated at $\mathbf{x}_0$ by the composition of the respective derivatives.

More precisely, $h = g \circ f$ is differentiable at $\mathbf{x}_0$ and the Jacobian of h at $\mathbf{x}_0$ equals the matrix product of the Jacobian matrices of f at $\mathbf{x}_0$ and g at $\mathbf{y}_0$:

$$\boxed{\nabla h(\mathbf{x}_0) = \nabla g(\mathbf{y}_0) \nabla f(\mathbf{x}_0)} \tag{1.13}$$

Formula (1.13) is known as the *chain rule* for differentiable functions. To verify it, it suffices to assume $k = 1$. Writing $\mathbf{A} = \nabla f(\mathbf{x}_0)$ and $\mathbf{B} = \nabla g(\mathbf{y}_0)$ for ease of

notation, the differentiability of f and g provides us with functions φ and ψ such that $\varphi(\mathbf{d}) \to \mathbf{0}$ as $\mathbf{d} \to \mathbf{0}$ and $\psi(\tilde{\mathbf{d}}) \to 0$ as $\tilde{\mathbf{d}} \to \mathbf{0}$ and

$$f(\mathbf{x}_0 + \mathbf{d}) = f(\mathbf{x}_0) + \mathbf{A}\mathbf{d} + \|\mathbf{d}\|\varphi(\mathbf{d})$$
$$g(\mathbf{y}_0 + \tilde{\mathbf{d}}) = g(\mathbf{y}_0) + \mathbf{B}\tilde{\mathbf{d}} + \|\tilde{\mathbf{d}}\|\psi(\tilde{\mathbf{d}}).$$

With $\tilde{\mathbf{d}} = \mathbf{A}\mathbf{d} + \|\mathbf{d}\|\varphi(\mathbf{d})$, we then obtain for all $\mathbf{d} \neq \mathbf{0}$,

$$\begin{aligned} h(\mathbf{x}_0 + \mathbf{d}) &= g(f(\mathbf{x}_0) + \tilde{\mathbf{d}}) \\ &= g(f(\mathbf{x}_0)) + \mathbf{B}\tilde{\mathbf{d}} + \|\tilde{\mathbf{d}}\|\psi(\tilde{\mathbf{d}}) \\ &= g(f(\mathbf{x}_0)) + \mathbf{B}\mathbf{A}\mathbf{d} + \mathbf{B}\|\mathbf{d}\|\varphi(\mathbf{d}) + \|\tilde{\mathbf{d}}\|\psi(\tilde{\mathbf{d}}) \\ &= h(\mathbf{x}_0) + \mathbf{B}\mathbf{A}\mathbf{d} + \|\mathbf{d}\|\left(\mathbf{B}\varphi(\mathbf{d}) + \frac{\|\tilde{\mathbf{d}}\|}{\|\mathbf{d}\|}\psi(\tilde{\mathbf{d}})\right). \end{aligned}$$

By the choice of φ and the continuity of the linear maps $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$, and $\mathbf{y} \mapsto \mathbf{B}\mathbf{y}$, we have both $\mathbf{B}\varphi(\mathbf{d}) \to 0$ and $\psi(\tilde{\mathbf{d}}) \to \mathbf{0}$ as $\mathbf{d} \to \mathbf{0}$.

In order to establish the differentiability of $h = g \circ f$ and the chain rule, it now suffices to show that the quotient $\|\tilde{\mathbf{d}}\|/\|\mathbf{d}\|$ is bounded as $\mathbf{d} \to \mathbf{0}$. From Ex. 1.12, however, we know $\|\mathbf{A}\mathbf{d}\| \leq \|\mathbf{A}\|_F\|\mathbf{d}\|$. In view of the triangle inequality (1.8), we therefore conclude

$$\|\tilde{\mathbf{d}}\|/\|\mathbf{d}\| \leq \|\mathbf{A}\|_F + \|\varphi(\mathbf{d})\| \ ,$$

which is bounded as $\mathbf{d} \to \mathbf{0}$.

The Product Rule. The chain rule is an often very useful tool for the computation of derivatives. Let, for example, $f_1,\ f_2 : (a, b) \to \mathbb{R}$ be differentiable on the open interval $(a, b) \subseteq \mathbb{R}$ and consider $h(t) = f_1(t) \cdot f_2(t)$.

With $F(t) = (f_1(t),\ f_2(t))^T$ and $H(x, y) = xy$, we have $h(t) = H(F(t))$. So

$$\nabla F(t) = \begin{pmatrix} f_1'(t) \\ f_2'(t) \end{pmatrix} \qquad \text{and} \qquad \nabla H(x, y) = (y, x)$$

yield

$$h'(t) = \nabla H(f_1(t),\ f_2(t))\ \nabla F(t) = (f_2(t),\ f_1(t)) \begin{pmatrix} f_1'(t) \\ f_2'(t) \end{pmatrix}$$

i.e.

$$\boxed{h'(t) = f_2(t)\, f_1'(t) + f_1(t)\, f_2'(t)}$$

Ex. 1.24. *Let $f_1,\ f_2 : (a, b) \to \mathbb{R}$ be differentiable and assume $f_2(t) \neq 0$ for all $t \in (a, b)$. Derive for $h(t) = f_1(t)/f_2(t)$ the* quotient rule*:*

$$\boxed{h'(t) = \frac{f_2(t)\, f_1'(t) - f_1(t)\, f_2'(t)}{f_2^2(t)}}$$

1.4.4. Second Derivatives and Taylor's Formula. The differentiable function $f : U \to \mathbb{R}$ gives rise to the function $\nabla f : U \to \mathbb{R}^n$ via the assignment $\mathbf{x} \mapsto [\nabla f(\mathbf{x})]^T$. Let us assume that also $\nabla f(\mathbf{x})$ is differentiable. Then the partial derivatives of the partial derivatives of f exist and define the second derivative matrix

$$\nabla^2 f(\mathbf{x}) = \left(\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right) ,$$

called the *Hessian* matrix of f at $\mathbf{x} \in U$.

REMARK. If all second partial derivatives are continuous on U, *i.e.,* f is a C^2-function, one can show for all $\mathbf{x} \in U$ and all i, j,

$$\frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} = \frac{\partial^2 f(\mathbf{x})}{\partial x_j \partial x_i} ,$$

which means that the Hessian $\nabla^2 f(\mathbf{x})$ is symmetric.

Ex. 1.25. *Let $\nabla f : U \to \mathbb{R}^n$ be differentiable. Show that the function $p_{\mathbf{u}}(t) = f(\mathbf{x}_0 + t\mathbf{u})$ is twice differentiable at t_0, if $\mathbf{x}_0 + t_0\mathbf{u} \in U$, and satisfies*

$$\boxed{p''_{\mathbf{u}}(t_0) = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\partial^2 f(\mathbf{x}_0 + t_0\mathbf{u})}{\partial x_i \partial x_j} u_i u_j}$$

Consider the case where all second partial derivatives of f exist and are continuous. Then Lemma 1.5 tells us that ∇f is differentiable. By Ex. 1.25, we know that $p_{\mathbf{u}}(t) = f(\mathbf{x}_0 + t\mathbf{u})$ is twice differentiable.

In the subsequent discussion, we consider vectors $\mathbf{u}$ of unit length $\|\mathbf{u}\| = 1$. Lemma 1.4 guarantees the existence of some $0 < \theta_{\mathbf{u}} < 1$, such that

$$p_{\mathbf{u}}(t) = p_{\mathbf{u}}(0) + p'_{\mathbf{u}}(0)t + \frac{t^2}{2} p''_{\mathbf{u}}(\theta_{\mathbf{u}} t) ,$$

provided $|t| > 0$ is so small that $U_{|t|}(\mathbf{x}_0) \subseteq U$. We want to derive an analogous representation for f.

Given $\varepsilon > 0$, the assumed continuity of the Hessian matrix $\nabla^2 f(\mathbf{x})$ allows us to choose $|t| > 0$ so small that for every $\mathbf{x} \in U$ with $\|\mathbf{x}_0 - \mathbf{x}\| < |t|$,

$$\left| \frac{\partial^2 f(\mathbf{x}_0)}{\partial x_i \partial x_j} - \frac{\partial^2 f(\mathbf{x})}{\partial x_i \partial x_j} \right| < \varepsilon .$$

Recalling $p''_{\mathbf{u}}(t) = \mathbf{u}^T \nabla^2 f(\mathbf{x}_0 + t\mathbf{u})\mathbf{u}$ and observing $|u_i u_j| \leq \|\mathbf{u}\|^2 = 1$ for every two components u_i and u_j of $\mathbf{u}$, we obtain

$$|p''_{\mathbf{u}}(0) - p''_{\mathbf{u}}(\theta_{\mathbf{u}} t)| \leq n^2 \varepsilon ,$$

which is valid for all $\mathbf{d} = t\mathbf{u}$ whenever the norm $\|\mathbf{d}\| = |t|$ is small enough (independent of the unit direction $\mathbf{u}$!). With $p'_{\mathbf{u}}(0) = \nabla f(\mathbf{x}_0)\mathbf{u}$, we thus arrive at *Taylor's formula* for real-valued functions in several variables:

$$\boxed{f(\mathbf{x}_0+\mathbf{d}) \;=\; f(\mathbf{x}_0)+\nabla f(\mathbf{x}_0)\mathbf{d}+\tfrac{1}{2}\mathbf{d}^T\nabla^2 f(\mathbf{x}_0)\mathbf{d}+o(\|\mathbf{d}\|^2)}$$

or with some $\tau \in (0,1)$:

$$\boxed{f(\mathbf{x}_0+\mathbf{d}) \;=\; f(\mathbf{x}_0)+\nabla f(\mathbf{x}_0)\mathbf{d}+\tfrac{1}{2}\mathbf{d}^T\nabla^2 f(\mathbf{x}_0+\tau\mathbf{d})\mathbf{d}}$$

CHAPTER 2

Linear Equations and Linear Inequalities

While Chapter 1 reviews general structural aspects of real vector spaces, we now discuss fundamental *computational techniques for linear systems* in this chapter. For convenience of the discussion, we generally assume that the coefficients of the linear systems are *real* numbers. It is important to note, however, that in practical computations we mostly deal with *rational* numbers as input parameters. We therefore point out right at the outset that *all* the algorithms of this chapter (Gaussian elimination, orthogonal projection, Fourier-Motzkin elimination) will compute rational output quantities if the input parameters are rational numbers, as the reader can easily check.

2.1. Gaussian Elimination

Let $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{m\times n}$ be an $(m \times n)$-matrix and $\mathbf{b} = (b_i) \in \mathbb{R}^m$ a vector. Can we represent $\mathbf{b}$ as a linear combination of the column vectors of $\mathbf{A}$? And if yes, how? To answer this question, we must find a vector $\mathbf{x} = (x_j) \in \mathbb{R}^n$ such that $\mathbf{Ax} = \mathbf{b}$ holds, *i.e.*, such that

$$\begin{array}{rcrcccrcl} a_{11}x_1 & + & a_{12}x_2 & + & \dots & + & a_{1n}x_n & = & b_1 \\ a_{21}x_1 & + & a_{22}x_2 & + & \dots & + & a_{2n}x_n & = & b_2 \\ \vdots & & & & & & & \vdots & \\ a_{m1}x_1 & + & a_{m2}x_2 & + & \dots & + & a_{mn}x_n & = & b_m \end{array} \tag{2.1}$$

We refer to (2.1) as a *system of linear equations* in variables $x_1, \dots, x_n$. A vector $\mathbf{x} = (x_j) \in \mathbb{R}^n$ satisfying (2.1) is called *feasible* or a *solution* for (2.1). The system is *infeasible* if no solution exists.

From a structural point of view, our problem is the following. We are given the linear function $f : \mathbb{R}^n \to \mathbb{R}^m$ via $f(\mathbf{x}) = \mathbf{Ax}$, and a vector $\mathbf{b} \in \mathbb{R}^m$. We are to determine a vector $\mathbf{x}$ in the "solution space" of $\mathbf{Ax} = \mathbf{b}$, *i.e.*, in the affine subspace (*cf.* Section 1.2)

$$S = f^{-1}(\{\mathbf{b}\}) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} = \mathbf{b}\} \subseteq \mathbb{R}^n \ .$$

In the computational approach to the problem, we try to transform the system (2.1) of linear equalities via elementary vector space operations that leave the solution space S unaltered until the system has attained an equivalent form from which a solution can be easily inferred.

If all coefficients occurring with the variable x_1 are zero, the column is irrelevant for any linear combination and we are reduced to solving the subsystem involving only the variables $x_2, \ldots, x_n$. Gaussian elimination wants to achieve a similar reduction even when some (or all) coefficients occurring with x_1 are non-zero.

Assume, for example, that $a_{11} \neq 0$. Then

$$x_1 = \frac{1}{a_{11}} (b_1 - a_{12}x_2 - \ldots - a_{1n}x_n) \ . \tag{2.2}$$

We can substitute this expression for x_1 in all other equations and obtain a new system $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$ that involves only the variables $x_2, \ldots, x_n$. In this sense, the variable x_1 has been "eliminated".

The systems $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$ of linear equations are very closely related. Each solution $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ for $\mathbf{Ax} = \mathbf{b}$ yields a solution $\mathbf{x}' = (x_2, \ldots, x_n)^T$ for $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$ (we just omit the variable x_1 in $\mathbf{x}$). Conversely, each solution $\mathbf{x}'$ for $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$ can be extended to a solution $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ for $\mathbf{Ax} = \mathbf{b}$ by computing the value of x_1 via *backward substitution* according to the formula (2.2) from $\mathbf{x}' = (x_2, \ldots, x_n)^T$.

REMARK. From a geometrical point of view, passing from the solution space S of $\mathbf{Ax} = \mathbf{b}$ to the solution space S' of $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$ amounts to projecting the vectors $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in S$ to $\mathbf{x}' = (x_2, \ldots, x_n) \in S'$.

We next eliminate another variable, say x_2, from the system $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$ in the same way *etc.* until all variables have been eliminated. Going all the way back, we can compute a solution for the original system $\mathbf{Ax} = \mathbf{b}$ via repeated backward substitution.

What does it mean in terms of algebraic operations to "eliminate" x_1 in the system $\mathbf{Ax} = \mathbf{b}$? It turns out that there is no need to actually remove x_1 from the system. The elimination process comes down to a suitable sequence of "pivoting" operations that successively transform our original system into a completely equivalent system which, however, has the virtue of being easily solvable.

Given a pair (i, j) of row and column indices such that $a_{ij} \neq 0$, let us call the following operation a *Gaussian (i, j)-pivot* (with *pivot row i* and *pivot column j*) on the rows (equations) of the system $\mathbf{Ax} = \mathbf{b}$:

(GP) For all rows $k > i$: Add $(-a_{kj}a_{ij}^{-1}) \times$ (row i) to row k.

EX. 2.1. *Assume that $\tilde{\mathbf{A}}\mathbf{x} = \tilde{\mathbf{b}}$ arises from $\mathbf{Ax} = \mathbf{b}$ via a Gaussian (i, j)-pivot. Show that both systems have the same solution space S.*

EX. 2.2. *Show that the system $\mathbf{A}'\mathbf{x}' = \mathbf{b}'$ in the Gaussian elimination step with respect to x_1 and $a_{11} \neq 0$ is exactly the subsystem we obtain when we first apply a $(1, 1)$-pivot to $\mathbf{Ax} = \mathbf{b}$ and then remove column 1 and row 1 from the system.*

Recall that a matrix $\mathbf{M} = (m_{ij})$ is said to be *lower triangular* if $m_{ij} = 0$ whenever $i < j$, and *upper triangular* if $m_{ij} = 0$ whenever $i > j$.

Ex. 2.3. *Let $\tilde{\mathbf{A}}\mathbf{x} = \tilde{\mathbf{b}}$ be the system arising from $\mathbf{A}\mathbf{x} = \mathbf{b}$ via a Gaussian (i, j)-pivot. Show that there exists an invertible lower triangular matrix $\mathbf{M} \in \mathbb{R}^{m\times m}$ such that $\tilde{\mathbf{A}} = \mathbf{M}\mathbf{A}$ and $\tilde{\mathbf{b}} = \mathbf{M}\mathbf{b}$.*

By interchanging rows if necessary in order to obtain a non-zero pivot element, we can transform $\mathbf{A}\mathbf{x} = \mathbf{b}$ into upper triangular form with Gaussian pivots:

Gaussian Elimination

INIT: Set $j = 1$, $i = 1$.

ITER: WHILE $i \le m$ and $j \le n$ DO

(1) Find a row index $k \ge i$ such that $a_{kj} \neq 0$;
If no such k exists, then $j \leftarrow j + 1$, GOTO (1);
Interchange row i and row k;
Perform a Gaussian (i, j)-pivot;
Update $j \leftarrow j + 1$ and $i \leftarrow i + 1$;

REMARK. Step (1) of the Gaussian Elimination algorithm does not specify which row k to choose in case several candidates are available. There are examples demonstrating that the numerical stability (with respect to rounding errors) of the algorithm very much depends on a good pivot choice. Practical experience shows very good results if k is chosen as to maximize the absolute value $|a_{kj}|$ of the pivot element . This rule is called *partial pivoting*. (*Complete pivoting* tries to enhance the numerical stability of the computations by allowing also column permutations in the search for a maximal pivot element. The result, however, is usually not worth the extra computational effort of complete pivoting).

Note that this Gaussian elimination algorithm does not necessarily "eliminate" all variables but just achieves an upper triangular form of the system of equations. If $(1, j_1), (2, j_2), \ldots, (r-1, j_{r-1}), (r, j_r)$ is the sequence of pivots during the algorithm, the final system $\tilde{\mathbf{A}}\mathbf{x} = \tilde{\mathbf{b}}$ of equations will have the form

$$
\begin{array}{ccccccccccccc}
\tilde{a}_{1j_1}x_{j_1}+ & \ldots & \tilde{a}_{1j_2}x_{j_2}+ & \ldots & \tilde{a}_{1j_{r-1}}x_{j_{r-1}}+ & \ldots & \tilde{a}_{1j_r}x_{j_r}+ & \ldots & = & \tilde{b}_1 \\
 & & \tilde{a}_{2j_2}x_{j_2}+ & \ldots & \tilde{a}_{2j_{r-1}}x_{j_{r-1}}+ & \ldots & \tilde{a}_{2j_r}x_{j_r}+ & \ldots & = & \tilde{b}_2 \\
 & & & \ddots & & & & & \vdots & \\
 & & & & \tilde{a}_{r-1j_{r-1}}x_{j_{r-1}}+ & \ldots & \tilde{a}_{r-1j_r}x_{j_r}+ & \ldots & = & \tilde{b}_{r-1} \\
 & & & & & & \tilde{a}_{rj_r}x_{j_r}+ & \ldots & = & \tilde{b}_r \\
 & & & & & & & & \vdots &
\end{array}
$$

This final form of a system of linear equations is also known as *Hermite normal form* (or *row echelon form*) of the system.

The Gaussian Elimination algorithm implies $\tilde{a}_{tj} = 0$ whenever $t > r$. So there cannot be any solution for $\mathbf{Ax} = \mathbf{b}$ if there exists some $\tilde{b}_t \neq 0$ with $t > r$. Otherwise, because all pivot elements $\tilde{a}_{pj_p}$ are non-zero, we can easily compute a solution for $\mathbf{Ax} = \mathbf{b}$ by backtracking the pivots and performing backward substitution:

$$\begin{aligned}
x_{j_r} &= \tilde{b}_r / \tilde{a}_{rj_r} \\
x_{j_{r-1}} &= (\tilde{b}_{r-1} - \tilde{a}_{r-1j_r} x_{j_r}) / \tilde{a}_{r-1j_{r-1}} \\
x_{j_{r-2}} &= (\tilde{b}_{r-2} - \tilde{a}_{r-2j_{r-1}} x_{j_{r-1}} - \tilde{a}_{r-2j_r} x_{j_r}) / \tilde{a}_{r-2j_{r-2}} \\
&\vdots \\
x_{j_1} &= (\tilde{b}_1 - \sum_{p=2}^{r} \tilde{a}_{1j_p} x_{j_p}) / \tilde{a}_{1j_1}
\end{aligned}$$

while the other variables x_j are set to zero. This procedure is correct because the operations during Gaussian Elimination leave the solution space S of $\mathbf{Ax} = \mathbf{b}$ unaltered.

REMARK [RECOVERING ALL SOLUTIONS]. The solution of $\mathbf{Ax} = \mathbf{b}$ just computed is a special ("basic") solution in the sense that all non-pivot variables x_j are set to zero. The backward substitution process can easily be generalized by first assigning arbitrary values to the non-pivot variables x_j and then computing (unique) corresponding values for the remaining variables recursively. This way one can, in principle, generate *every* feasible solution of $A\mathbf{x} = \mathbf{b}$.

The Gaussian elimination algorithm has some important matrix-theoretic implications. Recall that the matrix $\mathbf{P} = (p_{ij}) \in \mathbb{R}^{m \times m}$ is a *permutation matrix* if $p_{ij} \in \{0, 1\}$ and each row and each column of $\mathbf{P}$ contains exactly one coefficient 1. Note that $\mathbf{P}^T\mathbf{P} = \mathbf{I}$ holds for every permutation matrix $\mathbf{P}$, which implies $\mathbf{P}^{-1} = \mathbf{P}^T$ for the inverse matrix $\mathbf{P}^{-1}$ of $\mathbf{P}$.

EX. 2.4. *Let $\mathbf{P} = (p_{ij})$ be a $(m \times m)$-permutation matrix. Show for the matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$: Row i of $\mathbf{PA}$ equals row j of $\mathbf{A}$ if and only if $p_{ij} = 1$.* (In other words: $\mathbf{P}$ permutes the rows of $\mathbf{A}$ according to the coefficients p_{ij}.)

THEOREM 2.1. *For every matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, there exists an $(m \times m)$-permutation matrix $\mathbf{P}$ and an invertible lower triangular matrix $\mathbf{M} \in \mathbb{R}^{m \times m}$ such that $\mathbf{U} = \mathbf{MPA}$ is upper triangular.*

Proof. Run the Gaussian algorithm on the matrix $\mathbf{A}$ and record all the row permutations that occur in the permutation matrix $\mathbf{P}$. Then we obtain the same final upper triangular matrix $\tilde{\mathbf{A}}$ if we perform Gaussian Elimination on the matrix $\bar{\mathbf{A}} = \mathbf{PA}$ *without any row permutations*. Denote by $\mathbf{M}_1, \ldots, \mathbf{M}_r$ the matrices describing the Gaussian pivots.

By Ex. 2.3, each $\mathbf{M}_i$ is an invertible lower triangular matrix. Hence also the product $\mathbf{M} = \mathbf{M}_r\mathbf{M}_{r-1}\cdots\mathbf{M}_1$ is an invertible lower triangular matrix and we obtain for $\mathbf{U} = \tilde{\mathbf{A}}$:

$$\mathbf{U} = \tilde{\mathbf{A}} = \mathbf{M}\bar{\mathbf{A}} = \mathbf{MPA} \ .$$

◇

COROLLARY 2.1 (*LU*-factorization)**.** *For every matrix* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, there exists an* $(m \times m)$*-permutation matrix* $\mathbf{P}$*, and matrices* $\mathbf{L} \in \mathbb{R}^{m\times m}$, $\mathbf{U} \in \mathbb{R}^{m\times n}$ *such that* $\mathbf{L}$ *is invertible and lower triangular,* $\mathbf{U}$ *is upper triangular, and*

$$\mathbf{LU} = \mathbf{PA} \ .$$

Proof. With the matrices $\mathbf{M}_i$ as in the proof of Theorem 2.1, we take $\mathbf{L} = \mathbf{M}^{-1} = \mathbf{M}_1^{-1}\cdots\mathbf{M}_r^{-1}$. $\mathbf{L}$ is lower triangular (because it is the inverse of a lower triangular matrix), and $\mathbf{LU} = \mathbf{M}^{-1}\mathbf{MPA} = \mathbf{PA}$ follows.

◇

REMARK. If an *LU*-factorization $\mathbf{LU} = \mathbf{PA}$ of $\mathbf{A}$ is known, the system $\mathbf{Ax} = \mathbf{b}$ can be solved in three steps:

(1) Compute $\bar{\mathbf{b}} := \mathbf{Pb}$;
(2) Compute $\mathbf{y}$ as a solution of $\mathbf{Ly} = \bar{\mathbf{b}}$;
(3) Compute $\mathbf{x}$ as a solution of $\mathbf{Ux} = \mathbf{y}$.

Step (2) can always be carried out since $\mathbf{L}$ is invertible so that $\mathbf{y} = \mathbf{L}^{-1}\bar{\mathbf{b}}$. Step (3) can be successfully performed if and only if $\mathbf{Ax} = \mathbf{b}$ has a solution at all. Since $\mathbf{L}$ is triangular, it is usually more efficient not to determine $\mathbf{L}^{-1}$ explicitly but to compute both $\mathbf{x}$ and $\mathbf{y}$ by backward substitution.

EX. 2.5. *Compute an LU-factorization for matrix* $\mathbf{A} = \begin{pmatrix} 1 & 2 & 3 \\ 0 & 2 & 1 \\ 2 & 1 & 1 \end{pmatrix}$.

COROLLARY 2.2 (Gale's Theorem)**.** *Exactly one of the alternatives is true:*

(a) *The system* $\mathbf{Ax} = \mathbf{b}$ *has a solution.*
(b) *There exists a vector* $\mathbf{y} \in \mathbb{R}^m$ *such that* $\mathbf{y}^T\mathbf{A} = \mathbf{0}^T$ *and* $\mathbf{y}^T\mathbf{b} \neq 0$.

Proof. Because $\mathbf{y}^T\mathbf{Ax} = \mathbf{y}^T\mathbf{b}$, (a) and (b) cannot hold simultaneously. Assume now that $\mathbf{Ax} = \mathbf{b}$ has no solution. We show that then (b) is true.

Consider the final system $\tilde{\mathbf{A}}\mathbf{x} = \tilde{\mathbf{b}}$ computed by Gaussian Elimination from $\mathbf{Ax} = \mathbf{b}$, where $\tilde{\mathbf{A}} = \mathbf{MPA}$ and $\tilde{\mathbf{b}} = \mathbf{MPb}$. The system is infeasible if and only if there is some row index $i > r$, such that $\mathbf{0}^T$ is the ith row of $\tilde{\mathbf{A}}$ and $\tilde{b}_i \neq 0$.

Let $\mathbf{y}^T$ be the ith row vector of the matrix $\mathbf{MP}$. Then $\mathbf{y}^T\mathbf{A}$ yields the ith row vector of $\tilde{\mathbf{A}}$, while $\mathbf{y}^T\mathbf{b}$ yields the ith component $\tilde{b}_i$ of $\tilde{\mathbf{b}}$, and the Corollary follows.

◇

Let us take a vector space point of view at Gaussian Elimination with respect to the linear equality system $\mathbf{Ax} = \mathbf{b}$. The *row space* $V = \text{row } \mathbf{A}$ of $\mathbf{A}$ is the linear

hull of the row vectors of $\mathbf{A}$. By rank $\mathbf{A}$ we denote the *rank* of the matrix $\mathbf{A}$, *i.e.*, the maximal number of linearly independent rows of $\mathbf{A}$. So rank $\mathbf{A} = \dim \text{row } \mathbf{A}$.

Since pivot operations are, in particular, sequences of elementary vector space operations on the row vectors, the space row $\mathbf{A}$ will stay the same after each Gaussian pivot. From the upper triangular form of the final matrix $\tilde{\mathbf{A}}$ it follows immediately that rank $\tilde{\mathbf{A}} = r$, where r is the total number of Gaussian pivots. Hence

$$r = \text{rank } \tilde{\mathbf{A}} = \text{rank } \mathbf{A} \ .$$

So Gaussian Elimination provides a fast algorithm for computing a basis of the space row $\mathbf{A}$.

We emphasize that the *column space* col $\mathbf{A} = \text{row } \mathbf{A}^T$ *does* change when we apply (row) pivots to $\mathbf{A}$. Note, however, that the set of columns $\{\mathbf{a}_{j_1}, \ldots, \mathbf{a}_{j_l}\}$ of $\mathbf{A}$ is linearly independent if and only if the corresponding columns $\tilde{\mathbf{a}}_{j_1}, \ldots, \tilde{\mathbf{a}}_{j_l}$ of the transformed matrix $\tilde{\mathbf{A}}$ (obtained from $\mathbf{A}$ by row pivots) are linearly independent. So, in particular, Gaussian Elimination reveals that the 'pivot columns' $\mathbf{a}_{j_1}, \ldots, \mathbf{a}_{j_r}$ form a basis of the column space and we observe that

$$r = \text{rank } \mathbf{A} = \dim \text{ (col } \mathbf{A}) = \dim \text{ (row } \mathbf{A}) \ .$$

2.1.1. Gauss-Jordan Elimination. From a conceptual point of view, one might want to strengthen the Gaussian pivoting rule (GP) to

(GJP) For *all* rows $k \neq i$: Add $(-a_{kj}a_{ij}^{-1}) \times$ (row i) to row k ,

which transforms also the matrix elements above the pivot elements to zero.

If one applies the elimination algorithm with (GJP) instead of (GP), one obtains a system $\tilde{\mathbf{A}}\mathbf{x} = \tilde{\mathbf{b}}$ of equations with each pivot column $\mathbf{A}_{\cdot j_k}$ of $\tilde{\mathbf{A}}$ having a unique nonzero entry in the corresponding pivot position (k, j_k).

While this form of the system of equations would make backward substitution even easier, the elimination algorithm itself requires more computational effort. Therefore, Gauss-Jordan Elimination offers no practical advantage over Gaussian Elimination. Its virtues are more to be seen in being a theoretical tool for algorithmic analysis (see, for example, the simplex algorithm for linear programs in Chapter 4).

Ex. 2.6. *Assume that the matrix* $\mathbf{A} \in \mathbb{R}^{n \times n}$ *satisfies* rank $\mathbf{A} = n$. *Show that Gauss-Jordan Elimination transforms* $\mathbf{A}$ *into a diagonal matrix* $\tilde{\mathbf{A}}$ *all of whose diagonal elements are non-zero.*

2.1.2. Determinants. Let π be a permutation of the n distinct elements $1, 2, \ldots, n$. With π we associate the permutation matrix $\mathbf{P} = (p_{ij})$, where

$$p_{ij} \ = \ \begin{cases} 1 & \text{if } j = \pi(i) \\ 0 & \text{otherwise.} \end{cases}$$

When we run Gaussian Elimination on $\mathbf{P}$, the algorithm will only carry out row permutations (and thus will transform $\mathbf{P}$ into the $(n \times n)$-identity matrix $\mathbf{I}$). Suppose $s = s(\pi)$ is the number of proper row interchanges in the course of the algorithm. Then the number

$$\operatorname{sgn} \pi \;=\; (-1)^{s(\pi)}$$

is the *sign* of the permutation π. Depending on its sign, π is called either *even* or *odd*.

Given the permutation π, let π^{-1} be the inverse permutation $\pi(i) \mapsto i$, $i = 1, \ldots, n$. Then we apparently have

$$\operatorname{sgn} \pi = \operatorname{sgn} \pi^{-1} \; .$$

With the $(n \times n)$-matrix $\mathbf{A} = (a_{ij})$, one associates its *determinant* as the number

$$\det \mathbf{A} \;=\; \sum_{\pi} (\operatorname{sgn} \pi) a_{1\pi(1)} a_{2\pi(2)} \cdots a_{n\pi(n)} \; , \tag{2.3}$$

where the sum is taken over all $n!$ permutations π of the indices $1, 2, \ldots, n$.

Ex. 2.7. *Show:* $\det \mathbf{A} = \det \mathbf{A}^T$. *(Hint:* $\operatorname{sgn} \pi = \operatorname{sgn} \pi^{-1}$*).*

REMARK. Occasionally, it is helpful to think of $\det \mathbf{A}$ not just as a real parameter associated with a matrix $\mathbf{A}$ but to interpret $\det : \mathbb{R}^{n \times n} \to \mathbb{R}$ as a real-valued function on the n^2-dimensional vectors in $\mathbb{R}^{n^2}$. The expression (2.3) shows that $\det \mathbf{X}$ is a sum of products of components of $\mathbf{X} = (x_{ij})$ and hence clearly is a continuous (in fact, differentiable) function.

REMARK. It is well-known that $\det \mathbf{A}$ admits an intuitive interpretation as the change in volume of a body $K \subseteq \mathbb{R}^n$ under the influence ("deformation") of the linear map $\mathbf{x} \mapsto \mathbf{A}\mathbf{x}$:

$$\operatorname{vol} \mathbf{A}(K) = |\det \mathbf{A}| \cdot \operatorname{vol} K \; . \tag{2.4}$$

From the definition (2.3), it follows directly that

$$\det \mathbf{A} = a_{11} a_{22} \cdots a_{nn} \tag{2.5}$$

must hold if $\mathbf{A}$ is (upper or lower) triangular. One can, furthermore, deduce the *determinant multiplication rule* for all $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$:

$$\det \mathbf{AB} = \det \mathbf{A} \cdot \det \mathbf{B} \; . \tag{2.6}$$

Ex. 2.8.

(a) *Let* $\mathbf{P}$ *be the permutation matrix for the permutation* π.
Show: $\operatorname{sgn} \pi = \det \mathbf{P}$.

(b) *Use the determinant multiplication rule to show: If* $\mathbf{A}'$ *is the matrix obtained from* $\mathbf{A}$ *by performing a Gaussian* (i, j)*-pivot, then* $\det \mathbf{A}' = \det \mathbf{A}$.

Ex. 2.8 indicates that $\det \mathbf{A}$ can be efficiently computed via the Gaussian elimination algorithm: $\mathbf{A}$ is transformed into the triangular matrix $\tilde{\mathbf{A}} = (\tilde{a}_{ij})$ and, therefore

$$\det \mathbf{A} = (-1)^s \tilde{a}_{11}\tilde{a}_{22}\cdots\tilde{a}_{nn} \ , \tag{2.7}$$

where s is the number of proper row interchanges during the run of the algorithm.

Ex. 2.9. *Show:* $\det \mathbf{A} \neq 0$ *if and only if all rows of* $\mathbf{A}$ *are linearly independent.*

Cramer's Rule. Consider the system $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{A} \in \mathbb{R}^{n\times n}$ satisfies $\det \mathbf{A} \neq 0$. *Cramer's rule* provides a determinant formula for computing each component x_i of the (unique) solution vector $\mathbf{x}$:

$$x_i = \frac{\det \hat{\mathbf{A}}_i}{\det \mathbf{A}} \ , \tag{2.8}$$

where $\hat{\mathbf{A}}_i$ is the matrix we obtain from $\mathbf{A}$ upon replacing the ith column of $\mathbf{A}$ by the vector $\mathbf{b}$.

The validity of Cramer's rule is not difficult to check directly in the case where A is a diagonal matrix. In the general case, Gauss-Jordan Elimination will transform $\mathbf{A}$ into diagonal form if $\det \mathbf{A} \neq 0$ (see Ex. 2.6). Since the pivots will leave the determinants $\det \hat{\mathbf{A}}_i$ and $\det \mathbf{A}$ unchanged (see Ex. 2.8), the validity of Cramer's rule follows.

REMARK. Cramer's rule is only of theoretical value. Gaussian Elimination will compute a solution faster. The merit of (2.8) lies in the fact that it provides a means to estimate the numerical size of the solution vector $\mathbf{x}$ in theoretical algorithmic analysis (see Corollary **??**).

2.1.3. Symmetric and Positive Semidefinite Matrices. Recall that the matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ is said to be symmetric if $\mathbf{A} = \mathbf{A}^T$. We denote the set of (real) symmetric $n \times n$ matrices by $\mathbb{S}^{n\times n}$. We want to apply Gaussian Elimination to the rows *and* to the columns of the symmetric matrix $A = (a_{ij})$ with the goal of retaining symmetry after each elimination step. Thereby, we take advantage of the fact that matrix multiplication *from the left* has the same effect on the *rows* of a matrix as multiplication with the *transposed* matrix *from the right* has on the *columns* of a matrix.

Assume first $a_{11} \neq 0$. Then we can perform a Gaussian pivot with respect to a_{11} on the symmetric matrix $\mathbf{A}$. If this pivot is described by the matrix $\mathbf{M}_1$, say, then

$$\mathbf{A}' = \mathbf{M}_1\mathbf{A}\mathbf{M}_1^T$$

is again a symmetric matrix. (Note that $\mathbf{M}_1\mathbf{A}\mathbf{M}_1^T = (\mathbf{M}_1\mathbf{A})\mathbf{M}_1^T$ can be interpreted as the result of the symmetric Gaussian pivot with respect to a_{11} relative to the columns of $\mathbf{M}_1\mathbf{A}$).

If the diagonal element a'_{22} of $\mathbf{M}_1\mathbf{A}\mathbf{M}_1^T$ is non-zero, we can pivot on a'_{22} in the same way to obtain the symmetric matrix

$$\mathbf{M}_2(\mathbf{M}_1\mathbf{A}\mathbf{M}_1^T)\mathbf{M}_2^T = (\mathbf{M}_2\mathbf{M}_1)\mathbf{A}(\mathbf{M}_2\mathbf{M}_1)^T$$

and continue.

A problem occurs if a_{11} (or any subsequent diagonal element a_{ii}) equals zero. So assume that - after $i-1$ pivots - we have $a'_{ii} = 0$. If $a'_{kl} = 0$ for all $k, l \geq i$, we are done (the matrix is diagonalized). Hence assume this is not the case. If a diagonal element $a'_{kk} \neq 0$ $(k > i)$ exists, we may resolve the problem by simply permuting rows i and k and columns i and k, so as to *switch* $a'_{kk} \neq 0$ into position (i, i). If all diagonal elements a'_{kk} with $k \geq i$ are zero, let $a'_{kl} = a'_{lk} \neq 0$ for some $l > k \geq i$. We then add row l to row k and column l to column k so as to obtain $a'_{kl} + a'_{lk} = 2a'_{kl} \neq 0$ in position (k, k) and then switch i and k as before.

Let us refer to this operation as *switching a nonzero* into position (i, i). We may then state the algorithm transforming a symmetric matrix $\mathbf{A} \in \mathbb{S}^{n\times n}$ into a diagonal matrix as follows.

Diagonalization

FOR $i = 1, \ldots, n$ DO
    Switch a nonzero into position (i, i) if necessary;
    (If this is not possible, *i.e.*, $a'_{kl} = 0$ for $k, l \geq i$: STOP.)
    Perform a Gaussian (i, i)-pivot on the rows;
    Perform a Gaussian (i, i)-pivot on the columns;
NEXT i.

THEOREM 2.2. *Let $\mathbf{A} \in \mathbb{S}^{n\times n}$ be a symmetric matrix. Then there exists an invertible matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ such that*

$$\mathbf{D} = \mathbf{Q}\mathbf{A}\mathbf{Q}^T$$

is a diagonal matrix.

Proof. By construction, algorithm Diagonalization will produce a diagonal matrix. Moreover, each of the row operations performed by the algorithm can be described via multiplication by a suitable invertible matrix $\mathbf{M}$ from the left. Symmetrically, the corresponding column operation is given by the multiplication from the right with the transposed matrix $\mathbf{M}^T$.

Let $\mathbf{Q}$ be the product of these matrices $\mathbf{M}$. Then Diagonalization transforms $\mathbf{A}$ into the diagonal matrix $\mathbf{Q}\mathbf{A}\mathbf{Q}^T$. Because each of the row operations $\mathbf{M}$ is invertible, $\mathbf{Q}$ is invertible.

$\diamond$

EX. 2.10. *Find an invertible matrix* $\mathbf{Q} \in \mathbb{R}^{4\times 4}$ *such that* $\mathbf{QAQ}^T$ *is diagonal, where*

$$\mathbf{A} \;=\; \begin{pmatrix} 0 & 1 & 3 & 2 \\ 1 & -1 & 1 & 0 \\ 3 & 1 & 1 & 2 \\ 2 & 0 & 2 & 0 \end{pmatrix}$$

Relaxing the notion of *positive definiteness* we know from inner products and Gram matrices, we say that the symmetric matrix $\mathbf{A} = (a_{ij}) \in \mathbb{S}^{n\times n}$ is *positive semidefinite* ("p.s.d."), denoted by $\mathbf{A} \succeq \mathbf{0}$, if for every $\mathbf{x} = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}x_i x_j \;\geq\; 0\;. \tag{2.9}$$

Hence $\mathbf{A}$ is positive definite (denoted by $\mathbf{A} \succ \mathbf{0}$) if $\mathbf{A} \succeq \mathbf{0}$ and $\mathbf{x}^T\mathbf{A}\mathbf{x} = 0$ holds only for $\mathbf{x} = \mathbf{0}$.

COROLLARY 2.3. *Let* $\mathbf{A}$ *be a symmetric matrix and* $\mathbf{Q}$ *an invertible matrix such that* $\mathbf{D} = \mathbf{QAQ}^T$ *is diagonal. Then*

(a) $\mathbf{A}$ *is p.s.d. if and only if all diagonal elements of* $\mathbf{D}$ *are non-negative.*
(b) $\mathbf{A}$ *is positive definite if and only if all diagonal elements of* $\mathbf{D}$ *are strictly positive.*

Proof. Since $\mathbf{x} = \mathbf{Q}^T\mathbf{y}$ defines a 1-1 correspondence between $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, we conclude from

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \mathbf{y}^T\mathbf{Q}\mathbf{A}\mathbf{Q}^T\mathbf{y} = \mathbf{y}^T\mathbf{D}\mathbf{y} = \sum_{i=1}^{n} d_i y_i^2$$

that $\mathbf{A}$ is p.s.d. if and only if $\mathbf{D}$ is. The latter however is equivalent to $d_i \geq 0$ for $i = 1, \ldots, n$. (If $d_i < 0$ then $\mathbf{y} = \mathbf{e}_i$, the *ith* unit vector, yields $\mathbf{y}^T\mathbf{D}\mathbf{y} = d_i < 0$.) ◇

Corollary 2.3 is algorithmically very important. It implies that there is an efficient way of deciding whether a given matrix is positive (semi-)definite. One only needs to run the Diagonalization algorithm on the matrix and then read the result off the diagonalized matrix.

Furthermore, Corollary 2.3 explains how positive semidefinite matrices are constructed from other matrices. Let $\mathbf{A} \in \mathbb{R}^{m\times n}$ be an arbitrary matrix. Then the symmetric $(m \times m)$-matrix $\mathbf{S} = \mathbf{A}\mathbf{A}^T$ is p.s.d.. This is easily seen: We let $\mathbf{x} \in \mathbb{R}^m$ be arbitrary and consider $\mathbf{y} = \mathbf{A}^T\mathbf{x}$. Then

$$\mathbf{x}^T\mathbf{S}\mathbf{x} = \mathbf{x}^T\mathbf{A}\mathbf{A}^T\mathbf{x} = \mathbf{y}^T\mathbf{y} \geq 0\;.$$

Conversely, if $\mathbf{S}$ is symmetric, we can find some invertible matrix $\mathbf{Q}$ such that $\mathbf{D} = \mathbf{QSQ}^T$ is diagonal. If $\mathbf{S}$ is in addition p.s.d., Corollary 2.3 says that the elements of $\mathbf{D}$ are non-negative. So we may form $\sqrt{\mathbf{D}}$, the diagonal matrix whose diagonal elements are the square roots of the elements of D, and set

$$\mathbf{A} = \mathbf{Q}^{-1}\sqrt{\mathbf{D}}\;.$$

Then $\mathbf{A}\mathbf{A}^T = (\mathbf{Q}^{-1})\sqrt{\mathbf{D}}\sqrt{\mathbf{D}}(\mathbf{Q}^{-1})^T = \mathbf{Q}^{-1}\mathbf{D}(\mathbf{Q}^{-1})^T = \mathbf{S}$. So we arrive at

COROLLARY 2.4. *Let* $\mathbf{S}$ *be a symmetric matrix. Then*

(a) $\mathbf{S}$ *is p.s.d. if and only if there is a matrix* $\mathbf{A}$ *such that* $\mathbf{S} = \mathbf{A}\mathbf{A}^T$.
(b) $\mathbf{S}$ *is positive definite if and only if there is an invertible matrix* $\mathbf{A}$ *such that* $\mathbf{S} = \mathbf{A}\mathbf{A}^T$.

◇

EX. 2.11. *Prove part* (b) *of Corollary 2.4. Moreover, show: If the matrix* $\mathbf{A}$ *is positive definite, then* $\mathbf{A}^{-1}$ *exists and is positive definite.*

REMARK [INNER PRODUCTS]. Recall form Section 1.3 that each inner product $\langle\cdot|\cdot\rangle$ on $\mathbb{R}^n$ is defined by its values relative to the standard basis $\{\mathbf{e}_1, \ldots, \mathbf{e}_n\}$ and the (positive definite) Gram matrix $\mathbf{G}$, where

$$\mathbf{G} = (\langle \mathbf{e}_i | \mathbf{e}_j \rangle) \ .$$

If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, then

$$\langle \mathbf{x} | \mathbf{y} \rangle = \mathbf{x}^T \mathbf{G} \mathbf{y} \ .$$

Writing $\mathbf{G} = \mathbf{A}^T\mathbf{A}$, we obtain

$$\langle \mathbf{x} | \mathbf{y} \rangle = \mathbf{x}^T(\mathbf{A}^T\mathbf{A})\mathbf{y} = (\mathbf{A}\mathbf{x})^T(\mathbf{A}\mathbf{y}) \ . \tag{2.10}$$

Thus every inner product reduces to the standard Euclidean inner product *via* a suitable transformation $\mathbf{x} \to \mathbf{A}\mathbf{x}$.

REMARK. Corollary 2.4 may be *false* if we restrict ourselves to rational numbers! The reason is that we have to take square roots of numbers. $\sqrt{2}$, for example, is not in $\mathbb{Q}$. So the rational positive definite (1×1)-matrix $\mathbf{S} = [2]$ cannot be expressed in the form $\mathbf{S} = \mathbf{A}\mathbf{A}^T$ with a rational matrix $\mathbf{A}$.

EX. 2.12. *Let* $\mathbf{A} = (a_{ij}) \in \mathbb{S}^{n\times n}$ *be a positive definite matrix.*

(a) *Show that the diagonal elements* a_{ii} *of* $\mathbf{A}$ *are strictly positive.*
(b) *Show that the Diagonalization algorithm will always maintain* $a_{ii} \neq 0$.

REMARK [CHOLESKY FACTORIZATION]. As a consequence of Ex. 2.12, one finds that Diagonalization only performs Gaussian pivots when applied to a positive definite matrix $\mathbf{A}$. In particular, the matrix $\mathbf{Q}$ produced by Diagonalization is lower triangular. So $\mathbf{L} = \mathbf{Q}^{-1}\sqrt{\mathbf{D}}$ yields a LU-factorization with $\mathbf{U} = \mathbf{L}^\mathbf{T}$, the so-called *Cholesky factorization*:

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T \ . \tag{2.11}$$

2.2. Orthogonal Projection and Least Square Approximation

Given an inner product $\langle .|. \rangle$ on $\mathbb{R}^n$, we define for every $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ their *distance* via the norm

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{\langle \mathbf{x} - \mathbf{y} | \mathbf{x} - \mathbf{y} \rangle} \ . \tag{2.12}$$

REMARK. Although (2.12) involves the square root, none of the computations below would lead us outside the field $\mathbb{Q}$ of rational numbers since we actually work with the *squared* distance.

Given the vector $\mathbf{x} \in \mathbb{R}^n$ and a linear subspace $W \subseteq \mathbb{R}^n$, we want to find the *projection of* $\mathbf{x}$ *onto* W, *i.e.*, a vector $\hat{\mathbf{x}} \in W$ such that

$$\|\mathbf{x} - \hat{\mathbf{x}}\| = \min_{\mathbf{y} \in W} \|\mathbf{x} - \mathbf{y}\| \ . \tag{2.13}$$

The optimization problem (2.13) is equivalent with

$$\|\mathbf{x} - \hat{\mathbf{x}}\|^2 = \min_{\mathbf{y} \in W} \|\mathbf{x} - \mathbf{y}\|^2 \ . \tag{2.14}$$

Problem (2.13) is often called the *least square approximation* problem (see also the next subsection). Its solution is based on the following observation.

LEMMA 2.1. *Assume that $\mathbf{x}' \in W$ is such that the vector $\mathbf{x} - \mathbf{x}'$ is orthogonal with every $\mathbf{w} \in W$. Then $\hat{\mathbf{x}} = \mathbf{x}'$ is the unique optimal solution of (2.13).*

Proof. Let $\mathbf{y} \in W$ be an arbitrary vector and consider $\mathbf{w} = \mathbf{x}' - \mathbf{y}$. Because $\mathbf{w} \in W$, the Theorem of Pythagoras can be applied and yields

$$\|\mathbf{x} - \mathbf{y}\|^2 = \|\mathbf{x} - \mathbf{x}' + \mathbf{w}\|^2 = \|\mathbf{x} - \mathbf{x}'\|^2 + \|\mathbf{w}\|^2 \ .$$

Hence $\mathbf{y}$ is optimal for (2.13) if and only if $\mathbf{w} = \mathbf{0}$.

◇

It is not difficult to compute a vector $\mathbf{x}'$ satisfying the hypothesis of Lemma 2.1 if we know a basis $\{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ for W. Writing

$$\mathbf{x}' = z_1 \mathbf{a}_1 + \ldots + z_m \mathbf{a}_m \quad \text{with } z_i \in \mathbb{R},$$

observe first that $\mathbf{x} - \mathbf{x}'$ is orthogonal with every basis vector $\mathbf{a}_i$ of W exactly when for all $i = 1, \ldots, m$,

$$\langle \mathbf{x} - \mathbf{x}' | \mathbf{a}_i \rangle = \langle \mathbf{x} | \mathbf{a}_i \rangle - \langle \mathbf{x}' | \mathbf{a}_i \rangle = 0 \ . \tag{2.15}$$

From the linear expansion $\langle \mathbf{x}' | \mathbf{a}_i \rangle = z_1 \langle \mathbf{a}_1 | \mathbf{a}_i \rangle + \ldots + z_m \langle \mathbf{a}_m | \mathbf{a}_i \rangle$, we see that the equalities (2.15) give rise to the system of linear equations

$$\begin{array}{ccccccc}
\langle \mathbf{a}_1 | \mathbf{a}_1 \rangle z_1 & + & \ldots & + & \langle \mathbf{a}_m | \mathbf{a}_1 \rangle z_m & = & \langle \mathbf{x} | \mathbf{a}_1 \rangle \\
\langle \mathbf{a}_1 | \mathbf{a}_2 \rangle z_1 & + & \ldots & + & \langle \mathbf{a}_m | \mathbf{a}_2 \rangle z_m & = & \langle \mathbf{x} | \mathbf{a}_2 \rangle \\
\vdots & & & & & & \\
\langle \mathbf{a}_1 | \mathbf{a}_m \rangle z_1 & + & \ldots & + & \langle \mathbf{a}_m | \mathbf{a}_m \rangle z_m & = & \langle \mathbf{x} | \mathbf{a}_m \rangle,
\end{array}$$

which we can express more compactly with the help of the Gram matrix $\mathbf{G} = (\langle \mathbf{a}_j | \mathbf{a}_i \rangle) \in \mathbb{R}^{m \times m}$ as

$$\mathbf{G}\mathbf{z} = \mathbf{b} \ , \tag{2.16}$$

where $\mathbf{z} = (z_1, \ldots, z_m)^T \in \mathbb{R}^m$ and $\mathbf{b} = (\langle \mathbf{x} | \mathbf{a}_1 \rangle, \ldots, \langle \mathbf{x} | \mathbf{a}_m \rangle)^T \in \mathbb{R}^m$. Because the Gram matrix $\mathbf{G}$ is positive definite, the inverse matrix $\mathbf{G}^{-1}$ exists and yields an explicit formula for the solution

$$\mathbf{z} = \mathbf{G}^{-1}\mathbf{b} \ . \tag{2.17}$$

Let us consider the case of the standard inner product $\langle \mathbf{x} | \mathbf{y} \rangle = \mathbf{x}^T\mathbf{y}$. We form the matrix $\mathbf{A} \in \mathbb{R}^{n \times m}$ with columns $\mathbf{a}_i$. Then

$$W = \{\mathbf{y} \in \mathbb{R}^n \,|\, \mathbf{y} = \mathbf{A}\mathbf{z}, \mathbf{z} \in \mathbb{R}^m\} = \operatorname{col} \mathbf{A} \ ,$$

and we can write the least square approximation problem (2.14) in the form

$$\min_{\mathbf{z} \in \mathbb{R}^m} \|\mathbf{x} - \mathbf{A}\mathbf{z}\|^2 \ .$$

Here we have $\mathbf{G} = \mathbf{A}^T\mathbf{A}$ and $\mathbf{b} = \mathbf{A}^T\mathbf{x}$. So formula (2.17) implies for the orthogonal projection $\hat{\mathbf{x}}$ of $\mathbf{x}$ onto W:

$$\hat{\mathbf{x}} = \mathbf{A}\mathbf{z} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{b} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x} \ . \tag{2.18}$$

The same computational approach works when the linear subspace $W \subseteq \mathbb{R}^n$ of interest is given as the *orthogonal complement* $W = U^\perp$ of the linear subspace U generated by the columns $\mathbf{a}_1, \ldots, \mathbf{a}_m$ of $\mathbf{A}$, *i.e.*,

$$W = \{\mathbf{w} \in \mathbb{R}^n \,|\, \mathbf{a}_i^T\mathbf{w} = 0 \text{ for } i = 1, \ldots, m\} = \ker \mathbf{A}^T \ .$$

Let $\mathbf{x}'$ be the orthogonal projection of $\mathbf{x} \in \mathbb{R}^n$ onto U. Then

$$\hat{\mathbf{x}} = \mathbf{x} - \mathbf{x}' \tag{2.19}$$

is the projection of $\mathbf{x}$ onto W. By construction, namely, $\hat{\mathbf{x}}$ is orthogonal with every vector in U, which means $\hat{\mathbf{x}} \in W$. Moreover, $\mathbf{x} - \hat{\mathbf{x}} = \mathbf{x}' \in U$ is orthogonal with every $\mathbf{w} \in W$. So, by Lemma 2.1, $\hat{\mathbf{x}}$ is indeed the desired projection.

According to (2.19) and (2.18), the orthogonal projection $\hat{\mathbf{x}}$ of the vector $\mathbf{x} \in \mathbb{R}^n$ onto $W = \ker \mathbf{A}^T$ is the vector

$$\hat{\mathbf{x}} = \mathbf{x} - \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{x} \ . \tag{2.20}$$

Gradient Projection. To illustrate the usefulness of the concept of an orthogonal projection, let us consider the differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ at the point $\mathbf{x}_0 \in \mathbb{R}^n$ with $\nabla f(\mathbf{x}_0) \neq \mathbf{0}^T$. We have seen in Section 1.4.3 that the gradient vector $\mathbf{c} = [\nabla f(\mathbf{x}_0)]^T$ points into the direction of the largest marginal increase of f at $\mathbf{x}_0$.

Suppose we are interested in finding the direction $\mathbf{u}$ of largest marginal increase under the additional constraint $\mathbf{u} \in W$, where W is a fixed linear subspace of $\mathbb{R}^n$. This amounts to solving the problem

$$\max \{\mathbf{c}^T\mathbf{u} \,|\, \mathbf{u} \in W, \|\mathbf{u}\| = 1\} \,. \tag{2.21}$$

Let $\hat{\mathbf{c}}$ be the orthogonal projection of $\mathbf{c}$ onto W (and assume $\hat{\mathbf{c}} \neq \mathbf{0}$). We claim that $\hat{\mathbf{u}} = \hat{\mathbf{c}}/\|\hat{\mathbf{c}}\|$ solves the optimization problem (2.21). Indeed, if $\mathbf{u} \in W$, then $\mathbf{u}$ is orthogonal with $\mathbf{c} - \hat{\mathbf{c}}$. Hence

$$\mathbf{c}^T\mathbf{u} = (\mathbf{c} - \hat{\mathbf{c}})^T\mathbf{u} + \hat{\mathbf{c}}^T\mathbf{u} = \hat{\mathbf{c}}^T\mathbf{u} \,.$$

By the inequality of Cauchy-Schwarz, the latter is maximized exactly when $\mathbf{u}$ is a scalar multiple of $\hat{\mathbf{c}}$, which establishes the claim.

2.2.1. Least Square Approximation. A *linear model* tries to relate a vector $\mathbf{y} \in \mathbb{R}^m$ of m output parameters y_i to a vector $\mathbf{x} \in \mathbb{R}^n$ of input parameters x_j via the relation $\mathbf{y} = \mathbf{A}\mathbf{x}$, where the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ represents the structure of the linear model.

Suppose that upon the unknown input $\mathbf{x}$ in the model the output $\bar{\mathbf{y}}$ is observed. Then we can try to determine $\mathbf{x}$ by solving the system $\mathbf{A}\mathbf{x} = \bar{\mathbf{y}}$. Often, however, this system will have no solution (for example, because of measurement errors) and we will, more generally, content ourselves with an optimal solution $\hat{\mathbf{x}}$ for the problem

$$\min_{\mathbf{x}\in\mathbb{R}^n} \|\bar{\mathbf{y}} - \mathbf{A}\mathbf{x}\|^2 \,, \tag{2.22}$$

which can be solved by the method described in the previous section.

Best Fit. For illustration, assume that some quantity $y = y(t)$ is a function of some real parameter t. We do not know the function explicitly. As an approximation, we model it as a polynomial of degree n with $n+1$ unknown structural parameters $a_0, a_1, \ldots, a_n$:

$$y(t) = a_0 + a_1 t + a_2 t^2 + \ldots + a_n t^n \,.$$

If we have the data of $m \geq n+1$ measurements of the output $\bar{y}_i$ relative to the input t_i, $1 \leq i \leq m$, at our disposal, we can form the measurement matrix $\mathbf{M}$ with rows $(1, t_i, t_i^2, \ldots, t_i^n)$. We now wish to estimate $\mathbf{x} = (a_0, a_1, \ldots, a_n)^T$ as the solution that "fits best" the observed relation

$$\bar{\mathbf{y}} = \mathbf{M}\mathbf{x} \quad \text{where } \bar{\mathbf{y}} = (\bar{y}_1, \ldots, \bar{y}_m)^T \,.$$

REMARK. We should be aware that the important question whether (2.22) is indeed an appropriate measure for "best fit" can*not* be decided by mathematics but must be answered by the person who sets up the mathematical model for a concrete physical situation.

Ex. 2.13. *Find the line $y(t) = a + bt$ in the plane $\mathbb{R}^2$ that provides the best least square fit to the observed data $y(0) = -1$, $y(1) = 2$, and $y(2) = 1$.*

Quadratic Optimization. As a second example, consider the *quadratic optimization problem under linear equality constraints*

$$\min\ \mathbf{x}^T\mathbf{Q}\mathbf{x} = \sum_{i=1}^{n}\sum_{j=1}^{n} q_{ij}x_i x_j \quad s.t. \quad \mathbf{A}\mathbf{x} = \mathbf{b}\ , \tag{2.23}$$

where $\mathbf{Q} = (q_{ij}) \in \mathbb{R}^{n\times n}$, $\mathbf{A} \in \mathbb{R}^{m\times n}$, and $\mathbf{b} \in \mathbb{R}^m$ are given problem parameters. If the matrix $\mathbf{Q}$ is positive definite, we can solve (2.23) with the methods of this chapter in the following way. We define for any $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ a $\mathbf{Q}$-inner product and a $\mathbf{Q}$-norm:

$$\langle \mathbf{x}|\mathbf{y}\rangle_Q = \mathbf{x}^T\mathbf{Q}\mathbf{y} \qquad \text{and} \qquad \|\mathbf{x}\|_Q = \sqrt{\langle \mathbf{x}|\mathbf{x}\rangle_Q}\ .$$

With this terminology, (2.23) asks for an element with minimal $\mathbf{Q}$-norm in the affine subspace $L = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b}\}$.

If we now compute a feasible solution $\mathbf{p} \in \mathbb{R}^n$ for $\mathbf{A}\mathbf{x} = \mathbf{b}$, we obtain a representation $L = \mathbf{p} + \ker \mathbf{A}$. Minimizing $\|\mathbf{x}\|_Q$ over L becomes equivalent with

$$\min\ \{\|\mathbf{p} - \mathbf{w}\|_Q \,|\, \mathbf{w} \in \ker \mathbf{A}\}\ , \tag{2.24}$$

which is a particular case of (2.13) and can be solved by computing the $\mathbf{Q}$-orthogonal projection $\hat{\mathbf{p}}_Q$ of $\mathbf{p}$ onto $W = \ker \mathbf{A}$.

The quadratic optimization problem (2.23) occurs, for example, as a subproblem that has to be solved repeatedly during so-called SQP-algorithms (see Chapter 12). Another application arises from the fundamental *Gauss-Markov* model in the theory of statistical inference, which we briefly describe (see, *e.g.*, [67] for more details).

The Gauss-Markov Model. We generalize the linear model $\mathbf{y} = \mathbf{A}\mathbf{x}$ by allowing for random noise in the measurements. We assume not only that the output $\mathbf{y} \in \mathbb{R}^m$ depends linearly on the input $\mathbf{x} \in \mathbb{R}^n$ through the matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$ but also that each component y_i of $\mathbf{y}$ is disturbed by some random variable ε_i, which we express in matrix notation as

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \boldsymbol{\varepsilon}\ . \tag{2.25}$$

The model assumes that the noises ε_i have expected value $E(\varepsilon_i) = 0$, are *uncorrelated*, *i.e.*, satisfy $E(\varepsilon_i\varepsilon_j) = 0$ for $i \neq j$, and have the same (usually unknown) *variance* $\sigma^2 = E(\varepsilon_i^2) \geq 0$.

We seek an estimate $\hat{\mathbf{x}}$ for the unknown $\mathbf{x}$ that

(a) is *unbiased*, *i.e.*, satisfies $E(\hat{\mathbf{x}}) = \mathbf{x}$,
(b) depends linearly on the observation $\mathbf{y}$, *i.e.*, $\hat{\mathbf{x}} = \mathbf{Z}\mathbf{y}$ for some suitable matrix $\mathbf{Z} \in \mathbb{R}^{n\times m}$ (to be determined),
(c) minimizes $E(\|\hat{\mathbf{x}} - \mathbf{x}\|^2)$.

Because of the linearity of the expectation (which means that E is a linear function), we have

$$E(\hat{\mathbf{x}}) = E(\mathbf{Z}\mathbf{y}) = E(\mathbf{Z}\mathbf{A}\mathbf{x}) + E(\mathbf{Z}\boldsymbol{\varepsilon}) = \mathbf{Z}\mathbf{A}\mathbf{x}\ .$$

By (a), we want $E(\hat{\mathbf{x}}) = \mathbf{x}$. So $\mathbf{Z}$ should be a *generalized inverse* of $\mathbf{A}$, *i.e.*, satisfy $\mathbf{ZA} = \mathbf{I}$. In view of (c), $\mathbf{Z}$ should minimize the variance of the estimate $\hat{\mathbf{x}}$. Using again the linearity of expectation, we deduce

$$E(\|\hat{\mathbf{x}} - \mathbf{x}\|^2) = E(\|\mathbf{Zy} - \mathbf{x}\|^2) = E(\|\mathbf{Z}\varepsilon\|^2) = E(\varepsilon^T \mathbf{Z}^T \mathbf{Z}\varepsilon) = \sigma^2\, \mathbf{Z} \circ \mathbf{Z}\ ,$$

where the last equality follows from our assumption $E(\varepsilon_i \varepsilon_j) = 0$ for $i \neq j$ via

$$E(\varepsilon^T \mathbf{Z}^T \mathbf{Z}\varepsilon) = \sum_{i,j} \Big(\sum_k z_{ik} z_{jk}\Big) E(\varepsilon_i \varepsilon_j) = \sum_i \Big(\sum_k z_{ik} z_{ik}\Big) E(\varepsilon_i \varepsilon_i) = \sigma^2\, \mathbf{Z} \circ \mathbf{Z}\ .$$

Because σ^2 is fixed (although unknown), the problem of determining an unbiased linear estimator with least variance in the Gauss-Markov model reduces to minimizing the Frobenius norm $\|\mathbf{Z}\|_F = \sqrt{\mathbf{Z} \circ \mathbf{Z}}$ of the matrix $\mathbf{Z}$. So we want to solve the minimum norm problem

$$\min_{\mathbf{X} \in L} \|\mathbf{X}\|_F\ , \quad \text{where} \quad L = \{\mathbf{X} \in \mathbb{R}^{n \times m} \mid \mathbf{XA} = \mathbf{I}\}\ . \tag{2.26}$$

Identifying $\mathbf{X}$ with its $n \cdot m$-dimensional vector (x_{ij}), this is a quadratic problem (with $\mathbf{Q} = \mathbf{I}$) of type (2.23).

Ex. 2.14. *Observe that the constraints $\mathbf{X}_{i\cdot}\mathbf{A} = \mathbf{e}_i^T$ are "independent" of each other and conclude that (2.26) decomposes into n independent subproblems of type (2.23), each of dimension m.*

2.2.2. The Algorithm of Gram-Schmidt. Projections onto a subspace $W \subseteq \mathbb{R}^n$ are particularly easy to compute when $W = \operatorname{col} \mathbf{B}$, where $\mathbf{B} = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$ is such that the Gram matrix $\mathbf{G} = \mathbf{B}^T\mathbf{B} = (\langle \mathbf{a}_i | \mathbf{a}_j \rangle)$ is diagonal. Then formula (2.17) yields the coefficients $z_i = \langle \mathbf{a}_i | \mathbf{a}_i \rangle^{-1} \langle \mathbf{x} | \mathbf{a}_i \rangle$, $i = 1, \ldots, m$, for the projection $\hat{\mathbf{x}}$ of $\mathbf{x} \in \mathbb{R}^n$, *i.e.*,

$$\hat{\mathbf{x}} = \sum_{i=1}^m z_i \mathbf{a}_i = \sum_{i=1}^m \langle \mathbf{a}_i | \mathbf{a}_i \rangle^{-1} \langle \mathbf{x} | \mathbf{a}_i \rangle \mathbf{a}_i\ . \tag{2.27}$$

Let now $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{R}^n$ be arbitrary linearly independent vectors. We will construct vectors $\mathbf{b}_1, \ldots, \mathbf{b}_m$ that are pairwise orthogonal (in the sense that $\langle \mathbf{b}_i | \mathbf{b}_j \rangle = 0$ if $i \neq j$) such that for $k = 1, \ldots, m$,

$$V_k = \operatorname{span}\{\mathbf{a}_1, \ldots, \mathbf{a}_k\} = \operatorname{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}\ .$$

The procedure is straightforward. We start with $\mathbf{b}_1 = \mathbf{a}_1$. Assume we have already constructed $\mathbf{b}_1, \ldots, \mathbf{b}_k$. We then compute the projection $\hat{\mathbf{a}}_{k+1}$ of $\mathbf{a}_{k+1}$ onto V_k and take $\mathbf{b}_{k+1} = \mathbf{a}_{k+1} - \hat{\mathbf{a}}_{k+1}$. Since the $\mathbf{b}_1, \ldots, \mathbf{b}_k$ are pairwise orthogonal, the projection $\hat{\mathbf{a}}_{k+1}$ is easy to compute according to (2.27).

Gram-Schmidt
$\mathbf{b}_1 = \mathbf{a}_1$ and $k = 1$;
WHILE $k < m$ DO
$\mathbf{b}_{k+1} = \mathbf{a}_{k+1} - \sum_{i=1}^{k} \langle \mathbf{b}_i | \mathbf{b}_i \rangle^{-1} \langle \mathbf{a}_{k+1} | \mathbf{b}_i \rangle \mathbf{b}_i$;
$k \leftarrow k + 1$;

EX. 2.15. *Show (by induktion on k) that Gram-Schmidt produces pairwise orthogonal vectors* $\mathbf{b}_1, \ldots, \mathbf{b}_m$ *so that* $V_k = \text{span}\{\mathbf{b}_1, \ldots, \mathbf{b}_k\}$ *for* $k = 1, \ldots, m$. *Show furthermore:* $\langle \mathbf{b}_k | \mathbf{b}_k \rangle \leq \langle \mathbf{a}_k | \mathbf{a}_k \rangle$ *for* $k = 1, \ldots, m$.

EX. 2.16. *Extend the Gram-Schmidt algorithm to possibly linearly dependent vectors* $\mathbf{a}_1, \ldots, \mathbf{a}_m$. *(Hint: If* $\mathbf{a}_{k+1} \in V_k$, *set* $\mathbf{b}_{k+1} = \mathbf{0}$.*)*

It is instructive to look at the Gram-Schmidt algorithm from the point of view of matrix operations. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ be the matrix with rows $\mathbf{a}_1^T, \ldots, \mathbf{a}_m^T$. The algorithm of Gram-Schmidt then (just as Gaussian elimination) performs elementary row operations on $\mathbf{A}$ of the type

- Add (subtract) multiples of rows $1, \ldots, k$ to row $k + 1$.

Hence each iteration $k = 1, \ldots, m$ of Gram-Schmidt is achieved by multiplying $\mathbf{A}$ (from left) with a lower triangular $(m \times m)$-matrix $\mathbf{M}_k$ with all entries 1 on the diagonal. Letting $\mathbf{M}$ denote the product of the matrices $\mathbf{M}_k$, we obtain

$$\mathbf{B} = \mathbf{MA} = \mathbf{M}_m \ldots \mathbf{M}_1 \mathbf{A} \ ,$$

where $\mathbf{B}$ has rows $\mathbf{b}_1^T, \ldots, \mathbf{b}_m^T$ that are pairwise orthogonal. Note that each $\mathbf{M}_k$ has determinant 1. So the determinant multiplication rule says that also $\mathbf{M}$ has determinant 1. In the case where $\langle \cdot | \cdot \rangle$ is the standard inproduct $\langle \mathbf{x} | \mathbf{y} \rangle = \mathbf{x}^T \mathbf{y}$, this observation allows us to deduce the following estimate on the determinant of positive (semi-)definite matrices.

PROPOSITION 2.1 (Hadamard's Inequality)**.** *Let* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *be a matrix with row vectors* $\mathbf{a}_i^T$. *Then*

$$0 \leq \det(\mathbf{AA}^T) \leq \prod_{i=1}^{m} \mathbf{a}_i^T \mathbf{a}_i \ .$$

Proof. If the rows of $\mathbf{A}$ are linearly dependent, we have $\det \mathbf{AA}^T = 0$ and there is nothing to show. Otherwise we apply Gram-Schmidt to obtain $\mathbf{B} = \mathbf{MA}$ with $\det \mathbf{M} = \det \mathbf{M}^T = 1$ Hence

$$\det \mathbf{AA}^T = \det \mathbf{MAA}^T \mathbf{M}^T = \det \mathbf{BB}^T .$$

$\mathbf{B}$ has pairwise orthogonal rows $\mathbf{b}_i^T$. So the matrix $\mathbf{B}\mathbf{B}^T$ is a diagonal matrix with diagonal elements $\mathbf{b}_i^T\mathbf{b}_i \geq 0$. Hence (*cf.* Ex. 2.15),

$$0 \leq \det \mathbf{B}\mathbf{B}^T = \prod_{i=1}^{m} \mathbf{b}_i^T\mathbf{b}_i \leq \prod_{i=1}^{m} \mathbf{a}_i^T\mathbf{a}_i \ .$$

$\diamond$

2.2.3. Eigenvalues of Symmetric Matrices. Let $\mathbf{A} \in \mathbb{R}^{n\times n}$ be a square matrix. The number $\lambda \in \mathbb{R}$ is called an *eigenvalue of* $\mathbf{A}$ if there exists a vector $\mathbf{x} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ such that

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x} \quad (i.e., \ \mathbf{x} \in \ker(\mathbf{A} - \lambda\mathbf{I})) \ .$$

In other words, λ is an eigenvalue of $\mathbf{A}$ if $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$. The nonzero vectors $\mathbf{x} \in \ker(\mathbf{A} - \lambda\mathbf{I})$ are the *eigenvectors* corresponding to λ. Clearly, $\mathbf{x}$ is an eigenvector if and only if $\mathbf{x}' = \|\mathbf{x}\|^{-1}\mathbf{x}$ is an eigenvector. So we can restrict our attention to eigenvectors of unit length $\|\mathbf{x}\| = 1$.

Interest in eigenvalues and eigenvectors arises from the following consideration. Suppose there exists a basis $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of $\mathbb{R}^n$ with pairwise orthogonal eigenvectors of $\mathbf{A}$ of length $\|\mathbf{x}_i\| = 1$. Setting $\mathbf{Q} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, the orthogonality relations mean $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$, *i.e.*, $\mathbf{Q}^T = \mathbf{Q}^{-1}$, while the eigenvalue property yields the diagonalization $\mathbf{A}\mathbf{Q} = \mathbf{Q}\mathbf{D}$ or

$$\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \mathbf{D}, \quad \text{where } \mathbf{D} = \text{diag } (\lambda_1, \ldots, \lambda_n) \ .$$

REMARK. [SPECTRAL DECOMPOSITION]. The eigenvector basis $\mathbf{Q}$ implies in particular the so-called *spectral decomposition*

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i \mathbf{x}_i^T \tag{2.28}$$

of $\mathbf{A}$ as a (weighted) sum of the p.s.d. matrices $\mathbf{x}_i\mathbf{x}_i^T \in \mathbb{R}^{n\times n}$ of rank 1. The equality in (2.28) is straightforward to verify by checking for each basis vector

$$\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_j = \sum_{i=1}^{n} \lambda_i \mathbf{x}_i(\mathbf{x}_i^T\mathbf{x}_j) = \big(\sum_{i=1}^{n} \lambda_i \mathbf{x}_i\mathbf{x}_i^T\big)\mathbf{x}_j \ .$$

REMARK. Although eigenvalues are quite "natural" matrix parameters, not every (real) matrix admits (real) eigenvalues. Moreover, even when eigenvalues exist, they can*not* be calculated with elementary linear-algebraic operations and fall outside the realm of Gaussian elimination type methods. Already with positive definite matrices we face quadratic optimization problems subject to the nonlinear constraint $\|\mathbf{x}\| = 1$ (see (2.29) below). In this sense, Theorem 2.3 below is an "existence result". In practice, the numerical computation of eigenvalues is not exact.

We want to show that every *symmetric* matrix $\mathbf{A} \in \mathbb{S}^{n\times n}$ admits an orthogonal basis $\mathbf{Q}$ of eigenvectors. Observe first that the number

$$\lambda_1 = \min_{\|\mathbf{x}\|=1} \mathbf{x}^T\mathbf{A}\mathbf{x} \tag{2.29}$$

is well-defined since the continuous function $f(\mathbf{x}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$ attains its minimum on the compact set $\{\mathbf{x} \in \mathbb{R}^n \mid \|\mathbf{x}\| = 1\}$ (*cf.* Theorem 1.1). So there exists a vector $\mathbf{x}_1 \in \mathbb{R}^n$, $\|\mathbf{x}_1\| = 1$, such that $\lambda_1 = \mathbf{x}_1^T\mathbf{A}\mathbf{x}_1$. By definition of λ_1, we have $\mathbf{x}^T(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x} \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$, *i.e.*,

$$(\mathbf{A} - \lambda_1\mathbf{I}) \quad \text{is positive semidefinite} \quad \text{and} \quad \mathbf{x}_1^T(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x}_1 = 0\ .$$

Expressing the p.s.d. matrix as $\mathbf{A} - \lambda_1\mathbf{I} = \mathbf{C}^T\mathbf{C}$ (*cf.* Corollary 2.4), we find

$$0 = \mathbf{x}_1^T(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x}_1 = (\mathbf{C}\mathbf{x}_1)^T(\mathbf{C}\mathbf{x}_1) = \|\mathbf{C}\mathbf{x}_1\|^2$$

and hence $\mathbf{C}\mathbf{x}_1 = \mathbf{0}$. So $\mathbf{C}^T\mathbf{C}\mathbf{x}_1 = \mathbf{0}$ or, equivalently, $(\mathbf{A} - \lambda_1\mathbf{I})\mathbf{x}_1 = \mathbf{0}$, *i.e.*, $\mathbf{x}_1$ is an eigenvector of $\mathbf{A}$ with corresponding eigenvalue λ_1.

Starting with the eigenvector $\mathbf{x}_1$ corresponding to λ_1 we successively compute an orthonormal basis of eigenvectors $\mathbf{Q} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ as follows. We (arbitrarily) extend $\mathbf{x}_1$ to an orthonormal basis $\mathbf{Q}_1 = [\mathbf{x}_1, \mathbf{q}_2, \ldots, \mathbf{q}_n]$ and observe that

$$\mathbf{Q}_1^T\mathbf{A}\mathbf{Q}_1 = \begin{bmatrix} \lambda_1 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{A}_2 \end{bmatrix}$$

with a matrix $\mathbf{A}_2 \in \mathbb{S}^{(n-1)\times(n-1)}$. The same argument exhibits some orthogonal matrix $\overline{\mathbf{Q}}_2$ with $\overline{\mathbf{Q}}_2^T\mathbf{A}_2\overline{\mathbf{Q}}_2 = \begin{bmatrix} \lambda_2 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{A}_3 \end{bmatrix}$. Hence $\mathbf{Q}_2 = \begin{bmatrix} 1 & \mathbf{0}^T \\ \mathbf{0} & \overline{\mathbf{Q}}_2 \end{bmatrix}$ yields

$$\mathbf{Q}_2^T\mathbf{Q}_1^T\mathbf{A}\mathbf{Q}_1\mathbf{Q}_2 = \begin{bmatrix} \lambda_1 & 0 & \mathbf{0}^T \\ 0 & \lambda_2 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{0} & \mathbf{A}_3 \end{bmatrix}\ .$$

After (at most) n steps, the desired diagonalization is obtained by the orthogonal matrix $\mathbf{Q} = \mathbf{Q}_1 \cdots \mathbf{Q}_n$.

Summarizing, we have derived the following "spectral theorem".

THEOREM 2.3 (Spectral Theorem for Symmetric Matrices)**.** *Let $\mathbf{A} \in \mathbb{S}^{n\times n}$ be a symmetric matrix. Then there exists a matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ and eigenvalues $\lambda_1, \ldots, \lambda_n$ of $\mathbf{A}$ such that*

$$\mathbf{Q}^T\mathbf{Q} = \mathbf{I} \quad \textit{and} \quad \mathbf{Q}^T\mathbf{A}\mathbf{Q} = \text{diag}\ (\lambda_1, \ldots, \lambda_n)\ .$$

◇

REMARK. Our discussion exhibits the eigenvalue λ_1 of the symmetric matrix $\mathbf{A}$ as the optimal solution of the optimization problem

$$\min\ \lambda \quad \text{s.t.} \quad \mathbf{A} - \lambda\mathbf{I} \succeq \mathbf{0}\ . \tag{2.30}$$

This problem can in principle be solved approximately by using the Diagonalization algorithm (*cf.* Section 2.1.3): Suppose we have initial lower and upper bounds for λ_1, *i.e.*, $\underline{\lambda} \leq \lambda_1 \leq \overline{\lambda}$. We then approximate λ_1 by *binary search* :

```
WHILE  λ̄ − λ̲ > ε DO
BEGIN
        Let λ := (λ̄ + λ̲)/2.
        Check whether A − λI is p.s.d.
        If yes, update λ̲ := λ, otherwise λ̄ := λ .
END
```

In practice, other methods are more efficient (*e.g.*, the QR-algorithm *cf.* [33]).

2.3. Integer Solutions of Linear Equations

Often one may want to have solutions for systems of linear equations with each coordinate being an integer. This extra requirement adds some difficulty to the problem of solving linear equations. Consider, for example, the equation

$$3x_1 - 2x_2 \;=\; 1\ .$$

Gaussian Elimination will produce the rational solution $(x_1, x_2) = (1/3, 0)$ (or $(x_1, x_2) = (0, -1/2)$ if we re-order the variables) and miss the integral solution $(x_1, x_2) = (1, 1)$. Moreover, the example

$$2x = 1$$

shows that a linear equation may well have a rational solution while being infeasible with respect to the integer requirement. So we must approach the problem differently.

We assume that all coefficients of the linear equations we consider are rational. Hence we can multiply the equations by suitable integers so that we obtain an equivalent system with integral coefficients. The important point to make now comes from the following observation. For every $x_1, x_2 \in \mathbb{Z}$ and

$$b \;=\; a_1x_1 + a_2x_2\ ,$$

each divisor the integers a_1 and a_2 have in common must also divide b, or, to put it differently, b is an integral multiple of the greatest common divisor of a_1 and a_2. In fact, we have

LEMMA 2.2 (Euclid's Algorithm)**.** *Let $c = \gcd(a_1, a_2)$ be the greatest common divisor of the integers a_1 and a_2. Then*

$$L(a_1, a_2) := \{a_1\lambda_1 + a_2\lambda_2 \,|\, \lambda_1, \lambda_2 \in \mathbb{Z}\} \;=\; \{c\lambda \,|\, \lambda \in \mathbb{Z}\} =: L(c)\ .$$

Proof. We have already observed that every $b \in L(a_1, a_2)$ must be a multiple of $c = \gcd(a_1, a_2)$. So it suffices to show $c \in L(a_1, a_2)$, *i.e.*, it suffices to derive an explicit integer representation

$$c \;=\; a_1\lambda_1 + a_2\lambda_2\ .$$

We solve the latter problem with *Euclid's Algorithm*. The algorithm is based on the simple observation that, for any $k \in \mathbb{Z}$, we have

$$\gcd(a_1, a_2) = \gcd(a_1, a_2 - a_1) = \ldots = \gcd(a_1, a_2 - ka_1) \ .$$

Given a_1, a_2, we determine $\gcd(a_1, a_2)$ as follows. Assuming $|a_1| \leq |a_2|$, we first try $c = a_1$ as a candidate and check whether the quotient $\lambda = a_2/c = a_2/a_1$ is an integer. If yes, clearly $\gcd(a_1, a_2) = |c| = |a_1|$ holds and the algorithm stops.

If $\lambda \notin \mathbb{Z}$, we let $[\lambda] \in \mathbb{Z}$ denote the integer nearest to λ and write

$$a_2 = [\lambda]a_1 + \mu a_1,$$

noting

$$|\mu| = |\lambda - [\lambda]| \leq 1/2 \quad \text{and} \quad \mu a_1 = a_2 - [\lambda]a_1 \in \mathbb{Z} \ .$$

According to the basic observation above, it now suffices to determine

$$\gcd(a_1, \mu a_1) = \gcd(a_1, a_2 - [\lambda]a_1) \ .$$

Because $|\mu a_1| \leq |a_1|/2$, we try $c = \mu a_1$ as the next candidate for a greatest common divisor and proceed as before until the current λ satisfies $\lambda \in \mathbb{Z}$ (and hence $\mu = 0$).

Since the absolute value $|c|$ of our current candidate for $\gcd(a_1, a_2)$ is reduced by at least 50% in each iteration, the algorithm will stop after at most $\log|a_1|$ steps (log always denotes the logarithm to base 2) and output from the current c the result

$$|c| = \gcd(a_1, a_2) \ .$$

It is easy to update an expression for the current c as an integer combination

$$c = a_1\lambda_1 + a_2\lambda_2$$

of the original a_1 and a_2 because the parameters in each iteration are simple integer combinations of the parameters of the previous iteration and, hence, of the original a_1 and a_2.

$\diamond$

Euclid's Algorithm allows us to solve the integer equation

$$a_1x_2 + a_2x_2 = b$$

in a straightforward way. We first compute an integer representation for the greatest common divisor c of a_1 and a_2:

$$c = a_1\lambda_1 + a_2\lambda_2 \ .$$

If $\lambda = c^{-1}b \notin \mathbb{Z}$, then the equation does not have an integer solution. Otherwise the choice $x_1 = \lambda\lambda_1$ and $x_2 = \lambda\lambda_2$ yields a solution.

We now generalize Euclid's Algorithm with the goal of solving systems of linear equations in integer variables. To be specific, we assume that we are given n integral vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n \in \mathbb{Z}^m$ together with a prescribed right-hand-side vector $\mathbf{b} \in \mathbb{Z}^m$ and we want to find integers $x_j \in \mathbb{Z}$ such that

$$\mathbf{a}_1 x_1 + \mathbf{a}_2 x_2 + \ldots + \mathbf{a}_n x_n = \mathbf{b} \tag{2.31}$$

or assert that no integral solution of the system (2.31) of m linear equations exists.

REMARK. Seemingly more generally, we could admit rational data $\mathbf{a}_j, \mathbf{b} \in \mathbb{Q}^m$ for (2.31) as well. Multiplying then each equation in (2.31) by a suitable denominator, we easily obtain an equivalent problem with integral parameters.

Without loss of generality, we furthermore assume that the system (2.31) has full rank m (remove redundant equations) and that we have labeled the $\mathbf{a}_j$'s in such a way that the first m vectors $\mathbf{a}_1, \ldots, \mathbf{a}_m \in \mathbb{Z}^m$ are linearly independent.

Consider the set L of all feasible right-hand-sides $\mathbf{b}$ for (2.31), *i.e.,* all vectors that can be expressed as integral linear combinations of the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$:

$$L = L(\mathbf{a}_1, \ldots, \mathbf{a}_n) \;=\; \{\sum_{j=1}^{n} \mathbf{a}_j \lambda_j \,|\, \lambda_j \in \mathbb{Z}\} \;\subseteq \mathbb{R}^m \;.$$

L is said to be the *lattice* generated by the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$. It may happen that $L(\mathbf{a}_1, \ldots, \mathbf{a}_m)$ is a proper subset of $L(\mathbf{a}_1, \ldots, \mathbf{a}_n)$. Nevertheless, it turns out that one can find m linearly independent vectors $\mathbf{c}_1, \ldots, \mathbf{c}_m \in L(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ such that

$$L(\mathbf{c}_1, \ldots, \mathbf{c}_m) \;=\; L(\mathbf{a}_1, \ldots, \mathbf{a}_n) \;.$$

Such a set $C = \{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$ will be called a *lattice basis* for $L(\mathbf{a}_1, \ldots, \mathbf{a}_n)$. The key to our algorithmic approach for solving the system (2.31) will be the construction of a lattice basis.

Thinking of C as a matrix $\mathbf{C}$ with columns $\mathbf{c}_i$, we note that (2.31) has an integral solution if and only if $\mathbf{b} \in L(\mathbf{c}_1, \ldots, \mathbf{c}_m)$, *i.e.,* if and only if $\boldsymbol{\lambda} = \mathbf{C}^{-1}\mathbf{b} \in \mathbb{Z}^m$.

If $\mathbf{b} = \mathbf{C}\boldsymbol{\lambda}$ is a representation of $\mathbf{b}$ as an integral linear combination of the $\mathbf{c}_i$'s, and if we know how to express each $\mathbf{c}_i$ as an integral linear combination of the vectors $\mathbf{a}_1, \ldots, \mathbf{a}_n$, we can immediately compute an explicit integral solution $\mathbf{x}$ for (2.31).

The algorithm below constructs a lattice basis iteratively. In each iteration, we will be able to maintain an integral representation of the current $\mathbf{c}_i$'s in terms of the original $\mathbf{a}_j$'s. The next lemma is straightforward from the definitions. It tells us how to check whether $\{\mathbf{c}_1, \ldots, \mathbf{c}_m\}$ is a lattice basis.

LEMMA 2.3. *Let $\mathbf{c}_1, \ldots, \mathbf{c}_m \in L(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ be given vectors and consider $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_m]$. Then $L(\mathbf{c}_1, \ldots, \mathbf{c}_m) = L(\mathbf{a}_1, \ldots, \mathbf{a}_n)$ holds if and only if for all $j = 1, \ldots, n$, the linear equation*

$$\mathbf{C}\boldsymbol{\lambda} = \mathbf{a}_j$$

has an integral solution.

◇

The condition in Lemma 2.3 is easy to check if $\mathbf{C}$ yields a basis of $\mathbb{R}^m$ (and thus is invertible): We simply have to verify the property $\mathbf{C}^{-1}\mathbf{a}_j \in \mathbb{Z}^m$ for all $j = 1, \ldots, n$.

The algorithm for constructing a lattice basis now iterates two steps. The first step checks whether the current candidate basis is good. If some $\mathbf{C}^{-1}\mathbf{a}_j$ has a non-integral component, we modify our current basis in a second step similar to the adjustment of the candidate c in Euclid's Algorithm and return to the first step.

Lattice Basis

INIT: $\mathbf{C} = [\mathbf{c}_1, \ldots, \mathbf{c}_m] = [\mathbf{a}_1, \ldots, \mathbf{a}_m]$;
ITER: Compute $\mathbf{C}^{-1}$;
  If $\mathbf{C}^{-1}\mathbf{a}_j \in \mathbb{Z}^m$ for $j = 1, \ldots, n$, then STOP;
  If $\boldsymbol{\lambda} = \mathbf{C}^{-1}\mathbf{a}_j \notin \mathbb{Z}^m$ for some j, then
    Let $\mathbf{a}_j = \mathbf{C}\boldsymbol{\lambda} = \sum_{i=1}^m \lambda_i \mathbf{c}_i$ and compute
    $\mathbf{c} = \sum_{i=1}^m (\lambda_i - [\lambda_i])\mathbf{c}_i = \mathbf{a}_j - \sum_{i=1}^m [\lambda_i]\mathbf{c}_i$;
    Let k be the largest index i such that $\lambda_i \notin \mathbb{Z}$;
    Update $\mathbf{C}$ by replacing $\mathbf{c}_k$ with $\mathbf{c}$ in column k;
  NEXT ITERATION

Let us take a look at an iteration of the algorithm Lattice Basis. If $\mathbf{c}_1, \ldots, \mathbf{c}_m$ are elements of the lattice $L(\mathbf{a}_1, \ldots, \mathbf{a}_n)$, it is clear that $\mathbf{c}$ will also be a member of $L(\mathbf{a}_1, \ldots, \mathbf{a}_n)$. Moreover, if we have recorded how to express each of the vectors $\mathbf{c}_i$ as an integral linear combination of $\mathbf{a}_1, \ldots, \mathbf{a}_n$, then we can obtain such an explicit representation for $\mathbf{c}$ as well.

From these remarks, it is apparent that we can solve (2.31) if the algorithm Lattice Basis ever stops: We compute $\mathbf{y} = \mathbf{C}^{-1}\mathbf{b}$ and check whether $\mathbf{y} \in \mathbb{Z}^m$. If yes, substituting the $\mathbf{a}_j$'s for the $\mathbf{c}_i$'s in $\mathbf{b} = \mathbf{C}\mathbf{y}$ will produce the desired representation.

REMARK. In practical computation, it is preferable to solve $\mathbf{C}\boldsymbol{\lambda} = \mathbf{a}_j$ by Gaussian Elimination rather than to compute the inverse $\mathbf{C}^{-1}$ explicitly.

It remains to show that the algorithm is finite. In order to estimate the number of iterations of Lattice Basis, we follow the quantity

$$\Delta(\mathbf{c}_1, \ldots, \mathbf{c}_m) \;=\; |\det \mathbf{C}|$$

in each iteration. $\mathbf{C}$ will always be a basis of $\mathbb{R}^m$ and thus yields $|\det \mathbf{C}| > 0$. Since all coefficients in $\mathbf{C}$ are integers, we have $|\det \mathbf{C}| \in \mathbb{N}$ and hence conclude $|\det \mathbf{C}| \geq 1$.

Describing the replacement step in an iteration by matrix operations, we see that the update of $\mathbf{C}$ amounts to the computation of the matrix

$$\bar{\mathbf{C}} = \mathbf{C}\mathbf{M} \,,$$

where $\mathbf{M} = (\mu_{ij}) \in \mathbb{R}^{m\times m}$ is the matrix with

$$\mu_{ij} = \begin{cases} 1 & \text{if } i = j \neq k \\ \lambda_i - [\lambda_i] & \text{if } j = k \text{ and } i \leq k \\ 0 & \text{otherwise.} \end{cases}$$

Hence we obtain

$$|\det \bar{\mathbf{C}}| = |\det \mathbf{M}| \cdot |\det \mathbf{C}| = |\lambda_k - [\lambda_k]| \cdot |\det \mathbf{C}| \leq \frac{1}{2}\Delta(\mathbf{c}_1, \ldots, \mathbf{c}_m)$$

and thus conclude after K iterations

$$1 \leq \Delta(\mathbf{c}_1, \ldots, \mathbf{c}_m) \leq 2^{-K}\Delta(\mathbf{a}_1, \ldots, \mathbf{a}_m) ,$$

which implies the bound on the number of iterations:

$$K \leq \log \Delta(\mathbf{a}_1, \ldots, \mathbf{a}_m) .$$

EX. 2.17. *Compute integers $x, y, z \in \mathbb{Z}$ that solve the following system of linear equations:*

$$\begin{array}{rcrcrcr} 2x & + & 5y & + & 3z & = & 3 \\ 3x & + & 2y & + & z & = & -7 \end{array} .$$

The existence of lattice bases implies an integer analogue of Gale's Theorem for linear equations (see p. 27):

THEOREM 2.4. *Let $\mathbf{A} \in \mathbb{Z}^{m\times n}$ and $\mathbf{b} \in \mathbb{Z}^m$ be given. Then exactly one of the following statements is true:*

(a) *There exists some $\mathbf{x} \in \mathbb{Z}^n$ such that $\mathbf{Ax} = \mathbf{b}$.*
(b) *There exists some $\mathbf{y} \in \mathbb{R}^m$ such that $\mathbf{y}^T\mathbf{A} \in \mathbb{Z}^n$ and $\mathbf{y}^T\mathbf{b} \notin \mathbb{Z}$.*

Proof. If (a) is true, then $\mathbf{y}^T\mathbf{b} = \mathbf{y}^T\mathbf{Ax}$ for some $\mathbf{x} \in \mathbb{Z}^n$. So $\mathbf{y}^T\mathbf{A} \in \mathbb{Z}^n$ implies $\mathbf{y}^T\mathbf{b} \in \mathbb{Z}$, *i.e.*, (b) cannot be true.

If (a) is not true, then $\mathbf{C}^{-1}\mathbf{b} \notin \mathbb{Z}^m$, where we assume without loss of generality that $\mathbf{A}$ has full rank m and that $\mathbf{C}$ is a basis for the lattice generated by the columns of $\mathbf{A}$. In particular, $\mathbf{C}^{-1}$ contains a row $\mathbf{y}^T$, say, such that $\mathbf{y}^T\mathbf{b} \notin \mathbb{Z}$.

On the other hand, the fact that $\mathbf{C}$ is a lattice basis implies $\mathbf{C}^{-1}\mathbf{a}_j \in \mathbb{Z}^m$ for all $j = 1, \ldots, n$, *i.e.*, $\mathbf{C}^{-1}\mathbf{A} \in \mathbb{Z}^{m\times n}$. So, in particular, $\mathbf{y}^T\mathbf{A} \in \mathbb{Z}^n$ holds and proves statement (b) to be true.

◇

REMARK. One can show that lattice bases exist even if the vectors $\mathbf{a}_j$ are not rational (see Lekkerkerker [53]). So Theorem 2.4 remains true in this more general setting. However, our finiteness argument for the algorithm Lattice Bases will no longer be valid if the problem parameters are not rational. (This is no problem for practical applications, where the problem data are always rational).

2.4. Linear Inequalities

We now investigate the problem of computing a feasible vector $\mathbf{x} \in \mathbb{R}^n$ for a system $\mathbf{Ax} \leq \mathbf{b}$ of linear inequalities, with $\mathbf{A} = (a_{ij}) \in \mathbb{R}^{m\times n}$ and $\mathbf{b} = (b_i) \in \mathbb{R}^m$, which stands short for

$$\sum_{j=1}^{n} a_{ij}x_j \quad \leq \quad b_i \ , \qquad i = 1, \ldots, m \ . \tag{2.32}$$

We approach the problem with the same idea as in Gaussian elimination and eliminate variables one after the other until the system is either seen to be infeasible or a solution can be reconstructed *via* backward substitution. However, there is one important technical point to be observed:

- In the elementary row operations, only multiplications with *strictly positive scalars* are admitted.

REMARK. The restriction to operations with *positive* scalars comes from the fact that multiplication of an inequality with a negative scalar would *reverse* the inequality.

To see how we have to proceed, let us divide the i^{th} inequality in the system (2.32) by the positive number $|a_{i1}|$ whenever $a_{i1} \neq 0$, $i = 1, \ldots, m$, and investigate the equivalent system

$$\begin{aligned} x_1 + \quad & \sum_{j=2}^{n} a'_{rj}x_j \leq b'_r \ , \quad r = 1, \ldots, k \\ -x_1 + \quad & \sum_{j=2}^{n} a'_{sj}x_j \leq b'_s \ , \quad s = k+1, \ldots, \ell \\ & \sum_{j=2}^{n} a_{tj}x_j \leq b_t \ , \quad t = \ell+1, \ldots, m \end{aligned} \tag{2.33}$$

For clarity of the exposition, we assume here that the rows are indexed such that the first rows have coefficient $a_{i1} > 0$, then the rows with $a_{i1} < 0$ follow, and finally the rows with $a_{i1} = 0$ appear. Note that if either no $a_{i1} > 0$ or no $a_{i1} < 0$ occurs (*i.e.*, either $k = 0$ or $\ell = k$ in (2.33)), a solution $\mathbf{x} = (x_1, \ldots, x_n)$ is easily obtained recursively by solving the system of inequalities $t = \ell + 1, \ldots, m$ of (2.33) in the variables $x_2, \ldots, x_n$ and then chose x_1 sufficiently small resp. large so as to satisfy the inequalities involving x_1.

The first l inequalities in (2.33) can equivalently be written as

$$\min_{r=1,\ldots,k} \left(b'_r - \sum_{j=2}^{n} a'_{rj}x_j \right) \ \geq \ x_1 \ \geq \max_{s=k+1,\ldots,\ell} \left(\sum_{j=2}^{n} a'_{sj}x_j - b'_s \right). \tag{2.34}$$

with the understanding min $= +\infty$ if $k = 0$ (*i.e.*, there is no $a_{i1} > 0$) and max $= -\infty$ if $l = k$ (*i.e.*, there is no $a_{i1} < 0$).

Eliminating the variable x_1 in (2.34) and including the inequalities in which x_1 does not appear, we obtain the system

$$\begin{array}{rcll} \displaystyle\sum_{j=2}^{n} a'_{sj}x_j \;-\; b'_s & \leq & b'_r - \displaystyle\sum_{j=2}^{n} a'_{rj}x_j\,, & r = 1,\ldots,k;\ s = k+1,\ldots,\ell \\ \displaystyle\sum_{j=2}^{n} a_{tj}x_j & \leq & b_t\,, & t = \ell+1,\ldots,m. \end{array} \tag{2.35}$$

It is crucial to observe that for every feasible solution of (2.35) an x_1 can be found that satisfies the relation (2.34) because, by construction of the system (2.35), the min $\geq$ max property is guaranteed by all solutions. Re-ordering terms, we moreover see that (2.35) is equivalent with

$$\begin{array}{rcll} \displaystyle\sum_{j=2}^{n} (a'_{sj} + a'_{rj})x_j & \leq & b'_r + b'_s\,, & r = 1,\ldots,k;\ s = k+1,\ldots,\ell \\ \displaystyle\sum_{j=2}^{n} a_{tj}x_j & \leq & b_t\,, & t = \ell+1,\ldots,m. \end{array} \tag{2.36}$$

If $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ satisfies the system (2.33) then clearly $\mathbf{x}' = (x_2, \ldots, x_n)^T$ satisfies the linear inequality system (2.36) (or $\mathbf{A}'\mathbf{x}' \leq \mathbf{b}'$ for short). Moreover, whenever a vector $\mathbf{x}' = (x_2, \ldots, x_n)^T$ satisfies (2.36), then $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$ is feasible for (2.33) *if and only if* x_1 is chosen according to (2.34).

REMARK. (2.36) arises from (2.33) by adding the r and s rows in pairs. In particular, the system $\mathbf{A}'\mathbf{x}' \leq \mathbf{b}'$ in (2.36) can be understood to be of the form

$$\tilde{\mathbf{A}}\mathbf{x} \;=\; [\mathbf{0}, \mathbf{A}']\mathbf{x} \;\leq\; \mathbf{b}' \quad \text{(with } \mathbf{0} \text{ as the first column of } \tilde{\mathbf{A}}\text{)}.$$

As it was the case with Gaussian elimination, the preceding analysis says geometrically that the solutions of the system (2.36) are the projections of the solutions of the system (2.32) onto the variables $x_2, x_3, \ldots, x_n$. Iterating the construction, we thus observe:

THEOREM 2.5 (Projection Theorem)**.** *Let $P \subseteq \mathbb{R}^n$ be the set of feasible solutions of $\mathbf{Ax} \leq \mathbf{b}$. Then for all $k = 1, \ldots, n$, the projection*

$$P^{(k)} = \{(x_{k+1}, \ldots, x_n) \mid (x_1, \ldots, x_k, x_{k+1}, \ldots, x_n) \in P \text{ for suitable } x_i \in \mathbb{R}\}$$

is the solution set of a linear system $\mathbf{A}^{(k)}\mathbf{x}^{(k)} \leq \mathbf{b}^{(k)}$ in $n-k$ variables $\mathbf{x}^{(k)} = (x_{k+1}, \ldots, x_n)$.

◇

Given the inequality system (2.32), we can find a solution (or decide that no solution exists) recursively (eliminating one variable after the other until a solution can be obtained) *via* the following procedure.

Fourier-Motzkin
Compute a solution $(x_2, \ldots, x_n)$ for (2.36);
If no such solution exists, then STOP,
Compute x_1 *via* backward substitution from (2.34);

We note two important properties of Fourier-Motzkin elimination and its application to systems of linear inequalities:

- *Every* feasible solution of $\mathbf{Ax} \leq \mathbf{b}$ can, in principle, be obtained *via* suitable backward substitutions in the Fourier-Motzkin algorithm.
- If the coefficients of $\mathbf{A}$ and $\mathbf{b}$ are rational numbers, the Fourier-Motzkin algorithm will allow us to compute a solution with rational components (if $\mathbf{Ax} \leq \mathbf{b}$ is feasible at all).

EX. 2.18. *Eliminate x, y, z successively to solve the system*

$$\begin{array}{rcrcrcr} 3x & + & y & - & 2z & \leq & 1 \\ & - & 2y & - & 4z & \leq & -14 \\ -x & + & 3y & - & 2z & \leq & -2 \\ & & y & + & 4z & \leq & 13 \\ 2x & - & 5y & + & z & \leq & 0 \end{array}$$

REMARK. Fourier-Motzkin Elimination can be viewed as Gaussian Elimination with respect to the set of non-negative scalars. In contrast to Gaussian Elimination for linear equations, however, Fourier-Motzkin Elimination may increase the number of inequalities considerably in every elimination step. This is the reason why the Fourier-Motzkin algorithm is computationally not very efficient in general.

EX. 2.19. *Let m be the number of inequalities in the system (2.33). Establish the upper bound* $m' \leq m^2/4$ *on the number* m' *of inequalities in the system (2.36).*

The Satisfiability Problem. A fundamental model in artificial intelligence is concerned with *boolean functions* $\varphi : \{0,1\}^n \to \{0,1\}$. We consider such a function $\varphi = \varphi(x_1, \ldots, x_n)$ as a function of n logical (boolean) variables x_j and interpret the value "1" as "TRUE" and the value "0" as "FALSE". We say that φ is *satisfiable* if $\varphi(\mathbf{x}) = 1$ holds for at least one $\mathbf{x} \in \{0,1\}^n$. Every $\mathbf{x}$ with $\varphi(\mathbf{x}) = 1$ is called a *satisfying truth assignment* (as it assigns values to the logical variables that make φ become TRUE). Given a boolean function φ, we would like to find a satisfying truth assignment for φ (or decide that no such assignment exists).

Allowing also the *negation* $\bar{x}_j$ of a boolean variable x_j, it is well-known that each boolean function can be represented by a first order logic formula, or boolean formula, in *conjunctive normal form* (CNF). For example,

$$\varphi(x_1, x_2, x_3) = (x_1 \vee x_2) \wedge (\bar{x}_1 \vee x_2 \vee x_3) \wedge \bar{x}_3$$

is in CNF and has a satisfying truth assignment $\varphi(1, 1, 0) = 1$.

In other words, we can write $\varphi(x_1, \ldots, x_n)$ as a conjunction of *clauses* C_i, each of which is a disjunction of *literals*, namely unnegated and negated boolean variables. The *satisfiability problem* asks for an assignment that makes all clauses C_i of the system simultaneously TRUE.

The satisfiability problem for a CNF-system can be translated into the problem of solving a linear system in the $(0, 1)$-variables x_k , where each clause corresponds to an inequality of the system and the negated variable $\bar{x}_k$ is represented by $1 - x_k$. For example, the clause $C_i = x_2 \vee \bar{x}_5 \vee x_7$ is made TRUE if and only if we assign values 0 or 1 to the variables such that

$$x_2 + (1 - x_5) + x_7 \geq 1 \quad \text{i.e.} \quad -x_2 + x_5 - x_7 \leq 0 \ .$$

In general, we cannot solve the system by Fourier-Motzkin elimination since the Fourier-Motzkin solution may not be *binary*, *i.e.*, $x_k \in \{0, 1\}$ for all k (even if a binary solution exists). As a matter of fact, it is computationally difficult to find satisfying truth assignments for general CNF-systems (see Section **??**). However, there are classes of CNF-systems that can be solved efficiently.

Assume that $C_1 \wedge \ldots \wedge C_m$ is a CNF-formula in which each clause C_i consists of at most 2 literals. Then it is easy to compute a satisfying truth assignment (if one exists). Consider, for example, the variable x_k. If its negation $\bar{x}_k$ occurs in no clause, then we may set $x_k = 1$ and remove all clauses containing x_k from further consideration.

If there are clauses $C_1 = x_k \vee x_s$ and $C_2 = \bar{x}_k \vee x_t$, then a truth assignment satisfies C_1 and C_2 simultaneously if and only if it satisfies the so-called *resolution* $C = x_s \vee x_t$. That is, we can replace C_1 and C_2 by C and continue. The resolution step is equivalent with the elimination procedure in the Fourier-Motzkin algorithm when we eliminate x_k from the inequalities corresponding to C_1 and C_2. We add the inequalities

$$\begin{array}{rrrcr} -x_k & -x_s & & \leq & -1 \\ x_k & & -x_t & \leq & 0 \end{array}$$

in order to derive "$-x_s - x_t \leq -1$", which corresponds to C. Note that this resolution step does not increase the number of inequalities and hence gives rise to an efficient algorithm.

Ex. 2.20. *Use the Fourier-Motzkin algorithm to solve the satisfiability problem for the CNF-system with* $C_1 = x_1 \vee x_2$, $C_2 = \bar{x}_1 \vee x_3$, $C_3 = \bar{x}_1 \vee \bar{x}_2$, $C_4 = \bar{x}_3 \vee x_4$, $C_5 = x_1 \vee \bar{x}_4$.

2.4.1. Solvability of Linear Systems and Theorems of the Alternative. From a conceptual point of view, we can deal with the Fourier-Motzkin algorithm as we did with Gaussian Elimination. We imagine that the "eliminated" variable x_1 is actually still present in the derived system (2.36) but has coefficient 0. Moreover, the inequalities of the derived system are obtained according to the principle:

- Each inequality of (2.36) is a linear combination of the inequalities of (2.33) with non-negative scalars.

EX. 2.21. *Given the linear system* $\mathbf{Ax} \leq \mathbf{b}$*, we may consider arbitrary non-negative linear combinations of the inequalities* $\mathbf{A}_{i\cdot}\mathbf{x} \leq b_i$ *and obtain so-called* derived *inequalities of the form*

$$(\mathbf{y}^T\mathbf{A})\mathbf{x} \leq \mathbf{y}^T\mathbf{b}$$

for some $\mathbf{y} \geq \mathbf{0}$*. Show: Every non-negative linear combination of derived inequalities results again in a derived inequality.*

As a consequence of Ex. 2.21, we find:

- *Every* inequality in any iteration of the Fourier-Motzkin algorithm is of the form $(\mathbf{y}^T\mathbf{A})\mathbf{x} \leq \mathbf{y}^T\mathbf{b}$, where $\mathbf{y} \geq \mathbf{0}$.

Keeping this observation in mind, we arrive directly at the following generalization of Gale's Theorem on the solvability of linear equations, which is often referred to as the "Lemma of Farkas":

THEOREM 2.6 (Farkas Lemma)**.** *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{b} \in \mathbb{R}^m$ *be given. Then exactly one of the following alternatives is true:*

(I) $\mathbf{Ax} \leq \mathbf{b}$ *is feasible.*
(II) *There exists a vector* $\mathbf{y} \geq \mathbf{0}$ *such that* $\mathbf{y}^T\mathbf{A} = \mathbf{0}^T$ *and* $\mathbf{y}^T\mathbf{b} < 0$.

Proof. We apply Fourier-Motzkin elimination to the system $\mathbf{Ax} \leq \mathbf{b}$. After eliminating all variables we arrive at the system $\tilde{\mathbf{A}}\mathbf{x} \leq \tilde{\mathbf{b}}$ with coefficient matrix $\tilde{\mathbf{A}} = \mathbf{0}$, in which the ith inequality is of the type

$$0 = \mathbf{0}^T\mathbf{x} = [\mathbf{y}_i^T\mathbf{A}]\mathbf{x} \leq \mathbf{y}_i^T\mathbf{b} = \tilde{b}_i$$

for some vector $\mathbf{y}_i \geq \mathbf{0}$. This system is either trivially feasible (if $\tilde{b}_i \geq 0$ holds for all i) or infeasible. In the first case, (I) is true (and we can construct a feasible solution of $\mathbf{Ax} \leq \mathbf{b}$ via backward substitution). In the second case (II) holds (take $\mathbf{y} = \mathbf{y}_i$ if $\tilde{b}_i < 0$).

◇

EX. 2.22. *Show directly that (I) and (II) in Theorem 2.6 can not hold simultaneously.*

REMARK. It is usually quite straightforward to check computationally whether a given vector $\mathbf{x}$ is indeed a feasible solution of a given system of (linear or nonlinear) inequalities. In this sense, a system of inequalities has a "short proof" $\mathbf{x}$ for its feasibility. (Finding

such a "short proof" is, of course, a usually much more involved matter, see Chapter 8.) But how could one convince another person that a feasible solution does *not* exist? This question is generally very hard to answer. The Lemma of Farkas shows that *linear* systems enjoy the remarkable property of possessing also "short proofs" $\mathbf{y}$ for infeasibility.

EX. 2.23. *Find a vector* $\mathbf{y} \in \mathbb{R}^3$ *that exhibits the infeasibility of the system*

$$\begin{array}{rcrcrcr} x_1 & + & 2x_2 & + & 3x_3 & \leq & -1 \\ -2x_1 & + & x_2 & & & \leq & 2 \\ & & -5x_2 & - & 6x_3 & \leq & -1 \end{array} .$$

EX. 2.24. *Let* $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m, \mathbf{B} \in \mathbb{R}^{k\times n}, \mathbf{d} \in \mathbb{R}^k$ *be given. Show that exactly one of the following alternatives is true*

(I) $\mathbf{Ax} = \mathbf{b}, \mathbf{Bx} \leq \mathbf{d}$ *is feasible.*
(II) *There exist vectors* $\mathbf{u} \in \mathbb{R}^m$ *and* $\mathbf{v} \in \mathbb{R}^k$ *such that* $\mathbf{v} \geq \mathbf{0}$, $\mathbf{u}^T\mathbf{A} + \mathbf{v}^T\mathbf{B} = \mathbf{0}^T$, *and* $\mathbf{u}^T\mathbf{b} + \mathbf{v}^T\mathbf{d} < 0$.

In the same spirit, we can prove or disprove the existence of "non-trivial" solutions of systems of linear equations and inequalities. We give one example, where we use the notation

$$\mathbf{a} \ < \ \mathbf{b}$$

for any $\mathbf{a} = (a_j), \mathbf{b} = (b_j) \in \mathbb{R}^n$ that satisfy $a_j < b_j$ for all $j = 1, \ldots, n$.

COROLLARY 2.5 (Gordan)**.** *For every* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, exactly one of the following alternatives is true:*

(I) $\mathbf{Ax} = \mathbf{0}, \mathbf{x} \geq \mathbf{0}$ *has a* non-zero *solution.*
(II) $\mathbf{y}^T\mathbf{A} < \mathbf{0}^T$ *has a solution.*

Proof. If $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ satisfy $\mathbf{A}\bar{\mathbf{x}} = \mathbf{0}$, $\bar{\mathbf{x}} \geq \mathbf{0}$ and $\bar{\mathbf{y}}^T\mathbf{A} < \mathbf{0}^T$, we have $(\bar{\mathbf{y}}^T\mathbf{A})\bar{\mathbf{x}} = \bar{\mathbf{y}}^T(\mathbf{A}\bar{\mathbf{x}}) = 0$, and hence $\bar{\mathbf{x}} = \mathbf{0}$ because all components of $\bar{\mathbf{y}}^T\mathbf{A}$ are strictly negative. So (I) and (II) are mutually exclusive.

Assume now that (II) does not hold. Hence $\mathbf{A}^T\mathbf{y} \leq \mathbf{b}$ is infeasible for the particular choice $\mathbf{b} = -\mathbf{1}$. So the Lemma of Farkas guarantees the existence of a vector $\mathbf{x} \geq \mathbf{0}$ such that both $\mathbf{x}^T\mathbf{A}^T = \mathbf{0}^T$, *i.e.*, $\mathbf{Ax} = \mathbf{0}$, and $\mathbf{x}^T\mathbf{b} < 0$ (and hence $\mathbf{x} \neq \mathbf{0}$) are satisfied, which implies that statement (I) is true.

◇

REMARK. The results of Gordan [35] actually pre-date and imply the results of Farkas [20]. As we have seen, both are consequences of the Fourier-Motzkin algorithm that is essentially due to Fourier [26] even earlier (see also Motzkin [60]).

Stochastic Matrices. We illustrate the power of theorems of the alternative with an application in stochastics. A *probability distribution* on the finite set $S = \{1, 2, \ldots, n\}$ is a vector $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_n)^T$ such that

$$\pi_i \geq 0 \qquad \text{for all } i \in S \qquad \text{and} \qquad \sum_{i=1}^{n} \pi_i = 1 \ .$$

The matrix $\mathbf{P} = (p_{ij}) \in \mathbb{R}^{n\times n}$ is said to be *stochastic* if each row vector $\mathbf{P}_{i\cdot} = (p_{i1}, \ldots, p_{in})$ of P is a probability distribution.

Thinking of S as a system of *states* and of p_{ij} as the *transition probability* of the system to pass into state j, given it is in state i,

$$\pi'_k = \sum_{i=1}^{n} p_{ik}\pi_i$$

is the probability for the system to move into state k, assuming it is currently in state i with probability π_i. So

$$\boldsymbol{\pi}' \;=\; \mathbf{P}^T\boldsymbol{\pi}$$

is the probability distribution of the states after one transition.

Stochastic matrices arise in the study of random walks and Markov chains (see, *e.g.*, [21] for more details). A fundamental property of any stochastic matrix $\mathbf{P} = (p_{ij})$ is the existence of a *steady state distribution*, namely a probability distribution $\boldsymbol{\pi}$ such that $\boldsymbol{\pi} = \mathbf{P}^T\boldsymbol{\pi}$. To prove this fact, it is convenient to use matrix notation. Where $\mathbf{I}$ is the identity, we must prove that the system

$$(\mathbf{P} - \mathbf{I})^T\mathbf{x} = \mathbf{0},\ \mathbf{x} \geq \mathbf{0}$$

has a feasible solution $\mathbf{x} \neq \mathbf{0}$. Setting $\lambda = \sum_i x_i > 0$, the vector $\pi = \lambda^{-1}\mathbf{x}$ then yields the desired steady state distribution.

By Gordan's Theorem (Corollary 2.5), it suffices to show that the associated "dual" system

$$(\mathbf{P} - \mathbf{I})\mathbf{y} < \mathbf{0} \quad \text{or, equivalently,} \quad \mathbf{P}\mathbf{y} < \mathbf{y} \tag{2.37}$$

is infeasible. Consider therefore a potential feasible vector $\mathbf{y}^T = (y_1, \ldots, y_n)$. Let y_k be a smallest component of $\mathbf{y}$. From (2.37), we would then deduce

$$y_k > \sum_{j=1}^{n} p_{kj}y_j \geq y_k \sum_{j=1}^{n} p_{kj} = y_k\ ,$$

which is impossible.

2.4.2. Implied Inequalities. We say that an inequality $\mathbf{c}^T\mathbf{x} \leq z$ is *implied* by $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ if for all $\mathbf{x} \in \mathbb{R}^n$,

$$\mathbf{A}\mathbf{x} \leq \mathbf{b} \qquad \text{implies} \qquad \mathbf{c}^T\mathbf{x} \leq z$$

(in other words: there is no solution of $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ with $\mathbf{c}^T\mathbf{x} > z$). If $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ is infeasible then, by definition, *every* inequality is implied. Hence we will always assume in the following that $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ is feasible.

Implied inequalities are easily constructed: Multiplication of $\mathbf{A}_{i\cdot}\mathbf{x} \leq b_i$ with a scalar multiplier $y_i \geq 0$ yields the implied inequality $y_i\mathbf{A}_{i\cdot}\mathbf{x} \leq y_i b_i$. Adding all these inequalities, we obtain the implied inequality

$$(\mathbf{y}^T\mathbf{A})\mathbf{x} \leq \mathbf{y}^T\mathbf{b}\ .$$

Increasing the right hand side to any value $z \geq \mathbf{y}^T\mathbf{b}$ yields, of course, again an implied inequality $(\mathbf{y}^T\mathbf{A})\mathbf{x} \leq z$.

A fundamental and exceedingly useful property of linear systems is the converse: *Every* implied inequality arises this way. This statement is equivalent with Theorem 2.6 and also occasionally referred to as the "Lemma of Farkas".

COROLLARY 2.6 (Farkas)**.** *Assume that* $\mathbf{Ax} \leq \mathbf{b}$ *is feasible. Then the inequality* $\mathbf{c}^T\mathbf{x} \leq z$ *is implied by* $\mathbf{Ax} \leq \mathbf{b}$ *if and only if there exists a non-negative vector* $\mathbf{y} \geq \mathbf{0}$ *such that*

$$\mathbf{c}^T = \mathbf{y}^T\mathbf{A} \quad \textit{and} \quad \mathbf{y}^T\mathbf{b} \leq z. \tag{2.38}$$

Proof. We have seen that condition (2.38) is sufficient. To show the necessity, suppose that (2.38) has no non-negative solution. We claim that then $\mathbf{c}^T\mathbf{x} \leq z$ is not implied by $\mathbf{Ax} \leq \mathbf{b}$, *i.e.*, that there exists a solution of $\mathbf{Ax} \leq \mathbf{b}$ and $\mathbf{c}^T\mathbf{x} > z$. So suppose that the system (2.38) has no solution $\mathbf{y} \geq \mathbf{0}$, *i.e.*,

$$\begin{array}{rcl} \mathbf{A}^T\mathbf{y} & = & \mathbf{c} \\ \mathbf{b}^T\mathbf{y} & \leq & z \\ -\mathbf{I}\mathbf{y} & \leq & \mathbf{0} \end{array} \tag{2.39}$$

is infeasible. Then Theorem 2.6 implies (*cf.* Ex. 2.24) that the associated alternative system

$$\begin{array}{rcrcrcl} \mathbf{v}^T\mathbf{A}^T & + & u\mathbf{b}^T & - & \mathbf{w}^T\mathbf{I} & = & \mathbf{0}^T \\ \mathbf{v}^T\mathbf{c} & + & uz & & & < & 0 \end{array} \tag{2.40}$$

has a feasible solution $(\mathbf{v}, u, \mathbf{w})$ with $u \geq 0$ and $\mathbf{w} \geq \mathbf{0}$. We consider the two cases $u \neq 0$ and $u = 0$. Now $u > 0$ implies that $\bar{\mathbf{x}} = -u^{-1}\mathbf{v}$ satisfies $\mathbf{A}\bar{\mathbf{x}} \leq \mathbf{b}$ and $\mathbf{c}^T\bar{\mathbf{x}} > z$, which proves the claim.

If $u = 0$ then $\mathbf{Av} \geq \mathbf{0}$ and $(-\mathbf{c}^T\mathbf{v}) > 0$. Choosing some feasible solution $\mathbf{x}_0$ of $\mathbf{Ax} \leq \mathbf{b}$, we set $\mathbf{x}_t = \mathbf{x}_0 - t\mathbf{v}$ and find for $t > 0$,

$$\mathbf{Ax}_t \leq \mathbf{b} - t\mathbf{Av} \leq \mathbf{b} \quad \text{and} \quad \mathbf{c}^T\mathbf{x}_t = \mathbf{c}^T\mathbf{x}_0 - t\mathbf{c}^T\mathbf{v}\,.$$

Hence, for $t > 0$ sufficiently large, we obtain again $\mathbf{Ax}_t \leq \mathbf{b}$ and $\mathbf{c}^T\mathbf{x}_t > z$.

◇

EX. 2.25. *Show:* $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ *is bounded from above on* $P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} \leq \mathbf{0}\}$ *if and only if the inequality* $\mathbf{c}^T\mathbf{x} \leq 0$ *is implied by* $\mathbf{Ax} \leq \mathbf{0}$.

Corollary 2.6 allows us to check whether $\mathbf{c}^T\mathbf{x} \leq z$ is implied by $\mathbf{Ax} \leq \mathbf{b}$ by testing the feasibility of the system

$$\mathbf{y} \geq 0\,, \quad \mathbf{y}^T\mathbf{A} = \mathbf{c}^T\,, \quad \mathbf{y}^T\mathbf{b} \leq z\,.$$

Similarly, we can identify redundancies in a system of linear inequalities (*cf.* Ex. 2.26) and, more importantly, characterize optimal solutions of linear optimization problems (*cf.* Ex. 2.28).

Ex. 2.26. *An inequality* $\mathbf{A}_{i\cdot}\mathbf{x} \leq b_i$ *of* $\mathbf{Ax} \leq \mathbf{b}$ *is called* redundant *if its removal does not affect the set of feasible solutions. Explain how redundancy can be tested with the help of Fourier-Motzkin. Show by example: If* $\mathbf{A}_{i\cdot}\mathbf{x} \leq b_i$ *and* $\mathbf{A}_{j\cdot}\mathbf{x} \leq b_j$ *are both redundant, then removing* both *inequalities simultaneously may well alter the set of feasible solutions.*

Ex. 2.27. *Show by example that the hypothesis "*$\mathbf{Ax} \leq \mathbf{b}$ *is feasible" cannot be dropped in Corollary 2.6.*

Ex. 2.28. *In a* linear optimization problem *we are to maximize a linear function* $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ *on the set* $P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} \leq \mathbf{b}\}$. *Let* $\overline{\mathbf{x}} \in P$ *be given. Show:* $f(\overline{\mathbf{x}}) = \mathbf{c}^T\overline{\mathbf{x}} = \overline{z}$ *is optimal if and only if* $\mathbf{Ax} \leq \mathbf{b}$ *implies* $\mathbf{c}^T\mathbf{x} \leq \overline{z}$.
Conclude: The optimal solutions $\overline{\mathbf{x}}$ *(if they exist) arise precisely from the feasible solutions of the linear system (in the variables* $\mathbf{x}$, $\mathbf{y}$, z*):*

$$\mathbf{Ax} \leq \mathbf{b}\,, \quad \mathbf{c}^T\mathbf{x} = z\,, \quad \mathbf{y}^T\mathbf{A} = \mathbf{c}^T\,, \quad \mathbf{y} \geq \mathbf{0}\,, \quad \mathbf{y}^T\mathbf{b} \leq z\,.$$

CHAPTER 3

Polyhedra

A *polyhedron* $P \subseteq \mathbb{R}^n$ is, by definition, the solution set of some system $\mathbf{Ax} \leq \mathbf{b}$ of linear inequalities:

$$P = P(\mathbf{A}, \mathbf{b}) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} \leq \mathbf{b}\} .$$

In this chapter we study polyhedra as geometrical objects with the goal of providing some geometric intuition for the "algebraic" results about linear inequality systems in Chapter 2 (*e.g.*, the Farkas Lemma). Polyhedra can be looked at from two different ("dual") points of view. So this chapter also introduces the concept of *duality* for polyhedra. The duality principle will also play a fundamental role in our analysis of (linear and nonlinear) optimization problems in later chapters.

3.1. Polyhedral Cones and Polytopes

Geometrically speaking, an inequality $\mathbf{a}^T\mathbf{x} \leq \beta$ with $\mathbf{a} \neq \mathbf{0}$ defines a *halfspace* $H^{\leq} = P(\mathbf{a}^T, \beta) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} \leq \beta\}$ with

$$H = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}^T\mathbf{x} = \beta\}$$

as its associated *hyperplane*. Hence a polyhedron $P = P(\mathbf{A}, \mathbf{b}) \subseteq \mathbb{R}^n$ is the intersection of finitely many halfspaces. (In particular, $P = \mathbb{R}^n$ is the empty intersection of halfspaces).

Ex. 3.1. *Show by example that different inequality systems may define the same polyhedron.*

Every hyperplane $H \subseteq \mathbb{R}^n$ and, more generally, every linear or affine subspace $L \subseteq \mathbb{R}^n$ is the solution set of a linear system of inequalities and hence a polyhedron. The following type of polyhedron is of particular interest: A *polyhedral cone* is a polyhedron of the form $P = P(\mathbf{A}, \mathbf{0})$ (see also Ex. 3.2).

Linear Subspaces. To motivate the structural analysis of polyhedra, let us first take a look at the familiar case of linear subspaces (which form a particular class of polyhedral cones). A linear subspace $L \subseteq \mathbb{R}^n$ can be represented in two conceptually different ways as

$$L = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} = \mathbf{0}\} \quad \text{or} \quad L = \text{span}\,\{\mathbf{v}_1, \ldots, \mathbf{v}_k\},$$

where $\mathbf{A}$ is a suitable matrix and the vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in \mathbb{R}^n$ generate $L = \ker A$. We refer to these two as *implicit* resp. *explicit* representations of L. Both have

their advantages: The implicit representation allows us to check easily whether a given $\mathbf{x} \in \mathbb{R}^n$ belongs to L (by evaluating $\mathbf{Ax}$), while the explicit representation enables us to produce elements $\mathbf{x} = \Sigma\lambda_i \mathbf{v}_i \in L$ as linear combinations of the generators $\mathbf{v}_i$. .

Our definition of a polyhedron is based on the implicit representation $P = P(\mathbf{A}, \mathbf{b})$. We want to establish an explicit representation in terms of convex and conic hulls (Theorem 3.3 below).

Conic and Convex Hulls. A non-empty subset $S \subseteq \mathbb{R}^n$ is a *(convex) cone* if for scalars $\lambda_1, \lambda_2 \geq 0$,

$$\mathbf{x}_1, \mathbf{x}_2 \in S \implies \lambda_1 \mathbf{x} + \lambda_2 \mathbf{x}_2 \in S \, .$$

It is straightforward to check that a polyhedral cone $P(\mathbf{A}, \mathbf{0})$ is a cone.

Ex. 3.2. *Show: $P \subseteq \mathbb{R}^n$ is a polyhedral cone if and only if P is a polyhedron and a cone.*

REMARK. Some textbooks use the term "cone' for subsets $S \subseteq \mathbb{R}^n$ with the property "$\mathbf{x} \in S \Rightarrow \lambda \mathbf{x} \in S$ for all $\lambda \geq 0$". For us, however, a *cone* is always a *(convex) cone* as defined above.

A set $S \subseteq \mathbb{R}^n$ is *convex* if for scalars $\lambda_1, \lambda_2 \geq 0$, $\lambda_1 + \lambda_2 = 1$,

$$\mathbf{x}_1, \mathbf{x}_2 \in S \Rightarrow \lambda_1 \mathbf{x}_1 + \lambda_2 \mathbf{x}_2 \in S$$

In other words

$$\mathbf{x}_1, \mathbf{x}_2 \in S \Rightarrow (1 - \lambda)\mathbf{x}_1 + \lambda \mathbf{x}_2 \in S \qquad \text{for all } \lambda \in [0, 1].$$

Geometrically this means: If $\mathbf{x}_1, \mathbf{x}_2 \in S$ then S contains the whole *line segment* $[\mathbf{x}_1, \mathbf{x}_2] = \{(1 - \lambda)\mathbf{x}_1 + \lambda \mathbf{x}_2 \mid \lambda \in [0, 1]\}$.

Ex. 3.3. *Let $S \subseteq \mathbb{R}^n$ be a non-empty set. Show:*
(a) *S is a cone if and only if $\lambda_1 \mathbf{x}_1 + \ldots + \lambda_k \mathbf{x}_k \in S$ for all $k \geq 1$, $\mathbf{x}_1, \ldots, \mathbf{x}_k \in S$ and $\lambda_1, \ldots, \lambda_k \geq 0$.*
(b) *S is a convex set if and only if $\lambda_1 \mathbf{x}_1 + \ldots + \lambda_k \mathbf{x}_k \in S$ for all $k \geq 1$, $\mathbf{x}_1, \ldots, \mathbf{x}_k \in S$ and $\lambda_1, \ldots, \lambda_k \geq 0$ with $\lambda_1 + \ldots + \lambda_k = 1$.*
(Hint: Induction on k).

Clearly, intersections of convex sets (cones) are convex sets (cones) again. Hence for an arbitrary $S \subseteq \mathbb{R}^n$ we may define its *convex hull* conv S resp. its *conic hull* cone S as the smallest convex resp. conic set containing S. Explicitly, these sets are given by (*cf.* Ex. 3.4):

$$\text{cone } S \;:=\; \left\{ \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mid \mathbf{x}_i \in S,\ \lambda_i \geq 0,\ k \in \mathbb{N} \right\}$$

$$\text{conv } S \;:=\; \left\{ \sum_{i=1}^{k} \lambda_i \mathbf{x}_i \mid \mathbf{x}_i \in S,\ \lambda_i \geq 0, \sum_{i=1}^{k} \lambda_i = 1,\ k \in \mathbb{N} \right\} .$$

EX. 3.4. *Show for an arbitrary $S \subseteq \mathbb{R}^n$: The above given sets* conv *S resp.* cone *S are the smallest convex resp. conic sets containing S.*

The case of a finite set $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_k\} \subset \mathbb{R}^n$ is of particular interest. Letting $\mathbf{S} = [\mathbf{s}_1 \ldots \mathbf{s}_k] \in \mathbb{R}^{n \times k}$ be the matrix with columns $\mathbf{s}_i$, we also write

$$\begin{aligned} \text{cone } \mathbf{S} &= \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = \mathbf{S}\boldsymbol{\lambda},\ \boldsymbol{\lambda} \geq \mathbf{0}\} \\ \text{conv } \mathbf{S} &= \left\{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = \mathbf{S}\boldsymbol{\mu},\ \boldsymbol{\mu} \geq \mathbf{0}, \mathbf{1}^T\boldsymbol{\mu} = 1\right\} . \end{aligned}$$

If $|S| < \infty$, we say that cone S and conv S are *finitely generated*. A finitely generated set conv S is said to be a *polytope*. A polytope is always *bounded, i.e.*, there exists a number $r \in \mathbb{R}$ such that $\|\mathbf{x}\| \leq r$ holds for all $\mathbf{x} \in \text{conv } \{\mathbf{s}_1, \ldots, \mathbf{s}_k\}$. Indeed, the triangle inequality yields

$$\|\mu_1\mathbf{s}_1 + \ldots + \mu_k\mathbf{s}_k\| \leq \|\mathbf{s}_1\| + \ldots + \|\mathbf{s}_k\| = r$$

for all $0 \leq \mu_1, \ldots, \mu_k \leq 1$.

For arbitrary sets $A, B \subseteq \mathbb{R}^n$ the *Minkowski sum* is defined as

$$A + B = \{\mathbf{a} + \mathbf{b} \mid \mathbf{a} \in A,\ \mathbf{b} \in B\} .$$

We will show that each polyhedron $P \subseteq \mathbb{R}^n$ allows an explicit representation as a Minkowski sum $P = \text{conv } V + \text{cone } W$ with finite sets $V, W \subset \mathbb{R}^n$. We first establish the converse.

THEOREM 3.1 (Weyl)**.** *Let $V, W \subset \mathbb{R}^n$ be finite sets. Then*

$$P = \text{conv } V + \text{cone } W$$

is a polyhedron. In particular, the polytope conv *V is a polyhedron and the finitely generated cone* cone *W is a polyhedral cone.*

Proof. Assuming $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ and $W = \{\mathbf{w}_1, \ldots, \mathbf{w}_\ell\}$, consider the system of linear equations and inequalities:

$$\begin{aligned} \mathbf{z} &= \mathbf{v} + \mathbf{w} \\ \mathbf{v} &= \lambda_1\mathbf{v}_1 + \ldots + \lambda_k\mathbf{v}_k \\ \mathbf{w} &= \mu_1\mathbf{w}_1 + \ldots + \mu_\ell\mathbf{w}_\ell \\ \mathbf{1}^T\boldsymbol{\lambda} &= 1 \\ \boldsymbol{\lambda} &\geq \mathbf{0},\ \boldsymbol{\mu} \geq \mathbf{0} \end{aligned} \tag{3.1}$$

in variables $\mathbf{z}, \mathbf{v}, \mathbf{w}, \boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. Clearly, P is the projection of the set $\bar{P}$ of all feasible solutions in the $(\mathbf{z}, \mathbf{v}, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\mu})$-space onto the variables $\mathbf{z}$. Hence, P is a polyhedron by the Projection Theorem (Theorem 2.5).

The second statement follows from the special cases $W = \{\mathbf{0}\}$ and $V = \{\mathbf{0}\}$ respectively (*cf.* Ex. 3.2).

◇

REMARK. $P = \text{conv } V + \text{cone } W$ is an explicit representation of the polyhedron P. The proof of Weyl's Theorem shows that an implicit representation $P = P(\mathbf{A}, \mathbf{b})$ can

be obtained by applying Fourier-Motzkin elimination to the system (3.1) (eliminating all variables except $\mathbf{z}$).

Ex. 3.5. *Define the (k-dimensional)* standard cone *as* $\mathbb{R}^k_+ = \{\mathbf{x} \in \mathbb{R}^k \mid \mathbf{x} \geq \mathbf{0}\}$ *and the* standard simplex *in* $\mathbb{R}^k$ *as* $\Delta_k = \{\boldsymbol{\mu} \in \mathbb{R}^k \mid \boldsymbol{\mu} \geq \mathbf{0},\ \mathbf{1}^T\boldsymbol{\mu} = 1\}$. *Show:*
(a) $C \subseteq \mathbb{R}^n$ *is a finitely generated cone if and only if there is some* k *and a linear map* $f : \mathbb{R}^k \to \mathbb{R}^n$ *such that* $C = f(\mathbb{R}^k_+)$.
(b) $P \subseteq \mathbb{R}^n$ *is a polytope if and only if there is some* k *and a linear map* $f : \mathbb{R}^k \to \mathbb{R}^n$ *such that* $P = f(\Delta_k)$.

Ex. 3.6. *Show: The Minkowski sum* $P + Q$ *of the polyhedra* $P, Q \subseteq \mathbb{R}^n$ *is a polyhedron (Hint: Use the Projection Theorem).*

Ex. 3.7. *Show for the finite set* $V \subseteq \mathbb{R}^n$*:*
$$\text{conv}\,(V \cup \{\mathbf{0}\}) = \text{conv}\,[\mathbf{V}, \mathbf{0}] = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} = \mathbf{V}\boldsymbol{\lambda}, \mathbf{1}^T\boldsymbol{\lambda} \leq 1, \boldsymbol{\lambda} \geq \mathbf{0}\}.$$

Ex. 3.8. *Show that an affine map* $f(\mathbf{x}) = \mathbf{B}\mathbf{x} + \mathbf{d}$ *(*$\mathbf{B} \in \mathbb{R}^{m\times n}$ *and* $\mathbf{d} \in \mathbb{R}^m$*) maps each polyhedron* $P \subseteq \mathbb{R}^n$ *to a polyhedron* $P' = f(P) \subseteq \mathbb{R}^m$*. (Hint:* cf. *the proof of Weyl's Theorem 3.1.)*

Separating Hyperplanes. An inequality $\mathbf{c}^T\mathbf{x} \leq \gamma$ is said to be *valid* for $S \subseteq \mathbb{R}^n$ if $\mathbf{c}^T\mathbf{x} \leq \gamma$ holds for all $\mathbf{x} \in S$, *i.e.*, $S \subseteq H^{\leq} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} \leq \gamma\}$. If $S \subseteq H^{\leq}$ and the point $\mathbf{v} \in \mathbb{R}^n$ is not contained in $H^{\leq}$, we say that $H^{\leq}$ (or $\mathbf{c}^T\mathbf{x} \leq \gamma$) *separates* $\mathbf{v}$ from S and call
$$H = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} = \gamma\}$$
a *separating hyperplane*.

Ex. 3.9. *Give an example of a (convex) set* $S \subseteq \mathbb{R}^n$ *and a point* $\mathbf{v} \in \mathbb{R}^n \setminus S$ *that cannot be separated from* S *by a hyperplane.*

Let us illustrate the Farkas Lemma geometrically and point out its relation to Weyl's Theorem. Consider a polyhedron $P = P(\mathbf{A}, \mathbf{b})$ and $\mathbf{v} \in \mathbb{R}^n$. By definition, either $\mathbf{v} \in P$ or $\mathbf{v}$ can be separated from P (by some inequality $\mathbf{A}_{i\cdot}\mathbf{x} \leq b_i$). This is particularly true for $P = \text{cone}\ W$ (which is a polyhedral cone by Weyl's Theorem). In other words, given a finite set $W \subseteq \mathbb{R}^n$ and a vector $\mathbf{v} \in \mathbb{R}^n$, exactly one of the following holds:

(I) $\mathbf{v} \in \text{cone}\ W$
(II) $\mathbf{v}$ can be separated from cone W (and hence from W) by an inequality $\mathbf{a}^T\mathbf{x} \leq 0$.

"Algebraically", those two alternatives take the form

(I) $\mathbf{W}\boldsymbol{\lambda} = \mathbf{v},\ \boldsymbol{\lambda} \geq \mathbf{0}$ is feasible.
(II) There exists some $\mathbf{a} \in \mathbb{R}^n$ such that $\mathbf{a}^T\mathbf{W} \leq \mathbf{0}^T,\ \mathbf{a}^T\mathbf{v} > 0$,

which is exactly the Farkas Lemma (Theorem 2.6).

REMARK. Separating hyperplanes play an important role in mathematical programming, *e.g.*, as "cutting planes" in Chapter 9 as well as in the ellipsoid method of Chapter 10. Moreover, in Chapter 10, the closed convex sets are characterized as exactly those sets $S \subseteq \mathbb{R}^n$ that are intersections of (possibly infinitely many) halfspaces, (*i.e.*, every $\mathbf{v} \notin S$ can be separated from S).

3.2. Cone Duality

Each vector $\mathbf{c} \in \mathbb{R}^n$ corresponds to a unique real-valued linear function f *via* $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ (and conversely). Similarly, there is a one-to-one correspondence between points $\mathbf{c} \in \mathbb{R}^n$ with $\|\mathbf{c}\| = 1$ and hyperplanes $H = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} = 0\}$. More generally, each linear subspace $L \subseteq \mathbb{R}^n$ is uniquely determined by its orthogonal complement

$$L^\perp = \{\mathbf{c} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} = 0 \quad \text{for all } \mathbf{x} \in L\}.$$

If L is defined implicitly by the linear equality system $\mathbf{Ax} = \mathbf{0}$ with $\mathbf{A} \in \mathbb{R}^{m\times n}$, then $L^\perp$ is given explicitly as the row space of $\mathbf{A}$:

$$\begin{array}{rcccl} L & = & \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} = \mathbf{0}\} & = & \ker \mathbf{A} \\ L^\perp & = & \{\mathbf{c} \in \mathbb{R}^n \mid \mathbf{c}^T = \mathbf{y}^T\mathbf{A},\ \mathbf{y} \in \mathbb{R}^m\} & = & \text{row } \mathbf{A}. \end{array} \tag{3.2}$$

This duality relation is the core of Gale's Theorem (Corollary 2.2) for linear equality systems. We want to generalize it to a *cone duality* as a means to pass from explicit to implicit representations (and *vice versa*) of polyhedral cones.

Given a cone $C \subseteq \mathbb{R}^n$ we define its *dual* (or *polar*) cone as

$$C^0 = \{\mathbf{c} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} \leq 0 \text{ for all } \mathbf{x} \in C\}.$$

Clearly, $C^0 \subseteq \mathbb{R}^n$ is again a cone. It may be considered as the "cone of valid inequalities" (of type $\mathbf{c}^T\mathbf{x} \leq 0$) for C.

EX. 3.10. *Let $C \subseteq \mathbb{R}^n$ be a cone. Show that $\mathbf{c}^T\mathbf{x} \leq \gamma$ is valid for C if and only if $\mathbf{c}^T\mathbf{x} \leq 0$ is valid for C. (Hint: $\mathbf{0} \in C$ implies $\gamma \geq 0$.)*

EX. 3.11. *Show for the linear subspace $L \subseteq \mathbb{R}^n$: $L^\perp = L^0$.*

The Duality Relation. As C^0 is again a cone, its dual $C^{00} = (C^0)^0$ is well-defined. The following simple observation implies in particular that $C^{00} = C$ holds if $C \subseteq \mathbb{R}^n$ is a polyhedral cone (and generalizes the well-known relation $L = L^{\perp\perp}$ for linear subspaces).

PROPOSITION 3.1. *Let $C \subseteq \mathbb{R}^n$ be a cone. Then the following are equivalent:*

(i) $C = C^{00}$.

(ii) *C is the intersection of (possibly infinitely many) halfspaces.*

Proof. Assume that (ii) holds. Then C certainly equals the intersection of *all* halfspaces defined by valid inequalities. By Ex. 3.10 these are of the type $\mathbf{c}^T\mathbf{x} \leq 0$. So

$$C = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} \leq 0 \qquad \text{for all } \mathbf{c} \in C^0\}\ , \tag{3.3}$$

which is equivalent to saying that $C = C^{00}$. Conversely, (3.3) implies that C is an intersection of halfspaces.

◇

EX. 3.12. *Show: $C = C^{00}$ if and only if every $\mathbf{v} \notin C$ can be separated from C.*

REMARK. Corollary **??** of Chapter 10 says that $S \subseteq \mathbb{R}^n$ is the intersection of halfspaces if and only if S is a closed convex set. Combined with Proposition 3.1, this result implies that the cone C satisfies $C = C^{00}$ if and only if C is a closed cone.

Explicit Representations of Polyhedral Cones. We can now show that each polyhedral cone admits an explicit representation as a finitely generated cone. In view of Weyl's Theorem, this means that the finitely generated cones are exactly the polyhedral cones.

LEMMA 3.1. $P(\mathbf{A}, \mathbf{0})^0 = \text{cone } \mathbf{A}^T$.

Proof. By the Farkas Lemma (Corollary 2.6), we find for $C = P(\mathbf{A}, \mathbf{0})$:

$$\begin{aligned} C^0 &= \{\mathbf{c} \mid \mathbf{c}^T\mathbf{x} \leq 0 \quad \text{is implied by } \mathbf{A}\mathbf{x} \leq \mathbf{0}\} \\ &= \{\mathbf{c} \mid \mathbf{c}^T = \mathbf{y}^T\mathbf{A},\ \mathbf{y} \geq \mathbf{0}\} \\ &= \text{cone } \mathbf{A}^T. \end{aligned}$$

◇

THEOREM 3.2 (Weyl-Minkowski)**.** *Let $C = P(\mathbf{A}, \mathbf{0}) \subseteq \mathbb{R}^n$ be a polyhedral cone. Then there exists some finite set $W \subseteq \mathbb{R}^n$ such that $C = \text{cone } W$.*

Proof. Lemma 3.1 yields $C^0 = \text{cone } \mathbf{A}^T$. Moreover, Weyl's Theorem guarantees that $C^0 = P(\mathbf{B}, \mathbf{0})$ holds for some matrix $\mathbf{B}$. Consequently, Proposition 3.1 and again Lemma 3.1 yield

$$C = C^{00} = P(\mathbf{B}, \mathbf{0})^0 = \text{cone } \mathbf{B}^T\ ,$$

and the claim follows with W as the set of column vectors of $\mathbf{B}^T$.

◇

EX. 3.13. *Let* $\mathbf{W} = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & 0 \end{bmatrix}$. *Sketch $C = \text{cone } \mathbf{W}$ and its dual C^0 in $\mathbb{R}^2$ and compute an implicit representation $C = P(\mathbf{A}, \mathbf{0})$.*

EX. 3.14. *Show that $C = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{x} > \mathbf{0}\} \cup \{\mathbf{0}\}$ is a cone and $C^{00} = \mathbb{R}^n_+$.*

The Cone of Positive Semidefinite Matrices. If the cone $C \subseteq \mathbb{R}^n$ is contained in a proper subspace $L \subseteq \mathbb{R}^n$, we may wish to define its polar cone *relative* to L, *i.e.*,

$$C^0 = \{\mathbf{y} \in L \mid \mathbf{y}^T\mathbf{x} \leq 0 \text{ for all } \mathbf{x} \in C\}.$$

(We shall use the notation C^0 only if L is fixed in advance so that no misunderstanding is possible.)

A particularly interesting class of (non-polyhedral) cones is provided by the positive semidefinite matrices. Consider the subspace $L = \mathbb{S}^{n\times n} \subseteq \mathbb{R}^{n\times n}$ of symmetric $(n \times n)$-matrices $\mathbf{X} = (x_{ij})$ and the inner product

$$\mathbf{X} \circ \mathbf{Y} = \sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}y_{ij} \qquad \text{with} \quad \mathbf{X}, \mathbf{Y} \in \mathbb{S}^{n\times n}.$$

It is straightforward to verify that the set

$$K = \{\mathbf{X} \in \mathbb{S}^{n\times n} \mid \mathbf{X} \succeq \mathbf{0}\}$$

of positive semidefinite matrices is a cone. An explicit representation of K is provided by

PROPOSITION 3.2. $K = \text{cone}\,\{\mathbf{v}\mathbf{v}^T \mid \mathbf{v} \in \mathbb{R}^n\}$.

Proof. Every matrix of the form $\mathbf{X} = \mathbf{v}\mathbf{v}^T$ is p.s.d., which implies the inclusion "$\supseteq$". Conversely, assume $\mathbf{X} \succeq \mathbf{0}$ and express $\mathbf{X} = \mathbf{Z}\mathbf{Z}^T$ for some matrix $\mathbf{Z} = (z_{st})$ with columns, say, $\mathbf{z}_1, \ldots, \mathbf{z}_k \in \mathbb{R}^n$ (*cf.* Corollary 2.4). Now $\mathbf{X} = \mathbf{Z}\mathbf{Z}^T$ means

$$x_{ij} = \sum_{\ell=1}^{k} z_{i\ell}z_{j\ell} = \left(\sum_{\ell=1}^{k} \mathbf{z}_\ell\mathbf{z}_\ell^T\right)_{ij},$$

i.e., $\mathbf{X} = \sum_{\ell=1}^{k} \mathbf{z}_\ell\mathbf{z}_\ell^T$ is a non-negative combination of matrices of type $\mathbf{v}\mathbf{v}^T$. In other words, $\mathbf{X} \in \text{cone}\,\{\mathbf{v}\mathbf{v}^T \mid \mathbf{v} \in \mathbb{R}^n\}$.

◇

Let us consider the polar cone of K with respect to $\mathbb{S}^{n\times n}$:

$$K^0 = \{\mathbf{Y} \in \mathbb{S}^{n\times n} \mid \mathbf{X} \circ \mathbf{Y} \leq 0 \text{ for all } \mathbf{X} \in K\}\ .$$

COROLLARY 3.1. *K^0 is the cone of negative semidefinite matrices. In other words, $K^0 = -K$ and, consequently, $K^{00} = K$.*

Proof. From Proposition 3.2, we conclude for all $\mathbf{Y} \in \mathbb{S}^{n\times n}$:

$$\mathbf{Y} \in K^0 \quad \Longleftrightarrow \quad \mathbf{Y} \circ (\mathbf{v}\mathbf{v}^T) = \mathbf{v}^T\mathbf{Y}\mathbf{v} \leq 0 \quad \text{for every } \mathbf{v} \in \mathbb{R}^n\ .$$

◇

EX. 3.15. *What is the polar cone of K with respect to $\mathbb{R}^{n\times n}$?*

3.3. Polar Duality of Convex Sets

We want to extend the concept of polarity to arbitrary convex sets $C \subseteq \mathbb{R}^n$. For convenience, we assume $\mathbf{0} \in C$.

If $\mathbf{c}^T\mathbf{x} \leq z$ is a valid inequality for C, $\mathbf{0} \in C$ implies $z \geq 0$. If $z > 0$, we can scale the inequality to $\bar{\mathbf{c}}^T\mathbf{x} \leq 1$ (with $\bar{\mathbf{c}} = \mathbf{c}/z$). In analogy with cone duality, we now define the *polar* of C as

$$C^{pol} = \{\mathbf{c} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} \leq 1 \text{ is valid for all } \mathbf{x} \in C\} .$$

EX. 3.16. *Show: C^{pol} is a convex set. If $C \subseteq \mathbb{R}^n$ is a cone, then $C^{pol} = C^0$.*

EX. 3.17. *Determine the polars of $P = \{\mathbf{x} \in \mathbb{R}^2_+ \mid x_1 + x_2 \leq 1\}$ and $Q = \{\mathbf{x} \in \mathbb{R}^2 \mid |x_i| \leq 2,\ i = 1, 2\}$.*

EX. 3.18. *Let $V \subseteq \mathbb{R}^2$ consist of 5 equally spaced points on the unit circle in the Euclidean plane. Sketch the polar P^{pol} of the convex pentagon $P = \operatorname{conv} V$.*

By virtue of Ex. 3.16, our next observation generalizes Proposition 3.1.

PROPOSITION 3.3. *Let $C \subseteq \mathbb{R}^n$ be a convex set with $\mathbf{0} \in C$. Then the following are equivalent:*

(i) $C = C^{pol\,pol}$.
(ii) *C is an intersection of (possible infinitely many) halfspaces.*

Proof. Assume that (ii) holds, *i.e.*,

$$C = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{a}_i^T\mathbf{x} \leq \beta_i \text{ for all } i \in I\}.$$

$\mathbf{0} \in C$ implies $\beta_i \geq 0$. By scaling, we may therefore assume $\beta_i \in \{0, 1\}$ without loss of generality. Moreover, every inequality $\mathbf{a}_i^T\mathbf{x} \leq 0$ can be replaced by infinitely many inequalities $k\mathbf{a}_i^T\mathbf{x} \leq 1$ ($k \in \mathbb{N}$). So C admits an equivalent presentation of the form

$$C = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{c}_j^T\mathbf{x} \leq 1 \text{ for all } j \in J\} , \tag{3.4}$$

which implies

$$C = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} \leq 1 \text{ for all } \mathbf{c} \in C^{pol}\} = C^{pol\,pol} . \tag{3.5}$$

Conversely, of course, (3.5) exhibits C as an intersection of halfspaces. ◇

Explicit Representation of Polyhedra. Consider the special case of a polyhedron $P = P(\mathbf{A}, \mathbf{b}) \subseteq \mathbb{R}^n$. P is convex and $\mathbf{0} \in P$ is equivalent with $\mathbf{b} \geq \mathbf{0}$. By scaling, we may then assume $b_i \in \{0, 1\}$.

LEMMA 3.2. *Let $P \subseteq \mathbb{R}^n$ be a polyhedron of the form $P = P\left(\begin{bmatrix} \mathbf{A} \\ \mathbf{B} \end{bmatrix}, \begin{pmatrix} \mathbf{1} \\ \mathbf{0} \end{pmatrix}\right)$.*

Then $P^{pol} = \text{conv}\,[\mathbf{A}^T, \mathbf{0}] + \text{cone}\,\mathbf{B}^T$.

Proof. The Farkas Lemma (Corollary 2.6) yields

$$\begin{aligned} P^{pol} &= \{\mathbf{c} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} \leq 1 \text{ is implied by } \mathbf{Ax} \leq \mathbf{1},\ \mathbf{Bx} \leq \mathbf{0}\} \\ &= \{\mathbf{c} \in \mathbb{R}^n \mid \mathbf{c}^T = \mathbf{y}^T\mathbf{A} + \mathbf{z}^T\mathbf{B}, \mathbf{y} \geq \mathbf{0}, \mathbf{z} \geq \mathbf{0}, \mathbf{y}^T\mathbf{1} \leq 1\} \\ &= \{\mathbf{y}^T\mathbf{A} \mid \mathbf{y} \geq \mathbf{0}, \mathbf{y}^T\mathbf{1} \leq 1\} + \{\mathbf{z}^T\mathbf{B} \mid \mathbf{z} \geq \mathbf{0}\} \\ &= \text{conv}\,[\mathbf{A}^T, \mathbf{0}] + \text{cone}\,\mathbf{B}^T \qquad (\textit{cf.}\ \text{Ex. } 3.7). \end{aligned}$$

◇

We can now establish the full equivalence between explicit and implicit representations of polyhedra.

THEOREM 3.3 (Decomposition Theorem)**.** *The non-empty set $P \subseteq \mathbb{R}^n$ is a polyhedron if and only if there are finite sets $V, W \subseteq \mathbb{R}^n$ such that*

$$P = \text{conv}\ V + \text{cone}\ W\ .$$

(Hence, in particular, every bounded polyhedron is a polytope.)

Proof. By Weyl's Theorem, the condition is sufficient for P to be a polyhedron. We show that the polyhedron P indeed admits an explicit representation as claimed.

If $\mathbf{0} \in P$, the representation in Lemma 3.2 exhibits P^{pol} as a polyhedron. Since $\mathbf{0} \in P^{pol}$, Lemma 3.2 can be applied to the polyhedron $P' = P^{pol}$ and yields together with Proposition 3.3

$$P = (P')^{pol} = \text{conv}\ V + \text{cone}\ W$$

for suitable finite sets $V, W \subseteq \mathbb{R}^n$.

If $\mathbf{0} \notin P$, we choose some $\mathbf{x}_0 \in P$ and consider $P_0 = \{-\mathbf{x}_0\} + P$, which (by Ex. 3.6) is a polyhedron. Since $\mathbf{0} \in P_0$, there exist finite sets V_0 and W_0 with the property $P_0 = \text{conv}\ V_0 + \text{cone}\ W_0$. With $V = \{\mathbf{x}_0\} + V_0$ and $W = W_0$, it is now straightforward (*cf.* Ex. 3.19) to verify

$$P = \{\mathbf{x}_0\} + P_0 = \text{conv}\ V + \text{cone}\ W\ .$$

◇

EX. 3.19. *Show: $P = \text{conv}\ V + \text{cone}\ W$ if and only if $\mathbf{x}_0 + P = \text{conv}\ (\mathbf{x}_0 + V) + \text{cone}\ W$ for each $\mathbf{x}_0 \in \mathbb{R}^n$.*

EX. 3.20. *Let $P \subseteq \mathbb{R}^2$ be the unbounded polyhedron with boundary lines*

$$y = x + 3,\ y = -\frac{1}{2}x + 5 \quad \textit{and} \quad y = 10.$$

Express P as $P = P(\mathbf{A}, \mathbf{b})$ and $P = \text{conv}\ V + \text{cone}\ W$. Draw P, conv *V and* cone *W (separate pictures).*

The cone $C = \text{cone } W$ in the Decomposition Theorem 3.3 is called the *recession cone* of $P = P(\mathbf{A}, \mathbf{b})$ and is equal to $P(\mathbf{A}, \mathbf{0})$. It is uniquely determined by P (*cf.* Ex. 3.21).

Ex. 3.21. *Show that for* $\emptyset \neq P = P(\mathbf{A}, \mathbf{b}) = \text{conv } V + \text{cone } W$ *the following statements are equivalent:*

(i) $\mathbf{w} \in \text{cone } W$.
(ii) $\mathbf{w} \in P(\mathbf{A}, \mathbf{0})$.
(iii) $\mathbf{x}_0 + \lambda \mathbf{w} \in P$ *for each* $\mathbf{x}_0 \in P$ *and* $\lambda \geq 0$.
(iv) $\mathbf{x}_0 + \lambda \mathbf{w} \in P$ *for some* $\mathbf{x}_0 \in P$ *and* $\lambda \geq 0$.

3.4. Faces

Intuitively speaking, a "face" of a polyhedron $P \subseteq \mathbb{R}^n$ is a set F of the form $F = P \cap H$, where H is a hyperplane that "touches" P. We also call H a *supporting* hyperplane (supporting P in F). To formalize this idea, assume $P = P(\mathbf{A}, \mathbf{b})$ and recall that an inequality is valid for P if every point $\mathbf{x} \in P$ satisfies it. We say that the set $F \subseteq P$ is a *face* of P if there exists a valid inequality $\mathbf{c}^T\mathbf{x} \leq \gamma$ for P such that

$$F = \{\mathbf{x} \in P \mid \mathbf{c}^T\mathbf{x} = \gamma\} \ .$$

Note that this definition includes the empty set $\emptyset$ (take $\mathbf{0}^T\mathbf{x} \leq 1$) and the full polyhedron P itself (take $\mathbf{0}^T\mathbf{x} \leq 0$) as so-called "trivial" faces of P .

From the optimization point of view, a face of $P(\mathbf{A}, \mathbf{b})$ consists by definition of all points $\mathbf{x}$ of P that achieve the maximum value $f(\mathbf{x}) = \gamma$ (while all other points $\mathbf{x}' \in P$ yield $f(\mathbf{x}') < \gamma$) with respect to the linear function $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$.

Assume that the face $F = \{\mathbf{x} \in P \mid \mathbf{c}^T\mathbf{x} = \gamma\}$ is non-empty (and hence $\mathbf{Ax} \leq \mathbf{b}$ is feasible). Because $\mathbf{c}^T\mathbf{x} \leq \gamma$ is valid for P, it is implied by $\mathbf{Ax} \leq \mathbf{b}$. So the Farkas Lemma (Corollary 2.6) guarantees the existence of a vector $\mathbf{y} \geq \mathbf{0}$ such that $\mathbf{c}^T = \mathbf{y}^T\mathbf{A}$ and $\mathbf{y}^T\mathbf{b} \leq \gamma$. Hence we know that every $\mathbf{x} \in F$ satisfies

$$0 \leq \mathbf{y}^T(\mathbf{b} - \mathbf{Ax}) = \mathbf{y}^T\mathbf{b} - \mathbf{y}^T\mathbf{Ax} \leq \gamma - \mathbf{c}^T\mathbf{x} = 0 \ ,$$

which implies, for all $\mathbf{x} \in F$, the equality $\mathbf{y}^T\mathbf{b} = \mathbf{c}^T\mathbf{x} = \gamma$ as well as the so-called "complementary slackness" relation

$$0 = \mathbf{y}^T(\mathbf{b} - \mathbf{Ax}) = \sum_i y_i(b_i - \mathbf{A}_i\mathbf{x}) \ .$$

This relation says: If $y_i \neq 0$ holds for the ith component of the vector $\mathbf{y} \geq \mathbf{0}$, then the corresponding ith inequality $\mathbf{A}_i\mathbf{x} \leq b_i$ of the system $\mathbf{Ax} \leq \mathbf{b}$ must be *tight* (or *active*) for all $\mathbf{x} \in F$, *i.e.*,

$$\mathbf{A}_i\mathbf{x} = b_i \quad \text{for all} \quad \mathbf{x} \in F \ .$$

Therefore, in view of

$$\mathbf{c}^T = \mathbf{y}^T\mathbf{A} \ , \quad \mathbf{y}^T\mathbf{b} = \gamma \ , \tag{3.6}$$

the relation $\mathbf{c}^T\mathbf{x} = \gamma$ is seen to be implied by the set of those inequalities of $\mathbf{Ax} \leq \mathbf{b}$ that are tight for all $\mathbf{x} \in F$.

THEOREM 3.4. *The nonempty set F is a face of the polyhedron $P = P(\mathbf{A}, \mathbf{b})$ if and only if there exists a subsystem $\mathbf{A}'\mathbf{x} \leq \mathbf{b}'$ of $\mathbf{Ax} \leq \mathbf{b}$ such that*

$$F = \{\mathbf{x} \in P \mid \mathbf{A}'\mathbf{x} = \mathbf{b}'\} .$$

Proof. Assume $F = \{\mathbf{x} \in P \mid \mathbf{A}'\mathbf{x} = \mathbf{b}'\}$ for some subsystem $\mathbf{A}'\mathbf{x} \leq \mathbf{b}'$ of $\mathbf{Ax} \leq \mathbf{b}$. If the subsystem is empty we obtain the face $F = P$. Otherwise, choose $\mathbf{c}^T$ as the sum of the rows of $\mathbf{A}'$ and, correspondingly, γ as the sum of the coefficients of $\mathbf{b}'$ (*i.e.*, $\mathbf{c}^T = \mathbf{1}^T\mathbf{A}'$ and $\gamma = \mathbf{1}^T\mathbf{b}'$). Then $\mathbf{c}^T\mathbf{x} \leq \gamma$ is a valid inequality for $P(\mathbf{A}, \mathbf{b})$. Moreover, we observe for every $\mathbf{x} \in P(\mathbf{A}, \mathbf{b})$,

$$\mathbf{c}^T\mathbf{x} = \gamma \text{ if and only if } \mathbf{A}'\mathbf{x} = \mathbf{b}' .$$

So $F = \{\mathbf{x} \in P \mid \mathbf{c}^T\mathbf{x} = \gamma\}$ is a face of P.

Conversely, assume that $F = \{\mathbf{x} \in P \mid \mathbf{c}^T\mathbf{x} = \gamma\}$ is a face of P and let $\mathbf{A}'\mathbf{x} \leq \mathbf{b}'$ be the subsystem of $\mathbf{Ax} \leq \mathbf{b}$ consisting of all inequalities that are tight for every $\mathbf{x} \in F$. By definition, we then have $F \subseteq \{\mathbf{x} \in P \mid \mathbf{A}'\mathbf{x} = \mathbf{b}'\}$. We claim that, in fact, the equality $F = \{\mathbf{x} \in P \mid \mathbf{A}'\mathbf{x} = \mathbf{b}'\}$ holds.

Indeed, relation (3.6) shows that the linear equation $\mathbf{c}^T\mathbf{x} = \gamma$ can be obtained as a non-negative linear combination of equations in $\mathbf{A}'\mathbf{x} = \mathbf{b}'$. So every $\mathbf{x} \in P$ that satisfies $\mathbf{A}'\mathbf{x} = \mathbf{b}'$ must also satisfy $\mathbf{c}^T\mathbf{x} = \gamma$ and hence belong to F.

◇

Since a finite system of linear inequalities admits only a finite number of different subsystems, we immediately deduce from Theorem 3.4:

COROLLARY 3.2. *A polyhedron has only a finite number of faces.*

EX. 3.22. *Prove that the closed disk $S = \{\mathbf{x} \in \mathbb{R}^2 \mid \|\mathbf{x}\| \leq 1\}$ is* not *a polyhedron.*

Dimension. Recall from Chapter 1 (p. 4) that aff P denotes the smallest affine subspace of $\mathbb{R}^n$ that contains $P \subseteq \mathbb{R}^n$. Because affine subspaces are intersections of hyperplanes, aff P is the intersection of all hyperplanes that contain P.

Let $\mathbf{A}^=\mathbf{x} \leq \mathbf{b}^=$ denote the (possibly empty) subsystem of $\mathbf{Ax} \leq \mathbf{b}$ consisting of those inequalities that are tight for every $\mathbf{x} \in P = P(\mathbf{A}, \mathbf{b})$. Then $P \subseteq \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}^=\mathbf{x} = \mathbf{b}^=\}$ and, since aff P is the smallest affine subspace containing P, also aff $P \subseteq \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}^=\mathbf{x} = \mathbf{b}^=\}$. Actually, equality holds:

COROLLARY 3.3.

$$\text{aff } P(\mathbf{A}, \mathbf{b}) = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}^=\mathbf{x} = \mathbf{b}^=\} . \tag{3.7}$$

Proof. We have already observed that "$\subseteq$" holds. To establish the converse inclusion, it suffices to show that any hyperplane $H = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} = \gamma\}$ containing aff P also contains the solution set of $\mathbf{A}^=\mathbf{x} = \mathbf{b}^=$, *i.e.*, that $\mathbf{c}^T\mathbf{x} = \gamma$ is implied by $\mathbf{A}^=\mathbf{x} = \mathbf{b}^=$. Thus assume $H = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} = \gamma\}$ contains aff P, and hence P. As in the proof of Theorem 3.4 we thus find that $\mathbf{c}^T\mathbf{x} = \gamma$ is a (non-negative) linear combination of the equations in $\mathbf{A}^=\mathbf{x} = \mathbf{b}^=$ (which is the system $\mathbf{A}'\mathbf{x} = \mathbf{b}'$ corresponding to $F = P$) and the claim follows.

◇

EX. 3.23. *Let* $\mathbf{A}'\mathbf{x} \leq \mathbf{b}'$ *be the subsystem of* $\mathbf{Ax} \leq \mathbf{b}$ *of all inequalities that are tight for the given point* $\mathbf{x}_0 \in P = P(A, \mathbf{b})$. *Show:* $F = \{\mathbf{x} \in P \mid \mathbf{A}'\mathbf{x} = \mathbf{b}'\}$ *is the* unique *smallest face of* P *that contains* $\mathbf{x}_0$.

Let us define the *dimension* of a polyhedron P as

$$\dim P = \dim \text{ aff } P \ . \tag{3.8}$$

Then (3.7) implies a formula for the dimension:

COROLLARY 3.4. *If* $\mathbf{A} \in \mathbb{R}^{m\times n}$ *and* $P(\mathbf{A}, \mathbf{b}) \neq \emptyset$, *then*

$$\dim P(\mathbf{A}, \mathbf{b}) = n - \text{rank } \mathbf{A}^= \ . \tag{3.9}$$

Proof. $\dim P = \dim \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}^=\mathbf{x} = \mathbf{b}^=\} = \dim \ker \mathbf{A}^= = n - \text{rank } \mathbf{A}^=$.

◇

Facets. Since a face of a polyhedron is a polyhedron in its own right, it is meaningful to talk about the dimension of a face in general. We say that the face F of the polyhedron P is a *facet* if

$$\dim F = \dim P - 1 \ .$$

The example of an affine subspace shows that polyhedra without facets do exist. However, this example furnishes the only exception, as we shall prove in Corollary 3.5 below.

We say that $\mathbf{Ax} \leq \mathbf{b}$ is *irredundant* if no inequality in $\mathbf{Ax} \leq \mathbf{b}$ is implied by the remaining inequalities. Clearly, every polyhedron P can be defined by an irredundant system. (We only have to remove implied inequalities successively until an irredundant system is obtained.)

EX. 3.24. *Formulate an optimization problem whose solution would allow you to decide whether the* ith *inequality* $\mathbf{A}_i\mathbf{x} \leq b_i$ *is implied by the remaining inequalities of* $\mathbf{Ax} \leq \mathbf{b}$.

COROLLARY 3.5. *Assume that* $\mathbf{Ax} \leq \mathbf{b}$ *is irredundant and consider an inequality* $\mathbf{A}_i\mathbf{x} \leq b_i$ *which is not part of* $\mathbf{A}^=\mathbf{x} \leq \mathbf{b}^=$. *Then*

$$F = \{\mathbf{x} \mid \mathbf{Ax} \leq \mathbf{b}, \mathbf{A}_i\mathbf{x} = b_i\}$$

is a facet of $P = P(\mathbf{A}, \mathbf{b})$. Consequently, if P is not an affine subspace, then P has proper nonempty faces and each such face can be obtained as an intersection of facets.

Proof. Denote by $\mathbf{A}^*\mathbf{x} \leq \mathbf{b}^*$ the system of those inequalities of $\mathbf{Ax} \leq \mathbf{b}$ that are not in $\mathbf{A}^=\mathbf{x} \leq \mathbf{b}^=$. The hypothesis of the Corollary says that $\mathbf{A}_i\mathbf{x} \leq b_i$ is in $\mathbf{A}^*\mathbf{x} \leq \mathbf{b}^*$.

By definition, we can find for each $\mathbf{A}_s\mathbf{x} \leq b_s$ of $\mathbf{A}^*\mathbf{x} \leq \mathbf{b}^*$ some $\mathbf{x}^s \in P(\mathbf{A}, \mathbf{b})$ satisfying the strict inequality $\mathbf{A}_s\mathbf{x}^s < b_s$. Assuming there are k such inequalities, it follows that also the convex combination

$$\mathbf{x}^* = \frac{1}{k}\sum_{s=1}^{k} \mathbf{x}^s$$

lies in $P(\mathbf{A}, \mathbf{b})$. Moreover, one readily verifies that $\mathbf{A}^*\mathbf{x}^* < \mathbf{b}^*$ holds, *i.e.*, the average $\mathbf{x}^*$ of the k vectors $\mathbf{x}^s$ satisfies *each* inequality in $\mathbf{A}^*\mathbf{x} \leq \mathbf{b}^*$ *strictly*.

Since $\mathbf{Ax} \leq \mathbf{b}$ is an irredundant system, removing $\mathbf{A}_i\mathbf{x} \leq b_i$ from the system would result in a larger feasibility region. So there exists some vector $\mathbf{v} \in \mathbb{R}^n$ such that $\mathbf{A}_i\mathbf{v} > b_i$ holds while $\mathbf{v}$ satisfies all the other inequalities $\mathbf{A}_j\mathbf{x} \leq b_j$, $j \neq i$, of $\mathbf{Ax} \leq \mathbf{b}$. By the choice of $\mathbf{v}$, we have $0 < b_i - \mathbf{A}_i\mathbf{x}^* < \mathbf{A}_i\mathbf{v} - \mathbf{A}_i\mathbf{x}^*$.

Let $\lambda = (b_i - \mathbf{A}_i\mathbf{x}^*)(\mathbf{A}_i\mathbf{v} - \mathbf{A}_i\mathbf{x}^*)^{-1}$. Noting $0 < \lambda < 1$, we then obtain

$$\bar{\mathbf{x}} = \lambda\mathbf{v} + (1-\lambda)\mathbf{x}^* \ \in \ P(\mathbf{A}, \mathbf{b}) \ .$$

In particular, $\mathbf{A}_i\bar{\mathbf{x}} = b_i$ holds, while $\mathbf{A}_s\bar{\mathbf{x}} < b_s$ is true for all other inequalities in $\mathbf{A}^*\mathbf{x} \leq \mathbf{b}^*$. Hence the subsystem of $\mathbf{Ax} \leq \mathbf{b}$ of inequalities that are tight for the face F consists precisely of $\mathbf{A}^=\mathbf{x} \leq \mathbf{b}^=$ together with the one extra inequality $\mathbf{A}_i\mathbf{x} \leq b_i$, which yields $\dim F = \dim P - 1$, as claimed.

Consider finally an arbitrary non-trivial face F' of the polyhedron $P(\mathbf{A}, \mathbf{b})$. Each inequality of $\mathbf{Ax} \leq \mathbf{b}$ that is tight for F' is either already tight for $P(\mathbf{A}, \mathbf{b})$ or, as we have seen, induces a facet of $P(\mathbf{A}, \mathbf{b})$. So F' must be the intersection of the corresponding facets.

$\diamond$

REMARK. Our analysis in Corollary 3.5 exhibits "facet-generating" inequalities as the strongest inequalities for the description of a polyhedron. If $P = P(\mathbf{A}, \mathbf{b})$ is presented by an irredundant system $\mathbf{Ax} \leq \mathbf{b}$, then $\mathbf{A}^=\mathbf{x} = \mathbf{b}^=$ determines the affine subspace relative to which P has full dimension, while $\mathbf{A}^*\mathbf{x} \leq \mathbf{b}^*$ describes the facets of P. In particular, the number of facets of P equals the number of inequalities in $\mathbf{A}^*\mathbf{x} \leq \mathbf{b}^*$.

REMARK. Let $\mathcal{L} = \mathcal{L}(P)$ be the collection of all faces of the polyhedron P. Ordering the members of $\mathcal{L}$ by containment, we obtain the trivial face $\emptyset$ as the unique minimal and P as the unique maximal member of $\mathcal{L}$. As faces are precisely intersections of facets (*cf.* Corollary 3.5), an intersection of faces always yields a face. Hence $\mathcal{L}$ becomes a *lattice* relative to the binary operations for all $F_1, F_2 \in \mathcal{L}$,

$$F_1 \wedge F_2 = F_1 \cap F_2$$
$$F_1 \vee F_2 = \cap\{F \in \mathcal{L} \mid F_1, F_2 \subseteq F\} \ .$$

$(\mathcal{L}(P), \wedge, \vee)$ is called the *face lattice* of the polyhedron P and captures the combinatorial structure of the polyhedron P (see, *e.g.*, [79] for more details).

EX. 3.25. *Let F be a minimal nonempty face of $P = P(\mathbf{A}, \mathbf{b})$. Show: $F = \mathbf{x}_0 + \ker \mathbf{A}$ for some $\mathbf{x}_0 \in P$. (Hint: $L = \ker \mathbf{A}$ is the unique largest linear subspace contained in the recession cone $P(\mathbf{A}, \mathbf{0})$ of P.)*
Give an example of a full-dimensional polyhedron $P \subseteq \mathbb{R}^3$ whose minimal faces have dimension 1.

3.5. Vertices and Polytopes

The vector $\mathbf{v} \in \mathbb{R}^n$ is called a *vertex* (or *extreme point*) of the polyhedron $P \subseteq \mathbb{R}^n$ if $F = \{\mathbf{v}\}$ is a face of P (of dimension $\dim F = 0$). Relative to a representation of the polyhedron in terms of linear inequalities, Corollary 3.4 states that a point $\mathbf{v} \in P(\mathbf{A}, \mathbf{b})$ is a vertex if and only if there exists a subsystem $\mathbf{A}'\mathbf{x} \leq \mathbf{b}'$ of $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ so that $\operatorname{rank} \mathbf{A}' = n$ and $\mathbf{v}$ is the unique (feasible) solution of

$$\mathbf{A}'\mathbf{v} = \mathbf{b}' \ .$$

In other words, the rows of $\mathbf{A}'$ must contain a basis of $\mathbb{R}^n$. We therefore call a vertex of $P = P(\mathbf{A}, \mathbf{b})$ also *basic solution* or *vertex solution* of the system $\mathbf{A}\mathbf{x} \leq \mathbf{b}$.

A vertex $\mathbf{v}$ of a polyhedron P is defined "implicitly" by a hyperplane H that supports P exactly in the point $\mathbf{v}$. An "explicit" characterization is also possible:

THEOREM 3.5. *Let P be a polyhedron and $\mathbf{v} \in P$ arbitrary. Then $\mathbf{v}$ is a vertex of P if and only if $\mathbf{v}$ cannot be expressed as convex combination of other vectors in P.*

Proof. Let $P = P(\mathbf{A}, \mathbf{b})$ and assume that there are vectors $\mathbf{v}_1, \ldots, \mathbf{v}_k \in P$ and numbers $\lambda_i > 0$ such that $\sum_i \lambda_i = 1$ and $\mathbf{v} = \lambda_1\mathbf{v}_1 + \ldots + \lambda_k\mathbf{v}_k$. Let furthermore F be an arbitrary face of P. We claim:

$$\mathbf{v} \in F \qquad \text{if and only if} \qquad \mathbf{v}_1, \ldots, \mathbf{v}_k \in F \ .$$

Indeed, assume $F = \{\mathbf{x} \in P \mid \mathbf{A}'\mathbf{x} = \mathbf{b}'\}$ for some subsystem $\mathbf{A}'\mathbf{x} \leq \mathbf{b}'$ of $\mathbf{A}\mathbf{x} \leq \mathbf{b}$. Now $\mathbf{v}_i \in P$ implies in particular $\mathbf{A}'\mathbf{v}_i \leq \mathbf{b}'$. Hence $\mathbf{A}'\mathbf{v} = \mathbf{b}'$ or, equivalently, $\sum_i \lambda_i(\mathbf{A}'\mathbf{v}_i - \mathbf{b}') = \mathbf{0}$ is true if and only if $\lambda_i\mathbf{A}'\mathbf{v}_i = \lambda_i\mathbf{b}'$ is true for all i.

Consequently, if $F = \{\mathbf{v}\}$, then $\mathbf{v} = \mathbf{v}_1 = \ldots = \mathbf{v}_k$. Conversely, suppose that $\mathbf{v} \in P$ is not a vertex. Let $\mathbf{A}'\mathbf{x} \leq \mathbf{b}'$ consist of all inequalities in $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ that are tight at $\mathbf{v}$ and consider $F = \{\mathbf{x} \in P \mid \mathbf{A}'\mathbf{x} = \mathbf{b}'\}$. F is a face of P and contains $\mathbf{v}$. Since $\mathbf{v}$ is not a vertex, we know $\dim F \geq 1$, which yields $\operatorname{rank} \mathbf{A}' \leq n - 1$.

Hence we can find a vector $\mathbf{z} \neq \mathbf{0}$ such that $\mathbf{A}'\mathbf{z} = \mathbf{0}$. Choosing $\epsilon > 0$ sufficiently small, we can guarantee that both $\mathbf{v}_1 = \mathbf{v} + \epsilon\mathbf{z}$ and $\mathbf{v}_2 = \mathbf{v} - \epsilon\mathbf{z}$ are feasible for $\mathbf{A}\mathbf{x} \leq \mathbf{b}$. Hence we obtain

$$\mathbf{v} = \frac{1}{2}(\mathbf{v}_1 + \mathbf{v}_2)$$

as a non-trivial convex combination of points $\mathbf{v}_1, \mathbf{v}_2 \in P$.

◇

COROLLARY 3.6. *Let $P \subseteq \mathbb{R}^n$ be a polytope and let $V \subseteq P$ be the set of vertices of P. Then $P = \text{conv } V$.*

Proof. Assume that say $P = \text{conv } V'$ for some finite set $V' \subseteq \mathbb{R}^n$. We may assume that V' is minimal in the sense that $\mathbf{v} \notin \text{conv } (V' \setminus \{\mathbf{v}\})$ holds for every $\mathbf{v} \in V'$. Then no $\mathbf{v} \in V'$ can be expressed as a non-trivial convex combination of vectors in $P \setminus \{\mathbf{v}\}$. Hence V' must be the set of extreme points (vertices) of P.

◇

REMARK. The proof of Corollary 3.6 indicates that the list V of its vertices provides the smallest explicit representation of the polytope $P = P(\mathbf{A}, \mathbf{b})$. Note, however, that $|V|$ can be *exponentially large* with respect to the size of the implicit representation $\mathbf{Ax} \leq \mathbf{b}$ (see Ex. 3.26).

EX. 3.26. *Show that $P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{0} \leq \mathbf{x} \leq \mathbf{1}\}$ is a polytope with 2^n vertices.*

Vertices and Basic Solutions of $\mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}$. The simplex algorithm for linear programs in Chapter 4 refers to polyhedra P of the form

$$P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} = \mathbf{b}, \mathbf{x} \geq \mathbf{0}\} \ . \tag{3.10}$$

Observe that (due to the constraints $\mathbf{x} \geq \mathbf{0}$) P does not contain any affine space of dimension ≥ 1. So by virtue of Corollary 3.4, if P is nonempty, it must have vertices (every minimal nonempty face of P is one).

Assuming $\mathbf{A} \in \mathbb{R}^{m \times n}$ with rank $\mathbf{A} = r$ $(\leq m)$, we find in this case that a vector $\mathbf{x} \in P$ is a vertex if and only if there is a set N of $|N| = n - r$ indices j such that $\mathbf{x}$ is the unique solution of

$$\begin{aligned} \mathbf{Ax} &= \mathbf{b} \\ x_j = \mathbf{e}_j^T \mathbf{x} &= 0, \quad j \in N \ , \end{aligned}$$

or equivalently:

(i) $x_j = 0$ for all $j \in N$.
(ii) The submatrix $\mathbf{B}$ of those r columns $\mathbf{A}_{\cdot j}$ with $j \notin N$ has full rank r, *i.e.*, is a column basis for $\mathbf{A}$.

Thinking of a vertex of P algebraically as the unique solution of the linear system $\mathbf{Ax} = \mathbf{b}$ under the additional constraints (i) and (ii), we call a vertex of P also a *basic solution*. Computationally, solving $\mathbf{Ax} = \mathbf{b}$ under condition (i), (ii) simply amounts to applying Gaussian Elimination to $\mathbf{Bx} = \mathbf{b}$.

With the notion of a basic solution we can easily derive Carathéodory's Theorem on convex combinations.

THEOREM 3.6 (Carathéodory)**.** *Let $S \subseteq \mathbb{R}^n$. Then every $\mathbf{b} \in$ cone S can be expressed as a conic combination of at most n vectors and every $\mathbf{b} \in$ conv S can be written as a convex combination of at most $n+1$ vectors in S.*

Proof. Let $\mathbf{b} \in$ cone S. By definition, there exist vectors $\mathbf{s}_1, \ldots, \mathbf{s}_k \in S$ and coefficients $x_1, \ldots, x_k \geq 0$ such that $\mathbf{b} = \sum x_i \mathbf{s}_i$. Consider the matrix $\mathbf{A} \in \mathbb{R}^{n \times k}$ with the k column vectors $\mathbf{s}_i$ and let $P \subseteq \mathbb{R}^k$ be the polyhedron of all feasible solutions of the linear system

$$\begin{aligned} \mathbf{Ax} &= \mathbf{b} \\ \mathbf{x} &\geq \mathbf{0}. \end{aligned}$$

By the previous discussion, $P = P(\mathbf{A}, \mathbf{b})$ has a vertex $\mathbf{v}$. Interpreting $\mathbf{v}$ as a basic solution of the linear system, we find that $\mathbf{v}$ has at most n non-zero components v_i, which furnish the desired convex combination for $\mathbf{b}$.

Assume now $\mathbf{b} \in$ conv S and let $\mathbf{s}_1, \ldots, \mathbf{s}_k \in S$ and $x_1, \ldots, x_k \geq 0$ be such that $\mathbf{b} = \sum x_i \mathbf{s}_i$ and $\sum x_i = 1$. This means

$$\begin{pmatrix} \mathbf{b} \\ 1 \end{pmatrix} \in \text{cone} \left\{ \begin{pmatrix} \mathbf{s}_i \\ 1 \end{pmatrix} \mid i = 1, \ldots, k \right\} \subseteq \mathbb{R}^{n+1} ,$$

and the claim follows from the result for cones.

◇

COROLLARY 3.7. *Let $S \subset \mathbb{R}^n$ be a compact set. Then* conv (S) *is compact.*

Proof. Consider the standard simplex

$$\Delta_{n+1} = \left\{ (\lambda_1, \ldots, \lambda_{n+1}) \mid \lambda_j \geq 0, \ \sum_{j=1}^{n+1} \lambda_j = 1 \right\}$$

and define a continuous function $F : \Delta_{n+1} \times S \times \cdots \times S \to \mathbb{R}^n$ *via*

$$F(\lambda_1, \ldots, \lambda_{n+1}, \mathbf{a}_1, \ldots, \mathbf{a}_{n+1}) = \sum_{j=1}^{n+1} \lambda_j \mathbf{a}_j \ .$$

Carathéodory's Theorem implies $F(\Delta_{n+1} \times S \times \ldots \times S) =$ conv S. Since Δ_{n+1} and S are compact, $\Delta_{n+1} \times S \times \cdots \times S$ is compact. So conv (S) is the image of a compact set under a continuous function and, therefore, compact (*cf.* Section 1.4.2).

◇

EX. 3.27. *Give an example of a closed set $S \subset \mathbb{R}^n$ such that* conv *S is not closed. (Hint: Consider $S = \{(0,1)\} \cup \{(x_1, 0) \mid x_1 \in \mathbb{R}\}$.)*

EX. 3.28. *Give an example of a compact set $S \subset \mathbb{R}^n$ such that* cone *S is not closed. (Hint: Consider $S = \{\mathbf{x} \in \mathbb{R}^2 \mid x_1^2 + (x_2 - 1)^2 = 1\}$.)*

3.6. Rational Polyhedra

From a computational point of view it is reasonable to consider systems of inequalities $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ with only *rational* coefficients. Let us thus call the polyhedron $P(\mathbf{A}, \mathbf{b})$ *rational* if $\mathbf{A} \in \mathbb{Q}^{m\times n}$ and $\mathbf{b} \in \mathbb{Q}^m$. It then follows from Theorem 3.4 that all faces of a rational polyhedron are rational polyhedra. Moreover, for any $\mathbf{v}_1, \ldots, \mathbf{v}_m \in \mathbb{Q}^n$, we find that both cone $(\mathbf{v}_1, \ldots, \mathbf{v}_m)$ and conv $(\mathbf{v}_1, \ldots, \mathbf{v}_m)$ are rational polyhedra. Indeed, all our proofs are eventually based on the Fourier-Motzkin algorithm, for which we had noted that rationality of the parameters is preserved during the computation.

Similarly, the Decomposition Theorem of Weyl and Minkowski holds for rational polyhedra, *i.e.*, a set $P \subseteq \mathbb{R}^n$ is a rational polyhedron if and only if there are finite sets $V, W \subset \mathbb{Q}^n$ of vectors with rational components such that

$$P \;=\; \text{conv}\, V + \text{cone}\, W \;.$$

We leave the straightforward check to the reader.

CHAPTER 4

Lagrangian Duality

The present chapter pursues two goals. First, we take a (rather preliminary) look at nonlinear optimization problems by investigating to what extent fundamental concepts carry over from linear to general optimization problems and what kind of difficulties arise in the general context. Doing so, our second goal is to motivate much of the theory of nonlinear problems that are treated in more detail in subsequent chapters. The main points we want to make now are Lagrangian relaxation as a bounding technique for integer programs (*cf.* Chapter 9) and the optimality conditions (which we derive rather independently from the rest of the chapter in Section 4.4) that motivate many of the algorithmic approaches to non-linear problems.

4.1. Lagrangian Relaxation

A *nonlinear (constrained) optimization problem* is a problem of the type

$$\max \ f(\mathbf{x}) \qquad \text{s.t.} \quad g_j(\mathbf{x}) \le 0\,, \qquad j = 1, \ldots, m,$$

with *objective function* $f : \mathbb{R}^n \to \mathbb{R}$ and *constraint functions* $g_j : \mathbb{R}^n \to \mathbb{R}$. In terms of the vector-valued function $g(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))^T$, this problem can be stated more compactly as

$$\max \ f(\mathbf{x}) \quad \text{s.t.} \quad g(\mathbf{x}) \le \mathbf{0}\,. \tag{4.1}$$

REMARK. Usually, one assumes f and g to be (at least) continuous. In what follows, whenever a gradient ∇f or Jacobian ∇g are used, we implicitly assume that f and g are continuously differentiable.

The use of "max" resp. "min" is standard notation in nonlinear optimization although "sup" and "inf" would be more precise. For example,

$$\min \ x_1 \qquad \text{s.t.} \qquad x_1 x_2 \ge 1, \quad x_1 \ge 0$$

has "minimum value" 0, but optimal solutions do not exist.

Ex. 4.1. *Show that* $\bar{\mathbf{x}} = (1, 1)$ *(with* $f(\bar{\mathbf{x}}) = 6$*) is an optimal solution for*

$$\max \ f(\mathbf{x}) = 4(x_1 + x_2) - (x_1^2 + x_2^2) \quad \text{s.t.} \quad g(\mathbf{x}) = x_1 x_2 - 1 \ \le \ 0\,.$$

Clearly, a linear program, maximizing a linear objective $f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ under *linear constraints* $g(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b} \ \le \ \mathbf{0}$ is a special case of (4.1).

As in linear programming, we associate with the *primal problem* (4.1) a *dual* problem that wants to minimize certain upper bounds on the primal maximum

value. As in the linear case, we obtain such bounds from non-negative combinations of the constraints. Consider any $\mathbf{y} = (y_1, \ldots, y_m)^T \geq \mathbf{0}$. Then every *primal feasible* $\mathbf{x}$, *i.e.*, any $\mathbf{x} \in \mathbb{R}^n$ with $g(\mathbf{x}) \leq \mathbf{0}$, necessarily satisfies the "derived" inequality

$$\sum_{j=1}^{m} y_j g_j(\mathbf{x}) \;=\; \mathbf{y}^T g(\mathbf{x}) \;\leq\; 0\,. \tag{4.2}$$

So each $\mathbf{y} \geq \mathbf{0}$ gives rise to an upper bound $L(\mathbf{y})$:

$$\max_{g(\mathbf{x})\leq \mathbf{0}} f(\mathbf{x}) \;\leq\; \max_{\mathbf{x}} f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x}) \;=\; L(\mathbf{y}). \tag{4.3}$$

The (unconstrained) maximization problem defining the upper bound

$$L(\mathbf{y}) \;=\; \max_{\mathbf{x}} f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x})$$

is called the *Lagrangian relaxation* of (4.1) with *Lagrangian multipliers* $y_j \geq 0$ (which play the role of the dual variables in linear programming). So the Lagrangian relaxation is obtained by "moving the constraints into the objective function". We also say that we *relax* or *dualize* the constraints $g_j(\mathbf{x}) \leq 0$ with multipliers $y_j \geq 0$.

The problem of determining the best possible upper bound $L(\mathbf{y})$ is the *Lagrangian dual problem*

$$\min_{\mathbf{y}\geq \mathbf{0}} L(\mathbf{y}) \;=\; \min_{\mathbf{y}\geq \mathbf{0}} \max_{\mathbf{x}} \; f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x}). \tag{4.4}$$

We immediately observe the following relation between the primal problem (4.1) and its dual (4.4).

THEOREM 4.1. (Weak Duality)

$$\max_{\mathbf{g}(\mathbf{x})\leq \mathbf{0}} \; f(\mathbf{x}) \;\leq\; \min_{\mathbf{y}\geq \mathbf{0}} \; L(\mathbf{y})\,.$$

Consequently, if equality is attained with the primal feasible $\overline{\mathbf{x}}$ *and the (dual feasible)* $\overline{\mathbf{y}} \geq \mathbf{0}$*, then* $\overline{\mathbf{x}}$ *and* $\overline{\mathbf{y}}$ *are optimal primal resp. dual solutions. In this case* $\overline{\mathbf{x}}$ *and* $\overline{\mathbf{y}}$ *are necessarily* complementary *in the sense that* $\overline{\mathbf{y}}^T g(\overline{\mathbf{x}}) = 0$ *.*

Proof. In view of (4.2), we have the inequality

$$f(\overline{\mathbf{x}}) \leq\; f(\overline{\mathbf{x}}) - \overline{\mathbf{y}}^T g(\overline{\mathbf{x}}) \leq \max_{\mathbf{x}} \; f(\mathbf{x}) - \overline{\mathbf{y}}^T g(\mathbf{x}) = L(\overline{\mathbf{y}})\,.$$

Equality can only hold if $\overline{\mathbf{y}}^T g(\overline{\mathbf{x}}) = 0$.

◇

REMARK. As in the linear (programming) case, an equality constraint $g_j(\mathbf{x}) = 0$ is formally equivalent to two opposite inequalities and corresponds to a *sign-unrestricted* multiplier $y_j \in \mathbb{R}$ in the Lagrangian dual.

REMARK. An alternative view on Lagrangian relaxation is the following. Choosing all multipliers $y_j \geq 0$ "very large", we would expect the optimal solution of

$$\max_{\mathbf{x}} f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x})$$

to be primal feasible (since any violation of $g_j(\mathbf{x}) \leq 0$ is *penalized* by subtracting $y_j g_j(\mathbf{x}) > 0$ from the objective function). On the other hand, taking all y_j "very large" makes the objective function $f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x})$ almost unrelated to the "true" objective function $f(\mathbf{x})$. So the "best" choice of multipliers $\overline{\mathbf{y}} \geq \mathbf{0}$, *i.e.*, the solution of the dual problem, will usually *not* guarantee the corresponding maximizer $\mathbf{x}$ to be primal feasible.

At first sight, the Lagrangian dual (a so-called *min-max problem*) may appear more difficult than the original primal problem (4.1). Actually, it is often easier to solve. One reason is that the value of the Lagrangian relaxation

$$L(\mathbf{y}) = \max_{\mathbf{x}} f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x}) \tag{4.5}$$

can often be found by the extremum principle with respect to the maximization problem relative to $\mathbf{x}$ (see p. 17). In other words, a maximizer of (4.5) (for fixed $\mathbf{y} \geq \mathbf{0}$) must necessarily satisfy the (generally nonlinear) *critical equation*

$$\nabla f(\mathbf{x}) - \mathbf{y}^T \nabla g(\mathbf{x}) \;=\; \mathbf{0}^T \,. \tag{4.6}$$

Ex. 4.2. Consider the eigenvalue problem for the symmetric matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$:

$$\max_{\mathbf{x}^T\mathbf{x}=1} \mathbf{x}^T\mathbf{A}\mathbf{x} \quad \text{with dual} \quad \min_{\lambda\in\mathbb{R}} \max_{\mathbf{x}} \mathbf{x}^T\mathbf{A}\mathbf{x} - \lambda\,(\mathbf{x}^T\mathbf{x} - 1)\,.$$

The critical equation $\nabla f(\mathbf{x}) - \lambda \nabla g(\mathbf{x}) = 2\mathbf{x}^T\mathbf{A} - 2\lambda\,\mathbf{x}^T = \mathbf{0}^T$ is always solvable (with $\mathbf{x} = \mathbf{0}$ or – in case λ is an eigenvalue of $\mathbf{A}$ – a corresponding eigenvector). However, the relaxation

$$L(\lambda) \;=\; \max_{\mathbf{x}} \mathbf{x}^T\mathbf{A}\mathbf{x} - \lambda\,\mathbf{x}^T\mathbf{x} + \lambda = \lambda + \max_{\mathbf{x}} \mathbf{x}^T(\mathbf{A} - \lambda\mathbf{I})\mathbf{x}$$

has an optimal solution only for $\lambda \geq \lambda_{max}$, the maximum eigenvalue of $\mathbf{A}$, because the matrix $\mathbf{A} - \lambda\mathbf{I}$ is negative semidefinite for $\lambda \geq \lambda_{max}$ (so that $\mathbf{x} = \mathbf{0}$ is optimal). If $\lambda = \lambda_{max}$, also a corresponding eigenvector $\mathbf{x}_{max}$ is optimal.

Ex. 4.3. The Lagrangian relaxation for the problem in Ex. 4.1 is

$$L(y) = \max\; f(\mathbf{x}) - y g(\mathbf{x}) = 4(x_1 + x_2) - (x_1^2 + x_2^2) - y(x_1x_2 - 1)$$

with critical equation

$$\nabla f(\mathbf{x}) - y\nabla g(\mathbf{x}) = (4 - 2x_1 - yx_2,\, 4 - yx_1 - 2x_2) = (0,\,0)\,,$$

which is a linear system in variables x_1, x_2. If $y \neq 2$, then $\hat{\mathbf{x}} = (4/(2+y), 4/(2+y))$ is the unique solution (and can be shown to be optimal). If $y = 2$, every $\hat{\mathbf{x}} \in \mathbb{R}^2$ with $\hat{x}_1 + \hat{x}_2 = 2$ solves the critical equation and is optimal:

$$\max_{\mathbf{x}} f(\mathbf{x}) - 2g(\mathbf{x}) = \max_{\mathbf{x}} 4(x_1 + x_2) - (x_1 + x_2)^2 + 2 = \max_{t} 4t - t^2 + 2 = 6.$$

In particular, $\overline{\mathbf{x}} = (1, 1)$ and $\overline{y} = 2$ are optimal primal resp. dual solutions (see Theorem 4.1).

Strong Duality. The *duality gap* of problem (4.1) is the difference between the dual and primal optimal value. We say that *strong duality* holds if the inequality in Theorem 4.1 is an equality, *i.e.*, if the duality gap is zero. (We do not necessarily require the existence of optimal solutions $\overline{\mathbf{x}}$ resp. $\overline{\mathbf{y}}$ achieving equality.) Unfortunately, strong duality is generally not guaranteed (see Ex. 4.4 for an extreme case of a non-zero duality gap).

Ex. 4.4. *Show that the duality gap is infinite for*

$$\max\ x_1 \quad \text{s.t.} \quad x_1^3 + x_2 \leq 0,\ x_2 \geq 0\ .$$

Partial Relaxation. Often one may want to dualize not all of the constraints $g_j(\mathbf{x}) \leq 0$. Then one can partition the set of constraints as

$$\max_{\mathbf{x}}\ f(\mathbf{x}) \quad \text{s.t.} \quad g_1(\mathbf{x}) \leq \mathbf{0},\ g_2(\mathbf{x}) \leq \mathbf{0}$$

and only dualize the constraints $g_1(\mathbf{x}) \leq \mathbf{0}$ with multipliers $\mathbf{y}_1$. In the same way as before, one thus obtains a *partial relaxation* and the weak duality relation

$$\max_{\substack{g_1(\mathbf{x}) \leq \mathbf{0} \\ g_2(\mathbf{x}) \leq \mathbf{0}}} f(\mathbf{x}) \ \leq\ \min_{\mathbf{y}_1 \geq \mathbf{0}} \max_{g_2(\mathbf{x}) \leq \mathbf{0}} f(\mathbf{x}) - \mathbf{y}_1^T g_1(\mathbf{x})\ . \tag{4.7}$$

Ex. 4.5. *Show: The more constraints are dualized, the weaker are the bounds offered by the ("partial") Lagrangian dual in (4.7).*

4.2. Lagrangian Duality

In order to analyze the relationship between the primal problem (4.1) and its dual (4.4), we define (by slightly misusing our notation) the associated *Lagrangian function* as a function in the variables $\mathbf{x}$ and $\mathbf{y}$:

$$L(\mathbf{x}, \mathbf{y}) \ =\ f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x})\ . \tag{4.8}$$

So we regain the function from Section 4.1 as $L(\mathbf{y}) = \max_{\mathbf{x}} L(\mathbf{x}, \mathbf{y})$.

REMARK. It is occasionally convenient to allow a function to attain the "values" $\pm\infty$. We do this with the understanding $-\infty \leq x \leq +\infty$ for all $x \in \mathbb{R}$, $\lambda \cdot (+\infty) = +\infty$ for $\lambda > 0$, $(+\infty) + (+\infty) = +\infty$, *etc.* (Note, however, that $(+\infty) - (+\infty)$ is undefined).

4.2.1. Saddle Points.

For any $\overline{\mathbf{x}} \in \mathbb{R}^n$ and $\overline{\mathbf{y}} \geq \mathbf{0}$, we (trivially) observe

$$\min_{\mathbf{y} \geq \mathbf{0}} L(\overline{\mathbf{x}}, \mathbf{y}) \ \leq\ L(\overline{\mathbf{x}}, \overline{\mathbf{y}}) \ \leq\ \max_{\mathbf{x}} L(\mathbf{x}, \overline{\mathbf{y}}) \tag{4.9}$$

and, therefore, conclude

$$\max_{\mathbf{x}} \min_{\mathbf{y} \geq \mathbf{0}} L(\mathbf{x}, \mathbf{y}) \ \leq\ \min_{\mathbf{y} \geq \mathbf{0}} \max_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}) \ =\ \min_{\mathbf{y} \geq \mathbf{0}} L(\mathbf{y})\ . \tag{4.10}$$

Relation (4.10) is the Weak Duality Theorem in disguise. Indeed, the left hand side of (4.10) is equivalent with the primal problem (4.1) since (*cf.* Ex. 4.6)

$$\min_{\mathbf{y}\geq\mathbf{0}} L(\mathbf{x},\mathbf{y}) = \min_{\mathbf{y}\geq\mathbf{0}} f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } g(\mathbf{x}) \leq \mathbf{0} \\ -\infty & \text{otherwise.} \end{cases} \tag{4.11}$$

EX. 4.6. *Show:* $\min_{\mathbf{y}\geq\mathbf{0}} \mathbf{w}^T\mathbf{y} = 0$ *if* $\mathbf{w} \geq \mathbf{0}$, *and* $\min_{\mathbf{y}\geq\mathbf{0}} \mathbf{w}^T\mathbf{y} = -\infty$ *otherwise.*

So we arrive at the pair of *primal-dual Lagrangian problems* :

$$(P) \qquad \max_{\mathbf{x}} \min_{\mathbf{y}\geq\mathbf{0}} L(\mathbf{x},\mathbf{y}) \qquad\qquad (D) \qquad \min_{\mathbf{y}\geq\mathbf{0}} \max_{\mathbf{x}} L(\mathbf{x},\mathbf{y}) \ .$$

We are particularly interested in the case where strong duality holds. If $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ achieve equality in (4.9) (and hence equality holds in (4.10)), we call the pair $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ a *saddle point* of the Lagrangian function $L(\mathbf{x},\mathbf{y})$. In this case we also say that $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ *simultaneously* solve the primal and dual problem (in the sense that $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ solves the primal max-min problem and $(\bar{\mathbf{y}},\bar{\mathbf{x}})$ solves the dual min-max problem). In particular, $\bar{\mathbf{x}}$ is an optimal solution of the primal (4.1), $\bar{\mathbf{y}}$ is an optimal solution of the dual (4.4) and the duality gap is zero (*cf.* Theorem 4.1). Also the converse is true:

THEOREM 4.2. *For any* $\bar{\mathbf{x}}$ *and any* $\bar{\mathbf{y}} \geq \mathbf{0}$, *the following are equivalent:*

(i) $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ *is a saddle point of the Lagrangian function* $L(\mathbf{x},\mathbf{y})$.

(ii) $\bar{\mathbf{x}}$ *and* $\bar{\mathbf{y}}$ *are optimal solutions of (4.1) resp. (4.4) and the duality gap is zero.*

Proof. It remains to show that (ii) implies (i). Assume that $\bar{\mathbf{x}}$ is primal feasible. Then $\mathbf{y}^T g(\bar{\mathbf{x}}) \leq 0$ holds for any $\mathbf{y} \geq \mathbf{0}$. Hence (ii) yields

$$\min_{\mathbf{y}\geq\mathbf{0}} L(\bar{\mathbf{x}},\mathbf{y}) = \min_{\mathbf{y}\geq\mathbf{0}} f(\bar{\mathbf{x}}) - \mathbf{y}^T g(\bar{\mathbf{x}}) \geq f(\bar{\mathbf{x}}) = L(\bar{\mathbf{y}}) = \max_{\mathbf{x}} L(\mathbf{x},\bar{\mathbf{y}}),$$

i.e., equality must hold in (4.9).

◇

REMARK. As in the linear case (*cf.* Section **??**) the min-max relation (4.10) may be interpreted game-theoretically: $L(\mathbf{x},\mathbf{y})$ is the payoff (gain) of player 1 when he chooses his strategy $\mathbf{x} \in \mathbb{R}^n$ and player 2 chooses strategy $\mathbf{y} \in \mathbb{R}^m_+$. The primal problem (of player 1) is to maximize his gain in the "worst case" (against all possible strategies of player 2). Similarly, the dual problem (of player 2) is to minimize his loss (= gain of player 1). In this context, saddle points correspond to *equilibrium strategies*: None of the players can expect any gain from changing his strategy ($\bar{\mathbf{x}}$ resp. $\bar{\mathbf{y}}$), even if he knew his opponent's strategy.

We stress that the apparent symmetry between (P) and (D) is deceptive, as $L(\mathbf{x},\mathbf{y})$ is *not* symmetric in $\mathbf{x}$ and $\mathbf{y}$. Indeed the dual variables y_j occur *linearly* in $L(\mathbf{x},\mathbf{y})$ as opposed to the primal variables x_i. Only in the linear case, *i.e.*, when

$f(\mathbf{x}) = \mathbf{c}^T\mathbf{x}$ is a linear objective and $g(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b} \leq \mathbf{0}$ are linear constraints, the Lagrangian function

$$L(\mathbf{x}, \mathbf{y}) = \mathbf{c}^T\mathbf{x} - \mathbf{y}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = (\mathbf{c}^T - \mathbf{y}^T\mathbf{A})\mathbf{x} + \mathbf{b}^T\mathbf{y}$$

is linear in both $\mathbf{x}$ and $\mathbf{y}$.

This explains why, in the linear case, Lagrangian duality reduces to linear programming duality: Indeed, as in (4.11), we deduce

$$L(\mathbf{y}) = \max_{\mathbf{x}} L(\mathbf{x}, \mathbf{y}) = \max_{\mathbf{x}} (\mathbf{c}^T - \mathbf{y}^T\mathbf{A})\mathbf{x} + \mathbf{b}^T\mathbf{y} = \begin{cases} \mathbf{b}^T\mathbf{y} & \text{if } \mathbf{c}^T - \mathbf{y}^T\mathbf{A} = \mathbf{0}^T \\ +\infty & \text{otherwise.} \end{cases}$$

So the Lagrangian dual is equivalent with the linear programming dual:

$$\min_{\mathbf{y} \geq \mathbf{0}} L(\mathbf{y}) \quad \longleftrightarrow \quad \min \mathbf{b}^T\mathbf{y} \quad \text{s.t. } \mathbf{y}^T\mathbf{A} = \mathbf{c}^T,\ \mathbf{y} \geq \mathbf{0}\ .$$

The equivalence in Theorem 4.2 indicates that saddle points are generally not easy to find (if they exist at all). Assuming f and g to be differentiable, we can reduce the number of candidates by solving the critical equation (4.6):

COROLLARY 4.1. *Every saddle point $(\overline{\mathbf{x}}, \overline{\mathbf{y}})$ of $L(\mathbf{x}, \mathbf{y})$ satisfies the condition*

$$\nabla f(\overline{\mathbf{x}}) - \overline{\mathbf{y}}^T \nabla g(\overline{\mathbf{x}}) = \mathbf{0}^T\ .$$

Proof. If $\overline{\mathbf{x}}$ solves $L(\overline{\mathbf{y}}) = \max_{\mathbf{x}} f(\mathbf{x}) - \overline{\mathbf{y}}^T g(\mathbf{x})$, the extremum principle (*cf.* (1.12)) with respect to the function $\bar{f}(\mathbf{x}) = f(\mathbf{x}) - \overline{\mathbf{y}}^T g(\mathbf{x})$ says that $\overline{\mathbf{x}}$ must satisfy the critical equation

$$\nabla \bar{f}(\overline{\mathbf{x}}) = \nabla f(\overline{\mathbf{x}}) - \overline{\mathbf{y}}^T \nabla g(\overline{\mathbf{x}}) = \mathbf{0}^T\ .$$

◇

4.2.2. The Lagrangian Dual and Convexity. The fact that $\mathbf{y}$ occurs linearly in $L(\mathbf{x}, \mathbf{y})$ has important consequences: The Lagrangian dual is always a so-called convex optimization problem.

REMARK. Convex functions and convex optimization problems will be studied in detail in Chapter 10. For our present purpose it suffices to know the definition: If $S \subseteq \mathbb{R}^n$ is a convex set, then the function $f : S \to \mathbb{R}$ (or, more generally $f : S \to \mathbb{R} \cup \{\infty\}$) is convex if for all $\mathbf{x}_1, \mathbf{x}_2 \in S$ and $\lambda \in [0, 1]$,

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

A *convex optimization problem* is a problem of type

$$\min f(\mathbf{x}) \qquad \text{s.t.} \quad \mathbf{x} \in S\ ,$$

where f is a convex function on S and $S \subseteq \mathbb{R}^n$ is a closed convex set. (As mentioned earlier, closed convex sets are exactly the intersections of (possibly infinitely many) half-spaces, *cf.* Corollary **??**.)

PROPOSITION 4.1. *The function $L(\mathbf{y}) = \max_{\mathbf{x}} f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x})$ is convex.*

Proof. Assume to the contrary that there exist $\mathbf{y}_1, \mathbf{y}_2 \in \mathbb{R}^m_+$ and $\lambda \in [0, 1]$ such that

$$L(\lambda \mathbf{y}_1 + (1-\lambda)\mathbf{y}_2) > \lambda L(\mathbf{y}_1) + (1-\lambda) L(\mathbf{y}_2). \tag{4.12}$$

Let $\overline{\mathbf{y}} = \lambda \mathbf{y}_1 + (1-\lambda)\mathbf{y}_2$. By definition of $L(\mathbf{y})$, we have

$$L(\overline{\mathbf{y}}) = \max_{\mathbf{x}} \; f(\mathbf{x}) - \overline{\mathbf{y}}^T g(\mathbf{x}).$$

So (4.12) implies the existence of some $\overline{\mathbf{x}} \in \mathbb{R}^n$ so that

$$f(\overline{\mathbf{x}}) - \overline{\mathbf{y}}^T g(\overline{\mathbf{x}}) > \lambda L(\mathbf{y}_1) + (1-\lambda) L(\mathbf{y}_2). \tag{4.13}$$

Again, by definition of $L(\mathbf{y})$,

$$\begin{array}{rcccl} f(\overline{\mathbf{x}}) - \mathbf{y}_1^T g(\overline{\mathbf{x}}) & \leq & \max\limits_{\mathbf{x}} \; f(\mathbf{x}) - \mathbf{y}_1^T g(\mathbf{x}) & = & L(\mathbf{y}_1) \\ f(\overline{\mathbf{x}}) - \mathbf{y}_2^T g(\overline{\mathbf{x}}) & \leq & \max\limits_{\mathbf{x}} \; f(\mathbf{x}) - \mathbf{y}_2^T g(\mathbf{x}) & = & L(\mathbf{y}_2). \end{array}$$

Multiplying these two inequalities with λ resp. $(1-\lambda)$ and adding them yields a contradiction to (4.13).

◇

Proposition 4.1 reveals a fundamental difference between the primal problem (4.1) and its dual (4.4): The dual is always a convex optimization problem. In particular, the dual of the dual cannot be (equivalent to) the primal, unless the primal is a convex problem itself. In Section 4.3 we will see that this condition is (in some sense) also sufficient.

REMARK. Proposition 4.1 can be used to derive a geometric interpretation of the dual as a convexification of the primal. We only sketch the result, which will be presented (and proved) in detail in Chapter 10. Assume for simplicity that there is only a single constraint $g(\mathbf{x}) \leq 0$. Introducing the set $G \subseteq \mathbb{R}^2$ defined by

$$G := \{ \begin{pmatrix} f \\ g \end{pmatrix} \mid f = f(\mathbf{x}),\; g = g(\mathbf{x}),\; \mathbf{x} \in \mathbb{R}^n \},$$

the primal resp. dual optimum values are

$$\begin{array}{rcl} v_P & = & \max\{ f \mid \begin{pmatrix} f \\ g \end{pmatrix} \in G,\; g \leq 0 \} \quad \text{and} \\ v_D & = & \min\limits_{y \geq 0} \max\{ f - yg \mid \begin{pmatrix} f \\ g \end{pmatrix} \in G \} \end{array}$$

It turns out that in case G is compact the optimum dual value v_D can equivalently be obtained as the optimum value v_C of the following convexification of the primal (*cf.* Section **??**, Figure**??**):

$$v_C = \max\{ f \mid \begin{pmatrix} f \\ g \end{pmatrix} \in \text{conv } G,\; g \leq 0 \}\,.$$

As a consequence, strong duality (duality gap zero) can be guaranteed for so-called *compact convex problems* (*cf.* Section **??**).

4.2.3. Solving the Lagrangian Dual. Let us outline the basic idea for solving the Lagrangian dual to get a better understanding of the primal-dual relationship. In what follows we assume that we can compute $L(\mathbf{y})$ and a corresponding optimal solution $\hat{\mathbf{x}}$ of

$$L(\mathbf{y}) = \max_{\mathbf{x}} \; f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x}) \tag{4.14}$$

for any fixed $\mathbf{y} \geq \mathbf{0}$ (otherwise we cannot expect to solve $\min \; L(\mathbf{y})$ at all).

Starting with an arbitrary $\mathbf{y}_0 \geq \mathbf{0}$, we construct a sequence $\mathbf{y}_0, \mathbf{y}_1, \mathbf{y}_2, \ldots$ that hopefully converges to an optimal solution $\overline{\mathbf{y}}$ of the Lagrangian dual. We proceed as follows. Given $\mathbf{y}_k \geq \mathbf{0}$, we solve (4.14) for $\mathbf{y} = \mathbf{y}_k$ by computing a vector $\mathbf{x}_k$ with

$$L(\mathbf{y}_k) = \; f(\mathbf{x}_k) - \mathbf{y}_k^T g(\mathbf{x}_k). \tag{4.15}$$

How should we modify $\mathbf{y} = \mathbf{y}_k$ and (possibly) decrease $L(\mathbf{y})$ in the next iteration? Intuitively, (4.15) suggests to increase y_j in case $g_j(\mathbf{x}_k) > 0$ (thereby increasing the *penalty* for violating the constraint $g_j(\mathbf{x}) \leq 0$). Similarly, we would decrease y_j if $g_j(\mathbf{x}_k) < 0$. This intuition suggests the 'update'

$$\mathbf{y}_{k+1} = \mathbf{y}_k + \delta_k g(\mathbf{x}_k) \quad \text{for some stepsize } \delta_k > 0 \; .$$

On the other hand we want to ensure $\mathbf{y}_{k+1} \geq \mathbf{0}$. We therefore take $\mathbf{y}_{k+1}$ as

$$\mathbf{y}_{k+1} \; = \; \max\left\{\mathbf{y}_k + \delta_k g(\mathbf{x}_k) \; , \; \mathbf{0}\right\} \qquad \text{(componentwise)} \; . \tag{4.16}$$

This strategy is the essence of the so-called *subgradient method* (*cf.* Chapter 10).

Ex. 4.7. *Assume $\mathbf{y}_{k+1} = \mathbf{y}_k$ holds in (4.16). Show: $\mathbf{y}_k = \overline{\mathbf{y}}$ solves the dual problem. (Hint: Use (4.17) below.)*

Ex. 4.8. *Consider the Lagrangian $L(y) = \max\limits_{\mathbf{x}} \; 4(x_1 + x_2) - x_1^2 - x_2^2 - y(x_1 x_2 - 1)$ from Ex. 4.3. For any fixed $y \geq 0$, the* maximum *is attained in*

$$\hat{\mathbf{x}} = 4((2+y)^{-1}, (2+y)^{-1}) \quad \textit{with} \quad g(\hat{\mathbf{x}}) = \hat{x}_1 \hat{x}_2 - 1 \geq 0 \iff y \leq 2 \; .$$

So the subgradient method will decrease a current $y_k > 2$ and increase a current $y_k < 2$. Show that the step sizes $\delta_k = 1/k$ imply $y_k \to \overline{y}$, the optimum solution of the dual (for any initial value $y_0 \geq 0$).

REMARK. The term *subgradient method* is motivated by the following consideration. Let $\overline{\mathbf{y}} \geq \mathbf{0}$ with $\hat{\mathbf{x}}$ the corresponding solution of (4.14). Then by definition of L,

$$L(\overline{\mathbf{y}} + \mathbf{h}) \; \geq \; f(\hat{\mathbf{x}}) - (\overline{\mathbf{y}} + \mathbf{h})^T g(\hat{\mathbf{x}}) \; = \; L(\overline{\mathbf{y}}) - g(\hat{\mathbf{x}})^T \mathbf{h}. \tag{4.17}$$

If $L(\mathbf{y})$ is differentiable at $\overline{\mathbf{y}}$, then (4.17) implies that $\nabla L(\overline{\mathbf{y}}) = -g(\hat{\mathbf{x}})^T$ holds (*cf.* Ex. 4.9). So step (4.16) is a move in the direction of the largest marginal decrease of $L(\mathbf{y})$. In general, however, $L(\mathbf{y})$ is not differentiable and there is no reason to expect that $L(\mathbf{y}_{k+1}) \leq L(\mathbf{y}_k)$ should hold (*cf.* Section **??**).

EX. 4.9. *Suppose $\ell : \mathbb{R}^m \to \mathbb{R}$ is differentiable in $\bar{\mathbf{y}} \in \mathbb{R}^m$. Furthermore, assume there exists $\mathbf{g} \in \mathbb{R}^m$ such that*

$$\ell(\bar{\mathbf{y}} + \mathbf{h}) \geq \ell(\bar{\mathbf{y}}) + \mathbf{g}^T\mathbf{h}$$

for all $\mathbf{h} \in \mathbb{R}^m$. Show that $\mathbf{g}^T = \nabla\ell(\bar{\mathbf{y}})$ must hold.

4.3. Cone Duality

Weak duality (Theorem 4.1) rests on the basic fact

$$g(\mathbf{x}) \leq \mathbf{0} \quad \text{and} \quad \mathbf{y} \geq \mathbf{0} \quad \Longrightarrow \quad \mathbf{y}^T g(\mathbf{x}) \leq 0 \ .$$

This observation suggests to re-state the Weak Duality Theorem in a slightly more general setting: Instead of constraints $g(\mathbf{x}) \leq \mathbf{0}$ (as in the optimization model (4.1)) we allow constraints of the form $g(\mathbf{x}) \in K$, where $K \subseteq \mathbb{R}^m$ is a cone. Correspondingly, the dual variables $\mathbf{y}$ are then chosen in the dual cone K^0.

REMINDER. Recall from Section 3.2 that every cone $K \subseteq \mathbb{R}^m$ has an associated dual cone

$$K^0 \ = \{\mathbf{y} \in \mathbb{R}^m \mid \mathbf{y}^T\mathbf{g} \leq 0 \quad \text{for all } \mathbf{g} \in K\}.$$

Furthermore, $K = K^{00}$ holds if and only if K is the intersection of (possibly infinitely many) halfspaces, *i.e.*,

$$K = \{\mathbf{x} \in \mathbb{R}^m \mid \mathbf{a}_j^T\mathbf{x} \leq 0, \ \ j \in J\} \ .$$

THEOREM 4.3. *Let $K \subseteq \mathbb{R}^m$ be a cone. Then*

$$\max_{g(\mathbf{x}) \in K} f(\mathbf{x}) \quad \leq \quad \min_{\mathbf{y} \in K^0} L(\mathbf{y}) \ .$$

If equality is achieved at a primal feasible $\bar{\mathbf{x}}$ and a dual feasible $\bar{\mathbf{y}}$, then $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are primal resp. dual optimal and complementary, i.e., $\bar{\mathbf{y}}^T g(\bar{\mathbf{x}}) = 0$.

Proof. Assume $\bar{\mathbf{x}}$ is primal feasible, *i.e.*, $g(\bar{\mathbf{x}}) \in K$. Then $\mathbf{y}^T g(\bar{\mathbf{x}}) \leq 0$ for every $\mathbf{y} \in K^0$. Hence for every $\mathbf{y} \in K^0$ we have

$$f(\bar{\mathbf{x}}) \leq \ f(\bar{\mathbf{x}}) - \mathbf{y}^T g(\bar{\mathbf{x}}) \leq \max_{\mathbf{x}} \ f(\mathbf{x}) - \mathbf{y}^T g(\mathbf{x}) = L(\mathbf{y})$$

As this holds for each primal feasible $\bar{\mathbf{x}}$ (*i.e.*, $g(\bar{\mathbf{x}}) \in K$) and each dual feasible $\mathbf{y}$ (*i.e.*, $\mathbf{y} \in K^0$), the Theorem follows.

◇

REMARK. Purely formally, also problems of type $\max\{f(\mathbf{x}) \mid g(\mathbf{x}) \in K\}$ can be cast into the form (4.1). For example, one could define $\widetilde{g} : \mathbb{R}^n \to \mathbb{R}$ as

$$\widetilde{g}(\mathbf{x}) = \begin{cases} 0 & \text{if } g(\mathbf{x}) \in K \\ 1 & \text{otherwise} \end{cases} \qquad \text{or} \qquad \widetilde{g}(\mathbf{x}) = \min_{\mathbf{z} \in K} \|g(\mathbf{x}) - \mathbf{z}\|^2$$

and consider the equivalent problem $\max\{f(\mathbf{x}) \mid \widetilde{g}(\mathbf{x}) \leq 0\}$. Note, however, that the dual of a problem depends on the *constraint functions* rather than the *feasible set* they define. The practical solvability of a problem often depends critically on an "appropriate" formulation.

4.3.1. Examples. Interesting examples are obtained by "coning" the constraints in a primal-dual pair of linear programs,

$$\begin{array}{lll} \max\ \mathbf{c}^T\mathbf{x} & \text{resp.} & \min\ \mathbf{b}^T\mathbf{y} \\ s.t.\ \ \mathbf{A}\mathbf{x}-\mathbf{b}\le\mathbf{0} & & s.t.\ \ \mathbf{c}-\mathbf{A}^T\mathbf{y}\le\mathbf{0} \\ \mathbf{x}\ge\mathbf{0} & & \mathbf{y}\ge\mathbf{0} \end{array} \tag{4.18}$$

The resulting weak duality relation is the following.

COROLLARY 4.2. *Let $K\subseteq\mathbb{R}^n$ and $M\subseteq\mathbb{R}^m$ be arbitrary cones. Then*

$$\begin{array}{lll} \max\ \mathbf{c}^T\mathbf{x} & \le & \min\ \mathbf{b}^T\mathbf{y} \\ s.t.\ \ \mathbf{A}\mathbf{x}-\mathbf{b}\in K^0 & & s.t.\ \ \mathbf{c}-\mathbf{A}^T\mathbf{y}\in M^0 \\ \mathbf{x}\in M & & \mathbf{y}\in K \end{array} \tag{4.19}$$

Moreover, if $K=K^{00}$ and $M=M^{00}$, the two problems are dual to each other.

Proof. Let $\mathbf{x}$ and $\mathbf{y}$ be primal resp. dual feasible. Then

$$\mathbf{c}^T\mathbf{x}\ \le\ \mathbf{c}^T\mathbf{x}-\mathbf{y}^T(\mathbf{A}\mathbf{x}-\mathbf{b})\ =\ (\mathbf{c}^T-\mathbf{y}^T\mathbf{A})\mathbf{x}+\mathbf{y}^T\mathbf{b}\ \le\ \mathbf{y}^T\mathbf{b}\,.$$

The way the dual (right hand side in (4.19)) is constructed from the primal (left hand side in (4.19)) immediately implies that the dual of the dual equals the primal in case $K=K^{00}$ and $M=M^{00}$.

◇

EX. 4.10. *Show that for polyhedral cones $K=P(\mathbf{B},\mathbf{0})$ and $M=P(\mathbf{C},\mathbf{0})$ the two problems in (4.19) are a primal-dual pair of linear programs.*

Convex Problems. In general, if K and M are arbitrary cones with the property that $K=K^{00}$ and $M=M^{00}$, the two problems in (4.19) are convex optimization problems (as defined at the beginning of Section 4.2.2).

Conversely, consider an arbitrary convex optimization problem $\min\{f(\mathbf{x})\mid\mathbf{x}\in S\}$ where $f:\mathbb{R}^n\to\mathbb{R}$ is convex and $S\subseteq\mathbb{R}^n$ is a closed convex set. We may assume *w.l.o.g.* (*cf.* Ex. 4.11) that $f(\mathbf{x})=\mathbf{c}^T\mathbf{x}$ is linear. Since S is a closed convex set, it is the intersection of (possibly infinitely many) halfspaces, *i.e.*,

$$S=\{\mathbf{x}\in\mathbb{R}^n\mid\mathbf{a}_j^T\mathbf{x}\le b_j,\ \ j\in J\}.$$

Setting $M:=\{\binom{\mathbf{x}}{x_{n+1}}\in\mathbb{R}^{n+1}\mid\mathbf{a}_j^T\mathbf{x}-b_jx_{n+1}\le 0\}$ and $K:=\mathbb{R}$ with $K^0=\{0\}$, we can write our convex problem equivalently as

$$\max -\mathbf{c}^T\mathbf{x}\quad\text{s.t. } x_{n+1}-1\in K^0,\ \begin{pmatrix}\mathbf{x}\\ x_{n+1}\end{pmatrix}\in M.$$

Hence any convex problem can be stated (in a rather natural way) as a problem of type (4.19) with $K=K^{00}$ and $M=M^{00}$. In this sense the problems in (4.19) can be considered as the most general class of problems for which the dual of the dual is the primal.

Ex. 4.11. *Suppose $S \subseteq \mathbb{R}^n$ is a closed convex set and $f : S \to \mathbb{R}$ is a convex function. Show that $S' := \{\binom{\mathbf{x}}{z} \mid \mathbf{x} \in S,\ f(\mathbf{x}) \le z\}$ is a closed convex set and that* $\min\{f(\mathbf{x}) \mid \mathbf{x} \in S\}$ *equals* $\min\{z \mid \binom{\mathbf{x}}{z} \in S'\}$.

Semidefinite Programs. Particularly interesting examples are obtained from (4.19) by taking $K \subseteq \mathbb{S}^{k\times k}$ to be the cone of positive semidefinite $k \times k$ matrices with $K^0 = \{\mathbf{S} \in \mathbb{S}^{k\times k} \mid \mathbf{S} \preceq \mathbf{0}\}$ (*cf.* Section 3.2). So the dual variables $\mathbf{y}$ are considered as (vectors corresponding to) $k \times k$ matrices $\mathbf{Y} = (y_{ij}) \in \mathbb{S}^{k\times k}$ and, correspondingly we also interprete $\mathbf{b} = (b_{ij})$ as a matrix $\mathbf{B} \in \mathbb{S}^{k\times k}$ and every column $\mathbf{A}_{\cdot i}$ of $\mathbf{A}$ as a matrix $\mathbf{A}^{(i)} \in \mathbb{S}^{k\times k}$. Recalling our notation $\mathbf{B} \circ \mathbf{Y} = \sum_{i,j} b_{ij}y_{ij}$ for the "inner product" of matrices, (4.19) becomes (with $M = \mathbb{R}^n$)

$$
\begin{array}{lcll}
\max\ \mathbf{c}^T\mathbf{x} & \le & \min\ \mathbf{B}\circ\mathbf{Y} & \\
s.t.\ \ \sum_{i=1}^{n} \mathbf{A}^{(i)}x_i - \mathbf{B} \preceq \mathbf{0} & & s.t.\ \ \mathbf{A}^{(i)}\circ\mathbf{Y} = c_i & i = 1,\ldots,n \\
 & & \mathbf{Y} \succeq \mathbf{0} &
\end{array}
\tag{4.20}
$$

Such problems, maximizing or minimizing a linear objective under linear and semidefinite constraints, are called *semidefinite programs*.

REMARK. In Chapter 9 we will see how semidefinite programs arise in a natural way as Lagrangian relaxations of (certain) integer programming problems. We study semidefinite programs in more detail in Section **??**.

4.4. Optimality Conditions

We now return to *nonlinear* optimization problems of the form (4.1), *i.e.*,

$$\max_{\mathbf{x}\in\mathcal{F}} f(\mathbf{x}), \quad \text{where } \mathcal{F} = \{\mathbf{x} \in \mathbb{R}^n \mid g(\mathbf{x}) \le \mathbf{0}\}.$$

The objective function f and the constraint function g are (possibly) nonlinear functions. $\mathcal{F}$ is called the set of *feasible solutions* of the optimization problem. An *optimal solution* is, by definition, a feasible point $\bar{\mathbf{x}} \in \mathcal{F}$ with maximum objective value $f(\bar{\mathbf{x}})$. Computing an optimal solution can be extremely difficult. Even checking whether a given candidate vector $\bar{\mathbf{x}}$ is indeed optimal is generally a very hard task.

Since the computation of an overall optimal solution is so difficult, nonlinear optimization usually tries to at least identify *locally optimal solutions* (which is generally hard enough). We say that $\bar{\mathbf{x}} \in \mathcal{F}$ is a *local maximizer* (or simply a *maximizer*) if for some $\varepsilon > 0$:

$$f(\bar{\mathbf{x}}) \ge f(\mathbf{x}) \quad \text{holds for all } \mathbf{x} \in \mathcal{F} \ \text{ with } \ \|\bar{\mathbf{x}} - \mathbf{x}\| < \varepsilon. \tag{4.21}$$

If (4.21) is true for all $\varepsilon > 0$, $\bar{\mathbf{x}}$ is a *global maximizer*. Local resp. global *minimizers* are defined in the same way for minimization problems.

EX. 4.12. *Give an example of a polytope* $\mathcal{F} = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}\}$ *and a point* $\mathbf{x}_0 \in \mathbb{R}^n$ *so that each vertex of* $\mathcal{F}$ *is a local maximizer of the problem*

$$\max \ f(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}_0\|^2 \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b} \ .$$

4.4.1. Linear Constraints. We first take a look at the case of linear constraints $\mathbf{a}_j^T\mathbf{x} \leq b_j$ $(j = 1, \ldots, m)$. So we consider

$$\max \ f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b} \ , \tag{4.22}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the matrix with rows $\mathbf{a}_j^T$ and $f : \mathbb{R}^n \to \mathbb{R}$ is a differentiable function. The feasible set $\mathcal{F}$ is the polyhedron $P(\mathbf{A}, \mathbf{b})$.

Trying to decide whether $\bar{\mathbf{x}} \in \mathcal{F}$ is locally optimal, we are mainly interested in the constraints $\mathbf{a}_j^T\mathbf{x} \leq b_j$ that $\bar{\mathbf{x}}$ satisfies with equality (*cf.* Ex. 4.13). We call these constraints *tight* or *active* at $\bar{\mathbf{x}}$ and refer to

$$J(\bar{\mathbf{x}}) \ = \ \{j \mid \mathbf{a}_j^T\bar{\mathbf{x}} = b_j\} \ \subseteq \ \{1, \ldots, m\}$$

as the corresponding *active set* (of indices).

EX. 4.13. *Show: There exists some* $\varepsilon > 0$ *such that every* $\mathbf{x} \in \mathbb{R}^n$ *with* $\|\bar{\mathbf{x}} - \mathbf{x}\| < \varepsilon$ *satisfies all constraints nonactive at* $\bar{\mathbf{x}}$ *with strict inequality.*

A *feasible direction* at $\bar{\mathbf{x}}$ is a vector $\mathbf{d} \in \mathbb{R}^n$ such that

$$\mathbf{a}_j^T\mathbf{d} \leq 0 \qquad \text{for all } j \in J(\bar{\mathbf{x}}).$$

We denote by $D(\bar{\mathbf{x}}) \subseteq \mathbb{R}^n$ the (polyhedral) *cone of feasible directions*. Let $\mathbf{d} \in D(\bar{\mathbf{x}})$ with $\|\mathbf{d}\| = 1$. Then, in view of Ex. 4.13, we can find some $\varepsilon > 0$ so that

$$\bar{\mathbf{x}} + t\mathbf{d} \in \mathcal{F} \qquad \text{for all } 0 < t \leq \varepsilon. \tag{4.23}$$

If $\bar{\mathbf{x}}$ is a maximizer of f, then the differentiability of f yields

$$0 \ \geq \ f(\bar{\mathbf{x}} + t\mathbf{d}) - f(\bar{\mathbf{x}}) \ = \ t\nabla f(\bar{\mathbf{x}})\mathbf{d} + o(t).$$

Dividing by $t > 0$ and then letting $t \to 0$, we therefore conclude that $\nabla f(\bar{\mathbf{x}})\mathbf{d} \leq 0$ must hold. This is the necessary optimality condition we seek.

THEOREM 4.4 (Necessary Optimality Conditions)**.** *Every maximizer* $\bar{\mathbf{x}} \in \mathcal{F}$ *of (4.22) satisfies the following two equivalent conditions:*

(a) (Primal Condition)

$$\nabla f(\bar{\mathbf{x}})\mathbf{d} \leq 0 \quad \textit{for all} \quad \mathbf{d} \in D(\bar{\mathbf{x}}) \ .$$

(b) (Dual Condition) *There are multipliers* $y_j \geq 0$, $j \in J(\bar{\mathbf{x}})$, *such that*

$$\nabla f(\bar{\mathbf{x}}) - \sum_{j\in J(\bar{x})} y_j\mathbf{a}_j^T = \mathbf{0}^T.$$

Proof. We have seen that (a) is a necessary condition for optimality. We show that (b) is equivalent with (a). Now (a) means that the inequality $\nabla f(\overline{\mathbf{x}})\mathbf{d} \leq 0$ is implied by the systems of inequalities $\mathbf{a}_j^T\mathbf{d} \leq 0,\ j \in J(\overline{\mathbf{x}})$. By Farkas Lemma (*cf.* Corollary 2.6), this is equivalent with $\nabla f(\overline{\mathbf{x}})$ being a nonnegative combination of the vectors $\mathbf{a}_j,\ j \in J(\overline{\mathbf{x}})$, *i.e.*, with (b).

◇

REMARK. We emphasize that conditions (a) and (b) are just *necessary* conditions and at most exhibit *candidates* $\overline{\mathbf{x}}$ for being maximizers. Sometimes also *sufficient* conditions for optimality can be given (that guarantee $\overline{\mathbf{x}}$ to be indeed a maximizer). Such conditions typically require information about second order derivatives (see Chapter 12 for more details).

The necessary dual condition (b) is often stated in a slightly different form.

$$\overline{y}_j \ = \ \begin{cases} y_j & \text{if } j \in J(\overline{\mathbf{x}}), \\ 0 & \text{if } j \notin J(\overline{\mathbf{x}}) \ , \end{cases}$$

yields a vector $\overline{\mathbf{y}} \in \mathbb{R}^m_+$ of multipliers such that $\overline{\mathbf{x}}$ and $\overline{\mathbf{y}} \geq 0$ are *complementary* in the sense that for all j:

$$\mathbf{a}_j^T\overline{\mathbf{x}} < b_j \quad \Longrightarrow \quad \overline{y}_j = 0 \qquad (i.e.,\ \overline{\mathbf{y}}^T(\mathbf{A}\overline{\mathbf{x}} - \mathbf{b}) = 0) \ .$$

In other words, the dual condition (b) in Theorem 4.4 is equivalent with the so-called *Karush-Kuhn-Tucker* conditions (or KKT-*conditions*, for short):

$$\nabla f(\mathbf{x}) - \mathbf{y}^T\mathbf{A} = \mathbf{0}^T \ , \quad \mathbf{y} \geq \mathbf{0} \ , \quad \mathbf{y}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = 0. \tag{4.24}$$

We say that the feasible point $\overline{\mathbf{x}} \in \mathcal{F}$ is a *Kuhn-Tucker point* (or KKT-point for short) if $\overline{\mathbf{x}}$ satisfies (4.24) with suitable multipliers $\overline{\mathbf{y}} \in \mathbb{R}^m_+$.

REMARK. The reader may have noticed that the KKT-condition (4.24) for a local maximizer is a special case of the necessary condition $\nabla f(\mathbf{x}) - \mathbf{y}^T\nabla g(\mathbf{x}) = \mathbf{0}^T$ we derived for saddle points in Corollary 4.1 (because $g(\mathbf{x}) = \mathbf{A}\mathbf{x} - \mathbf{b}$ has the Jacobian $\nabla g(\mathbf{x}) = \mathbf{A}$). This is not surprising, indeed, a saddle point $(\overline{\mathbf{x}}, \overline{\mathbf{y}})$ of the Lagrangian $L(\mathbf{x}, \mathbf{y})$ always implies $\overline{\mathbf{x}}$ to be a local (even a global) maximizer.
On the other hand, a local maximizer usually is not even a kind of "local saddle point" of the Lagrangian $L(\mathbf{x}, \mathbf{y}) = f(\mathbf{x}) - \mathbf{y}^T(\mathbf{A}\mathbf{x} - \mathbf{b})$. The two concepts are quite different (in spite of the formal similarity of the necessary conditions they imply).

Ex. 4.14 ("Equality and Inequality Constraints")**.** *Show: Every maximizer of*

$$\max f(\mathbf{x}) \quad s.t. \quad \mathbf{B}\mathbf{x} = \mathbf{d}, \quad \mathbf{A}\mathbf{x} \leq \mathbf{b} \ ,$$

where $\mathbf{B} \in \mathbb{R}^{k\times n}$ *and* $\mathbf{A} \in \mathbb{R}^{m\times n}$*, satisfies the KKT-condition*

$$\nabla f(\mathbf{x}) - \boldsymbol{\lambda}^T\mathbf{B} - \boldsymbol{\mu}^T\mathbf{A} = \mathbf{0}^T \ , \quad \boldsymbol{\mu}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) = 0 \ , \ \boldsymbol{\lambda} \in \mathbb{R}^k \ , \ \boldsymbol{\mu} \in \mathbb{R}^m_+ \ .$$

4.4.2. General Constraints. In the presence of general nonlinear constraints $g(\mathbf{x}) \leq \mathbf{0}$ (where we assume $g : \mathbb{R}^n \to \mathbb{R}^m$ to be differentiable), we could try to take a similar approach: We first *linearize* the constraints, *i.e.*, replace them by their first order approximations, and then proceed as before.

Given $\bar{\mathbf{x}} \in \mathcal{F}$, we again consider the corresponding *active set*

$$J(\bar{\mathbf{x}}) \;=\; \{j \mid g_j(\bar{\mathbf{x}}) = 0\} \subseteq \{1, \ldots, m\} \; .$$

An (approximately) *feasible direction* at $\bar{\mathbf{x}}$ is then a vector $\mathbf{d} \in \mathbb{R}^n$ such that

$$\nabla g_j(\bar{\mathbf{x}})\mathbf{d} \;\leq\; 0 \qquad \text{for all } j \in J(\bar{\mathbf{x}}).$$

If $\mathbf{d} \neq \mathbf{0}$ is such a direction at the maximizer $\bar{\mathbf{x}}$ and (4.23) holds, the same argument as before yields $\nabla f(\bar{\mathbf{x}})\mathbf{d} \leq 0$. The problem is that (4.23) need no longer hold. Indeed, we may not be able to move into *any* feasible direction $\mathbf{d}$ without leaving the feasible set $\mathcal{F}$ immediately (*cf.* Ex. 4.15).

Under certain assumptions on the constraint functions $g_j(\mathbf{x})$ (so-called *constraint qualifications*), one can argue that the cone of feasible directions $D(\bar{\mathbf{x}})$ approximates the feasible set $\mathcal{F}$ (locally at $\bar{\mathbf{x}}$) sufficiently well so that this problem can be overcome by moving along a *feasible curve* in $\mathcal{F}$ leading approximately (rather than exactly) into direction $\mathbf{d} \in D(\bar{\mathbf{x}})$. These (truly nonlinear) phenomena are discussed in detail in Chapter 12. One obtains necessary conditions that again are formally the same as the saddle point conditions of KKT-type

$$\nabla f(\mathbf{x}) - \mathbf{y}^T \nabla g(\mathbf{x}) = \mathbf{0}^T \; , \quad \mathbf{y} \geq \mathbf{0} \; , \quad \mathbf{y}^T g(\mathbf{x}) = 0 \tag{4.25}$$

(which we already know to be necessary saddle point conditions without any assumptions on the constraints).

Ex. 4.15. *For* $\max\{x_1 \mid g(\mathbf{x}) = \|\mathbf{x}\|^2 \leq 1\}$ *or* $\max\{x_1 \mid h(\mathbf{x}) = \|\mathbf{x}\|^2 = 1\}$, *determine the cone of feasible directions* $D(\bar{\mathbf{x}})$ *in* $\bar{\mathbf{x}} \in \mathcal{F}$ *and find which directions have the property (4.23).*

Minimization Problems. In the case of minimization problems, where we maximize $(-f)$, the KKT-conditions for a minimizer $\bar{\mathbf{x}}$ are, of course, obtained when we replace $\nabla f(\bar{\mathbf{x}})$ in (4.25) by $(-\nabla f(\bar{\mathbf{x}}))$. After multiplication with (-1), the KKT-conditions for a minimizer $\bar{\mathbf{x}}$ therefore become

$$\nabla f(\bar{\mathbf{x}}) + \mathbf{y}^T \nabla g(\bar{\mathbf{x}}) = \mathbf{0}^T \; , \quad \mathbf{y} \geq \mathbf{0} \; , \quad \mathbf{y}^T g(\bar{\mathbf{x}}) = 0 \; . \tag{4.26}$$

CHAPTER 5

Integer Programming

An *integer linear program* (ILP) is, by definition, a linear program with the additional constraint that all variables take integer values:

$$\max\ \mathbf{c}^T\mathbf{x} \quad s.t. \quad \mathbf{Ax} \leq \mathbf{b}\ \ and\ \ \mathbf{x}\ \text{integral}\ . \tag{5.1}$$

Integrality restrictions occur in many situations. For example, the products in a linear production model (*cf.* p. **??**) might be "indivisible goods" that can only be produced in integer multiples of one unit. Many problems in operations research and combinatorial optimization can be formulated as ILPs. As integer programming is NP-hard (see Section **??**), every NP-problem can in principle be formulated as an ILP. In fact, such problems usually admit many different ILP formulations. Finding a particularly suited one is often a decisive step towards the solution of a problem.

5.1. Formulating an Integer Program

In this section we present a number of (typical) examples of problems with their corresponding ILP formulations.

Graph Coloring. Let us start with the combinatorial problem of coloring the nodes of a graph $G = (V, E)$ so that no two adjacent nodes receive the same color and as few colors as possible are used (*cf.* Section **??**). This problem occurs in many applications. For example, the nodes may represent "jobs" that can each be executed in one unit of time. An edge joining two nodes may indicate that the corresponding jobs cannot be executed in parallel (because they use perhaps common resources). In this interpretation, the graph G would be the *conflict graph* of the given set of jobs. The minimum number of colors needed to color its nodes equals the number of time units necessary to execute all jobs.

Formulating the *node coloring problem* as an ILP, we assume $V = \{1, \ldots, n\}$ and that we have n colors at our disposal. We introduce binary variables y_k, $k = 1, \ldots, n$, to indicate whether color k is used ($y_k = 1$) or not ($y_k = 0$). Furthermore, we introduce variables x_{ik} to indicate whether node i receives color k.

The resulting ILP is

$$
\begin{array}{lllll}
\min \sum_{k=1}^{n} y_k \quad s.t. & (1) & \sum_{k=1}^{n} x_{ik} & = & 1 \quad i = 1, \ldots, n \\
& (2) & x_{ik} - y_k & \leq & 0 \quad i, k = 1, \ldots, n \\
& (3) & x_{ik} + x_{jk} & \leq & 1 \quad (i, j) \in E,\ k = 1, \ldots, n \\
& (4) & 0 \leq x_{ik}, y_k & \leq & 1 \\
& (5) & x_{ik}, y_k & \in & \mathbb{Z}
\end{array} \tag{5.2}
$$

The constraints (4) and (5) ensure that the x_{ik} and y_k are binary variables. The constraints (1)–(3) guarantee (in this order) that each node is colored, node i receives color k only if color k is used at all, and any two adjacent nodes have different colors.

Ex. 5.1. *Show: If the integrality constraint (5) is removed, the resulting linear program has optimum value equal to 1.*

The Traveling Salesman Problem (TSP). This is one of the best-known combinatorial optimization problems: There are n towns and a "salesman", located in town 1, who is to visit each of the other $n-1$ towns exactly once and then return home. The tour (*traveling salesman tour*) has to be chosen so that the total distance traveled is minimized. To model this problem, consider the so-called *complete graph* K_n, *i.e.*, the graph $K_n = (V, E)$ with $n = |V|$ pairwise adjacent nodes. With respect to a given cost (distance) function $\mathbf{c} : E \to \mathbb{R}$ we then seek to find a *Hamilton circuit* $C \subseteq E$, *i.e.*, a circuit including every node, of minimal cost.

An ILP formulation can be obtained as follows. We introduce binary variables x_{ik} $(i, k = 1, \ldots, n)$ to indicate whether node i is the kth node visited. In addition, we introduce variables y_e $(e \in E)$ to record whether edge e is traversed:

$$
\begin{array}{lrcll}
& \min \sum_{e \in E} c_e y_e & & & \\
s.t. & x_{11} & = & 1 & \\
& \sum_{k=1}^{n} x_{ik} & = & 1 & i = 1, \ldots, n \\
& \sum_{i=1}^{n} x_{ik} & = & 1 & k = 1, \ldots, n \\
& \sum_{e} y_e & = & n & \\
& x_{i,k-1} + x_{jk} - y_e & \leq & 1 & e = (i, j), k \geq 2 \\
& x_{in} + x_{11} - y_e & \leq & 1 & e = (i, 1) \\
& 0 \leq x_{ik}, y_e & \leq & 1 & \\
& x_{ik}, y_e & \in & \mathbb{Z} &
\end{array} \tag{5.3}
$$

Ex. 5.2. *Show that each feasible solution of (5.3) corresponds to a Hamilton circuit and conversely.*

In computational practice, other TSP formulations have proved more efficient. To derive an alternative formulation, consider first the following simple program with edge variables y_e, $e \in E$:

$$\begin{array}{lrcll} \min \mathbf{c}^T\mathbf{y} \quad s.t. & \mathbf{y}(\delta(i)) & = & 2 & i = 1, \ldots, n \\ & \mathbf{0} \leq \mathbf{y} & \leq & \mathbf{1} & \\ & \mathbf{y} & & \text{integral}. & \end{array} \tag{5.4}$$

(Recall our shorthand notation $\mathbf{y}(\delta(i)) = \sum_{e \in \delta(i)} y_e$ for the sum of all $\mathbf{y}$-values on edges incident with node i.)

ILP (5.4) does *not* describe our problem correctly: We still must rule out solutions corresponding to disjoint circuits that cover all nodes. We achieve this by adding more inequalities, so-called *subtour elimination constraints*. To simplify the notation, we write for $\mathbf{y} \in \mathbb{R}^E$ and two disjoint subsets $S, T \subseteq V$

$$\mathbf{y}(S : T) = \sum_{\substack{e = (i,\, j) \\ i \in S,\, j \in T}} y_e.$$

The subtour elimination constraints

$$\mathbf{y}(S : \overline{S}) \geq 2$$

make sure that there will be at least two edges in the solution that lead from a proper nonempty subset $S \subset V$ to its complement $\overline{S} = V \setminus S$. So the corresponding tour is connected. A correct ILP-formulation is thus given by

$$\begin{array}{lrcll} \min \mathbf{c}^T\mathbf{y} \quad s.t. & \mathbf{y}(\delta(i)) & = & 2 & i = 1, \ldots, n \\ & \mathbf{y}(S : \overline{S}) & \geq & 2 & \emptyset \subset S \subset V \\ & \mathbf{0} \leq \mathbf{y} & \leq & \mathbf{1} & \\ & \mathbf{y} & & \text{integral}. & \end{array} \tag{5.5}$$

Note the contrast to our first formulation (5.3): ILP (5.5) has exponentially many constraints, one for each proper subset $S \subset V$. If $n = 30$, there are more than 2^{30} constraints. Yet, the way to solve (5.5) in practice is to add even more constraints! This approach of adding so-called cutting planes is presented in Sections 5.2 and 5.3 below.

REMARK. The mere fact that (5.5) has exponentially many constraints does not prevent us from solving it (without the integrality constraints) efficiently (*cf.* Section **??**).

Maximum Clique. This is another well-studied combinatorial problem, which we will use as a case study for integer programming techniques later. Consider again the complete graph $K_n = (V, E)$ on n nodes. This time, there are weights $\mathbf{c} \in \mathbb{R}^V$ and $\mathbf{d} \in \mathbb{R}^E$ given on both the vertices and the edges. We look for a set $C \subseteq V$ that maximizes the total weight of vertices and induced edges:

$$\max_{C \subseteq V} \mathbf{c}(C) + \mathbf{d}(E(C)). \tag{5.6}$$

As $K_n = (V, E)$ is the complete graph, each $C \subseteq V$ is a clique (set of pairwise adjacent nodes). Therefore, we call (5.6) the *maximum weighted clique problem*.

Ex. 5.3. *Given a graph $G = (V, E')$ with $E' \subseteq E$, choose $\mathbf{c} = \mathbf{1}$ and*

$$d_e = \begin{cases} 0 & e \in E' \\ -n & \text{otherwise} \end{cases}$$

Show: With these parameters (for $K_n = (V, E)$), (5.6) reduces to the problem of finding a clique C in G of maximum cardinality.

Problem (5.6) admits a rather straightforward ILP-formulation:

$$\begin{array}{rrcll} \max & \mathbf{c}^T\mathbf{x} + \mathbf{d}^T\mathbf{y} & & & \\ & y_e - x_i & \leq & 0 & e \in E, i \in e \\ & x_i + x_j - y_e & \leq & 1 & e = (i, j) \in E \\ & \mathbf{0} \leq \mathbf{x}, \mathbf{y} & \leq & \mathbf{1} & \\ & \mathbf{x}, \mathbf{y} & & \text{integer} & \end{array} \tag{5.7}$$

A vector $(\mathbf{x}, \mathbf{y})$ with all components $x_i, y_e \in \{0, 1\}$ that satisfies the constraints of (5.7) is the so-called *(vertex-edge) incidence vector* of the clique

$$C = \{i \in V \mid x_i = 1\} .$$

In other words, $\mathbf{x} \in \mathbb{R}^V$ is the incidence vector of C and $\mathbf{y} \in \mathbb{R}^E$ is the incidence vector of $E(C)$.

REMARK. The reader may have noticed that all ILPs we have formulated so far are binary programs, *i.e.*, the variables are restricted to take values in $\{0, 1\}$ only. This is not by pure accident. The majority of integer optimization problems can be cast in this setting. But of course, there are also others (*e.g.*, the integer linear production model mentioned in the introduction to this chapter).

5.2. Cutting Planes I

Consider the integer linear program

$$\max \ \mathbf{c}^T\mathbf{x} \quad s.t. \quad \mathbf{Ax} \leq \mathbf{b} \ \textit{and} \ \mathbf{x} \text{ integral} . \tag{5.8}$$

For the following structural analysis it is important (see Ex. 5.4) to assume that $\mathbf{A}$ and $\mathbf{b}$ are rational, *i.e.*, $\mathbf{A} \in \mathbb{Q}^{m \times n}$ and $\mathbf{b} \in \mathbb{Q}^m$. In this case, the polyhedron

$$P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} \leq \mathbf{b}\} \tag{5.9}$$

is a rational polyhedron (*cf.* Section 3.6). The set of integer vectors in P is a discrete set, whose convex hull we denote by

$$P_I = \text{conv}\,\{\mathbf{x} \in \mathbb{Z}^n \mid \mathbf{Ax} \leq \mathbf{b}\} . \tag{5.10}$$

Solving (5.8) is equivalent with maximizing $\mathbf{c}^T\mathbf{x}$ over the convex set P_I (Why?). Below, we shall prove that also P_I is a polyhedron and derive a system of inequalities describing P_I. We thus show how (at least in principle) the original problem (5.8) can be reduced to a linear program.

Ex. 5.4. *Give an example of a (non-rational) polyhedron $P \subseteq \mathbb{R}^n$ such that the set P_I is not a polyhedron.*

PROPOSITION 5.1. *Let $P \subseteq \mathbb{R}^n$ be a rational polyhedron. Then also P_I is a rational polyhedron. In case $P_I \neq \emptyset$, its recession cone equals that of P.*

Proof. The claim is trivial if P is bounded (as P then contains only finitely many integer points and the result follows by virtue of the discussion in Section 3.6). By the Weyl-Minkowski Theorem 3.2, a rational polyhedron generally decomposes into

$$P = \text{conv } V + \text{cone } W$$

with finite sets of rational vectors $V \subseteq \mathbb{Q}^n$ and $W \subseteq \mathbb{Q}^n$. By scaling, if necessary, we may assume that $W \subseteq \mathbb{Z}^n$. Denote by $\mathbf{V}$ and $\mathbf{W}$ the matrices whose columns are the vectors in V and W respectively. Thus each $\mathbf{x} \in P$ can be written as

$$\mathbf{x} = \mathbf{V}\boldsymbol{\lambda} + \mathbf{W}\boldsymbol{\mu}\ , \quad \text{where } \boldsymbol{\lambda}, \boldsymbol{\mu} \geq 0 \text{ and } \mathbf{1}^T\boldsymbol{\lambda} = 1.$$

Let $\lfloor \boldsymbol{\mu} \rfloor$ be the *integral part* of $\boldsymbol{\mu} \neq \mathbf{0}$ (obtained by rounding down each component $\mu_i \geq 0$ to the next integer $\lfloor \mu_i \rfloor$). Splitting $\boldsymbol{\mu}$ into its integral part $\lfloor \boldsymbol{\mu} \rfloor$ and its non-integral part $\overline{\boldsymbol{\mu}} = \boldsymbol{\mu} - \lfloor \boldsymbol{\mu} \rfloor$ yields

$$\mathbf{x} = \mathbf{V}\boldsymbol{\lambda} + \mathbf{W}\overline{\boldsymbol{\mu}} + \mathbf{W}\lfloor \boldsymbol{\mu} \rfloor = \overline{\mathbf{x}} + \mathbf{W}\lfloor \boldsymbol{\mu} \rfloor$$

with $\lfloor \boldsymbol{\mu} \rfloor \geq \mathbf{0}$ integral and $\overline{\mathbf{x}} \in \overline{P}$, where

$$\overline{P} = \{\mathbf{V}\boldsymbol{\lambda} + \mathbf{W}\overline{\boldsymbol{\mu}} \mid \boldsymbol{\lambda} \geq \mathbf{0},\ \mathbf{1}^T\boldsymbol{\lambda} = 1,\ \mathbf{0} \leq \overline{\boldsymbol{\mu}} < \mathbf{1}\}\ .$$

Because $W \subseteq \mathbb{Z}^n$, $\mathbf{x}$ is integral if and only if $\overline{\mathbf{x}}$ is integral. Hence

$$P \cap \mathbb{Z}^n = \overline{P} \cap \mathbb{Z}^n + \{\mathbf{W}\mathbf{z} \mid \mathbf{z} \geq \mathbf{0} \text{ integral}\}.$$

Taking convex hulls on both sides, we find (*cf.* Ex. 5.5)

$$P_I = \text{conv } (\overline{P} \cap \mathbb{Z}^n) + \text{cone } W.$$

Since $\overline{P}$ is bounded, $\overline{P} \cap \mathbb{Z}^n$ is finite. So the claim follows as before.

◇

EX. 5.5. *Show:* conv $(V + W) =$ conv $V +$ conv W *for all* $V, W \subseteq \mathbb{R}^n$.

We next want to derive a system of inequalities describing P_I. There is no loss of generality when we assume P to be described by a system $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ with $\mathbf{A}$ and $\mathbf{b}$ integral. The idea now is to derive new inequalities that are valid for P_I (but not necessarily for P) and to add these to the system $\mathbf{A}\mathbf{x} \leq \mathbf{b}$. Such inequalities are called *cutting planes* as they "cut off" parts of P that are guaranteed to contain no integral points.

Consider an inequality $\mathbf{c}^T\mathbf{x} \leq \beta$ that is valid for P. If $\mathbf{c} \in \mathbb{Z}^n$ but $\beta \notin \mathbb{Z}$, then each integral $\mathbf{x} \in P \cap \mathbb{Z}^n$ obviously satisfies the stronger inequality $\mathbf{c}^T\mathbf{x} \leq \lfloor \beta \rfloor$. Recall from Corollary 2.6 that valid inequalities for P can be derived from the system $\mathbf{A}\mathbf{x} \leq \mathbf{b}$ by taking nonnegative linear combinations. We therefore consider inequalities of the form

$$(\mathbf{y}^T\mathbf{A})\mathbf{x} \leq \mathbf{y}^T\mathbf{b}, \quad \mathbf{y} \geq \mathbf{0}\ . \tag{5.11}$$

If $\mathbf{y}^T\mathbf{A} \in \mathbb{Z}^n$, then every $\mathbf{x} \in P \cap \mathbb{Z}^n$ (and hence every $\mathbf{x} \in P_I$) satisfies

$$(\mathbf{y}^T\mathbf{A})\mathbf{x} \leq \lfloor \mathbf{y}^T\mathbf{b} \rfloor. \tag{5.12}$$

We say that (5.12) arises from (5.11) by *rounding* (if $\mathbf{y}^T\mathbf{A} \in \mathbb{Z}^n$). In particular, we regain the original inequalities $\mathbf{Ax} \leq \mathbf{b}$ taking as $\mathbf{y}$ all unit vectors. We conclude

$$P_I \subseteq P' = \{\mathbf{x} \in \mathbb{R}^n \mid (\mathbf{y}^T\mathbf{A})\mathbf{x} \leq \lfloor \mathbf{y}^T\mathbf{b} \rfloor,\ \mathbf{y} \geq \mathbf{0},\ \mathbf{y}^T\mathbf{A} \in \mathbb{Z}^n\} \subseteq P.$$

Searching for inequalities of type (5.12) with $\mathbf{y}^T\mathbf{A} \in \mathbb{Z}^n$, we may restrict ourselves to $\mathbf{0} \leq \mathbf{y} \leq \mathbf{1}$. Indeed, each $\mathbf{y} \geq \mathbf{0}$ splits into its integral part $\mathbf{z} = \lfloor \mathbf{y} \rfloor \geq \mathbf{0}$ and non-integral part $\bar{\mathbf{y}} = \mathbf{y} - \mathbf{z}$. The inequality (5.12) is then implied by the two inequalities

$$\begin{array}{rcll} (\mathbf{z}^T\mathbf{A})\mathbf{x} & \leq & \mathbf{z}^T\mathbf{b} & (\in \mathbb{Z}) \\ (\bar{\mathbf{y}}^T\mathbf{A})\mathbf{x} & \leq & \lfloor \bar{\mathbf{y}}^T\mathbf{b} \rfloor. & \end{array} \tag{5.13}$$

(Recall that we assume $\mathbf{A}$ and $\mathbf{b}$ to be integral.) The first inequality in (5.13) is implied by $\mathbf{Ax} \leq \mathbf{b}$. To describe P', it thus suffices to augment the system $\mathbf{Ax} \leq \mathbf{b}$ by all inequalities of the type (5.12) with $\mathbf{0} \leq \mathbf{y} < \mathbf{1}$, which describes

$$P' = \{\mathbf{x} \in \mathbb{R}^n \mid (\mathbf{y}^T\mathbf{A})\mathbf{x} \leq \lfloor \mathbf{y}^T\mathbf{b} \rfloor,\ \mathbf{0} \leq \mathbf{y} \leq \mathbf{1},\ \mathbf{y}^T\mathbf{A} \in \mathbb{Z}^n\}. \tag{5.14}$$

by a finite number of inequalities (see Ex. 5.6) and thus exhibits P' as a polyhedron.

EX. 5.6. *Show: There are only finitely many vectors* $\mathbf{y}^T\mathbf{A} \in \mathbb{Z}^n$ *with* $\mathbf{0} \leq \mathbf{y} \leq \mathbf{1}$.

EX. 5.7. *Show:* $P \subseteq Q$ *implies* $P' \subseteq Q'$. *(In particular,* P' *depends only on* P *and not on the particular system* $\mathbf{Ax} \leq \mathbf{b}$ *describing* P.*)*

Iterating the above construction, we obtain the so-called *Gomory sequence*

$$P \supseteq P' \supseteq P'' \supseteq \ldots \supseteq P^{(k)} \supseteq \ldots \supseteq P_I. \tag{5.15}$$

Remarkably (*cf.* Gomory [34], and also Chvatal [9]), Gomory sequences are always finite:

THEOREM 5.1. *The Gomory sequence is finite in the sense that* $P^{(t)} = P_I$ *holds for some* $t \in \mathbb{N}$.

Before giving the proof, let us examine in geometric terms what it means to pass from P to P'. Consider an inequality

$$(\mathbf{y}^T\mathbf{A})\mathbf{x} \leq \mathbf{y}^T\mathbf{b}$$

with $\mathbf{y} \geq \mathbf{0}$ and $\mathbf{y}^T\mathbf{A} \in \mathbb{Z}^n$. Assume that the components of $\mathbf{y}^T\mathbf{A}$ have greatest common divisor $d = 1$ (otherwise replace $\mathbf{y}$ by $d^{-1}\mathbf{y}$). Then the equation

$$(\mathbf{y}^T\mathbf{A})\mathbf{x} = \lfloor \mathbf{y}^T\mathbf{b} \rfloor$$

admits an integral solution $\mathbf{x} \in \mathbb{Z}^n$ (*cf.* Ex. 5.8). Hence passing from P to P' amounts to shifting all supporting hyperplanes H of P "towards" P_I until they "touch" $\mathbb{Z}^n$ in some point $\mathbf{x}$ (not necessarily in P_I).

FIGURE 5.1. *Moving a cutting plane towards P_I*

EX. 5.8. *Show: An equation $\mathbf{c}^T\mathbf{x} = \beta$ with $\mathbf{c} \in \mathbb{Z}^n$, $\beta \in \mathbb{Z}$ admits an integer solution if and only if the greatest common divisor of the components of $\mathbf{c}$ divides β (Hint: Section 2.3).*

The crucial step in proving Theorem 5.1 is the observation that the Gomory sequence (5.15) induces Gomory sequences on all faces of P simultaneously. More precisely, assume $F \subseteq P$ is a proper face. From Section 3.6, we know that $F = P \cap H$ holds for some *rational* hyperplane

$$H = \{\mathbf{x} \in \mathbb{R}^n \mid (\overline{\mathbf{y}}^T\mathbf{A})\mathbf{x} = \overline{\mathbf{y}}^T\mathbf{b}\}$$

with $\overline{\mathbf{y}} \in \mathbb{Q}^m_+$ (and hence $\overline{\mathbf{y}}^T\mathbf{A} \in \mathbb{Q}^n$ and $\overline{\mathbf{y}}^T\mathbf{b} \in \mathbb{Q}$).

LEMMA 5.1. *$F = P \cap H$ implies $F' = P' \cap H$.*

Proof. From Ex. 5.7 we conclude $F' \subseteq P'$. Since, furthermore, $F' \subseteq F \subseteq H$ holds, we conclude $F' \subseteq P' \cap H$. To prove the converse inclusion, note that F is the solution set of

$$\begin{aligned} \mathbf{A}\mathbf{x} &\leq \mathbf{b} \\ \overline{\mathbf{y}}^T\mathbf{A}\mathbf{x} &= \overline{\mathbf{y}}^T\mathbf{b}. \end{aligned}$$

Scaling $\overline{\mathbf{y}}$ if necessary, we may assume that $\overline{\mathbf{y}}^T\mathbf{A}$ and $\overline{\mathbf{y}}^T\mathbf{b}$ are integral. By definition, F' is described by the inequalities

$$(\mathbf{w}^T\mathbf{A} + \alpha\overline{\mathbf{y}}^T\mathbf{A})\mathbf{x} \leq \lfloor \mathbf{w}^T\mathbf{b} + \alpha\overline{\mathbf{y}}^T\mathbf{b} \rfloor \tag{5.16}$$

with $\mathbf{w} \geq \mathbf{0}$, $\alpha \in \mathbb{R}$ (not sign-restricted) and $\mathbf{w}^T\mathbf{A} + \alpha\overline{\mathbf{y}}^T\mathbf{A} \in \mathbb{Z}^n$. We show that each inequality (5.16) is also valid for $P' \cap H$ (and hence $P' \cap H \subseteq F'$).

If $\alpha < 0$, observe that for $\mathbf{x} \in H$ (and hence for $\mathbf{x} \in P' \cap H$) the inequality (5.16) remains unchanged if we increase α by an integer $k \in \mathbb{N}$: Since $\mathbf{x}$ satisfies $\overline{\mathbf{y}}^T\mathbf{A}\mathbf{x} =$

$\overline{\mathbf{y}}^T\mathbf{b} \in \mathbb{Z}$, both the left and right hand side will increase by $k\overline{\mathbf{y}}^T\mathbf{b}$ if α is increased to $\alpha + k$. Hence we can assume $\alpha \geq 0$ without loss of generality. If $\alpha \geq 0$, however, (5.16) is easily recognized as an inequality of type (5.12). (Take $\mathbf{y} = \mathbf{w} + \alpha\overline{\mathbf{y}} \geq \mathbf{0}$.) So the inequality is valid for P' and hence for $P' \cap H$.

◇

We are now prepared for the

Proof of Theorem 5.1. In case $P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{x} = \mathbf{b}\}$ is an affine subspace, the claim follows from Corollary 2.2 (*cf.* Ex. 5.9). In general, P is presented in the form

$$\begin{array}{rcl} \mathbf{A}\mathbf{x} & = & \mathbf{b} \\ \mathbf{A}'\mathbf{x} & \leq & \mathbf{b}' \end{array} \tag{5.17}$$

with $n - d$ equalities $\mathbf{A}_{i\cdot}\mathbf{x} = b_i$ and $s \geq 0$ facet inducing (*i.e.*, irredundant) inequalities $\mathbf{A}'_{j\cdot}\mathbf{x} \leq b'_j$.

CASE 1: $P_I = \emptyset$. Let us argue by induction on $s \geq 0$. If $s = 0$, P is an affine subspace and the claim is true. If $s \geq 1$, we remove the last inequality $\mathbf{A}'_{s\cdot}\mathbf{x} \leq b'_s$ in (5.17) and let $Q \subseteq \mathbb{R}^n$ be the corresponding polyhedron. By induction, we then have $Q^{(t)} = Q_I$ for some $t \in \mathbb{N}$. Now $P_I = \emptyset$ implies

$$Q_I \cap \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}'_{s\cdot}\mathbf{x} \leq b'_s\} = \emptyset\ .$$

Since $P^{(t)} \subseteq Q^{(t)}$ and (trivially) $P^{(t)} \subseteq \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{A}'_{s\cdot}\mathbf{x} \leq b'_s\}$, we conclude that $P^{(t)} = \emptyset$ holds, too.

CASE 2: $P_I \neq \emptyset$. We proceed now by induction on $\dim P$. If $\dim P = 0$, $P = \{\mathbf{p}\}$ is an affine subspace and the claim is true. In general, since P_I is a polyhedron, we can represent it as

$$\begin{array}{rcl} \mathbf{A}\mathbf{x} & = & \mathbf{b} \\ \mathbf{C}\mathbf{x} & \leq & \mathbf{d} \end{array}$$

with $\mathbf{C}$ and $\mathbf{d}$ integral.

We show that each inequality $\mathbf{c}^T\mathbf{x} \leq \delta$ of the system $\mathbf{C}\mathbf{x} \leq \mathbf{d}$ will eventually become valid for some $P^{(t)}$, $t \in \mathbb{N}$ (which establishes the claim immediately). So fix an inequality $\mathbf{c}^T\mathbf{x} \leq \delta$. Since P and P_I (and hence all $P^{(t)}$) have identical recession cones by Proposition 5.1, the values

$$\delta^{(t)} = \max_{\mathbf{x} \in P^{(t)}} \mathbf{c}^T\mathbf{x}$$

are finite for each $t \in \mathbb{N}$. The sequence $\delta^{(t)}$ is decreasing. Indeed, from the definition of the Gomory sequence we conclude that $\delta^{(t+1)} \leq \lfloor \delta^{(t)} \rfloor$. Hence the sequence $\delta^{(t)}$ reaches its limit

$$\overline{\delta} := \lim_{t \to \infty} \delta^{(t)} \geq \delta$$

in finitely many steps. If $\overline{\delta} = \delta$, there is nothing left to prove. Suppose therefore $\overline{\delta} = \delta^{(t)} > \delta$ and consider the face

$$F := \{\mathbf{x} \in P^{(t)} \mid \mathbf{c}^T\mathbf{x} = \overline{\delta}\}.$$

Then F_I must be empty since every $\mathbf{x} \in P_I \supseteq F_I$ satisfies $\mathbf{c}^T\mathbf{x} \leq \delta < \overline{\delta}$. If $\mathbf{c}^T \in$ row $\mathbf{A}$, then $\mathbf{c}^T\mathbf{x}$ is constant on $P \supseteq P^{(t)} \supseteq P_I$, so $\overline{\delta} > \delta$ is impossible. Hence $\mathbf{c}^T \notin$ row $\mathbf{A}$, *i.e.*, dim $F <$ dim P. By induction, we conclude from Lemma 5.1

$$F^{(k)} = P^{(t+k)} \cap \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{c}^T\mathbf{x} = \overline{\delta}\} = \emptyset$$

for some finite k. Hence $\delta^{(t+k)} < \overline{\delta}$, a contradiction.

◇

Ex. 5.9. *Assume* $P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} = \mathbf{b}\}$. *Show that either* $P = P_I$ *or* $P' = P_I = \emptyset$. *(Hint: Corollary 2.2 and Proposition 5.1)*

Ex. 5.10 (Matching Polytopes)**.** Let $G = (V, E)$ be a graph with an even number of nodes. A *perfect matching* in G is a set of pairwise disjoint edges covering all nodes. Perfect matchings in G are in one-to-one correspondence with integral (and hence binary) vectors $\mathbf{x} \in \mathbb{R}^E$ satisfying the constraints

(1) $\mathbf{x}(\delta(i)) = 1 \qquad (i \in V)$
(2) $\mathbf{0} \leq \mathbf{x} \leq \mathbf{1}$.

Let $P \subseteq \mathbb{R}^E$ be the polytope described by these constraints. The associated polytope P_I is called the *matching polytope* of G. Thus P_I is the convex hull of (incidence vectors of) perfect matchings in G. (For example, if G consists of two disjoint triangles, we have $\mathbb{R}^E \simeq \mathbb{R}^6$, $P = \{\frac{1}{2} \cdot \mathbf{1}\}$ and $P_I = \emptyset$).
To construct the Gomory polytope P', consider some $S \subseteq V$. When we add the constraints (1) for $i \in S$, every edge $e = (i, j)$ with $i, j \in S$ occurs twice. So the resulting equation is

(1') $\mathbf{x}(\delta(S)) + 2\mathbf{x}(E(S)) = |S|$

(Recall that $E(S) \subseteq E$ is the set of edges induced by S.) On the other hand, (2) implies

(2') $\mathbf{x}(\delta(S)) \geq 0$.

From (1') and (2') we conclude that $\mathbf{x}(E(S)) \leq \frac{1}{2}|S|$ is valid for P. Hence for $S \subseteq V$

(3) $\mathbf{x}(E(S)) \leq \lfloor \frac{1}{2}|S| \rfloor$

is valid for P'. It can be shown (*cf.* [12]) that the inequalities (1)-(3) describe P_I. So $P' = P_I$ and the Gomory sequence has length 1.

Gomory's Cutting Plane Method. Theorem 5.1 tells us that – at least in principle – integer programs can be solved by repeated application of linear programming. Conceptually, Gomory's method works as follows. Start with the integer linear program

$$\max \mathbf{c}^T\mathbf{x} \quad s.t. \quad \mathbf{Ax} \leq \mathbf{b},\ \mathbf{x} \text{ integral} \tag{5.18}$$

and solve its LP-*relaxation*, which is obtained by dropping the integrality constraint:

$$\max \mathbf{c}^T\mathbf{x} \quad s.t. \quad \mathbf{Ax} \leq \mathbf{b}\ . \tag{5.19}$$

So $\mathbf{c}^T\mathbf{x}$ is maximized over $P = \{\mathbf{x} \in \mathbb{R}^n \mid \mathbf{Ax} \leq \mathbf{b}\}$. If the optimal solution is integral, the problem is solved. Otherwise, determine P' and maximize $\mathbf{c}^T\mathbf{x}$ over P' *etc.*

Unfortunately, this approach is hopeless inefficient. In practice, if the optimum $\mathbf{x}^*$ of (5.19) is non-integral, one tries to find *cutting planes* (*i.e.*, valid inequalities for P_I that "cut off" a part of P containing $\mathbf{x}^*$) right away in order to add these to the system $\mathbf{Ax} \leq \mathbf{b}$ and then solves the new system *etc.*. This procedure is generally known as the *cutting plane method* for integer linear programs.

Of particular interest in this context are cutting planes that are best possible in the sense that they cut as much as possible off P. Ideally, one would like to add inequalities that define facets of P_I. Numerous classes of such *facet defining* cutting planes for various types of problems have been published in the literature. In Section 5.3, we discuss some techniques for deriving such cutting planes.

5.3. Cutting Planes II

The cutting plane method has been successfully applied to many types of problems. The most extensively studied problem in this context is the traveling salesman problem (see, *e.g.*, [12] for a detailed exposition). Here, we will take the max clique problem from Section 5.1 as our guiding example, trying to indicate some general techniques for deriving cutting planes. Moreover, we take the opportunity to explain how even more general (seemingly *nonlinear*) integer programs can be formulated as ILPs.

The following *unconstrained quadratic boolean* (*i.e.*, binary) *problem* was studied in Padberg [64] with respect to a symmetric matrix $\mathbf{Q} = (q_{ij}) \in \mathbb{R}^{n \times n}$:

$$\max \sum_{i,j=1}^{n} q_{ij} x_i x_j \ , \quad x_i \in \{0, 1\} \ . \tag{5.20}$$

As $x_i \cdot x_i = x_i$ holds for a binary variable x_i, the essential nonlinear terms in the objective function are the terms $q_{ij} x_i x_j$ $(i \neq j)$. These may be *linearized* with the help of new variables $y_{ij} = x_i x_j$. Since $x_i x_j = x_j x_i$, it suffices to introduce just $n(n-1)/2$ new variables y_e, one for each edge $e = (i, j) \in E$ in the complete graph $K_n = (V, E)$ with $V = \{1, \ldots, n\}$.

The salient point is the fact that the non-linear equation $y_e = x_i x_j$ is equivalent with the three linear inequalities

$$y_e \leq x_i, \ y_e \leq x_j, \ \text{and } x_i + x_j - y_e \leq 1$$

if x_i, x_j and y_e are binary variables.

With $c_i = q_{ii}$ and $d_e = q_{ij} + q_{ji}$ for $e = (i, j) \in E$, problem (5.20) can thus be written as an integer linear program:

$$\begin{array}{rrcll} \max & \sum_{i=1}^{n} c_i x_i + \sum_{e \in E} d_e y_e & & \text{s.t.} & \\ & y_e - x_i & \leq & 0 & e \in E, i \in e \\ & x_i + x_j - y_e & \leq & 1 & e = (i, j) \in E \\ & 0 \leq x_i, y_e & \leq & 1 & \\ & x_i, y_e & & \text{integer.} & \end{array} \tag{5.21}$$

Note that (5.21) is precisely our ILP formulation (5.7) of the weighted max clique problem.

Let $P \subseteq \mathbb{R}^{V \cup E}$ be the polytope defined by the inequality constraints of (5.21). As we have seen in Section 5.1, P_I is then the convex hull of the (vertex-edge) incidence vectors $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{V \cup E}$ of cliques (subsets) $C \subseteq V$.

The polytope $P \subseteq \mathbb{R}^{V \cup E}$ is easily seen to have full dimension $n + \binom{n}{2}$ (because, *e.g.*, $\mathbf{x} = \frac{1}{2} \cdot \mathbf{1}$ and $\mathbf{y} = \frac{1}{3} \cdot \mathbf{1}$ yields an interior point $(\mathbf{x}, \mathbf{y})$ of P). Even P_I is full-dimensional (see Ex. 5.11).

Ex. 5.11. *Show: $\mathbb{R}^{V \cup E}$ is the affine hull of the incidence vectors of the cliques of sizes 0,1 and 2.*

What cutting planes can we construct for P_I? By "inspection", we find that for any three vertices $i, j, k \in V$ and corresponding edges $e, f, g \in E$, the following *triangle* inequality

$$x_i + x_j + x_k - y_e - y_f - y_g \leq 1 \tag{5.22}$$

holds for any clique incidence vector $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{V \cup E}$.

Ex. 5.12. *Show: (5.22) can also be derived from the inequalities describing P by rounding.*

This idea can be generalized. To this end, we extend our general shorthand notation and write for $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{V \cup E}$ and $S \subseteq V$:

$$\mathbf{x}(S) = \sum_{i \in S} x_i \quad \text{and} \quad \mathbf{y}(S) = \sum_{e \in E(S)} y_e \ .$$

For example, (5.22) now simply becomes: $\mathbf{x}(S) - \mathbf{y}(S) \leq 1$ for $|S| = 3$.

For every $S \subseteq V$ and integer $\alpha \in \mathbb{N}$, consider the following *clique inequality*

$$\alpha \mathbf{x}(S) - \mathbf{y}(S) \leq \alpha(\alpha + 1)/2. \tag{5.23}$$

Proposition 5.2. *Each clique inequality is valid for P_I.*

Proof. Let $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{V \cup E}$ be the incidence vector of some clique $C \subseteq V$. We must show that $(\mathbf{x}, \mathbf{y})$ satisfies (5.23) for each $S \subseteq V$ and $\alpha \in \mathbb{N}$. Let $s = |S \cap C|$. Then $\mathbf{x}(S) = s$ and $\mathbf{y}(S) = s(s-1)/2$. Hence

$$\begin{aligned} \alpha(\alpha+1)/2 - \alpha\mathbf{x}(S) + \mathbf{y}(S) &= [\alpha(\alpha+1) - 2\alpha s + s(s-1)]/2 \\ &= (\alpha - s)(\alpha - s + 1)/2, \end{aligned}$$

which is nonnegative since both α and s are integers.

◇

A further class of inequalities can be derived similarly. For any two disjoint subsets $S, T \subseteq V$, the associated *cut inequality* is

$$\mathbf{x}(S) + \mathbf{y}(S) + \mathbf{y}(T) - \mathbf{y}(S : T) \geq 0 \tag{5.24}$$

(Recall from Section 5.1 that $\mathbf{y}(S : T)$ denotes the sum of all $\mathbf{y}$-values on edges joining S and T).

PROPOSITION 5.3. *Each cut inequality is valid for P_I.*

Proof. Assume that $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{V \cup E}$ is the clique incidence vector of $C \subseteq V$. With $s = |C \cap S|$ and $t = |C \cap T|$, we then find

$$\begin{aligned} \mathbf{x}(S) + \mathbf{y}(S) + \mathbf{y}(T) - \mathbf{y}(S : T) &= s + s(s-1)/2 + t(t-1)/2 - st \\ &= (s-t)(s-t+1)/2 \geq 0. \end{aligned}$$

◇

Multiplying a valid inequality with a variable $x_i \geq 0$, we obtain a new (nonlinear!) inequality. We can *linearize* it by introducing new variables as explained at the beginning of this section. Alternatively, we may simply use linear (lower or upper) bounds for the nonlinear terms, thus weakening the resulting inequality. For example, multiplying a clique inequality (5.23) by $2x_i$, $i \in S$, yields

$$2\alpha \sum_{j \in S} x_i x_j - 2x_i \mathbf{y}(S) \leq \alpha(\alpha+1)x_i.$$

Because of $x_i \mathbf{y}(S) \leq \mathbf{y}(S)$, $x_i^2 = x_i$ and $x_i x_j = y_e$, $e = (i, j) \in E$, the following so-called *i-clique inequality*

$$2\alpha \mathbf{y}(i : S \setminus \{i\}) - 2\mathbf{y}(S) - \alpha(\alpha-1)x_i \leq 0 \tag{5.25}$$

must be valid for P_I. (This may also be verified directly.)

REMARK. Most of the above inequalities actually define facets of P_I. Consider, *e.g.*, for some α, $1 \leq \alpha \leq n-2$, the clique inequality

$$\alpha\mathbf{x}(S) - \mathbf{y}(S) \leq \alpha(\alpha+1)/2\ ,$$

which is satisfied with equality by all incidence vectors of cliques $C \subseteq V$ with $|C \cap S| = \alpha$ or $|C \cap S| = \alpha + 1$. Let $H \subseteq \mathbb{R}^{V \cup E}$ be the affine hull of all these incidence vectors.
To prove that the clique inequality is facet defining, one has to show

$$\dim H = \dim P_I - 1\ ,$$

i.e., H is a hyperplane in $\mathbb{R}^{V \cup E}$. This is not too hard to do. (In the special case $S = V$ and $\alpha = 1$, it follows readily from Ex. 5.11).

The cutting plane method suffers from a difficulty we have not mentioned so far. Suppose we try to solve an integer linear program, starting with its LP-relaxation and repeatedly adding cutting planes. In each step, we then face the problem of finding a suitable cutting plane that cuts off the current non-integral optimum. This problem is generally difficult. *E.g.*, for the max clique problem one can show that it is NP-hard to check whether a given $(\mathbf{x}^*, \mathbf{y}^*) \in \mathbb{R}^{V \cup E}$ satisfies all clique inequalities and, if not, find a violated one to cut off $(\mathbf{x}^*, \mathbf{y}^*)$.

Moreover, one usually has only a limited number of different classes (types) of cutting planes to work with. In the max clique problem, for example, we could end up with a solution $(\mathbf{x}^*, \mathbf{y}^*)$ that satisfies all clique, i-clique and cut inequalities and yet is non-integral. The original system and these three classes of cutting planes namely describe P_I by no means completely.

The situation in practice, however, is often not so bad. Quite efficient heuristics can be designed that frequently succeed to find cutting planes of a special type. Macambira and de Souza [57], for example, solve max clique instances of up to 50 nodes with the above clique and cut inequalities and some more sophisticated generalizations thereof.

Furthermore, even when a given problem is not solved completely by cutting planes, the computation was not futile: Typically, the (non-integral) optimum obtained after having added hundreds of cutting planes provides a rather tight estimate of the true integer optimum. Such estimates are extremely valuable in a branch and bound method for solving ILPs as discussed in Section 5.4 below. For example, the combination of cutting planes and a branch and bound procedure has solved instances of the TSP with several thousand nodes to optimality (*cf.* [12]).

5.4. Branch and Bound

Any linear maximization program (ILP) with binary variables $x_1, \ldots, x_n$ can in principle be solved by *complete enumeration*: Check all 2^n possible solutions for feasibility and compare their objective values. To do this in a systematic fashion, one constructs an associated *tree of subproblems* as follows. Fixing, say the first variable x_1, to either $x_1 = 0$ or $x_1 = 1$, we generate two subproblems $(\text{ILP} \mid x_1 = 0)$ and $(\text{ILP} \mid x_1 = 1)$. These two subproblems are said to be obtained from (ILP) by *branching* on x_1.

Clearly, an optimal solution of (ILP) can be inferred by solving the two subproblems. Repeating the above branching step, we can build a *binary tree* whose nodes correspond to subproblems obtained by fixing some variables to be 0 or 1. (The term *binary* refers here to the fact that each node in the tree has exactly two *lower neighbors*.) The resulting tree may look as indicated in Figure 9.2 below.

$$(ILP)$$

$$(ILP \mid x_1 = 0) \qquad (ILP \mid x_1 = 1)$$

$$(ILP \mid x_1 = 0,\ x_3 = 0) \qquad (ILP \mid x_1 = 0,\ x_3 = 1)$$

FIGURE 5.2.

Having constructed the complete tree, we could solve (ILP) *bottom up* and inspect the 2^n leaves of the tree, which correspond to "trivial" (all variables fixed) problems. In contrast to this solution by complete enumeration, *branch and bound* aims at building only a small part of the tree, leaving most of the "lower part" unexplored. This approach is suggested by the following two obvious facts:

- If we can solve a particular subproblem, say (ILP $\mid x_1 = 0, x_3 = 1$), directly (*e.g.*, by cutting planes), there is no need to inspect the subproblems in the branch below (ILP $\mid x_1 = 0, x_3 = 1$) in the tree.
- If we obtain an upper bound $U(x_1 = 0, x_3 = 1)$ for the sub-problem (ILP $\mid x_1 = 0, x_3 = 1$) that is *less* than the objective value of some known feasible solution of the original (ILP), then (ILP $\mid x_1 = 0, x_3 = 1$) offers no optimal solution.

Only if neither of these circumstances occurs we have to explore the subtree rooted at (ILP $\mid x_1 = 0, x_3 = 1$) for possible optimal solutions. We do this by branching at (ILP $\mid x_1 = 0, x_3 = 1$) and creating two new subproblems in the search tree. An efficient branch and bound procedure tries to avoid such branching steps as much as possible. To this end, one needs efficient algorithms that produce

(1) "good" feasible solutions of the original (ILP).
(2) tight upper bounds for the subproblems.

There is a trade-off between the quality of the feasible solutions and upper bounds on the one hand and the size of the search tree we have to build on the other. As a rule of thumb, "good" solutions should be almost optimal and bounds should differ from the true optimum by less than 10%.

Algorithms for computing good feasible solutions usually depend very much on the particular problem at hand. So there is little to say in general. Quite often, however, simple and fast *heuristic procedures* for almost optimal solutions can be found. Such algorithms, also called *heuristics* for short, are known for many problem types. They have no guarantee for success, but work well in practice.

REMARK [LOCAL SEARCH]. In the max clique problem the following simple *local search* often yields surprisingly good solutions: We start with some $C \subseteq V$ and check

whether the removal of some node $i \in C$ or the addition of some node $j \notin C$ yields an improvement. If so, we add (delete) the corresponding node and continue this way until no such improvement is possible (in which case we stop with the current *local optimum* $C \subseteq V$). This procedure may be repeated with different initial solutions $C \subseteq V$.

Computing good upper bounds is usually more difficult. Often, one just solves the corresponding LP-relaxations. If these are too weak, one can try to improve them by adding cutting planes as outlined in Section 5.3 . An alternative is to obtain upper bounds from Lagrangian relaxation (see Section 5.5 below).

Search and Branching Strategies. For the practical execution of a branch and bound algorithm, one needs to specify how one should proceed. Suppose, for example, that we are in a situation as indicated in Figure 9.2, *i.e.*, that we have branched from (ILP) on variable x_1 and from (ILP | $x_1 = 0$) on variable x_3. We then face the question which subproblem to consider next, either (ILP | $x_1 = 1$) or one of the subproblems of (ILP | $x_1 = 0$). There are two possible (extremal) strategies: We either always go to one of the "lowest" (most restricted) subproblems or to one of the "highest" (least restricted) subproblems. The latter strategy, choosing (ILP | $x_1 = 1$) in our case, is called *breadth first search* while the former strategy is referred to as *depth first search*, as it moves down the search tree as fast as possible.

A second question concerns the way of branching itself. If LP-relaxation or cutting planes are used for computing upper bounds, we obtain a fractional optimum $\mathbf{x}^*$ each time we try to solve a subproblem. A commonly used branching rule then branches on the *most fractional* x_i^*. In the case of $(0,1)$-variables, this rule branches on the variable x_i for which x_i^* is closest to $1/2$. In concrete applications, we have perhaps an idea about the "relevance" of the variables. We may then alternatively decide to branch on the most relevant variable x_i. Advanced software packages for integer programming allow the user to specify the branching process and support various upper bounding techniques.

REMARK. The branch and bound approach can easily be extended to general integer problems. Instead of fixing a variable x_i to either 0 or 1, we may restrict it to $x_i \leq \alpha_i$ or $x_i \geq \alpha_i + 1$ for suitable $\alpha_i \in \mathbb{Z}$. Indeed, the general idea is to *partition* a given subproblem into a number of (possibly more than just two) subproblems of similar type.

5.5. Lagrangian Relaxation

In Section 4.1, Lagrangian relaxation was introduced as a means for calculating upper bounds for optimization problems. Thereby, one "relaxes" (dualizes) some (in)equality constraints by adding them to the objective function using Lagrangian multipliers $\mathbf{y} \geq \mathbf{0}$ (in case of inequality constraints) to obtain an upper bound $L(\mathbf{y})$.

The crucial question is which constraints to dualize. The more constraints are dualized, the weaker the bound becomes. On the other hand, dualizing more constraints facilitates the computation of $L(\mathbf{y})$. There is a trade-off between the

quality of the bounds we obtain and the effort necessary for their computation. Generally, one would dualize only the "difficult" constraints, *i.e.*, those that are difficult to deal with directly (see Section 5.5.2 for an example).

Held and Karp [39] were the first to apply the idea of Lagrangian relaxation to integer linear programs. Assume that we are given an integer program as

$$\max \{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b},\ \mathbf{B}\mathbf{x} \leq \mathbf{d},\ \mathbf{x} \in \mathbb{Z}^n\} \tag{5.26}$$

for given *integral* matrices $\mathbf{A}, \mathbf{B}$ and vectors $\mathbf{b}, \mathbf{c}, \mathbf{d}$ and let z^*_{IP} be the optimum value of (5.26). Dualizing the constraints $\mathbf{A}\mathbf{x} - \mathbf{b} \leq \mathbf{0}$ with multipliers $\mathbf{u} \geq \mathbf{0}$ yields the upper bound

$$\begin{aligned} L(\mathbf{u}) &= \max \{\mathbf{c}^T\mathbf{x} - \mathbf{u}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \mid \mathbf{B}\mathbf{x} \leq \mathbf{d}, \mathbf{x} \in \mathbb{Z}^n\} \\ &= \mathbf{u}^T\mathbf{b} + \max \{(\mathbf{c}^T - \mathbf{u}^T\mathbf{A})\mathbf{x} \mid \mathbf{B}\mathbf{x} \leq \mathbf{d}, \mathbf{x} \in \mathbb{Z}^n\} \end{aligned} \tag{5.27}$$

and thus the Lagrangian dual problem

$$z^*_D = \min_{\mathbf{u} \geq \mathbf{0}} L(\mathbf{u})\ . \tag{5.28}$$

EX. 5.13. *Show that $L(\mathbf{u})$ is an upper bound on z^*_{IP} for every $\mathbf{u} \geq \mathbf{0}$.*

It is instructive to compare (5.28) with the linear programming relaxation

$$z^*_{LP} = \max \{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{B}\mathbf{x} \leq \mathbf{d}\}, \tag{5.29}$$

which we obtain by dropping the integrality constraint $\mathbf{x} \in \mathbb{Z}^n$. We find that Lagrangian relaxation approximates the true optimum z^*_{IP} at least as well:

THEOREM 5.2. $z^*_{IP} \leq z^*_D \leq z^*_{LP}$.

Proof. The first inequality is clear (*cf.* Ex. 5.13). The second one follows from the fact that the Lagrangian dual of a linear program equals the linear programming dual. Formally, we may derive the second inequality by applying linear programming duality twice:

$$\begin{aligned} z^*_D &= \min_{\mathbf{u} \geq \mathbf{0}} L(\mathbf{u}) \\ &= \min_{\mathbf{u} \geq \mathbf{0}} [\mathbf{u}^T\mathbf{b} + \max_{\mathbf{x}} \{(\mathbf{c}^T - \mathbf{u}^T\mathbf{A})\mathbf{x} \mid \mathbf{B}\mathbf{x} \leq \mathbf{d}, \mathbf{x} \in \mathbb{Z}^n\}] \\ &\leq \min_{\mathbf{u} \geq \mathbf{0}} [\mathbf{u}^T\mathbf{b} + \max_{\mathbf{x}} \{(\mathbf{c}^T - \mathbf{u}^T\mathbf{A})\mathbf{x} \mid \mathbf{B}\mathbf{x} \leq \mathbf{d}\}] \\ &= \min_{\mathbf{u} \geq \mathbf{0}} [\mathbf{u}^T\mathbf{b} + \min_{\mathbf{v}} \{\mathbf{d}^T\mathbf{v} \mid \mathbf{v}^T\mathbf{B} = \mathbf{c}^T - \mathbf{u}^T\mathbf{A}, \mathbf{v} \geq \mathbf{0}\}] \\ &= \min_{\mathbf{u},\mathbf{v}} \{\mathbf{u}^T\mathbf{b} + \mathbf{v}^T\mathbf{d} \mid \mathbf{u}^T\mathbf{A} + \mathbf{v}^T\mathbf{B} = \mathbf{c}^T, \mathbf{u} \geq \mathbf{0}, \mathbf{v} \geq \mathbf{0}\} \\ &= \max_{\mathbf{x}} \{\mathbf{c}^T\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b}, \mathbf{B}\mathbf{x} \leq \mathbf{d}\} = z^*_{LP}. \end{aligned}$$

◇

REMARK. As the proof of Theorem 5.2 shows, $z_D^* = z_{LP}^*$ holds if and only if the integrality constraint $\mathbf{x} \in \mathbb{Z}^n$ is redundant in the Lagrangian dual problem defining z_D^*. In this case, the Lagrangian dual is said to have the *integrality property* (*cf.* Geoffrion [29]).

It turns out that solving the Lagrangian dual problem amounts to minimizing a "piecewise linear" function of a certain type. We say that a function $f : \mathbb{R}^n \to \mathbb{R}$ is *piecewise linear convex* if f is obtained as the maximum of a finite number of affine functions $f_i : \mathbb{R}^n \to \mathbb{R}$ (*cf.* Figure 5.3 below). (General convex functions are discussed in Chapter 10).

FIGURE 5.3. $f(\mathbf{u}) = \max\{ f_i(\mathbf{u}) \mid 1 \le i \le k\}$

PROPOSITION 5.4. *Let U be the set of vectors $\mathbf{u} \ge \mathbf{0}$ such that*

$$L(\mathbf{u}) = \mathbf{u}^T\mathbf{b} + \max_{\mathbf{x}} \{(\mathbf{c}^T - \mathbf{u}^T\mathbf{A})\mathbf{x} \mid \mathbf{B}\mathbf{x} \le \mathbf{d}, \mathbf{x} \in \mathbb{Z}^n\} < \infty \ . \tag{5.30}$$

Then L is a piecewise linear convex function on U.

Proof. For fixed $\mathbf{u} \ge \mathbf{0}$, the maximum in (5.30) is obtained by maximizing a linear function $f(\mathbf{x}) = (\mathbf{c}^T - \mathbf{u}^T\mathbf{A})\mathbf{x}$ over

$$P_I = \text{conv}\,\{\mathbf{x} \mid \mathbf{B}\mathbf{x} \le \mathbf{d},\ \mathbf{x} \in \mathbb{Z}^n\} = \text{conv}\ V + \text{cone}\ E\ ,$$

say, with finite sets $V \subseteq \mathbb{Z}^n$ and $E \subseteq \mathbb{Z}^n$ (*cf.* Proposition 5.1). If $L(\mathbf{u}) < \infty$, the maximum in (5.30) is attained at some $\mathbf{v} \in V$ (Why?). Hence

$$L(\mathbf{u}) = \mathbf{u}^T\mathbf{b} + \max\,\{(\mathbf{c}^T - \mathbf{u}^T\mathbf{A})\mathbf{v} \mid \mathbf{v} \in V\},$$

exhibiting the restriction of L to U as the maximum of the finitely many affine functions

$$\ell_i(\mathbf{u}) = \mathbf{u}^T(\mathbf{b} - \mathbf{A}\mathbf{v}_i) + \mathbf{c}^T\mathbf{v}_i \qquad (\mathbf{v}_i \in V).$$

$\diamond$

5.5.1. Solving the Lagrangian Dual. After these structural investigations, let us address the problem of computing (at least approximately) the best possible upper bound $L(\mathbf{u}^*)$ and solving the Lagrangian dual

$$z_D^* = \min_{\mathbf{u} \geq \mathbf{0}} L(\mathbf{u}).$$

To this end, we assume that we can evaluate (*i.e.*, efficiently solve) for any given $\overline{\mathbf{u}} \geq \mathbf{0}$:

$$L(\overline{\mathbf{u}}) = \max \{\mathbf{c}^T\mathbf{x} - \overline{\mathbf{u}}^T(\mathbf{A}\mathbf{x} - \mathbf{b}) \mid \mathbf{B}\mathbf{x} \leq \mathbf{d}, \mathbf{x} \in \mathbb{Z}^n\} . \tag{5.31}$$

REMARK. In practice this means that the constraints we dualize ($\mathbf{A}\mathbf{x} \leq \mathbf{b}$) have to be chosen appropriately so that the resulting $L(\mathbf{u})$ is easy to evaluate (otherwise we obviously cannot expect to solve the problem $\min L(\mathbf{u})$)

Suppose $\overline{\mathbf{x}} \in \mathbb{Z}^n$ is an optimal solution of (5.31). We then seek some $\mathbf{u} \geq \mathbf{0}$ such that $L(\mathbf{u}) < L(\overline{\mathbf{u}})$. Since $\overline{\mathbf{x}}$ is a feasible solution of the maximization problem in (5.31), $L(\mathbf{u}) < L(\overline{\mathbf{u}})$ implies

$$\mathbf{c}^T\overline{\mathbf{x}} - \mathbf{u}^T(\mathbf{A}\overline{\mathbf{x}} - \mathbf{b}) \leq L(\mathbf{u}) < L(\overline{\mathbf{u}}) = \mathbf{c}^T\overline{\mathbf{x}} - \overline{\mathbf{u}}^T(\mathbf{A}\overline{\mathbf{x}} - \mathbf{b}) \tag{5.32}$$

and hence

$$(\mathbf{u} - \overline{\mathbf{u}})^T(\mathbf{A}\overline{\mathbf{x}} - \mathbf{b}) > 0.$$

The Subgradient Method. The preceding argument suggests to try a vector $\mathbf{u} = \overline{\mathbf{u}} + \boldsymbol{\Delta}\mathbf{u}$ with

$$\boldsymbol{\Delta}\mathbf{u} = \mathbf{u} - \overline{\mathbf{u}} = \lambda(\mathbf{A}\overline{\mathbf{x}} - \mathbf{b})$$

for some small *step size* $\lambda > 0$.

Of course, we also want to have $\mathbf{u} = \overline{\mathbf{u}} + \boldsymbol{\Delta}\mathbf{u} \geq \mathbf{0}$. So we simply replace any negative component by 0, *i.e.*, we project the resulting vector $\mathbf{u}$ onto the set $\mathbb{R}^m_+$ of feasible multipliers and obtain

$$\mathbf{u} = \max\{\mathbf{0}\ ,\ \overline{\mathbf{u}} + \lambda(\mathbf{A}\overline{\mathbf{x}} - \mathbf{b})\} \qquad \text{(componentwise)} . \tag{5.33}$$

REMARK. This procedure appears intuitively reasonable: As our step size λ is small, a negative component can only occur if $\overline{u}_i \approx 0$ and $\mathbf{A}_{i\cdot}\overline{\mathbf{x}} < b_i$. This means that we do not need to enforce the constraint $\mathbf{A}_{i\cdot}\mathbf{x} \leq b_i$ by assigning a large penalty (Lagrangian multiplier) to it. Consequently, we try $u_i = 0$.

The above procedure is the *subgradient method* (*cf.* also Section 4.2.3) for solving the Lagrangian dual: We start with some $\mathbf{u}_0 \geq \mathbf{0}$ and compute a sequence $\mathbf{u}_1, \mathbf{u}_2, \ldots$ by iterating the above step with step sizes $\lambda_1, \lambda_2, \ldots$.

The appropriate choice of the step size λ_i is a delicate problem – both in theory and in practice. A basic result states that convergence takes place (in the sense of Theorem **??**) provided

$$\lim_{i \to \infty} \lambda_i = 0 \quad \text{and} \quad \sum_{i=0}^{\infty} \lambda_i = \infty.$$

5.5.2. Max Clique Revisited. How could Lagrangian relaxation be applied to the max clique problem? The first (and most crucial) step is to establish an appropriate ILP formulation of the max clique problem. This formulation should be such that dualizing a suitable subset of constraints yields upper bounds that are reasonably tight and efficiently computable. A bit of experimenting reveals our original formulation (5.7) resp. (5.21) to be inadequate. Below, we shall derive an alternative formulation that turns out to work better.

We start by passing from the underlying complete graph $K_n = (V, E)$ to the complete directed graph $D_n = (V, A)$, replacing each edge $e = (i, j) \in E$ by two oppositely directed arcs $(i, j) \in A$ and $(j, i) \in A$. To avoid confusion with the notation, we will always indicate whether a pair (i, j) is considered as an ordered or unordered pair and write $(i, j) \in A$ or $(i, j) \in E$, resp. With each arc $(i, j) \in A$, we associate a binary variable y_{ij}. The original edge weights d_e $(e \in E)$ are equally replaced by arc weights $q_{ij} = q_{ji} = d_e/2$ $(e = (i, j) \in E)$.

The original ILP formulation (5.7) can now be equivalently replaced by

$$
\begin{array}{lcccl}
 & \max\ \mathbf{c}^T\mathbf{x} + \mathbf{q}^T\mathbf{y} & s.t. & & \\
(1) & x_i + x_j - \frac{1}{2}(y_{ij} + y_{ji}) & \leq & 1 & (i, j) \in E \\
(2) & y_{ij} - y_{ji} & = & 0 & (i, j) \in E \\
(3) & y_{ij} - x_i & \leq & 0 & (i, j) \in A \\
(4) & \mathbf{x} \in \{0, 1\}^V, \mathbf{y} \in \{0, 1\}^A & & &
\end{array}
\tag{5.34}
$$

REMARK. (5.34) is a "directed version" of (5.7). The cliques (subsets) $C \subseteq V$ are now in one-to-one correspondence with the feasible solutions of (5.34), namely the *vertex-arc incidence vectors* $(\mathbf{x}, \mathbf{y}) \in \{0, 1\}^{V \cup A}$, defined by $x_i = 1$ if $i \in C$ and $y_{ij} = 1$ if $i, j \in C$.

The directed version (5.34) offers the following advantage over the formulation (5.7): After dualizing constraints (1) and (2) in (5.34), the remaining constraints (3) and (4) imply no "dependence" between different nodes i and j (*i.e.*, $y_{ij} = 1$ implies $x_i = 1$ but not $x_j = 1$) . The resulting Lagrangian relaxation can therefore be solved quite easily (*cf.* Ex. 5.14).

Ex. 5.14. Using Lagrangian multipliers $\mathbf{u} \in \mathbb{R}^E_+$ for dualizing constraints (1) and unrestricted multipliers $\mathbf{v} \in \mathbb{R}^E$ for dualizing the equality constraints (2) in (5.34), one obtains

$$
L(\mathbf{u}, \mathbf{v}) = \max\ \mathbf{c}^T\mathbf{x} + \mathbf{q}^T\mathbf{y} + \sum_{(i,j)\in E} u_{ij}\big(1 - x_i - x_j + \frac{1}{2}(y_{ij} + y_{ji})\big) + \sum_{(i,j)\in E} v_{ij}(y_{ij} - y_{ji})
$$

subject to (3)–(4) from (5.34) .

So for given $\mathbf{u} \in \mathbb{R}^E_+$ and $\mathbf{v} \in \mathbb{R}^E$, computing $L(\mathbf{u}, \mathbf{v})$ amounts to solving a problem of the following type (with suitable $\widetilde{\mathbf{c}} \in \mathbb{R}^V$ and $\widetilde{\mathbf{q}} \in \mathbb{R}^A$):

$$
\max\ \widetilde{\mathbf{c}}^T\mathbf{x} + \widetilde{\mathbf{q}}^T\mathbf{y} \quad \textit{subject to (3)–(4) from (5.34)}
$$

Show: A problem of the latter type is easy to solve because the constraints (3)–(4) imply no "dependence" between different nodes i and j.
(Hint: For $i \in V$, let $P_i = \{j \in V \mid \widetilde{q}_{ij} > 0\}$. Set $x_i = 1$ if $\widetilde{c}_i + \sum_{j\in P_i} \widetilde{q}_{ij} > 0$.)

Unfortunately, the Lagrangian bounds we obtain from the dualization of the constraints (1) and (2) in (5.34) are too weak to be useful in practice. To derive tighter bounds, we want to add more constraints to (5.34) while keeping the enlarged system still efficiently solvable after dualizing constraints (1) and (2). It turns out that one can add "directed versions" (*cf.* below) of the clique inequalities (5.23) and the i-clique inequalities (5.25) for $S = V$ without complicating things too much. The resulting formulation of the max clique problem is

$$
\begin{array}{llccll}
 & & \max \mathbf{c}^T\mathbf{x} + \mathbf{q}^T\mathbf{y} & \text{s.t.} & & \\
 & (1) & x_i + x_j - \frac{1}{2}(y_{ij} + y_{ji}) & \leq & 1 & (i,j) \in E \\
 & (2) & y_{ij} - y_{ji} & = & 0 & (i,j) \in E \\
(5.35) & (3) & y_{ij} - x_i & \leq & 0 & (i,j) \in A \\
 & (4) & 2\alpha\mathbf{x}(V) - \mathbf{y}(V) & \leq & \alpha(\alpha+1) & \alpha = 1, \ldots, n \\
 & (5) & 2\alpha\mathbf{y}(\delta^+(i)) - \mathbf{y}(V) - \alpha(\alpha-1)x_i & \leq & 0 & i \in V \\
 & (6) & \mathbf{x} \in \{0,1\}^V, \mathbf{y} \in \{0,1\}^A & & &
\end{array}
$$

where, in constraints (4) and (5), we used the straightforward extension of our general shorthand notation:

$$\mathbf{y}(V) = \sum_{(i,j)\in A} y_{ij} \qquad \text{and} \qquad \mathbf{y}(\delta^+(i)) = \sum_{j \neq i} y_{ij}.$$

Constraints (4) and (5) are "directed versions" of the original clique and i-clique inequalities (5.23) and (5.25).

Ex. 5.15. *Show that every incidence vector $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^{V \cup A}$ of a set (clique) $C \subseteq V$ satisfies the constraints in (5.35). (Hint: Section 5.3)*

To dualize constraints (1) and (2) in (5.35), we introduce Lagrangian multipliers $\mathbf{u} \in \mathbb{R}^E_+$ for the inequality constraints (1) and unrestricted multipliers $\mathbf{v} \in \mathbb{R}^E$ for the equality constraints (2). So we obtain for $L(\mathbf{u}, \mathbf{v})$ the expression

$$\max\ \mathbf{c}^T\mathbf{x} + \mathbf{q}^T\mathbf{y} + \sum_{(i,j)\in E} u_{ij}\big(1 - x_i - x_j + \frac{1}{2}(y_{ij} + y_{ji})\big) + \sum_{(i,j)\in E} v_{ij}(y_{ij} - y_{ji})$$
$$\text{subject to (3)–(6) from (5.35)}\ .$$

Given $\mathbf{u} \in \mathbb{R}^E_+$ and $\mathbf{v} \in \mathbb{R}^E$, the computation of $L(\mathbf{u}, \mathbf{v})$ amounts to solving a problem of the following type (for suitable $\widetilde{\mathbf{c}} \in \mathbb{R}^V$ and $\widetilde{\mathbf{q}} \in \mathbb{R}^A$):

$$\max\ \widetilde{\mathbf{c}}^T\mathbf{x} + \widetilde{\mathbf{q}}^T\mathbf{y} \quad \text{subject to (3)–(6) from (5.35)} \tag{5.36}$$

The integer linear program (5.36) appears to be more difficult, but can still be solved quickly.

For $p = 0, \ldots, n$, we determine the best solution satisfying $\mathbf{x}(V) = p$ as follows: For $p = 0$, set $\mathbf{x} = \mathbf{y} = \mathbf{0}$. Given $p \geq 1$, we choose for each $i \in V$ the $p-1$ most profitable arcs in $\delta^+(i)$, *i.e.*, those with the highest $\widetilde{q}$-values. Suppose their $\widetilde{q}$-values sum up to $\widetilde{q}_i$ for $i \in V$. We then let $x_i = 1$ for the p largest values of $\widetilde{c}_i + \widetilde{q}_i$. If $x_i = 1$, we let $y_{ij} = 1$ for the $p-1$ most profitable arcs in $\delta^+(i)$.

The optimal solution is then the best we found for $p = 0, \ldots, n$. This follows from

LEMMA 5.2. *Let* $(\mathbf{x}, \mathbf{y}) \in \{0,1\}^{V \cup A}$. *Then* $(\mathbf{x}, \mathbf{y})$ *is a feasible solution of (5.36) if and only if there exists some* $p \in \{0, \ldots, n\}$ *such that*

$$\text{(i)} \quad \mathbf{x}(V) = p \quad \text{and} \quad \text{(ii)} \quad \mathbf{y}(\delta^+(i)) = \begin{cases} p-1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 0 \end{cases} \qquad (i \in V)$$

Proof. Assume first that $(\mathbf{x}, \mathbf{y})$ satisfies (i) and (ii). Then $(\mathbf{x}, \mathbf{y})$ satisfies the constraints (3) and (6) of (5.35). Constraint (4) reduces to

$$\text{(4')} \qquad 2\alpha p - p(p-1) \leq \alpha(\alpha+1),$$

which holds for all $\alpha, p \in \mathbb{Z}$ since $(\alpha - p)^2 + (\alpha - p) \geq 0$. Constraint (5) is certainly satisfied if $x_i = 0$ (due to (ii)). For $x_i = 1$, constraint (5) becomes

$$2\alpha(p-1) - p(p-1) \leq \alpha(\alpha-1),$$

which is (4') again.

Conversely, assume that $(\mathbf{x}, \mathbf{y})$ is feasible for (5.36) and let $p = \mathbf{x}(V) = \sum_{i \in V} x_i$. Consider the constraints (5) of (5.36) for those i with $x_i = 1$. Adding the corresponding inequalities for any α, we find

$$2\alpha \mathbf{y}(V) - p\mathbf{y}(V) - p\alpha(\alpha-1) \leq 0.$$

Taking $\alpha = p$, we conclude $\mathbf{y}(V) \leq p(p-1)$.

On the other hand, letting $\alpha = p$ in (4), we have

$$2p^2 - \mathbf{y}(V) \leq p(p+1) \quad \text{and hence} \quad \mathbf{y}(V) \geq p(p-1)\,,$$

which proves $\mathbf{y}(V) = p(p-1)$. Substituting the latter equality into (5) (with $\alpha = p$) and dividing by p, we deduce for $i \in V$ with $x_i = 1$:

$$2\mathbf{y}(\delta^+(i)) \leq (p-1) + (p-1)x_i = 2(p-1).$$

In view of constraint (3) in (5.35), we thus have the inequalities

$$\mathbf{y}(\delta^+(i)) \leq \begin{cases} p-1 & \text{if } x_i = 1 \\ 0 & \text{if } x_i = 0. \end{cases}$$

Since $\mathbf{y}(V) = p(p-1)$, actually equality must hold.

◇

EX. 5.16. The Lagrangian bounds $L(\mathbf{u}, \mathbf{v})$ we obtain when solving (5.36) as explained above are generally better than the bound produced by the LP-relaxation of (5.36). Consider, for example, the complete directed graph $D_4 = (V, A)$ with $\widetilde{\mathbf{c}} = \mathbf{0} \in \mathbb{R}^V$ and symmetric arc weights $\widetilde{q}_{ij} = \widetilde{q}_{ji}$ as indicated in Figure 5.4 below.
An optimum integral solution of (5.36) can be obtained as follows: Choose any set $C \subseteq V$ with $|C| = 3$. Set $x_i = 1$ if $i \in C$. Furthermore, for each $i \in C$ choose two arcs in $\delta^+(i)$ with weight $\widetilde{q}_{ij} = 1$. Set $y_{ij} = 1$ on these two arcs. This solution guarantees an objective function value $\widetilde{\mathbf{q}}^T\mathbf{y} = 6$ (so the duality gap is zero).

In contrast, the LP-relaxation of (5.36) is solved by $x_1 = x_4 = 1$, $x_2 = x_3 = 2/3$, $y_{12} = y_{13} = y_{42} = y_{43} = 1$ and $y_{21} = y_{23} = y_{24} = y_{31} = y_{32} = y_{34} = 2/3$ with an objective value of 8. So Lagrangian relaxation (in this example) provides strictly better bounds than LP-relaxation. In other words, problem formulation (5.36) does not have the integrality property (*cf.* p. 105).

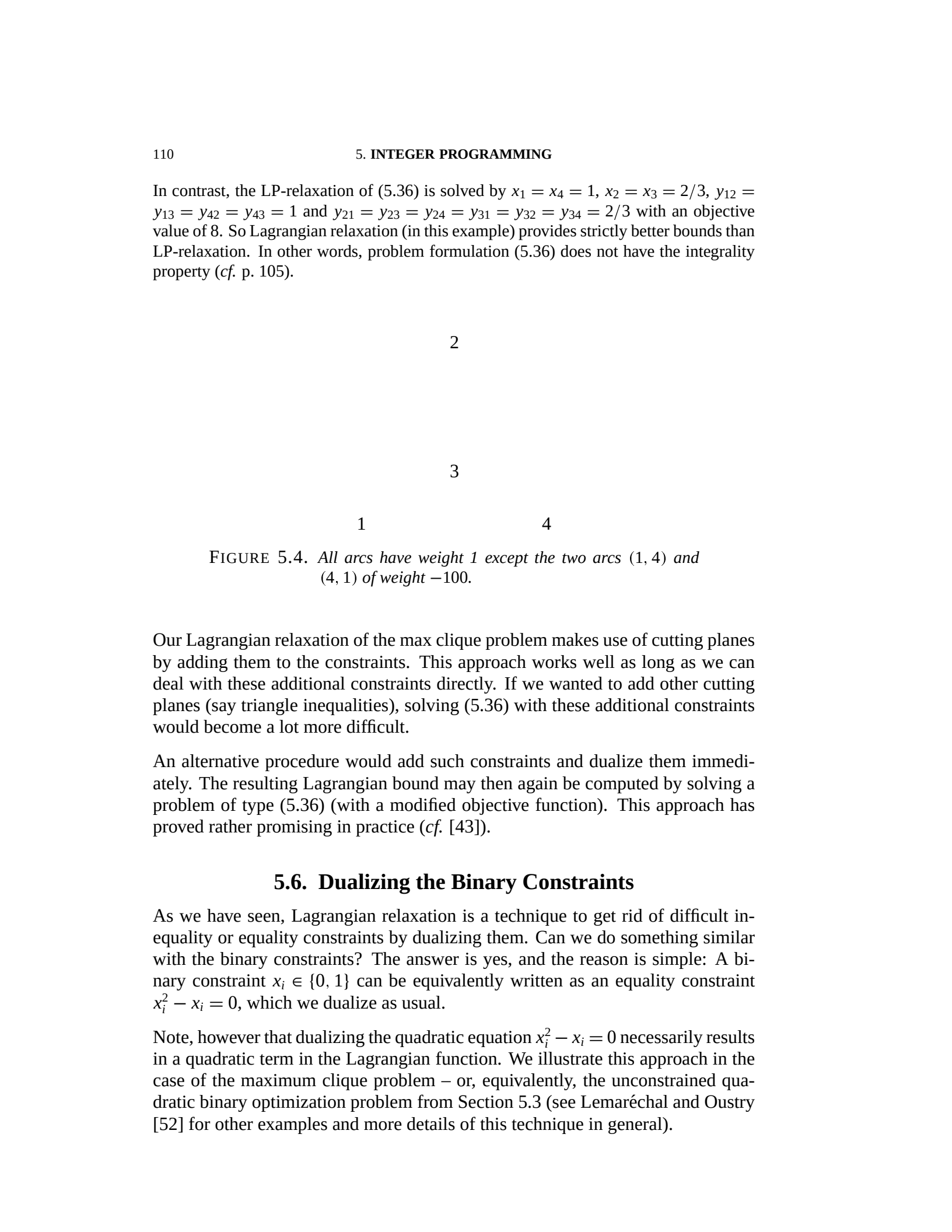

FIGURE 5.4. *All arcs have weight 1 except the two arcs* $(1, 4)$ *and* $(4, 1)$ *of weight* -100.

Our Lagrangian relaxation of the max clique problem makes use of cutting planes by adding them to the constraints. This approach works well as long as we can deal with these additional constraints directly. If we wanted to add other cutting planes (say triangle inequalities), solving (5.36) with these additional constraints would become a lot more difficult.

An alternative procedure would add such constraints and dualize them immediately. The resulting Lagrangian bound may then again be computed by solving a problem of type (5.36) (with a modified objective function). This approach has proved rather promising in practice (*cf.* [43]).

5.6. Dualizing the Binary Constraints

As we have seen, Lagrangian relaxation is a technique to get rid of difficult inequality or equality constraints by dualizing them. Can we do something similar with the binary constraints? The answer is yes, and the reason is simple: A binary constraint $x_i \in \{0, 1\}$ can be equivalently written as an equality constraint $x_i^2 - x_i = 0$, which we dualize as usual.

Note, however that dualizing the quadratic equation $x_i^2 - x_i = 0$ necessarily results in a quadratic term in the Lagrangian function. We illustrate this approach in the case of the maximum clique problem – or, equivalently, the unconstrained quadratic binary optimization problem from Section 5.3 (see Lemaréchal and Oustry [52] for other examples and more details of this technique in general).

Let $\mathbf{Q} \in \mathbb{R}^{n\times n}$ be a symmetric matrix and reconsider the unconstrained quadratic boolean problem

$$\max \{\mathbf{x}^T\mathbf{Q}\mathbf{x} \mid \mathbf{x} \in \{0,1\}^n\} . \tag{5.37}$$

Dualizing the constraints $x_i^2 - x_i = 0$ with Lagrangian multipliers $u_i \in \mathbb{R}$, we obtain the Lagrangian bound

$$L(\mathbf{u}) = \max_{\mathbf{x}\in\mathbb{R}^n} \mathbf{x}^T\mathbf{Q}\mathbf{x} + \sum_i u_i(x_i^2 - x_i). \tag{5.38}$$

Letting $\mathbf{U} \in \mathbb{R}^{n\times n}$ denote the diagonal matrix with diagonal $\mathbf{u} \in \mathbb{R}^n$, we can write

$$L(\mathbf{u}) = \max_{\mathbf{x}} \mathbf{x}^T(\mathbf{Q}+\mathbf{U})\mathbf{x} - \mathbf{u}^T\mathbf{x}. \tag{5.39}$$

Evaluating $L(\mathbf{u})$ amounts to solving the unconstrained quadratic optimization problem (5.39). Ex. 5.17 shows how to accomplish this.

Ex. 5.17. *For fixed $\mathbf{u} \in \mathbb{R}^n$, consider the function*

$$f(\mathbf{x}) = \mathbf{x}^T(\mathbf{Q}+\mathbf{U})\mathbf{x} - \mathbf{u}^T\mathbf{x}.$$

Show: If $\bar{\mathbf{x}}(\mathbf{Q}+\mathbf{U})\bar{\mathbf{x}} > 0$ holds for some $\bar{\mathbf{x}} \in \mathbb{R}^n$, then f has no finite maximum. Assume that $\mathbf{x}^T(\mathbf{Q}+\mathbf{U})\mathbf{x} \le 0$ always holds (i.e., $\mathbf{Q}+\mathbf{U}$ is negative semidefinite*). Show: $\bar{\mathbf{x}}$ is optimal for f if and only if $\nabla f(\bar{\mathbf{x}}) = 2\bar{\mathbf{x}}^T(\mathbf{Q}+\mathbf{U}) - \mathbf{u}^T = \mathbf{0}^T$. (Hint: Section* **??***). So f has a finite maximum if and only if $\mathbf{Q}+\mathbf{U}$ is negative semidefinite and $\nabla f(\mathbf{x}) = \mathbf{0}^T$ has a solution. The maximum is attained in each $\bar{\mathbf{x}} \in \mathbb{R}^n$ satisfying $2(\mathbf{Q}+\mathbf{U})\bar{\mathbf{x}} = \mathbf{u}$, which implies*

$$L(\mathbf{u}) = \max_{\mathbf{x}} f(\mathbf{x}) = \frac{1}{2}\bar{\mathbf{x}}^T\mathbf{u} - \mathbf{u}^T\bar{\mathbf{x}} = -\frac{1}{2}\mathbf{u}^T\bar{\mathbf{x}} .$$

The Lagrangian dual $\min_{\mathbf{u}} L(\mathbf{u})$ is called the *semidefinite relaxation* of the primal (5.37), as it can be reformulated as follows (with $\mathbf{u} \in \mathbb{R}^n$, $r \in \mathbb{R}$):

$$\begin{aligned}
\min_{\mathbf{u}} L(\mathbf{u}) &= \min_{\mathbf{u},r} \{r \mid L(\mathbf{u}) \le r\} \\
&= \min_{\mathbf{u},r} \{r \mid \mathbf{x}^T(\mathbf{Q}+\mathbf{U})\mathbf{x} - \mathbf{u}^T\mathbf{x} \le r \quad \forall \mathbf{x} \in \mathbb{R}^n\} \\
&= \min_{\mathbf{u},r} \{r \mid (1,\mathbf{x}^T)\begin{bmatrix} -r & -\frac{1}{2}\mathbf{u}^T \\ -\frac{1}{2}\mathbf{u} & (\mathbf{Q}+\mathbf{U}) \end{bmatrix}\begin{pmatrix}1\\ \mathbf{x}\end{pmatrix} \le 0 \quad \forall\, \mathbf{x} \in \mathbb{R}^n\} \\
&= \min_{\mathbf{u},r} \{\, r \mid \begin{bmatrix} -r & -\frac{1}{2}\mathbf{u}^T \\ -\frac{1}{2}\mathbf{u} & (\mathbf{Q}+\mathbf{U}) \end{bmatrix} \text{is negative semidefinite}\}.
\end{aligned}$$

Only the last step needs further explanation, which is given in Ex. 5.18 below.

Ex. 5.18. *Show for any $\mathbf{S} \in \mathbb{R}^{(n+1)\times(n+1)}$:*

$$(1,\mathbf{x}^T)\mathbf{S}\begin{pmatrix}1\\ \mathbf{x}\end{pmatrix} \le 0 \text{ for all } \mathbf{x} \in \mathbb{R}^n \quad \Longleftrightarrow \quad \mathbf{z}^T\mathbf{S}\mathbf{z} \le 0 \text{ for all } \mathbf{z} \in \mathbb{R}^{n+1}.$$

Our reformulation of the Lagrangian dual *via*

$$\min_{\mathbf{u}} L(\mathbf{u}) \quad = \quad \min_{r,\mathbf{u}} r \quad s.t. \quad \mathbf{S}_{r,\mathbf{u}} = \begin{bmatrix} -r & -\frac{1}{2}\mathbf{u}^T \\ -\frac{1}{2}\mathbf{u} & (\mathbf{Q}+\mathbf{U}) \end{bmatrix} \preceq \mathbf{0}. \tag{5.40}$$

is a special case of a *semidefinite program* (optimizing a linear objective under linear and semidefinite constraints, see also Section**??**).

REMARK. To understand how (and why) problem (5.40) can be solved at least approximately, consider the following "cutting plane approach": We first replace the condition of semidefiniteness for $\mathbf{S} = \mathbf{S}_{r,\mathbf{u}}$ by a *finite* number of linear inequalities

$$\mathbf{a}^T\mathbf{S}\mathbf{a} \leq 0 \qquad (\mathbf{a} \in A) \tag{5.41}$$

for some finite set $A \subseteq \mathbb{R}^{n+1}$. Note that, for each fixed $\mathbf{a} \in A$, the inequality $\mathbf{a}^T\mathbf{S}\mathbf{a} \leq 0$ is a *linear* inequality with variables r and $\mathbf{u}$.

We then minimize r subject to constraints (5.41). If the solution provides us with r and $\mathbf{u}$ such that $\mathbf{S}_{r,\mathbf{u}}$ is negative semidefinite, we have found a solution. Otherwise, if $\bar{\mathbf{a}}^T\mathbf{S}\bar{\mathbf{a}} > 0$ holds for some $\bar{\mathbf{a}} \in \mathbb{R}^{n+1}$, we add $\bar{\mathbf{a}}$ to A (*i.e.*, we add a *violated inequality*) and solve the modified problem *etc.* (Note that we can check whether $\mathbf{S} = \mathbf{S}_{r,\mathbf{u}}$ is negative semidefinite with the Diagonalization algorithm from Section 2.1. This also provides us with a suitable vector $\bar{\mathbf{a}}$ in case $\mathbf{S}$ is not negative semidefinite.)

The theoretical aspects of this approach will be discussed in the context of the ellipsoid method in Section **??**. In practice, analogues of the interior point method for linear programs (*cf.* Chapter **??**) solve semidefinite programs more efficiently.

We want to emphasize that the approach of dualizing the binary constraints in a general integer program

$$\max\ \mathbf{c}^T\mathbf{x} \qquad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \in \{0,1\}^n$$

is limited. If we dualize only the binary constraints $x_i^2 - x_i = 0$ using Lagrangian multipliers $u_i \in \mathbb{R}$, the Lagrangian function becomes

$$L(\mathbf{u}) = \max\ \mathbf{x}^T\mathbf{U}\mathbf{x} + (\mathbf{c}-\mathbf{u})^T\mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}\ .$$

In contrast to (5.38), this is a quadratic optimization problem with *inequality constraints*, which is in general difficult (NP-hard, *cf.* Section **??**).

List of frequently used Symbols

$\mathbb{R}$	set of real numbers
$\mathbb{N}, \mathbb{Z}, \mathbb{Q}$	set of natural, integer, rational numbers
$\mathbb{R}^n$	Euclidean n-space
$\mathbb{R}^n_+$	set of non-negative vectors in $\mathbb{R}^n$
$\mathbb{R}^E$	set of real vectors indexed by the set E
$\mathbb{R}^{m\times n}$	set of real $m \times n$ matrices
$\mathbb{S}^{n\times n}$	set of real symmetric $n \times n$ matrices
$\mathbf{x} \in \mathbb{R}^n$	column vector with components $x_1, \ldots, x_n$
$\|\mathbf{x}\|$	Euclidean norm
$U_\varepsilon(\mathbf{x})$	ε-neighborhood of $\mathbf{x}$
$\mathbf{e}_1, \ldots, \mathbf{e}_n$	standard unit vectors in $\mathbb{R}^n$
$\langle \mathbf{x} \vert \mathbf{y} \rangle$	inner product
$\mathbf{x}^T\mathbf{y}$	standard inner product in $\mathbb{R}^n$
$\mathbf{A} = (a_{ij}) \in \mathbb{R}^{m\times n}$	real $(m \times n)$ matrix
$\mathbf{A}^T$	transpose of $\mathbf{A}$
$\mathbf{A}_{i\cdot}, \ \mathbf{A}_{\cdot j}$	row vectors, column vectors of $\mathbf{A}$
$\mathbf{A} \circ \mathbf{B} = \sum_i \sum_j a_{ij}b_{ij}$	"inner product" of matrices
$\|\mathbf{A}\|_F = \sqrt{\mathbf{A} \circ \mathbf{A}}$	Frobenius norm of the matrix $\mathbf{A}$.
$\mathbf{A} \succeq \mathbf{0}$	positive semidefinite matrix (p.s.d.)
$\mathbf{A} \succ \mathbf{0}$	positive definite matrix
$\mathbf{A} \succeq \mathbf{B}$	$\mathbf{A} - \mathbf{B}$ is positive semidefinite
$\mathbf{I}$	unit matrix
diag $(d_1, \ldots, d_n)$	diagonal matrix
$\boldsymbol{\alpha}$	vector with all components equal to $\alpha \in \mathbb{R}$
span A	linear hull (span) of a set A
aff A	affine hull (affine span) of a set A
cone A	convex cone of a set A
conv A	convex hull of a set A
cl C	closure of a set C
int C	interior of a set C
$L^\perp$	orthogonal complement of L
C^0	dual cone of C
P^{pol}	polar of P

$[\mathbf{x}, \mathbf{y}]$	line segment between $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$
$P(\mathbf{A}, \mathbf{b})$	polyhedron of solutions of $\mathbf{Ax} \leq \mathbf{b}$
$\ker \mathbf{A}$	kernel of the matrix $\mathbf{A}$
$\nabla f(\mathbf{x}) = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ldots, \frac{\partial f(x)}{\partial x_n}\right)$	gradient of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x}$ (row vector)
$[\nabla f(\mathbf{x})]^T$ or $\nabla^T f(\mathbf{x})$	transpose of the gradient (column vector)
$\nabla f(\mathbf{x}) = \left(\frac{\partial f_i(\mathbf{x})}{\partial x_j}\right)$	Jacobian of $f : \mathbb{R}^n \to \mathbb{R}^m$ at $\mathbf{x}$
$[\nabla f(\mathbf{x})]^T$ or $\nabla^T f(\mathbf{x})$	transpose of the Jacobian
$\nabla_{\mathbf{x}} g(\mathbf{x}, \mathbf{y})$	partial derivative with respect to the $\mathbf{x}$-variable
$\nabla^2 f(\mathbf{x})$	Hessian of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x}$
$\partial f(\mathbf{x})$	subdifferential of $f : \mathbb{R}^n \to \mathbb{R}$ at $\mathbf{x}$
log	logarithm to base 2
ln	natural logarithm
$\langle q \rangle$	size of $q \in \mathbb{Q}$
$\langle \mathbf{x} \rangle$	size of $\mathbf{x} \in \mathbb{Q}^n$
$\langle I \rangle$	size of a problem instance I
$[\alpha]$	nearest integer
$\lceil \alpha \rceil$, $\lfloor \alpha \rfloor$	smallest integer $\geq \alpha$ resp. largest integer $\leq \alpha$
$w.l.o.g.$	without loss of generality
$\subseteq$	containment
$\subset$	proper containment

Bibliography

[1] A.V. Aho, J.E. Hopcraft and J.D. Ullman, The design and Analysis of Computer Algorithms, Addison-Wesley, Reading, (1974).

[2] R.K. Ahuja, T.L. Magnanti and J.B. Orlin, Network Flows, Prentice Hall, New Jersey, (1993).

[3] Atkinson K.E., Numerical Analysis, Wiley, New York, (1988).

[4] A. Bazaraa, H. Sherali and C. Shetty, Nonlinear Programming, John Wiley, New York, (1993).

[5] R.E. Bellman, *On a routing problem*, Quarterly of Applied Mathematics, 16, 87-90, (1958).

[6] R.C. Bland, *New finite pivoting rules for the simplex method*, Mathematics of Operations Research, 2, 103-107, (1977).

[7] L. Blum, F. Lucker, M. Shub and S. Smale, Complexity and Real Computation, Springer, (1997).

[8] Th. Bröcker, Differentiable germs and catastrophes, London Math. Soc. Lect. Notes Series 17, Cambridge University Press, (1975).

[9] V. Chvatal, *Edmonds polytopes and a hierarchy of combinatorial problems*, Discrete Mathematics 4, 305-337, (1973).

[10] A. Cohen, *Rate of convergence of several conjugate gradients algorithms*, SIAM Journal on Numerical Analysis, 9, 248-259, (1972).

[11] S. Cook, *The complexity of theorem-proving procedures*, Proc. 3rd Ann ACM Symp. on Theory of Computing, ACM, New York, (1971).

[12] W. Cook , W. Cunningham W. Pulleyblank and A. Schrijver, Combinatorial Optimization, John Wiley & Sons, New York, (1998).

[13] R.W. Cottle et al., The Linear Complementarity Problem, Academic Press, Boston, (1992).

[14] H.G. Daellenbach, Systems and Decision Making. John Wiley & Sons, Chichester (1994).

[15] M. Davis, Computability and Unsolvability, MacGraw-Hill, New York, (1958).

[16] J.E. Dennis and R.B. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, Prentice Hall, London, (1983).

[17] E. Dijkstra, *A note on two problems in connection with graphs*, Numerische Mathematik 1, 269-271, (1959).

[18] J. Edmonds, *Systems of Distinct Representation and Linear Algebra*, Journal of Research of the National Bureau of Standards, 71B,(4), (1967)

[19] U. Faigle, M. Hunting, W. Kern, R. Prakash and K.J. Supowit, *Simplices by point-sliding and the Yamnitsky-Levin algorithm*, Math. Methods of Operations Research 46, 131-142, (1997).

[20] J. Farkas, *Über die Theorie der einfachen Ungleichungen*, J. für die Reine und Angewandte Mathematik, 124, 1-27, (1902).

[21] W. Feller, *An Introduction to Probability Theory and Its Applications*, John Wiley, New York, 3rd. rev. ed., (1970).

[22] R. Fletcher, Practical Methods of Optimization, John Wiley & Sons, Chichester, (1987).

[23] L.R. Ford, Jr., *Network flow theory*, Paper P-923, RAND Corporation, Santa Monica, (1956).

[24] L.R. Ford and D.R. Fulkerson, Flows in Networks, Princeton University Press, Princeton, (1962).
[25] L.R. Ford and D.R. Fulkerson, *Maximal flow trough a network*, Canadian Hournal of Math.,8, 399–404, (1956).
[26] J.B. Fourier, Solution d'une question particuliere du calcul des inégalités, Oeuvres II, 317-328, (1826).
[27] B. Ganter and R. Wille, *Formal Concept Analysis*, Springer-Verlag, Berlin, (1999).
[28] M.R. Garey and D.S. Johnson, Computers and Intractibility: A Guided Tour to the Theory of NP-completeness, Freeman, San Francisco, (1979).
[29] A.M. Geoffrion, *Lagrangian relaxation for integer programming*, Mathematical Programming Study 2, 82-114, (1974).
[30] Ph.E. Gill, W. Murray, M.A. Saunders and M.A. Wright, *Inertia controlling methods for general quadratic problems*, SIAM Review 33, 1-36, (1991).
[31] M.X. Goemans and D.P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. Assoc. for Comp. Machinery 42, No. 6, 1115-1145, (1995).
[32] D. Goldfarb and A. Idnani, *A numerical stable dual method for solving strictly convex quadratic programs*, Math. Prog. 27, 1-33, (1983).
[33] G.H. Golub and C.F. Van Loan, *Matrix Computations*, The Johns Hopkins University Press 3rd. ed., (1996).
[34] R.F. Gomory, *An algorithm for integer solutions to linear problems*, in: Recent Advances in Mathematical Programming, (R.L. Graves and P. Wolfe, eds.), McGraw-Hill, New York, 262-302, (1963).
[35] P. Gordan, *Über die Auflösung linearer Gleichungen mit reellen Coefficienten*, Math. Ann. 6, 23-28, (1873).
[36] M. Grötschel, L. Lovasz and A. Schrijver, Geometric Algorithms and Combinatorial Optimization, 2nd edition, Springer, (1993)
[37] J. Gruska, Quantum Computing, McGraw Hill, London, (1999).
[38] Handbook of Semidefinite Programming: Theory, Algorithms and Applications, eds. Wolkowicz, Saigal and Vandenberghe, Kluwer, Boston, (2000).
[39] M. Held and R.M. Karp, *The travelling salesman problem and minimum spanning trees*, Operations Research 18, 1138-1162, (1970).
[40] R. Hettich and K. Kortanek, Semi-infinite programming: Theory, methods and applications, SIAM Review, vol 35, No.3, 380-429, (1993).
[41] R. Horst and H. Tuy, Global Optimization, Springer, Berlin, (1996).
[42] P. Lancaster and M. Tismenetsky, The Theory of Matrices, Academic Press, Boston, (1985).
[43] M. Hunting, U. Faigle and W. Kern, *A Lagrangian relaxation approach to the edge weighted clique problem*, European J. of Operational Research, 131, 119-131, (2001).
[44] R.G. Jeroslow, *There cannot be any algorithm for integer programming with quadratic constraints*, Operations Research 21,221-224, (1973).
[45] F. John, *Extremum problems with inequalities as subsidiaries conditions*. Studies and Essays, Presented to R. Courant on his 60th Birthday January 8, 1948, Interscience, New York, 187-204, (1948).
[46] N. Karmarkar, *A new polynomial time algorithm for linear programming*, Combinatorics, 4, 373-395, (1984).
[47] R.M. Karp, *Reducibility among combinatorial problems*, Complexity of Computer Computations (R.E. Miller and J.W. Thatcher, eds.), Plenum Press, New York, (1972).
[48] W. Kern and D. Paulusma, *The new FIFA rules are hard: Complexity aspects of sports competitions*, Discr. Appl. Math. 108, 317-323, (2001).

[49] L.G. Khachian, *Polynomial algorithms in linear programming* (in Russian), Doklady Akademii Nauk SSSR 244, 1093-1096. (English translation: Soviet Mathematics Doklady 20, 191-194), (1979).
[50] V. Klee and G.J. Minty, *How good is the simplex algorithm?*, in Inequalities III, ed. by O. Shisha, Acad. Press, New York, (1972).
[51] H.W. Kuhn, *Nonlinear programming: A historical note*. In: History of Mathematical Programming, J.K. Lenstra *et al.* eds., CWI, Amsterdam, 82-96, (1991).
[52] C. Lemaréchal and F. Oustry, *Semi-definite relaxations and Lagrangian duality with applications to combinatorical optimization*, Rapport de Recherche 3710, INRIA, (1999).
[53] G. Lekkerkerker, Geometry of Numbers. North-Holland, Amsterdam, (1969).
[54] L. Lovasz and M. Plummer, Matching Theory, North Holland, Amsterdam, (1986).
[55] D.G. Luenberger, Optimization by Vector Space Methods, John Wiley & Sons, New York, (1969).
[56] D.G. Luenberger, Introduction to Linear and Nonlinear Programming, Addison-Wesley, (1980).
[57] E.M. Macambira and C.C. de Souza, *The edge weighted clique problem: Valid inequalities, facets and polyhedral computations*, European Journal of Operational Research 123, 346-371, (1999).
[58] N. Maratos, *Exact penalty function algorithms for finite dimensional and control optimization problems*, Ph.D. Thesis, Imperial College Sci. Tech., University of London, (1978).
[59] R. Motwani and P. Raghavan, Randomized Algorithms, Cambridge Univ. Press, (1995).
[60] T.S. Motzkin, *Beiträge zur Theorie der Linearen Ungleichungen*, Dissertation, University of Basel, Jerusalem, (1936).
[61] J. von Neumann, *Zur Theorie der Gesellschaftsspiele*, Mathematische Annalen 100, 295-320, (1928).
[62] J. von Neumann, *A certain zero-sum game equivalent to the optimal assignment problem*. In: Contributions to the Theory of Games I. H.W. Kuhn and A.W. Tucker, eds., Annals of Mathematics Studies 28, 5-12, (1953).
[63] J. Nocedal and S.J. Wright, Numerical Optimization, Springer (1999).
[64] M. Padberg, *The boolean quadric polytope: some characteristics, facets and relatives*, Math. Progr. 45, 139-172, (1989).
[65] B.T. Polyak, *A general method for solving extremum problems*, Soviet Math. No 8, 593-597, (1966).
[66] V. Pratt, *Every prime has a succinct certificate*, SIAM Journal of Computing, 4, 214–220, (1975).
[67] C.R. Rao, Linear Statistical Inference and Its Applications, Wiley, New York, (1973).
[68] C. Roos, T. Terlaky and J.-Ph. Vial , Theory and Algorithms for Linear Optimization, John Wiley & Sons, Chichester, (1997).
[69] W. Rudin, *Principles of Mathematical Analysis*, (third edition), McGraw-Hill, (1976).
[70] A. Schrijver, Theory of Linear and Integer programming, John Wiley, New York, (1986).
[71] P. Spelucci, *Numerische Verfahren der nichtlinearen Optimierung*, Birkhäuser Verlag, Boston, (1993).
[72] D.M. Topkis and A.F. Veinott, *On the convergence of some feasible direction algorithms for nonlinear programming*, SIAM J. Control 5, 268-279, (1967).
[73] A. Turing, *On computable numbers, with application to the Entscheidungsproblem*, Proc. London Math. Soc. Ser. 2, 42, 230-265 and 43, 544-546, (1936).
[74] H. Tuy, *Convex Analysis and Global Optimization*, Kluwer, Dordrecht, (1998).
[75] S.A. Vavasis, Nonlinear Optimization (Complexity issues), Oxford University Press, New York, (1991).
[76] D. Welsh, *Codes and Cryptography*, Clardendon Press, Oxford, (1988).

[77] C.P. Williams and S.H. Clearwater, Explorations in Quantum Computing. Springer-Verlag, Heidelberg, (1998).
[78] P. Wolfe, *On the convergence of gradient methods under constraints*, IBM J. Research and Development, 16, 407-411, (1972).
[79] G.M. Ziegler, Lectures on Polytopes, Springer-Verlag, New York, (1999).
[80] G. Zoutendijk, *Methods of Feasible Directions*, Elsevier, Amsterdam, (1960).

Index