http://www.math.uni-konstanz.de/~schweigh/

Lecture notes

Real Algebraic Geometry, Positivity and Convexity

Academic Year 2016/2017

Markus Schweighofer

Version of Thursday 30th August, 2018, 22:11

Universität Konstanz, Germany

Preface. Chapters 1–4 are lecture notes of my course "Real Algebraic Geometry I" from the winter term 2016/2017. Chapters 5–8 are lecture notes of its continuation "Real Algebraic Geometry II" from the summer term 2017. Chapter 9 has been taught in my course "Geometry of Linear Matrix Inequalities" from the same summer term. Please report any ambiguities and errors (including typos) to:

markus.schweighofer@uni-konstanz.de

This document is to a large extent based on the work of other people. For the relevant scientific sources, we refer to the literature referenced at the end of this document as well as the bibliographies of the books [ABR, BCR, BPR, KS, Mar, PD]. I would like to thank the numerous people that helped to improve these lecture notes: First of all, I thank Tom-Lukas Kriel, especially for coauthoring Chapter 9. Thanks go also to Sebastian Gruler and María López Quijorna and to those participants that pointed out errors and typos (among them I mention especially Alexander Taveira Blomenhofer, Nicolas Daans, Carl Eggen, Rüdiger Grunwald and Emre Öztürk in alphabetical order).

Contents

Introduction

The study of polynomial equations is a canonical subject in mathematics education, as is illustrated by the following examples: Quadratic equations in one variable (high school), systems of linear equations (linear algebra), polynomial equations in one variable and their symmetries (algebra, Galois theory), diophantine equations (number theory) and systems of polynomial equations (algebraic geometry, commutative algebra).

In contrast to this, the study of polynomial inequalities (in the sense of "greater than" or "greater or equal than") is mostly neglected even though it is much more important for applications: Indeed, in applications one often searches for a real solution rather than a complex one (as in classical algebraic geometry) and this solution must not necessarily be exact but only approximate.

In a course about linear algebra there is frequently no time for linear optimization. An introductory course about algebra usually treats groups, rings and fields but disregards ordered and real closed fields as well as preorders or prime cones of rings. In a first course on algebraic geometry there is often no special attention paid to the real part of a variety and in commutative algebra quadratic modules are practically never treated.

Most algebraists do not even know the notion of a preorder although it is as important for the study of systems of polynomial inequalities as the notion of an ideal is for the study of systems of polynomial equations. People from more applied areas such as numerical analysis, mathematical optimization or functional analysis know often more about real algebraic geometry than some algebraists, but often do not even recognize that polynomials play a decisive role in what they are doing. There are for example countless articles from functional analysis which are full of equations with binomial coefficients which turn out to be just disguised simple polynomial identities.

In the same way as the study of polynomial systems of equations leads to the study of rings and their generalizations (such as modules), the study of systems of polynomial inequalities leads to the study of rings which are endowed with something that resembles an order. This additional structure raises many new questions that have to be clarified. These questions arise already at a very basic level so that we need as prerequisites only basic linear algebra, algebra and analysis. In particular, this course is really extremely well suited to students heading for a teaching degree. It includes several topics which are directly relevant for high school teaching.

To arouse the reader's curiosity, we present the following table. It contains on the left column notions we assume the reader is familiar with. On the right column we name

what could be seen more or less as their real counterparts mostly introduced in this course.

Algebra	Real Algebra
Algebraic Geometry	Real Algebraic Geometry
systems of polynomial equations	systems of polynomial inequalities
"$=$"	"$\geq$"
complex solutions	real solutions
$\mathbb{C}$	$\mathbb{R}$
algebraically closed fields	real closed fields
fields	ordered fields
ideals	preorders
prime ideals	prime cones
spectrum	real spectrum
Noetherian	quasi-compact
radical	real radical
fundamental theorem of algebra	fundamental theorem of algebra
Aachen, Aalborg, Aarhus, ...	Dortmund, Dublin, Innsbruck, ...
..., Zagreb, Zürich	..., Konstanz, Ljubljana, Rennes

It is intended that the fundamental theorem of algebra appears on both sides of the table. In its usual form, it says that each non-constant univariate complex polynomial has a complex root. In Section 1.4, we will formulate it in a "real" way. The difficulties one has to deal with in the "real world" become already apparent when one asks the corresponding "real question": When does a univariate complex polynomial have a real root? The answer to this will be given in Section 1.6 and requires already quite some thoughts.

Traditionally, Real Algebraic Geometry has many ties with fields like Model Theory, Valuation Theory, Quadratic Form Theory and Algebraic Topology. In this lecture, we mainly emphasize however connections to fields like Optimization, Functional Analysis and Convexity that came up during the recent years and are now fully established.

Throughout the lecture, $\mathbb{N} := \{1, 2, 3, \dots\}$ and $\mathbb{N}_0 := \{0\} \cup \mathbb{N}$ denote the set of positive and nonnegative integers, respectively.

§1 Ordered fields

1.1 Orders of fields

Reminder 1.1.1. Let M be a set. An *order* on M is a relation $\leq$ on M such that for all $a, b, c \in M$:

$$\begin{array}{rcll} & a \leq a & & \text{(reflexivity)} \\ & (a \leq b \,\&\, b \leq c) & \implies a \leq c & \text{(transitivity)} \\ & (a \leq b \,\&\, b \leq a) & \implies a = b & \text{(antisymmetry)} \\ \text{and} & a \leq b \text{ or } b \leq a & & \text{(linearity)} \end{array}$$

In this case, $(M, \leq)$ (or simply M if $\leq$ is clear from the context) is called an *ordered set*. For $a, b \in M$, one defines

$$\begin{aligned} a < b &: \iff a \leq b \,\&\, a \neq b, \\ a \geq b &: \iff b \leq a \end{aligned}$$

and so on.

Definition 1.1.2. Let $(M, \leq_1)$ and $(N, \leq_2)$ be ordered sets and $\varphi\colon M \to N$ be a map. Then φ is called a *homomorphism* (of ordered sets) or *monotonic* if

$$a \leq_1 b \implies \varphi(a) \leq_2 \varphi(b)$$

for all $a, b \in M$. If φ is $\left\{\begin{matrix}\text{injective}\\ \text{bijective}\end{matrix}\right\}$ and if

$$a \leq_1 b \iff \varphi(a) \leq_2 \varphi(b)$$

for all $a, b \in M$, then φ is called an $\left\{\begin{matrix}\textit{embedding}\\ \textit{isomorphism}\end{matrix}\right\}$ (of ordered sets).

Proposition 1.1.3. *Let $(M, \leq_1)$ and $(N, \leq_2)$ be ordered sets and $\varphi\colon M \to N$ a homomorphism. Then the following are equivalent:*

(a) *φ is an embedding*

(b) *φ is injective*

(c) $\forall a, b \in M : (\varphi(a) \leq_2 \varphi(b) \implies a \leq_1 b)$

Proof. (c) $\Longrightarrow$ (b) Suppose (c) holds and let $a, b \in M$ such that $\varphi(a) = \varphi(b)$. Then $\varphi(a) \leq_2 \varphi(b)$ and $\varphi(a) \geq_2 \varphi(b)$. Now (c) implies $a \leq_1 b$ and $a \geq_1 b$. Hence $a = b$.

(b) $\Longrightarrow$ (c) Suppose (b) holds and let $a, b \in M$ with $a \not\leq_1 b$. To show: $\varphi(a) \not\leq_2 \varphi(b)$. We have $a >_1 b$ and it suffices to show $\varphi(a) >_2 \varphi(b)$. From $a \geq_1 b$ it follows by the monotonicity of φ that $\varphi(a) \geq_2 \varphi(b)$. From $a \neq b$ and the injectivity of φ we get $\varphi(a) \neq \varphi(b)$.

From (b) $\Longleftrightarrow$ (c) and (a) $\Longleftrightarrow$ ((b)&(c)) [$\rightarrow$ 1.1.5] the claim now follows. □

Definition 1.1.4. Let K be a field. An *order* of K is an order $\leq$ on K such that for all $a, b, c \in K$ we have:

$$\begin{aligned} a \leq b &\implies a + c \leq b + c \qquad \text{(monotonicity of addition)} \\ \text{and} \quad (a \leq b \,\&\, c \geq 0) &\implies ac \leq bc \quad \text{(monotonicity of multiplication).} \end{aligned}$$

In this case, $(K, \leq)$ (or simply K when $\leq$ is clear from the context) is called an *ordered field*.

Definition 1.1.5. Let $(K, \leq_1)$ and $(L, \leq_2)$ be ordered fields.

A field homomorphism (or equivalently, field embedding!) $\varphi\colon K \to L$ is called a *homomorphism* or *embedding* of ordered fields if φ is monotonic (pay attention to 1.1.3 together with the fact that field homomorphisms are injective). If φ is moreover surjective, then φ is called an *isomorphism* of ordered fields.

If there exists an embedding of ordered fields from $(K, \leq_1)$ into $(L, \leq_2)$, then $(K, \leq_1)$ is called *embeddable* in $(L, \leq_2)$ and one denotes $(K, \leq_1) \hookrightarrow (L, \leq_2)$. If there is an isomorphism of ordered fields from $(K, \leq_1)$ to $(L, \leq_2)$, then $(K, \leq_1)$ and $(L, \leq_2)$ are called *isomorphic*. This is denoted by $(K, \leq_1) \cong (L, \leq_2)$.

$(K, \leq_1)$ is called an *ordered subfield* of $(L, \leq_2)$, or equivalently $(L, \leq_2)$ an *ordered extension field* of $(K, \leq_1)$, if $(K, \leq_1) \to (L, \leq_2)$, $a \mapsto a$ is an embedding, that is if K is a subfield of L and $(\leq_1) = (\leq_2) \cap K \times K$. For every subfield of L there is obviously a unique order making it into an ordered subfield of $(L, \leq_2)$. This order is called the order *induced* by $(L, \leq_2)$.

Proposition 1.1.6. *Let $(K, \leq)$ be an ordered field. Then $a^2 \geq 0$ for all $a \in K$.*

Proof. Let $a \in K$. When $a \geq 0$ this follows immediately from the monotonicity of multiplication [$\rightarrow$ 1.1.4]. When $a \leq 0$ the monotonicity of addition [$\rightarrow$ 1.1.4] yields $0 = a - a \leq -a$, whence $-a \geq 0$ and therefore $a^2 = (-a)^2 \geq 0$. □

Proposition 1.1.7. *Let $(K, \leq)$ be an ordered field. Then K is of characteristic 0 and the uniquely determined field homomorphism $\mathbb{Q} \to K$ is an embedding of ordered fields $(\mathbb{Q}, \leq_{\mathbb{Q}}) \hookrightarrow (K, \leq)$. Hence $(K, \leq)$ can be seen as an ordered extension field of $(\mathbb{Q}, \leq_{\mathbb{Q}})$. In particular, for $K = \mathbb{Q}$ it follows that $(\leq_{\mathbb{Q}}) = (\leq)$, i.e., $\mathbb{Q}$ can only be ordered in the familiar way.*

Proof. From 1.1.6 we have $0 \leq 1^2 = 1$ in $(K, \leq)$. Using the monotonicity of the addition, we deduce

$$(*) \qquad 0 \leq 1 \leq 1 + 1 \leq 1 + 1 + 1 \leq \ldots$$

If we had $\operatorname{char} K \neq 0$, then $(*)$ would give $0 \leq 1 \leq 0$ by the transitivity of $\leq$ which would imply $0 = 1$ in K by the antisymmetry of $\leq$, contradicting the definition of a field. Let φ denote the field homomorphism $\mathbb{Q} \to K$ and let $a, b \in \mathbb{Q}$ with $a \leq_{\mathbb{Q}} b$. To show: $\varphi(a) \leq \varphi(b)$. Write $a = \frac{k}{n}$ and $b = \frac{\ell}{n}$ with $k, \ell \in \mathbb{Z}$ and $n \in \mathbb{N}$. Then

$$\varphi(n) = \underbrace{1 + \cdots + 1}_{n \text{ times}} \overset{\substack{(*) \\ \operatorname{char} K = 0}}{>} 0$$

and, by the monotonicity of multiplication and Proposition 1.1.6, also

$$\frac{1}{\varphi(n)} = \left(\frac{1}{\varphi(n)}\right)^2 \varphi(n) \geq 0.$$

Hence it suffices to show that $\varphi(a)\varphi(n) \leq \varphi(b)\varphi(n)$. This reduces to $\varphi(an) \leq \varphi(bn)$, that is $\varphi(k) \leq \varphi(\ell)$, or equivalently $\varphi(\ell - k) \geq 0$. But due to $\ell - k \geq_{\mathbb{Q}} 0$ this follows from $(*)$. □

Proposition and Notation 1.1.8. *Let $(K, \leq)$ be an ordered field. Then for every $a \in K^\times$ there are uniquely determined* $\operatorname{sgn} a \in \{-1, 1\}$ *("sign" of a) and* $|a| \in K_{\geq 0} := \{x \in K \mid x \geq 0\}$ *("absolute value" of a) such that*

$$a = (\operatorname{sgn} a)|a|.$$

One declares $\operatorname{sgn} 0 := |0| := 0$. *It follows that* $|ab| = |a||b|$, $\operatorname{sgn}(ab) = (\operatorname{sgn} a)(\operatorname{sgn} b)$ *and* $|a + b| \leq |a| + |b|$ *for all* $a, b \in K$.

Proof. The first part is very easy. Let now $a, b \in K$. Then $ab = (\operatorname{sgn} a)(\operatorname{sgn} b)|a||b|$, implying $|ab| = |a||b|$ as well as $\operatorname{sgn}(ab) = (\operatorname{sgn} a)(\operatorname{sgn} b)$. For the claimed triangle inequality, we can suppose $a + b \geq 0$ (otherwise replace a by $-a$ and b by $-b$). Then $|a + b| = a + b \leq a + |b| \leq |a| + |b|$. □

Definition 1.1.9. Let $(K, \leq)$ be an ordered field.

(a) $(K, \leq)$ is called *Archimedean* if $\forall a \in K : \exists N \in \mathbb{N} : a \leq N$ (or equivalently, $\forall a \in K : \exists N \in \mathbb{N} : -N \leq a$).

(b) A sequence $(a_n)_{n \in \mathbb{N}}$ in K is called

- a *Cauchy sequence* if $\forall \varepsilon \in K_{>0} : \exists N \in \mathbb{N} : \forall m, n \geq N : |a_m - a_n| < \varepsilon$,
- *convergent to* $a \in K$ if $\forall \varepsilon \in K_{>0} : \exists N \in \mathbb{N} : \forall n \geq N : |a_n - a| < \varepsilon$ (one easily shows that a is then uniquely determined and writes $\lim_{n \to \infty} a_n = a$),
- *convergent* if there is some $a \in K$ such that $\lim_{n \to \infty} a_n = a$.

We call $(K, \leq)$ *Cauchy complete* if every Cauchy sequence converges in K (by the way it is immediate that every convergent sequence is a Cauchy sequence).

(c) We call a subset $A \subseteq K$ *bounded from above* if K contains an upper bound for A (meaning some $b \in K$ such that $\forall a \in A : a \leq b$). We call $(K, \leq)$ *complete* if every nonempty subset of K bounded from above possesses a lowest upper bound, i.e., a supremum.

Proposition 1.1.10. *Let $(K, \leq)$ be an ordered field. Then the following are equivalent:*

(a) $(K, \leq)$ *is Archimedean*

(b) $\forall a, b \in K : (a < b \implies \exists c \in \mathbb{Q} : a < c < b)$

Proof. (b) $\implies$ (a) Suppose (b) holds and let $a \in K$. To show: $\exists N \in \mathbb{N} : a \leq N$. WLOG $a > 0$. To show: $\exists N \in \mathbb{N} : \frac{1}{N} \leq \frac{1}{a}$. Choose $c \in \mathbb{Q}$ such that $0 < c < \frac{1}{a}$. Write $c = \frac{m}{N}$ for certain $m, N \in \mathbb{N}$. Then $\frac{1}{N} \leq \frac{m}{N} = c < \frac{1}{a}$.

(a) $\implies$ (b) Suppose (a) holds and let $a, b \in K$ such that $a < b$. Choose $N \in \mathbb{N}$ such that $\frac{1}{b-a} < N$. Then $\frac{1}{N} < b - a$ and hence $a + \frac{1}{N} < b$. Now choose the smallest $m \in \mathbb{Z}$ such that $a < \frac{m}{N}$. If we had $\frac{m}{N} \geq b$, then $a + \frac{1}{N} < \frac{m}{N}$ and therefore $a < \frac{m-1}{N}$, contradicting our choice of m. Therefore $a < \frac{m}{N} < b$. □

Lemma 1.1.11. Let $(K, \leq)$ be an Archimedean ordered field. Then

$$K = \left\{ \lim_{n\to\infty} a_n \mid (a_n)_{n\in\mathbb{N}} \text{ sequence in } \mathbb{Q} \text{ that converges in } K \right\}.$$

Proof. Let $a \in K$. We have to show that there is a sequence $(a_n)_{n\in\mathbb{N}}$ in $\mathbb{Q}$ that converges in K to a. Choose for every $n \in \mathbb{N}$ according to 1.1.10 some $a_n \in \mathbb{Q}$ such that $a \leq a_n < a + \frac{1}{n}$. Let $\varepsilon \in K_{>0}$. Choose $N \in \mathbb{N}$ such that $\frac{1}{\varepsilon} < N$. For $n \geq N$ we now have $|a_n - a| = a_n - a < \frac{1}{n} \leq \frac{1}{N} < \varepsilon$. □

Lemma 1.1.12. Suppose $(K, \leq)$ is an Archimedean ordered field and $(a_n)_{n\in\mathbb{N}}$ is a sequence in $\mathbb{Q}$. Then the following are equivalent:

(a) $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $(\mathbb{Q}, \leq_{\mathbb{Q}})$

(b) $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $(K, \leq)$

Proof. This follows easily from 1.1.10. □

Exercise 1.1.13. Suppose $(K, \leq)$ is an ordered field and $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ are convergent sequences in K. Then

$$\lim_{n\to\infty}(a_n + b_n) = \left(\lim_{n\to\infty} a_n\right) + \left(\lim_{n\to\infty} b_n\right) \qquad \text{and} \qquad \lim_{n\to\infty}(a_n b_n) = \left(\lim_{n\to\infty} a_n\right)\left(\lim_{n\to\infty} b_n\right).$$

Theorem 1.1.14. *Let $(K, \leq)$ be an ordered field. Then the following are equivalent:*

(a) $(K, \leq)$ *is Archimedean and Cauchy complete*

(b) $(K, \leq)$ *is complete*

Proof. (a) $\implies$ (b) Suppose (a) holds and let $A \subseteq K$ be a nonempty subset bounded from above. Choose for every $n \in \mathbb{N}$ the smallest $k_n \in \mathbb{Z}$ such that $\forall a \in A : a \leq \frac{k_n}{n}$ and set $a_n := \frac{k_n}{n} \in \mathbb{Q}$ (use the Archimedean property!). Using again the Archimedean property, one can show easily that $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence and therefore convergent

by hypothesis. We leave it as an exercise to the reader to show that $a := \lim_{n\to\infty} a_n$ is the lowest upper bound of A in $(K, \leq)$.

(b) $\Longrightarrow$ (a) We prove the contraposition.

First, suppose that $(K, \leq)$ is not Archimedean, i.e., the set

$$A := \{a \in K \mid \forall N \in \mathbb{N} : a \leq -N\}$$

is not empty. We claim that A does not have a lowest upper bound: Indeed, if $a \in K$ is an upper bound of A, then so is $a - 1 < a$ since $A = \{a \in K \mid \forall N \in \mathbb{Z} : a \leq N\} = \{a \in K \mid \forall N \in \mathbb{Z} : a + 1 \leq N\} = \{a - 1 \mid a \in K, \forall N \in \mathbb{N} : a \leq N\} = A - 1$.

Finally, suppose that $(K, \leq)$ is not Cauchy complete, say $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in K that does not converge. We claim that

$$A := \{a \in K \mid \exists N \in \mathbb{N} : \forall n \geq N : a \leq a_n\}$$

is nonempty and bounded from above but does not possess a lowest upper bound. We leave this as an exercise to the reader. □

Lemma 1.1.15. Suppose $(K, \leq)$ is an Archimedean ordered field and $(R, \leq_R)$ a complete ordered field. Then there is exactly one embedding $(K, \leq) \hookrightarrow (R, \leq_R)$. This embedding is an isomorphism if and only if $(K, \leq)$ is complete.

Proof. Exercise. □

Theorem 1.1.16. *There is a complete ordered field $(\mathbb{R}, \leq)$. It is essentially unique, for if $(K, \leq_K)$ is another complete ordered field, then there is exactly one isomorphism from $(K, \leq_K)$ to $(\mathbb{R}, \leq)$.*

Proof. The uniqueness is clear from 1.1.15 together with 1.1.14. We only sketch the proof of existence and leave the details as an exercise to the reader: Show that the Cauchy sequences in $\mathbb{Q}$ form a subring C of $\mathbb{Q}^{\mathbb{N}}$ and that

$$I := \left\{(a_n)_{n\in\mathbb{N}} \in C \mid \lim_{n\to\infty} a_n = 0\right\}$$

is a maximal ideal of C. Set $\mathbb{R} := C/I$. Show that

$$a \leq b :\Longleftrightarrow \exists (a_n)_{n\in\mathbb{N}}, (b_n)_{n\in\mathbb{N}} \text{ in } C : (a = \overline{(a_n)_{n\in\mathbb{N}}}^I \,\&\, b = \overline{(b_n)_{n\in\mathbb{N}}}^I \,\&\, \forall n \in \mathbb{N} : a_n \leq b_n)$$

$(a, b \in \mathbb{R})$ defines an order $\leq$ on $\mathbb{R}$. It is clear that $(\mathbb{R}, \leq)$ is Archimedean. By Theorem 1.1.14 it suffices to show that $(\mathbb{R}, \leq)$ is Cauchy complete. To this end, let $(a_n)_{n\in\mathbb{N}}$ be a Cauchy sequence in $(\mathbb{R}, \leq)$. By 1.1.10, there exists a sequence $(q_n)_{n\in\mathbb{N}}$ in $\mathbb{Q}$ such that $|a_n - q_n| < \frac{1}{n}$ for $n \in \mathbb{N}$. Now deduce from the fact that $(a_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in $(\mathbb{R}, \leq)$ that $(q_n)_{n\in\mathbb{N}}$ is such in $(\mathbb{R}, \leq)$ and hence also in $(\mathbb{Q}, \leq)$. Now $(q_n)_{n\in\mathbb{N}} \in C$. Set $a := \overline{(q_n)_{n\in\mathbb{N}}}^I$. It is enough to show $\lim_{n\to\infty} a_n = a$. Finally show that this is equivalent to $\lim_{n\to\infty} q_n = a$ in $(K, \leq)$ and prove the latter. □

Corollary 1.1.17. *$(\mathbb{R}, \leq)$ is an Archimedean ordered field into which every Archimedean ordered field can be embedded. Up to isomorphy it is the only such ordered field.*

Proof. The first statement is clear from 1.1.14, 1.1.15 and 1.1.16. Uniqueness: Let $(K, \leq_K)$ be another such ordered field. Then

$$(\mathbb{R}, \leq) \overset{\varphi}{\hookrightarrow} (K, \leq_K) \overset{\psi}{\hookrightarrow} (\mathbb{R}, \leq)$$

and $\psi \circ \varphi$ is the by 1.1.15 unique embedding $(\mathbb{R}, \leq) \hookrightarrow (\mathbb{R}, \leq)$, i.e., $\psi \circ \varphi = \mathrm{id}$. This implies that ψ is surjective. Hence $(K, \leq_K) \cong (\mathbb{R}, \leq)$. □

Notation 1.1.18. Let A be a ring. Then we often use suggestive notation to describe certain subsets of A such as the following:

- $A^2 = \{a^2 \mid a \in A\}$ ("squares")
- $\sum A^2 = \{\sum_{i=1}^{\ell} a_i^2 \mid \ell \in \mathbb{N}_0, a_i \in A\}$ ("sums of squares")
- $\sum A^2 T = \left\{\sum_{i=1}^{\ell} a_i^2 t_i \mid \ell \in \mathbb{N}_0, a_i \in A, t_i \in T\right\}$ $(T \subseteq A)$
  ("sums of elements of T weighted by squares")
- $T + T = \{s + t \mid s, t \in T\}$ $(T \subseteq A)$
- $TT = \{st \mid s, t \in T\}$ $(T \subseteq A)$
- $-T = \{-t \mid t \in T\}$ $(T \subseteq A)$
- $T + aT = \{s + at \mid s, t \in T\}$ $(T \subseteq A, a \in A)$

Proposition 1.1.19. *Let K be a field.*

(a) *If $\leq$ is an order of K [$\to$ 1.1.4], then $P := K_{\geq 0} = \{a \in K \mid a \geq 0\}$ has the following properties:*

$$(*) \qquad P + P \subseteq P, \quad PP \subseteq P, \quad P \cup -P = K \quad \text{and} \quad P \cap -P = \{0\}.$$

(b) *If P is a subset of K satisfying $(*)$, then the relation $\leq_P$ on K defined by*

$$a \leq_P b :\iff b - a \in P \qquad (a, b \in K)$$

is an order of K.

(c) *The correspondence*

$$(\leq) \mapsto K_{\geq 0}$$
$$(\leq_P) \mapsfrom P$$

defines a bijection between the set of all orders on K and the set of all subsets of K satisfying $(*)$.

Proof. (a) We get $P \dot{+} P \subseteq P$ from the monotonicity of $\left\{\begin{array}{c}\text{addition}\\ \text{multiplication}\end{array}\right\}$ [→ 1.1.4], $P \cup -P = K$ from the linearity [→ 1.1.1] and $P \cap -P = \{0\}$ from the antisymmetry [→ 1.1.1].

(b) We get reflexivity from $0 \in P$, transitivity from $P + P \subseteq P$, antisymmetry from $P \cap -P = \{0\}$, linearity from $P \cup -P = K$, monotonicity of addition from the definition of $\leq_P$ and monotonicity of multiplication $PP \subseteq P$.

(c) Suppose first that $\leq$ is an order of K and set $P := K_{\geq 0}$. Then $(\leq) = (\leq_P)$ since $a \leq b \iff b - a \geq 0 \iff b - a \in P \iff a \leq_P b$ for all $a, b \in K$. Conversely, let $P \subseteq K$ be given such that P satisfies $(*)$. We show $K_{\geq_P 0} = P$. Indeed,

$$K_{\geq_P 0} = \{a \in K \mid 0 \leq_P a\} = \{a \in K \mid a \in P\} = P.$$

□

Remark 1.1.20. 1.1.19(c) allows us to view orders of fields K as subsets of K. We reformulate some of the preceding notions and results in this new language:

(a) Definition 1.1.4: Let K be a field. An order of K is a subset P of K satisfying

$$P + P \subseteq P, \quad PP \subseteq P, \quad P \cup -P = K \quad \text{and} \quad P \cap -P = \{0\}.$$

(b) Definition 1.1.5: Let (K, P) and (L, Q) be ordered fields. A field homomorphism $\varphi\colon K \to L$ is called a homomorphism or an embedding of ordered fields if $\varphi(P) \subseteq Q$. One calls (K, P) an ordered subfield of (L, Q) if K is a subfield of L and $P = Q \cap K$ (or equivalently $P \subseteq Q$).

(c) Proposition 1.1.6: Let (K, P) be an ordered field. Then $K^2 \subseteq P$.

(d) Definition 1.1.9: An ordered field (K, P) is called *Archimedean* if

$$\forall a \in K : \exists N \in \mathbb{N} : N + a \in P,$$

$$(\iff P - \mathbb{N} = K \iff P + \mathbb{Z} = K \iff P + \mathbb{Q} = K).$$

1.2 Preorders

Definition 1.2.1. Let A be a commutative ring and $T \subseteq A$. Then T is called a *preorder* of A if $A^2 \subseteq T$, $T + T \subseteq T$ and $TT \subseteq T$. If moreover $-1 \notin T$, then T is called a *proper* preorder of A.

Example 1.2.2. (a) If A is a commutative ring, then $\sum A^2$ is the smallest preorder of A.

(b) Every order of a field is a proper preorder.

Proposition 1.2.3. *Let A be a commutative ring with $\frac{1}{2} \in A$ (i.e., $2 \in A^\times$). Then*

$$a = \left(\frac{a+1}{2}\right)^2 - \left(\frac{a-1}{2}\right)^2$$

for all $a \in A$. In particular, $A = A^2 - A^2$.

Definition and Proposition 1.2.4. *Let A be a commutative ring with $\frac{1}{2} \in A$ and $T \subseteq A$ a preorder. Then the* support $T \cap -T$ *of T is an ideal of A.*

Proof. $T \cap -T$ is obviously a subgroup of (the additive group of) A and we have

$$\begin{array}{rcl} A(T \cap -T) & \overset{1.2.3}{=} & (A^2 - A^2)(T \cap -T) \\ & \subseteq & (T - T)(T \cap -T) \\ & \subseteq & (T(T \cap -T)) - (T(T \cap -T)) \\ & \subseteq & ((TT) \cap (-TT)) + ((-TT) \cap TT) \\ & \subseteq & (T \cap -T) + ((-T) \cap T) = (T \cap -T) + (T \cap -T) \subseteq T \cap -T. \end{array}$$

□

Corollary 1.2.5. *Suppose A is a commutative ring with $\frac{1}{2} \in A$ and $T \subseteq A$ is a preorder. Then*

$$T \text{ is proper} \iff T \neq A.$$

Proof. "$\Longrightarrow$" trivial

"$\Longleftarrow$" Suppose $T \neq A$. Then of course also $T \cap -T \neq A$. Since $T \cap -T$ is an ideal, we have $1 \notin T \cap -T$. Since $1 = 1^2 \in T$, it follows that $1 \notin -T$, i.e., $-1 \notin T$. □

Example 1.2.6. In 1.2.3, 1.2.4 and 1.2.5, it is essential to require $\frac{1}{2} \in A$. Take for example $A = \mathbb{F}_2(X)$. Then $A^2 = \mathbb{F}_2(X^2)$ since $\mathbb{F}_2(X) \to \mathbb{F}_2(X)$, $p \mapsto p^2$ is a homomorphism (Frobenius). Therefore $A^2 - A^2 = \mathbb{F}_2(X^2) \neq \mathbb{F}_2(X)$. Moreover $T := \mathbb{F}_2(X^2) = \sum A^2$ is a preorder of A but $T \cap -T = \mathbb{F}_2(X^2)$ is not an ideal of A (since $1 \in T \cap -T \neq \mathbb{F}_2(X)$). Also T is not proper although $T \neq A$. The same is true for $\mathbb{F}_2[X]$ instead of $\mathbb{F}_2(X)$ and from this one can get similar examples in the ring $\mathbb{Z}[X]$ (exercise).

Proposition 1.2.7. *Let K be a field and $T \subseteq K$ a preorder. Then*

$$T \text{ is proper} \iff T \cap -T = \{0\}.$$

Proof. If $\operatorname{char} K = 2$, then $-1 = 1 \in T \cap -T$. Therefore suppose now $\operatorname{char} K \neq 2$. Then

$$-1 \notin T \overset{1 \in T}{\iff} 1 \notin T \cap -T \underset{\substack{K \text{ field} \\ \operatorname{char} K \neq 2}}{\overset{1.2.4}{\iff}} T \cap -T = \{0\}.$$

□

Lemma 1.2.8. Suppose A is a commutative ring, $T \subseteq A$ a preorder and $a \in A$. Then $T + aT$ is again a preorder.

Proof. $A^2 \subseteq T \subseteq T + aT$, $(T + aT) + (T + aT) \subseteq (T + T) + a(T + T) \subseteq T + aT$ and $(T + aT)(T + aT) \subseteq TT + aTT + aTT + a^2TT \subseteq T + aT + aT + A^2T \subseteq T + a(T + T) + TT \subseteq T + aT + T \subseteq T + aT$ □

Theorem 1.2.9. *Let K be a field and $P \subseteq K$. Then the following are equivalent:*

(a) *P is an order of K* [→ 1.1.20].

(b) *P is a proper preorder of K* [→ 1.2.1] *such that $P \cup -P = K$.*

(c) *P is a maximal proper preorder of K.*

Proof. (a) $\Longrightarrow$ (b) 1.2.2(b)

(b) $\Longrightarrow$ (c) Suppose (b) holds and let T be a proper preorder of K with $P \subseteq T$. To show: $T \subseteq P$. To this end, let $a \in T$. If a was not in P, then $-a \in P \subseteq T$ (since $P \cup -P = K$) and therefore $a \in T \cap -T \overset{1.2.7}{=} \{0\}$ in contradiction to $0 = 0^2 \in P$.

(c) $\Longrightarrow$ (a) Suppose (c) holds. Because of 1.2.7, we have to show only $P \cup -P = K$. Assume $P \cup -P \neq K$. Choose then $a \in K$ such that $a \notin P$ and $-a \notin P$. Then $P + aP$ and $P - aP$ are preorders according to Lemma 1.2.8 and both contain P properly (note that $0, 1 \in P$). Because of the maximality of P none of $P + aP$ and $P - aP$ is proper, i.e., $-1 \in P + aP$ and $-1 \in P - aP$. Write $-1 = s + as'$ and $-1 = t - at'$ for certain $s, s', t, t' \in P$. Then $-as' = 1 + s$ and $at' = 1 + t$. It follows that $-a^2s't' = 1 + s + t + st$ and therefore $-1 = s + t + st + a^2s't' \in P + P + PP + A^2PP \subseteq P$ ↯. □

Theorem 1.2.10. *Let K be a field and $T \subseteq K$ a proper preorder. Then there is an order P of K such that $T \subseteq P$ and we have $T = \bigcap\{P \mid P \text{ order of } K, T \subseteq P\}$.*

Proof. Consider the partially ordered set of all proper preorders of K containing T. In this partially ordered set, every chain has an upper bound (the empty chain has T as an upper bound and every nonempty chain possesses its union as an upper bound). By Zorn's lemma, the partially ordered set has a maximal element. Every such element is obviously a maximal proper preorder and therefore by 1.2.9 an order. Now we turn to the second statement:

"$\subseteq$" is clear.

"$\supseteq$" Let $a \in K \setminus T$. To show: There is an order P of K with $T \subseteq P$ and $a \notin P$. By 1.2.8, $T - aT$ is a preorder. It is proper for otherwise there would be $s, t \in T$ with $-1 = s - at$ and it would follow that $t \neq 0$, $at = 1 + s$ and $a = \left(\frac{1}{t}\right)^2 t(1 + s) \in K^2TT \subseteq T$. By the already proved, there is an order P of K with $T - aT \subseteq P$. If a lied in P, then $a \in P \cap -P = \{0\}$ in contradiction to $a \notin T$. □

Definition 1.2.11. A field is called *real* (in older literature mostly *formally real*) if it admits an order.

Theorem 1.2.12. *Let K be a field. Then the following are equivalent:*

(a) *K is real.*

(b) $-1 \notin \sum K^2$

(c) $\forall n \in \mathbb{N} : \forall a_1, \ldots, a_n \in K : (a_1^2 + \cdots + a_n^2 = 0 \implies a_1 = 0)$

Proof. (a) $\implies$ (b) follows from 1.1.6.

(b) $\implies$ (a) By 1.2.2, $\sum K^2$ is a preorder. If it is proper, then it is contained in an order by 1.2.10.

(b) $\iff$ (c)

$$\begin{aligned} -1 \in \sum K^2 &\iff \exists n \in \mathbb{N} : \exists a_2, \ldots, a_n \in K : -1 = a_2^2 + \cdots + a_n^2 \\ &\iff \exists n \in \mathbb{N} : \exists a_2, \ldots, a_n \in K : 1^2 + a_2^2 + \cdots + a_n^2 = 0 \\ &\iff \exists n \in \mathbb{N} : \exists a_1 \in K^\times : \exists a_2, \ldots, a_n \in K : a_1^2 + a_2^2 + \cdots + a_n^2 = 0 \end{aligned}$$

□

Example 1.2.13. Because of $-1 = \mathrm{i}^2 \in \sum \mathbb{C}^2$, the field $\mathbb{C} := \mathbb{R}(\mathrm{i})$ does not admit an ordering.

1.3 Extensions of orders

Definition 1.3.1. Let (K, P) be an ordered field and L an extension field of K (or in other words: let $L|K$ be a field extension and P be an order of K). We call Q an *extension* of the order P to L if the following equivalent conditions are fulfilled [→ 1.1.20(b)]:

(a) (L, Q) is an ordered extension field of (K, P).

(b) Q is an order of L such that $P \subseteq Q$.

(c) Q is an order of L such that $Q \cap K = P$.

Theorem 1.3.2. *Let (K, P) be an ordered field and L an extension field of K. Then the order P of K can be extended to L if and only if $-1 \notin \sum L^2 P$.*

Proof. Since every order is a preorder [→ 1.2.2], an order of L contains P if and only if it contains the preorder generated in L by P (i.e., the smallest preorder of L containing P, or in other words, the intersection of all preorders of L containing P), namely $\sum L^2 P$. If $\sum L^2 P$ is not proper, then there is of course no order of L containing it. On the contrary, if $\sum L^2 P$ is proper, then there is such an order by Theorem 1.2.10. □

Reminder 1.3.3. Let $L|K$ be a field extension with $\operatorname{char} K \neq 2$. Then

$$[L : K] \leq 2 \iff \exists d \in K : L = K(\sqrt{d})$$

since for $x \in L$ and $a, b, c \in K$ with $a \neq 0$ and $ax^2 + bx + c = 0$ we have $(2ax + b)^2 = 4a(ax^2 + bx) + b^2 = b^2 - 4ac =: d$ and therefore $K(x) = K(2ax + b) = K(\sqrt{d})$.

Theorem 1.3.4. *Let (K, P) be an ordered field and $d \in K$. The order P can be extended to $K(\sqrt{d})$ if and only if $d \in P$.*

Proof. If $\sqrt{d} \in K$, then $d = (\sqrt{d})^2 \in P$. Suppose now that $\sqrt{d} \notin K$. Because of $P + dP \subseteq \sum L^2 P \subseteq P + dP + K\sqrt{d}$, we have $-1 \notin \sum L^2 P \iff -1 \notin P + dP$. Since P is a maximal proper preorder by 1.2.9 and $P + dP$ is a preorder by 1.2.8, we obtain $-1 \notin P + dP \iff P = P + dP \iff d \in P$. Combining, we get $-1 \notin \sum L^2 P \iff d \in P$ and we conclude by Theorem 1.3.2. □

Example 1.3.5. In 1.3.3, the extension is in general not unique: $\mathbb{Q}(\sqrt{2})$ admits exactly two orders, namely the ones induced by the two field embeddings $\mathbb{Q}(\sqrt{2}) \hookrightarrow \mathbb{R}$ (in the one $\sqrt{2}$ is positive, in the other negative). That it does not admit a third one, follows from the fact that for every order P of $\mathbb{Q}(\sqrt{2})$ we have by 1.1.16 $(\mathbb{Q}(\sqrt{2}), P) \hookrightarrow (\mathbb{R}, \mathbb{R}_{\geq 0})$ because P is Archimedean since $\mathbb{Q}(\sqrt{2}) = \mathbb{Q} + \mathbb{Q}\sqrt{2}$ and

$$|\sqrt{2}|_P - 1 \overset{1.2.3}{=} \left(\frac{|\sqrt{2}|_P}{2}\right)^2 - \left(\frac{|\sqrt{2}|_P - 2}{2}\right)^2 \overset{1.1.6}{\leq}_P \left(\frac{|\sqrt{2}|_P}{2}\right)^2 = \frac{1}{2}$$

[→ 1.1.9(a)].

Theorem 1.3.6. *If $L|K$ is a field extension of odd degree, then each order of K can be extended to L.*

Proof. Assume the claim does not hold. Then there is a counterexample $L|K$ with $[L : K] = 2n + 1$ for some $n \in \mathbb{N}$. We choose the counterexample in a way such that n is as small as possible. We will now produce another counterexample $L'|K$ with $[L' : K] \leq 2n - 1$ which will contradict the minimality of n. Due to $\operatorname{char} K = 0$, we have that $L|K$ is separable. By the primitive element theorem, there is some $a \in L$ with $L = K(a) = K[a]$. The condition $-1 \in \sum L^2 P$ which is satisfied by 1.3.2 translates via the isomorphism $K[X]/(f) \to L$, $\overline{g} \mapsto g(a)$ in

$$1 + \sum_{i=1}^{\ell} a_i g_i^2 = hf \tag{$*$}$$

with $\ell \in \mathbb{N}$, $a_i \in P$, $g_i, h \in K[X]$, where f denotes the minimal polynomial of a over K (in particular $\deg f = [K(a) : K] = [L : K] = 2n + 1$) and the g_i are chosen in such a way that $\deg g_i \leq 2n$. Each of the $\ell + 1$ terms in the sum on the left hand side of $(*)$ has an *even* degree $\leq 4n$ and a leading coefficient from $PK^2 \subseteq P$ (except those terms that are zero of course). Since P is an order, the monomials of highest degree appearing on the left hand side of $(*)$ cannot cancel out. So the left hand side and therefore also the right hand side of $(*)$ has an *even* degree $\leq 4n$. It follows that h has an *odd* degree $\leq 2n - 1$. Choose an irreducible divisor h_1 of h in $K[X]$ of *odd* degree and a root b of h_1 in an extension field of K (e.g., in the splitting field of h_1 over K or in the algebraic closure of K). Set $L' := K(b)$. Substituting b in $(*)$ yields $-1 = \sum_{i=1}^{\ell} a_i g_i(b)^2 \in \sum PL'^2$. By 1.3.2, P cannot be extended to L'. Since $[L' : K] = [K(b) : K] = \deg \operatorname{irr}_K b = \deg h_1 \leq 2n - 1$ is *odd*, we gain the desired still smaller counterexample. □

Theorem 1.3.7. *Let K be a field. Then every order of K can be extended to $K(X)$.*

Proof. Let P be an order of K. Assume that P cannot be extended to $K(X)$. By 1.3.2 we then have $-1 \in \sum K(X)^2 P$. Because of $\#K = \infty$ [→ 1.1.7] we can plug in a suitable point from K ("avoid finitely many poles") and get $-1 \in \sum K^2 P = P$ ↯. □

Example 1.3.8. Due to 1.3.7 there is an order on $\mathbb{R}(X)$. If P is such an order, then by the completeness of $(\mathbb{R}, \leq)$ [→ 1.1.16], the set $\mathbb{R}_{\leq_P X} = \{a \in \mathbb{R} \mid a \leq_P X\}$ is either empty or not bounded from above (in which case it is $\mathbb{R}$) or it has a supremum t in $\mathbb{R}$ (in which case it equals $(-\infty, t)$ if $t >_P X$ or $(-\infty, t]$ if $t <_P X$). Hence

$$\mathbb{R}_{\leq_P X} = \{a \in \mathbb{R} \mid a \leq_P X\} \in \{\emptyset\} \cup \{(-\infty, t) \mid t \in \mathbb{R}\} \cup \{(-\infty, t] \mid t \in \mathbb{R}\} \cup \{\mathbb{R}\} =: C.$$

We claim now that the map

$$\begin{aligned} \Phi\colon \{P \mid P \text{ order of } \mathbb{R}(X)\} &\to C \\ P &\mapsto \mathbb{R}_{\leq_P X} \end{aligned}$$

is a bijection. It is easy to see that for all $I, J \in C$ there is a ring automorphism $\varphi_{I,J}$ of $\mathbb{R}(X)$ such that for all orders P of $\mathbb{R}(X)$, we have

$$\Phi(P) = I \iff \Phi(\varphi_{I,J}(P)) = J :$$

- $I = \mathbb{R}$ & $J = (-\infty, 0] \rightsquigarrow \varphi_{I,J}\colon X \mapsto \frac{1}{X}$
- $I = \emptyset$ & $J = (-\infty, 0) \rightsquigarrow \varphi_{I,J}\colon X \mapsto \frac{1}{X}$
- $I = (-\infty, t)$ & $J = (-\infty, 0) \rightsquigarrow \varphi_{I,J}\colon X \mapsto X + t$
- $I = (-\infty, t]$ & $J = (-\infty, 0] \rightsquigarrow \varphi_{I,J}\colon X \mapsto X + t$
- $I = (-\infty, 0)$ & $J = (-\infty, 0] \rightsquigarrow \varphi_{I,J}\colon X \mapsto -X$
- other I and $J \rightsquigarrow$ composition of the above automorphisms

From this we get the *surjectivity* of Φ, since as already mentioned there is an order P of $\mathbb{R}(X)$ and if we set $I := \Phi(P)$, then $\Phi(\varphi_{I,J}(P)) = J$ for all $J \in C$. For the *injectivity* of Φ, it suffices to show that *there is* some $I \in C$ having only one preimage under Φ since then

$$\begin{aligned} \#\{P \mid \Phi(P) = J\} &= \#\{P \mid \Phi(\varphi_{J,I}(P)) = I\} \\ &= \#\{\varphi_{I,J}(P) \mid \Phi(\varphi_{J,I}(P)) = I\} = \#\{P \mid \Phi(P) = I\} = 1 \end{aligned}$$

for all $J \in C$. We therefore fix $I := \mathbb{R} \in C$ and show that there at most (and therefore exactly) one order P of $\mathbb{R}(X)$ such that $\Phi(P) = I$. To this end, suppose $\Phi(P) = I$. If $f, g \in \mathbb{R}[X] \setminus \{0\}$, then one easily verifies that

$$\frac{f}{g} \in P \overset{\mathbb{R}(X)^2 \subseteq P}{\iff} fg \in P \iff \text{the leading coefficient of } fg \text{ is positive.}$$

This uniquely determines P. Consequently, Φ is a bijection. We fix the following notation:

$$\begin{aligned} P_{-\infty} &:= \Phi^{-1}(\emptyset) \\ P_{t-} &:= \Phi^{-1}((-\infty, t)) \text{ for } t \in \mathbb{R} \\ P_{t+} &:= \Phi^{-1}((-\infty, t]) \text{ for } t \in \mathbb{R} \\ P_{\infty} &:= \Phi^{-1}(\mathbb{R}) \end{aligned}$$

Now $\{P \mid P \text{ order of } \mathbb{R}(X)\} = \{P_{-\infty}\} \cup \{P_{t-}, P_{t+} \mid t \in \mathbb{R}\} \cup \{P_{\infty}\}$. By easy considerations one obtains,

$$\begin{aligned} P_{-\infty} &= \{r \in \mathbb{R}(X) \mid \exists c \in \mathbb{R} : \forall x \in (-\infty, c) : r(x) \geq 0\}, \\ P_{t-} &= \{r \in \mathbb{R}(X) \mid \exists \varepsilon \in \mathbb{R}_{>0} : \forall x \in (t-\varepsilon, t) : r(x) \geq 0\} \qquad (t \in \mathbb{R}), \\ P_{t+} &= \{r \in \mathbb{R}(X) \mid \exists \varepsilon \in \mathbb{R}_{>0} : \forall x \in (t, t+\varepsilon) : r(x) \geq 0\} \qquad (t \in \mathbb{R}), \\ P_{\infty} &= \{r \in \mathbb{R}(X) \mid \exists c \in \mathbb{R} : \forall x \in (c, \infty) : r(x) \geq 0\}. \end{aligned}$$

None of these orders is Archimedean.

1.4 Real closed fields

Proposition 1.4.1. *Let K be a field. Then the following are equivalent:*

(a) *K admits exactly one order.*

(b) *$\sum K^2$ is an order of K.*

(c) *$(\sum K^2) \cup (-\sum K^2) = K$ and $-1 \notin \sum K^2$*

Proof. (a) $\Longrightarrow$ (b) Suppose P is the unique order of K. By 1.2.2 and 1.2.10, we then get $\sum K^2 = P$.

(b) $\Longrightarrow$ (c) is trivial.

(c) $\Longrightarrow$ (a) Suppose (c) holds. Using 1.1.20(a) and 1.2.7, we see that $\sum K^2$ is an order of K, and it is the only one by 1.2.2 and 1.2.9(c). □

Example 1.4.2. $\mathbb{Q}$ and $\mathbb{R}$ possess exactly one order.

Convention 1.4.3. If K is a field admitting exactly one order, then we will often understand K as an ordered field, that is we speak of K but mean $(K, \sum K^2)$.

Definition 1.4.4. A field K is called *Euclidean* if K^2 is an order of K.

Remark 1.4.5. If K is Euclidean, then K^2 is the unique order of K.

Example 1.4.6. $\mathbb{R}$ is Euclidean but not $\mathbb{Q}$.

Notation and Remark 1.4.7. (a) Let $(K, \leq)$ be an ordered field. If $a, b \in K$ such that $a = b^2$, then we write $\sqrt{a} := |b| \in K_{\geq 0}$ [→ 1.1.8] (this is obviously well-defined). If $a \in K \setminus K^2$, we continue to denote by $\sqrt{a} \in \overline{K} \setminus K$ an arbitrary but fixed square root of a in the algebraic closure $\overline{K}$ of K. One shows easily that $a \leq b \iff \sqrt{a} \leq \sqrt{b}$ for all $a, b \in K^2$.

(b) If K is an Euclidean field (with order $\leq$ [→ 1.4.3, 1.4.5]), then in particular $\sqrt{a} \in K_{\geq 0}$ and $(\sqrt{a})^2 = a$ for *all* $a \in K_{\geq 0} = K^2 = \sum K^2$.

(c) We write $\mathbb{i} := \sqrt{-1}$. If K is a real field, then $K(\mathbb{i}) = K \oplus K\mathbb{i}$ as a K-vector space

Proposition 1.4.8. *Let K be a real field. Then the following are equivalent:*

(a) *K is Euclidean.*

(b) $K = -K^2 \cup K^2$

(c) $K(\mathbb{i}) = K(\mathbb{i})^2$

(d) *Every polynomial of degree 2 in $K(\mathbb{i})[X]$ has a root in $K(\mathbb{i})$.*

Proof. (d) $\Longrightarrow$ (c) is trivial.

(c) $\Longrightarrow$ (b) Suppose (c) holds and let $a \in K$. Write $a = (b + \mathbb{i}c)^2$ for some $b, c \in K$. Then $a = b^2 - c^2$ and $bc = 0$ [→ 1.4.7(c)]. Therefore $a = b^2$ or $a = -c^2$.

(b) $\Longrightarrow$ (a) Suppose (b) holds. It suffices to show $K^2 + K^2 \subseteq K^2$. For this purpose, let $a, b \in K$. To show: $a^2 + b^2 \in K^2$. If we had $a^2 + b^2 \notin K^2$, then $a^2 + b^2 \in -K^2$, say $a^2 + b^2 + c^2 = 0$ for some $c \in K$ and 1.2.12(c) would imply $c = 0$ ↯.

(a) $\Longrightarrow$ (c) Suppose (a) holds and let $a, b \in K$. By 1.4.7(c), we have to show $a + b\mathbb{i} \in K(\mathbb{i})^2$. Set $r := \sqrt{a^2 + b^2} \in K_{\geq 0}$ according to 1.4.7(b). Then $r^2 = a^2 + b^2 \geq a^2 = |a|^2$ and therefore $r \geq |a|$ by 1.4.7(a), i.e., $r \pm a \geq 0$. It follows that $\sqrt{\frac{r+a}{2}}, \sqrt{\frac{r-a}{2}} \in K_{\geq 0}$ and the calculation

$$\left(\sqrt{\frac{r+a}{2}} \pm \sqrt{\frac{r-a}{2}}\mathbb{i}\right)^2 = \frac{r+a}{2} \pm 2\sqrt{\frac{r^2-a^2}{2}}\mathbb{i} - \frac{r-a}{2} = a \pm 2\left|\frac{b}{2}\right|\mathbb{i} = a \pm |b|\mathbb{i}$$

shows $a + b\mathbb{i} \in K(\mathbb{i})^2$.

(c) $\Longrightarrow$ (d) follows from $X^2 + bX + c = (X + \frac{b}{2})^2 + (c - \frac{b^2}{4})$ for $b, c \in K(\mathbb{i})$. □

Definition 1.4.9. Let R be a field. Then R is called *real closed* if R is Euclidean [→ 1.4.4, 1.4.8] and every polynomial of *odd* degree from $R[X]$ has a root in R.

Example 1.4.10. $\mathbb{R}$ is real closed by the intermediate value theorem from calculus and by 1.4.4.

Remark 1.4.11. We now generalize the fundamental theorem of algebra from $\mathbb{C} = \mathbb{R}(\mathbb{i})$ to $R(\mathbb{i})$ for any real closed field R. The usual Galois theoretic proof goes through as we will see immediately.

Theorem 1.4.12 ("generalized fundamental theorem of algebra")**.** *Let R be a real closed field. Then $C := R(\mathbb{i})$ is algebraically closed.*

Proof. Let $z \in \overline{C}$. To show: $z \in C$. Choose an intermediate field L of $\overline{C}|C$ with $z \in L$ such that $L|R$ is a finite Galois extension (e.g., the splitting field of $(X^2+1)\operatorname{irr}_R z$ over R). We show $L = C$. Choose a 2-Sylow subgroup H of the Galois group $G := \operatorname{Aut}(L|R)$. From Galois theory we know that $[L^H : R] = [G : H]$ is odd. Hence $L^H = R$ since every element of L^H has over R a minimal polynomial of odd degree which has a root in R and therefore must be linear. Galois theory then implies $G = H$. Hence G is a 2-group. Therefore the subgroup $I := \operatorname{Aut}(L|C)$ of G is also a 2-group. By Galois theory, it is enough to show $I = \{1\}$. If we had $I \neq \{1\}$, then there would exist, as one knows from algebra, a subgroup J of I with $[I : J] = 2$. From this we get with Galois theory $[L^J : C] = [L^J : L^I] = [I : J] = 2$, contradicting 1.4.8(d). □

Theorem 1.4.13. *Let R be a field. Then the following are equivalent:*

(a) *R is real closed.*

(b) *$R \neq R(\mathbb{i})$ and $R(\mathbb{i})$ is algebraically closed.*

(c) *R is real but there is no real extension field $L \neq R$ of R such that $L|R$ is algebraic.*

Proof. (a) $\Longrightarrow$ (b) follows from 1.4.12.

(b) $\Longrightarrow$ (c) Suppose (b) holds. In order to show that R is real, it is enough to show by Theorem 1.2.12 that $\sum R^2 = R^2$ since $-1 \notin R^2$ because $R \neq R(\mathbb{i})$. To this end, let $a, b \in R$. To show: $a^2 + b^2 \in R^2$. Since $R(\mathbb{i})$ is algebraically closed, we have $a + b\mathbb{i} \in R(\mathbb{i})^2$, that is there are $c, d \in R$ such that $a + b\mathbb{i} = (c + d\mathbb{i})^2$ and it follows that $a^2 + b^2 = (a + b\mathbb{i})(a - b\mathbb{i}) = (c + d\mathbb{i})^2(c - d\mathbb{i})^2 = ((c + d\mathbb{i})(c - d\mathbb{i}))^2 = (c^2 + d^2)^2 \in R^2$. Now let $L|R$ be an algebraic field extension and suppose L is real. To show: $L = R$. Since $L(\mathbb{i})|R(\mathbb{i})$ is again algebraic and $R(\mathbb{i})$ is algebraically closed, we obtain $L(\mathbb{i}) = R(\mathbb{i})$. For this reason L is a real intermediate field of $R(\mathbb{i})|R$ and it follows that $L = R$.

(c) $\Longrightarrow$ (a) Suppose (c) holds. Choose an order P of R according to Definition 1.2.11. For all $d \in P$, $R(\sqrt{d})$ is real by 1.3.4 and therefore $R(\sqrt{d}) = R$. It follows that $P \subseteq R^2 \subseteq P$ and hence $P = R^2$, i.e., R is Euclidean. According to Definition 1.4.9 it remains to show that each polynomial $f \in R[X]$ of odd degree has a root in R. Let $f \in R[X]$ be of odd degree. Choose an irreducible divisor g of f in $R[X]$ of odd degree. Choose a root a of g in an extension field of R. Since $[R(a) : R] = \deg g$ is odd, $R(a)$ is real by 1.3.6 and therefore $R(a) = R$. Thus $a \in R$ satisfies $g(a) = 0$ and hence $f(a) = 0$. □

Theorem 1.4.14 ("real version of the generalized fundamental theorem of algebra")**.** *Let R be a field. Then the following are equivalent:*

(a) *R is real closed.*

(b) $\{f \in R[X] \mid f \text{ is irreducible and monic}\} =$
$\{X - a \mid a \in R\} \cup \{(X - a)^2 + b^2 \mid a, b \in R, b \neq 0\}$

Proof. (a) $\Longrightarrow$ (b) Suppose (a) holds.
"$\supseteq$" is clear since R is real.
"$\subseteq$" Let $f \in R[X]$ be irreducible and monic of degree ≥ 2. Since $R(\mathbb{i})$ is algebraically closed by 1.4.12, there are $a, b \in R$ such that $f(a+b\mathbb{i}) = 0$. Due to $R \neq R(\mathbb{i})$ we can apply the automorphism of the field extension $R(\mathbb{i})|R$ given by $\mathbb{i} \mapsto -\mathbb{i}$ in order to obtain $f(a-b\mathbb{i}) = 0$. By observing $a+b\mathbb{i} \neq a-b\mathbb{i}$ (since $b \neq 0$ because $f \in R[X]$ is irreducible of degree ≥ 2), we get

$$f = \underbrace{(X-(a+b\mathbb{i}))(X-(a-b\mathbb{i}))}_{(X-a)^2+b^2 \in R[X]} g$$

for some $g \in R(\mathbb{i})[X]$. But then $g \in R[X]$ and hence even $g = 1$.

(b) $\Longrightarrow$ (a) Suppose (b) holds. We will show 1.4.13(b), i.e., that $R \neq R(\mathbb{i})$ and $R(\mathbb{i})$ is algebraically closed. The first claim $R \neq R(\mathbb{i})$ follows from the irreducibility of $X^2+1 = (X-0)^2+1^2 \in R[X]$. Now suppose $f \in R(\mathbb{i})[X]$ is of degree ≥ 1. Consider the ring automorphism

$$R(\mathbb{i})[X] \to R(\mathbb{i})[X],\ p \mapsto p^*$$

given by $a^* = a$ for $a \in R$, $\mathbb{i}^* = -\mathbb{i}$ and $X^* = X$. Then $f^*f \in R[X]$. If f^*f has a root $a \in R$, then $f(a) = 0$ or $f^*(a) = 0$ and then again $f(a) = 0$. Suppose therefore that f^*f has no root in R. Then there exist $a, b \in R$ with $b \neq 0$ such that $(X-a)^2+b^2$ divides f^*f in $R[X]$. Because of $(X-a)^2+b^2 = (X-(a+b\mathbb{i}))(X-(a-b\mathbb{i}))$, $a+b\mathbb{i}$ is a root of f or of f^*. If $f^*(a+b\mathbb{i}) = 0$, then $f(a-\mathbb{i}b) = f^{**}((a+\mathbb{i}b)^*) = (f^*(a+\mathbb{i}b))^* = 0^* = 0$. Therefore $a+\mathbb{i}b$ or $a-\mathbb{i}b$ is a root of f in $R(\mathbb{i})$. □

Notation and Terminology 1.4.15. Let $(K, \leq)$ be an ordered field.

(a) We extend the order $\leq$ in the obvious way to the set $\{-\infty\} \cup K \cup \{\infty\}$ by declaring $-\infty < a < \infty$ for all $a \in K$.

(b) We adopt the usual notation for intervals

$$(a,b) := (a,b)_K := \{x \in K \cup \{\pm\infty\} \mid a < x < b\} \qquad (a, b \in K \cup \{\pm\infty\})$$
("interval from a to b without endpoints")
$$[a,b) := [a,b)_K := \{x \in K \cup \{\pm\infty\} \mid a \leq x < b\} \qquad (a, b \in K \cup \{\pm\infty\})$$
("interval from a to b with a and without b")

and so forth.

(c) We use terminology like

$$f \geq 0 \text{ on } S :\Longleftrightarrow \forall x \in S : f(x) \geq 0 \qquad (f \in K[X_1, \ldots, X_n], S \subseteq K^n)$$
("f is nonnegative on S")
$$f > 0 \text{ on } S :\Longleftrightarrow \forall x \in S : f(x) > 0 \qquad (f \in K[X_1, \ldots, X_n], S \subseteq K^n)$$
("f is positive on S").

Corollary 1.4.16 ("intermediate value theorem for polynomials"). *Let R be a real closed field, $f \in R[X]$ and $a, b \in R$ such that $a \leq b$ and $\operatorname{sgn}(f(a)) \neq \operatorname{sgn}(f(b))$. Then there is $c \in [a,b]_R$ with $f(c) = 0$.*

Proof. WLOG f is monic. By 1.4.14, all nonlinear monic irreducible polynomials in $R[X]$ are positive on R. Hence $f = g\prod_{i=1}^{k}(X-a_i)^{\alpha_i}$ with $k \in \mathbb{N}_0$, $a_i \in R$, $\alpha_i \in \mathbb{N}$, $a_1 < \cdots < a_k$ and some $g \in R[X]$ that is positive on R. On the sets $(-\infty, a_1)$, (a_1, a_2), $\ldots$, (a_{k-1}, a_k) and (a_k, ∞) each $X - a_i$ and therefore f has constant sign. Hence a and b cannot lie both in the same such set. Consequently, there is an $i \in \{1, \ldots, m\}$ such that $a_i \in [a,b]$. Set $c := a_i$. □

Corollary 1.4.17 ("Rolle's theorem for polynomials"). *Suppose R is a real closed field, $f \in R[X]$ and $a, b \in R$ with $a < b$ and $f(a) = f(b)$. Then there exists a $c \in (a,b)_R$ such that $f'(c) = 0$.*

Proof. WLOG $f \neq 0$, $f(a) = 0 = f(b)$ and $\nexists x \in (a,b) : f(x) = 0$. Write

$$f = (X-a)^{\alpha}(X-b)^{\beta}g$$

for some $\alpha, \beta \in \mathbb{N}$ and $g \in R[X]$ with $\forall x \in [a,b] : g(x) \neq 0$. We find

$$\begin{aligned} f' &= (X-a)^{\alpha}\beta(X-b)^{\beta-1}g + \alpha(X-a)^{\alpha-1}(X-b)^{\beta}g + (X-a)^{\alpha}(X-b)^{\beta}g' \\ &= (X-a)^{\alpha-1}(X-b)^{\beta-1}h \end{aligned}$$

where $h := \beta(X-a)g + \alpha(X-b)g + (X-a)(X-b)g'$. Hence it suffices to find $c \in (a,b)$ such that $h(c) = 0$. We can apply the intermediate value theorem 1.4.16 because

$$h(a) = \alpha(a-b)g(a) \quad \text{and} \quad h(b) = \beta(b-a)g(b)$$

and again by 1.4.16 we have $\operatorname{sgn}(g(a)) = \operatorname{sgn}(g(b))$. □

Corollary 1.4.18 ("mean value theorem for polynomials"). *Let R be a real closed field, $f \in R[X]$ and $a, b \in R$ with $a < b$. Then there is some $c \in (a,b)_R$ satisfying $f'(c) = \frac{f(b)-f(a)}{b-a}$.*

Proof. Setting $g := (X-a)(f(b)-f(a)) - (b-a)(f-f(a))$, we get $g(a) = 0 = g(b)$ and $g' = f(b) - f(a) - (b-a)f'$. Rolle's theorem 1.4.17 yields $c \in (a,b)$ such that $g'(c) = 0$. □

Definition 1.4.19. (a) Let $(M, \leq_1)$ and $(N, \leq_2)$ be ordered sets. A map $\varphi\colon M \to N$ is called *anti-monotonic* [→ 1.1.5] if

$$a \leq_1 b \implies \varphi(a) \geq_2 \varphi(b)$$

for all $a, b \in M$.

(b) If $(K,\leq)$ is an ordered field, $f\in K[X]$ and $I\subseteq K$, then we say that f is $\left\{\begin{array}{c}\text{monotonic}\\ \text{injective}\\ \text{anti-monotonic}\end{array}\right\}$ on I if $I\to K,\ x\mapsto f(x)$ is $\left\{\begin{array}{c}\text{monotonic}\\ \text{injective}\\ \text{anti-monotonic}\end{array}\right\}$.

Corollary 1.4.20. *Let R be a real closed field, $f\in R[X]$ and $a,b\in R$. If $\left\{\begin{array}{c}f'\geq 0\\ f'\leq 0\end{array}\right\}$ on (a,b) [→ 1.4.15(c)], then f is $\left\{\begin{array}{c}\\ \text{anti-}\end{array}\right\}$ monotonic on $[a,b]$. If f' has no root on (a,b), then f is injective on $[a,b]$.*

Proof. The statement is empty in case $a>b$, trivial in the case $a=b$ and it follows from the mean value theorem 1.4.18 in case $a<b$. □

1.5 Descartes' rule of signs

Notation 1.5.1. Let A be a commutative ring with $0\neq 1$ and $d\in\mathbb{R}$. We denote

$$A[X_1,\ldots,X_n]_d:=\{f\in A[X_1,\ldots,X_n]\mid \deg f\leq d\}$$

(where $\deg 0:=-\infty$).

Proposition 1.5.2 ("Taylor formula for polynomials")**.** *Suppose K is a field of characteristic 0, $d\in\mathbb{N}_0$, $f\in K[X]_d$ and $a\in K$. Then*

$$f=\sum_{k=0}^{d}\frac{f^{(k)}(a)}{k!}(X-a)^k.$$

Proof. Since $K[X]\to K[X],\ p\mapsto p'$ commutes with the ring automorphism $K[X]\to K[X],\ p\mapsto p(X+a)$, we can WLOG suppose $a=0$. But then the claim follows from the definition of the (formal) derivative. □

Lemma 1.5.3. Suppose $(K,\leq)$ is an ordered field, $k\in\mathbb{N}$, $c_1,\ldots,c_k\in K^\times$, $\alpha_1,\ldots,\alpha_k\in\mathbb{N}_0$, $\alpha_1<\ldots<\alpha_k$ and $f=\sum_{i=1}^k c_iX^{\alpha_i}$.

(a) $\operatorname{sgn}(f(x))=(\operatorname{sgn}x)^{\alpha_k}\operatorname{sgn}(c_k)$ for all $x\in K$ satisfying $|x|>\max\left\{1,\frac{|c_1|+\ldots+|c_{k-1}|}{|c_k|}\right\}$

(b) $\operatorname{sgn}(f(x))=(\operatorname{sgn}x)^{\alpha_1}\operatorname{sgn}(c_1)$ for all $x\in K^\times$ satisfying $\frac{1}{|x|}>\max\left\{1,\frac{|c_2|+\ldots+|c_k|}{|c_1|}\right\}$

Proof. (a) For all $x\in K$ with $|x|>\max\left\{1,\frac{|c_1|+\ldots+|c_{k-1}|}{|c_k|}\right\}$, we have

$$\left|\sum_{i=1}^{k-1}c_ix^{\alpha_i}\right|\overset{1.1.8}{\leq}\sum_{i=1}^{k-1}|c_i||x|^{\alpha_i}\overset{1\leq|x|}{\leq}\sum_{i=1}^{k-1}|c_i||x|^{\alpha_k-1}=|c_k|\left(\frac{\sum_{i=1}^{k-1}|c_i|}{|c_k|}\right)|x|^{\alpha_k-1}<|c_kx^{\alpha_k}|.$$

(b) For all $x \in K^\times$ with $\frac{1}{|x|} > \max\left\{1, \frac{|c_2|+\ldots+|c_k|}{|c_1|}\right\}$, we have

$$\left|\sum_{i=2}^{k} c_i x^{\alpha_i}\right| \overset{1.1.8}{\leq} \sum_{i=2}^{k} |c_i||x|^{\alpha_i} \overset{|x|\leq 1}{\leq} \sum_{i=2}^{k} |c_i||x|^{\alpha_1+1} = |c_1|\left(\frac{\sum_{i=2}^{k}|c_i|}{|c_1|}\right)|x|^{\alpha_1+1} < |c_1 x^{\alpha_1}|.$$

□

Reminder 1.5.4. Let K be a field, $f \in K[X]$ and $a \in K$. Then

$$\mu(a,f) := \sup\{k \in \mathbb{N}_0 \mid (X-a)^k \text{ divides } f \text{ in } K[X]\} \in \mathbb{N}_0 \cup \{\infty\}$$

is called the *multiplicity* of a in f. We have

$$\mu(a,f) = \infty \iff f = 0$$

and

$$\mu(a,f) \geq 1 \iff f(a) = 0.$$

We call a a *multiple* root of f if $\mu(a,f) \geq 2$ and we call it a k-fold root of f ($k \in \mathbb{N}$) if $\mu(a,f) = k$. In case $\operatorname{char} K = 0$, one has

$$\mu(a,f) = \sup\{k \in \mathbb{N}_0 \mid f^{(0)}(a) = \ldots = f^{(k-1)}(a) = 0\}$$

as one can see easily.

Definition 1.5.5. Let $(K, \leq)$ be an ordered field and $0 \neq f \in K[X]$.

(a) The *number of positive roots counted with multiplicity* of f is

$$\mu(f) := \sum_{a \in K_{>0}} \mu(a,f) \in \mathbb{N}_0.$$

Writing $f = g\prod_{i=1}^{m}(X-a_i)$ with $a_1,\ldots,a_m \in K_{>0}$ and $g \in K[X]$ with $g(x) \neq 0$ for all $x \in K_{>0}$, we therefore have $\mu(f) = m$.

(b) Writing $f = \sum_{i=1}^{k} c_i X^{\alpha_i}$ with $c_1,\ldots,c_k \in K^\times$ and $\alpha_1,\ldots,\alpha_k \in \mathbb{N}_0$ such that

$$\alpha_1 < \ldots < \alpha_k,$$

we define the *number of sign changes in the coefficients* of f

$$\sigma(f) := \#\{i \in \{1,\ldots,k-1\} \mid \operatorname{sgn}(c_i) \neq \operatorname{sgn}(c_{i+1})\} \in \mathbb{N}_0.$$

Proposition 1.5.6. *Let R be a real closed field and $f \in R[X] \setminus \{0\}$. Then $\mu(f)$ and $\sigma(f)$ have the same parity.*

Proof. Write $f = \sum_{i=1}^{k} c_i X^{\alpha_i} = g\prod_{i=1}^{m}(X-a_i)$ with $c_1,\ldots,c_k \in R^\times$, $\alpha_1,\ldots,\alpha_k \in \mathbb{N}_0$, $a_1,\ldots,a_m \in R_{>0}$ and $g \in R[X]$ such that $\alpha_1 < \ldots < \alpha_m$ and $g(x) \neq 0$ for all $x \in R_{>0}$. Since R is real closed, WLOG $g(x) > 0$ for all $x \in R_{>0}$ by the intermediate value theorem 1.4.16. But then by Lemma 1.5.3, both the lowest and highest coefficient of g is positive. Now the claim follows from $\mu(f) = m$, $\operatorname{sgn}(c_1) = (-1)^m$ and $\operatorname{sgn}(c_k) = 1$. □

Lemma 1.5.7. Let R be a real closed field and $f \in R[X] \setminus R$. Then $\mu(f) \leq \mu(f') + 1$ and $\sigma(f) \leq \sigma(f') + 1$.

Proof. The second statement is easy to prove. For the first statement, suppose $a_1, \ldots, a_m \in R$ are the positive roots of f and $a_1 < \ldots < a_m$. Since R is real closed, there exist roots $b_1, \ldots, b_{m-1} \in R$ of f' such that $a_1 < b_1 < a_2 < \ldots < b_{m-1} < a_m$ by Rolle's Theorem 1.4.17. Now $\mu(f') = \sum_{a \in K_{>0}} \mu(a, f') \geq \sum_{i=1}^{m} \mu(a_i, f') + \sum_{i=1}^{m-1} \mu(b_i, f') \geq \sum_{i=1}^{m} \mu(a_i, f') + m - 1 = \sum_{i=1}^{m} (\mu(a_i, f) - 1) + m - 1 = \sum_{i=1}^{m} \mu(a_i, f) - 1 = \mu(f) - 1$. □

Remark 1.5.8. In the situation of Lemma 1.5.7, $\sigma(f') \leq \sigma(f)$ holds trivially but $\mu(f') \leq \mu(f)$ fails in general as the example $f = (X-1)^2 + 1$ shows.

Theorem 1.5.9 (Descartes' rule of signs)**.** *Let R be a real closed field. Then $\mu(f) \leq \sigma(f)$ for all $f \in R[X] \setminus \{0\}$.*

Proof. Induction on $d := \deg f \in \mathbb{N}_0$.

$\underline{d = 0}$ $\quad \mu(f) = 0 = \sigma(f)$

$\underline{d - 1 \to d \quad (d \in \mathbb{N}_0)}$ $\quad \mu(f) \overset{1.5.7}{\leq} \mu(f') + 1 \underset{\text{hypothesis}}{\overset{\text{induction}}{\leq}} \sigma(f') + 1 \overset{1.5.8}{\leq} \sigma(f) + 1$ and

therefore $\mu(f) \leq \sigma(f)$ by Proposition 1.5.6. □

Example 1.5.10. Let R be a real closed field and $f := X^4 - 5X^3 - 21X^2 + 115X - 150 \in R[X]$. Then $\sigma(f) = 3$ and therefore $\mu(f) \in \{1, 3\}$ by 1.5.9 and 1.5.6. For $f(-X) = X^4 + 5X^3 - 21X^2 - 115X - 150$, we have $\sigma(f(-X)) = 1$ and therefore $\mu(f(-X)) = 1$. One can verify that $\mu((1+X)^{22} f) = 1$ from which we get $\mu(f) = \mu((1+X)^{22} f) = 1$. Hence f has exactly two roots in R, namely two simple (i.e., 1-fold [→ 1.5.4]) ones, one positive and one negative.

Definition 1.5.11. Let R be a real closed field. We call a polynomial $f \in R[X]$ *real-rooted* if it has no root in $R(\mathbb{i}) \setminus R$ [→ 1.4.12].

Proposition 1.5.12. *Let R be real closed field and $f \in R[X]$. Then the following are equivalent:*

(a) f is real-rooted.

(b) There are $d \in \mathbb{N}_0$, $c \in R^\times$ and $a_1, \ldots, a_d \in R$ such that $f = c \prod_{i=1}^{d} (X - a_i)$.

Proof. For (a) $\Longrightarrow$ (b) use the fundamental theorem 1.4.12 or 1.4.14. □

Theorem 1.5.13. [→ 1.5.8] *Suppose R is a real closed field and $f \in R[X] \setminus R$ is real-rooted. Then f' is real-rooted and $\mu(f') \leq \mu(f)$.*

Proof. Using 1.5.12, write $f = c \prod_{i=1}^{n} (X - a_i)^{\alpha_i}$ with $c, a_1, \ldots, a_n \in R$ and $\alpha_1, \ldots, \alpha_n \in \mathbb{N}$ such that $c \neq 0$ and

$$a_1 < \ldots < a_n.$$

Since R is real closed, there exist roots $b_1, \ldots, b_{n-1} \in R$ of f' such that

$$a_1 < b_1 < a_2 < \ldots < b_{n-1} < a_n$$

by Rolle's Theorem 1.4.17. We have $\mu(a_i, f) = \alpha_i$ and therefore

$$\mu(a_i, f') = \alpha_i - 1$$

for all $i \in \{1, \dots, n\}$. It follows that

$$\begin{aligned}\deg(f') &\geq \sum_{i=1}^{n} \mu(a_i, f') + \sum_{i=1}^{n-1} \mu(b_i, f') \geq \sum_{i=1}^{n} \mu(a_i, f') + n - 1 \\ &= \sum_{i=1}^{n} (\alpha_i - 1) + n - 1 = \deg(f) - 1 = \deg(f'),\end{aligned}$$

whence

$$\deg(f') = \sum_{i=1}^{n} \mu(a_i, f') + \sum_{i=1}^{n-1} \mu(b_i, f')$$

and

$$\mu(b_i, f') = 1$$

for all $i \in \{1, \dots, n-1\}$. It follows that

$$\{x \in R(\mathbb{i}) \mid f'(x) = 0\} \subseteq \{a_1, b_1, a_2, \dots, b_{n-1}, a_n\} \subseteq R,$$

in particular f' is real-rooted. Choose $k \in \{1, \dots, n+1\}$ such that $a_k, \dots, a_n$ are the positive roots of f'. Then

$$\{x \in R \mid f'(x) = 0, x > 0\} \begin{cases} \subseteq \{b_{k-1}, a_k, \dots, b_{n-1}, a_n\} & \text{if } k \geq 2 \\ = \{a_1, b_1, \dots, b_{n-1}, a_n\} & \text{if } k = 1 \end{cases}.$$

If $k \geq 2$, then

$$\mu(f') \leq \sum_{i=k}^{n} (\underbrace{\mu(b_{i-1}, f')}_{=1} + \underbrace{\mu(a_i, f')}_{=\mu(a_i, f)-1}) = \mu(f).$$

If $k = 1$, then one sees similarly that $\mu(f') = \mu(f) - 1 \leq \mu(f)$. □

Theorem 1.5.14 (Descartes' rule of signs for real-rooted polynomials)**.** *Let R be a real closed field. Then $\mu(f) = \sigma(f)$ for all real-rooted $f \in R[X]$.*

Proof. By Theorem 1.5.9, it is enough to show $\mu(f) \geq \sigma(f)$ for all real-rooted $f \in R[X]$ by induction on $d := \deg f \in \mathbb{N}_0$.

$\underline{d = 0}$ $\quad \mu(f) = 0 = \sigma(f)$

$\underline{d - 1 \to d \quad (d \in \mathbb{N}_0)}$ $\quad \mu(f) \overset{1.5.13}{\geq} \mu(f') \underset{1.5.13}{\overset{\text{induction hypothesis}}{\geq}} \sigma(f') \overset{1.5.7}{\geq} \sigma(f) - 1$ and therefore $\mu(f) \geq \sigma(f)$ by Proposition 1.5.6. □

Example 1.5.15.

$$\det\begin{pmatrix} 1-X & 0 & 1 \\ 0 & -2-X & 1 \\ 1 & 1 & -X \end{pmatrix} = (1-X)(2+X)X + 2 + X + X - 1$$
$$= (2 + X - 2X - X^2)X + 2X + 1 = -X^3 - X^2 + 4X + 1 \in \mathbb{R}[X]$$

is real-rooted since it is the characteristic polynomial of a symmetric matrix. By Descartes' rule 1.5.14, it has exactly one positive root.

1.6 Counting real zeros with Hermite's method

Reminder 1.6.1. (a) Let A be a commutative ring with $0 \neq 1$ and $f \in A[X_1, \ldots, X_n]$. Then f is called *homogeneous* if f is a an A-linear combination of monomials of the same degree. Moreover, f is called a *k-form* ($k \in \mathbb{N}_0$) if f is an A-linear combination of monomials of degree k (i.e., if $f = 0$ or f is homogeneous of degree k). One often says *linear form* instead of 1-form and *quadratic form* instead of 2-form.

(b) If K is a field, one can identify the K-vector subspace of $K[X_1, \ldots, X_n]$ consisting of the $\left\{\begin{smallmatrix}\text{linear}\\ \text{quadratic}\end{smallmatrix}\right\}$ forms introduced in (a) via the isomorphism $f \mapsto (x \mapsto f(x))$ with the K-vector space $\left\{\begin{smallmatrix}(K^n)^*\\ Q(K^n)\end{smallmatrix}\right\}$ introduced in linear algebra. Hence the notion of a linear or quadratic form introduced in (a) differs only insignificantly from the corresponding notion from linear algebra.

(c) Let A be a set and $M = (a_{ij})_{\substack{1\leq i\leq m\\ 1\leq j\leq n}} \in A^{m\times n}$ a matrix. Then $M^T := (a_{ji})_{\substack{1\leq j\leq n\\ 1\leq i\leq m}} \in A^{n\times m}$ is called the *transpose* of M. The elements of $SA^{n\times n} := \{M \in A^{n\times n} \mid M = M^T\}$ are called *symmetric* matrices.

(d) Let K be a field. Then $(a_1, \ldots, a_n) \mapsto a_1X_1 + \ldots + a_nX_n$ ($a_i \in K$) defines an isomorphism between $K^{1\times n} \cong K^n$ and the K-vector space of the linear forms in $K[X_1, \ldots, X_n]$. If $\operatorname{char} K \neq 2$, then $(a_{ij})_{1\leq i,j\leq n} \mapsto \sum_{i,j=1}^n a_{ij}X_iX_j$ ($a_{ij} \in K$) defines an isomorphism between $SK^{n\times n}$ and the K-vector space of the quadratic forms in $K[X_1, \ldots, X_n]$. If $f \in K[X_1, \ldots, X_n]$ is a linear of quadratic form, then we call the preimage $M(f)$ of f under the respective isomorphism the *representing matrix* of f. This is the representing matrix of f in the sense of linear algebra with respect to the canonical bases.

(e) Suppose K is a field satisfying $\operatorname{char} K \neq 2$, $q \in K[X_1, \ldots, X_n]$ a quadratic form,

$\ell_1, \dots, \ell_m \in K[X_1, \dots, X_n]$ linear forms and $\lambda_1, \dots, \lambda_m \in K$. Then

$$q = \sum_{k=1}^{m} \lambda_k \ell_k^2 \iff M(q) = P^T \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_m \end{pmatrix} P$$

where

$$P := \begin{pmatrix} M(\ell_1) \\ \vdots \\ M(\ell_m) \end{pmatrix} \in K^{m \times n}.$$

Here P is invertible if and only if $m = n$ and $\ell_1, \dots, \ell_m$ are linearly independent.

(f) Let K be a field satisfying $\operatorname{char} K \neq 2$ and $q \in K[X_1, \dots, X_n]$ a quadratic form. One can *easily* calculate linearly independent linear forms $\ell_1, \dots, \ell_m \in K[X_1, \dots, X_n]$ and $\lambda_1, \dots, \lambda_m \in K$ such that $q = \sum_{k=1}^{m} \lambda_k \ell_k^2$. Indeed, one can write

$$X_1^2 + a_2 X_1 X_2 + \dots + a_n X_1 X_n \qquad (a_i \in K)$$

as $\Big(\underbrace{X_1 + \frac{a_2}{2} X_2 + \dots + \frac{a_n}{2} X_n}_{\ell_1} \Big)^2 - \underbrace{\Big(\frac{a_2}{2} X_2 + \dots + \frac{a_n}{2} X_n \Big)^2}_{\in K[X_2, \dots, X_n]}$ and

$$X_1 X_2 + a_3 X_1 X_3 + \dots + a_n X_1 X_n + b_3 X_2 X_3 + \dots + b_n X_2 X_n$$

as

$$\big(\underbrace{X_1 + b_3 X_3 + \dots + b_n X_n}_{h_1} \big) \big(\underbrace{X_2 + a_3 X_3 + \dots + a_n X_n}_{h_2} \big) - \underbrace{(a_3 X_3 + \dots + a_n X_n)(b_3 X_3 + \dots + b_n X_n)}_{\in K[X_3, \dots, X_n]}$$

where $h_1 h_2 = \Big(\underbrace{\frac{h_1 + h_2}{2}}_{\ell_1} \Big)^2 - \Big(\underbrace{\frac{h_1 - h_2}{2}}_{\ell_2} \Big)^2$ [→ 1.2.3]. In this way one can in each step place one or two variables in one or two squares and the arising linear forms are obviously linearly independent. Consider $q := 2X_1X_2 + 2X_1X_3 + 2X_2X_3 + 2X_3X_4$ as an example:

$$\begin{aligned} q &:= 2(X_1X_2 + X_1X_3 + X_2X_3) + 2X_3X_4 \\ &= 2((\underbrace{X_1 + X_3}_{h_1})(\underbrace{X_2 + X_3}_{h_2})) - 2X_3^2 + 2X_3X_4 \\ &= \underbrace{\frac{1}{2}}_{\lambda_1 = -\lambda_2} ((\underbrace{h_1 + h_2}_{\ell_1})^2 - (\underbrace{h_1 - h_2}_{\ell_2})^2) \underbrace{-2}_{\lambda_3} \Big(\underbrace{X_3 - \frac{1}{2} X_4}_{\ell_3} \Big)^2 + \underbrace{\frac{2}{4}}_{\lambda_4} \underbrace{X_4}_{\ell_4}{}^2. \end{aligned}$$

Hence $q = \sum_{k=1}^{4} \lambda_k \ell_k^2 = \frac{1}{2}(X_1 + X_2 + 2X_3)^2 - \frac{1}{2}(X_1 - X_2)^2 - 2(X_3 - \frac{1}{2}X_4)^2 + \frac{1}{2}X_4^2$ and by (e)

$$\begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix} = P^T D P$$

where

$$P := \begin{pmatrix} 1 & 1 & 2 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -\frac{1}{2} \\ 0 & 0 & 0 & 1 \end{pmatrix} \qquad \text{and} \qquad D := \begin{pmatrix} \frac{1}{2} & 0 & 0 & 0 \\ 0 & -\frac{1}{2} & 0 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & \frac{1}{2} \end{pmatrix}.$$

(g) Translating (f) into the language of matrices, one obtains for each field K with $\operatorname{char} K \neq 2$ and each $M \in SK^{n\times n}$ the following: One can *easily* find a $P \in \mathrm{GL}_n(K) = (K^{n\times n})^\times$ and a diagonal matrix $D \in K^{n\times n}$ such that $M = \underline{\underline{P^T}} D P$. This is the diagonalization of M as a quadratic form which is much simpler than the diagonalization of M as an endomorphism where one wants to reach $M = \underline{\underline{P^{-1}}} D P$ (in case $K = \mathbb{R}$ perhaps even with $P^{-1} = P^T$).

(h) Let K be an Euclidean field [→ 1.4.4] and $q \in K[X_1, \ldots, X_n]$ a quadratic form. According to (f), one can then *easily* compute *linearly independent* linear forms

$$\ell_1, \ldots, \ell_s, \ell_{s+1}, \ldots, \ell_{s+t} \in K[X_1, \ldots, X_n]$$

satisfying $q = \sum_{i=1}^{s} \ell_i^2 - \sum_{j=1}^{t} \ell_{s+j}^2$. By completing $\ell_1, \ldots, \ell_{s+t}$ to a basis $\ell_1, \ldots, \ell_n$ of the vector space of all linear forms in $K[X_1, \ldots, X_n]$ and by writing $q = 1 \cdot \sum_{i=1}^{s} \ell_i^2 + (-1) \sum_{j=1}^{t} \ell_{s+j}^2 + 0 \cdot \sum_{k=t+1}^{n} \ell_k^2$, one sees for the *rank* $\operatorname{rk}(q) := \operatorname{rk} M(q)$ of q that $\operatorname{rk}(q) \overset{(e)}{=} s + t$. We define the *signature* of q as $\operatorname{sg}(q) := s - t$. This is well-defined by *Sylvester's law of inertia*: If $\ell'_1, \ldots, \ell'_{s'}, \ell'_{s'+1}, \ldots, \ell'_{s'+t'} \in K[X_1, \ldots, X_n]$ are other linearly independent linear forms satisfying $q = \sum_{i=1}^{s'} \ell_i'^2 - \sum_{j=1}^{t'} \ell_{s'+j}'^2$, then $s' + t' = \operatorname{rk}(q) = s + t$ and one sees again by completing to a basis and (e) that there are subspaces U, W, U', W' of K^n such that $q(U) \subseteq K_{\geq 0}$, $\dim U = n - t$, $q(W \setminus \{0\}) \subseteq K_{<0}$, $\dim W = t$, $q(U') \subseteq K_{\geq 0}$, $\dim U' = n - t'$, $q(W' \setminus \{0\}) \subseteq K_{<0}$, $\dim W' = t'$. One deduces $U \cap W' = \{0\}$ and $U' \cap W = \{0\}$, whence $(n - t) + t' \leq n$ and $(n - t') + t \leq n$. Therefore $t = t'$ and thus $s = s'$.

(i) Let K be a field and $f = X^d + a_{d-1}X^{d-1} + \ldots + a_0 \in K[X]$ with $d \in \mathbb{N}_0$ and $a_i \in K$. The *companion matrix* C_f of f is the representing matrix of the K-vector space endomorphism

$$K[X]/(f) \to K[X]/(f),\ \overline{p} \mapsto \overline{Xp} \qquad (p \in K[X])$$

with respect to the basis $\overline{1}, \dots, \overline{X^{d-1}}$, i.e.,

$$C_f = \begin{pmatrix} 0 & 0 & \cdots & 0 & -a_0 \\ 1 & 0 & \ddots & \vdots & -a_1 \\ 0 & 1 & \ddots & 0 & -a_2 \\ 0 & 0 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & & \\ 0 & 0 & \cdots & 1 & -a_{d-1} \end{pmatrix} \in K^{d\times d}.$$

One sees easily that f is the minimal polynomial and therefore for degree reasons also the characteristic polynomial of C_f. Now suppose furthermore that f splits into linear factors, i.e.,

$$f = \prod_{k=1}^{m} (X - x_k)^{\alpha_k}$$

for some $m \in \mathbb{N}_0$, $\alpha_k \in \mathbb{N}$ and pairwise distinct $x_1, \dots, x_m \in K$. Then C_f is similar to a triangular matrix with diagonal entries

$$\underbrace{x_1, \dots, x_1}_{\alpha_1}, \quad \underbrace{x_2, \dots, x_2}_{\alpha_2}, \quad \dots \quad , \quad \underbrace{x_m, \dots, x_m}_{\alpha_m}.$$

Then C_f^i is for every $i \in \mathbb{N}_0$ similar to a triangular matrix whose diagonal entries are

$$\underbrace{x_1^i, \dots, x_1^i}_{\alpha_1}, \quad \underbrace{x_2^i, \dots, x_2^i}_{\alpha_2}, \quad \dots \quad , \quad \underbrace{x_m^i, \dots, x_m^i}_{\alpha_m}.$$

In particular, we have $\operatorname{tr}(C_f^i) = \sum_{k=1}^{m} \alpha_k x_k^i$ for all $i \in \mathbb{N}_0$ and consequently

$$\operatorname{tr}(g(C_f)) = \sum_{k=1}^{m} \alpha_k g(x_k)$$

for all $g \in K[X]$.

(j) If K is a field and $x_1, \dots, x_m \in K$ are pairwise distinct, then the *Vandermonde* matrix

$$\begin{pmatrix} 1 & x_1 & \dots & x_1^{m-1} \\ \vdots & \vdots & & \vdots \\ 1 & x_m & \dots & x_m^{m-1} \end{pmatrix} \in K^{m\times m}$$

is invertible since it is the representing matrix of the injective and therefore bijective linear map

$$K[X]_{m-1} \to K^m,\ p \mapsto \begin{pmatrix} p(x_1) \\ \vdots \\ p(x_m) \end{pmatrix}$$

[→ 1.5.1] with respect to the canonical bases.

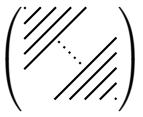
(k) Let K be a field and let $x_1, \dots, x_m \in K$ be pairwise distinct. Furthermore, let $d \in \mathbb{N}_0$ with $m \leq d$. Consider for $k \in \{1, \dots, m\}$ the linear forms $\ell_k := \sum_{i=1}^{d} x_k^{i-1} T_i \in K[T_1, \dots, T_d]$. Then $\ell_1, \dots, \ell_m$ are linearly independent. Indeed, because of (d) this is equivalent to the linear independence of the vectors $(x_k^0, \dots, x_k^{d-1})$ $(k \in \{1, \dots, m\})$ in K^d. But already the truncated vectors $(x_k^0, \dots, x_k^{m-1})$ $(k \in \{1, \dots, m\})$ are linearly independent by (j).

Definition 1.6.2. Let K be a field and $f, g \in K[X]$ where f is monic of degree d. Then the quadratic form

$$H(f,g) := \sum_{i,j=1}^{d} \operatorname{tr}(g(C_f) C_f^{i+j-2}) T_i T_j \in K[T_1, \dots, T_d]$$

is called the *Hermite form* of f with respect to g. The quadratic form $H(f) := H(f,1)$ is simply called the Hermite form of f.

Remark 1.6.3. Let K be a field with $\operatorname{char} K \neq 2$ and let $f, g \in K[X]$ where f is monic of degree d. Then $M(H(f,g))$ [→ 1.6.1(d)] is called the *Hermite matrix of f with respect to g*. This is a Hankel matrix, i.e., of the form



Furthermore, $M(H(f))$ is called the *Hermite matrix of f*.

Proposition 1.6.4. *Let K be a field and $f, g \in K[X]$. Suppose $x_1, \dots, x_m \in K$ and $\alpha_1, \dots, \alpha_m \in \mathbb{N}_0$ such that $f = \prod_{k=1}^{m} (X - x_k)^{\alpha_k}$ and $d := \deg f$. Then*

$$H(f,g) = \sum_{i,j=1}^{d} \left(\sum_{k=1}^{m} \alpha_k g(x_k) x_k^{i+j-2} \right) T_i T_j = \sum_{k=1}^{m} \alpha_k g(x_k) \left(\sum_{i=1}^{d} x_k^{i-1} T_i \right)^2.$$

Proof. 1.6.2 and 1.6.1(i). □

Theorem 1.6.5 (Counting roots with one side condition)**.** *Let R be a real closed field, $C := R(\mathbb{i})$, $f, g \in R[X]$ and f monic. Then*

$$\begin{aligned} \operatorname{rk} H(f,g) &= \#\{x \in C \mid f(x) = 0, g(x) \neq 0\} \qquad \text{and} \\ \operatorname{sg} H(f,g) &= \#\{x \in R \mid f(x) = 0, g(x) > 0\} \\ &\quad - \#\{x \in R \mid f(x) = 0, g(x) < 0\}. \end{aligned}$$

Proof. Denote by $p \mapsto p^*$ again the ring automorphism of $C[T_1, \dots, T_d]$ with $x^* = x$ for all $x \in \mathbb{R}$, $\mathbb{i}^* = -\mathbb{i}$ and $X^* = X$. Using the fundamental theorem of algebra 1.4.14 and this automorphism, we can write

$$f = \prod_{k=1}^{m} (X - x_k)^{\alpha_k} \prod_{t=1}^{n} (X - z_t)^{\beta_t} \prod_{t=1}^{n} (X - z_t^*)^{\beta_t}$$

for some $m, n \in \mathbb{N}_0$ $\alpha_k, \beta_t \in \mathbb{N}$, $x_k \in R$, $z_t \in C \setminus R$ and $x_1, \dots, x_m, z_1, \dots, z_n, z_1^*, \dots, z_n^*$ pairwise distinct. By renumbering the z_t, we can find $r \in \{0, \dots, n\}$ such that $g(z_1) \neq 0, \dots, g(z_r) \neq 0$ and $g(z_{r+1}) = 0, \dots, g(z_n) = 0$. By 1.6.4, 1.6.1(k) and 1.4.8(c), we obtain linear forms $\ell_1, \dots, \ell_m, g_1, \dots, g_r, h_1, \dots h_r \in R[T_1, \dots, T_d]$ such that

$$\begin{aligned} H(f,g) &= \sum_{k=1}^{m} \alpha_k g(x_k) \ell_k^2 + \sum_{t=1}^{r} (g_t + \mathbb{i} h_t)^2 + \sum_{t=1}^{r} (g_t - \mathbb{i} h_t)^2 \\ &= \sum_{k=1}^{m} \alpha_k g(x_k) \ell_k^2 + 2 \sum_{t=1}^{r} g_t^2 - 2 \sum_{t=1}^{r} h_t^2 \end{aligned}$$

where $\ell_1, \dots, \ell_m, g_1 + \mathbb{i} h_1, g_1 - \mathbb{i} h_1, \dots, g_r + \mathbb{i} h_r, g_r - \mathbb{i} h_r \in C[T_1, \dots, T_d]$ are linearly independent. Due to $C(g_i + \mathbb{i} h_i) + C(g_i - \mathbb{i} h_i) = C g_i + C h_i$, we have that

$$\ell_1, \dots, \ell_m, g_1, \dots, g_r, h_1, \dots, h_r$$

are also linearly independent in $C[T_1, \dots, T_d]$ and therefore also in $R[T_1, \dots, T_d]$. It follows that

$$\begin{aligned} \operatorname{rk} H(f,g) &= \#\{k \in \{1, \dots, m\} \mid g(x_k) \neq 0\} + 2r \\ &= \#\{k \in \{1, \dots, m\} \mid g(x_k) \neq 0\} + 2\#\{t \in \{1, \dots, n\} \mid g(z_t) \neq 0\} \\ &= \#\{x \in C \mid f(x) = 0, g(x) \neq 0\} \qquad \text{and} \\ \operatorname{sg} H(f,g) &= \#\{k \in \{1, \dots, m\} \mid g(x_k) > 0\} - \#\{k \in \{1, \dots, m\} \mid g(x_k) < 0\} + r - r \\ &= \#\{x \in R \mid f(x) = 0, g(x) > 0\} - \#\{x \in R \mid f(x) = 0, g(x) < 0\}. \end{aligned}$$

□

Corollary 1.6.6 (Counting roots without side conditions). *Let R be a real closed field, $C := R(\mathbb{i})$ and suppose $f \in R[X]$ is monic. Then*

$$\begin{aligned} \operatorname{rk} H(f) &= \#\{x \in C \mid f(x) = 0\} \qquad \text{and} \\ \operatorname{sg} H(f) &= \#\{x \in R \mid f(x) = 0\}. \end{aligned}$$

Corollary 1.6.7 (Counting roots with several side conditions). *Let R be a real closed field, $m \in \mathbb{N}_0$, $f, g_1, \dots, g_m \in R[X]$ and f monic. Then*

$$\frac{1}{2^m} \sum_{\alpha \in \{1,2\}^m} \operatorname{sg} H(f, g_1^{\alpha_1} \dots g_m^{\alpha_m}) = \#\{x \in R \mid f(x) = 0, g_1(x) > 0, \dots, g_m(x) > 0\}$$

Proof. The left hand side equals

$$\frac{1}{2^m} \sum_{\alpha \in \{1,2\}^m} \sum_{\substack{x \in R \\ f(x)=0}} \operatorname{sgn}((g_1^{\alpha_1} \dots g_m^{\alpha_m})(x)) = \frac{1}{2^m} \sum_{\substack{x \in R \\ f(x)=0}} \prod_{k=1}^{m} \underbrace{(\operatorname{sgn}(g_k(x)) + (\operatorname{sgn}(g_k(x)))^2)}_{= \begin{cases} 0 & \text{if } g_k(x) \leq 0 \\ 2 & \text{if } g_k(x) > 0 \end{cases}}.$$

□

1.7 The real closure

Definition 1.7.1. Let (K, P) be an ordered field. An extension field R of K is called a *real closure* of (K, P) if R is real closed, $R|K$ is algebraic and the order of R [$\to$ 1.4.3, 1.4.4] is an extension of P [$\to$ 1.3.1].

Proposition 1.7.2. *Let (R, P) be an ordered field. Then R is real closed if and only if there is no ordered extension field (L, Q) of (R, P) such that $L \neq R$ and $L|R$ is algebraic.*

Proof. One direction follows from 1.4.13(c). Conversely, suppose that every ordered extension field (L, Q) of (R, P) with $L|R$ algebraic satisfies $L = R$. To show:

(a) R is Euclidean.

(b) Every polynomial of odd degree from $R[X]$ has a root in R.

For (a), we show $P = R^2$. To this end, let $a \in P$. By 1.3.4, we can extend P to $R(\sqrt{a})$. Due to the hypothesis, this implies $R(\sqrt{a}) = R$ and therefore $a = (\sqrt{a})^2 \in R^2$.

To show (b), let $f \in R[X]$ be of odd degree. Choose in $R[X]$ an irreducible divisor g of f of odd degree. Choose a root x of g in some extension field of R. Then $R(x)$ is an extension field of R with odd $[R(x) : R]$ so that P can be extended to $R(x)$ by 1.3.6. By hypothesis, this gives $R(x) = R$. In particular, g and therefore f has a root in R. □

Theorem 1.7.3. *Every ordered field has a real closure.*

Proof. Let (K, P) be an ordered field. Consider the algebraic closure $\overline{K}$ of K and the set

$$M := \{(L, Q) \mid L \text{ subfield of } \overline{K}, Q \text{ order of } L, (K, P) \text{ is an ordered subfield of } (L, Q)\}$$

which is partially ordered by declaring

$$\begin{aligned}(L, Q) \preceq (L', Q') :\iff & (L, Q) \text{ is an ordered subfield of } (L', Q') \\ \overset{1.1.20(b)}{\iff} & (L \subseteq L' \,\&\, Q \subseteq Q')\end{aligned}$$

for all $(L, Q), (L', Q') \in M$. In M every chain possesses an upper bound: The empty chain has (K, P) as an upper bound. A nonempty chain $C \subseteq M$ has

$$\left(\bigcup\{L \mid (L, Q) \in C\}, \bigcup\{Q \mid (L, Q) \in C\}\right) \in M$$

as an upper bound. By Zorn's lemma, M possesses a maximal element (R, Q). Of course, $\overline{K}$ is also the algebraic closure of R and therefore each algebraic extension of R is (up to R-isomorphy) an intermediate field of $\overline{K}|R$. The maximality of (R, Q) in M signifies by 1.7.2 just that R is real closed. Because of $(R, Q) \in M$, the field extension $R|K$ is algebraic and the order Q is an extension of P. □

Lemma 1.7.4. Let (K, P) be an *ordered* subfield of the real closed fields R and R' [$\to$ 1.4.3] and $f \in K[X]$. Then f has the same number of roots in both R and R'.

Proof. WLOG f is monic. The number in question is by 1.6.6 equal to the signature of $H(f)$ that can be calculated already in (K,P) [→ 1.6.1(h)]. □

Theorem 1.7.5. *Let (K,P) be an ordered subfield of (L,Q) such that $L|K$ is algebraic. Let φ be a homomorphism of ordered fields from (K,P) into a real closed field R. Then there is exactly one homomorphism ψ of ordered fields from (L,Q) to R with $\psi|_K = \varphi$.*

Proof. Choose a real closure R' of (L,Q) according to 1.7.3.

Existence: Using Zorn's lemma, one reduces easily to the case where $L|K$ is finite. We denote the different field homomorphisms from L to R extending φ by $\psi_1,\dots,\psi_m$ ($m \in \mathbb{N}_0$). Assume that none of these is a homomorphism of ordered fields from (L,Q) to R (for example if $m = 0$). Then there are $b_1,\dots,b_m \in Q$ such that $\psi_1(b_1) \notin R^2,\dots,\psi_m(b_m) \notin R^2$. By the primitive element theorem there exists

$$a \in L' := L(\sqrt{b_1},\dots,\sqrt{b_m}) \overset{b_i \in Q \subseteq R'^2}{\subseteq} R'$$

such that $L' = K(a)$. The minimal polynomial of a over K has by 1.7.4 the same number of roots in R' and R and therefore in particular a root in R. Hence there is a field homomorphism $\psi\colon L' \to R$ extending φ. Choose $i \in \{1,\dots,m\}$ with $\psi|_L = \psi_i$ (in particular $m > 0$). Then $\psi_i(b_i) = \psi(b_i) = \psi(\sqrt{b_i})^2 \in R^2$ ↯.

Unicity: Let $a \in L$. Choose $f \in K[X] \setminus \{0\}$ with $f(a) = 0$. Choose $a_1,\dots,a_m \in R'$ with $a_1 < \dots < a_m$ such that $\{x \in R' \mid f(x) = 0\} = \{a_1,\dots,a_m\}$. Since $\varphi\colon K \to \varphi(K) \subseteq R$ is an isomorphism of ordered fields, we can suppose WLOG that (K,P) is an ordered subfield of R and $\varphi = \mathrm{id}$. By 1.7.4 there are $b_1,\dots,b_m \in R$ such that $b_1 < \dots < b_m$ and $\{x \in R \mid f(x) = 0\} = \{b_1,\dots,b_m\}$. Choose now $i \in \{1,\dots,m\}$ such that $a = a_i$. We show that each homomorphism ψ of ordered fields from (L,Q) to R with $\psi|_K = \mathrm{id}$ satisfies $\psi(a) = b_i$. To this end, fix such a ψ. By the already proved existence statement, there is a homomorphism of ordered fields $\varrho\colon R' \to R$ such that $\varrho|_L = \psi$. Since ϱ is an embedding, we have $\{\varrho(a_1),\dots,\varrho(a_m)\} = \{b_1,\dots,b_m\}$ and by the monotonicity we even get $\varrho(a_j) = b_j$ for all $j \in \{1,\dots,m\}$. We deduce $\psi(a) = \psi(a_i) = \varrho(a_i) = b_i$. □

Corollary 1.7.6. *Let R and R' be real closures of the ordered field (K,P). Then there is exactly one K-isomorphism from R to R'.*

Proof. The K-isomorphisms from R to R' are obviously exactly the isomorphisms of ordered fields from R to R' whose restriction to K is the identity. For this reason, the claim follows easily from 1.7.5 (for the surjectivity in the existence part use either 1.4.13(c) or the unicity of K-automorphisms of R and of R' [→ 1.7.5]). □

Notation and Terminology 1.7.7. Because of 1.7.6, we speak of *the* real closure $\overline{(K,P)}$ of (K,P). It contains by 1.7.5 (up to K-isomorphy) every ordered field extension (L,Q) of (K,P) with $L|K$ algebraic.

Theorem 1.7.8. *Suppose (K, P) is an ordered field, $L|K$ an algebraic extension, R a real closed field and φ a homomorphism of ordered fields from (K, P) to R. Then*

$$\begin{aligned}\{\psi \mid \psi\colon L \to R \text{ homomorphism}, \psi|_K = \varphi\} &\to \{Q \mid Q \text{ is an extension of } P \text{ to } L\}\\ \psi &\mapsto \psi^{-1}(R^2)\end{aligned}$$

is a bijection.

Proof. The well-definedness is easy to see. To verify the bijectivity, let Q be an extension of P to L. We have to show that there is exactly one homomorphism $\psi\colon L \to R$ with $\psi|_K = \varphi$ fulfilling the condition $\psi^{-1}(R^2) = Q$ that is equivalent to ψ being a homomorphism of ordered fields from (L, Q) to R since

$$\psi^{-1}(R^2) = Q \iff \psi^{-1}(R^2 \cap \psi(L)) = Q \overset{\text{bijective}}{\overset{\psi\colon L \to \psi(L)}{\iff}} R^2 \cap \psi(L) = \psi(Q)$$

$$\overset{\text{order of } \psi(L)}{\overset{R^2 \cap \psi(L)}{\iff}} \psi(Q) \subseteq R^2 \cap \psi(L) \iff \psi(Q) \subseteq R^2.$$

Hence we get the unicity and existence of ψ from 1.7.5. □

Corollary 1.7.9. *Suppose (K, P) is an ordered field, $R := \overline{(K, P)}$ and $L|K$ a finite extension. Let $a \in L$ with $L = K(a)$ and f be the minimal polynomial of a over K. Then*

$$\begin{aligned}\{x \in R \mid f(x) = 0\} &\to \{Q \mid Q \text{ is an extension of } P \text{ to } L\}\\ x &\mapsto \{g(a) \mid g \in K[X], g(x) \in R^2\}\end{aligned}$$

is a bijection.

Proof. By 1.7.8 it is enough to see that

$$\begin{aligned}\{x \in R \mid f(x) = 0\} &\to \{\psi \mid \psi\colon L \to R \text{ is a } K\text{-homomorphism}\}\\ x &\mapsto (g(a) \mapsto g(x)) \qquad (g \in K[X])\end{aligned}$$

is a bijection. This is easy to see. □

Example 1.7.10. Let (K, P) be an ordered field with $2 \notin K^2$. Denote by $\sqrt{2}$ one of the two square roots of 2 in the algebraic closure $\overline{K}$ of K [→ 1.4.7(a)]. Then there are exactly 2 orders of $K(\sqrt{2})$ that extend P, namely the two induced by the field embeddings $K(\sqrt{2}) \hookrightarrow \overline{(K, P)}$ (in one of which $\sqrt{2}$ is positive and in one of which it is negative). In particular, this is true if (K, P) is not Archimedean [→ 1.1.20(d)] and in this case we cannot argue with $\mathbb{R}$ instead of $\overline{(K, P)}$ as we did in 1.3.5.

Proposition 1.7.11. *Let R be a real closed field and K a subfield of R that is (relatively) algebraically closed in R (i.e., no element of R is algebraic over K). Then K is real closed.*

Proof. Apply the criterion from 1.7.2: Every ordered extension field (L, Q) of $(K, R^2 \cap K)$ such that $L|K$ is algebraic is contained in R up to K-isomorphy [→ 1.7.5, 1.7.7] and therefore equal to K. □

Example 1.7.12. The field $\mathbb{R}_{\text{alg}} := \{x \in \mathbb{R} \mid x \text{ algebraic over } Q\}$ of *real algebraic numbers* is the algebraic closure of $\mathbb{Q}$ in $\mathbb{R}$. By 1.7.11, $\mathbb{R}_{\text{alg}}$ is real closed and therefore the real closure of $\mathbb{Q}$ [$\to$ 1.4.3]. Hence $\mathbb{R}_{\text{alg}}$ is uniquely embeddable in every real closed field by 1.7.5. In this sense, $\mathbb{R}_{\text{alg}}$ is the smallest real closed field.

1.8 Real quantifier elimination

Remark 1.8.1. Let M, I and J_i for each $i \in I$ be sets and suppose $A_{ij} \subseteq M$ for all $i \in I$ and $j \in J_i$. Defining the empty intersection as M (that is $\bigcap_{i\in\emptyset} \ldots := \bigcap \emptyset := M$), one has

$$\bigcup_{i\in I} \bigcap_{j\in J_i} A_{ij} = \bigcap_{(j_i)_{i\in I}\in\prod_{i\in I} J_i} \bigcup_{i\in I} A_{ij_i},$$

$$\bigcap_{i\in I} \bigcup_{j\in J_i} A_{ij} = \bigcup_{(j_i)_{i\in I}\in\prod_{i\in I} J_i} \bigcap_{i\in I} A_{ij_i},$$

$$\complement \bigcup_{i\in I} \bigcap_{j\in J_i} A_{ij} = \bigcap_{i\in I} \bigcup_{j\in J_i} \complement A_{ij} \qquad \text{and}$$

$$\complement \bigcap_{i\in I} \bigcup_{j\in J_i} A_{ij} = \bigcup_{i\in I} \bigcap_{j\in J_i} \complement A_{ij}$$

where the *complement* of $A \subseteq M$ is given by $\complement A := \complement_M A := M \setminus A$.

Definition and Proposition 1.8.2. *Let M be a set and $\mathscr{P}(M)$ its power set.*

(a) We call $\mathscr{S} \subseteq \mathscr{P}(M)$ a Boolean algebra *on M if*

- $\emptyset \in \mathscr{S}$,
- $\forall S \in \mathscr{S} : \complement S \in \mathscr{S}$,
- $\forall S_1, S_2 \in \mathscr{S} : S_1 \cap S_2 \in \mathscr{S}$ *and*
- $\forall S_1, S_2 \in \mathscr{S} : S_1 \cup S_2 \in \mathscr{S}$.

(b) Let $\mathscr{G} \subseteq \mathscr{P}(M)$. Then the set of all finite $\left\{\begin{matrix} \text{unions} \\ \text{intersections} \end{matrix}\right\}$ of finite $\left\{\begin{matrix} \text{intersections} \\ \text{unions} \end{matrix}\right\}$ of elements of $\mathscr{G}$ and their complements (with $\bigcap \emptyset := M$) is obviously the smallest Boolean algebra $\mathscr{S}$ on M with $\mathscr{G} \subseteq \mathscr{S}$. It is called the Boolean algebra generated by $\mathscr{G}$ *(on M). Its elements are called the* Boolean combinations *of elements of $\mathscr{G}$.*

Definition and Remark 1.8.3. In the sequel, we let (K, P) always be an ordered field, for example $(K, P) = (\mathbb{Q}, \mathbb{Q}_{\geq 0})$ unless otherwise stated. Moreover, we let $\mathscr{R}$ be a set of real closed fields containing (K, P) as an ordered subfield. For $n \in \mathbb{N}_0$, we set

$$\mathscr{R}_n := \{(R, x) \mid R \in \mathscr{R}, x \in R^n\}.$$

Thereby we have $R^0 = \{\emptyset\} = \{0\}$ and we identify $\mathscr{R}_0$ with $\mathscr{R}$. A Boolean combination of sets of the form

$$\{(R, x) \in \mathscr{R}_n \mid p(x) \geq 0 \text{ (in } R)\} \qquad (p \in K[X_1, \ldots, X_n])$$

is called a

- *K-semialgebraic set in* R^n if $\mathscr{R} = \{R\}$, and

- an *n-ary* (K,P)*-semialgebraic class* if $\mathscr{R}$ is "potentially very big" (in any case big enough to contain all real closed ordered extension fields of (K,P) that are currently in the game).

We identify K-semialgebraic sets in R^n with subsets of R^n. Thus these are simply the subsets of R^n that can be defined by combining finitely many polynomial inequalities with coefficients in K by the logical connectives "not", "and" and "or". A *semialgebraic set* in R^n is an R-semialgebraic set in R^n. A *semialgebraic class* is a $\mathbb{Q}$-semialgebraic class.

Remark 1.8.4. (a) On the first reading, the reader might want to think of $\mathscr{R} = \{R\}$ or even of $\mathscr{R} = \{\mathbb{R}\}$ in order to have a good geometric perception. Initially one can therefore think of (K,P)-semialgebraic classes as K-semialgebraic sets.

(b) One can conceive $\mathscr{R}$ as the "set" of all real closed ordered extension fields of (K,P). Unfortunately, this is not a set (otherwise Zorn's lemma would yield real closed fields having no proper real closed extension field in contradiction to 1.3.7 combined with 1.7.3) but a proper *class*. But we do not want to get into the formal notion of a class and instead adopt a naïve point of view from which sets and classes are synonymous where "big" sets often tend to be called classes.

(c) Whoever gets vertiginous from (b), has several ways out: Our resort here is that $\mathscr{R}$ is a honest set that is at any one time sufficiently big (often $\#\mathscr{R} = 1$ is enough and almost always $\#\mathscr{R} = 2$ is enough). Alternatively, one could learn the subtle non-naïve handling of sets and classes. As a third option, one could work, instead of with (K,P)-semialgebraic classes, with formulas of first-order logic in the language of ordered fields with additional constants for the elements of K. The last two options are technically very involved.

Remark 1.8.5. Obviously, $\emptyset$ and $\mathscr{R}$ are the only 0-ary (K,P)-semialgebraic classes

Proposition 1.8.6. *Every* (K,P)*-semialgebraic class is of the form*

$$\bigcup_{i=1}^{k} \{(R,x) \in \mathscr{R}_n \mid f_i(x) = 0, g_{i1}(x) > 0, \ldots, g_{im}(x) > 0\}$$

for some $n,k,m \in \mathbb{N}_0$, $f_i, g_{ij} \in K[X_1,\ldots,X_n]$.

Proof. By 1.8.3 and 1.8.2(b) such a class is a finite union of classes of the form

$$\begin{aligned}
&\{(R,x) \in \mathscr{R}_n \mid h_1(x) \geq 0, \ldots, h_s(x) \geq 0, h_{s+1}(x) < 0, \ldots, h_{s+t}(x) < 0\} \\
&\qquad = \bigcup_{\delta \in \{0,1\}^s} \left\{ (R,x) \in \mathscr{R}_n \mid \begin{array}{r} \operatorname{sgn}(h_1(x)) = \delta_1, \ldots, \operatorname{sgn}(h_s(x)) = \delta_s, \\ -h_{s+1}(x) > 0, \ldots, -h_{s+t}(x) > 0 \end{array} \right\} \\
&\qquad\qquad = \bigcup_{\delta \in \{0,1\}^s} \left\{ (R,x) \in \mathscr{R}_n \mid \begin{array}{r} \left(\sum\limits_{\substack{i=1 \\ \delta_i = 0}}^{s} h_i^2 \right)(x) = 0, \mathop{\&}\limits_{\substack{i=1 \\ \delta_i = 1}}^{s} h_i(x) > 0, \\ -h_{s+1}(x) > 0, \ldots, -h_{s+t}(x) > 0 \end{array} \right\}
\end{aligned}$$

for some $s, t \in \mathbb{N}_0$ and $h_i \in K[X_1, \ldots, X_n]$ □

Proposition 1.8.7. *Let $m, n \in \mathbb{N}_0$, $h_1, \ldots, h_m \in K[X_1, \ldots, X_n]$ and $S \subseteq \mathscr{R}_m$ a (K,P)-semialgebraic class. Then $\{(R,x) \in \mathscr{R}_n \mid (R, (h_1(x), \ldots, h_m(x))) \in S\}$ is a (K,P)-semialgebraic class.*

Proof. If $S = \bigcup_{i=1}^{k} \{(R,y) \in \mathscr{R}_m \mid f_i(y) = 0, g_{i1}(y) > 0, \ldots, g_{i\ell}(y) > 0\}$ with $m, k, \ell \in \mathbb{N}_0$, $f_i, g_{ij} \in K[Y_1, \ldots, Y_m]$ so that

$$\begin{aligned}
&\{(R,x) \in \mathscr{R}_n \mid (h_1(x), \ldots, h_m(x)) \in S\} \\
&\qquad = \bigcup_{i=1}^{k} \{(R,x) \in \mathscr{R}_n \mid (f_i(h_1, \ldots, h_m))(x) = 0, \\
&\qquad\qquad (g_{i1}(h_1, \ldots, h_m))(x) > 0, \ldots, (g_{i\ell}(h_1, \ldots, h_m))(x) > 0\}.
\end{aligned}$$

□

Corollary 1.8.8. *Let R be a real closed field. Preimages of semialgebraic subsets of R^m under polynomial maps $R^n \to R^m$ are again semialgebraic in R^n.*

Lemma 1.8.9. For every $s \in \mathbb{N}_0$,

$$\left\{ (R,x) \in \mathscr{R}_{d+1} \mid \sigma\left(\sum_{i=0}^{d} x_i T^i \right) = s \text{ with respect to } R[T] \right\}$$

is a semialgebraic class.

Proof. The class in question equals

$$\bigcup_{\substack{\delta \in \{-1,0,1\}^n \\ \sigma\left(\sum_{i=0}^{d} \delta_i T^i\right) = s \text{ with respect to } \mathbb{R}[T]}} \{(R,x) \in \mathscr{R}_{d+1} \mid \operatorname{sgn}_R(x_0) = \delta_0, \ldots, \operatorname{sgn}_R(x_d) = \delta_d\}.$$

□

Remark 1.8.10. We will now need the *simultaneous* diagonalization of a symmetric matrix as a quadratic form and as an endomorphism [→ 1.6.1(g)]. The reader should know this over $\mathbb{R}$ from linear algebra but we will now need it more generally over an arbitrary real closed field. Later in this chapter, we will provide methods from which it becomes immediately clear that, for each fixed matrix size, the class of all fields $R \in \mathscr{R}$, over which the corresponding statement is true, is a 0-ary semialgebraic class. Since the statement is true over $\mathbb{R}$, it must then by 1.8.5 also hold true over every real closed field. In a similar way, we will soon be able to carry over a great many statements from $\mathbb{R}$ to all real closed fields. Unfortunately, we are not that far yet and therefore we have to check if the proof from linear algebra goes through over an arbitrary real closed field. Some of the proofs of the diagonalization in question use however proper analysis instead of just the fundamental theorem of algebra. Since the whole analysis is built on the completeness of $\mathbb{R}$ [→ 1.1.16], those proofs do not generalize without further ado. Thus we give a compact ad-hoc-proof.

Theorem 1.8.11. *Let R be a real closed field and $M \in SR^{n\times n}$. Then there is some $P \in \mathrm{GL}_n(R)$ satisfying $P^TP = I_n$ such that P^TMP is a diagonal matrix.*

Proof. Call a symmetric bilinear form $V \times V \to R$, $(v, w) \mapsto \langle v, w\rangle$ on an R-vector space V positive definite if $\langle v, v\rangle > 0$ for all $v \in V \setminus \{0\}$. Call an R-vector space together with a positive definite symmetric bilinear form a Euclidean R-vector space. Call an endomorphism f of a Euclidean R-vector space V self-adjoint if $\langle f(v), w\rangle = \langle v, f(w)\rangle$ for all $v, w \in V$.

Claim 1: Let V be a Euclidean R-vector space, $f \in \mathrm{End}(V)$ self-adjoint and v an eigenvector of f. Then $U := \{u \in V \mid \langle u, v\rangle = 0\}$ is a subspace of V with $v \notin U$ and $f(U) \subseteq U$.

Explanation. Choose $\lambda \in R$ with $f(v) = \lambda v$ and let $u \in U$. Then $\langle f(u), v\rangle = \langle u, f(v)\rangle = \langle u, \lambda v\rangle = \lambda\langle u, v\rangle = \lambda 0 = 0$.

Claim 2: Let $V \neq 0$ be a finite-dimensional Euclidean R-vector space and $f \in \mathrm{End}(V)$ self-adjoint. Then f possesses an eigenvalue in R.

Explanation. Assume f has no eigenvalue. By Caley-Hamilton and the fundamental theorem 1.4.14, there are $a, b \in R$ with $b \neq 0$ such that $(f - a\,\mathrm{id}_V)^2 + b^2\,\mathrm{id}_v$ has a non-trivial kernel. Since f is self-adjoint, $g := f - a\,\mathrm{id}_V$ is so. Choose $v \in V$ with $g^2(v) = -b^2v$. Then $0 \leq \langle g(v), g(v)\rangle = \langle g^2(v), v\rangle = \langle -b^2v, v\rangle = -b^2\langle v, v\rangle < 0$. ↯

Claim 3: Let V be a finite-dimensional Euclidean R-vector space and $f \in \mathrm{End}(V)$ self-adjoint. Then there is an eigenbasis $v_1, \ldots, v_n$ for f with $(\langle v_i, v_j\rangle)_{1\leq i,j\leq n} = I_n$.

Explanation. Use Claim 1, Claim 2 and induction over the dimension V.

In virtue of $\langle x, y\rangle := \sum_{i=1}^n x_iy_i$ $(x, y \in R^n)$, R^n is an Euclidean R-vector space and $f\colon R^n \to R^n$, $x \mapsto Mx$ is self-adjoint. By Claim 3, there is an eigenbasis $v_1, \ldots, v_n$ for f such that $(\langle v_i, v_j\rangle)_{1\leq i,j\leq n} = I_n$. Set $P := (v_1 \ldots v_n) \in \mathrm{GL}_n(R)$. Then

$$P^TP = \begin{pmatrix} v_1^T \\ \vdots \\ v_n^T \end{pmatrix} \begin{pmatrix} v_1 & \ldots & v_n \end{pmatrix} = I_n$$

and P is the change-of-basis matrix from $(v_1, \dots, v_n)$ to the standard basis. It follows that $P^T MP = P^{-1}MP$ is the representing matrix of f with respect to $(v_1, \dots, v_n)$. □

Corollary 1.8.12 (Determination of the signature using Descartes' rule of signs)**.** *Let R be a real closed field, $q \in R[T_1, \dots, T_d]$ a quadratic form and $h := \det(M(q) - XI_d) \in R[X]$ the characteristic polynomial of the representing matrix* [→ 1.6.1(d)] *of q. Then we have:*

(a) *h is real-rooted* [→ 1.5.11]

(b) $\operatorname{sg} q = \mu(h) - \mu(h(-X))$ [→ 1.6.1(h), 1.5.5(a)]

(c) $\operatorname{sg} q = \sigma(h) - \sigma(h(-X))$ [→ 1.5.5(b)]

Proof. Using 1.8.11, choose $P \in \mathrm{GL}_d(R)$ such that $P^T P = I_d$ and $P^T M(q) P$ is diagonal,

$$P^T M(q) P = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix}$$

with $\lambda_i \in R$. We have

$$\begin{aligned} h &= h \det(P^T P) \\ &= (\det(P^T))(\det(M(q) - XI_d))(\det P) \\ &= \det(P^T M(q) P - XP^T P) = \det \begin{pmatrix} \lambda_1 - X & & 0 \\ & \ddots & \\ 0 & & \lambda_d - X \end{pmatrix} = \prod_{i=1}^{d} (\lambda_i - X), \end{aligned}$$

from which (a) follows immediately. Because of

$$M(q) = (P^T)^T \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_d \end{pmatrix} P^T$$

and $P^T \in \mathrm{GL}_d(R)$, it follows from 1.6.1(e) that

$$\operatorname{sg} q = \#\{i \in \{1, \dots, d\} \mid \lambda_i > 0\} - \#\{i \in \{1, \dots, d\} \mid \lambda_i < 0\} = \mu(h) - \mu(h(-X)),$$

which proves (b). Finally, (c) follows from (a) and (b) due to the exactness of Descartes' rule of signs for real rooted polynomials [→ 1.5.14]. □

Remark 1.8.13. Combining 1.8.12 with 1.6.7, one can reduce the count of real roots of polynomials without multiplicity with side conditions by means of the Hermite method from §1.6 to the count of roots of real-rooted polynomials with multiplicity by means of Descartes' rule from §1.5.

Lemma 1.8.14. Let $m, n, d \in \mathbb{N}_0$ and $f, g_1, \dots, g_m \in K[X_1, \dots, X_{n+1}]$. Then

$$\{(R,x) \in \mathscr{R}_n \mid \deg f(x, X_{n+1}) = d \quad \& \\ \exists x_{n+1} \in R\colon (f(x,x_{n+1}) = 0 \,\&\, g_1(x,x_{n+1}) > 0 \,\&\, \dots \,\&\, g_m(x,x_{n+1}) > 0)\}$$

is a (K,P)-semialgebraic class.

Proof. Write $f = \sum_{i=0}^{D} h_i X_{n+1}^i$ for some $D \in \mathbb{N}_0$, $D \geq d$ and $h_i \in K[X_1, \dots, X_n]$. WLOG $h_d \neq 0$. Then

$$f_0 := \sum_{i=0}^{d} \frac{h_i}{h_d} X_{n+1}^i \in K(X_1, \dots, X_n)[X_n]$$

is monic of degree d. For every $\alpha \in \{1,2\}^m$, we consider also $g_1^{\alpha_1} \cdots g_m^{\alpha_m}$ as a polynomial in X_{n+1} with coefficients from the field $K(X_1, \dots, X_n)$ and set

$$h_\alpha := \det(M(H(f_0, g_1^{\alpha_1} \cdots g_m^{\alpha_m})) - XI_d) \in K(X_1, \dots, X_n)[X].$$

By construction [→ 1.6.1(i), 1.6.2, 1.6.3], there is some $N \in \mathbb{N}$ such that

$$h_d^N h_\alpha \in K[X_1, \dots, X_{n+1}]$$

for all $\alpha \in \{1,2\}^m$. Now the class from the claim can be written by 1.6.7 as

$$\left\{(R,x) \in \mathscr{R}_n \mid h_D(x) = \dots = h_{d+1}(x) = 0 \neq h_d(x) \quad \& \right. \\ \left. \sum_{\alpha \in \{1,2\}^m} \operatorname{sg} H(f_0(x, X_{n+1}), (g_1^{\alpha_1} \cdots g_m^{\alpha_m})(x, X_{m+1})) > 0 \right\}.$$

But

$$\left\{(R,x) \in \mathscr{R}_n \mid h_d(x) \neq 0 \,\&\, \sum_{\alpha \in \{1,2\}^m} \operatorname{sg} H(f_0(x, X_{n+1}), (g_1^{\alpha_1} \cdots g_m^{\alpha_m})(x, X_{m+1})) > 0 \right\} \\ \overset{1.8.12}{\underset{\text{⚠}}{=}} \left\{(R,x) \in \mathscr{R}_n \mid h_d(x) \neq 0 \,\&\, \sum_{\alpha \in \{1,2\}^m} (\sigma(h_\alpha(x,X)) - \sigma(h_\alpha(x,-X))) > 0 \right\} \\ = \bigcup_{\substack{(s_\alpha)_{\alpha \in \{1,2\}^m}, (t_\alpha)_{\alpha \in \{1,2\}^m} \in \{0,\dots,d\}^{\{1,2\}^m} \\ \sum_{\alpha \in \{1,2\}^m} (s_\alpha - t_\alpha) > 0}} \bigcap_{\alpha \in \{1,2\}^m} \left\{ \begin{array}{r} h_d(x) \neq 0, \\ (R,x) \in \mathscr{R}_n \mid \sigma((h_d^N h_\alpha)(x,X)) = s_\alpha, \\ \sigma((h_d^N h_\alpha)(x,-X)) = t_\alpha \end{array} \right\}$$

is (K,P)-semialgebraic by 1.8.9 and 1.8.7. Here the warning sign ⚠ indicates where an important argument flows in:

$$h_\alpha(x,X) = \det(M(H(f_0(x, X_{n+1}), (g_1^{\alpha_1} \cdots g_m^{\alpha_m})(x, X_{n+1}))) - XI_d)$$

since evaluating in x commutes with building companion matrices, Hermite forms and with taking determinants [→ 1.6.2, 1.6.1(i)]. □

Lemma 1.8.15. Let R be a real closed field, $m \in \mathbb{N}_0$ and $g_1, \ldots, g_m \in R[X]$. Setting $g := g_1 \cdots g_m$ and $f := (1-g^2)g'$, we have

(a) There is an $x \in R$ satisfying $g_1(x) > 0, \ldots, g_m(x) > 0$ if and only if there is such an $x \in R$ satisfying in addition $f(x) = 0$.

(b) If $f = 0$ and $g_1 \neq 0, \ldots, g_m \neq 0$, then $g_1, \ldots, g_m \in R$.

Proof. (b) Suppose $f = 0$. Then $g^2 = 1$ or $g' = 0$. In both cases it follows $g \in R$ and thus $g_1, \ldots, g_m \in R$ provided that $g_1 \neq 0, \ldots, g_m \neq 0$.

(a) Let $x \in R$ such that $g_1(x) > 0, \ldots, g_m(x) > 0$. Denote by $a_1, \ldots, a_r$ where $r \in \mathbb{N}_0$ and $a_1 < \ldots < a_r$ the roots of g in R.

First consider the case where $r = 0$. By the intermediate value theorem 1.4.16 each of the g_i is positive on R. It suffices therefore to show that f has a root in R. By Definition 1.4.9, g has even degree. If g has degree 0, then $g' = 0$ and we are done. So suppose now $\deg g \geq 2$. Then the degree of g' is odd so that g' and in particular f has a root in R by Definition 1.4.9.

From now on suppose that $r > 0$. By the intermediate value theorem 1.4.16 each of the g_i has constant sign on each of the intervals $(-\infty, a_1), (a_1, a_2), \ldots, (a_{r-1}, a_r), (a_r, \infty)$. It is therefore enough to show that f possesses in each of these sets a root. By Rolle's theorem 1.4.17, g' and therefore f has on each of the sets (a_i, a_{i+1}) $(1 \leq i \leq r-1)$ a root. WLOG $f \neq 0$. Then $g' \neq 0$ and g has degree ≥ 1. Consequently, $1 - g^2$ has a leading monomial of even degree with a negative leading coefficient. By Lemma 1.5.3(a), $(1-g^2)(y) < 0$ for all $y \in R$ with $|y|$ sufficiently big. On the other hand, $(1-g^2)(a_1) = 1 = (1-g^2)(a_r)$. By the intermediate value theorem 1.4.16, $1 - g^2$ and therefore f has a root on each of the sets $(-\infty, a_1)$ and (a_r, ∞). □

Lemma 1.8.16. Let $m, n \in \mathbb{N}_0$ and $g_1, \ldots, g_m \in K[X_1, \ldots, X_{n+1}]$. Then

$$\{(R, x) \in \mathscr{R}_n \mid \exists x_{n+1} \in R : (g_1(x, x_{n+1}) > 0 \,\&\, \cdots \,\&\, g_m(x, x_{n+1}) > 0)\}$$

is a (K, P)-semialgebraic class.

Proof. Set $g := g_1 \cdots g_m$ and $f := (1-g^2)\frac{\partial g}{\partial X_{n+1}}$. Denote by $D := \deg_{X_{n+1}} f \in \{-\infty\} \cup \mathbb{N}_0$ the degree of f considered as a polynomial in X_{n+1} with coefficients from $K[X_1, \ldots, X_n]$. The class in question equals because of 1.8.15

$$\bigcup_{d=0}^{D} \left\{ (R, x) \in \mathscr{R}_n \mid \deg f(x, X_{n+1}) = d \,\&\, \exists x_{n+1} \in R : \begin{pmatrix} f(x, x_{n+1}) = 0 \,\& \\ g_1(x, x_{n+1}) > 0 \,\& \\ \vdots \\ g_m(x, x_{n+1}) > 0 \end{pmatrix} \right\}$$
$$\cup \{(R, x) \in \mathscr{R}_n \mid f(x, X_{n+1}) = 0 \,\&\, g_1(x, 0) > 0 \,\&\, \cdots \,\&\, g_m(x, 0) > 0\}$$

and therefore is (K, P)-semialgebraic by 1.8.14. □

Theorem 1.8.17 (Real quantifier elimination). *Suppose $n \in \mathbb{N}_0$ and S is an $(n+1)$-ary (K,P)-semialgebraic class. Then $\{(R,x) \in \mathscr{R}_n \mid \exists x_{n+1} \in R : (R,(x,x_{n+1})) \in S\}$ and $\{(R,x) \in \mathscr{R}_n \mid \forall x_{n+1} \in R : (R,(x,x_{n+1})) \in S\}$ are n-ary (K,P)-semialgebraic classes.*

Proof. Because the second class is the complement of

$$\{(R,x) \in \mathscr{R}_n \mid \exists x_{n+1} \in R : (R,(x,x_{n+1})) \in \complement S\},$$

it is enough to consider the first class. By means of 1.8.6, one can assume WLOG that S is of the form

$$S = \{(R,(x,x_{n+1}) \in \mathscr{R}_{n+1} \mid f(x,x_{n+1}) = 0, g_1(x,x_{n+1}) > 0, \ldots, g_m(x,x_{n+1}) > 0\}$$

for some $f, g_i \in K[X_1,\ldots,X_{n+1}]$. Setting $D := \deg_{X_{n+1}} f$, we obtain

$$\begin{aligned}
&\{(R,x) \in \mathscr{R}_n \mid \exists x_{n+1} \in R : (R,(x,x_{n+1})) \in S\} \\
&= \bigcup_{d=0}^{D} \left\{ (R,x) \in \mathscr{R}_n \mid \deg f(x,X_{n+1}) = d \,\&\, \exists x_{n+1} \in R : \begin{pmatrix} f(x,x_{n+1}) = 0 \,\& \\ g_1(x,x_{n+1}) > 0 \,\& \\ \vdots \\ g_m(x,x_{n+1}) > 0 \end{pmatrix} \right\} \\
&\quad \cup \begin{pmatrix} \{(R,x) \in \mathscr{R}_n \mid f(x,X_{n+1}) = 0\} \cap \\ \{(R,x) \in \mathscr{R}_n \mid \exists x_{n+1} \in R : (g_1(x,x_{n+1}) > 0 \,\&\, \cdots \,\&\, g_m(x,x_{n+1}) > 0)\} \end{pmatrix}
\end{aligned}$$

which is (K,P)-semialgebraic by 1.8.14 and 1.8.16. □

Theorem 1.8.18. [→ 1.8.8] *Let R be a real closed field. Images of semialgebraic subsets of R^n under polynomial maps $R^n \to R^m$ are again semialgebraic in R^m.*

Proof. Let $S \subseteq R^n$ be semialgebraic and let $h_1,\ldots,h_m \in R[X_1,\ldots,X_n]$. We have to show that $\{y \in \mathbb{R}^m \mid \exists x \in R^n : (x \in S \,\&\, y_1 = h_1(x) \,\&\, \ldots \,\&\, y_m = h_m(x))\}$ is again semialgebraic. But this follows by applying n times the quantifier elimination 1.8.17. □

Example 1.8.19 (Tarski principle). The real quantifier elimination 1.8.17 can be used together with 1.8.5 to generalize many statements from $\mathbb{R}$ to other real closed fields. This has already been advertised in 1.8.10. To give the reader a sense of the type of statements admitting such a generalization, we give several examples.

(a) ("intermediate value theorem for rational functions") [→ 1.4.16] From analysis, we know for $R = \mathbb{R}$: If $f, g \in R[X]$, $a, b \in R$ with $a \leq b$, $g(c) \neq 0$ for all $c \in [a,b]$ and $\operatorname{sgn}\left(\frac{f(a)}{g(a)}\right) \neq \operatorname{sgn}\left(\frac{f(b)}{g(b)}\right)$, then there is a $c \in [a,b]$ with $f(c) = 0$. We claim that this

is valid even for all real closed fields R. To this end, it is enough to show that for each $d \in \mathbb{N}$

$$S_d := \left\{ R \in \mathscr{R} \mid \left(\begin{array}{l} \forall x_0, \dots, x_d, y_0, \dots, y_d, a, b \in R : \\ \left(\left\{ \begin{array}{l} (a \leq b \ \& \ \left(\forall c \in [a,b] : \sum_{i=0}^d y_i c^i \neq 0\right) \ \& \\ \operatorname{sgn}\left(\left(\sum_{i=0}^d x_i a^i\right)\left(\sum_{i=0}^d y_i b^i\right)\right) \neq \operatorname{sgn}\left(\left(\sum_{i=0}^d x_i b^i\right)\left(\sum_{i=0}^d y_i a^i\right)\right) \end{array} \right\}_{(*)} \right. \\ \left. \implies \exists c \in [a,b] : \sum_{i=0}^d x_i c^i = 0 \right\}_{(**)} \end{array} \right) \right\}_{(***)} \right\}$$

is a semialgebraic class because then $\mathbb{R} \in S_d$ implies by 1.8.5 $S_d = \mathscr{R}$. Fix $d \in \mathbb{N}$. Applying the quantifier elimination 1.8.17 $2d+4$ times, it is enough to show that the following class is semialgebraic:

$$\begin{aligned} \{(R, (x_0, \dots, x_d, y_0, \dots, y_d, a, b)) \in \mathscr{R}^{2d+4} \mid (***)\} = \\ \complement \underbrace{\{(R, (x_0, \dots, x_d, y_0, \dots, y_d, a, b)) \in \mathscr{R}^{2d+4} \mid (*)\}}_{S'} \\ \cup \underbrace{\{(R, (x_0, \dots, x_d, y_0, \dots, y_d, a, b)) \in \mathscr{R}^{2d+4} \mid (**)\}}_{S''} \end{aligned}$$

It is thus enough to show that S' and S'' are semialgebraic. We accomplish this in each case by applying the quantifier elimination 1.8.17. We explicate this only for S' since it is analogous and even simpler for S'':

$$\begin{aligned} S' = \{(R, (x_0, \dots, x_d, y_0, \dots, y_d, a, b)) \mid b - a \geq 0\} \cap \\ \{(R, (x_0, \dots, x_d, y_0, \dots, y_d, a, b)) \mid \forall c \in R : (c \in (a,b) \implies \overbrace{\sum_{i=0}^d y_i c^i \neq 0}^{(****)})\} \cap \\ \bigcup_{\substack{\delta, \varepsilon \in \{-1,0,1\} \\ \delta \neq \varepsilon}} \left\{ (R, (x_0, \dots, x_d, y_0, \dots, y_d, a, b)) \mid \begin{array}{l} \operatorname{sgn}\left(\left(\sum_{i=0}^d x_i a^i\right)\left(\sum_{i=0}^d y_i b^i\right)\right) = \delta, \\ \operatorname{sgn}\left(\left(\sum_{i=0}^d x_i b^i\right)\left(\sum_{i=0}^d y_i a^i\right)\right) = \varepsilon \end{array} \right\}. \end{aligned}$$

By quantifier elimination it is enough to show that

$$\{(R, (x_0, \dots, x_d, y_0, \dots, y_d, a, b)) \mid (****)\}$$

is semialgebraic. But this class equals

$$\begin{aligned} \{(R, (x_0, \dots, x_d, y_0, \dots, y_d, a, b)) \mid a \leq c, c \leq b\} \cup \\ \left\{ (R, (x_0, \dots, x_d, y_0, \dots, y_d, a, b, c)) \mid \sum_{i=0}^d y_i c^i \neq 0 \right\} \end{aligned}$$

(b) Let R be a real closed field and $f \in R[X]$ with $f \geq 0$ on R. We claim that the sum $g := f + f' + f'' + \ldots$ of all derivatives of f satisfies again $g \geq 0$ on R. We show this first for $R = \mathbb{R}$: In this case, we have for all $x \in \mathbb{R}$

$$\frac{dg(x)e^{-x}}{dx} = g'(x)e^{-x} - g(x)e^{-x} = (g'(x) - g(x))e^{-x} = -f(x)e^{-x} \leq 0,$$

from which it follows that $h\colon \mathbb{R} \to \mathbb{R},\ x \mapsto g(x)e^{-x}$ is anti-monotonic [→ 1.4.19]. From this and the fact that $\lim_{x\to\infty} h(x) = \lim_{x\to\infty}(g(x)e^{-x}) = 0$, we deduce that $h(x) \geq 0$ and therefore $g(x) \geq 0$ for all $x \in \mathbb{R}$. Thus the claim is proved for $R = \mathbb{R}$. To show it for all real closed fields R, it is now enough to show that for all $d \in \mathbb{N}$

$$S_d := \left\{ R \in \mathscr{R} \mid \forall a_0, \ldots, a_d \in R : \left(\left(\forall x \in R : \sum_{i=0}^{d} a_i x^i \geq 0 \right) \implies \forall x \in R : \sum_{k=0}^{d} \sum_{i=k}^{d} i(i-1)\cdots(i-k+1) a_i x^{i-k} \geq 0 \right) \right\}$$

is semialgebraic since then by 1.8.5 $\mathbb{R} \in S_d$ implies $S_d = \mathscr{R}$. This can be shown for each $d \in \mathbb{N}$ by applying the quantifier elimination $d + 3$ times.

(c) We can reprove 1.8.11 since for $R = \mathbb{R}$ it is already known from linear algebra and it suffices to show for fixed $n \in \mathbb{N}$ that

$$S_n := \left\{ R \in \mathscr{R} \mid \begin{array}{l} \forall a_{11}, a_{12}, \ldots, a_{nn} \in R : \\ \left((\forall i, j \in \{1, \ldots, n\} : a_{ij} = a_{ji}) \implies \right. \\ \left(\left(\begin{array}{l} \exists b_{11}, b_{12}, \ldots, b_{nn} \in R : \forall i, k \in \{1, \ldots, n\} : \\ \sum_{j=1}^{n} b_{ij} b_{jk} = \delta_{ik} \end{array} \right) \& \right. \\ \left. \left. \forall i, \ell \in \{1, \ldots, n\} : \left(i \neq \ell \implies \sum_{j,k=1}^{n} b_{ji} a_{jk} b_{k\ell} = 0 \right) \right) \right) \end{array} \right\}$$

is semialgebraic. We manage to do so by implementing the quantifications over i, j, k, ℓ as finite intersections of semialgebraic classes and by eliminating the quantification over $a_{11}, \ldots, b_{nn}$ by applying $2n^2$ times 1.8.17.

(d) By 1.8.5, $\{R \in \mathscr{R} \mid R \text{ archimedean}\}$ [→ 1.1.9(a)] is not a semialgebraic class (if $\mathscr{R}$ is big enough) since it contains $\mathbb{R}$ but not $\overline{(\mathbb{R}(X), P)}$ where P is an arbitrary order of $\mathbb{R}(X)$.

1.9 Canonical isomorphisms of Boolean algebras of semialgebraic sets and classes

In this section, we fix again an ordered field (K, P) and a set $\mathscr{R}$ of real closed extensions of (K, P) [→ 1.8.3].

Definition 1.9.1. Let M_1 and M_2 be sets, $\mathscr{S}_1$ a Boolean algebra on M_1 and $\mathscr{S}_2$ a Boolean algebra on M_2. Then $\Phi\colon \mathscr{S}_1 \to \mathscr{S}_2$ is called a *homomorphism of Boolean algebras* if $\Phi(\emptyset) = \emptyset$, $\Phi\left(\complement S\right) = \complement\Phi(S)$, $\Phi(S \cap T) = \Phi(S) \cap \Phi(T)$ and $\Phi(S \cup T) = \Phi(S) \cup \Phi(T)$ for all $S, T \in \mathscr{S}_1$. If Φ is in addition $\left\{\begin{array}{c}\text{injective}\\ \text{surjective}\\ \text{bijective}\end{array}\right\}$, then Φ is called an $\left\{\begin{array}{c}\text{embedding}\\ \text{epimorphismus}\\ \text{isomophism}\end{array}\right\}$ of Boolean algebras.

Lemma 1.9.2. Suppose $\mathscr{S}_1$ and $\mathscr{S}_2$ are Boolean algebras and $\Phi\colon \mathscr{S}_1 \to \mathscr{S}_2$ is a homomorphism. Then the following are equivalent:

(a) Φ is an embedding.

(b) $\forall S \in \mathscr{S}_1 : (\Phi(S) = \emptyset \implies S = \emptyset)$

Proof. (a) $\implies$ (b) Suppose (a) holds and consider $S \in \mathscr{S}_1$ such that $\Phi(S) = \emptyset$. Then $\Phi(S) = \emptyset = \Phi(\emptyset)$ and hence $S = \emptyset$ by the injectivity of Φ.

(b) $\implies$ (a) Suppose (b) holds and let $S, T \in \mathscr{S}_1$ such that $\Phi(S) = \Phi(T)$. Then $\Phi(S \setminus T) = \Phi\left(S \cap \complement T\right) = \Phi(S) \cap \complement\Phi(T) = \emptyset$ and therefore $S \setminus T = \emptyset$. Analogously, we obtain $T \setminus S = \emptyset$. Then $S = T$. □

Notation 1.9.3. Let $n \in \mathbb{N}_0$. From now on, we denote by $\mathscr{S}_n$ the Boolean algebra of all n-ary (K, P)-semialgebraic classes. For every $R \in \mathscr{R}$, we let furthermore $\mathscr{S}_{n,R}$ denote the Boolean algebra of all K-semialgebraic subsets of R^n (i.e., $\mathscr{S}_{n,R} = \mathscr{S}_n$ for $\mathscr{R} = \{R\}$). We call the map $\operatorname{Set}_R\colon \mathscr{S}_n \to \mathscr{S}_{n,R}$, $S \mapsto \{x \in R^n \mid (R, x) \in S\}$ the *setification* to R for every $R \in \mathscr{R}$.

Theorem and Definition 1.9.4. *Let $n \in \mathbb{N}_0$ and $R \in \mathscr{R}$. The setification*

$$\operatorname{Set}_R\colon \mathscr{S}_n \to \mathscr{S}_{n,R}$$

is an isomorphism of Boolean algebras. We call its inverse map

$$\operatorname{Class}_R := \operatorname{Set}_R^{-1}\colon \mathscr{S}_{n,R} \to \mathscr{S}_n$$

the classification.

Proof. It is clear that Set_R is an epimorphism. Suppose $\emptyset \neq S \in \mathscr{S}_n$. By Lemma 1.9.2, it suffices to show $\operatorname{Set}_R S \neq \emptyset$. By the quantifier elimination 1.8.17,

$$T := \{R' \in \mathscr{R} \mid \exists x \in R'^n : (R', x) \in S\}$$

is (K, P)-semialgebraic and hence by 1.8.5 either empty or $\mathscr{R}$. From $S \neq \emptyset$, we have of course $T \neq \emptyset$. Therefore $R \in \mathscr{R} = T$, i.e., there is some $x \in R^n$ with $(R, x) \in S$. Then $x \in \operatorname{Set}_R S$ and thus $\operatorname{Set}_R S \neq \emptyset$. □

Corollary and Definition 1.9.5. *Let $n \in \mathbb{N}_0$ and $R, R' \in \mathscr{R}$. Then there is exactly one isomorphism of Boolean algebras* $\mathrm{Transfer}_{R,R'} : \mathscr{S}_{n,R} \to \mathscr{S}_{n,R'}$ *satisfying*

$$\mathrm{Transfer}_{R,R'}(\{x \in R^n \mid p(x) \geq 0\}) = \{x \in R'^n \mid p(x) \geq 0\}$$

for all $p \in K[X_1, \ldots, X_n]$. We call $\mathrm{Transfer}_{R,R'}$ *the* transfer *from R to R'.*

Proof. The uniqueness is clear since $\mathscr{S}_{n,R}$ is generated by

$$\{\{x \in R^n \mid p(x) \geq 0\} \mid p \in K[X_1, \ldots, X_n]\}$$

[→ 1.8.2(b)]. Existence is established by setting $\mathrm{Transfer}_{R,R'} := \mathrm{Set}_{R'} \circ \mathrm{Class}_R$. Indeed, let $p \in K[X_1, \ldots, X_n]$ and set $S := \{(R, x) \in \mathscr{R}_n \mid p(x) \geq 0 \text{ in } R\}$. Then the claim is that $\mathrm{Transfer}_{R,R'}(\mathrm{Set}_R S) = \mathrm{Set}_{R'}(S)$ which is clear since $\mathrm{Transfer}_{R,R'}(\mathrm{Set}_R S) = (\mathrm{Set}_{R'} \circ \mathrm{Class}_R)(\mathrm{Set}_R S) = \mathrm{Set}_{R'}((\underbrace{\mathrm{Class}_R \circ \mathrm{Set}_R}_{\mathrm{id}_{\mathscr{S}_n}})(S))$. □

§2 Hilbert's 17th problem

2.1 Nonnegative polynomials in one variable

Theorem 2.1.1. *Suppose R is a real closed field and $f \in R[X]$. Then the following are equivalent:*

(a) $f \geq 0$ *on* R [→ 1.4.15]

(b) f *is a sum of two squares in* $R[X]$.

(c) $f \in \sum R[X]^2$ [→ 1.1.18]

Proof. (b) $\Longrightarrow$ (c) $\Longrightarrow$ (a) is trivial. In order to show (a) $\Longrightarrow$ (b), we set $C := \mathbb{R}(\mathbb{i})$ and consider the ring automorphism

$$C[X] \to C[X],\ p \mapsto p^*$$

given by $a^* = a$ for $a \in R$, $\mathbb{i}^* = -\mathbb{i}$ and $X^* = X$. WLOG $f \neq 0$. By the fundamental theorem of algebra 1.4.14, there exist $k, \ell \in \mathbb{N}_0$, $c \in R^\times$, $a_1, \ldots, a_k \in R$, $b_1, \ldots, b_k \in R^\times$, $\alpha_1, \ldots, \alpha_\ell \in \mathbb{N}$ and pairwise different $d_1, \ldots, d_\ell \in R$ such that

$$\begin{aligned} f &= c\left(\prod_{i=1}^{k}((X-a_i)^2+b_i^2)\right)\prod_{j=1}^{\ell}(X-d_j)^{\alpha_j} \\ &= c\left(\prod_{i=1}^{k}(X-(a_i+b_i\mathbb{i}))\right)\left(\prod_{i=1}^{k}(X-(a_i-b_i\mathbb{i}))\right)\prod_{j=1}^{\ell}(X-d_j)^{\alpha_j}. \end{aligned}$$

Suppose now $f \geq 0$ on R. Then we have $0 \leq \operatorname{sgn}(f(x)) = (\operatorname{sgn} c)\prod_{j=1}^{\ell}(\operatorname{sgn}(x-d_j))^{\alpha_j}$ for all $x \in R$. From this, we deduce easily $\alpha_j \in 2\mathbb{N}$ and $c \in R^2$. Setting

$$g := \sqrt{c}\left(\prod_{i=1}^{k}(X-(a_i+b_i\mathbb{i}))\right)\prod_{j=1}^{\ell}(X-d_j)^{\frac{\alpha_j}{2}} \in C[X],$$

we have now $f = g^*g$. Writing $g = p + \mathbb{i}q$ with $p, q \in R[X]$, this amounts to $f = (p-\mathbb{i}q)(p+\mathbb{i}q) = p^2+q^2$. □

Theorem 2.1.2 (Cassels). *Let $(K, \leq)$ be an ordered field. Suppose $\ell \in \mathbb{N}_0$, $f_1, \ldots, f_\ell \in K[X]$, $g_1, \ldots, g_\ell \in K[X] \setminus \{0\}$ and $a_1, \ldots, a_\ell \in K_{\geq 0}$ with $\sum_{i=1}^{\ell} a_i \left(\frac{f_i}{g_i}\right)^2 \in K[X]$. Then there are $p_1, \ldots, p_\ell \in K[X]$ such that*

$$\sum_{i=1}^{\ell} a_i \left(\frac{f_i}{g_i}\right)^2 = \sum_{i=1}^{\ell} a_i p_i^2.$$

Proof. WLOG $a_i > 0$ for all $i \in \{1, \ldots, \ell\}$ and $g_1 = \ldots = g_\ell$. It suffices to show: Let $h \in K[X]$ a polynomial for which there exists some $g \in K[X]$ of degree ≥ 1 and $f_1, \ldots, f_\ell \in K[X]$ satisfying $hg^2 = \sum_{i=1}^{\ell} a_i f_i^2$. Then there is some $G \in K[X] \setminus \{0\}$ with a degree that is smaller than that of g and $F_1, \ldots, F_\ell \in K[X]$ satisfying $hG^2 = \sum_{i=1}^{\ell} a_i F_i^2$. We prove this: Write $f_i = q_i g + r_i$ with $q_i, r_i \in K[X]$ and $\deg r_i < \deg g$ for all $i \in \{1, \ldots, \ell\}$. If $r_i = 0$ for all $i \in \{1, \ldots, \ell\}$, then we set $G := 1$ and $F_i := q_i$ for all $i \in \{1, \ldots, \ell\}$ and have

$$hG^2 = h = \frac{1}{g^2}(hg^2) = \frac{1}{g^2} \sum_{i=1}^{\ell} a_i f_i^2 = \sum_{i=1}^{\ell} a_i \left(\frac{f_i}{g}\right)^2 = \sum_{i=1}^{\ell} a_i q_i^2 = \sum_{i=1}^{\ell} a_i F_i^2.$$

In the sequel, we suppose that the set $I := \{i \in \{1, \ldots, \ell\} \mid r_i \neq 0\}$ is nonempty. Now we set $s := \sum_{i=1}^{\ell} a_i q_i^2 - h$, $t := \sum_{i=1}^{\ell} a_i f_i q_i - gh$, $F_i := sf_i - 2tq_i$ for $i \in \{1, \ldots, \ell\}$ and $G := sg - 2t$. Then we obtain

$$\begin{aligned}
hG^2 &= s^2 hg^2 - 4stgh + 4t^2 h \\
&= s^2 \sum_{i=1}^{\ell} a_i f_i^2 - 4st(t + gh) + 4t^2(s + h) \\
&= s^2 \sum_{i=1}^{\ell} a_i f_i^2 - 4st \sum_{i=1}^{\ell} a_i f_i q_i + 4t^2 \sum_{i=1}^{\ell} a_i q_i^2 \\
&= \sum_{i=1}^{\ell} a_i (sf_i - 2tq_i)^2 = \sum_{i=1}^{\ell} a_i F_i^2.
\end{aligned}$$

It remains to show that $G \neq 0$ and $\deg G < \deg g$. To this end, we calculate

$$\begin{aligned}
G &= g \sum_{i=1}^{\ell} a_i q_i^2 - gh - 2 \sum_{i=1}^{\ell} a_i f_i q_i + 2gh \\
&= \frac{1}{g} \left(g^2 \sum_{i=1}^{\ell} a_i q_i^2 + g^2 h - 2g \sum_{i=1}^{\ell} a_i f_i q_i \right) \\
&= \frac{1}{g} \left(g^2 \sum_{i=1}^{\ell} a_i q_i^2 + \sum_{i=1}^{\ell} a_i f_i^2 - 2g \sum_{i=1}^{\ell} a_i f_i q_i \right) \\
&= \frac{1}{g} \sum_{i=1}^{\ell} a_i (g^2 q_i^2 - 2(gq_i) f_i + f_i^2) \\
&= \frac{1}{g} \sum_{i=1}^{\ell} a_i (gq_i - f_i)^2 = \frac{1}{g} \sum_{i=1}^{\ell} a_i r_i^2 = \frac{1}{g} \sum_{i \in I} a_i r_i^2.
\end{aligned}$$

If we had $G = 0$, then this would mean $\sum_{i \in I} a_i r_i^2 = 0$. Since the leading coefficient of $a_i r_i^2$ is positive for all $i \in I \neq \emptyset$, this is impossible. Hence $G \neq 0$. Because of $\deg r_i < \deg g$ for all $i \in I$, we have $\deg G < 2 \deg g - \deg g = \deg g$. □

2.2 Homogenization and dehomogenization

Definition 2.2.1. [→ 1.6.1] Let A be commutative ring with $0 \neq 1$.

(a) If $k \in \mathbb{N}_0$ and $f \in A[X_1, \ldots, X_n]$, then the sum of all terms (i.e., monomials with their coefficients) of degree k of f is called the *k-th homogeneous part* of f. This is a k-form [→ 1.6.1(a)].

(b) If $f \in A[X_1, \ldots, X_n] \setminus \{0\}$ and $d := \deg f$, then the d-th homogeneous part of f is called the *leading form* $\mathrm{LF}(f)$ of f. We set $\mathrm{LF}(0) := 0$.

(c) If $f \in A[X_1, \ldots, X_n]$, $d := \deg f \in \mathbb{N}_0$ and $f = \sum_{k=0}^{d} f_k$ with a k-form f_k for all $k \in \{0, \ldots, d\}$, then the *homogenization* $f^* \in A[X_0, \ldots, X_n]$ of f (with respect to X_0) is given by $f^* := \sum_{k=0}^{d} X_0^{d-k} f_k = X_0^d f\left(\frac{X_1}{X_0}, \ldots, \frac{X_n}{X_0}\right)$. We set $0^* := 0$.

(d) For homogeneous $f \in A[X_0, \ldots, X_n]$, we call $\widetilde{f} := f(1, X_1, \ldots, X_n)$ the *dehomogenization* of f (with respect to X_0).

Remark 2.2.2. Let A be a commutative ring with $0 \neq 1$.

(a) $\mathrm{LF}(f) = f^*(0, X_1, \ldots, X_n)$ for all $f \in A[X_1, \ldots, X_n]$.

(b) For $f, g \in A[X_1, \ldots, X_n]$, we have $(f+g)^* = f^* + g^*$ in case $\deg f = \deg g = \deg(f+g)$ and we always have $(fg)^* = f^* g^*$.

(c) $A[X_0, \ldots, X_n] \to A[X_1, \ldots, X_n]$, $f \mapsto \widetilde{f}$ is a ring homomorphism.

(d) For all $f, g \in A[X_1, \ldots, X_n]$, we have $\mathrm{LF}(f+g) = \mathrm{LF}(f) + \mathrm{LF}(g)$ in case $\deg f = \deg g = \deg(f+g)$ and we always have $\mathrm{LF}(fg) = \mathrm{LF}(f)\,\mathrm{LF}(g)$.

(e) For all $f \in A[X_1, \ldots, X_n]$, we have $\widetilde{f^*} = f$.

(f) If $f \in A[X_0, \ldots, X_n] \setminus \{0\}$ is homogeneous and $m := \max\{k \in \mathbb{N}_0 \mid X_0^k \mid f\}$, then $X_0^m \widetilde{f}^{\,*} = f$.

Lemma 2.2.3. Suppose K is a field, $n, d \in \mathbb{N}_0$, $f \in K[X_1, \ldots, X_n]_d$ and let $I_1, \ldots, I_n \subseteq K$ be sets of cardinality at least $d+1$ each such that $f(x) = 0$ for all $x \in I_1 \times \ldots \times I_n$. Then $f = 0$.

Proof. Induction by n.

$\underline{n = 0}$ ✓

$\underline{n - 1 \to n \quad (n \in \mathbb{N})}$ Write $f = \sum_{k=0}^{d} f_k X_n^k$ with $f_k \in K[X_1, \ldots, X_{n-1}]_d$. For all

$$(x_1, \ldots, x_{n-1}) \in I_1 \times \ldots \times I_{n-1},$$

the polynomial $f(x_1, \ldots, x_{n-1}, X_n) = \sum_{k=0}^{d} f_k(x_1, \ldots, x_{n-1}) X_n^k \in K[X_n]_d$ is a polynomial with at $d+1$ roots. Thus $f_k(x_1, \ldots, x_{n-1}) = 0$ for all $k \in \{0, \ldots, d\}$ and $(x_1, \ldots, x_{n-1}) \in I_1 \times \ldots \times I_{n-1}$. By induction hypothesis, $f_k = 0$ for all $k \in \{0, \ldots, d\}$. □

Remark 2.2.4. Let K be a real field, $\ell, n \in \mathbb{N}_0$, $p_1, \ldots, p_\ell \in K[X_1, \ldots, X_n]$ and

$$f := \sum_{i=1}^{\ell} p_i^2.$$

(a) If $f = 0$, then $p_1 = \ldots = p_\ell = 0$. This follows from 2.2.3 together with 1.2.12(c). Instead of 2.2.3, one can alternatively employ the fact that $K(X_1, \ldots, X_n)$ is real which is clear by applying 1.3.7 n times.

(b) If $f \neq 0$, then $\deg f = 2d$ with $d := \max\{\deg(p_i) \mid i \in \{1, \ldots, \ell\}\}$ since otherwise $\sum_{i=1, \deg(p_i)=d}^{\ell} \mathrm{LF}(p_i)^2 = 0$, contradicting (a).

(c) If $d \in \mathbb{N}_0$ and f is a $2d$-form, then every p_i is a d-form. This can be seen similarly to (b) by considering the homogeneous parts of the p_i of smallest (instead of largest) degree.

(d) We have $f^* \in \sum K[X_0, \ldots, X_n]^2$. More precisely, f^* is a $2d$-form for some $d \in \mathbb{N}_0$ that is a sum of ℓ squares of d-forms since

$$f^* = X_0^{2d} f\left(\frac{X_1}{X_0}, \ldots, \frac{X_n}{X_0}\right) = \sum_{i=1}^{\ell} \left(X_0^d p_i\left(\frac{X_1}{X_0}, \ldots, \frac{X_n}{X_0}\right)\right)^2$$

and $X_0^d p_i\left(\frac{X_1}{X_0}, \ldots, \frac{X_n}{X_0}\right) = X_0^{d - \deg p_i} p_i^* \in K[X_0, \ldots, X_n]$ for all $i \in \{1, \ldots, \ell\}$ with $p_i \neq 0$ (note that $\deg p_i \leq d$ by (b)).

Proposition 2.2.5. *Let $(K, \leq)$ be an ordered field and $f \in K[X_1, \ldots, X_n]$ with $f \geq 0$ on K^n. Then f has an even degree except if $f = 0$, and we have $\mathrm{LF}(f) \geq 0$ on K^n.*

Proof. WLOG $f \neq 0$. Then $g := \mathrm{LF}(f) \neq 0$. Set $d := \deg g$. For all $x \in K^n$, $f_x := f(Tx) \in K[T]$ is a polynomial in one variable with $f_x \geq 0$ on K whose leading coefficient is $g(x)$ in case that $g(x) \neq 0$. Choose $x_0 \in K^n$ with $g(x_0) \neq 0$ [$\to$ 2.2.3]. Then f_{x_0} has degree d and because of $f_{x_0} \geq 0$ on K, it follows that $d \in 2\mathbb{N}_0$ by 1.5.3(a). Now let $x \in K^n$ be arbitrary such that $g(x) \neq 0$. Again by 1.5.3(a), it follows from $f_x \geq 0$ on K that $g(x) \geq 0$. □

Proposition 2.2.6. *Let $(K, \leq)$ be an ordered field and $f \in K[X_1, \ldots, X_n]$.*

(a) *$f \geq 0$ on $K^n \iff f^* \geq 0$ on K^{n+1}*

(b) *$f \in \sum K[X_1, \ldots, X_n]^2 \iff f^* \in \sum K[X_0, \ldots, X_n]^2$*

Proof. (a) "$\Longleftarrow$" If f^* is nonnegative on K^{n+1}, then also on $\{1\} \times K^n$.

"$\Longrightarrow$" Suppose $f \geq 0$ on K^n. WLOG $f \neq 0$. By 2.2.5, we can write $\deg f = 2d$ with $d \in \mathbb{N}_0$. Due to $f^* \overset{2.2.1(c)}{=} X_0^{2d} f\left(\frac{X_1}{X_0}, \ldots, \frac{X_n}{X_0}\right)$, we deduce $f^* \geq 0$ on $K^\times \times K^n$. It remains to show $f^* \geq 0$ on $\{0\} \times K^n$ which is equivalent by 2.2.2(a) to $\mathrm{LF}(f) \geq 0$ on K^n. The latter holds by 2.2.5.

(b) "$\Longrightarrow$" has been shown in 2.2.4(d).

"$\Longleftarrow$" follows from 2.2.2(c). □

2.3 Nonnegative quadratic polynomials

Definition 2.3.1. Let $(K, \leq)$ be an ordered field.

(a) If $f \in K[X_1, \ldots, X_n]$ is homogeneous [→ 1.6.1(a)], then f is called $\left\{\begin{matrix}\textit{positive semidefinite (psd)}\\ \textit{positive definite (pd)}\end{matrix}\right\}$ (over K) if f $\left\{\begin{matrix}\geq 0 \text{ on } K^n\\ > 0 \text{ on } K^n \setminus \{0\}\end{matrix}\right\}$.

(b) If $M \in SK^{n\times n}$, then M is called $\left\{\begin{matrix}\text{psd}\\ \text{pd}\end{matrix}\right\}$ (over K) if the quadratic form represented by M [→ 1.6.1(d)] is $\left\{\begin{matrix}\text{psd}\\ \text{pd}\end{matrix}\right\}$, i.e., $x^T M x$ $\left\{\begin{matrix}\geq 0 \text{ for all } x \in K^n\\ > 0 \text{ for all } x \in K^n \setminus \{0\}\end{matrix}\right\}$.

Proposition 2.3.2. *Let K be an Euclidean field and $q \in K[X_1, \ldots, X_n]$ a quadratic form. Then the following are equivalent:*

(a) *q is psd* [→ 2.3.1(a)]

(b) $q \in \sum K[X_1, \ldots, X_n]^2$ [→ 1.1.18]

(c) *q is a sum of n squares of linear forms* [→ 1.6.1(a)].

(d) $\operatorname{sg} q = \operatorname{rk} q$ [→ 1.6.1(h)].

Proof. (d) $\Longrightarrow$ (c) $\Longrightarrow$ (b) $\Longrightarrow$ (a) is trivial. Now suppose that (d) does not hold. We show that then (a) also fails. Write $q = \sum_{i=1}^{s} \ell_i^2 - \sum_{j=1}^{t} \ell_{s+j}^2$ with $s, t \in \mathbb{N}_0$ and linearly independent linear forms $\ell_1, \ldots, \ell_s, \ell_{s+1}, \ldots, \ell_{s+t} \in K[X_1, \ldots, X_n]$. Since $s - t = \operatorname{sg} q \neq \operatorname{rk} q = s + t$, we have $t \geq 1$. By linear algebra,

$$\varphi \colon K^n \to K^{s+t},\ x \mapsto \begin{pmatrix} \ell_1(x) \\ \vdots \\ \ell_{s+t}(x) \end{pmatrix}$$

is surjective. Choose $x \in K^n$ with $\varphi(x) = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}$. Then $q(x) = -1 < 0$. □

Proposition 2.3.3. *Let K be an Euclidean field and $M \in SK^{n\times n}$. Then the following are equivalent:*

(a) *M is psd* [→ 2.3.1(b)].

(b) $\exists s \in \mathbb{N}_0 : \exists A \in K^{s\times n} : M = A^T A$

(c) $\exists A \in K^{n\times n} : M = A^T A$

(d) *All eigenvalues of M in the real closure $\overline{(K, K^2)}$ are nonnegative.*

(e) *All coefficients of* $\det(M + XI_n) \in K[X]$ *are nonnegative.*

(f) *If* $M = (a_{ij})_{1 \le i,j \le n}$*, then for all* $I \subseteq \{1, \dots, n\}$*, we have* $\det((a_{ij})_{(i,j) \in I \times I}) \ge 0$.

Proof. Using 1.6.1(e) and 2.2.4(c), one sees that (a), (b) and (c) are nothing else than the corresponding statements in 2.3.2.

(a) $\Longrightarrow$ (f) follows from applying (a) $\Longrightarrow$ (c) to the submatrices of M in question.

(f) $\Longrightarrow$ (e) Each coefficients of $\det(M + XI_n)$ is a sum of certain determinants appearing in (f).

(e) $\Longrightarrow$ (d) is trivial.

(d) $\Longrightarrow$ (a) follows easily from 1.8.11. □

Terminology 2.3.4. [→ 1.5.1, 1.6.1(a)] Let A be a commutative ring with $0 \neq 1$. Polynomials from $A[X_1, \dots, X_n]_d$ [→ 1.5.1] are called *constant* for $d = 0$, *linear* for $d = 1$, *quadratic* for $d = 2$, *cubic* for $d = 3$, *quartic* for $d = 4$, *quintic* for $d = 5, \dots$

Proposition 2.3.5. *Let* K *be an Euclidean field and* $q \in K[X_1, \dots, X_n]_2$*. The following are equivalent:*

(a) $q \ge 0$ *on* K^n

(b) $q \in \sum K[X_1, \dots, X_n]^2$

(c) q *is a sum of* $n + 1$ *squares of linear polynomials.*

Proof. (a) $\overset{2.2.6(a)}{\Longrightarrow}$ $q^* \ge 0$ on K^{n+1} $\overset{2.3.2}{\Longrightarrow}$ (c) $\Longrightarrow$ (b) $\Longrightarrow$ (a) □

2.4 The Newton polytope

Definition and Proposition 2.4.1. *Let* $(K, \le)$ *be an ordered field,* V *a* K*-vector space and* $A \subseteq V$*. Then* A *is called* convex *if* $\forall x, y \in A : \forall \lambda \in [0,1]_K : \lambda x + (1 - \lambda) y \in A$*. The smallest convex superset of* A *is obviously*

$$\operatorname{conv} A := \left\{ \sum_{i=1}^{m} \lambda_i x_i \mid m \in \mathbb{N}, \lambda_i \in K_{\ge 0}, x_i \in A, \sum_{i=1}^{m} \lambda_i = 1 \right\},$$

called the convex set generated *by* A *or the* convex hull *of* A*. We call finitely generated convex sets, i.e., convex hulls of finite sets,* polytopes*. A polytope is thus of the form*

$$\operatorname{conv}\{x_1, \dots, x_m\} = \left\{ \sum_{i=1}^{m} \lambda_i x_i \mid \lambda_i \in K_{\ge 0}, \sum_{i=1}^{m} \lambda_i = 1 \right\}$$

for some $m \in \mathbb{N}_0$ *and* $x_1, \dots, x_m \in V$*. If* A *is a convex set, then a point* $x \in A$ *is called an* extreme point *of* A *if there are no* $y, z \in A$ *such that* $y \neq z$ *and* $x = \frac{y+z}{2}$*. Extreme points of polytopes are also called* vertices *of the polytope.*

Exercise 2.4.2. Suppose $(K, \leq)$ is an ordered field, V a K-vector space, $A \subseteq V$, $x \in A$ and $\lambda \in (0,1)_K$. Then the following are equivalent:

(a) x is an extreme point of A.

(b) There are no $y, z \in A$ such that $y \neq z$ and $x = \lambda y + (1-\lambda)z$.

Lemma 2.4.3. Let $(K, \leq)$ be an ordered field, V a K-vector space, $m \in \mathbb{N}_0$, $x_1, \ldots, x_m \in V$, $P := \operatorname{conv}\{x_1, \ldots, x_m\}$ and suppose $P \neq \operatorname{conv}(\{x_1, \ldots, x_m\} \setminus \{x_i\})$ for all $i \in \{1, \ldots, m\}$. Then P is a polytope and $x_1, \ldots, x_m$ are its vertices.

Proof. To show:

(a) Every vertex of P equals one of the x_i.

(b) Every x_i is a vertex of P.

For (a), let x be a vertex of P. Write $x = \sum_{i=1}^m \lambda_i x_i$ with $\lambda_i \in K_{\geq 0}$ and $\sum_{i=1}^m \lambda_i = 1$. WLOG $\lambda_1 \neq 0$. Then $\lambda_1 = 1$ for otherwise $\mu := \sum_{i=2}^m \lambda_i = 1 - \lambda_1 > 0$ and $x = \lambda_1 x_1 + \mu \underbrace{\left(\sum_{i=2}^m \frac{\lambda_i}{\mu} x_i \right)}_{\in \operatorname{conv}\{x_2, \ldots, x_m\}}$, contradicting 2.4.2(b).

To prove (b), we let $y, z \in P$ with $x_1 = \frac{y+z}{2}$. To show: $y = z$. Write $y = \sum_{i=1}^m \lambda_i x_i$ and $z = \sum_{i=1}^m \mu_i x_i$ with $\lambda_i, \mu_i \in K_{\geq 0}$ and $\sum_{i=1}^m \lambda_i = 1 = \sum_{i=1}^m \mu_i$. We show that $\lambda_1 = 1 = \mu_1$. It is enough to show $\frac{\lambda_1 + \mu_1}{2} = 1$. If we had $\frac{\lambda_1 + \mu_1}{2} < 1$, then it would follow from $(1 - \frac{\lambda_1 + \mu_1}{2}) x_1 = \sum_{i=2}^m \frac{\lambda_i + \mu_i}{2} x_i$ that $x_1 \in \operatorname{conv}\{x_2, \ldots, x_m\}$ and therefore $P = \operatorname{conv}\{x_1, \ldots, x_m\} = \operatorname{conv}\{x_2, \ldots, x_m\}$ ↯. □

Corollary 2.4.4. *Every polytope is the convex hull of its finitely many vertices.*

Definition and Proposition 2.4.5. *Suppose $(K, \leq)$ is an ordered field, V is a K-vector space and let A and B be subsets of V. Then $A + B := \{x + y \mid x \in A, y \in B\}$ is called the* Minkowski sum *of A and B. We have $(\operatorname{conv} A) + (\operatorname{conv} B) = \operatorname{conv}(A + B)$. Let now A and B be convex. Then $A + B$ is also convex. If z is an extreme point of $A + B$, then there are uniquely determined $x \in A$ and $y \in B$ such that $z = x + y$, and x is an extreme point of A and y is one of B.*

Proof. "$\subseteq$" Let $x_1, \ldots, x_m \in A$, $y_1, \ldots, y_n \in B$, $\lambda_1, \ldots, \lambda_m \in K_{\geq 0}$, $\mu_1, \ldots, \mu_n \in K_{\geq 0}$ and $\sum_{i=1}^m \lambda_i = 1 = \sum_{j=1}^n \mu_j$. Then $\sum_{i=1}^m \sum_{j=1}^n \lambda_i \mu_j = (\sum_{i=1}^m \lambda_i) \left(\sum_{j=1}^n \mu_j \right) = 1 \cdot 1 = 1$ and

$$\sum_{i=1}^m \lambda_i x_i + \sum_{j=1}^n \mu_j y_j = \left(\sum_{j=1}^n \mu_j \right) \sum_{i=1}^m \lambda_i x_i + \left(\sum_{i=1}^m \lambda_i \right) \sum_{j=1}^n \mu_j y_j = \sum_{i=1}^m \sum_{j=1}^n \lambda_i \mu_j (x_i + y_j).$$

"$\supseteq$" is trivial.

Let now A and B be convex. Then $A + B = (\operatorname{conv} A) + (\operatorname{conv} B) = \operatorname{conv}(A + B)$ is convex. Finally, let z be an extreme point of $A + B$ and let $x \in A$ and $y \in B$ with

$z = x + y$. Then x is an extreme point of A since if we had $x = \frac{x_1+x_2}{2}$ with different $x_1, x_2 \in A$, then it would follow that $z = \frac{(x_1+y)+(x_2+y)}{2}$ and $x_1 + y \neq x_2 + y$ ↯. In the same way, y is an extreme point of B. Suppose now that $x' \in A$ and $y' \in B$ such that $z = x' + y'$. Then $z = \frac{x+x'}{2} + \frac{y+y'}{2}$ and $\frac{x+x'}{2}$ is also an extreme point of A which is possible only for $x = x'$. Analogously, $y = y'$. □

Notation 2.4.6. Suppressing n in the notation, we denote by $\underline{X} := (X_1, \ldots, X_n)$ a tuple of variables and set $A[\underline{X}] := A[X_1, \ldots, X_n]$ for every commutative ring A with $0 \neq 1$ in A. Für $\alpha \in \mathbb{N}_0^n$, we write $|\alpha| := \alpha_1 + \cdots + \alpha_n$ and $\underline{X}^\alpha := X_1^{\alpha_1} \cdots X_n^{\alpha_n}$.

Definition 2.4.7. Let K be a field and $f \in K[\underline{X}]$. Write $f = \sum_{\alpha \in \mathbb{N}_0^n} a_\alpha \underline{X}^\alpha$ with $a_\alpha \in K$. Then the finite set $\operatorname{supp}(f) := \{\alpha \in \mathbb{N}_0^n \mid a_\alpha \neq 0\}$ is called the *support* of f and its convex hull $N(f) := \operatorname{conv}(\operatorname{supp}(f)) \subseteq \mathbb{R}^n$ the *Newton polytope* of f.

Definition 2.4.8. Let K be a field, $f \in K[\underline{X}]$ and $a \in K$. We say that a is a *vertex coefficient* of f if there is a vertex α of $N(f)$ such that $a\underline{X}^\alpha$ is a term of f.

Remark 2.4.9. Since every vertex of the Newton polytope of a polynomial lies by 2.4.3 in the support of the polynomial, vertex coefficients are always $\neq 0$.

Theorem 2.4.10. *Let K be a field and $f, g \in K[\underline{X}]$. Then $N(fg) = N(f) + N(g)$ and every vertex coefficient of fg is the product of a vertex coefficient of f with a vertex coefficient of g.*

Proof. "$\subseteq$" $\operatorname{supp}(fg) \subseteq \operatorname{supp}(f) + \operatorname{supp}(g) \subseteq N(f) + N(g)$ and therefore $N(fg) = \operatorname{conv}(\operatorname{supp}(fg)) \subseteq N(f) + N(g)$ since $N(f) + N(g)$ is convex by 7.4.19.

"$\supseteq$" By 7.4.19, $N(f) + N(g)$ is a polytope. By virtue of 2.4.4, it suffices to show that its vertices lie in $N(fg)$. Consider therefore a vertex γ of $N(f) + N(g)$. We even show that $\gamma \in \operatorname{supp}(fg)$. By 7.4.19, there are uniquely determined $\alpha \in N(f)$ and $\beta \in N(g)$ such that $\gamma = \alpha + \beta$, and α is a vertex of $N(f)$ and β a vertex of $N(g)$. By 2.4.9, we have $\alpha \in \operatorname{supp}(f)$ and $\beta \in \operatorname{supp}(g)$. Because of unicity of α and β, the coefficient of $\underline{X}^\gamma$ in fg equals the product of the respective coefficients of $\underline{X}^\alpha$ and $\underline{X}^\beta$ in f and g, respectively, and hence is in particular $\neq 0$. Thus $N(fg) = N(f) + N(g)$ is shown. Also the extra claim follows from the above. □

Proposition 2.4.11. *Let K be a field and $f, g \in K[\underline{X}]$. Then $N(f+g) \subseteq \operatorname{conv}(N(f) \cup N(g))$.*

Proof. $\operatorname{supp}(f+g) \subseteq \operatorname{supp}(f) \cup \operatorname{supp}(g) \subseteq N(f) \cup N(g)$ implies

$$N(f+g) = \operatorname{conv}(\operatorname{supp}(f+g)) \subseteq \operatorname{conv}(N(f) \cup N(g)).$$

□

Theorem 2.4.12. *Let $(K, \leq)$ be an ordered field and $f, g \in K[\underline{X}]$ such that all vertex coefficients of f and g have the same sign. Then $N(f+g) = \operatorname{conv}(N(f) \cup N(g))$ and all vertex coefficients of $f+g$ also have this sign.*

Proof. "$\subseteq$" is 2.4.11

"$\supseteq$" We have that $\operatorname{conv}(N(f) \cup N(g)) = \operatorname{conv}(\operatorname{supp}(f) \cup \operatorname{supp}(g))$ is a polytope. Let α be one of its vertices. By 2.4.4, it is enough to show that $\alpha \in N(f+g)$. We even show that $\alpha \in \operatorname{supp}(f+g)$. By 2.4.3, α lies in at least one of the sets $\operatorname{supp}(f)$ and $\operatorname{supp}(g)$. If α lies only in one of these two, then the claim is clear. If on the other hand α lies in both, then α is a vertex of both $\operatorname{conv}(\operatorname{supp}(f)) = N(f)$ and $\operatorname{conv}(\operatorname{supp}(g)) = N(g)$ and the coefficients of $\underline{X}^\alpha$ in f and in g and hence also in $f+g$ have the same sign, from which it follows again that $\alpha \in \operatorname{supp}(f+g)$. Thus $N(f+g) = \operatorname{conv}(N(f) \cup N(g))$ is proven. The extra claim follows from what was shown. □

Lemma 2.4.13. Let $(K, \leq)$ be an ordered field, V a K-vector space and A a convex subset of V. Then $A + A = 2A := \{2x \mid x \in A\}$.

Proof. "$\supseteq$" trivial

"$\subseteq$" Let $x, y \in A$. Then $x + y = 2\frac{x+y}{2} \in 2A$. □

Theorem 2.4.14. *Let $(K, \leq)$ be an ordered field and $f \in K[\underline{X}]$. Then $N(f^2) = 2N(f)$ and all vertex coefficients of f^2 are squares of vertex coefficients of f and therefore positive.*

Proof. $N(f^2) = 2N(f)$ follows from 2.4.10 and 2.4.13. Suppose γ is a vertex of $N(f^2) \overset{2.4.10}{=} N(f) + N(f)$. By 7.4.19, there are uniquely determined $\alpha, \beta \in N(f)$ with $\gamma = \alpha + \beta$. Due to $\gamma = \beta + \alpha$, it follows that $\alpha = \beta$. But then the coefficient of $\underline{X}^\gamma$ in f^2 is just the coefficient belonging to $\underline{X}^\alpha$ in f squared. □

Theorem 2.4.15. *Let $(K, \leq)$ be an ordered field, $\ell \in \mathbb{N}_0$, $p_1, \ldots, p_\ell \in K[\underline{X}]$ and $f := \sum_{i=1}^{\ell} p_i^2$. Then $N(f) = 2\operatorname{conv}(N(p_1) \cup \ldots \cup N(p_\ell))$ and all vertex coefficients of f are positive.*

Proof. For each $i \in \{1, \ldots, \ell\}$, we have by 2.4.14 that $N(p_i^2) = 2N(p_i)$ and that all vertex coefficients of p_i^2 are positive. By 2.4.12,

$$\begin{aligned} N(f) &= \operatorname{conv}(N(p_1^2) \cup \ldots \cup N(p_\ell^2)) = \operatorname{conv}(2N(p_1) \cup \ldots \cup 2N(p_\ell)) \\ &= 2\operatorname{conv}(N(p_1) \cup \ldots \cup N(p_\ell)) \end{aligned}$$

and all vertex coefficients of f are positive. □

Example 2.4.16. For the *Motzkin polynomial* $f := X^4Y^2 + X^2Y^4 - 3X^2Y^2 + 1 \in \mathbb{R}[X, Y]$, we have $f \geq 0$ on $\mathbb{R}^2$ but $f \notin \sum \mathbb{R}[X, Y]^2$. At first we show $f \geq 0$ on $\mathbb{R}^2$ in three different ways:

(1) From the inequality of arithmetic and geometric means known from analysis, it follows that $\sqrt[3]{abc} \leq \frac{1}{3}(a+b+c)$ for all $a, b, c \in \mathbb{R}_{\geq 0}$. Setting here $a := x^4y^2$, $b := x^2y^4$ and $c := 1$ for arbitrary $x, y \in \mathbb{R}$, we deduce $x^2y^2 \leq \frac{1}{3}(x^4y^2 + x^2y^4 + 1)$.

(2)

$$\begin{aligned} (1+X^2)f &= X^4Y^2 + X^2Y^4 - 3X^2Y^2 + 1 + X^6Y^2 + X^4Y^4 - 3X^4Y^2 + X^2 \\ &= 1 - 2X^2Y^2 + X^4Y^4 + X^2 - 2X^2Y^2 + X^2Y^4 + X^2Y^2 - 2X^4Y^2 + X^6Y^2 \\ &= (1 - X^2Y^2)^2 + X^2(1 - Y^2)^2 + X^2Y^2(1 - X^2)^2 \in \sum \mathbb{R}[X, Y]^2 \end{aligned}$$

(3)

$$\begin{aligned}
f(X^3,Y^3) &= X^{12}Y^6 + X^6Y^{12} - 3X^6Y^6 + 1 \\
&= X^4Y^2 - X^8Y^4 - X^6Y^6 + \frac{1}{4}X^{12}Y^6 + \frac{1}{2}X^{10}Y^8 + \frac{1}{4}X^8Y^{10} \\
&\quad + X^2Y^4 - X^6Y^6 - X^4Y^8 + \frac{1}{4}X^{10}Y^8 + \frac{1}{2}X^8Y^{10} + \frac{1}{4}X^6Y^{12} \\
&\quad + 1 - X^4Y^2 - X^2Y^4 + \frac{1}{4}X^8Y^4 + \frac{1}{2}X^6Y^6 + \frac{1}{4}X^4Y^8 \\
&\quad + \frac{3}{4}X^8Y^4 - \frac{3}{2}X^6Y^6 + \frac{3}{4}X^4Y^8 \\
&\quad + \frac{3}{4}X^{10}Y^8 - \frac{3}{2}X^8Y^{10} + \frac{3}{4}X^6Y^{12} \\
&\quad + \frac{3}{4}X^{12}Y^6 - \frac{3}{2}X^{10}Y^8 + \frac{3}{4}X^8Y^{10} \\
&= \left(X^2Y - \frac{1}{2}X^4Y^5 - \frac{1}{2}X^6Y^3\right)^2 \\
&\quad + \left(XY^2 - \frac{1}{2}X^3Y^6 - \frac{1}{2}X^5Y^4\right)^2 \\
&\quad + \left(1 - \frac{1}{2}X^2Y^4 - \frac{1}{2}X^4Y^2\right)^2 \\
&\quad + \frac{3}{4}\left(X^2Y^4 - X^4Y^2\right)^2 \\
&\quad + \frac{3}{4}(X^3Y^6 - X^5Y^4)^2 \\
&\quad + \frac{3}{4}(X^4Y^5 - X^6Y^3)^2
\end{aligned}$$

Now we show $f \notin \sum \mathbb{R}[X,Y]^2$:

$$\begin{aligned}
N(f) &= \operatorname{conv}(\operatorname{supp}(f)) = \operatorname{conv}\{(4,2),(2,4),(2,2),(0,0)\} \\
&= \operatorname{conv}\{(4,2),(2,4),(0,0)\}.
\end{aligned}$$

Assume $f = \sum_{i=1}^{\ell} p_i^2$ with $\ell \in \mathbb{N}_0$ and $p_1, \ldots, p_\ell \in \sum \mathbb{R}[X,Y]$. Then

$$N(p_i) \subseteq \operatorname{conv}(N(p_1) \cup \ldots \cup N(p_\ell)) = \frac{1}{2}N(f) = \operatorname{conv}\{(2,1),(1,2),(0,0)\}$$

by 2.4.15 and hence $\operatorname{supp}(p_i) \subseteq \mathbb{N}_0^2 \cap N(p_i) \subseteq \mathbb{N}_0^2 \cap \operatorname{conv}\{(2,1),(1,2),(0,0)\} = \{(0,0),(1,1),(2,1),(1,2)\}$ for all $i \in \{1,\ldots,\ell\}$. The coefficient of X^2Y^2 in p_i^2 is therefore the coefficient of XY in p_i squared and therefore nonnegative. Then the coefficient of X^2Y^2 in f is also nonnegative $\lightning$. This shows $f \notin \sum \mathbb{R}[X,Y]^2$. Thus one can neither generalize 2.1.1(a) $\Longrightarrow$ (c) to polynomials in several variables nor 2.3.5(a) $\Longrightarrow$ (b) to polynomials of arbitrary degree. Note also that exactly the same

proof shows even $f + c \notin \sum \mathbb{R}[X,Y]^2$ for all $c \in \mathbb{R}$. By 2.2.6, the *Motzkin form* $f^* := X^4Y^2 + X^2Y^4 - 3X^2Y^2Z^2 + Z^6$ is psd [→ 2.3.1] but is likewise no sum of squares of polynomials. Again by 2.2.6, the dehomogenizations $f^*(1,Y,Z) = Y^2 + Y^4 - 3Y^2Z^2 + Z^6$ and $f^*(X,1,Z) = X^4 + X^2 - 3X^2Z^2 + Z^6$ are also polynomials that are ≥ 0 on $\mathbb{R}^2$ but that are no sums of squares of polynomials.

2.5 Artin's solution to Hilbert's 17th problem

Lemma 2.5.1. Let R be a real closed field and $f, p, q \in R[\underline{X}]$. Suppose $q \neq 0$, $f = \frac{p}{q}$, $p \geq 0$ on R^n and $q \geq 0$ on R^n. Then $f \geq 0$ on R^n.

Proof. Using the Tarski principle 1.8.19, one can reduce to the case $R = \mathbb{R}$. But then the subset $\{x \in \mathbb{R}^n \mid f(x) < 0\}$ of $\{x \in \mathbb{R}^n \mid q(x) = 0\}$ is open in $\mathbb{R}^n$ and therefore empty since otherwise $q = 0$ would follow from 2.2.3. □

In the year 1900, Hilbert presented his famous list of 23 seminal problems at the International Congress of Mathematicians in Paris. In 1927, Artin gave a positive solution to the 17th of these problems. This corresponds to the case $K = \mathbb{R}$ in the following theorem.

Theorem 2.5.2 (Artin)**.** *Suppose R is a real closed field and $(K, \leq)$ an ordered subfield of R. Let $f \in K[\underline{X}]$. Then the following are equivalent:*

(a) $f \geq 0$ *on* R^n

(b) $f \in \sum K_{\geq 0} K(\underline{X})^2$

Proof. (b) $\Longrightarrow$ (a) follows from Lemma 2.5.1. We show (a) $\Longrightarrow$ (b) by contraposition. Suppose $f \notin \sum K_{\geq 0} K(\underline{X})^2$. To show: $\exists x \in R^n : f(x) < 0$. Since $\sum K_{\geq 0} K(\underline{X})^2$ is now a proper preorder of $K(\underline{X})$ [→ 1.2.1, 1.2.5], there is by 1.2.10 an order P of $K(\underline{X})$ with $f \notin P$. Set $R' := \overline{(K(\underline{X}), P)}$. Then there is an $x \in R'^n$ with $f(x) < 0$ namely $x := (X_1, \ldots, X_n)$ since $f(x) = f < 0$ in R'. Due to $K_{\geq 0} \subseteq P \subseteq R'^2$, $(K, \leq)$ is an *ordered* subfield of R'. Since the K-semialgebraic set $\{x \in R'^n \mid f(x) < 0\}$ is nonempty, its transfer $\{x \in R^n \mid f(x) < 0\}$ to R [→ 1.9.5] is also nonempty. □

Corollary 2.5.3. [→ 2.1.2] *Suppose R is a real closed field and $(K, \leq)$ an ordered subfield of R. Let $f \in K[X]$. Then the following are equivalent:*

(a) $f \geq 0$ *on* R

(b) $f \in \sum K_{\geq 0} K[X]^2$

Proof. (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (b) follows from 2.5.2 and 2.1.2. □

2.6 The Gram matrix method

Theorem 2.6.1. *Let K be an Euclidean field, $f \in K[\underline{X}]$ and $\frac{1}{2}N(f) \cap \mathbb{N}_0^n \subseteq \{\alpha_1, \dots, \alpha_m\} \subseteq \mathbb{N}_0^n$ (for instance set $\{\alpha_1, \dots, \alpha_m\}$ equal to $\frac{1}{2}N(f) \cap \mathbb{N}_0^n$ or to $\{\alpha \in \mathbb{N}_0^n \mid 2|\alpha| \leq \deg f\}$). Set $v := \begin{pmatrix} \underline{X}^{\alpha_1} \\ \vdots \\ \underline{X}^{\alpha_m} \end{pmatrix}$. Then the following are equivalent:*

(a) $f \in \sum K[\underline{X}]^2$

(b) *There is a* psd *matrix* [→ 2.3.1(b)] $G \in SK^{m\times m}$ *("Gram matrix") satisfying* $f = v^T G v$.

(c) f *is a sum of* m *squares in* $K[\underline{X}]$.

Proof. (a) $\Longrightarrow$ (b) Let $\ell \in \mathbb{N}_0$ and $p_1, \dots, p_\ell \in K[\underline{X}]$ with $f = \sum_{i=1}^{\ell} p_i^2$. By 2.4.14, we have $\operatorname{supp}(p_i) \subseteq \frac{1}{2}N(f) \cap \mathbb{N}_0^n \subseteq \{\alpha_1, \dots, \alpha_m\}$. Hence there is an $A \in K^{\ell\times m}$ such that

$$Av = \begin{pmatrix} p_1 \\ \vdots \\ p_\ell \end{pmatrix}.$$

It follows that $f = \begin{pmatrix} p_1 & \dots & p_\ell \end{pmatrix} \begin{pmatrix} p_1 \\ \vdots \\ p_\ell \end{pmatrix} = (Av)^T Av = v^T A^T A v = v^T G v$ where $G := A^T A \in SK^{m\times m}$. By 2.3.3, G is psd.

(b) $\Longrightarrow$ (c) Let $G \in SK^{m\times m}$ be psd with $f = v^T G v$. Choose according to 2.3.3 an $A \in K^{m\times m}$ satisfying $G = A^T A$. Write

$$Av = \begin{pmatrix} p_1 \\ \vdots \\ p_m \end{pmatrix}.$$

Then $p_1, \dots, p_m \in K[\underline{X}]$ and

$$v^T G v = v^T A^T A v = (Av)^T Av = \begin{pmatrix} p_1 & \dots & p_\ell \end{pmatrix} \begin{pmatrix} p_1 \\ \vdots \\ p_\ell \end{pmatrix} = \sum_{i=1}^{m} p_i^2.$$

(c) $\Longrightarrow$ (a) is trivial. □

Example 2.6.2. Let K be an Euclidean field and $f := 2X_1^4 + 5X_2^4 - X_1^2X_2^2 + 2X_1^3X_2 \in K[X_1, X_2]$. Then $N(f) = \operatorname{conv}\{(4,0),(0,4)\}$ and therefore

$$\frac{1}{2}N(f) \cap \mathbb{N}_0^2 = \{(2,0),(1,1),(0,2)\}.$$

Set $v := \begin{pmatrix} X_1^2 \\ X_1X_2 \\ X_2^2 \end{pmatrix}$. From $\{G \in SK^{3\times 3} \mid f = v^TGv\} = \left\{ \begin{pmatrix} 2 & 1 & a \\ 1 & -2a-1 & 0 \\ a & 0 & 5 \end{pmatrix} \mid a \in K \right\}$, we obtain

$$f \in \sum K[X_1, X_2]^2 \iff \exists a \in K : \begin{pmatrix} 2 & 1 & a \\ 1 & -2a-1 & 0 \\ a & 0 & 5 \end{pmatrix} \text{ psd.}$$

For all $a \in K$, we have

$$\det \begin{pmatrix} 2+T & 1 & a \\ 1 & T-2a-1 & 0 \\ a & 0 & 5+T \end{pmatrix} = (2+T)(T-2a-1)(5+T) - a^2(T-2a-1) - 5 - T$$
$$= (T^2 - 2aT + T - 4a - 2)(5+T) - (1+a^2)T + 2a^3 + a^2 - 5$$
$$= T^3 - 2aT^2 + T^2 - 4aT - 2T + 5T^2 - 10aT + 5T - 20a - 10 - (1+a^2)T + 2a^3 + a^2 - 5$$
$$= T^3 + (6-2a)T^2 + (2 - 14a - a^2)T - 15 - 20a + a^2 + 2a^3$$

and by 2.3.3(e), we obtain

$$\begin{pmatrix} 2 & 1 & a \\ 1 & -2a-1 & 0 \\ a & 0 & 5 \end{pmatrix} \text{ psd} \iff \begin{array}{r} 2a^3 + a^2 - 20a - 15 \geq 0 \\ \& \; -a^2 - 14a + 2 \geq 0 \\ \& \; -2a + 6 \geq 0 \end{array}.$$

Set $a := -3$. Then $2a^3 + a^2 - 20a - 15 = -2 \cdot 27 + 9 + 60 - 15 = -54 + 9 + 60 - 15 = 0$, $-a^2 - 14a + 2 = -9 + 42 + 2 = 35 \geq 0$ and $-2a + 6 = 12 \geq 0$. For this reason $f \in \sum K[X_1, X_2]^2$. The quadratic form

$$q := \begin{pmatrix} T_1 & T_2 & T_3 \end{pmatrix} \begin{pmatrix} 2 & 1 & a \\ 1 & -2a-1 & 0 \\ a & 0 & 5 \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \\ T_3 \end{pmatrix} \in K[T_1, T_2, T_3]$$

obviously satisfies

$$q(X_1^2, X_1X_2, X_2^3) = v^T \begin{pmatrix} 2 & 1 & a \\ 1 & -2a-1 & 0 \\ a & 0 & 5 \end{pmatrix} v = f.$$

Because of

$$\operatorname{sg} q \overset{2.3.2(d)}{=} \operatorname{rk} q = \operatorname{rk} \begin{pmatrix} 2 & 1 & -3 \\ 1 & 5 & 0 \\ -3 & 0 & 5 \end{pmatrix} = 2,$$

q is a sum of 2 squares of linear forms in $K[T_1, T_2, T_3]$ and thus f a sum of 2 squares of polynomials. To compute this representation explicitely, we employ the procedure

from 1.6.1(f):

$$\begin{aligned} q &= 2T_1^2 + 2T_1T_2 - 6T_1T_3 + 5T_2^2 + 5T_3^2 \\ &= 2\Big(\underbrace{T_1 + \frac{1}{2}T_2 - \frac{3}{2}T_3}_{\ell_1}\Big)^2 - 2\Big(\frac{1}{2}T_2 - \frac{3}{2}T_3\Big)^2 + 5T_2^2 + 5T_3^2 \\ &= 2\ell_1^2 + \frac{9}{2}T_2^2 + 3T_2T_3 + \frac{1}{2}T_3^2 \\ &= 2\ell_1^2 + \frac{9}{2}\Big(\underbrace{T_2 + \frac{1}{3}T_3}_{\ell_2}\Big)^2 = 2\ell_1^2 + \frac{9}{2}\ell_2^2 \\ &= \frac{1}{2}(2T_1 + T_2 - 3T_3)^2 + \frac{1}{2}(3T_2 + T_3)^2. \end{aligned}$$

Hence $f = \frac{1}{2}(2X_1^2 + X_1X_2 - 3X_2^2)^2 + \frac{1}{2}(3X_1X_2 + X_2^2)^2$.

§3 Prime cones and real Stellensätze

3.1 The real spectrum of a commutative ring

In this section, we let A, B and C always be commutative rings.

Reminder 3.1.1. An ideal $\mathfrak{p}$ of A is called a prime ideal of A if

$$1 \notin \mathfrak{p} \quad \text{and} \quad \forall a, b \in A : (ab \in \mathfrak{p} \implies (a \in \mathfrak{p} \text{ or } b \in \mathfrak{p})).$$

We call $\operatorname{spec} A = \{\mathfrak{p} \mid \mathfrak{p} \text{ prime ideal of } A\}$ the *spectrum* of A. If I is an ideal of A, then

$$I \in \operatorname{spec} A \iff A/I \text{ is an integral domain.}$$

Because every integral domain extends to a field (e.g., to its quotient field) and every field to an algebraically closed field (e.g., to its algebraic closure), $\operatorname{spec} A$ consists exactly of the kernels of ring homomorphisms of A in $\left\{\begin{array}{c}\text{integral domains}\\ \text{fields}\\ \text{algebraically closed fields}\end{array}\right\}$. Every ring homomorphism $\varphi\colon A \to B$ induces a map

$$\operatorname{spec}\varphi\colon \operatorname{spec} B \to \operatorname{spec} A, \mathfrak{q} \mapsto \varphi^{-1}(\mathfrak{q}),$$

for if $\mathfrak{q} \in \operatorname{spec} B$, then $\mathfrak{p} := \varphi^{-1}(\mathfrak{q}) \in \operatorname{spec} A$ since φ induces an embedding $A/\mathfrak{p} \hookrightarrow B/\mathfrak{q}$ by the homomorphism theorem. If $\varphi\colon A \to B$ and $\psi\colon B \to C$ are ring homomorphisms, then

$$\operatorname{spec}(\psi \circ \varphi) = (\operatorname{spec}\varphi) \circ (\operatorname{spec}\psi).$$

Notation 3.1.2. If A is an integral domain, then

$$\operatorname{qf} A := (A \setminus \{0\})^{-1}A = \left\{\frac{a}{b} \mid a, b \in A, b \neq 0\right\}$$

denotes its quotient field.

Definition 3.1.3. We call $\operatorname{sper} A := \{(\mathfrak{p}, \leq) \mid \mathfrak{p} \in \operatorname{spec} A,\ \leq \text{ order of } \operatorname{qf}(A/\mathfrak{p})\}$ the *real spectrum* of A.

Remark 3.1.4. Every ring homomorphism $\varphi\colon A \to B$ induces a map

$$\operatorname{sper}\varphi\colon \operatorname{sper} B \to \operatorname{sper} A,\ (\mathfrak{q}, \leq) \mapsto (\varphi^{-1}(\mathfrak{q}), \leq'),$$

where $\leq'$ denotes the order of $\operatorname{qf}(A/\mathfrak{p})$ with $\mathfrak{p} := \varphi^{-1}(\mathfrak{q})$ which makes the canonical embedding $\operatorname{qf}(A/\mathfrak{p}) \hookrightarrow \operatorname{qf}(B/\mathfrak{q})$ into an embedding $(\operatorname{qf}(A/\mathfrak{p}), \leq') \hookrightarrow (\operatorname{qf}(B/\mathfrak{q}), \leq)$ of ordered fields. If $\varphi\colon A \to B$ and $\psi\colon B \to C$ are ring homomorphisms, then we have again

$$\operatorname{sper}(\psi \circ \varphi) = (\operatorname{sper}\varphi) \circ (\operatorname{sper}\psi).$$

Example 3.1.5. Since $\mathbb{R}[X]$ is a principal ideal domain, the fundamental theorem 1.4.14 implies

$$\operatorname{spec}\mathbb{R}[X] = \{(0)\} \cup \{(X-a) \mid a \in \mathbb{R}\} \cup \{(\underbrace{(X-a)^2+b^2}_{=(X-(a+b\mathrm{i}))(X-(a-b\mathrm{i}))}) \mid a,b \in \mathbb{R}, b \neq 0\}$$

where $((X-a)^2+b^2) = ((X-a')^2+b'^2) \iff (a=a' \,\&\, |b| = |b'|)$ for all $a,a',b,b' \in \mathbb{R}$. The spectrum of $\mathbb{R}[X]$ therefore can be seen as consisting of

- one "generic point",
- the real numbers, and
- the unordered pairs of two distinct conjugated complex numbers.

Because of $\operatorname{qf}(\mathbb{R}[X]/(0)) \cong \operatorname{qf}(\mathbb{R}[X]) = \mathbb{R}(X)$, $\operatorname{qf}(\mathbb{R}[X]/(X-a)) = \mathbb{R}[X]/(X-a) \cong \mathbb{R}$ for all $a \in \mathbb{R}$ and $\operatorname{qf}(\mathbb{R}[X]/((X-a)^2+b^2) \cong \mathbb{R}[X]/((X-a)^2+b^2) \cong \mathbb{C}$ for $a,b \in \mathbb{R}$ with $b \neq 0$, we obtain in the notation of 1.3.8 (and with the identification $\mathbb{R}[X]/(0) = \mathbb{R}[X]$)

$$\operatorname{sper}\mathbb{R}[X] = \{((0),P_{-\infty}),((0),P_{\infty})\} \cup \{((0),P_{a-}) \mid a \in \mathbb{R}\} \cup \{((0),P_{a+}) \mid a \in \mathbb{R}\} \\ \cup \{((X-a),(\mathbb{R}[X]/(X-a))^2) \mid a \in \mathbb{R}\}.$$

The *real* spectrum of $\mathbb{R}[X]$ thus corresponds to an accumulation consisting of

- the two points at infinity,
- for each real number two points infinitely close, and
- the real numbers.

Definition 3.1.6. We call $\operatorname{supp}\colon \operatorname{sper} A \to \operatorname{spec} A$, $(\mathfrak{p},\leq) \mapsto \mathfrak{p}$ the *support map*.

Definition 3.1.7. [→ 1.1.19(a), 3.1.1] A subset P of A is called a *prime cone* of A if $P+P \subseteq P$, $PP \subseteq P$, $P \cup -P = A$, $-1 \notin P$ and $\forall a,b \in A : (ab \in P \implies (a \in P \text{ or } -b \in P))$.

Proposition 3.1.8. *Every prime cone of A is a proper preorder of A* [→ 1.2.1].

Proof. Suppose P is a prime cone of A and $a \in A$. To show: $a^2 \in P$. Due to $a \in A = P \cup -P$, we have $a \in P$ or $-a \in P$. In the first case we get $a^2 = aa \in PP \subseteq P$ and in the second $a^2 = (-a)^2 = (-a)(-a) \in PP \subseteq P$. □

Proposition 3.1.9. *Suppose $P \subseteq A$ satisfies $P+P \subseteq P$, $PP \subseteq P$ and $P \cup -P = A$. Then the following are equivalent:*

(a) *P is a prime cone of A.*

(b) *$-1 \notin P$ and $\forall a,b \in A : (ab \in P \implies (a \in P \text{ or } -b \in P))$*

(c) *$P \cap -P$ is a prime ideal of A*

Proof. (a) $\iff$ (b) is Definition 3.1.7.

(b) $\implies$ (c) Suppose (b) holds and set $\mathfrak{p} := P \cap -P$. Then $\mathfrak{p}$ is obviously a subgroup of A and we have $A\mathfrak{p} = (P \cup -P)\mathfrak{p} = P\mathfrak{p} \cup -P\mathfrak{p} = P(P \cap -P) \cup -P(P \cap -P) \subseteq (PP \cap -PP) \cup (-PP \cap PP) \subseteq (P \cap -P) \cup (-P \cap P) = P \cap -P = \mathfrak{p}$, i.e., $\mathfrak{p}$ is an ideal of A (if $\frac{1}{2} \in A$ this follows alternatively from 3.1.8 and 1.2.4). From $-1 \notin P$ we get $1 \notin \mathfrak{p}$. It remains to show $\forall a, b \in A\colon (ab \in \mathfrak{p} \implies (a \in \mathfrak{p} \text{ or } b \in \mathfrak{p}))$. To this end, let $a, b \in A$ with $a \notin \mathfrak{p}$ and $b \notin \mathfrak{p}$. To show: $ab \notin \mathfrak{p}$. WLOG $a \notin P$ and $-b \notin P$ (otherwise replace a by $-a$ and/or $-b$ by b, taking into account $-\mathfrak{p} = \mathfrak{p}$). By hypothesis, we obtain then $ab \notin P$ and thus $ab \notin \mathfrak{p}$.

(c) $\implies$ (b) Suppose (c) holds. Due to $P \cup -P = A$, we have $1 \in P$ or $-1 \in P$. If $-1 \in P$, then again $1 = (-1)(-1) \in PP \subseteq P$. Hence $1 \in P$. If we had $-1 \in P$, then $1 \in \mathfrak{p} := P \cap -P \in \operatorname{spec} A$ $\lightning$. Thus $-1 \notin P$. Let now $a, b \in A$ such that $a \notin P$ and $-b \notin P$. To show: $ab \notin P$. Because of $P \cup -P = A$, we have $a \in -P$ and $b \in P$ from which $-ab = (-a)b \in PP \subseteq P$. If we had in addition $ab \in P$, then $ab \in \mathfrak{p}$ and thus $a \in \mathfrak{p} \subseteq P$ or $b \in \mathfrak{p} \subseteq -P$ $\lightning$. Hence $ab \notin P$. □

Remark 3.1.10. If K is a field, then 3.1.9 signifies because of $\operatorname{spec} K = \{(0)\}$ just that the prime cones of K are exactly the orders of K [→ 1.1.20].

Lemma 3.1.11. Let P be a prime cone of A and $\mathfrak{p} := P \cap -P$ [→ 3.1.9(c)]. Then

$$P_{\mathfrak{p}} := \left\{ \frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} \mid a \in A, s \in A \setminus \mathfrak{p}, as \in P \right\}$$

is an order (i.e., a prime cone [→ 3.1.10]) of $\operatorname{qf}(A/\mathfrak{p})$.

Proof. To show [→ 1.1.20(a)]:

(a) $P_{\mathfrak{p}} + P_{\mathfrak{p}} \subseteq P_{\mathfrak{p}}$,

(b) $P_{\mathfrak{p}} P_{\mathfrak{p}} \subseteq P_{\mathfrak{p}}$,

(c) $P_{\mathfrak{p}} \cup -P_{\mathfrak{p}} = \operatorname{qf}(A/\mathfrak{p})$, and

(d) $P_{\mathfrak{p}} \cap -P_{\mathfrak{p}} = (0)$.

(a) Suppose that $a, b \in A$ and $s, t \in A \setminus \mathfrak{p}$ with $as, bt \in P$ define arbitrary elements $\frac{\overline{a}}{\overline{s}}, \frac{\overline{b}}{\overline{t}} \in P_{\mathfrak{p}}$. Then

$$\frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} + \frac{\overline{b}^{\mathfrak{p}}}{\overline{t}^{\mathfrak{p}}} = \frac{\overline{at}^{\mathfrak{p}}}{\overline{st}^{\mathfrak{p}}} + \frac{\overline{bs}^{\mathfrak{p}}}{\overline{st}^{\mathfrak{p}}} = \frac{\overline{at + bs}^{\mathfrak{p}}}{\overline{st}^{\mathfrak{p}}} \in P_{\mathfrak{p}},$$

since $at + bs \in A$, $st \in A \setminus \mathfrak{p}$ and $(at + bs)st = ast^2 + bts^2 \in PA^2 + PA^2 \subseteq PP + PP \subseteq P + P \subseteq P$.

(b) Let again $a, b \in A$ and $s, t \in A \setminus \mathfrak{p}$ satisfy $as, bt \in P$. Then

$$\frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} \frac{\overline{b}^{\mathfrak{p}}}{\overline{t}^{\mathfrak{p}}} = \frac{\overline{ab}^{\mathfrak{p}}}{\overline{st}^{\mathfrak{p}}} \in P_{\mathfrak{p}}$$

since $ab \in A$, $st \in A \setminus \mathfrak{p}$ and $abst = (as)(bt) \in PP \subseteq P$.

(c) Let $a \in A$ and $s \in A \setminus \mathfrak{p}$ define an arbitrary element $\frac{\overline{a}}{\overline{s}} \in \mathrm{qf}(A/\mathfrak{p})$. Because of $P \cup -P = A$, we have $as \in P$ or $-as \in P$, i.e., $-\frac{\overline{a}}{\overline{s}} = \frac{\overline{-a}}{\overline{s}} \in P_{\mathfrak{p}}$ or $\frac{\overline{a}}{\overline{s}} \in P_{\mathfrak{p}}$.

(d) Suppose $a, b \in A$ and $s, t \in A \setminus \mathfrak{p}$ with $as, bt \in P$ satisfy

$$\frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} = -\frac{\overline{b}^{\mathfrak{p}}}{\overline{t}^{\mathfrak{p}}}.$$

Then $at + bs \in \mathfrak{p}$ and therefore $ast^2 + bts^2 = st(at + bs) \in \mathfrak{p} \subseteq -P$, i.e., $-ast^2 - bts^2 \in P$. From $ast^2 = (as)t^2 \in PA^2 \subseteq P$ and $bts^2 = (bt)s^2 \in PA^2 \subseteq P$ we deduce $-ast^2, -bts^2 \in P$. Consequently, $ast^2, bts^2 \in \mathfrak{p}$ and thus $a, b \in \mathfrak{p}$. We obtain

$$\frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} = 0 = \frac{\overline{b}^{\mathfrak{p}}}{\overline{t}^{\mathfrak{p}}}$$

as desired. □

Lemma 3.1.12. [→ 1.1.19] Let $(\mathfrak{p}, \leq) \in \operatorname{sper} A$. Then $\{a \in A \mid \overline{a}^{\mathfrak{p}} \geq 0\}$ is a prime cone of A.

Proof. Set $P := \{a \in A \mid \overline{a}^{\mathfrak{p}} \geq 0\}$. Then $P + P \subseteq P$, $PP \subseteq P$, $P \cup -P = A$ and $P \cap -P = \mathfrak{p} \in \operatorname{spec} A$. Now P is a prime cone of A by 3.1.9(c). □

Proposition 3.1.13. [→ 1.1.19(c)] *The correspondence*

$$\begin{aligned} (\mathfrak{p}, \leq) &\mapsto \{a \in A \mid \overline{a}^{\mathfrak{p}} \geq 0\} \\ (P \cap -P, P_{P \cap -P}) &\leftarrow\!\shortmid P \end{aligned}$$

defines a bijection between sper A *and the set of all prime cones of* A.

Proof. The well-definedness of both maps follows from Lemmata 3.1.11 and 3.1.12. Now first let $(\mathfrak{p}, \leq) \in \operatorname{sper} A$ and $P := \{a \in A \mid \overline{a}^{\mathfrak{p}} \geq 0\}$. We show $(\mathfrak{p}, \leq) = (P \cap -P, P_{P \cap -P})$. It is clear that $\mathfrak{p} = P \cap -P$. Finally,

$$\begin{aligned} P_{P \cap -P} = P_{\mathfrak{p}} &= \left\{ \frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} \mid a \in A, s \in A \setminus \mathfrak{p}, as \in P \right\} \\ &= \left\{ \frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} \mid a \in A, s \in A \setminus \mathfrak{p}, \overline{as}^{\mathfrak{p}} \geq 0 \right\} \\ &= \left\{ \frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} \mid a \in A, s \in A \setminus \mathfrak{p}, \frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} \geq 0 \right\} = \{x \in \mathrm{qf}(A/\mathfrak{p}) \mid x \geq 0\}. \end{aligned}$$

Conversely, suppose that P is a prime cone of A and $\mathfrak{p} := P \cap -P$. We show

$$P = \{a \in A \mid \overline{a}^{\mathfrak{p}} \geq 0\}.$$

Here "$\subseteq$" is trivial. To show "$\supseteq$", let $a \in A$ such that $\overline{a}^{\mathfrak{p}} \geq 0$. Then there are $b \in A$ and $s \in A \setminus \mathfrak{p}$ such that $bs \in P$ and $\overline{a} = \frac{\overline{b}}{\overline{s}}$. It follows that $\overline{as^2}^{\mathfrak{p}} = \overline{bs}^{\mathfrak{p}}$ and thus $as^2 \in bs + \mathfrak{p} \subseteq P + \mathfrak{p} \subseteq P + P \subseteq P$. Since P is a prime cone, we deduce $a \in P$ or $-s^2 \in P$. If we had $-s^2 \in P$, then $s^2 \in P \cap -P = \mathfrak{p}$ (since $s^2 \in A^2 \subseteq P$) and therefore $s \in \mathfrak{p}$ ↯. □

Remark 3.1.14. [→ 1.1.20] As a result of 3.1.13, we can see elements of the real spectrum as prime cones. We reformulate some of the above in this new language:

(a) Remark 3.1.4: Let $\varphi\colon A \to B$ be a ring homomorphism. Then φ induces the map $\operatorname{sper}\varphi\colon \operatorname{sper} B \to \operatorname{sper} A,\ Q \mapsto \varphi^{-1}(Q)$. Suppose namely that $Q \in \operatorname{sper} B$, $\mathfrak{q} := Q \cap -Q$, $P := \varphi^{-1}(Q)$ and $\mathfrak{p} := P \cap -P$. Then $\varphi^{-1}(\mathfrak{q}) = \varphi^{-1}(Q) \cap -\varphi^{-1}(Q) = P \cap -P = \mathfrak{p}$ and the embedding $\operatorname{qf}(A/\mathfrak{p}) \hookrightarrow \operatorname{qf}(B/\mathfrak{q})$ induced by φ is an embedding of ordered fields $(\operatorname{qf}(A/\mathfrak{p}), P_{\mathfrak{p}}) \hookrightarrow (\operatorname{qf}(B/\mathfrak{q}), Q_{\mathfrak{q}})$ because for $a \in A$ and $s \in A \setminus \mathfrak{p}$ with $as \in P$ we have $\varphi(a) \in B$, $\varphi(s) \in B \setminus \mathfrak{q}$, $\varphi(a)\varphi(s) = \varphi(as) \in \varphi(P) \subseteq Q$.

(b) Definition 3.1.6: The support map is $\operatorname{supp}\colon \operatorname{sper} A \to \operatorname{spec} A,\ P \mapsto P \cap -P$ [→ 3.1.13]. In particular, the Definitions 3.1.6 and 1.2.4 are compatible.

Definition 3.1.15. For every $(\mathfrak{p}, \leq) \in \operatorname{sper} A$, we call the real closed field

$$R_{(\mathfrak{p},\leq)} := \overline{(\operatorname{qf}(A/\mathfrak{p}), \leq)}$$

the *representation field* of $(\mathfrak{p}, \leq)$ and the ring homomorphism

$$\varrho_{(\mathfrak{p},\leq)}\colon A \to R_{(\mathfrak{p},\leq)},\ a \mapsto \overline{a}^{\mathfrak{p}}$$

the *representation* of $(\mathfrak{p}, \leq)$.

Proposition 3.1.16. *Let $P \in \operatorname{sper} A$. Then $P = \varrho_P^{-1}(R_P^2)$ and $\operatorname{supp} P = \ker \varrho_P$.*

Proof. $\varrho_P^{-1}(R_P^2) = \{a \in A \mid \varrho_P(a) \geq 0 \text{ in } R_P\} = \{a \in A \mid \overline{a}^{\operatorname{supp} P} \in P_{\operatorname{supp} P}\} \overset{3.1.13}{=} P$ and therefore

$$\operatorname{supp} P = P \cap -P = \varrho_P^{-1}(R_P^2) \cap -\varrho_P^{-1}(R_P^2) = \varrho_P^{-1}(R_P^2 \cap -R_P^2) = \varrho_P^{-1}(\{0\}) = \ker \varrho_P.$$

□

Proposition 3.1.17. [→ 3.1.1] *Let P be a set. Then the following are equivalent:*

(a) $P \in \operatorname{sper} A$

(b) *There is an ordered field $(K, \leq)$ and a ring homomorphism $\varphi\colon A \to K$ such that $P = \varphi^{-1}(K_{\geq 0})$.*

(c) *There exists a real closed field R and a ring homomorphism $\varphi\colon A \to R$ such that $P = \varphi^{-1}(R^2)$.*

Proof. (a) $\overset{3.1.16}{\Longrightarrow}$ (c) $\overset{\text{trivial}}{\Longrightarrow}$ (b) $\overset{3.1.14(a)}{\Longrightarrow}$ (a) □

3.2 Preorders and maximal prime cones

Throughout this section, let A be a commutative ring.

Proposition 3.2.1. *Let T be a proper preorder of A [→ 1.2.1]. Then the following are equivalent:*

(a) *T is a prime cone of A.*

(b) $\forall a, b \in A : (ab \in T \implies (a \in T \text{ or } -b \in T))$

Proof. (a) $\implies$ (b) is trivial by Definition 3.1.7.

(b) $\implies$ (a) Suppose (b) holds. By Definition 3.1.7, it suffices to show $T \cup -T = A$. But for all $a \in A$ it follows from (b) that $a \in T$ or $-a \in T$ because of $aa = a^2 \in T$. □

Theorem 3.2.2. [→ 1.2.9] *Suppose T is a maximal proper preorder* [→ 1.2.1] *of A. Then T is a prime cone of A.*

Proof. We show 3.2.1(b). For this purpose let $a, b \in A$ satisfy $a \notin T$ and $-b \notin T$. Then $T + aT$ and $T - bT$ are preorders of A [→ 1.2.8] that properly contain T. Due to the maximality of T, therefore neither $T + aT$ nor $T - bT$ is proper as a preorder, i.e., $-1 \in T + aT$ and $-1 \in T - bT$. Choose $s, t \in T$ such that $-as \in 1 + T$ and $bt \in 1 + T$. Then $-abst \in (1 + T)(1 + T) \subseteq 1 + T$ and thus $-1 \in T + abst \subseteq T + abT$. Since T is proper, we conclude that $ab \notin T$ as desired. □

Corollary 3.2.3. *Every proper preorder of A is contained in a maximal prime cone of A.*

Proof. Use 3.2.2 and Zorn's lemma. □

Proposition 3.2.4. *Let $P, Q \in \operatorname{sper} A$ such that $P \subseteq Q$ and set $\mathfrak{q} := \operatorname{supp} Q$. Then $Q = P \cup \mathfrak{q}$.*

Proof. "$\supseteq$" is trivial.
"$\subseteq$" Let $a \in Q \setminus P$. To show: $a \in \mathfrak{q}$. From $-a \in P \subseteq Q$ we get $a \in Q \cap -Q = \mathfrak{q}$. □

Proposition and Terminology 3.2.5. *Let $P \in \operatorname{sper} A$. Then "the spear"*

$$\{Q \in \operatorname{sper} A \mid P \subseteq Q\}$$

is a chain in the partially ordered set sper A *that possesses a largest element ("a spearhead").*

Proof. Let $Q_1, Q_2 \in \operatorname{sper} A$ with $P \subseteq Q_1$ and $P \subseteq Q_2$. Suppose $Q_1 \not\subseteq Q_2$. To show: $Q_2 \subseteq Q_1$. Choose $a \in Q_1 \setminus Q_2$. Let $b \in Q_2$. To show $b \in Q_1$. We have $a - b \notin Q_2$ (or else $a \in Q_2$ ↯) and thus $a - b \notin P$ because of $P \subseteq Q_2$. Then $b - a \in P \subseteq Q_1$ and thus $b \in Q_1$. The existence of the "spearhead" follows now from 3.2.3. □

3.3 Quotients and localization

Throughout this section, we let A be a commutative ring.

Proposition 3.3.1. $\left\{\begin{matrix}\text{Preimages}\\ \text{Images}\end{matrix}\right\}$ *of preorders* [→ 1.2.1] *under* $\left\{\begin{matrix}\text{homomorphisms}\\ \text{epimorphisms}\end{matrix}\right\}$ *of commutative rings are again preorders.*

Proof. Exercise. □

Proposition 3.3.2. *Let I be an ideal of A. The correspondence*

$$\begin{aligned} T &\mapsto \overline{T}^I := \{\overline{a}^I \mid a \in T\} \\ \{a \in A \mid \overline{a}^I \in P\} &\hookleftarrow P \end{aligned}$$

defines a bijection between the set of $\left\{\begin{matrix}\text{preorders [→ 1.2.1]}\\ \text{prime cones [→ 3.1.7]}\end{matrix}\right\}$ *T of A with $I \subseteq T$ and the set of* $\left\{\begin{matrix}\text{preorders}\\ \text{prime cones}\end{matrix}\right\}$ *of A/I.*

Proof. Exercise. □

Lemma 3.3.3. Let $S \subseteq A$ be multiplicative and $T \subseteq A$ a preorder. Let

$$\iota\colon A \to S^{-1}A,\ a \mapsto \frac{a}{1}$$

denote the canonical homomorphism. Then the preorder generated by $\iota(T)$ in $S^{-1}A$ equals $S^{-2}T = \left\{\frac{a}{s^2} \mid a \in T, s \in S\right\}$. This preorder is proper if and only if $T \cap -S^2 = \emptyset$.

Proof. Exercise. □

Proposition 3.3.4. *Let $S \subseteq A$ be multiplicative. The correspondence*

$$\begin{aligned} P &\mapsto S^{-2}P \\ \left\{a \in A \mid \frac{a}{1} \in Q\right\} &\hookleftarrow Q \end{aligned}$$

gives rise to a bijection between $\{P \in \operatorname{sper} A \mid (\operatorname{supp} P) \cap S = \emptyset\}$ and $\operatorname{sper}(S^{-1}A)$.

Proof. Let $P \in \operatorname{sper} A$ with $(\operatorname{supp} P) \cap S = \emptyset$. By 3.3.3, $S^{-2}P$ is a proper preorder of $S^{-1}A$ since $P \cap -S^2 \subseteq (P \cap -A^2) \cap (-S) \subseteq (P \cap -P) \cap (-S) = (\operatorname{supp} P) \cap -S = -((\operatorname{supp} P) \cap S) = -\emptyset = \emptyset$. To show that $S^{-2}P$ is a prime cone of $S^{-1}A$, we verify the condition from 3.2.1(b) where we use that for any two fractions in $S^{-1}A$, one can find a common denominator from S^2. Let $a, b \in A$ and $s \in S$ with $\frac{a}{s^2} \cdot \frac{b}{s^2} \in S^{-2}P$. To show: $\frac{a}{s^2} \in S^{-2}P$ or $-\frac{b}{s^2} \in S^{-2}P$. Choose $c \in P$ and $u \in S$ with $\frac{ab}{s^4} = \frac{c}{u^2}$. Then there is $v \in S$ such that $abu^2v = cs^4v$ and therefore $(au^2)(bv^2) = abu^2v^2 = cs^4v^2 \in P$. Since

P is a prime cone, it follows that $au^2 \in P$ or $-bv^2 \in P$. Hence $\frac{a}{s^2} = \frac{au^2}{(su)^2} \in S^{-2}P$ or $-\frac{b}{s^2} = -\frac{bv^2}{(sv)^2} \in S^{-2}P$.

Conversely, let $Q \in \operatorname{sper}(S^{-1}A)$. For $\iota\colon A \to S^{-1}A,\ a \mapsto \frac{a}{1}$, we have [→ 3.1.14(a)]

$$\left\{a \in A \mid \frac{a}{1} \in Q\right\} = (\operatorname{sper}\iota)(Q) \in \operatorname{sper} A.$$

If we had $s \in S$ with $\frac{s}{1} \in Q \cap -Q$, then $1 = \frac{s}{s} = \frac{1}{s} \cdot \frac{s}{1} \in S^{-1}A(\operatorname{supp} Q) \subseteq \operatorname{supp} Q$ ↯.

It remains to show that the maps are inverse to each other:

(a) If $P \in \operatorname{sper} A$ with $(\operatorname{supp} P) \cap S = \emptyset$, then $P = \left\{a \in A \mid \frac{a}{1} \in S^{-2}P\right\}$.

(b) If $Q \in \operatorname{sper}(S^{-1}A)$, then $Q = \left\{\frac{a}{s^2} \mid a \in A, \frac{a}{1} \in Q, s \in S\right\}$.

To show (a), let $P \in \operatorname{sper} A$ with $(\operatorname{supp} P) \cap S = \emptyset$.
"$\subseteq$" is trivial.
"$\supseteq$" Let $a \in A$ with $\frac{a}{1} \in S^{-2}P$. Choose $b \in P$ and $s \in S$ with $\frac{a}{1} = \frac{b}{s^2}$. Then there is $t \in S$ such that $as^2t = bt$ and thus $as^2t^2 = bt^2 \in P$. It follows that $a \in P$ or $-s^2t^2 \in P$. The latter would lead to $s^2t^2 \in (\operatorname{supp} P) \cap S$ ↯. Hence $a \in P$.

To show (b), consider an arbitrary $Q \in \operatorname{sper}(S^{-1}A)$.
"$\supseteq$" is trivial.
"$\subseteq$" Let $b \in A$ and $s \in S$ with $\frac{b}{s} \in Q$. Then for $a := sb \in A$, we have $\frac{b}{s} = \frac{sb}{s^2} = \frac{a}{s^2}$ and $\frac{a}{1} = \frac{sb}{1} = \left(\frac{s}{1}\right)^2 \frac{b}{s} \in Q$. □

3.4 Abstract real Stellensätze

Definition 3.4.1. Let A be a commutative ring. We call the ring homomorphism

$$A \to \prod_{(\mathfrak{p},\leq) \in \operatorname{sper} A} R_{(\mathfrak{p},\leq)},\ a \mapsto (\widehat{a}\colon (\mathfrak{p},\leq) \mapsto \overline{a}^{\mathfrak{p}})$$

the *real representation* of A. For $a \in A$, we say that $\widehat{a}$ is the *function represented* by a on the real spectrum of A.

Theorem 3.4.2 (abstract real Stellensatz [Kri, Ste, Pre])**.** *Suppose A is a commutative ring, $I \subseteq A$ an ideal, $S \subseteq A$ a multiplicative set and $T \subseteq A$ a preorder. Then the following conditions are equivalent:*

(a) *There does* not *exist any $P \in \operatorname{sper} A$ satisfying*

$$\begin{aligned} &\forall a \in I : \widehat{a}(P) = 0, \\ &\forall s \in S : \widehat{s}(P) \neq 0 \quad \textit{and} \\ &\forall t \in T : \widehat{t}(P) \geq 0. \end{aligned}$$

(b) *There are $a \in I$, $s \in S$ and $t \in T$ such that $a + s^2 + t = 0$.*

Proof. (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (b) Replacing T by the preorder $T + I$, we can suppose WLOG $I = (0)$. Suppose (b) does not hold. By 3.3.3, $S^{-2}T$ is then a proper preorder of $S^{-1}A$. Consequently, $S^{-2}T$ is contained in a prime cone Q of $S^{-1}A$ by 3.2.3. Now 3.3.4 yields $P := \{a \in A \mid \frac{a}{1} \in Q\} \in \operatorname{sper} A$ and $(\operatorname{supp} P) \cap S = \emptyset$. For all $s \in S$, we have $\widehat{s}(P) = \overline{s}^{\operatorname{supp} P} \neq 0$ in R_P [$\to$ 3.1.15, 3.1.13] since $s \notin \operatorname{supp} P$. For all $t \in T$, we have $\widehat{t}(P) \geq 0$ because $t \in P$. □

Terminology and Notation 3.4.3. (a) We call a pair (A, T) consisting of a commutative ring A and a preorder T of A a *preordered ring*.

(b) If (A, T) is a preordered ring, then we define its *real spectrum*

$$\operatorname{sper}(A, T) := \{P \in \operatorname{sper} A \mid T \subseteq P\}.$$

(c) [$\to$ 1.4.15(c)] If A is a commutative ring, $a \in A$ and $S \subseteq \operatorname{sper} A$, then we write

$$\widehat{a} \geq 0 \text{ on } S :\Longleftrightarrow \forall P \in S : \widehat{a}(P) \geq 0,$$
$$\widehat{a} > 0 \text{ on } S :\Longleftrightarrow \forall P \in S : \widehat{a}(P) > 0,$$

and so forth.

Corollary 3.4.4 (abstract Positivstellensatz)**.** *Let (A, T) be a preordered ring and $a \in A$. Then the following are equivalent:*

(a) $\widehat{a} > 0$ *on* $\operatorname{sper}(A, T)$

(b) $\exists t \in T : ta \in 1 + T$

(c) $\exists t \in T : (1 + t)a \in 1 + T$.

Proof. (b) $\Longrightarrow$ (c) If $t, t' \in T$ satisfy $ta = 1 + t'$, then

$$(1 + t + t')a = ta + (1 + t')a = 1 + t' + ta^2 \in 1 + T.$$

(c) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (b) follows by applying 3.4.2 on the ideal (0), the multiplicative set $\{1\}$ and the preorder $T - aT$. □

Corollary 3.4.5 (abstract Nichtnegativstellensatz)**.** *Let (A, T) be a preordered ring and $a \in A$. Then the following are equivalent:*

(a) $\widehat{a} \geq 0$ *on* $\operatorname{sper}(A, T)$

(b) $\exists t \in T : \exists k \in \mathbb{N}_0 : ta \in a^{2k} + T$

Proof. (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (b) follows by applying 3.4.2 on the ideal (0), the multiplicative set $\{1, a, a^2, \dots\}$ and the preorder $T - aT$. □

Corollary 3.4.6 (abstract real Nullstellensatz [Kri, Du2, Ris, Efr]). *Let A be a commutative ring, $I \subseteq A$ an ideal and $a \in A$. Then the following are equivalent:*

(a) $\widehat{a} = 0$ *on* $\{P \in \operatorname{sper} A \mid I \subseteq \operatorname{supp} P\}$

(b) $\exists k \in \mathbb{N}_0 : \exists s \in \sum A^2 : a^{2k} + s \in I$

Proof. (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (b) follows by applying 3.4.2 on the ideal I, the multiplicative set $\{1, a, a^2, \dots\}$ and the preorder $\sum A^2$. □

3.5 The real radical ideal

Throughout this section, we let A be a commutative ring.

Definition 3.5.1. [→ 1.2.12(c)] A is called *real* (or *real reduced*) if

$$\forall n \in \mathbb{N} : \forall a_1, \dots, a_n \in A : (a_1^2 + \dots + a_n^2 = 0 \implies a_1 = 0).$$

Remark 3.5.2. We have

$$A \neq \{0\} \text{ real} \implies -1 \notin \sum A^2 \underset{1.2.2(a)}{\overset{3.2.3}{\iff}} \operatorname{sper} A \neq \emptyset.$$

Here "$\Longrightarrow$" cannot be replaced by "$\iff$" (in contrast to the case where A is a field [→ 1.2.12]) as the example of $A = \mathbb{R}[X]/(X^2)$ shows.

Definition 3.5.3. An ideal $I \subseteq A$ is called *real* (or *real radical ideal*) if A/I is real, i.e., $\forall n \in \mathbb{N} : \forall a_1, \dots, a_n \in A : (a_1^2 + \dots + a_n^2 \in I \implies a_1 \in I)$.

Proposition 3.5.4. *Let $\mathfrak{p} \in \operatorname{spec} A$. Then the following are equivalent:*

(a) $\mathfrak{p}$ *is real* [→ 3.5.3]

(b) $\operatorname{qf}(A/\mathfrak{p})$ *is real* [→ 1.2.11]

(c) $\exists P \in \operatorname{sper} A : \mathfrak{p} = \operatorname{supp} P$ [→ 3.1.14(b)]

Proof. (a) $\Longrightarrow$ (b) Suppose (a) holds and let $n \in \mathbb{N}$, $a_1, \dots, a_n, s \in A/\mathfrak{p}$ with $s \neq 0$ such that $\left(\frac{a_1}{s}\right)^2 + \dots + \left(\frac{a_n}{s}\right)^2 = 0$. Then $a_1^2 + \dots + a_n^2 = 0$. Since $A/\mathfrak{p}$ is real, it follows that $a_1 = 0$ and therefore $\frac{a_1}{s} = 0$.

(b) $\Longrightarrow$ (c) Suppose (b) holds. Then $\operatorname{qf}(A/\mathfrak{p})$ possesses an order $\leq$. According to Definition 3.1.3, we have $(\mathfrak{p}, \leq) \in \operatorname{sper} A$ and of course $\mathfrak{p} = \operatorname{supp}(\mathfrak{p}, \leq)$ by Definition 3.1.6.

(c) $\Longrightarrow$ (a) Suppose $\mathfrak{p} = \operatorname{supp} P$ for some $P \in \operatorname{sper} A$. Let $n \in \mathbb{N}$ and $a_1, \dots, a_n \in A$ satisfy $a_1^2 + \dots + a_n^2 \in \mathfrak{p}$. Then $\widehat{a}_1(P)^2 + \dots + \widehat{a}_n(P)^2 = 0$ and thus $\widehat{a}_1(P) = 0$, i.e., $a_1 \in \mathfrak{p}$. □

Definition 3.5.5. The *real radical* $\operatorname{rrad} I$ of an ideal $I \subseteq A$ is defined by

$$\operatorname{rrad} I := \bigcap \{\mathfrak{p} \in \operatorname{rspec} A \mid I \subseteq \mathfrak{p}\}$$

where $\operatorname{rspec} A := \{\mathfrak{p} \in \operatorname{spec} A \mid \mathfrak{p} \text{ is real}\}$ and $\bigcap \emptyset := A$.

Remark 3.5.6. Since every intersection of real ideals of A is obviously again a real ideal of A, for every ideal $I \subseteq A$, the set $\operatorname{rrad} I$ is a real ideal of I.

Theorem 3.5.7. [→ 3.4.6] *For every ideal I of A,*

$$\operatorname{rrad} I = \left\{a \in A \mid \exists k \in \mathbb{N}_0 : \exists s \in \sum A^2 : a^{2k} + s \in I\right\}.$$

Proof. 3.5.4 shows that this is just a reformulation of 3.4.6. □

Remark 3.5.8. Let $I \subseteq A$ be an ideal. Then by 3.5.6, $\operatorname{rrad} I$ is the smallest real ideal of A containing I.

Definition 3.5.9. We call $\operatorname{rnil} A := \bigcap \operatorname{rspec} A = \operatorname{rrad}(0)$ the *real nilradical* of A.

Corollary 3.5.10. *We have*

$$\{a \in A \mid \widehat{a} = 0\} = \operatorname{rnil} A = \{a \in A \mid \exists k \in \mathbb{N} : \exists s \in \sum A^2 : a^{2k} + s = 0\}.$$

3.6 Constructible sets

In this section, we let (A, T) always be a preordered ring [→ 3.4.3(a)]. At the moment it is a general one but after Proposition 3.6.2, we will further specialize (A, T).

Definition 3.6.1. [→ 1.8.3] A Boolean combination [→ 1.8.2(b)] of sets of the form

$$\{P \in \operatorname{sper}(A, T) \mid a \in P\} \qquad (a \in A)$$

is called a *constructible subset* of the real spectrum of (A, T). We denote the Boolean algebra of all constructible sets of $\operatorname{sper}(A, T)$ by $\mathscr{C}_{(A,T)}$. The analogous definition remains in force for a commutative ring instead of a preordered ring (A, T).

Proposition 3.6.2. [→ 1.8.6] *Every constructible subset of* $\operatorname{sper}(A, T)$ *is of the form*

$$\bigcup_{i=1}^{k} \left\{P \in \operatorname{sper}(A, T) \mid \widehat{a}_i(P) = 0, \widehat{b}_{i1}(P) > 0, \ldots, \widehat{b}_{im}(P) > 0\right\}$$

for some $k, m \in \mathbb{N}_0$, $a_i, b_{ij} \in A$.

Proof. Completely analogous to 1.8.6 using that $a \in P \overset{3.1.13}{\iff} \widehat{a}(P) \geq 0$ for all $a \in A$ and $P \in \operatorname{sper} A$. □

For the rest of this section, we fix an ordered field $(K, \leq)$, denote by $R := \overline{(K, \leq)}$ its real closure, we let $n \in \mathbb{N}_0$ and set $A := K[\underline{X}]$ and $T := \sum K_{\geq 0} A^2$. Then (A, T) is a preordered ring and for all $P \in \operatorname{sper}(A, T)$ there is by 1.7.5 exactly one homomorphism from R to the representation field R_P of P extending $\varrho_P|_K$ [→ 3.1.15]. In virtue of this homomorphism, which is of course an embedding of ordered fields, we interpret R as an (ordered) subfield of R_P. In particular, we write $R = R_P$ if it is an isomorphism.

Proposition and Notation 3.6.3. *The correspondence*

$$P \mapsto x_P := (\varrho_P(X_1), \ldots, \varrho_P(X_n))$$
$$\{f \in A \mid f(x) \geq 0\} =: P_x \hookleftarrow x$$

defines a bijection between $\{P \in \operatorname{sper}(A, T) \mid R_P = R\}$ *and* R^n.

Proof. We first show that both maps are well-defined. For every $P \in \operatorname{sper}(A, T)$ with $R_P = R$, we have $x_P \in R^n$ under the identification of R_P and R. Conversely, let $x \in R^n$. Consider the ring homomorphism

$$\varphi \colon A \to R,\ f \mapsto f(x).$$

Then $P_x = \varphi^{-1}(R^2) = (\operatorname{sper} \varphi)(R_{\geq 0}) \in \operatorname{sper} A$ [→ 3.1.4, 3.1.14(a)]. Obviously, $K_{\geq 0} \subseteq P_x$ and therefore $P_x \in \operatorname{sper}(A, T)$. In order to show $R_{P_x} = R$, we set $\mathfrak{p} := \operatorname{supp} P_x$ and consider the homomorphism of ordered fields

$$(\operatorname{qf}(A/\mathfrak{p}), (P_x)_{\mathfrak{p}}) \to (R, R^2),\ \frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} \mapsto \frac{a(x)}{s(x)} \qquad (a \in A, s \in A \setminus \mathfrak{p})$$

induced by φ according to 3.1.4 taking into account 3.1.14. Since R is real closed, this homomorphism extends (uniquely) to a homomorphism of (ordered) fields

$$\psi \colon R_{P_x} = \overline{(\operatorname{qf}(A/\mathfrak{p}), (P_x)_{\mathfrak{p}})} \to R.$$

We obviously have $\psi|_K = \mathrm{id}$ and therefore $\psi|_R$ is a K-endomorphism of the real closure R of $(K, \leq)$ which can only be the identity by 1.7.5. The injectivity of ψ now implies $R_{P_x} = R$ as desired. For later use we note that $\psi = \mathrm{id}_R$ implies

$$(*) \qquad \overline{f}^{\mathfrak{p}} = \psi^{-1}(\psi(\overline{f}^{\mathfrak{p}})) = \psi^{-1}(f(x)) = f(x)$$

for all $f \in A$.

It remains to show that both maps are inverse to each other. This means:

(a) $P = P_{x_P}$ for all $P \in \operatorname{sper}(A, T)$ with $R_P = R$

(b) $x = x_{P_x}$ for all $x \in R^n$

To show (a), let $P \in \operatorname{sper}(A, T)$ such that $R_P = R$. Then

$$\begin{aligned} P_{x_P} &= \{f \in A \mid f(\varrho_P(X_1), \ldots, \varrho_P(X_n)) \geq 0 \text{ in } R\} \\ &= \{f \in A \mid \varrho_P(f) \geq 0 \text{ in } R_P\} = \{f \in A \mid \widehat{f}(P) \geq 0\} = P. \end{aligned}$$

To show (b), we let $x \in R^n$. Then $\overline{X_i}^{\operatorname{supp} P_x} \overset{(*)}{=} x_i \in R$ for all $i \in \{1, \ldots, n\}$. Consequently, $x_{P_x} = (\varrho_{P_x}(X_1), \ldots, \varrho_{P_x}(X_n)) = (\overline{X_1}^{\operatorname{supp} P_x}, \ldots, \overline{X_n}^{\operatorname{supp} P_x}) = (x_1, \ldots, x_n) = x$. □

Theorem and Definition 3.6.4. [→ 1.9.3, 1.9.4] *Let $n \in \mathbb{N}_0$ and denote again by $\mathscr{S}_{n,R}$ the Boolean algebra of all K-semialgebraic subsets of R^n. Then*

$$\mathrm{Slim}\colon \mathscr{C}_{(A,T)} \to \mathscr{S}_{n,R},\ C \mapsto \{x \in R^n \mid P_x \in C\}$$

is an isomorphism of Boolean algebras. We call Slim *the* despectrification *or* slimming *(in German:* Entspeckung*) and*

$$\mathrm{Fatten} := \mathrm{Slim}^{-1}$$

the spectrification *or* fattening *(in German:* Verspeckung*). For all $f \in A$, one has*

$$\mathrm{Slim}(\{P \in \mathrm{sper}(A,T) \mid f \in P\}) = \{x \in R^n \mid f(x) \geq 0\}.$$

Proof. It is obvious that Slim is a homomorphism of Boolean algebras [→ 1.9.1] satisfying $\mathrm{Slim}(\{P \in \mathrm{sper}(A,T) \mid f \in P\}) = \{x \in R^n \mid f \in P_x\} = \{x \in R^n \mid f(x) \geq 0\}$ for all $f \in A$. Let $\mathscr{R} \supseteq \{R_P \mid P \in \mathrm{sper}(A,T)\}$ be a set of real closed fields that are ordered extension fields of $(K, \leq)$ [→ 1.8.4(b)]. Let $\mathscr{S}_n$ again denote the Boolean algebra of all $(K, \leq)$ -semialgebraic classes [→ 1.9.3] and consider

$$\Phi\colon \mathscr{S}_n \to \mathscr{C}_{(A,T)},\ S \mapsto \{P \in \mathrm{sper}(A,T) \mid (R_P, (\varrho_P(X_1), \ldots, \varrho_P(X_n))) \in S\}.$$

It is obvious that Φ is a homomorphism of Boolean algebras satisfying

$$\begin{aligned}\Phi(\{(R', x) \mid R' \in \mathscr{R}, x \in R'^n, f(x) \geq 0 \text{ in } R'\}) \\ = \{P \in \mathrm{sper}(A,T) \mid f(\varrho_P(X_1), \ldots, \varrho_P(X_n)) \geq 0 \text{ in } R_P\} \\ = \{P \in \mathrm{sper}(A,T) \mid \varrho_P(f) \geq 0 \text{ in } R_P\} \\ = \{P \in \mathrm{sper}(A,T) \mid \widehat{f}(P) \geq 0\} = \{P \in \mathrm{sper}(A,T) \mid f \in P\}\end{aligned}$$

for all $f \in A$. From this one sees, in the first place, that Φ is surjective and, secondly, that $\mathrm{Slim} \circ \Phi = \mathrm{Set}_R$ [→ 1.9.3] which is an isomorphism of Boolean algebras by 1.9.4. Along with Set_R, Φ is also injective. We conclude that Φ is an isomorphism and with it $\mathrm{Slim} = (\mathrm{Slim} \circ \Phi) \circ \Phi^{-1}$. □

Example 3.6.5. In 3.1.5, we have already described $\mathrm{sper}\,\mathbb{R}[X]$. Now we describe $\mathrm{sper}\,\mathbb{R}[X]$ as a set of prime cones [→ 3.1.14] while using 1.3.8: For $t \in \mathbb{R}$, we set

$$\begin{aligned}P_{t-} &:= \{f \in \mathbb{R}[X] \mid \exists \varepsilon \in \mathbb{R}_{>0} : \forall x \in (t-\varepsilon, t) : f(x) \geq 0\}, \\ P_t &:= \{f \in \mathbb{R}[X] \mid f(t) \geq 0\} \quad \text{and} \\ P_{t+} &:= \{f \in \mathbb{R}[X] \mid \exists \varepsilon \in \mathbb{R}_{>0} : \forall x \in (t, t+\varepsilon) : f(x) \geq 0\}\end{aligned}$$

Finally, we set

$$\begin{aligned}P_{-\infty} &:= \{f \in \mathbb{R}[X] \mid \exists c \in \mathbb{R} : \forall x \in (-\infty, c) : f(x) \geq 0\} \quad \text{and} \\ P_{\infty} &:= \{f \in \mathbb{R}[X] \mid \exists c \in \mathbb{R} : \forall x \in (c, \infty) : f(x) \geq 0\}.\end{aligned}$$

Then

$$\operatorname{sper}\mathbb{R}[X] = \{P_{-\infty}, P_\infty\} \cup \{P_{t-} \mid t \in \mathbb{R}\} \cup \{P_t \mid t \in \mathbb{R}\} \cup \{P_{t+} \mid t \in \mathbb{R}\}.$$

The fattening of the semialgebraic set $[0,1) \subseteq \mathbb{R}$ is the set

$$\begin{aligned} C := \{P_0, P_{0+}\} \cup \{P_{t-} \mid t \in (0,1)\} \cup \{P_t \mid t \in (0,1)\} \\ \cup \{P_{t+} \mid t \in (0,1)\} \cup \{P_{1-}\} \subseteq \operatorname{sper}\mathbb{R}[X]. \end{aligned}$$

In particular, C is constructible. In contrast to this, $C' := C \setminus \{P_{1-}\}$ is not constructible for otherwise C and C' would have the same slimming in contradiction to 3.6.4.

3.7 Real Stellensätze

Remark 3.7.1. Let A be a commutative ring.

(a) Since every intersection of $\left\{\begin{array}{c}\text{ideals}\\ \text{multiplicative sets}\\ \text{preorders}\end{array}\right\}$ of A is again $\left\{\begin{array}{c}\text{an ideal}\\ \text{a multiplicative set}\\ \text{a preorder}\end{array}\right\}$ of A, there exists for every subset $E \subseteq A$ $\left\{\begin{array}{c}\text{a smallest ideal}\\ \text{a smallest multiplicative set}\\ \text{a smallest preorder}\end{array}\right\}$ of A containing E. It is called the $\left\{\begin{array}{c}\text{ideal}\\ \text{multiplicative set}\\ \text{preorder}\end{array}\right\}$ *generated* by E.

(b) $\left\{\begin{array}{c}\text{An ideal}\\ \text{A multiplicative set}\\ \text{A preorder}\end{array}\right\}$ of A is called *finitely generated* if it is generated by a finite subset of A.

(c) The $\left\{\begin{array}{c}\text{ideal}\\ \text{multiplicative set}\\ \text{preorder}\end{array}\right\}$ generated by $a_1, \ldots, a_m \in A$ (i.e., by $\{a_1, \ldots, a_m\} \subseteq A$) is $\left\{\begin{array}{c} Aa_1 + \ldots + Aa_m \\ \{a_1^{\alpha_1} \cdots a_m^{\alpha_m} \mid \alpha \in \mathbb{N}_0^m\} \\ \sum_{\delta \in \{0,1\}^m} \sum A^2 a_1^{\delta_1} \cdots a_m^{\delta_m} \end{array}\right\}$.

(d) If $\left\{\begin{array}{c}\text{an ideal}\\ \text{a multiplicative set}\\ \text{a preorder}\end{array}\right\}$ of A is generated by $E \subseteq A$, then it is the union over all $\left\{\begin{array}{c}\text{ideals}\\ \text{multiplicative sets}\\ \text{preorders}\end{array}\right\}$ of A generated by a finite subset of E.

(e) If $\left\{\begin{array}{c}\text{an ideal } I\\ \text{a multiplicative set } S\\ \text{a preorder } T\end{array}\right\} \subseteq A$ is generated by $E \subseteq A$ and if $P \in \operatorname{sper} A$, then
$$\left\{\begin{array}{l}\forall a \in I : \widehat{a}(P) = 0\\ \forall s \in S : \widehat{s}(P) \neq 0\\ \forall t \in T : \widehat{t}(P) \geq 0\end{array}\right\} \iff \left\{\begin{array}{l}\forall a \in E : \widehat{a}(P) = 0\\ \forall s \in E : \widehat{s}(P) \neq 0\\ \forall t \in E : \widehat{t}(P) \geq 0\end{array}\right\}.$$

Remark 3.7.2. Let $(L, \leq)$ be an ordered field and K a subfield of L. If $\left\{\begin{array}{c}\text{an ideal } I\\ \text{a multiplicative set } S\\ \text{a preorder } T\end{array}\right\} \subseteq K[\underline{X}]$ is generated by $E \subseteq K[\underline{X}]$ and if $x \in L^n$, then
$$\left\{\begin{array}{l}\forall g \in I : g(x) = 0\\ \forall h \in S : h(x) \neq 0\\ \forall f \in T : f(x) \geq 0\end{array}\right\} \iff \left\{\begin{array}{l}\forall g \in E : g(x) = 0\\ \forall h \in E : h(x) \neq 0\\ \forall f \in E : f(x) \geq 0\end{array}\right\}.$$

Remark and Terminology 3.7.3. (a) "over B generated by E" stands for "generated by $B \cup E$"

(b) "over B finitely generated" stands for "generated by $B \cup E$ for some finite set E"

(c) If $(K, \leq)$ is an ordered field and $n \in \mathbb{N}_0$, then the preorder generated by $p_1, \ldots, p_m \in K[\underline{X}]$ over $K_{\geq 0}$ equals $\sum_{\delta \in \{0,1\}^m} \sum K_{\geq 0} K[\underline{X}]^2 p_1^{\delta_1} \cdots p_m^{\delta_m}$ [→ 3.7.1(c)].

Proposition 3.7.4. *Let $(K, \leq)$ be an ordered field, $R := \overline{(K, \leq)}$ and $n \in \mathbb{N}_0$ [→ 3.6.3]. Let I be an ideal, S a* finitely generated *multiplicative set and T a preorder of $K[\underline{X}]$* finitely generated over $K_{\geq 0}$. *Then*
$$\{P \in \operatorname{sper} K[\underline{X}] \mid (\forall g \in I : \widehat{g}(P) = 0), (\forall h \in S : \widehat{h}(P) \neq 0), (\forall f \in T : \widehat{f}(P) \geq 0)\}$$
is a constructible subset of $\operatorname{sper}(K[\underline{X}], \sum K_{\geq 0} K[\underline{X}]^2)$ *whose slimming is the K-semialgebraic set*
$$\{x \in R^n \mid (\forall g \in I : g(x) = 0), (\forall h \in S : h(x) \neq 0), (\forall f \in T : f(x) \geq 0)\}.$$

Proof. By Hilbert's basis theorem, I is finitely generated as well. Now use 3.7.1, 3.6.4 and 3.7.2. □

Theorem 3.7.5 (real Stellensatz [Kri, Ste, Pre])**.** [→ 3.4.2] *Let $(K, \leq)$ be an ordered subfield of the real closed field R, $n \in \mathbb{N}_0$, I an ideal of $K[\underline{X}]$, S a* finitely generated *multiplicative set of $K[\underline{X}]$ and T a preorder of $K[\underline{X}]$* finitely generated over $K_{\geq 0}$. *Then the following are equivalent:*

(a) *There does* not *exist any $x \in R^n$ satisfying*
$$\begin{array}{l}\forall g \in I : g(x) = 0,\\ \forall h \in S : h(x) \neq 0 \quad \textit{and}\\ \forall t \in T : t(x) \geq 0.\end{array}$$

(b) $0 \in I + S^2 + T$

Proof. (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (b) WLOG $R = \overline{(K, \leq)}$ [→ 1.7.11]. Because the fattening of the empty set is empty by 3.6.4 [→ 1.9.1], (a) implies Condition 3.4.2(a) from the abstract real Stellensatz applied to $A := K[\underline{X}]$. □

Corollary 3.7.6 (Positivstellensatz). [→ 3.4.4] *Let $(K, \leq)$ be an ordered subfield of the real closed field R, $n \in \mathbb{N}_0$, T a preorder of $K[\underline{X}]$* finitely generated over $K_{\geq 0}$,

$$S := \{x \in R^n \mid \forall p \in T : p(x) \geq 0\}$$

and $f \in K[\underline{X}]$. Then the following are equivalent:

(a) $f > 0$ *on* S

(b) $\exists t \in T : tf \in 1 + T$

(c) $\exists t \in T : (1 + t)f \in 1 + T$

Proof. Alternatively from 3.7.5 (as 3.4.4 from 3.4.2) or from 3.4.4 (as 3.7.5 from 3.4.2 using 3.6.4). □

Corollary 3.7.7 (Nichtnegativstellensatz). [→ 3.4.5] *Let $(K, \leq)$ be an ordered subfield of the real closed field R, $n \in \mathbb{N}_0$, T a preorder of $K[\underline{X}]$* finitely generated over $K_{\geq 0}$,

$$S := \{x \in R^n \mid \forall p \in T : p(x) \geq 0\}$$

and $f \in K[\underline{X}]$. Then the following are equivalent:

(a) $f \geq 0$ *on* S

(b) $\exists t \in T : \exists k \in \mathbb{N}_0 : tf \in f^{2k} + T$

Proof. Alternatively from 3.7.5 (as 3.4.5 from 3.4.2) or from 3.4.5 (as 3.7.5 from 3.4.2 using 3.6.4). □

Remark 3.7.8. In the special case $T = \sum K_{\geq 0} K[\underline{X}]^2$, the Nichtnegativstellensatz 3.7.7 is obviously a strengthening of Artin's solution 2.5.2 to Hilbert's 17th problem in which Condition (b) is refined. This refinement has the advantage that the proof of (b) $\Longrightarrow$ (a) does not require a real argument as it was the case in 2.5.2. The proof of 3.7.7 requires prime cones of rings instead of just preorders of fields and therefore is substantially more difficult as the proof of 2.5.2.

Corollary 3.7.9 (real Nullstellensatz [Kri, Du2, Ris, Efr]). [→ 3.4.6] *Let K be a Euclidean subfield of the real closed field R, $n \in \mathbb{N}_0$, I an ideal of $K[\underline{X}]$ and*

$$V := \{x \in R^n \mid \forall p \in I : p(x) = 0\}.$$

Then $\{f \in K[\underline{X}] \mid f = 0 \text{ on } V\} = \operatorname{rrad} I$.

Proof. Using the description of rrad I from 3.5.7, this follows alternatively from 3.7.5 (as 3.4.6 from 3.4.2) or from 3.4.6 (as 3.7.5 from 3.4.2 using 3.6.4). □

Definition 3.7.10. [→ 1.7.1] Let K be field. An extension field R of K is called *a* real closure of K if R is real closed and $R|K$ is algebraic.

Remark 3.7.11. For two fields K and R, the following are equivalent:

(a) R is a real closure of K.

(b) There is an order $\leq$ of K such that $R = \overline{(K, \leq)}$.

Theorem 3.7.12 (variant of the real Stellensatz)**.** [→ 3.7.5] *Let K be a field, $n \in \mathbb{N}_0$, I an ideal of $K[\underline{X}]$, S a* finitely generated *multiplicative set of $K[\underline{X}]$ and T a* finitely generated *preorder of $K[\underline{X}]$. Then the following are equivalent:*

(a) *There does* not *exist a real closure R of K and an $x \in R^n$ such that*

$$\begin{aligned} &\forall g \in I : g(x) = 0, \\ &\forall h \in S : h(x) \neq 0 \quad \textit{and} \\ &\forall f \in T : f(x) \geq 0. \end{aligned}$$

(b) $0 \in I + S^2 + T$

Proof. (b) $\Longrightarrow$ (a) is trivial.

We show (a) $\Longrightarrow$ (b) by contraposition. Suppose (b) does not hold. We have to show that (a) does not hold. By the abstract real Stellensatz 3.4.2, there is some $P \in \operatorname{sper} K[\underline{X}]$ such that $\forall g \in I : \widehat{g}(P) = 0$, $\forall h \in S : \widehat{h}(P) \neq 0$ and $\forall f \in T : \widehat{f}(P) \geq 0$ all hold at the same time. Now consider the real closure $R := \overline{(K, K \cap P)}$ of K and the preordered ring $(K[\underline{X}], \sum(K \cap P)K[\underline{X}]^2)$. The set

$$U := \{x \in R^n \mid (\forall g \in I : g(x) = 0), (\forall h \in S : h(x) \neq 0), (\forall f \in T : f(x) \geq 0)\}$$

is K-semialgebraic by 3.7.1(e) since I, S and T are finitely generated. We will show that $U \neq \emptyset$. We have chosen P to be an element of the constructible subset of the real spectrum of this preordered ring which is the fattening of U, i.e., $P \in \operatorname{Fatten}(U)$ in the notation of 3.6.4. In particular, $\operatorname{Fatten}(U) \neq \emptyset$ and thus $U \neq \emptyset$ by 3.6.4 □

Remark 3.7.13. Wherever the hypothesis "finitely generated" appears in this section, it cannot be omitted. For instance, assume that the Positivstellensatz 3.7.6 holds with the weaker hypothesis "$K_{\geq 0} \subseteq T$" instead of "T finitely generated over $K_{\geq 0}$". Consider then $K := R := \mathbb{R}$, $n := 1$ and the preorder of $\mathbb{R}[X]$ generated by

$$E := \{X - N \mid N \in \mathbb{N}\}.$$

Then $S := \{x \in \mathbb{R} \mid \forall p \in T : p(x) \geq 0\} = \emptyset$ and thus $f := -1 > 0$ on S. It follows that $\exists t \in T : tf \in 1 + T$ and thus by 3.7.1(d) even $\exists t \in T' : tf \in 1 + T'$ for a preorder $T' \subseteq T$ generated by a finite set $E' \subseteq E$. The trivial direction of 3.7.6 then yields $-1 > 0$ on $S' := \{x \in \mathbb{R} \mid \forall p \in T' : p(x) \geq 0\} \overset{3.7.2}{=} \{x \in \mathbb{R} \mid \forall p \in E' : p(x) \geq 0\} = [N, \infty)$ for some $N \in \mathbb{N}$. ↯

Remark 3.7.14. [→ 1.2.10] Let A be a commutative ring and $T \subseteq A$ a proper preorder. Exactly as in the field case, there exists some $P \in \operatorname{sper} A$ such that $T \subseteq P$ [→ 3.2.3]. In sharp contrast, to the field case we do in general however not have that $T = \bigcap \operatorname{sper}(A, T)$. As an example, take $A := \mathbb{R}[X, Y]$, $T := \sum \mathbb{R}[X, Y]^2$ and consider the Motzkin polynomial $f := X^4Y^2 + X^2Y^4 - 3X^2Y^2 + 1$. By 2.4.16, we have $f \notin T$ and $S := \{(x, y) \in \mathbb{R}^2 \mid f(x, y) \geq 0\} = \mathbb{R}^2$. By 3.6.4, the fattening

$$C := \{P \in \operatorname{sper} A \mid f \in P\} \subseteq \operatorname{sper} A = \operatorname{sper}(A, T)$$

of S equals the whole of $\operatorname{sper} A$, i.e., $f \in \bigcap \operatorname{sper}(A, T)$.

§4 Schmüdgen's Positivstellensatz

4.1 The abstract Archimedean Positivstellensatz

Definition 4.1.1. [→ 1.1.20(d)] A preordered ring (A,T) is called *Archimedean* if

$$\forall a \in A : \exists N \in \mathbb{N} : N + a \in T,$$

which is equivalent to $T - \mathbb{N} = A$ and also to $T + \mathbb{Z} = A$.

Definition 4.1.2. Let A be a commutative ring.

(a) A preorder T of A is called Archimedean if (A,T) is Archimedean.

(b) A is called Archimedean if $(A, \sum A^2)$ is Archimedean.

Theorem 4.1.3 (abstract Archimedean Positivstellensatz [Sto, Kad, Kri, Du1])**.** [→ 3.4.4] *Let (A,T) be an Archimedean preordered ring and $a \in A$. Then the following are equivalent:*

(a) $\widehat{a} > 0$ *on* $\operatorname{sper}(A,T)$

(b) $\exists N \in \mathbb{N} : Na \in 1 + T$

Proof. (b) $\Longrightarrow$ (a) is trivial

(a) $\Longrightarrow$ (b) For the multiplicative set $S := \mathbb{N} \cdot 1 \subseteq A$, $(S^{-1}A, S^{-2}T)$ is again an Archimedean preordered ring [→ 3.3.3] and we have [→ 3.3.4]

$$\widehat{a} > 0 \text{ on } \operatorname{sper}(A,T) \iff \widehat{\left(\frac{a}{1}\right)} > 0 \text{ on } \operatorname{sper}(S^{-1}A, S^{-2}T).$$

We can therefore suppose $\mathbb{N} \cdot 1 \subseteq A^\times$ and therefore have a homomorphism

$$\mathbb{Q} = \mathbb{N}^{-1}\mathbb{Z} \to A, \ \frac{p}{q} \mapsto \frac{p}{q} \qquad (p \in \mathbb{Z}, q \in \mathbb{N}).$$

Suppose now that (a) holds. By the abstract Positivstellensatz 3.4.4, there is some $t \in T$ such that $ta \in 1 + T$. Since T is Archimedean, there are $N \in \mathbb{N}$ with $N - t \in T$ and $r \in \mathbb{N}$ with $a + r \in T$. Now you can decrease $r \in \frac{1}{N}\mathbb{N}_0$ a finite number of times by $\frac{1}{N}$ until it gets negative since

$$a + \left(r - \frac{1}{N}\right) = \frac{N}{N^2}((\underbrace{N-t}_{\in T})(\underbrace{a+r}_{\in T}) + (\underbrace{ta-1}_{\in T}) + \underbrace{rt}_{\in T}) \in T$$

as long as $r \geq 0$. It follows $a - \frac{1}{N} \in T$ and thus $Na \in 1 + T$. □

Corollary 4.1.4. *Let A be a commutative ring and $P \in \operatorname{sper} A$. Then the following are equivalent:*

(a) *P is Archimedean and a maximal prime cone.* [→ 4.1.2(a)]

(b) *$(\operatorname{qf}(A/\mathfrak{p}), P_{\mathfrak{p}})$ is Archimedean where $\mathfrak{p} := \operatorname{supp} P$.* [→ 3.1.11]

(c) *R_P is Archimedean.* [→ 3.1.15]

(d) *There exists a homomorphism $\varphi\colon A \to \mathbb{R}$ such that $P = (\operatorname{sper}\varphi)(\mathbb{R}_{\geq 0})$.*

Proof. (a) $\Longrightarrow$ (b) Suppose (a) holds and let $a \in A$ and $s \in A \setminus \mathfrak{p}$. To show: $\exists N \in \mathbb{N} : \frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} + N \in P_{\mathfrak{p}}$. WLOG $s \in P$. Since P is maximal, we have $\operatorname{sper}(A,P) = \{P\}$ and thus $\widehat{s} > 0$ on $\operatorname{sper}(A,P)$. By 4.1.3, there is $N' \in \mathbb{N}$ such that $N's \in 1 + P$. Choose $N'' \in \mathbb{N}$ such that $a + N'' \in P$ and set $N := N'N''$. Then $a + Ns = a + N'N''s \in a + N'' + P \subseteq P + P \subseteq P$ and thus $(a + Ns)s \in PP \subseteq P$. It follows that $\frac{\overline{a}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} + N = \frac{\overline{a+Ns}^{\mathfrak{p}}}{\overline{s}^{\mathfrak{p}}} \in P_{\mathfrak{p}}$.

(b) $\Longrightarrow$ (c) If (b) holds, then $(\operatorname{qf}(A/\mathfrak{p}), P_{\mathfrak{p}}) \hookrightarrow (\mathbb{R}, \mathbb{R}_{\geq 0})$ by 1.1.17 [→ 1.1.5] and

$$R_P = \overline{(\operatorname{qf}(A/\mathfrak{p}), P_{\mathfrak{p}})} \hookrightarrow (\mathbb{R}, \mathbb{R}_{\geq 0})$$

by 1.7.5.

(c) $\Longrightarrow$ (d) Choose an embedding $\iota\colon R_P \hookrightarrow \mathbb{R}$ according to 1.1.17. We have $\iota^{-1}(\mathbb{R}_{\geq 0}) = (R_P)_{\geq 0}$ because ι is an embedding of ordered fields. Now set $\varphi := \iota \circ \varrho_P$. Then $\varphi^{-1}(\mathbb{R}_{\geq 0}) = \varrho_P^{-1}(\iota^{-1}(\mathbb{R}_{\geq 0})) = \varrho_P^{-1}((R_P)_{\geq 0}) \overset{3.1.16}{=} P$.

(d) $\Longrightarrow$ (a) Suppose $\varphi\colon A \to \mathbb{R}$ is a homomorphism with $P = \varphi^{-1}(\mathbb{R}_{\geq 0})$. Then P is Archimedean for if $a \in A$, then one can choose $N \in \mathbb{N}$ with $\varphi(a) + N \geq 0$ and it follows that $a + N \in \varphi^{-1}(\mathbb{R}_{\geq 0}) = P$. In order to show that P is maximal, let $Q \in \operatorname{sper} A$ with $P \subseteq Q$. To show: $P = Q$. If we had $a \in Q \setminus P$, then $\varphi(a) < 0$ and thus $\varphi(Na) \leq -1$ for some $N \in \mathbb{N}$ from which it would follow that $\varphi(-1 - Na) \geq 0$ and thus $-1 - Na \in P \subseteq Q$ and $-1 = (-1 - Na) + Na \in Q + Q \subseteq Q$ ↯. □

4.2 The Archimedean Positivstellensatz [→ §3.7]

Lemma 4.2.1. Suppose $(K, \leq)$ is an ordered subfield of $\mathbb{R}$, $n \in \mathbb{N}_0$ and $K_{\geq 0} \subseteq T \subseteq K[\underline{X}]$. Then the correspondence

$$\begin{aligned} x \mapsto \operatorname{ev}_x\colon K[\underline{X}] \to \mathbb{R},\ p \mapsto p(x) \\ (\varphi(X_1), \ldots, \varphi(X_n)) \mapsfrom \varphi \end{aligned}$$

defines a bijection between $S := \{x \in \mathbb{R}^n \mid \forall p \in T : p(x) \geq 0\}$ and the set of all ring homomorphisms $\varphi\colon K[\underline{X}] \to \mathbb{R}$ satisfying $\varphi(T) \subseteq \mathbb{R}_{\geq 0}$.

Proof. It is obviously enough to show that every ring homomorphism $\varphi\colon K[\underline{X}] \to \mathbb{R}$ with $\varphi(T) \subseteq \mathbb{R}_{\geq 0}$ is the identity on K. But this is clear by 1.1.15 since the identity is the *only* embedding of ordered fields $(K, \leq) \hookrightarrow (\mathbb{R}, \mathbb{R}_{\geq 0})$. □

Theorem 4.2.2 (Archimedean Positivstellensatz). [→ 4.1.3, 3.7.6] *Suppose $(K, \leq)$ is an ordered subfield of $\mathbb{R}$, $n \in \mathbb{N}_0$, $T \subseteq K[\underline{X}]$ is an* Archimedean *preorder containing $K_{\geq 0}$, $S := \{x \in \mathbb{R}^n \mid \forall p \in T : p(x) \geq 0\}$ and $f \in K[\underline{X}]$. Then the following are equivalent:*

(a) $f > 0$ *on* S

(b) $\exists N \in \mathbb{N} : f \in \frac{1}{N} + T$

Proof. (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (b) Suppose that (a) holds. It is enough to show that $\widehat{f} > 0$ on $\operatorname{sper}(K[\underline{X}], T)$ due to the abstract Archimedian Positivstellensatz 4.1.3 using $\frac{1}{N} = N\left(\frac{1}{N}\right)^2$. To this end, let $P \in \operatorname{sper}(K[\underline{X}], T)$. Choose a maximal prime cone Q of $K[\underline{X}]$ such that $P \subseteq Q$ by 3.2.3. Along with P, also Q is of course Archimedean. By 3.2.3, we have $Q = P \cup \mathfrak{q}$ where $\mathfrak{q} := \operatorname{supp} Q$. It follows that $Q \setminus -Q \subseteq P \setminus -P$ and therefore it is enough to show that $f \in Q \setminus -Q$. By 4.1.4(d) and 4.2.1, there is some $x \in S$ satisfying $Q = \operatorname{ev}_x^{-1}(\mathbb{R}_{\geq 0}) = \{p \in K[\underline{X}] \mid p(x) \geq 0\}$. From $f(x) > 0$, we deduce now $f \in Q \setminus -Q$ as desired. □

Remark 4.2.3. If T is finitely generated over $K_{\geq 0}$ in the situation of 4.2.2, then one can reduce 4.2.2 alternatively by fattening to 4.1.3. This ultimately uses unnecessarily the heavy artillery of real quantifier elimination 1.8.17 and is not applicable if T is not finitely generated over $K_{\geq 0}$. The principal reason why the real quantifier elimination is not needed here is 1.1.17.

4.3 Schmüdgen's characterization of Archimedean preorders of the polynomial ring

Definition and Proposition 4.3.1. *Let (A, T) be a preordered ring. Then*

$$B_{(A,T)} := \{a \in A \mid \exists N \in \mathbb{N} : N \pm a \in T\}$$

is a subring of A which we call the ring of with respect to T arithmetically bounded *elements of A.*

Proof. One sees immediately that $B_{(A,T)}$ is a subgroup of the additive group of A. It is clear that $1 \in B_{(A,T)}$. Finally, we have $B_{(A,T)}B_{(A,T)} \subseteq B_{(A,T)}$ as one sees immediately from the identity

$$3N^2 \pm ab = (N + a)(N \pm b) + N(N - a) + N(N \mp b) \qquad (N \in \mathbb{N}, a, b \in A).$$

□

Lemma 4.3.2. Let (A, T) be a preordered ring such that $\frac{1}{2} \in A$. Then

$$a^2 \in B_{(A,T)} \implies a \in B_{(A,T)}$$

for all $a \in A$.

Proof. Choose $N \in \mathbb{N}$ with $(N-1) - a^2 \in T$. Then

$$N \pm a = (N-1) - a^2 + \left(\frac{1}{2} \pm a\right)^2 + 3\left(\frac{1}{2}\right)^2 \in T.$$

□

Remark 4.3.3. If (A, T) is a preordered ring, then T is Archimedean if and only if $B_{(A,T)} = A$.

Lemma 4.3.4. Suppose $(K, \leq)$ is an ordered subfield of $\mathbb{R}$, $n \in \mathbb{N}_0$ and $T \subseteq K[\underline{X}]$ is a preorder containing $K_{\geq 0}$. Then the following are equivalent:

(a) T is Archimedean.

(b) $\exists N \in \mathbb{N} : N - \sum_{i=1}^n X_i^2 \in T$

(c) $\exists N \in \mathbb{N} : \forall i \in \{1, \ldots, n\} : N \pm X_i \in T$

Proof. (a) $\Longrightarrow$ (b) is trivial.

(b) $\Longrightarrow$ (c) If (b) holds, then $N - X_i^2 \in T$ and thus $X_i^2 \in B_{(A,T)}$ for all $i \in \{1, \ldots, n\}$. Now apply 4.3.2.

(c) $\Longrightarrow$ (a) Since $(K, \leq)$ is Archimedean and $K_{\geq 0} \subseteq T$, we have $K \subseteq B_{(A,T)}$. If now moreover (c) holds, then $K[\underline{X}] = B_{(A,T)}$. □

Theorem 4.3.5 (Schmüdgen's Theorem [Sch, BW])**.** *Suppose $(K, \leq)$ is an ordered subfield of $\mathbb{R}$, $n \in \mathbb{N}_0$ and T a preorder of $K[\underline{X}]$ which is finitely generated* over $K_{\geq 0}$. *Write*

$$S := \{x \in \mathbb{R}^n \mid \forall p \in T : p(x) \geq 0\}.$$

Then

$$T \text{ Archimedean} \iff S \text{ compact}.$$

Proof. [BW] "$\Longrightarrow$" Let T be Archimedean. By 4.3.4(b), there is some $N \in \mathbb{N}$ with $N - \sum_{i=1}^n X_i^2 \in T$. Then S is contained in the ball of radius $\sqrt{N}$ centered at the origin and thus bounded. Anyway S is already closed. Consequently, S is compact.

"$\Longleftarrow$" Let S be compact. Write $r := \sum_{i=1}^n X_i^2$ and choose $N \in \mathbb{N}$ such that $N - r > 0$ on S. By the Positivstellensatz 3.7.6, we find $t \in T$ such that $(1+t)(N-r) \in 1 + T \subseteq T$. We know that $T' := T + (N-r)T$ is a preorder of $K[\underline{X}]$ that is Archimedean by 4.3.4. We have $(1+t)T' \subseteq T$ and $N - r + Nt = (1+t)(N-r) + tr \in T + T \subseteq T$. Choose $N' \in \mathbb{N}$ with $N' - t \in T'$. Then

$$(1+N')(N'-t) = (1+t)(N'-t) + (N'-t)^2 \in (1+t)T' + T \subseteq T + T \subseteq T$$

from which $N' - t \in T$ follows because of $\frac{1}{1+N'} = (1+N')\left(\frac{1}{1+N'}\right)^2 \in T$. We conclude that

$$N(N'+1) - r = NN' + N - r = (N - r + tN) + N(N' - t) \in T + T \subseteq T.$$

Now 4.3.4 implies that T is Archimedean. □

Corollary 4.3.6 (Schmüdgen's Positivstellensatz)**.** [→ 4.2.2] *Suppose* $(K, \leq)$ *is an ordered subfield of* $\mathbb{R}$, $n \in \mathbb{N}_0$, T *a preorder of* $K[\underline{X}]$ *which is finitely generated* over $K_{\geq 0}$. *Suppose* $S := \{x \in \mathbb{R}^n \mid \forall p \in T : p(x) \geq 0\}$ *is* compact *and* $f \in K[\underline{X}]$. *Then the following are equivalent:*

(a) $f > 0$ *on* S

(b) $\exists N \in \mathbb{N} : f \in \frac{1}{N} + T$

Proof. By Schmüdgen's Theorem 4.3.5, T is Archimedean. But then the Archimedean Positivstellensatz 4.2.2 proves the equivalence of (a) and (b). □

Remark 4.3.7. (a) Exactly as in 3.7.13, one sees that the hypothesis "T finitely generated over $K_{\geq 0}$" cannot be replaced by the weaker hypothesis "$K_{\geq 0} \subseteq T$".

(b) If one drops the requirement that S is compact, then 4.3.6 gets wrong as the example $K := \mathbb{R}$, $n := 1$, $T := \sum \mathbb{R}[X]^2 + \sum \mathbb{R}[X]^2 X^3$ and $f := X + 1$ shows: We have $f > 0$ on $S = [0, \infty)$ but $f \notin T$ for degree reasons as one sees from 2.2.4(b).

(c) In the situation of 4.3.6, we unfortunately do not have in general

$$f \geq 0 \text{ on } S \iff f \in T.$$

For this, consider $K := \mathbb{R}$, $n := 1$, $T := \sum \mathbb{R}[X]^2 + \sum \mathbb{R}[X]^2 X^3(1 - X)$ and $f := X$. Then $f \geq 0$ on $S = [0, 1]$. Assume $f \in T$. Write $f = \sum_i p_i^2 + \sum_j q_j^2 X^3(1 - X)$ for some $p_i, q_j \in \mathbb{R}[X]$. Evaluating in 0, yields $0 = \sum_i p_i(0)^2$ and thus $p_i(0) = 0$ for all i. Write $p_i = X p_i'$ for some $p_i' \in \mathbb{R}[X]$. Then $X = f = X^2 \left(\sum_i p_i'^2 + \sum_j q_j^2 X(1 - X) \right)$ ↯.

§5 The real spectrum as a topological space

5.1 Tikhonov's theorem

Remark 5.1.1. Any finite intersection of unions of certain sets is a union of finite intersections of such sets [→ 1.8.1].

Reminder 5.1.2. [→ 1.8.2] Let M be a set.

(a) A set $\mathscr{O} \subseteq \mathscr{P}(M)$ is called a *topology* on M if
- $M \in \mathscr{O}$,
- $\forall A_1, A_2 \in \mathscr{O} : A_1 \cap A_2 \in \mathscr{O}$ and
- $\forall \mathscr{A} \subseteq \mathscr{O} : \bigcup \mathscr{A} \in \mathscr{O}$.

In this case, $(M, \mathscr{O})$ is called a *topological space* and the elements of $\mathscr{O}$ are called its *open sets*.

(b) Let $\mathscr{G} \subseteq \mathscr{P}(M)$. Then the set of all unions of finite intersections of elements of $\mathscr{G}$ (where $\bigcap \emptyset := M$) is obviously the smallest topology $\mathscr{O}$ on M such that $\mathscr{G} \subseteq \mathscr{O}$. It is called the topology *generated by* $\mathscr{G}$ (on M).

(c) If $\mathscr{O}$ and $\mathscr{O}'$ are topologies on M, then $\mathscr{O}$ is called $\left\{\begin{matrix} \textit{coarser} \\ \textit{finer} \end{matrix}\right\}$ than $\mathscr{O}'$ if $\left\{\begin{matrix} \mathscr{O} \subseteq \mathscr{O}' \\ \mathscr{O} \supseteq \mathscr{O}' \end{matrix}\right\}$.

(d) The finest topology on M is the *discrete topology* $\mathscr{O} := \mathscr{P}(M)$.

(e) The coarsest topology on M is the *trivial topology* (in German: *Klumpentopologie*) $\mathscr{O} := \{\emptyset, M\}$.

Reminder 5.1.3. Let $(M, \mathscr{O})$ and $(N, \mathscr{P})$ be topological spaces and $f : M \to N$ be a map. Then f is called *continuous* if $f^{-1}(B) \in \mathscr{O}$ for all $B \in \mathscr{P}$.

Reminder 5.1.4. Let M be a set, $(N_i, \mathscr{P}_i)_{i \in I}$ a family of topological spaces and $(f_i)_{i \in I}$ a family of maps $f_i : M \to N_i$ $(i \in I)$. Then there exists a coarsest topology $\mathscr{O}$ on M making all f_i $(i \in I)$ continuous. One calls $\mathscr{O}$ the *initial topology* (or *weak topology*) with respect to $(f_i)_{i \in I}$. If $I = \{1, \ldots, n\}$, then $\mathscr{O}$ is also called the initial topology with respect to $f_1, \ldots, f_n$. This topology $\mathscr{O}$ is generated by $\{f_i^{-1}(B_i) \mid i \in I, B \in \mathscr{P}_i\}$. More generally, the following holds: If $\mathscr{P}_i$ is generated by $\mathscr{G}_i$ for $i \in I$, then $\mathscr{O}$ is generated by $\{f_i^{-1}(B) \mid i \in I, B \in \mathscr{G}_i\}$. It holds that $\mathscr{O}$ is the unique topology on M having the following property: For every topological space $(M', \mathscr{O}')$ and every $g : M' \to M$, the map g is continuous if and only if all the maps $f_i \circ g$ with $i \in I$ are continuous.

Example 5.1.5. (a) Let $(N, \mathscr{P})$ be a topological space and $M \subseteq N$. Then one endows M with the initial topology $\mathscr{O}$ with respect to $M \to N$, $x \mapsto x$. One calls $\mathscr{O}$ the topology *induced* by $\mathscr{P}$ on M and $(M, \mathscr{O})$ a *subspace* of $(N, \mathscr{P})$. We have

$$\mathscr{O} = \{M \cap B \mid B \in \mathscr{P}\}.$$

(b) Let $(N_i, \mathscr{P}_i)_{i \in I}$ be a family of topological spaces. Then there exists a coarsest topology $\mathscr{O}$ on $N := \prod_{i \in I} N_i$ making all projections $\pi_i \colon N \to N_i$, $(x_j)_{j \in I} \mapsto x_i$ $(i \in I)$ continuous. One calls $\mathscr{O}$ the *product topology* of the $\mathscr{P}_i$ $(i \in I)$ on N and $(N, \mathscr{O})$ the *product space* of the (N_i, P_i) $(i \in I)$. The elements of $\mathscr{O}$ are exactly the unions of sets of the form $\prod_{i \in I} B_i$ where $B_i \in \mathscr{P}_i$ for $i \in I$ and $\#\{i \in I \mid B_i \neq N_i\} < \infty$.

Remark 5.1.6. The constructions (a) and (b) in Example 5.1.5 commute in the following sense: Let $(N_i, \mathscr{P}_i)_{i \in I}$ be a family of topological spaces and $(N, \mathscr{O})$ its product. Furthermore, let $(M_i)_{i \in I}$ be a family of sets such that $M_i \subseteq N_i$ for each $i \in I$ and set $M := \prod_{i \in I} M_i$. Then $\mathscr{O}$ induces on M the product topology of the topologies induced on the M_i by the $\mathscr{P}_i$.

Definition 5.1.7. Let M be a set and $\mathscr{S}$ a Boolean algebra on M [$\to$ 1.8.2] (for instance $\mathscr{S} = \mathscr{P}(M)$). A set $\mathscr{F} \subseteq \mathscr{S}$ is called a *filter* in $\mathscr{S}$ (or filter on M in case $\mathscr{S} = \mathscr{P}(M)$) if

- $\emptyset \notin \mathscr{F}, M \in \mathscr{F}$,
- $\forall U, V \in \mathscr{F} : U \cap V \in \mathscr{F}$ and
- $\forall U \in \mathscr{F} : \forall V \in \mathscr{S} : (U \subseteq V \implies V \in \mathscr{F})$.

If in addition $\forall U \in \mathscr{S} : (U \in \mathscr{F}$ or $\complement U \in \mathscr{F})$, then $\mathscr{F}$ is called an *ultrafilter*.

Proposition 5.1.8. *Let $\mathscr{S}$ be a Boolean algebra on the set M and $\mathscr{F}$ a filter in $\mathscr{S}$. Then the following are equivalent:*

(a) *$\mathscr{F}$ is an ultrafilter.*

(b) $\forall U, V \in \mathscr{S} : (U \cup V \in \mathscr{F} \implies (U \in \mathscr{F}$ *or* $V \in \mathscr{F}))$

Proof. (a) $\implies$ (b) Suppose that (a) holds and let $U, V \in \mathscr{S}$ such that $U \cup V \in \mathscr{F}$ and $U \notin F$. To show: $V \in \mathscr{F}$. Since $\mathscr{F}$ is an ultrafilter, we have $\complement U \in \mathscr{F}$ and thus $(U \cup V) \cap \complement U \in \mathscr{F}$. Because of $(U \cup V) \cap \complement U \subseteq V$ it then also holds that $V \in \mathscr{F}$.

(b) $\implies$ (a) is trivial. □

Example 5.1.9. Let $(M, \mathscr{O})$ be a topological space and $x \in M$. Then

$$\mathscr{U}_x := \{U \in \mathscr{P}(M) \mid \exists A \in \mathscr{O} : x \in A \subseteq U\}$$

is a filter on M. One calls $\mathscr{U}_x$ the *neighborhood filter* of x and its elements the *neighborhoods* of x. In general, $\mathscr{U}_x$ is not an ultrafilter since $[-1, 1] = [-1, 0] \cup [0, 1]$ is a neighborhood of 0 in $\mathbb{R}$ as opposed to $[-1, 0]$ and $[0, 1]$.

Definition 5.1.10. Let $(M, \mathscr{O})$ be a topological space and $\mathscr{F}$ a filter on M. For $x \in M$, one says that $\mathscr{F}$ *converges* to x and writes $\mathscr{F} \to x$ if $\mathscr{U}_x \subseteq \mathscr{F}$. If $\mathscr{F}$ converges to exactly one point x, one calls this the *limit* of $\mathscr{F}$ and writes $x = \lim \mathscr{F}$.

Example 5.1.11. Let $(M, \mathscr{O})$ be a topological space and $(a_n)_{n\in\mathbb{N}}$ a sequence in M. Then $\mathscr{F} := \{U \in \mathscr{P}(M) \mid \exists N \in \mathbb{N} : \{a_n \mid n \geq N\} \subseteq U\}$ is a filter on M. For $x \in M$, the sequence $(a_n)_{n\in\mathbb{N}}$ converges to x if and only if $\mathscr{F}$ converges to x.

Definition and Lemma 5.1.12. Suppose $f\colon M \to N$ is a map and $\mathscr{F}$ a filter on M. Then the *image filter* $f(\mathscr{F}) := \{V \in \mathscr{P}(N) \mid \exists U \in \mathscr{F} : f(U) \subseteq V\}$ is a filter on N. If $\mathscr{F}$ is an ultrafilter, then so is $f(\mathscr{F})$.

Proof. One sees immediately that $f(\mathscr{F})$ is a filter. Now let $\mathscr{F}$ be an ultrafilter. Suppose $V \subseteq N$ and $V \notin f(\mathscr{F})$. To show: $\complement V \in f(\mathscr{F})$. For $U := f^{-1}(V)$, one has $f(U) \subseteq V$ and thus $U \notin \mathscr{F}$. But then $f^{-1}(\complement V) = \complement U \in \mathscr{F}$. From $f(\complement U) \subseteq \complement V$, we obtain thus $\complement V \in f(\mathscr{F})$. □

Lemma 5.1.13. Let M be a set endowed with the initial topology with respect to a family $(f_i)_{i\in I}$ of maps $f_i\colon M \to N_i$ $(i \in I)$ into topological spaces N_i $(i \in I)$. Let $\mathscr{F}$ be a filter and $x \in M$. Then $\mathscr{F} \to x \iff \forall i \in I : f_i(\mathscr{F}) \to f_i(x)$.

Proof. "$\Longrightarrow$" Suppose $\mathscr{F} \to x$ and let $i \in I$. To show: $f_i(\mathscr{F}) \to f_i(x)$. Let $V \in \mathscr{U}_{f_i(x)}$. To show: $V \in f_i(\mathscr{F})$. Since f_i is continuous, we have $U := f_i^{-1}(V) \in \mathscr{U}_x$ and thus $U \in \mathscr{F}$. From $f_i(U) \subseteq V$, we get $V \in f_i(\mathscr{F})$.

"$\Longleftarrow$" Suppose $f_i(\mathscr{F}) \to f_i(x)$ for all $i \in I$. Let $U \in \mathscr{U}_x$. To show: $U \in \mathscr{F}$. Choose $n \in \mathbb{N}_0$, $i_1, \ldots, i_n \in I$ and V_k open in N_{i_k} $(k \in \{1, \ldots, n\})$ such that

$$x \in f_{i_1}^{-1}(V_1) \cap \ldots \cap f_{i_n}^{-1}(V_n) \subseteq U.$$

Since $\mathscr{F}$ is a filter, it is enough to show that $f_{i_k}^{-1}(V_k) \in \mathscr{F}$ for all $k \in \{1, \ldots, n\}$. Fix therefore $k \in \{1, \ldots, n\}$. Since V_k is an (open) neighborhood of $f_{i_k}(x)$ in N_{i_k}, the hypothesis yields $V_k \in f_{i_k}(\mathscr{F})$. Hence there is $U_0 \in \mathscr{F}$ such that $f_{i_k}(U_0) \subseteq V_k$. Now everything follows from $U_0 \subseteq f_{i_k}^{-1}(f_{i_k}(U_0)) \subseteq f_{i_k}^{-1}(V_k)$. □

Definition 5.1.14. Let $(M, \mathscr{O})$ be a topological space. Then $(M, \mathscr{O})$ is called a *Hausdorff* space if every two distinct points of M can be separated by disjoint neighborhoods, i.e.,

$$\forall x, y \in M : (x \neq y \implies \exists U \in \mathscr{U}_x : \exists V \in \mathscr{U}_y : U \cap V = \emptyset).$$

We call $(M, \mathscr{O})$ *quasicompact* if every open cover of M possesses a finite subcover, i.e.,

$$\forall \mathscr{A} \subseteq \mathscr{O} : \left(M = \bigcup \mathscr{A} \implies \exists \mathscr{B} \subseteq \mathscr{A} : \left(\#\mathscr{B} < \infty \ \&\ M = \bigcup \mathscr{B}\right)\right).$$

Furthermore, we call a quasicompact Hausdorff space *compact*.

Proposition 5.1.15. *Suppose M is a set, $\mathscr{S}$ a Boolean algebra on M and $\mathscr{U}$ a filter on $\mathscr{S}$. Then the following are equivalent:*

(a) $\mathscr{U}$ *is an ultrafilter in* $\mathscr{S}$.

(b) $\mathscr{U}$ *is a maximal filter in* $\mathscr{S}$.

Proof. (a) $\Longrightarrow$ (b) Suppose that (a) holds and let $\mathscr{F}$ be a filter in $\mathscr{S}$ such that $\mathscr{U} \subseteq \mathscr{F}$. In order to show $\mathscr{F} \subseteq \mathscr{U}$, we fix $U \in \mathscr{F}$. If we had $U \notin \mathscr{U}$, we would get $\complement U \in \mathscr{U} \subseteq \mathscr{F}$ and thus $\emptyset = U \cap \complement U \in \mathscr{F}$ ↯.

(b) $\Longrightarrow$ (a) Suppose that (b) holds and let $U \in \mathscr{S}$ satisfy $U \notin \mathscr{U}$. To show: $\complement U \in \mathscr{U}$. It is enough to show that $\mathscr{F} := \{V \in \mathscr{S} \mid \exists A \in \mathscr{U} : A \cap \complement U \subseteq V\}$ is a filter in $\mathscr{S}$ because then it follows from $\mathscr{U} \subseteq \mathscr{F}$ that $\complement U \in \mathscr{F} = \mathscr{U}$. For this, it suffices to show $\emptyset \notin \mathscr{F}$. If we had $\emptyset \in \mathscr{F}$, then there would be an $A \in \mathscr{U}$ satisfying $A \cap \complement U = \emptyset$ and from $A \subseteq U$ it would follow that $U \in \mathscr{U}$ ↯. □

Theorem 5.1.16. *Let M be a set, $\mathscr{S}$ a Boolean algebra on M and $\mathscr{F}$ a filter in $\mathscr{S}$. Then there is an ultrafilter $\mathscr{U}$ in $\mathscr{S}$ such that $\mathscr{F} \subseteq \mathscr{U}$.*

Proof. By 5.1.15, it suffices to show that the set $\{\mathscr{F}' \mid \mathscr{F}' \text{ Filter in } \mathscr{S}, \mathscr{F} \subseteq \mathscr{F}'\}$ partially ordered by inclusion has a maximal element. This follows from Zorn's lemma since the union of a nonempty chain of filters in $\mathscr{S}$ is again a filter in $\mathscr{S}$. □

Theorem 5.1.17. *A topological space M is quasicompact if and only if each ultrafilter on the set M converges in M.*

Proof. Let M be a topological space. We show the equivalence of the following statements:

(a) M is not quasicompact.

(b) There is an ultrafilter on M that does not converge.

(a) $\Longrightarrow$ (b) Suppose that (a) holds. Then for each $x \in M$, there is obviously an open set $A_x \subseteq M$ with $x \in A_x$ in such a way that $\bigcup_{x \in M} A_x = M$ and $A_{x_1} \cup \ldots \cup A_{x_n} \neq M$ for all $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in M$. Then

$$\mathscr{F} := \{U \in \mathscr{P}(M) \mid \exists n \in \mathbb{N} : \exists x_1, \ldots, x_n \in M : \left(\complement A_{x_1}\right) \cap \ldots \cap \left(\complement A_{x_n}\right) \subseteq U\}$$

is a filter on M that can be extended by 5.1.16 to an ultrafilter $\mathscr{U}$ on M. Let $x \in M$. We show that $\mathscr{U}$ does not converge to x. If we had $\mathscr{U} \to x$, then $A_x \in \mathscr{U}$ in contradiction to $\complement A_x \in \mathscr{U}$.

(b) $\Longrightarrow$ (a) Suppose that (b) holds. Choose an ultrafilter $\mathscr{U}$ on M that does not converge. Then for every $x \in M$ there is an $U_x \in \mathscr{U}_x$ such that $U_x \notin \mathscr{U}$. WLOG U_x is open for every $x \in M$. Of course $M = \bigcup_{x \in M} U_x$. If $n \in \mathbb{N}$ and $x_1, \ldots, x_n \in M$, then $\complement(U_{x_1} \cup \ldots \cup U_{x_n}) = \left(\complement U_{x_1}\right) \cap \ldots \cap \left(\complement U_{x_n}\right) \in \mathscr{U}$ and thus $\emptyset \neq \complement(U_{x_1} \cup \ldots \cup U_{x_n})$, i.e., $M \neq U_{x_1} \cup \ldots \cup U_{x_n}$. □

Theorem 5.1.18 (Tikhonov)**.** *Products of quasicompact topological spaces are quasicompact*

Proof. Let $(N_i)_{i\in I}$ be a family of quasicompact topological spaces and $M := \prod_{i\in I} N_i$ the product space [→ 5.1.5(b)]. Consider for each $i \in I$ the canonical projection $\pi_i\colon M \to N_i$. According to 5.1.17 it suffices to show that every ultrafilter on M converges. For this purpose, let $\mathscr{U}$ be an ultrafilter on M. By 5.1.12, the image filters $\pi_i(\mathscr{U})$ $(i \in I)$ are again ultrafilters and therefore converge. Accordingly, we choose $(x_i)_{i\in I}$ satisfying $\pi_i(\mathscr{U}) \to x_i$ for each $i \in I$. From 5.1.13, we now get $\mathscr{U} \to (x_i)_{i\in I}$. □

Corollary 5.1.19. *Products of compact spaces are compact.*

Remark 5.1.20. Let M be a topological space. $\left\{\begin{array}{c}\text{In 5.1.17, we have shown}\\ \text{Using 5.1.16, one shows as an exercise}\end{array}\right\}$ that M is $\left\{\begin{array}{c}\text{quasicompact}\\ \text{a Hausdorff space}\end{array}\right\}$ if and only if every ultrafilter on M converges to $\left\{\begin{array}{c}\text{at least}\\ \text{at most}\end{array}\right\}$ one point of M. Therefrom, M is compact if and only if each ultrafilter on M converges to exactly one point of M.

Reminder 5.1.21. Let M be a topological space and $A \subseteq M$. Then A is called *closed* in M if $\complement A$ is open in M. We call A $\left\{\begin{array}{c}\text{quasicompact}\\ \text{compact}\end{array}\right\}$ if A furnished with the subspace topology [→ 5.1.5(a)] is a $\left\{\begin{array}{c}\text{quasicompact}\\ \text{compact}\end{array}\right\}$ topological space. Consequently, A is quasicompact if and only if each open cover of A in M possesses a finite subcover, i.e.,

$$\forall \mathscr{A} \subseteq \mathscr{O} : \left(A \subseteq \bigcup \mathscr{A} \implies \exists \mathscr{B} \subseteq \mathscr{A} : \left(\#\mathscr{B} < \infty \;\&\; A \subseteq \bigcup \mathscr{B}\right)\right).$$

It follows immediately that closed subsets of $\left\{\begin{array}{c}\text{quasicompact}\\ \text{compact}\end{array}\right\}$ topological spaces are again $\left\{\begin{array}{c}\text{quasicompact}\\ \text{compact}\end{array}\right\}$.

5.2 Topologies on the real spectrum

Definition 5.2.1. Let A be a commutative ring. We call the topology generated by

$$\{\{P \in \operatorname{sper} A \mid \widehat{a}(P) > 0\} \mid a \in A\}$$

on $\operatorname{sper} A$ the *spectral topology* (or *Harrison-topology*) on $\operatorname{sper} A$. Moreover, we call the topology generated by $\mathscr{C}_A$ [→ 3.6.1] or, equivalently [→ 3.6.2], by

$$\{\{P \in \operatorname{sper} A \mid \widehat{a}(P) = 0\} \mid a \in A\} \cup \{\{P \in \operatorname{sper} A \mid \widehat{a}(P) > 0\} \mid a \in A\},$$

the *constructible topology* on $\operatorname{sper} A$. Unless otherwise indicated, we endow $\operatorname{sper} A$ always with the spectral topology. It is coarser than the constructible topology.

Reminder 5.2.2. Let M and N be topological spaces. A bijection $f\colon M \to N$ is called a *homeomorphism* if both f and f^{-1} are continuous. One calls M and N *homeomorphic* if there exists a homeomorphism from M to N.

Theorem 5.2.3. *Let A be a commutative ring. Then* sper A *is compact with respect to the constructible topology.*

Proof. We endow the two-element set $\{0,1\}$ with the discrete topology [→ 5.1.2(d)]. Then $\{0,1\}$ is compact and so is $\{0,1\}^A = \prod_{i\in A}\{0,1\}$ with respect to the product topology by Tikhonov's Theorem 5.1.18. For every $B \subseteq A$, we denote by

$$1_B\colon A \to \{0,1\},\ a \mapsto \begin{cases} 0 & \text{if } a \notin B \\ 1 & \text{if } a \in B \end{cases}$$

the corresponding characteristic function. Consider $S := \{1_P \mid P \in \operatorname{sper} A\} \subseteq \{0,1\}^A$ endowed with the subspace topology of the product topology. Obviously,

$$\operatorname{sper} A \to S,\ P \mapsto 1_P$$

is a homeomorphism. Since $\{0,1\}^A$ is compact by 5.1.19, it suffices to show that S is closed in $\{0,1\}^A$ since then S and consequently sper A is compact [→ 5.1.21]. Encoding 3.1.7 in characteristic functions, we obtain

$$\begin{aligned} S = &\bigcap_{a,b\in A} \Big\{\chi \in \{0,1\}^A \mid \chi(a) = 0 \text{ or } \chi(b) = 0 \text{ or } \chi(a+b) = 1\Big\} \cap \\ &\bigcap_{a,b\in A} \Big\{\chi \in \{0,1\}^A \mid \chi(a) = 0 \text{ or } \chi(b) = 0 \text{ or } \chi(ab) = 1\Big\} \cap \\ &\bigcap_{a\in A} \Big\{\chi \in \{0,1\}^A \mid \chi(a) = 1 \text{ or } \chi(-a) = 1\Big\} \cap \\ &\Big\{\chi \in \{0,1\}^A \mid \chi(-1) = 0\Big\} \cap \\ &\bigcap_{a\in A} \Big\{\chi \in \{0,1\}^A \mid \chi(ab) = 0 \text{ or } \chi(a) = 1 \text{ or } \chi(-b) = 1\Big\}. \end{aligned}$$

Being thus an intersection of closed sets, S is itself closed. □

Corollary 5.2.4. *Let A be a commutative ring. Then* sper A *is quasicompact.*

Proof. Every open cover of sper A is in particular an open cover with respect to the finer constructible topology. By 5.2.3, it possesses therefore a finite subcover. □

Reminder 5.2.5. Let M be a topological space and $A \subseteq M$. $\left\{\begin{matrix}\text{The } \textit{interior } A^\circ \\ \text{The } \textit{closure } \overline{A}\end{matrix}\right\}$ of A (in M) is the $\left\{\begin{matrix}\text{union} \\ \text{intersection}\end{matrix}\right\}$ over all $\left\{\begin{matrix}\text{open subsets} \\ \text{closed supersets}\end{matrix}\right\}$ of A in M, i.e., the $\left\{\begin{matrix}\text{largest open subset} \\ \text{smallest closed superset}\end{matrix}\right\}$ of A in M. One shows immediately

$$\begin{aligned} A^\circ &= \{x \in M \mid \exists U \in \mathscr{U}_x : U \subseteq A\} \quad \text{and} \\ \overline{A} &= \{x \in M \mid \forall U \in \mathscr{U}_x : U \cap A \neq \emptyset\}. \end{aligned}$$

Therefore one calls the elements of $\left\{\begin{matrix} A^\circ \\ \overline{A} \end{matrix}\right\}$ also $\left\{\begin{matrix} \textit{interior} \\ \textit{adherent} \end{matrix}\right\}$ *points* of A. One says that A is *dense* in M if $\overline{A} = M$ or, equivalently, if every nonempty open subset of M contains an element of A.

Remark 5.2.6. Let A be a commutative ring and let $P, Q \in \operatorname{sper} A$. Then

$$\begin{aligned} P \subseteq Q &\iff \forall a \in A : (\widehat{a}(P) \geq 0 \implies \widehat{a}(Q) \geq 0) \\ &\iff \forall a \in A : (\widehat{a}(Q) < 0 \implies \widehat{a}(P) < 0) \\ &\iff \forall a \in A : (\widehat{-a}(Q) < 0 \implies \widehat{-a}(P) < 0) \\ &\iff \forall a \in A : (\widehat{a}(Q) > 0 \implies \widehat{a}(P) > 0) \\ &\iff \forall U \in \mathscr{U}_Q : P \in U \\ &\iff \forall U \in \mathscr{U}_Q : U \cap \{P\} \neq \emptyset \\ &\iff Q \in \overline{\{P\}}. \end{aligned}$$

Thus if there are $P, Q \in \operatorname{sper} A$ with $P \subset Q$, then $\operatorname{sper} A$ ist not a Hausdorff space. For example, $\operatorname{sper} \mathbb{R}[X]$ is not a Hausdorff space [→ 3.6.5].

Remark 5.2.7. Suppose A and B are commutative rings and $\varphi\colon A \to B$ is a ring homomorphism. Then

$$\operatorname{sper} \varphi\colon \operatorname{sper} B \to \operatorname{sper} A,\ Q \mapsto \varphi^{-1}(Q)$$

is continuous with respect to the spectral topologies on both sides as well as with respect to the constructible topologies on both sides because for $a \in A$, we have

$$\begin{aligned} (\operatorname{sper} \varphi)^{-1}(\{P \in \operatorname{sper} A \mid \widehat{a}(P) > 0\}) &= \{Q \in \operatorname{sper} B \mid \widehat{a}((\operatorname{sper} \varphi)(Q)) > 0\} \\ &= \{Q \in \operatorname{sper} B \mid \widehat{a}(\varphi^{-1}(Q)) > 0\} \\ &= \{Q \in \operatorname{sper} B \mid a \in \varphi^{-1}(Q) \setminus -\varphi^{-1}(Q)\} \\ &= \{Q \in \operatorname{sper} B \mid a \in \varphi^{-1}(Q \setminus -Q)\} \\ &= \{Q \in \operatorname{sper} B \mid \varphi(a) \in Q \setminus -Q\} \\ &= \{Q \in \operatorname{sper} B \mid \widehat{\varphi(a)}(Q) > 0\} \end{aligned}$$

and analogously

$$(\operatorname{sper} \varphi)^{-1}(\{P \in \operatorname{sper} A \mid \widehat{a}(P) \geq 0\}) = \{Q \in \operatorname{sper} B \mid \widehat{\varphi(a)}(Q) \geq 0\}.$$

Remark 5.2.8. Let (A, T) be a preordered ring [→ 3.4.3]. Then

$$\operatorname{sper}(A, T) = \bigcap_{t \in T} \left\{P \in \operatorname{sper} A \mid \widehat{t}(P) \geq 0\right\},$$

as an intersection of closed sets, is again closed in $\operatorname{sper} A$, namely with respect to the spectral but also with respect to the constructible topology on $\operatorname{sper} A$. By 5.1.21, $\operatorname{sper}(A, T)$ is thus quasicompact with respect to the spectral and compact with respect to the constructible topology.

5.3 The real spectrum of polynomial rings

As in §3.6, we fix in this section an ordered field $(K, \leq)$, we denote by $R := \overline{(K, \leq)}$ its real closure, we let $n \in \mathbb{N}_0$, $A := K[\underline{X}] = K[X_1, \ldots, X_n]$ and $T := \sum K_{\geq 0} A^2$. Moreover, we denote by $\mathscr{S} := \mathscr{S}_{n,R}$ the Boolean algebra of all K-semialgebraic subsets of R^n [→ 1.8.3, 1.9.3] and by $\mathscr{C} := \mathscr{C}_{(A,T)}$ the Boolean algebra of all constructible subsets of $\operatorname{sper}(A, T)$ [→ 3.6.1]. Consider again the isomorphisms of Boolean algebras

$$\operatorname{Slim}\colon \mathscr{C} \to \mathscr{S}, C \mapsto \{x \in R^n \mid P_x \in C\}$$

and $\operatorname{Fatten} := \operatorname{Slim}^{-1}$ [→ 3.6.4].

Theorem 5.3.1. *Let $S \in \mathscr{S}$. Then $\operatorname{Fatten}(S)$ is the closure of $\{P_x \mid x \in S\}$ in $\operatorname{sper}(A, T)$ (or equivalently in $\operatorname{sper} A$ [→ 5.2.8]) with respect to the constructible topology.*

Proof. For the duration of this proof, we endow $\operatorname{sper}(A, T)$ with the constructible topology. Since $\complement \operatorname{Fatten}(S) \in \mathscr{C}$ is open, $\operatorname{Fatten}(S)$ is closed. Because of

$$S = \operatorname{Slim}(\operatorname{Fatten}(S)) \overset{3.6.4}{=} \{x \in R^n \mid P_x \in \operatorname{Fatten}(S)\},$$

we have $\{P_x \mid x \in S\} \subseteq \operatorname{Fatten}(S)$ and thus $\overline{\{P_x \mid x \in S\}} \subseteq \operatorname{Fatten}(S)$. In order to show $\operatorname{Fatten}(S) \subseteq \overline{\{P_x \mid x \in S\}}$, we let $P \in \operatorname{Fatten}(S)$ and $U \in \mathscr{U}_P$. To show: $U \cap \{P_x \mid x \in S\} \neq \emptyset$. WLOG U is open. WLOG $U \subseteq \operatorname{Fatten}(S)$ (because $\operatorname{Fatten}(S)$ is open and $P \in \operatorname{Fatten}(S)$, one can otherwise replace U by $U \cap \operatorname{Fatten}(S) \in \mathscr{U}_P$). WLOG $U = \{Q \in \operatorname{sper}(A, T) \mid \widehat{f}(Q) = 0, \widehat{g_1}(Q) > 0, \ldots, \widehat{g_n}(Q) > 0\}$ for certain $f, g_1, \ldots, g_n \in A$. Since Slim is an isomorphism of Boolean algebras by 3.6.4, it follows from $\emptyset \neq U \subseteq \operatorname{Fatten}(S)$ that $\emptyset \neq \operatorname{Slim}(U) \subseteq \operatorname{Slim}(\operatorname{Fatten}(S)) = S$. Taking into account that $\operatorname{Slim}(U) = \{x \in R^n \mid f(x) = 0, g_1(x) > 0, \ldots, g_n(x) > 0\}$, there is thus an $x \in S$ satisfying $f(x) = 0, g_1(x) > 0, \ldots, g_n(x) > 0$. This translates into $\widehat{f}(P_x) = 0, \widehat{g_1}(P_x) > 0, \ldots, \widehat{g_m}(P_x) > 0$ and thus into $P_x \in U$. □

Corollary 5.3.2. *$\{P_x \mid x \in R^n\}$ lies dense in $\operatorname{sper}(A, T)$ with respect to the constructible topology and thus also with respect to the spectral topology.*

Lemma 5.3.3. Let $\mathscr{F}$ be $\left\{\begin{smallmatrix}\text{a filter}\\ \text{an ultrafilter}\end{smallmatrix}\right\}$ in $\mathscr{S}$. Then $\{\operatorname{Fatten}(S) \mid S \in \mathscr{F}\}$ is $\left\{\begin{smallmatrix}\text{a filter}\\ \text{an ultrafilter}\end{smallmatrix}\right\}$ in $\mathscr{C}$ and $\bigcap\{\operatorname{Fatten}(S) \mid S \in \mathscr{F}\}$ is $\left\{\begin{smallmatrix}\text{nonempty}\\ \text{a singleton}\end{smallmatrix}\right\}$.

Proof. The first part follows immediately from the fact that Fatten is according to 3.6.4 an isomorphism of Boolean algebras combined with the definition of $\left\{\begin{smallmatrix}\text{a filter}\\ \text{an ultrafilter}\end{smallmatrix}\right\}$ 5.1.7. Since $\operatorname{Fatten}(S)$ is for each $S \in \mathscr{F}$ closed with respect to the constructible topology, it would follow from $\bigcap\{\operatorname{Fatten}(S) \mid S \in \mathscr{F}\} = \emptyset$ together with the compactness of $\operatorname{sper}(A, T)$ with respect to the constructible topology [→ 5.2.8] that there would be $n \in \mathbb{N}$ and $S_1, \ldots, S_n \in \mathscr{F}$ such that $\operatorname{Fatten}(S_1) \cap \ldots \cap \operatorname{Fatten}(S_n) = \emptyset$, which

would imply $\mathrm{Fatten}(S_1 \cap \ldots \cap S_n) = \emptyset$ and thus $\emptyset = S_1 \cap \ldots \cap S_n \in \mathscr{F}$ $\lightning$. Hence $\bigcap\{\mathrm{Fatten}(S) \mid S \in \mathscr{F}\} \neq \emptyset$. Finally, let $\mathscr{F}$ and thus $\{\mathrm{Fatten}(S) \mid S \in \mathscr{F}\}$ be an ultrafilter and fix $P, Q \in \bigcap\{\mathrm{Fatten}(S) \mid S \in \mathscr{F}\}$. Assume $P \neq Q$. Since $\mathrm{sper}(A,T)$ is a Hausdorff space with respect to the constructible topology, there is some $C \in \mathscr{C}$ such that $P \in C$ but $Q \notin C$. Since $\{\mathrm{Fatten}(S) \mid S \in \mathscr{F}\}$ is an ultrafilter in $\mathscr{C}$, we obtain $C = \mathrm{Fatten}(S)$ or $\complement C = \mathrm{Fatten}(S)$ for some $S \in \mathscr{F}$. In the first case, it follows that $Q \notin \mathrm{Fatten}(S)$ $\lightning$, in the second that $P \notin \mathrm{Fatten}(S)$ $\lightning$. □

Lemma 5.3.4. Let $\mathscr{U}$ be an ultrafilter in $\mathscr{S}$. Then

$$P_{\mathscr{U}} := \{f \in A \mid \{x \in R^n \mid f(x) \geq 0\} \in \mathscr{U}\} \in \mathrm{sper}(A,T)$$

and $\bigcap\{\mathrm{Fatten}(S) \mid S \in \mathscr{U}\} = \{P_{\mathscr{U}}\}$.

Proof. By Lemma 5.3.3, there is some $Q \in \mathrm{sper}(A,T)$ satisfying

$$\bigcap\{\mathrm{Fatten}(S) \mid S \in \mathscr{U}\} = \{Q\}.$$

We show $P_{\mathscr{U}} = Q$. If $f \in P_{\mathscr{U}}$, then $Q \in \mathrm{Fatten}(\{x \in R^n \mid f(x) \geq 0\})$, i.e., $\widehat{f}(Q) \geq 0$ and hence $f \in Q$. If on the other hand $f \in A \setminus P_{\mathscr{U}}$, then $\{x \in R^n \mid f(x) < 0\} \in \mathscr{U}$ (since $\mathscr{U}$ is an ultrafilter) and thus $Q \in \mathrm{Fatten}(\{x \in R^n \mid f(x) < 0\})$, i.e., $\widehat{f}(Q) < 0$ and hence $f \notin Q$. □

Lemma 5.3.5. Let $P \in \mathrm{sper}(A,T)$. Then

$$\begin{aligned}\mathscr{U}_P := \{S \in \mathscr{S} \mid\ & \exists f \in \mathrm{supp}\, P : \exists m \in \mathbb{N} : \exists g_1, \ldots, g_m \in P \setminus -P : \\ & \{x \in R^n \mid f(x) = 0, g_1(x) > 0, \ldots, g_m(x) > 0\} \subseteq S\}\end{aligned}$$

is an ultrafilter in $\mathscr{S}$ and we have $\{S \in \mathscr{S} \mid P \in \mathrm{Fatten}(S)\} = \mathscr{U}_P$.

Proof. Since $\{C \in \mathscr{C} \mid P \in C\}$ is an ultrafilter in $\mathscr{C}$ and $\mathrm{Slim}\colon \mathscr{C} \to \mathscr{S}$ is an isomorphism of Boolean algebras, $\{S \in \mathscr{S} \mid P \in \mathrm{Fatten}(S)\}$ is an ultrafilter in $\mathscr{S}$. From the description of K-semialgebraic subsets of R^n implied by Theorem 1.8.6, one gets that this ultrafilter equals

$$\begin{aligned}\{S \in \mathscr{S} \mid & \exists S' \subseteq S : \exists f, g_1, \ldots, g_m \in A : \\ & S' = \{x \in R^n \mid f(x) = 0, g_1(x) > 0, \ldots, g_m(x) > 0\} \,\&\, P \in \mathrm{Fatten}(S')\} = \mathscr{U}_P\end{aligned}$$

since Fatten is an isomorphism of Boolean algebras. □

Theorem 5.3.6 (Bröcker's ultrafilter theorem [Brö])**.** *The correspondence*

$$\begin{array}{rcll} \mathscr{U} & \mapsto & P_{\mathscr{U}} & [\to 5.3.4] \\ \mathscr{U}_P & \mapsto & P & [\to 5.3.5] \end{array}$$

defines a bijection between the set of ultrafilters in $\mathscr{S}$ *and* $\mathrm{sper}(A,T)$.

Proof. To show: (a) If $\mathscr{U}$ is an ultrafilter in $\mathscr{S}$, then $\mathscr{U} = \mathscr{U}_{P_{\mathscr{U}}}$.
(b) If $P \in \operatorname{sper}(A,T)$, then $P = P_{\mathscr{U}_P}$.

In order to show (a), we let $\mathscr{U}$ be an ultrafilter in $\mathscr{S}$. By 5.3.5, we have to show that $\{S \in \mathscr{S} \mid P_{\mathscr{U}} \in \operatorname{Fatten}(S)\} = \mathscr{U}$. Since Fatten is an isomorphism of Boolean algebras by 3.6.4, $\{S \in \mathscr{S} \mid P_{\mathscr{U}} \in \operatorname{Fatten}(S)\}$ is a filter in $\mathscr{S}$. Since $\mathscr{U}$ is a maximal filter in $\mathscr{S}$ [$\to$ 5.1.15], it suffices to show that $\mathscr{U} \subseteq \{S \in \mathscr{S} \mid P_{\mathscr{U}} \in \operatorname{Fatten}(S)\}$. To this end, let $S \in \mathscr{U}$. Then $\{P_{\mathscr{U}}\} \subseteq \operatorname{Fatten}(S)$ by 5.3.4 and thus $P_{\mathscr{U}} \in \operatorname{Fatten}(S)$.

For (b), we let $P \in \operatorname{sper}(A,T)$. By 5.3.4, $\bigcap\{\operatorname{Fatten}(S) \mid S \in \mathscr{U}_P\}$ consists of exactly one element, namely $P_{\mathscr{U}_P}$. Therefore it is enough to show $P \in \bigcap\{\operatorname{Fatten}(S) \mid S \in \mathscr{U}_P\}$. Thus fix $S \in \mathscr{U}_P$. By 5.3.5, we then obtain $P \in \operatorname{Fatten}(S)$. □

Proposition 5.3.7. *Every semialgebraic subset of R^n [$\to$ 1.8.3] is even K-semialgebraic.*

Proof. To begin with, we show that all one-element subsets of R are K-semialgebraic. For this, let $a \in R$. To show: $\{a\}$ is K-semialgebraic. Since $R|K$ is algebraic, there is $f \in K[\underline{X}] \setminus \{0\}$ with $f(a) = 0$. Set $k := \#\{x \in R \mid f(x) = 0\}$ and choose $j \in \{1,\dots,k\}$ such that a is the j-th root of f when the roots of f in R are arranged in increasing order with respect to the order $\leq_R$ of R. By applying the real quantifier elimination 1.8.17 k times, we obtain that

$$\{a\} = \{y \in R \mid \exists x_1,\dots,x_k \in R : (x_1 <_R \dots <_R x_k \,\&\, f(x_1) = \dots = f(x_k) = 0 \,\&\, x_j = y)\}$$

is K-semialgebraic. Now consider an arbitrary $p \in R[\underline{X}]$. It suffices to show that $\{x \in R^n \mid p(x) \geq 0\}$ is K-semialgebraic. Write $p = \sum_{\substack{\alpha \in \mathbb{N}^n \\ |\alpha| \leq d}} a_\alpha \underline{X}^\alpha$ [$\to$ 2.4.6] with $d := \deg p$ and $a_\alpha \in R$. Since all $\{a_\alpha\}$ are K-semialgebraic by what has already been shown, real quantifier elimination yields that

$$\{x \in R^n \mid p(x) \geq 0\} = \left\{ x \in R^n \mid \exists \text{ family } (y_\alpha)_{|\alpha| \leq d} \text{ in } R : \left(\mathop{\&}_{|\alpha| \leq d} y_\alpha \in \{a_\alpha\} \,\&\, \sum_{|\alpha| \leq d} y_\alpha x_1^{\alpha_1} \cdots x_n^{\alpha_n} \geq 0 \right) \right\}$$

is K-semialgebraic. □

Theorem 5.3.8. *$\operatorname{sper} R[\underline{X}] \to \operatorname{sper}(A,T)$, $P \mapsto P \cap A$ is bijective.*

Proof. Because of 5.3.7, we obtain from applying the ultrafilter theorem of Bröcker twice (once in the special case $K = R$) that

$$\begin{array}{c} \operatorname{sper} R[\underline{X}] \to \operatorname{sper}(A,T) \\ \{f \in R[\underline{X}] \mid \{x \in R^n \mid f(x) \geq 0\} \in \mathscr{U}\} \mapsto \{f \in A \mid \{x \in R^n \mid f(x) \geq 0\} \in \mathscr{U}\} \\ (\mathscr{U} \text{ an ultrafilter in } \mathscr{S}) \end{array}$$

is a bijection. □

Corollary 5.3.9. $\operatorname{sper} R(\underline{X}) \to \operatorname{sper}(K(\underline{X}), \sum K_{\geq 0} K(\underline{X})^2)$, $P \mapsto P \cap K(\underline{X})$ *is bijective.*

Proof. In the commutative diagram

$$\begin{array}{ccc} \operatorname{sper} R(\underline{X}) & \xrightarrow{P \mapsto P \cap K(\underline{X})} & \operatorname{sper}(K(\underline{X}), \sum K_{\geq 0} K(\underline{X})^2) \\ \Big\downarrow{\scriptstyle P \mapsto P \cap R[\underline{X}]} & & \Big\downarrow{\scriptstyle P \mapsto P \cap A} \\ \{P \in \operatorname{sper} R[\underline{X}] \mid \operatorname{supp} P = (0)\} & \xrightarrow[P \mapsto P \cap A]{} & \{P \in \operatorname{sper} A \mid \operatorname{supp} P = (0)\} \end{array}$$

both vertical arrows represent bijections by Proposition 3.3.4. It therefore suffices to show that the lower horizontal arrow represents a bijection. Because of the bijection from 5.3.8, it therefore suffices to show that every $P \in \operatorname{sper} R[\underline{X}]$ with $\operatorname{supp} P \neq (0)$ satisfies even $\operatorname{supp}(P \cap A) \neq (0)$. Thus fix $P \in \operatorname{sper} R[\underline{X}]$ and $f \in \mathfrak{p} := \operatorname{supp} P$ with $f \neq 0$. Since K has characteristic 0, there exists an extension field L of K containing all coefficients of f such that $L|K$ is a finite Galois extension. By extending the action of the Galois group $\operatorname{Aut}(L|K)$ from L to $L[\underline{X}]$, we obtain $h := \prod_{g \in \operatorname{Aut}(L|K)} gf \in A \setminus \{0\}$. Obviously, $h \in \mathfrak{p} \cap A = \operatorname{supp}(P \cap A)$. □

Theorem 5.3.10. *Let $(L, \leq')$ be an ordered extension field of $(K, \leq)$. Then*

$$\operatorname{sper}\left(L[\underline{X}], \sum L_{\geq' 0} L[\underline{X}]^2\right) \to \operatorname{sper}(A, T),\ P \mapsto P \cap A$$

is surjective.

Proof. Let $\mathscr{S}''$ denote the Boolean algebra of all L-semialgebraic subsets of R'^n where $R' := \overline{(L, L_{\geq' 0})}$. The Boolean algebra $\mathscr{S}' \subseteq \mathscr{S}''$ of all K-semialgebraic subsets of R'^n is isomorphic to $\mathscr{S}$ in virtue of the $\operatorname{Transfer}_{R,R'} \colon \mathscr{S} \to \mathscr{S}'$ [→ 1.9.5]. Now let $Q \in \operatorname{sper}(A, T)$ be given. We show that there is $P \in \operatorname{sper}(L[\underline{X}], \sum L_{\geq' 0} L[\underline{X}]^2)$ with $Q = P \cap A$. By 5.3.5, $\mathscr{U}_Q$ is an ultrafilter in $\mathscr{S}$. Since $\mathscr{U}_Q$ is a filter in $\mathscr{S}$,

$$\mathscr{F} := \{S'' \in \mathscr{S}'' \mid \exists S \in \mathscr{U}_Q : \operatorname{Transfer}_{R,R'}(S) \subseteq S''\}$$

is a filter in $\mathscr{S}''$. Choose by 5.1.16 an ultrafilter $\mathscr{U}$ in $\mathscr{S}''$ such that $\mathscr{F} \subseteq \mathscr{U}$. By Bröcker's ultrafilter theorem 5.3.6, there is $P \in \operatorname{sper}(L[\underline{X}], \sum L_{\geq' 0} L[\underline{X}]^2)$ such that $\mathscr{U} = \mathscr{U}_P$. We have

$$\begin{aligned} Q \overset{5.3.6}{=} P_{\mathscr{U}_Q} &= \{f \in A \mid \{x \in R^n \mid f(x) \geq 0\} \in \mathscr{U}_Q\} \\ &= \{f \in A \mid \operatorname{Transfer}_{R,R'}(\{x \in R^n \mid f(x) \geq 0\}) \in \{\operatorname{Transfer}_{R,R'}(S) \mid S \in \mathscr{U}_Q\}\} \\ &\overset{!}{=} \{f \in A \mid \{x \in R'^n \mid f(x) \geq'' 0\} \in \mathscr{U}\} = P_{\mathscr{U}} \cap A = P_{\mathscr{U}_P} \cap A \overset{5.3.6}{=} P \cap A \end{aligned}$$

where $\leq''$ denotes the unique order on R' and the equality flagged with an exclamation mark follows from the claim

$$\mathscr{U} \cap \mathscr{S}' = \{\operatorname{Transfer}_{R,R'}(S) \mid S \in \mathscr{U}_Q\}.$$

The inclusion "$\supseteq$" in this claim is trivial. The other inclusion "$\subseteq$" follows the fact that $\{\mathrm{Transfer}_{R,R'}(S) \mid S \in \mathscr{U}_Q\}$ is an ultrafilter and thus a maximal filter in $\mathscr{S}'$ and that $\mathscr{U} \cap \mathscr{S}'$ is a filter in $\mathscr{S}'$. □

5.4 The finiteness theorem for semialgebraic classes

In this section, we fix a real closed field R_0 (in the applications, one mostly has $R_0 = \mathbb{R}$ or $R_0 = \mathbb{R}_{\mathrm{alg}}$ [→ 1.7.12]). Moreover, we let $\mathscr{R}$ denote the class of all real closed extension fields of R_0 [→ 1.8.4(b)] (that is the class of all real closed fields in case $R_0 = \mathbb{R}_{\mathrm{alg}}$). Whoever gets vertiginous from this [→ 1.8.4(c)] can take for $\mathscr{R}$ a set of real closed extension fields of R_0 that is sufficiently large to contain all representation fields R_P of prime cones $P \in \operatorname{sper} R_0[\underline{X}]$ [→ 3.1.15] (which we perceive as an extension fields of R_0 in virtue of the representation ϱ_P of P, confer the discussion before 3.6.3).

Theorem 5.4.1 (Finiteness theorem for semialgebraic classes)**.** *Let $n \in \mathbb{N}_0$ and $\mathscr{E}$ a set of n-ary R_0-semialgebraic classes. Then the following are equivalent:*

(a) $\bigcup \mathscr{E} = \mathscr{R}_n$

(b) $\exists k \in \mathbb{N} : \exists S_1, \ldots, S_k \in \mathscr{E} : S_1 \cup \ldots \cup S_k = \mathscr{R}_n$.

(c) $\exists k \in \mathbb{N} : \exists S_1, \ldots, S_k \in \mathscr{E} : \mathrm{Set}_{R_0}(S_1) \cup \ldots \cup \mathrm{Set}_{R_0}(S_k) = R_0^n$ [→ 1.9.3].

Proof. (b) $\Longleftrightarrow$ (c) is clear because the setification $\mathrm{Set}_{R_0} : \mathscr{S}_n \to \mathscr{S}_{n,R_0}$ [→ 1.9.3] and thus also the classification $\mathrm{Class}_{R_0} = \mathrm{Set}_{R_0}^{-1} : \mathscr{S}_{n,R_0} \to \mathscr{S}_n$ is an isomorphism of Boolean algebras [→ 1.9.4].

(b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (b) Suppose that (a) holds. In 3.6.4, we have shown that

$$\Phi : \mathscr{S}_n \to \mathscr{C}_{R_0[\underline{X}]},\ S \mapsto \{P \in \operatorname{sper} R_0[\underline{X}] \mid (R_P, (\varrho_P(X_1), \ldots, \varrho_P(X_n))) \in S\}$$

is an isomorphism of Boolean algebras. Moreover, we have

$$\bigcup \{\Phi(S) \mid S \in \mathscr{E}\} = \operatorname{sper} R_0[\underline{X}]$$

by the definition of Φ. From 5.2.3, we get the existence of $k \in \mathbb{N}$ and $S_1, \ldots, S_k \in \mathscr{E}$ satisfying $\Phi(S_1) \cup \ldots \cup \Phi(S_k) = \operatorname{sper} R_0[\underline{X}]$. Since Φ is an isomorphism, we deduce $S_1 \cup \ldots \cup S_k = \mathscr{R}_n$. □

Corollary 5.4.2. *Let $n \in \mathbb{N}_0$ and $\mathscr{E}$ a set of n-ary R_0-semialgebraic classes satisfying*

$$\forall S_1, S_2 \in \mathscr{E} : \exists S_3 \in \mathscr{E} : S_1 \cup S_2 \subseteq S_3.$$

Then the following are equivalent:

(a) $\bigcup \mathscr{E} = \mathscr{R}_n$

(b) $S = \mathscr{R}_n$ *for some* $S \in \mathscr{E}$

(c) $\mathrm{Set}_{R_0}(S) = R_0^n$ *for some* $S \in \mathscr{E}$

Remark 5.4.3. In practice, 5.4.2 is mostly applied in the following context: One has a certain true statement about real numbers (for example that $\mathbb{R}$ is Archimedean [→ 1.1.9(a)]). Now one is interested in one of the following questions:

(a) Does the statement hold for all real closed extension fields of $\mathbb{R}$? (In our example: Is every real closed field extension of $\mathbb{R}$ Archimedean?)

(b) Does the statement hold in a strengthened form (with certain quantitative additional information, so called "bounds") for every real closed extension of $\mathbb{R}$? (In our example: Is there an $N \in \mathbb{N}$ such that we have for all real closed field extensions R of $\mathbb{R}$ and all $a \in R$ that $|a| \leq N$?)

(c) Does the statement hold in the strengthened from (that is "with bounds") for the real numbers? (In our example: Is there some $N \in \mathbb{N}$ such that for all $a \in \mathbb{R}$ one has $|a| \leq N$?)

5.4.2 establishes under certain circumstances a connection between these three questions. For this aim, one tries to express the statement in such a way that for n numbers a certain "semialgebraic event" occurs where the event is the existence of a bound. The set of events is $\mathscr{E}$.

Example 5.4.4. For $n := 1$, $R_0 := \mathbb{R}$ and $\mathscr{E} := \{\{(R, a) \in \mathscr{R}_1 \mid -N \leq a \leq N\} \mid N \in \mathbb{N}\}$, 5.4.2 says that the following are equivalent:

(a) For every real closed extension field R of $\mathbb{R}$ and every $a \in R$, there is some $N \in \mathbb{N}$ with $|a| \leq N$, i.e., every real closed extension field R of $\mathbb{R}$ is Archimedean.

(b) There is some $N \in \mathbb{N}$ such that for every real closed extension field R of $\mathbb{R}$ and every $a \in R$ we have $|a| \leq N$.

(c) There is some $N \in \mathbb{N}$ such that for every $a \in \mathbb{R}$ we have $|a| \leq N$.

Since (c) obviously fails, we see that (a) also fails. Thus we see (once more) that there are non-Archimedean real closed (extension) fields (of $\mathbb{R}$).

Theorem 5.4.5 (Existence of degree bounds for Hilbert's 17th problem)**.** *For all* $n, d \in \mathbb{N}_0$, *there is some* $D \in \mathbb{N}$ *such that for every real closed field* R *and every* $f \in R[\underline{X}]_d$ [→ 1.5.1] *with* $f \geq 0$ *on* R^n, *there are* $p_1, \ldots, p_D \in R[\underline{X}]_D$ *and* $q \in R[\underline{X}] \setminus \{0\}$ *with* $f = \sum_{i=1}^{D} \left(\frac{p_i}{q}\right)^2$.

Proof. Let $n, d \in \mathbb{N}_0$. Set $N := \dim \mathbb{R}[\underline{X}]_d$ and write $\{\alpha \in \mathbb{N}_0^n \mid |\alpha| \leq N\} = \{\alpha_1, \ldots, \alpha_N\}$.

Set $R_0 := \mathbb{R}_{\text{alg}}$ and

$$S_D := \left\{ (R,(a_1,\dots,a_N)) \in \mathscr{R}_N \;\middle|\; \begin{array}{l} \left(\forall x \in R^n : \sum_{i=1}^N a_i x_1^{\alpha_{i1}} \cdots x_n^{\alpha_{in}} \geq 0\right) \implies \\ \text{There are families } (b_{i\alpha})_{\substack{1\leq i\leq D \\ |\alpha|\leq D}} \text{ and } (c_\alpha)_{|\alpha|\leq D} \neq 0 \text{ in } R \\ \text{such that} \\ \left(\sum_{i=1}^N a_i \underline{X}^{\alpha_i}\right)\left(\sum_{|\alpha|\leq D} c_\alpha \underline{X}^\alpha\right)^2 = \sum_{i=1}^D \left(\sum_{|\alpha|\leq D} b_{i\alpha}\underline{X}^\alpha\right)^2 \end{array} \right\}$$

for each $D \in \mathbb{N}$. Obviously, S_D is for each $D \in \mathbb{N}$ an R_0-semialgebraic class since the polynomial identity in the last part of its specification can for example be expressed by finitely many polynomial equations in the a_i, $b_{i\alpha}$ and c_α, the requirement on the existence of the two finite families and the quantification "$\forall x \in R^n$" is allowed because of the real quantifier elimination 1.8.17. Set $\mathscr{E} := \{S_D \mid D \in \mathbb{N}\}$ and observe that $\forall D_1, D_2 \in \mathbb{N} : \exists D_3 \in \mathbb{N} : S_{D_1} \cup S_{D_2} \subseteq S_{D_3}$ (take $D_3 := \max\{D_1, D_2\}$). By Artin's solution to Hilbert's 17th problem 2.5.2, we have $\bigcup \mathscr{E} = \mathscr{R}_N$. Now 5.4.2 yields $S_D = \mathscr{R}_N$ for some $D \in \mathbb{N}$. □

Remark 5.4.6. Recently, Lombardi, Perrucci and Roy [LPR] managed to prove that one can choose in 5.4.5

$$D := 2^{2^{2^{d^{4^n}}}}.$$

We will neither use nor prove this in this lecture.

Definition and Proposition 5.4.7. *Let $(K,\leq)$ be an ordered extension field of $\mathbb{R}$. Then*

$$\mathscr{O}_{(K,\leq)} := B_{(K,K_{\geq 0})} = \{a \in K \mid \exists N \in \mathbb{N} : |a| \leq N\}$$

is a subring of K [→ 4.3.1] with a single maximal ideal

$$\mathfrak{m}_{(K,\leq)} := \left\{ a \in K \mid \forall N \in \mathbb{N} : |a| \leq \frac{1}{N} \right\}$$

with group of units

$$\mathscr{O}^\times_{(K,\leq)} = \mathscr{O}_{(K,\leq)} \setminus \mathfrak{m}_{(K,\leq)} = \left\{ a \in K \mid \exists N \in \mathbb{N} : \frac{1}{N} \leq |a| \leq N \right\}.$$

We call the elements of $\left\{\begin{array}{c} \mathscr{O}_{(K,\leq)} \\ \mathfrak{m}_{(K,\leq)} \\ K \setminus \mathscr{O}_{(K,\leq)} \end{array}\right\}$ *the* $\left\{\begin{array}{c} \text{finite} \\ \text{infinitesimal} \\ \text{infinit} \end{array}\right\}$ *elements of $(K,\leq)$. For every $a \in \mathscr{O}_{(K,\leq)}$, there is exactly one* $\mathrm{st}(a) \in \mathbb{R}$, *called the* standard part *of a, such that*

$$a - \mathrm{st}(a) \in \mathfrak{m}_{(K,\leq)}.$$

The map $\mathscr{O}_{(K,\leq)} \to \mathbb{R}$, $a \mapsto \operatorname{st}(a)$ is a ring homomorphism with kernel $\mathfrak{m}_{(K,\leq)}$. If $a, b \in \mathscr{O}_{(K,\leq)}$ satisfy $\operatorname{st}(a) < \operatorname{st}(b)$, then $a < b$. The standard part $\operatorname{st}(p)$ of a polynomial $p \in \mathscr{O}_{(K,\leq)}[\underline{X}]$ arises by replacing each coefficient of p by its standard part. Also $\mathscr{O}_{(K,\leq)}[\underline{X}] \to \mathbb{R}[\underline{X}]$, $p \mapsto \operatorname{st}(p)$ is a ring homomorphism.

Proof. The existence of the standard part follows easily from the completeness of $\mathbb{R}$ [→ 1.1.16] and its uniqueness is trivial. The rest is also easy. We show exemplarily:

(a) $\operatorname{st}(ab) = (\operatorname{st}(a))(\operatorname{st}(b))$ for all $a, b \in \mathscr{O}_{(K,\leq)}$

(b) $\operatorname{st}(a) < \operatorname{st}(b) \implies a < b$ for all $a, b \in \mathscr{O}_{(K,\leq)}$

To show (a), let $a, b \in \mathscr{O}_{(K,\leq)}$. Because of $a - \operatorname{st}(a), b - \operatorname{st}(b) \in \mathfrak{m}_{(K,\leq)}$, we have

$$\begin{aligned} ab - (\operatorname{st}(a))(\operatorname{st}(b)) &= (ab - (\operatorname{st}(a))b) + ((\operatorname{st}(a))b - (\operatorname{st}(a))(\operatorname{st}(b))) \\ &= (a - \operatorname{st}(a))b + (\operatorname{st}(a))(b - \operatorname{st}(b)) \in \mathfrak{m}_{(K,\leq)} + \mathfrak{m}_{(K,\leq)} \subseteq \mathfrak{m}_{(K,\leq)} \end{aligned}$$

For (b), we fix again $a, b \in \mathscr{O}_{(K,\leq)}$ with $\operatorname{st}(a) < \operatorname{st}(b)$. Choose $N \in \mathbb{N}$ with

$$\operatorname{st}(b) - \operatorname{st}(a) > \frac{1}{N}.$$

Then $|a - \operatorname{st}(a)| \leq \frac{1}{2N}$ and $|b - \operatorname{st}(b)| \leq \frac{1}{2N}$ and thus

$$\begin{aligned} a = a - \operatorname{st}(a) + \operatorname{st}(a) &\leq |a - \operatorname{st}(a)| + \operatorname{st}(a) \leq \frac{1}{2N} + \operatorname{st}(a) - \operatorname{st}(b) + \operatorname{st}(b) \\ &< \frac{1}{2N} - \frac{1}{N} + \operatorname{st}(b) = -\frac{1}{2N} + \operatorname{st}(b) - b + b \\ &\leq -\frac{1}{2N} + |b - \operatorname{st}(b)| + b \leq -\frac{1}{2N} + \frac{1}{2N} + b = b \end{aligned}$$

□

Example 5.4.8 (Nonexistence of degree bounds for Schmüdgen's Positivstellensatz [→ 4.3.6])**.** For every $\varepsilon \in \mathbb{R}_{>0}$, we have $X + \varepsilon > 0$ on $[0, 1]$ so that Schmüdgen's Positivstellensatz 4.3.6 together with 2.1.1(b) yields $p_1, p_2, q_1, q_2 \in \mathbb{R}[X]$ such that

$$(*) \qquad X + \varepsilon = p_1^2 + p_2^2 + (q_1^2 + q_2^2)X^3(1 - X).$$

One can ask the question if there is in analogy to 5.4.5 a $D \in \mathbb{N}$ such that for all $\varepsilon \in \mathbb{R}_{>0}$ there are $p_1, p_2, q_1, q_2 \in \mathbb{R}[X]_D$ satisfying $(*)$. To this end, consider for each $D \in \mathbb{N}$

$$S_D := \left\{ (R, \varepsilon) \in \mathscr{R}_1 \;\middle|\; \begin{array}{l} \varepsilon > 0 \implies \exists b_0, \ldots, b_D, b_0', \ldots, b_D', c_0, \ldots, c_D, c_0', \ldots, c_D' \in R : \\ X + \varepsilon = \left(\sum_{i=0}^{D} b_i X^i\right)^2 + \left(\sum_{i=0}^{D} b_i' X^i\right)^2 + \\ \qquad \left(\left(\sum_{i=0}^{D} c_i X^i\right)^2 + \left(\sum_{i=0}^{D} c_i' X^i\right)^2\right) X^3(1 - X) \end{array} \right\}$$

As in the proof of 5.4.5, one shows that S_D is for each $D \in \mathbb{N}$ an $\mathbb{R}$-semialgebraic class. Set $\mathscr{E} := \{S_D \mid D \in \mathbb{N}\}$. We claim that the answer to the above question is no. Assume it would be yes. Then $\mathrm{Set}_{\mathbb{R}}(S_D) = \mathbb{R}$ for some $D \in \mathbb{N}$ and thus $\bigcup \mathscr{E} = \mathscr{R}_1$ by 5.4.2. Choose a non-Archimedean real closed extension field R of $\mathbb{R}$ and an $\varepsilon > 0$ which is infinitesimal in R. Then there are $p_1, p_2, q_1, q_2 \in R[X]$ satisfying $(*)$. It suffices to show that all coefficients of these four polynomials are finite in R [$\to$ 5.4.7] since then $X = \mathrm{st}(X+\varepsilon) = (\mathrm{st}(p_1))^2 + (\mathrm{st}(p_2))^2 + ((\mathrm{st}(q_1))^2 + (\mathrm{st}(q_2))^2)X^3(1-X)$ in contradiction to 4.3.7. It therefore suffices to show that the coefficient c of biggest absolute value among all coefficients of the four polynomials is finite. Assume it were infinite. Then $\frac{1}{c}$ would be infinitesimal and

$$\begin{aligned} 0 = \mathrm{st}\left(\frac{X+\varepsilon}{c^2}\right) &= \mathrm{st}\left(\left(\frac{p_1}{c}\right)^2 + \left(\frac{p_2}{c}\right)^2 + \left(\left(\frac{q_1}{c}\right)^2 + \left(\frac{q_2}{c}\right)^2\right) X^3(1-X)\right) \\ &= \Big(\underbrace{\mathrm{st}\left(\frac{p_1}{c}\right)}_{\widetilde{p}_1}\Big)^2 + \Big(\underbrace{\mathrm{st}\left(\frac{p_2}{c}\right)}_{\widetilde{p}_2}\Big)^2 + \Big(\Big(\underbrace{\mathrm{st}\left(\frac{q_1}{c}\right)}_{\widetilde{q}_1}\Big)^2 + \Big(\underbrace{\mathrm{st}\left(\frac{q_2}{c}\right)}_{\widetilde{q}_2}\Big)^2\Big) X^3(1-X). \end{aligned}$$

It follows that $\widetilde{p}_1 = \widetilde{p}_2 = \widetilde{q}_1 = \widetilde{q}_2 = 0$ on $(0,1)$ and thus $\widetilde{p}_1 = \widetilde{p}_2 = \widetilde{q}_1 = \widetilde{q}_2 = 0$, contradicting the choice of c $\lightning$.

Remark 5.4.9. Completely analogous to 5.4.5, one can prove the existence of degree bounds for the real Stellensätze 3.7.5, 3.7.6 and 3.7.7 *in the case* $K = R$.

§6 Semialgebraic geometry

Throughout this chapter, we let R be a real closed field and K a subfield of R. Moreover, $\mathscr{S}_n$ denotes for each $n \in \mathbb{N}_0$ the Boolean algebra of all K-semialgebraic subsets of R^n [→ 1.9.3, 1.8.3].

6.1 Semialgebraic sets and functions

Reminder 6.1.1. [→ 1.8.6, 1.8.4(a)] Every K-semialgebraic subset of R^n is of the form

$$\bigcup_{i=1}^{k} \{x \in R^n \mid f_i(x) = 0, g_{i1}(x) > 0, \ldots, g_{im}(x) > 0\}$$

for some $k, m \in \mathbb{N}_0$, $f_i, g_{ij} \in K[X_1, \ldots, X_n]$.

Reminder 6.1.2. [→ 1.8.17] For all $n \in \mathbb{N}_0$ and $S \in \mathscr{S}_{n+1}$,

$$\{x \in R^n \mid \exists y \in R : (x,y) \in S\}, \{x \in R^n \mid \forall y \in R : (x,y) \in S\} \in \mathscr{S}_n.$$

Definition 6.1.3. Let $m, n \in \mathbb{N}_0$ and $A \subseteq R^m$. A map $f\colon A \to R^n$ is called *K-semialgebraic* if its graph

$$\Gamma_f := \{(x,y) \in A \times R^n \mid y = f(x)\} \subseteq R^{m+n}$$

is K-semialgebraic. We say "semialgebraic" for "R-semialgebraic".

Remark 6.1.4. The domains of K-semialgebraic functions are K-semialgebraic. Indeed, if $A \subseteq R^m$ and $f\colon A \to R^n$ is K-semialgebraic, then by 6.1.2 also

$$\{x \in R^m \mid \exists y \in R^n : (x,y) \in \Gamma_f\} = A$$

is K-semialgebraic.

Definition 6.1.5. We equip R with the *order topology* which is generated [→ 5.1.2(b)] by the intervals $(a,b)_R$ with $a, b \in R$ [→ 1.4.15(b)]. Moreover, we endow R^n with the corresponding product topology [→ 5.1.5(b)] which is generated according to 5.1.4 by the sets $\prod_{i=1}^{n}(a_i, b_i)_R$ with $a_i, b_i \in R$.

Remark 6.1.6. For $R = \mathbb{R}$, the topology introduced in 5.1.4 on $R^n = \mathbb{R}^n$ is obviously the usual Euclidean topology on $\mathbb{R}^n$.

Exercise 6.1.7. Let $m, n \in \mathbb{N}_0$, $A \subseteq R^m$ and $f\colon A \to R^n$ a map. Then f is continuous [→ 5.1.3, 5.1.5(a)] if and only if

$$\forall x \in A : \forall \varepsilon \in R_{>0} : \exists \delta \in R_{>0} : \forall y \in A : (\|x-y\|_\infty < \delta \implies \|f(x)-f(y)\|_\infty < \varepsilon)$$

where

$$\|x\|_\infty := \begin{cases} 0 & \text{if } k = 0 \\ \max\{|x_1|, \ldots, |x_k|\} & \text{if } k > 0 \end{cases}$$

for $x \in R^k$.

Proposition 6.1.8. *The maps*

$$\begin{aligned} R^2 &\to R,\ (a,b) \mapsto a+b, \\ R^2 &\to R,\ (a,b) \mapsto ab, \\ R \setminus \{0\} &\to R,\ a \mapsto a^{-1}, \\ R &\to R,\ a \mapsto |a| \qquad [\to 1.1.8], \\ R_{\geq 0} &\to R,\ a \mapsto \sqrt{a} \qquad [\to 1.4.7] \end{aligned}$$

are $\mathbb{Q}$*-semialgebraic and continuous.*

Proof. It is clear that these maps are $\mathbb{Q}$-semialgebraic. Because of the real quantifier elimination 1.8.17, the class of all real closed fields for which the claim holds is semialgebraic [→ 1.8.3]. Since the claim is known to hold for $R = \mathbb{R}$, it holds also for all real closed fields [→ 1.8.5]. □

Corollary 6.1.9. *Polynomial maps $R^m \to R^n$ are continuous.*

Corollary 6.1.10. $R^n \to R,\ x \mapsto \|x\| := \|x\|_2 := \sqrt{x_1^2 + \ldots + x_n^2}$ *is continuous.*

Remark 6.1.11. Because of 6.1.10 and 6.1.7, there is to every $\varepsilon \in R_{>0}$ some $\delta \in R_{>0}$ such that $\forall x \in R^n : (\|x\|_\infty < \delta \implies \|x\| < \varepsilon)$. On the other hand, $\|x\|_\infty \leq \|x\|$ for all $x \in R^n$. It follows that the topology on R^n is also generated by the open balls $\{x \in R^n \mid \|x-y\| < \varepsilon\}$ $(y \in R^n, \varepsilon > 0)$ and that 6.1.7 holds also with $\|.\|$ instead of $\|.\|_\infty$.

Remark 6.1.12. (a) By 6.1.9 and 6.1.11, R^n is obviously endowed with the initial topology with respect to all maps $R^n \to R,\ x \mapsto p(x)$ $(p \in R[\underline{X}])$ [→ 5.1.4].

(b) Because of (a), the topology on R^n is obviously generated by the sets

$$\{x \in R^n \mid p(x) > 0\} \qquad (p \in R[\underline{X}]).$$

(c) Viewing R^n in virtue of the injective map

$$R^n \to \operatorname{sper} R[\underline{X}],\ x \mapsto P_x = \{f \in R[\underline{X}] \mid f(x) \geq 0\}$$

as a subset of $\operatorname{sper} R[\underline{X}]$, the topology on R^n is due to (b) induced by the spectral topology [→ 5.2.1] on $\operatorname{sper} R[\underline{X}]$.

Theorem 6.1.13. (a) *If $A \subseteq R^m$ and $f\colon A \to R^n$ is K-semialgebraic, then $f(B) \in \mathscr{S}_n$ for all $B \in \mathscr{S}_m$ with $B \subseteq A$ and $f^{-1}(C) \in \mathscr{S}_m$ for all $C \in \mathscr{S}_n$.*

(b) *If $A \subseteq R^\ell$, $B \subseteq R^m$, $f\colon A \to B$ and $g\colon B \to R^n$ are K-semialgebraic, then $g \circ f\colon A \to R^n$ is again K-semialgebraic.*

(c) *If $A \in \mathscr{S}_n$, then the K-semialgebraic functions $A \to R$ form a subring of the ring R^A of all functions $A \to R$.*

Proof. (a) Let $A \subseteq R^m$ and $f\colon A \to R^n$ be K-semialgebraic. By 6.1.2, with Γ_f also $f(B) = \{y \in R^n \mid \exists x \in R^m : (x \in B \,\&\, (x,y) \in \Gamma_f)\}$ is for all $B \in \mathscr{S}_m$ with $B \subseteq A$ K-semialgebraic, and $f^{-1}(C) = \{x \in R^m \mid \exists y \in R^n : (y \in C \,\&\, (x,y) \in \Gamma_f)\}$ is for all $C \in \mathscr{S}_n$ also K-semialgebraic.

(b) Suppose $A \subseteq R^\ell$, $B \subseteq R^m$ and $f\colon A \to B$ as well as $g\colon B \to R^n$ are K-semialgebraic. Then $\Gamma_f \in \mathscr{S}_{\ell+m}$ and $\Gamma_g \in \mathscr{S}_{m+n}$ and thus

$$\Gamma_{g\circ f} = \{(x,z) \in A \times R^n \mid \exists y \in R^m : ((x,y) \in \Gamma_f \,\&\, (y,z) \in \Gamma_g)\} \in \mathscr{S}_{\ell+n}.$$

Hence $g \circ f$ is K-semialgebraic.

(c) If $A \in \mathscr{S}_n$ and $f_1, f_2\colon A \to R$ are K-semialgebraic, then also

$$A \to R^2,\ x \mapsto (f_1(x), f_2(x))$$

is K-semialgebraic. Now apply 6.1.8 and (b).

□

Example 6.1.14. If R is a non-Archimedean (real closed) extension of $\mathbb{R}$, then $[0,1]_R$ is not compact [→ 5.1.14]. Indeed, if $\varepsilon \in \mathfrak{m}_R$ [→ 5.4.7] with $\varepsilon > 0$, then

$$[0,1]_R \subseteq \bigcup_{a \in [0,1]_R} (a-\varepsilon, a+\varepsilon)_R,$$

but there is no $N \in \mathbb{N}$ and $a_1, \ldots, a_N \in [0,1]_R$ with $[0,1]_R \subseteq \bigcup_{k=1}^N (a_k - \varepsilon, a_k + \varepsilon)_R$ (for otherwise $[0,1]_{\mathbb{R}} = \mathrm{st}([0,1]_R) \subseteq \{\mathrm{st}(a_1), \ldots, \mathrm{st}(a_N)\}$ ↯).

Definition 6.1.15. Let $A \subseteq R^n$. We call A *bounded* if there is $b \in R$ with $\|x\| \leq b$ for all $x \in A$ [→ 6.1.10]. Moreover, A is called *K-semialgebraically compact* if $A \in \mathscr{S}_n$ and A is bounded and closed. We simply say "semialgebraically compact" instead of "R-semialgebraically compact".

Remark 6.1.16. From analysis, one knows for $R = \mathbb{R}$: A K-semialgebraic set $A \subseteq \mathbb{R}^n$ is compact if and only if it is K-semialgebraically compact.

Proposition 6.1.17. *Let $A \in \mathscr{S}_n$. Then the following are equivalent:*

(a) *A is bounded.*

(b) $\exists b \in R : \forall x \in A : \|x\| \leq b$

(c) $\exists b \in R : \forall x \in A : \|x\|_\infty \leq b$

(d) $\exists b \in K : \forall x \in A : \|x\| \leq b$

(e) $\exists b \in K : \forall x \in A : \|x\|_\infty \leq b$

Proof. WLOG $A \neq \emptyset$. We have (a) $\overset{6.1.15}{\iff}$ (b) $\overset{6.1.11}{\iff}$ (c) $\impliedby$ (e) $\impliedby$ (d). It remains to show (b) $\implies$ (d). Suppose therefore that (b) holds. The set

$$S := \{\|x\| \mid x \in A\} \subseteq R_{\geq 0}$$

is K-semialgebraic by real quantifier elimination [$\to$ 1.8.17]. Hence S can be defined by finitely many polynomials [$\to$ 1.8.6] with coefficients in K and by Lemma 1.5.3(a) we find some $b \in K_{>1}$ such that each of these polynomials has constant sign on the interval $(b, \infty)_R$. Then either $(b, \infty)_R \cap S = \emptyset$ or $(b, \infty)_R \subseteq S$. But the latter is impossible due to (b). Hence $\forall x \in A : \|x\| \leq b$. □

Theorem 6.1.18. *Let $A \subseteq R^m$ and suppose $f\colon A \to R^n$ is K-semialgebraic and continuous. Then for every K-semialgebraically compact set $B \subseteq A$, the set $f(B)$ is also K-semialgebraically compact.*

Proof. If $B \in \mathscr{S}_m$ with $B \subseteq A$, then $f(B) \in \mathscr{S}_n$ by 6.1.13(a) since f is K-semialgebraic. For the rest of the claim we can suppose that $K = R$. We fix a "complexity bound" $N \in \mathbb{N}$ and fix $m, n \in \mathbb{N}_0$ but no longer fix A and f. By 6.1.1, it suffices to show the following:

$(*)$ For all $f_1, \ldots, f_N, g_{11}, g_{12}, \ldots, g_{NN} \in R[X_1, \ldots, X_m, Y_1, \ldots, Y_n]_N$ and $\widetilde{f}_1, \ldots, \widetilde{f}_N, \widetilde{g}_{11}, \widetilde{g}_{12}, \ldots, \widetilde{g}_{NN} \in R[X_1, \ldots, X_m]_N$, if we set

$$\begin{aligned}
\Gamma &:= \bigcup_{i=1}^{N} \{(x,y) \in R^m \times R^n \mid f_i(x,y) = 0, g_{i1}(x,y) > 0, \ldots, g_{iN}(x,y) > 0\},\\
A &:= \{x \in R^m \mid \exists y \in R^n : (x,y) \in \Gamma\} \text{ and}\\
B &:= \bigcup_{i=1}^{N} \{x \in R^m \mid \widetilde{f}_i(x) = 0, \widetilde{g}_{i1}(x) > 0, \ldots, \widetilde{g}_{iN}(x) > 0\},
\end{aligned}$$

then

- Γ is not the graph of a continuous function from A to R^n or
- B is not a subset of A or
- B is not closed in R^m or
- B is not bounded in R^m or
- $\{y \in R^n \mid \exists x \in R^m : (x \in B \,\&\, (x,y) \in \Gamma)\}$ is closed and bounded in R^n.

We now in addition no longer fix R. One can easily figure out why the class of all real closed fields R for which $(*)$ holds is semialgebraic. For this aim, one applies many times the real quantifier elimination 1.8.17, for example for introducing the finitely many coefficients of the $f_i, g_{ij}, \widetilde{f}_i, \widetilde{g}_{ij}$ by universal quantifiers. By 1.8.5, $(*)$ now holds either for all or for no real closed field R. Therefore it is enough to show $(*)$ for $R = \mathbb{R}$. But we know this from analysis due to 6.1.16. □

Exercise 6.1.19. (a) The $\left\{\begin{matrix}\text{open}\\ \text{closed}\end{matrix}\right\}$ semialgebraic subsets of R are exactly the finite unions of pairwise disjoint sets of the form $\left\{\begin{matrix}(-\infty,\infty)_R, (-\infty,a)_R, (a,\infty)_R \text{ and } (a,b)_R\\ (-\infty,\infty)_R, (-\infty,a]_R, [a,\infty)_R \text{ and } [a,b]_R\end{matrix}\right\}$ with $a, b \in R$.

(b) The semialgebraically compact subsets of R are exactly the finite unions of pairwise disjoint sets of the form $[a,b]_R$ with $a, b \in R$.

6.2 The Łojasiewicz inequality

Proposition 6.2.1. *Let $a \in K$ and suppose $h\colon (a,\infty)_R \to R$ is K-semialgebraic. Then there is $b \in K \cap [a,\infty)_R$ and $N \in \mathbb{N}$ such that $|h(x)| \leq x^N$ for all $x \in (b,\infty)_R$.*

Proof. Using 6.1.1, we write

$$\Gamma_h = \bigcup_{i=1}^{k} \{(x,y) \in R^2 \mid f_i(x,y) = 0, g_{i1}(x,y) > 0, \ldots, g_{im}(x,y) > 0\}$$

with $k, m \in \mathbb{N}_0$ and $f_i, g_{ij} \in K[X,Y]$ where we suppose each of the k sets contributing to this union to be nonempty. We must have $k > 0$ and $\deg_Y f_i > 0$ for all $i \in \{1,\ldots,k\}$ (for otherwise there would be $x, c, d \in R$ with $c < d$ and $\{x\} \times (c,d)_R \subseteq \Gamma_h$ which is impossible since Γ_h is the graph of a function). Write $\prod_{i=1}^{k} f_i = \sum_{i=0}^{d} p_i Y^i$ with $d > 0$, $p_0, \ldots, p_d \in K[X]$ and $p_d \neq 0$. By rescaling one of the f_i if necessary, we can suppose that the leading coefficient of p_d is greater than 1. Choose $c \in K \cap [a,\infty)_R$ such that $p_d > 1$ on $(c,\infty)_R$ [→ 1.5.3(a)]. Because of $\sum_{i=0}^{d} p_i(x) h(x)^i = 0$ and $p_d(x) \neq 0$ for all $x \in (c,\infty)_R$, we have

$$|h(x)| \leq \max\left\{1, \frac{|p_0(x)| + \ldots + |p_{d-1}(x)|}{|p_d(x)|}\right\} \leq 1 + |p_0(x)| + \ldots + |p_{d-1}(x)|$$

for all $x \in (c,\infty)_R$ [→ 1.5.3(a)]. Now the existence of b is easy to see. □

Theorem 6.2.2 (Łojasiewicz inequality). *Let $n \in \mathbb{N}_0$ and suppose $A \subseteq R^n$ is K-semialgebraically compact and $f, g\colon A \to R$ are continuous K-semialgebraic functions satisfying*

$$\forall x \in A : (f(x) = 0 \implies g(x) = 0).$$

Then there is $N \in \mathbb{N}$ and $C \in K_{\geq 0}$ such that

$$\forall x \in A : |g(x)|^N \leq C|f(x)|.$$

Proof. With A also $A_t := \{x \in A \mid |g(x)| = \frac{1}{t}\}$ is K-semialgebraically compact for each $t \in R_{>0}$. Set $I := \{t \in R_{>0} \mid A_t \neq \emptyset\}$. For each $t \in I$,

$$f_t := \min\{|f(x)| \mid x \in A_t\}$$

exists by 6.1.18 and 6.1.19(b). Apparently, we have to show that there exist $N \in \mathbb{N}$ and $C \in K_{\geq 0}$ such that $\forall t \in I : \left(\frac{1}{t}\right)^N \leq C f_t$. By hypothesis, we have $f_t > 0$ for all $t \in I$. Furthermore,

$$R_{>0} \to R,\ t \mapsto \begin{cases} 0 & \text{if } t \notin I \\ \frac{1}{f_t} & \text{if } t \in I \end{cases}$$

is K-semialgebraic. Thus, by 6.2.1 there are $b \in K_{>0}$ and $N \in \mathbb{N}$ such that

$$(*) \qquad \frac{1}{f_t} \leq t^N$$

for all $t \in I \cap (b, \infty)_R$. Since

$$B := \left\{x \in A \mid |g(x)| \geq \frac{1}{b}\right\} = \bigcup_{t \in I \cap (0,b]_R} A_t$$

is K-semialgebraically compact, we can choose according to 6.1.18 and 6.1.17 some $C \in K_{\geq 1}$ satisfying

$$\frac{|g(x)|^N}{|f(x)|} \leq C$$

for all $x \in B$ (note that $f(x) \neq 0$ for all $x \in B$). We deduce

$$(**) \qquad \frac{1}{f_t} \leq C t^N$$

for all $t \in I \cap (0, b]_R$. Together with $(*)$, we obtain $(**)$ even for all $t \in I$ as desired. □

Lemma 6.2.3. ("shrinking map", in German: "Schränkungstranformation") Let $n \in \mathbb{N}_0$, $B := \{x \in R^n \mid \|x\| < 1\}$ and $S := \{x \in R^n \mid \|x\| = 1\}$. The maps

$$\varphi\colon R^n \to B,\ x \mapsto \frac{x}{\sqrt{1 + \|x\|^2}} \qquad \text{and}$$
$$\psi\colon B \to R^n,\ y \mapsto \frac{y}{\sqrt{1 - \|y\|^2}}$$

are $\mathbb{Q}$-semialgebraic, continuous and inverse to each other. For all $A \in \mathscr{S}_n$, we have

$$A \text{ closed} \iff \varphi(A) \cup S \text{ is } K\text{-semialgebraically compact.}$$

Proof. From 6.1.8, the $\mathbb{Q}$-semialgebraicity and the continuity are clear. For all $x \in R^n$, we have

$$\psi(\varphi(x)) = \frac{\frac{x}{\sqrt{1+\|x\|^2}}}{\sqrt{1-\frac{\|x\|^2}{1+\|x\|^2}}} = \frac{\frac{x}{\sqrt{1+\|x\|^2}}}{\frac{1}{\sqrt{1+\|x\|^2}}} = x.$$

For all $y \in B$, we have

$$\varphi(\psi(y)) = \frac{\frac{y}{\sqrt{1-\|y\|^2}}}{\sqrt{1+\frac{\|y\|^2}{1-\|y\|^2}}} = \frac{\frac{y}{\sqrt{1-\|y\|^2}}}{\sqrt{\frac{1}{1-\|y\|^2}}} = y.$$

Now let $A \in \mathscr{S}_n$. To show: A closed $\iff$ $\varphi(A) \cup S$ closed.

"$\Longleftarrow$" Suppose $\varphi(A) \cup S$ is closed. Then $\varphi(A) = (\varphi(A) \cup S) \cap B$ is closed in B (with respect to the topology induced from R^n) and thus also $A = \varphi^{-1}(\varphi(A))$ in R^n.

"$\Longrightarrow$" Let A be closed. Then $\varphi(A) = \psi^{-1}(A)$ is closed in B and hence $\varphi(A) = C \cap B$ for some closed set $C \subseteq R^n$. WLOG $C \subseteq B \cup S$ (otherwise replace C by $C \cap (B \cup S)$). WLOG $S \subseteq C$ (otherwise replace C by $C \cup S$). Now $\varphi(A) \cup S \subseteq C \subseteq (C \cap B) \cup (C \cap S) = \varphi(A) \cup S$. Hence $\varphi(A) \cup S = C$ is closed. □

Corollary 6.2.4. *Let $n \in \mathbb{N}_0$ and suppose that $A \subseteq R^n$ is closed and $f, g\colon A \to R$ are continuous K-semialgebraic functions satisfying*

$$\forall x \in A : (f(x) = 0 \implies g(x) = 0).$$

Then there are $N, k \in \mathbb{N}$ and $C \in K_{\geq 0}$ such that

$$\forall x \in A : |g(x)|^N \leq C(1 + \|x\|^2)^k |f(x)|.$$

Proof. By 6.1.4, A is K-semialgebraic. If A is bounded, then A is K-semialgebraically compact and the claim follows (with $k := 1$) from the Łojasiewicz inequality 6.2.2. Now suppose that A is unbounded. Since $\{\|x\| \mid x \in A\} \subseteq R$ is K-semialgebraic, there is then some $a \in K$ such that $(a, \infty)_R \subseteq \{\|x\| \mid x \in A\}$. The functions

$$\mathring{f}\colon (a,\infty)_R \to R, t \mapsto \max\{|f(x)| \mid x \in A, \|x\| = t\} \qquad \text{and}$$
$$\mathring{g}\colon (a,\infty)_R \to R, t \mapsto \max\{|g(x)| \mid x \in A, \|x\| = t\}$$

are semialgebraic. By 6.2.1, there are $b \in K \cap [a,\infty)_R$ with $b \geq 1$ and $\ell \in \mathbb{N}$ such that $\mathring{f}(t) \leq (1+t^2)^\ell$ and $\mathring{g}(t) \leq (1+t^2)^\ell$ for all $t \in (b,\infty)_R \subseteq R_{\geq 1}$. Now consider the continuous K-semialgebraic functions

$$f_0\colon A \to R,\ x \mapsto \frac{f(x)}{(1+\|x\|^2)^{\ell+1}} \qquad \text{and} \qquad g_0\colon A \to R,\ x \mapsto \frac{g(x)}{(1+\|x\|^2)^{\ell+1}}.$$

We have $\forall x \in A : (f_0(x) = 0 \implies g_0(x) = 0)$ and obviously it is enough to show that there are $N \in \mathbb{N}$ and $C \in K_{\geq 0}$ such that $\forall x \in A : |g_0(x)|^N \leq C|f_0(x)|$ (set then $k := \max\{1, (N-1)(\ell+1)\}$). The advantage of f_0 and g_0 over f and g is that there

is for all $\varepsilon \in R_{>0}$ a semialgebraically compact set $B \subseteq A$ such that $|f_0(x)| < \varepsilon$ and $|g_0(x)| < \varepsilon$ for all $x \in A \setminus B$. With the notation of Lemma 6.2.3, the K-semialgebraic functions

$$\widetilde{f}\colon \varphi(A) \cup S \to R, y \mapsto \begin{cases} 0 & \text{if } y \in S \\ f_0(\psi(y)) & \text{if } y \in \varphi(A) \end{cases} \quad \text{and}$$

$$\widetilde{g}\colon \varphi(A) \cup S \to R, y \mapsto \begin{cases} 0 & \text{if } y \in S \\ g_0(\psi(y)) & \text{if } y \in \varphi(A) \end{cases}$$

are continuous. For example, for $\widetilde{f}$ one sees this as follows: Since $f_0 \circ \psi|_{\varphi(A)}$ is continuous and $\varphi(A) = (\varphi(A) \cup S) \cap B$ is open in $\varphi(A) \cup S$, it suffices to show by 6.1.7 and 6.1.11 that

$$\forall y_0 \in S : \forall \varepsilon \in R_{>0} : \exists \delta \in R_{>0} : \forall y \in \varphi(A) : (\|y_0 - y\| < \delta \implies |f_0(\psi(y))| < \varepsilon).$$

To this end, let $y_0 \in S$ and $\varepsilon \in R_{>0}$. Choose a semialgebraically compact set $B \subseteq A$ with $|f_0(x)| < \varepsilon$ for all $x \in A \setminus B$. Then $\varphi(B)$ is semialgebraically compact by 6.1.18 and consequently $S \cup \varphi(A \setminus B) = (S \cup \varphi(A)) \setminus \varphi(B)$ is open in $\varphi(A) \cup S$. Thus there is $\delta \in R_{>0}$ with $\{y \in \varphi(A) \cup S \mid \|y_0 - y\| < \delta\} \subseteq S \cup \varphi(A \setminus B)$, i.e.,

$$\{y \in \varphi(A) \mid \|y_0 - y\| < \delta\} \subseteq \varphi(A \setminus B).$$

Now let $y \in \varphi(A)$ with $\|y_0 - y\| < \delta$. Then $y \in \varphi(A \setminus B)$ and thus $\psi(y) \in A \setminus B$. Hence $|f_0(\psi(y))| < \varepsilon$. This shows the continuity of $\widetilde{f}$. For all $y \in \varphi(A)$, we have obviously

$$\widetilde{f}(y) = 0 \implies f_0(\psi(y)) = 0 \implies g_0(\psi(y)) = 0 \implies \widetilde{g}(y) = 0.$$

Altogether, $\forall y \in \varphi(A) \cup S : (\widetilde{f}(y) = 0 \implies \widetilde{g}(y) = 0)$. Since $\varphi(A) \cup S$ is K-semialgebraically compact by 6.2.3, we get from the Łojasiewicz inequality 6.2.2 $N \in \mathbb{N}$ and $C \in R_{\geq 0}$ with $\forall y \in \varphi(A) \cup S : |\widetilde{g}(y)|^N \leq C|\widetilde{f}(y)|$. In particular, we obtain $\forall y \in \varphi(A) : |g_0(\psi(y))|^N \leq C|f_0(\psi(y))|$ which means $\forall x \in A : |g_0(x)|^N \leq C|f_0(x)|$ as desired. □

6.3 The finiteness theorem for semialgebraic sets

Definition 6.3.1. Let $n \in \mathbb{N}_0$. A subset S of R^n is called K-*basic* $\left\{\begin{matrix} open \\ closed \end{matrix}\right\}$ if there are $m \in \mathbb{N}_0$ and $g_1, \ldots, g_m \in K[\underline{X}]$ satisfying $S = \{x \in R^n \mid g_1(x) \left\{\begin{matrix} > \\ \geq \end{matrix}\right\} 0, \ldots, g_m(x) \left\{\begin{matrix} > \\ \geq \end{matrix}\right\} 0\}$.

Remark 6.3.2. Every K-basic $\left\{\begin{matrix} \text{open} \\ \text{closed} \end{matrix}\right\}$ subset of R^n is K-semialgebraic and $\left\{\begin{matrix} \text{open} \\ \text{closed} \end{matrix}\right\}$ in R^n.

Theorem 6.3.3 (Finiteness theorem for semialgebraic sets). *Let $n \in \mathbb{N}_0$ and $S \in \mathscr{S}_n$ $\left\{\begin{matrix}\text{open}\\\text{closed}\end{matrix}\right\}$. Then S is a finite union of K-basic $\left\{\begin{matrix}\text{open}\\\text{closed}\end{matrix}\right\}$ subsets of R^n.*

Proof.

S is a finite union of K-basic open subsets of R^n

$\iff$ S is a finite union of finite intersections of sets of the form $\{x \in R^n \mid g(x) > 0\}$ $(g \in K[\underline{X}])$

$\iff$ $\complement S$ is a finite intersection of finite unions of sets of the form $\{x \in R^n \mid g(x) \geq 0\}$ $(g \in K[\underline{X}])$

$\overset{1.8.1}{\iff}$ $\complement S$ is a finite union of finite intersections of sets of the form $\{x \in R^n \mid g(x) \geq 0\}$ $(g \in K[\underline{X}])$

$\iff$ $\complement S$ is a finite union of K-basic closed subsets of R^n.

It is thus enough to show the claim for open S. Write

$$S = \bigcup_{i=1}^{k} \{x \in R^n \mid f_i(x) = 0, g_{i1}(x) > 0, \ldots, g_{im}(x) > 0\}$$

according to 6.1.1 with $k, m \in \mathbb{N}_0$, $f_i, g_{ij} \in K[\underline{X}]$. Fix $i \in \{1, \ldots, k\}$. It is enough to find a K-basic open set $U \subseteq R^n$ such that

$$\{x \in R^n \mid f_i(x) = 0, g_{i1}(x) > 0, \ldots, g_{im}(x) > 0\} \subseteq U \subseteq S.$$

Consider the closed set $A := R^n \setminus S \in \mathscr{S}_n$ and the continuous K-semialgebraic functions

$$f\colon A \to R, x \mapsto (f_i(x))^2 \qquad \text{and}$$
$$g\colon A \to R, x \mapsto \prod_{j=1}^{m} (|g_{ij}(x)| + g_{ij}(x)).$$

We have $\forall x \in A : (f(x) = 0 \implies g(x) = 0)$. By 6.2.4, there thus exist $N, k \in \mathbb{N}$ and $C \in K_{\geq 0}$ such that $\forall x \in A : |g(x)|^N \leq C(1 + \|x\|^2)^k f(x)$. For all $x \in A$ satisfying $g_{i1}(x) > 0, \ldots, g_{im}(x) > 0$, we thus have $(2^m \prod_{j=1}^{m} g_{ij}(x))^N \leq C(1 + \sum_{j=1}^{n} x_j^2)^k f_i(x)^2$. Set

$$U := \left\{ x \in R^n \mid C\left(1 + \sum_{j=1}^{n} x_j^2\right)^k f_i(x)^2 < \left(2^m \prod_{j=1}^{m} g_{ij}(x)\right)^N, g_{i1}(x) > 0, \ldots, g_{im}(x) > 0 \right\}.$$

Then $U \cap A = \emptyset$ and $\{x \in R^n \mid f_i(x) = 0, g_{i1}(x) > 0, \ldots, g_{im}(x) > 0\} \subseteq U \subseteq S$. □

Example 6.3.4. The "slashed square" $S := (-1,1)_R^2 \setminus ([0,1]_R \times \{0\})$ is K-semialgebraic and open. By 6.3.3, it is thus a finite union of K-basic open subsets of R^2. Indeed,

$$S = \{(x,y) \in R^2 \mid -1 < x < 1, -(y+1)y^2(y-1) > 0\} \cup \left\{(x,y) \in R^2 \mid \left(x + \frac{1}{2}\right)^2 + y^2 < \left(\frac{1}{2}\right)^2\right\}$$

is a union of two K-basic open sets. However, S is not K-basic open. To show this, we assume

$$S = \{(x,y) \in R^2 \mid g_1(x,y) > 0, \ldots, g_m(x,y) > 0\}$$

with $m \in \mathbb{N}_0$, $g_i \in K[X,Y]$. For continuity reasons, we have $g_i(x,0) \geq 0$ for all $x \in [0,1]_R$ and $i \in \{1,\ldots,m\}$. Because of $([0,1]_R \times \{0\}) \cap S = \emptyset$, we have thus $[0,1]_R = \bigcup_{i=1}^m \{x \in [0,1]_R \mid g_i(x,0) = 0\}$. WLOG $\#\{x \in [0,1]_R \mid g_1(x,0) = 0\} = \infty$. Then $g_1(X,0) = 0$ and consequently $(R \times \{0\}) \cap S = \emptyset$ in contradiction to $(-1,0)_R \times \{0\} \subseteq S$.

Theorem 6.3.5 (Abstract version of the finiteness theorem for semialgebraic sets)**.** *Let $R|K$ be algebraic, i.e., R be the real closure of $(K, K \cap R^2)$. Let $n \in \mathbb{N}_0$ and write $A := K[\underline{X}]$ and $T := \sum K_{\geq 0} A^2$ so that we are in the setting described before 3.6.3. Denote by*

$$\text{Fatten}\colon \mathscr{S}_n \to \mathscr{C} := \mathscr{C}_{(A,T)}$$

again the fattening [→ 3.6.4, 5.3.1]. *Let $S \in \mathscr{S}_n$. Then*

$$S \left\{ \begin{matrix} \text{open} \\ \text{closed} \end{matrix} \right\} \text{ in } R^n \iff \text{Fatten}(S) \left\{ \begin{matrix} \text{open} \\ \text{closed} \end{matrix} \right\} \text{ in } \operatorname{sper}(A,T).$$

Proof. It is enough to show: S open $\iff$ Fatten(S) open.

"$\Longleftarrow$" By definition of the spectral topology [→ 5.2.1], Fatten(S) is a union of sets of the form $\{P \in \operatorname{sper}(A,T) \mid \widehat{g}_1(P) > 0, \ldots, \widehat{g}_m(P) > 0\}$ ($m \in \mathbb{N}_0, g_1, \ldots, g_m \in A$). By 5.2.3 and 5.1.21, Fatten(S) is quasicompact [→ 5.1.14] with respect to the constructible topology [→ 5.2.1]. Hence Fatten(S) is a finite union of sets of the described form, i.e.,

$$(**) \qquad \text{Fatten}(S) = \bigcup_{i=1}^{k} \{P \in \operatorname{sper}(A,T) \mid \widehat{g}_{i1}(P) > 0, \ldots, \widehat{g}_{im}(P) > 0\}$$

with $k, m \in \mathbb{N}_0$, $g_{ij} \in A$. It follows by 3.6.4 that

$$(*) \qquad S = \bigcup_{i=1}^{k} \{x \in R^n \mid g_{i1}(x) > 0, \ldots, g_{im}(x) > 0\}.$$

In particular, S is open.

"$\Longrightarrow$" By the finiteness theorem for semialgebraic sets 6.3.3, we can find $k, m \in \mathbb{N}_0$ and $g_{ij} \in A$ such that $(*)$ holds. It follows that $(**)$ holds. In particular, Fatten(S) is open. □

Remark 6.3.6. The description of 6.3.5 as an abstract version of 6.3.3 is motivated by the fact that one can easily retrieve the latter from the first: Note first that one can reduce in 6.3.3 to the case where $R|K$ is algebraic by using the transfer between R and $\overline{(K, K \cap R_{\geq 0})}$ [→ 1.9.5]. For this, one has to argue that this transfer preserves openness which can be accomplished by real quantifier elimination 1.8.17. Thus let now $R|K$ be algebraic, $n \in \mathbb{N}_0$ and $S \in \mathscr{S}_n$ open (by the first part of the proof of Theorem 6.3.3, it

suffices to treat the case of open sets). We have to show that S is a finite union of K-basic open subsets of R^n. As seen in the easy part "$\Longleftarrow$" of the proof of 6.3.5, for this purpose, it suffices to show that $\mathrm{Fatten}(S)$ is open. This follows from the difficult part "$\Longrightarrow$" of 6.3.5.

Corollary 6.3.7 (Strengthening of 5.3.8)**.** *Let $R|K$ be algebraic, i.e., R be the real closure of $(K, K \cap R_{\geq 0})$. Let $n \in \mathbb{N}_0$ and write $A := K[\underline{X}]$ and $T := \sum K_{\geq 0} A^2$. Then*

$$\operatorname{sper} R[\underline{X}] \to \operatorname{sper}(A, T),\ P \mapsto P \cap A$$

is a homeomorphism with respect to both, the spectral as well as the constructible topology on both sides.

Proof. The map is continuous with respect to both topologies by 5.2.7 and bijective by 5.3.8. According to the definition of a homeomorphism 5.2.2 and the definition of both topologies in 5.2.1, it suffices to show that for all $C \in \mathscr{C}_{R[\underline{X}]}$ we have $\{P \cap A \mid P \in C\} \in \mathscr{C}_{(A,T)}$ and that this latter is set is open in $\operatorname{sper}(A, T)$ whenever C is open in $\operatorname{sper} R[\underline{X}]$. For this purpose, let $C \in \mathscr{C}_{R[\underline{X}]}$. The slimming $\{x \in R^n \mid P_x \in C\}$ [→ 3.6.4] of C is then a semialgebraic subset of R^n and thus even K-semialgebraic by 5.3.7 since $R|K$ is algebraic. By 6.1.1, we thus find $k, m \in \mathbb{N}_0$ and $f_i, g_{ij} \in K[\underline{X}]$ such that

$$\{x \in R^n \mid P_x \in C\} = \bigcup_{i=1}^{k} \{x \in R^n \mid f_i(x) = 0, g_{i1}(x) > 0, \ldots, g_{im}(x) > 0\},$$

where one can even choose $f_1 = \ldots = f_k = 0$ by the finiteness theorem for semialgebraic sets 6.3.3 in the case where C is open. Fattening this, we obtain

$$C = \bigcup_{i=1}^{k} \{P \in \operatorname{sper} R[\underline{X}] \mid \widehat{f}_i(P) = 0, \widehat{g}_{i1}(P) > 0, \ldots, \widehat{g}_{im}(P) > 0\}$$

and therefore [→ 5.2.7]

$$\{P \cap A \mid P \in C\} = \bigcup_{i=1}^{k} \{P \in \operatorname{sper}(A, T) \mid \widehat{f}_i(P) = 0, \ldots, \widehat{g}_{i1}(P) > 0, \ldots, \widehat{g}_{im}(P) > 0\}$$
$$\in \mathscr{C}_{(A,T)}.$$

If C is open, then so is $\{P \cap A \mid P \in C\}$ because of the choice of $f_i = 0$. □

Remark 6.3.8. In the situation of 6.3.5, one can obviously generalize 6.1.12 as follows:

(a) R^n is equipped with the initial topology with respect to all maps $R^n \to R,\ x \mapsto p(x)$ $(p \in A)$.

(b) The topology on R^n is generated by the sets $\{x \in R^n \mid p(x) > 0\}$ $(p \in A)$.

(c) Viewing R^n in virtue of the injective map [→ 3.6.3]

$$R^n \to \operatorname{sper} A,\ x \mapsto P_x = \{f \in A \mid f(x) \geq 0\}$$

as a subset of $\operatorname{sper} A$, the topology on R^n is induced by the spectral topology on $\operatorname{sper} A$ [→ 6.3.7].

§7 Convex sets in vector spaces

In this chapter, K denotes always a subfield of $\mathbb{R}$ equipped with the order and the subspace topology [→ 5.1.5(a)] induced by $\mathbb{R}$ unless otherwise specified.

7.1 The isolation theorem for cones

Definition 7.1.1. Let V be a K-vector space. A subset $C \subseteq V$ is called a *(convex) cone* (in V) if $0 \in C$, $C + C \subseteq C$ and $K_{\geq 0}C \subseteq C$ [→ 1.1.18]. A cone $C \subseteq V$ is called *proper* if $C \neq V$.

Example 7.1.2. Let T be a preorder [→ 1.2.1] of $K[\underline{X}]$ with $K_{\geq 0} \subseteq T$. Then T is a cone. Moreover, T is proper as a preorder [→ 1.2.5] if and only if T is proper as a cone.

Proposition 7.1.3. *Let V be a K-vector space and $C \subseteq V$. Then the following are equivalent:*

(a) *C is a cone.*

(b) *C is convex* [→ 2.4.1]*, $C \neq \emptyset$ and $K_{\geq 0}C \subseteq C$.*

Proof. (a) $\Longrightarrow$ (b) is trivial.

(b) $\Longrightarrow$ (a) Suppose that (b) holds. From $C \neq \emptyset$ and $0C \subseteq C$, we get $0 \in C$. To show: $C + C \subseteq C$. Let $x, y \in C$. Then $\frac{x}{2} + \frac{y}{2} \in C$ and thus $x + y = 2\left(\frac{x}{2} + \frac{y}{2}\right) \in C$. □

Definition 7.1.4. Let C be a cone in the K-vector space V and $u \in V$. Then u is called a *unit* for C (in V) if for every $x \in V$ there is some $N \in \mathbb{N}$ with $Nu + x \in C$.

Example 7.1.5. [→ 7.1.2] Let T be a preorder of $K[\underline{X}]$ with $K_{\geq 0} \subseteq T$. Then T is Archimedean [→ 4.1.2(a)] if and only if 1 is a unit for T.

Proposition 7.1.6. *Let C be a cone on the K-vector space V and $u \in V$. Then the following are equivalent:*

(a) *u is a unit for C.*

(b) $V = C - \mathbb{N}u$

(c) $V = C - K_{\geq 0}u$

(d) *$u \in C$ and $V = C + \mathbb{Z}u$*

(e) *$u \in C$ and $V = C + Ku$*

(f) $\forall x \in V : \exists \varepsilon \in K_{>0} : u + \varepsilon x \in C$

Proof. (a) $\Longrightarrow$ (b) $\Longrightarrow$ (c) is clear.

(c) $\Longrightarrow$ (d) Suppose that (c) holds. Then $u \in C - K_{\geq 0}u$ and thus $(1 + K_{\geq 0})u \in C$ and so $u \in C$. Fix now $x \in V$. To show: $x \in C + \mathbb{Z}u$. Choose $\lambda \in K_{\geq 0}$ with $x \in C - \lambda u$. Choose $N \in \mathbb{N}$ with $\lambda \leq N$. Then $(N - \lambda)u \in C$ and hence

$$x = (x - (N - \lambda)u) + (N - \lambda)u \in (C - \lambda u - (N - \lambda)u) + C \subseteq C - Nu \subseteq C - \mathbb{N}u \subseteq C + \mathbb{Z}u.$$

(d) $\Longrightarrow$ (e) is trivial.

(e) $\Longrightarrow$ (f) Suppose that (e) holds and let $x \in V$. Choose $\lambda \in K$ such that $x \in C - \lambda u$. If $\lambda \leq 0$, then $x \in C$ and consequently $u + \varepsilon x = u + x \in C + C \subseteq C$ with $\varepsilon := 1$. If $\lambda > 0$, then set $\varepsilon := \frac{1}{\lambda} > 0$. Then $u + \varepsilon x \in \varepsilon C \subseteq C$.

(f) $\Longrightarrow$ (a) Suppose that (f) holds and let $x \in V$. To show: $\exists N \in \mathbb{N} : Nu + x \in C$. Choose $\varepsilon \in K_{>0}$ with $u + \varepsilon x \in C$. Choose $N \in \mathbb{N}$ with $\frac{1}{\varepsilon} \leq N$. From (f), it follows also that $u \in C$ and hence $(N - \frac{1}{\varepsilon})u \in C$. Now $Nu + x = (N - \frac{1}{\varepsilon})u + \frac{1}{\varepsilon}u + x \in C + \frac{1}{\varepsilon}(u + \varepsilon x) \subseteq C + \frac{1}{\varepsilon}C \subseteq C + C \subseteq C$. □

Corollary 7.1.7. *Let u be a unit for the cone C in the K-vector space V. Then $u \in C$ and $V = C - C$.*

Remark 7.1.8. The units for a cone in K^n are exactly its interior points [→ 5.2.5, 7.1.6(f)].

Definition 7.1.9. Let V be a K-vector space, $C \subseteq V$ and $u \in V$. A *state* of (V, C, u) is a K-linear function $\varphi\colon V \to \mathbb{R}$ satisfying $\varphi(C) \subseteq \mathbb{R}_{\geq 0}$ and $\varphi(u) = 1$. We refer to the set $S(V, C, u) \subseteq \mathbb{R}^V$ of all states of (V, C, u) as the *state space* of (V, C, u).

Example 7.1.10. Set $K := \mathbb{R}$, $V := \mathbb{R}[X]$, $C := P_\infty \in \operatorname{sper} \mathbb{R}[X]$. Then the cone C does not possess a unit in V and we have $S(V, C, u) = \emptyset$ for all $u \in V$. Indeed, let $u \in V$. Choose $d \in \mathbb{N}$ with $d > \deg u$. Then $u - \varepsilon X^d \notin C$ for all $\varepsilon > 0$. By 7.1.6(f), u is thus not a unit for C. Assume $\varphi \in S(V, C, u)$. Then $\varepsilon\varphi(X^d) - 1 = \varphi(\varepsilon X^d - u) \in \varphi(C) \subseteq \mathbb{R}_{\geq 0}$ for all $\varepsilon > 0$ ↯.

Example 7.1.11. Set $K := \mathbb{Q}$, $V := \mathbb{Q}^2$, $C := \{(x, y) \in \mathbb{Q}^2 \mid y \geq \sqrt{2}x\}$. All elements of C except 0 are units for C [→ 7.1.8]. There is no $\varphi \in V^* \setminus \{0\}$ satisfying $\varphi(C) \subseteq \mathbb{Q}_{\geq 0}$ but for each $u \in C \setminus \{0\}$, we have $\#S(V, C, u) = 1$.

Lemma 7.1.12. Let u be a unit for a proper cone C in the K-vector space V. Then

$$\varrho\colon V \to \mathbb{R},\ x \mapsto \sup\{\lambda \in K \mid x - \lambda u \in C\}$$

is well-defined and we have $\varrho(x) + \varrho(y) \leq \varrho(x + y)$ as well as $\varrho(\lambda x) = \lambda\varrho(x)$ for all $x, y \in V$ and $\lambda \in K_{\geq 0}$.

Proof. Let $x, y \in V$ and $\lambda \in K_{\geq 0}$. For the well-definedness of ϱ, we have to show that $I := \{\lambda \in K \mid x - \lambda u \in C\}$ is nonempty and bounded from above [→ 1.1.9, 1.1.16]. Since u is a unit for C, we have $I \neq \emptyset$ and furthermore there is $N \in \mathbb{N}$ such that $-x + Nu \in C$. Then $\lambda < N+1$ for all $\lambda \in I$ since otherwise, if $\lambda \in I$ satisfied $\lambda \geq N+1$, then

$$\begin{aligned} -u &= Nu - (N+1)u = (-x + Nu) + x - (N+1)u \\ &\in C + x - \lambda u + (\lambda - (N+1))u \\ &\subseteq C + C + K_{\geq 0}u \subseteq C. \end{aligned}$$

But now $-u \notin C$ for otherwise $C \overset{7.1.6(b)}{=} V$. Now choose sequences $(\lambda_n)_{n\in\mathbb{N}}$ and $(\mu_n)_{n\in\mathbb{N}}$ in K such that $x - \lambda_n u, y - \mu_n u \in C$ for all $n \in \mathbb{N}$ and $\lim_{n\to\infty} \lambda_n = \varrho(x)$ as well as $\lim_{n\to\infty} \mu_n = \varrho(y)$. Then we have $(x+y) - (\lambda_n + \mu_n)u \in C + C \subseteq C$ and thus $\lambda_n + \mu_n \leq \varrho(x+y)$ for all $n \in \mathbb{N}$. It follows that

$$\varrho(x) + \varrho(y) = \left(\lim_{n\to\infty} \lambda_n\right) + \left(\lim_{n\to\infty} \mu_n\right) = \lim_{n\to\infty} (\lambda_n + \mu_n) \leq \varrho(x+y).$$

Moreover, $\lambda x - \lambda\lambda_n u \in \lambda C \subseteq C$ and thus $\lambda\lambda_n \leq \varrho(\lambda x)$ for all $n \in \mathbb{N}$. It follows that $\lambda\varrho(x) = \lambda \lim_{n\to\infty} \lambda_n = \lim_{n\to\infty} \lambda\lambda_n \leq \varrho(\lambda x)$ and analogously $\frac{1}{\lambda}\varrho(\lambda x) \leq \varrho\left(\frac{1}{\lambda}(\lambda x)\right)$ if $\lambda \neq 0$, i.e., $\lambda\varrho(x) = \varrho(\lambda x)$. □

Theorem 7.1.13 (Isolation theorem for cones)**.** *Let u be a unit for the proper cone C in the K-vector space V. Then $S(V, C, u) \neq \emptyset$.*

Proof. Since the union of a nonempty chain of cones in V is again a cone in V, we can use Zorn's lemma to enlarge C to a cone of V that is maximal with respect to the property of not containing $-u$. WLOG suppose that C has already this maximality property.

Claim 1: $C \cup -C = V$

Explanation. Let $x \in V$ with $x \notin -C$. To show: $x \in C$. Due to the maximality of C it is enough to show that the cone $C + K_{\geq 0}x$ does not contain $-u$. But if we had $-u = y + \lambda x$ for some $y \in C$ and $\lambda \in K_{\geq 0}$, then $\lambda > 0$ and $x = \frac{1}{\lambda}(-u - y) \in -C$ ↯.

Consider for each $x \in V$, the sets

$$I_x := \{\lambda \in K \mid x - \lambda u \in C\} \text{ and } J_x := \{\lambda \in K \mid x - \lambda u \in -C\}.$$

Claim 2: $\forall x \in V : \forall \lambda \in I_x : \forall \mu \in J_x : \lambda \leq \mu$

Explanation. Let $x \in V$, $\lambda \in I_x$ and $\mu \in J_x$. Then $x - \lambda u \in C$ and $\mu u - x \in C$. Thus, $(\mu - \lambda)u = (\mu u - x) + (x - \lambda u) \in C + C \subseteq C$. If we had $\mu < \lambda$, then we had $-u \in C$ ↯.

Consider now $\varphi\colon V \to \mathbb{R},\ x \mapsto \sup I_x$ [→ 7.1.12].

Claim 3: $-\varphi(x) = \sup\{\lambda \in K \mid x - \lambda(-u) \in -C\}$ for all $x \in V$

Explanation. Let $x \in V$. From $I_x \cup J_x \overset{\text{Claim 1}}{=} K$ and Claim 2, we get

$$\varphi(x) = \sup I_x = \inf J_x$$

and hence

$$-\varphi(x) = -\inf J_x = \sup\{-\lambda \mid \lambda \in K, x - \lambda u \in -C\} = \sup\{\lambda \in K \mid x + \lambda u \in -C\}.$$

From 7.1.12, we obtain $\varphi(x) + \varphi(y) \leq \varphi(x+y)$ and $\varphi(\lambda x) = \lambda\varphi(x)$ for all $x, y \in V$ and $\lambda \in K_{\geq 0}$. Since $-u$ is a unit for the proper cone $-C$, 7.1.12 and Claim 3 yield also $-\varphi(x) - \varphi(y) \leq -\varphi(x+y)$ for all $x, y \in V$. It follows that

$$\varphi(x) + \varphi(y) \leq \varphi(x+y) \leq \varphi(x) + \varphi(y)$$

and therefore $\varphi(x) + \varphi(y) = \varphi(x+y)$ for all $x, y \in V$. In particular, $\varphi(x) + \varphi(-x) = \varphi(0) = 0$ and hence $\varphi(-x) = -\varphi(x)$ for all $x \in V$ from which we deduce

$$\varphi((-\lambda)x) = \varphi(-\lambda x) = -\varphi(\lambda x) = -\lambda\varphi(x) = (-\lambda)\varphi(x)$$

for all $x \in V$ and $\lambda \in K_{\geq 0}$. Altogether, $\varphi(\lambda x) = \lambda\varphi(x)$ for all $x \in V$ and $\lambda \in K_{\geq 0} \cup K_{\leq 0} = K$, i.e., φ is K-linear. Obviously, $\varphi(C) \subseteq \mathbb{R}_{\geq 0}$ and $\varphi(u) = 1$. Therefore $\varphi \in S(V, C, u)$. □

Lemma 7.1.14. Let C be a cone in the K-vector space V and $x \in V$. Then

$$x \in C \iff x \in C - K_{\geq 0}x.$$

Proof. "$\Longrightarrow$" is trivial.

"$\Longleftarrow$" Let $x \in C - K_{\geq 0}x$, for instance $x = y - \lambda x$ with $y \in C$ and $\lambda \in K_{\geq 0}$. Then

$$x = \frac{1}{1+\lambda}y \in C.$$

□

Corollary 7.1.15. *Suppose u is a unit for the cone C in the K-vector space V and $x \in V$. If $\varphi(x) > 0$ for all $\varphi \in S(V, C, u)$, then $x \in C$.*

Proof. Suppose $x \notin C$. To show: $\exists \varphi \in S(V, C, u) : \varphi(x) \leq 0$. By 7.1.14, the cone $C - K_{\geq 0}x$ is proper. Since u is a unit for C, it is of course also a unit for $C - K_{\geq 0}x$. By the isolation theorem 7.1.13, there is $\varphi \in S(V, C - K_{\geq 0}x, u)$. We have $\varphi \in S(V, C, u)$ and $\varphi(x) \leq 0$. □

Exercise 7.1.16. [→ 7.1.9] Let V be a K-vector space, $C \subseteq V$ and $u \in V$. We equip the $\mathbb{R}$-vector space $\mathbb{R}^V$ of all functions from V to $\mathbb{R}$ with the product topology [→ 5.1.5(b)]. Then $S(V, C, u)$ is a closed convex subset of $\mathbb{R}^V$ which we equip with the subspace topology [→ 5.1.5(a)]. This topology is at the same time also the initial topology [→ 5.1.4] with respect to the functions

$$S(V, C, u) \to \mathbb{R},\ \varphi \mapsto \varphi(x) \qquad (x \in V)$$

[→ 5.1.6].

Theorem 7.1.17. *Let u be a unit for the cone C in the K-vector space V. Then the state space $S(V,C,u)$ is compact* [→ 5.1.14].

Proof. Choose for each $x \in V$ an $N_x \in \mathbb{N}$ such that $\pm x + N_x u \in C$. Then we have for all $\varphi \in S(V,C,u)$ and $x \in V$ that $\pm\varphi(x) + N_x = \varphi(\pm x + N_x u) \geq 0$ and thus

$$\varphi(x) \in [-N_x, N_x].$$

Thus $S(V,C,u) \subseteq \prod_{x\in V}[-N_x, N_x]$. From analysis (cf. 6.1.16) and Tikhonov's theorem 5.1.18, $\prod_{x\in V}[-N_x, N_x]$ is compact with respect to the product topology. But the product topology on $\prod_{x\in V}[-N_x, N_x]$ is induced by the topology of $\mathbb{R}^V$ [→ 5.1.6]. By 7.1.16, $S(V,C,u)$ is thus closed in the compact space $\prod_{x\in V}[-N_x, N_x]$ and hence is compact itself [→ 5.1.21]. □

Exercise 7.1.18. Let M and N be topological spaces and $f\colon M \to N$ be continuous. If M is quasicompact [→ 5.1.14], then so is $f(M)$ [→ 5.1.21]

Corollary 7.1.19. *Let M be a nonempty quasicompact topological space and $f\colon M \to \mathbb{R}$ be continuous. Then f takes on a minimum and a maximum, i.e., there are $x, y \in M$ with*

$$f(x) \leq f(z) \leq f(y)$$

for all $z \in M$.

Proof. $f(M)$ is compact by 7.1.18. Hence $f(M)$ is nonempty, bounded and closed. From the first two properties, it follows that $\inf f(M), \sup f(M) \in \mathbb{R}$ exist [→ 1.1.9(c), 1.1.16]. The last property yields $\inf f(M) = \min f(M)$ and $\sup f(M) = \max f(M)$. □

Theorem 7.1.20 (Strengthening of 7.1.15)**.** [→ 4.2.2] *Let u be a unit for the cone C in the K-vector space V and $x \in V$. Then the following are equivalent:*

(a) $\forall \varphi \in S(V,C,u) : \varphi(x) > 0$

(b) $\exists N \in \mathbb{N} : x \in \frac{1}{N}u + C$

(c) *x is a unit for C.*

Proof. (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (b) Suppose that (a) holds. If $S(V,C,u) = \emptyset$, then $C = V$ by 7.1.13 and we can choose $N \in \mathbb{N}$ arbitrarily. Suppose therefore that $S(V,C,u) \neq \emptyset$. Then the continuous function $S(V,C,u) \to \mathbb{R}$, $\varphi \mapsto \varphi(x)$ takes on by 7.1.17 and 7.1.19 a minimum μ for which $\mu > 0$ holds by (a). Choose $N \in \mathbb{N}$ such that $\frac{1}{N} < \mu$. Then $\varphi\left(x - \frac{1}{N}u\right) = \varphi(x) - \frac{1}{N} \geq \mu - \frac{1}{N} > 0$ for all $\varphi \in S(V,C,u)$. Now 7.1.15 yields that $x - \frac{1}{N}u \in C$.

(b) $\Longrightarrow$ (c) Suppose that (b) holds and let $y \in V$. To show: $\exists N \in \mathbb{N} : Nx + y \in C$. Choose $N', N'' \in \mathbb{N}$ with $x \in \frac{1}{N'}u + C$ and $N''u + y \in C$. Setting $N := N'N''$, we obtain $Nx + y \in N''N'\left(\frac{1}{N'}u + C\right) + y \subseteq N''(u + C) + y \subseteq N''u + y + C \subseteq C + C \subseteq C$.

(c) $\Longrightarrow$ (a) Suppose that (c) holds and let $\varphi \in S(V,C,u)$. To show: $\varphi(x) > 0$. Choose $N \in \mathbb{N}$ with $Nx - u \in C$. Then $N\varphi(x) - 1 = \varphi(Nx - u) \geq 0$ and thus $\varphi(x) \geq \frac{1}{N} > 0$ for all $\varphi \in S(V,C,u)$. □

7.2 Separating convex sets in topological vector spaces

Definition 7.2.1. A K-vector space V together with a topology on V [→ 5.1.2(a)] is called a *topological K-vector space* if $V \times V \to V, (x,y) \mapsto x+y$ and $K \times V \to V,\ (\lambda, x) \mapsto \lambda x$ are continuous and $\{0\}$ is a closed set in V.

Example 7.2.2. (a) If I is a set, then K^I (endowed with the product topology [→ 5.1.5(b)]) is a topological K-vector space.

(b) A K-vector space V together with the discrete topology on V is a topological vector space if and only if $V = \{0\}$. Indeed, if $y \in V \setminus \{0\}$, then

$$\{(\lambda, x) \in K \times V \mid \lambda x = y\} = \{(\lambda, \lambda^{-1}y) \mid \lambda \in K^\times\}$$

is not open in $K \times V$.

(c) From analysis, one knows that every normed $\mathbb{R}$-vector space, in particular every $\mathbb{R}$-vector space with scalar product, is a topological $\mathbb{R}$-vector space.

Lemma 7.2.3. Let V be a K-vector space and $A \subseteq V$ be convex. If $0 \notin A \neq \emptyset$, then A generates a proper convex cone, i.e., $\sum_{x \in A} K_{\geq 0} x \neq V$.

Proof. Suppose that $A \neq \emptyset$ and $\sum_{x \in A} K_{\geq 0} x = V$. We show $0 \in A$. Choose $y \in A$ and write $-y = \sum_{i=1}^m \lambda_i x_i$ with $\lambda_1, \ldots, \lambda_m \in K_{\geq 0}$ and $x_1, \ldots, x_m \in A$. Setting $\mu := 1 + \sum_{i=1}^m \lambda_i > 0$, we have then $0 = \frac{1}{\mu} y + \sum_{i=1}^m \frac{\lambda_i}{\mu} x_i \in A$ since $\frac{1}{\mu} + \sum_{i=1}^m \frac{\lambda_i}{\mu} = \frac{\mu}{\mu} = 1$. □

Lemma 7.2.4. [→ 7.1.8] Let V be a topological K-vector space, $C \subseteq V$ a convex cone and $u \in C^\circ$ [→ 5.2.5]. Then u is a unit for C [→ 7.1.4].

Proof. We show $\forall x \in V : \exists \varepsilon \in K_{>0} : u + \varepsilon x \in C$ [→ 7.1.6(f)]. For this aim, fix $x \in V$. From Definition 7.2.1, it follows that $K \times V, \lambda \mapsto u + \lambda x$ is continuous. Choose an open set $A \subseteq V$ such that $u \in A \subseteq C$. Then $\{\lambda \in K \mid u + \lambda x \in A\}$ is open and contains 0. In particular, there is $\varepsilon \in K_{>0}$ such that $u + \varepsilon x \in A \subseteq C$. □

Example 7.2.5. Consider the $\mathbb{R}$-vector space $V := C([0,1], \mathbb{R})$ of all continuous real valued functions on the interval $[0,1] \subseteq \mathbb{R}$ together with the scalar product defined by

$$\langle f, g \rangle := \int_0^1 f(x)g(x)dx \qquad (f, g \in V).$$

By 7.2.2(c), this is a topological vector space. The constant function $u\colon [0,1] \to \mathbb{R},\ x \mapsto 1$ is a unit for the cone $C := C([0,1], \mathbb{R}_{\geq 0})$ of all functions nonnegative on $[0,1]$ by 7.1.19 (since $[0,1]$ is compact by 6.1.16). But u does not lie in C° since for every $\varepsilon > 0$ there is some $f \in V$ with $\|u - f\| = \sqrt{\int_0^1 (u(x) - f(x))^2 dx} < \varepsilon$ and $f \notin C$.

Remark 7.2.6. From Definition 7.2.1, it follows that for every topological K-vector space V the maps $V \to V,\ x \mapsto \lambda x + y$ $(\lambda \in K^\times, y \in V)$ are homeomorphisms [→ 5.2.2].

Lemma 7.2.7. Suppose V is a topological K-vector space and $\varphi\colon V \to \mathbb{R}$ is K-linear. Then the following are equivalent:

(a) φ is continuous.

(b) $\varphi^{-1}(\mathbb{R}_{>0})$ is open.

(c) $\varphi^{-1}(\mathbb{R}_{\geq 0})$ is closed.

Proof. (b) $\iff$ (c) follows from $\varphi^{-1}(\mathbb{R}_{\geq 0}) = -\varphi^{-1}(\mathbb{R}_{\leq 0}) = -(V \setminus \varphi^{-1}(\mathbb{R}_{>0}))$ since $V \to V$, $x \mapsto -x$ is a homeomorphism by 7.2.6.

(a) $\implies$ (b) is trivial.

(b) $\implies$ (a) WLOG $\varphi \neq 0$. WLOG choose $u \in V$ in such a way that $\varphi(u) = 1$ (otherwise scale φ). Suppose that (b) holds. Then the set $\varphi^{-1}(\mathbb{R}_{>a}) = au + \varphi^{-1}(\mathbb{R}_{>0})$ is open and hence $\varphi^{-1}(\mathbb{R}_{<-a}) = -\varphi^{-1}(\mathbb{R}_{>a})$ is open for all $a \in K$ [$\to$ 7.2.6]. So the set $\varphi^{-1}((a,b)_{\mathbb{R}}) = \varphi^{-1}(\mathbb{R}_{>a}) \cap \varphi^{-1}(\mathbb{R}_{<b})$ is open for all $a, b \in K$. Since every open subset of $\mathbb{R}$ is a union of intervals $(a,b)_{\mathbb{R}}$ with $a, b \in K$, the continuity of φ follows. □

Lemma 7.2.8. Let V be a topological K-vector space and $\varphi\colon V \to \mathbb{R}$ be K-linear map. Then φ is continuous if and only if $\varphi^{-1}(\mathbb{R}_{\geq 0})$ has an interior point.

Proof. WLOG $\varphi \neq 0$. If φ is continuous, then $\varphi^{-1}(\mathbb{R}_{>0})$ is open and because of $\varphi \neq 0$ nonempty. Conversely, let u be an interior point of $\varphi^{-1}(\mathbb{R}_{\geq 0})$. By 7.2.7, it is enough to show that $\varphi^{-1}(\mathbb{R}_{>0})$ is open. For this, consider $x \in \varphi^{-1}(\mathbb{R}_{>0})$. We have to show that there is an open set $A \subseteq V$ such that $x \in A \subseteq \varphi^{-1}(\mathbb{R}_{>0})$. Choose u in the interior of $\varphi^{-1}(\mathbb{R}_{\geq 0})$. Choose an open set $B \subseteq V$ with $u \in B \subseteq \varphi^{-1}(\mathbb{R}_{\geq 0})$. Choose $\lambda \in K_{>0}$ such that $\lambda\varphi(u) < \varphi(x)$. Then $A := x + \lambda(B - u)$ is open by 7.2.6, and we have $x = x + \lambda(u - u) \in A$ and

$$\varphi(A) = \varphi(x) + \lambda(\varphi(B) - \varphi(u)) \subseteq \varphi(x) + \mathbb{R}_{\geq 0} - \lambda\varphi(u) \subseteq \mathbb{R}_{>0}.$$

□

Example 7.2.9. Let $V := C([0,1], \mathbb{R})$ be the topological K-vector space from 7.2.5 and $x \in [0,1]$. Then $V \to \mathbb{R}$, $f \mapsto f(x)$ is not continuous.

Theorem 7.2.10 (Separation theorem for topological vector spaces)**.** *Let A and B be convex sets in the topological K-vector space V with $A^\circ \neq \emptyset \neq B$ and $A \cap B = \emptyset$. Then there is a continuous K-linear function $\varphi\colon V \to \mathbb{R}$ with $\varphi \neq 0$ and $\varphi(x) \leq \varphi(y)$ for all $x \in A$ and $y \in B$.*

Proof. Since A is convex, also $-A$ is convex and thus the Minkowski sum $B - A = B + (-A)$ [$\to$ 7.4.19] is also convex. By hypothesis, we have $0 \notin B - A \neq \emptyset$, for which reason there is according to 7.2.3 a proper cone $C \subseteq V$ such that $B - A \subseteq C$. Due to $A^\circ \neq \emptyset$ and $B \neq \emptyset$, 7.2.6 yields $(B - A)^\circ \neq \emptyset$ and thus $C^\circ \neq \emptyset$. Choose $u \in C^\circ$. By 7.2.4, u is a unit for C. By the isolation theorem 7.1.13, there exists a state φ of (V, C, u). Because of $\varphi(u) = 1$, we have $\varphi \neq 0$ and because of $\varphi(B - A) \subseteq \mathbb{R}_{\geq 0}$, we have $\varphi(x) \leq \varphi(y)$ for all $x \in A$ and $y \in B$. Finally, φ is continuous by 7.2.8 since u is an interior point of C and a fortiori of $\varphi^{-1}(\mathbb{R}_{\geq 0})$. □

Corollary 7.2.11. *Let A and B be nonempty convex sets in the topological K-vector space V satisfying $A \cap B = \emptyset$. Suppose A is open. Then there is a continuous K-linear function $\varphi\colon V \to \mathbb{R}$ and an $r \in \mathbb{R}$ such that $\varphi(x) < r \leq \varphi(y)$ for all $x \in A$ and $y \in B$.*

Proof. Choose by 7.2.10 a continuous K-linear function $\varphi\colon V \to \mathbb{R}$ with $\varphi \neq 0$ and $\varphi(x) \leq \varphi(y)$ for all $x \in A$ and $y \in B$. The set $\{\varphi(x) \mid x \in A\} \subseteq \mathbb{R}$ is nonempty because of $A \neq \emptyset$ and bounded from above because of $B \neq \emptyset$. It thus possesses a supremum $r \in \mathbb{R}$. We have $\varphi(x) \leq r \leq \varphi(y)$ for all $x \in A$ and $y \in B$. Let $x \in A$. It remains to show that $\varphi(x) < r$. For this purpose, choose $z \in V$ such that $\varphi(z) > 0$. The function $K \to V$, $\lambda \mapsto x + \lambda z$ is continuous and together with 0, a whole neighborhood of 0 lies in the preimage of A under this function. In particular, there is an $\varepsilon \in K_{>0}$ such that $x + \varepsilon z \in A$. Then $\varphi(x) < \varphi(x) + \varepsilon\varphi(z) = \varphi(x + \varepsilon z) \leq r$. □

Lemma 7.2.12. Let V be a topological K-vector space, $A \subseteq V$ be convex, $x \in A^\circ$, $y \in A$ and $\lambda \in K$ with $0 < \lambda \leq 1$. Then $\lambda x + (1-\lambda)y \in A^\circ$.

Proof. Choose an open neighborhood B of x with $B \subseteq A$. Setting $z := \lambda x + (1-\lambda)y$, $C := z + \lambda(B - x)$ is by 7.2.6 an open neighborhood of z. It is enough to show $C \subseteq A$. To this end, let $c \in C$. Because of $B = x + \frac{1}{\lambda}(C - z)$, we have then $b := x + \frac{1}{\lambda}(c - z) \in B \subseteq A$. Consequently, $c = \lambda(b - x) + z = \lambda b - \lambda x + \lambda x + (1-\lambda)y = \lambda b + (1-\lambda)y \in A$. □

Proposition 7.2.13. *Suppose V is a topological K-vector space and $A \subseteq V$ is convex. Then both A° and $\overline{A}$ are convex.*

Proof. It follows immediately from Lemma 7.2.12 that A° is convex. In order to show that $\overline{A}$ is convex, fix $x, y \in \overline{A}$ and $\lambda \in [0,1]_K$. To show: $z := \lambda x + (1-\lambda)y \in \overline{A}$. Let B be a neighborhood of z in V. To show: $B \cap A \neq \emptyset$. Since

$$V \times V \to V,\ (x', y') \mapsto \lambda x' + (1-\lambda)y'$$

is continuous, there are neighborhoods C of x and D of y in V such that

$$\lambda C + (1-\lambda)D \subseteq B.$$

Due to $x, y \in \overline{A}$, we find $x_0 \in C \cap A$ and $y_0 \in D \cap A$. Then

$$z_0 := \lambda x_0 + (1-\lambda)y_0 \in B \cap A.$$

□

Definition 7.2.14. Let V be a K-vector space and $A \subseteq V$ a set. Then A is called *balanced* if $\lambda x \in A$ for all $x \in A$ and $\lambda \in K$ with $|\lambda| \leq 1$.

Proposition 7.2.15. *Suppose V be a topological K-vector space and B is a neighborhood of 0 in V. Then there is a balanced open neighborhood A of 0 in V with $A \subseteq B$.*

Proof. WLOG B is open [→ 5.1.9]. Since the scalar multiplication is continuous by 7.2.1, there is an $\varepsilon \in K_{>0}$ and an open neighborhood C of 0 in V such that

$$\forall \lambda \in (-\varepsilon, \varepsilon)_K : \forall x \in C : \lambda x \in B.$$

By 7.2.6, each λC with $\lambda \in K^\times$ is open. Thus also $A := \bigcup_{\lambda \in (-\varepsilon,\varepsilon)_K \setminus \{0\}} \lambda C \subseteq B$ is open. Moreover, we have $0 \in A$ and A is obviously balanced. □

Exercise 7.2.16. In a Hausdorff space [→ 5.1.14], every compact subset [→ 5.1.21] is closed.

Definition 7.2.17. Let V be a K-vector space. We call a topology on V making V into a topological vector space [→ 7.2.1] a *vector space topology* on V.

Remark 7.2.18. Up to now the condition $\overline{\{0\}} = \{0\}$ from Definition 7.2.1 has been used nowhere. From now on, we will however need it. We will show that each finite-dimensional $\mathbb{R}$-vector space carries exactly one vector space topology which would be false without the condition $\overline{\{0\}} = \{0\}$ since otherwise the trivial topology [→ 5.1.2(e)] would also be a vector space topology.

Proposition 7.2.19. *Every topological K-vector space is a Hausdorff space.*

Proof. Let V be a topological K-vector space [→ 7.2.1] and let $x, y \in V$ with $x \neq y$. Set $z := x - y \neq 0$. by Definition 7.2.1, $\{0\}$ and thus by 7.2.6 also $\{z\}$ is closed. Hence $V \setminus \{z\}$ is an open neighborhood of 0. Since $V \times V \to V$, $(v, w) \mapsto v - w$ is continuous by 7.2.1, there is a neighborhood U of 0 such that $U - U \subseteq V \setminus \{z\}$. Then $(x + U) \cap (y + U) = \emptyset$ for otherwise there would be $u, v \in U$ with $x + u = y + v$ from which it would follow $z = x - y = v - u \in U - U$ ↯. □

Proposition 7.2.20. *Let V be a finite-dimensional $\mathbb{R}$-vector space. Then there is exactly one vector space topology* [→ 7.2.17] *on V.*

Proof. Choose a basis $v_1, \ldots, v_n$ of V. Then $f\colon \mathbb{R}^n \to V$, $x \mapsto \sum_{i=1}^n x_i v_i$ is a vector space isomorphism. With $\mathbb{R}^n$ [→ 7.2.2] also V possesses therefore a vector space topology. This shows existence. For uniqueness, endow now V with any vector space topology. We show that f is a homeomorphism. By 7.2.1, f is certainly continuous. It is enough to show that images of open sets under f are again open. For this purpose, it suffices to show that for all open balls in $\mathbb{R}^n$ the image of their center is an interior point of their image because if $A \subseteq \mathbb{R}^n$ is open then every point in $f(A)$ is the image of the center of an open ball contained in A. Due to 7.2.6, it suffices to consider the ball $B := \{x \in \mathbb{R}^n \mid \|x\| < 1\}$ around the origin of radius 1. In order to show that $0 \in (f(B))^\circ$, we take the sphere $S := \{x \in \mathbb{R}^n \mid \|x\| = 1\}$. By 6.1.16, S is compact and hence so is by 7.1.18 and 7.2.19 also $f(S)$. According to 7.2.16, $f(S)$ is thus closed in V. Hence $V \setminus f(S)$ is a neighborhood of 0 in V. By 7.2.15, there is a balanced open neighborhood A of 0 in V with $A \subseteq V \setminus f(S)$, i.e., $A \cap f(S) = \emptyset$. Since f is continuous, $f^{-1}(A)$ is an open neighborhood of 0 in $\mathbb{R}^n$. Due to the linearity of f, with A also $f^{-1}(A)$ is balanced according to Definition 7.2.14. Since $f^{-1}(A)$ is disjoint to S, it follows that $f^{-1}(A) \subseteq B$ and thus $A \subseteq f(B)$. Hence $0 \in (f(B))^\circ$ as desired. □

7.3 Convex sets in locally convex vector spaces

Definition 7.3.1. A *locally convex K-vector space* is a topological K-vector space [→ 7.2.1] in which for every $x \in V$ each neighborhood of x contains a convex neighborhood of x.

Remark 7.3.2. Because of 7.2.6, one can restrict oneself in 7.3.1 to $x = 0$.

Example 7.3.3. [→ 7.2.2]

(a) If I is a set, then K^I is a locally convex K-vector space.

(b) If a K-vector space V is endowed with the initial topology [→ 5.1.4] with respect to a family $(f_i)_{i\in I}$ of K-linear functions $f_i\colon V \to \mathbb{R}$ in such a way that to each $x \in V \setminus \{0\}$ there is some $i \in I$ with $f_i(x) \neq 0$, then V is a locally convex K-vector space.

(c) Every normed $\mathbb{R}$-vector space V, in particular every $\mathbb{R}$-vector space with scalar product, is a locally convex $\mathbb{R}$-vector space since

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

for all $\varepsilon > 0$ and $x, y \in V$ satisfying $\|x\|, \|y\| < \varepsilon$ ("balls are convex").

Lemma 7.3.4. Suppose V is a topological K-vector space, $A \subseteq V$ is closed and $C \subseteq V$ is compact. Then $A + C$ is closed.

Proof. Let $x \in V \setminus (A + C)$. We have to show that there is a neighborhood U of the origin satisfying $(x + U) \cap (A + C) = \emptyset$.

Claim: For each $y \in C$, there exists a neighborhood U_y of the origin such that

$$(x + U_y) \cap (y + U_y + A) = \emptyset.$$

Explanation. Let $y \in C$. Then $V \times V \to V$, $(x', y') \mapsto x - y + x' - y'$ is continuous and $(0, 0)$ lies in the preimage of the open set $V \setminus A$ since $x - y \notin A$ (otherwise we would have $x \in A + y \subseteq A + C$). Hence there is a neighborhood U_y with

$$x - y + U_y - U_y \subseteq V \setminus A,$$

i.e., $(x + U_y - y - U_y) \cap A = \emptyset$.
By compactness of C, there is a finite subset $D \subseteq C$ such that $C \subseteq \bigcup_{y\in D}(y + U_y)$. Now $U := \bigcap_{y\in D} U_y$ is a neighborhood of the origin. In order to show that

$$(x + U) \cap (A + C) = \emptyset,$$

it is enough to prove that $(x + U) \cap (A + y + U_y) = \emptyset$ for all $y \in D$. For this purpose, it suffices to show that $(x + U_y) \cap (y + U_y + A) = \emptyset$ for all $y \in D$. But this holds even for all $y \in C$ by the above claim. □

Theorem 7.3.5 (Separation theorem for locally convex vector spaces). [→ 7.2.10, 7.2.11] *Let A and C be nonempty convex sets in the locally convex K-vector space V with $A \cap C = \emptyset$. Let A be closed and C be compact. Then there is a continuous K-linear function $\varphi\colon V \to \mathbb{R}$ and $r, s \in \mathbb{R}$ with $\varphi(x) \leq r < s \leq \varphi(y)$ for all $x \in A$ and $y \in C$.*

Proof. $B := C - A$ is by 7.3.4 closed and by hypothesis we have $0 \notin B$. Since V is locally convex, there is in view of 7.2.13 a convex open set $D \subseteq V$ with $0 \in D$ and $D \cap B = \emptyset$. Since B is also convex, there is by Corollary 7.2.11 a continuous K-linear function $\varphi\colon V \to \mathbb{R}$ and an $\varepsilon \in \mathbb{R}$ such that $\varphi(x) < \varepsilon \leq \varphi(y)$ for all $x \in D$ and $y \in B$. In particular, $\varepsilon > \varphi(0) = 0$ and $\varphi(x) + \varepsilon \leq \varphi(y)$ for all $x \in A$ and $y \in C$. Because of $A \neq \emptyset \neq C$, $r := \sup\{\varphi(x) \mid x \in A\} \in \mathbb{R}$ and $s := \inf\{\varphi(y) \mid y \in C\} \in \mathbb{R}$ exist. Moreover, we have $r + \varepsilon \leq s$, i.e., $r < s$. □

Definition 7.3.6. Let V be a K-vector space and $A \subseteq V$ be convex. Then a convex set $F \subseteq A$ is called a *face* of A if for all $x, y \in A$ with $\frac{x+y}{2} \in F$, we have also $x, y \in F$.

Proposition 7.3.7. *Suppose V is a K-vector space, $A \subseteq V$ is convex and $x \in A$. Then x is an extreme point of A* [→ 2.4.1] *if and only if $\{x\}$ is a face of A.*

Proof.

$$
\begin{aligned}
& x \text{ is an extreme point of } A \\
\overset{2.4.1}{\iff} & \nexists y, z \in A : \left(y \neq z \,\&\, x = \frac{y+z}{2} \right) \\
\iff & \forall y, z \in A : \left(x = \frac{y+z}{2} \implies y = z \right) \\
\iff & \forall y, z \in A : \left(x = \frac{y+z}{2} \implies y = z = x \right) \\
\iff & \forall y, z \in A : \left(\frac{y+z}{2} \in \{x\} \implies y, z \in \{x\} \right)
\end{aligned}
$$

□

Proposition 7.3.8. [→ 2.4.2] *Suppose V is a K-vector space, $A \subseteq V$ is convex, $F \subseteq A$ is convex and $\lambda \in (0,1)_K$. Then the following are equivalent:*

(a) *F is a face of A*

(b) $\forall x, y \in A : (\lambda x + (1-\lambda)y \in F \implies x, y \in F)$

Proof. (b) $\implies$ (a) is an easy exercise.

(a) $\implies$ (b) Assume that F is a face of A but there are $x, y \in A$ such that

$$\lambda x + (1-\lambda)y \in F$$

and WLOG (otherwise permute x and y and replace λ by $1-\lambda$) $x \notin F$. If $\lambda < \frac{1}{2}$, one then can replace (x,λ) by (x',λ') where $x' := \frac{x+y}{2}$ and $\lambda' := 2\lambda \in (0,1)_K$ because we then have $x' \in A \setminus F$ (since A is convex and F is a face of A), $\lambda' \in (0,1)_K$ and

$$\lambda' x' + (1-\lambda')y = 2\lambda\frac{x+y}{2} + (1-2\lambda)y = \lambda x + (1-\lambda)y \in F.$$

By iterating this in case of need finitely many times, one can suppose $\lambda \geq \frac{1}{2}$. Then

$$z := x + 2((\lambda x + (1-\lambda)y) - x) = (2\lambda - 1)x + 2(1-\lambda)y \in A$$

since $2\lambda - 1 \geq 0$, $2(1-\lambda) \geq 0$ and $(2\lambda-1) + 2(1-\lambda) = 1$. Now

$$\frac{x+z}{2} = x + (\lambda x + (1-\lambda)y) - x = \lambda x + (1-\lambda)y \in F$$

and thus $x, z \in F$ since F is a face of A $\lightning$. □

Example 7.3.9. (a) If V is a K-vector space and $A \subseteq V$ is convex, then both $\emptyset$ and A are faces of A. We call these the *trivial* faces of A.

(b) The faces of $[0,1]^2 \subseteq \mathbb{R}^2$ are $\emptyset$, $\{(0,0)\}$, $\{(0,1)\}$, $\{(1,0)\}$, $\{(1,1)\}$, $\{0\} \times [0,1]$, $\{1\} \times [0,1]$, $[0,1] \times \{0\}$, $[0,1] \times \{1\}$, $[0,1]^2$.

(c) The faces of $B := \{x \in \mathbb{R}^2 \mid \|x\| \leq 1\}$ are $\emptyset$, $\{x\}$ ($x \in B \setminus B^\circ$) and B.

(d) $\{x \in \mathbb{R}^2 \mid \|x\| < 1\}$ has only the trivial faces.

Definition and Proposition 7.3.10. *Let V be a K-vector space and suppose $A \subseteq V$ is convex. We call F an* exposed face *of A if there is a K-linear function $\varphi\colon V \to \mathbb{R}$ such that*

$$F = \{x \in A \mid \forall y \in A : \varphi(x) \leq \varphi(y)\}.$$

Every exposed face of A is a face of A.

Proof. Let F be an exposed face of A. To show: F is a face of A. It is easy to show that F is convex. Choose a K-linear $\varphi\colon V \to \mathbb{R}$ such that $F = \{x \in A \mid \forall y \in A : \varphi(x) \leq \varphi(y)\}$. Let $x, y \in A$ such that $\frac{x+y}{2} \in F$. To show: $x, y \in F$. It is obviously enough to show that $\varphi(x) = \varphi\left(\frac{x+y}{2}\right) = \varphi(y)$. We have that

$$\varphi(x) + \varphi(y) = \varphi\left(\frac{x+y}{2}\right) + \varphi\left(\frac{x+y}{2}\right) \underset{\frac{x+y}{2} \in F}{\overset{x,y \in A}{\leq}} \varphi(x) + \varphi(y)$$

where the inequality would be strict if one of $\varphi(x)$ and $\varphi(y)$ were different from $\varphi\left(\frac{x+y}{2}\right)$. □

Example 7.3.11. [→ 7.3.9]

(a) If V is a K-vector space and $A \subseteq V$ is convex, then A is an exposed face of A while $\emptyset$ might be exposed [→ 7.3.9(d)] or non-exposed [7.3.9(c)].

(b) All faces of $[0,1]^2 \subseteq \mathbb{R}^2$ are exposed except $\emptyset$.

(c) All faces of $\{x \in \mathbb{R}^2 \mid \|x\| \leq 1\}$ are exposed except $\emptyset$.

(d) All faces of $\{x \in \mathbb{R}^2 \mid \|x\| < 1\}$ are exposed.

(e) $((-\infty, 0] \times [0, \infty)) \cup \{(x,y) \in \mathbb{R}^2_{\geq 0} \mid y \geq x^2\}$ has exactly one nonexposed face, namely $\{0\}$.

Proposition 7.3.12. *Suppose V is a K-vector space, $A \subseteq V$ is convex, F is a face of A and $G \subseteq F$. Then the following holds:*

$$G \text{ is a face of } F \iff G \text{ is a face of } A.$$

Proof. "$\Longrightarrow$" Let G be a face of F and let $x, y \in A$ with $\frac{x+y}{2} \in G$. To show: $x, y \in G$. Because of $\frac{x+y}{2} \in G \subseteq F$, we have $x, y \in F$. Since G is a face of F, it follows that $x, y \in G$.

"$\Longleftarrow$" Let G be a face of A and let $x, y \in F$ with $\frac{x+y}{2} \in G$. Because of $x, y \in F \subseteq A$, we then have $x, y \in G$. □

Remark 7.3.13. Every intersection of faces of a convex set in a K-vector space V is obviously again a face of this convex set.

Lemma 7.3.14. Let $C \neq \emptyset$ be a compact convex subset of a locally convex K-vector space V. Then C possesses an extreme point.

Proof. Every intersection of a nonempty chain of closed nonempty faces of C is again a closed nonempty face of C. Indeed, if the intersection were empty, then a finite subintersection would be empty by the compactness of C [→ 5.1.14] which is impossible since we dealt with a chain. By Zorn's lemma there is thus a minimal closed nonempty face F of C. Being a closed subset of a compact set, F is compact itself [→ 5.1.21]. By 7.3.7, it suffices to show that $\#F = 1$. Due to $F \neq \emptyset$, it suffices to exclude $\#F \geq 2$. Assume $x, y \in F$ such that $x \neq y$. By 7.3.5, there is a continuous K-linear function $\varphi \colon V \to \mathbb{R}$ such that $\varphi(x) < \varphi(y)$. Then

$$G := \{v \in F \mid \forall w \in F : \varphi(v) \leq \varphi(w)\}$$

is nonempty by 7.1.19 because F is compact and nonempty and φ is continuous. According to 7.3.10, G is an (exposed) face of F. Hence G is a face of C by 7.3.12. From the continuity of φ, we deduce that

$$G = F \cap \bigcap_{w \in F} \varphi^{-1}((-\infty, \varphi(w)])$$

is closed. Moreover, $y \notin G$ since $\varphi(y) \not\leq \varphi(x)$. Therefore G is a closed nonempty face of C that is properly contained in F, contradicting the minimality of F. □

Notation 7.3.15. Let A be a convex set in a K-vector space V. Then we write

$$\operatorname{extr} A$$

for the set of extreme points of A.

Theorem 7.3.16. [→ 7.1.19] *Suppose C is a nonempty compact convex subset of a locally convex K-vector space V and $\varphi\colon V \to \mathbb{R}$ is a continuous K-linear function. Then φ attains on C a minimum and a maximum in an extreme point of C. In other words, there are $x, y \in \operatorname{extr} C$ such that*

$$\varphi(x) \leq \varphi(z) \leq \varphi(y)$$

for all $z \in C$.

Proof. Since one could replace φ by $-\varphi$, we show only the existence of $x \in \operatorname{extr} C$ such that $\varphi(x) \leq \varphi(z)$ for all $z \in C$. By 7.1.19, there is $y \in C$ such that $\varphi(y) \leq \varphi(z)$ for all $z \in C$, i.e., the exposed face [→ 7.3.10]

$$F := \{y \in C \mid \forall z \in C : \varphi(y) \leq \varphi(z)\}$$

of C is nonempty. Since φ is continuous,

$$F = C \cap \bigcap_{z \in C} \varphi^{-1}((-\infty, \varphi(z)]_{\mathbb{R}})$$

is a closed subset of the compact set C and hence compact itself. By Lemma 7.3.14, F possesses an extreme point x which is by 7.3.12 and 7.3.7 also an extreme point of C. □

Corollary 7.3.17 (Krein–Milman theorem)**.** *Suppose C is a compact convex subset of a locally convex K-vector space V. Then C is the closure of the convex hull of the set of its extreme points, i.e.,*

$$C = \overline{\operatorname{conv}(\operatorname{extr} C)}.$$

Proof. "$\supseteq$" From $\operatorname{extr} C \subseteq C$ and the convexity of C, we get $\operatorname{conv}(\operatorname{extr} C) \subseteq C$. Being a compact subset of a Hausdorff space [→ 7.2.19], C is closed [→ 7.2.16] which entails even $\overline{\operatorname{conv}(\operatorname{extr} C)} \subseteq C$.

"$\subseteq$" WLOG $C \neq \emptyset$. $A := \overline{\operatorname{conv}(\operatorname{extr} C)}$ is closed, nonempty by Lemma 7.3.14 and convex by 7.2.13. We show $V \setminus A \subseteq V \setminus C$. Let $x \in V \setminus A$. By the separation theorem for locally convex vector spaces 7.3.5, there is a continuous K-linear function $\varphi\colon V \to \mathbb{R}$ such that $\varphi(y) < \varphi(x)$ for all $y \in A$. By 7.3.16, there is $y \in \operatorname{extr} C \subseteq A$ satisfying $\varphi(z) \leq \varphi(y)$ for all $z \in C$. It follows that $\varphi(z) \leq \varphi(y) < \varphi(x)$ for all $z \in C$. Therefore $x \notin C$. □

Definition 7.3.18. Let V be a K-vector space, $C \subseteq V$ and $u \in V$. We call an extreme point [→ 2.4.1] of the state space $S(V, C, u)$ [→ 7.1.9, 7.1.16] a *pure state* of (V, C, u).

Theorem 7.3.19 (Strengthening of 7.1.20)**.** *Suppose u is a unit for the cone C in the K-vector space V and let $x \in V$. Then the following are equivalent:*

(a) $\forall \varphi \in \operatorname{extr} S(V,C,u) : \varphi(x) > 0$

(b) $\forall \varphi \in S(V,C,u) : \varphi(x) > 0$

(c) $\exists N \in \mathbb{N} : x \in \frac{1}{N}u + C$

(d) *x is a unit for C.*

Proof. (b) $\iff$ (c) $\iff$ (d) is 7.1.20.

(b) $\implies$ (a) is trivial.

(a) $\implies$ (b) WLOG $S(V,C,u) \neq \emptyset$. It suffices to show that the function

$$S(V,C,u) \to \mathbb{R},\ \varphi \mapsto \varphi(x)$$

attains a minimum in an extreme point of $S(V,C,u)$. But this follows from 7.3.16 because $S(V,C,u)$ is a nonempty compact [→ 7.1.17] convex [→ 7.1.16] subset of the locally convex [→ 7.2.2(a)] $\mathbb{R}$-vector space $\mathbb{R}^V$ and

$$\mathbb{R}^V \to \mathbb{R},\ \varphi \mapsto \varphi(x)$$

is continuous [→ 5.1.5(b)]. □

Corollary 7.3.20 (Strengthening of 7.1.15)**.** *Suppose u is a unit for the cone C in the K-vector space V and let $x \in V$. If $\varphi(x) > 0$ for all pure states φ of (V,C,u), then $x \in C$.*

7.4 Convex sets in finite-dimensional vector spaces

Lemma 7.4.1. Let C be a cone in a finite-dimensional K-vector space V. Then $U := C - C$ is a subspace of V and C possesses *in* U a unit [→ 7.1.4].

Proof. On the basis of Definition 7.1.1, it is easy to see that U is a subspace of V. Choose a basis $u_1, \ldots, u_m$ of U and write $u_i = v_i - w_i$ with $v_i, w_i \in C$ for $i \in \{1, \ldots, m\}$. We show that $u := \sum_{i=1}^m v_i + \sum_{i=1}^m w_i \in C$ is a unit for C *in* U. For this purpose, fix $v \in U$. To show: $\exists N \in \mathbb{N} : Nu + v \in C$. Write $v = \sum_{i=1}^m \lambda_i u_i$ with $\lambda_i \in K$ for $i \in \{1, \ldots, m\}$. Choose $N \in \mathbb{N}$ with $N \geq |\lambda_i|$ for $i \in \{1, \ldots, m\}$. Then

$$Nu + v = \sum_{i=1}^m \underbrace{(N + \lambda_i)}_{\geq 0} v_i + \sum_{i=1}^m \underbrace{(N - \lambda_i)}_{\geq 0} w_i \in C.$$

□

Theorem 7.4.2 (Finite-dimensional isolation theorem)**.** [→ 7.1.13] *Let C be a proper cone in the finite-dimensional K-vector space V. Then there is a K-linear function $\varphi\colon V \to \mathbb{R}$ with $\varphi \neq 0$ and $\varphi(C) \subseteq \mathbb{R}_{\geq 0}$.*

Proof. $U := C - C$ is by 7.4.1 a subspace of V.

Case 1: $C = U$

Then U is a proper subspace of V and by linear algebra it is easy to see that there is some $\varphi \in V^* \setminus \{0\}$ such that $\varphi(U) = \{0\}$.

Case 2: $C \neq U$

By 7.4.1, there exists a unit u for C in U. The isolation theorem 7.1.13 provides us with some $\varphi_0 \in S(U, C, u)$. Extend φ_0 by linear algebra to a K-linear function $\varphi\colon V \to \mathbb{R}$. □

Remark 7.4.3. Example 7.1.10 shows that one cannot omit the hypothesis $\dim V < \infty$ in 7.4.1 and 7.4.2.

Theorem 7.4.4 (Separation theorem for finite-dimensional vector spaces)**.** [→ 7.2.10] *Let A and B be convex sets in the finite-dimensional K-vector space V such that $A \neq \emptyset \neq B$ and $A \cap B = \emptyset$. Then there is a K-linear function $\varphi\colon V \to \mathbb{R}$ such that $\varphi \neq 0$ and $\varphi(x) \leq \varphi(y)$ for all $x \in A$ and $y \in B$.*

Proof. Completely analogous to the proof of 7.2.10. □

Definition 7.4.5. [→ 2.4.1] Let V be a K-vector space and $A \subseteq V$. Then A is called an *affine subspace of V* if $\forall x, y \in A : \forall \lambda \in K : \lambda x + (1-\lambda)y \in A$. The smallest affine subspace of V containing A is obviously

$$\operatorname{aff} A := \left\{ \sum_{i=1}^{m} \lambda_i x_i \mid m \in \mathbb{N}, \lambda_i \in K, x_i \in A, \sum_{i=1}^{m} \lambda_i = 1 \right\},$$

called the affine subspace generated by A or the *affine hull* of A.

Definition and Proposition 7.4.6. *Let V be a K-vector space. Then for each $A \subseteq V$, the following are equivalent:*

(a) *A is a nonempty affine subspace of V.*

(b) *There is an $x \in V$ and a subspace U of V such that $A = x + U$.*

If these conditions are met, then U as in (b) *is uniquely determined and is called the* direction *of A. Then* $\dim A := \dim U \in \mathbb{N}_0 \cup \{\infty\}$ *is the* dimension *of A. We set* $\dim \emptyset := -1$.

Proof. (b) $\Longrightarrow$ (a) is easy.

(a) $\Longrightarrow$ (b) Suppose that (a) holds. Choose $x \in A$. Set $U := A - x$. To show: $U + U \subseteq U$ and $KU \subseteq U$. Let $u, v \in U$ and $\lambda \in K$. To show: $u + v \in U$ and $\lambda u \in U$. Choose $a, b \in A$ such that $u = a - x$ and $v = b - x$. Then $u + v = (1a + 1b + (-1)x) - x \in (\operatorname{aff} A) - x = A - x = U$ and $\lambda u = (\lambda a + (1-\lambda)x) - x \in (\operatorname{aff} A) - x = A - x = U$.

Uniqueness claim Whenever $x, y \in V$ and U and W are subspaces of V satisfying $x + U = y + W$, then $x - y \in W$ and thus $U = (y - x) + W = W$. □

Definition 7.4.7. Let V be a K-vector space and $A \subseteq V$ be convex. Then

$$\dim A := \dim \operatorname{aff} A \in \{-1\} \cup \mathbb{N}_0 \cup \{\infty\}$$

is the *dimension* of A.

Proposition 7.4.8. [→ 7.3.8] *Suppose that V is a K-vector space, $A \subseteq V$ is convex and F is a face of A. Let $m \in \mathbb{N}$, $x_1, \ldots, x_m \in A$ and $\lambda_1, \ldots, \lambda_m \in K_{>0}$ such that $\sum_{i=1}^m \lambda_i = 1$ and $\sum_{i=1}^m \lambda_i x_i \in F$. Then $x_1, \ldots, x_m \in F$.*

Proof. WLOG $m \geq 2$. Let $i \in \{1, \ldots, m\}$. To show: $x_i \in F$. WLOG $i = 1$. We have $0 < \lambda_1 < 1$ and $y := \sum_{i=2}^m \frac{\lambda_i}{1-\lambda_1} x_i \in A$ since $\sum_{i=2}^m \frac{\lambda_i}{1-\lambda_1} = \frac{1-\lambda_1}{1-\lambda_1} = 1$. From $\sum_{i=1}^m \lambda_i x_i = \lambda_1 x_1 + (1-\lambda_1) y$ it follows thus by 7.3.8 that $x_1, y \in F$. □

Proposition 7.4.9. *Suppose V is a K-vector space, $A \subseteq V$ is convex and F is a face of A. Then $F = (\operatorname{aff} F) \cap A$.*

Proof. "$\subseteq$" is trivial.

"$\supseteq$" Let $x \in (\operatorname{aff} F) \cap A$. To show: $x \in F$. Write $x = \sum_{i=1}^m \lambda_i y_i - \sum_{j=1}^n \mu_j z_j$ with $m, n \in \mathbb{N}_0$, $\lambda_i, \mu_j \in K_{>0}$, $y_i, z_j \in F$ and $\sum_{i=1}^m \lambda_i - \sum_{j=1}^n \mu_j = 1$. Setting $\lambda := \sum_{i=1}^m \lambda_i$ and $\mu := \sum_{j=1}^n \mu_j$, it follows that $\frac{1}{1+\mu} x + \sum_{j=1}^n \frac{\mu_j}{1+\mu} z_j = \sum_{i=1}^m \frac{\lambda_i}{\lambda} y_i \in F$ and thus $x \in F$ by 7.4.8. □

Proposition 7.4.10. *Let V be a finite-dimensional K-vector space.*

(a) *If A and B are affine subspaces of V with $A \subseteq B$, then*

$$A = B \iff \dim A = \dim B.$$

(b) *If F and G are faces of the convex set $A \subseteq V$ with $F \subseteq G$, then*

$$F = G \iff \dim F = \dim G.$$

Proof. (a) follows from 7.4.6 by linear algebra and (b) follows hereof by 7.4.7 and 7.4.9. □

Remark 7.4.11. Suppose V is a topological K-vector space, K' is a subfield of K and V' a K'-vector subspace of the K'-vector space V. Then V induces on V' a vector space topology. This is easy to see since $V \times V$ induces on $V' \times V'$ the product topology of the induced topologies and $K \times V$ induces on $K' \times V'$ the product topology of the induced topologies.

Definition and Proposition 7.4.12. *Let A be a convex set in the topological K-vector space V. The interior of A in the topological space* aff *A (endowed with the topology induced by V) is called the* relative interior *of A, denoted by* relint *A. This is a convex set.*

Proof. WLOG $A \neq \emptyset$. Write $\operatorname{aff} A = x + U$ for some $x \in V$ and some subspace U of V [→ 7.4.6]. WLOG $x = 0$ [→ 7.2.6]. WLOG $U = V$ [→ 7.4.11]. Then $\operatorname{relint} A = A^\circ$ is convex by 7.2.13. □

Remark 7.4.13. Let V be a finite-dimensional K-vector space. Choose a basis $v_1, \dots, v_n$ of V. Then $f\colon K^n \to V,\ x \mapsto \sum_{i=1}^n x_i v_i$ is a vector space isomorphism that is continuous with respect to every vector space topology of V [→ 7.2.17] and that is a homeomorphism with respect to exactly one vector space topology of V [→ 7.2.2(a)]. Consequently, there is a finest [→ 5.1.2(c)] vector space topology on V (cf. also 7.2.20). With respect to this topology on V, we have for all $A \subseteq V$ that

$$A \text{ open in } V \iff f^{-1}(A) \text{ open in } K^n,$$

independently of the choice of the basis $v_1, \dots, v_n$. It is easy to see that K^n carries the initial topology with respect to all $\left\{\begin{smallmatrix}\text{linear forms on } K^n\\ K\text{-linear functions } K^n \to \mathbb{R}\end{smallmatrix}\right\}$. The finest vector space topology on V is therefore also the initial topology [→ 5.1.4] with respect to all $\left\{\begin{smallmatrix}\text{linear forms on } V\\ K\text{-linear functions } V \to \mathbb{R}\end{smallmatrix}\right\}$. If U is a subspace of V, then the finest vector space topology of V induces on U again the finest vector space topology because one can extend every linear form on U to one on V.

Example 7.4.14. Since $\mathbb{R}$ is a topological $\mathbb{R}$-vector space and thus also a topological $\mathbb{Q}$-vector space, also $\mathbb{Q} + \sqrt{2}\mathbb{Q} \subseteq \mathbb{R}$ is a topological $\mathbb{Q}$-vector space with respect to the induced topology [→ 7.4.11]. One sees easily that

$$\mathbb{Q} + \sqrt{2}\mathbb{Q} \to \mathbb{Q},\ x + \sqrt{2}y \mapsto x \qquad (x, y \in \mathbb{Q})$$

is not continuous.

Lemma 7.4.15. Let $A \subseteq K^n$ be convex. Then $A^\circ = \emptyset \implies \operatorname{aff} A \neq K^n$.

Proof. Suppose that $\operatorname{aff} A = K^n$. We show that $A^\circ \neq \emptyset$. Denote by $e_1, \dots, e_n$ the standard basis of K^n and set $e_0 := 0 \in K^n$. Write $e_i = \sum_{j=1}^m \lambda_{ij} x_{ij}$ with $m \in \mathbb{N}$, $\lambda_{ij} \in K$, $x_{ij} \in A$ and $\sum_{j=1}^m \lambda_{ij} = 1$ for $i \in \{0, \dots, n\}$. We show that

$$x := \sum_{i=0}^{n} \sum_{j=1}^{m} \frac{1}{m(n+1)} x_{ij} \in A^\circ.$$

Since A is convex, we have $x \in A$ and it suffices to show that for each $i \in \{1, \dots, n\}$, there is an $\varepsilon > 0$ such that $x \pm \varepsilon e_i \in A$ (cf. also 7.1.8). For this purpose, fix $i \in \{1, \dots, n\}$. From $e_i = e_i - 0 = e_i - e_0 = \sum_{j=1}^m \lambda_{ij} x_{ij} + \sum_{j=1}^m (-\lambda_{0j}) x_{0j}$ and $\sum_{j=1}^m \lambda_{ij} - \sum_{j=1}^m \lambda_{0j} = 1 - 1 = 0$, the existence of such an $\varepsilon > 0$ easily follows. □

Theorem 7.4.16. *Suppose V is a finite-dimensional topological K-vector space that is equipped with the finest vector space topology* [→ 7.4.13] *and $A \subseteq V$ is convex. Then $A \subseteq \overline{\operatorname{relint} A}$.*

Proof. WLOG $A \neq \emptyset$. Write $\operatorname{aff} A = x + U$ for some $x \in V$ and some subspace U of V. Obviously, $\operatorname{aff}(A - x) \overset{7.4.5}{=} (\operatorname{aff} A) - x = U$, $\operatorname{relint}(A - x) \overset{7.2.6}{=} (\operatorname{relint} A) - x$ and $\overline{\operatorname{relint}(A - x)} = \overline{\operatorname{relint} A} - x$. Replacing A by $A - x$, we can thus suppose that $\operatorname{aff} A = U$.

Using the last remark from 7.4.13, we can therefore suppose that $\operatorname{aff} A = V$ (otherwise replace V by U). According to 7.4.13, we can reduce to the case where $V = K^n$ (with the product topology). We have to show $A \subseteq \overline{A^\circ}$. Choose $y \in A^\circ$ with Lemma 7.4.15. Let $x \in A$. To show: $x \in \overline{A^\circ}$. By 7.2.12, we have $(1-\lambda)x + \lambda y \in A^\circ$ for all $\lambda \in (0,1]_K$. Obviously, we have $x \in \overline{\{(1-\lambda)x + \lambda y \mid \lambda \in (0,1]_K\}} \subseteq \overline{A^\circ}$. □

Theorem 7.4.17. *Let V be a finite-dimensional K-vector space that is equipped with the finest vector space topology* [→ 7.4.13]. *Let $A \subseteq V$ be convex and $x \in A \setminus \operatorname{relint} A$. Then there is an exposed face F of A satisfying* $\dim F < \dim A$ *and $x \in F$.*

Proof. Similarly to the proof of 7.4.16, we reduce again to the case $\operatorname{aff} A = V$. Note that A° is convex [→ 7.4.12] and nonempty [→ 7.4.16]. The separation theorem 7.4.4 yields a K-linear function $\varphi\colon V \to \mathbb{R}$ with $\varphi \neq 0$ and $\varphi(x) \leq \varphi(y)$ for all $y \in A^\circ$. Since φ is continuous [→ 7.4.13], the set $\varphi^{-1}([\varphi(x), \infty)_{\mathbb{R}})$ is closed and contains with A° also $\overline{A^\circ}$ and hence by 7.4.16 also A, i.e., $\varphi(x) \leq \varphi(y)$ for all $y \in A$. In other words, x is an element of the exposed face [→ 7.3.10] $F := \{z \in A \mid \forall y \in A : \varphi(z) \leq \varphi(y)\}$ of A. By 7.4.10, it is enough to show $F \neq A$. If we had $F = A$, we would have $\varphi|_A = \varphi(x)$ and hence by linearity of φ also $\varphi = \varphi|_{\operatorname{aff} A} = \varphi(x)$, i.e., $\varphi = 0$ ↯. □

Remark 7.4.18. If we use topological notions in a finite-dimensional $\mathbb{R}$-vector space V, then we tacitly furnish V with its unique vector space topology [→ 7.2.20] which is the initial topology with respect to the family of all linear forms on V [→ 7.4.13].

Theorem 7.4.19 (Minkowski's theorem)**.** [→ 2.4.4, 7.3.17] *Let V be a finite-dimensional $\mathbb{R}$-vector space. Let $A \subseteq V$ be convex and compact. Then*

$$A = \operatorname{conv}(\operatorname{extr} A).$$

Proof. Since A is closed [→ 7.2.16], all faces of A are also closed [→ 7.4.9, 7.4.6, 7.2.6] and therefore compact [→ 5.1.21]. By induction, we can thus assume that $F = \operatorname{conv}(\operatorname{extr} F)$ for all faces F of A that satisfy $\dim F < \dim A$.

"$\supseteq$" is trivial.

"$\subseteq$" Let $x \in A$. To show: $x \in \operatorname{conv}(\operatorname{extr} A)$. WLOG $x \notin \operatorname{extr} A$. Choose $y, z \in A$ with $y \neq z$ and $x \in \operatorname{conv}\{y, z\}$. Because of the assumptions on A, WLOG $y, z \in A \setminus \operatorname{relint} A$. By 7.4.17, there are (exposed) faces F and G of A such that $\dim F < \dim A$, $\dim G < \dim A$, $y \in F$ and $z \in G$. From 7.3.7 and 7.3.12, we get $\operatorname{extr} F \subseteq \operatorname{extr} A$ and $\operatorname{extr} G \subseteq \operatorname{extr} A$. Consequently, $y \in F = \operatorname{conv}(\operatorname{extr} F) \subseteq \operatorname{conv}(\operatorname{extr} A)$ and $z \in G = \operatorname{conv}(\operatorname{extr} G) \subseteq \operatorname{conv}(\operatorname{extr} A)$ where the equalities follow from the induction hypothesis. Finally, $x \in \operatorname{conv}\{y, z\} \subseteq \operatorname{conv}(\operatorname{extr} A)$. □

Theorem 7.4.20. *Let $(K, \leq)$ be an arbitrary ordered field, let V be a K-vector space with $n := \dim V < \infty$. Suppose that $E \subseteq V$ is a finite set that generates V and $x \in V$. Then exactly one of the following conditions occurs:*

(a) *E contains a basis of V that generates a cone containing x.*

(b) *There is some* $\ell \in V^*$ *with* $\ell(E) \subseteq K_{\geq 0}$ *and* $\ell(x) < 0$ *and a linear independent set* $F \subseteq E \cap \ker \ell$ *with* $\#F = n - 1$.

Proof. It is easy to see that (a) and (b) cannot occur both at the same time. Indeed, from (a) it follows that $x \in \sum_{v \in E} K_{\geq 0} v$ which is not compatible with (b) because if $\ell \in V^*$ with $\ell(E) \subseteq K_{\geq 0}$, then $\ell(x) \in \ell\left(\sum_{v \in E} K_{\geq 0} v\right) \subseteq K_{\geq 0}$.

We choose an order $\leq$ on E [$\rightarrow$ 1.1.1] and a basis $B \subseteq E$ of V. We show that the following algorithm always terminates:

(1) Write $x = \sum_{v \in B} \lambda_v v$ with $\lambda_v \in K$ for all $v \in B$.

(2) If $\lambda_v \geq 0$ for all $v \in B$, then stop since (a) occurs.

(3) $u := \min\{v \in B \mid \lambda_v < 0\}$

(4) Define $\ell \in V^*$ by $\ell(u) = 1$ and $\ell(v) = 0$ for all $v \in B \setminus \{u\}$ (so that $\ell(x) = \lambda_u < 0$).

(5) If $\ell(E) \subseteq K_{\geq 0}$, then stop since (b) occurs.

(6) $w := \min\{v \in E \mid \ell(v) < 0\}$

(7) Replace B by the new basis $(B \setminus \{u\}) \cup \{w\}$ and go to (1).

Observe first of all that in Step (7) the set $(B \setminus \{u\}) \cup \{w\}$ is again a basis since B is one. Indeed, w does not lie in the subspace generated by $B \setminus \{u\}$ since ℓ vanishes according to its choice in (4) on this subspace while it does not vanish on w by the choice of w in (6).

To show that this algorithm terminates, we assume that this is not the case. Let then denote by (B_k, u_k, w_k, ℓ_k) the value of (B, u, w, ℓ) after Step (6) in the k-th iteration of the loop. We first argue that the existence of $s, t \in \mathbb{N}$ with

$$(*) \qquad u_t \leq u_s = w_t \text{ and } \{v \in B_s \mid v > u_s\} = \{v \in B_t \mid v > u_s\}$$

causes a contradiction. For this purpose, let $x = \sum_{v \in B_s} \lambda_v v$ with $\lambda_v \in K$ for all $v \in B_s$ be the representation of x from the s-th iteration of the loop. We will apply ℓ_t to this representation of x. For that matter, observe the following:

- For all $v \in B_s$ with $v < u_s = w_t$, we have by the assignment to u_s in (3) that $\lambda_v \geq 0$.
- For all $v \in E \supseteq B_s$ with $v < u_s = w_t$, we have by the assignement to w_t in (6) that $\ell_t(v) \geq 0$.
- $\lambda_{u_s} < 0$ according to (3)
- $\ell_t(u_s) = \ell_t(w_t) < 0$ according to (6)
- For all $v \in B_s$ with $v > u_s = w_t$, we have $\ell_t(v) = 0$ since for these v we have by $(*)$ that $v \in B_t \setminus \{u_t\}$ (using that $u_t \leq u_s$) and thus $\ell_t(v) = 0$ by (4).

It thus follows that

$$0 \overset{(4)}{>} \ell_t(x) = \underbrace{\sum_{\substack{v \in B_s \\ v < u_s}} \underbrace{\lambda_v}_{\geq 0} \underbrace{\ell_t(v)}_{\geq 0}}_{\geq 0} + \underbrace{\underbrace{\lambda_{u_s}}_{<0} \underbrace{\ell_t(u_s)}_{<0}}_{>0} + \underbrace{\sum_{\substack{v \in B_s \\ v > u_s}} \lambda_v \underbrace{\ell_t(v)}_{=0}}_{=0} > 0$$

which is the desired contradiction.

Finally, we show the existence of $s, t \in \mathbb{N}$ with $(*)$. For clarity, we first generalize the algorithm by looking at the following more abstract version of it:

Suppose E is a finite set, $\leq$ an order on E and B a subset of E.

(1′) Choose $u \in B$.

(2′) Choose $w \in E \setminus B$.

(3′) Replace B by $(B \setminus \{u\}) \cup \{w\}$ and go to (1′).

Denote by (B_k, u_k, w_k) the value of (B, u, w) after Step (2′) in the k-th iteration of the algorithm. We show the existence of $s, t \in \mathbb{N}$ satisfying $(*)$. Since E is finite, the power set of E is also finite. Consequently, there are $p, q \in \mathbb{N}$ such that $p < q$ and $B_p = B_q$. Because of (3′), it then obviously holds that $\{u_s \mid p \leq s < q\} = \{w_t \mid p \leq t < q\}$. Set $v_0 := \max\{u_s \mid p \leq s < q\} = \max\{w_t \mid p \leq t < q\}$. Then

$$\{v \in B_s \mid v > v_0\} = \{v \in B_t \mid v > v_0\}$$

for all $s, t \in \{p, \ldots, q-1\}$. Choose $s, t \in \{p, \ldots, q-1\}$ with $u_s = v_0 = w_t$ (note that $s < t$ or $t < s$ but certainly not $s = t$). Now $(*)$ holds. □

Corollary 7.4.21. [→ 2.3.2] *Let $(K, \leq)$ be an arbitrary ordered field. Let $m, n \in \mathbb{N}_0$, $f, \ell_1, \ldots, \ell_m \in K[X_1, \ldots, X_n]$ be linear forms* [→ 1.6.1(a)] *and set*

$$S := \{x \in K^n \mid \ell_1(x) \geq 0, \ldots, \ell_m(x) \geq 0\}.$$

Then the following are equivalent:

(a) $f \geq 0$ *on* S

(b) $f \in K_{\geq 0}\ell_1 + \ldots + K_{\geq 0}\ell_m$

(c) *There are $i_1, \ldots, i_s \in \{1, \ldots, m\}$ such that $\ell_{i_1}, \ldots, \ell_{i_s}$ are linearly independent and*

$$f \in K_{\geq 0}\ell_{i_1} + \ldots + K_{\geq 0}\ell_{i_s}.$$

Proof. (c) $\Longrightarrow$ (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (c) Suppose that (a) holds.

Claim: $f \in V := K\ell_1 + \ldots + K\ell_m$

Explanation. Assume $f \notin V$. Then there is some $\varphi \in (KX_1 + \ldots + KX_n)^*$ such that $\varphi(\ell_1) = \ldots = \varphi(\ell_m) = 0$ and $\varphi(f) = -1$. Set $x := (\varphi(X_1), \ldots, \varphi(X_n)) \in K^n$. Then $\ell_i(x) = \varphi(\ell_i) = 0$ for all $i \in \{1, \ldots, m\}$. Hence $x \in S$ and $f(x) = \varphi(f) = -1 < 0$. $\lightning$

Now apply 7.4.20 to V and $E := \{\ell_1, \ldots, \ell_m\}$ (taking account of the claim). Then it suffices to show that for all $\varphi \in V^*$ with $\varphi(E) \subseteq K_{\geq 0}$ also $\varphi(f) \geq 0$ holds. Thus let $\varphi \in V^*$ with $\varphi(E) \subseteq K_{\geq 0}$. Choose $\psi \in (KX_1 + \ldots + KX_n)^*$ with $\psi|_V = \varphi$. Set $x := (\psi(X_1), \ldots, \psi(X_n)) \in K^n$. Then $\ell_i(x) = \psi(\ell_i) = \varphi(\ell_i) \geq 0$ for all $i \in \{1, \ldots, m\}$ and thus $x \in S$ and $\varphi(f) = \psi(f) = f(x) \geq 0$. □

Corollary 7.4.22 (Linear Nichtnegativstellensatz). [→ 2.3.5, 3.7.7] *Let $(K, \leq)$ be an arbitrary ordered field. Let $m, n \in \mathbb{N}_0$, $f, \ell_1, \ldots, \ell_m \in K[X_1, \ldots, X_n]_1$ [→ 1.5.1] and suppose*

$$S := \{x \in K^n \mid \ell_1(x) \geq 0, \ldots, \ell_m(x) \geq 0\} \neq \emptyset.$$

Then the following are equivalent:

(a) $f \geq 0$ *on* S

(b) $f \in K_{\geq 0} + K_{\geq 0}\ell_1 + \ldots + K_{\geq 0}\ell_m$

(c) *There are* $i_1, \ldots, i_s \in \{0, \ldots, m\}$ *such that* $\ell_{i_1}, \ldots, \ell_{i_s}$ *are linearly independent and*

$$f \in K_{\geq 0}\ell_{i_1} + \ldots + K_{\geq 0}\ell_{i_s}$$

where $\ell_0 := 1$.

Proof. (c) $\Longrightarrow$ (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (c) Suppose that (a) holds. Due to 7.4.21, it suffices to show that $f^* \geq 0$ on $S^* := \{x = (x_0, \ldots, x_n) \in K^{n+1} \mid x_0 \geq 0, \ell_1^*(x) \geq 0, \ldots, \ell_m^*(x) \geq 0\}$ [→ 2.2.1(c)(d), 2.2.2(e)]. To this end, let $(x_0, \ldots, x_n) \in S^*$.

Case 1: $x_0 > 0$

Then $\left(1, \frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right) = \frac{1}{x_0}(x_0, \ldots, x_n) \in S^*$ and hence $\left(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right) \in S$. From (a), it follows that $f^*\left(1, \frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right) = f\left(\frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right) \geq 0$ and hence also $f^*(x_0, \ldots, x_n) = x_0 f^*\left(1, \frac{x_1}{x_0}, \ldots, \frac{x_n}{x_0}\right) \geq 0$

Case 2: $x_0 = 0$

Then $(\mathrm{LF}(\ell_i))(x_1, \ldots, x_n) \overset{2.2.2(a)}{=} \ell_i^*(x_0, \ldots, x_n) \geq 0$ and therefore

$$(\mathrm{LF}(\ell_i))(\lambda x_1, \ldots, \lambda x_n) \geq 0$$

for all $i \in \{1, \ldots, m\}$ and $\lambda \in K_{\geq 0}$. Because of $S \neq \emptyset$, we can choose $(y_1, \ldots, y_n) \in S$. Then $\ell_i(y_1 + \lambda x_1, \ldots, y_n + \lambda x_n) \geq 0$ for all $\lambda \in K_{\geq 0}$ and $i \in \{1, \ldots, m\}$. Due to (a), we have thus $f(y_1 + \lambda x_1, \ldots, y_n + \lambda x_n) \geq 0$ for all $\lambda \in K_{\geq 0}$. It follows that $(\mathrm{LF}(f))(x_1, \ldots, x_n) \geq 0$. Hence

$$f^*(x_0, \ldots, x_n) \overset{2.2.2(a)}{=} (\mathrm{LF}(f))(x_1, \ldots, x_n) \geq 0.$$

□

Definition 7.4.23. Let V be a K-vector space and $C \subseteq V$ a cone [→ 7.1.1].

(a) The sets of the form $K_{\geq 0}x$ with $x \in C \setminus \{0\}$ are called the *rays* of C.

(b) Rays of C that are at the same time faces [→ 7.3.6] of C are called *extreme rays* of C.

(c) A set $B \subseteq C \setminus \{0\}$ is called a *base* of C, if for each $x \in C \setminus \{0\}$ there is exactly one $\lambda \in K_{>0}$ such that $\lambda x \in B$, i.e., if every ray of C hits the set B in exactly one point.

Proposition 7.4.24. *Suppose V is a K-vector space and $C \subseteq V$ is a cone with convex base B. Then for all $x \in V$,*

$$K_{\geq 0}x \text{ is an extreme ray of } C \iff \exists \lambda \in K_{>0} : \lambda x \in \operatorname{extr} B.$$

Proof. Let $x \in V$.

"$\Longrightarrow$" Let $K_{\geq 0}x$ be an extreme ray of C. Then it follows that $x \in C \setminus \{0\}$. Hence there is exactly one $\lambda \in K_{>0}$ such that $\lambda x \in B$. We claim $\lambda x \in \operatorname{extr} B$. For this purpose, consider $y, z \in B$ with $\frac{y+z}{2} = \lambda x$. To show: $y = z = \lambda x$. From $y, z \in C$ and $\frac{y+z}{2} \in K_{\geq 0}x$, we deduce $y, z \in K_{\geq 0}x$. Due to $y, z \in B$ and $0 \notin B$, we get $y, z \in K_{>0}x$. Again from $y, z \in B$ and the uniqueness of λ, we get $y = \lambda x = z$.

"$\Longleftarrow$" WLOG let $x \in \operatorname{extr} B$. To show: $K_{\geq 0}x$ is an extreme ray of C. Since $x \in B \subseteq C \setminus \{0\}$, $K_{\geq 0}x$ is a ray of C. Let $y, z \in C$ with $\frac{y+z}{2} \in K_{\geq 0}x$. To show: $y, z \in K_{\geq 0}x$. WLOG $y \neq 0$ and $z \neq 0$. If we had $y + z = 0$, then one could easily show $0 \in B$ ↯. WLOG $y + z = x$. Choose $\mu, \nu \in K_{>0}$ such that $\mu y, \nu z \in B$. Then

$$x = y + z = (\mu^{-1} + \nu^{-1}) \underbrace{\left(\frac{\mu^{-1}}{\mu^{-1} + \nu^{-1}} (\mu y) + \frac{\nu^{-1}}{\mu^{-1} + \nu^{-1}} (\nu z) \right)}_{\in B}$$

and thus $\mu^{-1} + \nu^{-1} = 1$. Since $x = \mu^{-1}(\mu y) + \nu^{-1}(\nu z)$, $\mu y, \nu z \in B$ and $x \in \operatorname{extr} B$, we have $\mu y = x = \nu y$. □

Theorem 7.4.25. [→ 7.4.19] *Every convex cone with compact* [→ 7.4.18] *convex base in a finite-dimensional $\mathbb{R}$-vector space is the sum of its extreme rays.*

Proof. Suppose V is a finite-dimensional $\mathbb{R}$-vector space and $C \subseteq V$ is a convex cone with compact convex base B. Let $x \in C$. To show: x is a sum of elements of extreme rays of C. WLOG $x \in B$. By Minkowski's theorem 7.4.19, we have $x \in \operatorname{conv}(\operatorname{extr} B)$, say $x = \sum_{i=1}^{m} \lambda_i x_i$ with $m \in \mathbb{N}$, $\lambda_1, \ldots, \lambda_m \in K_{\geq 0}$, $\lambda_1 + \ldots + \lambda_m = 1$ and $x_i \in \operatorname{extr} B$. According to 7.4.24, $K_{\geq 0}x_i$ is for all $i \in \{1, \ldots, m\}$ an extreme ray of C. □

Proposition 7.4.26. *Every convex cone with compact* [→ 7.4.18] *base in a finite-dimensional $\mathbb{R}$-vector space is closed.*

Proof. Suppose V is a finite-dimensional $\mathbb{R}$-vector space and $C \subseteq V$ is a convex cone with compact base B. By Tikhonov's theorem 5.1.18, also $[0,1]_{\mathbb{R}} \times B$ is compact. From 7.1.18 together with the continuity of the scalar multiplication, we obtain that

$$A := \{\lambda x \mid \lambda \in [0,1]_{\mathbb{R}}, x \in B\}$$

is again compact. WLOG $V = \mathbb{R}^n$ by 7.2.20. WLOG $B \neq \emptyset$. Set

$$d := \min\{\|y\| \mid y \in B\} > 0.$$

In order to show that C is closed, we now let $x \in V \setminus C$. WLOG $\|x\| < \frac{d}{2}$ [→ 7.2.6]. Since A is closed by 7.2.16, there is an $\varepsilon > 0$ such that $\{y \in V \mid \|x - y\| < \varepsilon\} \cap A = \emptyset$. From $0 \in A$, we get $\varepsilon \leq \|x\| < \frac{d}{2}$. Then $\{y \in V \mid \|x - y\| < \varepsilon\} \cap C = \emptyset$ for if $y \in C \setminus A$, then there is $\lambda \in K$ with $0 < \lambda < 1$ and $\lambda y \in B$ and it follows that $\|y\| = \frac{1}{\lambda}\|\lambda y\| \geq \frac{1}{\lambda} d > d$ which is incompatible with $\|x - y\| < \frac{d}{2}$ (which would imply contrarily $\|y\| \leq \|y - x\| + \|x\| < \frac{d}{2} + \frac{d}{2} = d$). This shows that C is closed. □

7.5 Application to ternary quartics

A *ternary quartic* is a 4-form (also called quartic form [→ 2.3.4]) in 3 variables.

Lemma 7.5.1. Let $(K, \leq)$ be an ordered field and $G \in SK^{m \times m}$. Then G is psd [→ 2.3.1] if and only if $x^T G x \geq 0$ for all $x \in (K^\times)^m$.

Proof. Suppose $x^T G x \geq 0$ for all $x \in (K^\times)^m$. Let $z \in K^m$. We have to show that

$$z^T G z \geq 0.$$

Choose $y \in (K^\times)^m$ arbitrary. Then $z + \lambda y \in (K^\times)^m$ and therefore

$$z^T G z + 2\lambda y^T G z + \lambda^2 y^T G y = (z + \lambda y)^T G (z + \lambda y) \geq 0$$

for all but finitely many $\lambda \in K$. For example, by 1.5.3(b) applied to the polynomial

$$z^T G z + 2 y^T G z T + y^T G y T^2 \in K[T],$$

it follows that $z^T G z \geq 0$. □

Lemma 7.5.2. Let K be an Euclidean field and $f \in K[X, Y, Z]$ a 4-form. Suppose that there are linearly independent $v_1, v_2, v_3 \in K^3$ such that $f(v_1) = f(v_2) = f(v_3) = 0$. Then the following are equivalent:

(a) f is psd [→ 2.3.1(a)]

(b) $f \in \sum K[X, Y, Z]^2$

(c) f is a sum of 3 squares of quadratic forms in $K[X, Y, Z]$.

Proof. Denote by e_1, e_2, e_3 the standard basis of K^3. Set $A := \begin{pmatrix} v_1 & v_2 & v_3 \end{pmatrix} \in \mathrm{GL}_3(K)$ and $g := f\left(A \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}\right) \in K[X, Y, Z]$. Then g is a 4-form satisfying $g(e_1) = g(e_2) = g(e_3) = 0$. Since A defines a permutation (even a vector space isomorphism)

$$K^3 \to K^3, \ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \mapsto A \begin{pmatrix} x \\ y \\ z \end{pmatrix}$$

on K^3, we have that

$$f \text{ is psd} \iff g \text{ is psd.}$$

Since A induces on the other hand a ring automorphism

$$K[X,Y,Z] \to K[X,Y,Z],\ h \mapsto h\left(A\begin{pmatrix}X\\Y\\Z\end{pmatrix}\right),$$

we obtain

$$f \in \sum K[X,Y,Z]^2 \iff g \in \sum K[X,Y,Z]^2.$$

Since this ring automorphism permutes the quadratic forms in $K[X,Y,Z]$, we have that

$$\text{(c)} \iff g \text{ is a sum of 3 squares of quadratic forms.}$$

Replacing f by g, we can henceforth suppose that $v_1 = e_1$, $v_2 = e_2$ and $v_3 = e_3$.

(c) $\Longrightarrow$ (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (c) It is easy to see that each polynomial $g \in K[T]$ with $g \geq 0$ on K and $g(0) = 0$ lies in the ideal (T^2) [→ 1.5.3(b)]. Suppose now that (a) holds. The vanishing at 0 and the nonnegativity of the polynomials

$$f(1,T,0),\ f(1,0,T),\ f(T,1,0),\ f(0,1,T),\ f(0,T,1),\ f(T,0,1) \in K[T]$$

therefore forces the coefficients of

$$X^4,\quad X^3Y,\quad X^3Z,\quad Y^4,\quad Y^3X,\quad Y^3Z,\quad Z^4,\quad Z^3X,\quad Z^3Y$$

in f to vanish. For example, the first polynomial forces the coefficients of X^4 and X^3Y to vanish, and the second one the coefficients of again X^4 and of X^3Z. It follows that

$$\begin{aligned} N(f) &\subseteq \operatorname{conv}\{(2,2,0),\cancel{(2,1,1)},(2,0,2),(0,2,2),\cancel{(1,2,1)},\cancel{(1,1,2)}\}, \text{ i.e.,}\\ \frac{1}{2}N(f) &\subseteq \operatorname{conv}\{(1,1,0),(1,0,1),(0,1,1)\} \quad \text{and thus}\\ \frac{1}{2}N(f) \cap \mathbb{N}_0^3 &\subseteq \{(1,1,0),(1,0,1),(0,1,1)\}. \end{aligned}$$

By the Gram matrix method 2.6.1, we have to show that there is a *psd* matrix $G \in SK^{3\times 3}$ satisfying

$$(*) \qquad f = \begin{pmatrix}XY & XZ & YZ\end{pmatrix} G \begin{pmatrix}XY\\XZ\\YZ\end{pmatrix}.$$

Since every monomial occurring in f is a product of two entries of $\begin{pmatrix}XY & XZ & YZ\end{pmatrix}$, there is certainly a $G \in SK^{3\times 3}$ satisfying $(*)$ (actually one sees easily that there is a *unique* such G which does however not play an immediate role). But from $(*)$ it follows *automatically* that G is *psd* since f is psd. In order to see this, let $v \in K^3$. We have to show

that $v^T G v \geq 0$. Using 7.5.1, one reduces to the case $v \in (K^\times)^3$. Then set $\lambda := v_1 v_2 v_3$ and $x := \frac{1}{v_3}$, $y := \frac{1}{v_2}$ and $z := \frac{1}{v_1}$. Now $v = \lambda \begin{pmatrix} xy \\ xz \\ yz \end{pmatrix}$ and therefore

$$v^T G v = \lambda^2 \begin{pmatrix} xy & xz & yz \end{pmatrix} G \begin{pmatrix} xy \\ xz \\ yz \end{pmatrix} \overset{(*)}{=} \lambda^2 f(x,y,z) \geq 0.$$

□

Lemma 7.5.3. Let K be an Euclidean field and $f \in K[X,Y,Z]$ a 4-form. Suppose there are linearly independent $v_1, v_2, v_3 \in K^3$ satisfying $f(v_1 + Tv_2) \in (T^3)$ and $f(v_3) = 0$. Then the following are equivalent:

(a) f is psd

(b) $f \in \sum K[X,Y,Z]^2$

(c) f is a sum of 3 squares of quadratic forms in $K[X,Y,Z]$.

Proof. Almost exactly as in the proof of 7.5.2, one sees that one can suppose WLOG $v_1 = e_1$, $v_2 = e_2$ and $v_3 = e_3$.

(c) $\Longrightarrow$ (b) $\Longrightarrow$ (a) is again trivial.

(a) $\Longrightarrow$ (c) One sees easily that a polynomial $g \in K[T]$ with $g \geq 0$ on K and $g \in (T^{2k-1})$ lies in (T^{2k}) for $k \in \mathbb{N}$. Suppose now that (a) holds. By considering the polynomials

$$f(1,T,0),\ f(1,0,T),\ f(0,T,1),\ f(T,0,1) \in K[T],$$

one sees easily that the coefficients of

$$X^4,\quad X^3Y,\quad X^2Y^2,\quad XY^3,\quad X^3Z,\quad Z^4,\quad Z^3Y,\quad Z^3X$$

in f must vanish. More precisely, the first polynomial ist responsible for the first four of these coefficients, the second for the coefficients of X^4 (again) and X^3Z, the third for the coefficients of Z^4 and Z^3Y, and the last for the coefficients of Z^4 (again) and Z^3X. It follows that

$$N(f) \subseteq \operatorname{conv}\{(2,0,2),(2,1,1),\cancel{(1,1,2)},\cancel{(1,2,1)},(0,2,2),\cancel{(0,3,1)},(0,4,0)\}, \text{ i.e.,}$$
$$\frac{1}{2}N(f) \subseteq \operatorname{conv}\left\{(1,0,1),\left(1,\frac{1}{2},\frac{1}{2}\right),(0,1,1),(0,2,0)\right\} \quad \text{and thus}$$
$$\frac{1}{2}N(f) \cap \mathbb{N}_0^3 \subseteq \{(1,0,1),(0,1,1),(0,2,0)\}.$$

By the Gram matrix method 2.6.1, we have to show that there is a *psd* matrix $G \in SK^{3\times 3}$ satisfying

$$(*) \qquad f = \begin{pmatrix} XZ & YZ & Y^2 \end{pmatrix} G \begin{pmatrix} XZ \\ YZ \\ Y^2 \end{pmatrix}.$$

If the monomial X^2YZ actually appeared in f, we would now run into a big problem that we did not have in the proof of 7.5.2 because this monomial is not a product of two entries of $\begin{pmatrix} XZ & YZ & Y^2 \end{pmatrix}$. But this coefficient vanishes as one easily shows since for all $y \in K$, the leading coefficient of $f(X, y, 1) \in K[X]$ is nonnegative since this polynomial is nonnegative on K. As in the proof of 7.5.2, one sees again that there exists $G \in SK^{3\times 3}$ satisfying $(*)$ (one could again see easily that G is unique). From $(*)$ it follows *automatically* that G is psd since f is psd. To see this, let $v \in K^3$. To show: $v^T G v \geq 0$. Using 7.5.1, one reduces to the case $v \in K \times (K^\times)^2$. Then set $\lambda := v_2^2 v_3$ and $x := \frac{v_1}{v_2^2}$, $y := \frac{1}{v_2}$, $z := \frac{1}{v_3}$. Now $v = \lambda \begin{pmatrix} xz \\ yz \\ y^2 \end{pmatrix}$ and therefore

$$v^T G v = \lambda^2 \begin{pmatrix} xz & yz & y^2 \end{pmatrix} G \begin{pmatrix} xz \\ yz \\ y^2 \end{pmatrix} \overset{(*)}{=} \lambda^2 f(x, y, z) \geq 0.$$

□

Lemma 7.5.4. Let K be an Euclidean field and $f \in K[X, Y, Z]$ a 4-form. Suppose there are linearly independent $v_1, v_2 \in K^3$ satisfying $f(v_1 + Tv_2) \in (T^3)$ and $f(v_2) = 0$. Then the following are equivalent:

(a) f is psd

(b) $f \in \sum K[X, Y, Z]^2$

(c) f is a sum of 3 squares of quadratic forms in $K[X, Y, Z]$.

Proof. One can again suppose WLOG $v_1 = e_1$ and $v_2 = e_2$.

(c) $\Longrightarrow$ (b) $\Longrightarrow$ (a) is again trivial.

(a) $\Longrightarrow$ (c) One uses again that a polynomial $g \in K[T]$ with $g \geq 0$ on K and $g \in (T^{2k-1})$ lies in (T^{2k}) for $k \in \mathbb{N}$. Suppose now that (a) holds. By considering the polynomials

$$f(1, T, 0),\ f(1, 0, T),\ f(T, 1, 0),\ f(0, 1, T) \in K[T],$$

one sees easily that the coefficients of

$$X^4, \quad X^3Y, \quad X^2Y^2, \quad XY^3, \quad X^3Z, \quad Y^4, \quad Y^3Z$$

in f must vanish. More precisely, the first polynomial ist responsible for the first four of these coefficients, the second for the coefficients of X^4 (again) and X^3Z, the third for

the coefficients of Y^4 and XY^3 (again), and the last for the coefficients of Y^4 (again) and Y^3Z. It follows that

$N(f) \subseteq \operatorname{conv}\{(2,0,2),(2,1,1),(1,2,1),(0,2,2),\cancel{(0,1,3)},(0,0,4),\cancel{(1,0,3)},\cancel{(1,1,2)}\}$, i.e.,

$\frac{1}{2}N(f) \subseteq \operatorname{conv}\left\{(1,0,1),\left(1,\frac{1}{2},\frac{1}{2}\right),\left(\frac{1}{2},1,\frac{1}{2}\right),(0,1,1),(0,0,2)\right\}$ and thus

$$\frac{1}{2}N(f) \cap \mathbb{N}_0^3 \subseteq \{(1,0,1),(0,1,1),(0,0,2)\}.$$

By the Gram matrix method 2.6.1, we have to show that there is a *psd* matrix $G \in SK^{3\times 3}$ satisfying

$$(*) \qquad f = \begin{pmatrix} XZ & YZ & Z^2 \end{pmatrix} G \begin{pmatrix} XZ \\ YZ \\ Z^2 \end{pmatrix}.$$

If one of the monomials X^2YZ and XY^2Z actually appeared in f, we would have trouble since these monomials are not a product of two entries of $\begin{pmatrix} XZ & YZ & Z^2 \end{pmatrix}$. But these coefficients vanish as one easily shows since for all $x, y \in K$, the leading coefficients of $f(X,y,1) \in K[X]$ and $f(x,Y,1) \in K[Y]$ are nonnegative since these polynomials are nonnegative on K. One sees again that there exists $G \in SK^{3\times 3}$ satisfying $(*)$ (one could again see easily that G is unique). From $(*)$ it follows *automatically* that G is psd since f is psd. To see this, let $v \in K^3$. To show: $v^TGv \geq 0$. Using 7.5.1, one reduces to the case $v \in K \times (K^\times)^2$. Then set $\lambda := v_2^2 v_3$ and $x := \frac{v_1}{v_2 v_3}$, $y := \frac{1}{v_3}$, $z := \frac{1}{v_2}$. Now $v = \lambda \begin{pmatrix} xz \\ yz \\ z^2 \end{pmatrix}$ and therefore

$$v^TGv = \lambda^2 \begin{pmatrix} xz & yz & z^2 \end{pmatrix} G \begin{pmatrix} xz \\ yz \\ z^2 \end{pmatrix} \overset{(*)}{=} \lambda^2 f(x,y,z) \geq 0.$$

□

Lemma 7.5.5. Let $f \in \mathbb{R}[X,Y,Z]$ be a psd 4-form that is not a sum of 3 squares of quadratic forms in $\mathbb{R}[X,Y,Z]$ and that has two linear independent zeros in $\mathbb{R}^3$. Then there is a linear form $\ell \in \mathbb{R}[X,Y,Z] \setminus \{0\}$ such that $f - \ell^4$ is psd.

Proof. By Lemma 7.5.2, the zeros of f span a two-dimensional subspace of $\mathbb{R}^3$. By a change of coordinates, we can thus achieve that $f(e_2) = f(e_3) = 0$ and

$$f > 0 \text{ on } \mathbb{R}^\times \times \mathbb{R}^2.$$

We now claim that there is some $\varepsilon \in \mathbb{R}_{>0}$ such that f$-\varepsilon X^4$ is psd. By homogeneity, it suffices to find $\varepsilon > 0$ such that $f - \varepsilon X^4 \geq 0$ holds on the compact set

$$[-1,1]^3_{\mathbb{R}} \setminus (-1,1)^3_{\mathbb{R}}.$$

For this purpose, it is enough show that for each two-dimensional face F of the polytope $[-1,1]^3$ (i.e., for each side of the cube $[-1,1]^3$) there is some $\varepsilon > 0$ such that $f - \varepsilon X^4 \geq 0$ on F. On the two sides $\{-1\} \times [-1,1]^2$ and $\{1\} \times [-1,1]^2$, f is positive so that the existence of such an ε for them follows from 7.1.19. After a further change of coordinates, it suffices to consider from the remaining four sides just $[-1,1]^2 \times \{1\}$. Consider therefore $\widetilde{f} := f(X,Y,1) \in \mathbb{R}[X,Y]$ [→ 2.2.1(d)]. From Lemma 7.5.4, we deduce

$$\frac{\partial^2 \widetilde{f}}{\partial Y^2}(0,y) > 0$$

for all $y \in \mathbb{R}$ that satisfy $\widetilde{f}(0,y) = 0$ (apply 7.5.4 to f, $v_1 := (0,y,1)$ and $v_2 := (0,1,0)$, taking into account that $\frac{\partial \widetilde{f}}{\partial Y}(0,y) = 0$ due to $\widetilde{f} \geq 0$ on $\mathbb{R}^2$). In the same way, Lemma 7.5.3 implies that for each $y \in \mathbb{R}$ satisfying $\widetilde{f}(0,y) = 0$ all other directional derivatives of $\widetilde{f}$ in $(0,y)$ are also positive. Altogether, $\widetilde{f}$ has thus only zeros in $\mathbb{R}^2$ at which the second derivative (i.e., the Hessian) is *pd* (recall that all zeros of $\widetilde{f}$ lie on the y-axis). From analysis we know know that each zero of the nonnegative polynomial $\widetilde{f}$ (in $\mathbb{R}^2$, or equivalently $\{0\} \times \mathbb{R}$ since all zeros lie on the y-axis) is an isolated *global* minimizer. Therefore

$$\{(x,y) \in \mathbb{R}^2 \mid \widetilde{f}(x,y) = 0\} = \{(0,y_1), \ldots, (0,y_m)\}$$

for some $m \in \mathbb{N}$ and $y_1, \ldots, y_m \in \mathbb{R}$ (one of the y_i is 0). Since $-X^4$ as well as its first and second derivative vanishes on the y-axis (since $\frac{\partial X^4}{\partial X} = 4X^3$, $\frac{\partial X^4}{\partial Y} = 0$, $\frac{\partial^2 X^4}{\partial X^2} = 12X^2$, $\frac{\partial^2 X^4}{\partial X \partial Y} = 0$ and $\frac{\partial^2 X^4}{\partial Y^2} = 0$), every $(0,y_i)$ is a zero and an isolated *local* minimizer of $\widetilde{f} - X^4$. Choose for each $i \in \{1,\ldots,m\}$ an open neighborhood U_i of $(0,y_i)$ such that $\widetilde{f} - X^4 > 0$ on $U_i \setminus \{(0,y_i)\}$. Then of course also $\widetilde{f} - \varepsilon X^4 > 0$ on $U_i \setminus \{(0,y_i)\}$ for all $\varepsilon \leq 1$ and $i \in \{1,\ldots,m\}$. Since $\widetilde{f}$ is positive on the compact set $[-1,1]^2 \setminus (U_1 \cup \cdots \cup U_m)$, there is an $\varepsilon \in (0,1)_{\mathbb{R}}$ such that $\widetilde{f} - \varepsilon X^4 > 0$ on $[-1,1]^2 \setminus (U_1 \cup \cdots \cup U_m)$. Altogether, $\widetilde{f} - \varepsilon X^4 > 0$ on $[-1,1]^2 \setminus \{(0,y_1), \ldots, (0,y_m)\}$ and $\widetilde{f} - \varepsilon X^4 = 0$ on $\{(0,y_1), \ldots, (0,y_m)\}$. □

Lemma 7.5.6. Suppose f lies on an extreme ray [→ 7.4.23(b)] of the cone P of the psd 4-forms in $\mathbb{R}[X,Y,Z]$. Then there are linear independent $v_1, v_2 \in \mathbb{R}^3$ such that $f(v_1) = f(v_2) = 0$.

Proof. If f were pd, then the forms $f \pm \varepsilon X^4$ would be psd for some $\varepsilon > 0$ (choose ε for instance as the minimum of f on the compact unit sphere of $\mathbb{R}^3$) and because of $f = \frac{1}{2}(f - \varepsilon X^4) + \frac{1}{2}(f + \varepsilon X^4)$ it would follow that $f + \varepsilon X^4 \in \mathbb{R}_{\geq 0} f$ and thus $f \in \mathbb{R} X^4$ ↯. Hence f has at least one zero $v_1 \in \mathbb{R}^3 \setminus \{0\}$. After a change of coordinates, we can without loss of generality achieve $v_1 = e_1$. Since $(0,0)$ is a local (even a global) minimizer of $f(1,Y,Z) \in \mathbb{R}[Y,Z]$, we know from analysis that $\frac{\partial f}{\partial Y}(1,0,0) = \frac{\partial f}{\partial Z}(1,0,0) = 0$. It follows that there are $a, b, c \in \mathbb{R}[Y,Z]$ such that

$$f = aX^2 + bX + c.$$

We have to show that there exists $v_2 \in \mathbb{R} \times (\mathbb{R}^2 \setminus \{0\})$ such that $f(v_2) = 0$. We make a case distinction by $\mathrm{rk}(a)$ [→ 1.6.1(h)].

Case 1: $\mathrm{rk}(a) = 0$

Then $a = 0$ and thus $b(y,z) = 0$ for all $(y,z) \in \mathbb{R}^2$ from which $b = 0$ follows by 2.2.3. If $f = c \in \mathbb{R}[Y,Z]$ was pd, then $c \pm \varepsilon Y^4 \in \mathbb{R}[Y,Z]$ would be psd for some $\varepsilon > 0$ and it would follow that $c + \varepsilon Y^4 \in \mathbb{R}_{\geq 0} c$ and thus $c \in \mathbb{R} Y^4$ ↯.

Case 2: $\mathrm{rk}(a) = 1$

By a coordinate change in the y-z-plane WLOG $a = Y^2$. Then $b(0,z) = 0$ for all $z \in \mathbb{R}$ and hence $b(0,Z) = 0$, i.e., $b = Yb'$ for some $b \in \mathbb{R}[X,Y]$. It follows that $f = X^2Y^2 + b'XY + c = \left(XY + \frac{b'}{2}\right)^2 + \left(c - \frac{b'^2}{4}\right)$. For all $(y,z) \in \mathbb{R}^\times \times \mathbb{R}$, we find some $x \in \mathbb{R}$ satisfying $xy + \frac{b'(y,z)}{2} = 0$ from which $c(y,z) - \frac{b'(y,z)^2}{4} = f(x,y,z) \geq 0$ follows. Hence $c - \frac{b'^2}{4} \in P$. Aside from that, we have of course $(XY + \frac{b'}{2})^2 \in P$. Since f lies on an extreme ray of P, it follows that $(XY + \frac{b'}{2})^2 \in \mathbb{R} f$ (and $c - \frac{b'^2}{4} \in \mathbb{R} f$). Now choose $(y,z) \in \mathbb{R}^\times \times \mathbb{R}$ arbitrary and with it $x \in \mathbb{R}$ such that $xy + \frac{b'(y,z)}{2} = 0$. Then $f(x,y,z) = 0$.

Case 3: $\mathrm{rk}(a) = 2$

By a coordinate change in the y-z-plane WLOG $a = Y^2 + Z^2$. Since f is psd, also the 6-form $4ac - b^2 \in \mathbb{R}[Y,Z]$ is psd. We have to show that there is $(y,z) \in \mathbb{R}^2 \setminus \{0\}$ such that there exists $x \in \mathbb{R}$ satisfying $a(y,z)x^2 + b(y,z) + c(y,z) = 0$. Because of $a(y,z) \neq 0$ for all $(y,z) \in \mathbb{R}^2 \setminus \{0\}$, this is equivalent to the existence of $(y,z) \in \mathbb{R}^2 \setminus \{0\}$ with $(b^2 - 4ac)(y,z) \geq 0$, i.e., $(4ac - b^2)(y,z) = 0$ (since $4ac - b^2$ is psd). We have thus to show that $4ac - b^2$ is not pd. Aiming for a contradiction, assume that $4ac - b^2$ is psd. Then also the 6-forms $4a(c \pm \varepsilon Y^4) - b^2$ are psd for some $\varepsilon > 0$ (choose for example 4ε as the minimum of $4ac - b^2$ on the compact unit sphere of $\mathbb{R}^2$ and take into account that $a = Y^2 + Z^2$). It follows that $f \pm \varepsilon Y^4 \in P$. From $f = \frac{1}{2}(f + \varepsilon Y^4) + \frac{1}{2}(f - \varepsilon Y^4)$, we obtain $f + \varepsilon Y^4 \in \mathbb{R}_{\geq 0} f$ and thus $f \in \mathbb{R} Y^4$ ↯. □

Lemma 7.5.7. Let $d, n \in \mathbb{N}_0$ and let V be the $\mathbb{R}$-vector space of all $2d$-forms in $\mathbb{R}[\underline{X}] = \mathbb{R}[X_1, \ldots, X_n]$ and $P \subseteq V$ be the cone of all psd forms in V. Then P is a closed cone with compact convex base [→ 7.4.23(c)].

Proof. As an intersection of closed sets, $P = \bigcap_{x \in \mathbb{R}^n} \{p \in V \mid p(x) \geq 0\}$ is closed. By 2.2.3,

$$\|p\| := \sum_{x_1=-d}^{d} \cdots \sum_{x_n=-d}^{d} |p(x_1, \ldots, x_n)| \qquad (p \in V)$$

defines a norm on V. Then

$$B := \{p \in P \mid \|p\| = 1\} = \left\{ p \in V \mid \sum_{x_1=-d}^{d} \cdots \sum_{x_n=-d}^{d} p(x_1, \ldots, x_n) = 1 \right\}$$

is a compact convex base of P. □

Lemma 7.5.8. Let V denote the $\mathbb{R}$-vector space of all 4-forms in $\mathbb{R}[X,Y,Z]$ and $P \subseteq V$ the cone of all psd forms in V. Suppose that f lies on an extreme ray of P. Then f is a square of a quadratic form.

Proof. It is enough to show that f is a *sum of* squares of quadratic forms for if $f = \sum_{i=1}^m q_i^2 \neq 0$ with 2-forms $q_i \in \mathbb{R}[X,Y,Z]$, then

$$f = \frac{1}{2}\underbrace{2q_1^2}_{\in P} + \frac{1}{2}\underbrace{2\sum_{i=2}^m q_i^2}_{\in P}$$

and thus $q_1^2 \in \mathbb{R}_{\geq 0} f$. If there is a linear form $\ell \in \mathbb{R}[X,Y,Z] \setminus \{0\}$ such that $f - \ell^4$ is psd, then $\ell^4 \in \mathbb{R}_{\geq 0} f$ and $f = (c\ell^2)^2$ for some $c \in \mathbb{R}^\times$ so that we are done. From now on therefore suppose that such a linear form does not exist. From the Lemmata 7.5.5 and 7.5.6, it follows now that f is a sum of 3 squares of 2-forms in $\mathbb{R}[X,Y,Z]$. □

Theorem 7.5.9. *Let R be a real closed field and $f \in R[X,Y,Z]$ a 4-form. Then the following are equivalent:*

(a) *f is psd.*

(b) *$f \in \sum R[X,Y,Z]^2$*

(c) *f is a sum of squares of quadratic forms in $R[X,Y,Z]$.*

Proof. (c) $\Longrightarrow$ (b) $\Longrightarrow$ (a) is trivial.

(a) $\Longrightarrow$ (c) follows for $R = \mathbb{R}$ from 7.5.8 together with the conic version 7.4.25 of Minkowski's theorem. Using the Gram matrix method 2.6.1 (or 7.4.20), one sees that the class of all real closed fields R for which (a) $\Longrightarrow$ (c) holds for all $f \in R[X,Y,Z]$, is semialgebraic. By 1.8.5, every real closed field belongs to this class. In short, the statement follows thus from the case $R = \mathbb{R}$ by the Tarski principle 1.8.19. □

Corollary 7.5.10 (dehomogenized version of 7.5.9)**.** *Let R be a real closed field and $f \in R[X,Y]_4$. Then*

$$f \geq 0 \text{ on } R^2 \iff f \in \sum R[X,Y]^2.$$

Proof. "$\Longleftarrow$" is trivial.

"$\Longrightarrow$" Suppose $f \geq 0$ on R^2. WLOG $f \notin R$. Then $\deg f = 2$ or $\deg f = 4$ by 2.2.4(b). For $\deg f = 2$, the claim follows from 2.3.5. Suppose therefore $\deg f = 4$. Then $f^* := Z^4 f\left(\frac{X}{Z}, \frac{Y}{Z}\right) \in R[X,Y,Z]$ is the homogenization of f with respect to Z [→ 2.2.1(c)] and f^* is psd by 2.2.6(a). Now 7.5.9 yields $f^* \in \sum R[X,Y,Z]^2$. By dehomogenization [→ 2.2.1(d), 2.2.2], it follows that $f \in \sum R[X,Y]^2$. □

Remark 7.5.11. A posteriori, we see now that in the situation of Lemma 7.5.6, there actually exist even infinitely many pairwise distinct zeros of f. This follows from 7.5.8. Indeed, if $f = q^2$ with a 2-form $q \in \mathbb{R}[X,Y,Z]$ with $\operatorname{rk} q = 3$, then WLOG $\operatorname{sg} q \geq 0$ (otherwise replace q by $-q$) and thus $\operatorname{sg} q \in \{3,1\}$. If $\operatorname{sg} q = 3$, then q and thus f is pd. If $\operatorname{sg} q = 1$, then q and thus f have infinitely many pairwise linearly independent zeros.

Remark 7.5.12. We will neither use nor prove the following:

(a) In 1888, Hilbert showed a strengthening of 7.5.9 ("sum of *three* squares" instead of "sum of squares", cf. also 7.5.2, 7.5.3, 7.5.4 and 7.5.5) [Hil]. A very long and tedious elementary proof for this has been given by Scheiderer and Pfister in 2012 [PS]..

(b) Scheiderer showed in 2016 that

$$X^4 + XY^3 + Y^4 - 3X^2YZ - 4XY^2Z + 2X^2Z^2 + XZ^3 + YZ^3 + Z^4$$

is psd but does not belong to $\sum \mathbb{Q}[X, Y, Z]^2$ [S2]. In the same year, Henrion, Naldi, Safey El Din gave an elementary proof for this [HNS].

§8 Nonnegative polynomials with zeros

Throughout this chapter, K denotes again always a subfield of $\mathbb{R}$ with the induced order. Moreover, we let A always be a commutative ring (e.g., $A = K[X_1, \ldots, X_n]$).

8.1 Modules over semirings

Definition 8.1.1. Let $T \subseteq A$. Then we call T a *semiring* of A if $\{0,1\} \subseteq T$, $T + T \subseteq T$ and $TT \subseteq T$ [→ 1.2.1]. If T is a semiring of A, then $M \subseteq A$ is called a *T-module* of A if $0 \in M$, $M + M \subseteq M$ and $TM \subseteq M$.

Remark 8.1.2. (a) T is a preorder of $A \iff (T$ is a semiring of A & $A^2 \subseteq T)$

(b) If T is a semiring of A, then $T - T$ is a subring of A.

(c) If T is a semiring of A and M a T-module of A, then $M - M$ is a $(T - T)$-module of A.

(d) If T is a semiring of A, then T is a T-module of A.

Definition 8.1.3. Let T be a semiring of A and M a T-module of A. Then M is called *Archimedean* (in A) if $\forall a \in A : \exists N \in \mathbb{N} : N + a \in M$ [→ 4.1.2(a)].

Remark 8.1.4. Due to 8.1.2(d), the notion of an Archimedean semiring is also defined by 8.1.3. Because of 8.1.2(a), this generalizes the notion of an Archimedean preorder of A [→ 4.1.2(a)].

Definition 8.1.5. [→ 4.3.1] Let T be a semiring of A, M a T-module of A and $u \in A$. Then

$$B_{(A,M,u)} := \{a \in A \mid \exists N \in \mathbb{N} : Nu \pm a \in M\}$$

the set of with respect to M by *u arithmetically bounded* elements of A. If $u = 1$, then we write $B_{(A,M,u)} := B_{(A,M)}$ and omit the specification "by u".

Proposition 8.1.6. *Suppose T is a semiring of A, M_1 and M_2 are T-modules of A, $u_1 \in M_1$ and $u_2 \in M_2$. Then $\sum M_1 M_2$ is also a T-module of A and we have*

$$B_{(A,M_1,u_1)} B_{(A,M_2,u_2)} \subseteq B_{(A,\sum M_1 M_2, u_1 u_2)}.$$

Proof. Let $a_i \in B_{(A,M_i,u_i)}$, say $Nu_i \pm a_i \in M_i$ for $i \in \{1,2\}$ with $N \in \mathbb{N}$. Then (cf. the proof of 4.3.1)

$$3N^2 u_1 u_2 \pm a_1 a_2 = (Nu_1 + a_1)(Nu_2 \pm a_2) + Nu_2(Nu_1 - a_1) + Nu_1(Nu_2 \mp a_2).$$

□

Corollary 8.1.7. *Let T be a semiring of A, M a T-module of A, $u \in T$ and $v \in M$. Then $B_{(A,T,u)}B_{(A,M,v)} \subseteq B_{(A,M,uv)}$.*

Proof. Apply 8.1.6 to $M_1 := T$, $M_2 := M$, $u_1 := u$, $u_2 := v$ and observe $\sum M_1M_2 = \sum TM = M$. □

Corollary 8.1.8. [→ 4.3.1] *Let T be a semiring of A. Then $B_{(A,T)}$ is a subring of A. Moreover, if M a T-module of A and $u \in M$, then $B_{(A,M,u)}$ is a $B_{(A,T)}$-module of A.*

Remark 8.1.9. [→ 8.1.3, 4.3.3] If $T \subseteq A$ is a semiring and $M \subseteq A$ a T-module, then M is Archimedean if and only if $B_{(A,M)} = A$.

Theorem 8.1.10. [→ 4.3.4] *Let $n \in \mathbb{N}_0$ and $T \subseteq K[\underline{X}]$ a semiring with $K_{\geq 0} \subseteq T$. Then the following are equivalent:*

(a) *T is Archimedean.*

(b) $\exists N \in \mathbb{N} : \forall i \in \{1, \ldots, n\} : N \pm X_i \in T$

(c) $\exists m \in \mathbb{N} : \exists \ell_1, \ldots, \ell_m \in T \cap K[\underline{X}]_1 : \exists N \in \mathbb{N} :$
$\emptyset \neq \{x \in K^n \mid \ell_1(x) \geq 0, \ldots, \ell_m(x) \geq 0\} \subseteq [-N, N]_K^n$

Proof. Write $A := K[\underline{X}]$. From $K_{\geq 0} \subseteq T$, it follows that $K \subseteq B_{(A,T)}$. Hence we have $B_{(A,T)} = A \iff X_1, \ldots, X_n \in B_{(A,T)}$ which shows (a) $\iff$ (b). The implication (b) $\implies$ (c) is trivial and (c) $\implies$ (b) is an easy consequence of the linear Nichtnegativstellensatz 7.4.22. □

Lemma 8.1.11. [→ 4.3.2] Suppose that $\frac{1}{2} \in A$ (i.e., $2 \in A^\times$), let $M \subseteq A$ be a $(\sum A^2)$-module with $1 \in M$ and let $a \in A$. Then

$$a^2 \in B_{(A,M)} \iff a \in B_{(A,M)}.$$

Proof. "$\implies$" If $N \in \mathbb{N}$ with $(N-1) - a^2 \in M$, then

$$N \pm a = (N-1) - a^2 + \left(\frac{1}{2} \pm a\right)^2 + 3\left(\frac{1}{2}\right)^2 \in M$$

(exactly as in the proof of 4.3.2).

"$\impliedby$" If $N \in \mathbb{N}$ with $(2N-1) \pm a \in M$, then

$$N^2(2N-1) - a^2 = 2\left(\frac{1}{2}\right)^2 ((N-a)^2(2N-1+a) + (N+a)^2(2N-1-a)) \in M.$$

□

Proposition 8.1.12. *Suppose $\frac{1}{2} \in A$, $T \subseteq A$ is a preorder and $M \subseteq A$ is a T-module with $1 \in M$. Then $B_{(A,M)}$ is a subring of A and $B_{(A,M,u)}$ a $B_{(A,M)}$-module of A for each $u \in T$.*

Proof. It obviously suffices to show $B_{(A,M)}B_{(A,M,u)} \subseteq B_{(A,M,u)}$ for all $u \in T$ (since this means $B_{(A,M)}B_{(A,M)} \subseteq B_{(A,M)}$ for $u = 1$). If $a \in B_{(A,M)}$, then we have

$$a = \left(\frac{1}{2}\right)^2 ((a+1)^2 - (a-1)^2)$$

and because of $1 \in M$ also $a+1, a-1 \in B_{(A,M)}$. Therefore it is enough to show $a^2 B_{(A,M,u)} \subseteq B_{(A,M,u)}$ for all $a \in B_{(A,M)}$ and $u \in T$. For this purpose, fix $a \in B_{(A,M)}$, $u \in T$ and $b \in B_{(A,M,u)}$. To show: $a^2 b \in B_{(A,M,u)}$. From 8.1.11, we get $a^2 \in B_{(A,M)}$. Choose $N \in \mathbb{N}$ such that $N - a^2, Nu \pm b \in M$. Due to $a^2, u \in T$, we get now $Nu - ua^2, Nua^2 \pm a^2 b \in M$. Consequently,

$$N^2 u \pm a^2 b = (N^2 u - Nua^2) + (Nua^2 \pm a^2 b) \in M + M \subseteq M.$$

□

Theorem 8.1.13. [→ 4.3.4, 8.1.10] *Suppose $n \in \mathbb{N}_0$ and $M \subseteq K[\underline{X}]$ is a $(\sum K_{\geq 0} K[\underline{X}]^2)$-module with $1 \in M$. Then the following are equivalent:*

(a) *M is Archimedean.*

(b) $\exists N \in \mathbb{N} : N - \sum_{i=1}^n X_i^2 \in M$

(c) $\exists N \in \mathbb{N} : \forall i \in \{1, \ldots, n\} : N \pm X_i \in M$

(d) $\exists m \in \mathbb{N} : \exists \ell_1, \ldots, \ell_m \in M \cap K[\underline{X}]_1 : \exists N \in \mathbb{N} :$
$\emptyset \neq \{x \in K^n \mid \ell_1(x) \geq 0, \ldots, \ell_m(x) \geq 0\} \subseteq [-N, N]_K^n$

Proof. (a) $\Longrightarrow$ (b) is trivial.

(b) $\Longrightarrow$ (c) If (b) holds, then $N - X_i^2 \in M$ and thus $X_i^2 \in B_{(K[\underline{X}],M)}$ for all $i \in \{1, \ldots, n\}$. Apply now 8.1.11.

(c) $\Longrightarrow$ (d) is trivial and (d) $\Longrightarrow$ (c) follows again from the linear Nichtnegativstellensatz 7.4.22.

(c) $\Longrightarrow$ (a) follows from 8.1.12. □

8.2 Pure states on rings and ideals

In this section, we always suppose that the field K is a subring of A. In particular, $\mathbb{Q} \subseteq A$ and A is a K-vector space.

Remark 8.2.1. Under the just made mild hypothesis $\mathbb{Q} \subseteq A$, one can reformulate the abstract Archimedean Positivstellensatz 4.1.3 as follows:

For arbitrary A and K as above, let T be an Archimedean preorder of A such that $K_{\geq 0} \subseteq T$ and $a \in A$. Then the following are equivalent:

(a) $\varphi(a) > 0$ for all K-linear ring homomorphisms $\varphi\colon A \to \mathbb{R}$ with $\varphi(T) \subseteq \mathbb{R}_{\geq 0}$.

(b) $\exists N \in \mathbb{N} : a \in \frac{1}{N} + T$

To see this, first note that in (a), one can omit the K-linearity of φ since it just means that $\varphi|_K = \mathrm{id}_K$ which follows from 1.1.15 by $K_{\geq 0} \subseteq T$ since the identity is the *only* embedding of ordered fields from K to $\mathbb{R}$ (cf. the proof of 4.2.1). But then the theorem becomes strongest for $K = \mathbb{Q}$ and we can thus assume $K = \mathbb{Q}$ which makes redundant the hypothesis $K_{\geq 0} \subseteq T$ since for all $m, n \in \mathbb{N}$, we have $\frac{m}{n} = mn\left(\frac{1}{n}\right)^2 \in \sum A^2 \subseteq T$. This last fact also shows (b) $\iff$ (b′) where we denote by (a′) and (b′) the corresponding conditions from 4.1.3, namely:

(a′) $\widehat{a} > 0$ on $\mathrm{sper}(A, T)$

(b′) $\exists N \in \mathbb{N} : Na \in 1 + T$

It remains to show that (a) $\iff$ (a′). To this end, it suffices by 4.1.4(d) to show that (a) is equivalent to

(a″) $\widehat{a}(Q) > 0$ for all maximal elements Q of $\mathrm{sper}(A, T)$.

It is clear that (a′) $\implies$ (a″). To show (a″) $\implies$ (a′), suppose that (a″) holds and let $P \in \mathrm{sper}(A, T)$. To show: $\widehat{a}(P) > 0$. Using 3.2.3 or 3.2.5, we find a maximal element Q of $\mathrm{sper}(A, T)$ such that $P \subseteq Q$. By 3.2.4, we have $Q = P \cup \mathrm{supp}(Q)$. Due to (a″), we have $a \in Q \setminus -Q$, i.e., $a \in Q \setminus \mathrm{supp}(Q) \subseteq P$, and because of $a \notin -P$ (for otherwise $a \in -Q$) it follows that $a \in P \setminus -P$, i.e., $\widehat{a}(P) > 0$. This shows (a′) $\iff$ (a″). These arguments were implicitly present already in the proof of 4.2.2.

Remark 8.2.2. Suppose T is a semiring of A with $K_{\geq 0} \subseteq T$ and M a T-module of A. Then M is a cone in the K-vector space A and we have:

$$M \text{ is Archimedean } [\to 8.1.3] \iff 1 \text{ is a unit for } M \ [\to 7.1.4]$$

Motivation 8.2.3. If T is an Archimedean preorder of A with $K_{\geq 0} \subseteq T$, then the Archimedean Positivstellensatz 4.1.3 in the version of 8.2.1 amounts to the equivalence of

$$\exists N \in \mathbb{N} : a \in \frac{1}{N} + T$$

with

$$(*) \qquad \varphi(a) > 0 \text{ for all } (K\text{-linear}) \text{ ring homomorphisms } \varphi\colon A \to \mathbb{R} \text{ with } \varphi(T) \subseteq \mathbb{R}_{\geq 0}$$

while 7.3.19, paying attention to 8.2.2, tells that the same condition is equivalent to

$$(**) \qquad \varphi(a) > 0 \text{ for all pure states } \varphi \text{ of } (A, T, 1).$$

The following imprecise questions arise:

(a) What do pure states "on rings" have to do with ring homomorphisms?

(b) Can the Archimedean Positivstellensatz be generalized from preorders to semirings or even to modules over semirings?

(c) If $(*)$ holds only with "$\geq$" instead of "$>$", then $\exists N \in \mathbb{N} : a \in \frac{1}{N} + T$ can of course not hold anymore but one would still want to prove that $a \in T$. In this case, is it possible to find an ideal $I \subseteq A$ (e.g., the kernel of a ring homomorphism φ from $(*)$ with $\varphi(a) = 0$) such that $I \cap T$ possesses in the K-vector space I a unit u in such a way that $a \in I$ and $(**)$ holds for $(I, I \cap T, u)$ instead of $(A, T, 1)$? Then one could apply 7.3.19 or 7.3.20 in order to finally still show that $a \in T$ (even $a \in \frac{1}{N}u + (I \cap T)$).

(d) What can one say about pure states "on ideals"? This question generalizes (a) and is motivated by (c).

Reminder 8.2.4. For $z \in \mathbb{C}$ and $k \in \mathbb{N}_0$, the binomial coefficient

$$\binom{z}{k} := \prod_{i=1}^{k} \frac{z - i + 1}{i}$$

is declared. From analysis, one knows that

$$\sqrt{1+t} = (1+t)^{\frac{1}{2}} = \sum_{i=0}^{\infty} \binom{\frac{1}{2}}{i} t^i$$

for all $t \in \mathbb{R}$ with $|t| < 1$.

Lemma 8.2.5. For all $k \in \mathbb{N}$, the coefficients of

$$p_k := \left(\sum_{i=0}^{k} \binom{\frac{1}{2}}{i} (-T)^i \right)^2 - (1 - T) \in \mathbb{Q}[T]$$

are nonnegative.

Proof. In the ring $\mathbb{Q}[[T]]$ of formal power series, we have because of 8.2.4 and the identity theorem for power series from analysis that

$$\left(\sum_{i=0}^{\infty} \binom{\frac{1}{2}}{i} (-T)^i \right)^2 = 1 - T.$$

Now let $k \in \mathbb{N}$ be fixed. For $i \in \mathbb{N}_0$ with $i \leq k$, the coefficient of T^i in p_k obviously equals the coefficient of T^i in

$$\left(\sum_{i=0}^{\infty} \binom{\frac{1}{2}}{i} (-T)^i \right)^2 - (1 - T)$$

which is zero. The binomial coefficient $\binom{\frac{1}{2}}{i}$ is positive for $i \in \{0, 1, 3, 5, \dots\}$ and negative for $i \in \{2, 4, 6, \dots\}$. The only positive coefficient of

$$\sum_{i=0}^{\infty} \binom{\frac{1}{2}}{i} (-T)^i$$

is thus the constant term. Hence, for $i \in \mathbb{N}_0$ with $i > k$, the coefficient of T^i in p_k is thus a sum of products of two nonpositive reals and therefore nonnegative. □

Lemma 8.2.6. Suppose I is an ideal of A, T is a preorder of A with $K_{\geq 0} \subseteq T$, $M \subseteq I$ is a T-module of A, u is a unit for M in I, $a \in T$ and $(1-2a)u \in M$. Then [→ 7.1.9]

$$S(I, M, u) \subseteq S(I, (1-a)M, u).$$

Proof. Let $\varphi \in S(I, M, u)$. To show: $\varphi((1-a)M) \subseteq \mathbb{R}_{\geq 0}$. Let $b \in M$. To show:

$$\varphi((1-a)b) \geq 0.$$

WLOG $u - b \in M$ (otherwise choose $N \in \mathbb{N}$ with $Nu - b \in M$ and replace b by $\frac{1}{N}b \in M$). We show $\varphi((1-a)b) > -\varepsilon$ for all $\varepsilon > 0$. To this end, let $\varepsilon > 0$. It is enough to show that there is a $k \in \mathbb{N}$ satisfying

$$\varphi((1-a)b) > \varphi\left(\left(\sum_{i=0}^{k} \binom{\frac{1}{2}}{i} (-a)^i\right)^2 b\right) - \varepsilon$$

since $A^2M \subseteq TM \subseteq M \subseteq \varphi^{-1}(\mathbb{R}_{\geq 0})$. Because of $a \in T$, we have

$$(1-(2a)^i)u = \sum_{j=0}^{i-1}((2a)^j - (2a)^{j+1})u = \sum_{j=0}^{i-1}(2a)^j(1-2a)u \in M$$

for all $i \in \mathbb{N}_0$, i.e.,

$$(\square) \qquad \left(\frac{1}{2^i} - a^i\right) u \in M$$

for all $i \in \mathbb{N}_0$. By 8.2.4, we can choose $k \in \mathbb{N}$ such that

$$\left(\sum_{i=0}^{k} \binom{\frac{1}{2}}{i} \left(-\frac{1}{2}\right)^i\right)^2 < \left(1 - \frac{1}{2}\right) + \varepsilon,$$

i.e., $p_k\left(\frac{1}{2}\right) < \varepsilon$ with p_k as in Lemma 8.2.5. We show that $\varphi(p_k(a)b) < \varepsilon$. Since $p_k(a) \in T$ holds by Lemma 8.2.5, it is enough to show that $\varphi(p_k(a)u) < \varepsilon$ since $\varphi(p_k(a)b) \leq \varphi(p_k(a)u)$ holds due to $p_k(a)(u-b) \in M$. But we have

$$\varphi(p_k(a)u) \leq \varphi(p_k\left(\frac{1}{2}\right)u) = p_k\left(\frac{1}{2}\right)\varphi(u) = p_k\left(\frac{1}{2}\right) < \varepsilon$$

due to $(p_k\left(\frac{1}{2}\right) - p_k(a))u \in M$ (use 8.2.5 and $(\square)$). □

Theorem 8.2.7 (Burgdorf, Scheiderer, Schweighofer [BSS]). [→ 8.2.3(d)] *Suppose that I is an ideal of A, T is a preorder or an Archimedean semiring of A, $K_{\geq 0} \subseteq T$, $M \subseteq I$ is a T-module of A, u is a unit for M in I and φ is a pure state of (I, M, u). Then*

$$(*) \qquad \varphi(ab) = \varphi(au)\varphi(b)$$

for all $a \in A$ and $b \in I$.

Proof. Due to $T - T = A$ [→ 1.2.3, 8.1.3] it suffices to show $(*)$ for all $a \in T$ and $b \in I$. If T is $\left\{\begin{smallmatrix}\text{an Archimedean semiring}\\ \text{a preorder}\end{smallmatrix}\right\}$, then one can here suppose by scaling a that $\left\{\begin{smallmatrix}1-a \in T\\ u-2au \in M\end{smallmatrix}\right\}$ and thus because of $\left\{\begin{smallmatrix}TM \subseteq M\\ \text{Lemma 8.2.6}\end{smallmatrix}\right\}$ that

$$S(I,M,u) \subseteq S(I,(1-a)M,u).$$

Moreover, we can suppose that $\varphi(au) < 1$. Fix therefore $a \in T$ with $S(I,M,u) \subseteq S(I,(1-a)M,u)$ and $\varphi(au) < 1$. We have to show $(*)$ for all $b \in I$.

Case 1: $\varphi(au) = 0$

Then we have to show that $\varphi(ab) = 0$ for all $b \in I$. For this purpose, fix $b \in I$. Choose $N \in \mathbb{N}$ such that $Nu \pm b \in M$. Then $Nau \pm ab \in TM \subseteq M$ and therefore $|\varphi(ab)| \leq N\varphi(au) = 0$. Hence $\varphi(ab) = 0$.

Case 2: $\varphi(au) \neq 0$

Then $\varphi(au) > 0$ because of $au \in TM \subseteq M$. Furthermore, we have $\varphi((1-a)u) > 0$ since $\varphi(au) < 1 = \varphi(u)$. For each $c \in A$ with $\varphi(cu) > 0$ and $\varphi \in S(I,cM,u)$,

$$\varphi_c \colon I \to \mathbb{R},\ b \mapsto \frac{\varphi(cb)}{\varphi(cu)}$$

is a state of (I,M,u). In particular, $\varphi_a, \varphi_{1-a} \in S(I,M,u)$. Because of $\varphi = \varphi(au)\varphi_a + \varphi((1-a)u)\varphi_{1-a}$, $\varphi(au) > 0$, $\varphi((1-a)u) > 0$ and $\varphi(au) + \varphi((1-a)u) = \varphi(u) = 1$, we have by 2.4.2 or 7.3.8 that $\varphi = \varphi_a$ (and $\varphi = \varphi_{1-a}$). □

Corollary 8.2.8. [→ 8.2.3(a)] *Let T be an Archimedean semiring of A such that $K_{\geq 0} \subseteq T$ and M a T-module of A with $1 \in M$. Then every pure state of $(A,M,1)$ is a ring homomorphism.*

Corollary 8.2.9. [→ 8.2.3(a)] *Let M be an Archimedean $\left(\sum K_{\geq 0}A^2\right)$-module of A. Then every pure state of $(A,M,1)$ is a ring homomorphism.*

Corollary 8.2.10 (Becker, Schwartz [BS], first generalization of the abstract Archimedean Positivstellensatz 4.1.3 in the version of 8.2.1)**.** [→ 8.2.3(b)] *Let T be an Archimedean semiring of A with $K_{\geq 0} \subseteq T$, M a T-module of A with $1 \in M$ and $a \in A$. Then the following are equivalent:*

(a) $\varphi(a) > 0$ *for all (K-linear) ring homomorphisms* $\varphi \colon A \to \mathbb{R}$ *with* $\varphi(M) \subseteq \mathbb{R}_{\geq 0}$.

(b) $\exists N \in \mathbb{N} : a \in \frac{1}{N} + M$

Proof. 7.3.19, 8.2.2, 8.2.8 □

Corollary 8.2.11 (Jacobi [Jac], second generalization of the abstract Archimedean Positivstellensatz 4.1.3 in the version of 8.2.1)**.** [→ 8.2.3(b)] *Let M be an Archimedean $\left(\sum K_{\geq 0}A^2\right)$-module of A. Then (a) and (b) from 8.2.10 are equivalent.*

Remark 8.2.12. Using Lemma 4.2.1, one gets for the polynomial ring $K[\underline{X}]$ concrete geometric versions of 8.2.10 and 8.2.11 which are completely analogous to 4.2.2 (first and second generalization of the Archimedean Positivstellensatz). Instead of stating them, we give immediately concrete examples.

Example 8.2.13. [→ 8.2.10] Let $\ell_1, \dots, \ell_m \in \mathbb{R}[\underline{X}]_1$ such that

$$\{x \in \mathbb{R}^n \mid \ell_1(x) \geq 0, \dots, \ell_m(x) \geq 0\}$$

is nonempty and compact. Moreover, let $g_1, \dots, g_\ell \in \mathbb{R}[\underline{X}]$ and set

$$S := \{x \in \mathbb{R}^n \mid \ell_1(x) \geq 0, \dots, \ell_m(x) \geq 0, g_1(x) \geq 0, \dots, g_\ell(x) \geq 0\}.$$

Then for each $f \in \mathbb{R}[\underline{X}]$ with $f > 0$ on S, we have

$$f \in \sum_{i=0}^{\ell} \sum_{\alpha \in \mathbb{N}_0^m} \mathbb{R}_{\geq 0} \ell_1^{\alpha_1} \cdots \ell_m^{\alpha_m} g_i =: M$$

where $g_0 := 1$. This is because M is a T-module with $1 \in M$ for the semiring

$$T := \sum_{\alpha \in \mathbb{N}_0^m} \mathbb{R}_{\geq 0} \ell_1^{\alpha_1} \cdots \ell_m^{\alpha_m}$$

which is Archimedean by 8.1.10(c).

Example 8.2.14 (Putinar). [→ 8.2.11] Let $R \in \mathbb{R}_{\geq 0}$ and let $g_1, \dots, g_m \in \mathbb{R}[\underline{X}]$. Set

$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, \dots, g_m(x) \geq 0, \|x\| \leq R\}.$$

Then for every $f \in \mathbb{R}[\underline{X}]$ with $f > 0$ on S, we have

$$f \in \sum_{i=0}^{m+1} \sum \mathbb{R}[\underline{X}]^2 g_i$$

with $g_0 := 1$ and $g_{m+1} := R^2 - \sum_{i=1}^n X_i^2$ [→ 8.1.13(b)].

Example 8.2.15 (Pólya [Pól]). [→ 8.2.10] Let $k \in \mathbb{N}_0$ and suppose $f \in \mathbb{R}[\underline{X}]$ a k-form such that $f(x) > 0$ for all $x \in \mathbb{R}^n_{\geq 0} \setminus \{0\}$. Then there is some $N \in \mathbb{N}$ such that

$$(X_1 + \cdots + X_n)^N f \in \sum_{\substack{\alpha \in \mathbb{N}_0^n \\ |\alpha| = N+k}} \mathbb{R}_{>0} \underline{X}^\alpha.$$

This can be shown as follows: We have $f > 0$ on $\Delta := \{x \in \mathbb{R}^n_{\geq 0} \mid x_1 + \ldots + x_n = 1\}$. By 8.2.10, we obtain analogously to 8.2.13 that

$$f - \varepsilon \in \sum_{\alpha \in \mathbb{N}_0^{n+2}} \mathbb{R}_{\geq 0} X_1^{\alpha_1} \cdots X_n^{\alpha_n} (1 - (X_1 + \cdots + X_n))^{\alpha_{n+1}} (X_1 + \cdots + X_n - 1)^{\alpha_{n+2}}.$$

By substituting $X_i \mapsto \frac{X_i}{X_1 + \cdots + X_n}$ and clearing denominators, one gets the claim due to homogeneity of f.

8.3 Dichotomy of pure states on ideals

In this section, we let K again be a subring of A so that we consider A also as a K-vector space.

Proposition 8.3.1. *Let I be a ideal of A and $u \in I$. Let $\varphi \in S(I, \emptyset, u)$ [→ 7.1.9]. Then the following are equivalent:*

(a) $\forall a \in A : \forall b \in I : \varphi(ab) = \varphi(au)\varphi(b)$ [→ 8.2.7(∗)]

(b) *There is a ring homomorphism $\Phi\colon A \to \mathbb{R}$ such that*

$$(**) \qquad \forall a \in A : \forall b \in I : \varphi(ab) = \Phi(a)\varphi(b).$$

In Condition (b), Φ is uniquely determined since $(**)$ implies $\Phi(a) = \varphi(au)$ for all $a \in A$ and we call Φ the ring homomorphism belonging to *or* associated to *φ (on A). Note that Φ does not depend on u for if $v \in I$ with $\varphi(v) = 1$ then $(**)$ of course also implies $\Phi(a) = \varphi(av)$. Exactly one of the following alternatives occurs:*

(1) $\Phi(u) \neq 0$ *and* $\forall b \in I : \varphi(b) = \frac{\Phi(b)}{\Phi(u)}$

(2) $\Phi|_I = 0$

Proof. (a) $\Longrightarrow$ (b) If (a) holds, then $\Phi\colon A \to \mathbb{R}$, $a \mapsto \varphi(au)$ is a ring homomorphism since $\Phi(a)\Phi(b) = \varphi(au)\varphi(bu) \overset{(*)}{=} \varphi(abu) = \Phi(ab)$ holds for all $a, b \in A$.

(b) $\Longrightarrow$ (a) is clear.

Because of $u \in I$ it is clear that (1) and (2) exclude each other. If $\varphi(u^2) \neq 0$, then (1) occurs since $(*)$ implies $\varphi(bu) = \varphi(u^2)\varphi(b)$ for all $b \in I$. If $\varphi(u^2) = 0$, then $\varphi(bu) = \varphi(u^2)\varphi(b) = 0\varphi(b) = 0$ for all $b \in I$. □

Theorem 8.3.2 (Dichotomy)**.** *Under the hypotheses of 8.2.7, exactly one of the following cases occurs:*

(1) *φ is the restriction of a scaled ring homomorphism: There is a ring homomorphism $\Phi\colon A \to \mathbb{R}$ such that $\Phi(u) \neq 0$ and $\varphi = \frac{1}{\Phi(u)}\Phi|_I$.*

(2) *There is a ring homomorphism $\Phi\colon A \to \mathbb{R}$ with $\Phi|_I = 0$ such that $(**)$ from 8.3.1(b) holds.*

We have (1) $\iff$ $\varphi(u^2) \neq 0$ and (2) $\iff$ $\varphi(u^2) = 0$. In both (1) and (2), Φ is uniquely determined, namely it is the ring homomorphism that according to 8.3.1 belongs to φ. We have $\Phi(T) \subseteq \mathbb{R}_{\geq 0}$. If $u \in T$, then additionally $\Phi(M) \subseteq \mathbb{R}_{\geq 0}$.

Proof. Easy with 8.2.7 and 8.3.1. □

Corollary 8.3.3. *Let M be a $\left(\sum K_{\geq 0}A^2\right)$-module of A with $1 \in M$. If M has a unit in A, then M is Archimedean.*

Proof. Let u be a unit for M in A. By 7.3.19, it is enough to show that $\varphi(1) > 0$ for all $\varphi \in \operatorname{extr} S(A, M, u)$. Now let φ be a pure state of (A, M, u) with the associated ring homomorphism $\Phi\colon A \to \mathbb{R}$. Due to $\Phi(1) = 1 \neq 0$, in the Dichotomy 8.3.2 only case (1) can occur, i.e., $\Phi(u) \neq 0$ and $\varphi = \frac{1}{\Phi(u)}\Phi$. Because of $\Phi(u) = \varphi(u^2) = \varphi(u^2 \cdot 1) \in \varphi(M) \subseteq \mathbb{R}_{\geq 0}$, we have $\Phi(u) > 0$. It follows that $\varphi(1) > 0$. □

Example 8.3.4. Consider the semiring $T := \sum_{\alpha,\beta,\gamma \in \mathbb{N}_0} K_{\geq 0} X^\alpha Y^\beta (1 - X - Y)^\gamma$ of $K[X, Y]$ and

$$S := \{(x, y) \in \mathbb{R}^2 \mid \forall p \in T : p(x, y) \geq 0\} = \{(x, y) \in \mathbb{R}^2 \mid x \geq 0, y \geq 0, x + y \leq 1\}.$$

Since S is bounded and $X, Y, 1 - X - Y$ are linear, T is Archimedean by 8.1.10(c). Consider the ideal $I := (X, Y)$ and the T-module $M := T \cap I$ of $K[X, Y]$. Then $u := X + Y$ is a unit for M in I because $B_{(A,T)} \overset{8.1.9}{=} K[X, Y]$ and thus by 8.1.8 $B_{(A,M,u)}$ is an ideal of $K[X, Y]$ that contains X, Y and thus I since $u \pm X, u \pm Y \in M$. The ring homomorphisms

$$\Phi\colon K[X, Y] \to \mathbb{R}$$

satisfying $\Phi(T) \subseteq \mathbb{R}_{\geq 0}$ are obviously exactly the evaluations ev_x in points $x \in S$ (compare Lemma 4.2.1). Now let φ be a pure state of (I, M, u). By the Dichotomy 8.3.2, exactly one of the following cases occurs:

(1) There is some $(x, y) \in S \setminus \{(0, 0)\}$ with $\varphi(p) = \frac{p(x,y)}{x+y}$ for all $p \in I$.

(2) $\varphi(pX + qY) = \varphi(pX) + \varphi(qY) = p(0, 0)\varphi(X) + q(0, 0)\varphi(Y)$ for all $p, q \in K[X, Y]$.

In Case (2), one can set $\lambda_1 := \varphi(X) \geq 0$ and $\lambda_2 := \varphi(Y) \geq 0$ and one obtains $\lambda_1 + \lambda_2 = \varphi(X + Y) = \varphi(u) = 1$ as well as $\varphi = \lambda_1\varphi_1 + \lambda_2\varphi_2$ with $\varphi_1\colon I \to \mathbb{R},\ p \mapsto \frac{\partial p}{\partial X}(0, 0)$ and $\varphi_2\colon I \to \mathbb{R},\ p \mapsto \frac{\partial p}{\partial Y}(0, 0)$. Since every polynomial in M vanishes in the origin and is nonnegative on S, we obtain $\varphi_1, \varphi_2 \in S(I, M, u)$. Because of $\varphi \in \operatorname{extr} S(I, M, u)$, in Case (2) we have $\varphi = \varphi_1$ or $\varphi = \varphi_2$. Using 7.3.19, we now obtain: If $f \in K[X, Y]$ with $f > 0$ on $S \setminus \{0\}$, $f(0) = 0$, $\frac{\partial f}{\partial X}(0) > 0$ and $\frac{\partial f}{\partial Y}(0) > 0$, then $f \in T$.

Example 8.3.5. Let T and S be as in Example 8.3.4. Consider the ideal $I := (X)$ and the T-module $M := T \cap I$ of $K[X, Y]$. Then $u := X$ is a unit for M in I since $B_{(I,M,u)}$ is an ideal of $K[X, Y]$ by 8.1.8 that contains X and thus I because $u \pm X \in M$. Let φ be a pure state of (I, M, u). By the Dichotomy 8.3.2, exactly one of the following cases occurs:

(1) There is some $(x, y) \in S \setminus (\{0\} \times \mathbb{R})$ with $\varphi(p) = \frac{p(x,y)}{x}$ for all $p \in I$.

(2) There is some $y \in [0, 1]$ such that $\varphi(pX) = p(0, y)\varphi(X) = p(0, y)\varphi(u) = p(0, y)$ for alle $p \in K[X, Y]$.

In Case (2), there is obviously a $y \in [0, 1]$ such that $\varphi(p) = \frac{\partial p}{\partial X}(0, y)$ for all $p \in K[X, Y]$. Observe that each $f \in K[X, Y]$ with $f = 0$ on $S \cap (\{0\} \times \mathbb{R})$ satisfies $f(0, Y) = 0$ and thus $f \in I$. Now 7.3.19 yields: If $f \in K[X, Y]$ with $f > 0$ on $S \setminus (\{0\} \times \mathbb{R})$, $f = 0$ on

$S \cap (\{0\} \times \mathbb{R})$ and $\frac{\partial f}{\partial X}(0,y) > 0$ for all $y \in [0,1]$, then $f \in T$. At first glance, it might irritate that one would have to check here that $\frac{\partial f}{\partial X} f(0,1) > 0$. However, note that for $y = 1$ and in fact for every $y \in \mathbb{R}$, $\frac{\partial f}{\partial X} f(0,y)$ is the derivative of f in *every* direction $(1,z)$ with $z \in \mathbb{R}$ since $\frac{\partial f}{\partial Y} f(0,y) = 0$.

Example 8.3.6. Let T and S again be as in 8.3.4 and 8.3.5. Consider the ideal $I := (X^2, XY)$ and the T-module $M := T \cap I$ of $K[X,Y]$. Then $u := X^2 + XY$ is a unit for M in I since $u \pm X^2, u \pm XY \in M$. Let φ be a pure state of (I,M,u). By the Dichotomy 8.3.2, exactly one of the following cases occurs:

(1) There is some $(x,y) \in S \setminus (\{0\} \times \mathbb{R})$ with $\varphi(p) = \frac{p(x,y)}{x(x+y)}$ for all $p \in I$.

(2) There is some $y \in [0,1]$ such that $\varphi(pX^2 + qXY) = p(0,y)\varphi(X^2) + q(0,y)\varphi(XY)$ for all $p, q \in K[X,Y]$.

Suppose now that (2) holds and fix $y \in [0,1]$ accordingly. Consider $\lambda_1 := \varphi(X^2) \geq 0$, $\lambda_2 := \varphi(XY) \geq 0$. Then $\lambda_1 + \lambda_2 = \varphi(u) = 1$.

Consider first the case $y > 0$. From $0 = \varphi(YX^2 - X(XY)) = \lambda_1 y - \lambda_2 0 = \lambda_1 y$ we get $\lambda_1 = 0$. Then $\frac{1}{y}\frac{\partial(pX^2+qXY)}{\partial X}(0,y) = \frac{1}{y} q(0,y) y = q(0,y) = \lambda_1 p(0,y) + \lambda_2 q(0,y) = \varphi(pX^2 + qXY)$ for all $p, q \in K[X,Y]$. Hence $\varphi = \varphi_y$ with

$$\varphi_y \colon I \to \mathbb{R},\ p \mapsto \frac{1}{y}\frac{\partial p}{\partial X}(0,y).$$

Consider now the case $y = 0$. Then $\frac{1}{2}\frac{\partial^2(pX^2+qXY)}{\partial X^2}(0,0) = p(0,0) = p(0,y)$ and $\frac{\partial^2(pX^2+qXY)}{\partial X \partial Y}(0,0) = q(0,0) = q(0,y)$ for all $p, q \in K[X,Y]$. Hence $\varphi = \lambda_1\psi_1 + \lambda_2\psi_2$ with

$$\psi_1 \colon I \to \mathbb{R},\ p \mapsto \frac{1}{2}\frac{\partial^2 p}{\partial X^2}(0,0) \qquad \text{and} \qquad \psi_2 \colon I \to \mathbb{R},\ p \mapsto \frac{\partial^2 p}{\partial X \partial Y}(0,0).$$

Before we give a summary, we observe that

$$I = \left\{ f \in K[X,Y] \mid f = 0 \text{ on } S \cap (\{0\} \times \mathbb{R}), \frac{\partial f}{\partial X}(0) = 0 \right\}$$

where "$\subseteq$" is clear since the right hand side forms obviously an ideal and "$\supseteq$" can be seen as follows: If $f \in K[X,Y]$ with $f = 0$ on $S \cap (\{0\} \times \mathbb{R})$, then $f(0,Y) = 0$ and thus $f \in (X)$. If $f = Xg \in K[X,Y]$ with $\frac{\partial f}{\partial X}(0) = 0$, then $g(0) = 0$, hence $g \in (X,Y)$ and consequently $f \in (X^2, XY)$. Taking into account that each polynomial in M is nonnegative on S, one obtains $\varphi_y \in S(I,M,u)$ for all $y \in (0,1]_{\mathbb{R}}$ and $\psi_1, \psi_2 \in S(I,M,u)$. The above considerations therefore yield

$$\operatorname{extr} S(I,M,u) \subseteq \{\varphi_y \mid y \in (0,1]_{\mathbb{R}}\} \cup \{\psi_1, \psi_2\}$$

from which one obtains with 7.3.19: If $f \in K[X,Y]$ with

- $f > 0$ on $S \setminus (\{0\} \times \mathbb{R})$,
- $f = 0$ on $S \cap (\{0\} \times \mathbb{R})$,
- $\frac{\partial f}{\partial X}(0, y) > 0$ for $y \in (0, 1]_{\mathbb{R}}$,
- $\frac{\partial f}{\partial X}(0, 0) = 0$,
- $\frac{\partial^2 f}{\partial X^2}(0, 0) > 0$ and
- $\frac{\partial^2 f}{\partial X \partial Y}(0, 0) > 0$,

then $f \in T$.

8.4 A local-global-principle

Proposition 8.4.1. *Let T be a semiring of A with $K_{\geq 0} \subseteq T$, M a T-module of A, $n \in \mathbb{N}_0$ and $a_1, \ldots, a_n \in A$. Set $I := (a_1, \ldots, a_n)$. Moreover, let u be a unit for $\left\{ \begin{matrix} T \\ M \end{matrix} \right\}$ in A and suppose $a_1, \ldots, a_n \in \left\{ \begin{matrix} M \\ T \end{matrix} \right\}$. Then $u(a_1 + \ldots + a_n)$ is a unit for $M \cap I$ in I.*

Proof. Let $b \in I$ and set $v := u(a_1 + \cdots + a_n)$. To show: $\exists N \in \mathbb{N} : Nv + b \in M \cap I$. Write $b = \sum_{i=1}^{n} c_i a_i$ with $c_1, \ldots, c_n \in A$. Choose $N \in \mathbb{N}$ such that $Nu \pm c_i \in \left\{ \begin{matrix} T \\ M \end{matrix} \right\}$ for $i \in \{1, \ldots, n\}$. Then $Nv \pm b = \sum_{i=1}^{n} (Nua_i \pm c_i a_i) = \sum_{i=1}^{n} (Nu \pm c_i) a_i \in M$. □

Theorem 8.4.2 (Burgdorf, Scheiderer, Schweighofer [BSS])**.** *Let T be an Archimedean semiring of A with $K_{\geq 0} \subseteq T$ and M a T-module of A. Let $a \in A$ such that there is for each maximal ideal $\mathfrak{m}$ of A some $t \in T \setminus \mathfrak{m}$ with $ta \in M$. Then $a \in M$.*

Proof. Set $I := (\{b \in M \mid \exists t \in T : ta - b \in M\})$. By hypothesis and by the definition of I, for each maximal ideal $\mathfrak{m}$ of A there exists $t \in T \setminus \mathfrak{m}$ such that $ta \in M \cap I$.

Claim: $a \in I$

Explanation. $(\{t \in T \mid ta \in I\})$ is not contained in any maximal ideal of A. Thus there are $m \in \mathbb{N}$ and $t_1, \ldots, t_m \in T$ with $t_1 a, \ldots, t_m a \in I$ and $d_1, \ldots, d_m \in A$ with $1 = \sum_{i=1}^{m} d_i t_i$. It follows that

$$a = 1 \cdot a = \sum_{i=1}^{m} d_i \underbrace{(t_i a)}_{\in I} \in I.$$

By the just proven claim, there are $m \in \mathbb{N}$, $b_i, c_i \in M$ and $t_i \in T$ such that $t_i a - b_i = c_i$ for $i \in \{1, \ldots, m\}$ and $a \in (b_1, \ldots, b_m) \subseteq (b_1, \ldots, b_m, c_1, \ldots, c_m) =: J$. By the (first version of) 8.4.1, $u := \sum_{i=1}^{m} (b_i + c_i) = a \sum_{i=1}^{m} t_i = at$ with $t := \sum_{i=1}^{m} t_i \in T$ is a unit for

$M \cap J$ in J. To show that $a \in M$, we will now apply 7.3.20. So let φ be a pure state of $(J, J \cap M, u)$. To show: $\varphi(a) > 0$. Denote by Φ the ring homomorphism that belongs to φ according to 8.3.1. We have $\Phi(T) \subseteq \mathbb{R}_{\geq 0}$ [$\to$ 8.3.2]. Now

$$1 = \varphi(u) = \varphi(at) = \varphi(ta) = \underbrace{\Phi(t)}_{\geq 0} \varphi(a).$$

Thus $\varphi(a) > 0$. □

§9 Nonnegative polynomials and truncated quadratic modules

9.1 Pure states and polynomials over real closed fields

Throughout this section, we let R be a real closed extension field of $\mathbb{R}$, we set $\mathscr{O} := \mathscr{O}_R$, $\mathfrak{m} := \mathfrak{m}_R$ and we make extensive use of the standard part maps $\mathscr{O} \to \mathbb{R}$, $a \mapsto \mathrm{st}(a)$, $\mathscr{O}[\underline{X}] \to \mathbb{R}[\underline{X}]$, $p \mapsto \mathrm{st}(p)$ [→ 5.4.7] and $\mathscr{O}^n \to \mathbb{R}^n$, $x \mapsto \mathrm{st}(x) := (\mathrm{st}(x_1), \ldots, \mathrm{st}(x_n))$ which are surjective ring homomorphisms.

Definition 9.1.1. [→ 1.2.1, 4.1.2(a)] Let A be a commutative ring and $M \subseteq A$. Then M is called a *quadratic module* of A if M is a $\sum A^2$-module of A containing 1 [→ 8.1.1], or in other words, if $\{0,1\} \subseteq M$, $M + M \subseteq M$ and $A^2M \subseteq M$. We call a quadratic module M of A *Archimedean* if $B_{(A,M)} = A$ [→ 4.3.1, 4.3.3, 8.1.5].

Proposition 9.1.2. [→ 8.1.13] *Suppose $n \in \mathbb{N}_0$ and M is a quadratic module of $\mathscr{O}[\underline{X}]$. Then the following are equivalent:*

(a) *M is Archimedean.*

(b) $\exists N \in \mathbb{N} : N - \sum_{i=1}^n X_i^2 \in M$

(c) $\exists N \in \mathbb{N} : \forall i \in \{1, \ldots, n\} : N \pm X_i \in M$

Proof. (a) $\Longrightarrow$ (b) is trivial.

(b) $\Longrightarrow$ (c) If (b) holds, then $N - X_i^2 \in M$ and thus $X_i^2 \in B_{(\mathscr{O}[\underline{X}],M)}$ for all $i \in \{1, \ldots, n\}$. Apply now 8.1.11.

(c) $\Longrightarrow$ (a) follows from 8.1.12 since $\mathscr{O} \subseteq B_{(\mathscr{O}[\underline{X}],M)}$. □

Remark 9.1.3. In contrast to 8.1.13(d), one cannot add

$$\exists m \in \mathbb{N} : \exists \ell_1, \ldots, \ell_m \in M \cap \mathscr{O}[\underline{X}]_1 : \exists N \in \mathbb{N} : \\ \varnothing \neq \{x \in R^n \mid \ell_1(x) \geq 0, \ldots, \ell_m(x) \geq 0\} \subseteq [-N, N]_R^n$$

as another equivalent condition in 9.1.2. Indeed, choose R non-Archimedean [→ 5.4.4] and $\varepsilon \in \mathfrak{m} \setminus \{0\}$. Then $\varnothing \neq \{0\} = \{x \in R \mid \varepsilon x \geq 0, -\varepsilon x \geq 0\} \subseteq [-1,1]_R$ but the quadratic module

$$\sum \mathscr{O}[X]^2 + \sum \mathscr{O}[X]^2 \varepsilon X + \sum \mathscr{O}[X]^2 (-\varepsilon X) \overset{1.2.3}{=} \sum \mathscr{O}[X]^2 + \mathscr{O}[X] \varepsilon X$$

generated by εX and $-\varepsilon X$ in $\mathscr{O}[X]$ is not Archimedean for if we had $N \in \mathbb{N}$ with

$$N - X^2 \in \sum \mathscr{O}[X]^2 + \mathscr{O}[X]\varepsilon X,$$

then taking standard parts would yield $N - X^2 \in \sum \mathbb{R}[X]^2$ which contradicts 2.2.4(b).

Definition 9.1.4. [→ 4.2.1] For every $x \in \mathscr{O}^n$, we define the ring homomorphism

$$\mathrm{ev}_x \colon \mathscr{O}[\underline{X}] \to \mathscr{O},\ p \mapsto p(x)$$

and set $I_x := \ker \mathrm{ev}_x$.

Proposition 9.1.5. *Let $x \in \mathscr{O}^n$. Then $I_x = (X_1 - x_1, \ldots, X_n - x_n)$.*

Proof. It is trivial that $J := (X_1 - x_1, \ldots, X_n - x_n) \subseteq I_x$. Conversely, $p \equiv_J p(x) = 0$ for all $p \in I_x$. This shows the converse inclusion $I_x \subseteq J$. □

Notation 9.1.6. Suppose A is a commutative ring and I is an ideal of A. As it is customary in commutative algebra, we will in the following often denote by I^2 the product of the ideal I with itself which in our suggestive notation [→ 1.1.18] would be written $\sum II$. From the context, the reader should be able to avoid misinterpreting I^2 as what it would mean in this suggestive notation, namely $\{a^2 \mid a \in I\}$. The same applies to I^3 and so on. Another source of confusion could be that, we will often use the notation $\mathfrak{m}^n$ to denote the Cartesian power

$$\underbrace{\mathfrak{m} \times \ldots \times \mathfrak{m}}_{n \text{ times}}.$$

Lemma 9.1.7. Suppose $x, y \in \mathscr{O}^n$ with $x - y \notin \mathfrak{m}^n$. Then I_x and I_y are coprime, i.e., $1 \in I_x + I_y$.

Proof. WLOG $x_1 - y_1 \notin \mathfrak{m}$. Then $x_1 - y_1 \in \mathscr{O}^\times$ and

$$1 = \frac{x_1 - X_1}{x_1 - y_1} + \frac{X_1 - y_1}{x_1 - y_1} \in I_x + I_y.$$

□

Lemma 9.1.8. Let M be an Archimedean quadratic module of $\mathscr{O}[\underline{X}]$ and $x \in \mathscr{O}^n$. Then

$$u := (X_1 - x_1)^2 + \ldots + (X_n - x_n)^2$$

is a unit for $M \cap I_x^2$ in the real vector space I_x^2 [→ 7.1.4].

Proof. Using the ring automorphism

$$\mathscr{O}[\underline{X}] \to \mathscr{O}[\underline{X}],\ p \mapsto p(X_1 - x_1, \ldots, X_n - x_n),$$

which is also an isomorphism of real vector spaces, we can reduce to the case $x = 0$. Since $u \in I_0^2$, it suffices to show that $I_0^2 \subseteq B_{(\mathscr{O}[\underline{X}], M, u)}$. Since M is Archimedean,

8.1.12 yields that $B_{(\mathscr{O}[\underline{X}],M,u)}$ is an $\mathscr{O}[\underline{X}]$-module of $\mathscr{O}[\underline{X}]$ [→ 8.1.1], i.e., an ideal of $\mathscr{O}[\underline{X}]$. Because of

$$I_0^2 = (X_iX_j \mid i,j \in \{1,\dots,n\}),$$

it suffices therefore to show that $X_iX_j \in B_{(\mathscr{O}[\underline{X}],M,u)}$ for all $i,j \in \{1,\dots,n\}$. Thus fix $i,j \in \{1,\dots,n\}$. Then $\frac{1}{2}(X_i^2+X_j^2) \pm X_iX_j = \frac{1}{2}(X_i \pm X_j)^2 \in M$ and thus $\frac{1}{2}u \pm X_iX_j \in M$. Since $u \in M$, this implies $u \pm X_iX_j \in M$. □

Notation 9.1.9. We use the symbols ∇ and Hess to denote the gradient and the Hessian of a real-valued function of n real variables, respectively. For a *polynomial* $p \in \mathbb{R}[\underline{X}]$, we understand its gradient ∇p as a column vector from $\mathbb{R}[\underline{X}]^n$, i.e., as a vector of polynomials. Similarly, its Hessian Hess p is a symmetric matrix polynomial of size n, i.e., a symmetric matrix from $\mathbb{R}[\underline{X}]^{n\times n}$. Using formal partial derivatives, we more generally define $\nabla p \in R[\underline{X}]^n$ and Hess $p \in R[\underline{X}]^{n\times n}$ even for $p \in R[\underline{X}]$.

Lemma 9.1.10. Let $x \in \mathscr{O}^n$ and set $u := (X_1-x_1)^2+\ldots+(X_n-x_n)^2 \in I_x^2$. Suppose $\varphi \in S(I_x^2, \sum \mathscr{O}[\underline{X}]^2 \cap I_x^2, u)$ [→ 7.1.9] such that $\varphi|_{I_x^3} = 0$. Then there exist $v_1,\dots,v_n \in \mathbb{R}^n$ such that $\sum_{i=1}^n v_i^Tv_i = 1$ and

$$\varphi(p) = \frac{1}{2}\operatorname{st}\left(\sum_{i=1}^n v_i^T(\operatorname{Hess} p)(x)v_i\right)$$

for all $p \in I_x^2$.

Proof. As in the proof of Lemma 9.1.8, one easily reduces to the case $x = 0$.

Claim 1: $\varphi(au) = 0$ for all $a \in \mathfrak{m}$.

Explanation. Let $a \in \mathfrak{m}$. WLOG $a \geq 0$. Then $a \in \mathscr{O} \cap R_{\geq 0} = \mathscr{O}^2$ and thus $au \in \sum \mathscr{O}[\underline{X}]^2 \cap I_0^2$. This shows $\varphi(au) \geq 0$. It remains to show that $\varphi(au) \leq \frac{1}{N}$ for all $N \in \mathbb{N}$. For this purpose, fix $N \in \mathbb{N}$. Then $\frac{1}{N} - a \in \mathscr{O} \cap R_{\geq 0} = \mathscr{O}^2$ and thus $\left(\frac{1}{N}-a\right)u \in \sum \mathscr{O}[\underline{X}]^2 \cap I_0^2$. It follows that $\varphi\left(\left(\frac{1}{N}-a\right)u\right) \geq 0$, i.e., $\varphi(au) \leq \frac{1}{N}$.

Claim 2: $\varphi(aX_i^2) = 0$ for all $a \in \mathfrak{m}$ and $i \in \{1,\dots,n\}$.

Explanation. Let $a \in \mathfrak{m}$. WLOG $a \geq 0$ and thus $a \in \mathscr{O}^2$. Then

$$\sum_{i=1}^n \underbrace{\varphi(\overbrace{aX_i^2}^{\in \mathscr{O}[\underline{X}]^2 \cap I_0^2})}_{\geq 0} = \varphi(au) \overset{\text{Claim 1}}{=} 0.$$

Claim 3: $\varphi(aX_iX_j) = 0$ for all $a \in \mathfrak{m}$ and $i,j \in \{1,\dots,n\}$.

Explanation. Fix $i,j \in \{1,\dots,n\}$ and $a \in \mathfrak{m}$. If $i = j$, then we are done by Claim 2. So suppose $i \neq j$. WLOG $a \geq 0$ and thus $a \in \mathscr{O}^2$. Then

$$a(X_i^2+X_j^2 \pm 2X_iX_j) = a(X_i \pm X_j)^2 \in \mathscr{O}[\underline{X}]^2 \cap I_0^2$$

and thus $\pm 2\varphi(aX_iX_j) \underset{\text{Claim 2}}{=} \varphi(aX_i^2) + \varphi(aX_j^2) \pm 2\varphi(aX_iX_j) \geq 0$.

Claim 4: $\varphi(p) = \frac{1}{2}\operatorname{st}(\operatorname{tr}((\operatorname{Hess} p)(0)A))$ for all $p \in I_0^2$ where

$$A := \begin{pmatrix} \varphi(X_1X_1) & \dots & \varphi(X_1X_n) \\ \vdots & \ddots & \vdots \\ \varphi(X_nX_1) & \dots & \varphi(X_nX_n) \end{pmatrix}.$$

Explanation. Let $p \in I_0^2$. By $\varphi|_{I_0^3}$, we can reduce to the case $p = aX_iX_j$ with $i, j \in \{1, \dots, n\}$ and $a \in \mathscr{O}$. Using Claim 3, we can assume $a = 1$. Comparing both sides, yields the result.

Claim 5: A is psd [→ 2.3.1(b)].

Explanation. If $x \in \mathbb{R}^n$, then $x^TAx = \varphi((x_1X_1 + \dots + x_nX_n)^2) \geq 0$ since

$$(x_1X_1 + \dots + x_nX_n)^2 \in \mathbb{R}[\underline{X}]^2 \cap I_0^2 \subseteq \sum \mathscr{O}[\underline{X}]^2 \cap I_0^2.$$

By 2.3.3(c), we can choose $B \in \mathbb{R}^{n\times n}$ such that $A = B^TB$. Denote by v_i the i-th row of B for $i \in \{1, \dots, n\}$. Then by Claim 4, we get

$$\begin{aligned}\varphi(p) &= \frac{1}{2}\operatorname{st}(\operatorname{tr}((\operatorname{Hess} p)(0)A)) = \frac{1}{2}\operatorname{st}(\operatorname{tr}((\operatorname{Hess} p)(0)B^TB)) \\ &= \frac{1}{2}\operatorname{st}(\operatorname{tr}(B(\operatorname{Hess} p)(0)B^T)) = \frac{1}{2}\operatorname{st}\left(\sum_{i=1}^n v_i^T(\operatorname{Hess} p)(0)v_i\right)\end{aligned}$$

for all $p \in I_0^2$. In particular, we obtain $1 = \varphi(u) = \sum_{i=1}^n v_i^Tv_i$. □

Lemma 9.1.11. Let $\Phi\colon \mathscr{O}[\underline{X}] \to \mathbb{R}$ be a ring homomorphism. Then there is some $x \in \mathbb{R}^n$ such that $\Phi(p) = \operatorname{st}(p(x))$ for all $p \in \mathscr{O}[\underline{X}]$.

Proof. By 1.1.15, we have $\Phi|_{\mathbb{R}} = \mathrm{id}_{\mathbb{R}}$. It is easy to see that $\Phi|_{\mathfrak{m}} = 0$. Indeed, for each $N \in \mathbb{N}$ and $a \in \mathfrak{m}$, we have $\frac{1}{N} \pm a \in R_{\geq 0} \cap \mathscr{O} = \mathscr{O}^2$ and therefore $\frac{1}{N} \pm \Phi(a) \in \mathbb{R}_{\geq 0}$. Finally set

$$x := (\Phi(X_1), \dots, \Phi(X_n)) \in \mathbb{R}^n$$

and use that $\Phi|_{\mathbb{R}} = \mathrm{id}_{\mathbb{R}}$, $\Phi|_{\mathfrak{m}} = 0$ and that Φ is a ring homomorphism. □

Theorem 9.1.12. [→ 8.3.2] *Let M be an Archimedean quadratic module of $\mathscr{O}[\underline{X}]$ and set*

$$S := \{x \in \mathbb{R}^n \mid \forall p \in M : \operatorname{st}(p(x)) \geq 0\}.$$

Moreover, suppose $k \in \mathbb{N}_0$ and let $x_1, \dots, x_k \in \mathscr{O}^n$ satisfy $x_i - x_j \notin \mathfrak{m}^n$ for $i, j \in \{1, \dots, k\}$ with $i \neq j$. Set

$$u_i := (X_1 - x_{i1})^2 + \dots + (X_n - x_{in})^2 \in \mathscr{O}[\underline{X}]$$

for $i \in \{1, \dots, k\}$. Then $u := u_1 \cdots u_k$ is a unit for the $\sum \mathscr{O}[\underline{X}]^2$-module $M \cap I$ in

$$I := I_{x_1}^2 \cdots I_{x_k}^2 = I_{x_1}^2 \cap \dots \cap I_{x_k}^2$$

and for all pure states φ of $(I, M \cap I, u)$ (where I is understood as a real vector space), exactly one of the following cases occurs:

(1) *There is an* $x \in S \setminus \{\mathrm{st}(x_1), \dots, \mathrm{st}(x_k)\}$ *such that*

$$\varphi(p) = \mathrm{st}\left(\frac{p(x)}{u(x)}\right)$$

for all $p \in I$.

(2) *There is an* $i \in \{1, \dots, k\}$ *and* $v_1, \dots, v_n \in \mathbb{R}^n$ *such that* $\sum_{\ell=1}^n v_\ell^T v_\ell = 1$ *and*

$$\varphi(p) = \mathrm{st}\left(\frac{\sum_{\ell=1}^n v_\ell^T (\mathrm{Hess}\, p)(x_i) v_\ell}{2 \prod_{\substack{j=1 \\ j \neq i}}^k u_j(x_i)}\right)$$

for all $p \in I$.

Proof. The Chinese remainder theorem from commutative algebra shows that

$$I = I_{x_1}^2 \cdots I_{x_k}^2 = I_{x_1}^2 \cap \ldots \cap I_{x_k}^2$$

since I_{x_i} and I_{x_j} and thus also $I_{x_i}^2$ and $I_{x_j}^2$ are coprime for all $i, j \in \{1, \dots, k\}$ with $i \neq j$. By 9.1.8, u_i is a unit for $M \cap I_{x_i}^2$ in $I_{x_i}^2$ for each $i \in \{1, \dots, k\}$. To show that u is a unit for the cone $M \cap I$ in the real vector space I, it suffices to find for all $a_1, b_1 \in I_{x_1}, \dots, a_k, b_k \in I_{x_k}$ an $N \in \mathbb{N}$ such that $Nu + ab \in M$ where we set $a := a_1 \cdots a_k$ and $b := b_1 \cdots b_k$. Because of $Nu + ab = (Nu - \frac{1}{2}a^2 - \frac{1}{2}b^2) + \frac{1}{2}(a+b)^2$, it is enough to find $N \in \mathbb{N}$ with $Nu - a^2 \in M$ and $Nu - b^2 \in M$. By symmetry, it suffices to find $N \in \mathbb{N}$ with $Nu - a^2 \in M$. Choose $N_i \in \mathbb{N}$ with $N_i u_i - a_i^2 \in M$ for $i \in \mathbb{N}$. We now claim that $N := N_1 \cdots N_k$ does the job. Indeed, the reader shows easily by induction that actually

$$N_1 \cdots N_i u_1 \cdots u_i - a_1^2 \cdots a_i^2 \in M$$

for $i \in \{1, \dots, k\}$. Now let φ be a pure state of $(I, M \cap I, u)$. Denote by $\Phi\colon \mathscr{O}[\underline{X}] \to \mathbb{R}$ the ring homomorphism belonging to φ, i.e., $\Phi(p) = \varphi(up)$ for all $p \in \mathscr{O}[\underline{X}]$. By Lemma 9.1.11, we can choose $x \in \mathbb{R}^n$ such that

$$\Phi(p) = \mathrm{st}(p(x))$$

for all $p \in \mathscr{O}[\underline{X}]$. Since $u \in \sum \mathscr{O}[\underline{X}]^2$, we have $\Phi(M) \subseteq \mathbb{R}_{\geq 0}$ by 8.3.2. This means $x \in S$.

Now first suppose that Case (1) in the Dichotomy 8.3.2 occurs. We show that x satisfies (1). Note that $\Phi(u) \neq 0$ by 8.3.2. This means $\mathrm{st}(u_i(x)) \neq 0$ and therefore $\mathrm{st}(x) \neq \mathrm{st}(x_i)$ for all $i \in \{1, \dots, k\}$. The rest follows from 8.3.2.

Now suppose that Case (2) in the Dichotomy 8.3.2 occurs. We show that then (2) holds. Then $\prod_{i=1}^k \Phi(u_i) = \Phi(u) = 0$ because $u \in I$ and $\Phi|_I = 0$. Choose $i \in \{1, \dots, k\}$ such that $\mathrm{st}(u_i(x)) = \Phi(u_i) = 0$. Then $x = \mathrm{st}(x_i)$. Define

$$\psi\colon I_{x_i}^2 \to \mathbb{R},\ p \mapsto \varphi\left(p \prod_{\substack{j=1 \\ j \neq i}}^k u_j\right).$$

Since $u_j \in \sum \mathscr{O}[\underline{X}]^2 \cap I_{x_j}^2$ for all $j \in \{1,\dots,k\}$, it follows that $\psi \in S(I_{x_i}^2, M \cap I_{x_i}^2, u_i)$. If $p \in I_{x_i}$ and $q \in I_{x_i}^2$, then

$$\psi(pq) = \varphi\left(pq\prod_{\substack{j=1\\ j\neq i}}^{k} u_j\right) \underset{8.3.1(b)}{\overset{(**)}{=}} \Phi(p)\varphi\left(q\prod_{\substack{j=1\\ j\neq i}}^{k} u_j\right) = 0$$

since $\Phi(p) = \mathrm{st}(p(x)) = (\mathrm{st}(p))(x) = (\mathrm{st}(p))(\mathrm{st}(x_i)) = \mathrm{st}(p(x_i)) = \mathrm{st}(0) = 0$. It follows that $\psi|_{I_{x_i}^3} = 0$. We can thus apply Lemma 9.1.10 to ψ and obtain $v_1,\dots,v_n \in \mathbb{R}^n$ such that $\sum_{\ell=1}^n v_\ell^T v_\ell = 1$ and

$$\psi(p) = \frac{1}{2}\,\mathrm{st}\left(\sum_{\ell=1}^{n} v_\ell^T (\mathrm{Hess}\, p)(x_i) v_\ell\right)$$

for all $p \in I_{x_i}^2$. Because of $\mathrm{st}(x_i) \neq \mathrm{st}(x_j)$ for $j \in \{1,\dots,k\} \setminus \{i\}$, we have

$$c := \Phi\left(\prod_{\substack{j=1\\ j\neq i}}^{k} u_i\right) = \prod_{\substack{j=1\\ j\neq i}}^{k} \Phi(u_i) = \prod_{\substack{j=1\\ j\neq i}}^{k} (\mathrm{st}(u_j))(\mathrm{st}(x_i)) \neq 0.$$

Hence we obtain

$$c\varphi(p) \underset{8.3.1(b)}{\overset{(**)}{=}} \psi(p)$$

for all $p \in I$.

It only remains to show that (1) and (2) cannot occur both at the same time. If (1) holds, then we have obviously $\varphi(u^2) \neq 0$. If (2) holds, then $\varphi(u^2) = 0$ since $\mathrm{Hess}(u^2)(x_i) = 0$ for all $i \in \{1,\dots,k\}$ as one easily shows. □

Lemma 9.1.13. For all $x \in \mathscr{O}^n$, we have

$$I_x^2 = \{p \in \mathscr{O}[\underline{X}] \mid p(x) = 0, \nabla p(x) = 0\}.$$

Proof. For $x = 0$ it is easy. One reduces the general case to the case $x = 0$ as in the proof of 9.1.8. □

Theorem 9.1.14. *Let M be an Archimedean quadratic module of $\mathscr{O}[\underline{X}]$ and set*

$$S := \{x \in \mathbb{R}^n \mid \forall p \in M : \mathrm{st}(p(x)) \geq 0\}.$$

Moreover, suppose $k \in \mathbb{N}_0$ and let $x_1,\dots,x_k \in \mathscr{O}^n$ satisfy $x_i - x_j \notin \mathfrak{m}^n$ for $i,j \in \{1,\dots,k\}$ with $i \neq j$. Let $f \in \mathscr{O}[\underline{X}]$ such that

$$f(x_1) = \dots = f(x_k) = 0 \qquad \textit{and} \qquad \nabla f(x_1) = \dots = \nabla f(x_k) = 0.$$

Suppose

$$\operatorname{st}(f(x)) > 0$$

for all $x \in S \setminus \{\operatorname{st}(x_1), \ldots, \operatorname{st}(x_k)\}$ *and*

$$\operatorname{st}(v^T (\operatorname{Hess} f)(x_i) v) > 0$$

for all $i \in \{1, \ldots, k\}$ *and* $v \in \mathbb{R}^n \setminus \{0\}$. *Then* $f \in M$.

Proof. Define I and u as in Theorem 9.1.12. By Lemma 9.1.13, we have $f \in I$. We will apply 7.3.20 to the real vector space I, the cone $M \cap I$ in I and the unit u for $M \cap I$. From Theorem 9.1.12, we see indeed easily that $\varphi(f) > 0$ for all $\varphi \in \operatorname{extr} S(I, M \cap I, u)$. □

Corollary 9.1.15. *Let M be an Archimedean quadratic module of $\mathscr{O}[\underline{X}]$ and set*

$$S := \{x \in \mathbb{R}^n \mid \forall p \in M : \operatorname{st}(p(x)) \geq 0\}.$$

Moreover, let $k \in \mathbb{N}_0$ and $x_1, \ldots, x_k \in \mathscr{O}^n$ such that the standard parts $\operatorname{st}(x_1), \ldots, \operatorname{st}(x_k) \in \mathbb{R}^n$ are pairwise distinct and lie in the interior of S. Let $f \in \mathscr{O}[\underline{X}]$ such that

$$f(x_1) = \ldots = f(x_k) = 0 \qquad \text{and} \qquad \nabla f(x_1) = \ldots = \nabla f(x_k) = 0.$$

Define $u \in \mathscr{O}[\underline{X}]$ as in Theorem 9.1.12. Suppose there is $\varepsilon \in \mathbb{R}_{>0}$ such that

$$f \geq \varepsilon u \text{ on } S.$$

Then $f \in M$.

Proof. By 9.1.14, we have to show:

(a) $\forall x \in S \setminus \{\operatorname{st}(x_1), \ldots, \operatorname{st}(x_k)\} : \operatorname{st}(f(x)) > 0$

(b) $\forall i \in \{1, \ldots, k\} : \forall v \in \mathbb{R}^n \setminus \{0\} : \operatorname{st}(v^T (\operatorname{Hess} f)(x_i) v) > 0$

It is easy to show (a). To show (b), fix $i \in \{1, \ldots, k\}$. Because of $f - \varepsilon u \geq 0$ on S and

$$(f - \varepsilon u)(x_i) = f(x_i) - \varepsilon u(x_i) = 0 - 0 = 0,$$

$\operatorname{st}(x_i)$ is a local minimum of $\operatorname{st}(f - \varepsilon u) \in \mathbb{R}[\underline{X}]$ on $\mathbb{R}^n$. From elementary analysis, we know therefore that $(\operatorname{Hess} \operatorname{st}(f - \varepsilon u))(\operatorname{st}(x_i))$ is psd. Because of $u_i(x_i) = 0$ and $\nabla u_i(x_i) = 0$, we get

$$\operatorname{Hess} u(x_i) = \left(\prod_{\substack{j=1 \\ j \neq i}}^{k} u_j(x_i) \right) \operatorname{Hess} u_i(x_i) = 2 \left(\prod_{\substack{j=1 \\ j \neq i}}^{k} u_j(x_i) \right) I_n.$$

Therefore

$$\operatorname{st}(v^T (\operatorname{Hess} f)(x_i) v) \geq \varepsilon \operatorname{st}(v^T (\operatorname{Hess} u)(x_i) v) = 2\varepsilon v^T v \operatorname{st} \left(\prod_{\substack{j=1 \\ j \neq i}}^{k} u_j(x_i) \right) > 0$$

for all $v \in \mathbb{R}^n \setminus \{0\}$. □

Corollary 9.1.16. *Let $n, m \in \mathbb{N}_0$ and suppose $g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]$ generate an Archimedean quadratic module in $\mathbb{R}[\underline{X}]$ [→ 8.1.13]. Set*

$$S := \{x \in R^n \mid g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}.$$

Moreover, let $k \in \mathbb{N}_0$ and $x_1, \ldots, x_k \in \mathscr{O}^n$ and $\varepsilon \in \mathbb{R}_{>0}$ such that the sets $x_1 + \varepsilon B, \ldots, x_k + \varepsilon B$ are pairwise disjoint and all contained in S where [→ 6.1.10]

$$B := \{x \in R^n \mid \|x\|_2 < 1\} \subseteq \mathscr{O}^n.$$

Define $u \in \mathscr{O}[\underline{X}]$ as in Theorem 9.1.12. Let $f \in \mathscr{O}[\underline{X}]$ such that $f \geq \varepsilon u$ on S and

$$f(x_1) = \ldots = f(x_k) = 0.$$

Then f lies in the quadratic module generated by $g_1, \ldots, g_m$ in $\mathscr{O}[\underline{X}]$.

Proof. This follows easily from 9.1.15 once we show that

$$\nabla f(x_1) = \ldots = \nabla f(x_k) = 0.$$

Since $f \geq \varepsilon u \geq 0$ on S and thus $f \geq 0$ on $x_i + \varepsilon B$ for all $i \in \{1, \ldots, k\}$, it suffices to prove the following: If $p \in R[\underline{X}]$, $x \in R^n$, $\delta \in R_{>0}$ such that $p \geq 0$ on $x + \delta B$ and $p(x) = 0$, then $\nabla p(x) = 0$. To see this, we employ the Tarski principle [→ 1.8.19]: For each fixed number of variables n and $d \in \mathbb{N}$, the class of all $R \in \mathscr{R}$ [→ 1.8.3] such that this holds true for all $p \in R[\underline{X}]_d$ is obviously a 0-ary semialgebraic class by real quantifier elimination. By elementary analysis, $\mathbb{R}$ is an element of this class. We conclude thus by 1.8.5. □

9.2 Degree bounds and quadratic modules

Definition 9.2.1. Let $d, m \in \mathbb{N}_0$, $g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]$ and set $g_0 := 1 \in \mathbb{R}[\underline{X}]$. For $i \in \{0, \ldots, m\}$, set $r_i := \frac{d - \deg g_i}{2}$ if $g_i \neq 0$ and $r_i := -\infty$ if $g_i = 0$. Then we denote by $M(g_1, \ldots, g_m)$ the quadratic module generated by $g_1, \ldots, g_m$ in $\mathbb{R}[\underline{X}]$. Moreover, we define the *d-truncated quadratic module* $M_d(g_1, \ldots, g_m)$ associated to $g_1, \ldots, g_m$ by

$$M_d(g_1, \ldots, g_m) := \left\{ \sum_{i=0}^{m} \sum_j p_{ij}^2 g_i \mid p_{ij} \in \mathbb{R}[\underline{X}]_{r_i} \right\} \subseteq M(g_1, \ldots, g_m) \cap \mathbb{R}[\underline{X}]_d.$$

Remark 9.2.2. Let $m \in \mathbb{N}_0$ and $g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]$. Set again $g_0 := 1 \in \mathbb{R}[\underline{X}]$.

(a) $M(g_1, \ldots, g_m) = \bigcup_{d \in \mathbb{N}_0} M_d(g_1, \ldots, g_m)$

(b) For all $d \in \mathbb{N}_0$,

$$M_d(g_1, \ldots, g_m) = \sum_{i=0}^{m} \left(\left(\sum \mathbb{R}[\underline{X}]^2 g_i \right) \cap \mathbb{R}[\underline{X}]_d \right)$$

by 2.2.4(b).

(c) In general, the inclusion $M_d(g_1,\dots,g_m) \subseteq M(g_1,\dots,g_m) \cap \mathbb{R}[\underline{X}]_d$ is proper as 5.4.8 shows. In fact, the validity of Schmüdgen's and Putinar's Positivstellensätze 4.3.5 and 8.2.14 strongly relies on this.

Theorem 9.2.3 (Putinar's Positivstellensatz with zeros and degree bounds)**.** *Let $n, m \in \mathbb{N}_0$ and $g_1,\dots,g_m \in \mathbb{R}[\underline{X}]$ such that $M(g_1,\dots,g_m)$ is Archimedean. Set*

$$B := \{x \in \mathbb{R}^n \mid \|x\| < 1\} \qquad \text{and}$$
$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0,\dots,g_m(x) \geq 0\}.$$

Moreover, let $k \in \mathbb{N}_0$, $N \in \mathbb{N}$ and $\varepsilon \in \mathbb{R}_{>0}$. Then there exists

$$d \in \mathbb{N}_0$$

such that for all $f \in \mathbb{R}[\underline{X}]_N$ with all coefficients in $[-N,N]_{\mathbb{R}}$ and $\#\{x \in S \mid f(x) = 0\} = k$, we have: Denoting by $x_1,\dots,x_k$ the distinct zeros of f on S, if the sets $x_1+\varepsilon B,\dots,x_k+\varepsilon B$ are pairwise disjoint and contained in S and if we have $f \geq \varepsilon u$ on S where $u \in \mathbb{R}[\underline{X}]$ is defined as in Theorem 9.1.12, then

$$f \in M_d(g_1,\dots,g_m).$$

Proof. (cf. the proof of Theorem 5.4.5) Set $\nu := \dim \mathbb{R}[\underline{X}]_N$. For each $d \in \mathbb{N}_0$, the class S_d of all pairs (R,a) where R is a real closed extension field of $\mathbb{R}$ and $a \in R^\nu$ such that the following holds is obviously a ν-ary $\mathbb{R}$-semialgebraic class [→ 1.8.3]: If $a \in [-N,N]_R^\nu$ and if a is the vector of coefficients (in a certain fixed order) of a polynomial $f \in R[\underline{X}]_N$ with exactly k zeros $x_1,\dots,x_k$ on $S' := \{x \in R^n \mid g_1(x) \geq 0,\dots,g_m(x) \geq 0\}$, then at least one of the following conditions (a), (b) and (c) is fulfilled:

(a) The sets $x_1+\varepsilon B',\dots,x_k+\varepsilon B'$ are not pairwise disjoint or not all contained in S' where $B' := \{x \in R^n \mid \|x\|_2 < 1\}$.

(b) $f \geq \varepsilon u$ on S' is violated where $u \in R[\underline{X}]$ is defined as in Theorem 9.1.12.

(c) f is not a sum of d elements from $R[\underline{X}]$ where each term in the sum is of degree at most d and is of the form $p^2 g_i$ with $p \in R[\underline{X}]$ and $i \in \{0,\dots,m\}$ where $g_0 := 1 \in R[\underline{X}]$.

Set $\mathscr{E} := \{S_d \mid d \in \mathbb{N}_0\}$ and observe that $\forall d_1,d_2 \in \mathbb{N}_0 : \exists d_3 \in \mathbb{N}_0 : S_{d_1} \cup S_{d_2} \subseteq S_{d_3}$ (take $d_3 := \max\{d_1,d_2\}$). By 9.1.16, we have $\bigcup \mathscr{E} = \mathscr{R}_\nu$. Now 5.4.2 yields $S_d = \mathscr{R}_\nu$ for some $d \in \mathbb{N}_0$. □

Corollary 9.2.4 (Putinar's Positivstellensatz with degree bounds [NS])**.** [→ 8.2.14] *Let $n, m \in \mathbb{N}_0$ and $g_1,\dots,g_m \in \mathbb{R}[\underline{X}]$ such that $M(g_1,\dots,g_m)$ is Archimedean. Set*

$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0,\dots,g_m(x) \geq 0\}.$$

Moreover, let $N \in \mathbb{N}$ and $\varepsilon \in \mathbb{R}_{>0}$. Then there exists

$$d \in \mathbb{N}_0$$

such that for all $f \in \mathbb{R}[\underline{X}]_N$ with all coefficients in $[-N,N]_{\mathbb{R}}$ and with $f \geq \varepsilon$ on S, we have

$$f \in M_d(g_1,\dots,g_m).$$

Proposition 9.2.5. *Suppose $S \subseteq \mathbb{R}^n$ is compact, $x_1, \ldots, x_k \in S^\circ$ are pairwise distinct, $u \in \mathbb{R}[\underline{X}]$ is defined as in Theorem 9.1.12 and $f \in \mathbb{R}[\underline{X}]$ with $f(x_1) = \ldots = f(x_k) = 0$. Then the following are equivalent:*

(a) *$f > 0$ on $S \setminus \{x_1, \ldots, x_k\}$ and $\operatorname{Hess} f(x_1), \ldots, \operatorname{Hess} f(x_k)$ are pd.*

(b) *There is some $\varepsilon \in \mathbb{R}_{>0}$ such that $f \geq \varepsilon u$ on S.*

Proof. (b) $\Longrightarrow$ (a) is easy to show (cf. the proof of 9.1.15).

(a) $\Longrightarrow$ (b) It is easy to show that one can WLOG assume that $S = \dot{\bigcup}_{i=1}^{k}(x_i + \varepsilon B)$ for some $\varepsilon > 0$ where B is the closed unit ball in $\mathbb{R}^n$. Then one finds easily an Archimedean quadratic module M of $\mathbb{R}[\underline{X}]$ such that

$$S = \{x \in \mathbb{R}^n \mid \forall p \in M : p(x) \geq 0\}.$$

A strengthened version of Theorem 9.1.14 now yields $f - \varepsilon u \in M$ for some $\varepsilon \in \mathbb{R}_{>0}$ and thus $f - \varepsilon u \geq 0$ on S. One gets this strengthened version of Theorem 9.1.14 by applying (a) $\Longrightarrow$ (c) from 7.3.19 instead of 7.3.20 in its proof. Alternatively, we leave it as an exercise to the reader to give a direct proof using only basic multivariate analysis. □

Corollary 9.2.6 (Putinar's Positivstellensatz with zeros [S1])**.** *Let $g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]$ such that $M(g_1, \ldots, g_m)$ is Archimedean. Set*

$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}.$$

Moreover, suppose $k \in \mathbb{N}_0$ and $x_1, \ldots, x_k \in S^\circ$ are pairwise distinct. Let $f \in \mathbb{R}[\underline{X}]$ such that $f(x_1) = \ldots = f(x_k) = 0$, $f > 0$ on $S \setminus \{x_1, \ldots, x_k\}$ and $\operatorname{Hess} f(x_1), \ldots, \operatorname{Hess} f(x_k)$ are pd. Then

$$f \in M(g_1, \ldots, g_m).$$

Proof. This follows from 9.2.3 by Proposition 9.2.5. □

Remark 9.2.7. Because of Proposition 9.2.5, Theorem 9.2.3 is really a quantitative version of Corollary 9.2.6.

Remark 9.2.8. (a) In Condition (c) from the proof of Theorem 9.2.3, we speak of "a sum of d elements" instead of "a sum of elements" (which would in general be strictly weaker). Our motivation to do this was that this is the easiest way to make sure that we can formulate (c) in a "semialgebraic way". A second motivation could have been to formulate Theorem 9.2.3 in stronger way, namely by letting d be a bound not only on the degree of the quadratic module representation but also on the number of terms in it. This second motivation is however not interesting because we get also from the Gram matrix method 2.6.1 a bound on this number of terms (a priori bigger than d but after readjusting d we can again assume it to be d). We could have used the Gram matrix method already to see that "a sum of elements" (instead of "a sum of d elements") can also be expressed semialgebraically.

(b) We could strengthen condition (c) from the proof of Theorem 9.2.3, by writing "with $p \in R[\underline{X}]$ all of whose coefficients lie in $[-d,d]_R$" instead of just "with $p \in R[\underline{X}]$". Then $\bigcup \mathscr{E} = \mathscr{R}_v$ would still hold since Corollary 9.1.16 states that f lies in the quadratic module generated by $g_1,\dots,g_m$ even in $\mathscr{O}[\underline{X}]$ not just in $\mathbb{R}[\underline{X}]$. This would lead to a real strengthening of Theorem 9.2.3, namely we could ensure that d is a bound not only on the degree of the quadratic module representation but also on the size of the coefficients in it. However, we do currently not know of any application of this and therefore renounced to carry this out.

9.3 Concavity and Lagrange multipliers

Definition 9.3.1. [→ 2.4.1] Suppose $(K,\leq)$ is an ordered field, $A \subseteq K^n$ is convex and $f \in K[\underline{X}]$. Then f is called $\left\{\begin{matrix} convex \\ concave \end{matrix}\right\}$ on A if for all $x,y \in A$ and $\lambda \in [0,1]_K$, we have

$$f(\lambda x + (1-\lambda)y) \left\{\begin{matrix} \leq \\ \geq \end{matrix}\right\} \lambda f(x) + (1-\lambda) f(y),$$

Exercise 9.3.2. Suppose $(K,\leq)$ is an ordered field, $A \subseteq K^n$ is convex and $f \in K[\underline{X}]$. Then the following are equivalent:

(a) f is $\left\{\begin{matrix} \text{convex} \\ \text{concave} \end{matrix}\right\}$ on A.

(b) The $\left\{\begin{matrix} \text{epigraph} \\ \text{hypograph} \end{matrix}\right\}$ $\left\{(x,y) \in \mathbb{R}^n \times \mathbb{R} \mid f(x) \left\{\begin{matrix} \leq \\ \geq \end{matrix}\right\} y\right\}$ is convex.

(c) For all $\ell \in \mathbb{N}$, $x_1,\dots,x_\ell \in A$ and $\lambda_1,\dots,\lambda_\ell \in K_{\geq 0}$ with $\lambda_1 + \ldots + \lambda_\ell = 1$, we have

$$f\left(\sum_{i=1}^{\ell} \lambda_i x_i\right) \left\{\begin{matrix} \leq \\ \geq \end{matrix}\right\} \sum_{i=1}^{\ell} \lambda_i f(x_i).$$

Lemma 9.3.3 (Existence of Lagrange multipliers)**.** Let $u \in \mathbb{R}^n$, $f, g_1,\dots,g_m \in \mathbb{R}[\underline{X}]$, let U be a convex subset of $\mathbb{R}^n$ containing u and set

$$S := \{x \in U \mid g_1(x) \geq 0,\dots,g_m(x) \geq 0\}.$$

Suppose f is convex on U, $g_1,\dots,g_m$ are concave on U,

$$f(u) = g_1(u) = \ldots = g_m(u) = 0,$$

S has nonempty interior and $f \geq 0$ on S. Then there are $\lambda_1,\dots,\lambda_m \in \mathbb{R}_{\geq 0}$ such that $f - \sum_{i=1}^m \lambda_i g_i \geq 0$ on U.

Proof. WLOG $g_i \neq 0$ for all $i \in \{1, \dots, m\}$. Consider the set

$$A := \operatorname{conv}\{(-f(x), g_1(x), \dots, g_m(x)) \mid x \in U\} \subseteq \mathbb{R}^{m+1}.$$

Claim: $A \cap \mathbb{R}^{m+1}_{>0} = \emptyset$

Explanation. Assume that $x_1, \dots, x_\ell \in U$ and $\lambda_1, \dots, \lambda_\ell \in \mathbb{R}_{\geq 0}$ with $\lambda_1 + \ldots + \lambda_\ell = 1$ such that

$$w := \sum_{j=1}^{\ell} \lambda_i(-f(x_j), g_1(x_j), \dots, g_m(x_j)) \in \mathbb{R}^{m+1}_{>0}.$$

Setting $x := \sum_{j=1}^{\ell} \lambda_j x_j \in U$, we get

$$(*) \qquad f(x) \leq \sum_{j=1}^{\ell} \lambda_j f(x_j) < 0$$

since f is convex on U and

$$(**) \qquad g_i(x) \geq \sum_{j=1}^{\ell} \lambda_j g_i(x_j) > 0$$

since g_i is concave on U for $i \in \{1, \dots, m\}$. Hence $x \in S$ by $(**)$ but $f(x) < 0$ by $(*)$ $\lightning$.

By the separation theorem for finite-dimensional vector spaces [$\to$ 7.4.4], we find a linear $\varphi\colon \mathbb{R}^{m+1} \to \mathbb{R}$ such that $\varphi \neq 0$ and $\varphi(x) \leq \varphi(y)$ for all $x \in A$ and $y \in \mathbb{R}^{m+1}_{>0}$. By continuity, it follows that $\varphi(x) \leq \varphi(0) = 0$ for all $x \in A$ and $0 = \varphi(0) \leq \varphi(y)$ for all $y \in \mathbb{R}^{m+1}_{\geq 0}$ (use that $0 \in A$). Choosing $\lambda := (\lambda_0, \dots, \lambda_m) \in \mathbb{R}^{m+1} \setminus \{0\}$ such that $\varphi(y_0, \dots, y_m) = \sum_{i=0}^{m} \lambda_i y_i$ for all $(y_0, \dots, y_m) \in \mathbb{R}^{m+1}$, we thus have $(\lambda_0, \dots, \lambda_m) \in \mathbb{R}^{m+1}_{\geq 0}$ due to $\varphi(\mathbb{R}^{m+1}_{\geq 0}) \subseteq \mathbb{R}_{\geq 0}$ and $\lambda_0 f - \sum_{i=1}^{m} \lambda_i g_i \geq 0$ on U because of $\varphi(A) \subseteq \mathbb{R}_{\leq 0}$. It only remains to show $\lambda_0 \neq 0$.

So assume $\lambda_0 = 0$. Then $\sum_{i=1}^{m} \lambda_i g_i \leq 0$ on U and hence on S. For all

$$i \in I := \{i \in \{1, \dots, m\} \mid \lambda_i \neq 0\} \overset{\lambda \neq 0}{\neq} \emptyset,$$

it follows that $g_i = 0$ on S and hence $g_i = 0$ since S has nonempty interior. This is impossible because $I \neq \emptyset$ and $g_i \neq 0$ for all $i \in \{1, \dots, m\}$. □

Definition 9.3.4. [$\to$ 2.3.1(b)] Let $(K, \leq)$ be an ordered field and $A \in SK^{n \times n}$. We write $A \begin{Bmatrix} \succeq \\ \succ \end{Bmatrix} 0$ to express that A is psd, i.e., A is symmetric and $x^T M x \begin{Bmatrix} \geq \\ > \end{Bmatrix} 0$ for all $x \in \begin{Bmatrix} K^n \\ K^n \setminus \{0\} \end{Bmatrix}$. If $B \in K^{n \times n}$ is another matrix, we write $A \begin{Bmatrix} \succeq \\ \succ \end{Bmatrix} B$ or $B \begin{Bmatrix} \preceq \\ \prec \end{Bmatrix} A$ to express that $A - B \begin{Bmatrix} \succeq \\ \succ \end{Bmatrix} 0$. We say that A is $\begin{Bmatrix} \textit{negative semidefinite (nsd)} \\ \textit{negative definite (nd)} \end{Bmatrix}$ if $A \preceq 0$.

Definition 9.3.5. Suppose $(K, \leq)$ be an ordered field and $f \in K[\underline{X}]$. If $x \in K^n$, then we call f *strictly* $\left\{\begin{matrix} \text{convex} \\ \text{concave} \end{matrix}\right\}$ at x if $(\operatorname{Hess} f)(x) \left\{\begin{matrix} \succ \\ \prec \end{matrix}\right\} 0$ and *strictly* $\left\{\begin{matrix} \text{quasiconvex} \\ \text{quasiconcave} \end{matrix}\right\}$ at x if

$$((\nabla f)(x))^T v = 0 \implies v^T (\operatorname{Hess} f)(x) v \left\{\begin{matrix} > \\ < \end{matrix}\right\} 0$$

for all $v \in K^n \setminus \{0\}$. If $A \subseteq K^n$, we call f *strictly (quasi-)*$\left\{\begin{matrix} \text{convex} \\ \text{concave} \end{matrix}\right\}$ on A is f is strictly (quasi-)$\left\{\begin{matrix} \text{convex} \\ \text{concave} \end{matrix}\right\}$ at every point of A.

Proposition 9.3.6. $\mathbb{R}^{n\times n}_{\succeq 0} := \{A \in \mathbb{R}^{n\times n} \mid A \succeq 0\}$ *is a cone in the vector space* $S\mathbb{R}^{n\times n}$ *whose interior* [→ 7.4.18] *is* $\mathbb{R}^{n\times n}_{\succ 0} := \{A \in \mathbb{R}^{n\times n} \mid A \succ 0\}$.

Proof. Equip $S\mathbb{R}^{n\times n}$ with the norm defined by

$$\|A\| := \max_{\substack{x\in\mathbb{R}^n \\ \|x\|\leq 1}} x^T A x$$

for $A \in S\mathbb{R}^{n\times n}$. By 7.2.2(c), this norm (as any other norm) induces the unique vector space topology on $S\mathbb{R}^{n\times n}$.

If A is an interior point of $\mathbb{R}^{n\times n}_{\succeq 0}$, then there exists $\varepsilon \in \mathbb{R}_{>0}$ such that $A - \varepsilon I_n \succeq 0$ and thus $A \succeq \varepsilon I_n \succ 0$.

Conversely, let $A \in \mathbb{R}^{n\times n}$ satisfy $A \succ 0$. We show that A is an interior point of $\mathbb{R}^{n\times n}_{\succeq 0}$. By 2.3.3, the lowest eigenvalue ε of A is nonnegative since $A \succeq 0$. Actually, we have even $\varepsilon > 0$ since A has trivial kernel due to $A \succ 0$. Now $A - \varepsilon I_n$ has only nonnegative eigenvalues and thus $A - \varepsilon I_n \succeq 0$ by 2.3.3. It suffices to show that a ball around A with radius ε in $S\mathbb{R}^{n\times n}$ is contained in $\mathbb{R}^{n\times n}_{\succeq 0}$. For this purpose, let $B \in S\mathbb{R}^{n\times n}$ with $\|B - A\| \leq \varepsilon$ and fix $x \in \mathbb{R}^n$ with $\|x\| = 1$. We have to show that $x^T B x \geq 0$. But we have $x^T B x = x^T A x + x^T (B - A) x \geq x^T A x - \|B - A\| \geq \varepsilon x^T I_n x - \varepsilon = 0$. □

Lemma 9.3.7. Suppose $g \in \mathbb{R}[\underline{X}]$ and $S \subseteq \mathbb{R}^n$ is compact. Then the following are equivalent:

(a) g is strictly quasiconcave on S.

(b) $\exists \lambda \in \mathbb{R} : \forall x \in S : \lambda (\nabla g(x))(\nabla g(x))^T \succ (\operatorname{Hess} g)(x)$

Proof. (b) $\implies$ (a) WLOG $S = \{x\}$. If $\lambda \in \mathbb{R}$ such that $\lambda (\nabla g(x))(\nabla g(x))^T \succ \operatorname{Hess} g$ and $v \in K^n \setminus \{0\}$ such that $((\nabla g)(x))^T v = 0$, then

$$0 = \lambda v^T (\nabla g(x))(\nabla g(x))^T v > v^T (\operatorname{Hess} g)(x) v.$$

(a) $\implies$ (b) Consider the unit ball $U := \{x \in \mathbb{R}^n \mid \|x\| = 1\}$. It is easy to show that (a) is equivalent to

$$\forall x \in S : \forall v \in U : \exists \lambda \in \mathbb{R} : \lambda v^T (\nabla g(x))(\nabla g(x))^T v > v^T (\operatorname{Hess} g)(x) v.$$

Suppose that this holds. We have to show

$$\exists \lambda \in \mathbb{R} : \forall x \in S : \forall v \in U : \lambda v^T (\nabla g(x))(\nabla g(x))^T v > v^T (\operatorname{Hess} g)(x) v.$$

For this purpose, we will use the compactness of $S \times U$ which is due to Tikhonov's theorem 5.1.18. For all $(x, v) \in S \times U$, we choose $\lambda_{(x,v)} \in \mathbb{R}$ such that

$$\lambda_{(x,v)} v^T (\nabla g(x))(\nabla g(x))^T v > v^T (\operatorname{Hess} g)(x) v.$$

Then $S \times U = \bigcup_{(x,v)\in S\times U} A_{(x,v)}$ where

$$A_{(x,v)} := \{(y, u) \in S \times U \mid \lambda_{(y,u)} u^T (\nabla g(y))(\nabla g(y))^T u > u^T (\operatorname{Hess} g)(y) u\}$$

is open in $S \times U$ for each $(x, v) \in S \times U$. WLOG $n > 0$. By compactness, there is a nonempty finite subset $F \subseteq S \times U$ such that $S \times U = \bigcup_{(x,v)\in F} A_{(x,v)}$. Now

$$\lambda := \max\{\lambda_{(x,v)} \mid (x, v) \in F\}$$

will do the job. □

Lemma 9.3.8. Let $g \in \mathbb{R}[\underline{X}]$. If g is strictly $\left\{\begin{array}{c}\text{concave}\\ \text{quasiconcave}\end{array}\right\}$ at $x \in \mathbb{R}^n$, then there is a neighborhood A of x such that g is $\left\{\begin{array}{c}\text{strictly concave}\\ \text{strictly quasiconcave}\end{array}\right\}$ on A.

Proof. The first statement follows from the openness of $\mathbb{R}^{n\times n}_{\succ 0}$ [→ 9.3.6] by the continuity of $\mathbb{R}^n \to S\mathbb{R}^{n\times n}$, $x \mapsto (\operatorname{Hess} g)(x)$. The second statement follows similarly by using the equivalence of (a) and (b) in 9.3.7. □

Lemma 9.3.9. If $A \subseteq \mathbb{R}^n$ be convex and $g \in \mathbb{R}[\underline{X}]$ is strictly concave on A, then g is concave on A.

Proof. One easily reduces to the case $n = 1$. Hence let $x, y \in A \subseteq \mathbb{R}$ with $x < y$ and $\lambda \in (0, 1)_{\mathbb{R}}$. We have to show that $g(z) \geq \lambda g(x) + (1 - \lambda) g(y)$ where $z := \lambda x + (1 - \lambda) y$. This is equivalent to

$$\frac{g(z) - g(x)}{z - x} \geq \frac{g(y) - g(z)}{y - z}$$

since $y - z = \lambda(y - x)$ and $z - x = (1 - \lambda)(y - x)$. By applying 1.4.18 twice to g, it suffices thus to show that g' is anti-monotonic on $[x, y]_{\mathbb{R}}$ [→ 1.4.19(b)] since $x < z < y$. By 1.4.20, this follows from the nonpositivity of g'' on $[x, y]_{\mathbb{R}}$ (actually g'' is even *negative* on $A \supseteq [x, y]_{\mathbb{R}}$). □

Exercise 9.3.10. Let R be a real closed field. Let $g \in R[\underline{X}]$ and $x \in R^n$ with $g(x) = 0$. Then

$$(\nabla(g(1 - g)^k))(x) = (\nabla g)(x) \qquad \text{and}$$
$$(\operatorname{Hess}(g(1 - g)^k))(x) = (\operatorname{Hess} g - 2k(\nabla g)(\nabla g)^T)(x).$$

Lemma 9.3.11. Suppose $g \in \mathbb{R}[\underline{X}]$, $S \subseteq \mathbb{R}^n$ is compact and $g = 0$ on S. Then the following are equivalent:

(a) g is strictly quasiconcave on S.

(b) There exists $k \in \mathbb{N}$ such that $g(1-g)^k$ is strictly concave on S.

(c) There exists $k \in \mathbb{N}$ such that for all $\ell \in \mathbb{N}$ with $\ell \geq k$, we have that $g(1-g)^\ell$ is strictly concave on S.

Proof. Combine 9.3.10 and 9.3.7. □

Definition 9.3.12. [→ 5.2.5] Let M be a topological space and $A \subseteq M$. We call

$$\partial A := \overline{A} \setminus A^\circ = \overline{A} \cap \overline{M \setminus A}$$

the *boundary* of A.

Notation and Terminology 9.3.13. Let $S \subseteq \mathbb{R}^n$. We call

$$\operatorname{convbd} S := S \cap \partial \operatorname{conv} S$$

the *convex boundary* of S. Obviously,

$$\operatorname{convbd} S = \{x \in S \mid \forall U \in \mathscr{U}_x : U \not\subseteq \operatorname{conv} S\}.$$

We say that S *has nonempty interior near its convex boundary* if $\operatorname{convbd} S \subseteq \overline{S^\circ}$.

Proposition 9.3.14. *Let $S \subseteq \mathbb{R}^n$. Then*

$$\operatorname{convbd} S = \{u \in S \mid \exists \varphi \in (\mathbb{R}^n)^* \setminus \{0\} : \forall x \in S : \varphi(u) \leq \varphi(x)\}.$$

Proof. "$\supseteq$" Let $u \in S$ and $\varphi \in (\mathbb{R}^n)^* \setminus \{0\}$ such that $\forall x \in S : \varphi(u) \leq \varphi(x)$. Then even $\forall x \in \operatorname{conv} S : \varphi(u) \leq \varphi(x)$. Choose $v \in \mathbb{R}^n$ such that $\varphi(v) > 0$. Then $\varphi(u - \varepsilon v) < \varphi(u)$ and hence $u - \varepsilon v \notin \operatorname{conv} S$ for each $\varepsilon \in \mathbb{R}_{>0}$. It follows that every neighborhood of u intersects the complement of $\operatorname{conv} S$. Hence $u \in \operatorname{convbd} S$.

"$\subseteq$" If $\dim \operatorname{conv} S < n$ [→ 7.4.7], we have $\partial \operatorname{conv} S = \overline{\operatorname{conv} S}$ and hence $\operatorname{convbd} S = S$ and one easily finds $\varphi \in (\mathbb{R}^n)^* \setminus \{0\}$ that is constant on $\operatorname{conv} S$. So now suppose that $\dim \operatorname{conv} S = n$. Let $u \in \operatorname{convbd} S$. By Theorem 7.4.17, we get an exposed face F of $\operatorname{conv} S$ with $\dim F < n$ and $u \in F$. Choose $\varphi \colon \mathbb{R}^n \to \mathbb{R}$ linear such that

$$F = \{y \in \operatorname{conv} S \mid \forall x \in \operatorname{conv} S : \varphi(y) \leq \varphi(x)\}.$$

Since $\dim F < n$, we have obviously $\varphi \neq 0$. □

Notation 9.3.15. For $g \in \mathbb{R}[\underline{X}]$, set $Z(g) := \{x \in \mathbb{R}^n \mid g(x) = 0\}$.

Lemma 9.3.16. Let $B \subseteq \mathbb{R}^n$ be a closed ball in $\mathbb{R}^n$ and suppose that $g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]$ are strictly quasiconcave on B. Then the following hold:

(a) $S := \{x \in B \mid g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}$ is convex.

(b) Every linear form from $\mathbb{R}[\underline{X}] \setminus \{0\}$ [→ 1.6.1(a)] has at most one minimizer on S.

(c) Let u be a minimizer of the linear form $f \in \mathbb{R}[\underline{X}] \setminus \{0\}$ on S and set

$$I := \{i \in \{1, \ldots, m\} \mid g_i(u) = 0\}.$$

Then u is also minimizer of f on $S' := \{x \in B \mid \forall i \in I : g_i(x) \geq 0\} \supseteq S$.

Proof. (a) Let $x, y \in S$ with $x \neq y$ and $i \in \{1, \ldots, m\}$. The polynomial

$$f := g_i(Tx + (1-T)y) \in \mathbb{R}[T]$$

attains a minimum a on $[0,1]_{\mathbb{R}}$ [→ 7.1.19]. We have to show $a \geq 0$. Because of $f(0) = g_i(y) \geq 0$ and $f(1) = g_i(x) \geq 0$, it is enough to show that this minimum is not attained in a point $t \in (0,1)_{\mathbb{R}}$. Assume it is. Then $f'(t) = 0$, i.e., $((\nabla g_i)(z))^T v = 0$ for $z := tx + (1-t)y$ and $v := x - y \neq 0$. Since $z \in B$ and hence g_i is strictly quasiconcave at z, it follows that $v^T((\operatorname{Hess} g_i)(z))v < 0$, i.e., $f''(t) < 0$. Then $f < a$ on a neighborhood of t [→ 1.5.3(b)] ↯.

(b) Suppose x and y are minimizers of the linear form $f \in \mathbb{R}[\underline{X}] \setminus \{0\}$ on S. Then $x, y \in \operatorname{convbd} S$ by 9.3.14. Since f is linear, it is constant on $\operatorname{aff}\{x, y\}$. Hence even

$$\operatorname{conv}\{x,y\} \overset{\text{(a)}}{\subseteq} \operatorname{aff}\{x,y\} \cap S \overset{9.3.14}{\subseteq} \operatorname{convbd} S \overset{\text{(a)}}{=} S \cap \partial S = S \cap (\overline{S} \setminus S^\circ) \overset{S \text{ closed}}{=} S \setminus S^\circ = \partial S.$$

Since $\operatorname{conv}\{x,y\} \setminus \{x,y\} \subseteq B^\circ$, we have then that $\operatorname{conv}\{x,y\} \setminus \{x,y\} \subseteq Z(g_1 \cdots g_m)$. Assume now for a contradiction that $x \neq y$. Then this implies that at least one of the g_i vanishes on $\operatorname{aff}\{x,y\}$. Fix a corresponding i. Setting $v := y - x$, we have then $((\nabla g_i)(x))^T v = 0$ and $v^T((\operatorname{Hess} g_i)(x))v = 0$. Since g_i is strictly quasiconcave at x, this implies $v = 0$, i.e., $x = y$ as desired.

(c) By definition of I, the sets S and S' coincide on a neighborhood of u in $\mathbb{R}^n$. Hence u is a *local* minimizer of f on S'. Since S' is convex by (a) and f is linear, u is also a (*global*) minimizer of f on S'. □

Lemma 9.3.17. Suppose B is a closed ball in $\mathbb{R}^n$, $g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]$ are strictly quasiconcave on B and

$$S := \{x \in B \mid g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}$$

has nonempty interior. Then the following hold:

(a) For every real closed extension field R of $\mathbb{R}$ and all linear forms $f \in R[\underline{X}] \setminus \{0\}$, f has a unique minimizer on $\operatorname{Transfer}_{\mathbb{R},R}(S)$.

(b) For every real closed extension field R of $\mathbb{R}$, all linear forms $f \in R[\underline{X}]$ with

$$\|\nabla f\|_2 = 1$$

[→ 6.1.10] (note that $\nabla f \in R^n$ as f is linear) and every $u \in \mathrm{Transfer}_{\mathbb{R},R}(B^\circ)$ which minimizes f on $\mathrm{Transfer}_{\mathbb{R},R}(S)$, there are $\lambda_1, \ldots, \lambda_m \in \mathscr{O}_R \cap R_{\geq 0}$ with

$$\lambda_1 + \ldots + \lambda_m \notin \mathfrak{m}_R$$

such that both $f - f(u) - \sum_{i=1}^m \lambda_i g_i$ and its gradient vanish at u.

Proof. (a) Consider the class of all real closed extension fields R of $\mathbb{R}$ such that all linear forms from $R[\underline{X}] \setminus \{0\}$ have a unique minimizer on $\mathrm{Transfer}_{\mathbb{R},R}(S)$. By real quantifier elimination [→ 1.8.17], this is easily seen to be a 0-ary $\mathbb{R}$-semialgebraic class [→ 1.8.3]. By 1.8.5, this class is either empty or consists of all real closed extensions fields of $\mathbb{R}$. Hence it suffices to prove the statement in the case $R = \mathbb{R}$ [→ 1.8.19]. But then the unicity part follows from Lemma 9.3.16(b) and the existence part from 7.1.19.

(b) By a scaling argument, we can suppose WLOG $g_i < 1$ on B for all $i \in \{1, \ldots, m\}$. Now let R be a real closed field extension of $\mathbb{R}$, $f \in R[\underline{X}]$ a linear form with $\|\nabla f\|_2 = 1$ and u a minimizer of f on $\mathrm{Transfer}_{\mathbb{R},R}(S)$ which lies in $\mathrm{Transfer}_{\mathbb{R},R}(B^\circ)$. Set

$$I := \{i \in \{1, \ldots, m\} \mid g_i(u) = 0\}$$

and define

$$S' := \{x \in B \mid \forall i \in I : g_i(x) \geq 0\} \supseteq S.$$

Using the Tarski principle 1.8.19, one shows easily that u is a minimizer of f on $\mathrm{Transfer}_{\mathbb{R},R}(S')$ by Lemma 9.3.16(c). Note also that of course $u \in \mathscr{O}_R^n$ and $\mathrm{st}(u) \in S$. Using Lemma 9.3.11, choose $k \in \mathbb{N}$ such that

$$h_i := g_i(1 - g_i)^k$$

is strictly concave at $\mathrm{st}(u)$ for $i \in I$. In particular, we note for later use that of course $h_i \neq 0$ for all $i \in I$. Choose an $\varepsilon \in \mathbb{R}_{>0}$ such that h_i is strictly concave and therefore concave on

$$U := \{x \in B \mid \|x - \mathrm{st}(u)\| < \varepsilon\}$$

for all $i \in I$ [→ 9.3.8, 9.3.9]. Now Theorem 7.4.16 implies that $S \cap U$ and hence $S' \cap U \supseteq S \cap U$ has nonempty interior since $\mathrm{st}(u) \in S$, the set S is convex by Lemma 9.3.16(a) and has nonempty interior (so that $S^\circ = \mathrm{relint}\, S$).

Now Lemma 9.3.3 says in particular that for all linear forms $\widetilde{f} \in \mathbb{R}[\underline{X}]$ and minimizers $\widetilde{u}$ of $\widetilde{f}$ on $S' \cap U$ with $\forall i \in I : g_i(\widetilde{u}) = 0$, there is a family $(\lambda_i)_{i \in I}$ in $\mathbb{R}_{\geq 0}$ such that

$$\widetilde{f} - \widetilde{f}(\widetilde{u}) - \sum_{i \in I} \lambda_i h_i \geq 0 \text{ on } U.$$

Using the Tarski priniciple [→ 1.8.19], we see that actually for all real closed extension fields $\widetilde{R}$ of $\mathbb{R}$, all linear forms $\widetilde{f} \in \widetilde{R}[\underline{X}]$ and all minimizers $\widetilde{u}$ of $\widetilde{f}$ on $\mathrm{Transfer}_{\mathbb{R},R}(S' \cap U)$ with $\forall i \in I : g_i(\widetilde{u}) = 0$, there is a family $(\lambda_i)_{i \in I}$ in $R_{\geq 0}$ such that

$$\widetilde{f} - \widetilde{f}(\widetilde{u}) - \sum_{i \in I} \lambda_i h_i \geq 0 \text{ on } \mathrm{Transfer}_{\mathbb{R},\widetilde{R}}(U).$$

We apply this to $\widetilde{R} := R$, $\widetilde{u} := u$, $\widetilde{f} := f$ and thus obtain a family $(\lambda_i)_{i\in I}$ in $R_{\geq 0}$ such that

$$(*) \qquad f - f(u) - \sum_{i\in I} \lambda_i h_i \geq 0 \text{ on } \operatorname{Transfer}_{\mathbb{R},R}(U).$$

Since $h := \prod_{i\in I} h_i \neq 0$ and $S \cap U$ has nonempty interior (as seen above), there is a point

$$x \in S \cap U \subseteq U \subseteq \operatorname{Transfer}_{\mathbb{R},R}(U) \subseteq \mathscr{O}_R^n$$

with $h(x) \neq 0$ [→ 2.2.3] and thus $h_i(x) > 0$ for all $i \in I$. Evaluating $(*)$ in x, we get

$$f(x) - f(u) - \sum_{i\in I} \lambda_i h_i(x) \geq 0.$$

Since $f(x) \in \mathscr{O}_R$ and $f(u) \in \mathscr{O}_R$, this implies

$$\sum_{i\in I} \underbrace{\lambda_i}_{\in R_{\geq 0}} \underbrace{h_i(x)}_{\in \mathbb{R}_{>0}} \in \mathscr{O}_R$$

and thus $\lambda_i \in \mathscr{O}_R \cap R_{\geq 0}$ for all $i \in I$.

It now suffices to show that $\sum_{i\in I} \lambda_i \notin \mathfrak{m}_R$ and that the gradient of the polynomial $f - f(u) - \sum_{i\in I} \lambda_i g_i$ vanishes at u (it is clear that the polynomial itself vanishes at u). From $u \in \operatorname{Transfer}_{\mathbb{R},R}(B^\circ)$ and $u \in U$, it follows that $u \in \operatorname{Transfer}_{\mathbb{R},R}(U^\circ)$. Together with $(*)$ and the Tarski principle, this implies

$$\nabla f = \sum_{i\in I} \lambda_i (\nabla h_i)(u) \overset{9.3.10}{=} \sum_{i\in I} \lambda_i (\nabla g_i)(u)$$

since the gradient of any real polynomial vanishes at each of its local minimizers. In particular, we get

$$1 = \|\nabla f\|_2 \leq \sum_{i\in I} \lambda_i \|(\nabla h_i)(u)\|_2 \leq \left(\sum_{i\in I} \lambda_i\right) \max_{i\in I} \|(\nabla h_i)(u)\|_2$$

(note that $I \neq \emptyset$ by the first inequality) which readily implies $\sum_{i\in I} \lambda_i \notin \mathfrak{m}_R$. □

9.4 Linear polynomials and truncated quadratic modules

Lemma 9.4.1. Let $g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]$ define a compact set

$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}$$

that has nonempty interior near its convex boundary. Suppose that g_i is strictly quasi-concave on $(\text{convbd } S) \cap Z(g_i)$ for each $i \in \{1, \ldots, m\}$. Let R be real closed extension field of $\mathbb{R}$ and $f \in R[\underline{X}]$ be a linear form with $\|\nabla f\|_2 = 1$. Then the following hold:

(a) $F := \{u \in S \mid \forall x \in S : \operatorname{st}(f(u)) \leq \operatorname{st}(f(x))\}$ is a finite subset of convbd S.

(b) $S' := \mathrm{Transfer}_{\mathbb{R},R}(S) \subseteq \mathscr{O}_R^n$ and f has a unique minimizer x_u on

$$\{x \in S' \mid \mathrm{st}(x) = u\}$$

for each $u \in F$.

(c) For every $u \in F$, there are $\lambda_{u1}, \ldots, \lambda_{um} \in \mathscr{O}_R \cap R_{\geq 0}$ with

$$\lambda_{u1} + \ldots + \lambda_{um} \notin \mathfrak{m}_R$$

such that both $f - f(x_u) - \sum_{i=1}^m \lambda_{ui} g_i$ and its gradient vanish at x_u.

Proof. (a) Obviously $\mathrm{st}(f) \neq 0$ and hence

$$F = \{u \in S \mid \forall x \in S : (\mathrm{st}(f))(u) \leq (\mathrm{st}(f))(x)\} \subseteq \mathrm{convbd}\, S$$

by Proposition 9.3.14. We now prove that F is finite. WLOG $S \neq \emptyset$. Set [→ 7.1.19]

$$a := \min\{(\mathrm{st}(f))(x) \mid x \in S\}$$

so that

$$F = \{u \in S \mid (\mathrm{st}(f))(u) = a\}.$$

By compactness of S, it is enough to show that every $x \in S$ possesses a neighborhood U in S such that $U \cap F \subseteq \{x\}$. This is trivial for the points of $S \setminus F$. So consider an arbitrary point $x \in F$. Since $x \in \mathrm{convbd}\, S$, each g_i is positive or strictly quasiconcave at x. According to 9.3.8, we can choose a closed ball B of positive radius around x in $\mathbb{R}^n$ such that each g_i is positive or strictly quasiconcave even on B. By Lemma 9.3.16(b), $\mathrm{st}(f)$ has at most one minimizer on $U := S \cap B$, namely x, i.e., $U \cap F \subseteq \{x\}$.

(b) First observe that $S' := \mathrm{Transfer}_{\mathbb{R},R}(S) \subseteq \mathscr{O}_R^n$ since the transfer from $\mathbb{R}$ to R is an isomorphism of Boolean algebras [→ 1.9.5]: Choosing $N \in \mathbb{N}$ with $S \subseteq [-N, N]_{\mathbb{R}}^n$, we have $S' \subseteq \mathrm{Transfer}_{\mathbb{R},R}([-N, N]_{\mathbb{R}}^n) = [-N, N]_R^n \subseteq \mathscr{O}_R^n$.

Now we fix $u \in F$ and we show that f has a unique minimizer on

$$A := \{x \in S' \mid \mathrm{st}(x) = u\}.$$

Choose $\varepsilon \in \mathbb{R}_{>0}$ such that each g_i is strictly quasiconcave or positive on the ball

$$B := \{v \in \mathbb{R}^n \mid \|v - u\| \leq \varepsilon\}.$$

Since $u \in \mathrm{convbd}\, S \subseteq \overline{S^\circ}$, Lemma 9.3.17(a) says that f has a unique minimizer x on $\mathrm{Transfer}_{\mathbb{R},R}(S \cap B)$. Because of $A \subseteq \mathrm{Transfer}_{\mathbb{R},R}(S \cap B)$, it is thus enough to show $x \in A$. Note that $u \in F \cap B \subseteq S \cap B \subseteq \mathrm{Transfer}_{\mathbb{R},R}(S \cap B)$ and thus $f(x) \leq f(u)$. This implies $\mathrm{st}(f(\mathrm{st}(x))) = \mathrm{st}(f(x)) \leq \mathrm{st}(f(u))$ which yields together with $\mathrm{st}(x) \in S$ that $\mathrm{st}(x) \in F$ (and $\mathrm{st}(f(\mathrm{st}(x))) = \mathrm{st}(f(u))$). Again by Lemma 9.3.17(a), $\mathrm{st}(f)$ has a unique minimizer on $S \cap B$. But u and $\mathrm{st}(x)$ are both a minimizer of $\mathrm{st}(f)$ on $S \cap B$ (note that $\mathrm{st}(x) \in S \cap B$). Hence $u = \mathrm{st}(x)$ and thus $x \in A$ as desired.

(c) Fix $u \in F$. Choose again $\varepsilon \in \mathbb{R}_{>0}$ such that each g_i is strictly quasiconcave or positive on the ball $B := \{v \in \mathbb{R}^n \mid \|v - u\| \leq \varepsilon\}$ and such that $B \cap F = \{u\}$. Since $x_u \in \mathrm{Transfer}_{\mathbb{R},R}(B^\circ)$ obviously minimizes f on $\mathrm{Transfer}_{\mathbb{R},R}(S \cap B)$, we get the necessary Lagrange multipliers by Lemma 9.3.17(b). □

Exercise 9.4.2. For all $k \in \mathbb{N}$ and $x \in [0,1]_{\mathbb{R}}$, we have $x(1-x)^k \leq \frac{1}{k}$.

Theorem 9.4.3. *Let $g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]$ such that $M(g_1, \ldots, g_m)$ is Archimedean and suppose that*

$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}$$

has nonempty interior near its convex boundary. Suppose that g_i is strictly quasiconcave on $(\operatorname{convbd} S) \cap Z(g_i)$ for each $i \in \{1, \ldots, m\}$. Let R be a real closed extension field of $\mathbb{R}$ and $\ell \in \mathscr{O}_R[\underline{X}]_1$ such that $\ell \geq 0$ on $\operatorname{Transfer}_{\mathbb{R},R}(S)$. Then ℓ lies in the quadratic module generated by $g_1, \ldots, g_m$ in $\mathscr{O}_R[\underline{X}]$.

Proof. We will apply Theorem 9.1.14. Since S is compact, we can rescale the g_i and suppose WLOG that

$$g_i \leq 1 \text{ on } S$$

for $i \in \{1, \ldots, m\}$. Let M denote the quadratic module generated by $g_1, \ldots, g_m$ in $\mathscr{O}_R[\underline{X}]$. Since $M(g_1, \ldots, g_m)$ is Archimedean, also M is Archimedean by 8.1.13(b) and 9.1.2(b). Moreover, S could now alternatively be defined from M as in Theorem 9.1.14. Write

$$\ell = f - c$$

with a linear form $f \in \mathscr{O}_R[\underline{X}]$ and $c \in \mathscr{O}_R$. By a rescaling argument, we can suppose that at least one of the coefficients of ℓ lies in $\mathscr{O}_R^\times$ [→ 5.4.7]. If $\operatorname{st}(\ell(x)) > 0$ for all $x \in S$, then Theorem 9.1.14 applied to ℓ with $k = 0$ yields $\ell \in M$ and we are done. Hence we can from now on suppose that there is some $u \in S$ with $\operatorname{st}(\ell(u)) = 0$. For such an u, we have $\operatorname{st}(c) = \operatorname{st}(f(u))$ so that at least one coefficient of f must lie $\mathscr{O}_R^\times$. By another rescaling, we now can suppose WLOG that $\|\nabla f\|_2 = 1$. Now we are in the situation of Lemma 9.4.1 and we define

$$F, \quad (x_u)_{u \in F} \quad \text{and} \quad (\lambda_{ui})_{(u,i) \in F \times \{1,\ldots,m\}}$$

accordingly. Note that

$$F = \{u \in S \mid \operatorname{st}(\ell(u)) = 0\} \neq \emptyset$$

since $\operatorname{st}(\ell(x)) \geq 0$ for all $x \in S$. We have $f(x_u) - c = \ell(x_u) \geq 0$ and

$$\operatorname{st}(f(x_u) - c) = \operatorname{st}(\ell(u)) = 0$$

for all $u \in F$. Hence $f(x_u) - c \in \mathfrak{m}_R \cap R_{\geq 0}$ for all $u \in F$. We thus have

$$\ell - \underbrace{(f(x_u) - c)}_{=:\lambda_{u0} \in \mathfrak{m}_R \cap R_{\geq 0}} - \sum_{i=1}^{m} \underbrace{\lambda_{ui}}_{\in \mathscr{O}_R \cap R_{\geq 0}} g_i \in I^2_{x_u}$$

for all $u \in F$ by 9.4.1(c) and 9.1.13. Evaluating this in x_u (and using $g_i(x_u) \geq 0$) yields

$$(*) \qquad g_i(x_u) \neq 0 \implies \lambda_{ui} = 0 \qquad \text{and thus}$$

$$(**) \qquad \lambda_{ui} g_i \equiv_{I^2_{x_u}} \lambda_{ui} g_i (1 - g_i)^k$$

for all $u \in F$, $i \in \{1,\dots,m\}$ and $k \in \mathbb{N}$. By the Chinese remainder theorem, we find polynomials $s_0,\dots,s_m \in \mathscr{O}_R[\underline{X}]$ such that $s_i \equiv_{I^3_{x_u}} \sqrt{\lambda_{ui}} \in \mathscr{O}_R$ for all $u \in F$ and $i \in \{0,\dots,m\}$ because the ideals $I^3_{x_u}$ ($u \in F$) are pairwise coprime [$\to$ 9.1.7] (use that $\mathrm{st}(x_u) = u \neq v = \mathrm{st}(x_v)$ for all $u,v \in F$ with $u \neq v$). By an easy scaling argument, we can even guarantee that the coefficients of s_0 lie in $\mathfrak{m}_R$ since $\sqrt{\lambda_{u0}} \in \mathfrak{m}_R$. Then we have

$$(***) \qquad s_i^2 \equiv_{I^3_{x_u}} \lambda_{ui}$$

which means in other words

$$s_i^2(x_u) = \lambda_{ui}, \qquad (\nabla(s_i^2))(x_u) = 0 \qquad \text{and} \qquad (\mathrm{Hess}(s_i^2))(x_u) = 0$$

for all $i \in \{0,\dots,m\}$ and $k \in \mathbb{N}$. It suffices to show that there is $k \in \mathbb{N}$ such that the polynomial

$$\ell - s_0^2 - \sum_{i=1}^m s_i^2(1-g_i)^{2k}g_i \overset{\substack{(***)\\(**)}}{\in} \bigcap_{u\in F} I^2_{x_u}$$

lies in M since this implies immediately $\ell \in M$. By Theorem 9.1.14, this task reduces to find $k \in \mathbb{N}$ such that $f_k > 0$ on $S \setminus F$ and $(\mathrm{Hess}(f_k))(u) \succ 0$ for all $u \in F$ where

$$f_k := \mathrm{st}(\ell) - \sum_{i=1}^m \mathrm{st}(s_i^2)(1-g_i)^{2k}g_i \in \mathbb{R}[\underline{X}]$$

is the standard part of this polynomial. Note for later use that f_k and ∇f_k vanish on F for all $k \in \mathbb{N}$. In order to find such a k, we calculate

$$\begin{aligned}(\mathrm{Hess}\, f_k)(u) &\overset{(***)}{=} -\sum_{i=1}^m \mathrm{st}(\lambda_{ui})\,\mathrm{Hess}((1-g_i)^{2k}g_i)(u)\\ &\overset{\substack{9.3.10\\(*)}}{=} \sum_{i=1}^m \mathrm{st}(\lambda_{ui})(4k(\nabla g_i)(\nabla g_i)^T - \mathrm{Hess}\, g_i)(u)\end{aligned}$$

for $u \in F$ and $k \in \mathbb{N}$. By Lemma 9.3.11 we can choose $k \in \mathbb{N}$ such that $g_i(1-g_i)^{2k}$ is strictly concave on $\{x \in F \mid g_i(x) = 0\}$ for $i \in \{1,\dots,m\}$. Since $\mathrm{st}(\lambda_1) + \ldots + \mathrm{st}(\lambda_m) > 0$ [$\to$ 9.4.1(c)], we get together with $(*)$ and 9.3.10 that for all sufficiently large k, we have $(\mathrm{Hess}\, f_k)(u) \succ 0$ for all $u \in F$. In particular, we can choose $k_0 \in \mathbb{N}$ such that $\mathrm{Hess}(f_{k_0})(u) \succ 0$ for all $u \in F$. Since f_{k_0} and ∇f_{k_0} vanish on F, we have by elementary analysis that there is an open subset U of $\mathbb{R}^n$ containing F such that $f_{k_0} \geq 0$ on U. Then $S \setminus U$ is compact so that we can choose $N \in \mathbb{N}$ with $\mathrm{st}(\ell) \geq \frac{1}{N}$ and $\mathrm{st}(s_i^2) \leq N$ on $S \setminus U$. Then $f_k \geq \frac{1}{N} - m\frac{N}{2k}$ on $S \setminus U$ by Exercise 9.4.2 since $0 \leq g_i \leq 1$ on S for all $i \in \{1,\dots,m\}$. For all sufficiently large $k \in \mathbb{N}$ with $k \geq k_0$, we now have $f_k > 0$ on $S \setminus U$ and because of $f_k \geq f_{k_0} > 0$ on $S \cap U$ (use again that $0 \leq g_i \leq 1$ on S) even $f_k > 0$ on S. □

Corollary 9.4.4. *Let $g_1,\dots,g_m \in \mathbb{R}[\underline{X}]$ such that $M(g_1,\dots,g_m)$ is Archimedean and suppose that*

$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0,\dots,g_m(x) \geq 0\}$$

has nonempty interior near its convex boundary. Suppose that g_i is strictly quasiconcave on $(\operatorname{convbd} S) \cap Z(g_i)$ *for each* $i \in \{1, \ldots, m\}$. *Let R be a real closed extension field of $\mathbb{R}$ and* $\ell \in R[\underline{X}]_1$ *such that* $\ell \geq 0$ *on* $\operatorname{Transfer}_{\mathbb{R},R}(S)$. *Then ℓ lies in the quadratic module generated by* $g_1, \ldots, g_m$ *in* $R[\underline{X}]$.

Corollary 9.4.5. *Let* $g_1, \ldots, g_m \in \mathbb{R}[\underline{X}]$ *such that* $M(g_1, \ldots, g_m)$ *is Archimedean and suppose that*

$$S := \{x \in \mathbb{R}^n \mid g_1(x) \geq 0, \ldots, g_m(x) \geq 0\}$$

has nonempty interior near its convex boundary. Suppose that g_i is strictly quasiconcave on $(\operatorname{convbd} S) \cap Z(g_i)$ *for each* $i \in \{1, \ldots, m\}$. *Then there exists*

$$d \in \mathbb{N}$$

such that for all $\ell \in \mathbb{R}[\underline{X}]_1$ *with* $\ell \geq 0$ *on S, we have*

$$\ell \in M_d(g_1, \ldots, g_m).$$

Proof. (cf. the proofs of Theorems 5.4.5 and 9.2.3) For each $d \in \mathbb{N}$, consider the class S_d of all pairs $(R, a_0, a_1, \ldots, a_n)$ where R is a real closed extension field of $\mathbb{R}$ and $a_0, a_1, \ldots, a_n \in R$ such that whenever

$$\forall x \in \operatorname{Transfer}_{\mathbb{R},R}(S) : a_1 x_1 + \ldots + a_n x_n + a_0 \geq 0$$

holds, the polynomial $a_1 X_1 + \ldots + a_n X_n + a_0$ is a sum of d elements from $R[\underline{X}]$ where each term in the sum is of degree at most d and is of the form $p^2 g_i$ with $p \in R[\underline{X}]$ and $i \in \{0, \ldots, m\}$ where $g_0 := 1 \in R[\underline{X}]$ [$\rightarrow$ 9.2.8(a)]. By real quantifier elimination 1.8.17, it is easy to see that this is an $(n+1)$-ary $\mathbb{R}$-semialgebraic class. Set $\mathscr{E} := \{S_d \mid d \in \mathbb{N}\}$ and observe that $\forall d_1, d_2 \in \mathbb{N} : \exists d_3 \in \mathbb{N} : S_{d_1} \cup S_{d_2} \subseteq S_{d_3}$ (take $d_3 := \max\{d_1, d_2\}$). By 9.4.4, we have $\bigcup \mathscr{E} = \mathscr{R}_{n+1}$. Now 5.4.2 yields $\operatorname{Set}_{\mathbb{R}}(S_d) = \mathbb{R}^{n+1}$ for some $d \in \mathbb{N}$. □

Bibliography

[ABR] C. Andradas, L. Bröcker, J. M. Ruiz: Constructible sets in real geometry, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) 33, Springer-Verlag, Berlin, 1996 ii

[BCR] J. Bochnak, M. Coste, M.-F. Roy: Real algebraic geometry, Translated from the 1987 French original, Revised by the authors, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), 36, Springer-Verlag, Berlin, 1998 ii

[BPR] S. Basu, R. Pollack, M.-F. Roy: Algorithms in real algebraic geometry, Algorithms and Computation in Mathematics 10, Springer-Verlag, Berlin, 2003 ii

[Brö] L. Bröcker: Real spectra and distributions of signatures, Real algebraic geometry and quadratic forms (Rennes, 1981), pp. 249–272, Lecture Notes in Math., 959, Springer, Berlin-New York, 1982 89

[BS] E. Becker, N. Schwartz: Zum Darstellungssatz von Kadison-Dubois, Arch. Math. (Basel) 40 (1983), no. 5, 421–428 147

[BSS] S. Burgdorf, C. Scheiderer, M. Schweighofer: Pure states, nonnegative polynomials and sums of squares, Comment. Math. Helv. 87 (2012), no. 1, 113–140 146, 152

[BW] R. Berr, T. Wörmann: Positive polynomials on compact sets, Manuscripta Math. 104 (2001), no. 2, 135–143 78

[Du1] D.W. Dubois: A note on David Harrison's theory of preprimes, Pacific J. Math. 21, 1967, 15–19 75

[Du2] D.W. Dubois: A nullstellensatz for ordered fields, Ark. Mat. 8 (1969), 111–114 66, 72

[Efr] G. Efroymson: Local reality on algebraic varieties, J. Algebra 29 (1974), 133–142 66, 72

[Hil] D. Hilbert: Über die Darstellung definiter Formen als Summe von Formenquadraten, Math. Ann. 32 (1888), no. 3, 342–350 140

[HNS] D. Henrion, S. Naldi, M. Safey El Din: Exact algorithms for linear matrix inequalities, SIAM J. Optim. 26 (2016), no. 4, 2512–2539 140

[Jac] T. Jacobi: A representation theorem for certain partially ordered commutative rings, Math. Z. 237 (2001), no. 2, 259–273 147

[Kad] R.V. Kadison: A representation theory for commutative topological algebra, Mem. Amer. Math. Soc., No. 7 (1951), 39 pp. 75

[KS] M. Knebusch, C. Scheiderer: Einführung in die reelle Algebra, Vieweg Studium: Aufbaukurs Mathematik 63, Friedr. Vieweg & Sohn, Braunschweig, 1989 ii

[Kri] J.-L. Krivine: Anneaux préordonnés, J. Analyse Math. 12, 1964, 307–326 64, 66, 71, 72, 75

[LPR] H. Lombardi, D. Perrucci, M.-F. Roy: An elementary recursive bound for effective Positivstellensatz and Hilbert 17-th problem, preprint [https://arxiv.org/abs/1404.2338] 94

[Mar] M. Marshall: Positive polynomials and sums of squares, Mathematical Surveys and Monographs, 146. American Mathematical Society, Providence, RI, 2008 ii

[NS] J. Nie, M. Schweighofer: On the complexity of Putinar's Positivstellensatz, J. Complexity 23 (2007), no. 1, 135–150 163

[PD] A. Prestel, C. Delzell: Positive polynomials, From Hilbert's 17th problem to real algebra, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2001 ii

[Pre] A. Prestel: Lectures on formally real fields, Monografias de Matemática, Instituto de Matemática Pura e Aplicada, Rio de Janeiro, vol. 22 (1975); reprinted as: Lecture Notes in Mathematics, 1093, Springer-Verlag, Berlin (1984) 64, 71

[Pól] G. Pólya: Über positive Darstellung von Polynomen, Vierteljahresschrift der Naturforschenden Gesellschaft in Zürich 73 (1928), 141–145 [http://www.ngzh.ch/archiv/1928_73/73_1-2/73_4.pdf] 148

[PS] A. Pfister, C. Scheiderer: An elementary proof of Hilbert's theorem on ternary quartics, J. Algebra 371 (2012), 1–25 140

[Put] M. Putinar: Positive polynomials on compact semi-algebraic sets, Indiana Univ. Math. J. 42 (1993), no. 3, 969–984

[Ris] J.-J. Risler: Une caractérisation des idéaux des variétés algébriques réelles, C. R. Acad. Sci. Paris Sér. A-B 271,1970, A1171–A1173 66, 72

[S1] C. Scheiderer: Distinguished representations of non-negative polynomials, J. Algebra 289 (2005), no. 2, 558–573 164

[S2] C. Scheiderer: Sums of squares of polynomials with rational coefficients, J. Eur. Math. Soc. (JEMS) 18 (2016), no. 7, 1495–1513 140

[Sch] K. Schmüdgen: The K-moment problem for compact semi-algebraic sets, Math. Ann. 289 (1991), no. 2, 203–206 78

[Ste] G. Stengle: A nullstellensatz and a positivstellensatz in semialgebraic geometry, Math. Ann. 207 (1974), 87–97 64, 71

[Sto] M.H. Stone: A general theory of spectra I, Proc. Nat. Acad. Sci. U. S. A. 26, (1940). 280–283 75