Hilbert's Seventeenth Problem: sums of squares

Is a rational function with real coefficients that only takes non-negative values a sum of squares of rational functions with real coefficients?

1 Introduction

We begin with an example. Let $f(x)$ is the polynomial in one variable $f(x) = x^2 + bx + c$, with $b, c \in \mathbb{R}$ and suppose that we want to know if, for every $\alpha \in \mathbb{R}$, the evaluation $f(\alpha)$ is non-negative. Completing the square, we find

$$f(x) = \left(x + \frac{b}{2}\right)^2 + \frac{4c - b^2}{4}.$$

Since squares of real numbers are always non-negative, we deduce that $f(x)$ only has non-negative evaluations if and only if $4c - b^2$ is non-negative. Moreover, if the inequality $4c - b^2 \geq 0$ holds, then the identity

$$f(x) = \left(x + \frac{b}{2}\right)^2 + \left(\frac{\sqrt{4c - b^2}}{2}\right)^2$$

shows that f is a sum of squares of polynomials (in this case, one linear and one constant).

We now generalize the previous example. Let n be a non-negative integer and let $f(x_1, \ldots, x_n) \in \mathbb{R}[x_1, \ldots, x_n]$ be a polynomial with real coefficients in n variables. Suppose that we would like to know if for every choice $\alpha_1, \ldots, \alpha_n \in \mathbb{R}^n$ the evaluation of f at the n-tuple $(\alpha_1, \ldots, \alpha_n)$ is non-negative:

$$\text{for all } \alpha_1, \ldots, \alpha_n \in \mathbb{R}^n \qquad \Longrightarrow \qquad f(\alpha_1, \ldots, \alpha_n) \geq 0.$$

Of course, if the polynomial f is a sum of squares of polynomials, then all the evaluations of f are non-negative. Yet, not every polynomial that only assumes non-negative values is a sum of squares of polynomials. Polynomials with this property were implicitly known to exist for a long time, probably also with explicit examples. Nevertheless, the first published non-negative polynomial that is not a sum of squares is due to Motzkin [Mot67] in 1967.

Example 1.1 (Motzkin)**.** Let $m(x, y)$ be the polynomial

$$m(x, y) = x^4y^2 + x^2y^4 + 1 - 3x^2y^2.$$

The polynomial m satisfies the identities

$$\begin{aligned} m(x,y) &= \frac{x^2y^2(x^2+y^2-2)^2(x^2+y^2+1)+(x^2-y^2)^2}{(x^2+y^2)^2} \\ &= \frac{x^4y^2(x^2+y^2-2)^2}{(x^2+y^2)^2}+\frac{x^2y^4(x^2+y^2-2)^2}{(x^2+y^2)^2}+ \\ &\qquad +\frac{(x^2+y^2-2)^2}{(x^2+y^2)^2}+\frac{(x^2-y^2)^2}{(x^2+y^2)^2} \\ &= \left(\frac{x^2y(x^2+y^2-2)}{(x^2+y^2)}\right)^2+\left(\frac{xy^2(x^2+y^2-2)}{(x^2+y^2)}\right)^2+ \\ &\qquad +\left(\frac{(x^2+y^2-2)}{(x^2+y^2)}\right)^2+\left(\frac{(x^2-y^2)}{(x^2+y^2)}\right)^2, \end{aligned}$$

showing that m is a sum of squares of rational functions with real coefficients. We deduce that m only takes non-negative values: if $(\alpha,\beta)\neq(0,0)$, then the evaluation $m(\alpha,\beta)$ is a sum of squares of real numbers; if $(\alpha,\beta)=(0,0)$, then $m(0,0)=1>0$. We shall see that m is not a sum of squares of *polynomials*.

Thus, Hilbert's Seventeenth Problem asks whether for every polynomial p with real coefficients and non-negative evaluations, there are rational functions $f_1,\ldots,f_r$ with real coefficients satisfying the identity $p=\sum f_i^2$.

E. Artin gave a positive answer to this question: a polynomial in n variables and real coefficients all of whose evaluations are non-negative is a sum of squares of rational functions in n variables with real coefficients.

We follow closely the treatment of Artin's Theorem appearing in [Pfi95, Chapter 6]. First, we concentrate on sums of squares in general fields k. For our applications, the most important fields will be the field $\mathbb{R}$ of real numbers, the field $\mathbb{R}(x_1,\ldots,x_n)$ of rational functions in n variables over $\mathbb{R}$, and the field $\mathbb{Q}$ of rational numbers. We then establish strict relations between sums of squares in a field and orderings of the same field. Next, we find ways to enlarge an ordered field by adding elements to the field and extending the order to the new elements. Finally, we show how Hilbert's Seventeenth Problem follows from the Artin-Lang Homomorphism Theorem.

2 Formally real fields and orders

We begin here our study of orders on fields and sums of squares. The main motivation to keep in mind is that we would like to have a notion of *positive* elements of a field and that we also want (sums of) non-zero squares to be positive.

Definition 2.1. A field k is *formally real* if $-1\in k$ is not a sum of squares of elements of k.

It is easy to check that the following are examples/non-examples:

(1) the field $\mathbb{R}$ of real numbers is formally real;

(2) the field $\mathbb{Q}$ of rational numbers is formally real;

(3) the field $\mathbb{C}$ of complex numbers is *not* formally real;

(4) the finite field $\mathbb{F}_2$ with two elements is *not* formally real;

(5) more generally, no finite is formally real;

(6) more generally still, no field of positive characteristic is formally real.

A slightly more elaborate example involves rational functions. Recall that a rational function with coefficients in a field k is the ratio of two polynomials with coefficients in k, whose denominator is not the zero polynomial. Note that we only require that the polynomial in the denominator be non-zero: it may evaluate to zero at some points. For instance, $\frac{3x^2-2}{x(x^2+1)(x^2-7)}$ is a rational function over the real numbers and the denominator vanishes for $x \in \{0, \pm\sqrt{7}\}$. The set of rational functions over a field k in one variable x forms a field that we denote by $k(x)$.

Example 2.2. The field $\mathbb{R}(x)$ of rational functions in one variable over the real numbers is formally real. Indeed, suppose that $f_1(x), \ldots, f_r(x) \in \mathbb{R}(x)$ are rational functions satisfying the identity

$$f_1(x)^2 + \cdots + f_r(x)^2 = -1. \tag{1}$$

Let $d(x)$ be the product of the denominators of $f_1, \ldots, f_r$. Thus, $d(x)$ is a non-zero polynomial and hence there is a real number α such that $d(\alpha) \neq 0$. It follows that we can evaluate both sides of Equation (1) at α and obtain

$$f_1(\alpha)^2 + \cdots + f_r(\alpha)^2 = -1.$$

This equation is impossible, since the left-hand side is a sum of squares of real numbers and therefore it is non-negative.

Notation 2.3. Let k be a field. We set

$$\begin{aligned} \textstyle\sum k &= \{a \in k \quad : \quad a \text{ is a sum of squares in } k\}, \\ \textstyle\sum k^\times &= \textstyle\sum k \smallsetminus \{0\}. \end{aligned}$$

For any subset $A \subset k$ and any $a \in A$, we set

$$\begin{aligned} A + A &= \{a_1 + a_2 \quad : \quad a_1, a_2 \in A\}, \\ A \cdot A &= \{a_1 a_2 \quad : \quad a_1, a_2 \in A\}, \\ aA &= \{aa' \quad : \quad a' \in A\}. \end{aligned}$$

We leave the proof of the following lemma as an exercise.

Lemma 2.4. *Let k be a field.*

(1) The set $\sum k$ is closed under addition and multiplication.

(2) The set $\sum k^\times$ is a multiplicative group.

(3) The field k is formally real if and only if -1 is not in $\sum k$.

We now come to the definition of a (pre)order. We want to construct orders, and we build them by extending preorders. We use the notation P for preorders, since they share many properties of positive real numbers.

Definition 2.5. Let k be a field.

(1) A *preorder* of k is a subset $P \subset k$ satisfying

$$P + P \subset P, \qquad P \cdot P \subset P, \qquad \sum k \subset P, \qquad -1 \notin P.$$

(2) An *order* of k is a preorder P satisfying

$$P \cup -P = k, \qquad P \cap -P = \{0\}.$$

It follows from this definition that if -1 is a sum of squares in k, then k admits no preorder. If, instead, -1 is not a sum of squares, then $\sum k$ is a preorder of k and any preorder of k must contain $\sum k$.

Lemma 2.6. *Let P be a preorder of k and let $a, b \in k$. If $ab \in P$, then $P + aP$ or $P - bP$ is a preorder of k.*

Proof. Let $p_1, \ldots, p_4$ be elements of P. The identities

$$\begin{aligned} (p_1 + ap_2) + (p_3 + ap_4) &= (p_1 + p_3) + a(p_2 + p_4) \\ (p_1 + ap_2)(p_3 + ap_4) &= (p_1p_3 + a^2p_2p_4) + a(p_1p_4 + p_2p_3) \end{aligned}$$

show that $P + aP$ is closed under addition and multiplication, since P is closed under addition and multiplication, and contains all (sums of) squares. Moreover, $P + aP$ contains P and hence contains $\sum k$. Similarly, $P - bP$ is closed under addition and multiplication, and contains $\sum k$. Thus, to show that $P + aP$ or $P - bP$ is a preorder, it remains to show that -1 cannot belong to both $P + aP$ and $P - bP$.

Proceed by contradiction and suppose that -1 belongs to both $P + aP$ and $P - bP$. Thus, there are $p_1, \ldots, p_4 \in P$ such that

$$-1 = p_1 + ap_2 = p_3 - bp_4.$$

We compute

$$(ap_2)(-bp_4) = (1 + p_1)(1 + p_3) = 1 + p_1 + p_3 + p_1p_3,$$

and we obtain

$$-1 = p_1 + p_3 + p_1p_3 + abp_2p_4. \tag{2}$$

Since the product ab is in P by assumption, the right-hand side of Equation (2) is in P, contradicting the assumption that $-1 \notin P$, as P is a preorder. □

Lemma 2.7. *A preorder of k that is maximal with respect to inclusion is an order.*

Proof. Let P be a maximal preorder of k. We show that the two equalities $P \cup -P = k$ and $P \cap -P = \{0\}$ hold.

First, let $a \in k$ be any element. We apply Lemma 2.6 with $a = b$, noting that $a^2 \in P$, since P contains all squares. We deduce that $P + aP$ or $P - aP$ is a preorder. By maximality of P, at least one of the inclusions $P + aP \subset P$ or $P - aP \subset P$ holds. Therefore, either a or $-a$ belongs to P. Since a is arbitrary, we obtain that $k = P \cup -P$.

To show that the intersection $P \cap -P$ consists only of 0, suppose, by contradiction, that $a \in k$ belongs to $(P \cap -P) \smallsetminus \{0\}$. Thus, $a \in P$ is non-zero and there is an $a' \in P$ satisfying $a = -a' \in P$. Using the chain of identities $-1 = \frac{a'}{a} = \frac{aa'}{a^2}$, we deduce that -1 belongs to P, since P, being a preorder, contains all squares and is closed under multiplication. Yet, preorders do not contain -1 and we reach a contradiction, as required. □

Let k be a field such that $-1 \notin \sum k$, so that there is at least the preorder $P_0 = \sum k$ of k. As an application of Zorn's Lemma (an equivalent of the Axiom of Choice), we deduce that there is a maximal preorder containing P_0: the field k admits an order.

Corollary 2.8 (Artin-Schreier)**.** *A field k is formally real if and only if k has an order.*

The following proposition implies a strong relation between orders and preorders.

Proposition 2.9. *Let T be a preorder of k. Then $T = \cap P$ is the intersection of all the orders P of k containing T.*

Proof. The inclusion $T \subset \cap P$ is clear. Suppose that $a \in k \smallsetminus T$ is an element not in T.

We check that -1 does not belong to $T - aT$. Otherwise, there exist $t_1, t_2 \in T$ satisfying the identity $-1 = t_1 + at_2$. The fact that T is a preorder implies that t_2 cannot be 0. Thus, we obtain the identity $a = \frac{1+t_1}{t_2} \in T$, since sums and ratios of elements of T are in T, contradicting our hypothesis.

Thus, $T - aT$ is a preorder of k. By Lemma 2.7, a maximal preorder P containing $T - aT$ is an order of k containing $-a$. We deduce that P cannot contain a, since a is non-zero and $P \cap -P = \{0\}$. It follows that $a \notin \cap P$ and, since a is an arbitrary element not in T, we conclude that the equality $T = \cap P$ holds. □

Corollary 2.10. *Let k be a formally real field. The intersection of all preorders on k is the set $\sum k$.*

Proof. As k is formally real, the set $\sum k$ is a preorder. Since any preorder contains the preorder $\sum k$, we conclude applying Proposition 2.9. □

Let P be an order of k. Recall that our guiding principle is that P is the set of positive elements of k. Thus, we define an order relation $\leq_P$ on k using P. For $a, b \in k$, we say that a is P-less than or equal to b, and denote it by $a \leq_P b$, whenever $b - a$ is in P: in formulas,

$$a \leq_P b \qquad \Longleftrightarrow \qquad b - a \in P.$$

Usually, when the order P is clear from the context, we remove the subscript P from $\leq_P$ and simply write $\leq$ and say that a is less than or equal to b. For all $a, b, c \in P$, the relation $\leq$ satisfies the following properties:

(1)	reflexivity	$a \leq a$;		
(2)	anti-symmetry	$a \leq b$ and $b \leq a$	$\Longrightarrow$	$a = b$;
(3)	transitivity	$a \leq b$ and $b \leq c$	$\Longrightarrow$	$a \leq c$;
(4)	total order	$a \leq b$ or $b \leq a$;		
(5)	compatibility with addition	$a \leq b$	$\Longrightarrow$	$a + c \leq b + c$;
(6)	compatibility with multiplication by non-negative numbers	$a \leq b$ and $0 \leq c$	$\Longrightarrow$	$ac \leq bc$.

A binary relation satisfying properties (1), (2), (3) is called an *order relation*. An order relation satisfying property (4) is called at *total ordering* or a *linear ordering*: it means that any two elements of k are comparable. Together, the six properties (1-6) are the axioms for an order relation on a field k.

We usually write

- $a \geq b$ for $b \leq a$, and
- $a < b$ or $b > a$ for $a \leq b$ and $a \neq b$.

Conversely, an order relation $\leq$ on a field k satisfying properties (1-6) determines an order of k by setting $P = \{a \in k \mid a \geq 0\}$. We call the elements of P *positive*.

Thus, we call an *ordered field* either a pair (K, P) or a pair $(K, \leq)$ and interchange the two as we need it.

Definition 2.11. An element a of k is *totally positive* if a is a sum of squares.

Corollary 2.10 shows that the totally positive elements of k are precisely the elements of k that are positive with respect to every order relation on k.

Example 2.12. The fields $\mathbb{R}$, of real numbers, and $\mathbb{Q}$, of rational numbers, admit a unique order. It follows that, for these two fields, the totally positive elements are precisely the sums of squares.

We now give an extended example: quadratic fields.

2.1 Quadratic fields

Let $d \in \mathbb{Q}$ be a rational number. Denote by $\mathbb{Q}\left(\sqrt{d}\right)$ denote the subfield of $\mathbb{C}$ generated by $\sqrt{d}$. It is an easy exercise that the set of all complex numbers of the form $a + b\sqrt{d}$, for $a, b \in \mathbb{Q}$ is the field $\mathbb{Q}\left(\sqrt{d}\right)$. If d is the square of a rational number, then $\mathbb{Q}\left(\sqrt{d}\right)$ coincides with $\mathbb{Q}$. We exclude this case and assume that d is not a square.

Case 1: $d < 0$. If d is negative, then it follows from one of the exercises that $-d$ is a sum of squares in $\mathbb{Q}$. Combining an expression

$$-d = \sum_{i=1}^{r} a_i^2, \qquad\qquad a_1, \ldots, a_r \in \mathbb{Q},$$

of $-d > 0$ as a sum of squares and the equality $\left(\sqrt{d}\right)^2 = d$, we obtain the identities

$$-1 = \frac{\sum_i a_i^2}{d} = \sum_i \left(\frac{a_i}{\sqrt{d}}\right)^2.$$

Hence, the field $\mathbb{Q}\left(\sqrt{d}\right)$ is not formally real, if $d < 0$.

Case 2: d is not a square and $d > 0$. As we saw, the equality

$$\mathbb{Q}\left(\sqrt{d}\right) = \left\{a + b\sqrt{d} \quad \middle| \quad a, b \in \mathbb{Q}\right\}$$

holds. We now define two distinct orders P_+ and P_- on $\mathbb{Q}\left(\sqrt{d}\right)$.

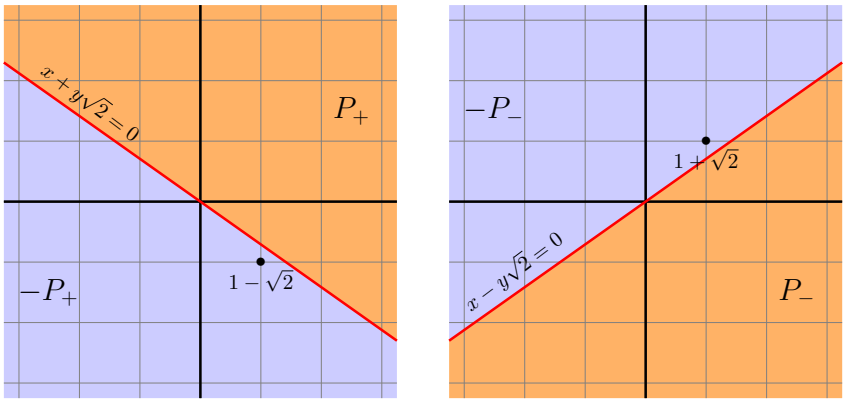
The order P_+ is obtained by viewing $\mathbb{Q}\left(\sqrt{d}\right)$ as a subfield of $\mathbb{R}$, identifying, as usual, $\sqrt{d}$ with the *positive* square root of d. Thus, in this case, the subset P_+ of $\mathbb{Q}\left(\sqrt{d}\right)$ of positive elements is

$$P_+ = \left\{a + b\sqrt{d} \quad \middle| \quad a + b\sqrt{d} \geq 0\right\}.$$

The order P_- is again obtained by viewing $\mathbb{Q}\left(\sqrt{d}\right)$ as a subfield of $\mathbb{R}$, but in a different way: this time, we identify $\sqrt{d}$ with the *negative* square root of d. Thus, in this case, the subset P_- of $\mathbb{Q}\left(\sqrt{d}\right)$ of positive elements is

$$P_- = \left\{a + b\sqrt{d} \quad \middle| \quad a - b\sqrt{d} \geq 0\right\}.$$

Example 2.13. In the case $d = 2$, we give a drawing of the two orders P_+ and P_-.



The order P_+ The order P_-

The two possibilities for the positive elements of $\mathbb{Q}(\sqrt{2})$, in orange

- The number $1 + \sqrt{2}$ is contained in P_+ (as $1 + \sqrt{2} \simeq 2.4142\ldots > 0$) and it is therefore positive with respect to the order determined by P_+. However, $1 + \sqrt{2}$ is not contained in P_- (as $1 - \sqrt{2} \simeq -0.4142\ldots < 0$) and it is therefore negative with respect to the order determined by P_-. We deduce that $1 + \sqrt{2}$ is not a sum of squares in $\mathbb{Q}(\sqrt{2})$.
- The number $2 + \sqrt{2}$ lies in the intersection $P_+ \cap P_-$ and indeed it is a sum of squares:

$$2 + \sqrt{2} = \left(1 + \frac{1}{\sqrt{2}}\right)^2 + \frac{1}{2} = \left(1 + \frac{1}{\sqrt{2}}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2.$$

Going back to the general case, we leave the proof of the following proposition as an exercise.

Proposition 2.14. *The only orders of* $\mathbb{Q}\left(\sqrt{d}\right)$ *are* P_+ *and* P_-.

Proof. Exercise. □

We deduce that the sums of squares in $\mathbb{Q}\left(\sqrt{d}\right)$ are the numbers in the intersection $P_+ \cap P_-$, that is the numbers $a + b\sqrt{d}$, with $a, b \in \mathbb{Q}$, satisfying the inequalities

$$a + b\sqrt{d} \geq 0 \qquad \text{and} \qquad a - b\sqrt{d} \geq 0,$$

or, equivalently, $a \geq |b|\sqrt{d}$.

2.2 Adding square roots of positive elements

In our argument towards establishing Hilbert's Seventeenth Problem, it is useful to be able to extend a field k by adding a square root of an element. We give this construction here. In the case of a formally real field, we prove that adding a square root of a positive element yields a formally real field.

Begin by recalling the familiar extension going from $\mathbb{R}$ to $\mathbb{C}$: in this case, we are adding the square root of -1 to $\mathbb{R}$. We may introduce the complex numbers as formal, real linear combinations of 1 and a symbol i. Thus, every element of $\mathbb{C}$ is of the form $a+bi$, with $a, b \in \mathbb{R}$. We define addition componentwise and multiplication using distributivity and the identity $i^2 = -1$: for $a, b, a', b' \in \mathbb{R}$, we set

$$\begin{aligned}(a+bi)+(a'+b'i) &= (a+a')+(b+b')i \\ (a+bi)(a'+b'i) &= (aa'-bb')+(ab'+a'b)i.\end{aligned}$$

The missing ingredient is the computation of multiplicative inverses. Their existence is a consequence of the identities

$$\frac{1}{a+bi} = \frac{a-bi}{(a+bi)(a-bi)} = \frac{a-bi}{(a^2+b^2)} = \frac{a}{(a^2+b^2)} + \frac{-b}{(a^2+b^2)}i,$$

allowing us to write the inverse of $a+bi$, with $(a,b) \neq (0,0)$, in the required form as a formal, real linear combination of $1, i$. It is a routine check to verify that these definitions really produce a field.

In the general case, we proceed analogously. Let k be a field and let $d \in k$ be any element. We construct a field $k\left(\sqrt{d}\right)$, containing k and also an element $\sqrt{d}$ satisfying the identity $\left(\sqrt{d}\right)^2 = d$. If d is a square in k, then we set $k\left(\sqrt{d}\right) = k$ and we denote by $\sqrt{d}$ any fixed square root of d. Suppose, therefore, that k is not a square in k. Set

$$k\left(\sqrt{d}\right) = \left\{a + b\sqrt{d} \quad \middle| \quad a, b \in k\right\}.$$

For all $a, b, a', b' \in k$, we define

$$\begin{aligned}\left(a+b\sqrt{d}\right)+\left(a'+b'\sqrt{d}\right) &= (a+a')+(b+b')\sqrt{d} \\ \left(a+b\sqrt{d}\right)\left(a'+b'\sqrt{d}\right) &= (aa'+bb'd)+(ab'+a'b)\sqrt{d},\end{aligned}$$

and, if $(a,b) \neq (0,0)$, we also define

$$\frac{1}{a+b\sqrt{d}} = \frac{a}{(a^2-b^2d)} + \frac{-b}{(a^2-b^2d)}\sqrt{d}.$$

We leave it as an exercise to check that $k\left(\sqrt{d}\right)$ is a field containing k and that the element $\sqrt{d} \in k\left(\sqrt{d}\right)$ satisfies the identity $\left(\sqrt{d}\right)^2 = d$.

Now that we know how to extend a field by adding a square root of an element, we give a criterion to decide when a formally real field extends to a formally real field.

Theorem 2.15. *Let k be a formally real field, let P be a preorder of k and let $d \in P$ be a positive element. The field $k\left(\sqrt{d}\right)$ is also formally real and there is a preorder of $k\left(\sqrt{d}\right)$ containing P.*

Proof. If d is a square in k, then $k\left(\sqrt{d}\right) = k$ and there is nothing to prove. Thus, we assume that d is not a square in k. Set

$$P' = \left\{ \sum_{i=1}^{r} c_i \gamma_i^2 \;\middle|\; \begin{array}{l} r \in \mathbb{N}, \\ c_1, \ldots, c_r \in P, \\ \gamma_1, \ldots, \gamma_r \in k\left(\sqrt{d}\right) \end{array} \right\}.$$

It is enough to check that P' is a preorder on $k\left(\sqrt{d}\right)$ containing P. It is clear that P' is closed under addition and multiplication, and that P' contains P and $\sum k\left(\sqrt{d}\right)$. Thus, to check that P' is a preorder, and hence conclude, it suffices to show that P' does not contain -1.

Proceed by contradiction. Suppose that there are an integer $r \geq 0$ and an identity

$$-1 = \sum_{i=1}^{r} c_i \gamma_i^2, \tag{3}$$

with $c_1, \ldots, c_r \in P$, and $\gamma_1, \ldots, \gamma_r \in k\left(\sqrt{d}\right)$. For $i \in \{1, \ldots, r\}$, we write $\gamma_i = a_i + b_i\sqrt{d}$, with $a_i, b_i \in k$. Equation (3) becomes

$$-1 = \sum_{i=1}^{r} c_i \left(a_i + b_i\sqrt{d}\right)^2 = \sum_{i=1}^{r} c_i \left(a_i^2 + b_i^2 d\right) + \left(2\sum_{i=1}^{r} c_i a_i b_i\right)\sqrt{d},$$

implying the identity

$$-1 = \sum_{i=1}^{r} c_i (a_i^2 + b_i^2 d).$$

Since the sum $\sum c_i(a_i^2 + b_i^2 d)$ belongs to P, the last identity contradicts the assumption that P is a preorder. We conclude that P' does not contain -1 and hence P' is indeed a preorder extending P. □

The main tool for the resolution of Hilbert's 17th problem that we use is the Artin-Lang Homomorphism. The statement that we give below uses the concept of finitely generated $\mathbb{R}$-algebra, zero-divisors, field of fractions.

Theorem 2.16 (Artin-Lang Homomorphism)**.** *Let $\mathbb{R}[y_1, \ldots, y_n]$ be a finitely generated $\mathbb{R}$-algebra with no non-zero zero-divisors. If the field of fractions $K(\mathbb{R}[y_1, \ldots, y_n])$ is a formally real field, then there is an $\mathbb{R}$-algebra homomorphism*

$$\varphi\colon \mathbb{R}[y_1, \ldots, y_n] \longrightarrow \mathbb{R}.$$

We do not give a proof of the Artin-Lang Homomorphism Theorem. Rather, we show how we can find a solution to Hilbert's 17th Problem using it.

Theorem 2.17. *Let n be a non-negative integer, let $K = \mathbb{R}(x_1, \ldots, x_n)$ be the rational function field in n variables $x_1, \ldots, x_n$ over $\mathbb{R}$ and let $f \in K$ be a rational function. Suppose that, for all $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ where f is defined, the evaluation $f(a)$ is non-negative. Then, there are a positive integer r and rational functions $f_1, \ldots, f_r \in K$ satisfying the identity*

$$f = \sum_{i=1}^{r} f_i^2.$$

Proof. First, we show that -1 is not a sum of squares of K, that is, the field K is formally real. Indeed, suppose by contradiction that there is an identity

$$-1 = \sum_{i=1}^{r} g_i^2, \tag{4}$$

with $g_1, \ldots, g_r$ rational functions. Let $\alpha \in \mathbb{R}^n$ be an n-tuple where the product of the denominators of $g_1, \ldots, g_r$ does not vanish. Evaluating the identity (4) at α we deduce an expression of -1 as a sum of squares of real numbers. This contradiction shows that K is formally real.

Since K is formally real, the set $\sum K$, the sums of squares of K, is a preorder of K. Our goal is to show that f belongs to $\sum K$.

Recall that $\sum K$ is the intersection of all the orders of K. Thus, to prove the theorem, it suffices to show that f belongs to every order of K.

We proceed by contradiction and assume that there is an order P of K such that $f \notin P$. Since P is an order, the equality $P \cup -P = K$ holds and hence $-f$ is in P. Let $K\left(\sqrt{-f}\right)$ be the field obtained from K by adding a square root of $-f$. Denote $\sqrt{-f}$ by w, so that we have the identity $w^2 = -f$. By Theorem 2.15, the field $K\left(\sqrt{-f}\right)$ is formally real.

Write $f = \frac{g}{h}$, with $g, h \in \mathbb{R}[x_1, \ldots, x_n]$ polynomials and $h \neq 0$. Apply the Artin-Lang Homomorphism Theorem 2.16 to the $\mathbb{R}$-algebra

$$A = \mathbb{R}\left[x_1, \ldots, x_n, \frac{1}{h}, w, \frac{1}{w}\right].$$

We obtain a homomorphism $\varphi\colon A \to \mathbb{R}$. Set $a_1 = \varphi(x_1), \ldots, a_n = \varphi(x_n)$. Evaluating φ at f, we obtain

$$\varphi(f) = \frac{\varphi(g)}{\varphi(h)} = \frac{g(\alpha_1, \ldots, \alpha_n)}{h(\alpha_1, \ldots, \alpha_n)} = f(\alpha_1, \ldots, \alpha_n).$$

By construction, the homomorphism φ is defined at $\frac{1}{h}$, so that the real number $h(\alpha_1, \ldots, \alpha_n) \in \mathbb{R}$ is non-zero. By the same reasoning, the real number $\varphi(w)$ is non-zero, since φ is defined at $\frac{1}{w}$ and $\varphi\left(\frac{1}{w}\right)$ is an inverse of $\varphi(w)$. The equation $w^2 = -f$ shows that the identity

$$f(\alpha_1, \ldots, \alpha_n) = \varphi(f) = -\varphi(w)^2$$

holds. This contradicts the assumption that all the evaluations of f are non-negative. $\square$

Newton polygons and Motzkin's example

We give here an alternative proof of the fact that Motzkin's polynomial of Example 1.1 is not a sum of square of polynomials. To this end, we recall the notion of *convex hulls* of subsets of $\mathbb{R}^n$.

Definition 2.18. Let $A \subset \mathbb{R}^n$ be a set. The *convex hull of* A is the set

$$\operatorname{conv}(A) = \left\{ \sum_{a \in A'} t_a a \;\middle|\; \begin{array}{l} A' \subset A \text{ is finite,} \\ \text{for all } a \in A',\ t_a \geq 0, \\ \displaystyle\sum_{a \in A'} t_a = 1. \end{array} \right\}.$$

The set A if *convex* is the equality $A = \operatorname{conv} A$ holds.

A *polytope in* $\mathbb{R}^n$ is the convex hull of a finite subset of $\mathbb{R}^n$. In particular, a polytope is closed, bounded, convex and its boundary is contained in finitely many affine linear subspaces.

Definition 2.19 (Vertices)**.** Let $P \subset \mathbb{R}^n$ be a polytope. A point $v \in \mathbb{R}^n$ is a *vertex* of P if the only points p and q of P satisfying the identity $v = \frac{1}{2}(p+q)$ are the points $p = q = v$. We denote by $\operatorname{Vert}(P)$ the set of all vertices of P.

We use the following basic fact about polytopes.

Theorem 2.20. *Let $P \subset \mathbb{R}^n$ be a polytope. The set* $\operatorname{Vert}(P)$ *is finite and if P is the convex hull of a set $A \subset \mathbb{R}^n$, then A contains* $\operatorname{Vert}(P)$.

Let n be a non-negative integer. We associate a polytope in $\mathbb{R}^n$ to each polynomial in n variables with coefficients in a field k.

Definition 2.21 (Newton polytope)**.** Let

$$f(x_1, \ldots, x_n) = \sum f_{i_1 \ldots i_n} x_1^{i_1} \cdots x_n^{i_n}$$

be a polynomial in n variables over a field k. The *Newton polytope of* f is the polytope

$$\operatorname{Newt}(f) = \operatorname{conv}\left(\left\{(i_1, \ldots, i_n) \mid f_{i_1 \ldots i_n} \neq 0\right\}\right),$$

obtained as the convex hull in $\mathbb{R}^n$ of the set of exponent vectors of the monomials appearing in f with non-zero coefficient.

The Newton polytope of any polynomial f is a polytope in $\mathbb{R}^n$, regardless of what the base field k of the polynomial f is. If f is a polynomial in two variables, we call the Newton polytope of f, the Newton *polygon*.

For most of our applications, the main consequence of defining the Newton polytope of a polynomial f is that it highlights a finite set of monomials: the *vertices* of $\operatorname{Newt}(f)$.

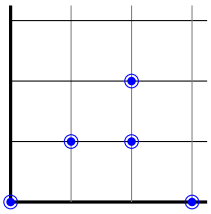
Example 2.22. Let $f(x,y) \in \mathbb{R}[x,y]$ be the polynomial

$$f(x,y) = e \cdot x^3 - \sqrt{2} \cdot x^2y^2 - 7 \cdot x^2y + (\log 2) \cdot xy + \pi.$$

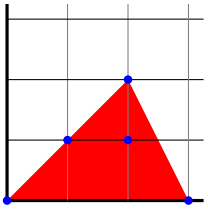
The exponent vectors of the polynomial f are the vectors $(3,0)$, $(2,2)$, $(2,1)$, $(1,1)$, $(0,0)$ in $\mathbb{R}^2$.

exponent vector	$(3,0)$	$(2,2)$	$(2,1)$	$(1,1)$	$(0,0)$
corresponding to	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$	$\updownarrow$
the monomial	x^3y^0	x^2y^2	x^2y^1	x^1y^1	x^0y^0

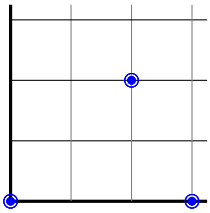
The Newton polygon of f is the convex hull of the exponent vectors of f.



Exponent vectors of f



The Newton polygon of f



The vertices of Newt(f)

Thus, the terms of f corresponding to the vertices of Newt(f) are

$$e \cdot x^3, \qquad -\sqrt{2} \cdot x^2y^2, \qquad \pi \cdot 1.$$

All that matters for the Newton polygon of f is that the coefficients of the monomials $x^3, x^2y^2, 1$ are non-zero and that the polynomial involves those monomials and an arbitrary linear combination of xy, x^2y, x, x^2.

Proposition 2.23. *Let k be a field and let $f \in k[x_1, \ldots, x_n]$ be a polynomial in n variables and coefficients in k. The coefficients of the monomials corresponding to vertices of* Newt(f^2) *are squares.*

Proof. Let m be a monomial in $k[x_1, \ldots, x_n]$. Denote by $e(m) \in \mathbb{R}^n$ the exponent vector of m and by $f_m \in k$ the coefficient of m in the polynomial f. Thus, we have an equality

$$f_m = \sum_{\substack{m_1, m_2 \text{ monomials} \\ m_1 m_2 = m}} f_{m_1} f_{m_2}. \tag{5}$$

The sum need only range over the pairs of monomials m_1, m_2 such that the vectors $e(m_1)$ and $e(m_2)$ belong to P and satisfy $e(m_1) + e(m_2) = e(m)$, since the monomials with exponent vectors not in P have zero coefficient in f. In particular, if m is a monomial such that $e(m)$ is a vertex of Newt(f), then the sum in (5) reduces to the single contribution f_m^2, as needed. □

Corollary 2.24. *Let $f \in \mathbb{R}[x_1, \ldots, x_n]$ be a real polynomial in n variables. If there are r polynomials $f_1, \ldots, f_r \in \mathbb{R}[x_1, \ldots, x_n]$ satisfying the identity $f = f_1^2 + \cdots + f_r^2$, then the polytopes*

$$\operatorname{conv}\left(\bigcup_i \operatorname{Newt}(f_i^2)\right) \qquad \textit{and} \qquad \operatorname{Newt}(f)$$

coincide.

Proof. By assumption, the only monomials that can appear in f with non-zero coefficient are the monomials appearing in at least one of the polynomials $f_1^2, \ldots, f_r^2$. This implies that there is an inclusion

$$\operatorname{Newt}(f) \quad \subset \quad \operatorname{conv}\left(\bigcup_i \operatorname{Newt}(f_i^2)\right).$$

To prove the reverse inclusion, first, observe that there is an equality

$$\operatorname{conv}\left(\bigcup_i \operatorname{Newt}(f_i^2)\right) = \operatorname{conv}\left(\bigcup_i \operatorname{Vert}\left(\operatorname{Newt}(f_i^2)\right)\right). \tag{6}$$

Let v be a vertex of the polytope in (6). By Theorem 2.20, v is therefore a vertex of the Newton polytope of one of the polynomials $f_1^2, \ldots, f_r^2$. Suppose that $i \in \{1, \ldots, r\}$ is an index such that $\operatorname{Newt}(f_i^2)$ contains v. By definition, the vertex v is also a vertex of $\operatorname{Newt}(f_i^2)$. Let m_v be the monomial whose exponent vector is v. Using Proposition 2.23, we deduce that the coefficient f_{m_v} of m_v is a sum of squares of non-zero real numbers. We deduce that f_{m_v} is positive and, in particular, non-zero. We obtain that v is a point in $\operatorname{Newt}(f)$ and we conclude that the containment

$$\operatorname{conv}\left(\bigcup_i \operatorname{Newt}(f_i^2)\right) \quad \subset \quad \operatorname{Newt}(f)$$

also holds. The result follows. □

Our immediate application of Newton polytopes is to verify that an explicit polynomial in 2 variables with real coefficients is not a sum of squares of polynomials. The example is due to Motzkin and the argument uses the following easy lemma.

Recall Motzkin's example: let $m(x, y)$ be the polynomial

$$m(x, y) = x^4y^2 + x^2y^4 + 1 - 3x^2y^2.$$

Lemma 2.25. *The polynomial $m(x, y)$ is not a sum of squares of polynomials with real coefficients.*

Proof. Proceed by contradiction and suppose that there exist polynomials $f_1, \ldots, f_r \in \mathbb{R}[x, y]$ such that the identity

$$m(x, y) = x^4y^2 + x^2y^4 + 1 - 3x^2y^2 = \sum_{i=1}^{r} f_i(x, y)^2$$

holds. By Corollary 2.24, for each $i \in \{1, \ldots, r\}$, the Newton polygon of f_i^2 is contained in the Newton polygon of m. We deduce that the polynomials $f_1, \ldots, f_r$ are linear combinations of the monomials $x^2y, xy^2, 1, xy$ and we can write

$$x^4y^2 + x^2y^4 + 1 - 3x^2y^2 = \sum_{i=1}^{r} (a_i x^2 y + b_i xy^2 + c_i xy + d_i)^2, \tag{7}$$

for some real numbers a_i, b_i, c_i, d_i and $i \in \{1, \ldots, r\}$. Comparing the coefficients of x^2y^2 in (7), we obtain the equation

$$-3 = \sum_{i=1}^{r} c_i^2,$$

which is impossible, since $c_1, \ldots, c_r$ are real numbers. □

Thus, a real polynomial that only takes non-negative values, need not be a sum of squares of real polynomials.

References

[Mot67] T. S. Motzkin, *The arithmetic-geometric inequality*, Inequalities (Proc. Sympos. Wright-Patterson Air Force Base, Ohio, 1965), Academic Press, New York, 1967, pp. 205–224.

[Pfi95] Albrecht Pfister, *Quadratic forms with applications to algebraic geometry and topology*, London Mathematical Society Lecture Note Series, vol. 217, Cambridge University Press, Cambridge, 1995.