CHAPTER XII

Polynomial rings and ideals

12.1. Introduction.

We introduce Zorn's lemma, which is reformulated in the definition of Noetherian rings, and use these rings in turn in the proof of Hilbert's basis theorem. Hilbert's Nullstellensatz is stated and proved. We also discuss Gröbner bases [1BW93], [2Wa07]. The theory used in this chapter does not provide the theory begun by Grothendieck and developed by this school of thought on the idea of schemes. Its contents are the starting point for a discussion of noncommutative sheaves without Galois theory in *Number, space and logic* [Ad18].

12.2. Zorn's lemma, well-ordering and the axiom of choice.

We will introduce Hilbert's basis theorem, which uses Zorn's lemma. In order to compare Zorn's lemma, well-ordering and the axiom of choice, we need to say a little about logic and the foundations of set theory. Leśniewski's ideas probably do not make an alternative [Ur12].

The Zermelo-Fraenkel axioms of set theory with the axiom of choice (ZFC) are expressible in first order logic, and have only one type of object, which is a set.

In first order logic a *predicate* takes an entity as input and outputs either True or False. There are two key parts of first order logic. The syntax determines which collections of symbols are legal expressions in first order logic, while the semantics determine the meanings behind these expressions. Some of the symbols stand for NOT, AND, OR, implication ($\Rightarrow$), equivalence ($\Leftrightarrow$) and equality (=). In some logics these sets of symbols may be reduced, for instance (x OR y) to NOT(NOT x AND NOT y), ($x \Rightarrow y$) to (NOT x) OR y, also ($x \Leftrightarrow y$) to ($x \Rightarrow y$) AND ($y \Rightarrow x$). We can use first order language quantifiers $\forall x$, "for every x", which can also be evaluated as NOT $\exists$ NOT x, and $\exists x$ interpreted as "for some x".

The language of set theory has one binary predicate symbol $\in$. We adopt a modification of ZFC which we introduced in chapter III section 2, and call mZFC (*modified* ZFC). In mZFC, if the axiom of comprehension holds, that if P is a predicate there exists a set $Y = \{x: Px\}$, which can be applied to the set $\{X: X \notin X\}$ (Russell's paradox), then we allow $X = \varnothing$, the empty set, which is also the void set $\odot$ in mZFC and satisfies a false predicate, for instance

$$\odot = \{x: x \in \varnothing \text{ AND } x \notin \varnothing\}.$$

A universal set (or universe) is then the set for a true predicate

$$\mathbb{V} = \{x: x \in \mathbb{V} \text{ OR } x \notin \mathbb{V}\},$$

or as another example

$$\mathbb{V} = \{x: x = x\}.$$

$\odot$ can be expressed indirectly in terms of a true predicate as the complement of $\mathbb{V}$. Nonempty sets obey a predicate which is somewhere true.

A model of set theory provides a universe M and a binary relation E on M that interprets $\in$.

Axiom 12.2.1. *Axiom of choice*. If $x \in X$ and $y \in Y \neq \varnothing$, then there exists a function f on x with $f(y) \in Y$ for all $y \in X$.

Definition 12.2.2. *Total order*. A *partial order* is a binary relation B (a set of ordered pairs, given for example by $\leq$ between them) on a set W when it is reflexive (c B c), antisymmetric (c B d and d B c implies c = d), and transitive (c B d and d B e implies c B e). A total or linear order is a partial order existing for all pairs (c B d).

Definition 12.2.3. *Well-ordering*. A well-ordering on a set W is a total order on W with the property that every non-empty subset of W has a least element $k \in W$ such that $k \leq w$ for every $w \in W$ in this ordering. The set W together with the well-order relation is then called a well-ordered set.

For example, the set $\mathbb{N}_{\cup 0}$ is well-ordered given by $n \leq m$ if and only if $(m - n) \in \mathbb{N}_{\cup 0}$. The set $\mathbb{Z}$ has no minimal element, but may be given a well-ordering under the mapping

$$z \to -2z \text{ for } z \leq 0$$
$$z \to 2z - 1 \text{ for } 1 \leq z. \qquad (1)$$

Definition 12.2.4. Let $\leq$ be a partial order on a set W. An element $m \in W$ is a *maximal element* of W if there is no element $w \in W$, $w \neq m$, such that $m \leq w$.

Axiom 12.2.5. *Zorn's lemma*. Suppose a partially ordered set X has the property that every totally ordered subset (called a *chain*) has an upper bound in X. Then the set X contains at least one maximal element.

Note that under the orderings already selected, X is not $\mathbb{N}$ or $\mathbb{Z}$, since totally ordered subsets of $\mathbb{N}$ or $\mathbb{Z}$ may be constructed $\geq n$ with no upper bound. However, they may be given a partial ordering under a map given by the negative of (1), so that they satisfy Zorn's lemma.

In first order logic, the standard logic for model theory of infinite structures, there are limited things we can say. For example, we might want to express facts about the structure of the set $\mathbb{N}$ of natural numbers, which are defined by the Peano arithmetic of chapter III, section 3, together with arithmetical operations and relations. We include in the language a constant symbol 0 for the number zero, a one place function symbol S for the successor operation which applied to a natural number gives the next one, a two place predicate symbol < for the ordering relation < on $\mathbb{N}$, and two place function symbols + and × for addition and multiplication, respectively, giving the structure $(\mathbb{N}; 0, S, <, +, \times)$.

With this language we can symbolise many of the facts we know to be true about the natural numbers. So we can form the sentence $\forall x(x < Sx)$ expressing that each number is smaller than the next one.

A difficulty arises if we want to express the well-ordering property that any nonempty set of natural numbers has a smallest member. If P is a new one place predicate symbol, then

$$\exists x\, Px \Rightarrow \exists x(Px \text{ AND } \forall y(Py \Rightarrow (y = x \text{ OR } x < y))) \qquad (2)$$

states that P is true of some smallest number, if it is true of any numbers at all. This formula is true in the structure $(\mathbb{N}; 0, S, <, +, \times)$ when we interpret the predicate symbol P as being true of the numbers in some particular set, no matter what that set is. It says that the set has a least member, if it is nonempty. By adding the quantifier $\forall P$

$$\forall P[\exists x\, Px \Rightarrow \exists x(Px \text{ AND } \forall y(Py \Rightarrow (y = x \text{ OR } x < y)))] \qquad (3)$$

we get a formalisation of the well-ordering property.

Nonstandard models of arithmetic exist. To show this, a set of axioms A* is defined in a language including the language of Peano arithmetic together with a new highest constant

symbol $\mathbf{\Omega}_{\mathbb{N}}$. The axioms consist of the axioms of Peano arithmetic A, together with another infinite set of axioms. For each numeral n, the axiom $\mathbf{\Omega}_{\mathbb{N}} > n$ is included. Any finite subset of these axioms is satisfied by a model that is the standard model of arithmetic plus the constant $\mathbf{\Omega}_{\mathbb{N}}$ interpreted as some number larger than any numeral mentioned in the finite subset of A*.

The language of second order logic extends the language of first order logic by allowing quantification of predicate symbols and function symbols. Second order logics contain first order logic. A first order logic obtained from a second order logic is called a *scheme*. As example (3) shows, in a second order language for arithmetic we can say the natural numbers are well-ordered.

We know that the well-ordering property is not expressible by any first order sentence, except in the trivial cases in which a set X has a fixed finite bound on its cardinality, because of the existence of nonstandard models in the first order theory of $(\mathbb{N}; 0, S, <, +, \times)$ that are not well-ordered, for example for $\{\mathbf{\Omega}_{\mathbb{N}} - n; n \in \mathbb{N}\}$. So second order logic is a genuine extension. We can translate some natural language sentences, such as "the relation $\leq$ is a well-ordering" in second order logic that are not translatable into the language of first order logic. □

The *well-ordering principle* (to be distinguished from well-ordering) in second order logic is equivalent to the second order logic version of the axiom of choice. From the well-ordering principle we may deduce the axiom of choice, and from the axiom of choice we can infer the well-ordering principle, which says

for every set X, *there exists a well-ordering with domain* X.

Theorem 12.2.6. *In second order logic Zorn's lemma implies the well-ordering principle.*

Proof. Take the set A of all well-orderings of subsets of X: an element of A is an ordered pair (b, c) where b is a subset of X and c is a well-ordering of b. A can be partially ordered by continuation. That means, define $E \leq F$ if E is an initial segment of F and the ordering of the members in E is the same as their ordering in F.

If E is a chain in A, then the union of the sets in E can be ordered in a way that makes it a continuation of any set in E; this ordering is a well-ordering, and therefore, an upper bound of E in A. We may therefore apply Zorn's Lemma to conclude that A has a maximal element, say (M, S). The set M must be equal to X, because if X has an element *x* not in M, then the set $M \cup \{x\}$ has a well-ordering that restricts to S on M, in which *x* is larger than all elements of M. This well-ordered set is a continuation of (M, S), contradicting its maximality, therefore M = X. Now S is a well-ordering of X. □

In second order set theory, AC is a rather strong principle. It entails what is called global choice, the existence of a single choice function for the entire universe. Indeed, it follows from an instance of the comprehension scheme that there is a binary relation E such that Exy holds if and only if either x is empty or $y \in x$. Then $\forall x \exists$ yExy. An application of AC yields a single function whose value at any non-empty x is a member of x. The graph of this global choice function can be countable in mZFC.

Theorem 12.2.7. *In second order logic the well-ordering principle implies the axiom of choice.*

Proof. To make a choice function for a collection of non-empty sets, E, take the union of the sets in E and call it X. There exists a well-ordering of X. Let S be such an ordering. The function that to each set W of E associates the smallest element of W, as ordered by the restriction to W of S, is a choice function for the collection E.

An essential point of this proof is that it involves only a single arbitrary choice, that of S. Applying the well-ordering theorem to each member W of E separately would not work, since the theorem only asserts the existence of a well-ordering, and choosing for each W a well-ordering would not be easier than choosing an element. □

Theorem 12.2.8. *In second order logic the axiom of* $\mathbf{\Omega}_{\mathbb{N}}$- (*respectively* $\mathbf{\Omega}_{\mathbb{R}}$-) *choice implies the well-ordering principle for a set with cardinality* $\mathbb{N}$ (*respectively* $\mathbb{R}$).

Proof. This result differs from [Sh91], because of our nonstandard countability results. If x and y are sets, define y to have x-choice if, for every function f whose value at each $z \in x$ is a non-empty subset of y, there is a function gf such that if $z \in x$ then gf(z) is a member of f(z). So y has x-choice if there is a choice function for every set of non-empty subsets of y that is 'indexed' by x, and y has $\mathbf{\Omega}_{\mathbb{N}}$-choice if there is a choice function for every countable set of subsets of y. Say that y has continuum-choice if y has $2^{\mathbf{\Omega}}$-choice. These are the same in mZFC. For any reordering of $\mathbb{N}$ the original ordering of $\mathbb{N}$ can be restored in a countable decision tree. Take an ordered subset $\{n_j\}$ of a countable $\mathbb{N}$ not in standard ordering and for a subsequent value m of $\mathbb{N}$ allocate it in $\{n_j\}$ in standard ordering. This is a well-ordering. □

12.3. Applications to rings and ideals. [2Wa07]

Definition 12.3.1. Let $m, n \in \mathbb{Z}$. We say that m *divides* n if $n = rm$ for some $r \in \mathbb{Z}$. The *highest common factor*, h.c.f. (sometimes called the *greatest common divisor*, g.c.d.), of a and $b \in \mathbb{Z}$ is the greatest number $c \in \mathbb{N}$ such that c divides a and c divides b. The *least common multiple* or l.c.m. of a and $b \in \mathbb{Z}$ is the least number $d \in \mathbb{N}$ so that a divides d and b divides d.

Definition 12.3.2. A nonzero ring R (all rings we consider in this chapter are commutative) is called an *integral domain* if R has no other zero divisors than 0.

Theorem 12.3.3. *Any field is an integral domain.*

Proof. Let $\mathbb{F}$ be a field. If $a \in \mathbb{F}$ is a zero divisor in $\mathbb{F}$ then $ab = 0$ for some $b \neq 0 \in \mathbb{F}$, but since $\mathbb{F}$ is a field, b has an inverse b^{-1} in $\mathbb{F}$, and we can write

$$a = a.1 = a\,(b\,b^{-1}) = (ab)b^{-1} = 0b^{-1} = 0,$$

so the only zero divisor in $\mathbb{F}$ is zero. □

Example 12.3.4. If $\mathbb{F}$ is a field then the polynomial ring $\mathbb{F}[x]$ with $f, g \in \mathbb{F}[x]$ is an integral domain, in which

$$f = a_n x^n + \ldots + a_1 x + a_0$$
$$g = b_m x^m + \ldots + b_1 x + b_0,$$

since if $f \neq 0$ and $g \neq 0$, then $fg = 0$ implies

$$b_m a_n x^{m+n} + \ldots + (b_1 a_0 + b_0 a_1)x + b_0 a_0 = 0,$$

so that the coefficients in the product are zero, which means that $\mathbb{F}$ is an integral domain. □

We now discuss principal, prime and maximal ideals. We have

maximal ideals $\subseteq$ *prime ideals* $\subseteq$ *principal ideals.*

Definition 12.3.5. For a ring R with a ∈ R, then as we defined in chapter III, the ideal I given by {ra: r ∈ R} is called a *principal ideal*. We say that the ideal I is *generated* by a, and we will usually write it as (a), or alternatively as Ra or aR if we wish to emphasise that we are multiplying the element a by elements of a particular ring R.

Definition 12.3.6. Let P be a proper ideal of a ring R, that is, P ≠ R. Then the ideal is *prime* if for a, b ∈ R

ab ∈ P *implies* a ∈ P *or* b ∈ P.

For example, (0) is a prime ideal, also

Theorem 12.3.7. *If* n *is a positive integer, then the principal ideal* (n) *is a prime ideal of* ℤ *if and only if* n *is a prime number.*

Proof. Assume n is not prime and n = ab ∈ ℤ. If there are no elements a, b ≠ 1 so that ab cannot be selected with a highest common factor of 1, then a is a nonzero power of b. But if this is the case, then a ∉ (n) and b ∉ (n). If, however, the highest common factor of a and b is 1, then ab ∈ (n), but again a ∉ (n) and b ∉ (n).

Conversely, assume that n is a prime number. Let ab ∉ (n) with a, b ∈ ℤ. We can write ab = rn for some r ∈ ℤ. Thus n divides ab, but n is a prime number, so n divides a or n divides b. □

Theorem 12.3.8. *If* I *is an ideal of a ring* R *and* a, b ∈ R, *then the cosets of* I *satisfy*

(i) a + I = I *if and only if* a ∈ I
(ii) a + I = b + I *if and only if* (a – b) ∈ I
(iii) a + I = b + I *if and only if* (a – b) + I = I
(iv) *two cosets* a + I *and* b + I *are either disjoint or equal.*

Proof. For (i) on assuming a + I = I, we will show that a ∈ I. Now a = a + 0 ∈ a + I, therefore a ∈ I, and conversely if a ∈ I, we will show a + I = I. We can do this if we can demonstrate that the sets a + I ⊆ I and simultaneously I ⊆ a + I. For the first case, if we take a typical element b ∈ I, then a + b ∈ a + I, and since I is an ideal, which means if a and b ∈ I then a + b ∈ I, we must have a + I ⊆ I. For the second part of the proof of (i), let c ∈ I. Then we need to show that c ∈ a + I, but since I is an ideal, -a ∈ I, which implies c – a ∈ I, and so as we wish to prove c = a + (c – a) ∈ a + I. □

For (ii) we will assume that a + I = b + I. Since a = a + 0 ∈ a + I = b + I, we have a ∈ b + I. Thus we can say for some j ∈ I that a = b + j. Then a – b = j ∈ I. On the other hand, if a – b ∈ I then a – b = j for some j ∈ I. As in the similar proof of (i) we will show these two sets a + I and b + I are included in each other, and so are equal. On taking c ∈ I for a + c ∈ a + I, because j + c ∈ I we have a + c = (b + j) + c = b + (j + c) ∈ I. Therefore, a + I ⊆ b + I. To prove b + I ⊆ a + I, we note the symmetry between a and b, in that if a – b ∈ I then b – a ∈ I, and so the first part of the argument carries through again. □

To prove (iii) we merely note that this follows directly from (i) and (ii). □

For the proof of item (iv) suppose a + I and b + I intersect somewhere, at m, say, so they are not disjoint. On writing m = a + j for some j ∈ I and also m = b + k for some k ∈ I, clearly we have a + j = b + k, so that a – b = k – j ∈ I. From condition (ii) a – b ∈ I means a + I = b + I, and these two cosets are equal. □

Definition 12.3.9. If I is an ideal of a ring R, the *quotient ring* or *residue class ring* R/I (spoken as R mod I) is the ring of cosets of I, where addition and multiplication of cosets were given in chapter III by

$$(a + I) + (b + I) = (a + b) + I$$
$$(a + I)(b + I) = (ab) + I.$$

To take an example, the ideal generated by, say, $I = (7) = 7\mathbb{Z}$, maps bijectively addition and multiplication of cosets to the finite congruence arithmetic denoted by (mod 7)

$$[a \pmod 7] + [b \pmod 7] = (a + b) \pmod 7,$$
$$[a \pmod 7][b \pmod 7] = ab \pmod 7.$$

Theorem 12.3.10. *If* P *is an ideal of a ring* R, *then* P *is a prime ideal if and only if* R/P *is an integral domain*.

Proof. We will assume firstly that the ideal P is prime. We need to show that R/P has no zero divisors other than zero, so that it is an integral domain. We will write the zero element of the ring R/P as P, which is 0 + P. Suppose a + P, $a \in R$, is a zero divisor of R/P. Then there exists another $b \in R$ such that $b + P \neq P$, which means $b \notin P$, satisfying $(a + P)(b + P) = 0 + P = P = ab + P$, implying $ab \in P$. Since P is a prime ideal, by definition $a \in P$ or $b \in P$, and we have proved that $b \notin P$, so $a \in P$. Thus, $a + P = P$ and the only zero divisor of R/P is P, which is a statement that R/P is an integral domain.

For the converse, if we assume R/P is an integral domain, we will need to show that P is a prime ideal. Suppose a, $b \in R$ with $ab \in P$. Then $(a + P)(b + P) = ab + P = P$, and because integral domains have no nonzero divisors, we find either $a + P = 0 + P = P$ or $b + P = P$, which is the definition of a prime ideal. □

Definition 12.3.11. Let M be a proper ideal in a ring R, that is, $M \neq R$. M is a *maximal ideal* if the only ideal properly containing M is the whole ring R itself.

For example (3), denoted by $3\mathbb{Z}$, and $7\mathbb{Z}$ are maximal ideals of $\mathbb{Z}$, since there is no ideal between them and $\mathbb{Z}$, but $6\mathbb{Z}$ is not maximal, which is included in $2\mathbb{Z}$ in turn included in $\mathbb{Z}$.

All maximal ideals are prime ideals, which we will soon prove.

Theorem 12.3.12. M *is a maximal ideal of a ring* R *if and only if* R/M *is a field*.

Proof. Suppose M is a maximal ideal. This implies $M \neq R$, with R/M a nonzero ring. We can demonstrate R/M is a field by showing every nonzero element of R/M has a multiplicative inverse. To do this, let $a + M \neq 0 \in R/M$, so $a \notin M$. Now let the ideal T be generated by M and a. Because $a \notin M$, $M \subset T$, but M is maximal so this implies T = R. But R has an element 1, so we can find an $m \in M$ and $r \in R$ such that $1 = m + ar$. The inverse we are looking for is in fact r + M, since

$$(a + M)(r + M) = ar + M = 1 - m + m = 1 + M, \qquad (1)$$

from which it follows that R/M is a field.

To prove the converse, now assume R/M is a field. To prove M is a maximal ideal, note that R/M is nonzero, so $M \neq R$. If $M \subset T$, where T is an ideal, we have to demonstrate that R = T. Say $a \in T$ and $a \notin M$. Then $a + M \neq 0 \in R/M$ has an inverse b + M, because R/M is a field, so

$$1 + M = (a + M)(b + M) = ab + M. \qquad (2)$$

Then by property (ii) of theorem 12.3.8, we can define m by $1 - ab = m \in M$. The elements a and $m \in T$, so $1 = ab + m \in T$, from which T = R and M is a maximal ideal. □

Corollary 12.3.13. *Every maximal ideal is a prime ideal.*

Proof. For M a maximal ideal of a ring R, by theorem 12.3.12, we have shown R/M is a field, and since by theorem 12.3.3 any field is an integral domain, by theorem 12.3.10, M is a prime ideal. □

By definition, any ideal in a principal ideal domain is generated by a single element, hence the union of an ascending chain of ideals is itself an ideal that is generated by a single element, which, in turn, must be one of the ideals in the chain, forcing the chain to terminate at this ideal.

More generally, if R is a ring such that any ideal R is generated by a finite number of elements, then R satisfies the *ascending chain* condition:

every ascending chain of ideals

$$I_1 \subseteq I_2 \subseteq I_3 \subseteq ...$$

terminates, that is, there is an integer k *such that*

$$I_1 \subseteq I_2 \subseteq ... \subseteq I_k = I_{k+1} = ...$$

A ring R is called *Noetherian* if it satisfies the ascending chain condition.

Zorn's lemma can be used to show the property, N, of being a Noetherian ring is equivalent to the statement that every nonzero ring has a maximal ideal, or alternatively that every ideal is finitely generated.

Example 12.3.14. (0) is a maximal ideal of $7\mathbb{Z}$, but (0) is not a maximal ideal of $\mathbb{Z}$, since $(0) \subset (2) \subset \mathbb{Z}$.

Theorem 12.3.15. *Every nonzero ring has at least one maximal ideal*.

Proof. If R is a nonzero ring, let T be the set of all proper ideals of R with a partial order given by set inclusion. As we have noted in example 12.3.14, T is nonempty because (0) is a proper ideal of R. A maximal element of the set T is a maximal ideal of R.

To apply Zorn's lemma, we will show that any chain in T has an upper bound in T, which is the union of all the ideals in the chain, C.

In order to show that this union, U, is a proper ideal of R, we need to demonstrate that if a and $b \in U$ then $a - b \in U$. Now a is in some ideal I in the chain C, as is $b \in J$ for some ideal J. Since C is totally ordered, either $I \subseteq J$ or $J \subseteq I$, and arbitrarily assuming the first, $a \in I \subseteq J$ with $b \in J$, so $a - b \in J$, which because U contains J implies that $a - b \in U$.

Next, if $a \in U$ and $r \in R$, we must demonstrate that $ra \in U$. Because $a \in I$ for some ideal I in the chain C, so $ra \in I \subseteq U$, U is an ideal of R.

Note that U is a proper ideal of R since $1 \notin I$ for any I in the chain C. Applying Zorn's lemma, T has a maximal element, which is a maximal ideal of R. □

12.4. Hilbert's basis theorem. [2Wa07]

Hilbert's basis theorem states that if R is a commutative Noetherian ring, then R[x] also has property N of being Noetherian. This is an existence statement and not an algorithm.

We can summarise Hilbert's first step as

if a ring R *has property* N, *then so does the ring* R[x].

Hilbert wanted then to show that the ring of polynomials in two variables x and y with coefficients from the ring R also has property N. We represent this ring of polynomials in two variables by R[x, y].

He saw that this ring of polynomials R[x, y] could be thought of as the ring of polynomials in a single variable, x, concluding that the ring of polynomials R[x, y] also has property N.

In this way, by adding one variable at a time, Hilbert proved that any finite number of variables has the property N. He did not construct a basis, but showed if there were no finite bases, then a contradiction results.

Definition 12.4.1. The *leading coefficient* of a polynomial is the coefficient of its highest power term.

Lemma 12.4.2. *The leading coefficients of all the polynomials in an ideal form an ideal.*

Proof. Let a_m and b_n respectively be two leading coefficients for terms with degrees m for polynomial f and n for polynomial g in a polynomial ideal I, and J be the set of all leading coefficients in I. Assume $m \leq n$. Then $x^{n-m}f - g \in I$ and has leading coefficient $a_m - b_n$. Now let $a_m \in J$ and $r \in R$. Then $rf \in I$, so ra_m is the leading coefficient of a polynomial in I, giving $ra_m \in J$. □

Theorem 12.4.3. (Hilbert's basis theorem). *If a ring* R *has the Noetherian property, then so has the ring* R[x].

Proof. For an ideal I in R[x] we wish to show that this is finitely generated. For each positive natural number n, let J_n be its corresponding leading coefficient ideal in I. Then

$$J_0 \subseteq J_1 \subseteq J_2 \subseteq ...$$

is an ascending chain of ideals in the Noetherian ring R, and so terminates as

$$J_0 \subseteq J_1 \subseteq ... \subseteq J_k = J_{k+1} = ...$$

so the J_n are finitely generated, which gives a finite set of polynomials, which generates the polynomial ideal I. □

12.5. Hilbert's Nullstellensatz. [Ta07]

This is a development of Terry Tao's decision tree proof of the Nullstellensatz (zero locus theorem), using symbolic logic and zero algebras instead. Let $\mathbb{F}$ be a field, for instance the complex numbers, and $\mathbb{F}[x]$ the ring of polynomials in the variable x with coefficients in $\mathbb{F}$. We state the Nullstellensatz in a fairly concrete way, in terms of solving a set of simultaneous polynomial equations $T_1(x) = ... = T_n(x) = 0$ in a number of variables $x = (x_1, ..., x_d) \in \mathbb{F}^d$ over $\mathbb{F}$, with polynomials $T_1, ..., T_n \in \mathbb{F}[x]$ in d variables. One obvious obstruction to solvability of this system of equations is if they are *inconsistent,* in the sense that they can be used to imply $1 = 0$.

If we can find polynomials $U_1, \ldots, U_n \in \mathbb{F}[x]$ so that $T_1U_1 + \ldots + T_nU_n = 1$, then clearly by substitution we cannot solve $T_1(x) = \ldots = T_n(x) = 0$ simultaneously. The *weak Nullstellensatz* asserts that this is, in fact, the only obstruction.

Proposition 12.5.1 (Weak Nullstellensatz). *Let* $T_1, \ldots, T_n \in \mathbb{F}[x]$ *be polynomials. Exactly one of the following statements holds*:

A. *The system of equations* $T_1(x) = \ldots = T_n(x) = 0$ *has a solution* $x = (x_1, \ldots, x_d) \in \mathbb{F}^d$.

B. *There exist polynomials* $U_1, \ldots, U_n \in \mathbb{F}[x]$ *such that* $T_1U_1 + \ldots + T_nU_n = 1$. □

Note that a hypothesis that $\mathbb{F}$ is a fixed algebraically closed field, which means it contains a root for every non-constant polynomial, is not crucial. For instance, if $\mathbb{F}$ is the Eudoxus (real, which are countable) numbers then the equation $x^2 + 1 = 0$ has no solution, but there is a polynomial $U(x) = 1$ such that $(x^2 + 1)U(x) = 1$, namely the solution $x = 0$.

Like many results of "the only obstructions are the obvious obstructions" type, the power of the Nullstellensatz lies in the ability to take a hypothesis about *nonexistence*, in this case nonexistence of solutions to $T_1(x) = \ldots = T_n(x) = 0$, and deduce a conclusion about *existence*, in this case existence of $U_1, \ldots, U_n$ such that $T_1U_1 + \ldots + T_nU_n = 1$. The ability to get "something from nothing" is clearly going to be both nontrivial and useful. In particular, the Nullstellensatz offers an important correspondence between algebraic geometry – conclusion A is an assertion that a certain algebraic variety is empty, and commutative algebra – conclusion B is an assertion that a certain ideal is not proper.

In order to prove the Nullstellensatz, we first collect together a number of propositions using truth tables in symbolic logic.

Lemma 12.5.2. *For sentences* A, B, X *and* Y, *let* T *indicate* True *and* F False. *The following truth tables in propositional calculus are defined, or applied give the following results.*

NOT A	A
F	T
T	F

NOT

X AND Y	X	Y
T	T	T
F	F	T
F	T	F
F	F	F

AND

A ⇒ B	B ⇒ A	A	B
T	T	T	T
T	F	F	T
F	T	T	F
T	T	F	F

IMPLIES

A ⇔ B	A	B
T	T	T
F	F	T
F	T	F
T	F	F

IF AND ONLY IF

A XOR B	A	B
F	T	T
T	F	T
T	T	F
F	F	F

EXCLUSIVE OR

1. A XOR B = NOT(A ⇔ B)
2. A ⇔ B = (A ⇒ B) AND (B ⇒ A)
3. If X is False, then X AND Y is False. □

Theorem 12.5.3. *The weak Nullstellensatz fails if and only if statements* A *and* B *of proposition* 12.5.1 *are false simultaneously.*

Proof. Statements A and B for the weak Nullstellensatz are True or False and are mutually exclusive, so they satisfy the XOR truth table of the lemma, A XOR B. To give a proof, we will assume the contrary of the Nullstellensatz and prove the condition for a contradiction. This means we assume NOT(A XOR B), which is the same as A $\Leftrightarrow$ B, which in turn is the same as (A $\Rightarrow$ B) AND (B $\Rightarrow$ A). But we have already stated that if $T_1(x) = ... = T_n(x) = 0$ then $T_1U_1 + ... + T_nU_n = 1$ cannot hold, in other words if A is True then the only value of B is False, so that the statement (A $\Rightarrow$ B) is False. Likewise if B is True then the only value of A is False and (B $\Rightarrow$ A) is False. But this means (A $\Leftrightarrow$ B) is either restricted to the middle pair of rows in its truth table and can only be False, which is a required contradiction, or A and B are both False. □

Suppose we now try to solve the more complicated system $T_1(x) = ... = T_n(x) = 0$; $W(x) \neq 0$ for some polynomials $T_1, ... , T_n, W$. Again, any identity of the form $T_1U_1 + ... + T_nU_n = 1$ will be an obstruction to solvability, but now more obstructions are possible: any identity of the form $T_1U_1 + ... + T_nU_n = W^r$ for some nonnegative integer r will also obstruct solvability. The *strong Nullstellensatz* asserts that this is the only obstruction.

Proposition 12.5.4 (Strong Nullstellensatz). *Let* $T_1, ... , T_n, W \in \mathbb{F}[x]$ *be polynomials. Then exactly one of the following statements holds*:

C. *The system of equations* $T_1(x) = ... = T_n(x) = 0, W(x) \neq 0$ *has a solution* $x \in \mathbb{F}^d$.

D. *There exist polynomials* $U_1, ... , U_n \in \mathbb{F}[x]$ *and a non-negative integer* r *such that* $T_1U_1 + ... + T_nU_n = W^r$. □

We could consider generalising the Nullstellensatz a bit further by considering systems of the form $T_1(x) = ... = T_n(x) = 0, W_1(x), ... , W_m(x) \neq 0$, but this is not a significant generalisation, since we can string together all the inequations $W_1(x) \neq 0, ... , W_m(x) \neq 0$ into a single inequation $W_1(x) ... W_m(x) \neq 0$.

Theorem 12.5.5. *The weak Nullstellensatz is equivalent to the strong Nullstellensatz.*

Proof. The strong Nullstellensatz implies the weak Nullstellensatz when W = 1.

To prove the weak Nullstellensatz implies the strong Nullstellensatz we use the *Rabinowitsch trick* of introducing an extra variable. Suppose the polynomial W in $\mathbb{F}[x_1, ... , x_d]$ vanishes whenever all polynomials $T_1, ... , T_n$ vanish. Then the polynomials $T_1, ... , T_n, 1 - x_0W$ have no common zeros (we have introduced a new variable x_0), thus by the weak Nullstellensatz for $\mathbb{F}[x_0, ... , x_d]$ they generate the unit ideal of $\mathbb{F}[x_0 , ... , x_d]$.

Spelt out, this means there are polynomials $U_0, U_1, ... , U_n \in \mathbb{F}[x_0, x_1, ... , x_d]$ such that

$$U_0(x_0, x_1, \cdots, x_d)\left(1 - x_0W(x_1, \cdots, x_d)\right) + \sum_{i=1}^{n} U_i(x_0, x_1, \cdots, x_d)T_i(x_1, \cdots, x_d) = 1,$$

as an equality of elements of the polynomial ring $\mathbb{F}[x_0, x_1, ..., x_d]$. Since $x_0, x_1, ..., x_d$ are free variables, this equality continues to hold if expressions are substituted for some of the variables; in particular, it follows from substituting $x_0 = 1/W(x_1, \cdots, x_d)$ that

$$1 = \sum_{i=1}^{n} U_i(1/W(x_1, \cdots, x_d), x_1, \cdots, x_d)T_i(x_1, \cdots, x_d),$$

as elements of the field of rational functions $\mathbb{F}(x_1, \ldots, x_d)$, the field of fractions of the polynomial ring $\mathbb{F}[x_1, \ldots, x_d]$. Moreover, the only expressions that occur in the denominators of the right hand side are W and powers of W, so rewriting that right hand side to have a common denominator results in an equality of the form

$$1 = \frac{\sum_{i=1}^{n} V_i(x_1, \cdots, x_d) T_i(x_1, \cdots, x_d)}{W(x_1, \cdots, x_d)^r}$$

for some natural number r and polynomials $V_1, \ldots, V_n \in \mathbb{F}[x_1, \ldots, x_d]$. Hence

$$W(x_1, \cdots, x_d)^r = \sum_{i=1}^{n} V_i(x_1, \cdots, x_d) T_i(x_1, \cdots, x_d),$$

which literally states that W^r lies in the ideal generated by $T_1, \ldots, T_n$. This is the strong version of the Nullstellensatz for $\mathbb{F}[x_1, \ldots, x_d]$. □

Theorem 12.5.6. *The Nullstellensatz of equivalent propositions of* 12.5.1 *and* 12.5.4 *holds.*

Proof. We will use statements A and B from the weak Nullstellensatz proposition 12.5.1 and C and D from the strong Nullstellensatz proposition 12.5.4. If statement A is always false for n simultaneous equations then it must be true for a rearranged first $T_1(x), \ldots, T_{n-k}(x) = 0$, with $0 < k < n$, ($k = n$ is not a statement of the theorem) and false for the remainder, which means the statement C holds for the corresponding set of $(n - k)$ simultaneous equations and the inequality $W(x) \neq 0$. If statement B is false, so there does not exist even one

$$T_1U_1 + \ldots + T_nU_n = 1$$

then dividing by any Eudoxus number $1/v \neq 0$ for any such v there do not exist solutions

$$T_1V_1 + \ldots + T_nV_n = v,$$

in other words there can only exist solutions $1/v = 0$, or using the zero algebra ultrainfinity Ч of chapter III section 4

$$T_1V_1 + \ldots + T_nV_n = Ч, \qquad (1)$$

and solutions do not exist in a field. Now since we are assuming C holds, if the strong Nullstellensatz does not hold then statement D also holds where

$$T_1V_1 + \ldots + T_{n-k}V_{n-k} = W^r,$$

so that by (1), there only exist solutions

$$Ч - T_{n-k+1}V_{n-k+1} - \ldots - T_nV_n = W^r,$$

which are outside the values in a field. This contradicts statement C, since W(x) is in a field.

Thus A and B cannot be false simultaneously, and by theorem 11.10.3 the Nullstellensatz holds. □

12.6. Gröbner bases.

Gröbner bases were introduced in 1965 with an algorithm to compute them (Buchberger's algorithm), by Bruno Buchberger in his Ph.D. thesis. He named them after his thesis advisor Wolfgang Gröbner, who had been using similar ideas. However, they were first discovered by N. M. Gjunter in 1913 [Gj1913], who published in various Russian mathematical journals.

The basic idea is, firstly, given a finite set of polynomials in one variable $\{f, g_1, \ldots g_m\}$ over a field, to decide membership of f in the ideal generated by the g_k. This is done by computing the greatest common divisor (g.c.d.) of the g_k using the polynomial Euclidean algorithm. Any given polynomial lies in the ideal in question if and only if its remainder on division of f by the g.c.d. of the g_k is zero.

Gröbner basis theory imitates this procedure for polynomials in several variables. Given a finite set of such polynomials over a field, the Buchberger algorithm computes a new set of

polynomials, called a Gröbner basis, which generates the same ideal as the original one and is an analogue of the Euclidean division algorithm for polynomials in one variable. A given polynomial lies in the ideal generated by the Gröbner basis if and only if a suitably defined normal form of the polynomial with respect to the Gröbner basis equals zero, the computation of which is a straightforward generalisation of the Euclidean case.

A complex polynomial in one variable, x, may not be solvable by radicals, yet for any such polynomial, on substituting for x the variable y + iz, we obtain two polynomials in two variables, each of which, as we have seen in section 5, can be reduced to a polynomial in one variable, w, say. Since there exist solutions in which y and z are indeed Eudoxus (so that iz is complex), there exist two polynomials in w with Eudoxus solutions, and there exist solutions in which the roots of w are Eudoxus, even if these are not computable by radicals.

It is the case that any selection of complex variables for the original equation in x can be reduced to an approximation in which its coefficients are rational numbers in both their real and imaginary parts. By an extension of the method which follows for polynomials in many variables, it is possible to obtain a polynomial remainder on division by another polynomial of not greater degree which approximates with arbitrary precision a general complex polynomial, and the limit of this exists. The coefficients which we use next will therefore be rational, or on multiplying out, integers, and this describes in the limit complex polynomials.

We wish to be able to divide a polynomial in many variables by another polynomial of lesser degree. We have met in chapter II, section 9 how to divide one integer by another leaving a quotient and a remainder – this is known as the Euclidean algorithm. So in the terminology we have been using, we have a polynomial $\mathbb{F}[x_1, \ldots, x_d]$ and we want to factorise it as

$$\mathbb{F}[x_1, \ldots, x_d] = I_1[x_1, \ldots, x_d] \ldots I_n[x_1, \ldots, x_d] + r[x_1, \ldots, x_d],$$

where the product $I_1, \ldots I_n$ is a sequence of ideals = I, polynomials are not of greater degree than $\mathbb{F}$, and r is a remainder polynomial. We will develop an extended Euclidean algorithm for polynomials, but find that the remainder using this process is not unique. In order to restore uniqueness, we first start by introducing an ordering of a polynomial, which can be of various types, and then further restrict the division process using Gröbner bases – specific types of polynomials in which the ideal I is represented additively by the Gröbner bases.

Two difficulties arise, which we will solve.

(i) There does not exist a single generator of the ideal in general, and so the polynomial has to be divided by a set of polynomials.

(ii) The terms have to be ordered in a way that replaces the natural ordering of polynomials in one variable.

A *monomial* in a polynomial is a term without the coefficient. Thus for $(x^2y + 3xy – 4y^2)$ its three monomials are x^2y, xy and y^2.

The order selected for terms in an additive polynomial depends on two features, the order of the variables, x, y, z etc. and the degree of each variable, like x^2. We wish to combine the two types of order into one type, called a *monomial order*.

A *lexicographical order*, $<_{lex}$, between the monomials, say, $x^ay^bz^c$ and $x^dy^ez^f$ satisfies

$$x^ay^bz^c <_{lex} x^dy^ez^f$$

whenever, starting from the left, the variables x, y and z are in alphabetic order and the first power that differs for a variable satisfies, say if we have a = d for the variable x, that b < e for the next variable, y say. Lexicographic order is one of many such monomial orders.

The corresponding order, $>_{lex}$, between the monomials $x^a y^b z^c$ and $x^d y^e z^f$ satisfies

$$x^a y^b z^c >_{lex} x^d y^e z^f$$

whenever, starting from the left, the variables x, y and z are in alphabetic order and the first power that differs for a variable satisfies, say if we have a = d for the variable x, that b > e for the next variable, y.

Example 12.6.1. We will divide the polynomial $(x + 2y + 1)$ into $(x^2y + 3xy - 3y^2)$, where we have arranged the polynomials in $>_{lex}$ order of the variables x first and y second with the order of the additive terms determined by

$$x >_{lex} y, \qquad y >_{lex} 1,$$
$$x^2y >_{lex} xy, \qquad xy >_{lex} y^2.$$

We start dividing with the leading term, x, of $(x + 2y + 1)$. Since

$$xy(x + 2y + 1) = (x^2y + 2xy^2 + xy)$$

we obtain a remainder in $>_{lex}$ order

$$(x^2y + 3xy - 3y^2) - (xy)(x + 2y + 1) = (-2xy^2 + 2xy - 3y^2).$$

Continuing this process

$$(-2xy^2 + 2xy - 3y^2) - (-2y^2)(x + 2y + 1) = (2xy + 4y^3 - y^2)$$

and finally

$$(2xy + 4y^3 - y^2) - (2y)(x + 2y + 1) = (4y^3 - 5y^2 - 2y).$$

We can no longer continue dividing by x since under $>_{lex}$ order x is greater than $4y^3$. This terminates the process leaving the remainder $(4y^3 - 2y^2 - 2y)$. Thus

$$(x^2y + 3xy - 3y^2) = (xy - 2y^2 + 2y)(x + 2y + 1) + (4y^3 - 5y^2 - 2y). \square$$

We will now divide the polynomial $x^2y + x^2 + xy^2$ by two polynomials in two different ways, and get conflicting results. The polynomials by which we will be dividing will be $(x^2 + y)$ and $(xy - 1)$ respectively, and we will find the result depends on the order in which we do this division.

Example 12.6.2. We will divide the polynomials $(x^2 + y)$ and $(xy - 1)$ into $(x^2y + x^2 + xy^2)$ using the lexicographic order $>_{lex}$ again, starting with the polynomial $(x^2 + y)$. We begin by noting that

$$(x^2y + x^2 + xy^2) - (y)(x^2 + y) = (x^2 + xy^2 - y^2),$$

then

$$(x^2 + xy^2 - y^2) - (1)(x^2 + y) = (xy^2 - y^2 - y),$$

for which on finding $(x^2 + y) >_{lex} (xy^2 - y^2 - y)$, we see that $(xy^2 - y^2 - y)$ is the remainder.

Continuing with

$$(y + 1)(x^2 + y) + (xy^2 - y^2 - y),$$

we will now divide the remainder $(xy^2 - y^2 - y)$ by $(xy - 1)$. Then

$$(xy^2 - y^2 - y) - (y)(xy - 1) = -y^2,$$

and this is the final remainder, since $(xy - 1) >_{lex} (-y^2)$. Summarising

$$(x^2y + x^2 + xy^2) = (y + 1)(x^2 + y) + (y)(xy - 1) + (-y^2). \qquad (1)$$

We will now begin the division in the swapped order $(xy - 1)$ first and $(x^2 + y)$ second. We start with the factorisation

$$(x^2y + x^2 + xy^2) - (x)(xy - 1) = (x^2 + xy^2 + x),$$

then there is no way (xy) can be divided into the leading term of $(x^2 + xy^2 + x)$, so this is the remainder under division by $(xy - 1)$. We now switch to division by $(x^2 + y)$.

$$(x^2 + xy^2 + x) - (1)(x^2 + y) = (xy^2 + x - y),$$

and $(xy^2 + x - y)$ is immediately the remainder, since $(x^2 + y) >_{lex} (xy^2 + x - y)$.

If we now think we have finished, we can observe that, having done this division, we can go back to dividing the remainder $(xy^2 + x – y)$ by $(xy – 1)$. Then

$$(xy^2 + x – y) – (y)(xy – 1) = x,$$

and this is the final remainder, so

$$(x^2y + x^2 + xy^2) = (x + y)(xy – 1) + (1)(x^2 + y) + x. \qquad (2)$$

It is now clear that (1) and (2) are two different factorisations of the same polynomial. □

Not only is factorisation not unique, as we will now demonstrate the question which we can answer using the Euclidean algorithm as to whether on division there is any nonzero remainder has no such answer for the forms of division we have been using for polynomials.

Example 12.6.3. We will now divide the polynomial $(y^3 – xy)$ by the polynomials $(y – 1)$ and $(xy – 1)$, using $>_{lex}$ order, and ask whether $(y^3 – xy)$ belongs to the ideal generated by the polynomials written as $(y – 1, xy – 1)$. Then starting with division by $(y – 1)$ we get the sequence

$$(-xy + y^3) – (-x)(y – 1) = -x + y^3$$
$$(-x + y^3) – (y^2)(y – 1) = -x + y^2$$
$$(-x + y^2) – (y)(y – 1) = -x + y$$
$$(-x + y) – (1)(y – 1) = -x + 1,$$

so the remainder by this process is $(-x + 1)$.

Now starting with $(xy – 1)$ we obtain

$$(-xy + y^3) – (-1)(xy – 1) = y^3 – 1,$$

which does not contain xy, so we switch to $(y^3 – 1)$, giving a zero remainder from

$$(y^3 – 1) – (y^2 + y + 1)(y – 1) = 0.$$ □

A Gröbner basis G_B of an ideal I in a polynomial ring $\mathbb{F}[x]$ over a field $\mathbb{F}^d$ satisfies any one of the following equivalent properties, stated relative to some monomial ordering:

(i) The leading term of any polynomial in I is divisible by the leading term of some polynomial in the basis G_B.

(ii) The ideal given by the leading terms of polynomials in I is itself generated by the leading terms of the basis G_B.

(iii) Multivariate division of any polynomial in the ideal I by G_B gives remainder 0.

(iv) Multivariate division of any polynomial in the polynomial ring $\mathbb{F}[x]$ by G_B gives a unique remainder.

We will go through these reformulations in sequence, taking the first item as a definition of a Gröbner basis G_B.

Definition 12.6.4. (i) Let $\mathbb{F}[x_1, ..., x_d]$ be a ring of polynomials in d variables in monomial order over a field $\mathbb{F}$. A set of polynomials $\{g_1, ..., g_n\}$ forms a *Gröbner basis* G_B for the ideal $I = (g_1, ..., g_n)$ if for each nonzero polynomial $f \in I$, there is some g_i with the leading term of g_i divisible by the leading term of f.

Theorem 12.6.5. (ii) *In monomial order, let* $\mathbb{F}[x_1, ..., x_d]$ *be a ring of polynomials in* d *variables over a field* $\mathbb{F}$. *If* I *is an ideal in this ring, a set of polynomials* $\{g_1, ..., g_n\}$ *forms a Gröbner basis for the ideal if for each nonzero polynomial* $f \in \mathbb{F}[x_1, ..., x_d]$, *with*

$$f = p_1g_1 + ... + p_ng_n + r$$

where no leading term in $g_1, ..., g_n$ *divides any term in the remainder r, or* $r = 0$, *then*

$f \in I$ *if and only if* $r = 0$.

Proof. Let r = 0, then f ∈ $\{g_1, ..., g_n\} \subseteq$ I. In the other direction, if f ∈ I then f – r ∈ I, which gives r ∈ I. Consequently r = 0, because if it did not then leading terms of $g_1, ..., g_n$ would divide any term of r, a contradiction of a remainder for a Gröbner basis. □

Theorem 12.6.6. (iii) If $\{g_1, ..., g_n\}$ *is a Gröbner basis for the ideal* I, *then* $(g_1, ..., g_n)$ = I.

Proof. We have already shown in theorem 11.11.5 that I $\subseteq (g_1, ..., g_n)$, since the remainder, r, is zero for f ∈ I, but also $(g_1, ..., g_n) \subseteq$ I. □

Theorem 12.6.7. (iv) *In monomial order, let* $\mathbb{F}[x_1, ..., x_d]$ *be a ring of polynomials in* d *variables over a field* $\mathbb{F}$. *If* I *is an ideal in this ring, let a set of polynomials* $\{g_1, ..., g_n\}$ *form a Gröbner basis for the ideal. If* f ∈ $\mathbb{F}[x_1, ..., x_d]$, *then division of* f *by* $g_1, ..., g_n$ *using the analogue of the Euclidean division algorithm gives a unique remainder* r.

Proof. Consider two possible remainders

$$f = p_1g_1 + ... + p_ng_n + r = p_1g_1 + ... + p_ng_n + s.$$

Since for an ideal r – s ∈ I, for a Gröbner basis if r – s ≠ 0 then for some k the leading term of g_k divides the leading term of r – s, which contradicts the definition of a Gröbner basis. Thus r = s. □

Theorem 12.6.8. *In monomial order, let* $\mathbb{F}[x_1, ..., x_d]$ *be a ring of polynomials in* d *variables over a field* $\mathbb{F}$. *If* I *is an ideal in this ring, let a set of polynomials* $\{g_1, ..., g_n\}$ *form a Gröbner basis for the ideal. If* f, g ∈ $\mathbb{F}[x_1, ..., x_d]$, *with respective remainders* r *and* s, *then*

f + I = g + I *if and only* if r = s.

Proof. If r = s, by the definition of an ideal f – g ∈ I, implying f + I = g + I. Conversely, if we let f + I = g + I, which gives f – g and r – s ∈ I, then by the previous theorem provided r ≠ 0 and s ≠ 0, by the definition of a Gröbner basis, no leading term of a g_k divides the leading term of r – s, so r – s = 0. □

Example 12.6.9. Consider the polynomial f. We can prove by contradiction on assuming the opposite that {x + z, y – z} is a Gröbner basis in $\mathbb{F}[x, y, z]$ given under the lexicographic order x > y > z, since if this were not a Gröbner basis no leading term, x or y, divides the nonzero polynomial f, so the leading term must contain z, f ∈ [z] and

$$f = f_1(x + z) + f_2(y – z)$$

where f_1 and $f_2 \in \mathbb{F}[x, y, z]$. If in this expression we replace every y by a z the left hand side does not change, since f ∈ [z], and now

$$f = g_1(x + z) + g_2(z – z) = g_1(x + z),$$

so we arrive at the contradiction that (x + z) divides f, since f ∈ $\mathbb{F}[z]$.

On putting f = xyz and dividing by x, the leading term of x + z, we obtain

$$xyz = yz(x + z) – yz^2,$$

then dividing by (y – z) gives

$$xyz = yz(x + z) – z^2(y – z) – z^3.$$

If we do the division in the reverse order, then we get

$$xyz = xz(y – z) + z^2(x + z) – z^3.$$

In this example the remainder stays the same as we expect, but the quotients are not unique, even for a Gröbner basis. □

12.7. S-polynomials.

There exists an algorithm to find a Gröbner basis of I, called Buchberger's algorithm. We can view it as a generalisation of the Euclidean algorithm for finding the greatest common divisor in one variable and of Gaussian elimination for linear systems.

The reader is referred to the excellent books [1BW93] and [2Wa07] for a description of the structure and properties of Gröbner bases, and algorithms to find them.

We could consider this algorithm in the following way.
(i) Prove the existence of a Gröbner basis.
(ii) Construct a Gröbner basis.
(iii) Find an efficient algorithm to check for a Gröbner basis.
(iv) Reduce the amount of computation in the construction of a Gröbner basis.

In what follows in this section, $\mathbb{F}[x_1, ..., x_d]$ is a ring of polynomials with a monomial order, in a number d of variables over a field $\mathbb{F}$. $G = \{g_1, ..., g_m\} \neq \varnothing$ is a set of polynomials that generates a nonzero ideal $I = (g_1, ... g_m)$. $G_B = \{g_1, ..., g_n\}$ is the Gröbner basis that will be constructed from $\{g_1, ..., g_m\}$.

Theorem 12.7.1. *A Gröbner basis exists*.

Proof. If J is the ideal generated by all leading terms of $\mathbb{F}$, then by the Hilbert basis theorem, it is finitely generated, so $J = (g_1, ..., g_k)$ for some finite subset of polynomials in $\mathbb{F}$. It is the case that the leading terms of $\{g_1, ..., g_k\}$ also generate J, because J is generated by all leading terms from $\mathbb{F}$.

Conversely, $f = f_1g_1 + ... + f_kg_k$ is in the ideal J. Then the leading term of f is in the ideal generated by leading terms $\{g_1, ..., g_k\}$, so if we subtract the leading term of f, then the remaining terms are also in J, and so on, inductively.

The leading term of f is now divisible by one of the leading terms $\{g_1, ..., g_k\}$, which is the definition of a Gröbner basis. □

Algorithm 12.7.2. *An algorithm to find a Gröbner basis for an ideal* I *of a polynomial ring* $\mathbb{F}[x]$ *can proceed as follows*.

Input A set of polynomials $G = \{g_1, ..., g_m\}$ that generates I.

Output A Gröbner basis G_B for I.

(a) $G := \{g_1, ..., g_m\}$.
(b) For every g_i, g_j in G, denote by a_i the leading term of g_i with respect to the given ordering, and by c_{ij} the least common multiple (l.c.m.) of the monomial part of a_i and a_j.
(c) Choose two polynomials in G and let $S(g_i, g_j) = (c_{ij} / a_i)g_i - (c_{ij} / a_j)g_j$ (*note that the leading terms here will cancel by construction*).
(d) Reduce $S(g_i, g_j)$, with the many variable division algorithm relative to the set G until the result is not further reducible. If the result is non-zero, add it to G.
(e) Repeat steps a-d until all possible pairs are considered, including those involving the new polynomials added in step d.
(f) Output $G_B := G$. □

The polynomial $S(g_i, g_j)$ is usually referred to as the S-polynomial, where Buchberger uses S to refer to *subtraction*. Its essential feature is that for two polynomials g_i and g_j, the combined S-polynomial has a leading term which cancels, leaving the remaining terms to determine whether or not the g_i and g_j form a Gröbner basis. The property of being an S-polynomial determines a necessary condition to create a Gröbner basis. When an S-polynomial is divided by a Gröbner basis, the remainder is 0. As was proved by Buchberger, this is also a sufficient condition.

Example 12.7.3. Consider the polynomials $g_1 = 4x^2y + xy$ and $g_2 = 7x^3 - y^2$ under the lexicographic order $x >_{lex} y$. Then the l.c.m. of the leading ordered monomial terms of g_1 and g_2 is x^3y and

$$S(g_1, g_2) = (x^3y/4x^2y)g_1 - (x^3y/7x^3)g_2 = (x/4y)g_1 - (y/7)g_2$$
$$= x^3y + (x^2y/4) - x^3y + (y^3/7) = (x^2y/4) + (y^3/7). \square$$

Example 12.7.4. Let the polynomials $\{g_1, ..., g_m\}$ all have identical leading monomials $x_a{}^a x_b{}^b ... x_d{}^d$, where there are in number n of $\{a, b, ... d\} \in \mathbb{F}$, and define $f = ag_1 + ... + dg_m$, where the leading polynomial of f in the chosen monomial order is less than $x_a{}^a x_b{}^b ... x_d{}^d$. We will show it is possible to express f as a linear combination of S-polynomials.

We first note that, on putting p_k as the leading coefficient of g_k for each k, that because the monomial of f is less than that of the g_k, its leading coefficient is

$$ap_1 + bp_2 + ... + dp_m = 0.$$

Then, since the leading monomials of the g_k are the same, the S-polynomial is simply

$$S(g_i, g_j) = \frac{1}{p_i} g_i - \frac{1}{p_j} g_j.$$

We can now express f as

$$f = ag_1 + ... + dg_m$$
$$= ap_1 \left(\frac{1}{p_1} g_1 - \frac{1}{p_2} g_2\right) + (ap_1 + bp_2)\left(\frac{1}{p_2} g_2 - \frac{1}{p_3} g_3\right) + \cdots$$
$$+ (ap_1 + \cdots + dp_m)\left(\frac{1}{p_m} g_m\right),$$

and since the coefficient of the last term is zero, we have expressed f as a linear combination of S-polynomials. $\square$

Lemma 12.7.5. *Let*

$$g_1 = jx_a{}^a x_b{}^b ... x_c{}^c + ... \text{ and } g_2 = kx_d{}^d x_e{}^e ... x_f{}^f + ...$$

have leading terms $jx_a{}^a x_b{}^b ... x_c{}^c$ *and* $kx_d{}^d x_e{}^e ... x_f{}^f$, *and say* s = max(a, d) *for the powers* a, d *etc., so the* l.c.m. *of these leading terms is* $x_a{}^s x_b{}^t ... x_c{}^u$. *Then if the remainder is* 0 *when the* S-*polynomial* $S(g_1, g_2)$ *is divided by the polynomials* $g_1, ..., g_m$, *then the remainder is also* 0 *for the* S-*polynomial* $S(x_a{}^{a'} x_b{}^{b'} ... x_c{}^{c'} g_1, x_d{}^{d'} x_e{}^{e'} ... x_f{}^{f'} g_2)$.

Proof.

$$S(x_a{}^{a'} x_b{}^{b'} ... x_c{}^{c'} g_1, x_d{}^{d'} x_e{}^{e'} ... x_f{}^{f'} g_2) = \left(\tfrac{1}{j}\right)(x_a{}^{a'} x_b{}^{b'} ... x_c{}^{c'} g_1) - \left(\tfrac{1}{k}\right)(x_d{}^{d'} x_e{}^{e'} ... x_f{}^{f'} g_2)$$
$$= \frac{x_a^p x_b^q \cdots x_c^r}{jx_a^a x_b^b \cdots x_c^c} g_1 - \frac{x_a^p x_b^q \cdots x_c^r}{kx_d^d x_e^e \cdots x_f^f} g_2$$
$$= \frac{x_a^p x_b^q \cdots x_c^r}{x_a^s x_b^t \cdots x_c^u}\left(\frac{x_a^s x_b^t \cdots x_c^u}{jx_a^a x_b^b \cdots x_c^c} g_1 - \frac{x_a^s x_b^t \cdots x_c^u}{kx_d^d x_e^e \cdots x_f^f} g_2\right)$$
$$= \frac{x_a^p x_b^q \cdots x_c^r}{x_a^s x_b^t \cdots x_c^u} S(g_1, g_2). \square$$

Theorem 12.7.6. *If for every* i, j, $i \neq j$, *whenever the* S-*polynomial* $S(g_i, g_j)$ *is divided by the polynomials* $g_1, \ldots g_m$, *the remainder is* 0, *then* $\{g_1, \ldots, g_m\}$ *is a Gröbner basis for* I.

Proof. Let $f = S(g_i, g_j) \in I$. By the definition of a Gröbner basis we must show the leading term of g_k divides the leading term of f for some k. Then let us represent f by

$$f = h_1g_1 + \ldots + h_mg_m, \quad (1)$$

but we have not defined any bound on the leading terms on the right-hand side of (1). Define the largest such term by

$$X = \max\{\text{leading monomial of } h_kg_k,\ k = 1, \ldots m\}. \quad (2)$$

Now choose the order of the h_kg_k terms in (1) so that the term containing X is minimal under this order. Note that the existence of a minimal term follows from the well-ordering principle of section 2.

The first possibility is that the leading monomial on the right of the equals sign in (1) is also the leading monomial for $f = S(g_i, g_j)$, and this means by definition that $\{g_1, \ldots, g_m\}$ is a Gröbner basis for f, since one of the g_k's divides the leading term of f.

Now assume that this is not the case. We will prove that this is a contradiction by showing that X is not then a minimally ordered term.

Under this scenario, let us collect together the indices of these terms that are not responsible for a Gröbner basis

$$k_X = \{k \in \{1, \ldots, m\}: \text{leading monomials for } h_kg_k = X\},$$

where for $k \in k_X$, X_k is the leading monomial of h_kg_k, and we will write

$$h_k = c_kX_k + h_k',$$

with $c_k \in \mathbb{F}$ and X is greater in value than the leading term of h_k'.

Define

$$g = \sum_{k \in k_X} c_k(X_kg_k),$$

which has the same leading part as for f in equation (1), so this completely cancels out. This means the leading monomial of g is less in value than X, even though the leading monomial of X_kg_k equals X for every k. By example 7.4, g is a linear combination of S-polynomials and by lemma 12.5, division of g by $g_1, \ldots, g_m$ has remainder 0, so every expression on the right hand side of equation (1) has a leading term less in value than X, which contradicts the definition of X in (2), and hence its minimal ordering. □

Example 12.7.7. as an application of the previous result, consider the polynomials x + z and y – z, using the lexicographic order $>_{lex}$, where x > y > z. We will show these polynomials form a Gröbner basis. Since there are only two polynomials in this set, there is only one S-polynomial, which we compute as

$$\begin{aligned} S(x + z,\ y - z) &= \frac{xy}{x}(x + z) - \frac{xy}{y}(y - z) \\ &= xy + yz - xy + xz = xz + yz \\ &= z(x + z) + z(y - z), \end{aligned}$$

which has remainder 0 when divided by (x + z) and (y – z), so these form a Gröbner basis. □

Definition 12.7.8. G_R *is called a reduced Gröbner basis if and only if the leading coefficient of each* g_i *is* 1, *and if no term in any* g_i *is divisible by the leading term of any of the other* g_i's.

Theorem 12.7.9. *The ideal* I *has a unique reduced Gröbner basis.*

Proof. First we will convert the Gröbner basis to a *minimal Gröbner basis*, where all the leading coefficients of the polynomials in the basis are 1, and in no cases does the leading term of one of the polynomials in the basis divide the leading term of any other polynomial in the basis.

To turn a minimal Gröbner basis $\{g_1, ... g_n\}$ to a reduced Gröbner basis, for each g_k in turn, divide it by all other polynomials in the basis. Then replace g_k by the remainder polynomial resulting from this division. Since the remainder has the same leading polynomial as g_k, the new basis will be a Gröbner basis too.

This reduced basis is unique. Let us assume the contrary. Suppose $F_R = \{f_1, ..., f_j\}$ and $G_R = \{g_1, ..., g_k\}$ are two reduced Gröbner bases for I. Firstly we note that the leading terms of the two polynomials are equal, so consider the leading term of $f_1 \in I$. Then the leading term of a polynomial in G_R, g_h, say, divides the leading term of f_1. Similarly, the leading term of a polynomial in F_R divides the leading term of this g_h, and this can only be f_1, since we cannot have another polynomial in F_R dividing the leading term of f_1. Thus the leading terms of f_1 and g_h are equal. By continuing this reasoning for all the polynomials in the bases, we conclude that the leading terms of all polynomials in F_R and G_R are equal, so we deduce their numbers of elements are equal.

To prove that $f_1 = g_h$ for some h, consider the polynomial $f_1 - g_h$. Because the leading terms of f_1 and g_h are equal, they cancel in this polynomial. But if f_1 and g_h differ in any way then they cannot be reduced to all the polynomials in F_R, which contradicts the fact that F_R is a reduced Gröbner basis. Thus $F_R = G_R$. □

12.8. Exercises.

(A) Show the following logical operations are equivalence relations described in chapter III, section 3: AND, OR and ⇔, using ⇔ for IF and only IF (the truth table for which is given in section 10).

(B) For statements A and B, A implies B is written A ⇒ B and is only false when A is true and B is false. Using truth tables or otherwise, show A ⇒ B is the same as (NOT A) OR B, and (A AND B) ⇔ A.

We transform a statement A to a set A′ through the identification of A with $a \in A'$. What are the corresponding Venn diagrams for 'implies' in sets? See the Prologue for a description of Venn diagrams, or look at the internet.

(C) Show ⇒ defines a partial order of definition 12.2.2, (which is repeated in chapter III section 5).

(D) Can you describe 'includes' as a Venn diagram? What is its corresponding mapping back to symbolic logic? Is this a partial order? Is the Venn diagram for (B) different from or just an extension of the Venn diagram for includes?

(E) A is sufficient for B means A implies B. A is necessary for B means A is implied by B (the same as B implies A). What does the internet say about modal logics? Do you think it is a good or a bad idea to have 'necessary' and 'sufficient' as primitive terms?