

## Chapter 2

# Formal Methods for PDE Systems

**Abstract** This chapter discusses formal methods which transform a system of partial differential equations (PDEs) into an equivalent form that allows to determine its power series solutions. Janet's algorithm deals with the case of systems of linear PDEs. The first section presents a generalization of this algorithm to linear functional equations defined over Ore algebras. As a byproduct of a Janet basis computation, a generalized Hilbert series enumerates either a vector space basis for the linear equations that are consequences of the given system or those Taylor coefficients of a power series solution of the PDE system which can be chosen arbitrarily. Systems of polynomially nonlinear PDEs are treated in the second section from the same point of view. A Thomas decomposition of such a system consists of finitely many so-called simple differential systems whose sets of solutions form a partition of the solution set of the given system. Each simple differential system admits a straightforward method of determining its power series solutions. If the given PDE system generates a prime differential ideal, then exactly one of the simple differential systems is the most generic one in a precise sense. Both Janet's and Thomas' algorithm also solve certain elimination problems as described and employed in the following chapter.

### 2.1 Janet's Algorithm

*A system of linear functional equations*

$$Ru = 0 \tag{2.1}$$

for a vector  $u$  of  $p$  unknown functions is given by a matrix  $R \in D^{q \times p}$  with entries in a ring  $D$  of linear operators. We outline an algebraic approach of handling such a linear system. This approach assumes that the set  $\mathcal{F}$  of functions which are candidates for solutions of (2.1) is chosen as a left  $D$ -module, the left action of  $D$  being the one used in (2.1). In particular, the assumption that the result of applying any

operator in  $D$  to any function in  $\mathcal{F}$  is a function in  $\mathcal{F}$  is unavoidable as soon as algebraic manipulations are performed on the equations of system (2.1).

In this section the focus is on the case of linear partial differential equations (PDEs). Then we choose  $D$  as a ring of differential operators with left action on smooth functions by partial differentiation. (We will concentrate on analytic functions.)

Every solution of (2.1) satisfies all consequences of (2.1); we restrict our attention here to consequences which are obtained from (2.1) by multiplying a matrix with  $q$  columns and entries in  $D$  from the left. The condition that a vector of  $p$  functions solves (2.1) can be restated as follows. Let  $(e_1, \dots, e_p)$  be the standard basis of the free left  $D$ -module  $D^{1 \times p}$ . Then every homomorphism  $\phi: D^{1 \times p} \rightarrow \mathcal{F}$  of left  $D$ -modules is uniquely determined by its values  $u_1, \dots, u_p$  for  $e_1, \dots, e_p$ , and every choice of values for  $e_1, \dots, e_p$  defines such a homomorphism. Now,  $(u_1, \dots, u_p)$  solves (2.1) if and only if the corresponding homomorphism  $\phi$  factors over  $D^{1 \times p}/D^{1 \times q}R$ , i.e., is well-defined on residue classes modulo  $D^{1 \times q}R$ . In other words, we have

$$\text{hom}_D(D^{1 \times p}/D^{1 \times q}R, \mathcal{F}) \cong \{u \in \mathcal{F}^{p \times 1} \mid Ru = 0\} \quad (2.2)$$

as abelian groups. (We attribute this remark to B. Malgrange [Mal62, Subsect. 3.2]; it is a basic principle of algebraic analysis, cf., e.g., [Kas03]. For recent work combining algebraic analysis with systems and control theory, cf., e.g., [PQ99], [Pom01], [CQR05], [CQ08], [Qua10b], [Rob14], and the references therein.)

Understanding the structure of the solution set of (2.1) therefore requires at least being able to compute in the residue class module

$$M := D^{1 \times p}/D^{1 \times q}R.$$

Moreover, the left  $D$ -module  $M$  is an intrinsic description of the given system of linear functional equations in the following sense. Let  $Sv = 0$  with  $S \in D^{s \times r}$  be another system of linear functional equations (defined over the same ring) and assume that  $Ru = 0$  and  $Sv = 0$  are equivalent, i.e., there exist  $T \in D^{r \times p}$  and  $U \in D^{p \times r}$  such that the homomorphisms of abelian groups

$$\mathcal{F}^{p \times 1} \longrightarrow \mathcal{F}^{r \times 1}: u \longmapsto Tu, \quad \mathcal{F}^{r \times 1} \longrightarrow \mathcal{F}^{p \times 1}: v \longmapsto Uv$$

induce isomorphisms between  $\{u \in \mathcal{F}^{p \times 1} \mid Ru = 0\}$  and  $\{v \in \mathcal{F}^{r \times 1} \mid Sv = 0\}$  which are inverse to each other. If the set  $\mathcal{F}$  is chosen appropriately (viz. an injective cogenerator for the category of left  $D$ -modules, cf. Remark 3.1.52, p. 154), then this implies that the homomorphisms of left  $D$ -modules

$$D^{1 \times p} \longrightarrow D^{1 \times r}: a \longmapsto aU, \quad D^{1 \times r} \longrightarrow D^{1 \times p}: b \longmapsto bT$$

induce isomorphisms between  $D^{1 \times p}/D^{1 \times q}R$  and  $D^{1 \times r}/D^{1 \times s}S$  which are inverse to each other. Hence, the left  $D$ -modules which are associated with two equivalent systems of linear functional equations are isomorphic.

Computation in  $M = D^{1 \times p} / D^{1 \times q} R$  for certain rings  $D$  of linear operators is made possible by (a generalization of) an algorithm named after the French mathematician Maurice Janet (1888–1983). Having computed a special generating set for the left  $D$ -module  $D^{1 \times q} R$ , called *Janet basis*, a (unique) normal form for the representatives of each residue class in  $M$  is defined and can be computed effectively.

The origin of Janet bases can roughly be described as follows. In work of C. Méray [Mér80] and C. Riquier [Riq10] in the second half of the 19th century the analytic solvability of systems of PDEs was investigated and a generalization of the Cauchy-Kovalevskaya Theorem was obtained. A typical formulation of this classical theorem (cf., e.g., [RR04, Sect. 2.2], [Eva10, Thm. 2 in Subsect. 4.6.3]) assumes a Cauchy problem for unknown functions  $u_1, \dots, u_m$  of  $x_1, \dots, x_n$  in a neighborhood of the origin of the following form. The given PDEs are solved for the partial derivatives of  $u_1, \dots, u_m$  with respect to the first argument  $x_1$ , say, their right hand sides being linear in the other (first order) partial derivatives of  $u_1, \dots, u_m$  with coefficients which are analytic in  $x_2, \dots, x_n$  and  $u_1, \dots, u_m$ , and boundary data for  $u_i(0, x_2, \dots, x_n)$ ,  $i = 1, \dots, m$ , are given by analytic functions. Then this Cauchy problem has a unique analytic solution. Other common formulations of this theorem allow higher order of differentiation (which can be reduced to the above situation by introducing further unknown functions). Analytic coordinate changes may be used to transform boundary data on an analytic hypersurface which is non-characteristic for the first order PDE system to the hypersurface  $x_1 = 0$ .

Riquier's Existence Theorem asserts the existence of analytic solutions to systems of PDEs of a certain class (cf. also [Tho28, Tho34], [Rit34, Chap. IX], [Rit50, Chap. VIII]). The equations are solved for certain distinct partial derivatives and their right hand sides are analytic functions of  $x_1, \dots, x_n$  and of partial derivatives of  $u_1, \dots, u_m$  which are less than the ones on the respective left hand side with respect to some total ordering<sup>1</sup>. Moreover, the system is assumed to incorporate all integrability conditions in some sense (i.e., to be passive as in Definition 2.1.40, p. 28, or Definition 2.2.48, p. 94).

First the existence of formal power series solutions is investigated (formal integrability). Given appropriate boundary conditions, convergence is considered as a second step. Confining ourselves, for the moment, to systems of linear PDEs, the first problem can be solved by transforming any given system into an equivalent one whose formal power series solutions can readily be determined. More precisely, the resulting system allows to partition the set of Taylor coefficients of a power series solution into two sets: coefficients which can be chosen arbitrarily and coefficients which are then uniquely determined by these choices. M. Janet developed an effective procedure which accomplishes such a transformation into a formally integrable system of PDEs (cf. [Jan29, Jan20]; certainly Janet was influenced by work of D. Hilbert [Hil90] as well). The result is now called a Janet basis.

More details on Riquier's Existence Theorem and, in particular, a version for differential regular chains can be found in [PG97, Sect. I.2], [Lem02, Chap. 3]. For

<sup>1</sup> The ordering is assumed to be a Riquier ranking as discussed in Remark 3.1.39, p. 142, and is assumed to respect the differentiation order; a PDE system of this form is called orthonomic.

applications of the theory of Riquier and Janet to the study of Bäcklund transformations, symmetries of differential equations, and related questions, we refer, e.g., to [Sch84], [Sch08a], [RWB96], [MRC98], [Dra01].

Around 1990 the similarity of Janet bases and Gröbner bases became evident to several researchers (cf., e.g., [Wu91], [Pom94, pp. 16–17], [ZB96]). In the case of a commutative polynomial algebra  $D$ , the result of Janet’s algorithm is actually a Gröbner basis for the ideal of  $D$  which is generated by the input. The development of the notion of *involutive division* and *involutive basis* by V. P. Gerdt, Y. A. Blinkov, A. Y. Zharkov, and others (cf. [Ger05] and the references therein) turned Janet’s algorithm into an efficient alternative to Buchberger’s algorithm [Buc06] for computing Gröbner bases, cf. also Remark 2.1.49 below. In fact, the decomposition of multiple-closed sets of monomials into disjoint cones in a computation of an involutive basis (cf. Subsect. 2.1.1) allows to neglect many S-polynomials that are dealt with by Buchberger’s original algorithm (cf. also [Ger05, Sect. 5]).

Janet’s and Buchberger’s algorithms solve the problem of constructing a convergent (i.e., confluent and terminating) rewriting system for the representatives of residue classes of a multivariate polynomial ring modulo an ideal. In other words, given a representative of a residue class, reduction modulo a Janet or Gröbner basis constructs the unique irreducible representative of the same residue class in finitely many steps. In fact, a unification of Buchberger’s algorithm and the Knuth-Bendix completion procedure [KB70] can be achieved (cf. [Buc87, pp. 24–25], [BG94]), e.g., by incorporating constraints for coefficients into term rewriting (the inverse of a non-zero element of the ground field being the solution of an equation). In contrast to the general Knuth-Bendix completion procedure, Janet’s and Buchberger’s algorithms always terminate. For a study of rewriting systems for free associative algebras over commutative rings, we refer to [Ber78].

Generalizations of Gröbner bases to non-commutative algebras have been studied since a couple of decades, cf., e.g., [KRW90], [Kre93], [Mor94], [Lev05], [GL11]; for rings of differential operators, cf., e.g., [CJ84], [Gal85], [IP98], [SST00]. Buchberger’s algorithm was adapted to Ore algebras by F. Chyzak (cf. [Chy98], [CS98], where it is also applied to the study of special functions and combinatorial sequences). Involutive divisions were studied for the Weyl algebra in [HSS02] and were extended to non-commutative rings in [EW07]. However, we follow a more direct approach below, in order to develop Janet’s algorithm for Ore algebras.

The following presentation generalizes earlier descriptions that were given in [PR05, Rob06, Rob07]. In Subsect. 2.1.1 we discuss the combinatorics on which Janet’s algorithm is based. In each non-zero polynomial a unique term is selected as the most significant one in a certain sense, and it is the technique of forming a partition of the set of monomials arising in this way which directs Janet’s algorithm to new polynomials to be included in the resulting Janet basis. The same technique will be used in Sect. 2.2 for the computation of Thomas decompositions of systems of nonlinear partial differential equations and inequations.

After recalling the concept of Ore algebra in Subsect. 2.1.2, Janet’s algorithm is adapted to a certain class of Ore algebras in Subsect. 2.1.3. The relation between

Janet bases and Gröbner bases is described in Subsect. 2.1.4, where we also comment on the complexity of their computation.

Subsection 2.1.5 develops the notion of generalized Hilbert series and applies this combinatorial device to the construction of a Noether normalization of a finitely generated commutative algebra over a field and to the solution of systems of linear partial differential equations. Subsection 2.1.6 summarizes work by the author of this monograph which resulted in implementations of the involutive basis technique in Maple and C++ and refers to related software.

### 2.1.1 Combinatorics of Janet Division

Janet's algorithm constructs a distinguished generating set, called Janet basis, for an ideal of a commutative polynomial algebra, or for a left ideal of a ring of differential operators, or, more generally, for finitely generated left modules over certain Ore algebras. Following Maurice Janet [Jan29], this method examines in a precise sense the highest terms occurring in the generators and their divisibility relations. We therefore restrict our attention in this subsection to the combinatorial properties of certain sets of monomials which are relevant for Janet's algorithm.

Let  $X := \{x_1, \dots, x_n\}$  be a set of  $n$  symbols. For any subset  $Y = \{y_1, \dots, y_r\}$  of  $X$  we denote by

$$\text{Mon}(Y) := \left\{ \prod_{i=1}^r y_i^{\alpha_i} \mid \alpha \in (\mathbb{Z}_{\geq 0})^r \right\}$$

the monoid of monomials in  $y_1, \dots, y_r$ , which is the free commutative semigroup with identity element generated by  $y_1, \dots, y_r$  with the usual divisibility relation  $\mid$ . For  $m = y_1^{\alpha_1} \cdots y_r^{\alpha_r}$  we define  $\deg_{y_i}(m) := \alpha_i$ ,  $i = 1, \dots, r$ . We will often write  $m$  as  $y^\alpha$ , and we denote by  $|\alpha|$  the length  $\alpha_1 + \dots + \alpha_r$  of the multi-index  $\alpha$ .

**Definition 2.1.1.** A set  $S \subseteq \text{Mon}(X)$  is said to be  $\text{Mon}(X)$ -multiple-closed, if

$$ms \in S \quad \text{for all } m \in \text{Mon}(X), \quad s \in S.$$

Every set  $G \subseteq \text{Mon}(X)$  satisfying

$$\text{Mon}(X) \cdot G = \{mg \mid m \in \text{Mon}(X), g \in G\} = S$$

is called a *generating set* for  $S$ .

Janet used the following lemma for his “calcul inverse de la dérivation” (cf. [Jan29]). It can be seen as the special case of Hilbert's Basis Theorem dealing with ideals generated by monomials, which amounts to the statement that every sequence of monomials in which no monomial has a divisor among the previous ones is finite. This combinatorial fact is also referred to as Dickson's Lemma and is proved by induction on  $n$ .

**Lemma 2.1.2.** *Every  $\text{Mon}(X)$ -multiple-closed subset of  $\text{Mon}(X)$  has a finite generating set. Equivalently, every ascending chain of  $\text{Mon}(X)$ -multiple-closed subsets of  $\text{Mon}(X)$  terminates.*

**Remark 2.1.3.** Every  $\text{Mon}(X)$ -multiple-closed set has a unique minimal generating set, which is obtained from any generating set  $G$  by removing all elements which have a proper divisor in  $G$ .

We are going to partition multiple-closed sets (and, more importantly, their complements in  $\text{Mon}(X)$ ) into cones of monomials, one instrumental fact being that the latter are again  $\text{Mon}(\mu)$ -multiple-closed sets for some  $\mu \subseteq X$ .

**Definition 2.1.4.** a) A set  $C \subseteq \text{Mon}(X)$  is called a (*monomial*) *cone* if there exist  $m \in C$  and  $\mu \subseteq \{x_1, \dots, x_n\}$  such that  $\text{Mon}(\mu)m = C$ . The monomial  $m$  is uniquely determined by  $C$  and is called the *generator* of the cone  $C$ , and the elements of  $\mu$  (of  $\bar{\mu} := \{x_1, \dots, x_n\} - \mu$ ) are called the *multiplicative* (resp. *non-multiplicative*) *variables* for  $C$ . Geometrically speaking, the extremal rays of the cone are parallel to the coordinate axes corresponding to multiplicative variables when monomials are visualized as points in the positive orthant (cf. Ex. 2.1.7). We often refer to such a cone  $C$  by the pair  $(m, \mu)$ .

b) Let  $S \subseteq \text{Mon}(X)$  be a set of monomials. A *cone decomposition* of  $S$  is a finite set  $\{(m_1, \mu_1), \dots, (m_r, \mu_r)\}$  of monomial cones such that the sets  $C_i := \text{Mon}(\mu_i)m_i$ ,  $i = 1, \dots, r$ , satisfy  $C_1 \cup \dots \cup C_r = S$  and  $C_i \cap C_j = \emptyset$  for all  $i \neq j$ .

Given a finite set  $M = \{m_1, \dots, m_r\}$  of monomials, there may exist in general no or many ways of arranging sets of multiplicative variables  $\mu_1, \dots, \mu_r$  such that  $\{(m_1, \mu_1), \dots, (m_r, \mu_r)\}$  is a cone decomposition of the  $\text{Mon}(X)$ -multiple-closed set  $S$  generated by  $M$ . After enlarging the set  $M$  by elements of  $S$ , cone decompositions of  $S$  of this form exist. The possible strategies generating such cone decompositions are addressed by the notion of *involutive division*, studied, e.g., by Gerdt, Blinkov [GB98a, GB98b], Apel [Ape98], Seiler [Sei10] and others; cf. [Ger05] for a survey<sup>2</sup>. We restrict our attention to the strategy developed by Janet:

**Definition 2.1.5.** [GB98a] Let  $M \subset \text{Mon}(X)$  be finite. For each  $m \in M$ , the *Janet division* defines the set  $\mu$  of multiplicative variables for the cone with generator  $m$  as follows. Let  $m = x^\alpha = x_1^{\alpha_1} \cdot \dots \cdot x_n^{\alpha_n} \in M$ . For  $1 \leq i \leq n$ , let

$$x_i \in \mu \quad :\Longleftrightarrow \quad \alpha_i = \max \{ \beta_i \mid x^\beta \in M, \beta_j = \alpha_j \text{ for all } j < i \},$$

i.e.,  $x_i$  is a multiplicative variable for the cone with generator  $m$  if and only if its exponent in  $m$  is maximal among the corresponding exponents of all monomials in  $M$  whose sequence of exponents of  $x_1, x_2, \dots, x_{i-1}$  coincides with that of  $m$ .

<sup>2</sup> At the time of this writing, computer experiments have been carried out by Y. A. Blinkov and V. P. Gerdt which indicate that an involutive division which is computationally superior to the one of Janet can be defined by determining the non-multiplicative variables for a generator  $m_i$  as the union of those necessary for separating each two cones  $\text{Mon}(X)m_i, \text{Mon}(X)m_j$ ,  $i \neq j$ , and by deciding the latter using a suitable term ordering on  $\text{Mon}(X)$ , cf. [GB11].

There are also other common involutive divisions. For instance, J. M. Thomas proposed to define  $x_i$  to be a multiplicative variable for the cone with generator  $m$  if and only if  $\alpha_i = \max \{ \beta_i \mid x^\beta \in M \}$  (cf. [Tho37, § 36]). *Pommaret division* defines the set of multiplicative variables for the cone with generator  $m \neq 1$  to be  $\{x_1, \dots, x_k\}$ , where  $k = \min \{ j \mid \alpha_j \neq 0 \}$  is the *class* of  $m$  (cf. [Pom94, p. 90], [Jan29, no. 58]).

We mention that, in the context of combinatorics, cone decompositions as defined above are referred to as Stanley decompositions, cf., e.g., [SW91].

Given a finite generating set for a  $\text{Mon}(X)$ -multiple-closed set  $S$ , the following algorithm constructs a cone decomposition of  $S$  using the strategy proposed by Janet division. A given total ordering  $>$  on  $X$  determines the order in which exponents of monomials are compared. (In Definition 2.1.5 we have  $x_1 > x_2 > \dots > x_n$ .)

**Algorithm 2.1.6** (*Decompose*).

**Input:** A finite subset  $G$  of  $\text{Mon}(X)$ , a subset  $\eta$  of  $X$ , and a total ordering  $>$  on  $X = \{x_1, \dots, x_n\}$

**Output:** A cone decomposition of the  $\text{Mon}(\eta)$ -multiple-closed subset of  $\text{Mon}(X)$  generated by  $G$

**Algorithm:**

```

1: if  $|G| \leq 1$  or  $\eta = \emptyset$  then
2:   return  $\{(g, \eta) \mid g \in G\}$ 
3: else
4:   let  $y$  be the maximal element of  $\eta$  with respect to  $>$ 
5:    $d \leftarrow \max \{ \deg_y(g) \mid g \in G \}$ 
6:   for  $i = 0, \dots, d$  do
7:      $C^{(i)} \leftarrow \text{Decompose}(\bigcup_{j=0}^i \{y^{i-j}g \mid g \in G, \deg_y(g) = j\}, \eta - \{y\}, >)$ 
8:   end for
9:   replace each  $(m, \mu)$  in  $C^{(d)}$  with  $(m, \mu \cup \{y\})$ 
10:  return  $\bigcup_{i=0}^d C^{(i)}$ 
11: end if

```

*Proof.* Termination follows from the fact that the cardinality of  $\eta$  decreases in recursive calls of the algorithm.

We show the correctness by induction on  $|\eta|$ . First of all, a  $\text{Mon}(\eta)$ -multiple-closed set which is generated by a single element or is empty admits a trivial cone decomposition. If  $\eta = \emptyset$ , then each element of  $G$  is the generator of a cone without multiplicative variables. In any other case, the sets of multiples of elements of  $G$  with a fixed degree in the maximal variable  $y$  in  $\eta$  are treated separately. Let us assume that Algorithm 2.1.6 is correct if the input is any  $\text{Mon}(\eta')$ -multiple-closed

subset of  $\text{Mon}(X)$ , where  $\eta' \subset X$  has cardinality less than  $|\eta|$ . Then we assert that the monomial cones in each  $C^{(i)}$ ,  $i = 0, \dots, d$ , in step 10 are mutually disjoint. This assertion holds for  $i = d$  because mutually disjoint cones with generators of the same degree  $d$  in  $y$  for which  $y$  is a non-multiplicative variable are still mutually disjoint after  $y$  has been added as a multiplicative variable (in step 9). By the induction hypothesis, the assertion is true for  $i = 0, \dots, d - 1$ . Since  $y$  is only chosen to be multiplicative for the cones in  $C^{(d)}$ , and since cones in different  $C^{(i)}$  contain only monomials of distinct degrees in  $y$ , it is clear that the cones in  $\bigcup_{i=0}^d C^{(i)}$  are mutually disjoint. Finally, we show that we have

$$\bigcup_{(m, \mu) \in \bigcup_{i=0}^d C^{(i)}} \text{Mon}(\mu)m = \text{Mon}(\eta)G$$

after step 9. The inclusion “ $\subseteq$ ” is obvious. By the induction hypothesis, for each  $i = 0, \dots, d$ , the cones in  $C^{(i)}$  resulting from step 7 form a partition of

$$\text{Mon}(\eta - \{y\}) \cdot \bigcup_{j=0}^i \{y^{i-j}g \mid g \in G, \deg_y(g) = j\}.$$

Every element  $s$  in  $\text{Mon}(\eta)G$  can be written as  $my^k g$  for some  $m \in \text{Mon}(\eta - \{y\})$ ,  $k \in \mathbb{Z}_{\geq 0}$ , and  $g \in G$ . If the degree  $i$  of  $s$  in  $y$  is at most  $d$ , then  $s$  is an element of a unique cone in  $C^{(i)}$ . If  $i$  is greater than  $d$ , then  $s$  is an element of the cone in  $C^{(d)}$  resulting from step 9 which contains  $my^{k-(i-d)}g$ .  $\square$

Algorithm 2.1.6 will be applied both in Subsect. 2.1.3 and Subsect. 2.2.2 for the construction of Janet bases for Ore algebras and of Thomas decompositions of differential systems, respectively. Whenever possible, Algorithm 2.1.6 should be applied to the minimal generating set for the multiple-closed set under consideration (also in recursive calls). This is easily achieved by an additional preliminary step which removes all elements from  $G$  which have a proper divisor in  $G$ . The algorithms discussed in Subsects. 2.1.3 and 2.2.2 produce a more compact result when making use of this modification. It is not incorporated into Algorithm 2.1.6 because for the computation of Janet bases over the ring of integers, the numeric coefficients of highest terms of polynomials must be taken into account, so that an adaptation of (auto-) reduction of a generating set is required (cf. also Def. 2.1.33).

**Example 2.1.7.** Let  $R$  denote the commutative polynomial algebra  $K[x_1, x_2, x_3]$  over a field  $K$  and define  $X := \{x_1, x_2, x_3\}$ . We are going to apply the previous algorithm with the total ordering  $x_1 > x_2 > x_3$ . Let  $\eta = X$  and let  $S \subset \text{Mon}(X)$  be the  $\text{Mon}(X)$ -multiple-closed set generated by  $\{x_1x_2, x_1^3x_3\}$ . Then Algorithm 2.1.6 sets  $d = 3$  and is applied recursively to

$$(\emptyset, \{x_2, x_3\}), \quad (\{x_1x_2\}, \{x_2, x_3\}), \quad (\{x_1^2x_2\}, \{x_2, x_3\}), \quad (\{x_1^3x_2, x_1^3x_3\}, \{x_2, x_3\}),$$

where the first component in each pair is a generating set for a  $\text{Mon}(\{x_2, x_3\})$ -multiple-closed set. Only the last recursive run starts new recursions; the respective



(minimized) arguments are  $(\{x_1^3x_3\}, \{x_3\})$ ,  $(\{x_1^3x_2\}, \{x_3\})$ . The final result is

$$\{(x_1^3x_2, \{x_1, x_2, x_3\}), (x_1^3x_3, \{x_1, x_3\}), (x_1^2x_2, \{x_2, x_3\}), (x_1x_2, \{x_2, x_3\})\}.$$

We also display this decomposition in the following form, where the symbol  $*$  indicates a non-multiplicative variable and does not represent an element of the set of multiplicative variables:

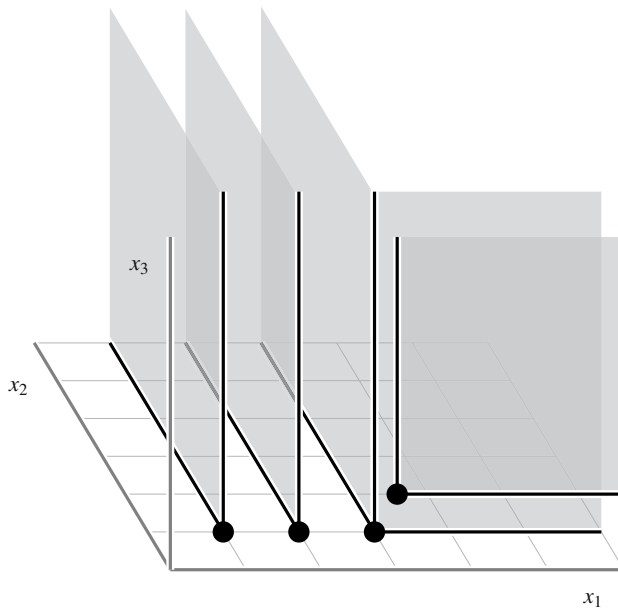
$$x_1^3x_2, \{x_1, x_2, x_3\},$$

$$x_1^3x_3, \{x_1, *, x_3\},$$

$$x_1^2x_2, \{*, x_2, x_3\},$$

$$x_1x_2, \{*, x_2, x_3\}.$$

The cones of this decomposition may also be visualized in the positive orthant of a coordinate system whose axes specify the exponents of  $x_1, x_2, x_3$  in monomials:



**Fig. 2.1** A visualization of the cone decomposition in Example 2.1.7

Next we give a similar algorithm which produces a cone decomposition for the complement of a  $\text{Mon}(X)$ -multiple-closed set  $S$  in  $\text{Mon}(X)$ . Decompositions produced by this algorithm will be used later in the case of the set of leading monomials of a submodule of  $D^{1 \times q}$ , where  $D$  is an Ore algebra, viz. to get a partition of the set of “standard monomials”, and in the case of the set of leaders of a system of polynomial differential equations (cf. also Remarks 2.1.67 and 2.2.79).

**Algorithm 2.1.8** (*DecomposeComplement*).

**Input:** A finite subset  $G$  of  $\text{Mon}(X)$ , a subset  $\eta$  of  $X$ , and  $v \in \text{Mon}(X)$  such that  $G \subseteq \text{Mon}(\eta)v$ , and a total ordering  $>$  on  $X = \{x_1, \dots, x_n\}$

**Output:** A cone decomposition of  $\text{Mon}(\eta)v - S$ , where  $S$  is the  $\text{Mon}(\eta)$ -multiple-closed subset of  $\text{Mon}(X)$  generated by  $G$

**Algorithm:**

```

1: if  $G = \emptyset$  then                                // the complement equals  $\text{Mon}(\eta)v$ , which is a cone
2:   return  $\{(v, \eta)\}$ 
3: else if  $\eta = \emptyset$  then                          // thus,  $G = S = \{v\}$ 
4:   return  $\emptyset$ 
5: else
6:   let  $y$  be the maximal element of  $\eta$  with respect to  $>$ 
7:    $d \leftarrow \max \{\deg_y(g) \mid g \in G\}$ ;  $e \leftarrow \deg_y(v)$ 
8:   for  $i = e, \dots, d$  do
9:      $C^{(i)} \leftarrow \text{DecomposeComplement}(\bigcup_{j=e}^i \{y^{i-j}g \mid g \in G, \deg_y(g) = j\}, \eta - \{y\},$ 
        $y^{i-e}v, >)$ 
10:   end for
11:   replace each  $(m, \mu)$  in  $C^{(d)}$  with  $(m, \mu \cup \{y\})$ 
12:   return  $\bigcup_{i=e}^d C^{(i)}$ 
13: end if

```

*Proof.* It is clear that Algorithm 2.1.8 terminates. If  $G$  is empty, then  $S = \emptyset$ , and  $\{(v, \eta)\}$  is a trivial cone decomposition of  $\text{Mon}(\eta)v$ . Otherwise, if  $\eta$  is empty, then  $S = \{v\}$ , and  $\text{Mon}(\eta)v - S$  is empty. If  $|\eta| = 1$ , then the algorithm enumerates the monomials in  $\text{Mon}(\eta)v - \text{Mon}(\eta)G$ , which are finitely many. These monomials are generators of cones without multiplicative variables. The rest of Algorithm 2.1.8 is similar to Algorithm 2.1.6. The only difference is the additional argument  $v$ , which comprises the information in which set  $\text{Mon}(\eta)v$  the complement is to be taken. The recursive treatment of the sets of multiples of elements of  $G$  with a fixed degree in  $y$  needs to consider only monomials of degree at least  $e = \deg_y(v)$ .  $\square$

**Remark 2.1.9.** The result of applying Algorithm 2.1.8 to a finite generating set  $G$  for a  $\text{Mon}(X)$ -multiple-closed subset  $S$  of  $\text{Mon}(X)$  and  $v = 1$  is a cone decomposition of  $\text{Mon}(X) - S$ . An additional preliminary step removing all elements from  $G$  which have a proper divisor in  $G$  reduces the number of unnecessary recursive calls.

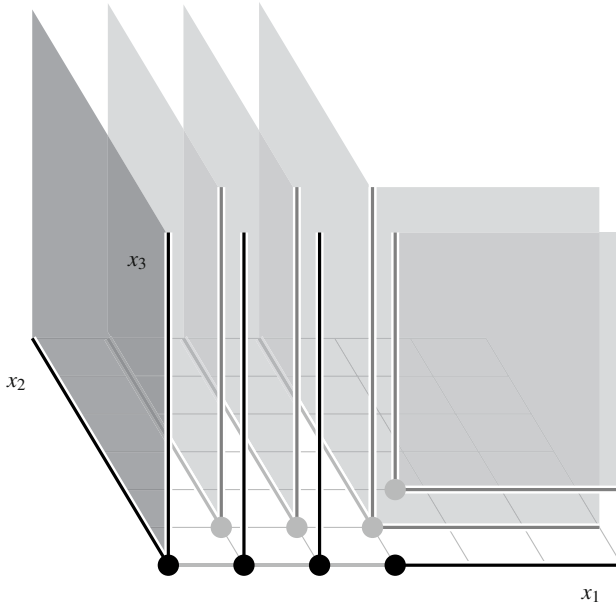
**Example 2.1.10.** Applying Algorithm 2.1.8 to the same data as in Example 2.1.7 and  $v = 1$  leads again to  $d = 3$  and the same recursive calls with additional arguments  $v = 1, x_1, x_1^2$ , and  $x_1^3$ , respectively. After additional recursive runs, the results are

$$\{(1, \{x_2, x_3\})\}, \quad \{(x_1, \{x_3\})\}, \quad \{(x_1^2, \{x_3\})\}, \quad \text{and} \quad \{(x_1^3, \emptyset)\},$$

respectively. The final result is:  $\{(1, \{x_2, x_3\}), (x_1, \{x_3\}), (x_1^2, \{x_3\}), (x_1^3, \{x_1\})\}$ . An alternative representation of the result is the following, where, as in Example 2.1.7, the symbol  $*$  replaces a non-multiplicative variable in the set of all variables and is not to be understood as an element of the set:

$$\begin{aligned} &1, \{*, x_2, x_3\}, \\ &x_1, \{*, *, x_3\}, \\ &x_1^2, \{*, *, x_3\}, \\ &x_1^3, \{x_1, *, *\}. \end{aligned}$$

A visualization of the cone decompositions of both the multiple-closed set  $S$  and its complement in the same orthant is given as follows:



**Fig. 2.2** A visualization of the cone decompositions in Examples 2.1.7 and 2.1.10

**Definition 2.1.11.** Let  $S$  be a  $\text{Mon}(X)$ -multiple-closed subset of  $\text{Mon}(X)$  with finite generating set  $G$ , and let  $>$  be a total ordering on  $X$ . We call the cone decomposition of  $S$  (of  $\text{Mon}(X) - S$ ) which is constructed by Algorithm 2.1.6 (resp. 2.1.8) a *Janet decomposition* of  $S$  (resp. of  $\text{Mon}(X) - S$ ). (If Algorithms 2.1.6 and 2.1.8 reduce  $G$  to the minimal generating set in the beginning, then this notion only depends on  $S$  and  $>$ .) The set of generators of the cones is called the *Janet completion* of  $G$ .

### 2.1.2 Ore Algebras

Ore algebras form a large class of algebras, many instances of which are encountered in applications as algebras of linear operators. The name refers to Ø. Ore, who studied non-commutative rings of polynomials under the assumption that the degree of a product of two non-zero polynomials is the sum of their degrees [Ore33]. Under the same assumption E. Noether and W. Schmeidler proved earlier that one-sided ideals of such rings are finitely generated and investigated the decompositions of such ideals as intersections of irreducible ones [NS20].

For instance, the Weyl algebra  $A_1(\mathbb{R})$  consists of the polynomials in  $\frac{d}{dt}$  whose coefficients are real polynomials in  $t$ , and the structure of  $A_1(\mathbb{R})$  as a (non-commutative) algebra is defined in such a way that its elements represent ordinary differential operators with polynomial coefficients (i.e., the commutation rules in  $A_1(\mathbb{R})$  are determined by the product rule of differentiation; cf. Ex. 2.1.18 a)). Many types of systems of linear equations can be analyzed structurally by viewing them as (left) modules over appropriate Ore algebras. The Ore algebra is chosen to contain all polynomials in the operators occurring in the system equations (cf. also the introduction to this section).

An Ore algebra is obtained as an iterated *Ore extension* of another algebra. An Ore extension forms a skew polynomial ring by adjoining one indeterminate, which does not necessarily commute with the specified algebra of coefficients. After giving the definition of skew polynomial rings and Ore algebras following [CS98], several examples of Ore algebras are discussed. At the end of this subsection important properties of Ore algebras are recalled.

In what follows, let  $K$  be a field (of any characteristic) or  $K = \mathbb{Z}$ , and let  $A$  be a (not necessarily commutative)  $K$ -algebra which is a domain, i.e., an associative and unital<sup>3</sup> algebra over  $K$  without zero divisors.

**Definition 2.1.12** ([MR01], [Coh71]). Let  $\partial$  be an indeterminate,  $\sigma: A \rightarrow A$  a  $K$ -algebra endomorphism and  $\delta: A \rightarrow A$  a  $\sigma$ -derivation, i.e., a  $K$ -linear map which satisfies

$$\delta(ab) = \sigma(a)\delta(b) + \delta(a)b \quad \text{for all } a, b \in A.$$

The *skew polynomial ring*  $A[\partial; \sigma, \delta]$  is the (not necessarily commutative)  $K$ -algebra generated by  $A$  and  $\partial$  obeying the commutation rules

$$\partial a = \sigma(a)\partial + \delta(a) \quad \text{for all } a \in A.$$

( $K$ -linearity of both  $\sigma$  and  $\delta$  implies that  $\partial$  commutes with every element of  $K$ .)

**Remark 2.1.13.** If  $\sigma$  is injective, then  $A[\partial; \sigma, \delta]$  is a domain because the maximum multiplicity of  $\partial$  as a factor in the terms of an element  $p$  of  $A[\partial; \sigma, \delta]$  is then referred

---

<sup>3</sup> All algebra homomorphisms are assumed to map the multiplicative identity element to the multiplicative identity element.

to as the degree of  $p$ , and the degree of a product of two non-zero elements of  $A[\partial; \sigma, \delta]$  equals the sum of their degrees.

We recall the notion of Ore algebra as defined in [Chy98, CS98], which is an iterated skew polynomial ring with commuting indeterminates.

**Definition 2.1.14.** Let  $A$  be a  $K$ -algebra which is a domain and  $\partial_1, \dots, \partial_l$  indeterminates,  $l \in \mathbb{Z}_{\geq 0}$ . The *Ore algebra*  $D = A[\partial_1; \sigma_1, \delta_1][\partial_2; \sigma_2, \delta_2] \dots [\partial_l; \sigma_l, \delta_l]$  is the (not necessarily commutative)  $K$ -algebra generated by  $A$  and  $\partial_1, \dots, \partial_l$  subject to the relations

$$\partial_i d = \sigma_i(d) \partial_i + \delta_i(d), \quad d \in A[\partial_1; \sigma_1, \delta_1] \dots [\partial_{i-1}; \sigma_{i-1}, \delta_{i-1}], \quad i = 1, \dots, l, \quad (2.3)$$

where the map  $\sigma_i$  is a  $K$ -algebra monomorphism of  $A[\partial_1; \sigma_1, \delta_1] \dots [\partial_{i-1}; \sigma_{i-1}, \delta_{i-1}]$  and  $\delta_i$  is a  $\sigma_i$ -derivation of  $A[\partial_1; \sigma_1, \delta_1] \dots [\partial_{i-1}; \sigma_{i-1}, \delta_{i-1}]$  (cf. Def. 2.1.12) satisfying for all  $1 \leq j < i \leq l$

$$\begin{cases} \sigma_i(\partial_j) = \partial_j, \\ \delta_i(\partial_j) = 0 \end{cases} \quad (2.4)$$

and

$$\begin{cases} \sigma_i \circ \sigma_j = \sigma_j \circ \sigma_i, \\ \delta_i \circ \delta_j = \delta_j \circ \delta_i, \\ \sigma_i \circ \delta_j = \delta_j \circ \sigma_i, \\ \sigma_j \circ \delta_i = \delta_i \circ \sigma_j \end{cases} \quad (2.5)$$

as restrictions to  $A[\partial_1; \sigma_1, \delta_1] \dots [\partial_{j-1}; \sigma_{j-1}, \delta_{j-1}]$ . Moreover, we require that (2.3) holds for all  $d \in D$  by extending  $\sigma_i$  and  $\delta_i$  to  $D$  as  $K$ -algebra monomorphism and  $\sigma_i$ -derivation, respectively, subject to  $\sigma_i(\partial_j) = \partial_j$  and  $\delta_i(\partial_j) = 0$  for all  $1 \leq i < j \leq l$ .

**Remark 2.1.15.** Conditions (2.4) imply that the indeterminates  $\partial_i$  and  $\partial_j$  commute in  $D$  for all  $1 \leq i, j \leq l$ , and conditions (2.5) ensure that this postulation is compatible with associativity of the multiplication in  $D$ . Indeed, for all  $1 \leq j < i \leq l$  and all  $d \in A[\partial_1; \sigma_1, \delta_1] \dots [\partial_{j-1}; \sigma_{j-1}, \delta_{j-1}]$  we have

$$\begin{aligned} \partial_i(\partial_j d) &= \partial_i(\sigma_j(d) \partial_j + \delta_j(d)) \\ &= \sigma_i(\sigma_j(d) \partial_j + \delta_j(d)) \partial_i + \delta_i(\sigma_j(d) \partial_j + \delta_j(d)) \\ &= \sigma_i(\sigma_j(d)) \partial_j \partial_i + \sigma_i(\delta_j(d)) \partial_i + \delta_i(\sigma_j(d)) \partial_j + \delta_i(\delta_j(d)) \\ &= \sigma_j(\sigma_i(d)) \partial_i \partial_j + \sigma_j(\delta_i(d)) \partial_j + \delta_j(\sigma_i(d)) \partial_i + \delta_j(\delta_i(d)) \\ &= \sigma_j(\sigma_i(d) \partial_i + \delta_i(d)) \partial_j + \delta_j(\sigma_i(d) \partial_i + \delta_i(d)) \\ &= \partial_j(\sigma_i(d) \partial_i + \delta_i(d)) \\ &= \partial_j(\partial_i d). \end{aligned}$$

Moreover, since all maps  $\sigma_i$  and  $\delta_i$  are  $K$ -linear, each indeterminate  $\partial_i$  commutes with every element of  $K$ . Extending Remark 2.1.13 we note that, since every  $\sigma_i$  is a  $K$ -algebra monomorphism,  $D$  is a domain.

We will concentrate on  $K$ -algebras  $A$  that are either fields (e.g., a field of rational functions over a field  $K$  or a field of meromorphic functions on a connected open subset of  $K^n$ , where  $K = \mathbb{C}$ ) or commutative polynomial algebras over  $K$  (where  $K$  is a field or  $\mathbb{Z}$ ) with finitely many indeterminates. The definition of a monomial in an Ore algebra depends on the type of the  $K$ -algebra  $A$  in this sense.

**Definition 2.1.16.** Let  $D = A[\partial_1; \sigma_1, \delta_1] \dots [\partial_l; \sigma_l, \delta_l]$  be an Ore algebra.

- a) In case  $A = K[z_1, \dots, z_n]$  is a commutative polynomial algebra over a field  $K$  or over  $K = \mathbb{Z}$ , then the set of *indeterminates* of  $D$  is defined by

$$\text{Indet}(D) := \{z_1, \dots, z_n, \partial_1, \dots, \partial_l\}.$$

A *monomial* of  $D$  is then defined to be an element of the form  $z^\alpha \partial^\beta$ , where  $z^\alpha := z_1^{\alpha_1} \dots z_n^{\alpha_n}$ ,  $\partial^\beta := \partial_1^{\beta_1} \dots \partial_l^{\beta_l}$ ,  $\alpha \in (\mathbb{Z}_{\geq 0})^n$ ,  $\beta \in (\mathbb{Z}_{\geq 0})^l$ , and we set

$$\text{Mon}(D) := \{z^\alpha \partial^\beta \mid \alpha \in (\mathbb{Z}_{\geq 0})^n, \beta \in (\mathbb{Z}_{\geq 0})^l\}.$$

- The *total degree* of  $z^\alpha \partial^\beta$  is defined to be  $|\alpha| + |\beta| = \alpha_1 + \dots + \alpha_n + \beta_1 + \dots + \beta_l$ .  
b) If  $A$  is a field, then we define

$$\text{Indet}(D) := \{\partial_1, \dots, \partial_l\}, \quad \text{Mon}(D) := \{\partial^\beta \mid \beta \in (\mathbb{Z}_{\geq 0})^l\},$$

and the total degree of  $\partial^\beta$  is defined to be  $|\beta| = \beta_1 + \dots + \beta_l$ .

We denote the total degree of a monomial  $m \in \text{Mon}(D)$  by  $\deg(m)$ .

For any subset  $Y$  of  $\text{Indet}(D)$ , let  $\text{Mon}(Y)$  be the subset of elements of  $\text{Mon}(D)$  which do not involve any indeterminate in  $\text{Indet}(D) - Y$ .

Let  $q \in \mathbb{N}$  and denote by  $e_1, \dots, e_q$  the standard basis vectors of the free left  $D$ -module  $D^{1 \times q}$ . We set

$$\text{Mon}(D^{1 \times q}) := \bigcup_{k=1}^q \text{Mon}(D)e_k.$$

**Remark 2.1.17.** The definition of the commutation rules of  $D$  implies that  $D^{1 \times q}$  is a free left  $A$ -module with basis

$$\{\partial^\beta e_i \mid \beta \in (\mathbb{Z}_{\geq 0})^l, 1 \leq i \leq q\}. \quad (2.6)$$

Moreover, if  $A = K[z_1, \dots, z_n]$ , then  $\text{Mon}(D^{1 \times q})$  is a basis of  $D^{1 \times q}$  as a free left  $K$ -module. In other words, every  $p \in D^{1 \times q}$  has a unique representation

$$p = \sum_{k=1}^q \sum_{m \in \text{Mon}(D)} c_{k,m} m e_k \quad (2.7)$$

as linear combination of the elements of  $\text{Mon}(D^{1 \times q})$  with coefficients  $c_{k,m} \in K$ , where only finitely many  $c_{k,m}$  are non-zero. In case  $A$  is a field the same holds true with  $c_{k,m} \in A$  (because the basis in (2.6) equals  $\text{Mon}(D^{1 \times q})$ ).

Since  $D$  is a non-commutative ring in general, elements  $p \in D^{1 \times q}$  may have more than one representation as sum of terms with unspecified order of the indeterminates. However, by the previous definition of monomials and the choice to write coefficients in  $A$  on the left, we distinguish a *normal form* (2.7) for the elements of  $D^{1 \times q}$ . For any  $p \in D - \{0\}$ , we define the *total degree* of  $p$  by

$$\deg(p) := \max \{ \deg(m) \mid c_{m,k} \neq 0 \},$$

using the representation (2.7). (Note that, if  $\sigma_1, \dots, \sigma_l$  are injective, then the maximum of the total degrees of monomials with non-zero coefficient is the same for any representation of  $p$  as sum of terms.)

We list important examples of Ore algebras.

**Examples 2.1.18.** a) If  $A = K[z_1, \dots, z_n]$ ,  $\sigma_i = \text{id}_D$  and  $\delta_i = 0$  for all  $i = 1, \dots, l$ , then the Ore algebra  $D = K[z_1, \dots, z_n][\partial_1; \sigma_1, \delta_1] \dots [\partial_l; \sigma_l, \delta_l]$  is the commutative polynomial algebra over  $K$  in  $n + l$  indeterminates.  
b) For  $n \in \mathbb{N}$ , the *Weyl algebra*

$$A_n(K) := K[z_1, \dots, z_n][\partial_1; \sigma_1, \delta_1] \dots [\partial_n; \sigma_n, \delta_n]$$

over  $K$  is defined by

$$\sigma_i = \text{id}_{A_n(K)}, \quad \delta_i = \left( a \mapsto \frac{\partial a}{\partial z_i} \right), \quad i = 1, \dots, n.$$

In  $A_n(K)$  the commutation rules

$$\partial_j z_i = z_i \partial_j + \delta_{i,j}, \quad 1 \leq i, j \leq n,$$

hold, where  $\delta_{i,j}$  is the Kronecker symbol, i.e.,  $\delta_{i,j} = 1$  if  $i = j$  and  $\delta_{i,j} = 0$  otherwise. Let  $K$  be  $\mathbb{R}$  or  $\mathbb{C}$ . We may interpret  $z_1, \dots, z_n$  as coordinates of the smooth manifold  $K^n$ . Then the indeterminate  $z_i$  in  $A_n(K)$  can be understood as a name for the linear operator acting from the left on the  $K$ -vector space of smooth functions on  $K^n$  by multiplication with  $z_i$ , and the indeterminate  $\partial_j$  represents the partial differential operator with respect to  $z_j$ . (The indeterminates  $\partial_i$  and  $\partial_j$  commute, cf. Rem. 2.1.15, which is required by Schwarz' Theorem in this context.)

Another variant of the Weyl algebra is the *algebra of differential operators with rational function coefficients*

$$B_n(K) := K(z_1, \dots, z_n)[\partial_1; \sigma_1, \delta_1] \dots [\partial_n; \sigma_n, \delta_n],$$

where  $\sigma_i$  and  $\delta_i$ ,  $i = 1, \dots, n$ , are defined in the same way as above, but the elements of  $A = K(z_1, \dots, z_n)$  are rational functions in  $z_1, \dots, z_n$ .

c) For  $h \in \mathbb{R}$  let  $S_h := \mathbb{R}[t][\delta_h; \sigma, \delta]$  be the *algebra of shift operators*, where

$$\sigma = (a(t, \delta_h) \mapsto a(t - h, \delta_h)), \quad \delta = (a \mapsto 0), \quad a = a(t, \delta_h) \in S_h.$$

This implies the commutation rule

$$\delta_h t = (t - h) \delta_h$$

in  $S_h$ . Hence,  $\delta_h$  represents the linear operator which shifts the argument of a function of  $t$  by the amount  $h$ .

d) For  $h \in \mathbb{R}$  define  $D_h := \mathbb{R}[t][\partial; \sigma_1, \delta_1][\delta_h; \sigma_2, \delta_2]$ , where  $\sigma_1 = \text{id}_{D_h}$  is the identity map,  $\delta_1$  is defined by formal differentiation with respect to  $t$ , and

$$\sigma_2 = (a(t, \partial, \delta_h) \mapsto a(t - h, \partial, \delta_h)), \quad \delta_2 = (a \mapsto 0), \quad a = a(t, \partial, \delta_h) \in D_h.$$

This algebra consists of linear operators which are relevant for differential time-delay systems.

e) Let  $D = K[z_1, \dots, z_n][\partial_1; \sigma_1, \delta_1] \dots [\partial_n; \sigma_n, \delta_n]$ , where

$$\sigma_i(a) = a(z_1, \dots, z_{i-1}, z_i - 1, z_{i+1}, \dots, z_n, \partial_1, \dots, \partial_n), \quad a \in D,$$

and  $\delta_i = 0$ ,  $i = 1, \dots, n$ . This algebra is used for the algebraic treatment of (multidimensional) discrete systems. Of course, the direction of the shifts can be reversed.

We only recall the essential property of Ore algebras, studied by Ore [Ore33], which ensures the existence of left skew fields of fractions. (All concepts dealing with left multiplication, left ideals, etc., can of course be translated into analogous concepts for right multiplication, right ideals and so on.)

**Definition 2.1.19.** A ring  $D$  is said to satisfy the *left Ore condition* if for all  $a_1, a_2 \in D - \{0\}$  there exist  $b_1, b_2 \in D - \{0\}$  such that  $b_1 a_2 = b_2 a_1$ .

If the left Ore condition is satisfied, then every right-fraction  $a_1 \cdot \frac{1}{a_2}$  has a representation<sup>4</sup> as left-fraction  $\frac{1}{b_2} \cdot b_1$ . Thus, if  $D - \{0\}$  is multiplicatively closed, non-commutative localization with set of denominators  $D - \{0\}$  is made possible.

**Proposition 2.1.20 ([MR01], Cor. 2.1.14).** *Let  $D$  be a domain. A left skew field of fractions of  $D$  exists if and only if  $D$  satisfies the left Ore condition.*

In fact, if we confine ourselves to left Noetherian rings, i.e., rings for which every ascending chain of left ideals terminates, then every domain has this property.

**Proposition 2.1.21 ([MR01], Thm. 2.1.15).** *If  $D$  is a left Noetherian domain, then  $D$  satisfies the left Ore condition.*

<sup>4</sup> If Janet bases can be computed over  $D$ , as explained in the next subsection, then pairs  $(b_1, b_2)$  can be determined effectively as syzygies of  $(a_2, a_1)$ , cf. Subsect. 3.1.5, p. 147.



Moreover, in analogy to Hilbert's Basis Theorem, we have the following important proposition.

**Proposition 2.1.22** ([MR01], Thm. 1.2.9 (iv)). *If  $A$  is a left Noetherian domain and  $\sigma$  is an automorphism of  $A$ , then  $A[\partial; \sigma, \delta]$  is also a left Noetherian domain.*

All Ore algebras in the Examples 2.1.18 are left Noetherian (with bijective twist).

### 2.1.3 Janet Bases for Ore Algebras

In this subsection we present Janet's algorithm for a certain class of Ore algebras  $D$ . Given a submodule  $M$  of the free left  $D$ -module  $D^{1 \times q}$ ,  $q \in \mathbb{N}$ , in terms of a finite generating set, a distinguished generating set for  $M$  is constructed, which, in particular, allows to decide effectively whether a given element of  $D^{1 \times q}$  is in  $M$  and to read off important invariants of  $M$ . In case  $D$  is a commutative polynomial algebra over a field, Janet's algorithm can be viewed as a simultaneous generalization of Euclid's algorithm (dealing with univariate polynomials) and Gaussian elimination (dealing with linear polynomials).

We deal at the same time with both cases of  $D$  being an iterated Ore extension of either a field  $K$  (whose elements do not necessarily commute with every element of  $D$ ) or of a commutative polynomial algebra (over a field or over  $\mathbb{Z}$ ) with finitely many indeterminates (cf. Def. 2.1.16).

In the former case we define

$$D = K[\partial_1; \sigma_1, \delta_1] \dots [\partial_l; \sigma_l, \delta_l],$$

where for some subfield  $K_0$  of  $K$ , each monomorphism  $\sigma_i$  is assumed to be a  $K_0$ -algebra automorphism and each  $\delta_i$  is a  $K_0$ -linear  $\sigma_i$ -derivation as in Definition 2.1.14,  $i = 1, \dots, l$ . We set  $n := 0$  in this case. Note that  $K$  plays the role of the algebra  $A$  in the previous subsection, so that elements of  $K$  do not necessarily commute with the elements  $\partial_1, \dots, \partial_l$  in  $D$ . We also use the notation  $K\langle \partial_1, \dots, \partial_l \rangle$  for such a skew polynomial ring.

In the latter case we define

$$D = K[z_1, \dots, z_n][\partial_1; \sigma_1, \delta_1] \dots [\partial_l; \sigma_l, \delta_l], \quad n \in \mathbb{Z}_{\geq 0},$$

where each monomorphism  $\sigma_i$  is assumed to be a  $K$ -algebra automorphism and each  $\delta_i$  is a  $\sigma_i$ -derivation as in Definition 2.1.14,  $i = 1, \dots, l$ , and where  $K$  is a field (of any characteristic) or  $K = \mathbb{Z}$ , and  $K[z_1, \dots, z_n]$  is the commutative polynomial algebra over  $K$  with standard grading. Moreover, in order to be able to develop Janet's algorithm for Ore algebras employing the notion of multiple-closed sets of monomials (as discussed in Subsect. 2.1.1), we restrict ourselves to the following class of Ore algebras. Let  $K^*$  denote the group of multiplicatively invertible elements of  $K$ . (In case  $n = 0$  the next assumption is vacuous.)

**Assumption 2.1.23.** The automorphisms  $\sigma_1, \dots, \sigma_l$  are of the form

$$\sigma_i(z_j) = c_{ij}z_j + d_{ij}, \quad c_{ij} \in K^*, \quad d_{ij} \in K, \quad j = 1, \dots, n, \quad i = 1, \dots, l,$$

and each  $\sigma_i$ -derivation  $\delta_i$  satisfies

$$\delta_i(z_j) = 0 \quad \text{or} \quad \deg(\delta_i(z_j)) \leq 1, \quad j = 1, \dots, n, \quad i = 1, \dots, l,$$

where  $\deg(\delta_i(z_j))$  denotes the total degree of the polynomial  $\delta_i(z_j) \in K[z_1, \dots, z_n]$ .

By Proposition 2.1.22, in both of the above cases  $D$  is a left Noetherian domain. We assume that the operations in  $D$  which are necessary for executing the algorithms described below can be carried out effectively, e.g., arithmetic in  $D$  and deciding equality of elements in  $D$ .

Let  $q \in \mathbb{N}$ , and denote by  $e_1, \dots, e_q$  the standard basis vectors of the free left  $D$ -module  $D^{1 \times q}$ . Recall from Remark 2.1.17 that every  $p \in D^{1 \times q}$  has a unique representation

$$p = \sum_{k=1}^q \sum_{m \in \text{Mon}(D)} c_{k,m} m e_k \quad (2.8)$$

as linear combination of monomials in  $\text{Mon}(D^{1 \times q})$  with coefficients  $c_{k,m} \in K$ , where only finitely many  $c_{k,m}$  are non-zero.

**Definition 2.1.24.** A *term ordering*  $>$  on  $\text{Mon}(D^{1 \times q})$  (or on  $D^{1 \times q}$ ) is a total ordering on  $\text{Mon}(D^{1 \times q})$  which satisfies the following two conditions.

- a) For all  $1 \leq i \leq n$ ,  $1 \leq j \leq l$ , and  $1 \leq k \leq q$ , we have  $z_i e_k > e_k$  and  $\partial_j e_k > e_k$ .
- b) For all  $m_1 e_k, m_2 e_l \in \text{Mon}(D^{1 \times q})$  the following implications hold:

$$m_1 e_k > m_2 e_l \implies z_i m_1 e_k > z_i m_2 e_l \quad \text{for all } i = 1, \dots, n$$

and

$$m_1 e_k > m_2 e_l \implies m_1 \partial_j e_k > m_2 \partial_j e_l \quad \text{for all } j = 1, \dots, l.$$

If a term ordering  $>$  on  $\text{Mon}(D^{1 \times q})$  is fixed, then for every non-zero  $p \in D^{1 \times q}$  the  $>$ -greatest monomial occurring (with non-zero coefficient) in the representation (2.8) of  $p$  as left  $K$ -linear combination of monomials is uniquely determined and is called the *leading monomial* of  $p$ , denoted by  $\text{lm}(p)$ . The coefficient of  $\text{lm}(p)$  in this representation of  $p$  is called the *leading coefficient* of  $p$ , denoted by  $\text{lc}(p)$ . For any subset  $S \subseteq D^{1 \times q}$  we define

$$\text{lm}(S) := \{ \text{lm}(p) \mid 0 \neq p \in S \}.$$

**Remark 2.1.25.** By Lemma 2.1.2, every term ordering on  $D^{1 \times q}$  is a well-ordering, i.e., every non-empty subset of  $\text{Mon}(D^{1 \times q})$  has a least element. Equivalently, every descending sequence of elements of  $\text{Mon}(D^{1 \times q})$  terminates.

**Example 2.1.26.** Let  $\pi: \{1, \dots, n+l\} \rightarrow \text{Indet}(D)$  be a bijection. The *lexicographical ordering* (*lex*) on  $\text{Mon}(D)$  (which extends the total ordering  $\pi(1) > \pi(2) > \dots > \pi(n+l)$  of the indeterminates) is defined for monomials  $m_1, m_2 \in \text{Mon}(D)$  by

$$m_1 > m_2 \quad :\Longleftrightarrow \quad \begin{cases} m_1 \neq m_2 \quad \text{and} \quad \deg_{\pi(j)}(m_1) > \deg_{\pi(j)}(m_2) \quad \text{for} \\ j = \min \{ 1 \leq i \leq n+l \mid \deg_{\pi(i)}(m_1) \neq \deg_{\pi(i)}(m_2) \}. \end{cases}$$

**Example 2.1.27.** Let  $\pi: \{1, \dots, n+l\} \rightarrow \text{Indet}(D)$  be a bijection. The *degree-reverse lexicographical ordering* (*degrevlex*) on  $\text{Mon}(D)$  (extending the total ordering  $\pi(1) > \pi(2) > \dots > \pi(n+l)$  of the indeterminates) is defined for monomials  $m_1, m_2 \in \text{Mon}(D)$  by

$$m_1 > m_2 \quad :\Longleftrightarrow \quad \begin{cases} \deg(m_1) > \deg(m_2) \quad \text{or} \\ ( \deg(m_1) = \deg(m_2) \quad \text{and} \quad m_1 \neq m_2 \quad \text{and} \\ \deg_{\pi(j)}(m_1) < \deg_{\pi(j)}(m_2) \quad \text{for} \\ j = \max \{ 1 \leq i \leq n+l \mid \deg_{\pi(i)}(m_1) \neq \deg_{\pi(i)}(m_2) \} ). \end{cases}$$

**Example 2.1.28.** Two ways of extending a given term ordering  $>_1$  on  $\text{Mon}(D)$  to  $\text{Mon}(D^{1 \times q})$  for  $q > 1$  are often used. The *term-over-position ordering* (extending  $>_1$  and the total ordering  $e_1 > \dots > e_q$  of the standard basis vectors) is defined for  $m_1, m_2 \in \text{Mon}(D)$  by

$$m_1 e_i > m_2 e_j \quad :\Longleftrightarrow \quad m_1 >_1 m_2 \quad \text{or} \quad (m_1 = m_2 \quad \text{and} \quad i < j).$$

Accordingly, the *position-over-term ordering* (extending  $>_1$  and  $e_1 > \dots > e_q$ ) is defined by

$$m_1 e_i > m_2 e_j \quad :\Longleftrightarrow \quad i < j \quad \text{or} \quad (i = j \quad \text{and} \quad m_1 >_1 m_2).$$

In order to apply Janet's method of partitioning multiple-closed sets of monomials into cones, we make the following assumption. It ensures that left multiplication by  $\partial_j$  has an easily predictable effect on leading monomials, namely multiplication of the leading monomial by  $\partial_j$  yields the leading monomial of the product.

**Assumption 2.1.29.** The term ordering  $>$  on  $\text{Mon}(D^{1 \times q})$  has the property that for all  $i = 1, \dots, n$  and  $j = 1, \dots, l$  such that  $\delta_j(z_i) \neq 0$ , and all  $k = 1, \dots, q$  we have

$$z_i \partial_j e_k > \text{lm}(\delta_j(z_i) e_k)$$

(where  $\text{lm}$  is defined with respect to  $>$ ). We call such a term ordering *admissible*.

**Example 2.1.30.** If  $D$  satisfies Assumption 2.1.23, then every degree-reverse lexicographical ordering  $>$  on  $\text{Mon}(D)$  is admissible. If, in addition,  $\delta_j(z_i)$  is a polynomial in  $K[z_i]$  of total degree at most one for all  $i = 1, \dots, n$ ,  $j = 1, \dots, l$ , then

every lexicographical ordering  $>$  on  $\text{Mon}(D)$  is admissible. (For a common generalization of both types of term orderings, cf. Definition 3.1.4, p. 123.) If  $>$  is an admissible term ordering on  $\text{Mon}(D)$ , then its extensions to a term-over-position or a position-over-term ordering on  $\text{Mon}(D^{1 \times q})$  are admissible.

**Remark 2.1.31.** Let  $D$  be an Ore algebra as above, satisfying Assumption 2.1.23, and let  $>$  be a term ordering on  $D^{1 \times q}$ . Then, for every non-zero  $p \in D^{1 \times q}$ , the monomials which occur with non-zero coefficient in the representation (2.8) of  $p$  form a finite sequence that is sorted with respect to  $>$ . Left multiplication of these monomials by any non-zero element of  $D$  produces a sequence of non-zero elements of  $D^{1 \times q}$ . If the term ordering  $>$  satisfies Assumption 2.1.29, then the sequence that is obtained from the sequence of products by extracting the leading monomial of each element is necessarily sorted with respect to  $>$ . In particular, the leading monomial of every non-zero left multiple of  $p$  can easily be determined as a result of combining Assumptions 2.1.23 and 2.1.29. Moreover, in this situation the combinatorics of Janet division discussed in Subsect. 2.1.1 become applicable as follows.

Let  $X := \{x_1, \dots, x_{n+l}\}$  serve as the set of symbols used in Subsect. 2.1.1 and let

$$\Xi: \text{Mon}(D) \longrightarrow \text{Mon}(X)$$

be any bijection of the set  $\text{Mon}(D)$  onto the monoid  $\text{Mon}(X)$  satisfying that  $\Xi(z^{\alpha_1} \partial^{\beta_1})$  divides  $\Xi(z^{\alpha_2} \partial^{\beta_2})$  if and only if  $\alpha_1$  and  $\beta_1$  are componentwise less than or equal to  $\alpha_2$  and  $\beta_2$ , respectively, where  $\alpha_1, \alpha_2 \in (\mathbb{Z}_{\geq 0})^n$ ,  $\beta_1, \beta_2 \in (\mathbb{Z}_{\geq 0})^l$ . This implies that  $\Xi$  maps  $\text{Indet}(D)$  onto  $X$ .

Suppose that  $L$  is a subset of  $D - \{0\}$ . Let  $S$  be the set of leading monomials of all left multiples of elements of  $L$  by non-zero elements of  $D$ . Then  $\Xi(S)$  is a multiple-closed set of monomials in  $X$ .

In order to apply Algorithms 2.1.6 and 2.1.8, which construct Janet decompositions of multiple-closed sets of monomials in  $X$  and of their complements, respectively, a total ordering on  $X$  is assumed to be chosen (independently of the choice of term ordering on  $D^{1 \times q}$ ).

**Definition 2.1.32.** Let  $\Xi$  be a bijection as defined in the previous remark and let  $S \subseteq \text{Mon}(D^{1 \times q})$ . For  $k \in \{1, \dots, q\}$  we define  $S_k := \{m \in \text{Mon}(D) \mid me_k \in S\}$ .

- a) We call the set  $S$  *multiple-closed* if  $\Xi(S_1), \dots, \Xi(S_q)$  are  $\text{Mon}(X)$ -multiple-closed. A set  $G \subseteq \text{Mon}(D^{1 \times q})$  such that  $\Xi(G_1), \dots, \Xi(G_q)$  are generating sets for  $\Xi(S_1), \dots, \Xi(S_q)$ , respectively, where  $G_k := \{m \in \text{Mon}(D) \mid me_k \in G\}$ , is called a *generating set* for  $S$ . In other words, the multiple-closed set generated by  $G$  is

$$[G] := \bigcup_{k=1}^q \Xi^{-1}(\text{Mon}(X) \cdot \Xi(G_k))e_k.$$

- b) Let  $S$  be multiple-closed. For  $k = 1, \dots, q$ , let

$$\{(m_1^{(k)}, \mu_1^{(k)}), \dots, (m_{t_k}^{(k)}, \mu_{t_k}^{(k)})\}$$

be a Janet decomposition of  $\Xi(S_k)$  (or of  $\text{Mon}(X) - \Xi(S_k)$ ) with respect to the chosen total ordering on  $X$  (cf. Def. 2.1.11). Then

$$\bigcup_{k=1}^q \left\{ \left( \Xi^{-1}(m_1^{(k)})e_k, \Xi^{-1}(\mu_1^{(k)}) \right), \dots, \left( \Xi^{-1}(m_{t_k}^{(k)})e_k, \Xi^{-1}(\mu_{t_k}^{(k)}) \right) \right\}$$

is called a *Janet decomposition* of  $S$  (resp. of  $\text{Mon}(D^{1 \times q}) - S$ ). The *cones* of the Janet decomposition are given by

$$\Xi^{-1}(\text{Mon}(\mu_i^{(k)})m_i^{(k)})e_k, \quad i = 1, \dots, t_k, \quad k = 1, \dots, q.$$

If the Janet decomposition is constructed from the generating set  $G$  for  $S$ , then we call the set of generators  $\Xi^{-1}(m_i^{(k)})e_k$  of the cones the *Janet completion* of  $G$ .

For the rest of this section, let  $D$  be an Ore algebra as described in the beginning of this subsection which satisfies Assumption 2.1.23, and let  $>$  be an admissible term ordering on  $D^{1 \times q}$  (i.e., satisfying Assumption 2.1.29). We fix a bijection  $\Xi: \text{Mon}(D) \rightarrow \text{Mon}(X)$  as above and a total ordering on  $X$  such that the Janet completion of any set  $G \subseteq \text{Mon}(D^{1 \times q})$  is uniquely defined.

Let  $M$  be a submodule of  $D^{1 \times q}$ . Starting with a finite generating set  $L$  of  $M$ , Janet's algorithm possibly removes elements from  $L$  and inserts new elements of  $M$  into  $L$  repeatedly in order to finally achieve that

$$[\text{lm}(L)] = \text{lm}(M).$$

An element  $p \in L$  is removed if it is reduced to zero by subtraction of suitable left multiples of other elements of  $L$ . Before describing the process of auto-reduction we define when a coefficient in  $K$  is reducible modulo another one. This notion depends on whether  $K$  is a field or not.

**Definition 2.1.33.** Let  $a, b \in K$ ,  $b \neq 0$ . If  $K$  is a field, then  $a$  is said to be *reducible modulo  $b$*  if  $a \neq 0$ . If  $K = \mathbb{Z}$ , then  $a$  is said to be *reducible modulo  $b$*  if  $|a| \geq |b|$ . In both cases, if  $a$  is not reducible modulo  $b$ , then the element  $a$  is also said to be *reduced modulo  $b$* .

**Definition 2.1.34.** A subset  $L$  of  $D^{1 \times q}$  is said to be *auto-reduced* if  $0 \notin L$  holds, and for every  $p_1, p_2 \in L$ ,  $p_1 \neq p_2$ , there exists no monomial  $m \in \text{Mon}(D^{1 \times q})$  such that the following two conditions are satisfied.

- a) We have  $\Xi(\text{lm}(p_2)) \mid \Xi(m)$ .
- b) The coefficient  $c$  of  $m$  in the representation of  $p_1$  as left  $K$ -linear combination of monomials is reducible modulo  $\text{lc}(p_2)$ .

Given any finite subset  $L$  of  $D^{1 \times q}$ , there is an obvious way of computing an auto-reduced subset  $L'$  of  $D^{1 \times q}$  which generates the same submodule of  $D^{1 \times q}$  as  $L$ , namely by subtracting suitable left multiples of elements of  $L$  from other elements of  $L$ . We denote by  ${}_D\langle L \rangle$  the submodule of  $D^{1 \times q}$  generated by  $L$ .

**Algorithm 2.1.35** (*Auto-reduce*).**Input:**  $L \subseteq D^{1 \times q}$  finite and an admissible term ordering  $>$  on  $D^{1 \times q}$ **Output:**  $L' \subseteq D^{1 \times q} - \{0\}$  finite such that  $_D\langle L' \rangle = _D\langle L \rangle$  and  $L'$  is auto-reduced**Algorithm:**

- 1:  $L' \leftarrow L - \{0\}$
- 2: **while** there exist  $p_1, p_2 \in L'$ ,  $p_1 \neq p_2$  and  $m \in \text{Mon}(D^{1 \times q})$  occurring with coefficient  $c$  in the representation (2.8) of  $p_1$  such that  $\Xi(\text{lm}(p_2)) \mid \Xi(m)$  and  $c$  is reducible modulo  $\text{lc}(p_2)$  **do**
- 3:    $L' \leftarrow L' - \{p_1\}$
- 4:   subtract a suitable left multiple of  $p_2$  from  $p_1$  such that the coefficient of  $m$  in the representation (2.8) of the result  $r$  is reduced modulo  $\text{lc}(p_2)$
- 5:   **if**  $r \neq 0$  **then**
- 6:      $L' \leftarrow L' \cup \{r\}$
- 7:   **end if**
- 8: **end while**
- 9: **return**  $L'$

**Remark 2.1.36.** Termination and the result of Algorithm 2.1.35 depend on the order in which reductions are performed. Our intention is to construct any auto-reduced set  $L'$  satisfying  $[\text{lm}(L)] \subseteq [\text{lm}(L')]$  (and  $_D\langle L' \rangle = _D\langle L \rangle$ ). By the choice of reductions, the result of Algorithm 2.1.35 is auto-reduced. Since only elements are removed from or replaced in  $L'$  whose leading monomial  $m$  satisfies  $\Xi(\text{lm}(p_2)) \mid \Xi(m)$  for a different element  $p_2 \in L'$ , the property  $[\text{lm}(L)] \subseteq [\text{lm}(L')]$  is ensured as well (cf. also Def. 2.1.32 a)). Clearly, the assertion  $_D\langle L' \rangle = _D\langle L \rangle$  also holds. Moreover, it is easy to see that, if in each round of the loop, the monomial  $m$  in step 2 is chosen as large as possible with respect to  $>$ , then Algorithm 2.1.35 terminates because  $>$  is a well-ordering. In fact, if  $K$  is a field, then step 4 can be understood as replacing the term  $c \cdot m$  of  $p_1$  with a sum of terms whose monomials are smaller than  $m$  with respect to  $>$ . In case  $K = \mathbb{Z}$ , either the same kind of substitution takes place or this substitution also adds a term with monomial  $m$ , whose coefficient, however, is smaller in absolute value than  $c$ ; this can be repeated only finitely many times.

In case  $K = \mathbb{Z}$ , computing the coefficient of  $m$  in  $r$  amounts to Euclidean division for integers. If  $m$  is the leading monomial of  $p_1$ , then it is more efficient in practice to apply the extended Euclidean algorithm to  $\text{lc}(p_1)$  and  $\text{lc}(p_2)$  in order to obtain a representation of the greatest common divisor  $g$  of  $\text{lc}(p_1)$  and  $\text{lc}(p_2)$  as linear combination of these. If  $\text{lc}(p_1)$  is not a multiple of  $\text{lc}(p_2)$ , then the corresponding linear combination  $r$  of  $p_1$  and  $p_2$  is computed such that the leading coefficient of  $r$  equals  $g$ . Then both  $p_1$  and  $r$  are inserted into  $L'$ . In this context,  $p_2$  in step 2 should be chosen with the least possible absolute value of  $\text{lc}(p_2)$  among the candidates with the same leading monomial. In this way, Euclid's algorithm and polynomial division in the sense of Janet are interwoven.

Next we describe a reduction process which takes the Janet division into account. If a divisor of the leading monomial exists in a set defined as follows, then it is uniquely determined due to the disjointness of a cone decomposition.

**Definition 2.1.37.** Let  $T = \{(b_1, \mu_1), (b_2, \mu_2), \dots, (b_t, \mu_t)\}$ , where  $b_i \in D^{1 \times q} - \{0\}$  and  $\mu_i \subseteq \text{Indet}(D)$ ,  $i = 1, \dots, t$ .

- a) The set  $T$  is said to be *Janet complete* if  $\{\text{lm}(b_1), \text{lm}(b_2), \dots, \text{lm}(b_t)\}$  equals its Janet completion<sup>5</sup> and, for each  $i \in \{1, \dots, t\}$ ,  $\mu_i$  is the set of multiplicative variables of the cone with generator  $\text{lm}(b_i)$  in the Janet decomposition  $\{(\text{lm}(b_1), \mu_1), \dots, (\text{lm}(b_t), \mu_t)\}$  of  $[\text{lm}(b_1), \dots, \text{lm}(b_t)]$  (cf. Def. 2.1.32).
- b) An element  $p \in D^{1 \times q}$  is said to be *Janet reducible modulo  $T$*  if there exist  $(b, \mu) \in T$  and a monomial  $m \in \text{Mon}(D^{1 \times q})$  which occurs with coefficient  $c$  in the representation of  $p$  as left  $K$ -linear combination of monomials such that

$$\Xi(m) \in \text{Mon}(\Xi(\mu)) \Xi(\text{lm}(b))$$

and  $c$  is reducible modulo  $\text{lc}(b)$ . In this case,  $(b, \mu)$  is called a *Janet divisor* of  $p$ . Otherwise,  $p$  is also said to be *Janet reduced modulo  $T$* .

The following algorithm subtracts suitable multiples of Janet divisors from a given element  $p \in D^{1 \times q}$  as long as a term in  $p$  is Janet reducible modulo  $T$ .

**Algorithm 2.1.38** (*Janet-reduce*).

**Input:**  $p \in D^{1 \times q}$ ,  $T = \{(b_1, \mu_1), \dots, (b_t, \mu_t)\}$ , and an admissible term ordering  $>$  on  $D^{1 \times q}$ , where  $T$  is Janet complete (with respect to  $>$ , cf. Def. 2.1.37)

**Output:**  $r \in D^{1 \times q}$  such that  $r + {}_D\langle b_1, \dots, b_t \rangle = p + {}_D\langle b_1, \dots, b_t \rangle$  and  $r$  is Janet reduced modulo  $T$

**Algorithm:**

- 1:  $p' \leftarrow p$ ;  $r \leftarrow 0$
- 2: **while**  $p' \neq 0$  **do**
- 3:   **if** there exists a Janet divisor  $(b, \mu)$  of  $\text{lc}(p') \text{lm}(p')$  in  $T$  **then**
- 4:     subtract a suitable left multiple of  $b$  from  $p'$  such that the coefficient of  $\text{lm}(p')$  in the result is reduced modulo  $\text{lc}(b)$ ; replace  $p'$  with this result
- 5:   **else**
- 6:     subtract the term of  $p'$  with monomial  $\text{lm}(p')$  from  $p'$  and add it to  $r$
- 7:   **end if**
- 8: **end while**
- 9: **return**  $r$

---

<sup>5</sup> More generally, a set of monomials is said to be *complete* (with respect to an involutive division), if it consists of the generators of the cones in a cone decomposition of the multiple-closed set they generate, where multiplicative variables for each cone are defined according to the involutive division (cf. Def. 2.1.5 for the case of Janet division). Here we confine ourselves to the complete sets of monomials which are constructed by Algorithm 2.1.6.

- Remarks 2.1.39.** a) Algorithm 2.1.38 terminates because, as long as  $p'$  is non-zero, the leading monomial of  $p'$  decreases with respect to the term ordering  $>$ , which is a well-ordering, or, if  $K = \mathbb{Z}$ , the absolute value of its coefficient decreases. Its correctness is clear. The result  $r$  of Algorithm 2.1.38 is uniquely determined for the given input because every monomial has at most one Janet divisor in  $T$ , and also the course of Algorithm 2.1.38 is uniquely determined as opposed to reduction procedures which apply multivariate polynomial division without distinguishing between multiplicative and non-multiplicative variables.
- b) Let  $p_1, p_2 \in D^{1 \times q}$  and  $T$  be as in the input of Algorithm 2.1.38. In general, the equality  $p_1 + {}_D\langle b_1, \dots, b_t \rangle = p_2 + {}_D\langle b_1, \dots, b_t \rangle$  does not imply that the results of applying *Janet-reduce* to  $p_1$  and  $p_2$ , respectively, are equal. But later on (cf. Thm. 2.1.43 d)) it is shown that, if  $T$  is a Janet basis, then the result of *Janet-reduce* constitutes a unique representative for every coset in  $D^{1 \times q} / {}_D\langle b_1, \dots, b_t \rangle$ . This unique representative of  $p_1 + {}_D\langle b_1, \dots, b_t \rangle$  is called the *Janet normal form* of  $p_1$  modulo  $T$ . For the sake of conciseness, we write  $\text{NF}(p, T, >)$  for *Janet-reduce*( $p, T, >$ ), even if  $T$  is not a Janet basis.

**Definition 2.1.40.** A Janet complete set  $T = \{(b_1, \mu_1), \dots, (b_t, \mu_t)\}$  (as in Definition 2.1.37 a)) is said to be *passive* if

$$\text{NF}(v \cdot b_i, T, >) = 0 \quad \text{for all } v \in \overline{\mu_i}, \quad i = 1, \dots, t \quad (2.9)$$

(where we recall that  $\text{NF}(p, T, >)$  is the result of Algorithm 2.1.38 (*Janet-reduce*) applied to  $p, T, >$ ). If  $T$  is passive, then it is called a *Janet basis* for  ${}_D\langle b_1, \dots, b_t \rangle$ , and  $\{b_1, \dots, b_t\}$  is often referred to as a Janet basis for  ${}_D\langle b_1, \dots, b_t \rangle$  as well.

The term “passive” can be understood as the property of  $T$  which ensures that taking left  $D$ -linear combinations of  $b_1, \dots, b_t$  does not produce any  $p \in D^{1 \times q} - \{0\}$  such that  $\text{lm}(p) \notin [\text{lm}(b_1), \dots, \text{lm}(b_t)]$  (cf. also Remark 2.1.41 below).

More generally, an *involutive basis* is defined by replacing the reference to Janet completeness in the previous definition with a possibly different way of partitioning multiple-closed sets of monomials into cones, as determined by an involutive division (cf. the paragraphs before and after Def. 2.1.5, p. 10). For instance, *Pommaret bases*, i.e., involutive bases with respect to Pommaret division (cf. [Pom94, p. 90], [Jan29, no. 58]) are investigated, e.g., in [Sei10]. Pommaret bases are guaranteed to be finite only in coordinate systems that are sufficiently generic (so-called  $\delta$ -regular coordinates). Essentially the same technique has been applied to a study of homogeneous ideals in a commutative algebra context by Mutsumi Amasaki in [Ama90] (where this form of Gröbner basis in generic coordinates is referred to as a *system of Weierstraß polynomials*).

**Remark 2.1.41.** Let  $M$  be a submodule of  $D^{1 \times q}$ . Then  $\text{lm}(M)$  is multiple-closed (cf. Rem. 2.1.31 and Def. 2.1.32). Janet’s algorithm (cf. Alg. 2.1.42 below) constructs an ascending chain of multiple-closed subsets of  $\text{lm}(M)$ , which terminates by Lemma 2.1.2. In each round, a cone decomposition is computed for the current



multiple-closed set  $S$  generated by the leading monomials of an auto-reduced generating set  $G$  for  $M$ . Note that, if  $K$  is a field, these leading monomials form not just any generating set, but the minimal generating set for  $S$ .

The Janet decomposition is constructed by applying Algorithm 2.1.6, p. 11, directly to  $G$ , in the sense that its run is determined by the monomials  $\Xi(\text{lm}(g))$ ,  $g \in G$ , but left multiplications of such a monomial by  $y$  are replaced with left multiplications of  $g$  by  $\Xi^{-1}(y)$ . Accordingly, the result  $J = \{(b_1, \mu_1), \dots, (b_t, \mu_t)\}$  consists of pairs of a non-zero element  $b_i$  of  $D^{1 \times q}$  and a subset  $\mu_i$  of  $\text{Indet}(D)$ . In the following algorithm, this adapted version of Algorithm 2.1.6 (*Decompose*) will be applied (using the given total ordering on  $X$ ).

Since  $\{b_1, \dots, b_t\}$  is a generating set for  $M$ , every element of  $M$  is a left  $D$ -linear combination of  $b_1, \dots, b_t$ . We assume that  $J$  is passive. Let  $k_i m_i b_i$  be a summand in such a linear combination, where  $k_i \in K$  and  $m_i \in \text{Mon}(D)$ . If  $m_i$  involves some variable which is non-multiplicative for  $b_i$ , then this summand can be replaced with a left  $K$ -linear combination of elements in  $\text{Mon}(\mu_1)b_1, \dots, \text{Mon}(\mu_t)b_t$ . Using (2.9), this can be achieved by applying successively Algorithm 2.1.38 to terms involving only one non-multiplicative variable. This substitution process should deal with the largest term with respect to  $>$  first. Elimination of all non-multiplicative variables demonstrates that the leading monomial of every element of  $M - \{0\}$  has a Janet divisor in  $J$ . We conclude that passivity of the Janet complete set  $J$  is equivalent to

$$[\text{lm}(b_1), \dots, \text{lm}(b_t)] = \text{lm}(M).$$

Now Janet's algorithm is presented, which computes a Janet basis for a submodule of  $D^{1 \times q}$ , given in terms of a finite generating set. Note that we ignore efficiency issues in favor of a concise formulation of the algorithm (cf. also Subsect. 2.1.6).

For any set  $S$  we denote by  $\mathcal{P}(S)$  the power set of  $S$ .

**Algorithm 2.1.42** (*JanetBasis*).

**Input:** A finite set  $L \subseteq D^{1 \times q}$ , an admissible term ordering  $>$  on  $D^{1 \times q}$ , and a total ordering on  $\Xi(\text{Indet}(D)) = X$  (used by *Decompose*)

**Output:** A finite subset  $J$  of  $D^{1 \times q} \times \mathcal{P}(\text{Indet}(D))$  which is a Janet basis for the left  $D$ -module  ${}_D\langle p \mid (p, \mu) \in J \rangle = {}_D\langle L \rangle$  (and  $J = \emptyset$  if and only if  ${}_D\langle L \rangle = \{0\}$ )

**Algorithm:**

- 1:  $G \leftarrow L$
- 2: **repeat**
- 3:    $G \leftarrow \text{Auto-reduce}(G, >)$  // cf. Alg. 2.1.35
- 4:    $J \leftarrow \text{Decompose}(G)$  // cf. Rem. 2.1.41
- 5:    $P \leftarrow \{\text{NF}(v \cdot p, J, >) \mid (p, \mu) \in J, v \in \overline{\mu}\}$  // cf. Alg. 2.1.38
- 6:    $G \leftarrow \{p \mid (p, \mu) \in J\} \cup P$
- 7: **until**  $P \subseteq \{0\}$
- 8: **return**  $J$

**Theorem 2.1.43.** a) *Algorithm 2.1.42 terminates and is correct.*

b) A  $K$ -basis of  ${}_D\langle L \rangle$  is given by  $\bigsqcup_{(p,\mu) \in J} \text{Mon}(\mu)p$ , where  $J$  is the result of Algorithm 2.1.42. In particular, every  $r \in {}_D\langle L \rangle$  has a unique representation

$$r = \sum_{(p,\mu) \in J} c_{(p,\mu)} p,$$

where each  $c_{(p,\mu)} \in D$  is a left  $K$ -linear combination of elements in  $\text{Mon}(\mu)$ .

c) The cosets in  $D^{1 \times q} / {}_D\langle L \rangle$  with representatives in

$$\text{Mon}(D^{1 \times q}) - [\text{lm}(p) \mid (p, \mu) \in J, 1 \text{ is reducible modulo } \text{lc}(p)]$$

form a generating set for the left  $K$ -module  $D^{1 \times q} / {}_D\langle L \rangle$ , and the cosets with representatives in  $C := \text{Mon}(D^{1 \times q}) - [\text{lm}(p) \mid (p, \mu) \in J]$  form the unique maximal  $K$ -linearly independent subset.

Let  $C_1, \dots, C_k$  be the cones of a Janet decomposition of  $C$  (cf. Fig. 2.2, p. 15, for an illustration). If  $K$  is a field, then the cosets with representatives in  $C_1 \sqcup \dots \sqcup C_k$  form a basis for the left  $K$ -vector space  $D^{1 \times q} / {}_D\langle L \rangle$ .

d) For every  $r_1, r_2 \in D^{1 \times q}$  the following equivalence holds.

$$r_1 + {}_D\langle L \rangle = r_2 + {}_D\langle L \rangle \iff \text{NF}(r_1, J, >) = \text{NF}(r_2, J, >).$$

*Proof.* a) First we show that *JanetBasis* terminates. For the result  $G$  of *Auto-reduce* in step 3,  $[\text{lm}(G)]$  contains the multiple-closed set generated by the leading monomials of the previous set  $G$ . *Decompose* only augments the generating set  $G$  by elements  $p \in D^{1 \times q}$  satisfying  $\text{lm}(p) \in [\text{lm}(G)]$  if it is necessary for the chosen strategy of decomposing  $[\text{lm}(G)]$  into disjoint cones. In any case it ensures  $[\text{lm}(p) \mid (p, \mu) \in J] = [\text{lm}(G)]$ . If all Janet normal forms in step 5 are zero, then the algorithm terminates. If  $P \not\subseteq \{0\}$ , then  $G' := G \cup P$  satisfies  $[\text{lm}(G)] \subsetneq [\text{lm}(G')]$ . By Lemma 2.1.2, after finitely many steps we have  $[\text{lm}(G)] = [\text{lm}(G')]$ , which is equivalent to  $P \subseteq \{0\}$  (cf. Rem. 2.1.41). Therefore, *JanetBasis* terminates in any case.

In order to prove correctness of *JanetBasis*, we note that the result  $J$  of step 4 is Janet complete. Therefore  $\text{NF}(v \cdot p, J, >)$  in step 5 is well-defined. Once  $P \subseteq \{0\}$  holds in step 7, the set  $J$  is passive, thus a Janet basis. The equality of left  $D$ -modules  ${}_D\langle p \mid (p, \mu) \in J \rangle = {}_D\langle L \rangle$  is an invariant of the loop.

b) Set  $B := \bigcup_{(p,\mu) \in J} \text{Mon}(\mu)p$ .

For the  $K$ -linear independence of  $B$  we note first that  $0 \notin B$  holds because  $J$  is constructed as Janet completion of an auto-reduced set. Furthermore, we have  $\text{lm}(p_1) \neq \text{lm}(p_2)$  for all  $p_1, p_2 \in B$  with  $p_1 \neq p_2$  because  $J$  is Janet complete, which proves that  $B$  is  $K$ -linearly independent.

We are going to show that  $B$  is a generating set for the free left  $K$ -module  ${}_D\langle L \rangle$ . Let  $0 \neq r \in {}_D\langle L \rangle$ . Then  $r$  is Janet reducible modulo  $J$  by Remark 2.1.41. Now,  $\text{NF}(r, J, >) \in {}_D\langle L \rangle$ , and  $\text{NF}(r, J, >)$  is not Janet reducible modulo  $J$ . The previous argument implies  $\text{NF}(r, J, >) = 0$ . Hence, we have  $r \in {}_K\langle B \rangle$ .

- c) Let  $0 \neq r + {}_D\langle L \rangle \in D^{1 \times q} / {}_D\langle L \rangle$ . Then  $\text{NF}(r, J, >)$  is a representative of the residue class  $r + {}_D\langle L \rangle$  as well, and  $\text{NF}(r, J, >) \neq 0$  because otherwise  $r \in {}_D\langle L \rangle$ . Janet reduction (Alg. 2.1.38) ensures that if a term  $c \cdot m$  in  $\text{NF}(r, J, >)$ , where  $c \in K$  and  $m \in \text{Mon}(D^{1 \times q})$ , has a Janet divisor  $(p, \mu)$  in  $J$ , then  $c$  is reduced modulo  $\text{lc}(p)$ . Therefore, the cosets represented by those monomials  $\text{lm}(p)$ ,  $(p, \mu) \in J$ , for which 1 is reducible modulo  $\text{lc}(p)$ , are not needed to generate  $D^{1 \times q} / {}_D\langle L \rangle$  as a left  $K$ -module.
- Due to the equality  $[\text{lm}(p) \mid (p, \mu) \in J] = \text{lm}({}_D\langle L \rangle)$  (cf. Rem. 2.1.41) we have  $C = \text{Mon}(D^{1 \times q}) - \text{lm}({}_D\langle L \rangle)$ . The cosets with representatives in  $C$  are  $K$ -linearly independent because no  $K$ -linear combination of these has a non-zero representative with leading monomial in  $\text{lm}({}_D\langle L \rangle)$ . The rest is clear.
- d) It remains to show that the normal form  $\text{NF}(r, J, >)$  is uniquely determined by the coset  $r + {}_D\langle L \rangle \in D^{1 \times q} / {}_D\langle L \rangle$ . But if  $n_1, n_2 \in D^{1 \times q}$  are Janet normal forms of the same coset  $r + {}_D\langle L \rangle$ , then  $n_1 - n_2 \in {}_D\langle L \rangle$ , and  $n_1 - n_2$  is Janet reduced modulo  $J$  because  $n_1$  and  $n_2$  are so. The same argument as in the last part of b) shows that  $n_1 - n_2 = 0$ .  $\square$

We present a couple of examples demonstrating Janet's algorithm.

**Example 2.1.44.** Let  $D = K[x, y]$  be the commutative polynomial algebra over a field  $K$  of arbitrary characteristic or over  $K = \mathbb{Z}$ . We choose the degree-reverse lexicographical ordering on  $\text{Mon}(D)$  satisfying  $x > y$  (cf. Ex. 2.1.27). Let the ideal  $I$  of  $D$  be generated by

$$g_1 := \underline{x^2} - y, \quad g_2 := \underline{xy} - y.$$

Then the method of Subsect. 2.1.1 (using the total ordering on  $\{x, y\}$  for which  $x$  is greater than  $y$ ) constructs the following cone decomposition of the multiple-closed set which is generated by the (underlined) leading monomials of  $g_1$  and  $g_2$ :

$$\{(x^2, \{x, y\}), (xy, \{y\})\}.$$

This result indicates that we need to check whether  $f := x \cdot g_2$  can be written as

$$f = c_1 \cdot (x^2 - y) + c_2 \cdot (xy - y), \quad c_1 \in K[x, y], \quad c_2 \in K[y]. \quad (2.10)$$

The monomials appearing in  $f = x^2 y - xy \in I$  lie in the cones  $(x^2, \{x, y\})$  and  $(xy, \{y\})$ , respectively. Reduction yields  $g_3 := y^2 - y \in I$ , which does not have a representation as in (2.10). So, we include  $g_3$  in our list of generators, and for this example, we already arrive at the (minimal) Janet basis

$$\{(g_1, \{x, y\}), (g_2, \{y\}), (g_3, \{y\})\}$$

for  $I$ .

No division by any coefficient was necessary to arrive at a Janet basis for  $I$ . The statements above therefore hold for a field  $K$  of any characteristic and for  $K = \mathbb{Z}$ . In Example 2.1.47, the relevance of Janet bases with integer coefficients for constructing matrix representations of finitely presented groups is demonstrated.

**Example 2.1.45.** Let us consider the system

$$\frac{\partial u}{\partial x} = x \frac{\partial u}{\partial y}, \quad u(x-1, y) = u(x, y) \quad (2.11)$$

of one linear partial differential equation and one linear delay equation for one unknown smooth function  $u$  of two independent variables  $x$  and  $y$ . According to the types of functional operators occurring in the system (2.11) we define the Ore algebra<sup>6</sup>  $D = \mathbb{Q}[x, y][\partial_x; \text{id}, \delta_1][\partial_y; \text{id}, \delta_2][\delta_x; \sigma, \delta_3]$ , where the derivations  $\delta_1$  and  $\delta_2$  are defined by partial differentiation with respect to  $x$  and  $y$ , respectively, where  $\sigma$  is the  $\mathbb{Q}$ -algebra automorphism of  $D$  defined by

$$a(x, y, \partial_x, \partial_y, \delta_x) \mapsto a(x-1, y, \partial_x, \partial_y, \delta_x),$$

and where  $\delta_3$  is the zero map. By writing the equations in (2.11) as

$$(\partial_x - x \partial_y) u = 0, \quad (\delta_x - 1) u = 0,$$

we are led to study the left ideal  $I$  of  $D$  which is generated by

$$\{\partial_x - x \partial_y, \delta_x - 1\}.$$

Janet's algorithm can be applied for determining all linear consequences of (2.11). We choose the degree-reverse lexicographical ordering (cf. Ex. 2.1.27) on  $\text{Mon}(D)$  satisfying  $\partial_x > \partial_y > \delta_x > x > y$ . The multiple-closed set which is generated by the leading monomials  $\partial_x$  and  $\delta_x$  is partitioned into cones. Using the total ordering on  $\{\partial_x, \partial_y, \delta_x, x, y\}$  which is defined by  $\partial_x > \partial_y > \delta_x > x > y$ , all indeterminates are assigned as multiplicative variables to the first generator  $g_1$ , whereas  $\partial_x$  is a non-multiplicative variable for the second generator  $g_2$ . Janet reduction of  $\partial_x(\delta_x - 1)$  yields

$$g_3 := \partial_x(\delta_x - 1) - (\delta_x - 1)(\partial_x - x \partial_y) - (x-1)\partial_y(\delta_x - 1) = -\partial_y.$$

After adding  $-g_3$  to the generating set and updating the Janet decomposition, Janet reduction replaces  $g_1$  with  $g_1 - x g_3 = \partial_x$ . It can be easily checked that the resulting Janet complete set is passive. Therefore, the (minimal) Janet basis is given by

$$\frac{\partial u}{\partial x} = 0, \quad \{\partial_x, \partial_y, \delta_x, x, y\},$$

$$\frac{\partial u}{\partial y} = 0, \quad \{*, \partial_y, \delta_x, x, y\},$$

$$u(x-1, y) - u(x, y) = 0, \quad \{*, *, \delta_x, x, y\}.$$

The system (2.11) only admits constant solutions.

<sup>6</sup> The computations performed in this example do not change if we replace  $\mathbb{Q}[x, y]$  with its field of fractions  $\mathbb{Q}(x, y)$  in the definition of  $D$ .

The following example applies Janet's algorithm to a system of linear partial differential equations which arises from linearization of a nonlinear PDE system. Linearization is a common simplification technique for studying differential equations. Being an approximation, the linearized system reflects only a few properties of the original system in general. It can be understood as the first order term of the Taylor expansion of the nonlinear system around a chosen solution (e.g., a critical point for ordinary differential equations) using the Fréchet derivative (cf., e.g., [Olv93, Sect. 5.2], [Rob06, Sect. 3.2]). By computing these derivatives symbolically, we do not need any particular solution of the nonlinear system, but work with a symbol which is subject to the nonlinear equations. We refer to the resulting linear system as the *general linearization*. (Alternatively, the linearization of algebraic differential equations can also be expressed in terms of Kähler differentials, cf. also Subsect. 3.3.3.)

In the given example all real analytic solutions of the nonlinear PDE system are available explicitly, which is obviously a very special case, but which allows a comparison of the solutions of the linearized system and those of the original one. (For notation that concerns notions of differential algebra, we refer to Sect. A.3. More details on the notion of general linearization can be found in [Rob06, Sect. 3.2].)

**Example 2.1.46.** [Rob06, Ex. 3.3.9] We consider the system of nonlinear PDEs

$$\frac{\partial u}{\partial x} - u^2 = 0, \quad \frac{\partial^2 u}{\partial y^2} - u^3 = 0 \quad (2.12)$$

for one unknown real analytic function  $u$  of two independent variables  $x$  and  $y$ . Note that

$$u(x, y) = \frac{2}{-2x \pm \sqrt{2y} + c}, \quad c \in \mathbb{R}, \quad (2.13)$$

are explicit solutions of (2.12). We are going to apply Janet's algorithm to the general linearization

$$\frac{\partial U}{\partial x} - 2uU = 0, \quad \frac{\partial^2 U}{\partial y^2} - 3u^2U = 0 \quad (2.14)$$

of (2.12), which is a system of linear PDEs for an unknown real analytic function  $U$  of  $x$  and  $y$ , and where the function  $u$  is subject to (2.12). Since Janet's algorithm has to decide whether coefficients of polynomials are equal to zero or not, it is required to bring the nonlinear system (2.12) into a form that allows effective computation with  $u$ . Applying the techniques to be discussed in Sect. 2.2 to (2.12) yields a Thomas decomposition of that system (where subscripts indicate differentiation):

$\begin{aligned} \underline{u_x} - u^2 &= 0 \{ \partial_x, \partial_y \} \\ 2\underline{u_y^2} - u^4 &= 0 \{ *, \partial_y \} \\ u &\neq 0 \end{aligned}$	$u = 0 \{ \partial_x, \partial_y \}$
---	--------------------------------------

We are interested in the first case, solve the second equation for  $u_y$ , and use

$$u_x = u^2, \quad u_y = \pm \frac{\sqrt{2}}{2} u^2 \quad (2.15)$$

as rewriting rules for the coefficients in Janet's algorithm. For the second rule a choice of sign should be made and used consistently in what follows. Let  $\mathbb{Q}(\sqrt{2})\{u\}$  be the differential polynomial ring in one differential indeterminate  $u$  with coefficients in  $\mathbb{Q}(\sqrt{2})$  and commuting derivations  $\delta_x, \delta_y$  (cf. Sect. A.3). Moreover, let  $I$  be the differential ideal of  $\mathbb{Q}(\sqrt{2})\{u\}$  which is generated by  $u_x - u^2$  and  $u_y \mp \frac{\sqrt{2}}{2} u^2$ . Then  $\mathbb{Q}(\sqrt{2})\{u\}/I$  is a domain because it is isomorphic to  $\mathbb{Q}(\sqrt{2})[u]$ . We denote by  $K$  the field of fractions of  $\mathbb{Q}(\sqrt{2})\{u\}/I$ , which is a differential field with derivations extending  $\delta_x$  and  $\delta_y$  (using the quotient rule). Now the left hand sides of the input (2.14) for Janet's algorithm are to be understood as elements of the skew polynomial ring  $K\langle \partial_x, \partial_y \rangle = K[\partial_x; \text{id}, \delta_1][\partial_y; \text{id}, \delta_2]$ , where the derivations  $\delta_1$  and  $\delta_2$  are defined as the extensions of  $\delta_x$  and  $\delta_y$  to  $K\langle \partial_x, \partial_y \rangle$  satisfying  $\delta_1(\partial_x) = \delta_1(\partial_y) = 0$  and  $\delta_2(\partial_x) = \delta_2(\partial_y) = 0$ , respectively (cf. also Def. 2.1.14).

In this example the passivity check only involves the partial derivative with respect to  $x$  of the second equation in (2.14), whose normal form is computed by subtracting the second partial derivative with respect to  $y$  of the first equation. After simplification using the rewriting rules (2.15) we obtain

$$\pm 2\sqrt{2}u^2 U_y - 4u^3 U = 0,$$

and since  $u \neq 0$ , a Janet basis for the linearized system is

$$\begin{aligned} U_x - 2uU &= 0, \{ \partial_x, \partial_y \}, \\ U_y \mp \sqrt{2}uU &= 0, \{ *, \partial_y \}. \end{aligned}$$

Substituting (2.13) for  $u$  in this Janet basis results in a system of linear PDEs for  $U$  whose analytic solutions are given by

$$U(x, y) = \frac{C}{(-2x \pm \sqrt{2}y + c)^2}, \quad C \in \mathbb{R}. \quad (2.16)$$

If we consider the map (between Banach spaces) which associates with  $\varepsilon$  in a small real interval containing 0 the explicit solution in (2.13) with constant  $c + \varepsilon$ , then the solution (2.16) for a certain value of  $C$  coincides, as expected, with the coefficient of  $\varepsilon$  in the Taylor expansion of this (sufficiently differentiable) map around 0. (We refer to [Rob06, Sect. 3.2] for more details.)

For applications of Janet bases with integer coefficients, e.g., for constructing matrix representations of finitely presented groups (without specifying the characteristic of the field in advance), for a constructive version of the Quillen-Suslin Theorem, and for primary decomposition, we refer to [PR06], [Fab09], [FQ07], [Jam11]. The following example is an application of the first kind.

**Example 2.1.47.** [Rob07, Ex. 5.1] We would like to construct matrix representations of degree 3 over various fields  $K$  of the finitely presented group

$$G_{2,3,13;4} := \langle a, b \mid a^2, b^3, (ab)^{13}, [a, b]^4 \rangle,$$

where  $[a, b] := aba^{-1}b^{-1}$ . To this end, we write the images of (the residue classes of)  $a$  and  $b$  under such a representation as  $3 \times 3$  matrices  $A$  and  $B$  with indeterminate entries. The relators  $a^2, b^3, (ab)^{13}, [a, b]^4$  of the above presentation are translated into relations for commutative polynomials obtained from the entries of the matrix equations

$$A^2 = I_3, \quad B^3 = I_3, \quad (AB)^{13} = I_3, \quad [A, B]^4 = I_3, \quad (2.17)$$

where  $I_3$  is the identity matrix in  $\text{GL}(3, K)$ . (We refer to [PR06] for more details on this approach.) We may choose a  $K$ -basis  $(v_1, v_2, v_3)$  of  $K^{3 \times 1}$  with respect to which the  $K$ -linear action on  $K^{3 \times 1}$  of (the residue classes of)  $a$  and  $b$  in  $G_{2,3,13;4}$  is represented by  $A$  and  $B$ , respectively. We let  $v_1$  be an eigenvector of  $A \cdot B$  with eigenvalue  $\lambda$ , possibly in an algebraic extension field of  $K$ , and let  $v_2 := Bv_1$  and  $v_3 := Bv_2$ . We confine ourselves to finding irreducible representations, which implies that  $(v_1, v_2, v_3)$  is  $K$ -linearly independent. By using  $(v_1, v_2, v_3)$  as a basis for  $K^{3 \times 1}$ , we may assume without loss of generality that  $A$  and  $B$  have the form

$$A := \begin{pmatrix} 0 & c_2 & c_3 \\ c_1 & 0 & c_4 \\ 0 & 0 & c_5 \end{pmatrix}, \quad B := \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix},$$

where  $c_1 = \lambda^{-1}$ ,  $c_2 = \lambda$ , because  $ABv_1 = \lambda v_1$  implies  $v_2 = Bv_1 = A^2Bv_1 = \lambda Av_1$  and we have  $B^3v_1 = v_1$  due to the given relations (2.17). Moreover, we derive from (2.17) a system of algebraic equations for  $c_1, \dots, c_5$ . We compute

$$A^2 = \begin{pmatrix} c_1 c_2 & 0 & c_2 c_4 + c_3 c_5 \\ 0 & c_1 c_2 & c_1 c_3 + c_4 c_5 \\ 0 & 0 & c_5^2 \end{pmatrix}.$$

The determinant of  $A$  equals  $-c_1 c_2 c_5$ . Now, by (2.17) we have  $\det(A^2) = 1$ ,  $\det(B) = 1$ , and  $\det((AB)^{13}) = 1$ , which implies  $\det(A) = 1$ . Due to  $c_1 c_2 = 1$  we have  $c_5 = -1$ . Hence, we substitute  $-1$  for  $c_5$  in  $A$  and we are left with four unknowns. We define  $L$  to be the set of all entries of the matrices  $A^2 - I_3$ ,  $(AB)^{13} - I_3$ ,  $(ABAB^2)^2 - (BAB^2A)^2$ . Thus  $L$  consists of polynomials in  $c_1, c_2, c_3, c_4$  with integer coefficients. We compute a Janet basis for the ideal of  $\mathbb{Q}[c_1, c_2, c_3, c_4]$  which is generated by  $L$ . The result consists of the polynomial 1 only. This shows that the above system of algebraic equations for  $c_1, c_2, c_3, c_4$  has no solution in  $\mathbb{C}^4$ , hence there exists no irreducible matrix representation  $G_{2,3,13;4} \rightarrow \text{GL}(3, \mathbb{C})$ .

Next we check whether there are such matrix representations of  $G_{2,3,13;4}$  in positive characteristic. To this end, we compute a Janet basis  $J$  with respect to the degree-reverse lexicographical ordering extending  $c_1 > c_2 > c_3 > c_4$  (and using the same total ordering of variables for determining Janet decompositions) for the ideal

of  $\mathbb{Z}[c_1, c_2, c_3, c_4]$  which is generated by  $L$ :

$$\begin{array}{ll}
 15, & \{ *, *, *, * \}, \\
 15c_4, & \{ *, *, *, * \}, \\
 15c_3, & \{ *, *, *, * \}, \\
 15c_2, & \{ *, *, *, * \}, \\
 c_1 + 4c_2 + c_3 + 4c_4, & \{ c_1, c_2, c_3, c_4 \}, \\
 15c_4^2, & \{ *, *, *, * \}, \\
 15c_3c_4, & \{ *, *, *, * \}, \\
 c_2c_4 - c_3, & \{ *, *, *, c_4 \}, \\
 c_3^2 + 4c_3c_4 + c_4^2 + c_3 + c_4 + 4, & \{ *, *, c_3, c_4 \}, \\
 c_2c_3 - 4c_4^2 - 4c_3 - 1, & \{ *, *, c_3, c_4 \}, \\
 c_2^2 + c_4^2 + 2c_3 - 7, & \{ *, c_2, c_3, c_4 \}, \\
 c_4^3 + 2c_3c_4 + 4c_4^2 + 4c_3 - 7c_4 + 1, & \{ *, *, *, c_4 \}, \\
 c_3c_4^2 - 4c_3c_4 - c_4^2 + c_2 + 7c_3 - 2c_4 - 4, & \{ *, *, *, c_4 \}.
 \end{array}$$

We find that solutions of the system of algebraic equations exist only if  $15 = 0$ , i.e., possibly in characteristic 3 or 5. We are going to check both possibilities. It turns out that replacing each coefficient of the above polynomials with its remainder modulo 3 (resp. 5) yields (after removing zero polynomials) the minimal Janet basis for the algebraic system over  $\mathbb{Z}/3\mathbb{Z}$  (resp.  $\mathbb{Z}/5\mathbb{Z}$ ) with the same multiplicative variables. A Janet decomposition of the complement in  $\text{Mon}(\{c_1, c_2, c_3, c_4\})$  of the multiple-closed set generated by the leading monomials is given by

$$\{(1, \emptyset), (c_4, \emptyset), (c_3, \emptyset), (c_2, \emptyset), (c_4^2, \emptyset), (c_3c_4, \emptyset)\}.$$

Denoting by  $F$  either  $\mathbb{Z}/3\mathbb{Z}$  or  $\mathbb{Z}/5\mathbb{Z}$ , and by  $I$  the ideal of  $F[c_1, c_2, c_3, c_4]$  which is generated by  $L$  (modulo 3 resp. 5), we conclude that  $R := F[c_1, c_2, c_3, c_4]/I$  is 6-dimensional as an  $F$ -vector space. By the Chinese Remainder Theorem, the residue class ring of  $R$  modulo its radical is isomorphic to a direct sum of at most six fields, which define at most six solutions to the above algebraic system over an algebraic closure of  $F$ , the bound being attained precisely if  $R$  has no nilpotent elements. For the present example we obtain quickly the Janet basis

$$\begin{array}{ll}
 c_4^6 + c_4^4 + c_4 + 1, & \{ *, *, *, c_4 \}, \\
 c_3 + 2c_4^5 + 2c_4^4 + c_4^3 + 2c_4 + 2, & \{ *, *, c_3, c_4 \}, \\
 c_2 + c_4^5 + 2c_4^4 + c_4^2, & \{ *, c_2, c_3, c_4 \}, \\
 c_1 + 2c_4^4 + 2c_4^3 + 2c_4^2 + 2c_4 + 1, & \{ c_1, c_2, c_3, c_4 \}
 \end{array}$$

for  $I$  with respect to the lexicographical ordering extending  $c_1 > c_2 > c_3 > c_4$ , which allows to solve for  $c_1, c_2, c_3$  in terms of  $c_4$ . In  $(\mathbb{Z}/3\mathbb{Z})[c_4]$  we have

$$c_4^6 + c_4^4 + c_4 + 1 = (c_4^3 + c_4^2 + 2)(c_4^3 + 2c_4^2 + 2c_4 + 2),$$

the factors on the right hand side being irreducible. Hence, we have found matrix representations of  $G_{2,3,13;4}$  of degree 3 over the fields  $(\mathbb{Z}/3\mathbb{Z})[\xi]/(\xi^3 + \xi^2 + 2)$  and



$(\mathbb{Z}/3\mathbb{Z})[\xi]/(\xi^3 + 2\xi^2 + 2\xi + 2)$ . For instance, in the first case we obtain

$$A = \begin{pmatrix} 0 & 2\xi + 1 & 2\xi^2 + \xi \\ \xi^2 + 2\xi + 2 & 0 & \xi \\ 0 & 0 & 2 \end{pmatrix}.$$

For  $F = \mathbb{Z}/5\mathbb{Z}$  an analogous computation yields irreducible matrix representations over the fields  $(\mathbb{Z}/5\mathbb{Z})[\zeta]/(\zeta^2 + 2\zeta + 4)$  and  $(\mathbb{Z}/5\mathbb{Z})[\zeta]/(\zeta^4 + 3\zeta^3 + \zeta^2 + 2\zeta + 4)$ ; e.g., we may choose  $A$  as

$$A = \begin{pmatrix} 0 & 4\zeta^3 + \zeta^2 + 3 & 4\zeta^3 + \zeta^2 + 4 \\ \zeta + 4 & 0 & \zeta \\ 0 & 0 & 4 \end{pmatrix}$$

in the second case.

### 2.1.4 Comparison and Complexity

In this subsection we comment on the relationship between Janet bases and Gröbner bases and on complexity results. For surveys on the latter topic, cf., e.g., [May97], [vzGG03, Sect. 21.7].

We use the same notation as in the previous subsection.

**Definition 2.1.48.** Let  $M$  be a submodule of the free left  $D$ -module  $D^{1 \times q}$ . A finite subset  $G \subseteq M - \{0\}$  is said to be a *Gröbner basis* for  $M$  (with respect to the term ordering  $>$  on  $D^{1 \times q}$ ) if the leading monomial of every non-zero element of  $M$  is the leading monomial of a left multiple of some element of  $G$ .

**Remark 2.1.49.** If  $J = \{(p_1, \mu_1), \dots, (p_t, \mu_t)\}$  is a Janet basis for the submodule  $M = {}_D\langle p_1, \dots, p_t \rangle$  of  $D^{1 \times q}$ , then the multiple-closed set  $[\text{lm}(p_1), \dots, \text{lm}(p_t)]$  equals  $\text{lm}(M)$  (cf. also Rem. 2.1.41). More generally, this equality is used as a criterion for the termination of algorithms constructing involutive bases, cf., e.g., [Ger05], and is also well-known from Buchberger's algorithm computing Gröbner bases (cf., e.g., his PhD thesis of 1965, [Buc06]). In fact, for this reason, every involutive basis is also a Gröbner basis, whenever both notions exist in the same context, but the former reflects a lot more combinatorial information about the ideal or module (cf. Subsect. 2.1.5). More precisely, in general the reduced Gröbner basis of a module (cf., e.g., [CLO07, § 2.7]) is a proper subset of a Janet basis for the same module (and with respect to the same term ordering). For another comparison of Janet and Gröbner bases, cf. also [CJMF03].

Janet's algorithm can be understood as a refinement of the original version of Buchberger's algorithm (cf., e.g., [CLO07, § 2.7], [Eis95, Sect. 15.4], [vzGG03, Sect. 21.5]). For simplicity we assume that the module is an ideal of  $\mathbb{Q}[x_1, \dots, x_n]$ . Given a finite generating set  $L$ , Buchberger's algorithm forms the *S-polynomial*

$$S(p_1, p_2) := \frac{\text{lcm}(\text{lm}(p_1), \text{lm}(p_2))}{\text{lc}(p_1) \text{lm}(p_1)} p_1 - \frac{\text{lcm}(\text{lm}(p_1), \text{lm}(p_2))}{\text{lc}(p_2) \text{lm}(p_2)} p_2$$

for each unordered pair of (non-zero) generators  $p_1, p_2$  in  $L$  and reduces it modulo  $L$  using multivariate polynomial division. Non-zero remainders are added to  $L$ , and this process is repeated until every S-polynomial reduces to zero. Janet's algorithm decomposes the multiple-closed set generated by the leading monomials of the generators in  $L$  into disjoint cones as described in Subsect. 2.1.1 and considers the S-polynomials which are determined by the non-multiplicative variables  $v$  of generators  $p$  and the (uniquely determined) Janet divisor of  $v \cdot p$  in the current generating set (if any). This strategy avoids many S-polynomials which are examined by Buchberger's original algorithm (cf. also [Ger05, Sect. 5]). However, Buchberger's algorithm was enhanced as well by incorporating criteria which allow to neglect certain S-polynomials (cf., e.g., [Buc79], [CLO07, § 2.9]).

Algorithm 2.1.42 constructs a Janet basis which is minimal with respect to inclusion for the fixed total ordering on  $\mathcal{E}(\text{Indet}(D)) = X$ . This Janet basis  $J$  is uniquely determined under the assumption that no term in any of its elements is Janet reducible modulo  $J$  and that, if  $K$  is a field, the coefficient of each leading monomial equals one, say. In case  $K = \mathbb{Z}$ , a choice for the systems of residues modulo integers, e.g., the symmetric one, should be fixed to ensure uniqueness.

Redundancy of a Janet basis (compared with the reduced Gröbner basis) is diminished by the concept of *Janet-like Gröbner basis* [GB05a, GB05b]. For each generator the partition of the set of indeterminates into sets of multiplicative and non-multiplicative variables is replaced with a map of this set into  $\mathbb{Z}_{\geq 0} \cup \{\infty\}$  indicating the multiplicative degree for each indeterminate. If the image of each of these maps is required to be a subset of  $\{0, \infty\}$ , then the Janet division is recovered as a special case. The number of left multiples of generators by non-multiplicative variables to be included for completion is often reduced when applying Janet-like division.

Let  $D = \mathbb{Q}[x_1, \dots, x_n]$  be a commutative polynomial algebra in  $n$  variables.

The complexity of the problem to decide whether a given polynomial is an element of an ideal of  $D$  (the latter being given in terms of a finite generating set) was studied by G. Hermann [Her26]. Her result states the following.

**Theorem 2.1.50.** *Let  $G \subset D - \{0\}$  be a finite generating set of cardinality  $m$  for an ideal  $I$  of  $D$ , and let  $p \in D$ . Let  $d$  be the maximum total degree of the elements of  $G$ . If  $p$  is an element of  $I$ , then  $p$  is a linear combination of the generators in  $G$  with coefficients that are either zero or polynomials of total degree at most*

$$\deg(p) + (md)^{2^n}.$$

The following upper bound on the degrees of the elements of a reduced Gröbner basis is proved, e.g., in [Dub90], where techniques of partitioning sets of monomials similar to Subsect. 2.1.1 are used.

**Theorem 2.1.51.** *Let  $G \subset D - \{0\}$  be a finite generating set for an ideal  $I$  of  $D$ . Let  $d$  be the maximum total degree of the elements of  $G$ . Then the total degree of the elements of the reduced Gröbner basis for  $I$  with respect to any term ordering on  $\text{Mon}(D)$  is bounded by*

$$2 \left( \frac{d^2}{2} + d \right)^{2^{n-1}}.$$

Better bounds for special cases are also known. For instance, if  $n = 3$ , then the total degree of the polynomials constructed by Buchberger's algorithm computing a Gröbner basis for  $I$  (including the elements of  $G$ ) is bounded by  $(8d + 1) \cdot 2^\delta$ , where  $\delta$  is the minimum total degree of the elements of  $G$  [Win84].

A corresponding doubly exponential degree bound for Janet bases over the Weyl algebras was obtained in [GC08] by reducing the problem to solving linear systems over a variant of the Weyl algebra whose commutation rules have been made homogeneous by introducing an additional commuting variable (cf. also [Gri91], [Gri96]).

E. W. Mayr and A. R. Meyer constructed a family of ideals (generated by binomials) for which the doubly exponential upper bound is attained [MM82]. Further work by E. W. Mayr showed that the computation of a Gröbner basis for a general polynomial ideal is an EXPSPACE-complete problem.

**Remark 2.1.52.** In practice a behavior much better than the worst case has been observed for algorithms computing Gröbner or Janet bases when applied to, e.g., problems arising in algebraic geometry or systems of linear partial equations with origin in physics or the engineering sciences. In the algebraic geometry context a growth measure was introduced for the degrees of (iterated) syzygies (cf. Subsect. 3.1.5) for a given ideal  $I$  of a commutative polynomial algebra, which reflects the difficulty of computing Gröbner or Janet bases for  $I$ . This measure, called Castelnuovo-Mumford regularity, denoted by  $\text{reg}(I)$ , is significant not only from the computational, but also from the geometric and algebraic point of view, cf., e.g., [Eis95, Eis05].

The regularity of the ideal generated by  $\text{lm}(I)$  is an upper bound for the regularity of  $I$ . In generic coordinates, the maximum total degree of the elements of the reduced Gröbner basis for  $I$  equals  $\text{reg}(\langle \text{lm}(I) \rangle)$ . If leading monomials are determined with respect to the degree-reverse lexicographical ordering, then we have, in generic coordinates,  $\text{reg}(\langle \text{lm}(I) \rangle) = \text{reg}(I)$ , cf. [BS87]. Therefore, this term ordering is preferably used.

For lack of space, we do not discuss here recent approaches by J.-C. Faugère (cf., e.g., [Fau99]) to compute Gröbner bases using linear algebra techniques, which result in very efficient programs and applications to cryptography.

### 2.1.5 The Generalized Hilbert Series

In this subsection we extend the notion of generalized Hilbert series (cf. [PR05], [Rob06]) to Ore algebras and present applications of this combinatorial invariant.

Throughout this subsection, let  $K$  be a field.

Let  $D$  be an Ore algebra as described in the beginning of Subsect. 2.1.3 which satisfies Assumption 2.1.23 (p. 22). Moreover, let  $q \in \mathbb{N}$  and  $>$  be an admissible term ordering on  $\text{Mon}(D^{1 \times q})$  (cf. Assumption 2.1.29, p. 23). For combinatorial purposes we introduce a totally ordered set

$$X := \{x_1, \dots, x_{n+l}\}$$

of indeterminates and we choose a bijection  $\Xi: \text{Mon}(D) \rightarrow \text{Mon}(X)$  as in Remark 2.1.31 (p. 24), where it was used to apply the combinatorics of Janet division to a set of monomials of the Ore algebra  $D$ .

**Definition 2.1.53.** For any subset  $S$  of  $\text{Mon}(D^{1 \times q})$ , the *generalized Hilbert series* of  $S$  is defined by

$$H_S(x_1, \dots, x_{n+l}) := \sum_{s \in S} \Xi(s) f_k \in \bigoplus_{k=1}^q \mathbb{Z}[[x_1, \dots, x_{n+l}]] f_k,$$

where the symbols  $f_1, \dots, f_q$  form a basis of a free left  $\mathbb{Z}[[x_1, \dots, x_{n+l}]]$ -module of rank  $q$ . For  $k = 1, \dots, q$ , we define  $H_{S,k}(x_1, \dots, x_{n+l})$  by

$$H_S(x_1, \dots, x_{n+l}) = \sum_{k=1}^q H_{S,k}(x_1, \dots, x_{n+l}) f_k,$$

and we identify  $H_S(x_1, \dots, x_{n+l})$  with  $H_{S,1}(x_1, \dots, x_{n+l})$  in case  $q = 1$ .

**Remark 2.1.54.** Let  $M$  be a submodule of  $D^{1 \times q}$  and let  $J$  be a Janet basis for  $M$  with respect to some term ordering on  $D^{1 \times q}$ . We denote by  $S$  the multiple-closed set generated by  $\{\text{lm}(p) \mid (p, \mu) \in J\}$  (cf. Def. 2.1.32).

- a) According to Theorem 2.1.43 b), the (disjoint) union of the cones  $\text{Mon}(\mu)p$ ,  $(p, \mu) \in J$ , is a  $K$ -basis of  $M$ . Thus, the generalized Hilbert series  $H_S(x_1, \dots, x_{n+l})$  enumerates a  $K$ -basis of  $M$ , in the sense that its terms enumerate the leading monomials of the above  $K$ -basis via  $\Xi^{-1}$ .
- b) Similarly, by Theorem 2.1.43 c), a  $K$ -basis of the factor module  $D^{1 \times q}/M$  is given by the cosets in  $D^{1 \times q}/M$  which are represented by the monomials in  $C_1 \uplus \dots \uplus C_k$ , where  $C_1, \dots, C_k$  are the cones of a Janet decomposition of  $\bar{S} := \text{Mon}(D^{1 \times q}) - S$ . Therefore, the generalized Hilbert series  $H_{\bar{S}}(x_1, \dots, x_{n+l})$  enumerates a  $K$ -basis of  $D^{1 \times q}/M$  via  $\Xi^{-1}$ .

When the Janet basis  $J$  for  $M$  is clear from the context, we also call  $H_S$  and  $H_{\bar{S}}$  the *generalized Hilbert series* of  $M$  and  $D^{1 \times q}/M$ , respectively.

The next remark shows that the generalized Hilbert series of a set  $S$  of monomials has a succinct representation in finite terms if a cone decomposition of  $S$  is available.

**Remark 2.1.55.** Let  $(C, \mu)$  be a monomial cone, i.e.,  $C \subseteq \text{Mon}(X)$ ,  $\mu \subseteq X$ , and

$$S := C = \text{Mon}(\mu) \cdot v$$

for some  $v \in C$ . We use the (formal) geometric series

$$\frac{1}{1-x} = \sum_{i \in \mathbb{Z}_{\geq 0}} x^i$$

to write down the generalized Hilbert series  $H_S(x_1, \dots, x_{n+l})$  as follows:

$$H_S(x_1, \dots, x_{n+l}) = \frac{v}{\prod_{x \in \mu} (1-x)}.$$

More generally, every cone decomposition of a multiple-closed set  $S$  allows to compute the generalized Hilbert series of  $S$  by adding the generalized Hilbert series of the cones. In an analogous way this remark applies to the complements of multiple-closed sets.

**Example 2.1.56.** Let  $R$  be the commutative polynomial algebra  $K[x_1, x_2, x_3]$  over any field  $K$  and  $S$  the multiple-closed set generated by  $\{x_1x_2, x_1^3x_3\}$ . The Janet decomposition of  $S$  which is computed in Example 2.1.7 yields the generalized Hilbert series

$$\begin{aligned} H_S(x_1, x_2, x_3) &= \frac{x_1^3x_2}{(1-x_1)(1-x_2)(1-x_3)} + \frac{x_1^3x_3}{(1-x_1)(1-x_3)} \\ &\quad + \frac{x_1^2x_2}{(1-x_2)(1-x_3)} + \frac{x_1x_2}{(1-x_2)(1-x_3)}. \end{aligned}$$

The Janet decomposition of the complement  $\bar{S} = \text{Mon}(\{x_1, x_2, x_3\}) - S$  of  $S$  obtained in Example 2.1.10 yields

$$H_{\bar{S}}(x_1, x_2, x_3) = \frac{1}{(1-x_2)(1-x_3)} + \frac{x_1}{1-x_3} + \frac{x_1^2}{1-x_3} + \frac{x_1^3}{1-x_1}.$$

Note that the sum of the two Hilbert series equals  $1/((1-x_1)(1-x_2)(1-x_3))$ , i.e.,

$$H_S + H_{\bar{S}} = H_{\text{Mon}(\{x_1, x_2, x_3\})}.$$

Next we describe the relationship between the generalized Hilbert series and the Hilbert series of filtered and graded modules.

**Definition 2.1.57.** Let  $A$  be a (not necessarily commutative)  $K$ -algebra and assume that  $F = (F_i)_{i \in \mathbb{Z}_{\geq 0}}$  is an (exhaustive) increasing filtration of  $A$  (cf., e.g., [Bou98b]),

i.e., each  $F_i$  is a (left)  $K$ -subspace of  $A$  and we have  $1 \in F_0$ ,

$$\bigcup_{i \in \mathbb{Z}_{\geq 0}} F_i = A, \quad \text{and} \quad F_i \subseteq F_{i+1}, \quad F_i \cdot F_j \subseteq F_{i+j} \quad \text{for all } i, j \in \mathbb{Z}_{\geq 0}.$$

Moreover, let  $M$  be a left  $A$ -module endowed with an (exhaustive) increasing  $F$ -filtration  $\Phi = (\Phi_i)_{i \in \mathbb{Z}}$ , i.e., each  $\Phi_i$  is a (left)  $K$ -subspace of  $M$  such that

$$\bigcup_{i \in \mathbb{Z}} \Phi_i = M \quad \text{and} \quad \Phi_i \subseteq \Phi_{i+1}, \quad F_i \cdot \Phi_j \subseteq \Phi_{i+j} \quad \text{for all } i \in \mathbb{Z}_{\geq 0}, \quad j \in \mathbb{Z}.$$

We assume that  $M$  is finitely generated and that each  $\Phi_i$  has finite  $K$ -dimension. Then the *Hilbert series of  $M$  with respect to  $\Phi$*  is defined by the (formal) Laurent series

$$H_{M, \Phi}(\lambda) := \sum_{i \in \mathbb{Z}} (\dim_K \Phi_i) \lambda^i \in \mathbb{Z}((\lambda)).$$

The map

$$\mathbb{Z} \longrightarrow \mathbb{Z}_{\geq 0}: i \longmapsto \dim_K \Phi_i$$

is called the *Hilbert function of  $M$  with respect to  $\Phi$* .

**Definition 2.1.58.** Let  $A$  be a (not necessarily commutative)  $K$ -algebra and assume that  $A$  is positively graded, i.e., it is endowed with a family  $G = (G_i)_{i \in \mathbb{Z}_{\geq 0}}$  of (left)  $K$ -subspaces of  $A$  such that

$$A = \bigoplus_{i \in \mathbb{Z}_{\geq 0}} G_i \quad \text{and} \quad G_i \cdot G_j \subseteq G_{i+j} \quad \text{for all } i, j \in \mathbb{Z}_{\geq 0}.$$

Moreover, let  $M$  be a left  $A$ -module with  $G$ -grading  $\Gamma = (\Gamma_i)_{i \in \mathbb{Z}}$ , i.e., a family of (left)  $K$ -subspaces of  $M$  such that

$$M = \bigoplus_{i \in \mathbb{Z}} \Gamma_i \quad \text{and} \quad G_i \cdot \Gamma_j \subseteq \Gamma_{i+j} \quad \text{for all } i \in \mathbb{Z}_{\geq 0}, \quad j \in \mathbb{Z}.$$

We assume that  $M$  is finitely generated and that each  $\Gamma_i$  has finite  $K$ -dimension. Then the *Hilbert series of  $M$  with respect to  $\Gamma$*  is defined by the (formal) Laurent series

$$H_{M, \Gamma}(\lambda) := \sum_{i \in \mathbb{Z}} (\dim_K \Gamma_i) \lambda^i \in \mathbb{Z}((\lambda)).$$

The map

$$\mathbb{Z} \longrightarrow \mathbb{Z}_{\geq 0}: i \longmapsto \dim_K \Gamma_i$$

is called the *Hilbert function of  $M$  with respect to  $\Gamma$* .

We recall that every grading defines an increasing filtration on the same module and that every increasing filtration defines an associated graded module over the associated graded ring.

**Remark 2.1.59.** Given a  $K$ -algebra  $A$  with grading  $G = (G_i)_{i \in \mathbb{Z}_{\geq 0}}$  and a left  $A$ -module  $M$  with  $G$ -grading  $\Gamma = (\Gamma_i)_{i \in \mathbb{Z}}$  as in the previous definition, an (exhaustive) increasing filtration  $F = (F_i)_{i \in \mathbb{Z}_{\geq 0}}$  of  $A$  is defined by

$$F_i := \bigoplus_{j \leq i} G_j, \quad i \in \mathbb{Z}_{\geq 0},$$

and an (exhaustive) increasing  $F$ -filtration  $\Phi = (\Phi_i)_{i \in \mathbb{Z}}$  of  $M$  is defined by

$$\Phi_i := \bigoplus_{j \leq i} \Gamma_j, \quad i \in \mathbb{Z}.$$

If  $M$  is finitely generated and each  $\Gamma_i$  has finite  $K$ -dimension, then each  $\Phi_i$  has finite  $K$ -dimension, and we have

$$H_{M, \Phi}(\lambda) = H_{M, \Gamma}(\lambda) \cdot \frac{1}{1 - \lambda}.$$

Conversely, in the situation of Definition 2.1.57, the *associated graded ring* is defined to be the  $K$ -algebra

$$\text{gr}(A) := \bigoplus_{i \in \mathbb{Z}_{\geq 0}} (F_i / F_{i-1}) \quad (\text{where } F_{-1} := \{0\})$$

with multiplication

$$(p_1 + F_{i-1}) \cdot (p_2 + F_{j-1}) := p_1 \cdot p_2 + F_{i+j-1}, \quad p_1 \in F_i, \quad p_2 \in F_j,$$

and the *associated graded module* is defined by

$$\text{gr}(M) := \bigoplus_{i \in \mathbb{Z}} (\Phi_i / \Phi_{i-1})$$

with left  $\text{gr}(A)$ -action

$$(p + F_{i-1})(m + \Phi_{j-1}) = p \cdot m + \Phi_{i+j-1}, \quad p \in F_i, \quad m \in \Phi_j.$$

The grading of  $\text{gr}(M)$  defines again an increasing filtration of  $\text{gr}(M)$ , but since  $A$  and  $\text{gr}(A)$  are non-isomorphic rings in general, the resulting filtration reflects only partial information about  $M$ . (Note also that, even if  $M$  is assumed to be finitely generated and each  $\Phi_i$  has finite  $K$ -dimension,  $\text{gr}(M)$  is not a finitely generated  $\text{gr}(A)$ -module in general; cf., e.g., [Bjö79, Sect. 1.2] or [Cou95, Sect. 8.3].)

The following two remarks establish a link between the Hilbert series of certain graded modules and filtered modules, respectively, and the generalized Hilbert series, which is computable via Janet bases. (The first remark will be applied in Remark 2.1.64, the second one in Remark 3.2.17.)

**Remark 2.1.60.** Let  $D = K[x_1, \dots, x_n]$  be a commutative polynomial algebra over the field  $K$  and assume  $D$  is positively graded. We denote by  $\deg(x_i)$  the degree of  $x_i$ ,  $i = 1, \dots, n$ , with respect to this grading  $G$ . Let  $q \in \mathbb{N}$  and let  $\Gamma = (\Gamma_i)_{i \in \mathbb{Z}}$  be a  $G$ -grading of  $D^{1 \times q}$  such that each  $\Gamma_i$  is a finite-dimensional  $K$ -vector space. For any submodule  $M$  of  $D^{1 \times q}$  such that  $\Gamma'_i := \Gamma_i \cap M$ ,  $i \in \mathbb{Z}$ , defines a  $G$ -grading  $\Gamma'$  of  $M$ , a Janet basis  $J$  for  $M$  (with respect to any term ordering) provides via the generalized Hilbert series a  $K$ -basis of  $M$  (cf. Rem. 2.1.54 a)). Then the Hilbert series of  $M$  with respect to  $\Gamma'$  is obtained from the generalized Hilbert series by substitution:

$$H_{M, \Gamma'}(\lambda) = \sum_{k=1}^q H_{S, k}(\lambda^{\deg(x_1)}, \dots, \lambda^{\deg(x_n)}),$$

where  $S$  is the multiple-closed set generated by the leading monomials of elements of  $J$ .

In this case,  $D^{1 \times q}/M$  has the  $G$ -grading  $\Gamma'' = (\Gamma''_i)_{i \in \mathbb{Z}}$ , where  $\Gamma''_i$  is defined as the image of  $\Gamma_i$  under the canonical projection  $D^{1 \times q} \rightarrow D^{1 \times q}/M$ . The generalized Hilbert series of the complement  $\bar{S}$  of  $S$  in  $\text{Mon}(D^{1 \times q})$  yields the Hilbert series of  $D^{1 \times q}/M$  with respect to  $\Gamma''$ :

$$H_{D^{1 \times q}/M, \Gamma''}(\lambda) = \sum_{k=1}^q H_{\bar{S}, k}(\lambda^{\deg(x_1)}, \dots, \lambda^{\deg(x_n)})$$

(cf. Rem. 2.1.54 b)).

The maximum number of multiplicative variables of cones in a Janet decomposition of  $D^{1 \times q}/M$ , and therefore the order of 1 as a pole of the corresponding generalized Hilbert series, equals the Krull dimension of  $D^{1 \times q}/M$  (cf., e.g., [Sta96, I.5] or [SW91]).

**Remark 2.1.61.** Let  $D$  be an Ore algebra as before,  $q \in \mathbb{N}$ , and  $M$  a submodule of  $D^{1 \times q}$ . We define an (exhaustive) increasing filtration  $F = (F_i)_{i \in \mathbb{Z}_{\geq 0}}$  on  $D$  by

$$F_i := \{p \in D \mid p = 0 \text{ or } \deg(p) \leq i\}, \quad i \in \mathbb{Z}_{\geq 0},$$

where  $\deg$  denotes the total degree, and an (exhaustive) increasing  $F$ -filtration on  $D^{1 \times q}$  by

$$\Phi_i := \{t \in D^{1 \times q} \mid t = 0 \text{ or } \deg(t) \leq i\}, \quad i \in \mathbb{Z},$$

where  $\deg(t)$  is defined as the maximum of the total degrees of the non-zero components of the tuple  $t$ . (In case of the Weyl algebras  $A_n(K)$  this filtration is known as the Bernstein filtration; cf., e.g., [Bjö79] or [Cou95].) Intersecting with  $M$  defines an  $F$ -filtration  $\Phi' := (\Phi'_i \cap M)_{i \in \mathbb{Z}}$  of  $M$ .

Assumption 2.1.23 implies that the associated graded ring  $\text{gr}(D)$  is isomorphic to the commutative polynomial algebra  $K[\xi_1, \dots, \xi_n, \eta_1, \dots, \eta_l]$  with standard grading  $G$  (since the degrees of the indeterminates of  $D$  are all equal to one), and the  $\text{gr}(D)$ -module  $\text{gr}(M)$  is isomorphic to a graded  $K[\xi_1, \dots, \xi_n, \eta_1, \dots, \eta_l]$ -module. Let  $\Gamma$  be the  $G$ -grading of  $\text{gr}(M)$ .



Let  $J$  be a Janet basis for  $M$  with respect to an admissible term ordering which is compatible with the total degree. Then the corresponding generalized Hilbert series  $H_S(x_1, \dots, x_{n+l})$  enumerates a  $K$ -basis of  $M$  (cf. Rem. 2.1.54 a)), where  $S$  is the multiple-closed set generated by the leading monomials of elements of  $J$ . Since the term ordering is compatible with the total degree,

$$\bigsqcup_{(p, \mu) \in J} \text{Mon}(\mu) (\text{lm}(p) + \Phi'_{\deg(p)-1})$$

is a  $K$ -basis of  $\text{gr}(M)$ , so that the coefficient of  $\lambda^i$  in the formal power series  $H_S(\lambda, \dots, \lambda)$  equals  $\dim_K \Gamma_i$  for all  $i \in \mathbb{Z}_{\geq 0}$ . (Of course, we have  $\dim_K \Gamma_i = 0$  for all  $i \in \mathbb{Z}_{< 0}$ .) Therefore, we have

$$H_{\text{gr}(M), \Gamma}(\lambda) = \sum_{k=1}^q H_{S, k}(\lambda, \dots, \lambda).$$

More generally, if degrees (not necessarily equal to one) are assigned to the indeterminates  $z_1, \dots, z_n, \partial_1, \dots, \partial_l$  of  $D$ , the corresponding Hilbert series of  $\text{gr}(M)$  with respect to  $\Gamma$  is obtained from the generalized Hilbert series of  $S$  by substituting  $\lambda^{\deg(z_i)}$  for  $x_i, i = 1, \dots, n$ , and  $\lambda^{\deg(\partial_j)}$  for  $x_{n+j}, j = 1, \dots, l$ .

An (exhaustive) increasing  $F$ -filtration of the factor module  $D^{1 \times q}/M$  is given by  $\Phi'' := (\Phi''_i)_{i \in \mathbb{Z}}$ , where  $\Phi''_i$  is the image of  $\Phi_i$  under the canonical projection  $D^{1 \times q} \rightarrow D^{1 \times q}/M$ . Note that  $\text{gr}(D^{1 \times q}/M)$  is a finitely generated  $\text{gr}(D)$ -module. Let  $\Gamma''$  be the grading of  $\text{gr}(D^{1 \times q}/M)$ . If  $C$  is a cone decomposition of the complement  $\bar{S}$  of  $S$  in  $\text{Mon}(D^{1 \times q})$  (cf. Rem. 2.1.54 b)), then

$$\bigsqcup_{(t, v) \in C} \text{Mon}(v) ((t + M) + \Gamma''_{\deg(t)-1})$$

is a  $K$ -basis of  $\text{gr}(D^{1 \times q}/M)$  and we obtain the Hilbert series of  $\text{gr}(D^{1 \times q}/M)$  with respect to  $\Gamma''$  as follows:

$$H_{\text{gr}(D^{1 \times q}/M), \Gamma''}(\lambda) = \sum_{k=1}^q H_{\bar{S}, k}(\lambda, \dots, \lambda).$$

**Remark 2.1.62.** Let  $K$  be a field,  $D = K[x_1, \dots, x_n]$  a commutative polynomial algebra over  $K$  which is positively graded, and  $q \in \mathbb{N}$ . We denote by  $G$  the grading of  $D$  and let  $\Gamma = (\Gamma_i)_{i \in \mathbb{Z}}$  be a  $G$ -grading of the free left  $D$ -module  $D^{1 \times q}$  such that each  $\Gamma_i$  is a finite-dimensional  $K$ -vector space. Let  $M$  be a submodule of  $D^{1 \times q}$  such that

$$\Gamma'_i := \Gamma_i \cap M, \quad i \in \mathbb{Z},$$

defines a  $G$ -grading  $\Gamma'$  of  $M$ . Moreover, let

$$J = \{(p_1, \mu_1), \dots, (p_t, \mu_t)\}$$

be a Janet basis for  $M$  with respect to any admissible term ordering on  $\text{Mon}(D^{1 \times q})$ . Then the Hilbert series of  $M$  with respect to  $\Gamma'$  is given by

$$\begin{aligned} H_{M, \Gamma'}(\lambda) &= \sum_{i \in \mathbb{Z}} (\dim_K \Gamma'_i) \lambda^i \\ &= \sum_{k=1}^t \frac{\lambda^{\deg(p_k)}}{(1-\lambda)^{|\mu_k|}} \\ &= \sum_{k=1}^t \lambda^{\deg(p_k)} \sum_{j \geq 0} \binom{|\mu_k| + j - 1}{j} \lambda^j. \end{aligned} \quad (2.18)$$

For  $i \geq \max \{ \deg(p_k) \mid k = 1, \dots, t \}$ , we have

$$\dim_K \Gamma'_i = \sum_{k=1}^t \binom{|\mu_k| + i - \deg(p_k) - 1}{i - \deg(p_k)} = \sum_{k=1}^t \binom{|\mu_k| + i - \deg(p_k) - 1}{|\mu_k| - 1},$$

which is a polynomial in  $i$  of degree less than  $n + l$ . In other words, the Hilbert function of  $M$  with respect to  $\Gamma'$  is a polynomial function when restricted to integers greater than or equal to  $\max \{ \deg(p_k) \mid k = 1, \dots, t \}$ . This polynomial is called the *Hilbert polynomial of  $M$  with respect to  $\Gamma'$* . In a similar way the notion of Hilbert polynomial is defined for a residue class module of  $D^{1 \times q}$  with respect to some  $G$ -grading and for submodules and residue class modules of  $D^{1 \times q}$  with respect to (exhaustive) increasing filtrations.

Note that if non-standard degrees are assigned to the indeterminates of  $D$ , these have to be taken into account in the right hand side of (2.18) in terms of the corresponding powers of  $\lambda$ . Thus, in general, the Hilbert function is asymptotically polynomial on residue classes (also called quasipolynomial, cf. [Sta96, Sect. 0.1]).

**Remark 2.1.63.** In the situation of the previous remark let  $I$  be a homogeneous ideal of the commutative polynomial algebra  $D$  with standard grading, i.e.,  $q = 1$ . Then the degree  $d$  of the Hilbert polynomial of  $D/I$  equals the dimension of the corresponding projective variety in projective  $(n-1)$ -space defined over an algebraic closure  $\bar{K}$  of  $K$  (cf., e.g., [Eis95]). The product of the leading coefficient of the Hilbert polynomial and  $d!$  is called the degree of the corresponding projective variety and coincides with the number of points in which the variety intersects a generic projective subspace of dimension  $n-1-d$ .

For the case of an algebra  $D$  of differential operators with rational function coefficients (cf. Ex. 2.1.18 b), p. 19) and general  $q$  an upper bound for this product in terms of the numbers of independent and dependent variables, the number of equations, the maximum differential order, and the degree of the Hilbert polynomial of the given system was derived in [Gri05].

In the following remark we outline one application of the generalized Hilbert series to commutative algebra and algebraic geometry. This application provides a constructive and deterministic approach to the Noether normalization lemma.

**Remark 2.1.64.** For a given finitely generated commutative algebra over a field, the Noether normalization lemma (cf., e.g., [Eis95], [Vas98]) ensures the existence of a maximal subset of algebraically independent elements such that the given algebra is an integral extension of the polynomial ring generated by this system of parameters. An affine variety whose coordinate ring is isomorphic to the given algebra is therefore shown to be a branched covering of some affine space.

The normalization lemma can be proved in a constructive manner, but most of the computational approaches today perform a random change of coordinates producing very large polynomials, which are difficult to handle afterwards.

Given an ideal  $I$  of a commutative polynomial algebra  $D = K[x_1, \dots, x_n]$  over a field  $K$ , the generalized Hilbert series of  $D/I$  can be used effectively to construct sparse coordinate changes which achieve Noether normal position for the given ideal [Rob09].

The maximum number of multiplicative variables of cones in a Janet decomposition of  $D/I$  equals the Krull dimension  $d$  of  $D/I$  (cf. Rem. 2.1.60). Let  $\mathbf{v}$  be the union of all sets of multiplicative variables of the cones of such a decomposition. The crucial observation is that a variable is not an element of  $\mathbf{v}$  if and only if a power of that variable is a leading monomial of an element of the Janet basis  $J$  for  $I$ . If  $|\mathbf{v}| = d$ , then the set  $Y$  of residue classes of the elements of  $\mathbf{v}$  in  $D/I$  is a maximal subset of algebraically independent elements of  $D/I$  such that  $D/I$  is an integral extension of  $K[Y]$ . Otherwise, we have  $|\mathbf{v}| > d$  and coordinates should be changed in such a way that the Janet basis for the transformed ideal has more elements whose leading monomial is a power of a variable. It turns out that a good strategy is to investigate an element  $p \in J$  such that  $\text{lm}(p) \in \text{Mon}(\mathbf{v})$  and  $\text{lm}(p)$  involves the least number of variables and is maximal with respect to the chosen term ordering  $>$  among these candidates. Then the coordinate transformation is chosen in such a way that each variable dividing  $\text{lm}(p)$  is mapped to a linear combination of variables in which the  $>$ -greatest variable in  $\mathbf{v}$  has non-zero coefficient<sup>7</sup>. We demonstrate this procedure in the following example and refer to [Rob09] for more details.

**Example 2.1.65.** Let  $D = \mathbb{Q}[w, x, y, z]$  be the commutative polynomial algebra and choose the degree-reverse lexicographical ordering  $>$  on  $D$  which extends the ordering  $w > x > y > z$ . Let  $I$  be the ideal of  $D$  which is generated by

$$L := \{ wxy^2 - y^2z, xyz - wz^2, y^2z - wx^2yz \}.$$

It is not radical and has five minimal associated primes of dimensions 1, 2, 2, 2, 2, respectively, and one embedded associated prime of dimension 1. All the following computations were done in the computer algebra system Maple in a couple of seconds using the package `Involutive` [BCG<sup>+</sup>03a] (cf. also Subsect. 2.1.6).

Let  $J_1$  be a Janet basis for  $I$  with respect to the term ordering  $>$  (using the total ordering  $w > x > y > z$  for determining the multiplicative variables). The Janet decomposition of the complement of  $\text{lm}(I)$  in  $\text{Mon}(\{w, x, y, z\})$  (determined by Alg. 2.1.8,

<sup>7</sup> If the ground field  $K$  is finite and not large enough, it may be necessary to use polynomials of higher degree to define the coordinate transformation.

p. 14) yields the generalized Hilbert series<sup>8</sup> of  $D/I$ :

$$\begin{aligned} & \frac{1}{1-z} + \frac{y}{1-z} + \frac{x}{(1-x)(1-z)} + \frac{w}{1-z} + \frac{y^2}{1-y} + \frac{xy}{(1-x)(1-y)} + wy + \frac{wx}{(1-x)(1-z)} + \\ & w^2 + \frac{y^2 z}{1-y} + wyz + w^2 z + wy^2 + \frac{wxy}{1-x} + w^2 y + \frac{w^2 x}{1-x} + \frac{w^3}{1-w} + \frac{y^2 z^2}{1-y} + wyz^2 + w^2 z^2 + \\ & wy^2 z + w^2 yz + \frac{w^2 xz}{1-x} + \frac{w^3 z}{1-w} + \frac{wy^3}{1-y} + w^2 y^2 + \frac{w^2 xy}{1-x} + \frac{w^3 y}{1-w} + \frac{w^3 x}{(1-w)(1-x)} + \\ & \frac{y^2 z^3}{1-y} + w^2 z^3 + wy^2 z^2 + \frac{w^3 z^2}{1-w} + w^2 y^2 z + \frac{w^3 yz}{1-w} + \frac{w^3 xz}{(1-w)(1-x)} + \frac{w^2 y^3}{1-y} + \frac{w^3 y^2}{1-w} + \\ & \frac{w^3 xy}{(1-w)(1-x)} + \frac{y^2 z^4}{1-y} + \frac{w^3 y^2 z}{1-w} + \frac{w^3 y^3}{(1-w)(1-y)}. \end{aligned}$$

Hence, the sets of multiplicative variables  $\mu_i$  of the Janet decomposition are among the following ones:

$$\emptyset, \quad \{w\}, \quad \{x\}, \quad \{y\}, \quad \{z\}, \quad \{w, x\}, \quad \{w, y\}, \quad \{x, y\}, \quad \{x, z\}.$$

The Krull dimension  $d$  of  $D/I$  equals 2. We have  $v_1 := \bigcup \mu_i = \{w, x, y, z\}$ , and so  $|v_1| = 4 > d$ . In order to keep the coordinate transformation sparse, it is advisable to choose

$$p_1 = w^2 z^4 - wy^2 z^2 \in J_1,$$

whose leading monomial  $\text{lm}(p_1) = w^2 z^4$  involves only two variables. We choose the automorphism  $\psi_1 : D \rightarrow D$  (restricting to the identity on  $K$ ) which maps  $z$  to  $z - w$  and fixes all other variables.

Let  $J_2$  be a Janet basis for  $\psi_1(I)$  (with the same specifications as above). The generalized Hilbert series of  $D/\psi_1(I)$  is given by

$$\begin{aligned} & \frac{1}{(1-y)(1-z)} + \frac{x}{1-z} + \frac{w}{1-z} + \frac{xy}{1-z} + \frac{wy}{1-z} + \frac{x^2}{1-z} + \frac{wx}{(1-x)(1-z)} + \frac{w^2}{1-z} + xy^2 + \\ & \frac{wy^2}{1-z} + \frac{x^2 y}{1-z} + \frac{wxy}{(1-x)(1-z)} + \frac{w^2 y}{1-z} + \frac{x^3}{(1-x)(1-z)} + \frac{w^2 x}{1-z} + xy^2 z + xy^3 + \frac{wy^3}{1-y} + \\ & x^2 y^2 + w^2 y^2 + \frac{x^3 y}{(1-x)(1-z)} + \frac{w^2 xy}{1-z} + \frac{w^2 x^2}{(1-x)(1-z)} + xy^3 z + \frac{wy^3 z}{1-y} + x^2 y^2 z + w^2 y^2 z + \\ & \frac{xy^4}{1-y} + \frac{x^2 y^3}{1-y} + \frac{x^3 y^2}{(1-x)(1-y)} + \frac{wy^3 z^2}{1-y} + w^2 y^2 z^2 + \frac{wy^3 z^3}{1-y}. \end{aligned}$$

In particular, the Janet decomposition of  $\text{Mon}(\{w, x, y, z\}) - \text{lm}(\psi_1(I))$  has sets of multiplicative variables among the following ones:

$$\emptyset, \quad \{y\}, \quad \{z\}, \quad \{x, y\}, \quad \{x, z\}, \quad \{y, z\}.$$

We have  $v_2 := \{x, y, z\}$ , and therefore  $|v_2| = 3 > d$ . Now we choose the polynomial

<sup>8</sup> Using the package *Involutive* (cf. Subsect. 2.1.6), the generalized Hilbert series can be obtained with the command `FactorModuleBasis`, after applying `InvolutiveBasis` to  $L$ .

$$p_2 = xy^2z^2 - w^2y^2 + wy^3 + 3wy^2z - y^3z - 2y^2z^2 \in J_2$$

with  $\text{lm}(p_2) = xy^2z^2$ , and the automorphism  $\psi_2$  of  $D$  mapping  $y$  to  $y - x$ ,  $z$  to  $z - x$ , and fixing  $w$  and  $x$ .

Finally, we compute a Janet basis  $J_3$  for  $(\psi_2 \circ \psi_1)(I)$ . The generalized Hilbert series of  $D/(\psi_2 \circ \psi_1)(I)$  is given by

$$\begin{aligned} & \frac{1}{(1-y)(1-z)} + \frac{x}{(1-y)(1-z)} + \frac{w}{(1-y)(1-z)} + \frac{x^2}{(1-y)(1-z)} + \frac{wx}{(1-y)(1-z)} + \\ & \frac{w^2}{(1-y)(1-z)} + \frac{x^3}{(1-y)(1-z)} + \frac{wx^2}{1-z} + \frac{w^2x}{1-z} + \frac{wx^2y}{1-y} + \frac{w^2xy}{1-z} + x^4 + w^2x^2 + \\ & \frac{wx^2yz}{1-y} + x^4z + w^2x^2z + w^2xy^2 + \frac{wx^2yz^2}{1-y} + w^2x^2z^2 + \frac{wx^2yz^3}{1-y}. \end{aligned}$$

The Janet decomposition of the complement of  $\text{lm}((\psi_2 \circ \psi_1)(I))$  in  $\text{Mon}(\{w, x, y, z\})$  consists of cones having sets of multiplicative variables among the following ones:

$$\emptyset, \quad \{y\}, \quad \{z\}, \quad \{y, z\}.$$

Thus,  $v_3 := \{y, z\}$  and  $|v_3| = d$ , and we are done<sup>9</sup>. The coordinate change  $\psi_2 \circ \psi_1$  is defined by

$$w \mapsto w, \quad x \mapsto x, \quad y \mapsto y - x, \quad z \mapsto z - x - w.$$

The maximum number of summands of a polynomial in  $J_3$  is 102. The coefficient in  $J_3$  of largest absolute value equals 40.

A typical coordinate transformation to Noether normal position returned by the (randomized) command `noetherNormal` of the computer algebra system Singular (version 3-1-6) [DGPS12] is defined by

$$w \mapsto w, \quad x \mapsto 10w + x, \quad y \mapsto 6w + 10x + y, \quad z \mapsto 8w + 4x + 3y + z,$$

which in this case results in a Gröbner basis of the transformed ideal with coefficient of largest absolute value of more than 30 decimal digits and maximum number of summands 123.

For more details about this approach to Noether normalization and a more systematic comparison of some existing implementations, we refer to [Rob09].

In the rest of this subsection we discuss the relevance of the generalized Hilbert series for systems of linear partial differential equations. Computation of a Janet basis for such a system produces an equivalent system which is ensured to be *formally integrable*, i.e., it admits a straightforward method of determining all formal power series solutions from the equations of the system (which is in some sense similar to back substitution applied to the result of Gaussian elimination). In general, two distinct equations may yield a non-trivial consequence of lower differentiation order

<sup>9</sup> Note that neither of the Janet bases  $J_1, J_2, J_3$  is a Pommaret basis, i.e., Noether normalization is achieved using a sparse transformation which does not define  $\delta$ -regular coordinates.

when the highest terms in a suitable linear combination of certain of their derivatives cancel. If the system is not formally integrable, the computation of a power series solution from the given equations may miss the conditions implied by such consequences. Since Janet's algorithm determines the multiple-closed set of monomials which occur as leading monomials of consequences of the system, a Janet basis reveals all conditions on Taylor coefficients of a solution.

We recall that the vector space which is dual to a polynomial algebra over a field is given by the algebra of formal power series in the same number of indeterminates. This relationship will be generalized to Ore algebras in the following remark.

**Remark 2.1.66.** Let  $D := A[\partial_1; \sigma_1, \delta_1] \dots [\partial_l; \sigma_l, \delta_l]$  be an Ore algebra, where the domain  $A$  is an algebra over the field  $K$ , and define

$$\mathcal{F} := \text{hom}_K(D, K).$$

Since multiplication in  $K$  is commutative, the set  $\mathcal{F}$  of all homomorphisms from the left  $K$ -vector space  $D$  to  $K$  is a left  $K$ -vector space. Moreover,  $\mathcal{F}$  is a left  $D$ -module in virtue of

$$D \times \mathcal{F} \longrightarrow \mathcal{F}: (d, f) \longmapsto (a \mapsto f(a \cdot d)),$$

and this left action of  $D$  restricts to the left action of  $K$  because every element of  $K$  commutes with every element of  $D$ . We have a pairing of  $D$  and  $\mathcal{F}$ , i.e., a  $K$ -bilinear form

$$(\ , \ ) : D \times \mathcal{F} \longrightarrow K: (d, f) \longmapsto f(d) \quad (2.19)$$

which is non-degenerate in both arguments. With respect to this pairing,  $D$  and  $\mathcal{F}$  can be considered as dual to each other. Moreover, the linear map  $D \rightarrow D$  defined by right multiplication by a fixed element  $d \in D$  and the linear map  $\mathcal{F} \rightarrow \mathcal{F}$  given by left multiplication by the same element  $d$  are adjoint to each other:

$$(a \cdot d, f) = f(a \cdot d) = (d \cdot f)(a) = (a, d \cdot f), \quad a \in D, \quad f \in \mathcal{F}. \quad (2.20)$$

Since every homomorphism  $f \in \mathcal{F} = \text{hom}_K(D, K)$  is uniquely determined by its values for the elements of the  $K$ -basis  $\text{Mon}(D)$  of  $D$ , we can write  $f$  in a unique way as a (not necessarily finite) formal sum

$$\sum_{m \in \text{Mon}(D)} (m, f) m. \quad (2.21)$$

Due to (2.20), for every  $d \in D$  the representation of  $d \cdot f$  can be obtained from

$$\sum_{m \in \text{Mon}(D)} (m, d \cdot f) m = \sum_{m \in \text{Mon}(D)} (m \cdot d, f) m. \quad (2.22)$$

It is reasonable to write the monomials in the sum (2.21) using new indeterminates, which will be done in the following remark dealing with the case of commutative polynomial algebras.

**Remark 2.1.67.** Let  $D$  be the commutative polynomial algebra  $K[\partial_1, \dots, \partial_n]$  over a field  $K$  of characteristic zero and  $>$  a term ordering on  $D$ . Then  $(\partial^\beta \mid \beta \in (\mathbb{Z}_{\geq 0})^n)$  is a  $K$ -basis of  $D$ . We define  $\mathcal{F} := \text{hom}_K(D, K)$  with (left)  $D$ -module structure and the pairing in (2.19) as in the previous remark. The discussion leading to (2.21) shows that  $\mathcal{F}$  can be considered as the  $K$ -algebra  $K[[z_1, \dots, z_n]]$  of formal power series in the same number  $n$  of indeterminates. Moreover, it follows from (2.22) that the (left) action on  $\mathcal{F}$  of any monomial in  $D$  effects a shift of the coefficients of the power series according to the exponent vector of the monomial, which is the same action as the one defined by partial differentiation. Therefore, we establish the identification of  $\mathcal{F}$  with  $K[[z_1, \dots, z_n]]$  in such a way that

$$(z^\alpha / \alpha! \mid \alpha \in (\mathbb{Z}_{\geq 0})^n) \quad \text{and} \quad (\partial^\beta \mid \beta \in (\mathbb{Z}_{\geq 0})^n)$$

are dual to each other with respect to the pairing (2.19), i.e.,

$$\left( \partial^\beta, \sum_{\alpha \in (\mathbb{Z}_{\geq 0})^n} c_\alpha \frac{z^\alpha}{\alpha!} \right) = c_\beta, \quad \beta \in (\mathbb{Z}_{\geq 0})^n, \quad \alpha! := \alpha_1! \cdots \alpha_n!.$$

Suppose that a system of (homogeneous) linear PDEs with constant coefficients for one unknown function of  $n$  arguments is given. We compute a Janet basis  $J$  for the ideal of  $D$  which is generated by the left hand sides  $p$  of these equations with respect to the term ordering  $>$ . The differential equations are considered as linear equations for  $(\partial^\beta, f)$ ,  $\beta \in (\mathbb{Z}_{\geq 0})^n$ , where  $f \in \mathcal{F}$  is a formal power series solution, and using the term ordering  $>$ , we may solve each of these equations for  $(\text{lm}(p), f)$ . Then Janet's algorithm partitions  $\text{Mon}(D)$  into a set of monomials  $m$  for which  $(m, f) \in K$  can be chosen arbitrarily and a set  $S$  of monomials for which  $(\text{lm}(p), f) \in K$  is uniquely determined by these choices. The latter set is the multiple-closed subset

$$S := [\text{lm}(p) \mid (p, \mu) \in J]$$

of  $\text{Mon}(D)$ . In particular, the  $K$ -dimension of the space of formal power series solutions, if finite, can be computed as the number of monomials in the complement  $C$  of  $S$  in  $\text{Mon}(D)$ . In fact, the generalized Hilbert series  $H_C(\partial_1, \dots, \partial_n)$  of  $C$  enumerates a basis for the Taylor coefficients  $(\partial^\beta, f)$  of  $f$  whose values can be assigned freely.

M. Janet called the monomials  $\partial^\beta$  in  $\text{Mon}(D) - S$  *parametric derivatives* because the corresponding Taylor coefficients  $(\partial^\beta, f)$  of a formal power series solution  $f$  can be chosen arbitrarily. The monomials in  $S$  are called *principal derivatives* [Jan29, e.g., no. 22, no. 38]. The Taylor coefficients  $(\partial^\beta, f)$  which correspond to principal derivatives  $\partial^\beta$  are uniquely determined by  $K$ -linear equations in terms of the Taylor coefficients of parametric derivatives. Of course, the extension of this method of determining the formal power series solutions of a system of linear partial differential equations is extended to the case of more than one unknown function in a straightforward way by using submodules of  $D^{1 \times q}$  instead of ideals of  $D$ .

Note that convergence of series solutions is to be investigated separately.

For a similar treatment of partial difference equations, we refer to [OP01].

**Example 2.1.68.** [Jan29, no. 23] The left hand side of the heat equation

$$\frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \quad (2.23)$$

for an unknown real analytic function  $u$  of  $t$  and  $x$  is represented by the polynomial

$$p := \partial_t - \partial_x^2 \in D := K[\partial_t, \partial_x],$$

where  $K = \mathbb{Q}$  or  $\mathbb{R}$ . Choosing a degree-reverse lexicographical term ordering on the polynomial algebra  $D$ , the leading monomial of  $p$  is  $\partial_x^2$ . The polynomial  $p$  forms a Janet basis for the ideal of  $D$  it generates, and the parametric derivatives are given by  $\partial_t^i, \partial_t^j \partial_x, i, j \in \mathbb{Z}_{\geq 0}$ . Hence, any choice of formal power series in  $t$  for  $u(t, 0)$  and  $\frac{\partial u}{\partial x}(t, 0)$  uniquely determines a formal power series solution  $u$  to (2.23). In this case, every choice of convergent power series yields a convergent series solution  $u$ . On the other hand, using the lexicographical term ordering extending  $t > x$ , the parametric derivatives are given by  $\partial_x^i, i \in \mathbb{Z}_{\geq 0}$ . Now, the choice

$$u(0, x) = \sum_{i \geq 0} x^i$$

determines a divergent series solution  $u$ .

The following example demonstrates how a Janet decomposition (and the resulting generalized Hilbert series) of the complement of the set of principal derivatives in  $\text{Mon}(D)$  allows to collect the parametric derivatives in such a way as to express the solutions in terms of arbitrary functions and constants.

**Example 2.1.69.** For illustrative reasons, we consider the system of linear partial differential equations for one unknown analytic function  $u$  of  $x, y, z$  which corresponds to the set of monomials dealt with in Examples 2.1.7, p. 12, and 2.1.10:

$$\frac{\partial^2 u}{\partial x \partial y} = 0, \quad \frac{\partial^4 u}{\partial x^3 \partial z} = 0. \quad (2.24)$$

The Janet completion already yields the (minimal) Janet basis

$$\begin{aligned} \frac{\partial^2 u}{\partial x \partial y} &= 0, \{ *, \partial_y, \partial_z \}, \\ \frac{\partial^3 u}{\partial x^2 \partial y} &= 0, \{ *, \partial_y, \partial_z \}, \\ \frac{\partial^4 u}{\partial x^3 \partial z} &= 0, \{ \partial_x, *, \partial_z \}, \\ \frac{\partial^4 u}{\partial x^3 \partial y} &= 0, \{ \partial_x, \partial_y, \partial_z \} \end{aligned}$$



and we obtain the following Janet decomposition of the set of parametric derivatives (cf. also Ex. 2.1.10 and Fig. 2.2, p. 15):

$$\begin{aligned} &1, \{ *, \partial_y, \partial_z \}, \\ &\partial_x, \{ *, *, \partial_z \}, \\ &\partial_x^2, \{ *, *, \partial_z \}, \\ &\partial_x^3, \{ \partial_x, *, * \}. \end{aligned}$$

The corresponding generalized Hilbert series is

$$\frac{1}{(1 - \partial_y)(1 - \partial_z)} + \frac{\partial_x}{1 - \partial_z} + \frac{\partial_x^2}{1 - \partial_z} + \frac{\partial_x^3}{1 - \partial_x}.$$

Accordingly, a formal power series solution  $u$  of (2.24) is uniquely determined as

$$u(x, y, z) = f_0(y, z) + x f_1(z) + x^2 f_2(z) + x^3 f_3(x)$$

by any choice of formal power series  $f_0, f_1, f_2, f_3$  of the indicated variables.

In general, the expression of the solutions in terms of arbitrary functions and constants depends on the choices of the coordinate system, the term ordering, and the total ordering which is used for determining the Janet decomposition. However, the maximum number of arguments of functions which occur in such an expression is invariant because it is the Krull dimension of the corresponding (graded) module (over the associated graded ring defined in Rem. 2.1.61), cf. also Rem. 2.1.60. The number of cones in a Janet decomposition having a fixed number of multiplicative variables and generator of a certain degree is also referred to as *Cartan character*.

**Remark 2.1.70.** The statements of Remark 2.1.67 also apply to systems of linear PDEs whose coefficients are rational functions in  $z_1, \dots, z_n$ , i.e.,  $D = K[\partial_1, \dots, \partial_n]$  is replaced with  $B_n(K) = K(z_1, \dots, z_n)\langle \partial_1, \dots, \partial_n \rangle$  (introduced also in Ex. 2.1.18 b) using Ore algebra notation), where  $K$  is the subfield of constants of  $K(z_1, \dots, z_n)$ .

Let  $M$  be the submodule of  $B_n(K)^{1 \times p}$  which is generated by the left hand sides of the equations (for  $p$  unknown functions) and let  $J$  be a Janet basis for  $M$ . Now, a formal power series solution is determined by any choice of Taylor coefficients for the parametric derivatives, if the left hand sides of the given PDE system are also defined over  $A\langle \partial_1, \dots, \partial_n \rangle^{1 \times p}$ , where  $A$  is a  $K$ -subalgebra of  $K(z_1, \dots, z_n)$  whose elements do not have a pole in  $0 \in K^n$ , if  $J$  is computed within  $A\langle \partial_1, \dots, \partial_n \rangle^{1 \times p}$ , and if  $0$  is not a zero of the leading coefficient of any element of  $J$ . In other words, if  $0$  is not a zero of any denominator arising in the course of Janet's algorithm applied to the PDE system and is not a zero of any leading coefficient, then all power series solutions are obtained in this way. Accordingly, having computed a Janet basis for  $M$ , the center  $c \in K^n$  for the Taylor series expansion of an analytic solution has to be chosen in such a way that the previous conditions are met with  $0$  replaced with  $c$ .

Similar remarks hold for the case of coefficients in a field of meromorphic functions on a connected open subset of  $\mathbb{C}^n$ .

## 2.1.6 Implementations

Work by the author of this monograph on implementations of techniques related to Janet's algorithm is summarized in this subsection. We also give references to software serving similar purposes. However, because of the large number of implementations of Buchberger's algorithm, a complete review is not aimed for.

The formulation of Algorithm 2.1.42, p. 29, computing Janet bases (and of the algorithms on which it depends) ignores the matter of realizing these techniques as efficient computer programs. For instance, the alternating use of *Auto-reduce* and *Decompose* in Algorithm 2.1.42 clearly removes left multiples of generators by non-multiplicative variables which may be added again if required by the Janet decomposition. Moreover, the computation of the Janet normal form of left multiples of generators by non-multiplicative variables should not be performed a second time when it is clear that the reduction steps will not differ from the previous computation.

The *involutive basis algorithm* (cf., e.g., [Ger05]), developed in work of V. P. Gerdt, Y. A. Blinkov, and A. Y. Zharkov, provides an efficient method to compute Janet bases. It builds on the more general concept of involutive division, which allows for other ways of defining multiplicative variables for generators than the pattern named after Janet. (We refer to [GB11] and ongoing work for a recent development of an even more efficient involutive division.) Using a very small part of the history of an involutive basis computation, the reduced Gröbner basis for the same ideal or module (and with respect to the same term ordering) can be extracted as a subset of the involutive basis without further computation. Moreover, analogues of Buchberger's criteria [Buc79] in the context of involutive division avoid unnecessary passivity checks. Heuristic strategies determining in which order the left multiples of generators by non-multiplicative variables should be considered for reduction are incorporated into the involutive approach.

A software package ALLTYPES realizing Riquier's and Janet's theory in the computer algebra system REDUCE [Hea99] has been developed by F. Schwarz (cf. [Sch08b, Sch84]). Another implementation in the programming language REFAL was described in [Top89]. The REDUCE package INVSYS, which computes Janet bases for ideals of commutative polynomial algebras, was developed by A. Y. Zharkov and Y. A. Blinkov [ZB96]. A program computing involutive bases for monomial ideals in Mathematica [Wol99] was reported on in [GBC98]. Moreover, involutive basis techniques have been implemented in MuPAD [CGO04] by M. Hausdorf and W. M. Seiler [HS02].

Symmetry analysis of systems of partial differential equations (cf., e.g., [Olv93], [Pom78], [Vin84], [BCA10], [Sch08a]) is an important area of application of Riquier's and Janet's theory. G. Reid and collaborators have been developing the *rif* algorithm and have been applying it in the symmetry analysis context, cf. [RWB96] and the references therein, and also [MRC98], where it had been combined with the differential Gröbner basis package of E. Mansfield [Man91]. By repeated prolongation and elimination steps as described in geometric approaches to differential

systems (cf., e.g., [Pom78] and the references therein), the *rif* algorithm transforms a system of nonlinear PDEs into *reduced involutive form*, which is formally integrable. An implementation by A. Wittkopf is available as a Maple package. For a review of further symbolic software for symmetry analysis of differential equations, cf. also [Her97].

Implementations of Buchberger's algorithm for computing Gröbner bases are available in many computer algebra systems and more specialized software, e.g., in AXIOM [JS92], Maple [MAP], Mathematica [Wol99], Magma [BCP97], REDUCE [Hea99], Singular [DGPS12], Macaulay2 [GS], CoCoA [CoC] (cf. also [CLO07, Appendix C] for a further discussion of such implementations). A variant of Buchberger's algorithm for systems of linear differential or difference equations and algorithms for computing Hilbert polynomials along with an implementation in the programming language REFAL were described in [Pan89].

An implementation of the involutive basis technique for commutative polynomial algebras over fields and for  $K\langle\partial_1, \dots, \partial_n\rangle$  as packages *Involutive* and *Janet*, respectively, for the computer algebra system Maple has been started by C. F. Cid at Lehrstuhl B für Mathematik, RWTH Aachen, in 2000. Here,  $K\langle\partial_1, \dots, \partial_n\rangle$  is the skew polynomial ring of differential operators with coefficients in a differential field  $K$  (of characteristic zero), whose arithmetic is implemented in Maple.

Since the year 2001 the author of this monograph has been adapting the packages *Involutive* and *Janet* to more recent versions of the involutive basis algorithm (with the help of V. P. Gerdt and Y. A. Blinkov) and has been extending these packages with new features.

Starting in 2003, the author of this monograph has been developing a Maple package *JanetOre*, which implements the involutive basis technique for certain iterated Ore extensions of a commutative polynomial algebra (as in Subsect. 2.1.3).

In collaboration with V. P. Gerdt a Maple package *LDA* (for “linear difference algebra”) has been developed since 2005, which computes involutive bases for left ideals of (and left modules over) rings of difference operators with coefficients in a difference field (of characteristic zero), whose arithmetics are supported by Maple. For applications of the package *LDA* to formal computational consistency checks of finite difference approximation of linear PDE systems, cf. [GR10].

We refer to [BCG<sup>+</sup>03a, BCG<sup>+</sup>03b, Rob07, GR06, GR12], the Maple help pages accompanying these packages, and the related web pages for more information.

The package *Involutive* computes Janet bases and Janet-like Gröbner bases (cf. Rem. 2.1.49, p. 37) for submodules of finitely generated free modules over commutative polynomial algebras with coefficients in  $\mathbb{Z}$  or finitely generated extension fields of  $\mathbb{Q}$  or finite fields that are supported by Maple. The implementation of the involutive basis algorithm has the additional feature that its computations may be performed in a parallel way on auxiliary data, which yields a means to record the history of a Janet basis computation. Syzygies and free resolutions, cf. Subsect. 3.1.5, can be computed by *Involutive*. Further procedures implementing module-theoretic constructions build on this possibility. (We also refer to the package *homalg* [BR08], which implements methods of homological algebra in

an abstract way, and to which `Involutive` can be connected. Delegating ring arithmetics to separate software, the package `homalg` provides an additional layer of abstraction. Meanwhile, `homalg` has been redesigned by M. Barakat as a package in GAP4 [GAP] and has been widely extended, e.g., being capable now of computing certain spectral sequences.) The package `Involutive` has been extended with functionality improving computation with rational function coefficients by M. Schröer and with procedures dealing with localizations at maximal ideals by M. Lange-Hegermann.

The Maple package `Janet` computes Janet bases and Janet-like Gröbner bases for submodules of finitely generated free left modules over the skew polynomial ring  $K\langle\partial_1, \dots, \partial_n\rangle$  of partial differential operators. Apart from implementing the counterpart of the module-theoretic methods of `Involutive` for the ring  $K\langle\partial_1, \dots, \partial_n\rangle$ , it provides, e.g., procedures which compute (truncated) formal power series solutions and polynomial solutions up to a given degree of systems of linear PDEs. The package `Janet` uses some data structures and procedures of the Maple package `jets` developed by M. Barakat [Bar01] and can be combined with `jets`, in order to compute symmetries of differential equations (cf., e.g., [Olv93]).

The functionality of the packages `JanetOre` and `LDA`, although handling different types of algebras, is analogous to that of `Involutive` and `Janet`, respectively.

Each of the above mentioned Maple packages provides combinatorial tools like the generalized Hilbert series (cf. Subsect. 2.1.5), Hilbert polynomials, Cartan characters, etc.

A very useful feature of these packages is the possibility to collect all expressions (typically arising as coefficients of polynomials) by which a Janet basis computation divided. Hence, the applicability of the performed computation for special values of parameters can be checked and singular configurations can be determined afterwards (cf. also Rem. 2.1.70).

The open source software package `ginv` implements the involutive basis technique in C++, using Python as an interpreter in addition [BG08]. Its development was initiated by V. P. Gerdt and Y. A. Blinkov. Contributions have been made at Lehrstuhl B für Mathematik, RWTH Aachen, during the last seven years, in particular by S. Jambor and the author of this monograph.

Interfaces between the Maple package `Involutive` and `ginv` are available including the possibility to delegate involutive basis computations during the current session of `Involutive` to the considerably faster C++ routines.

The author of this monograph implemented some parts of the involutive basis technique as a package `InvolutiveBases` [Rob] in Macaulay2 [GS].

Another Maple implementation of the involutive basis technique for linear PDEs is described in [ZL04].

## 2.2 Thomas Decomposition of Differential Systems

*A system of polynomial partial differential equations and inequations*

$$p_1 = 0, \quad \dots, \quad p_s = 0, \quad q_1 \neq 0, \quad \dots, \quad q_t \neq 0 \quad (s, t \in \mathbb{Z}_{\geq 0}) \quad (2.25)$$

for  $m$  unknown smooth functions of independent variables  $z_1, \dots, z_n$  is given by differential polynomials  $p_1, \dots, p_s, q_1, \dots, q_t$  in  $u_1, \dots, u_m$ , i.e., elements of the differential polynomial ring  $K\{u_1, \dots, u_m\}$  with commuting derivations  $\partial_1, \dots, \partial_n$ , where  $K$  is a differential field of characteristic zero. (For definitions of these notions of differential algebra, cf. Sect. A.3.) Similarly to Sect. 2.1 we will concentrate on analytic solutions.

Every solution of (2.25) satisfies all consequences of (2.25); the consequences we consider here are given by linear combinations of arbitrary partial derivatives of system equations with coefficients in  $K\{u_1, \dots, u_m\}$ , polynomial factors of (left hand sides of) such equations and of inequations, and quotients of (the respective left hand sides of) equations by inequations. Leaving aside for a moment the inequations, we then deal with the radical differential ideal of  $K\{u_1, \dots, u_m\}$  which is generated by  $p_1, \dots, p_s$ . Taking the inequations into account, the present section is concerned with an effective procedure which constructs a finite set of differential systems as in (2.25), whose sets of solutions form a partition of the solution set of (2.25), and such that all consequences of each resulting system can easily be described.

Let us assume for simplicity that (2.25) already has the same quality as each of these resulting systems. Then all consequences of (2.25) in terms of equations are

$$\{p = 0 \mid p \in \sqrt{E : q^\infty}\}, \quad (2.26)$$

where  $E$  is the differential ideal of  $R := K\{u_1, \dots, u_m\}$  which is generated by the polynomials  $p_1, \dots, p_s$ , the differential polynomial  $q$  is the product of  $q_1, \dots, q_t$ , and

$$E : q^\infty := \{p \in R \mid q^r \cdot p \in E \text{ for some } r \in \mathbb{Z}_{\geq 0}\}$$

is the *saturation* of  $E$  with respect to  $q$ . The solutions of (2.25) form an open subset of the set of solutions of (2.26) with respect to a certain topology. (It is, in fact, a dense subset, cf. Lemma 2.2.62.)

Let  $\mathcal{F}$  be a differential algebra over  $K$ , whose elements we think of as candidates for solutions of (2.26). Every homomorphism  $\varphi: K\{u_1, \dots, u_m\} \rightarrow \mathcal{F}$  of differential algebras over  $K$  is uniquely determined by its values  $f_1, \dots, f_m$  for  $u_1, \dots, u_m$ , and every choice of these values defines such a homomorphism. Now,  $(f_1, \dots, f_m)$  solves (2.26) if and only if the corresponding homomorphism  $\varphi$  of differential algebras factors over  $K\{u_1, \dots, u_m\}/\sqrt{E : q^\infty}$ . Thus, the set of homomorphisms

$$K\{u_1, \dots, u_m\}/\sqrt{E : q^\infty} \longrightarrow \mathcal{F}$$

of differential algebras over  $K$  is in one-to-one correspondence with the set of solutions  $(f_1, \dots, f_m) \in \mathcal{F}^m$  of (2.26).

This structural description of the solutions of (2.26) is analogous to the linear case (cf. the introduction to Sect. 2.1). However, only to some extent does it incorporate the conditions on solutions of (2.25) that are imposed by the given inequations. Moreover, even if no inequations are present in the given system, inequations emerge naturally. As it turns out, an equivalent form of (2.25) which allows to keep track of all of its consequences effectively, requires splittings into complementary systems. The approach we pursue here introduces inequations, which results in a partition of the solution set.

In this section we describe a method introduced by the American mathematician Joseph Miller Thomas (1898–1979) to deal in an effective way with systems of polynomial differential equations and inequations [Tho37, Tho62]. It belongs to the class of triangular decomposition methods (cf., e.g., the survey papers [Hub03a, Hub03b] by Evelyne Hubert) and can be used to compute characteristic sets (cf. also Subsect. A.3.2). Each system in the resulting decomposition admits an effective membership test for the corresponding differential ideal. A first implementation of this decomposition method was realized in the computer algebra system Maple by Dongming Wang [Wan98, LW99, Wan01, Wan04].

While the development of differential algebra following Joseph Fels Ritt in the twentieth century, in particular the work by Ellis R. Kolchin, did not seem to adapt the ideas of Thomas, they have been revived in recent years by Vladimir P. Gerdt [Ger08]. In the context of algebraic equations, Wilhelm Plesken introduced a univariate polynomial which is a counting invariant of a quasi-affine or quasi-projective variety (in given coordinates) in the sense that it counts the (closed) points using the indeterminate  $\infty$  for the cardinality of the affine line [Ple09a]. Markus Lange-Hegermann defined a differential counting polynomial and generalized the differential dimension polynomial, which had been introduced by E. R. Kolchin [Kol64] for prime differential ideals and which had been elaborated by Joseph Johnson [Joh69a], to differential systems which result from Thomas' method [LH14]. For further applications of the algebraic Thomas decomposition to algebraic varieties, to algebraic groups, and to linear codes and hyperplane arrangements we refer to [Ple09b], [PB14], [Bäc14]. An application of the differential Thomas decomposition to nonlinear control systems is developed in [LHR13].

In joint work of T. Bächler, V. P. Gerdt, M. Lange-Hegermann, and the author of this monograph the algorithmic details of J. M. Thomas' method of decomposing algebraic and differential systems into simple systems in combination with the notion of passive differential system following M. Janet have been worked out [BGL<sup>+</sup>10, BGL<sup>+</sup>12]. Implementations in Maple have been developed by T. Bächler and M. Lange-Hegermann (cf. also Subsect. 2.2.6).

The characteristic set method developed by J. F. Ritt and Wen-tsün Wu provides another decomposition algorithm, which, however, depends on the possibility to factor polynomials (cf., e.g., [Rit50] and also Subsect. A.3.2 for the rudiments of this theory, and [Wu89] for another variant). For algebraic systems, Wu's method competes with Janet and Gröbner basis techniques and has been applied to automated proving of theorems in geometry (cf., e.g., [Wan04]).

For applications of the characteristic set method to systems theory, we refer to work by Sette Diop, cf., e.g., [Dio92].

Abraham Seidenberg developed an elimination method for differential algebra [Sei56] by using the same splitting technique as J. M. Thomas. As a result, a constructive analog of Hilbert's Nullstellensatz for differential algebra was obtained. For the case of ordinary differential equations an algorithm with improved complexity was given by Dmitry Grigoryev in [Gri89].

Combining Seidenberg's theory and Buchberger's algorithm, the Rosenfeld-Gröbner algorithm, described in [BLOP95, BLOP09], computes a representation of a radical differential ideal as finite intersection of certain differential ideals, each of which also allows an effective membership test. The interactions of the relevant differential and algebraic constructions were investigated in [Hub00]. This approach is based on Rosenfeld's Lemma in differential algebra [Ros59], which is also applicable in the context of Thomas' theory. However, the assumption of coherence of an auto-reduced set of differential polynomials is replaced here with a passivity condition in the sense of Janet (cf. Sect. 2.1). The Rosenfeld-Gröbner algorithm is implemented in the Maple package `DifferentialAlgebra` (formerly `diffalg`). A description of its foundation based on Kolchin's book [Kol73] was given in [Sad00]. Another approach to characteristic sets using Gröbner bases was presented in [BKRM01].

Yet another direction of research tries to adapt the notion of Gröbner basis to the case of a differential polynomial ring, cf., e.g., [CF07]. In general a differential ideal may admit only infinite differential Gröbner bases as defined by Giuseppa Carrà Ferro or infinite standard bases as defined by François Ollivier in this context [Oll91]. Elizabeth L. Mansfield developed an algorithm for the computation of a different kind of (finite) differential Gröbner basis (cf. [Man91]), which applies pseudo-reductions, but does not analyze the initials of divisors, and which therefore may result in a basis which cannot be used to decide membership to the given differential ideal.

Subsection 2.2.1 is devoted to the Thomas decomposition of systems of algebraic equations and inequations, its geometric properties, and its construction. Subsection 2.2.2 builds on the algebraic techniques of the previous subsection and develops Thomas' algorithm for systems of differential equations and inequations. The combinatorics of Janet's algorithm (cf. Subsect. 2.1.1) are used here to ensure formal integrability for each simple system in the resulting Thomas decomposition. After defining and discussing the notion of the generic simple system of a Thomas decomposition of a prime (algebraic or differential) ideal in Subsect. 2.2.3, which will be an essential ingredient for the elimination methods in Sect. 3.3, the following subsection comments on the relationship of simple systems and other types of triangular sets and on the complexity of differential elimination. Subsection 2.2.5 introduces the generalized Hilbert series for simple differential systems. In the last subsection implementations of J. M. Thomas' ideas are discussed and references to related packages are given.



### 2.2.1 Simple Algebraic Systems

Let  $K$  be a computable field of characteristic zero and  $R := K[x_1, \dots, x_n]$  a commutative polynomial algebra with standard grading. We assume that the set  $\{x_1, \dots, x_n\}$  is totally ordered, without loss of generality

$$x_1 > x_2 > \dots > x_n,$$

and we denote by  $\bar{K}$  an algebraic closure of  $K$ .

This subsection presents the approach of J. M. Thomas [Tho37] transforming a given system of polynomial equations and inequations in  $x_1, \dots, x_n$ , defined over  $K$ , into a finite collection of so-called simple systems, each of which can in principle be solved recursively by determining roots of univariate polynomials according to the recursive structure of the solution set as finite-sheeted covering. In other words, the set  $V$  of solutions in  $\bar{K}^n$  of the given system is partitioned into finitely many subsets  $V_1, \dots, V_m$  in such a way that, for each  $i$ , the projection of the last  $k+1$  coordinates of  $V_i$  onto the last  $k$  coordinates has fibers of the same finite or co-finite cardinality (where the cardinality may depend on  $i$  and where  $k$  ranges from  $n-1$  down to 1).

The corresponding decomposition of differential systems (cf. Subsect. 2.2.2) is based on the decomposition of algebraic systems discussed here, but the algebraic part is interesting and of high value in itself.

In the present context we adopt a recursive representation of the elements of  $R = K[x_1, \dots, x_n]$  as follows.

**Definition 2.2.1.** For  $p \in R - K$  we denote by  $\text{ld}(p)$  the  $>$ -greatest variable such that  $p$  is a non-constant polynomial in that variable. According to standard terminology in differential algebra we call it the *leader* of  $p$  (although often *main variable* is also used when dealing with algebraic systems). The coefficient of the highest power of  $\text{ld}(p)$  occurring in  $p$  is called the *initial* of  $p$  and denoted by  $\text{init}(p)$ . Finally, the *discriminant* of  $p$  is defined in terms of the resultant of  $p$  and its partial derivative with respect to its leader as

$$\text{disc}(p) := (-1)^{d(d-1)/2} \cdot \text{res} \left( p, \frac{\partial p}{\partial \text{ld}(p)}, \text{ld}(p) \right) / \text{init}(p),$$

where  $d$  is the degree of  $p$  in  $\text{ld}(p)$  and  $\text{res}(p_1, p_2, x)$  denotes the resultant of the polynomials  $p_1$  and  $p_2$  with respect to the indeterminate  $x$ . Recall that the above resultant is divisible by  $\text{init}(p)$  because every entry of the first column of the Sylvester matrix of  $p$  and  $\partial p / \partial \text{ld}(p)$  is so. The discriminant of  $p$  is used to determine those values for the indeterminates smaller than  $\text{ld}(p)$  with respect to  $>$  for which  $p$  as a polynomial in  $\text{ld}(p)$  has zeros of multiplicity greater than one.

Every non-constant polynomial  $p \in R$  is now considered as univariate polynomial in  $\text{ld}(p)$ , whose coefficients are univariate polynomials in their leaders (if not constant) and so on, i.e., we regard  $R$  as  $K[x_n][x_{n-1}] \dots [x_1]$ .



**Definition 2.2.2.** Let

$$S = \{p_i = 0, q_j \neq 0 \mid i \in I, j \in J\}, \quad p_i, q_j \in R,$$

be a system of algebraic equations and inequations, where  $I$  and  $J$  are index sets. We define the *set of solutions* or *variety of  $S$  in  $\overline{K}^n$*  by

$$\text{Sol}_{\overline{K}}(S) := \{a \in \overline{K}^n \mid p_i(a) = 0, q_j(a) \neq 0 \text{ for all } i \in I, j \in J\}$$

( $a_1, \dots, a_n$  are substituted for  $x_1, \dots, x_n$ , respectively). For  $k \in \{0, 1, \dots, n-1\}$  let

$$\pi_k: \overline{K}^n \longrightarrow \overline{K}^{n-k}: (a_1, a_2, \dots, a_n) \longmapsto (a_{k+1}, a_{k+2}, \dots, a_n)$$

be the projection onto the last  $n-k$  components (i.e., the first  $k$  components are dropped).

**Remark 2.2.3.** By Hilbert's Basis Theorem (cf., e.g., [Eis95]), the index set  $I$  may be assumed to be finite without loss of generality. In general, the set of inequations cannot be replaced with an equivalent finite set of inequations. Since we aim at effective methods for dealing with algebraic systems, both index sets  $I$  and  $J$  will be assumed to be finite. The subsets of affine space  $\overline{K}^n$  which are of the form  $\text{Sol}_{\overline{K}}(S)$  for systems  $S$  of algebraic equations defined over  $\overline{K}$ , i.e.,  $J = \emptyset$ , are the closed sets of the *Zariski topology* on  $\overline{K}^n$ .

The notion of simple system, central for constructing partitions of varieties as proposed by J. M. Thomas, can now be defined using the projections  $\pi_k$  as follows.

**Definition 2.2.4.** A system  $S$  of algebraic equations and inequations

$$p_1 = 0, \quad \dots, \quad p_s = 0, \quad q_1 \neq 0, \quad \dots, \quad q_t \neq 0,$$

where  $p_1, \dots, p_s, q_1, \dots, q_t \in R - K$ ,  $s, t \in \mathbb{Z}_{\geq 0}$ , is said to be *simple* if the following three conditions are satisfied.

- a) The leaders of  $p_1, \dots, p_s, q_1, \dots, q_t$  are pairwise distinct.
- b) For every  $r \in \{p_1, \dots, p_s, q_1, \dots, q_t\}$ , if  $\text{ld}(r) = x_k$ , then the equation  $\text{init}(r) = 0$  has no solution in  $\pi_k(\text{Sol}_{\overline{K}}(S))$ .
- c) For every  $r \in \{p_1, \dots, p_s, q_1, \dots, q_t\}$ , if  $\text{ld}(r) = x_k$ , then the equation  $\text{disc}(r) = 0$  has no solution in  $\pi_k(\text{Sol}_{\overline{K}}(S))$ .

(In b) and c), we have  $\text{init}(r), \text{disc}(r) \in K[x_{k+1}, \dots, x_n]$ .)

**Remark 2.2.5.** A set of polynomials satisfying condition a) is called *triangular set* (cf., e.g., [Hub03a]). This condition implies that  $s + t \leq n$ .

Furthermore, a simple system  $S$  admits the following recursive solution procedure. We introduce the notations  $S_{< x_k}$  and  $S_{\leq x_k}$  for the subsets of  $S$  consisting of the equations and inequations with leader smaller than  $x_k$  and with leader smaller than or equal to  $x_k$ , respectively. For every  $k \in \{1, 2, \dots, n-1\}$ , every tuple

$$(a_{k+1}, a_{k+2}, \dots, a_n) \in \bar{K}^{n-k}$$

which is a solution of  $S_{<x_k}$  can be extended to a solution

$$(a_k, a_{k+1}, \dots, a_n) \in \bar{K}^{n-(k-1)}$$

of  $S_{\leq x_k}$ , and every solution  $a \in \bar{K}^n$  of  $S$  with projection  $\pi_k(a) = (a_{k+1}, a_{k+2}, \dots, a_n)$  is obtained through this process. The possible values of  $a_k$  are determined exactly by the equation or inequation in  $S$  with leader  $x_k$  if it exists, and  $a_k$  may take an arbitrary value in  $\bar{K}$  otherwise. Condition b) of Definition 2.2.4 implies that the degree of the equation or inequation in  $S$  in its leader  $x_k$ , if it exists, does not depend on the choice of the values  $a_{k+1}, a_{k+2}, \dots, a_n$  of  $x_{k+1}, x_{k+2}, \dots, x_n$ . The result of substituting  $x_{k+1} = a_{k+1}, \dots, x_n = a_n$  into the left hand side of the equation or inequation is a square-free polynomial by condition c). Therefore, the fibers of the projection of  $\pi_{k-1}(\text{Sol}_{\bar{K}}(S))$  onto  $\pi_k(\text{Sol}_{\bar{K}}(S))$  have the same finite or co-finite cardinality, which is given by the degree in  $x_k$  of the equation or inequation, respectively.

Geometrically speaking, the solution set of  $S$  is identified recursively as a branched covering. If the variety of interest has a non-trivial ramification locus as a branched covering, then the Thomas decomposition represents it as a partition into solution sets of several simple systems.

Before giving a precise definition and describing the algorithmic construction of a Thomas decomposition, we draw an algebraic consequence that will also be relevant for the differential case. First we recall the notion of vanishing ideal.

**Definition 2.2.6.** For any  $X \subseteq \bar{K}^n$  we define the *vanishing ideal of  $X$  in  $R$*  by

$$\mathcal{I}_R(X) := \{ p \in R \mid p(x) = 0 \text{ for all } x \in X \}.$$

It is a radical ideal of  $R = K[x_1, \dots, x_n]$ . By Hilbert's Nullstellensatz (cf., e.g., [Eis95]), the closed sets of the Zariski topology on  $\bar{K}^n$  are in one-to-one and inclusion-reversing correspondence with the radical ideals of  $\bar{K}[x_1, \dots, x_n]$ . Therefore,  $\text{Sol}_{\bar{K}}(\mathcal{I}_R(X))$  is the closure of  $X$  in  $\bar{K}^n$  with respect to the Zariski topology.

**Proposition 2.2.7.** *Let a simple algebraic system  $S$  over  $R$  be given by*

$$p_1 = 0, \quad \dots, \quad p_s = 0, \quad q_1 \neq 0, \quad \dots, \quad q_t \neq 0.$$

*Let  $E$  be the ideal of  $R$  which is generated by  $p_1, \dots, p_s$  and define  $q$  to be the product of all  $\text{init}(p_i)$ ,  $i = 1, \dots, s$ . Then we have the equality*

$$E : q^\infty := \{ p \in R \mid q^r \cdot p \in E \text{ for some } r \in \mathbb{Z}_{\geq 0} \} = \mathcal{I}_R(\text{Sol}_{\bar{K}}(S)).$$

*In particular,  $E : q^\infty$  is a radical ideal. A polynomial  $p \in R$  is an element of  $E : q^\infty$  if and only if the remainder of pseudo-reduction of  $p$  modulo  $p_1, \dots, p_s$  is zero.*

**Remark 2.2.8.** Since  $\text{Sol}_{\bar{K}}(E : q^\infty) = \text{Sol}_{\bar{K}}(\mathcal{I}_R(\text{Sol}_{\bar{K}}(S)))$  is the closure of  $\text{Sol}_{\bar{K}}(S)$  in  $\bar{K}^n$  with respect to the Zariski topology, the inequations  $q_1 \neq 0, \dots, q_t \neq 0$  do not figure on the left hand side of the equality asserted in Proposition 2.2.7.

*Proof (of Proposition 2.2.7).* From the discussion in Remark 2.2.5 it follows that  $\text{Sol}_{\overline{K}}(S)$  is not empty. Hence, the vanishing ideal  $\mathcal{J} := \mathcal{J}_R(\text{Sol}_{\overline{K}}(S))$  is contained in a maximal ideal of  $R$ , and, in particular, we have  $\mathcal{J} \cap K = \{0\}$ . The inclusion “ $\subseteq$ ” in the assertion of the proposition is clear. Moreover, if  $\mathcal{J} = \{0\}$ , then the reverse inclusion is also clear. Otherwise, let  $p \in \mathcal{J} - \{0\}$  and  $x_k := \text{ld}(p)$ . Let

$$(a_{k+1}, \dots, a_n) \in \overline{K}^{n-k}$$

be a solution of  $S_{<x_k}$  (possibly the empty tuple). As in Remark 2.2.5, this tuple can be extended to a solution

$$(a_k, a_{k+1}, \dots, a_n) \in \overline{K}^{n-(k-1)}$$

of  $S_{\leq x_k}$ . If  $S$  contains no equation with leader  $x_k$  or contains an inequation with that leader, then the set of possible  $a_k$  is infinite, which is a contradiction to the fact that the equation  $p = 0$  allows only  $\deg_{x_k}(p)$  values for  $a_k$ . Hence,  $S$  contains an equation  $p_i = 0$  with  $\text{ld}(p_i) = x_k$  and  $\deg_{x_k}(p_i) \leq \deg_{x_k}(p)$ . Now, pseudo-division of  $p$  modulo  $p_i$  (i.e., Euclidean division of  $c \cdot p$  modulo  $p_i$  for a suitable power  $c$  of  $\text{init}(p_i)$ ) yields a polynomial  $p'$  which is either zero or has smaller degree in  $x_k$  than  $p$  and which is an element of  $\mathcal{J}$ . Iteration of this argument shows that pseudo-reduction of  $p$  modulo equations in  $S$  yields the zero polynomial. Hence,  $p \in E : q^\infty$ , which proves the inclusion “ $\supseteq$ ”.  $\square$

**Remark 2.2.9.** The same argument as in the proof of Proposition 2.2.7 shows that the residue classes in  $R/(E : q^\infty)$  of the variables in  $\{x_1, \dots, x_n\}$  that are not leaders of an equation in a simple system  $S$  form a maximal subset of  $R/(E : q^\infty)$  that is algebraically independent over  $K$ . In other words, these residue classes form a system of parameters for the coordinate ring  $R/(E : q^\infty)$  of the Zariski closure  $V$  of  $\text{Sol}_{\overline{K}}(S)$ , in the sense that their number equals the dimension of the affine variety  $V$  and any choice of values for these “coordinates” defines a point on (one branch of) the variety.

**Definition 2.2.10.** Let

$$S = \{p_i = 0, q_j \neq 0 \mid i \in I, j \in J\}, \quad p_i, q_j \in R,$$

be a system of algebraic equations and inequations, where  $I$  and  $J$  are index sets and  $J$  is finite. A *Thomas decomposition* of  $S$  or of  $\text{Sol}_{\overline{K}}(S)$  is a finite collection of simple systems  $S_1, \dots, S_k$  such that

$$\text{Sol}_{\overline{K}}(S) = \text{Sol}_{\overline{K}}(S_1) \uplus \dots \uplus \text{Sol}_{\overline{K}}(S_k)$$

is a partition of  $\text{Sol}_{\overline{K}}(S)$ .

**Remark 2.2.11.** We outline *Thomas’ algorithm* (for algebraic systems), which computes a Thomas decomposition for any given system of finitely many algebraic equations and inequations (defined over the computable field  $K$ ) in finitely many steps. A more precise description will be given on pages 67–87.

First of all, systems containing an equation whose left hand side is a non-zero constant or an inequation with zero left hand side are inconsistent and will be discarded. On the other hand, an equation with zero left hand side and an inequation whose left hand side is a non-zero constant are supposed to be removed from each system. In what follows, we therefore assume that the left hand side of every equation and inequation is a non-constant polynomial.

According to the recursive representation of polynomials, Euclidean pseudo-division is applied to (the left hand sides of) pairs of distinct equations with the same leader, i.e., if  $p_1 = 0$ ,  $p_2 = 0$  are distinct equations of the system satisfying  $\text{ld}(p_1) = \text{ld}(p_2) =: x$  and  $\deg_x(p_1) \geq \deg_x(p_2)$ , then usual Euclidean division is performed on  $c \cdot p_1$  modulo  $p_2$ , where the polynomial  $c$  is chosen as (a suitable power of) the initial of  $p_2$  such that division without fractions is made possible.

Let the result of the pseudo-division be  $p_3$ . When  $p_1 = 0$  is replaced with  $p_3 = 0$  in the system  $S$  under consideration, a sufficient condition for the set of solutions of  $S$  to be unaltered is that  $c$  does not vanish for any solution of  $S$ . In order to guarantee that the solution set is not changed, the algorithm actually replaces  $S$  with two systems  $S'$  and  $S''$  and continues to work with  $S'$  and  $S''$  separately in the same way as it did with  $S$ . The systems  $S'$  and  $S''$  are obtained from  $S$  by replacing  $p_1 = 0$  with  $p_3 = 0$  and inserting the inequation  $c \neq 0$  in case of  $S'$  and the equation  $c = 0$  in case of  $S''$ .

For each pair  $p = 0$ ,  $q \neq 0$  in  $S$  with  $\text{ld}(p) = \text{ld}(q)$ , the greatest common divisor<sup>10</sup>  $r$  of  $p$  and  $q$  is computed. To this end, pseudo-divisions are performed, assuming that the initials of the divisors do not vanish<sup>11</sup>, which possibly generates new case distinctions. If  $q$  is a multiple of  $p$ , then  $S$  is inconsistent and will be discarded. If  $r$  is a non-zero constant, then  $q \neq 0$  is removed from  $S$ . Otherwise,  $p = 0$  is replaced with  $p/r = 0$ .

If  $q_1 \neq 0$ ,  $q_2 \neq 0$  are two inequations in  $S$  with  $\text{ld}(q_1) = \text{ld}(q_2)$ , then these are replaced with  $q_3 \neq 0$ , where  $q_3$  is the least common multiple of  $q_1$  and  $q_2$ . The computation of the least common multiple involves pseudo-divisions and case distinctions according to vanishing of initials as above.

In the same way as Euclid's algorithm terminates with a single polynomial (being the greatest common divisor of the input polynomials), after finitely many steps the systems produced by Thomas' algorithm will be triangular sets (i.e., condition a) in Def. 2.2.4 will be satisfied), and initials of equations and inequations of each system will not vanish for any solution of the respective system (condition b)). Condition c) in Def. 2.2.4 is accomplished as follows. Since the field of definition  $K$  is of characteristic zero, the square-free part of a non-constant polynomial  $r$  can be determined as quotient of  $r$  by the greatest common divisor of  $r$  and the partial derivative of  $r$

<sup>10</sup> The terms *greatest common divisor* and *least common multiple* should actually be used with care here because the coefficients of the polynomials in question will be considered subject to equations and inequations with smaller leader so that these notions may not be uniquely defined. For a more precise description, we refer to pages 67–87.

<sup>11</sup> Using subresultant polynomial remainder sequences (cf., e.g., [Mis93]) to compute greatest common divisors often reduces the growth of initials and therefore the number of case distinctions. For more details, cf. [BGL<sup>+</sup>12, Sect. 2].

with respect to its leader. Again, coefficients of  $r$  must be handled with care, and computation of this greatest common divisor usually involves case distinctions. By equating some element of the polynomial remainder sequence with zero, the possible cases for the greatest common divisor are dealt with separately, which in general produces new systems to be treated again in the same way as above.

There are a number of possible ways how to combine these steps. One strategy is to deal in each system with the polynomials of least leader first. For each variable  $x$  at most one equation or inequation with leader  $x$  is registered which is guaranteed to have non-vanishing initial in the above sense, where equations are preferred to inequations. The next equation or inequation in the current algebraic system to be processed is reduced modulo the registered equations. If the resulting left hand side is not a constant and if an equation or inequation with the same leader is registered, then this pair is treated as discussed above. Splittings of systems regarding initials and square-free parts result in new equations and inequations with smaller leader. Since a registered equation is only replaced with an equation of smaller degree (in the same leader) and since inequations are replaced with equations if possible or with the least common multiple of inequations with the same leader, this strategy terminates after finitely many steps.

The result of the algorithm is a Thomas decomposition of the given algebraic system. It depends on the chosen ordering of the variables  $x_1, \dots, x_n$  and on the order in which the steps of Thomas' algorithm are carried out. Moreover, polynomial factorization of left hand sides of equations is often favorable because proper factors lead to a splitting of the system into smaller systems, each of which is obtained by replacing the original equation with one of its factors of smaller degree.

Thomas' algorithm returns an empty result if and only if no solution (defined over  $\overline{K}$ ) exists for the input system. The result being  $\{\emptyset\}$  (i.e., a set consisting of one empty system) is equivalent to the solution set being  $\overline{K}^n$ .

**Example 2.2.12.** [BGL<sup>+</sup>12, Ex. 2.5] Let us examine

$$ax^2 + bx + c = 0, \quad (2.27)$$

a quadratic equation in  $x$  with parameters  $a, b, c$ . In order to discuss the well-known types of solution sets (in an algebraic closure of  $\mathbb{Q}$  or in  $\mathbb{C}$ ) such an equation can have, we consider the left hand side  $p$  of (2.27) as element of  $\mathbb{Q}[x, a, b, c]$ , where  $x > c > b > a$ , and apply Thomas' algorithm to this algebraic system.

The initial of  $p$  equals  $a$ . The given system is therefore replaced with

$$S_1 := \{p = 0, a \neq 0\}, \quad S_2 := \{p = 0, a = 0\}.$$

Conditions a) and b) in Definition 2.2.4 are already satisfied for  $S_1$ . Euclid's algorithm applied to  $p$  and  $\frac{\partial p}{\partial \text{ld}(p)}$  (as polynomials in  $\text{ld}(p) = x$ ) computes the polynomial remainder sequence

$$p, \quad \frac{\partial p}{\partial \text{ld}(p)}, \quad 4ac - b^2.$$

Multiplication by  $a$  for pseudo-division is harmless because  $a$  is assumed not to vanish. Note that the last polynomial equals the discriminant of  $p$  (up to sign). Therefore, we replace  $S_1$  with

$$S_{1,1} := \{p = 0, 4ac - b^2 \neq 0, a \neq 0\}, \quad S_{1,2} := \{2ax + b = 0, 4ac - b^2 = 0, a \neq 0\},$$

where  $2ax + b$  is the square-free part of  $p$  in case  $4ac - b^2 = 0$ . These two systems are simple.

On the other hand,  $S_2$  is not a triangular set. Euclidean division simplifies  $p = 0$  to  $bx + c = 0$ , whose initial equals  $b$ . Thus  $S_2$  is split into two systems

$$S_{2,1} := \{bx + c = 0, b \neq 0, a = 0\}, \quad S_{2,2} := \{bx + c = 0, b = 0, a = 0\},$$

which are easily dealt with. The final result is given by the following four simple systems, where leaders of polynomials are underlined, where not obvious:

$\underline{a}\underline{x}^2 + \underline{b}\underline{x} + c = 0$ $4a\underline{c} - b^2 \neq 0$ $a \neq 0$	$2a\underline{x} + b = 0$ $4a\underline{c} - b^2 = 0$ $a \neq 0$	$\underline{b}\underline{x} + c = 0$ $b \neq 0$ $a = 0$	$c = 0$ $b = 0$ $a = 0$
---	--	---	-------------------------

We give another example, which shows that an algebraic system may be simple, although it contains no inequations.

**Example 2.2.13.** Let  $R = \mathbb{Q}[x, y]$  and  $x > y$ . Then

$$(y + 1)x = 0, \quad y(y - 1) = 0$$

is a simple algebraic system  $S$  over  $R$ . Using the factorization of the second equation, a splitting of this system into

$$\{(y + 1)x = 0, y = 0\}, \quad \{(y + 1)x = 0, y - 1 = 0\}$$

makes further reductions possible, which results in another Thomas decomposition

$$\{x = 0, y = 0\}, \quad \{x = 0, y = 1\}$$

of the same system  $S$ . Using the factorization of the first equation yields the same answer after removing inconsistent systems.

**Remark 2.2.14.** Let us assume that a system  $S$  of algebraic equations and inequations, defined over  $\mathbb{Z}$ , is given. A variant of Thomas' algorithm (neglecting square-freeness) allows to compute a finite collection of systems from which partitions of the solution sets  $\text{Sol}_{\overline{\mathbb{Q}}}(S)$  and  $\text{Sol}_{\overline{\mathbb{F}_p}}(S)$  can be extracted for algebraic closures  $\overline{\mathbb{Q}}$  of  $\mathbb{Q}$  and  $\overline{\mathbb{F}_p}$  of  $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ , where  $p$  is a prime number. To this end, vanishing of

initials has to be checked also if these are integers, and division by non-invertible integers must be prevented. This leads to new splittings, in particular when the greatest common divisor of two polynomials is an integer of absolute value at least 2. For instance, a system could be split into two systems which include a new equation  $6 = 0$  and a new inequation  $6 \neq 0$ , respectively. Integer factorization can be used to split the first system again.

**Example 2.2.15.** We consider the system of algebraic equations

$$y\underline{x}^2 - \underline{x} + 1 = 0, \quad y^2\underline{x} - y^3 + 2 = 0, \quad \underline{x} + y = 0,$$

which is defined over  $\mathbb{Z}$  and where  $x > y$ . Euclidean division of the first and the second modulo the third polynomial yields

$$-2\underline{y}^3 + 2 = 0, \quad \underline{y}^3 + \underline{y} + 1 = 0, \quad \underline{x} + y = 0. \quad (2.28)$$

The result of applying Euclidean division to the first polynomial modulo the second one is  $2y + 4$ . In order to be able to replace the second polynomial by its pseudo-remainder modulo the new polynomial without changing the solution set of the system, we assume that  $2 \neq 0$  holds. Then the pseudo-division yields  $18 = 0$ , which is equivalent to  $9 = 0$ . Since we consider the solutions in an algebraic closure of a field  $\mathbb{F}_p$ , the final result in this case is

$$\underline{x} + 1 = 0, \quad \underline{y} + 2 = 0, \quad 3 = 0.$$

If  $2 = 0$ , only the second and third equation in (2.28) remain, and the final result in this case is

$$\underline{x} + y = 0, \quad \underline{y}^3 + \underline{y} + 1 = 0, \quad 2 = 0.$$

For a different approach to decomposing algebraic systems into simple systems in positive characteristic, cf. [LMW10, MLW13].

We finish this subsection by giving a more precise description of the algebraic part of Thomas' algorithm, ignoring, however, efficiency issues. The total ordering  $>$  on the set of indeterminates  $\{x_1, \dots, x_n\}$  of  $R$  is part of the input. It determines the leader of each non-constant polynomial in  $x_1, \dots, x_n$ .

**Definition 2.2.16.** Let  $p \in R$ ,  $q \in R - K$ , and  $G \subseteq R - K$ .

- The polynomial  $p$  is said to be *reduced with respect to  $q$*  if  $p \in K$  or if  $p \in R - K$  and we have  $\deg_v(p) < \deg_v(q)$  for  $v := \text{ld}(q)$ .
- The polynomial  $p$  is said to be *reduced with respect to  $G$*  if  $p \in K$  or if  $p \in R - K$  and  $p$  is reduced with respect to each element of  $G$ , and if each coefficient of  $p$  (as a polynomial in its leader) is reduced with respect to each element of  $G$ .
- An equation or inequation (with zero right hand side) is said to be *reduced with respect to  $q$*  or *reduced with respect to  $G$*  if its left hand side is so.

Given a polynomial  $r$  in  $R$  and a finite set  $G$  of non-constant polynomials in  $R$ , the following algorithm subtracts from  $r$  suitable multiples of polynomials in  $G$  with the same leader as  $r$  until the result is reduced with respect to each polynomial in  $G$  with that leader. It treats the coefficients of the result, which are polynomials with smaller leader, if not constant, in the same way. This recursive reduction is essential for the description of Thomas' algorithm below because we suppose that the highest term of the left hand side of  $p = 0$  (or of  $p \neq 0$ ) will be canceled if  $\text{init}(p) = 0$  is an equation of the same algebraic system. However, the reduction of coefficients of terms of lower degree could be omitted (which would require an adaptation of Definition 2.2.16).

**Algorithm 2.2.17** (*Reduce*).

**Input:**  $r \in R$ ,  $G = \{p_1, p_2, \dots, p_s\} \subseteq R - K$ , and a total ordering  $>$  on  $\{x_1, \dots, x_n\}$

**Output:**  $r' \in R$  and an element  $b$  of the multiplicatively closed set generated by  $\bigcup_{i=1}^s \{\text{init}(p_i)\} \cup \{1\}$  such that  $r'$  is reduced with respect to  $G$ , and such that  $r' = r$ ,  $b = 1$  if  $G = \emptyset$ , and  $r' + \langle p_1, \dots, p_s \rangle = b \cdot r + \langle p_1, \dots, p_s \rangle$  otherwise

**Algorithm:**

- 1:  $r' \leftarrow r$
- 2:  $b \leftarrow 1$
- 3: **if**  $r' \notin K$  **then**
- 4:    $v \leftarrow \text{ld}(r')$
- 5:   **while**  $r' \notin K$  and there exists  $p \in G$  with  $\text{ld}(p) = v$ ,  $\deg_v(r') \geq \deg_v(p)$  **do**
- 6:      $r' \leftarrow \text{init}(p) \cdot r' - \text{init}(r') \cdot v^{d-d'} \cdot p$ , where  $d := \deg_v(r')$  and  $d' := \deg_v(p)$
- 7:      $b \leftarrow \text{init}(p) \cdot b$
- 8:   **end while**
- 9:   **while** there exists a coefficient  $c$  of  $r'$  (as a polynomial in  $v$ ) which is not reduced with respect to  $G$  **do**
- 10:      $(r'', b') \leftarrow \text{Reduce}(c, G, >)$
- 11:     replace the coefficient  $b' \cdot c$  in  $b' \cdot r'$  with  $r''$  and replace  $r'$  with this result
- 12:      $b \leftarrow b' \cdot b$
- 13:   **end while**
- 14: **end if**
- 15: **return**  $(r', b)$

**Remarks 2.2.18.** a) The loop in steps 5–8 ensures that  $r'$  is reduced with respect to each  $p \in G$  with  $\text{ld}(p) = v$ . Termination of Algorithm 2.2.17 follows from the facts that the coefficients  $c$  which are dealt with recursively in step 10 are either constant or have leaders which are smaller than  $v$  with respect to  $>$  and that the property of  $r'$  which is achieved by the loop in steps 5–8 is retained by the recursion. The asserted equation follows recursively from the updates of  $b$ . Note that in general, if  $b \neq 1$ , then  $r$  and  $r'$  are not in the same residue class of  $R/(\langle p_1, \dots, p_s \rangle : q^\infty)$  (cf. also the following example).



b) Let  $r_1, r_2 \in R$  and  $G = \{p_1, p_2, \dots, p_s\}$  be as in the input of Algorithm 2.2.17, and define  $q$  to be the product of all  $\text{init}(p_i), i = 1, \dots, s$ . In general, the equality

$$r_1 + \langle p_1, \dots, p_s \rangle : q^\infty = r_2 + \langle p_1, \dots, p_s \rangle : q^\infty$$

does not imply that the results of applying *Reduce* to  $r_1$  and  $r_2$ , respectively, are equal. However, Proposition 2.2.7 shows that, if  $p_1 = 0, p_2 = 0, \dots, p_s = 0$  are the equations of a simple algebraic system, then the result  $r'$  of applying *Reduce* to  $r_1$  is zero if and only if we have  $r_1 \in \langle p_1, \dots, p_s \rangle : q^\infty$ .

**Example 2.2.19.** Let  $R = \mathbb{Q}[x, y]$  and  $x > y$ . Then

$$yx - 1 = 0, \quad y \neq 0$$

is a simple algebraic system over  $R$ . Algorithm 2.2.17 (*Reduce*) applied to  $r := x, G := \{yx - 1\}$ , and  $>$  computes

$$r' := yr - (yx - 1) = 1,$$

and the output is  $(r', b) = (1, y)$ . Note that  $r$  and  $r'$  are not in the same residue class of  $R/\langle yx - 1 \rangle : q^\infty$ , where  $q := y$ , but  $b \cdot r$  and  $r'$  are. Moreover, the result of applying *Reduce* to  $yx$ , which is in the same residue class as 1, is  $(y, y)$ . Hence, for different representatives of the same residue class, the first component of the output of *Reduce* may be different in general.

The following description of the algebraic part of Thomas' algorithm deals with triples  $(L, M, N)$  of finite algebraic systems over  $R$  which are gathered in a set  $Q$ . Initially this set contains only the triple  $(S, \emptyset, \emptyset)$ , where  $S$  is the input system, more triples will usually be inserted into  $Q$  as such triples are processed, and after finitely many steps the set  $Q$  will be empty. Another set  $T$  collects the simple algebraic systems of the Thomas decomposition to be constructed.

The second and third component of every triple have the following properties throughout the algorithm. The left hand side  $p$  of every equation and inequation in  $M$  is non-constant and  $\text{init}(p) \neq 0$  holds if a solution of the algebraic system  $L \cup M \cup N$  is substituted for  $x_1, \dots, x_n$ . Similarly, the left hand side  $p$  of every equation and inequation in  $N$  is non-constant and both  $\text{init}(p) \neq 0$  and  $\text{disc}(p) \neq 0$  hold if a solution of  $L \cup M \cup N$  is substituted for  $x_1, \dots, x_n$ . Moreover, for every  $v \in \{x_1, \dots, x_n\}$ ,  $M \cup N$  contains at most one equation or inequation with leader  $v$ .

For any algebraic system

$$S = \{p_i = 0, q_j \neq 0 \mid i \in I, j \in J\}, \quad p_i, q_j \in R,$$

where  $I$  and  $J$  are index sets, we denote by

$$S^\equiv := \{p_i \mid i \in I\}$$

the set of left hand sides of equations in  $S$ .

**Algorithm 2.2.20** (*AlgebraicThomasDecomposition*).**Input:** A finite algebraic system  $S$  over  $R$  and a total ordering  $>$  on  $\{x_1, \dots, x_n\}$ **Output:** A Thomas decomposition of  $S$ **Algorithm:**

```

1:  $Q \leftarrow \{(S, \emptyset, \emptyset)\}$ 
2:  $T \leftarrow \emptyset$ 
3: repeat
4:   choose  $(L, M, N) \in Q$  and remove  $(L, M, N)$  from  $Q$ 
5:   replace the left hand side  $p$  of each equation and inequation in  $L$  with the first
     entry of the result of  $Reduce(p, M^\perp \cup N^\perp, >)$  // cf. Alg. 2.2.17
6:   remove  $0 = 0$  and  $p \neq 0$  from  $L$  for any  $p \in K - \{0\}$ 
7:   if  $L$  does neither contain  $p = 0$  with  $p \in K - \{0\}$  nor  $0 \neq 0$  then
8:     if  $L = \emptyset$  then
9:       if  $M = \emptyset$  then
10:        insert  $N$  into  $T$ 
11:       else
12:          $Q \leftarrow ProcessDiscriminant((L, M, N), Q, >)$  // cf. Alg. 2.2.23
13:       end if
14:     else
15:        $Q \leftarrow ProcessInitial((L, M, N), Q, >)$  // cf. Alg. 2.2.21
16:     end if
17:   end if
18: until  $Q = \emptyset$ 
19: return  $T$ 

```

The proof that Algorithm 2.2.20 terminates and is correct will be given after the description of the algorithms on which it depends (cf. Thm. 2.2.32, p. 79).

The following terminology will be useful for the rest of this subsection. For a triple  $(L, M, N)$  of algebraic systems over  $R$  we refer to  $Sol_{\overline{K}}(L \cup M \cup N)$  as the *solution set* of the triple  $(L, M, N)$ , and for a set  $Q$  of such triples we denote by

$$Sol_{\overline{K}}(Q) := \bigcup_{(L, M, N) \in Q} Sol_{\overline{K}}(L \cup M \cup N)$$

the union of the solution sets of all triples in  $Q$ .

Moreover, let

$$S = \{p_i = 0, q_j \neq 0 \mid i \in I, j \in J\}, \quad p_i, q_j \in R,$$

be an algebraic system, where no  $p_i$  and no  $q_j$  is constant, and let  $v \in \{x_1, \dots, x_n\}$ . Then  $S_{\geq v}$  (resp.  $S_{< v}$ ) is a notation for the subset of  $S$  which consists of the equations and inequations with leader greater than or equal to (resp. smaller than)  $v$  with respect to  $>$ . We are also going to write  $S_{< v}^\perp$  instead of  $(S_{< v})^\perp$  (in Remarks 2.2.26).

**Algorithm 2.2.21** (*ProcessInitial*).

**Input:** A triple  $(L, M, N)$  of finite algebraic systems over  $R$ , a finite set  $P$  of such triples, and a total ordering  $>$  on  $\{x_1, \dots, x_n\}$ , where  $L \neq \emptyset$ , the left hand sides of elements of  $L \cup M \cup N$  are non-constant, those of  $M \cup N$  having pairwise distinct leaders, those of  $L$  being reduced with respect to  $M \cup N =$  (cf. Def. 2.2.16 b)), where  $\text{Sol}_{\bar{K}}(L \cup M \cup N)$  and the solution sets of triples in  $P$  are pairwise disjoint

**Output:** A finite set  $Q \supseteq P$  of triples as in  $P$  whose solution sets form a partition of  $\text{Sol}_{\bar{K}}(L \cup M \cup N) \uplus \text{Sol}_{\bar{K}}(P)$  such that either

- a) each triple in  $Q - P$  has the property that all of its solutions satisfy  $\text{init}(p) \neq 0$  or all of its solutions satisfy  $\text{init}(p) = 0$ , where  $p$  is the left hand side of the equation or inequation in  $L$  chosen in step 2, or
- b) the triples in  $Q - P$  have been inserted by Algorithm 2.2.27 (*LCMSplit*)

**Algorithm:**

```

1:  $Q \leftarrow P$ 
2: among the elements of  $L$  with least possible leader  $v$  with respect to  $>$  choose
   one with left hand side  $p$  of least possible degree in  $v$ , preferably an equation
3: if the equation  $p = 0$  is chosen then
4:   insert  $((L - \{p = 0\}) \cup M_{\geq v} \cup N_{\geq v} \cup \{\text{init}(p) \neq 0\},$ 
       $(M - M_{\geq v}) \cup \{p = 0\}, N - N_{\geq v})$  into  $Q$ 
5:   insert  $(L \cup \{\text{init}(p) = 0\}, M, N)$  into  $Q$ 
6: else // the inequation  $p \neq 0$  is chosen
7:   if  $M \cup N$  contains an equation  $q = 0$  with  $\text{ld}(q) = v$  then
8:      $Q \leftarrow \text{GCDSplit}(q, p, (L, M, N), Q, >)$  // cf. Alg. 2.2.25
9:   else if  $M \cup N$  contains an inequation  $q \neq 0$  with  $\text{ld}(q) = v$  then
10:    if  $\deg_v(p) \geq \deg_v(q)$  then
11:       $Q \leftarrow \text{LCMSplit}(p, q, (L, M, N), Q, >)$  // cf. Alg. 2.2.27
12:    else
13:      insert  $((L - \{p \neq 0\}) \cup \{q \neq 0, \text{init}(p) \neq 0\},$ 
         $(M - \{q \neq 0\}) \cup \{p \neq 0\}, N - \{q \neq 0\})$  into  $Q$ 
14:      insert  $(L \cup \{\text{init}(p) = 0\}, M, N)$  into  $Q$ 
15:    end if
16:  else
17:    insert  $((L - \{p \neq 0\}) \cup \{\text{init}(p) \neq 0\}, M \cup \{p \neq 0\}, N)$  into  $Q$ 
18:    insert  $(L \cup \{\text{init}(p) = 0\}, M, N)$  into  $Q$ 
19:  end if
20: end if
21: return  $Q$ 

```

**Remark 2.2.22.** Termination of Algorithm 2.2.21 follows from the fact that Algorithm 2.2.25 and Algorithm 2.2.27 terminate (cf. Lemma 2.2.28 and Lemma 2.2.29). An inspection of steps 4, 5, 13, 14, 17, and 18 and of the specifications of Algorithms 2.2.25 and 2.2.27 shows that the solution sets of triples in  $Q$  form a partition of  $\text{Sol}_{\bar{K}}(L \cup M \cup N) \uplus \text{Sol}_{\bar{K}}(P)$ . In order to show the last assertion stated in the description of the output, we observe that in each of these steps as well as in steps 6 and 7 in Algorithm 2.2.25, where  $r_{i+2}$  is equal to  $p_2$  in the first round of the loop, either the inequation  $\text{init}(p) \neq 0$  or the equation  $\text{init}(p) = 0$  is imposed.

**Algorithm 2.2.23** (*ProcessDiscriminant*).

**Input:**  $(L, M, N)$ ,  $P$ , and  $>$  with the same specification as in Algorithm 2.2.21 and satisfying  $L = \emptyset$  and  $M \neq \emptyset$

**Output:** A finite set  $Q \supseteq P$  of triples as in  $P$  whose solution sets form a partition of  $\text{Sol}_{\bar{K}}(M \cup N) \uplus \text{Sol}_{\bar{K}}(P)$  such that either

- a) each triple in  $Q - P$  has the property that all solutions satisfy  $\text{disc}(p) \neq 0$ , where  $p$  is the left hand side of the equation or inequation in  $M$  with least leader with respect to  $>$ , or
- b) the triples in  $Q - P$  have been inserted by Algorithm 2.2.30 (*SquarefreeSplit*)

**Algorithm:**

```

1:  $Q \leftarrow P$ 
2: let  $p = 0$  or  $p \neq 0$  be the equation or inequation in  $M$  with least leader with
   respect to  $>$  and let  $v$  be its leader
3: if  $\deg_v(p) = 1$  then
4:   if  $M$  contains  $p = 0$  then
5:     insert  $(\emptyset, M - \{p = 0\}, N \cup \{p = 0\})$  into  $Q$ 
6:   else //  $M$  contains  $p \neq 0$ 
7:     insert  $(\emptyset, M - \{p \neq 0\}, N \cup \{p \neq 0\})$  into  $Q$ 
8:   end if
9: else
10:   $Q \leftarrow \text{SquarefreeSplit}(p, (\emptyset, M, N), Q, >)$  // cf. Alg. 2.2.30
11: end if
12: return  $Q$ 

```

**Remark 2.2.24.** Termination of Algorithm 2.2.23 follows from the fact that Algorithm 2.2.30 terminates (cf. Lemma 2.2.31). It is easily checked by considering steps 5 and 7 and the specification of Algorithm 2.2.30 that the solution sets of triples in  $Q$  form a partition of  $\text{Sol}_{\bar{K}}(M \cup N) \uplus \text{Sol}_{\bar{K}}(P)$ . The last assertion which is stated in the description of the output is shown as follows. A solution of a triple in  $Q - P$  satisfies  $\text{disc}(p) = 0$  if and only if the univariate polynomial  $\bar{p}$  which is obtained by substituting this solution for  $x_1, \dots, x_n$  except  $\text{ld}(p)$  in  $p$  has multiple roots. But in steps 5 and 7 the polynomial  $\bar{p}$  has degree one.

**Algorithm 2.2.25** (*GCDSplit*).

**Input:**  $p_1, p_2 \in R - K$  with the same leader  $v$  and  $(L, M, N), P, >$  with the same specification as in Algorithm 2.2.21, where  $p_1 = 0$  is in  $M \cup N$ ,  $p_2 \neq 0$  is in  $L$ ,  $\deg_v(p_1) \geq \deg_v(p_2)$ , and  $p_2$  is reduced with respect to  $M^\# \cup N^\#$

**Output:** A finite set  $Q \supseteq P$  of triples as in  $P$  whose solution sets form a partition of  $\text{Sol}_{\bar{K}}(L \cup M \cup N) \uplus \text{Sol}_{\bar{K}}(P)$  such that for each triple in  $Q - P$  we have either

- a) the polynomials  $\bar{p}_1$  and  $\bar{p}_2$  which are obtained from  $p_1$  and  $p_2$  by substituting a solution of the triple for  $x_1, \dots, x_n$  except  $v$  have a greatest common divisor whose degree does not depend on the choice of the solution of the triple, or
- b) the triple has been inserted in step 6

**Algorithm:**

```

1:  $Q \leftarrow P; U \leftarrow \emptyset$ 
2:  $v \leftarrow \text{ld}(p_1); i \leftarrow 0$ 
3:  $r_1 \leftarrow p_1; c_1 \leftarrow 0$ 
4:  $r_2 \leftarrow p_2; c_2 \leftarrow 1$ 
5: repeat
6:   insert  $(L \cup \{\text{init}(r_{i+2}) = 0\} \cup U, M, N)$  into  $Q$ 
7:    $U \leftarrow U \cup \{\text{init}(r_{i+2}) \neq 0\}$ 
8:    $i \leftarrow i + 1$ 
9:    $r_{i+2} \leftarrow a_i \cdot r_i - q_i \cdot r_{i+1}$ , where  $a_i$  is a power of  $\text{init}(r_{i+1})$  and  $q_i \in R$  such that
      $r_{i+2} = 0$  or  $\deg_v(r_{i+2}) < \deg_v(r_{i+1})$ 
10:   $(r_{i+2}, b_{i+2}) \leftarrow \text{Reduce}(r_{i+2}, M^\# \cup N^\#, >)$  // cf. Alg. 2.2.17
11:   $c_{i+2} \leftarrow b_{i+2} \cdot (a_i \cdot c_i + q_i \cdot c_{i+1})$ 
12: until  $r_{i+2} = 0$  or  $\deg_v(r_{i+2}) = 0$ 
13: insert  $(L \cup \{c_{i+2} = 0, r_{i+2} = 0\} \cup U, M - \{p_1 = 0\}, N - \{p_1 = 0\})$  into  $Q$ 
14: insert  $((L - \{p_2 \neq 0\}) \cup \{r_{i+2} \neq 0\} \cup U, M, N)$  into  $Q$ 
15: return  $Q$ 

```

The proof that Algorithm 2.2.25 terminates and is correct (cf. Lemma 2.2.28) is based on the following remarks.

- Remarks 2.2.26.** a) The triples which are inserted into  $Q$  in steps 6, 13, and 14 in Algorithm 2.2.25 define inconsistent algebraic systems if  $\text{init}(r_{i+2})$  in step 6 or  $r_{i+2}$  in step 13 is a non-zero constant or if  $r_{i+2}$  is the zero polynomial in step 14. These triples should be discarded right away. For the sake of conciseness these case distinctions are omitted here.
- b) Since we have  $r_1 = p_1$  and  $\deg_v(r_{i+2}) < \deg_v(r_{i+1})$  after step 9 and since  $p_1 = 0$  is the unique equation with leader  $v$  in  $M \cup N$ , the reduction in step 10 considers only left hand sides of equations with leader smaller than  $v$  as pseudo-divisors.

- c) Algorithm 2.2.25 is a variant of Euclid's Algorithm with bookkeeping, where (coefficients of) intermediate results are also reduced with respect to  $M_{<v}^{\equiv} \cup N_{<v}^{\equiv}$ . Steps 9–11 ensure that the following congruence holds for all  $i \in \mathbb{Z}_{\geq 0}$ :

$$c_{i+2} \cdot r_{i+1} \equiv \left( \prod_{j=1}^i a_j \right) \cdot \left( \prod_{k=3}^{i+2} b_k \right) \cdot p_1 - c_{i+1} \cdot r_{i+2} \pmod{\langle M_{<v}^{\equiv} \cup N_{<v}^{\equiv} \rangle}. \quad (2.29)$$

Its significance derives from the following special case. If, for all  $i \in \mathbb{Z}_{\geq 0}$ , both sides are not merely congruent modulo  $\langle M_{<v}^{\equiv} \cup N_{<v}^{\equiv} \rangle$ , but equal, and if  $i$  is minimal with the property that  $r_{i+2}$  is the zero polynomial, then  $r_{i+1}$  is the greatest common divisor of  $p_1$  and  $p_2$  in  $\text{Quot}(K[x \mid v > x])[v]$ , where we denote by  $\text{Quot}(K[x \mid v > x])$  the field of fractions of the polynomial ring  $K[x \mid v > x]$ . Then  $c_{i+2}$  is the quotient of  $a_1 \cdot a_2 \cdot \dots \cdot a_i \cdot b_3 \cdot b_4 \cdot \dots \cdot b_{i+2} \cdot p_1$  divided by  $r_{i+1}$ .

We prove (2.29) by induction on  $i$ . Indeed, for  $i = 0$  we have  $c_2 \cdot r_1 = p_1$  by steps 3 and 4 (where an empty product is equal to 1 by convention). Let  $i > 0$ . After step 11 we have

$$\left. \begin{aligned} r_{i+2} &\equiv b_{i+2} \cdot (a_i \cdot r_i - q_i \cdot r_{i+1}) \pmod{\langle M_{<v}^{\equiv} \cup N_{<v}^{\equiv} \rangle}, \\ c_{i+2} &= b_{i+2} \cdot (a_i \cdot c_i + q_i \cdot c_{i+1}). \end{aligned} \right\} \quad (2.30)$$

The induction hypothesis states that we have

$$c_{i+1} \cdot r_i \equiv \left( \prod_{j=1}^{i-1} a_j \right) \cdot \left( \prod_{k=3}^{i+1} b_k \right) \cdot p_1 - c_i \cdot r_{i+1} \pmod{\langle M_{<v}^{\equiv} \cup N_{<v}^{\equiv} \rangle}. \quad (2.31)$$

Using (2.30) and (2.31), we deduce

$$\begin{aligned} c_{i+2} r_{i+1} &\equiv b_{i+2} (a_i c_i + q_i c_{i+1}) r_{i+1} \\ &\equiv b_{i+2} a_i c_i r_{i+1} + c_{i+1} b_{i+2} q_i r_{i+1} \\ &\equiv b_{i+2} a_i c_i r_{i+1} + c_{i+1} (b_{i+2} a_i r_i - r_{i+2}) \\ &\equiv b_{i+2} a_i c_i r_{i+1} - c_{i+1} r_{i+2} + b_{i+2} a_i \left( \left( \prod_{j=1}^{i-1} a_j \right) \left( \prod_{k=3}^{i+1} b_k \right) p_1 - c_i r_{i+1} \right) \\ &\equiv \left( \prod_{j=1}^i a_j \right) \left( \prod_{k=3}^{i+2} b_k \right) p_1 - c_{i+1} r_{i+2} \end{aligned}$$

modulo  $\langle M_{<v}^{\equiv} \cup N_{<v}^{\equiv} \rangle$ , which proves (2.29).

Similarly, if we set  $d_2 := 0$  and  $d_3 := 1$  and update, if  $i > 1$ ,

$$d_{i+2} \leftarrow b_{i+2} \cdot (a_i \cdot d_i + q_i \cdot d_{i+1})$$

after step 11, then the following congruence holds for all  $i \in \mathbb{Z}_{\geq 1}$ :

$$d_{i+2} \cdot r_{i+1} \equiv \left( \prod_{j=2}^i a_j \right) \cdot \left( \prod_{k=4}^{i+2} b_k \right) \cdot p_2 - d_{i+1} \cdot r_{i+2} \pmod{\langle M_{<v}^= \cup N_{<v}^= \rangle}. \quad (2.32)$$

This is proved in the same way as (2.29).

The next algorithm applies a reduction, analogous to the one used in the previous algorithm, to a pair of inequations  $p_1 \neq 0$ ,  $p_2 \neq 0$  instead of  $p_1 = 0$  and  $p_2 \neq 0$ .

**Algorithm 2.2.27** (*LCMSplit*).

**Input:**  $p_1, p_2 \in R - K$  with the same leader  $v$  and  $(L, M, N)$ ,  $P, >$  with the same specification as in Algorithm 2.2.21, where  $p_1 \neq 0$  is in  $L$ ,  $p_2 \neq 0$  is in  $M \cup N$ , and  $\deg_v(p_1) \geq \deg_v(p_2)$

**Output:** A finite set  $Q \supseteq P$  of triples as in  $P$  whose solution sets form a partition of  $\text{Sol}_{\bar{K}}(L \cup M \cup N) \uplus \text{Sol}_{\bar{K}}(P)$  such that for each triple in  $Q - P$  we have either

- a) the polynomials  $\bar{p}_1$  and  $\bar{p}_2$  which are obtained from  $p_1$  and  $p_2$  by substituting a solution of the triple for  $x_1, \dots, x_n$  except  $v$  have a least common multiple whose degree does not depend on the choice of the solution of the triple, or
- b) the triple has been inserted in step 11

**Algorithm:**

- 1:  $Q \leftarrow P$ ;  $U \leftarrow \emptyset$
- 2:  $v \leftarrow \text{ld}(p_1)$ ;  $i \leftarrow 0$
- 3:  $r_1 \leftarrow p_1$ ;  $c_1 \leftarrow 0$
- 4:  $r_2 \leftarrow p_2$ ;  $c_2 \leftarrow 1$
- 5: **repeat**
- 6:    $i \leftarrow i + 1$
- 7:    $r_{i+2} \leftarrow a_i \cdot r_i - q_i \cdot r_{i+1}$ , where  $a_i$  is a power of  $\text{init}(r_{i+1})$  and  $q_i \in R$  such that  $r_{i+2} = 0$  or  $\deg_v(r_{i+2}) < \deg_v(r_{i+1})$
- 8:    $(r_{i+2}, b_{i+2}) \leftarrow \text{Reduce}(r_{i+2}, M^= \cup N^=, >)$  // cf. Alg. 2.2.17
- 9:    $c_{i+2} \leftarrow b_{i+2} \cdot (a_i \cdot c_i + q_i \cdot c_{i+1})$
- 10:   **if**  $r_{i+2} \neq 0$  **and**  $\deg_v(r_{i+2}) > 0$  **then**
- 11:     insert  $(L \cup \{\text{init}(r_{i+2}) = 0\} \cup U, M, N)$  into  $Q$
- 12:      $U \leftarrow U \cup \{\text{init}(r_{i+2}) \neq 0\}$
- 13:   **end if**
- 14: **until**  $r_{i+2} = 0$  **or**  $\deg_v(r_{i+2}) = 0$
- 15: insert  $((L - \{p_1 \neq 0\}) \cup \{c_{i+2} \cdot p_2 \neq 0, r_{i+2} = 0\} \cup U, M - \{p_2 \neq 0\}, N - \{p_2 \neq 0\})$  into  $Q$
- 16: insert  $((L - \{p_1 \neq 0\}) \cup \{p_1 \cdot p_2 \neq 0, r_{i+2} \neq 0\} \cup U, M - \{p_2 \neq 0\}, N - \{p_2 \neq 0\})$  into  $Q$
- 17: **return**  $Q$

**Lemma 2.2.28.** *Algorithm 2.2.25 (on page 73) terminates and is correct.*

*Proof.* Termination of Algorithm 2.2.25 follows from the fact that the degree in  $v$  of the elements of the sequence  $r_2, r_3, r_4, \dots$  is decreasing.

The solution set of  $(L, M, N)$  is partitioned into solution sets of several triples in the result  $Q$  due to steps 6, 13, and 14. In the beginning of each round of the loop the splitting of the current triple into the one defined in step 6 and complementary ones incorporating the update of  $U$  in step 7 ensures that in step 9 the inequation  $\text{init}(r_{i+1}) \neq 0$  holds if a solution of the current triple is substituted for  $x_1, \dots, x_n$ . This also implies  $a_i \neq 0$ . Since the initials of left hand sides of elements of  $M \cup N$  do not vanish on solutions of the current triple, the inequation  $b_{i+2} \neq 0$  holds as well.

In step 13 the condition  $r_{i+2} = 0$  is imposed, which is complemented by  $r_{i+2} \neq 0$  in the triple defined in step 14. Note that the first component of the triple in step 13 contains the inequation  $p_2 \neq 0$ , so that the inequation  $r_{i+1} \neq 0$  holds for all solutions of this triple because of (2.32),  $r_{i+2} = 0$ , and  $a_2 \cdot a_3 \cdot \dots \cdot a_i \cdot b_4 \cdot b_5 \cdot \dots \cdot b_{i+2} \neq 0$ . Then, by (2.29), the equation  $c_{i+2} = 0$  holds for all solutions of the triple. Conversely, the equations  $c_{i+2} = 0$  and  $r_{i+2} = 0$  and the inequation  $a_1 \cdot a_2 \cdot \dots \cdot a_i \cdot b_3 \cdot b_4 \cdot \dots \cdot b_{i+2} \neq 0$  imply  $p_1 = 0$ . Therefore, the solution set of  $L \cup U \cup M \cup N$  is not changed if  $p_1 = 0$  is replaced with  $c_{i+2} = 0$ .

Finally, in step 14 the condition  $r_{i+2} \neq 0$  is imposed. Since  $r_{i+2}$  is an  $R$ -linear combination of  $p_2$  and the left hand sides of the equations in  $M \cup N$ , this condition implies  $p_2 \neq 0$ , so that the latter inequation is dispensable for the updated triple.

Let  $\bar{p}_1$  and  $\bar{p}_2$  be obtained from  $p_1$  and  $p_2$ , respectively, by substituting a solution of the triple in step 13 or 14 for  $x_1, \dots, x_n$  except  $v$ . The same substitution specializes the sequence of polynomials  $r_1, r_2, r_3, \dots$  to the one (up to non-zero constant factors) which is computed by Euclid's algorithm for the univariate polynomials  $\bar{p}_1$  and  $\bar{p}_2$ , because  $\text{init}(r_i)$  does not vanish for any polynomial  $r_i$  preceding the final one. This shows the last assertion stated in the description of the output.  $\square$

**Lemma 2.2.29.** *Algorithm 2.2.27 (on page 75) terminates and is correct.*

*Proof.* Termination is shown exactly as in the proof of Lemma 2.2.28.

The solution set of  $(L, M, N)$  is partitioned into solution sets of several triples in the result  $Q$  due to steps 11, 15, and 16. As opposed to the input of Algorithm 2.2.25, the inequation  $p_2 \neq 0$  is an element of  $M \cup N$  rather than  $L$ . This ensures that  $\text{init}(r_{i+1}) \neq 0$  holds if a solution of the current triple in step 7 in the first round of the loop is substituted for  $x_1, \dots, x_n$ . The splitting of algebraic systems in step 11 arranges for the corresponding property in the next round.

Similarly to step 13 in Algorithm 2.2.25, in step 15 the condition  $r_{i+2} = 0$  is imposed, which is complemented by  $r_{i+2} \neq 0$  in the triple defined in step 16. Again, the inequation  $r_{i+1} \neq 0$  holds for all solutions of the triple in step 15 because of (2.32),  $r_{i+2} = 0$ ,  $a_1 \cdot a_2 \cdot \dots \cdot a_i \cdot b_3 \cdot b_4 \cdot \dots \cdot b_{i+2} \neq 0$ , and  $p_2 \neq 0$  (imposed by the first entry of the triple). Given these conditions, the inequations  $c_{i+2} \neq 0$  and  $p_1 \neq 0$  are equivalent by (2.29). Hence, replacing  $p_1 \neq 0$  and  $p_2 \neq 0$  with  $c_{i+2} \cdot p_2 \neq 0$  in step 15 does not change the solution set of  $L \cup U \cup M \cup N$ . Replacing  $p_1 \neq 0$  and  $p_2 \neq 0$  with  $p_1 \cdot p_2 \neq 0$  does not change the solution set in step 16 either.



The last assertion stated in the description of the output is proved in the same way as the corresponding one for Algorithm 2.2.25.  $\square$

Finally, the same reduction technique is applied to determine square-free parts.

**Algorithm 2.2.30** (*SquarefreeSplit*).

**Input:**  $p \in R - K$  with degree at least 2 in its leader  $v$  and  $(L, M, N)$ ,  $P, >$  with the same specification as in Algorithm 2.2.21, where  $L = \emptyset$  and  $p$  is the left hand side of an equation or inequation in  $M$

**Output:** A finite set  $Q \supseteq P$  of triples as in  $P$  whose solution sets form a partition of  $\text{Sol}_{\overline{K}}(M \cup N) \uplus \text{Sol}_{\overline{K}}(P)$  such that for each triple in  $Q - P$  we have either

- a) the two polynomials which are obtained from  $p$  and  $\frac{\partial p}{\partial v}$  by substituting a solution of the triple for  $x_1, \dots, x_n$  except  $v$  have a greatest common divisor whose degree does not depend on the choice of the solution of the triple, or
- b) the triple has been inserted in step 10

**Algorithm:**

```

1:  $Q \leftarrow P$ ;  $U \leftarrow \emptyset$ ;  $v \leftarrow \text{ld}(p)$ ;  $i \leftarrow 0$ 
2:  $r_1 \leftarrow p$ ;  $c_1 \leftarrow 0$ 
3:  $r_2 \leftarrow \frac{\partial p}{\partial v}$ ;  $c_2 \leftarrow 1$ 
4: repeat
5:    $i \leftarrow i + 1$ 
6:    $r_{i+2} \leftarrow a_i \cdot r_i - q_i \cdot r_{i+1}$ , where  $a_i$  is a power of  $\text{init}(r_{i+1})$  and  $q_i \in R$  such that
      $r_{i+2} = 0$  or  $\deg_v(r_{i+2}) < \deg_v(r_{i+1})$ 
7:    $(r_{i+2}, b_{i+2}) \leftarrow \text{Reduce}(r_{i+2}, M^\# \cup N^\#, >)$  // cf. Alg. 2.2.17
8:    $c_{i+2} \leftarrow b_{i+2} \cdot (a_i \cdot c_i + q_i \cdot c_{i+1})$ 
9:   if  $r_{i+2} \neq 0$  and  $\deg_v(r_{i+2}) > 0$  then
10:    insert  $(\{\text{init}(r_{i+2}) = 0\} \cup U, M, N)$  into  $Q$ 
11:     $U \leftarrow U \cup \{\text{init}(r_{i+2}) \neq 0\}$ 
12:   end if
13: until  $r_{i+2} = 0$  or  $\deg_v(r_{i+2}) = 0$ 
14: if  $M$  contains  $p = 0$  then
15:   insert  $(\{c_{i+2} = 0, r_{i+2} = 0\} \cup U, M - \{p = 0\}, N)$  into  $Q$ 
16:   insert  $(\{r_{i+2} \neq 0\} \cup U, M - \{p = 0\}, N \cup \{p = 0\})$  into  $Q$ 
17: else //  $M$  contains  $p \neq 0$ 
18:   insert  $(\{c_{i+2} \neq 0, r_{i+2} = 0\} \cup U, M - \{p \neq 0\}, N)$  into  $Q$ 
19:   insert  $(\{r_{i+2} \neq 0\} \cup U, M - \{p \neq 0\}, N \cup \{p \neq 0\})$  into  $Q$ 
20: end if
21: return  $Q$ 

```

**Lemma 2.2.31.** *Algorithm 2.2.30 terminates and is correct.*

*Proof.* Again, the same argument as in the proof of Lemma 2.2.28 shows that Algorithm 2.2.30 terminates.

The solution set of  $(L, M, N)$  is partitioned into solution sets of several triples in the result  $Q$  due to steps 10 and 15, 16 or 18, 19. Since  $p$  has degree at least two in  $v$ , the initial of  $\frac{\partial p}{\partial v}$  is a constant multiple of  $\text{init}(p)$ , and since the inequation  $p \neq 0$  is an element of  $M$ , the inequation  $\text{init}(r_{i+1}) \neq 0$  holds if a solution of the current triple in step 6 in the first round of the loop is substituted for  $x_1, \dots, x_n$ . For further rounds the inequation  $\text{init}(r_{i+1}) \neq 0$  has been added to  $U$  in the previous round.

After step 8 the congruence (2.29) holds with  $p_1$  replaced with  $p$ , and if the sequence  $d_2, d_3, d_4, \dots$  defined in Remark 2.2.26 c) is also computed, then the congruence (2.32) holds with  $p_2$  replaced with  $\frac{\partial p}{\partial v}$ .

In steps 15 and 18 the condition  $r_{i+2} = 0$  is imposed, which is complemented by  $r_{i+2} \neq 0$  in the triple defined in step 16 or 19, respectively. We claim that replacing the equation  $p = 0$  with  $c_{i+2} = 0$  does not change the solution set of the triple in step 15. First of all, by (2.29), the equations  $c_{i+2} = 0$  and  $r_{i+2} = 0$  and the inequation  $a_1 \cdot a_2 \cdot \dots \cdot a_i \cdot b_3 \cdot b_4 \cdot \dots \cdot b_{i+2} \neq 0$  imply  $p = 0$ . Conversely, we show that every solution of  $(L \cup U \cup \{r_{i+2} = 0\}, M, N)$  is a solution of  $c_{i+2} = 0$ . Let  $\bar{p}, \bar{c}_{i+2}, \bar{r}_{i+1}, \bar{a}_j$ , and  $\bar{b}_k$  be obtained from  $p, c_{i+2}, r_{i+1}, a_j$ , and  $b_k$ , respectively, by substituting such a solution for  $x_1, \dots, x_n$  except  $v$ . Then (2.29) specializes to

$$\bar{c}_{i+2} \cdot \bar{r}_{i+1} = \left( \prod_{j=1}^i \bar{a}_j \right) \cdot \left( \prod_{k=3}^{i+2} \bar{b}_k \right) \cdot \bar{p}, \quad (2.33)$$

where the degree in  $v$  of each factor is the same as the degree in  $v$  of the corresponding factor in (2.29) because  $\text{init}(p)$  and  $\text{init}(r_{i+1})$  do not vanish. Let  $\eta \in \bar{K}$  be the component of the solution which corresponds to  $v$ . If  $\bar{r}_{i+1}(\eta) = 0$ , then (2.33) implies  $\bar{c}_{i+2}(\eta) = 0$ , which proves the claim in this case. Otherwise, the corresponding specialization of (2.32) shows that  $\eta$  is a common root of  $\bar{p}$  and its derivative. Then  $\eta$  is a root of  $\bar{p}$  of multiplicity greater than one. Since  $\bar{r}_{i+1}$  divides both  $\bar{p}$  and its derivative, we conclude that  $\eta$  is a root of  $\bar{p}/\bar{r}_{i+1}$  and hence of  $\bar{c}_{i+2}$ .

Next we show that replacing the inequation  $p \neq 0$  with  $c_{i+2} \neq 0$  does not change the solution set of the triple in step 18. Clearly, by (2.29), the equation  $r_{i+2} = 0$  and the inequations  $p \neq 0$  and  $a_1 \cdot a_2 \cdot \dots \cdot a_i \cdot b_3 \cdot b_4 \cdot \dots \cdot b_{i+2} \neq 0$  imply  $c_{i+2} \neq 0$ . Conversely, we show that the inequation  $p \neq 0$  holds for all solutions of the triple in step 18. Using the same notation as above, we have  $\bar{r}_{i+1}(\eta) = 0$  or  $\bar{r}_{i+1}(\eta) \neq 0$ . In the former case we conclude in the same way as above that  $\eta$  is a common root of  $\bar{p}$  and its derivative and therefore a root of  $\bar{c}_{i+2}$ , which is a contradiction. Hence, we have  $\bar{r}_{i+1}(\eta) \neq 0$  and therefore,  $p \neq 0$  holds.

The last assertion stated in the description of the output follows by the same argument as in the proof of Lemma 2.2.28. Finally, in order to justify the transfer of  $p = 0$  or  $p \neq 0$  from the second to the third component of the triple in step 16 or 19, we note that either  $r_{i+2}$  is the zero polynomial and the triple has no solution, or the greatest common divisor of  $\bar{p}$  and its derivative is the non-zero constant which

is obtained from  $r_{i+2}$  by substituting the solution that defines  $\bar{p}$ , which shows that  $\bar{p}$  and its derivative have no common root.  $\square$

**Theorem 2.2.32.** *Algorithm 2.2.20, p. 70, terminates and is correct.*

*Proof.* In order to prove correctness, we note first that step 5 in Algorithm 2.2.20 (*AlgebraicThomasDecomposition*) ensures that the left hand sides of elements of  $L$  in step 15 are reduced with respect to  $M^\infty \cup N^\infty$ , and steps 6 and 7 guarantee that they are not constant. The property that  $M \cup N$  contains at most one equation or inequation with a given leader is retained throughout.

An equation or inequation with left hand side  $p$  is only inserted into the second component  $M$  of a triple  $(L, M, N)$  if all solutions of the updated triple satisfy  $\text{init}(p) \neq 0$ , namely in steps 4, 13, and 17 in Algorithm 2.2.21 (*ProcessInitial*). Similarly, an equation or inequation with left hand side  $p$  is only inserted into the third component  $N$  of such a triple if it is moved there from the second component  $M$  and if all solutions of the updated triple satisfy  $\text{disc}(p) \neq 0$ , namely in steps 5 and 7 in Algorithm 2.2.23 (*ProcessDiscriminant*) and in steps 16 and 19 in Algorithm 2.2.30 (*SquarefreeSplit*) (cf. the end of Remark 2.2.24 and the end of the proof of Lemma 2.2.31 for justifications).

As a result of the above discussion, if an algebraic system  $N$  is inserted into  $T$  in step 10, this system is simple. The output  $T$  is a Thomas decomposition of the input system  $S$  because the solution sets of triples in  $Q$  are pairwise disjoint throughout the algorithm, the solution sets of algebraic systems in  $T$  are pairwise disjoint, and the union of  $\text{Sol}_{\bar{K}}(Q)$  and the solution sets of algebraic systems in  $T$  equals  $\text{Sol}_{\bar{K}}(S)$  (cf. Remarks 2.2.22 and 2.2.24, Lemmas 2.2.28, 2.2.29, and 2.2.31).

Termination of Algorithm 2.2.20 follows if we show that after finitely many steps the set  $Q$  is empty. Since every triple in  $Q$  arises from splittings of algebraic systems, whose common origin is the triple  $(S, \emptyset, \emptyset)$ , it is sufficient to prove that every triple is removed after finitely many steps and that no triple has infinitely many descendants. In fact, we are going to argue for each splitting that the further treatment of a new triple  $(L', M', N')$  leads to a modification of  $M'$  or  $N'$  and that only finitely many consecutive modifications are possible for each triple and its descendants.

Each triple  $(L, M, N)$  in  $Q$  is either discarded or is dealt with by Algorithm 2.2.21 (*ProcessInitial*) or Algorithm 2.2.23 (*ProcessDiscriminant*). The first case occurs if an equation or inequation with constant left hand side reveals that the algebraic system is inconsistent, or if  $L$  and  $M$  are empty, in which case  $N$  is inserted into the set  $T$ . Algorithms 2.2.21 and 2.2.23, using also Algorithms 2.2.25 (*GCDSplit*), 2.2.27 (*LCMSplit*), and 2.2.30 (*SquarefreeSplit*), insert further triples into  $Q$  whose solution sets form a partition of  $\text{Sol}_{\bar{K}}(L \cup M \cup N)$ .

A modification of  $M$  or  $N$  is possible precisely in the following ways:

- a) An equation  $p = 0$  with leader  $v$  is transferred from  $L$  to  $M$  after equations and inequations with leader greater than or equal to  $v$  have been transferred from  $M \cup N$  to  $L$  (Alg. 2.2.21, step 4). If  $M \cup N$  contained an equation with leader  $v$  before, then  $p$  has smaller degree in  $v$  than the left hand side of the old equation because  $p$  was reduced with respect to  $M^\infty \cup N^\infty$  before the insertion.

- b) An inequation  $p \neq 0$  with leader  $v$  is transferred from  $L$  to  $M$  only if  $M \cup N$  does not contain an equation with leader  $v$  (Alg. 2.2.21, steps 13 and 17). If  $M \cup N$  contained an inequation with leader  $v$  before, then  $p$  has smaller degree in  $v$  than the left hand side of the old inequation, and the old inequation is transferred to  $L$ .
- c) An equation  $p_1 = 0$  with leader  $v$  is removed from  $M$  or from  $N$  and an equation  $c_{i+2} = 0$  is inserted into  $L$ , where  $c_{i+2}$  is constant, but non-zero, or the leader of  $c_{i+2}$  is  $v$ , and  $\deg_v(c_{i+2})$  is less than  $\deg_v(p_1)$ , and  $\text{init}(c_{i+2})$  is not in the ideal  $\langle M_{<v}^- \cup N_{<v}^- \rangle$  (Alg. 2.2.25, step 13). Finitely many inequations whose left hand sides are constant, but non-zero, or have leaders which are smaller than  $v$  may be inserted into  $L$  as well.

In order to confirm these properties, we note that  $c_3$  in Algorithm 2.2.25 is constant if and only if we have  $\deg_v(p_1) = \deg_v(p_2)$ , that the degree in  $v$  of the entries of the sequence  $c_3, c_4, c_5, \dots$  is increasing and the degree in  $v$  of those in  $r_2, r_3, r_4, \dots$  is decreasing. If  $c_3$  is constant, then it is non-zero because  $c_1$  is zero, but  $b_3, q_1$ , and  $c_2$  are not, so that the new triple defines an inconsistent algebraic system. Otherwise,  $c_{i+2}$  has leader  $v$  and degree in  $v$  less than  $\deg_v(p_1)$  due to (2.29), p. 74, and because of the properties of the above sequences. The initial of  $c_{i+2}$  is not in the ideal  $\langle M_{<v}^- \cup N_{<v}^- \rangle$  because the initial of  $p_1$  is not.

The set  $L$  in the input of Algorithm 2.2.25 contains neither equations with leader  $v$  nor equations or inequations with smaller leader. When the new triple defined in step 13 with  $\deg_v(c_{i+2}) > 0$  will be further processed, inequations with leader smaller than  $v$ , contributed by the set  $U$  in step 13, if any, and equations and inequations with leader smaller than  $v$  produced by this process will be dealt with. Further splittings may occur. Since  $\text{init}(c_{i+2})$  is not in  $\langle M_{<v}^- \cup N_{<v}^- \rangle$ , a reduction may decrease the degree of  $c_{i+2}$  in  $v$  only if an equation with leader smaller than  $v$  has been inserted into the second component of the triple in question. Otherwise, (a reduced form of) the new equation  $c_{i+2} = 0$  will be inserted into the second component. In all cases a modification of type a) along with the generation of a new triple as in g) below will occur.

- d) Inequations  $p_1 \neq 0$  and  $p_2 \neq 0$  with the same leader  $v$  are removed from  $L$  and  $M$  or  $N$ , respectively, and an inequation with leader  $v$  is inserted into  $L$  (Alg. 2.2.27, steps 15 and 16). Finitely many equations and inequations whose left hand sides are constant or have leaders smaller than  $v$  may be inserted into  $L$  as well.
- e) An equation or inequation with left hand side  $p$  is removed from  $M$  or  $N$  and, correspondingly, an equation or inequation with left hand side  $c_{i+2}$  is inserted into  $L$ , where  $c_{i+2}$  is constant, but non-zero, or the leader of  $c_{i+2}$  is  $v$ , and  $\deg_v(c_{i+2})$  is less than  $\deg_v(p)$ , and  $\text{init}(c_{i+2})$  is not in the ideal  $\langle M_{<v}^- \cup N_{<v}^- \rangle$  (Alg. 2.2.30, steps 15 and 18). An equation and finitely many inequations whose left hand sides are constant or have leaders smaller than  $v$  may be inserted into  $L$  as well. The fact that  $c_{i+2}$  has the above property follows in the same way as in c).
- f) An equation or inequation with left hand side  $p$  is transferred from  $M$  to  $N$  (Alg. 2.2.23, steps 5 and 7, Alg. 2.2.30, steps 16 and 19). Finitely many inequations whose left hand sides are constant or have leaders which are smaller than  $v$  may also be inserted into  $L$ .

New triples with unmodified second and third component arise as follows:

- g) An equation is inserted into  $L$  whose left hand side is the initial of a polynomial whose coefficients are reduced with respect to  $M^\infty \cup N^\infty$  (Alg. 2.2.21, steps 5, 14, and 18, Alg. 2.2.25, step 6, Alg. 2.2.27, step 11, or Alg. 2.2.30, step 10). A similar argument about the insertion of equations into  $M$  as given in c) applies in this case (if the left hand side is not constant).
- h) An inequation  $p_2 \neq 0$  in  $L$  is replaced with finitely many inequations whose left hand sides are constant, but non-zero, or have leaders which are smaller than  $v$  (Alg. 2.2.25, step 14).

Every new triple which is inserted into  $Q$  arises in exactly one of the cases a)–h). Modifications of type c) and g) entail, after finitely many steps, modifications of type a) for each resulting triple and the creation of a new triple as in g). In this way only finitely many triples are generated because the vector  $(d_1, \dots, d_n)$  defined by

$$d_i := \begin{cases} \deg_v(p), & \text{if } M \cup N \text{ contains the equation } p = 0 \text{ with leader } v, \\ \infty, & \text{if } M \cup N \text{ contains no equation with leader } v, \end{cases}$$

where  $v$  is the  $i$ -th smallest variable with respect to  $>$ , decreases with respect to the lexicographical ordering as a result of a) and also as an indirect result of c) or g). Moreover, the leader of left hand sides of equations dealt with in g) decreases with respect to  $>$ . We claim that modifications of type b), d), e), f), and h) can be repeated (in any order) only finitely many times before a modification of type a) is applied or the algorithm stops. If an inequation in  $M \cup N$  is replaced with an inequation with the same leader  $v$ , then the new inequation has smaller degree in  $v$  (cf. b)). This shows that a sequence of modifications as in the assertion contains types b) and f) only finitely many times. Modifications of the remaining types either replace two inequations with the same leader  $v$  with one inequation with leader  $v$  in  $L \cup M \cup N$  (cf. d)) or remove one inequation from  $L \cup M \cup N$  (cf. e) and h)), besides possibly inserting finitely many inequations into  $L$  whose left hand sides are constant or have smaller leader than the left hand side(s) of the removed inequation(s). Since no infinite sequence of non-constant polynomials exists in which each polynomial is followed by one with smaller leader, after finitely many steps either  $L$  and  $M$  will be empty or the element which is chosen in step 2 in Algorithm 2.2.21 will be an equation. Termination of Algorithm 2.2.20 now follows from the fact that modifications of type b), d), e), f), and h) do not change the vector  $(d_1, \dots, d_n)$ .  $\square$

**Remark 2.2.33.** In order to prevent a large growth of expressions and to simplify the final result, two strategies should be included at certain stages of the above algorithms. The left hand side of each equation and inequation should be divided by its numerical content, i.e., to obtain a primitive (multivariate) polynomial. Moreover, and also more generally, if all coefficients of the left hand side of an equation  $p = 0$  or inequation  $p \neq 0$  are divisible by a non-trivial factor  $r$  of the left hand side of an inequation  $q \neq 0$ , then they should be replaced with their quotients by  $r$ . In particular, it is worthwhile to apply this simplification, if possible, after  $\text{init}(p) \neq 0$  has been inserted into the first component of a triple (cf. steps 4, 13, and 17 in Alg. 2.2.21).

For instance, the following two simple algebraic systems over  $\mathbb{Q}[x, y]$  are equivalent, where  $x > y$ .

$$\boxed{\begin{array}{c} 2y\underline{x} + 4y^2 = 0 \\ 2y \neq 0 \end{array}} \iff \boxed{\begin{array}{c} \underline{x} + 2y = 0 \\ y \neq 0 \end{array}}$$

In Algorithms 2.2.25 (*GCDSplit*), 2.2.27 (*LCMSplit*), and 2.2.30 (*SquarefreeSplit*) it is not specified in step 9, 7, and 6, respectively, which power  $a_i$  of  $\text{init}(r_{i+1})$  should be chosen. The power with exponent  $\deg_v(r_i) - \deg_v(r_{i+1}) + 1$  allows a polynomial division without fractions in any case because the polynomial division involves at most  $\deg_v(r_i) - \deg_v(r_{i+1}) + 1$  subtractions, but a proper divisor of this power may allow this as well for a particular pair  $r_i, r_{i+1}$  of polynomials. Using subresultant polynomial remainder sequences (cf., e.g., [Mis93]) is a considerable improvement (cf. also [BGL<sup>+</sup>12, Sect. 2]).

Furthermore, in order to avoid repeated computations, for each equation and inequation information about whether the initial and discriminant of its left hand side are ensured not to vanish on the solution set of the algebraic system should be recorded. If the equation or inequation is inserted into  $M$  and its initial is known not to vanish, the insertion of  $\text{init}(p) \neq 0$  can be neglected in steps 4, 13, and 17 in Algorithm 2.2.21 (*ProcessInitial*) and step 5, 14, or 18, respectively, can be skipped. Similarly, in Algorithm 2.2.23 (*ProcessDiscriminant*) step 5 or 7 can be applied to an equation or inequation, respectively, which is chosen in step 2 and which is known to have non-vanishing discriminant.

We demonstrate Algorithm 2.2.20 (*AlgebraicThomasDecomposition*) on two examples.

**Example 2.2.34.** We revisit Example 2.2.12, p. 65, where  $R = \mathbb{Q}[x, a, b, c]$  and the total ordering  $>$  on the set of variables is given by  $x > c > b > a$ . In step 1 we have

$$Q = \{(\{ax^2 + bx + c = 0\}, \emptyset, \emptyset)\}.$$

Steps 4 and 5 in Algorithm 2.2.21 (*ProcessInitial*) insert two triples into  $Q$  whose solution sets form a partition of the solution set of the initial triple:

$$Q = \{(\{a = 0, ax^2 + bx + c = 0\}, \emptyset, \emptyset), (\{a \neq 0\}, \{ax^2 + bx + c = 0\}, \emptyset)\}.$$

The first triple in this enumeration is dealt with by Algorithm 2.2.21, which moves the equation  $a = 0$  to the second component. We omit both the inequation  $1 \neq 0$  and the inconsistent algebraic system containing the equation  $1 = 0$ , which arise from the splitting in steps 4 and 5. Similarly, the inequation in the second triple is moved to the second component, and we omit inequations with constant left hand sides and inconsistent systems here:

$$Q = \{(\{ax^2 + bx + c = 0\}, \{a = 0\}, \emptyset), (\emptyset, \{a \neq 0, ax^2 + bx + c = 0\}, \emptyset)\}.$$

The left hand side of the first equation in the first triple is replaced with  $bx + c$  by Algorithm 2.2.17 (*Reduce*), and steps 4 and 5 in Algorithm 2.2.21 split this triple according to the vanishing or non-vanishing of the initial of the modified equation. Algorithm 2.2.23 (*ProcessDiscriminant*) is applied to the second triple, which moves the inequation to the third component in step 7:

$$Q = \{ (\{b = 0, bx + c = 0\}, \{a = 0\}, \emptyset), (\{b \neq 0\}, \{a = 0, bx + c = 0\}, \emptyset), (\emptyset, \{ax^2 + bx + c = 0\}, \{a \neq 0\}) \}.$$

The equation  $b = 0$  in the first triple is moved to the second component and a subsequent reduction replaces  $bx + c$  with  $c$ . The inequation  $b \neq 0$  in the second triple is moved to the second component. Steps 15 and 16 in Algorithm 2.2.30 split the third triple and add the inequation  $4ac - b^2 \neq 0$  and the equation  $4ac - b^2 = 0$ , respectively:

$$\begin{aligned} Q = \{ (\{c = 0\}, \{a = 0, b = 0\}, \emptyset), (\emptyset, \{a = 0, b \neq 0, bx + c = 0\}, \emptyset), \\ (\{4ac - b^2 \neq 0\}, \emptyset, \{a \neq 0, ax^2 + bx + c = 0\}), \\ (\{4ac - b^2 = 0, 2ax + b = 0\}, \emptyset, \{a \neq 0\}) \}. \end{aligned} \quad (2.34)$$

The equation  $c = 0$  in the first triple is moved to the second component and subsequently all three equations are moved to the third component. Similarly, all elements of the second component of the second triple are moved to the third component in steps 5 and 7 in Algorithm 2.2.23. These two triples give rise to the following simple algebraic systems (cf. also the end of Ex. 2.2.12):

$b\bar{x} + c = 0$   $b \neq 0$  $a = 0$	$c = 0$  $b = 0$  $a = 0$
---	---------------------------------------

Algorithm 2.2.21 (*ProcessInitial*) splits the third triple in (2.34) according to the vanishing or non-vanishing initial of  $4ac - b^2$  (steps 17 and 18):

$$\begin{aligned} (\{a \neq 0\}, \{4ac - b^2 \neq 0\}, \{a \neq 0, ax^2 + bx + c = 0\}), \\ (\{a = 0, 4ac - b^2 \neq 0\}, \emptyset, \{a \neq 0, ax^2 + bx + c = 0\}). \end{aligned}$$

The equation  $a = 0$  in the first component of the second triple is moved to the second component and the inequation  $a \neq 0$  from the third to the first one. A subsequent reduction shows that this triple defines an inconsistent algebraic system.

The first triple is dealt with by applying Algorithm 2.2.27 (*LCMSplit*) to the pair of inequations  $a \neq 0, a \neq 0$ . Step 15 produces the triple

$$(\emptyset, \{a \neq 0, 4ac - b^2 \neq 0\}, \{ax^2 + bx + c = 0\}),$$

which yields the simple system

$$\begin{array}{c} a\underline{x}^2 + b\underline{x} + c = 0 \\ 4a\underline{c} - b^2 \neq 0 \\ a \neq 0 \end{array}$$

after the inequations have been moved from the second to the third component by Algorithm 2.2.23 (*ProcessDiscriminant*).

The fourth triple in (2.34) is split into two triples by steps 4 and 5 in Algorithm 2.2.21 (*ProcessInitial*):

$$\begin{aligned} &(\{4a \neq 0, 2ax + b = 0\}, \{4ac - b^2 = 0\}, \{a \neq 0\}), \\ &(\{4a = 0, 4ac - b^2 = 0, 2ax + b = 0\}, \emptyset, \{a \neq 0\}). \end{aligned} \quad (2.35)$$

Again, a reduction reveals that the second triple has an empty solution set. After another application of Algorithm 2.2.27 (*LCMSplit*), we obtain the remaining simple algebraic system of the Thomas decomposition (cf. also the end of Ex. 2.2.12):

$$\begin{array}{c} 2a\underline{x} + b = 0 \\ 4a\underline{c} - b^2 = 0 \\ a \neq 0 \end{array}$$

We give an outline of a computation of a Thomas decomposition which is a little bit more involved. Advantage is taken of simplifications as described in Remark 2.2.33.

**Example 2.2.35.** Let  $R = \mathbb{Q}[x, y, z]$ . The *Steiner quartic surface* (cf., e.g., [Bak10, p. 221]) is defined by the equation

$$x^2 y^2 + x^2 z^2 + y^2 z^2 - xyz = 0. \quad (2.36)$$

We choose the total ordering  $x > y > z$  on the set of variables.

Algorithm 2.2.20 (*AlgebraicThomasDecomposition*) starts by splitting the original algebraic system according to vanishing or non-vanishing of the initial  $y^2 + z^2$ . In the former case the original equation is reduced to  $zyx + z^4 = 0$ . The updated system is split again according to vanishing or non-vanishing of the initial  $zy$ . Again, in the former case the analogous case distinction yields, after application of Algorithm 2.2.30 (*SquarefreeSplit*), the simple system

$$\{z = 0, y = 0\}$$



and produces only inconsistent algebraic systems otherwise. In case of the algebraic system containing the inequation  $zy \neq 0$ , Algorithm 2.2.25 (*GCDSplit*) is applied to  $p_1 = y^2 + z^2$  and  $p_2 = zy$ . Three new algebraic systems are generated, one in step 6 containing the equation  $z = 0$ , which is inconsistent, one in step 13 containing  $z \neq 0$  and  $z^4 = 0$ , which is also inconsistent, and one in step 14, which after applying Algorithms 2.2.27 (*LCMSplit*) and 2.2.30 (*SquarefreeSplit*) yields the simple system

$$\{z \neq 0, \underline{y}^2 + z^2 = 0, y\underline{x} + z^3 = 0\}.$$

The branch emerging from the case  $y^2 + z^2 \neq 0$  remains to be dealt with. Application of *SquarefreeSplit* to this inequation splits the algebraic system into one containing  $z^2 = 0$  and  $y \neq 0$  and another one containing  $z^2 \neq 0$  and  $y^2 + z^2 \neq 0$ . In the former case (2.36) is reduced to  $y^2x^2 - zyx = 0$ , which simplifies to  $yx^2 - zx = 0$  because of  $y \neq 0$ . After computing the square-free part of  $z^2$ , the reduced form of the simplified equation modulo  $z$ , and the square-free part of  $x^2$ , we obtain the simple system

$$\{z = 0, y \neq 0, x = 0\}.$$

In the latter case *SquarefreeSplit* is applied to  $z^2 \neq 0$  and then to (2.36). This generates two new branches to which the equation or inequation with left hand side

$$4z^2y^4 + (4z^4 - z^2)y^2 \quad (2.37)$$

is added, respectively. The first branch, where the original equation is replaced with

$$2(y^2 + z^2)x - zy = 0, \quad (2.38)$$

produces two simple systems. The essential steps amount to applying *SquarefreeSplit* to  $z^2 \neq 0$  and to the equation with left hand side (2.37). Thus, the cases of vanishing or non-vanishing of  $4z^2 - 1$  and in the latter case that of  $16z^4 - 8z^2 + 1$  are investigated. If  $4z^2 - 1 = 0$  is imposed, then (2.37) is reduced to  $y^4$  and (2.38) is reduced to  $(4y^2 + 1)x - 2zy = 0$ . Applying *GCDSplit* to  $4z^2 - 1 = 0$  and  $z \neq 0$ , *SquarefreeSplit* to  $y^4 = 0$ , and reduction modulo  $y$  yields the simple system

$$\{4z^2 - 1 = 0, y = 0, x = 0\}.$$

Since  $4z^2 - 1$  divides  $16z^4 - 8z^2 + 1$ , the algebraic system containing  $4z^2 - 1 \neq 0$  and  $16z^4 - 8z^2 + 1 = 0$  does not contribute to the Thomas decomposition. In the case of the algebraic system containing  $4z^2 - 1 \neq 0$  and  $16z^4 - 8z^2 + 1 \neq 0$  the equation with left hand side (2.37) is replaced with  $4y^3 + (4z^2 - 1)y = 0$ . After application of *SquarefreeSplit* to the least common multiple  $16z^5 - 8z^3 + z \neq 0$  of the inequations with leader  $z$  which have been encountered before, we obtain the simple system

$$\{4z^3 - z \neq 0, 4\underline{y}^3 + (4z^2 - 1)\underline{y} = 0, 2(y^2 + z^2)\underline{x} - zy = 0\}.$$

The branch addressing the inequation with left hand side (2.37) yields the rest of the Thomas decomposition. It is treated by first applying *LCMSplit* to this inequation

and  $y^2 + z^2 \neq 0$ , which essentially reveals that, in presence of the inequation  $z \neq 0$ , the least common multiple of their left hand sides is

$$4z^2y^6 + (8z^4 - z^2)y^4 + (4z^6 - z^4)y^2.$$

Again, due to the inequation  $z \neq 0$ , only the case of non-vanishing initial  $4z^2$  is relevant. Application of *SquarefreeSplit* to the simplified inequation

$$4y^6 + (8z^2 - 1)y^4 + (4z^4 - z^2)y^2 \neq 0 \quad (2.39)$$

produces five algebraic systems. One of them contains  $8z^2 - 1 = 0$  and the others contain the corresponding inequation. Among the latter systems one contains  $16z^4 - 4z^2 + 1 = 0$  and the others contain the complementary condition. The equation  $4z^4 - z^2 = 0$  is imposed in exactly one of the complementary systems and the two remaining ones incorporate the inequation with the same left hand side. One of these two contains  $z^4(2z+1)^2(2z-1)^2(8z^2-1) = 0$  and the other one the corresponding inequation.

In the very first case (2.36) is reduced to  $(8y^2 + 1)x^2 - 8zyx + y^2 = 0$  and (2.39) is reduced to  $64y^6 - y^2 \neq 0$ . After applying *SquarefreeSplit* to  $8z^2 - 1 = 0$  and to  $64y^6 - y^2 \neq 0$ , we obtain the simple system

$$\{8z^2 - 1 = 0, 64y^5 - y \neq 0, (8y^2 + 1)\underline{x}^2 - 8zy\underline{x} + y^2 = 0\}.$$

The algebraic system containing  $16z^4 - 4z^2 + 1 = 0$  leads after the application of *GCDSplit* to this equation and the least common multiple of  $8z^2 - 1 \neq 0$  and  $z \neq 0$  and the application of *SquarefreeSplit* to the equation  $16z^4 - 4z^2 + 1 = 0$  and to the inequation  $16y^6 + (32z^2 - 4)y^4 - y^2 \neq 0$  to the simple system

$$\{16z^4 - 4z^2 + 1 = 0, 16\underline{y}^5 + 4(8z^2 - 1)\underline{y}^3 - \underline{y} \neq 0, (y^2 + z^2)\underline{x}^2 - zy\underline{x} + z^2y^2 = 0\}.$$

The third of the five systems mentioned previously is dealt with by first applying *LCMSplit* to  $8z^2 - 1 \neq 0$  and  $z \neq 0$ . The equation  $4z^4 + z^2 = 0$  is used to reduce (2.39) to  $4y^6 + (8z^2 - 1)y^4 \neq 0$ . Then *GCDSplit* is applied to  $4z^4 + z^2 = 0$  and  $8z^3 - z \neq 0$ , which replaces the equation with  $4z^3 - z = 0$ , and the inequation is reduced to  $z \neq 0$ . Subsequently, *GCDSplit* replaces the equation with  $4z^2 - 1 = 0$  and removes  $z \neq 0$ . Then (2.36) is reduced to  $(4y^2 + 1)x^2 - 4zyx + y^2 = 0$  and  $4y^6 + (8z^2 - 1)y^4 \neq 0$  to  $4y^6 + y^4 \neq 0$ . After application of *SquarefreeSplit*, we get

$$\{4z^2 - 1 = 0, 4y^3 + y \neq 0, (4y^2 + 1)\underline{x}^2 - 4zy\underline{x} + y^2 = 0\}.$$

Finally, the inequations with leader  $z$  in the last of the five systems are combined by a series of calls of *LCMSplit* resulting in the inequation

$$z^4(2z+1)^2(2z-1)^2(8z^2-1)(16z^4-4z^2+1) \neq 0. \quad (2.40)$$

The inequation (2.39) had been replaced by *SquarefreeSplit* with an inequation whose left hand side has degree five in  $y$ . After imposing the condition that its initial

does not vanish, this inequation simplifies to  $4y^5 + (8z^2 - 1)y^3 + (4z^4 - z^2)y \neq 0$ . The square-free part of the inequation with leader  $z$  is determined next. Then *SquarefreeSplit* is applied to the inequation with leader  $y$ , distinguishing the cases of vanishing or non-vanishing of  $8z^2 - 1$ , of  $32z^4 - 8z^2 + 3$ , of  $4z^4 - z^2$ , and of  $z^4(2z+1)^2(2z-1)^2(8z^2-1)$ . After further applications of *GCDSplit*, *LCMSplit*, and *SquarefreeSplit*, the algebraic system containing  $32z^4 - 8z^2 + 3 = 0$  and the one containing  $z^4(2z+1)^2(2z-1)^2(8z^2-1) \neq 0$  each yield one simple system.

We conclude by displaying the constructed Thomas decomposition of (2.36), listing the simple systems in order of increasing dimension of their solution sets.

$(2z+1)(2z-1) = 0$ $y = 0$ $x = 0$	$z(2z+1)(2z-1) \neq 0$ $\underline{y}(4\underline{y}^2 + 4z^2 - 1) = 0$ $2(y^2 + z^2)\underline{x} - zy = 0$	
$z = 0$ $y = 0$	$z = 0$ $y \neq 0$ $x = 0$	$8z^2 - 1 = 0$ $y(8y^2 + 1)(8y^2 - 1) \neq 0$ $(8y^2 + 1)\underline{x}^2 - 8zy\underline{x} + y^2 = 0$
$(2z+1)(2z-1) = 0$ $y(4y^2 + 1) \neq 0$ $(4y^2 + 1)\underline{x}^2 - 4zy\underline{x} + y^2 = 0$	$16z^4 - 4z^2 + 1 = 0$ $\underline{y}(16\underline{y}^4 + 4(8z^2 - 1)\underline{y}^2 - 1) \neq 0$ $(y^2 + z^2)\underline{x}^2 - zy\underline{x} + z^2y^2 = 0$	
$z \neq 0$ $\underline{y}^2 + z^2 = 0$ $y\underline{x} + z^3 = 0$	$32z^4 - 8z^2 + 3 = 0$ $\underline{y}(32\underline{y}^4 + 8(8z^2 - 1)\underline{y}^2 - 3) \neq 0$ $(y^2 + z^2)\underline{x}^2 - zy\underline{x} + z^2y^2 = 0$	
$z(2z+1)(2z-1)(8z^2-1)(16z^4-4z^2+1)(32z^4-8z^2+3) \neq 0$ $\underline{y}(4\underline{y}^4 + (8z^2-1)\underline{y}^2 + z^2(4z^2-1)) \neq 0$ $(y^2 + z^2)\underline{x}^2 - zy\underline{x} + z^2y^2 = 0$ <span style="float: right;">(2.41)</span>		

The simple system (2.41) is the generic simple system for the prime ideal of  $R$  generated by (2.36) as discussed in Subsect. 2.2.3 (cf. also Ex. 2.2.68, p. 107).

Note that some unnecessary case distinctions are avoided when using subresultants. For more details about this technique, we refer to [BGL<sup>+</sup>12].

### 2.2.2 Simple Differential Systems

This subsection gives a modern description of the method of J. M. Thomas of decomposing systems of finitely many partial differential equations and inequations into finitely many so-called simple systems. The set of solutions of the given system is thereby partitioned into the solution sets of the simple systems, and using the simple systems, e.g., an effective membership test for the radical differential ideal defined by the given system is made possible. We restrict our attention to analytic solutions on connected open subsets of  $\mathbb{C}^n$ . Before stating the definition of a simple differential system, we elaborate on certain formal manipulations of differential polynomials, on which Thomas' algorithm is based.

Let  $R := K\{u_1, \dots, u_m\}$  be the differential polynomial ring in  $u_1, \dots, u_m$  with commuting derivations  $\partial_1, \dots, \partial_n$ , where  $K$  is a computable differential field of characteristic zero (with derivations  $\partial_1|_K, \dots, \partial_n|_K$ ). We define the set

$$\Delta := \{\partial_1, \dots, \partial_n\}$$

and the (commutative) monoid  $\text{Mon}(\Delta)$  consisting of the monomials in  $\partial_1, \dots, \partial_n$ . For  $\theta \in \text{Mon}(\Delta)$  we denote by  $\deg(\theta)$  the total degree of the monomial  $\theta$ . If  $L$  is a subset of  $R$ , then  $\langle L \rangle$  is defined to be the differential ideal of  $R$  generated by  $L$ .

In what follows, we fix a ranking  $>$  on  $K\{u_1, \dots, u_m\}$  (i.e., a total ordering on

$$\text{Mon}(\Delta)u := \{(u_k)_J \mid 1 \leq k \leq m, J \in (\mathbb{Z}_{\geq 0})^n\} \quad (2.42)$$

which respects the action of the derivations and which is a well-ordering; cf. Subsect. A.3.2, p. 249, for more details). Then for every non-constant differential polynomial  $p \in R - K$ , the *leader*  $\text{ld}(p)$ , the *initial*  $\text{init}(p)$ , and the *discriminant*  $\text{disc}(p)$  are defined as in Definition 2.2.1 by considering  $p$  as polynomial in the finitely many indeterminates  $(u_k)_J$  which occur in it, totally ordered by the ranking  $>$ . For any subset  $P \subseteq R$  we define

$$\text{ld}(P) := \{\text{ld}(p) \mid p \in P, p \notin K\}.$$

**Remark 2.2.36.** For any non-constant differential polynomial  $p \in R - K$  and any  $i \in \{1, \dots, n\}$ , the defining properties of a ranking imply that the leader of  $\partial_i p$  equals  $\partial_i \text{ld}(p)$ . In fact,  $\partial_i p$  is a polynomial of degree one in  $\partial_i \text{ld}(p)$ , i.e., every proper partial derivative of a differential polynomial is *quasi-linear*. This observation implies important relations among differential polynomials in terms of polynomial division, which we discuss next.

**Definition 2.2.37.** Let  $p \in R - K$ . The *separant* of  $p$  is defined to be the differential polynomial

$$\text{sep}(p) := \frac{\partial p}{\partial \text{ld}(p)},$$

i.e., the formal partial derivative of  $p$  with respect to its leader. It is the coefficient of the leader  $\partial_i \text{ld}(p)$  of the derivative  $\partial_i p$  for any  $i \in \{1, \dots, n\}$ .

**Remark 2.2.38.** Let  $p_1 \in R$  and  $p_2 \in R - K$ . Proper partial derivatives of the leader of  $p_2$  can be eliminated from  $p_1$  by applying Euclidean pseudo-division in an appropriate way, using the fact that any proper derivative of  $p_2$  is quasi-linear (cf. Rem. A.3.6, p. 250, for details). In order to avoid to deal with a partial derivative of  $\text{ld}(p_2)$  twice, these derivatives should be processed in decreasing order with respect to the ranking. Apart from this *differential reduction*, which multiplies  $p_1$  by the separant of  $p_2$  to realize the desired cancelation, Euclidean pseudo-division modulo  $p_2$  eliminates powers of  $\text{ld}(p_2)$  in  $p_1$  whose exponents are greater than or equal to the degree of  $p_2$  in  $\text{ld}(p_2)$ . This *algebraic reduction* multiplies  $p_1$  by the initial of  $p_2$ . In both cases, the computation is performed in such a way that no fractions of non-constant differential polynomials are involved.

**Remark 2.2.39.** Using the algebraic reduction technique from the previous remark, we apply the algebraic version of Thomas' algorithm (cf. Rem. 2.2.11, Alg. 2.2.20) to a finite differential system  $S$  over  $R$ , i.e., a finite set of equations and inequations whose left hand sides are elements of  $R$  and whose right hand sides are zero. This set is viewed as an algebraic system in the finitely many indeterminates  $(u_k)_J$  which occur in it. Let us consider  $(m$ -tuples of)  $F$ -valued analytic functions as candidates for solutions, where  $F$  is an extension field of the subfield of constants of  $K$ . A solution of the system consists of one analytic function  $f_k$  of  $z_1, \dots, z_n$  for each differential indeterminate  $u_k$  such that every equation and inequation of the system is satisfied upon substitution of  $\frac{\partial^{|J|} f_k}{\partial z^J}$  for  $(u_k)_J$ ,  $J \in (\mathbb{Z}_{\geq 0})^n$ . Taylor expansion translates the problem into algebraic equations and inequations for the Taylor coefficients of a solution. It is therefore convenient to assume that  $F$  is algebraically closed. The defining properties of a simple algebraic system (cf. Def. 2.2.4) ensure that a sequence of Taylor coefficients defining a solution of the *algebraic* system corresponding to  $S_{<(u_k)_J}$ , i.e., the equations and inequations in  $S$  with leader smaller than  $(u_k)_J$ , can be adjusted to be a sequence defining a solution of the algebraic system corresponding to  $S_{\leq(u_k)_J}$ . However, differential consequences of  $S$  must also be taken into account, which may again be equations with a smaller leader (cf. also the discussion leading to Remark 2.1.67).

Recall that Thomas' algorithm splits systems according to vanishing or non-vanishing initials so that pseudo-divisions do not change the total solution set. It also splits systems according to the possible square-free parts until every left hand side in each system is a square-free polynomial (for every possible specialization). Let us assume that a differential system is a simple algebraic system in the above sense. Then the discriminant of each equation is non-zero when evaluated at any solution of the system. Note that in the differential algebra context the discriminant is essentially the resultant of the differential polynomial and its separant (cf. Def. 2.2.1). Moreover, the initial of any partial derivative of a differential polynomial is equal to the separant. Therefore, pseudo-division modulo partial derivatives of equations of a system that is simple in the above sense transforms a differential polynomial into

an equivalent one. (For comments about singular solutions of a differential system, cf. Remark 2.2.59.)

We adopt the following piece of notation from the case of algebraic systems. For any differential system

$$S = \{p_i = 0, q_j \neq 0 \mid i \in I, j \in J\}, \quad p_i, q_j \in R,$$

where  $I$  and  $J$  are index sets, we denote by

$$S^= := \{p_i \mid i \in I\}, \quad S^\neq := \{q_j \mid j \in J\}$$

the set of left hand sides of equations and inequations in  $S$ , respectively.

Given a set of differential polynomials which are the left hand sides of the equations of a simple algebraic system, the following algorithm performs differential reductions in order to eliminate leaders which are proper partial derivatives of other leaders in the system. This is a preparatory step for computing a cone decomposition of the multiple-closed set (with respect to the action of  $\text{Mon}(\Delta)$ ) generated by the leaders, which is discussed afterwards.

**Algorithm 2.2.40** (*Auto-reduce for differential polynomials*).

**Input:**  $L \subset R - K$  finite and a ranking  $>$  on  $R$  such that  $L = S^=$  for some finite differential system  $S$  which is simple as an algebraic system (in the finitely many indeterminates  $(u_k)_J$  which occur in it, totally ordered by  $>$ )

**Output:**  $a \in \{\text{true}, \text{false}\}$  and  $L' \subset R - K$  finite such that

$$\langle L' \rangle : q^\infty = \langle L \rangle : q^\infty, \quad q := \prod_{p \in L} \text{sep}(p), \quad (2.43)$$

and, in case  $a = \text{true}$ , there exists no  $p_1, p_2 \in L'$ ,  $p_1 \neq p_2$ , such that we have  $\text{ld}(p_1) \in \text{Mon}(\Delta)\text{ld}(p_2)$

**Algorithm:**

- 1:  $L' \leftarrow L$
- 2: **while** there exist  $p_1, p_2 \in L'$ ,  $p_1 \neq p_2$  and  $\theta \in \text{Mon}(\Delta)$  such that we have  $\text{ld}(p_1) = \theta \text{ld}(p_2)$  **do**
- 3:    $L' \leftarrow L' - \{p_1\}$ ;  $v \leftarrow \text{ld}(p_1)$
- 4:    $r \leftarrow \text{sep}(p_2) \cdot p_1 - \text{init}(p_1) \cdot v^{d-1} \cdot \theta p_2$ , where  $d := \deg_v(p_1)$
- 5:   **if**  $r \neq 0$  **then**
- 6:     **return**  $(\text{false}, L' \cup \{r\})$
- 7:   **end if**
- 8: **end while**
- 9: **return**  $(\text{true}, L')$

- Remarks 2.2.41.** a) Since  $L$  is the set of left hand sides of equations in a simple algebraic system, we have  $L \cap K = \emptyset$ . For the same reason,  $L'$  is a triangular set with respect to the ranking  $>$  in the first round of the loop and, while  $r = 0$ , also in later rounds. Therefore,  $\deg(\theta) > 0$  holds inside the loop, and step 4 eliminates  $\text{ld}(p_1)$  from  $p_1$  (cf. also Rem. A.3.6 a), p. 250). Since this process can be understood as replacing the largest term (possibly multiplied by a polynomial with smaller leader) with a sum of terms that are smaller with respect to  $>$ , and since  $>$  is a well-ordering, the algorithm terminates. By Remark 2.2.39,  $\text{sep}(p_2)$  does not vanish when evaluated at any solution of the system. Hence, (2.43) holds. Correctness of the algorithm is clear.
- b) Note that, once we have  $r \neq 0$ , the equality (2.43) still holds, but further reductions as in step 4 would not be guaranteed to respect the solution set (when  $p_2 = r$  is chosen as a divisor). Therefore,  $L' \cup \{r\}$  is returned immediately in this case with the intention that the algebraic consequences of this system are examined by the algebraic version of Thomas' algorithm, which also takes care of the initials and separants of the system.
- c) For efficiency reasons it is desirable to find pseudo-remainders in step 4 with least possible leader with respect to the ranking (if not constant), because these lend themselves to be divisors of many other polynomials of the system. Therefore,  $p_2$  and then  $p_1$  should be chosen with least possible leaders<sup>12</sup> in step 2.

**Remark 2.2.42.** We apply the combinatorics of Janet division (cf. Subsect. 2.1.1) in the present context in order to construct a generating set (in an appropriate sense) of all differential polynomial consequences of a finite system of polynomial partial differential equations (ignoring for a moment necessary splittings of systems). Let  $p_1, \dots, p_s \in R - K$  be non-constant differential polynomials. The chosen ranking on  $R$  uniquely determines  $\theta_1, \dots, \theta_s \in \text{Mon}(\Delta)$  and  $k_1, \dots, k_s \in \{1, \dots, m\}$  such that

$$\text{ld}(p_i) = \theta_i u_{k_i}, \quad i = 1, \dots, s.$$

We interpret  $p_1, \dots, p_s$  as left hand sides of PDEs. Then every partial derivative of each  $p_i$  is the left hand side of a consequence of the system. Therefore, for each  $k \in \{1, \dots, m\}$ , the set of  $\theta \in \text{Mon}(\Delta)$  such that  $\theta u_k$  is the leader of an equation that is a consequence of the system is  $\text{Mon}(\Delta)$ -multiple-closed. Hence,  $\Delta$  serves as the set  $X$  of symbols referred to in Subsect. 2.1.1. We assume that a total ordering on  $\Delta$  is chosen which is used by Algorithms 2.1.6 and 2.1.8 to construct Janet decompositions of multiple-closed sets of monomials in  $\Delta$  and their complements, respectively. The choice of the ranking  $>$  on  $R$  and the choice of the total ordering on  $\Delta$  are independent, the former one singling out the leader of each non-constant differential polynomial, the latter one determining Janet decompositions. The symbol  $>$  will continue to refer to the ranking on  $R$ .

---

<sup>12</sup> More information about the polynomials at hand should be taken into account to enhance this basic heuristic strategy, because it turns out that an implementation is slowed down drastically when polynomials get too large (measured in number of terms, say), so that it may be reasonable to trade compactness against rank.

**Definition 2.2.43.** Let  $M \subseteq \text{Mon}(\Delta)u$  (cf. (2.42)). For  $k \in \{1, \dots, m\}$  we define  $M_k := \{\theta \in \text{Mon}(\Delta) \mid \theta u_k \in M\}$ .

- a) We call the set  $M$  *multiple-closed* if  $M_1, \dots, M_m$  are  $\text{Mon}(\Delta)$ -multiple-closed. A set  $G \subseteq \text{Mon}(\Delta)u$  such that  $G_1, \dots, G_m$  are generating sets for  $M_1, \dots, M_m$ , respectively, where

$$G_k := \{\theta \in \text{Mon}(\Delta) \mid \theta u_k \in G\},$$

is called a *generating set* for  $M$ . The multiple-closed set generated by  $G$  is denoted by

$$[G] := \text{Mon}(\Delta)G = \bigcup_{k=1}^m \text{Mon}(\Delta)G_k u_k.$$

- b) Let  $M$  be multiple-closed. For  $k = 1, \dots, m$ , let

$$\{(\theta_1^{(k)}, \mu_1^{(k)}), \dots, (\theta_{t_k}^{(k)}, \mu_{t_k}^{(k)})\}$$

be a Janet decomposition of  $M_k$  (or of  $\text{Mon}(\Delta) - M_k$ , cf. Def. 2.1.11, p. 15). Then

$$\bigcup_{k=1}^m \{(\theta_1^{(k)} u_k, \mu_1^{(k)}), \dots, (\theta_{t_k}^{(k)} u_k, \mu_{t_k}^{(k)})\}$$

is called a *Janet decomposition* of  $M$  (resp. of  $\text{Mon}(\Delta)u - M$ ). The *cones* of the Janet decomposition are given by  $\text{Mon}(\mu_i^{(k)})\theta_i^{(k)} u_k$ ,  $i = 1, \dots, t_k$ ,  $k = 1, \dots, m$ . If the Janet decomposition is constructed from the generating set  $G$  for  $M$ , then we call the set of generators  $\theta_i^{(k)} u_k$  of the cones the *Janet completion* of  $G$ .

For the rest of this section, we fix a total ordering on  $\Delta$  such that the Janet completion of any set  $G \subseteq \text{Mon}(\Delta)u$  is uniquely defined.

**Definition 2.2.44.** Let  $T = \{(p_1, \mu_1), \dots, (p_s, \mu_s)\}$ ,  $p_i \in R - K$ ,  $\mu_i \subseteq \Delta$ ,  $i = 1, \dots, s$ .

- a) The set  $T$  is said to be *Janet complete* if

$$\{\text{ld}(p_1), \text{ld}(p_2), \dots, \text{ld}(p_s)\}$$

equals its Janet completion and, for each  $i \in \{1, \dots, s\}$ ,  $\mu_i$  is the set of multiplicative variables of the cone with generator  $\text{ld}(p_i)$  in the Janet decomposition  $\{(\text{ld}(p_1), \mu_1), \dots, (\text{ld}(p_s), \mu_s)\}$  of  $[\text{ld}(p_1), \dots, \text{ld}(p_s)]$  (cf. Def. 2.2.43).

- b) An element  $r \in R$  is said to be *Janet reducible modulo  $T$*  if there exist a jet variable  $v \in \text{Mon}(\Delta)u$  and  $(p, \mu) \in T$  such that  $v$  occurs in  $r$  and  $v \in \text{Mon}(\mu)\text{ld}(p)$ . In this case,  $(p, \mu)$  is called a *Janet divisor* of  $r$ . If  $r$  is not Janet reducible modulo  $T$ , then  $r$  is also said to be *Janet reduced modulo  $T$* .

The following algorithm applies differential and algebraic reductions to a given differential polynomial in such a way that the remainder of these pseudo-reductions is Janet reduced modulo a given Janet complete set.



**Algorithm 2.2.45** (*Janet-reduce for differential polynomials*).

**Input:**  $r \in R$ ,  $T = \{(p_1, \mu_1), (p_2, \mu_2), \dots, (p_s, \mu_s)\}$ , and a ranking  $>$  on  $R$ , where  $T$  is Janet complete (with respect to  $>$ , cf. Def. 2.2.44)

**Output:**  $r' \in R$  and an element  $b$  of the multiplicatively closed set generated by  $\bigcup_{i=1}^s \{\text{init}(p_i), \text{sep}(p_i)\} \cup \{1\}$  such that  $r'$  is Janet reduced modulo  $T$ , and such that  $r' = r$ ,  $b = 1$  if  $T = \emptyset$ , and  $r' + \langle p_1, \dots, p_s \rangle = b \cdot r + \langle p_1, \dots, p_s \rangle$  otherwise

**Algorithm:**

```

1:  $r' \leftarrow r$ 
2:  $b \leftarrow 1$ 
3: if  $r' \notin K$  then
4:    $v \leftarrow \text{ld}(r')$ 
5:   while  $r' \notin K$  and there exist  $(p, \mu) \in T$  and  $\theta \in \text{Mon}(\mu)$  with  $\deg(\theta) > 0$ 
     such that  $v = \theta \text{ld}(p)$  do
6:      $r' \leftarrow \text{sep}(p) \cdot r' - \text{init}(r') \cdot v^{d-1} \cdot \theta p$ , where  $d := \deg_v(r')$ 
7:      $b \leftarrow \text{sep}(p) \cdot b$ 
8:   end while
9:   while  $r' \notin K$  and there exists  $(p, \mu) \in T$  with  $\text{ld}(p) = v$ ,  $\deg_v(r') \geq \deg_v(p)$ 
     do
10:     $r' \leftarrow \text{init}(p) \cdot r' - \text{init}(r') \cdot v^{d-d'} \cdot p$ , where  $d := \deg_v(r')$  and  $d' := \deg_v(p)$ 
11:     $b \leftarrow \text{init}(p) \cdot b$ 
12:   end while
13:   while there exists a coefficient  $c$  of  $r'$  (as a polynomial in  $v$ ) which is not
     Janet reduced modulo  $T$  do
14:      $(r'', b') \leftarrow \text{Janet-reduce}(c, T, >)$ 
15:     replace the coefficient  $b' \cdot c$  in  $b' \cdot r'$  with  $r''$  and replace  $r'$  with this result
16:      $b \leftarrow b' \cdot b$ 
17:   end while
18: end if
19: return  $(r', b)$ 

```

The following remarks are analogous to Remarks 2.2.18 for the algebraic case.

**Remarks 2.2.46.** a) Algorithm 2.2.45 terminates because for the recursive calls in step 14 each coefficient  $c$  of  $r'$  is either constant or has a leader which is smaller than  $v$  with respect to  $>$ , which is a well-ordering, and the properties of  $r'$  which are achieved by steps 5–12 are retained by the recursion. The uniqueness of the Janet divisor of a jet variable implies that the result of Algorithm 2.2.45 is uniquely determined for the given input, so that remarks similar to Remark 2.1.39 a), p. 28, can be made. However, as opposed to Algorithm 2.1.38, the differential and the algebraic reductions are pseudo-reductions in general.

- b) Let  $r_1, r_2 \in R$  and  $T$  be as in the input of Algorithm 2.2.45, and define  $q$  to be the product of all  $\text{init}(p_i)$  and all  $\text{sep}(p_i)$ ,  $i = 1, \dots, s$ . In general, the equality

$$r_1 + \langle p_1, \dots, p_s \rangle : q^\infty = r_2 + \langle p_1, \dots, p_s \rangle : q^\infty$$

does not imply that the results of applying *Janet-reduce* to  $r_1$  and  $r_2$ , respectively, are equal. However, later on (cf. Prop. 2.2.50) it is shown that, if  $T$  is defined by the subset  $S^-$  of equations of a simple differential system  $S$  (cf. Def. 2.2.49), then the result of applying *Janet-reduce* to  $r \in R$  is zero if and only if we have  $r \in \langle p_1, \dots, p_s \rangle : q^\infty$ . In this case, we refer to the first component  $r'$  of the output of *Janet-reduce* applied to  $r$  as the *Janet normal form of  $r$  modulo  $T$* . In order to simplify notation, we denote the result  $r'$  of *Janet-reduce* applied to  $r$ ,  $T$ ,  $>$  by  $\text{NF}(r, T, >)$ , even if  $T$  does not have the properties mentioned above.

The polynomial  $q$  defined in part b) of the previous remarks will play an important role in what follows.

**Remark 2.2.47.** The method of J. M. Thomas for treating a differential system  $S$  applies the algebraic decomposition technique (cf. Rem. 2.2.11, Alg. 2.2.20), which in general causes a splitting of the system. Restricting attention to one of these simple systems and assuming that this system is not split further, an ascending chain of multiple-closed subsets of  $\text{Mon}(\Delta)u$  is produced, which terminates by Lemma 2.1.2, p. 10. The current multiple-closed set is generated by the leaders of the equations of the differential system, from which dispensable equations have been removed by Algorithm 2.2.40 (*Auto-reduce*). If the latter algorithm finds a new differential consequence, Thomas' algorithm for algebraic systems is applied first to the augmented system, and this process is iterated until Algorithm 2.2.40 (*Auto-reduce*) confirms that the leaders of the differential equations form the minimal generating set for the multiple-closed set they generate.

Let  $G$  be the set of left hand sides of these equations. The Janet decomposition of the multiple-closed set  $[\text{Id}(G)]$  is constructed as described in Subsect. 2.1.1. To this end, Algorithm 2.1.6 (*Decompose*), p. 11, is applied, but in a slightly modified way (cf. also Rem. 2.1.41, p. 28, for the corresponding adaptation of *Decompose* for Janet's algorithm). This algorithm is applied directly to  $G$ , in the sense that its run is determined by  $\text{Id}(g)$  for  $g \in G$ , but the application of  $d \in \Delta$  to  $\text{Id}(g)$  is replaced with the application of the derivation  $d$  to  $g$ . The result  $\{(p_1, \mu_1), \dots, (p_s, \mu_s)\}$  consists of pairs of a non-constant differential polynomial  $p_i$  in  $R$  and a subset  $\mu_i$  of  $\Delta$ . The elements of  $\mu_i$  (of  $\Delta - \mu_i$ ) are called the (*non-*) *admissible derivations* for  $p_i = 0$ . In the differential version of Thomas' algorithm, *Decompose* will be applied in this adapted version.

Using the Janet decomposition, Thomas' algorithm tries to find new differential consequences by applying derivations to an equation for which they are non-admissible and computing the Janet reductions of these derivatives.

**Definition 2.2.48.** A Janet complete set  $T = \{(p_1, \mu_1), \dots, (p_s, \mu_s)\}$  (as in Definition 2.2.44 a)) is said to be *passive*, if

$$\text{NF}(d p_i, T, >) = 0 \quad \text{for all } d \in \overline{\mu_i} = \Delta - \mu_i, \quad i = 1, \dots, s$$

(where we recall that  $\text{NF}(r, T, >)$  is the result of Algorithm 2.2.45 (*Janet-reduce*) applied to  $r, T, >$ ). A system of partial differential equations  $\{p_1 = 0, \dots, p_s = 0\}$ , where  $p_i \in R - K$ ,  $i = 1, \dots, s$ , is said to be *passive* if the Janet completion of  $\{p_1, \dots, p_s\}$  (using the fixed ranking on  $R$  and the fixed total ordering on  $\Delta$ ) and the corresponding sets of admissible derivations define a passive Janet complete set.

The result of Thomas' algorithm for differential systems is a finite set of differential systems which are simple, a notion that is defined next.

**Definition 2.2.49.** Let a ranking  $>$  on  $K\{u_1, \dots, u_m\}$  and a total ordering on the set  $\Delta = \{\partial_1, \dots, \partial_n\}$  be fixed. A system  $S$  of polynomial partial differential equations and inequations

$$p_1 = 0, \quad \dots, \quad p_s = 0, \quad q_1 \neq 0, \quad \dots, \quad q_t \neq 0,$$

where  $p_1, \dots, p_s, q_1, \dots, q_t \in R - K$ ,  $s, t \in \mathbb{Z}_{\geq 0}$ , is said to be *simple* if the following three conditions are satisfied.

- a) The system  $S$  is simple as an algebraic system (in the finitely many indeterminates  $(u_k)_J$  which occur in it, totally ordered by the ranking  $>$ ).
- b) The system of partial differential equations  $\{p_1 = 0, \dots, p_s = 0\}$  is passive.
- c) The left hand sides of the inequations  $q_1 \neq 0, \dots, q_t \neq 0$  are Janet reduced modulo the left hand sides of the passive system  $\{p_1 = 0, \dots, p_s = 0\}$ .

Janet division associates (according to the chosen ordering of  $\Delta$ ) with each equation  $p_i = 0$  a set  $\mu_i \subseteq \Delta$  of *admissible derivations* in the sense that the monomials in the derivations in  $\mu_i$  are those elements of  $\text{Mon}(\Delta)$  which are potentially applied to  $p_i$  for Janet reduction of some differential polynomial. The complement  $\overline{\mu_i}$  of  $\mu_i$  in  $\Delta$  consists of the *non-admissible derivations* for  $p_i = 0$ . We refer to  $\theta p_i$ , where  $\theta \in \text{Mon}(\mu_i)$ , as an *admissible derivative* of  $p_i$ .

The next proposition gives a description in terms of a radical differential ideal of all differential equations that are consequences of a simple differential system. Janet reduction modulo the simple differential system decides membership of a differential polynomial to the corresponding radical differential ideal. The statements are analogous to Proposition 2.2.7 in the algebraic case, which is used in the proof.

**Proposition 2.2.50.** *Let a simple differential system  $S$  over  $R$  be given by*

$$p_1 = 0, \quad \dots, \quad p_s = 0, \quad q_1 \neq 0, \quad \dots, \quad q_t \neq 0.$$

*Let  $E := \langle P \rangle$  be the differential ideal of  $R$  generated by  $P := \{p_1, \dots, p_s\}$  and define the product  $q$  of the initials and separants of all elements of  $P$ . Then*

$$E : q^\infty := \{p \in R \mid q^r \cdot p \in E \text{ for some } r \in \mathbb{Z}_{\geq 0}\}$$

*is a radical differential ideal. A differential polynomial  $p \in R$  is an element of  $E : q^\infty$  if and only if the Janet normal form of  $p$  modulo  $p_1, \dots, p_s$  is zero.*

**Remark 2.2.51.** Similarly to the algebraic case (cf. Rem. 2.2.8, p. 62), the assertion of Proposition 2.2.50 does not depend on the inequations  $q_1 \neq 0, \dots, q_t \neq 0$  because it describes the radical differential ideal of all differential polynomials in  $R$  vanishing on the analytic solutions of the given simple differential system, which is not influenced by inequations (cf. also p. 98).

*Proof (of Proposition 2.2.50).* By definition of the saturation  $E : q^\infty$ , every element  $p \in R$  for which Algorithm 2.2.45 (*Janet-reduce*) yields pseudo-remainder zero is an element of  $E : q^\infty$ . Conversely, let  $p \in E : q^\infty$  be arbitrary. Then there exist  $r \in \mathbb{Z}_{\geq 0}$ ,  $k_1, \dots, k_s \in \mathbb{Z}_{\geq 0}$ , and  $c_{i,j} \in R - \{0\}$ ,  $\theta_{i,j} \in \text{Mon}(\Delta)$ ,  $j = 1, \dots, k_i$ ,  $i = 1, \dots, s$ , such that

$$q^r \cdot p = \sum_{i=1}^s \left( \sum_{j=1}^{k_i} c_{i,j} \theta_{i,j} \right) p_i. \quad (2.44)$$

Our aim is to replace each term  $c_{i,j} \theta_{i,j} p_i$  on the right hand side for which  $\theta_{i,j}$  is divisible by a derivation that is non-admissible for  $p_i = 0$ , with a suitable linear combination of derivatives of  $p_1, \dots, p_s$  not involving any non-admissible derivations (cf. also Rem. 2.1.41, p. 28). Passivity of  $\{p_1 = 0, \dots, p_s = 0\}$  (cf. Def. 2.2.49 b)) guarantees that Janet reduction (Alg. 2.2.45) computes such a linear combination if  $\theta_{i,j}$  involves only one non-admissible derivation. This computation is a pseudo-reduction in general, so that substitution of the term in question may require multiplying equation (2.44) by a suitable power of  $q$  first. Iterating this substitution process and dealing with terms as above in decreasing order with respect to the ranking constructs a representation as in (2.44), possibly with larger  $r$ , in which no  $\theta_{i,j}$  is divisible by any derivation that is non-admissible for  $p_i = 0$ . This shows that for every non-zero element  $p$  of  $E : q^\infty$  there exists a Janet divisor of  $\text{ld}(p)$  in the passive set defined by  $p_1 = 0, \dots, p_s = 0$ . Consequently, the last part of the assertion holds. (Moreover, the uniqueness of the Janet divisor of a jet variable implies the uniqueness of the representation of  $q^r \cdot p$  as in (2.44) with admissible derivations only and further conditions on  $c_{i,j}$  (for fixed  $r$ ).)

In order to prove that  $E : q^\infty$  is a radical differential ideal, let us first define, for any polynomial algebra  $K[V] \subset R$ , where  $V$  is a finite subset of  $\text{Mon}(\Delta)u$  such that  $S^\infty \subset K[V]$  and  $S^\neq \subset K[V]$ , the (non-differential) ideal  $I_V$  of  $K[V]$  which is generated by  $p_1, \dots, p_s$ . Since  $S$  is simple as an algebraic system (cf. Def. 2.2.49 a)), Proposition 2.2.7, p. 62, implies that  $I_V : q^\infty$  is a radical ideal of  $K[V]$ .

Assume that  $p \in R$  satisfies  $p^k \in E : q^\infty$  for some  $k \in \mathbb{N}$ . Using the first part of the proof, the Janet normal form of  $p^k$  modulo  $p_1, \dots, p_s$  is zero. Hence, we obtain an equation of the form (2.44), where  $p$  is replaced with  $p^k$ . Let  $p'$  be the Janet normal form of  $p$  modulo  $p_1, \dots, p_s$ . Then, using the passivity again, no proper derivative of any  $\text{ld}(p_i)$  occurs in  $p'$ . We raise the equation

$$q^{r'} \cdot p = p' + \sum_{i=1}^s \left( \sum_{j=1}^{k'_i} c'_{i,j} \theta'_{i,j} \right) p_i, \quad (2.45)$$

which is constructed by Janet reduction, to the  $k$ -th power. After arranging for the left hand sides of this power and of the equation for  $p^k$  to be equal by multiplying by a suitable power of  $q$ , the difference of the right hand sides expresses  $q^l \cdot (p')^k$  for some  $l \in \mathbb{Z}_{\geq 0}$  as an  $R$ -linear combination of  $p_1, \dots, p_s$  because the proper admissible derivatives of the  $\text{Id}(p_i)$  and hence the proper admissible derivatives of the  $p_i$  cancel. By defining the polynomial algebra  $K[V]$  appropriately (such that  $K[V]$  contains all relevant jet variables), we conclude that we have  $q^l \cdot (p')^k \in I_V$ , thus  $(p')^k \in I_V : q^\infty$ . It follows that  $p' \in I_V : q^\infty \subseteq E : q^\infty$ , and therefore,  $p \in E : q^\infty$ .  $\square$

In order to define a Thomas decomposition for differential systems, we first need to discuss the notion of solution of a differential system.

Recall from Definition 2.2.2 that the set  $\text{Sol}_{\bar{K}}(S)$  of solutions of a given algebraic system  $S$  is the set of tuples in  $\bar{K}^n$  which satisfy the equations and inequations of  $S$ . Correspondingly, we are going to define now the set of analytic solutions  $\text{Sol}_\Omega(S)$  on a certain subset  $\Omega$  of  $\mathbb{C}^n$  of a differential system  $S$ .

From now on we focus on differential equations with (complex) analytic or meromorphic coefficients and we will consider analytic solutions. Let  $\Omega$  be an open and connected subset of  $\mathbb{C}^n$  with coordinates  $z_1, \dots, z_n$  and  $K$  the field of meromorphic functions on  $\Omega$ . The differential polynomial ring  $R := K\{u_1, \dots, u_m\}$  is defined with meromorphic coefficients and with commuting derivations  $\partial_1, \dots, \partial_n$  extending partial differentiation with respect to  $z_1, \dots, z_n$  on  $K$ . Let a ranking  $>$  on  $R$  be fixed. We assume that input to the algorithms is provided in such a way that the arithmetic operations can be carried out effectively when computing with coefficients in  $K$ , that equality of such coefficients can be decided, etc.

Given a differential system, an appropriate choice of the set  $\Omega$  may often be difficult to make before the algebraic and differential consequences of the system have been analyzed. The latter task is achieved by the methods discussed in this section. The defining properties of a simple differential system imply that each PDE of such a system can locally be solved for the highest derivative of some  $u_k$ . Therefore, analytic solutions exist in some open neighborhood of any point that is sufficiently generic. (It is sufficient to exclude those points which are poles of meromorphic functions occurring in the given PDEs and those which are zeros of meromorphic functions  $f$  for which the resolution process uses division by  $f$ , cf. also Rem. 2.1.70, p. 53, for the linear case.) *Usually, we assume that  $\Omega$  is chosen in such a way that the given system has analytic solutions on  $\Omega$ .*

The following example shows that a prior choice of  $\Omega$  may in general exclude certain solutions (depending also on initial or boundary conditions).

**Example 2.2.52.** The analytic solutions of the ordinary differential equation (ODE)  $u' + u^2 = 0$  for an unknown function  $u$  of  $z$  are uniquely determined by the choice of  $u(z_0)$  for any  $z_0 \in \mathbb{C}$ . Let us choose  $z_0 = 0$ . For  $u(0) = 0$  the solution is identically zero. Given  $u(0) = \frac{1}{c}$  with  $c \in \mathbb{C} - \{0\}$ , the solution  $u(z) = \frac{1}{z+c}$  is analytic in an open neighborhood of any point in  $\mathbb{C} - \{-c\}$  and has a pole at  $z = -c$ . The open neighborhood and  $\Omega$  have to avoid the point  $-c$ . Alternatively, one may allow meromorphic solutions.

**Definition 2.2.53.** Let  $\Omega \subseteq \mathbb{C}^n$  be open and connected,  $K$  the differential field of meromorphic functions on  $\Omega$ , and  $R := K\{u_1, \dots, u_m\}$ . Let

$$S = \{p_i = 0, q_j \neq 0 \mid i \in I, j \in J\}, \quad p_i, q_j \in R,$$

where  $I$  and  $J$  are index sets. We define the *set of (complex analytic) solutions (on  $\Omega$ ) or differential variety*<sup>13</sup> of  $S$  (defined on  $\Omega$ ) by

$$\begin{aligned} \text{Sol}_\Omega(S) := \{f = (f_1, \dots, f_m) \mid f_k: \Omega \rightarrow \mathbb{C} \text{ analytic, } k = 1, \dots, m, \\ p_i(f) = 0, q_j(f) \neq 0, i \in I, j \in J\}, \end{aligned}$$

where  $p_i(f)$  and  $q_j(f)$  are obtained from  $p_i$  and  $q_j$ , respectively, by substituting  $f_k$  for  $u_k$  and the partial derivatives of  $f_k$  for the corresponding jet variables in  $u_k$ . For any set  $V$  of  $m$ -tuples of analytic functions  $\Omega \rightarrow \mathbb{C}$  the set

$$\mathcal{I}_R(V) := \{p \in R \mid p(v) = 0 \text{ for all } v \in V\}$$

is called the *vanishing ideal of  $V$  in  $R$* .

**Remark 2.2.54.** Usually, we assume that  $I$  and  $J$  are finite index sets. By the Basis Theorem of Ritt-Raudenbush (cf., e.g., Thm. A.3.22, p. 256, or [Kol73, Sect. III.4]), every system of polynomial PDEs is equivalent to a finite one, which shows that the assumption on  $I$  is without loss of generality. However, similarly to the algebraic case (cf. Rem. 2.2.3), in general, an infinite set of inequations cannot be reduced to a finite set of inequations with the same solution set. (If  $F$  is a differentially closed differential field (cf. [Kol99, pp. 580–583]), then the subsets of  $F^m$  which are sets of solutions of systems  $S$  of polynomial differential equations defined over  $F$ , i.e.,  $J = \emptyset$ , are the closed sets of the *Kolchin topology* on  $F^m$ .)

**Definition 2.2.55.** Let

$$S = \{p_i = 0, q_j \neq 0 \mid i \in I, j \in J\}, \quad p_i, q_j \in R,$$

be a system of partial differential equations and inequations, where  $I$  and  $J$  are index sets and  $J$  is finite. A *Thomas decomposition* of  $S$  or of  $\text{Sol}_\Omega(S)$  is a finite collection of simple differential systems  $S_1, \dots, S_k$  such that

$$\text{Sol}_\Omega(S) = \text{Sol}_\Omega(S_1) \uplus \dots \uplus \text{Sol}_\Omega(S_k)$$

is a partition of  $\text{Sol}_\Omega(S)$ .

The following algorithm constructs a Thomas decomposition of a given finite differential system  $S$  in finitely many steps. Note that we give a succinct presentation of such an algorithm and ignore efficiency issues. For other variants and details

<sup>13</sup> The term *differential variety* is used here mainly in contrast to the term *variety* in Definition 2.2.2 and should not be confused with the solution set of a differential system in an infinite jet space, which is also referred to as a *diffiety*, cf., e.g., [Vin84].

about the latter point we refer to [Ger08], [Ger09], [BGL<sup>+</sup>10], [BGL<sup>+</sup>12, Sect. 3]. Some remarks about implementations are also given in Subsect. 2.2.6.

Similarly to the algebraic case, a Thomas decomposition of a differential system is by no means uniquely determined. Its algorithmic construction may be enhanced by using factorization of polynomials (cf. also Remark 2.2.11). Since this possibility depends on the properties of the differential field  $K$ , we will not use factorization.

**Algorithm 2.2.56** (*DifferentialThomasDecomposition*).

**Input:** A finite differential system  $S$  over  $R$ , a ranking  $>$  on  $R$ , and a total ordering on  $\Delta$  (used by *Decompose*)

**Output:** A Thomas decomposition of  $S$

**Algorithm:**

```

1:  $Q \leftarrow \{S\}; T \leftarrow \emptyset$ 
2: repeat
3:   choose  $L \in Q$  and remove  $L$  from  $Q$ 
4:   compute a Thomas decomposition  $\{A_1, \dots, A_r\}$  of  $L$  considered as an algebraic system (cf. Rem. 2.2.11 or Alg. 2.2.20, and Rem. 2.2.39)
5:   for  $i = 1, \dots, r$  do
6:     if  $A_i = \emptyset$  then                                     // no equation and no inequation
7:       return  $\{\emptyset\}$ 
8:     else
9:        $(a, G) \leftarrow \text{Auto-reduce}(A_i^=, >)$                 // cf. Alg. 2.2.40
10:      if  $a = \text{true}$  then
11:         $J \leftarrow \text{Decompose}(G)$                             // cf. Rem. 2.2.47
12:         $P \leftarrow \{\text{NF}(d p, J, >) \mid (p, \mu) \in J, d \in \overline{\mu}\}$  // cf. Alg. 2.2.45
13:        if  $P \subseteq \{0\}$  then                                     //  $J$  is passive
14:          replace each inequation  $q \neq 0$  in  $A_i$  with  $\text{NF}(q, J, >) \neq 0$ 
15:          if  $0 \notin A_i^{\neq}$  then
16:            insert  $\{p = 0 \mid (p, \mu) \in J\} \cup \{q \neq 0 \mid q \in A_i^{\neq}\}$  into  $T$ 
17:          end if
18:          else if  $P \cap K \subseteq \{0\}$  then
19:            insert  $\{p = 0 \mid (p, \mu) \in J\} \cup \{p = 0 \mid p \in P - \{0\}\} \cup$ 
               $\{q \neq 0 \mid q \in A_i^{\neq}\}$  into  $Q$ 
20:          end if
21:        else
22:          insert  $\{p = 0 \mid p \in G\} \cup \{q \neq 0 \mid q \in A_i^{\neq}\}$  into  $Q$ 
23:        end if
24:      end if
25:    end for
26:  until  $Q = \emptyset$ 
27: return  $T$ 

```

**Theorem 2.2.57.** a) Algorithm 2.2.56 terminates and is correct.

b) Let

$$S = \{p_1 = 0, \dots, p_s = 0, q_1 \neq 0, \dots, q_t \neq 0\}$$

be a simple system in the result  $T$  of Algorithm 2.2.56. Define  $q$  to be the product of all  $\text{init}(p_i)$  and all  $\text{sep}(p_i)$ ,  $i = 1, \dots, s$ . Moreover, let  $I := \langle p_1, \dots, p_s \rangle : q^\infty$ , and let  $\mu_1, \dots, \mu_s \subseteq \Delta$  be the sets of admissible derivations of  $p_1, \dots, p_s$ , respectively, and  $J := \{(p_i, \mu_i) \mid i = 1, \dots, s\}$ . Then we have

$$\bigoplus_{i=1}^s \text{Mon}(\mu_i) \text{ld}(p_i) = \text{ld}(I).$$

For any  $r \in R$  we have

$$r \in I \iff \text{NF}(r, J, >) = 0.$$

c) Let  $C_1, \dots, C_k$  be the cones of a Janet decomposition of the complement of  $[\text{ld}(p_1), \dots, \text{ld}(p_s)]$  in  $\text{Mon}(\Delta)u$ . Then the cosets in  $R/I$  with representatives in the disjoint union  $C_1 \uplus \dots \uplus C_k$  form a maximal subset of  $R/I$  that is algebraically independent over  $K$ .

*Proof.* In order to prove that *DifferentialThomasDecomposition* terminates, it is sufficient to show that we have  $Q = \emptyset$  after finitely many steps. Apart from step 1, new elements are inserted into  $Q$  in steps 19 and 22.

In case of step 22, differential reduction in Algorithm 2.2.40 (*Auto-reduce*) computed in step 9 a non-zero differential polynomial, which is the left hand side of an equation in the new system that is inserted into  $Q$ . The algebraic version of Thomas' algorithm in step 4 will apply algebraic reductions to this system and possibly split this system. Hence, steps 4 and 9 apply algebraic and differential reductions alternately until the result  $G$  is simple as an algebraic system and for every pair  $(p, p')$  of distinct elements of  $G$ ,  $\text{ld}(p)$  is reduced with respect to  $p'$ . Similarly to the auto-reduction method without case distinctions (cf. Rem. A.3.6 c), p. 250), each system constructed in step 4 will, after finitely many steps, either be recognized as inconsistent or be turned into a system  $G$  having the above property. During this process, the differentiation order of leaders in such a system is bounded by the maximum of the differentiation orders of the leaders in the system from which the algebraic version of Thomas' algorithm started. The generation of systems is therefore governed by the algebraic splitting method for a polynomial ring in finitely many variables.

In case of step 19, Algorithm 2.2.45 (*Janet-reduce*) returned a non-constant differential polynomial  $p'$  in step 12 which is Janet reduced modulo  $J$ . Therefore, we have either  $\text{ld}(p') \notin [\text{ld}(G)]$ , or  $\text{ld}(p')$  equals  $\text{ld}(p)$  for some  $(p, \mu) \in J$  and the degree of  $p'$  in  $\text{ld}(p')$  is smaller than the corresponding one of  $p$ . In the former case, the new system  $N$  inserted into  $Q$  in step 19 will define a multiple-closed set which properly contains  $[\text{ld}(G)]$ . In the latter case, we distinguish two kinds of systems that are derived from  $N$  by the algebraic version of Thomas' algorithm in step 4 (e.g., by steps 4 and 5 of Algorithm 2.2.21). A system of the first kind contains



an equation (e.g.,  $p' = 0$ ) with leader  $\text{ld}(p) = \text{ld}(p')$ , different from  $p = 0$ , whose initial is guaranteed not to vanish, so that a pseudo-reduction of  $p$  is performed. Then the degree of  $p$  in  $\text{ld}(p)$  decreases. In a system of the second kind such a pseudo-reduction of  $p$  (as a polynomial in  $\text{ld}(p) = \text{ld}(p')$ ) is prevented as a result of equating  $\text{init}(p')$  with zero. If  $\text{init}(p')$  is constant, such a system will be discarded. Otherwise, it contains a new equation with leader  $\text{ld}(\text{init}(p'))$ . Since  $\text{init}(p')$  is Janet reduced modulo  $J$ , a pseudo-reduction of  $\text{init}(p')$  could at most be performed modulo equations originating from other elements of  $P$ . We may assume that  $p' \in P$  is chosen such that no pseudo-reduction of  $\text{init}(p')$  is possible. Again, we have either  $\text{ld}(\text{init}(p')) \notin [\text{ld}(G)]$ , or a pseudo-reduction of the left hand side of an equation with leader  $\text{ld}(\text{init}(p'))$  is performed, or the initial of the new equation is equated with zero. By iterating this argument, we conclude that either we obtain an equation whose leader is not contained in  $[\text{ld}(G)]$  or for some generator  $v$  of  $[\text{ld}(G)]$  the minimal degree of  $v$  as a leader in equations in  $G$  is decreased. Since the latter situation can occur only finitely many times, in any case the multiple-closed set  $[\text{ld}(G)]$  will be enlarged after finitely many steps.

Hence, termination follows from the termination of the algebraic version of Thomas' algorithm and Dickson's Lemma (cf. Lemma 2.1.2, p. 10).

We are going to show the correctness of *DifferentialThomasDecomposition*. The algebraic version of Thomas' algorithm only performs algebraic pseudo-reductions and splittings while maintaining the total solution set. Hence, the solution sets of the systems  $A_1, \dots, A_r$  in step 4 form a partition of the solution set of  $L$ .

In step 9, Algorithm 2.2.40 (*Auto-reduce*) applies differential reductions to equations in a system  $A_i$  which is simple as an algebraic system. An equation is replaced with its pseudo-remainder if the latter is non-zero, and Algorithm 2.2.40 stops after the first proper replacement. From the discussion in Remark 2.2.39 we conclude that this transformation does not change the solution set of  $A_i$  because the separant which is used for pseudo-division does not vanish on the solution set.

The set  $J$  constructed in step 11 is Janet complete. Therefore,  $\text{NF}(d p, J, >)$  in step 12 and  $\text{NF}(q, J, >)$  in step 14 are well-defined and are realized by applying Algorithm 2.2.45 (*Janet-reduce*). Step 12 computes left hand sides of equations that are consequences of  $A_i$ . If some of them are non-zero constants, then these consequences reveal an inconsistent differential system, which is therefore discarded. If some non-constant consequences are obtained and all constant consequences are zero, then the former ones are inserted into the system to be processed again in a later round. If  $P \subseteq \{0\}$  holds in step 13, then  $J$  is a passive set. Then conditions a) and b) of Definition 2.2.49 are satisfied, and after step 14 condition c) is also ensured if the system is not inconsistent. Thus, every differential system which is inserted into  $T$  in step 16 is simple.

In every step of *DifferentialThomasDecomposition* the solution sets of the differential systems in  $Q$  and in  $T$  form a partition of the solution set of  $S$ . Hence, the result  $T$  is a Thomas decomposition of  $S$ .

The statements in parts b) and c) of the theorem are immediate consequences of Proposition 2.2.50.  $\square$

**Remark 2.2.58.** The result of Algorithm 2.2.56 is empty if and only if the input system  $S$  is inconsistent. If it equals  $\{\emptyset\}$  (i.e., a set consisting of one empty system), then the input system admits all analytic functions on  $\Omega$  as solutions.

**Remark 2.2.59.** The notion of a *singular solution* of a differential equation dates back to the 18th century and research by, e.g., A. C. Clairaut, J.-L. Lagrange, P.-S. Laplace, and S. D. Poisson (cf., e.g., [Inc56, footnote on p. 87], [Kol99, Sect. 1.8]). The intuitive idea of this concept is that the solutions of a differential equation form a number of families, each of which is parametrized by a number of constants (or functions in case of underdetermined systems) which can be chosen arbitrarily. If one family is identified as constituting the *general solution* (obtained by a generic choice of the constants or functions), then all solutions which do not belong to this family are said to be singular. A more rigorous definition was given by J.-G. Darboux [Dar73, p. 158]: A solution of a differential equation is said to be singular if it is also a solution of the separant of the equation.

Thomas' algorithm splits systems according to vanishing or non-vanishing initials of equations and their partial derivatives, the initials of the latter being the separants of the equations (cf. Rem. 2.2.39). A Thomas decomposition of a differential system therefore allows to detect singular solutions of the system. More generally, the problem of determining how singular solutions are distributed among irreducible components of a differential variety was a major motivation for J. F. Ritt to develop differential algebra. For more details and contributions to this problem we refer to [Ham93], [Rit36], [Hub97, Hub99], [Kol99, Sect. 1.8], and the references therein.

**Example 2.2.60.** In order to investigate the singular solutions of the ordinary differential equation (with non-constant coefficients)

$$\frac{dU^2}{dt} - 4t \frac{dU}{dt} - 4U + 8t^2 = 0,$$

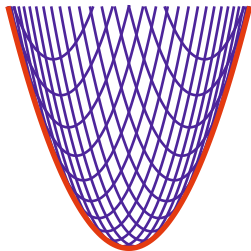
we are going to compute a Thomas decomposition of this system. We denote by  $R$  the differential polynomial ring  $\mathbb{Q}(t)\{u\}$  in one differential indeterminate  $u$  with derivation  $\partial_t$ , which restricts to formal differentiation with respect to  $t$  on  $\mathbb{Q}(t)$ . Let the differential polynomial corresponding to the left hand side be

$$p := u_t^2 - 4t u_t - 4u + 8t^2.$$

No case distinction is necessary for the initial. The separant of  $p$  equals  $2u_t - 4t$ . Euclidean division applied to  $p$  and  $\text{sep}(p)$  (as polynomials in  $u_t$ ) yields  $u - t^2$ , which is, up to a constant factor, the discriminant of  $p$  as a polynomial in  $u_t$ . We obtain the following Thomas decomposition (where the set of admissible derivations is also recorded for each equation):

$\underline{u}_t^2 - 4t \underline{u}_t - 4u + 8t^2 = 0 \quad \{ \partial_t \}$ $\underline{u} - t^2 \neq 0$	$\underline{u} - t^2 = 0 \quad \{ \partial_t \}$
--	--

The analytic solutions of the first system are given by  $U(t) = 2((t+c)^2 + c^2)$ , where  $c$  is an arbitrary (real or complex) constant. The solution  $U(t) = t^2$  of the second system is an essential singular solution. Considering all real analytic solutions at the same time, the singular solution is distinguished as an envelope of the general solution.



**Fig. 2.3** A visualization of the essential singular solution in Example 2.2.60 as an envelope

**Example 2.2.61.** We are going to compute a Thomas decomposition of

$$\frac{\partial U}{\partial t} - 6U \frac{\partial U}{\partial x} + \frac{\partial^3 U}{\partial x^3} = 0, \quad U \frac{\partial^2 U}{\partial t \partial x} - \frac{\partial U}{\partial t} \frac{\partial U}{\partial x} = 0,$$

the differential system given by the Korteweg-de Vries equation (cf., e.g., [BC80]) and another partial differential equation for  $U(t, x)$  to be discussed in Sect. 3.3 (cf. also Ex. 3.3.49, p. 228).

Let  $R := K\{u\}$  be the differential polynomial ring in one differential indeterminate  $u$  with commuting derivations  $\partial_t, \partial_x$  over a differential field  $K$  of characteristic zero (with derivations  $\partial_t|_K, \partial_x|_K$ ). The jet variable  $u_{(i,j)}$ ,  $i, j \in \mathbb{Z}_{\geq 0}$ , will also be denoted by  $u_{t^i, x^j}$ . We set

$$p := u_t - 6uu_x + u_{x,x,x}, \quad q := uu_{t,x} - u_t u_x$$

and choose the degree-reverse lexicographical ranking on  $R$  satisfying  $u_t > u_x$  (cf. Ex. A.3.3, p. 250).

We have  $\text{ld}(p) = u_{x,x,x}$ ,  $\text{ld}(q) = u_{t,x}$ ,  $\text{init}(p) = 1$ , and  $\text{init}(q) = u$ . Hence,

$$\{p = 0, q = 0\}$$

is a triangular set. We replace this system with two systems

$$\{p = 0, q = 0, u = 0\}, \quad \{p = 0, q = 0, u \neq 0\}$$

according to vanishing or non-vanishing initial of  $q$ . (No case distinctions are necessary for the separants.) The first system is equivalent to the simple differential

system

$$S_1 := \{u = 0\}.$$

The second system is simple as an algebraic system (cf. Def. 2.2.49 a)), but not passive. We define  $\Delta := \{\partial_t, \partial_x\}$  and give  $\partial_t$  priority over  $\partial_x$  for Janet division (cf. Alg. 2.1.6 and Def. 2.2.43). Then the admissible derivations for  $p$  and  $q$  are given by  $\{\partial_x\}$  and  $\{\partial_t, \partial_x\}$ , respectively. Janet reduction of  $\partial_t p$  modulo  $\{(p, \{\partial_x\}), (q, \{\partial_t, \partial_x\})\}$  yields the following non-zero pseudo-remainder:

$$r := u(u p_t - q_{x,x}) - u u_t p + u_x q_x = u^2 \underline{u_{t,t}} - u(6u^2 - u_{x,x}) u_{t,x} - u_t u_x u_{x,x} - u u_t^2.$$

The augmented system  $\{p = 0, q = 0, r = 0, u \neq 0\}$  is simple as an algebraic system, and the passivity check only involves Janet reduction of  $\partial_t q$  modulo  $\{(p, \{\partial_x\}), (q, \{\partial_t, \partial_x\}), (r, \{\partial_t, \partial_x\})\}$ . The result is:

$$\begin{aligned} s &:= u((u q_t - r_x) - (6u^2 - u_{x,x}) q_x + q p) + 3u_x r + 3(2u^2 u_x - u u_t - u_x u_{x,x}) q \\ &= 6u^3 u_t \underline{u_{x,x}}. \end{aligned}$$

We have  $\text{init}(s) = 6u^3 u_t$ . Now,  $\text{init}(s) \neq 0$  implies  $u_{x,x} = 0$ , which results in the simple system

$$S_2 := \{u_t - 6u u_x = 0, u_{x,x} = 0, u \neq 0\}.$$

On the other hand,  $\text{init}(s) = 0$  implies  $u_t = 0$ , hence the simple system

$$S_3 := \{u_t = 0, u_{x,x,x} - 6u u_x = 0, u_{x,x} \neq 0, u \neq 0\}.$$

$u = 0 \{ \partial_t, \partial_x \}$	$\underline{u_t} - 6u u_x = 0 \{ \partial_t, \partial_x \}$ $u_{x,x} = 0 \{ *, \partial_x \}$ $u \neq 0$	$u_t = 0 \{ \partial_t, \partial_x \}$ $\underline{u_{x,x,x}} - 6u u_x = 0 \{ *, \partial_x \}$ $u_{x,x} \neq 0$ $u \neq 0$
--------------------------------------	--	---

For an explicit integration of these simple systems, cf. Example 3.3.49, p. 228.

### 2.2.3 The Generic Simple System for a Prime Ideal

In this subsection we prove that in every Thomas decomposition of a prime (algebraic or differential) ideal there exists a unique simple system that is in a precise sense the most generic one in the decomposition. Moreover, a corollary to Theorem 2.2.57 of the previous subsection is obtained, which shows how membership to a radical (algebraic or differential) ideal can be decided using a Thomas decomposition.

The statements below will be at the same time about algebraic and differential systems using the following notation.

For the rest of this section,  $R$  denotes either the commutative polynomial algebra  $K[x_1, \dots, x_n]$  over a field  $K$  of characteristic zero or the differential polynomial ring  $K\{u_1, \dots, u_m\}$ , where  $K$  is the field of meromorphic functions on a connected open subset  $\Omega$  of  $\mathbb{C}^n$ , as in the previous subsection. If  $S$  is an algebraic or differential system, we will write  $\text{Sol}(S)$  referring to either  $\text{Sol}_{\overline{K}}(S)$  (cf. Def. 2.2.2) or  $\text{Sol}_{\Omega}(S)$  (cf. Def. 2.2.53). We will also use  $\langle P \rangle$  to denote the ideal and the differential ideal, respectively, generated by the set  $P$  depending on the context, and  $\mathcal{I}_R$  is a notation for the vanishing ideal in both cases.

Recall that for both algebraic and differential systems

$$S = \{p_1 = 0, \dots, p_s = 0\}$$

of equations over  $R$  a theorem holds, called Nullstellensatz, which states that, if  $p \in R$  vanishes on  $\text{Sol}(S)$ , then some power of  $p$  is an element of  $E := \langle p_1, \dots, p_s \rangle$ , i.e., we have

$$\mathcal{I}_R(\text{Sol}(S)) = \sqrt{E}. \quad (2.46)$$

(The theorem is due to D. Hilbert in the algebraic case and to J. F. Ritt and H. W. Raudenbush in the differential case; cf. also Thm. A.3.24, p. 258). In particular, if  $\langle p_1, \dots, p_s \rangle$  is a radical ideal, then  $\mathcal{I}_R(\text{Sol}(S)) = \langle p_1, \dots, p_s \rangle$  holds.

The following lemma generalizes (2.46) and will be essential in what follows.

**Lemma 2.2.62.** *Let*

$$S = \{p_i = 0, q_j \neq 0 \mid i \in I, j \in J\}, \quad p_i, q_j \in R,$$

*be a (not necessarily simple) system, where  $I$  and  $J$  are index sets and  $J$  is finite. Define  $E := \langle p_i \mid i \in I \rangle$  and  $q := \prod_{j \in J} q_j$ . Then we have*

$$\mathcal{I}_R(\text{Sol}(S)) = \sqrt{E : q^\infty}.$$

*Proof.* Let  $f \in R$ . If  $q^r f^s \in E$  for some  $r \in \mathbb{Z}_{\geq 0}$  and  $s \in \mathbb{N}$ , then  $f(x)^s = 0$  for all  $x \in \text{Sol}(S)$  because  $q(x)^r \neq 0$  for all  $x \in \text{Sol}(S)$ . Since  $\text{Sol}(S)$  is a subset of an integral domain, we have  $f(x) = 0$  for all  $x \in \text{Sol}(S)$ , i.e.,  $f \in \mathcal{I}_R(\text{Sol}(S))$ . Conversely,  $f(x) = 0$  for all  $x \in \text{Sol}(S)$  implies  $(qf)(x) = 0$  for all  $x \in \text{Sol}(\{p_i = 0 \mid i \in I\})$ . By the Nullstellensatz, there exists  $s \in \mathbb{N}$  such that  $(qf)^s \in E$ . It follows that  $f \in \sqrt{E : q^\infty}$ .  $\square$

For any set  $V$  of elements of  $\overline{K}^n$  or of  $m$ -tuples of analytic functions on  $\Omega$ , we define

$$\overline{V} := \text{Sol}(\{p = 0 \mid p \in \mathcal{I}_R(V)\}),$$

which is a shorthand notation used in the next corollary and in what follows, and which is reminiscent of the closure with respect to the Zariski topology. (We expect no confusion with the notation  $\overline{K}$  for an algebraic closure of  $K$ .)

**Corollary 2.2.63.** *Let  $S$  be a (not necessarily simple) system as in Lemma 2.2.62,  $E := \langle p_i \mid i \in I \rangle$ , and  $q := \prod_{j \in J} q_j$ . Then we have*

$$\overline{\text{Sol}(S)} - \text{Sol}(S) = \text{Sol}(\{p = 0 \mid p \in \sqrt{E : q^\infty}\} \cup \{q = 0\}).$$

*In particular,  $\overline{\text{Sol}(S)} - \text{Sol}(S)$  is closed (i.e., equals its closure).*

*Proof.* Since on the one hand all polynomials in  $\sqrt{E : q^\infty}$  vanish on  $\text{Sol}(S)$  and on the other hand  $p_i \in \sqrt{E : q^\infty}$  for all  $i \in I$ , we have

$$\text{Sol}(S) = \text{Sol}(\{p = 0 \mid p \in \sqrt{E : q^\infty}\} \cup \{q \neq 0\}).$$

A reformulation of the statement of the previous lemma is:

$$\overline{\text{Sol}(S)} = \text{Sol}(\{p = 0 \mid p \in \sqrt{E : q^\infty}\}).$$

Now an elementary observation shows that the claim holds.  $\square$

A central result of this subsection is the corollary to the next proposition. The proof of the corollary uses the following lemma.

**Lemma 2.2.64.** *Let  $S$  and  $S'$  be systems as in Lemma 2.2.62 satisfying  $\text{Sol}(S') \neq \emptyset$ ,  $\text{Sol}(S) \cap \text{Sol}(S') = \emptyset$ , and  $\overline{\text{Sol}(S)} \subseteq \overline{\text{Sol}(S')}$ . Then we have  $\overline{\text{Sol}(S)} \neq \overline{\text{Sol}(S')}$ .*

*Proof.* It follows from  $\text{Sol}(S) \cap \text{Sol}(S') = \emptyset$  and

$$\text{Sol}(S) \subseteq \overline{\text{Sol}(S')} = \text{Sol}(S') \uplus (\overline{\text{Sol}(S')} - \text{Sol}(S'))$$

that  $\text{Sol}(S)$  is a subset of  $\overline{\text{Sol}(S')} - \text{Sol}(S')$ . By Corollary 2.2.63,  $\overline{\text{Sol}(S')} - \text{Sol}(S')$  is closed. Since it is a proper subset of  $\overline{\text{Sol}(S')}$ , the claim follows.  $\square$

**Proposition 2.2.65.** *Suppose that  $p_1, \dots, p_s \in R$  generate a prime (algebraic or differential) ideal of  $R$  and that*

$$V := \text{Sol}(\{p_1 = 0, \dots, p_s = 0\})$$

*is the union of finitely many non-empty sets  $V_1, \dots, V_k$  of the form  $V_i = \text{Sol}(S_i)$  for (not necessarily simple) systems  $S_i$  of equations and inequations over  $R$ . Then we have  $V = \overline{V_i}$  for some  $i \in \{1, \dots, k\}$ .*

*Proof.* By the Nullstellensatz and since  $p_1, \dots, p_s$  generate a radical ideal, we have

$$\langle p_1, \dots, p_s \rangle = \mathcal{I}_R(V) = \mathcal{I}_R(V_1) \cap \dots \cap \mathcal{I}_R(V_k).$$

Each vanishing ideal  $\mathcal{I}_R(V_j)$  has a representation as intersection of finitely many prime ideals, all of which clearly contain  $\mathcal{I}_R(V)$ . The uniqueness of the minimal representation of a radical ideal in such a form implies that the prime ideal  $\langle p_1, \dots, p_s \rangle$  occurs in the minimal representation of at least one  $\mathcal{I}_R(V_i)$ , and we have  $\mathcal{I}_R(V) = \mathcal{I}_R(V_i)$ . It follows that  $V = \overline{V_i}$ .  $\square$

**Corollary 2.2.66.** *Suppose that  $p_1, \dots, p_s \in R$  generate a prime ideal of  $R$  and let  $S_1, \dots, S_k$  be a Thomas decomposition of  $V := \text{Sol}(\{p_1 = 0, \dots, p_s = 0\})$ . Then there exists a unique  $i \in \{1, \dots, k\}$  such that  $\text{Sol}(S_i) = V$ . Moreover, the prime ideal  $\mathcal{J}_R(\text{Sol}(S_i)) = \langle p_1, \dots, p_s \rangle$  is a proper subset of  $\mathcal{J}_R(\text{Sol}(S_j))$  for every  $j \neq i$ .*

*Proof.* The existence of  $i$  follows from Proposition 2.2.65 applied to  $V_j := \text{Sol}(S_j)$ ,  $j = 1, \dots, k$ . On the other hand, let  $j \in \{1, \dots, k\}$ ,  $j \neq i$ . Then we have  $\text{Sol}(S_i) \neq \emptyset$ ,  $\text{Sol}(S_j) \neq \emptyset$ , and  $\text{Sol}(S_i) \cap \text{Sol}(S_j) = \emptyset$ . By Lemma 2.2.64, equality of  $\text{Sol}(S_i)$  and  $\text{Sol}(S_j)$  is impossible. This proves the uniqueness.

We have  $\mathcal{J}_R(\text{Sol}(S_j)) = \mathcal{J}_R(\text{Sol}(S_i))$  for all  $j \in \{1, \dots, k\}$  and furthermore  $\mathcal{J}_R(\text{Sol}(S_i)) = \mathcal{J}_R(V)$ . Since  $V$  is the union of  $\text{Sol}(S_1), \dots, \text{Sol}(S_k)$ , we have

$$\mathcal{J}_R(\text{Sol}(S_i)) = \mathcal{J}_R(\text{Sol}(S_1)) \cap \dots \cap \mathcal{J}_R(\text{Sol}(S_k)).$$

Therefore,  $\mathcal{J}_R(\text{Sol}(S_i))$  is properly contained in each  $\mathcal{J}_R(\text{Sol}(S_j))$ ,  $j \neq i$ . □

**Definition 2.2.67.** Let  $S_1, \dots, S_k$  be a Thomas decomposition of a system of equations whose left hand sides generate a prime (algebraic or differential) ideal of  $R$ . The simple system  $S_i$  in Corollary 2.2.66 is called the *generic simple system* of the given Thomas decomposition<sup>14</sup>.

**Example 2.2.68.** A Thomas decomposition of the Steiner quartic surface defined by

$$x^2 y^2 + x^2 z^2 + y^2 z^2 - xyz = 0 \tag{2.47}$$

is constructed in Example 2.2.35, p. 84. This surface is an irreducible variety, i.e., the ideal of  $\overline{\mathbb{Q}}[x, y, z]$  generated by the left hand side of (2.47) is prime. The generic simple system of the Thomas decomposition in Example 2.2.35 is system (2.41) because it is the only one whose solution set has Zariski closure of dimension two.

The generic simple system of a Thomas decomposition can be determined if the pairwise inclusion relations among the radical ideals  $\mathcal{J}_R(\text{Sol}(S_i)) = \langle S_i^- \rangle : q_i^\infty$  can be checked effectively, where  $q_i$  is the product of the initials (and separants in the differential case) of all elements of  $S_i^-$ . Corollary 2.2.71 below solves this problem.

As before, we deal at the same time with algebraic and differential systems  $S$ , implying that pseudo-reductions modulo  $S^\infty$  are understood to be pseudo-reductions modulo the elements of  $S^\infty$  in the algebraic case and pseudo-reductions modulo the elements of  $S^\infty$  and their derivatives in the differential case.

**Proposition 2.2.69.** *Let  $S_1$  and  $S_2$  be simple systems over  $R$ . For  $i = 1, 2$ , we define  $I_i := \langle S_i^\infty \rangle : q_i^\infty$ , where  $q_i$  is the product of the initials (and separants in the differential case) of all elements of  $S_i^\infty$ . If  $I_1$  is a prime ideal and if we have  $I_1 \subseteq I_2$  and  $\text{ld}(I_1) = \text{ld}(I_2)$ , then the equality  $I_1 = I_2$  holds.*

<sup>14</sup> The notion of *generic simple system* should not be confused with the notion of *general component* of an irreducible differential polynomial  $p$  (cf. [Kol73, Sect. IV.6] or [Rit50, p. 167]), which is a certain prime differential ideal not containing the separant of  $p$ .

*Proof.* First of all, we have the equality  $I_1 : q_2^\infty = I_1$ . In fact, if  $q_2^k \cdot p \in I_1$  for some  $k > 0$  and  $p \in R - I_1$ , then we have  $q_2 \in I_1$  because  $I_1$  is a prime ideal; thus,  $q_2$  vanishes on  $\text{Sol}(I_1)$  and hence on  $\text{Sol}(I_2)$ , which is a contradiction to the fact that  $q_2 \notin \mathcal{I}_R(\text{Sol}(S_2))$  because  $S_2$  is a simple system.

In order to prove the proposition, we show that the pseudo-remainder of every  $p_2 \in S_2^\infty$  modulo the equations in  $S_1$  (and, in the differential case, their consequences obtained by applying admissible derivations) is zero. Then it follows  $\langle S_2^\infty \rangle \subseteq I_1$  (by Thm. 2.2.57 b)) and  $I_2 = \langle S_2^\infty \rangle : q_2^\infty \subseteq I_1 : q_2^\infty = I_1$ .

Let us assume that the pseudo-remainder modulo  $S_1^\infty$  of some element of  $S_2^\infty$  is non-zero, and let  $p_2 \in S_2^\infty$  be such an element with minimal possible  $\text{ld}(p_2)$ .

The hypothesis  $\text{ld}(I_1) = \text{ld}(I_2)$  implies that there exists (an admissible derivative of) an element of  $S_1^\infty$  with the same leader as  $p_2$ . Let us denote this algebraic pseudo-divisor by  $p_1$  and the variable  $\text{ld}(p_1) = \text{ld}(p_2)$  by  $x$ . In case  $\deg_x(p_1) \leq \deg_x(p_2)$  algebraic pseudo-division of  $p_2$  modulo  $p_1$  is possible resulting in an element of  $I_2$  which is either zero or has leader smaller than  $x$  or has leader  $x$  and smaller degree in  $x$  than  $p_1$ . The first two cases are contradictions to the choice of  $p_2$ . Hence, it is sufficient to consider the case  $\deg_x(p_1) > \deg_x(p_2)$ .

The pseudo-remainder of  $p_1$  modulo  $S_2^\infty$  is zero because  $I_1 \subseteq I_2$ . Let  $q$  be the product of the initials of (the admissible derivatives of) the elements of  $S_2^\infty$  which are involved in the pseudo-reduction of  $p_1$  modulo  $S_2^\infty$ . Then we have  $q \in K$  or  $\text{ld}(q) < x$ . There exist  $k \in \mathbb{Z}_{\geq 0}$ ,  $c \in R$ , and  $r \in I_2$  with  $\text{ld}(r) < x$  such that  $q^k \cdot p_1 = c \cdot p_2 + r$ . Now  $\deg_x(p_1) > \deg_x(p_2)$  implies  $\text{ld}(c) = x$ . By the choice of  $p_2$ , the pseudo-remainder modulo  $S_1^\infty$  of (every admissible derivative of) every element of  $S_2^\infty$  with leader smaller than  $x$  is zero. Therefore, we have  $q^k \cdot p_1 - r = c \cdot p_2 \in I_1$ . The assumption  $p_2 \notin I_1$  and the fact that  $I_1$  is prime imply  $c \in I_1$ . But the pseudo-remainder of  $c$  modulo  $S_1^\infty$  is non-zero due to  $\deg_x(c) < \deg_x(p_1)$ , which is a contradiction.  $\square$

**Remark 2.2.70.** In the differential case the pseudo-divisor  $p_1$  in the proof of the previous proposition is actually not a proper derivative of an element of  $S_1^\infty$  because every partial derivative of a differential polynomial has degree one in its leader.

**Corollary 2.2.71.** *Let  $S_1, \dots, S_k$  be a Thomas decomposition of a system of equations whose left hand sides generate a prime ideal of  $R$ . Set inclusion defines a partial order on  $L := \{\text{ld}(I_1), \dots, \text{ld}(I_k)\}$ , where  $I_i := \langle S_i^\infty \rangle : q_i^\infty$  and  $q_i$  is defined to be the product of the initials (and separants in the differential case) of all elements of  $S_i^\infty$ ,  $i = 1, \dots, k$ . Then  $L$  has a unique least element  $\text{ld}(I_i)$ . It determines the generic simple system  $S_i$  among  $S_1, \dots, S_k$ .*

*Proof.* Let  $i \in \{1, \dots, k\}$  be such that  $S_i$  is the generic simple system of the given Thomas decomposition. By Corollary 2.2.66,  $I_i$  is a proper subset of  $I_j$  for every  $j \neq i$ . Hence, we have  $\text{ld}(I_i) \subseteq \text{ld}(I_j)$  for every  $j \neq i$ . By Proposition 2.2.69, each of these inclusions is strict.  $\square$

In the differential case, each set  $\text{ld}(I_j)$  is infinite. An effective method which determines the generic simple system of a Thomas decomposition of a prime differential ideal will be given in Proposition 2.2.82.

We draw another important conclusion from Lemma 2.2.62.



**Proposition 2.2.72.** *Let a (not necessarily simple) system be given by*

$$S = \{p_i = 0, q_j \neq 0 \mid i \in I, j \in J\}, \quad p_i, q_j \in R,$$

where  $I$  and  $J$  are index sets and  $J$  is finite. Define  $E := \langle p_i \mid i \in I \rangle$  and  $q := \prod_{j \in J} q_j$ . Moreover, let  $S_1, \dots, S_k$  be a Thomas decomposition of  $S$ , and for  $i = 1, \dots, k$ , define  $E^{(i)} := \langle S_i^- \rangle$  and the product  $q^{(i)}$  of the initials (and separants in the differential case) of all elements of  $S_i^-$ . Then we have

$$\sqrt{E : q^\infty} = (E^{(1)} : (q^{(1)})^\infty) \cap \dots \cap (E^{(k)} : (q^{(k)})^\infty).$$

*Proof.* The Thomas decomposition defines a partition  $\text{Sol}(S_1) \uplus \dots \uplus \text{Sol}(S_k)$  of  $\text{Sol}(S)$ . As a consequence of the Nullstellensatz and Propositions 2.2.7 and 2.2.50, we have  $\mathcal{I}_R(\text{Sol}(S_i)) = E^{(i)} : (q^{(i)})^\infty$  for all  $i = 1, \dots, k$ . Now, Lemma 2.2.62 implies

$$\begin{aligned} \sqrt{E : q^\infty} &= \mathcal{I}_R(\text{Sol}(S)) = \mathcal{I}_R(\text{Sol}(S_1) \uplus \dots \uplus \text{Sol}(S_k)) \\ &= \mathcal{I}_R(\text{Sol}(S_1)) \cap \dots \cap \mathcal{I}_R(\text{Sol}(S_k)) \\ &= (E^{(1)} : (q^{(1)})^\infty) \cap \dots \cap (E^{(k)} : (q^{(k)})^\infty). \end{aligned}$$

□

Membership to a radical (algebraic or differential) ideal can therefore be decided by computing the pseudo-remainder modulo every simple system in a Thomas decomposition.

**Corollary 2.2.73.** *Let  $p_1, \dots, p_s \in R$  and let  $S_1, \dots, S_k$  be a Thomas decomposition of  $\{p_1 = 0, \dots, p_s = 0\}$  (with respect to any total ordering of the indeterminates or any ranking on  $R$ ). For  $r \in R$  let  $r_i$  be the pseudo-remainder of  $r$  modulo (the equations of)  $S_i$ ,  $i = 1, \dots, k$ . Then we have*

$$r \in \sqrt{\langle p_1, \dots, p_s \rangle} \iff r_i = 0 \text{ for all } i = 1, \dots, k.$$

*Proof.* The claim follows from Theorem 2.2.57 b) and Proposition 2.2.72. □

Finally, the disjointness of the solution sets of the simple systems in a Thomas decomposition implies a statement about the corresponding prime ideals which is more general than the one given in Corollary 2.2.66.

**Proposition 2.2.74.** *Let  $S_1$  and  $S_2$  be two different simple systems in a Thomas decomposition. Moreover, for  $i = 1, 2$ , let  $\mathcal{I}_R(\text{Sol}(S_i)) = P_{i,1} \cap \dots \cap P_{i,r_i}$  be the minimal representation of the vanishing ideal as intersection of prime (differential) ideals. If a prime (differential) ideal  $P$  of  $R$  satisfies*

$$\mathcal{I}_R(\text{Sol}(S_1)) \subseteq P \quad \text{and} \quad \mathcal{I}_R(\text{Sol}(S_2)) \subseteq P,$$

*then some  $P_{i,j}$  is properly contained in  $P$ . In particular, the sets of prime ideals  $\{P_{1,1}, \dots, P_{1,r_1}\}$  and  $\{P_{2,1}, \dots, P_{2,r_2}\}$  are disjoint.*

*Proof.* Let us define  $V := \text{Sol}(\{p = 0 \mid p \in P\})$  and  $V_{i,j} := \text{Sol}(\{p = 0 \mid p \in P_{i,j}\})$ ,  $i \in \{1, 2\}$ ,  $j \in \{1, \dots, r_i\}$ . First of all, we have  $V_{i,j} \cap \text{Sol}(S_i) \neq \emptyset$  for each  $j$ . Otherwise,  $\text{Sol}(\{p = 0 \mid p \in \bigcap_{k \neq j} P_{i,k}\})$  would be a closed set containing  $\text{Sol}(S_i)$  and contained in  $\overline{\text{Sol}(S_i)}$ , and  $P_{i,j}$  would be redundant in  $P_{i,1} \cap \dots \cap P_{i,r_i}$ .

By the hypothesis of the proposition, we have  $V \subseteq \overline{\text{Sol}(S_i)}$  for  $i = 1, 2$ . According to the definition of a Thomas decomposition,  $\text{Sol}(S_1)$  and  $\text{Sol}(S_2)$  are disjoint. Using the notation  $C_i := \overline{\text{Sol}(S_i)} - \text{Sol}(S_i)$ ,  $i = 1, 2$ , we therefore have

$$V \subseteq \overline{\text{Sol}(S_1)} \cap \overline{\text{Sol}(S_2)} \subseteq C_1 \cup C_2$$

and

$$(V \cap C_1) \cup (V \cap C_2) = V.$$

Now,  $V \cap C_1$  and  $V \cap C_2$  are closed sets by Corollary 2.2.63 and because  $V$  is closed. Since the vanishing ideal of the closed set  $V$  is a prime ideal, we conclude that we have  $V \cap C_1 = V$  or  $V \cap C_2 = V$ . Hence, there exists  $i \in \{1, 2\}$  such that

$$V \subseteq C_i = \overline{\text{Sol}(S_i)} - \text{Sol}(S_i).$$

Since  $C_i$  and  $V_{i,j}$  are closed,  $C_i \cap V_{i,j}$  is closed. Moreover,  $C_i \cap V_{i,j}$  is a proper subset of  $V_{i,j}$ , because otherwise  $V_{i,j} \cap \text{Sol}(S_i) \neq \emptyset$  implies  $C_i \cap \text{Sol}(S_i) \neq \emptyset$ , which is a contradiction. Now  $P_{i,1} \cap \dots \cap P_{i,r_i} \subseteq P$  implies that there exists  $j \in \{1, \dots, r_i\}$  such that  $P_{i,j} \subseteq P$ , because  $P_{i,k}$  and  $P$  are prime. Then  $V \subseteq C_i \cap V_{i,j} \subsetneq V_{i,j}$  implies  $P_{i,j} \subsetneq P$ . The last claim of the proposition follows from the minimality of the representation of  $\mathcal{S}_R(\text{Sol}(S_i))$ ,  $i = 1, 2$ .  $\square$

## 2.2.4 Comparison and Complexity

Some comments about the relationship of simple systems in the sense of Thomas and other types of triangular sets and about the complexity of computing a Thomas decomposition are given in this subsection.

**Remark 2.2.75.** Conditions a) and b) of Definition 2.2.4, p. 61, of a simple algebraic system imply that the set  $S^\infty$  of equations of such a system  $S$  is a consistent triangular set, i.e., a triangular set admitting solutions (cf. Rem. 2.2.5). The radical ideal generated by  $S^\infty$  is a characterizable ideal, as membership to it can be decided effectively by applying pseudo-reductions (cf. [Hub03a, Hub03b]). Moreover, the set  $S^\infty$  of equations of a simple system is a regular chain (cf. [ALMM99]).

In addition to these properties, the solution sets of the simple systems in a Thomas decomposition are pairwise disjoint. In practice, achieving this requirement often is at the cost of a more involved computation. However, this strong geometric property may be used both in theory (cf., e.g., the previous subsection) and for concrete applications (e.g., for counting solutions; cf. [Ple09a] for the case of algebraic systems).

Recall that the Nullstellensatz for analytic functions (cf. Thm. A.3.24, p. 258) states that a differential polynomial  $q$  which vanishes on all analytic solutions of a system  $\{p_1 = 0, \dots, p_s = 0\}$  of polynomial partial differential equations is an element of the radical differential ideal  $I$  generated by  $p_1, \dots, p_s$ . Now by Corollary 2.2.73, based on Theorem 2.2.57 b), p. 100, membership to  $I$  can be decided using a Thomas decomposition. In order to rate the complexity of this membership problem, we would like to know estimates for the degrees and differential orders of differential polynomials occurring in a representation of  $q$  as element of  $I$ .

Let us assume that an upper bound is known for the maximum number of differentiations which need to be applied to any of the  $p_i$  such that some power of  $q$  is in the algebraic ideal generated by the  $p_i$  and their derivatives of bounded order (in a polynomial ring with finitely many indeterminates). Then the (constructive) membership problem reduces to an effective version of Hilbert's Nullstellensatz. A well-known result in this direction can be stated as follows (cf. [Bro87]).

**Theorem 2.2.76.** *Let  $p_1, \dots, p_s$  be non-zero elements of a commutative polynomial algebra over  $\mathbb{Q}$  in  $r$  indeterminates,  $d$  the maximum degree of  $p_1, \dots, p_s$ , and let  $\mu := \min\{r, s\}$ . If  $q$  is in the radical ideal generated by  $p_1, \dots, p_s$ , then there exist  $e \in \mathbb{N}$  and elements  $c_1, \dots, c_s$  of the same polynomial algebra such that*

$$q^e = \sum_{i=1}^s c_i p_i, \quad e \leq (\mu + 1)(r + 2)(d + 1)^{\mu+1}, \quad \deg(c_j) \leq (\mu + 1)(r + 2)(d + 1)^{\mu+2}$$

for all  $j = 1, \dots, s$  such that  $c_j \neq 0$ .

An upper bound for the number of differentiations necessary for the above reduction was obtained in [GKOS09]. We may assume that  $p_1, \dots, p_s, q$  are non-zero.

**Theorem 2.2.77.** *Let  $d$  be the maximum of the degrees of  $p_1, \dots, p_s, q$ , and let  $h$  be the maximum of the differential orders (of jet variables in any differential indeterminate) of the same polynomials. If  $q$  is an element of the radical differential ideal generated by  $p_1, \dots, p_s$ , then some power of  $q$  is a linear combination of  $p_1, \dots, p_s$  and their derivatives up to order at most*

$$A(n + 8, \max\{m, h, d\}),$$

where  $A$  is the Ackermann function recursively defined by

$$A(0, m) = m + 1, \quad A(n + 1, 0) = A(n, 1), \quad A(n + 1, m + 1) = A(n, A(n + 1, m)).$$

This bound, of course, allows an extreme growth of the polynomials involved in a computation of a Thomas decomposition. However, on the other hand, it applies to every other differential elimination method of this kind, e.g., the Rosenfeld-Gröbner algorithm, and no smaller bound is known up to now for the general case of partial differential equations.

For systems of ordinary differential equations, Seidenberg's elimination method (cf. [Sei56]) was improved by D. Grigoryev in [Gri89], where upper bounds for the

time complexity, the differential orders, the degrees, and the bit sizes of the resulting differential polynomials are given in terms of the corresponding data for the input.

In [DJS14] the statement of Theorem 2.2.77 was improved for the case of ordinary differential equations with constant coefficients in a field of characteristic zero. An upper bound  $L$  for the order of derivatives is given by  $(mhd)$  raised to the power  $2^{c(mh)^3}$  for some universal constant  $c > 0$ , and the exponent of  $q$  may be bounded by  $d^{m(h+L+1)}$ , where  $h$  is the maximum of 2 and the differential orders as above.

### 2.2.5 Hilbert Series for Simple Differential Systems

In this subsection we define the generalized Hilbert series for simple differential systems. It allows, in particular, to determine effectively the generic simple system in a Thomas decomposition of a prime differential ideal, which is our main application in Sect. 3.3. More benefits of the generalized Hilbert series (e.g., as indicated by the corresponding notion in the case of linear differential polynomials, cf. Subsect. 2.1.5) will be studied in the future.

We mention that alternative notions capturing in some sense the dimension of the solution set of a system of polynomial differential equations were developed by, e.g., E. R. Kolchin (cf. [Kol64]), J. Johnson (cf. [Joh69a]), A. Levin (cf., e.g., [Lev10]), and recently by M. Lange-Hegermann (cf. [LH14]).

Let  $\Omega$  be an open and connected subset of  $\mathbb{C}^n$  with coordinates  $z_1, \dots, z_n$  and  $K$  the differential field of meromorphic functions on  $\Omega$ . We denote by  $R$  the differential polynomial ring  $K\{u_1, \dots, u_m\}$  in the differential indeterminates  $u_1, \dots, u_m$  with commuting derivations  $\partial_1, \dots, \partial_n$  extending partial differentiation with respect to  $z_1, \dots, z_n$  on  $K$ . We define

$$\text{Mon}(\Delta)u := \{ \partial^J u_i \mid 1 \leq i \leq m, J \in (\mathbb{Z}_{\geq 0})^n \}$$

and fix a ranking  $>$  on  $R$ .

**Definition 2.2.78.** For any subset  $M$  of  $\text{Mon}(\Delta)u$  the *generalized Hilbert series* of  $M$  is defined by

$$H_M(\partial_1, \dots, \partial_n) := \sum_{\partial^J u_i \in M} \partial^J u_i \in \bigoplus_{i=1}^m \mathbb{Z}[[\partial_1, \dots, \partial_n]] u_i,$$

where for simplicity the differential indeterminates  $u_1, \dots, u_m$  are used here also as generators of a free  $\mathbb{Z}[[\partial_1, \dots, \partial_n]]$ -module of rank  $m$ . For  $i = 1, \dots, m$ , we define  $H_{M,i}(\partial_1, \dots, \partial_n)$  by

$$H_M(\partial_1, \dots, \partial_n) = \sum_{i=1}^m H_{M,i}(\partial_1, \dots, \partial_n) u_i,$$

and we identify  $H_M(\partial_1, \dots, \partial_n)$  with  $H_{M,1}(\partial_1, \dots, \partial_n)$  in case  $m = 1$ .

**Remark 2.2.79.** Let  $S$  be a simple differential system with set of equations

$$\{p_1 = 0, \dots, p_s = 0\}$$

and corresponding sets of admissible derivations  $\mu_1, \dots, \mu_s \subseteq \Delta$  (cf. Def. 2.2.48). We define  $q$  to be the product of all  $\text{init}(p_i)$  and all  $\text{sep}(p_i)$ ,  $i = 1, \dots, s$ . The simple differential system provides a Janet decomposition of the multiple-closed set  $[\text{ld}(p_1), \dots, \text{ld}(p_s)]$ , which defines a partition of  $M := \text{ld}(\langle p_1, \dots, p_s \rangle : q^\infty)$  by Theorem 2.2.57 b):

$$M = \bigsqcup_{i=1}^s \text{Mon}(\mu_i) \text{ld}(p_i).$$

Accordingly, the generalized Hilbert series of  $M$  is obtained from the simple system  $S$  as

$$H_M(\partial_1, \dots, \partial_n) = \sum_{i=1}^s \left( \prod_{d \in \mu_i} \frac{1}{1-d} \right) \text{ld}(p_i).$$

The simple differential system  $S$  defines a partition of  $\text{Mon}(\Delta)u$  into the set  $M$  of jet variables which occur as leaders of equations that are consequences of  $S$  and the set  $\text{Mon}(\Delta)u - M$  of remaining jet variables. We call the elements of  $M$  the *principal jet variables* and the elements of  $\text{Mon}(\Delta)u - M$  the *parametric jet variables* of the simple differential system  $S$ .

In a similar manner, a Janet decomposition of  $\overline{M} := \text{Mon}(\Delta)u - M$  with cones  $C_j = \text{Mon}(v_j)\theta_j u_{i_j}$ ,  $j = 1, \dots, k$ , allows to write the generalized Hilbert series of  $\overline{M}$  in the form

$$H_{\overline{M}}(\partial_1, \dots, \partial_n) = \sum_{j=1}^k \left( \prod_{d \in v_j} \frac{1}{1-d} \right) \theta_j u_{i_j},$$

which enumerates the parametric jet variables of  $S$  via expansion of the (formal) geometric series.

**Definition 2.2.80.** Let  $S$  be a simple differential system and  $M := \text{ld}(\langle S^\infty \rangle : q^\infty)$ , where  $q$  is the product of the initials and separants of all elements of the set  $S^\infty$  and  $\overline{M} := \text{Mon}(\Delta)u - M$ . Then the *Hilbert series counting the principal jet variables* or the *parametric jet variables* of  $S$  is the formal power series in  $\lambda$  with non-negative integer coefficients defined by

$$H_S(\lambda) := \sum_{i=1}^m H_{M,i}(\lambda, \dots, \lambda), \quad H_{\overline{S}}(\lambda) := \sum_{i=1}^m H_{\overline{M},i}(\lambda, \dots, \lambda), \text{ respectively.}$$

**Remark 2.2.81.** In the same way as a Janet basis for a system of linear partial differential equations allows to determine all (formal) power series solutions for the system (cf. Rem. 2.1.67, p. 50), a simple differential system is a formally integrable system of PDEs. Whereas in the linear case the Taylor coefficients corresponding to the principal derivatives are uniquely determined by any choice of values for the Taylor coefficients corresponding to the parametric derivatives, the equations of a simple differential system allow a finite number of values for the Taylor coeffi-

cients that are associated with the principal jet variables. If a principal jet variable is the leader of an equation of the simple system, then its degree in that equation coincides with this number because the left hand side is a square-free polynomial. Taylor coefficients for proper derivatives of such leaders are uniquely determined due to quasi-linearity (cf. Rem. 2.2.36).

By part c) of Theorem 2.2.57, p. 100, values for all Taylor coefficients corresponding to the parametric jet variables can be chosen independently. Any choice yields finitely many (formal) power series solutions of the simple system.

The following proposition describes a method which determines the generic simple system in a Thomas decomposition of a system of differential equations whose left hand sides generate a prime differential ideal (cf. Def. 2.2.67).

**Proposition 2.2.82.** *Suppose that  $p_1, \dots, p_s \in R$  generate a prime differential ideal and let  $S_1, \dots, S_k$  be a Thomas decomposition of the system  $\{p_1 = 0, \dots, p_s = 0\}$ . For  $i = 1, \dots, k$ , let  $H_{S_i}$  (and  $H_{\overline{S_i}}$ ) be the Hilbert series counting the principal jet variables (the parametric jet variables, respectively) of  $S_i$ . Comparing the sequences of Taylor coefficients lexicographically, the unique index  $i \in \{1, \dots, k\}$  for which  $H_{S_i}$  (or  $H_{\overline{S_i}}$ ) is the least (the greatest, respectively) among these Hilbert series determines the generic simple system  $S_i$  of this Thomas decomposition.*

*Proof.* For every  $j \in \mathbb{Z}_{\geq 0}$ , the coefficient of  $\lambda^j$  in the Taylor expansion of  $H_{S_i}$  equals the number of elements of  $\text{Id}(\langle S_i^\infty \rangle : q_i^\infty)$  which are jet variables of differentiation order  $j$ , where  $q_i$  is the product of the initials and separants of all elements of  $S_i^\infty$ . Obviously, the set inclusions referred to in Corollary 2.2.71 can be checked by comparing the sequences of these coefficients lexicographically.  $\square$

**Example 2.2.83.** Let  $R = K\{u\}$  be the differential polynomial ring in one differential indeterminate  $u$  with commuting derivations  $\partial_w, \partial_x, \partial_y, \partial_z$  over a differential field  $K$  of characteristic zero (with derivations  $\partial_v|_K, v \in \{w, x, y, z\}$ ). We choose the degree-reverse lexicographical ranking  $>$  on  $R$  which extends the ordering  $\partial_w u > \partial_x u > \partial_y u > \partial_z u$  (cf. Ex. A.3.3, p. 250).

Let us consider the differential system given by

$$\left\{ \begin{array}{l} u_{w,y} = u_{w,z} = u_{x,y} = u_{x,z} = 0, \quad \begin{vmatrix} u & u_w & u_y \\ u_x & u_{w,x} & 0 \\ u_z & 0 & u_{y,z} \end{vmatrix} = 0, \\ \begin{vmatrix} u_w & u_{w,x} \\ u_{w,w} & u_{w,w,x} \end{vmatrix} = \begin{vmatrix} u_x & u_{x,x} \\ u_{w,x} & u_{w,x,x} \end{vmatrix} = \begin{vmatrix} u_y & u_{y,z} \\ u_{y,y} & u_{y,y,z} \end{vmatrix} = \begin{vmatrix} u_z & u_{z,z} \\ u_{y,z} & u_{y,z,z} \end{vmatrix} = 0. \end{array} \right. \quad (2.48)$$

For reasons that will become clear in Sect. 3.3, the differential ideal of  $R$  which is generated by the left hand sides of these equations is prime. (In fact, the left hand sides of the equations in (2.48) generate the prime differential ideal of  $K\{u\}$  consisting of all differential polynomials in  $u$  which vanish under substitution of  $f_1(w) \cdot f_2(x) + f_3(y) \cdot f_4(z)$  for  $u$ , where  $f_1, \dots, f_4$  are analytic functions.)

Using the Maple package `DifferentialThomas` (cf. Subsect. 2.2.6), we obtain a Thomas decomposition of (2.48) consisting of simple systems  $S_1, \dots, S_8$ . Their Hilbert series counting the parametric jet variables are

$$\begin{aligned} H_{S_1}(\lambda) &= 1 + 4\lambda + 5\lambda^2 + \frac{4\lambda^3}{1-\lambda}, \\ H_{S_2}(\lambda) &= H_{S_3}(\lambda) = 1 + 3\lambda + 4\lambda^2 + \frac{3\lambda^3}{1-\lambda}, \\ H_{S_4}(\lambda) &= H_{S_5}(\lambda) = H_{S_6}(\lambda) = H_{S_7}(\lambda) = H_{S_8}(\lambda) = 1 + 2\lambda + \frac{2\lambda^2}{1-\lambda}. \end{aligned}$$

Proposition 2.2.82 implies that  $S_1$  is the generic simple system of the Thomas decomposition. This system is the following one:

$$\left\{ \begin{array}{l} (uu_{y,z} - u_y u_z) \underline{u_{w,x}} - u_w u_x u_{y,z} = 0, \{ \partial_w, \partial_x, \partial_y, \partial_z \}, \\ u_{w,y} = 0, \{ \partial_w, *, \partial_y, \partial_z \}, \\ u_{w,z} = 0, \{ \partial_w, *, *, \partial_z \}, \\ u_{x,y} = 0, \{ *, \partial_x, \partial_y, \partial_z \}, \\ u_{x,z} = 0, \{ *, \partial_x, *, \partial_z \}, \\ u_y \underline{u_{y,y,z}} - u_{y,y} u_{y,z} = 0, \{ *, *, \partial_y, \partial_z \}, \\ u_z \underline{u_{y,z,z}} - u_{y,z} u_{z,z} = 0, \{ *, *, *, \partial_z \}, \\ u \neq 0, \\ u_y \neq 0, \\ u_z \neq 0, \\ u \underline{u_{y,z}} - u_y u_z \neq 0. \end{array} \right. \quad (2.49)$$

For illustrative purposes we also give the generalized Hilbert series for  $S_1$ . We denote by  $E$  the differential ideal of  $R$  which is generated by the left hand sides of the equations in (2.49) and we define  $q := u_y u_z (uu_{y,z} - u_y u_z)$ . Then the generalized Hilbert series  $H_M(\partial_w, \partial_x, \partial_y, \partial_z)$  of  $M := \text{ld}(E : q^\infty)$  is determined along the lines of Remark 2.2.79:

$$\begin{aligned} & \frac{\partial_w \partial_x u}{(1-\partial_w)(1-\partial_x)(1-\partial_y)(1-\partial_z)} + \frac{\partial_w \partial_y u}{(1-\partial_w)(1-\partial_y)(1-\partial_z)} + \frac{\partial_w \partial_z u}{(1-\partial_w)(1-\partial_z)} \\ & + \frac{\partial_x \partial_y u}{(1-\partial_x)(1-\partial_y)(1-\partial_z)} + \frac{\partial_x \partial_z u}{(1-\partial_x)(1-\partial_z)} + \frac{\partial_y^2 \partial_z u}{(1-\partial_y)(1-\partial_z)} + \frac{\partial_y \partial_z^2 u}{1-\partial_z}. \end{aligned}$$

A Janet decomposition of the complement  $\overline{M}$  of  $M$  in  $\text{Mon}(\{\partial_w, \partial_x, \partial_y, \partial_z\})u$  yields the generalized Hilbert series of  $\overline{M}$ :

$$\frac{1}{1 - \partial_z} + \partial_y + \frac{\partial_x}{1 - \partial_x} + \frac{\partial_w}{1 - \partial_w} + \partial_y \partial_z + \frac{\partial_y^2}{1 - \partial_y}. \quad (2.50)$$

Expansion of this geometric series enumerates the parametric jet variables of  $S_1$ . The representation (2.50) of the generalized Hilbert series as a rational function indicates that (the essential part of) the set of analytic solutions of (2.49) can be parametrized by four arbitrary analytic functions  $f_1(w)$ ,  $f_2(x)$ ,  $f_3(y)$ ,  $f_4(z)$ . The conditions  $f'_3 \neq 0$ ,  $f'_4 \neq 0$ , which are implied by the inequations  $u_y \neq 0$ ,  $u_z \neq 0$  in  $S_1$ , are not reflected by the generalized Hilbert series (2.50). A direct comparison of the set of parametric jet variables (enumerated in (2.50)) with the Taylor coefficients of  $f_1, \dots, f_4$ , which may be chosen arbitrarily to define a solution of (2.49) of the form

$$u(w, x, y, z) = f_1(w) \cdot f_2(x) + f_3(y) \cdot f_4(z),$$

is hindered, e.g., because the Taylor coefficients of  $f_1, \dots, f_4$  of order zero add up to the Taylor coefficient of  $u$  of order zero. (Note also that (2.49) admits further solutions, e.g., certain analytic functions of the form  $f_1(w) + f_2(x) + f_3(y) + f_4(z)$ .)

## 2.2.6 Implementations

This subsection is devoted to implementations of Thomas' algorithm and also refers to related packages.

Thomas' algorithm has been implemented in the packages `AlgebraicThomas` and `DifferentialThomas` for the computer algebra system Maple by T. Bächler and M. Lange-Hegermann, respectively, at Lehrstuhl B für Mathematik, RWTH Aachen [BLH] (with the help of V. P. Gerdt and the author of this monograph). As important efficiency issues have been ignored in the above presentation of Thomas' algorithm, the implementation of these packages is not along the lines of the algorithms in the previous subsections. For instance, the packages avoid to apply pseudo-reduction to the same pair of polynomials repeatedly (which occurs in different branches of the splittings of systems). In the algebraic case, handling of the square-free part of a polynomial, being often a very expensive part of the computation, may be postponed. The growth of polynomials during the computation of a Thomas decomposition is especially severe in the differential case due to the product rule of differentiation. Strategies which counteract this growth and heuristics for choosing the equation or inequation to be treated next are needed for an efficient implementation. Factorization of polynomials (whenever possible) leads to further splittings, but the gain in simplification is usually significant. The remarks in the beginning of Subsect. 2.1.6 concerning Janet division also apply to the differential version of Thomas' algorithm. For more details, we refer to [BGL<sup>+</sup>10, BGL<sup>+</sup>12].

The package `AlgebraicThomas` computes Thomas decompositions of algebraic systems. It includes procedures which determine counting polynomials of quasi-affine or quasi-projective varieties as defined in [Ple09a]. Set-theoretic con-



structions can be applied to solution sets, e.g., forming complements and intersections. Moreover, comprehensive Thomas decompositions can be computed, i.e., Thomas decompositions of parametric systems such that specialization of parameters respects the structure of the Thomas decomposition.

The package `DifferentialThomas` implements Thomas' algorithm for differential systems. The field of coefficients for the differential polynomial ring can be any differential field supported by Maple, and the ranking may be chosen from a list of standard rankings or may be specified in terms of a matrix as discussed in Remark 3.1.39, p. 142. Building on the concept of Janet division (cf. Rem. 2.2.42 and Rem. 2.2.47), the package provides combinatorial data given by the constructed Janet decomposition, such as, e.g., Hilbert series. Moreover, it includes procedures that solve differential systems in terms of (truncated) power series or via the built-in solvers of Maple.

A very useful feature is the possibility to stop the computation of a Thomas decomposition as soon as a given number of simple systems have been constructed by the program and to output these simple systems. Options given to the program determine whether special branches or the generic branch of the splittings of systems should be preferred. The computation of the Thomas decomposition may then be continued, starting from the point where the previous computation stopped. This feature is used in applications presented in Subsect. 3.3.5.

The Maple package `epsilon` (by Dongming Wang) [Wan04] is a collection of several implementations of triangular decomposition methods. In particular, computation of Thomas decompositions of algebraic systems is possible using `epsilon`.

The packages `RegularChains` (by F. Lemaire, M. Moreno Maza, and Y. Xie) [LMMX05] and `DifferentialAlgebra` (by F. Boulier and E. S. Cheb-Terrab), formerly `diffalg` (by F. Boulier and E. Hubert, cf. also [Hub00]), are part of the standard Maple library. The latter one is now based on the BLAD libraries (written by F. Boulier in the programming language C) [Bou]. These packages compute decompositions of algebraic varieties and systems of polynomial ordinary and partial differential equations, respectively. In the former case, the decomposition is constructed in terms of regular chains [ALMM99], in the latter case the Rosenfeld-Gröbner algorithm [BLOP09] is applied, which computes finitely many characterizable ideals [Hub03a, Hub03b] whose intersection equals the radical differential ideal generated by the input. In both cases, the decomposition of the solution set is not a disjoint one, in general, but can be made disjoint in principle.

For a further comparison of these packages including timings for benchmark examples, we refer to [BGL<sup>+</sup>12, Sect. 4].

Formal Algorithmic Elimination for PDEs

Robertz, D.

2014, VIII, 283 p. 6 illus., 3 illus. in color., Softcover

ISBN: 978-3-319-11444-6