Algebraic Theory of Linear Systems: A Survey

Werner M. Seiler and Eva Zerz

Abstract An introduction into the algebraic theory of several types of linear systems is given. In particular, linear ordinary and partial differential and difference equations are covered. Special emphasis is given to the formulation of formally well-posed initial value problem for treating solvability questions for general, i. e. also under- and overdetermined, systems. A general framework for analysing abstract linear systems with algebraic and homological methods is outlined. The presentation uses throughout Gröbner bases and thus immediately leads to algorithms.

Key words: Linear systems, algebraic methods, symbolic computation, Gröbner bases, over- and underdetermined systems, initial value problems, autonomy and controllability, behavioral approach
Mathematics Subject Classification (2010): 13N10, 13P10, 13P25, 34A09, 34A12, 35G40, 35N10, 68W30, 93B25, 93B40, 93C05, 93C15, 93C20, 93C55

1 Introduction

We survey the algebraic theory of linear differential algebraic equations and their discrete counterparts. Our focus is on the use of methods from symbolic computation (in particular, the theory of Gröbner bases is briefly reviewed in Section 7) for studying structural properties of such systems, e. g., autonomy and controllability, which are important concepts in systems and control theory. Moreover, the formulation of a well-posed initial value problem is a fundamental issue with differential

Werner M. Seiler
Institut für Mathematik, Universität Kassel, 34109 Kassel, Germany
e-mail: seiler@mathematik.uni-kassel.de

Eva Zerz
Lehrstuhl D für Mathematik, RWTH Aachen, 52062 Aachen, Germany
e-mail: eva.zerz@math.rwth-aachen.de

algebraic equations, as it leads to existence and uniqueness theorems, and Gröbner bases provide a unified approach to tackle this question for both ordinary and partial differential equations, and also for difference equations.

Here are the key ideas of the algebraic approach: Given a linear differential or difference equation, we first identify a ring $\mathscr{D}$ of operators and a set $\mathscr{F}$ of functions where the solutions are sought. For instance, the equation $\ddot{f}+f=0$ can be modeled by setting $\mathscr{D}=\mathbb{R}[\partial]$ (the ring of polynomials in the indeterminate ∂ with real coefficients) and $\mathscr{F}=\mathscr{C}^\infty(\mathbb{R},\mathbb{R})$. The operator $d=\partial^2+1\in\mathscr{D}$ acts on the function $f\in\mathscr{F}$ and the given equation takes the form $df=0$. Partial differential equations with constant coefficients can be described by $\mathscr{D}=\mathbb{R}[\partial_1,\ldots,\partial_r]$ and, say, $\mathscr{F}=\mathscr{C}^\infty(\mathbb{R}^r,\mathbb{R})$. Similarly, difference equations such as the Fibonacci equation $f(t+2)=f(t+1)+f(t)$ can be put into this framework by setting $\mathscr{D}=\mathbb{R}[\sigma]$, where σ denotes the shift operator defined by $(\sigma f)(t)=f(t+1)$. Then $d=\sigma^2-\sigma-1\in\mathscr{D}$ acts on $f\in\mathscr{F}=\mathbb{R}^{\mathbb{N}_0}$, which is the set of all functions from $\mathbb{N}_0$ to $\mathbb{R}$. Again, this can easily be extended to partial difference equations by admitting several shift operators. The situation becomes more complicated when variable coefficients are involved, because then the coefficients do not necessarily commute with the operators ∂_i or σ_i, respectively. However, this setting can still be modeled by using appropriate noncommutative operator rings $\mathscr{D}$ such as the Weyl algebra (the ring of linear differential operators with polynomial coefficients). The function set $\mathscr{F}$ is supposed to have the structure of a left $\mathscr{D}$-module. This means that we may apply the operators ∂_i or σ_i arbitrarily often, and that for $f\in\mathscr{F}$, any df belongs again to $\mathscr{F}$, where $d\in\mathscr{D}$. Thus, the set of smooth functions or the set of distributions are the prototypes of such function sets in the continuous setting.

Having identified a suitable pair $(\mathscr{D},\mathscr{F})$, one may just as well treat nonscalar equations $Df=0$ with a matrix $D\in\mathscr{D}^{g\times q}$ and a vector $f\in\mathscr{F}^q$, where $(Df)_i=\sum_{j=1}^q D_{ij}f_j$ as usual. The set $S=\{f\in\mathscr{F}^q\mid Df=0\}$ is the *solution set*, in $\mathscr{F}^q$, of the linear system of equations $Df=0$. Associated with the system is the *row module* $N=\mathscr{D}^{1\times g}D$ consisting of all $\mathscr{D}$-linear combinations of the rows of the matrix D and the *system module* $M=\mathscr{D}^{1\times q}/N$, the corresponding factor module. Any $f\in\mathscr{F}^q$ gives rise to a $\mathscr{D}$-linear map $\phi(f):\mathscr{D}^{1\times q}\to\mathscr{F}$ which maps $d\in\mathscr{D}^{1\times q}$ to $df=\sum_{j=1}^q d_jf_j\in\mathscr{F}$. Now if $f\in S$ is an arbitrary solution, then any vector $d\in N$ in the row module belongs to the kernel of $\phi(f)$. Thus $\phi(f)$ induces a well-defined $\mathscr{D}$-linear map $\psi(f):M\to\mathscr{F}$ on the system module. An important observation by Malgrange [32] says that there is a bijection (actually even an isomorphism of Abelian groups with respect to addition) between the solution set S and the set of all $\mathscr{D}$-linear maps from the system module M to $\mathscr{F}$, that is,

$$S=\{f\in\mathscr{F}^q\mid Df=0\}\cong\mathrm{Hom}_{\mathscr{D}}(M,\mathscr{F}),\quad f\mapsto\psi(f).$$

One of the nice features of this correspondence is the fact that it separates the information contained in the system S into a purely algebraic object (the system module M, which depends only on the chosen operator ring and the matrix representing the system) and an analytic object (the function set $\mathscr{F}$). Thus the study of the system module is important for all possible choices of $\mathscr{F}$. This makes it possible to consider

S for successively larger function sets (smooth functions, distributions, hyperfunctions etc.) as proposed by the "algebraic analysis" school of Sato, Kashiwara et al. (see e. g. [21, 22, 34] and references therein).

This contribution is structured as follows. In the next three sections three particularly important classes of linear systems are studied separately: ordinary differential equations, difference equations and partial differential equations. The main emphasis here lies on an existence and uniqueness theory via the construction of formally well-posed initial value problems. Section 5 shows how the concept of an index of a differential algebraic equation can be recovered in the algebraic theory. Then Section 6 provides a general algebraic framework for studying abstract linear systems in a unified manner, using a common language for all the classes of linear systems considered in this paper. Here the main emphasis lies on systems theoretic aspects such as autonomy and controllability. The algebraic characterization of these properties is used throughout the paper, thus sparing the necessity of individual proofs for each system class. Finally, an appendix briefly recapitulates Gröbner bases as the main algorithmic tool for algebraic systems.

2 Linear Ordinary Differential Equations

Consider a linear ordinary differential equation

$$c_n(t)\frac{d^n f}{dt^n}(t) + \ldots + c_1(t)\frac{df}{dt}(t) + c_0(t)f(t) = 0. \tag{1}$$

The time-varying coefficients c_i are supposed to be real-meromorphic functions. Thus, the differential equation is defined on $\mathbb{R} \setminus \mathbb{E}_c$, where $\mathbb{E}_c$ is a discrete set (the collection of all poles of the functions c_i), and it is reasonable to assume that any solution f is smooth on the complement of some discrete set $\mathbb{E}_f \subset \mathbb{R}$.

For the algebraic approach, it is essential to interpret the left hand side of (1) as the result of applying a differential operator to f. For this, let $\Bbbk$ denote the field of meromorphic functions over the reals. Let $\mathscr{D}$ denote the ring of linear ordinary differential operators with coefficients in $\Bbbk$, that is, $\mathscr{D}$ is a polynomial ring over $\Bbbk$ in the formal indeterminate ∂ which represents the derivative operator, i.e., $\partial \in \mathscr{D}$ acts on f via $\partial f = \frac{df}{dt}$. Then (1) takes the form $df = 0$, where

$$d = c_n \star \partial^n + \ldots + c_1 \star \partial + c_0 \in \mathscr{D}. \tag{2}$$

Due to the Leibniz rule $\frac{d}{dt}(cf) = c\frac{df}{dt} + \frac{dc}{dt}f$, the multiplication $\star$ in $\mathscr{D}$ satisfies

$$\partial \star c = c \star \partial + \frac{dc}{dt} \quad \text{for all } c \in \Bbbk.$$

For simplicity, we will write $\partial c = c\partial + \dot{c}$ below. Thus $\mathscr{D}$ is a noncommutative polynomial ring in which the indeterminate ∂ commutes with a coefficient c according to this rule. In the language of Ore algebras (see Section 7), we have $\mathscr{D} = \Bbbk[\partial; \mathrm{id}, \frac{d}{dt}]$.

The algebraic properties of $\mathscr{D}$ can be summarized as follows: The ring $\mathscr{D}$ is a left and right Euclidean domain, that is, the product of two nonzero elements is nonzero, and we have a left and right division with remainder. The Euclidean function is given by the degree which is defined as usual, that is, $\deg(d) = \max\{i \mid c_i \neq 0\}$ for $d \neq 0$ as in (2). Thus $\mathscr{D}$ is also a left and right principal ideal domain, and it possesses a skew field $\mathscr{K}$ of fractions $k = ed^{-1}$ with $e, d \in \mathscr{D}$ and $d \neq 0$. Therefore, the *rank* of a matrix with entries in $\mathscr{D}$ can be defined as usual (i.e., as the dimension of its row or column space over $\mathscr{K}$ [26]). Moreover, $\mathscr{D}$ is also a simple ring, that is, it has only trivial two-sided ideals. These properties of $\mathscr{D}$ (see [7, 13]) imply that every matrix $E \in \mathscr{D}^{g\times q}$ can be transformed into its Jacobson form [17] by elementary row and column operations. Thus there exist invertible $\mathscr{D}$-matrices U and V such that

$$UEV = \begin{bmatrix} D & 0 \\ 0 & 0 \end{bmatrix} \tag{3}$$

with $D = \operatorname{diag}(1,\ldots,1,d) \in \mathscr{D}^{r\times r}$ for some $0 \neq d \in \mathscr{D}$, where r is the rank of E (over $\mathscr{K}$). The matrix on the right hand side of (3) is called the *Jacobson normal form* of E. (For its computation, see e.g. [30].) The existence of this noncommutative analogue of the Smith form makes the algebraic theory of linear ordinary differential equations over $\Bbbk$ very similar to the constant coefficient case (for this, one mainly uses the fact that D is a diagonal matrix and not its special form given above). The main analytical difference is that over $\Bbbk$, one has to work locally due to the presence of singularities of the coefficients and solutions. Therefore, let $\mathscr{F}$ denote the set of functions that are smooth up to a discrete set of exception points. Then $\mathscr{F}$ is a left $\mathscr{D}$-module. In [57], it was shown that $\mathscr{F}$ is even an injective cogenerator (see Subsection 6.1 for the definition). Thus the algebraic framework outlined in Section 6 can be applied to systems of linear ordinary differential equations which take the form $Ef = 0$, where E is a $\mathscr{D}$-matrix and f is a column vector with entries in $\mathscr{F}$.

Let $S = \{f \in \mathscr{F}^q \mid Ef = 0\}$ for some $E \in \mathscr{D}^{p\times q}$. Due to the Jacobson form, one may assume without loss of generality that E has full row rank. Two representation matrices E_i of S, both with full row rank, differ only by a unimodular left factor, that is, $E_2 = UE_1$ for some invertible matrix U [57]. According to Theorem 6.1, the system S is autonomous (i.e., it has no free variables) if and only if any representation matrix E has full column rank. Combining this with the observation from above, we obtain that an autonomous system always possesses a square representation matrix with full rank. Given an arbitrary representation matrix E with full row rank, we can select a square submatrix P of E of full rank. Up to a permutation of the columns of E, we have $E = [-Q, P]$. Partitioning the system variables accordingly, the system law $Ef = 0$ reads $Py = Qu$, where u is a vector of free variables, that is,

$$\forall u \in \mathscr{F}^{q-p} \exists y \in \mathscr{F}^p : Py = Qu \tag{4}$$

and it is maximal in the sense that $Py = 0$ defines an autonomous system. The number $m := q - p$, where $p = \operatorname{rank}(E)$, is sometimes called the *input dimension* of S. With this terminology, a system is autonomous if and only if its input dimension

is zero. To see that (4) holds, note that the homomorphism (of left $\mathscr{D}$-modules) $\mathscr{D}^{1\times p} \to \mathscr{D}^{1\times p}$, $x \mapsto xP$ is injective, because P has full row rank. Since $\mathscr{F}$ is an injective $\mathscr{D}$-module, the induced homomorphism (of Abelian groups w.r.t. addition) $\mathscr{F}^p \to \mathscr{F}^p$, $y \mapsto Py$ is surjective. This implies (4).

Such a representation $Py = Qu$ is called an *input-output decomposition* of S (with input u and output y). Note that it is not unique since it depends on the choice of the p linearly independent columns of E that form the matrix P. Once a specific decomposition $E = [-Q, P]$ is chosen, the input u is a free variable according to (4). For a fixed input u, any two outputs $y, \tilde{y}$ belonging to u must satisfy $P(y - \tilde{y}) = 0$. Since $Py = 0$ is an autonomous system, none of the components of y is free, and each component y_i of y can be made unique by an appropriate choice of initial conditions.

Consider an autonomous system, that is, $Py = 0$, where $P \in \mathscr{D}^{p\times p}$ has full rank. Our goal is to formulate a well-posed initial value problem. For this, one computes a minimal Gröbner basis of the row module of P with respect to an ascending POT term order (see Example 7.2), for instance

$$\partial^n \mathbf{e}_i \prec_{\mathrm{POT}} \partial^m \mathbf{e}_j \quad \Leftrightarrow \quad i < j \text{ or } (i = j \text{ and } n < m), \tag{5}$$

where $\mathbf{e}_i$ denotes the ith standard basis row. According to Definition 7.12, a Gröbner basis G is called *minimal* if for all $g \neq h \in G$, $\mathrm{lt}(g)$ does not divide $\mathrm{lt}(h)$. This means that none of the elements of a minimal Gröbner basis is superfluous, that is, we have $\langle \mathrm{lt}(G \setminus \{g\})\rangle \subsetneq \langle \mathrm{lt}(G)\rangle$ for all $g \in G$. In the next paragraph, we show that – possibly after re-ordering the generators – the result of this Gröbner basis computation is a lower triangular matrix $P' = UP$ with nonzero diagonal entries, where U is invertible.

To see this, let G be a minimal Gröbner basis of the row module of P. The minimality of G implies that there exist no $g \neq h \in G$ with $\mathrm{lt}(g) = \partial^n \mathbf{e}_i$ and $\mathrm{lt}(h) = \partial^m \mathbf{e}_i$. So for every $1 \leq i \leq p$ there is at most one $g_i \in G$ with $\mathrm{lt}(g_i) = \partial^{n_i} \mathbf{e}_i$ for some n_i. By the choice of the POT order (5), the last $p - i$ components of g_i must be zero. On the other hand, since $\mathscr{D}^{1\times p} \to \mathscr{D}^{1\times p}$, $x \mapsto xP$ is injective, the row module of P is isomorphic to the free $\mathscr{D}$-module $\mathscr{D}^{1\times p}$ and hence, it cannot be generated by less than p elements. Thus $G = \{g_1, \ldots, g_p\}$ and the matrix $P' \in \mathscr{D}^{p\times p}$ that has g_i as its ith row is lower triangular with nonzero diagonal entries.

The fact that P and P' have the same row module implies that $P' = UP$ with an invertible matrix U. Clearly, $Py = 0$ holds if and only if $P'y = 0$. Let $\rho_i := \deg(P'_{ii})$ for all $1 \leq i \leq p$. Then there exists an exception set $\mathbb{E}$ such that for all open intervals $I \subset \mathbb{R} \setminus \mathbb{E}$ and all $t_0 \in I$, the differential equation $Py = 0$ together with the initial data $y_i^{(j_i)}(t_0)$ for $1 \leq i \leq p$ and $0 \leq j_i < \rho_i$ determines $y|_I$ uniquely. This also shows that the set of solutions to $Py = 0$ on such an interval is a finite-dimensional real vector space (of dimension $\rho = \sum_{i=1}^p \rho_i$). The number ρ is also equal to the degree of the polynomial d that appears in the Jacobson form $D = \mathrm{diag}(1, \ldots, 1, d)$ of P.

Example 2.1. Consider [24, Ex. 3.1]

$$\begin{bmatrix} -t & t^2 \\ -1 & t \end{bmatrix} \dot{f} = \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix} f.$$

Writing the system in the form $Ef = 0$, the operator matrix E acting on $f = [f_1, f_2]^T$ is given by

$$E = \begin{bmatrix} -t\partial + 1 & t^2\partial \\ -\partial & t\partial + 1 \end{bmatrix} \in \mathscr{D}^{2\times 2}.$$

We compute a Gröbner basis of the row module of E with respect to the term order in (5) and obtain $E' = [1, -t]$. Indeed, E and E' have the same row module, as

$$E' = [1, -t]E \quad \text{and} \quad E = \begin{bmatrix} -t\partial + 1 \\ -\partial \end{bmatrix} E'.$$

The system S given by $Ef = 0$ is not autonomous, since $\operatorname{rank}(E) = \operatorname{rank}(E') = 1$, and thus, E does not have full column rank. In fact, the connection between the matrices E and E' shows that

$$S = \{f \in \mathscr{F}^2 \mid Ef = 0\} = \{[tf_2, f_2]^T \mid f_2 \in \mathscr{F}\},$$

that is, S has an image representation and is therefore controllable. One may interpret f_2 as the system's input and f_1 as its output (or conversely). In this example, the output is uniquely determined by the input. The Jacobson form of E is $\operatorname{diag}(1, 0)$.

Example 2.2. Consider [24, Ex. 3.2]

$$\begin{bmatrix} 0 & 0 \\ 1 & -t \end{bmatrix} \begin{bmatrix} \dot{f}_3 \\ \dot{f}_4 \end{bmatrix} = \begin{bmatrix} -1 & t \\ 0 & 0 \end{bmatrix} \begin{bmatrix} f_3 \\ f_4 \end{bmatrix} + \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}.$$

Writing the system in the form $Ef = 0$, where $f = [f_1, \ldots, f_4]^T$, one gets

$$E = \begin{bmatrix} -1 & 0 & 1 & -t \\ 0 & -1 & \partial & -t\partial \end{bmatrix} \in \mathscr{D}^{2\times 4}.$$

Proceeding as above, we obtain

$$U = \begin{bmatrix} -t\partial + 1 & t \\ -\partial & 1 \end{bmatrix} \quad \text{and} \quad E' = UE = \begin{bmatrix} t\partial - 1 & -t & 1 & 0 \\ \partial & -1 & 0 & 1 \end{bmatrix}.$$

We may choose f_1 and f_2 as inputs, and then the outputs

$$f_3 = -t\dot{f}_1 + f_1 + tf_2 \quad \text{and} \quad f_4 = -\dot{f}_1 + f_2$$

are uniquely determined according to $E'f = 0$. Note that $E = [-I, E_1]$ with an invertible matrix E_1, where $U = E_1^{-1}$, and $E' = UE = [-E_1^{-1}, I]$. Thus

$$S = \{f \in \mathscr{F}^4 \mid Ef = 0\} = \left\{ \begin{bmatrix} E_1 \\ I \end{bmatrix} u \mid u \in \mathscr{F}^2 \right\} = \left\{ \begin{bmatrix} I \\ E_1^{-1} \end{bmatrix} v \mid v \in \mathscr{F}^2 \right\}.$$

Again, S has an image representation and is therefore controllable. The Jacobson form of E is $[I, 0] \in \mathscr{D}^{2\times 4}$.

Example 2.3. Consider the system

$$\dot{f}_1 + t f_2 = 0, \quad \dot{f}_2 + t f_1 = 0.$$

Writing the system in the form $Ef = 0$, the operator matrix E acting on $f = [f_1, f_2]^T$ is given by

$$E = \begin{bmatrix} \partial & t \\ t & \partial \end{bmatrix} \in \mathscr{D}^{2\times 2}.$$

Proceeding as above, we obtain

$$U = \begin{bmatrix} \partial - \frac{1}{t} & -t \\ \frac{1}{t} & 0 \end{bmatrix} \quad \text{and} \quad E' = UE = \begin{bmatrix} \partial^2 - \frac{1}{t}\partial - t^2 & 0 \\ \frac{1}{t}\partial & 1 \end{bmatrix}.$$

Clearly, the system S given by $Ef = 0$ is autonomous and has vector space dimension 2. The Jacobson form of E is $\mathrm{diag}(1, \partial^2 - \frac{1}{t}\partial - t^2)$.

3 Linear Difference Equations

Consider a linear ordinary difference equation

$$c_n f(t+n) + \ldots + c_1 f(t+1) + c_0 f(t) = 0 \quad \text{for all } t \in \mathbb{N}_0.$$

The coefficients c_i are supposed to be elements of a commutative quasi-Frobenius ring R (see e.g. [25]). This means that (i) R is Noetherian, and (ii) $\mathrm{Hom}_R(\cdot, R)$ is an exact functor. For instance, all fields are quasi-Frobenius rings, but also the residue class rings $\mathbb{Z}/k\mathbb{Z}$ for an integer $k \geq 2$. The ring of operators is $\mathscr{D} = R[\sigma]$, a univariate polynomial ring with coefficients in R. The action of $\sigma \in \mathscr{D}$ on a sequence $f : \mathbb{N}_0 \to R$ is given by the left shift $(\sigma f)(t) = f(t+1)$. We set $\mathscr{F} = R^{\mathbb{N}_0}$, which is the set of all functions from $\mathbb{N}_0$ to R. Then $\mathscr{F}$ is a left $\mathscr{D}$-module and an injective cogenerator [31, 36, 58].

Example 3.1. Consider the Fibonacci equation

$$f(t+2) = f(t+1) + f(t)$$

over R, which can be written as $df = 0$ with $d = \sigma^2 - \sigma - 1 \in R[\sigma]$. Over the real numbers, one of its solutions is the famous Fibonacci sequence $0, 1, 1, 2, 3, 5, 8, \ldots$ Over a finite ring however, its solutions are periodic functions, because there exists a positive integer p such that $\sigma^2 - \sigma - 1$ divides $\sigma^p - 1$. (This is due to the fact that the element $\bar{\sigma} \in S := R[\sigma]/\langle \sigma^2 - \sigma - 1 \rangle$ belongs to the group of units of the finite ring S and hence, it has finite order.) Thus any solution to $df = 0$ must satisfy $(\sigma^p - 1)f = 0$, that is, $f(t+p) = f(t)$ for all t.

The discrete setting can easily be generalized to the multivariate situation. A linear partial difference equation takes the form

$$\sum_{\nu\in\mathbb{N}_0^r} c_\nu f(t+\nu) = 0 \quad \text{for all } t \in \mathbb{N}_0^r,$$

where only finitely many of the coefficients $c_\nu \in R$ are nonzero. The relevant operator ring is $\mathscr{D} = R[\sigma_1,\ldots,\sigma_r]$, where σ_i acts on $f : \mathbb{N}_0^r \to R$ via $(\sigma_i f)(t) = f(t+\mathbf{e}_i)$. We also use the multi-index notation $(\sigma^\nu f)(t) = f(t+\nu)$ for $t,\nu \in \mathbb{N}_0^r$. Let $\mathscr{F}$ denote the set of all functions from $\mathbb{N}_0^r$ to R. Then $\mathscr{F}$ is again a left $\mathscr{D}$-module and an injective cogenerator [31, 36, 58]. Finally, let $E \in \mathscr{D}^{g\times q}$ be given and consider $S = \{f \in \mathscr{F}^q \mid Ef = 0\}$. The system S is autonomous (i.e., it has no free variable) if and only if there exists a $\mathscr{D}$-matrix X such that $XE = \operatorname{diag}(d_1,\ldots,d_q)$ with $0 \neq d_i \in \mathscr{D}$ for all i. In general, the input number (or "input dimension") of S is defined as the maximal m for which there exists a permutation matrix Π such that $S \to \mathscr{F}^m$, $f \mapsto [I_m, 0]\Pi f$ is surjective. Partitioning $\Pi f =: \begin{bmatrix} u \\ y \end{bmatrix}$ accordingly, this means that for all "inputs" $u \in \mathscr{F}^m$, there exists an "output" $y \in \mathscr{F}^{q-m}$ such that $f \in S$. Via the injective cogenerator property, the input number m is also the largest number such that there exists a permutation matrix Π with $\mathscr{D}^{1\times g}E\Pi \cap (\mathscr{D}^{1\times m} \times \{0\}) = \{0\}$. For simplicity, suppose that the columns of E have already been permuted such that $\mathscr{D}^{1\times g}E \cap (\mathscr{D}^{1\times m} \times \{0\}) = \{0\}$, where m is the input number of S. Let $E = [-Q, P]$ be partitioned accordingly such that the system law reads $Py = Qu$ with input u and output y. Moreover, the system given by $Py = 0$ is autonomous (otherwise, we'd have a contradiction to the maximality of m). By construction, we have $\ker(\cdot P) \subseteq \ker(\cdot Q)$ and this guarantees that $P\mathscr{F}^{q-m} \supseteq Q\mathscr{F}^m$. If R is a domain, then $\mathscr{D}$ has a quotient field $\mathscr{K}$, and we may simply set $m := q - \operatorname{rank}(E)$ and choose P as a submatrix of E whose columns are a basis of the column space $E\mathscr{K}^q$ of E. Then $Q = PH$ holds for some $\mathscr{K}$-matrix H, which clearly implies $\ker(\cdot P) \subseteq \ker(\cdot Q)$.

Example 3.2. Let $R = \mathbb{Z}/8\mathbb{Z}$ and consider the system given by

$$4f_1(t) = 2f_2(t+1)$$

for all $t \in \mathbb{N}_0$. Then $E = [-4, 2\sigma]$. Since $\ker(\cdot 2\sigma) = \operatorname{im}(\cdot 4) \subseteq \ker(\cdot 4)$, we may choose $Q = 4$ and $P = 2\sigma$, that is, we may interpret $u := f_1$ as an input, and $y := f_2$ as an output. For any choice of u, the left hand side of the system law is in $\{0,4\}$, and hence, the equation is always solvable for $y(t+1)$. The autonomous system $2y(t+1) = 0$ consists of all sequences with $y(t) \in \{0,4\}$ for all $t \geq 1$ (with $y(0)$ being arbitrary). Conversely, f_2 is not a free variable, since the system law implies (by multiplying both sides by 2) that $4f_2(t+1) = 0$ for all $t \in \mathbb{N}_0$, and hence $f_2(t) \in \{0,2,4,6\}$ for all $t \geq 1$.

For the rest of this section, let $R = \Bbbk$ be a field. Let $Py = 0$ define an autonomous system, that is, let P have full column rank q. Then none of the components of y is free, and it can be made unique by an appropriate choice of initial conditions. The theory of Gröbner bases can be used to set up a well-posed initial value problem.

For this, one computes a Gröbner basis G of the row module $N = \mathscr{D}^{1\times g}P$ of P. If $d \in \mathscr{D}^{1\times q}$ has leading term $\sigma^\mu \mathbf{e}_i$, then we write $\mathrm{lead}(d) = (\mu, i) \in \mathbb{N}_0^r \times \{1,\dots,q\}$. For a set $D \subseteq \mathscr{D}^{1\times q}$ with $D \setminus \{0\} \neq \emptyset$, we put $\mathrm{lead}(D) := \{\mathrm{lead}(d) \mid 0 \neq d \in D\}$. With this notation, the fact that G is a Gröbner basis of N reads

$$\mathrm{lead}(N) = \mathrm{lead}(G) + (\mathbb{N}_0^r \times \{0\}).$$

Define

$$\Gamma := \mathbb{N}_0^r \times \{1,\dots,q\} \setminus \mathrm{lead}(N).$$

Then the initial value problem $Py = 0$, $y|_\Gamma = z$ (that is, $y_i(\mu) = z(\mu, i)$ for all (μ, i) in Γ) has a unique solution $y \in \mathscr{F}^q$ for every choice of the initial data $z \in \mathbb{k}^\Gamma$ [36]. The solution y can be computed recursively, proceeding in the specific order on $\mathbb{N}_0^r \times \{1,\dots,q\}$ that was used to compute the Gröbner basis. (Clearly, ordering this set is equivalent to ordering $\{\sigma^\mu \mathbf{e}_i \mid \mu \in \mathbb{N}_0^r, 1 \le i \le q\}$.) For $(\mu, i) \in \mathrm{lead}(N)$, there exists a $g \in G$ such that $(\mu, i) = \mathrm{lead}(g) + (\nu, 0) = \mathrm{lead}(\sigma^\nu g)$, and thus, the value $y_i(\mu)$ can be computed from values $y_j(\kappa)$ with $(\kappa, j) < (\mu, i)$, that is, from values that are already known. Uniqueness follows by induction, since at each (μ, i), we may assume that all consequences d of the system law (that is, all equations $dy = 0$, where $d \in N$) with $\mathrm{lead}(d) < (\mu, i)$ are satisfied by the values that are already known. On the other hand, the values of y on Γ are unconstrained by the system law. More formally, the unique solvability of the initial value problem can be shown as follows: There is a $\mathbb{k}$-vector space isomorphism between S and $\mathrm{Hom}_\mathbb{k}(\mathscr{D}^{1\times q}/N, \mathbb{k})$. Clearly, a linear map on a vector space is uniquely determined by choosing the image of a basis. However, the set $\{[\sigma^\mu \mathbf{e}_i] \mid (\mu, i) \in \Gamma\}$ is indeed a $\mathbb{k}$-basis of $\mathscr{D}^{1\times q}/N$. This shows that each element of S is uniquely determined by fixing its values on Γ. Note that the set Γ is infinite, in general. We remark that Γ is finite if and only if the system module is finite-dimensional as a $\mathbb{k}$-vector space [54].

Example 3.3. Let $\mathbb{k} = \mathbb{R}$ and consider the autonomous system given by

$$\begin{aligned} y(t_1+2, t_2) + y(t_1, t_2+2) &= 0 \\ y(t_1+3, t_2) + y(t_1, t_2+3) &= 0 \end{aligned}$$

for all $t = (t_1, t_2) \in \mathbb{N}_0^2$. A Gröbner basis of $N = \langle \sigma_1^2 + \sigma_2^2, \sigma_1^3 + \sigma_2^3 \rangle$ with respect to the lexicographic order with $\sigma_1 > \sigma_2$ is given by $G = \{\sigma_1^2 + \sigma_2^2, \sigma_1\sigma_2^2 - \sigma_2^3, \sigma_2^4\}$. Therefore, each solution $y : \mathbb{N}_0^2 \to \mathbb{R}$ is uniquely determined by its values on the complement of $\mathrm{lead}(N) = \{(2,0),(1,2),(0,4)\} + \mathbb{N}_0^2$, that is, on the set $\Gamma = \{(0,0),(0,1),(0,2),(0,3),(1,0),(1,1)\}$. Thus, the solution set is a real vector space of dimension $|\Gamma| = 6$.

4 Linear Partial Differential Equations

The theory of partial differential equations shows some notable differences compared to ordinary differential equations. For general systems (i. e. including under-

or overdetermined ones), it is much harder to prove the existence of at least formal solutions, as now integrability conditions may be of higher order than the original system[1]. The reason is a simple observation. In systems of ordinary differential equations, only one mechanism for the generation of integrability conditions exist: (potentially after some algebraic manipulations) the system contains equations of different order and differentiation of the lower-order ones may lead to new equations. In the case of partial differential equations, such differences in the order of the individual equations are less common in practice. Here the dominant mechanism for the generation of integrability conditions are (generalised) cross-derivatives and these lead generally to equations of higher order. A comprehensive discussion of general systems of partial differential equations and the central notions of involution and formal integrability can be found in the monograph [52]. Within this article, we will consider exclusively linear systems where again standard Gröbner basis techniques can be applied. A somewhat more sophisticated approach using the formal theory of differential equations and involutive bases is contained in [16]; it provides intrinsic results independent of the used coordinates.

Example 4.1. We demonstrate the appearance of integrability conditions with a system of two second-order equations for one unknown function u of three independent variable (x,y,z) due to Janet [18, Ex. 47]:

$$\begin{bmatrix} \partial_z^2 + y\partial_x^2 \\ \partial_y^2 \end{bmatrix} f = 0\,. \tag{6}$$

It hides two integrability conditions of order 3 and 4, respectively, namely

$$\partial_x^2\partial_y f = 0\,, \qquad \partial_x^4 f = 0\,. \tag{7}$$

It is important to note that they do not represent additionally imposed equations but that *any* solution of (6) will automatically satisfy (7).

A systematic derivation of these conditions is easily possible with Gröbner bases. As we have only one unknown function in the system, the row module becomes here the row ideal $I \lhd \mathscr{D} = \mathbb{R}(x,y,z)[\partial_x,\partial_y,\partial_z]$ in the ring of linear differential operators with rational coefficients generated by the operators $D_1 = \partial_z^2 + y\partial_x^2$ and $D_2 = \partial_y^2$. We use the reverse lexicographic order with $\partial_z \succ \partial_y \succ \partial_x$ and follow the Buchberger Algorithm 6 for the construction of a Gröbner basis. The S-polynomial (see (24)) of the two given operators is $S(D_1,D_2) = \partial_y^2 \cdot D_1 - \partial_z^2 \cdot D_2 = y\partial_x^2\partial_y^2 + 2\partial_x^2\partial_y$. Reduction modulo D_2 eliminates the first term so that $D_3 = \partial_x^2\partial_y$ and we have found the first integrability condition. The second one arises similarly from the S-polynomial $S(D_1,D_3) = \partial_x^2\partial_y \cdot D_1 - \partial_z^2 \cdot D_3$ leading to the operator $D_4 = \partial_x^4$ after reduction, whereas the S-polynomial $S(D_2,D_3)$ immediately reduces to zero. Since also all S-

[1] For arbitrary systems, not even an a priori bound on the maximal order of an integrability condition is known. In the case of linear equations, algebraic complexity theory provides a double exponential bound which is, however, completely useless for computations, as for most systems appearing in applications it grossly overestimates the actual order.

polynomials $S(D_i, D_4)$ reduce to zero, the set $\{D_1, D_2, D_3, D_4\}$ represents a Gröbner basis of the ideal I and there are no further hidden integrability conditions.

There are two reasons why one should make the integrability conditions (7) explicit. First of all, only after the construction of all hidden conditions, we can be sure that the system (6) indeed possesses solutions; in general, a condition like $1 = 0$ can be hidden which shows that the system is inconsistent.[2] Secondly, the knowledge of these conditions often significantly simplifies the integration of the system and provides information about the size of the solution space. As we will show later in this section, in our specific example one can immediately recognise from the combined system (6,7) that the solution space of our problem is finite-dimensional (more precisely, 12-dimensional)—something rather unusual for partial differential equations. In fact, once (7) is taken into account, one easily determines the general solution of the system, a polynomial with 12 arbitrary parameters:

$$\begin{aligned} f(x,y,z) = &- a_1 xyz^3 + a_1 x^3 z - 3a_2 xyz^2 - \frac{1}{3} a_3 yz^3 + a_2 x^3 + \\ &a_3 x^2 z - a_4 yz^2 + a_5 xyz + a_4 x^2 + a_6 xy + a_7 yz + \\ &a_8 xz + a_9 x + a_{10} y + a_{11} z + a_{12}\,. \end{aligned} \tag{8}$$

An important notion for the formulation of existence and uniqueness results for differential equations is *well-posedness*. According to an informal definition due to Hadamard, an initial or boundary value problem is well-posed, if (i) a solution exists for arbitrary initial or boundary data, (ii) this solution is unique and (iii) it depends continuously on the data. For a mathematically rigorous definition, one would firstly have to specify function spaces for the data and the solution and secondly define topologies on these spaces in order to clarify what continuous dependency should mean. In particular this second point is highly non-trivial and application dependent. For this reason, we will use in this article a simplified version which completely ignores (iii) and works with formal power series.

We will consider here exclusively initial value problems, however in a more general sense as usual. For notational simplicity, we assume in the sequel that there is only one unknown function f. Since we work with formal power series solutions, we further assume that some expansion point $\hat{\mathbf{x}} = (\hat{x}_1, \ldots, \hat{x}_n)$ has been chosen. For any subset $X' \subseteq X$ of variables, we introduce the $(n - |X'|)$-dimensional coordinate space $H_{X'} = \{x_i = \hat{x}_i \mid x_i \in X'\}$ through $\hat{\mathbf{x}}$. An inital condition then prescribes some derivative $f_\mu = \partial^{|\mu|} f / \partial x^\mu$ for a multi index $\mu \in \mathbb{N}_0^n$ of the unknown function f restricted to such a coordinate space $H_{X'}$. If several conditions of this kind are imposed with coordinate spaces H_1, H_2, …, then one obtains an initial value problem in a strict sense only, if (after a suitable renumbering) the coordinate spaces form a chain $H_1 \subseteq H_2 \subseteq \cdots$. However, in general a formally well-posed initial value problem in this strict sense will exist only after a linear change of coordinates. Therefore, we will not require such a chain condition.

[2] Here we are actually dealing with the special case of a homogeneous linear system where consistency simply follows from the fact that $u = 0$ is a solution.

Definition 4.2. An initial value problem for a differential equation is *formally well-posed*, if it possesses a unique formal power series solution for arbitrary formal power series as initial data.

In principle, we approach the algorithmic construction of formally well-posed initial value problems for a given linear partial differential operator L in the same manner as described in the last section for linear difference operators: we compute a Gröbner basis of the row module N—which here is actually an ideal $I \trianglelefteq \mathscr{D}$ because of our assumption that there is only one unknown function—and then use the complement of the leading module. However, in order to be able to translate this complement into initial conditions, we need a complementary decomposition of it (see Definition 7.10) which can be constructed with the help of Algorithm 5 in Section 7. In fact, it was precisely this problem of constructing formally well-posed problems which lead to the probably first appearance of such decompositions in the works of Riquier [45] and Janet [18].

A leading "term" can now be identified with a derivative f_μ. All derivatives in $\mathrm{lt}\, I$ are traditionally called *principal derivatives*, all remaining ones *parametric derivatives*. Assume that we want to compute a formel power series solution with expansion point $\hat{\mathbf{x}}$. Making the usual ansatz

$$f(\mathbf{x}) = \sum_{\mu \in \mathbb{N}_0^n} \frac{a_\mu}{\mu!} (\mathbf{x} - \hat{\mathbf{x}})^\mu \,, \tag{9}$$

we may identify by Taylor's theorem the coefficient a_μ with the value of the derivative f_μ at $\hat{\mathbf{x}}$.

Let the leading term of the scalar equation $Df = 0$ be f_μ. Because of the monotonicity of term orders, the leading term of the differentiated equation $\partial_i Df = 0$ is $f_{\mu+1_i} = \partial_i f_\mu$. Hence by differentiating the equations in our system sufficiently often, we can generate for any principal derivative f_μ a differential consequence of our system with f_μ as leading term. Prescribing initial data such that unique values are provided for all parametric derivatives (and no values for any principal derivative), we obtain a formally well-posed initial value problem, as its unique power series solution is obtained by determining each principal derivative via the equation that has it as leading term. Such initial data can be systematically constructed via complementary decompositions of the leading ideal.

A complementary decomposition of the monomial ideal $\mathrm{lt}\, I$ is defined by pairs (f_ν, X_ν) where f_ν is a parametric derivative and X_ν the associated set of multiplicative variables. Differentiating f_ν arbitrarily often with respect to variables contained in X_ν, we always obtain again a parametric derivative. Denoting by $\overline{X}_\nu = X \setminus X_\nu$ the set of all *non*-multiplicative variables of f_ν, we associate with the pair (f_ν, X_ν) the initial condition that f_ν restricted to the coordinate space $H_\nu = H_{\overline{X}_\nu}$ is some prescribed function $\rho_\nu(X_\nu)$. In this way, each complementary decomposition induces an initial value problem.[3]

[3] An initial value problem in the strict sense is obtained, if one starts with a complementary Rees decomposition (see Definition 7.10).

Theorem 4.3 ([52, Thm. 9.3.5]). *For any complementary decomposition of* $\mathrm{lt}I$*, the above constructed initial value problem for the linear differential operator* D *is formally well-posed.*

If both the coefficients of D and the prescribed initial data ρ_v are analytic functions and a degree compatible term order has been used, then *Riquier's Theorem* [37, 45] even guarantees the convergence of the unique formal power series solution to the thus constructed initial value problem. Under these assumptions, we therefore obtain an existence and uniqueness theorem in the analytic category.

Example 4.4. Let us begin with a scalar first-order equation like the simple advection equation $(\partial_2 - \partial_1)f = 0$. If we assume that for the chosen term order f_{x_2} is the leading derivative, then it is easy to see that a complementary decomposition is defined by the single pair $\big(f, \{x_1\}\big)$, as the only parametric derivatives are f and all its pure x_1-derivatives. Hence choosing an expansion point $(\hat{x}_1, \hat{x}_2)$, we recover the familiar initial condition $f(x_1, \hat{x}_2) = \rho(x_1)$ with an arbitrary function ρ.

Advancing to a second-order equation like the wave equation $(\partial_2^2 - \partial_1^2)f = 0$ with leading derivative $f_{x_2x_2}$, we find that a simple complementary decomposition is defined by the two pairs $\big(f, \{x_1\}\big)$ and $\big(f_{x_2}, \{x_1\}\big)$. Hence our construction yields again the classical initial conditions $f(x_1, \hat{x}_2) = \rho(x_1)$ and $f_{x_2}(x_1, \hat{x}_2) = \sigma(x_1)$. This decomposition is shown in Figure 1 on the left hand side. The small black dots depict the terms ∂^μ of the underlying ring $\mathscr{D}$ or alternatively the derivatives f_μ. All dots lying in the blue two-dimensional cone correspond to principal derivatives contained in the ideal I and are multiples of the leading derivative $f_{x_2x_2}$ shown as a large blue dot. All parametric derivatives are contained in the two red one-dimensional cones whose vertices are shown as large red dots. Obviously, the complementary decomposition provides a disjoint partitioning of the complement of the "blue" ideal.

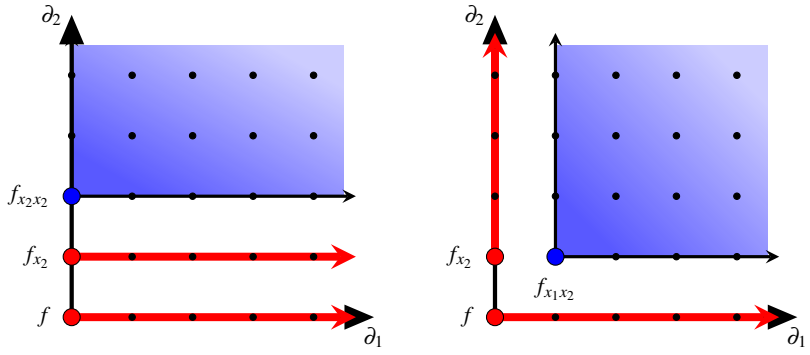

Fig. 1 Complementary decompositions for the wave equation

If we consider the wave equation in characteristic coordinates $\partial_1\partial_2 f = 0$, then we have two natural options for a complementary decomposition: $\big(f, \{x_1\}\big)$, $\big(f_{x_2}, \{x_2\}\big)$

(shown in Figure 1 on the right hand side) or $\big(f,\{x_2\}\big)$, $\big(f_{x_1},\{x_1\}\big)$. Both correspond to the classical characteristic initial value problem (which is *not* an initial value problem in the strict sense) in which usually both $f(\hat{x}_1,x_2)=\rho(x_2)$ and $f(x_1,\hat{x}_2)=\sigma(x_1)$ are prescribed. However, this classical formulation is not formally well-posed, as the initial data must satisfy the consistency condition $\rho(0)=\sigma(0)$. The formulations induced by the above complementary decompositions avoid such a restriction by substituting in one of the initial conditions f by a first derivative. The quite different character of the standard and the characteristic initial value problem of the wave equation is here encoded in the fact that on the left hand side of Figure 1 the two red one-dimensional cones are pointing in the same direction, whereas on the right hand side they have different directions.

The results become less obvious when we proceed to larger overdetermined systems. As a simple example, we take the "monomial" system with

$$I=\langle\partial_3^2,\partial_2^2\partial_3,\partial_1\partial_2\partial_3\rangle \tag{10}$$

(any linear system with a Gröbner basis consisting of three operators with these generators as leading terms leads to the same initial conditions). Here a complementary decomposition (shown in Figure 2 where again the blue colour marks the ideal I and the red colour the different cones of the complementary decomposition) is given by the three pairs $\big(f,\{x_1,x_2\}\big)$, $\big(f_{x_3},\{x_1\}\big)$ and $(f_{x_2x_3},\emptyset)$. Hence a formally well-posed initial value problem is given by

$$f(x_1,x_2,\hat{x}_3)=\rho(x_1,x_2)\,,\quad f_{x_3}(x_1,\hat{x}_2,\hat{x}_3)=\sigma(x_1)\,,\quad f_{x_2x_3}(\hat{x}_1,\hat{x}_2,\hat{x}_3)=\tau \tag{11}$$

where the initial data consist of two functions (of one and two arguments, resp.) and one constant. Note that this time the red cones are of different dimensions which is typical and unavoidable for overdetermined systems. We are dealing here with a non-characteristic initial value problem, as the directions of the various cones form a flag: all directions defining a lower-dimensional cone are also contained in any higher-dimensional cone.

The considerations above also yield the simple Algorithm 1 for constructing any Taylor coefficient a_λ in the unique solution of the above constructed initial value problem. It computes the normal form (cf. Definition 7.7) of ∂^λ with respect to a Gröbner basis of I. If ∂^λ corresponds to a parametric derivative, it remains unchanged and we obtain in Line /4/ directly the value of a_λ from the appropriate initial condition. Otherwise the normal form computation expresses the principal derivative ∂^λ as a linear combination of parametric derivatives ∂^μ. We determine for each appearing ∂^μ in which unique cone (∂^ν,X_ν) of the given complementary decomposition it lies and then compute the required derivative of the corresponding initial data ρ_ν. Evaluating at the expansion point $\hat{\mathbf{x}}$, we obtain first the values of all required parametric coefficients a_μ and then compute the principal coefficient a_λ from these.

Remark 4.5. For notational simplicity, we restricted in the above discussion to the ideal case. The extension to the module case is straightforward; a separate comple-

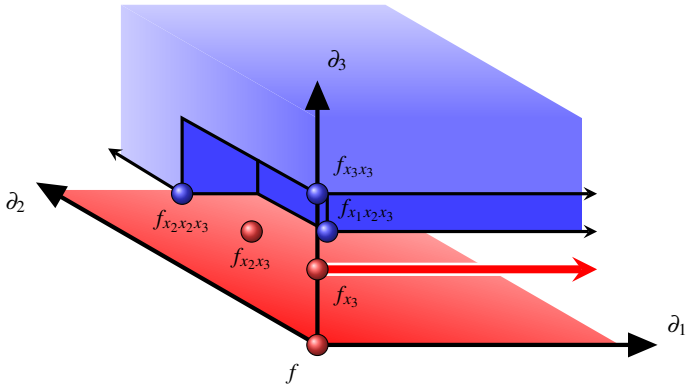

Fig. 2 Complementary decomposition for the monomial system (10)

Algorithm 1 Taylor coefficient of formal solution

Input: Gröbner basis G of I for term order $\prec$, complementary decomposition $\mathscr{T}$ with corresponding initial data for each cone $(\partial^\nu, X_\nu) \in \mathscr{T}$, expansion point $\hat{\mathbf{x}}$, multi index $\lambda \in \mathbb{N}_0^n$

Output: Taylor coefficient a_λ of unique formal power series solution of initial value problem corresponding to given complementary decomposition

1: $h \leftarrow \texttt{NormalForm}_\prec(\partial^\lambda, \mathscr{G})$
2: **for all** $\partial^\mu \in \operatorname{supp} h$ **do**
3: find unique $(\partial^\nu, X_\nu) \in \mathscr{T}$ such that ∂^μ lies in the cone of ∂^ν
4: $a_\mu \leftarrow \partial^{\mu-\nu} \rho_\nu(\hat{\mathbf{x}})$
5: **end for**
6: write $h = \sum_\mu c_\mu(\mathbf{x})\partial^\mu$
7: **return** $\sum_\mu c_\mu(\hat{\mathbf{x}}) a_\mu$

mentary decomposition is now needed for each vector component (corresponding to one unknown function). Only in the application of Riquier's Theorem a slight technical complication arises. One must assume that the used module term order is not only degree compatible but also *Riquier* which means that if for one value of α the inequality $t\mathbf{e}_\alpha \prec s\mathbf{e}_\alpha$ with terms $s,t \in \mathbb{T}$ holds, then it holds for all values of α. This property does not appear in the usual theory of Gröbner bases, as it is only relevant in applications to differential equations. It has the following meaning. Using a classical trick due to Drach [10] (see [52] for an extensive discussion), any system of differential equations with m unknown functions can be transformed via the introduction of m additional independent variables into a system for only one unknown function (at the expense of raising the order by one). Algebraically, this trick transforms a submodule $M \subseteq \mathscr{D}^m$ defined over a skew polynomial ring $\mathscr{D}$ in n variables into an ideal I_M living in a skew polynomial ring $\mathscr{D}'$ in $n+m$ variables. Now the chosen module term order on $\mathscr{D}^m$ induces an ordering on the terms in $\mathscr{D}'$ which is guaranteed to be a term order only if the module term order is Riquier. Lemaire [28] constructed a concrete counter example to Riquier's Theorem where all of its assumptions are satisfied except that the chosen module term order is not Riquier and

showed that the unique formal solution of the corresponding initial value problem diverges. Note that any TOP lift of a term order is automatically Riquier.

We have used above already several times the terms "under-" and "overdetermined" systems without actually having defined these notions. Somewhat surprisingly, it is not so easy to provide a rigorous definition for them (and this topic is not much considered in the literature). The linear algebra inspired approach of comparing the number of unknown functions and of equations fails in some instances, as we will show below with an explicit example. The main problem is that it is not clear what should be taken as the "right" number of equations. In linear algebra, one must count the number of linearly independent equations. In the case of linear differential equations, we are working with modules and ideals where almost never a linearly independent set of generators exists and where minimal generating sets may possess different cardinalities.

Example 4.6. Consider the following linear system for two unknown functions u, v of two independent variables x, t:

$$\begin{bmatrix} \partial_t^2 & -\partial_x\partial_t \\ \partial_x\partial_t & -\partial_x^2 \end{bmatrix} \begin{bmatrix} f \\ g \end{bmatrix} = 0\,. \tag{12}$$

One may call it the *two-dimensional $U(1)$ Yang-Mills equations*, as it represents a gauge theoretic model of electromagnetism in one spatial and one temporal dimension. At first sight, one would call (12) a well-determined system, as it comprises as many equations as unknown functions. However, let $f = \phi(x,t)$ and $g = \psi(x,t)$ be a solution of (12) and choose an arbitrary function $\Lambda(x,t)$; then $f = \phi + \partial_x\Lambda$ and $g = \psi + \partial_t\Lambda$ is also a solution. Hence one of the unknown functions f, g may be chosen arbitrarily—the typical behaviour of an underdetermined system! From a physical point of view, this phenomenon is a consequence of the invariance of (12) under gauge transformations of the form $f \to f + \partial_x\Lambda$ and $g \to g + \partial_t\Lambda$ and lies at the heart of modern theoretical physics.

Using our above considerations about the construction of formally well-posed initial value problems for a linear partial differential operator D with row module N, it is fairly straightforward to provide a rigorous definition of under- and overdeterminacy. At first sight, it is not clear that the definition presented below is independent of the chosen complementary decomposition. But this follows from the remarks after Definition 7.10, as it is only concerned with the cones of maximal dimension and their dimension and number are the same for any decomposition (in the Cartan-Kähler approach this maximal dimension is sometimes called the Cartan genus of the operator and the number of such cones the degree of arbitrariness; see [48] or [52, Chapt. 8] for a more detailed discussion of these notions).

Definition 4.7. Let D be a linear differential operator of order q in n independent variables operating on m unknown functions and choose a complementary decomposition of the associated monomial module $\mathrm{lt}N$ (for some term order $\prec$). The operator D defines an *underdetermined* system, if the decomposition contains at least

one cone of dimension n. It defines a *welldetermined* system, if a complementary decomposition exists consisting of mq cones of dimension $n-1$ (and no other cones). In any other case, the system is *overdetermined*.

The definition of underdeterminacy should be clear from the discussion above. If $k \geq 1$ cones of dimension n appear in the complementary decomposition, then $k \geq 1$ unknown functions can be chosen completely arbitrarily and thus are not constrained by the differential operator. This behaviour represents exactly what one intuitively expects from an underdetermined system. Note that such cones will always appear, if there are less equations than unknown functions, as in this case there exists at least one unknown function to which no leading derivative belongs and for which thus all derivatives are parametric.

Example 4.8. It is less obvious that there exist underdetermined systems comprising as many (or even more) equations as unknowns. If we go back to Example 4.6 and use the TOP extension of the reverse lexicographic order, then the two equations in (12) induce already a Gröbner basis of N with leading terms f_{tt} and f_{xt}. Thus there are no leading terms corresponding to derivatives of g and all g-derivatives are parametric. Hence the complementary decomposition consists for g of just one two-dimensional cone with vertex at g.

The well-determined case corresponds to what is called in the theory of partial differential equations a system in *Cauchy-Kovalevskaya form*. Note that for being in this form it is necessary but *not* sufficient that the system comprises as many equations as unknown functions. In addition, a distinguished independent variable t must exist such that each equation in the system can be solved for a pure t-derivative of order q (possibly after a coordinate transformation[4]). We may assume that the αth equation is solved for $\partial^q f_\alpha/\partial t^q$ and a formally well-posed initial value problem is then given by prescribing $\partial^k f_\alpha/\partial t^k(t=\hat{t})$ for $0 \leq k < q$ and hence the initial data indeed consists of mq functions of $n-1$ variables. The advection and the wave equation in Example 4.4 are of this type with $t = x_2$. If the wave equation is given in characteristic coordinates, then one must first transform to non-characteristic ones to find a suitable t.

Remark 4.9. There is a certain arbitrariness in Definition 4.7. We give precedence to under- and welldeterminacy; overdetermined systems are the remaining ones. This could have been done the other way round. Consider the system

$$\begin{bmatrix} \partial_1 & -\partial_2 & 0 \\ 0 & 0 & \partial_1 \\ 0 & 0 & \partial_2 \end{bmatrix} \begin{bmatrix} f \\ g \\ h \end{bmatrix} = 0\,. \tag{13}$$

Obviously, it decouples into one underdetermined equation for the two unknown functions f, g and an overdetermined system for the unknown function h. According to our definition, the combined system is underdetermined, although one could

[4] It is quite instructive to try to transform (12) into such a form: one will rapidly notice that this is not possible!

consider it equally well as an overdetermined one. One reason for our choice is simply that underdeterminacy is the more intuitive notion (some unknown functions are unconstrained) compared to the rather technical concept of overdeterminacy.

Another reason lies in the system theoretic interpretation of our results. As we will discuss in Section 6.3 in a more abstract setting, a system is *autonomous*, if and only if it is not underdetermined. Those variables for which the complementary decomposition contains n-dimensional cones are *free* and thus represent the *inputs* of the system. Different complementary decompositions generally lead to different possible choices of the inputs. But as already remarked above, all complementary decompositions contain the same number of cones of dimension n; thus this number represents the *input dimension* of the system.

5 The Index

Both in the theory and in the numerical analysis of differential algebraic equations, the notion of an index plays an important role. One interpretation is that it provides a measure for the difficulties to expect in a numerical integration, as higher-index systems show a high sensitivity to perturbations (e. g. through numerical errors). It also shows up in the development of an existence and uniqueness theory.

Over the years, many different index concepts have been introduced. One may roughly distinguish two main types of indices. *Differentiation indices* basically count how many times some equations in the system must be differentiated, until a certain property of the system becomes apparent. E.g., "the" differentiation index tells us when we can decide that the system is not underdetermined (therefore it is called determinacy index in [49]). By contrast, *perturbation indices* are based on estimates for the difference of solutions of the given system and of a perturbed form of it and count how many derivatives of the perturbations have to be taken into account. Generally, differentiation indices are easier to compute, whereas the relevance of perturbation indices—e. g. for a numerical computation—is more obvious. In many cases, all the different approaches lead to the same index value. But it is not difficult to produce examples where the differences can become arbitrarily large. An overview of many index notions and their relationship is contained in [5]; some general references for differential algebraic equations are [2, 24, 27, 44].

We will now use Gröbner bases to introduce some index notions for linear differential equations. While these are strictly speaking differentiation indices, the main result about them is an estimate for a perturbation index. We will first consider the case of ordinary differential equations. Here the theory of Gröbner bases becomes of course rather trivial, as we are dealing with a univariate polynomial ring. Afterwards, we will discuss the extension to partial differential equations. From the point of view of mere definitions, this extension is straightforward and provides—in contrast to most results in the literature—an approach to introduce indices directly for partial differential equations without resorting to some sort of semi-discretisation or other forms of reduction to the ordinary case. Using a more geometric approach

to differential equations, one can also handle nonlinear equations in an analogous manner. For details we refer to [15, 49]. One could say that the here presented material represents a concrete algorithmic version of the ideas in [15, 49] specialised to linear systems.

Let a homogeneous system $Df = 0$ be given where for the moment it does not matter whether D is an ordinary or a partial differential operator. We introduce for each equation in the system a new auxiliary unknown function δ_i and consider the inhomogeneous system $Df = \delta$. Depending on the context, there are different ways to interpret the new unknowns δ_i. One may consider them as perturbations of the original system or as residuals obtained by entering some approximate solution of the original system. For us, they are mainly a convenient way to keep track of what is happening during a Gröbner basis computation. Obviously, the following analysis will lead to the same results for an inhomogeneous system $Df = g$, as the right hand side g can be absorbed into the perturbation δ.

We compute a Gröbner basis of the row module of the operator D. The most natural choice for the term order is the TOP lift of a degree compatible order. In the case of ordinary differential equations, this choice uniquely determines the term order. For partial differential equations, we will discuss further properties of the underlying order below. The outcome of this computation can be decomposed into two subsystems $\tilde{D}f = F\delta$ and $0 = G\delta$. Here the rows of $\tilde{D}$ form the Gröbner basis of the row module of D and we have the equality $\tilde{D} = FD$. Thus F tells us how the elements of the Gröbner basis have been constructed from the rows of D.

The rows of G form a generating set of the first syzygy module of D (see (25) for an explanation of syzygies). Indeed, the equations in the subsystem $0 = G\delta$ arise when some S-polynomial reduces to zero (that is the zero on the left hand side!) and thus represent syzygies. The Schreyer Theorem 7.15 asserts that they actually generate the whole syzygy module. This second subsystem represents a necessary condition for the existence of solutions in the inhomogeneous case: only for right hand sides δ satisfying it, the system can possess solutions. According to the fundamental principle discussed in the next section, it is also a sufficient condition (for "good" function spaces). Alternatively, we may consider δ as the residuals which arise when an approximate solution $\hat{f}$ is entered into the original system. Now the subsystem $0 = G\delta$ describes what is often called the *drift* which appears in the numerical integration of overdetermined systems. In particular, a stability analysis of the trivial solution $\delta = 0$ provides valuable information about the behaviour of the drift, i. e. whether it is automatically damped or whether it accelerates itself.

Definition 5.1. The *first Gröbner index* γ_1 of $Df = 0$ is the order[5] of the operator F; the *second Gröbner index* γ_2 the order of G.

Example 5.2. For linear differential equations, most index concepts coincide (sometimes with a shift of 1). Hence it is not surprising that the first Gröbner index γ_1 yields often the same value as other approaches. Many applications lead to systems in *Hessenberg form*. In our operator notation a (perturbed) Hessenberg system of index 3 has the form

[5] We define the order of an operator matrix as the maximal order of an entry.

$$\begin{bmatrix} \partial - a_{11} & -a_{12} & -a_{13} \\ -a_{21} & \partial - a_{22} & 0 \\ 0 & -a_{32} & 0 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \\ f_3 \end{bmatrix} = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \delta_3 \end{bmatrix}$$

where all coefficients a_{ij} and the right hand sides δ_i are functions of the independent variable t and where we assume that $a_{13}(t)a_{21}(t)a_{32}(t) \neq 0$ at every point t considered. According to [24, Thm. 4.23] such a system has strangeness index 2 and according to [2, Prop. 2.4.1] the classical differentiation index is 3.

Computing a Gröbner basis of the row module for the descending TOP lift corresponds exactly to the classical steps for computing the differentiation index. In the first two rows, the leading term is $\partial \mathbf{e}_1$ and $\partial \mathbf{e}_2$. In the last row we have only one non-vanishing term $-a_{32}\mathbf{e}_2$ so that it may be considered as an algebraic equation determining f_2 since our assumption implies that its coefficient $a_{32} \neq 0$.

Since the second and the third row have their leading term in the same component, we can compute an S-polynomial by adding ∂ times the third row to a_{32} times the second row – in other words: we differentiate the last equation and then simplify it with respect to the second one. This yields a new row $\begin{bmatrix} -a_{21}a_{32} & \dot{a}_{32} - a_{22}a_{32} & 0 \end{bmatrix}$ with right hand side $\dot{\delta}_3 + a_{32}\delta_2$ (note how it "records" the performed operation). We could use the third equation for further simplifications (in fact, elimination of the second component). For our purposes, it suffices to observe that the new row yields an algebraic equation for f_1, since again our assumption implies that its coefficient $a_{21}a_{32}$ does not vanish.

As the leading term of the new row is $-a_{21}a_{32}\mathbf{e}_1$ for the descending TOP lift, we must now form its S-polynomial with the first row: we add ∂ times the new row to $a_{21}a_{32}$ times the first row. As the result still contains a term $\partial \mathbf{e}_2$, it must be reduced with respect to the second equation. We omit the explicit form of the result (which also should be simplified further with respect to the other algebraic equations) and only observe that it represents an algebraic equation for f_3 with nonvanishing coefficient $a_{13}a_{21}a_{32}$. The right hand side of the new row is given by $\ddot{\delta}_3 + a_{32}\dot{\delta}_2 + 2\dot{a}_{32}\delta_2$.

Now we are done: a reduced Gröbner basis of the row module consists of the simplified form of the three algebraic equations. Their precise form is not relevant for us, but we see that we have $\gamma_1 = 2$, as the right hand side for the last obtained row contains the second derivative of δ_3. Thus the first Gröbner index coincides here with the strangeness index. The classical differentiation index is one higher because for its determination one wants a *differential* equation for f_3 and hence multiplies the last row once more with ∂.

Example 5.3. We go back to the partial differential system (6) of Janet. If we add there a right hand side with entries δ_1, δ_2 and redo the calculations described above in Example 4.1, then we obtain the following perturbed system:

$$\begin{bmatrix} \partial_z^2 + y\partial_x^2 \\ \partial_y^2 \\ \partial_x^2\partial_y \\ \partial_x^4 \end{bmatrix} f = \begin{bmatrix} \delta_1 \\ \delta_2 \\ \frac{1}{2}\partial_y^2\delta_1 - \left(y\partial_x^2 + \partial_z^2\right)\delta_2 \\ \left(\frac{1}{2}y\partial_x^2\partial_y^2 - \frac{1}{2}\partial_y^2\partial_z^2 + \partial_x^2\partial_y\right)\delta_1 + \left(\frac{1}{2}y^2\partial_x^4 + y\partial_x^2\partial_y^2 + \frac{1}{2}\partial_z^4\right)\delta_2 \end{bmatrix}.$$

Hence the first Gröbner index is $\gamma_1 = 4$ for this example. In order to obtain also the second Gröbner index, we must also record how the remaining four S-polynomials ($S(D_2,D_3)$, $S(D_1,D_4)$, $S(D_2,D_4)$, $S(D_3,D_4)$) reduce to zero. This leads to the following four equations:

$$\begin{aligned}
0 &= -\frac{1}{2}\partial_y^3\delta_1 + \Big(\frac{1}{2}y\partial_x^2\partial_y + \frac{1}{2}\partial_y\partial_z^2 + \frac{3}{2}\partial_x^2\Big)\delta_2\,,\\
0 &= \Big(\frac{1}{2}y^2\partial_x^4\partial_y^2 + y\partial_x^2\partial_y^2\partial_z^2 + \frac{1}{2}\partial_y^2\partial_z^4 - y\partial_x^4\partial_y - \partial_x^2\partial_y\partial_z^2 + \partial_x^4\Big)\delta_1 -\\
&\quad \Big(\frac{1}{2}y^3\partial_x^6 + \frac{3}{2}y^2\partial_x^4\partial_z^2 + \frac{3}{2}y\partial_x^2\partial_z^4 + \frac{1}{2}\partial_z^6\Big)\delta_2\,,\\
0 &= \Big(\frac{1}{2}y\partial_x^2\partial_y^4 + \frac{1}{2}\partial_y^4\partial_z^2\Big)\delta_1 -\\
&\quad \Big(\frac{1}{2}y^2\partial_x^4\partial_y^2 + y\partial_x^2\partial_y^2\partial_z^2 + \frac{1}{2}\partial_y^2\partial_z^4 + 2y\partial_x^4 + \frac{3}{2}\partial_x^2\partial_z^2\Big)\delta_2\,,\\
0 &= \Big(\frac{1}{2}y\partial_x^2\partial_y^3 + \frac{1}{2}\partial_x^3\partial_z^2\Big)\delta_1 -\\
&\quad \Big(\frac{1}{2}y^2\partial_x^4\partial_y + y\partial_x^2\partial_y\partial_z^2 + \frac{1}{2}\partial_y\partial_z^4 + \frac{3}{2}y\partial_x^4 + \frac{3}{2}\partial_x^2\partial_z^2\Big)\delta_2\,.
\end{aligned}$$

Thus we conclude that the second Gröbner index is 6 here.

For the definition of the Gröbner indices, it does not matter whether we are dealing with ordinary or partial differential equations. One should, however, note that in the case of partial differential equations the values γ_1 and γ_2 will generally depend on the used term order (in the above example the system was so simple and in addition homogeneous so that most term orders yield the same results). The whole interpretation of the indices is less obvious in this case and depends on certain properties of the considered system. For this reason, we discuss first the case of ordinary differential equations.

Definition 5.4 ([14]). Let $f(t)$ be a given solution of the linear system $Df = 0$ and $\hat{f}(t)$ an arbitrary solution of the perturbed system $D\hat{f} = \delta$ both defined on the interval $[0,T]$. The system $Df = 0$ has the *perturbation index* ν along the given solution $f(t)$, if ν is the smallest integer such that the following estimate holds for all $t \in [0,T]$ provided the right hand side is sufficiently small:

$$|f(t) - \hat{f}(t)| \leq C\big(|f(0) - \hat{f}(0)| + \|\delta\|_{\nu-1}\big)\,. \tag{14}$$

Here the constant C may only depend on the operator D and the length T of the considered interval. $|\cdot|$ denotes some norm on $\mathbb{R}^m$ and $\|\delta\|_k = \sum_{i=0}^k \|\delta^{(i)}\|$ for some norm $\|\cdot\|$ on $C\big([0,T],\mathbb{R}\big)$ (common choices are the L^1 or the L^∞ norm).

Note that the perturbation index is well-defined only for systems which are not underdetermined, as an estimate of the form (14) can never hold, if there are free variables present. And obviously it is hard to determine, as it is firstly defined with respect to a specific solution and secondly, given an estimate of the form (14), it is

very difficult to prove that no other estimate with a lower value of ν exists. On the other hand, Definition 5.4 shows why a solution with a perturbation index greater than one can be difficult to obtain with a standard numerical integration: it does not suffice to control only the size of the residuals δ—as automatically done by any reasonable numerical method—but one must also separately control the size of some derivatives $\delta^{(i)}$ (and thus for example ensure that the residuals do not oscillate with small amplitude but high frequency). The following result asserts that the first Gröbner index provides a uniform bound for the perturbation index along *any* solution of the considered linear system.

Proposition 5.5. *Let the linear system $Df = 0$ of ordinary differential equations be not underdetermined. Along any solution $f(t)$ of it, the perturbation index ν and the first Gröbner index γ_1 satisfy the inequality $\nu \leq \gamma_1 + 1$.*

Proof. As discussed above, we compute for the perturbed system $Df = \delta$ a Gröbner basis with respect to the TOP lift of the ordering by degree and consider the subsystem $\tilde{D}f = F\delta$. We may assume without loss of generality that the operator $\tilde{D}$ is of first order, since a transformation to first order by introducing derivatives of f as additional unknown functions does not affect the value of γ_1 as the maximal order on the right hand side. Because of the use of a TOP lift, the leading term tells us whether or not an equation in the system $\tilde{D}f = F\delta$ is algebraic. The algebraic equations are differentiated once in order to produce an underlying differential equation of the form $\dot{f} = Kf + \tilde{F}\delta$ where K is of order zero and $\tilde{F}$ of order $\gamma_1 + 1$.

This underlying equation leads immediately to an estimate of the form

$$|f(t) - \hat{f}(t)| \leq |f(0) - \hat{f}(0)| + \Big| \int_0^t K(\tau)\big(f(\tau) - \hat{f}(\tau)\big)d\tau \Big| + \Big| \int_0^t \tilde{F}\delta(\tau)d\tau \Big| .$$

An application of the well-known Gronwall Lemma yields now the estimate

$$|f(t) - \hat{f}(t)| \leq C\Big(|f(0) - \hat{f}(0)| + \Big| \int_0^t \tilde{F}\delta(\tau)d\tau \Big|\Big) .$$

Since the integration kills one order of differentiation, our claim follows. □

Example 5.6. Continuing Example 5.2, we see why one speaks of Hessenberg systems of index 3. The perturbation index of such a system is indeed 3, as we obtained above an algebraic equation for every component f_i and the one for f_3 contains a second derivative of δ_3. This observation leads immediately to an estimate of the form (14) with $\nu = 3$.

A similar result can be obtained in some cases of partial differential equations. We do not give a rigorous theorem, but merely describe a situation where the above approach works with only minor modifications. The main assumption is that the system is of an evolutionary character (so that it makes sense to consider an initial value problem), as eventually we want to consider our partial differential equation as an abstract ordinary differential equation. Thus we assume that the independent variable x_n plays the role of time $x_n = t$ and choose on our ring $\mathscr{D} = \mathbb{k}[\partial_1, \ldots, \partial_n]$

of linear differential operators the TOP lift of a term order which gives precedence to terms containing the derivative $\partial_n = \partial_t$. Furthermore, we will always assume that the hyperplane $t = 0$ is not characteristic.

As a first step, we must generalise Definition 14 to partial differential equations. In principle, this is straighforward; however, one has a vast choice of possible function spaces and norms. We will restrict ourselves for simplicity to smooth functions; other choices should not lead to fundamentally different results. Let $C^\infty(\Omega, \mathbb{R}^m)$ denote the space of smooth $\mathbb{R}^m$-valued functions defined on some domain $\Omega \subseteq \mathbb{R}^n$. As in the ordinary case above, we introduce for functions $f \in C^\infty(\Omega, \mathbb{R}^m)$ and for any integer $\ell \in \mathbb{N}$ the Sobolev type norms

$$\|f\|_\ell = \sum_{0 \leq |\mu| \leq \ell} \Big\| \frac{\partial^{|\mu|} f}{\partial x^\mu} \Big\| \tag{15}$$

where $\|\cdot\|$ denotes some norm on $C^\infty(\Omega, \mathbb{R}^m)$.

Partial differential equations are usually accompanied by initial or boundary conditions which we accommodate by the following simple approach. Let $\Omega' \subset \Omega$ be a subdomain (typically of lower dimension) and introduce on $C^\infty(\Omega', \mathbb{R}^m)$ similar norms denoted by $\|\cdot\|'_\ell$ etc. The conditions are then written in the form $(Kf)|_{\Omega'} = 0$ for some linear differential operator K of order k. This setting comprises most kinds of initial or boundary conditions typically found in applications.

Definition 5.7. Let $Df = 0$ be a linear system of partial differential equations and let $f(\mathbf{x})$ be a smooth solution of it defined on the domain $\Omega \subseteq \mathbb{R}^m$ and satisfying some initial/boundary conditions $Kf = 0$ of order k on a subdomain $\Omega' \subset \Omega$. The system has the *perturbation index* ν along this solution, if ν is the smallest integer such that for any smooth solution $\hat{f}(\mathbf{x})$ of the perturbed system $D\hat{f} = \delta$ defined on Ω there exists an estimate

$$\big\|\hat{f}(\mathbf{x}) - f(\mathbf{x})\big\| \leq C\left(\|\hat{f} - f\|'_k + \|\delta\|_{\nu-1}\right), \tag{16}$$

whenever the right hand side is sufficiently small. The constant C may depend only on the domains Ω, Ω' and on the operator D.

Given the assumed evolutionary character of our linear system, we consider for simplicity a pure initial value problem with initial conditions prescribed on the hyperplane $t = 0$. As in the proof above, we first compute a Gröbner basis for the perturbed system $D\hat{f} = \delta$ and assume without loss of generality that it is of first order. If the system is not underdetermined, then our choice of the term order entails that (after differentiating potentially present algebraic equations) we find a subsystem of the form

$$\frac{\partial \hat{f}}{\partial t} = A\hat{f} + F\delta$$

where the linear differential operator A comprises only derivatives with respect to the remaining "spatial" variables $x_1, \ldots, x_{n-1}$. Our final assumption is that this operator A generates a strongly continuous semigroup $S(t)$ acting on an appropriately

chosen Banach space of functions of $x_1, \ldots, x_{n-1}$ and we consider $\hat{f}$ and δ as functions of t alone taking values in this Banach space. Essentially, we have thus transformed our partial differential equation into an abstract ordinary one.

By an elementary result in the theory of semigroups (see e. g. [39]), every solution of this subsystem can be written in the form

$$\hat{f}(t) = S(t)\hat{f}(0) + \int_0^t S(t-\tau)F\delta(\tau)d\tau$$

for an arbitrary element $\hat{f}(0)$ of the domain of A. Now a straightforward comparison with a solution $f(t)$ of the unperturbed system yields an estimate

$$\|f(t) - \hat{f}(t)\| \leq C\Big(\|f(0) - \hat{f}(0)\| + \int_0^t \|F\delta(\tau)\|d\tau\Big)$$

with a constant C, as the norms of elements $S(t)$ of a strongly continuous semigroup are bounded on any finite interval. Since the operator F is of order $\gamma_1 + 1$ with γ_1 the first Gröbner index, we finally obtain as bound for the perturbation index $\nu \leq \gamma_1 + 2$. We obtain a slightly worse result than in the case of an ordinary differential equation, as here we cannot necessarily argue that the integration kills one order of differentiation.[6]

Note that in both cases our basic strategy is the same: we derive a subsystem which can be taken as *underlying equation*, i. e. as a differential equation in Cauchy-Kovalevskaya form (in the ordinary case this simply corresponds to an explicit equation $f' = Af + g$) such that every solution of our system is also a solution of it. For such equations many results and estimates are known; above we used the Gronwall lemma and the theory of semigroups, respectively. The crucial assumption for obtaining the above estimates on the perturbation index is that we may consider all other equations merely as constraints on the permitted initial data. In the ordinary case, this is a triviality. For overdetermined systems of partial differential equations, one must be more careful. Under certain non-degenericity assumption on the used coordinates, one can show that this is indeed always the case (strictly speaking, we must assume that our Gröbner basis is in fact even a Pommaret basis). A detailed discussion of this point can be found in [52, Sect. 9.4].

6 Abstract Linear Systems

In his pioneering work, Kalman [19] developed an algebraic approach to control systems given by linear ordinary differential equations, using module theory and constructive methods. Independently, the school of Sato, Kashiwara et al. [21, 22, 34] was putting forward their algebraic analysis approach to linear systems of partial dif-

[6] In applications, it is actually quite rare that systems of partial differential equations contain algebraic equations. In this case, no differentiations are required and F is of order γ_1 so that we obtain the same estimate $\nu \leq \gamma_1 + 1$ as in the case of ordinary differential equations.

ferential equations, using homological algebra and sheaf theory. This work highly inspired Malgrange [32], Oberst [36], and Pommaret [41, 42, 43], when they were developing their seminal contributions to linear partial differential and difference systems with constant or variable coefficients. One of the successes of the resulting new line of research in mathematical systems theory was to realize that the behavioral approach to dynamical systems developed by the school of Willems [53] provided just the right language to tackle also control theoretic questions within this framework. For a recent survey on these topics, see [46]. In this section, we present a unified algebraic approach to the classes of linear systems discussed so far, based on ideas and results from the papers mentioned above and others. Our aim is to extract the common features of these system classes on a general and abstract level.

6.1 Some homological algebra

Let $\mathscr{D}$ be a ring (with 1) and let M,N,P be left $\mathscr{D}$-modules. Let $f : M \to N$ and $g : N \to P$ be $\mathscr{D}$-linear maps. The sequence

$$M \xrightarrow{f} N \xrightarrow{g} P$$

is called *exact* if $\mathrm{im}(f) = \ker(g)$. Let $\mathscr{F}$ be a left $\mathscr{D}$-module. The contravariant functor $\mathrm{Hom}_{\mathscr{D}}(\cdot,\mathscr{F})$ assigns to each left $\mathscr{D}$-module M the Abelian group (w.r.t. addition of functions) $\mathrm{Hom}_{\mathscr{D}}(M,\mathscr{F})$ consisting of all $\mathscr{D}$-linear maps from M to $\mathscr{F}$. Moreover, any $\mathscr{D}$-linear map $f : M \to N$ induces a group homomorphism

$$f' : \mathrm{Hom}_{\mathscr{D}}(N,\mathscr{F}) \to \mathrm{Hom}_{\mathscr{D}}(M,\mathscr{F}), \quad \varphi \mapsto \varphi \circ f. \tag{17}$$

The left $\mathscr{D}$-module $\mathscr{F}$ is said to be a *cogenerator* if the functor $\mathrm{Hom}_{\mathscr{D}}(\cdot,\mathscr{F})$ is *faithful*, that is, if $f \neq 0$ implies $f' \neq 0$, where f' is as in (17). In other words, for any $f \neq 0$ there exists φ such that $\varphi \circ f \neq 0$.

The functor $\mathrm{Hom}_{\mathscr{D}}(\cdot,\mathscr{F})$ is *left exact*, that is, if

$$M \xrightarrow{f} N \xrightarrow{g} P \to 0$$

is exact (i.e., both $M \to N \to P$ and $N \to P \to 0$ are exact), then the induced sequence

$$\mathrm{Hom}_{\mathscr{D}}(M,\mathscr{F}) \xleftarrow{f'} \mathrm{Hom}_{\mathscr{D}}(N,\mathscr{F}) \xleftarrow{g'} \mathrm{Hom}_{\mathscr{D}}(P,\mathscr{F}) \leftarrow 0$$

is again exact.

The left $\mathscr{D}$-module $\mathscr{F}$ is said to be *injective* if the functor $\mathrm{Hom}_{\mathscr{D}}(\cdot,\mathscr{F})$ is *exact*, that is, if exactness of

$$M \xrightarrow{f} N \xrightarrow{g} P$$

implies exactness of

$$\mathrm{Hom}_{\mathscr{D}}(M,\mathscr{F}) \xleftarrow{f'} \mathrm{Hom}_{\mathscr{D}}(N,\mathscr{F}) \xleftarrow{g'} \mathrm{Hom}_{\mathscr{D}}(P,\mathscr{F}).$$

6.2 Application to systems theory

Let $\mathscr{D}$ be a ring (with 1) and let $\mathscr{F}$ be a left $\mathscr{D}$-module. One may think of $\mathscr{F}$ as a set of functions, and of $\mathscr{D}$ as a set of operators acting on them. Any operator $d \in \mathscr{D}$ can be applied to any function $f \in \mathscr{F}$ to yield another function $df \in \mathscr{F}$. Given a positive integer q, this action naturally extends to $d \in \mathscr{D}^{1\times q}$ and $f \in \mathscr{F}^q$ via $df = \sum_{i=1}^{q} d_i f_i$. To any subset $D \subseteq \mathscr{D}^{1\times q}$, one may associate

$$D^{\perp} := \{f \in \mathscr{F}^q \mid \forall d \in D : df = 0\}.$$

This is the solution set, in $\mathscr{F}^q$, of the homogeneous linear system of equations given in terms of the coefficient rows $d \in D$. Note that $D^{\perp} = \langle D\rangle^{\perp}$, where $\langle D\rangle$ denotes the submodule of $\mathscr{D}^{1\times q}$ generated by the set D. In general, the set $D^{\perp} \subseteq \mathscr{F}^q$ has no particular algebraic structure besides being an Abelian group with respect to addition. Conversely, given a set $F \subseteq \mathscr{F}^q$, one considers

$${}^{\perp}F := \{d \in \mathscr{D}^{1\times q} \mid \forall f \in F : df = 0\}$$

which formalizes the set of all equations (given by their coefficient rows) satisfied by the given solutions $f \in F$. It is easy to check that ${}^{\perp}F$ is a left submodule of $\mathscr{D}^{1\times q}$. Thus we have a *Galois correspondence* between the left submodules of $\mathscr{D}^{1\times q}$ on the one hand, and the Abelian subgroups of $\mathscr{F}^q$ on the other. This means that $(\cdot)^{\perp}$ and ${}^{\perp}(\cdot)$ are both inclusion-reversing, and that

$$D \subseteq {}^{\perp}(D^{\perp}) \quad \text{and} \quad F \subseteq ({}^{\perp}F)^{\perp}$$

hold for all D and F as above. This implies that

$$({}^{\perp}(D^{\perp}))^{\perp} = D^{\perp} \quad \text{and} \quad {}^{\perp}(({}^{\perp}F)^{\perp}) = {}^{\perp}F.$$

A *linear system* $S \subseteq \mathscr{F}^q$ takes the form $S := E^{\perp}$ for some finite subset $E \subseteq \mathscr{D}^{1\times q}$. We may identify E with a matrix in $\mathscr{D}^{g\times q}$ and write $S = \{f \in \mathscr{F}^q \mid Ef = 0\}$. One calls E a *(kernel) representation matrix* of S. Consider the exact sequence

$$\mathscr{D}^{1\times g} \xrightarrow{\cdot E} \mathscr{D}^{1\times q} \to M := \mathscr{D}^{1\times q}/\mathscr{D}^{1\times g}E \to 0.$$

Applying the contravariant functor $\mathrm{Hom}_{\mathscr{D}}(\cdot,\mathscr{F})$, and using the standard isomorphism $\mathrm{Hom}_{\mathscr{D}}(\mathscr{D}^{1\times k},\mathscr{F}) \cong \mathscr{F}^k$ which relates a linear map defined on the free module $\mathscr{D}^{1\times k}$ to the image of the standard basis, one obtains an exact sequence

$$\mathscr{F}^g \xleftarrow{E} \mathscr{F}^q \leftarrow \mathrm{Hom}_{\mathscr{D}}(M,\mathscr{F}) \leftarrow 0.$$

This proves the *Malgrange isomorphism* [32]

$$S \cong \mathrm{Hom}_{\mathscr{D}}(M, \mathscr{F}),$$

which is an isomorphism of Abelian groups. One calls M the *system module* of S.

From now on, we shall assume that $\mathscr{D}$ is a left Noetherian ring. Then every submodule of $\mathscr{D}^{1\times q}$ is finitely generated, and thus, $D^{\perp}$ is a linear system for every subset $D \subseteq \mathscr{D}^{1\times q}$, because $D^{\perp} = \langle D\rangle^{\perp} = E^{\perp}$ for some finite set E.

For any linear system S, we have $({}^{\perp}S)^{\perp} = S$. If the left $\mathscr{D}$-module $\mathscr{F}$ is a cogenerator, then we also have ${}^{\perp}(D^{\perp}) = D$ for every submodule $D \subseteq \mathscr{D}^{1\times q}$. To see this, we show that $d \notin D$ implies $d \notin {}^{\perp}(D^{\perp})$. If $d \notin D$, then $0 \neq [d] \in M = \mathscr{D}^{1\times q}/D$ and thus the $\mathscr{D}$-linear map $g : \mathscr{D} \to M$ with $g(1) = [d]$ is nonzero. By the cogenerator property, there exists $\varphi \in \mathrm{Hom}_{\mathscr{D}}(M, \mathscr{F})$ such that $\varphi \circ g$ is nonzero, that is, $\varphi([d]) \neq 0$. Using the Malgrange isomorphism, this φ corresponds to $f \in S := D^{\perp}$ with $df \neq 0$. Thus $d \notin {}^{\perp}S = {}^{\perp}(D^{\perp})$.

Summing up: If $\mathscr{D}$ is left Noetherian and $\mathscr{F}$ a cogenerator, we have a duality between linear systems in $\mathscr{F}^q$ and submodules of $\mathscr{D}^{1\times q}$, that is, $(\cdot)^{\perp}$ and ${}^{\perp}(\cdot)$ are bijections and inverse to each other. More concretely, since $\mathscr{D}$ is left Noetherian, any submodule of $\mathscr{D}^{1\times q}$ can be written in the form $D = \mathscr{D}^{1\times g}E$ for some matrix $E \in \mathscr{D}^{g\times q}$. Then its associated system is $S = \{f \in \mathscr{F}^q \mid Ef = 0\}$, and we have both $S = D^{\perp}$ and $D = {}^{\perp}S$. For instance, let $E_i \in \mathscr{D}^{g_i\times q}$ be representation matrices of two systems $S_i \subseteq \mathscr{F}^q$. Then we have $S_1 \subseteq S_2$ if and only if $\mathscr{D}^{1\times g_1}E_1 \supseteq \mathscr{D}^{1\times g_2}E_2$, that is, if $E_2 = XE_1$ holds for some $X \in \mathscr{D}^{g_2\times g_1}$. In particular, $S_1 = \{0\}$ holds if and only if E_1 is left invertible.

If $\mathscr{F}$ is injective, then the exactness of

$$\mathscr{D}^{1\times m} \xrightarrow{\cdot E} \mathscr{D}^{1\times n} \xrightarrow{\cdot G} \mathscr{D}^{1\times p} \tag{18}$$

implies the exactness of

$$\mathscr{F}^m \xleftarrow{E} \mathscr{F}^n \xleftarrow{G} \mathscr{F}^p. \tag{19}$$

Thus the inhomogeneous system of equations $Gg = f$ (where $G \in \mathscr{D}^{n\times p}$ and $f \in \mathscr{F}^n$ are given) has a solution $g \in \mathscr{F}^p$ if and only if $f \in \mathrm{im}(G) = \ker(E)$, that is, if and only if $Ef = 0$. Since $EG = 0$ by construction, it is clear that $Gg = f$ implies $Ef = 0$. The crucial aspect of injectivity is that $Ef = 0$ is also sufficient for the existence of g. The resulting solvability condition for $Gg = f$ is often called the "fundamental principle": One computes a matrix E whose rows generate the left kernel $\ker(\cdot G) \subseteq \mathscr{D}^{1\times n}$ of G. In other words, one computes the "syzygies" of the rows of G as in Equation (25). Then the inhomogeneous system $Gg = f$ has a solution g if and only if the right hand side f satisfies the "compatibility condition" $Ef = 0$. If $\mathscr{F}$ is an injective cogenerator, then the exactness of (18) is in fact equivalent to the exactness of (19).

6.3 Autonomy

Let $\mathscr{D}$ be left Noetherian and let $\mathscr{F}$ be an injective cogenerator. Let q be a positive integer and let $S \subseteq \mathscr{F}^q$ be a linear system. The system S is said to be autonomous if it has no free variables, that is, if none of the projection maps

$$\pi_i : S \to \mathscr{F}, \quad f \mapsto f_i$$

is surjective for $1 \leq i \leq q$. Let E be a representation matrix of S and let $M = \mathscr{D}^{1\times q}/\mathscr{D}^{1\times g}E$ be the system module of S.

Theorem 6.1. *The following are equivalent:*

1. *S is autonomous.*
2. *For $1 \leq i \leq q$, there exist $0 \neq d_i \in \mathscr{D}$ and $X \in \mathscr{D}^{q\times g}$ such that* $\operatorname{diag}(d_1,\ldots,d_q) = XE$.

If $\mathscr{D}$ is a domain, then these conditions are also equivalent to:

3. *M is a torsion module, that is, for all $m \in M$ there exists $0 \neq d \in \mathscr{D}$ such that $dm = 0$.*
4. *E has full column rank.*

Note that the rank of E is well-defined, since a left Noetherian domain $\mathscr{D}$ possesses a quotient (skew) field $\mathscr{K} = \{d^{-1}n \mid d,n \in \mathscr{D}, d \neq 0\}$. Recall that the number $m := q - \operatorname{rank}(E)$ is known as the input dimension of S. Thus, the theorem above describes the situation where $m = 0$. In other words, S is not underdetermined.

Proof. Via the injective cogenerator property, a surjection $\pi_i : S \to \mathscr{F}$ exists if and only if there is an injection $j_i : \mathscr{D} \to M$ with $j_i(1) = [\mathbf{e}_i]$, where $\mathbf{e}_i \in \mathscr{D}^{1\times q}$ denotes the ith standard basis row. Hence autonomy is equivalent to j_i being noninjective for all $1 \leq i \leq q$, that is, $j_i(d_i) = [d_i\mathbf{e}_i] = 0 \in M$ for some $0 \neq d_i \in \mathscr{D}$, and then $d_i\mathbf{e}_i \in \mathscr{D}^{1\times g}E$.

Now let $\mathscr{D}$ be a domain. The following implications are straightforward: "3 $\Rightarrow$ 2 $\Rightarrow$ 4". We show "4 $\Rightarrow$ 2". Assume that $\operatorname{rank}(E) = q \leq g$. Let E_1 denote the matrix E after deleting the first column. Then $\operatorname{rank}(E_1) = q - 1 < g$ and thus there exists $0 \neq x \in \mathscr{K}^{1\times g}$ with $xE_1 = 0$. This x can be chosen such that $xE = [k,0,\ldots,0]$ for some $0 \neq k \in \mathscr{K}$. Writing $x = d^{-1}n_1$ for some $n_1 \in \mathscr{D}^{1\times g}$, we get $n_1E = [d_1,0,\ldots,0]$ with $0 \neq d_1 \in \mathscr{D}$. Proceeding like this with the remaining columns, we get $XE = \operatorname{diag}(d_1,\ldots,d_q)$ as desired. Finally, "2 $\Rightarrow$ 3" follows from the fact that in a left Noetherian domain, any two nonzero elements have a nonzero common left multiple, see [13, Ex. 4N] or [25, Cor. 10.23]. □

To test whether condition 2 is satisfied, define the augmented matrix

$$E_i' := \begin{bmatrix} \mathbf{e}_i \\ E \end{bmatrix} \in \mathscr{D}^{(1+g)\times q}.$$

Consider the left $\mathscr{D}$-module $\ker(\cdot E_i') := \{x \in \mathscr{D}^{1\times(1+g)} \mid xE_i' = 0\}$. Let

$$\pi : \mathscr{D}^{1\times(1+g)} \to \mathscr{D}, \quad x \mapsto x_1$$

denote the projection on the first component. Then S is autonomous if and only if $\pi(\ker(\cdot E_i')) \neq 0$ holds for all $1 \leq i \leq q$. Thus autonomy can be tested constructively if $\mathscr{D}$ admits the computation of generating sets of kernels of matrices (or, in other words, "syzygies" as described in Equation (25)).

Algorithm 2 Test for autonomy of system S

Input: kernel representation matrix E of S
Output: message "autonomous" or "not autonomous"
1: $q \leftarrow$ number of columns of E
2: **for all** $1 \leq i \leq q$ **do**
3: $E_i' \leftarrow \begin{bmatrix} \mathbf{e}_i \\ E \end{bmatrix}$
4: $F \leftarrow$ matrix whose rows generate $\ker(\cdot E_i')$
5: $F_1 \leftarrow$ set of first column entries of F
6: **if** $F_1 = \{0\}$ **then**
7: **return** "not autonomous" and **stop**
8: **end if**
9: **end for**
10: **return** "autonomous"

For $\mathscr{D} = \Bbbk[s_1, \ldots, s_r]$, that is, for partial differential or difference equations with constant coefficients, the notion of autonomy can be refined as follows. The *autonomy degree* of S [54] is defined to be $r - d$, where d denotes the maximal dimension of a cone in the complementary decomposition of the row module of E (note that the number d coincides with the Krull dimension of the system module M). Thus, nonautonomous systems have autonomy degree zero, and nonzero autonomous systems have autonomy degrees between 1 and r. The value r corresponds to systems which are finite-dimensional as $\Bbbk$-vector spaces, which is the strongest autonomy notion. Systems whose autonomy degree is at least 2 are always overdetermined. Analytic characterizations of this property in terms of the system trajectories are given in [47, 59].

6.4 Controllability

Let $\mathscr{D}$ be left and right Noetherian and let $\mathscr{F}$ be an injective cogenerator. Let q be a positive integer and let $S \subseteq \mathscr{F}^q$ be a linear system. The system S is said to be controllable if it has an image representation, that is, if there exists a matrix $L \in \mathscr{D}^{q\times l}$ such that

$$S = \{Lv \mid v \in \mathscr{F}^l\}.$$

The motivation for this definition comes from behavioral systems theory. There, controllability corresponds to concatenability of trajectories. Roughly speaking, this

amounts to being able to join any given "past" trajectory with any desired "future" trajectory by a connecting trajectory. For many relevant system classes, a properly defined version of this concatenability property has been shown to be equivalent to the existence of an image representation [40, 47, 53, 57].

Theorem 6.2. *Let E be a kernel representation matrix of S and let M be the system module of S. The following are equivalent:*

1. *S has an image representation.*
2. *E is a left syzygy matrix, that is, for some matrix $L \in \mathscr{D}^{q\times l}$, we have*

$$\mathrm{im}(\cdot E) = \{xE \mid x \in \mathscr{D}^{1\times g}\} = \ker(\cdot L) := \{y \in \mathscr{D}^{1\times q} \mid yL = 0\}.$$

3. *M is torsionless, that is, for every $0 \neq m \in M$ there exists $\varphi \in \mathrm{Hom}_{\mathscr{D}}(M, \mathscr{D})$ with $\varphi(m) \neq 0$.*

If $\mathscr{D}$ is a domain, then these conditions are also equivalent to:

4. *M is torsionfree, that is, $dm = 0$ with $d \in \mathscr{D}$ and $m \in M$ implies that $d = 0$ or $m = 0$.*

Proof. The sequence $\mathscr{D}^{1\times g} \xrightarrow{\cdot E} \mathscr{D}^{1\times q} \xrightarrow{\cdot L} \mathscr{D}^{1\times l}$ is exact if and only if $\mathscr{F}^g \xleftarrow{E} \mathscr{F}^q \xleftarrow{L} \mathscr{F}^l$ is exact. This proves "1 $\Leftrightarrow$ 2". For "2 $\Rightarrow$ 3", we use that $M = \mathscr{D}^{1\times q}/\mathrm{im}(\cdot E) = \mathscr{D}^{1\times q}/\ker(\cdot L) \cong \mathrm{im}(\cdot L) \subseteq \mathscr{D}^{1\times l}$ by the homomorphism theorem, and submodules of free modules are torsionless. For "3 $\Rightarrow$ 2", let K be a $\mathscr{D}$-matrix such that

$$\ker(E) = \{x \in \mathscr{D}^q \mid Ex = 0\} = \mathrm{im}(K) = \{Ky \mid y \in \mathscr{D}^k\}$$

and let $\bar{E}$ be a $\mathscr{D}$-matrix such that

$$\ker(\cdot K) = \{y \in \mathscr{D}^{1\times q} \mid yK = 0\} = \mathrm{im}(\cdot \bar{E}) = \{z\bar{E} \mid z \in \mathscr{D}^{1\times \bar{g}}\}.$$

Since $EK = 0$, we have $\mathrm{im}(\cdot E) \subseteq \mathrm{im}(\cdot \bar{E})$. The proof is finished if we can show that this inclusion is in fact an equality, because then $\mathrm{im}(\cdot E) = \mathrm{im}(\cdot \bar{E}) = \ker(\cdot K)$ and we may take $L = K$ in condition 2. Assume conversely that $d \in \mathrm{im}(\cdot \bar{E}) \setminus \mathrm{im}(\cdot E)$. Then $0 \neq [d] \in M$. Any homomorphism $\phi : M \to \mathscr{D}$ takes the form $\phi([d]) = dx$ for some $x \in \mathscr{D}^q$ which must satisfy $Ex = 0$ for well-definedness, and this implies $x = Ky$. Since $d = z\bar{E}$, we have $\phi([d]) = dx = z\bar{E}Ky = 0$. This contradicts the assumption of torsionlessness. Now let $\mathscr{D}$ be a domain. For "3 $\Rightarrow$ 4", suppose that $dm = 0$ and $0 \neq d$. Then any $\varphi \in \mathrm{Hom}_{\mathscr{D}}(M, \mathscr{D})$ satisfies $d\varphi(m) = 0$ and thus $\varphi(m) = 0$. By condition 3, this implies that $m = 0$. The converse implication holds as well, since M is finitely generated [13, Prop. 7.19]. □

To test whether condition 2 is satisfied, one proceeds as in the proof. One computes the $\mathscr{D}$-matrices K and $\bar{E}$ as described above. This requires that generating sets of kernels of matrices (that is, "syzygies" as in Equation (25)) can be computed over $\mathscr{D}$. The assumption that $\mathscr{D}$ is left and right Noetherian guarantees that left and right kernels are finitely generated. Then E is a left syzygy matrix if and only if

$$\mathrm{im}(\cdot E) = \mathrm{im}(\cdot \bar{E}).$$

This can be shown similarly as in the proof. The condition can be tested if $\mathscr{D}$ allows to decide module membership: It holds if and only if each row of $\bar{E}$ is contained in the row module of E. If $\mathscr{D}$ admits a Gröbner basis theory, then a row $d \in \mathscr{D}^{1\times q}$ belongs to a submodule $D \subseteq \mathscr{D}^{1\times q}$ if and only if its normal form with respect to a Gröbner basis of D is zero.

Algorithm 3 Test for controllability of system S

Input: kernel representation matrix E of S
Output: message "controllable" or "not controllable"
1: $\mathscr{G} \leftarrow$ Gröbner basis of row module of E
2: $K \leftarrow$ matrix whose columns generate $\ker(E)$
3: $\bar{E} \leftarrow$ matrix whose rows generate $\ker(\cdot K)$
4: **for all** rows d of $\bar{E}$ **do**
5: **if** `NormalForm`$(d,\mathscr{G}) \neq 0$ **then**
6: **return** "not controllable" and **stop**
7: **end if**
8: **end for**
9: **return** "controllable"

The data K and $\bar{E}$ computed by this algorithm are useful on their own: Due to $\mathrm{im}(\cdot\bar{E}) = \ker(\cdot K)$, we have

$$S_c := \{f \in \mathscr{F}^q \mid \bar{E}f = 0\} = \{Kv \mid v \in \mathscr{F}^k\}$$

because of the injectivity of $\mathscr{F}$. It turns out that S_c is the largest controllable subsystem of S, and that $S = S_c$ holds if and only if S is controllable. Thus, the factor group S/S_c (sometimes called the obstruction to controllability [56]) measures how far S is from being controllable.

The controllability test described above has been developed, for certain operator domains $\mathscr{D}$, in a series of papers [41, 42, 43] by J.F. Pommaret and A. Quadrat, see also [6]. These authors also introduced a concept of controllability degrees, similar to the autonomy degrees. However, the definition of the controllability degrees is more involved and uses extension functors (for an introduction in the systems theoretic setting, see e.g. [55]). The controllability test has been generalized to a large class of noncommutative rings with zero-divisors in [60].

7 Appendix: Gröbner Bases

Gröbner bases are a fundamental tool in constructive algebra, as they permit to perform many basic algebraic constructions algorithmically. They were formally introduced for ideals in a polynomial ring in the Ph.D. thesis of Buchberger [4] (written under the supervision of Gröbner); modern textbook presentations can be found e. g.

in [1, 8]. Most general purpose computer algebra systems like Maple or Mathematica provide an implementation. However, the computation of a Gröbner basis can be quite demanding (with respect to both time and space) and for larger examples the use of specialised systems like CoCoA,[7] Macaulay 2[8], Magma[9] or Singular[10] is recommended.

Gröbner bases were originally introduced for ideals in the standard commutative polynomial ring $\mathscr{P} = \mathbb{k}[x_1, \ldots, x_n]$ where $\mathbb{k}$ is any field (e. g. $\mathbb{k} = \mathbb{R}$ or $\mathbb{k} = \mathbb{C}$). Since then, several generalisations to non-commutative rings have been studied. We will present here one such extension covering all rings appearing in this survey: *polynomial rings of solvable type*. This class of algebras was first introduced in [20]; further studies can be found in [23, 29, 50, 52]. Furthermore, we will also discuss besides ideals the case of submodules of a free module $\mathscr{P}^m$. A general introduction to modules over certain non-commutative rings covering also algorithmic aspects and Gröbner bases was recently given by Gómez-Torrecillas [12].

Gröbner bases are always defined with respect to a *term order*. In the polynomial ring $\mathscr{P}$ there does not exist a natural ordering on the set $\mathbb{T}$ of all terms x^μ with an exponent vector $\mu \in \mathbb{N}_0^n$ (only in the univariate case the degree provides a canonical ordering). However, the use of such an ordering is crucial for many purposes like extending the familiar polynomial division to the multivariate case. Elements of the free module $\mathscr{P}^m$ can be represented as m-dimensional column (or row) vectors where each component is a polynomial. Here a "term" $\mathbf{t}$ is a vector where all components except one vanish and the non-zero component consists of a term $t \in \mathbb{T}$ in the usual sense and thus can be written as $\mathbf{t} = t\mathbf{e}_i$ where $\mathbf{e}_i$ denotes the ith vector in the standard basis of $\mathscr{P}^m$. We denote the set of all such vector terms by $\mathbb{T}^m$. We say that $\mathbf{t}$ *divides* another term $\mathbf{s} = s\mathbf{e}_j$, written $\mathbf{t} \mid \mathbf{s}$, if $i = j$ and $t \mid s$, i. e. only terms living in the same component can divide each other.

Definition 7.1. A total order $\prec$ on $\mathbb{T}$ is a *term order*, if it satisfies: (i) given three terms $r, s, t \in \mathbb{T}$ such that $s \prec t$, we also have $rs \prec rt$ (monotonicity), and (ii) any term $t \in \mathbb{T}$ different from 1 is greater than 1. A term order for which additionally terms of higher degree are automatically greater than terms of lower degree is called *degree compatible*.

A total order $\prec$ on $\mathbb{T}^m$ is a *module term order*, if it satisfies: (i) for two vector terms $\mathbf{s}, \mathbf{t} \in \mathbb{T}^m$ with $\mathbf{s} \prec \mathbf{t}$ and an ordinary term $r \in \mathbb{T}$, we also have $r\mathbf{s} \prec r\mathbf{t}$ and (ii) for any $\mathbf{t} \in \mathbb{T}^m$ and $s \in \mathbb{T}$, we have $\mathbf{t} \prec s\mathbf{t}$.

Given a (non-zero) polynomial $f \in \mathscr{P}$ and a term order $\prec$, we can sort the finitely many terms actually appearing in f according to $\prec$. We call the largest one the *leading term* $\operatorname{lt} f$ of f and its coefficient is the *leading coefficient* $\operatorname{lc} f$; finally, the

[7] `http://cocoa.dima.unige.it`

[8] `http://www.math.uiuc.edu/Macaulay2`

[9] `http://magma.usyd.edu.au`

[10] `http://www.singular.uni-kl.de`

leading monomial [11] of f is then the product $\operatorname{lm} f = \operatorname{lc}(f)\operatorname{lt}(f)$. The same notations are also used for polynomial vectors $\mathbf{f} \in \mathscr{P}^m$.

Example 7.2. Usually, term orders are introduced via the exponent vectors. The *lexicographic order* is defined as follows: $x^\mu \prec_{\text{lex}} x^\nu$, if the first non-vanishing entry of $\mu - \nu$ is negative. Thus it implements the ordering used for words in a dictionary: if we take $x_1 = a$, $x_2 = b$ etc, then $ab^2 \prec_{\text{lex}} a^2$ and a^2 is ordered before ab^2, although it is of lower degree. The lexicographic order is very useful in elimination problems and for solving polynomial equations. Unfortunately, it tends to be rather inefficient in Gröbner bases computations.

An example of a degree compatible order is the *reverse lexicographic order*: $x^\mu \prec_{\text{revlex}} x^\nu$, if $\deg x^\mu < \deg x^\nu$ or $\deg x^\mu = \deg x^\nu$ and the last non-vanishing entry of $\mu - \nu$ is positive. Now we find $a^2 \prec_{\text{revlex}} ab^2$ because of the different degrees and $a^3 \prec_{\text{revlex}} ab^2$ because only the latter term contains b. Usually, the reverse lexicographic order is the most efficient order for Gröbner bases computations.

Given a term order $\prec$ on $\mathbb{T}$, there exist two natural ways to lift it to a module term order on $\mathbb{T}^m$. In the *(ascending) TOP lift*, we put *t*erm *o*ver *p*osition and define $s\mathbf{e}_i \prec_{\text{TOP}} t\mathbf{e}_j$, if $s \prec t$ or $s = t$ and $i < j$ (in the descending TOP lift one uses $i > j$). The *(ascending) POT lift* works the other way round: $s\mathbf{e}_i \prec_{\text{POT}} t\mathbf{e}_j$, if $i < j$ or $i = j$ and $s \prec t$ (again we use $i > j$ for the descending version). Such lifts are the most commonly used module term orders.

Finally, assume that $F = \{\mathbf{f}_1, \ldots, \mathbf{f}_r\} \subset \mathscr{P}^m$ is a finite set of r (non-zero) vectors and $\prec$ a module term order on $\mathbb{T}^m$. Then the set F induces a module term order $\prec_F$ on $\mathbb{T}^r \subset \mathscr{P}^r$ as follows: $s\mathbf{e}_i \prec_F t\mathbf{e}_j$, if $\operatorname{lt}(s\mathbf{f}_i) \prec \operatorname{lt}(t\mathbf{f}_j)$ or $\operatorname{lt}(s\mathbf{f}_i) = \operatorname{lt}(t\mathbf{f}_j)$ and $i > j$ (no, the direction of this relation is not a typo!). As we will see later, this induced order is very important for computing the syzygies of the set F.

Let $\star$ be a non-commutative multiplication on $\mathscr{P}$. We allow both that our variables x_i do no longer commute, i. e. $x_i \star x_j \neq x_j \star x_i$, and that the variables act on the coefficients $c \in \Bbbk$, i. e. $x_i \star c \neq c \star x_i$. However, we do not permit that the coefficients act on the variables, i. e. $c \star x_i = cx_i$ where on the right hand side the usual product is used. A prototypical example are linear differential operators where we may choose $\Bbbk = \mathbb{R}(t_1, \ldots, t_n)$, the fields of real rational functions in some unknowns $t_1, \ldots, t_n$, and take as "variables" for the polynomials the partial derivatives with respect to these unknowns, $x_i = \partial/\partial t_i$. Here we still find $x_i \star x_j = x_j \star x_i$, as for smooth functions partial derivatives commute, but $x_i \star c = cx_i + \partial c/\partial t_i$ for any rational function $c \in \Bbbk$. Non-commutative variables occur e. g. in the Weyl algebra or in "quantised algebras" like q-difference operators.

For the definition of Gröbner bases in such non-commutative polynomial rings, it is important that the product does not interfere with the chosen term order. This motivates the following definition of a special class of polynomial rings.

[11] For us a *term* is a pure power product x^μ whereas a *monomial* is of the form cx^μ with a coefficient $c \in \Bbbk$; beware that some text books on Gröbner bases use the words term and monomial with exactly the opposite meaning.

Definition 7.3. Let $\prec$ be a term order and $\star$ a non-commutative product on the polynomial ring $\mathscr{P} = \Bbbk[x_1,\ldots,x_n]$. The triple $(\mathscr{P},\star,\prec)$ defines a *solvable polynomial ring*, if the following three conditions are satisfied:

(i) $(\mathscr{P},\star)$ is a ring;
(ii) $c\star f = cf$ for all coefficients $c\in\Bbbk$ and polynomials $f\in\mathscr{P}$;
(iii) $\mathrm{lt}(f\star g) = \mathrm{lt}(f)\cdot\mathrm{lt}(g)$ for all polynomials $f,g\in\mathscr{P}\setminus\{0\}$ (note the use of the ordinary commutative product on the right hand side!).

The first condition ensures that the arithmetics in $\mathscr{P}$ satisfies all the usual rules like associativity or distributivity. The second condition implies that $\mathscr{P}$ is still a $\Bbbk$-linear space (as long as we multiply with field elements only from the left). The third condition enforces the compatibility of the new product $\star$ with the chosen term order: the non-commutativity does not affect the leading terms. If such a compatibility holds, then the usual commutative Gröbner bases theory remains valid in our more general setting without any changes.

Example 7.4. All non-commutative rings appearing in this article belong to a subclass of the solvable polynomial rings, namely the Ore algebras which may be considered as generalisations of the ring of linear differential operators. This class was first considered by Noether und Schmeidler [35] and then more extensively by Ore [38]; our presentation follows [3].

Let $\sigma:\Bbbk\to\Bbbk$ be an automorphism of the field $\Bbbk$. A *pseudo-derivation* with respect to σ is a map $\delta:\Bbbk\to\Bbbk$ such that $\delta(c+d)=\delta(c)+\delta(d)$ and $\delta(cd)=\sigma(c)\delta(d)+\delta(c)d$ for all $c,d\in\Bbbk$. If $\sigma=\mathrm{id}$, the identity map, the second condition is the standard Leibniz rule for derivations. $\sigma(c)$ is called the *conjugate* and $\delta(c)$ the *derivative* of $c\in\Bbbk$.

Given the maps σ and δ, the ring $\Bbbk[\partial;\sigma,\delta]$ of univariate Ore polynomials consists of all polynomials $\sum_{i=0}^{d} c_i\partial^i$ in ∂ with coefficients $c_i\in\Bbbk$. The addition is defined as usual. The "variable" ∂ operates on an element $c\in\Bbbk$ according to $\partial\star c=\sigma(c)\partial+\delta(c)$. Note that we may interpret this equation as a rewrite rule which tells us how to bring a ∂ from the left to the right of a coefficient. This rewriting can be used to define the multiplication $\star$ on the whole ring $\Bbbk[\partial;\sigma,\delta]$: given two elements $\theta_1,\theta_2\in\Bbbk[\partial;\sigma,\delta]$, we can transform the product $\theta_1\star\theta_2$ to the normal form of a polynomial (coefficients to the left of the variable) by repeatedly applying our rewrite rule. The product of two linear polynomials evaluates then to

$$\begin{aligned}(f_1+f_2\partial)\star(g_1+g_2\partial) = f_1g_1+f_2\delta(g_1)+{}&\\ \big[f_1g_2+f_2\sigma(g_1)+f_2\delta(g_2)\big]\partial+f_2\sigma(g_2)\partial^2\,.&\end{aligned}\tag{20}$$

The fact that σ is an automorphism ensures that $\deg(\theta_1\star\theta_2)=\deg\theta_1+\deg\theta_2$. We call $\Bbbk[\partial;\sigma,\delta]$ the *Ore algebra* generated by σ and δ.

A simple familiar example is given by $\Bbbk=\mathbb{Q}(x)$, $\delta=\frac{d}{dx}$ and $\sigma=\mathrm{id}$ yielding linear ordinary differential operators with rational functions as coefficients. Similarly, we can obtain recurrence and difference operators. We set $\Bbbk=\mathbb{C}(n)$, the field of sequences $(s_n)_{n\in\mathbb{Z}}$ with complex elements $s_n\in\mathbb{C}$, and take for σ the shift operator,

i. e. the automorphism mapping s_n to s_{n+1}. Then $\Delta = \sigma - \mathrm{id}$ is a pseudo-derivation. $\Bbbk[E;\sigma,0]$ consists of linear ordinary recurrence operators, $\Bbbk[E;\sigma,\Delta]$ of linear ordinary difference operators.

For multivariate Ore polynomials, we take a set $\Sigma = \{\sigma_1,\ldots,\sigma_n\}$ of automorphisms and a set $\Delta = \{\delta_1,\ldots,\delta_n\}$ where each δ_i is a pseudo-derivation with respect to σ_i. For each pair (σ_i,δ_i) we introduce a "variable" ∂_i satisfying a commutation rule as in the univariate case. If we require that all the maps σ_i,δ_j commute with each other, one easily checks that $\partial_i \star \partial_j = \partial_j \star \partial_i$, i. e. the "variables" ∂_i also commute. Setting $D = \{\partial_1,\ldots,\partial_n\}$, we denote by $\Bbbk[D;\Sigma,\Delta]$ the ring of multivariate Ore polynomials. Because of the commutativity of the variables ∂_i, we may write the terms as ∂^μ with multi indices $\mu \in \mathbb{N}_0^n$, so that it indeed makes sense to speak of a polynomial ring. The proof that $\big(\Bbbk[D;\Sigma,\Delta],\star,\prec\big)$ is a solvable polynomial ring for any term order $\prec$ is trivial.

From now on, we always assume that we have chosen a fixed solvable polynomial algebra $(\mathscr{P},\star,\prec)$. All references to a leading term etc. are then meant with respect to the term order $\prec$ contained in this choice. In a non-commutative ring we must distinguish between left, right and two-sided ideals. In this appendix, we exclusively deal with left ideals: if $F = \{f_1,\ldots,f_r\} \subset \mathscr{P}$ is some finite set of polynomials, then the left ideal generated by the basis F is the set

$$\langle F\rangle = \Big\{ \sum_{\alpha=1}^{r} g_\alpha \star f_\alpha \mid g_\alpha \in \mathscr{P} \Big\} \tag{21}$$

of all left linear combinations of the elements of F and it satisfies $g\star f \in \langle F\rangle$ for any $f \in \langle F\rangle$ and any $g \in \mathscr{P}$. Right ideals are defined correspondingly[12] and a two-sided ideal is simultaneously a left and a right ideal.

A for computational purposes highly relevant property of the commutative polynomial ring is that it is *Noetherian*, i. e. any ideal in it possesses a finite basis (Hilbert's Basis Theorem). For solvable polynomial rings, the situation is more complicated. [52, Sect. 3.3] collects a number of possible approaches to prove for large classes of such rings that they are Noetherian, too. In particular, this is the case for all Ore algebras. In [52, Prop. 3.2.10], it is also shown that all solvable algebras (over a coefficient field) satisfy the left Ore condition so that one can define a left quotient skew field [25].

Remark 7.5. A complication in the treatment of non-commutative polynomial rings is given by the fact that in general the product of two terms is no longer a term. Hence the notion of a monomial ideal makes no longer sense. In the sequel, we will use the convention that when we speak about the divisibility of terms $s,t \in \mathbb{T}$, this is always to be understood within the *commutative* polynomial ring, i. e. $s \mid t$, if and only if a further term $r \in \mathbb{T}$ exists such that $r\cdot s = s\cdot r = t$. In other words, we consider $(\mathbb{T},\cdot)$ as an Abelian monoid. Given a left ideal $I \lhd \mathscr{P}$ in a solvable polynomial ring, we then introduce within this monoid the *leading ideal* $\mathrm{lt}\, I = \langle \mathrm{lt}\, f \mid f \in I \setminus \{0\}\rangle$. Thus

[12] Beware that the left and the right ideal generated by a set F are generally different.

$\operatorname{lt} I$ is always to be understood as a monomial ideal in the *commutative* polynomial ring, even if the considered polynomial ring $\mathscr{P}$ has a non-commutative product $\star$.

Definition 7.6. Let $I \lhd \mathscr{P}$ be a left ideal in the polynomial ring $\mathscr{P}$. A finite set $G \subset I$ is a *Gröbner basis* of I, if for every non-zero polynomial $f \in I$ in the ideal a generator $g \in G$ exists such that $\operatorname{lt} g \mid \operatorname{lt} f$ and thus $\operatorname{lt} I = \langle \operatorname{lt} G \rangle$.

Above definition is rather technical and it is neither evident that Gröbner bases exist nor that they are useful for anything. We will now show that they allow us to solve effectively the *ideal membership problem*: given an ideal $I = \langle F \rangle \lhd \mathscr{P}$ and a polynomial $f \in \mathscr{P}$, decide whether or not $f \in I$. A solution of this problem is for example mandatory for an effective arithmetics in the factor ring $\mathscr{P}/I$. As a by-product, we will see that a Gröbner basis is indeed a basis, i. e. $I = \langle G \rangle$.

Definition 7.7. Let $G = \{g_1, \ldots, g_r\} \subset \mathscr{P} \setminus \{0\}$ be a finite set of non-zero polynomials. A further polynomial $f \in \mathscr{P} \setminus \{0\}$ is called *reducible* with respect to G, if f contains a term $t \in \mathbb{T}$ for which a polynomial $g_i \in G$ exists such that $\operatorname{lt} g_i \mid t$. If this is the case, then a term $s \in \mathbb{T}$ exists such that $\operatorname{lt}(s \star g_i) = t$ and we can perform a *reduction step*: $f \to_{g_i} \tilde{f} = f - (c/\operatorname{lc}(s \star g_i)) s \star g_i$ where $c \in \Bbbk$ denotes the coefficient of t in f.[13] A polynomial $h \in \mathscr{P}$ is a *normal form* of f modulo the set G, if we can find a sequence of reduction steps

$$f \to_{g_{i_1}} h_1 \to_{g_{i_2}} h_2 \to_{g_{i_3}} \cdots \to_{g_{i_s}} h_s = h \tag{22}$$

and h is not reducible with respect to G. In this case, we also write shortly $f \to_G^+ h$ for (22) or $h = \operatorname{NF}(f, G)$.

It should be noted that the notation $h = \operatorname{NF}(f, G)$ is somewhat misleading. A polynomial f may have many different normal forms with respect to some set G, as generally different sequences of reduction steps lead to different results. If h is some normal form of the polynomial f, then (non-unique) *quotients* $q_1, \ldots, q_r \in \mathscr{P}$ exist such that $f = \sum_{i=1}^r q_i \star g_i + h$. Thus $h = 0$ immediately implies $f \in \langle G \rangle$; however, the converse is not necessarily true: even if $f \in \langle G \rangle$, it may possess non-vanishing normal forms. In the univariate case (and for $r = 1$), we recover here the familar polynomial division from high school where both the normal form (or *remainder*) h and the quotient q are unique. A concrete multivariate *division algorithm* for computing a normal form together with some quotients is shown in Algorithm 4.

In this article, we are always assuming that our polynomials are defined over a field $\Bbbk$. This assumption entails that the divisibility of two monomials cx^μ and dx^ν is decided exclusively by the contained terms x^μ and x^ν. Over a coefficient ring $\mathscr{R}$, this is no longer the case: note that in Line /5/ of Algorithm 4 we must perform the division $\operatorname{lc} f / \operatorname{lc}(s \star g_i)$ which may not be possible in a coefficient ring. Under

[13] The term "reduction" refers to the fact that the monomial ct in f is replaced by a linear combination of terms which are all smaller than t with respect to the used term order. It does *not* imply that $\tilde{f}$ is simpler in the sense that it has less terms than f. In fact, quite often the opposite is the case!

Algorithm 4 Multivariate polynomial division

Input: finite set $G = \{g_1, \ldots, g_r\} \subset \mathscr{P} \setminus \{0\}$, polynomial $f \in \mathscr{P}$
Output: a normal form $h = \mathrm{NF}(f, G)$, quotients $q_1, \ldots, q_r$

$h \leftarrow 0$; $q_1 \leftarrow 0$; $\ldots$ $q_r \leftarrow 0$
while $f \neq 0$ **do**
  if $\exists\, i : \mathrm{lt}\, g_i \mid \mathrm{lt}\, f$ **then**
    choose smallest index i with this property and $s \in \mathbb{T}$ such that $\mathrm{lt}\,(s \star g_i) = t$
    $m \leftarrow \frac{\mathrm{lc}\, f}{\mathrm{lc}\,(s \star g_i)} s$; $f \leftarrow f - m \star g_i$; $q_i \leftarrow q_i + m$
  else
    $h \leftarrow h + \mathrm{lm}\, f$; $f \leftarrow f - \mathrm{lm}\, f$
  end if
end while
return $h, q_1, \ldots, q_r$

certain technical assumptions on the ring $\mathscr{R}$, it is still possible to set up a theory of Gröbner bases which, however, becomes more complicated. For details on this extension, we refer to the literature, see e. g. [1].

The following fundamental characterisation theorem collects a number of equivalent definitions for Gröbner bases. It explains their distinguished position among all possible bases of a given ideal. In particular, (ii) already solves the ideal membership problem. Furthermore, (iii) implies that for a Gröbner basis G the notation $NF(f, G)$ is well-defined, as in this case any sequence of reduction steps leads to same final result.

Theorem 7.8. *Let $0 \neq I \lhd \mathscr{P}$ be an ideal and $G \subset I$ a finite subset. Then the following statements are equivalent.*

(i) *G is a Gröbner basis of I.*
(ii) *Given a polynomial $f \in \mathscr{P}$, ideal membership $f \in I$ is equivalent to $f \to_G^+ 0$.*
(iii) *$I = \langle G \rangle$ and every $f \in \mathscr{P}$ has a* unique *normal form with respect to G.*
(iv) *A polynomial $f \in \mathscr{P}$ is contained in the ideal I, if and only if it possesses a* standard representation *with respect to G, i. e. there are coefficients $h_g \in \mathscr{P}$ such that $f = \sum_{g \in G} h_g \star g$ and $\mathrm{lt}\,(h_g \star g) \preceq \mathrm{lt}\, f$ whenever $h_g \neq 0$.*

Obviously, we may also consider the ring $\mathscr{P}$, any ideal $I \lhd \mathscr{P}$ and the corresponding factor space $\mathscr{P}/I$ as $\Bbbk$-linear spaces. The following observation shows why it is of interest to know the leading ideal $\mathrm{lt}\, I$.

Theorem 7.9 (Macaulay). *Let $I \lhd \mathscr{P}$ be an ideal and $\prec$ an arbitrary term order. Then $\mathscr{P}/I$ and $\mathscr{P}/\mathrm{lt}\, I$ are isomorphic as $\Bbbk$-linear spaces.*

Proof (Sketch). Denote by $\mathscr{B} = \mathbb{T} \setminus \mathrm{lt}\, I$ the set of all terms *not* contained in the leading ideal. One now shows that the respective equivalence classes of the elements of $\mathscr{B}$ define a $\Bbbk$-linear basis of $\mathscr{P}/I$ and $\mathscr{P}/\mathrm{lt}\, I$, respectively. The linear independence is fairly obvious and $\mathscr{B}$ induces generating sets, as the normal form of any polynomial $f \in \mathscr{P}$ with respect to a Gröbner basis is a $\Bbbk$-linear combination of elements of $\mathscr{B}$.

The use of the term "basis" in commutative algebra is a bit misleading, as one does not require linear independence. Opposed to the situation in linear algebra, elements of an ideal generally do not possess a unique representation as linear combination of the basis. For homogeneous ideals, which may also be considered as graded vector spaces, the following concept of a combinatorial composition leads again to unique representations.

Definition 7.10. Let $I \lhd \mathscr{P}$ be a homogeneous ideal. A *Stanley decomposition* of I is an isomorphism as graded vector spaces

$$I \cong \bigoplus_{t \in \mathscr{T}} \mathbb{k}[X_t] \cdot t \tag{23}$$

with a finite set $\mathscr{T} \subset \mathbb{T}$ of terms and subsets $X_t \subseteq \{x_1, \ldots, x_n\}$. The elements of X_t are called the *multiplicative variables* of the generator t. One speaks of a *Rees decomposition*, if all sets of multiplicative variables are of the form $X_t = \{x_1, x_2, \ldots, x_{k_t}\}$ where $0 \leq k_t \leq n$ is called the *class* of t. A *complementary (Stanley) decomposition* is an analogous isomorphism for the factor space $\mathscr{P}/I$.

Vector spaces of the form $\mathbb{k}[X_t] \cdot t$ are called *cones* and the number of multiplicative variables is the *dimension* of such a cone. While Stanley decompositions are anything but unique and different decompositions may consist of differently many cones, one can show that the highest appearing dimension of a cone is always the same (the dimension of the ideal I) and also the number of cones with this particular dimension (algebraically it is given by the multiplicity or degree of the ideal) is an invariant. This observation is a simple consequence of the connection between complementary decompositions and Hilbert polynomials (or functions) which, however, cannot be discussed here (see e. g. [52] and references given there). Complementary decomposition got their name from the simple observation that in the monomial case they are equivalent to expressing the complement $\mathscr{P} \setminus I$ as a direct sum of cones. Concrete examples are shown in Figures 1 and 2 on pages 13 and 15.

Because of Macaulay's Theorem 7.9, it indeed suffices for complementary decompositions to consider $\mathscr{P}/\operatorname{lt} I$ and thus monomial ideals. This observation reduces the task of their construction to a purely combinatorial problem. A simple solution is provided by the recursive Algorithm 5. It takes as input the minimal basis of I and returns a set of pairs (t, X_t) consisting of a generator t and its multiplicative variables. The recursion is on the number n of variables in the polynomial ring $\mathscr{P}$. If $\nu = (\nu_1, \ldots, \nu_n) \in \mathbb{N}_0^n$ is an exponent vector, we denote by $\nu' = [\nu_1, \ldots, \nu_{n-1}]$ its truncation to the first $n-1$ entries and write $\nu = [\nu', \nu_n]$. We remark that a special type of Gröbner bases, the *involutive bases* (first introduced by Gerdt and Blinkov [11]), is particularly adapted to this problem [52, Sect. 5.1]. We refer to [50, 51] and [52, Chapt. 3] for an extensive treatment of these bases and further references.

Despite its great theoretical importance, Theorem 7.8 still does not settle the question of the existence of Gröbner bases, as none of the given characterisations is effective. For a constructive approach, we need syzygies. The fundamental tool is the *S-polynomial* of two polynomials $f, g \in \mathscr{P}$. Let $x^\mu = \operatorname{lt} f$, $x^\nu = \operatorname{lt} g$ be the

Algorithm 5 Complementary decomposition of monomial ideal

Input: minimal basis $\mathscr{B}$ of monomial ideal $I \subset \mathscr{P}$
Output: finite complementary decomposition $\mathscr{T}$ of $\bar{I}$

if n=1 **then** {in this case $\mathscr{B} = \{x^\nu\}$ with $\nu \in \mathbb{N}$}
  $q_0 \leftarrow \nu$; $\mathscr{T} \leftarrow \{([x^0], \emptyset), \ldots, ([x^{q_0-1}], \emptyset)\}$
else
  $q_0 \leftarrow \max\{\nu_n \mid x^\nu \in \mathscr{B}\}$; $\mathscr{T} \leftarrow \emptyset$
  for q **from** 0 **to** q_0 **do**
    $\mathscr{B}'_q \leftarrow \{x^{\nu'} \in \mathbb{N}_0^{n-1} \mid x^\nu \in \mathscr{B},\ \nu_n \le q\}$
    $\mathscr{T}'_q \leftarrow$ `ComplementaryDecomposition`$(\mathscr{B}'_q)$
    if $q < q_0$ **then**
      $\mathscr{T} \leftarrow \mathscr{T} \cup \{(x^{[\nu',q]}, X_{\nu'}) \mid (x^{\nu'}, X_{\nu'}) \in \mathscr{T}'_q\}$
    else
      $\mathscr{T} \leftarrow \mathscr{T} \cup \{(x^{[\nu',q]}, X_{\nu'} \cup \{n\}) \mid (x^{\nu'}, X_{\nu'}) \in \mathscr{T}'_q\}$
    end if
  end for
end if
return $\mathscr{T}$

corresponding leading terms and $x^\rho = \operatorname{lcm}(x^\mu, x^\nu)$ their least common multiple. If $x^\rho = x^{\bar{\mu}} x^\mu = x^{\bar{\nu}} x^\nu$ (in the commutative sense, cf. Remark 7.5), then the S-polynomial of f and g is defined as the difference

$$S(f,g) = \frac{x^{\bar{\mu}} \star f}{\operatorname{lc}(x^{\bar{\mu}} \star f)} - \frac{x^{\bar{\nu}} \star g}{\operatorname{lc}(x^{\bar{\nu}} \star g)}. \tag{24}$$

Note that the coefficients are chosen in such a way that the leading monomials cancel in the subtraction. With the help of this construction, one can provide an effective criterion for a set to be a Gröbner basis of an ideal.

Theorem 7.11 (Buchberger). *A finite set $G \subset \mathscr{P}$ of polynomials is a Gröbner basis of the left ideal $I = \langle G \rangle$ generated by it, if and only if for every pair $f, g \in G$ the S-polynomial $S(f,g)$ reduces to zero with respect to G.*

This theorem translates immediately into a simple algorithm for the effective construction of Gröbner bases, the *Buchberger Algorithm 6*, and also ensures its correctness. The termination is guaranteed, if the solvable polynomial ring $\mathscr{P}$ is Noetherian (which is the case for all rings appearing in this article). It should be noted that the basic form of the Buchberger algorithm shown here can handle only very small examples. A version able to handle substantial problems requires many optimisations and the development and improvement of efficient implementations is still an active field of research.

Gröbner bases are anything but unique: if G is a Gröbner basis of the ideal I for some term order $\prec$, then we may extend G by arbitrary elements of I and still have a Gröbner basis. For obtaining uniqueness, one must impose further conditions on the basis. It is easy to show that a monomial ideal I (in the commutative polynomial ring) always possesses a unique *minimial* basis $\mathscr{B}$ consisting entirely of monomials. Minimial means here that no element of $\mathscr{B}$ divides another element.

Algorithm 6 Gröbner basis (Buchberger)

Input: finite set $F \subset \mathscr{P}$, term order $\prec$
Output: Gröbner basis G of the ideal $\langle F \rangle$
1: $G \leftarrow F$
2: $\mathscr{S} \leftarrow \big\{\{g_1, g_2\} \mid g_1, g_2 \in G, g_1 \neq g_2\big\}$
3: **while** $\mathscr{S} \neq \emptyset$ **do**
4: choose $\{g_1, g_2\} \in \mathscr{S}$
5: $\mathscr{S} \leftarrow \mathscr{S} \setminus \big\{\{g_1, g_2\}\big\}$; $\bar{g} \leftarrow \mathrm{NF}\big(S(g_1, g_2), G\big)$
6: **if** $\bar{g} \neq 0$ **then**
7: $\mathscr{S} \leftarrow \mathscr{S} \cup \big\{\{\bar{g}, g\} \mid g \in G\big\}$; $G \leftarrow G \cup \{\bar{g}\}$
8: **end if**
9: **end while**
10: **return** G

Definition 7.12. A Gröbner basis G of an ideal $I \lhd \mathscr{P}$ is called *minimal*, if the set $\mathrm{lt}\, G$ is the minimal basis of the monomial ideal $\mathrm{lt}\, I$. We call G a *reduced Gröbner basis*, if every generator $g \in G$ is in normal form with respect to $G \setminus \{g\}$ and every leading coefficient $\mathrm{lc}\, g$ equals 1.

It is not difficult to show that augmenting Algorithm 6 by autoreductions of the set G (i. e. every element of G is reduced with respect to all other elements) leads to an algorithm that always returns a reduced Gröbner basis. With a little bit more effort, one obtains in addition the following uniqueness result which allows for effectively deciding whether two ideals are equal.

Proposition 7.13 ([1, Thm. 1.8.7]). *Every ideal $I \lhd \mathscr{P}$ possesses for any term order $\prec$ a unique reduced Gröbner basis.*

Although there exist infinitely many different term orders, one can show that any given ideal I has only finitely many different reduced Gröbner bases [33, Lemma 2.6].

All the presented material on Gröbner bases is readily translated to left submodules $\mathscr{M} \subseteq \mathscr{P}^m$ of a free polynomial module using the module term orders introduced in Definition 7.1 and all results remain true in this more general situation. The only slight difference concerns the definition of the S-polynomial. In the case of two elements $\mathbf{f}, \mathbf{g} \in \mathscr{P}^m$, their S-"polynomial" (which now is of course also a vector in $\mathscr{P}^m$) is set to zero, if $\mathrm{lt}\, \mathbf{f}$ and $\mathrm{lt}\, \mathbf{g}$ live in different components, as then our construction of the S-"polynomial" (24) makes no sense (recall that in a free module terms are only divisible, if they are in the same component, and thus we can speak of a least common multiple only in this case).

Remark 7.14. The Buchberger algorithm may be considered as a simultaneous generalisation of the Gauß algorithm for linear systems of equations and of the Euclidean algorithm for determining the greatest common divisor of two univariate polynomials. One can easily verify that the S-polynomial of two polynomials with relatively prime leading terms always reduces to zero. Hence, in the case of linear polynomials it suffices to consider pairs of polynomials with the same leading

term (variable) for which the construction of the S-polynomial amounts to a simple Gaußian elimination step.

In the case of two univariate polynomials, the construction of their S-polynomial and its subsequent reduction with respect to the two polynomials is equivalent to the familiar polynomial division. Hence computing the Gröbner basis of a set F amounts simply to determine the greatest common divisor of the elements of F and any reduced Gröbner basis consists of a single polynomial (this observation may be considered as an alternative proof that univariate polynomials define a principal ideal domain). By the same reasoning, we conclude that a reduced Gröbner basis of a submodule of a free module $\mathscr{P}^m$ over a univariate polynomial ring $\mathscr{P}$ may have at most m elements, the leading terms of which are all in different components.

The terminology "S-polynomial" is actually an abbreviation of "syzygy polynomial." Recall that a *syzygy* of a finite set $F = \{\mathbf{f}_1, \ldots, \mathbf{f}_r\} \subset \mathscr{P}^m$ is a vector $\mathbf{S} \in \mathscr{P}^r$ with components $S_i \in \mathscr{P}$ such that

$$S_1 \star \mathbf{f}_1 + \cdots + S_r \star \mathbf{f}_r = 0\,. \tag{25}$$

All syzygies of F together form again a left submodule $\mathrm{Syz}(F) \subseteq \mathscr{P}^r$. Note that this submodule may be understood as the solution set of a linear system of equations over the ring $\mathscr{P}$ or—in a more abstract terminology—as the kernel of a linear map. Thus the effective determination of syzygy modules represents a natural and important problem, if one wants to do linear algebra over a ring.

The Schreyer Theorem shows that, by retaining information that is automatically computed during the determination of a Gröbner basis G with the Buchberger Algorithm 6, one obtains for free a Gröbner basis of the syzygy module $\mathrm{Syz}(G)$. More precisely, assume that $G = \{\mathbf{g}_1, \ldots, \mathbf{g}_r\} \subset \mathscr{P}^m$ is a Gröbner basis and let $\mathbf{g}_i, \mathbf{g}_j \in G$ be two generators with leading terms in the same component. According to (24), their S-"polynomial" can be written in the form $S(\mathbf{g}_i, \mathbf{g}_j) = m_i \star \mathbf{g}_i - m_j \star \mathbf{g}_j$ for suitable monomials m_i, m_j, and Theorem 7.8 implies the existence of coefficients $h_{ijk} \in \mathscr{P}$ such that $\sum_{k=1}^{r} h_{ijk} \star \mathbf{g}_k$ is a standard representation of $S(\mathbf{g}_i, \mathbf{g}_j)$. Combining these two representations, we obtain a syzygy

$$\mathbf{S}_{ij} = m_i \mathbf{e}_i - m_j \mathbf{e}_j - \sum_{k=1}^{r} h_{ijk} \mathbf{e}_k \tag{26}$$

where $\mathbf{e}_k$ denote the vectors of the standard basis of $\mathscr{P}^r$. Recalling the module term order introduced at the end of Example 7.2, we obtain now the following fundamental result on the syzygy module of a Gröbner basis.

Theorem 7.15 (Schreyer [9, Thm. 3.3]). *Let $G \subset \mathscr{P}^m$ be a Gröbner basis for the term order $\prec$ of the submodule generated by it. Then the set of all the syzygies $\mathbf{S}_{ij}$ defined by (26) is a Gröbner basis of* $\mathrm{Syz}(G) \subseteq \mathscr{P}^r$ *for the induced term order $\prec_G$.*

For a general finite set $F \subset \mathscr{P}^m$, one can determine $\mathrm{Syz}(F)$ by first computing a Gröbner basis G of $\langle F \rangle$, then using Theorem 7.15 to obtain a generating set $\mathscr{S}$ of

$\mathrm{Syz}(G)$ and finally transforming $\mathscr{S}$ into a generating set of $\mathrm{Syz}(F)$ essentially by linear algebra. Details can be found e. g. in [1].

By iterating Schreyer's Theorem 7.15, one obtains a *free resolution* of the submodule $\langle G\rangle \subseteq \mathscr{P}^m$ (although this is not necessarily the most efficient way to do this), i. e. an exact sequence

$$0 \longrightarrow \mathscr{P}^{r_n} \longrightarrow \cdots \longrightarrow \mathscr{P}^{r_1} \longrightarrow \mathscr{P}^{r_0} \longrightarrow \langle G\rangle \longrightarrow 0 \tag{27}$$

(Hilbert's Syzygy Theorem guarantees that the length of the resolution is at most the number n of variables in the polynomial ring $\mathscr{P}$). The *minimal free resolution* which can be constructed from any free resolution via some linear algebra gives access to many important invariants of the submodule $\langle G\rangle$ like Betti numbers. However, it is beyond the scope of this article to discuss this application of Gröbner bases.

References

[1] Adams, W., Loustaunau, P.: An Introduction to Gröbner Bases. Graduate Studies in Mathematics 3. American Mathematical Society, Providence (1994)

[2] Brenan, K., Campbell, S., Petzold, L.: Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations. Classics in Applied Mathematics 14. SIAM, Philadelphia (1996)

[3] Bronstein, M., Petkovšek, M.: An introduction to pseudo-linear algebra. Theor. Comp. Sci. **157**, 3–33 (1996)

[4] Buchberger, B.: Ein Algorithmus zum Auffinden der Basiselemente des Restklassenringes nach einem nulldimensionalen Polynomideal. Ph.D. thesis, Universität Innsbruck (1965). (Engl. translation: J. Symb. Comput. 41 (2006) 475–511)

[5] Campbell, S., Gear, C.: The index of general nonlinear DAEs. Numer. Math. **72**, 173–196 (1995)

[6] Chyzak, F., Quadrat, A., Robertz, D.: Effective algorithms for parametrizing linear control systems over Ore algebras. Appl. Algebra Eng. Commun. Comput. **16**, 319–376 (2005)

[7] Cohn, P.: Free Rings and Their Relations. Academic Press (1971)

[8] Cox, D., Little, J., O'Shea, D.: Ideals, Varieties, and Algorithms. Undergraduate Texts in Mathematics. Springer-Verlag, New York (1992)

[9] Cox, D., Little, J., O'Shea, D.: Using Algebraic Geometry. Graduate Texts in Mathematics 185. Springer-Verlag, New York (1998)

[10] Drach, J.: Sur les systèmes complètement orthogonaux dans l'espace à n dimensions et sur la réduction des systèmes différentielles les plus généraux. Compt. Rend. Acad. Sci. **125**, 598–601 (1897)

[11] Gerdt, V., Blinkov, Y.: Involutive bases of polynomial ideals. Math. Comp. Simul. **45**, 519–542 (1998)

[12] Gómez-Torrecillas, J.: Basic module theory over non-commutative rings with computational aspects of operator algebras. In: M. Barkatou, T. Cluzeau, G. Regensburger, M. Rosenkranz (eds.) Algebraic and Algorithmic Aspects of Differential and Integral Operators, Lecture Notes in Computer Science 8372, pp. 23–82. Springer-Verlag, Heidelberg (2014)
[13] Goodearl, K., Warfield, R.: An Introduction to Noncommutative Noetherian Rings, 2^{nd} edn. London Mathematical Society Student Texts 61. Cambridge University Press, Cambridge (2004)
[14] Hairer, E., Lubich, C., Roche, M.: The Numerical Solution of Differential-Algebraic Equations by Runge-Kutta Methods. Lecture Notes in Mathematics 1409. Springer-Verlag, Berlin (1989)
[15] Hausdorf, M., Seiler, W.: Perturbation versus differentiation indices. In: V. Ghanza, E. Mayr, E. Vorozhtsov (eds.) Computer Algebra in Scientific Computing — CASC 2001, pp. 323–337. Springer-Verlag, Berlin (2001)
[16] Hausdorf, M., Seiler, W.: An efficient algebraic algorithm for the geometric completion to involution. Appl. Alg. Eng. Comm. Comp. **13**, 163–207 (2002)
[17] Jacobson, N.: The Theory of Rings. Americal Mathematical Society (1943)
[18] Janet, M.: Leçons sur les Systèmes d'Équations aux Dérivées Partielles. Cahiers Scientifiques, Fascicule IV. Gauthier-Villars, Paris (1929)
[19] Kalman, R.: Algebraic structure of linear dynamical systems. Proc. Nat. Acad. Sci. USA **54**, 1503–1508 (1965)
[20] Kandry-Rody, A., Weispfenning, V.: Non-commutative Gröbner bases in algebras of solvable type. J. Symb. Comp. **9**, 1–26 (1990)
[21] Kashiwara, M., Kawai, T., Kimura, T.: Foundations of Algebraic Analysis. Princeton University Press, Princeton (1986)
[22] Kato, G., Struppa, D.: Fundamentals of Algebraic Microlocal Analysis. Pure and Applied Mathematics 217. Marcel Dekker, New York (1999)
[23] Kredel, H.: Solvable Polynomial Rings. Verlag Shaker, Aachen (1993)
[24] Kunkel, P., Mehrmann, V.: Differential-Algebraic Equations: Analysis and Numerical Solution. EMS Textbooks in Mathematics. EMS Publishing House, Zürich (2006)
[25] Lam, T.: Lectures on Modules and Rings. Graduate Texts in Mathematics 189. Springer, New York (1999)
[26] Lam, T.: On the equality of row rank and column rank. Expo. Math. **18**, 161–163 (2000)
[27] Lamour, R., März, R., Tischendorf, C.: Differential-Algebraic Equations: A Projector Based Analysis. Differential-Algebraic Equations Forum. Springer-Verlag, Berlin/Heidelberg (2013)
[28] Lemaire, F.: An orderly linear PDE with analytic initial conditions with a non-analytic solution. J. Symb. Comp. **35**, 487–498 (2003)
[29] Levandovskyy, V.: Non-commutative computer algebra for polynomial algebras: Gröbner bases, applications and implementation. Ph.D. thesis, Fachbereich Mathematik, Universität Kaiserslautern (2005)

[30] Levandovskyy, V., Schindelar, K.: Computing diagonal form and Jacobson normal form of a matrix using Gröbner bases. J. Symb. Comput. **46**, 595–608 (2011)
[31] Li, P., Liu, M., Oberst, U.: Linear recurring arrays, linear systems and multidimensional cyclic codes over quasi-Frobenius rings. Acta Appl. Math. **80**, 175–198 (2004)
[32] Malgrange, B.: Systemes différentiels à coefficients constants. Semin. Bourbaki **15**, 1–11 (1964)
[33] Mora, T., Robbiano, L.: The Gröbner fan of an ideal. J. Symb. Comp. **6**, 183–208 (1988)
[34] Morimoto, M.: An Introduction to Sato's Hyperfunctions. Transl. Math. Monogr. 129. American Mathematical Society, Providence (1993)
[35] Noether, E., Schmeidler, W.: Moduln in nichtkommutativen Bereichen, insbesondere aus Differential- und Differenzausdrücken. Math. Zeit. **8**, 1–35 (1920)
[36] Oberst, U.: Multidimensional constant linear systems. Acta Appl. Math. **20**, 1–175 (1990)
[37] Oberst, U., Pauer, F.: The constructive solution of linear systems of partial difference and differential equations with constant coefficients. Multidim. Syst. Signal Proc. **12**, 253–308 (2001)
[38] Ore, O.: Theory of non-commutative polynomials. Ann. Math. **34**, 480–508 (1933)
[39] Pazy, A.: Semigroups of Linear Operators and Applications to Partial Differential Equations. Applied Mathematical Sciences 44. Springer-Verlag, New York (1983)
[40] Pillai, H., Shankar, S.: A behavioral approach to control of distributed systems. SIAM J. Control Optim. **37**, 388–408 (1999)
[41] Pommaret, J., Quadrat, A.: Generalized Bezout identity. Appl. Algebra Eng. Commun. Comput. **9**, 91–116 (1998)
[42] Pommaret, J., Quadrat, A.: Algebraic analysis of linear multidimensional control systems. IMA J. Math. Control Inf. **16**, 275–297 (1999)
[43] Pommaret, J., Quadrat, A.: Localization and parametrization of linear multidimensional control systems. Syst. Control Lett. **37**, 247–260 (1999)
[44] Rabier, P., Rheinboldt, W.: Theoretical and numerical analysis of differential-algebraic equations. In: P. Ciarlet, J. Lions (eds.) Handbook of Numerical Analysis, vol. VIII, pp. 183–540. North-Holland, Amsterdam (2002)
[45] Riquier, C.: Les Systèmes d'Équations aux Derivées Partielles. Gauthier-Villars, Paris (1910)
[46] Robertz, D.: Recent progress in an algebraic analysis approach to linear systems. Multidimensional Syst. Signal Process. (2014). To appear.
[47] Rocha, P., Zerz, E.: Strong controllability and extendibility of discrete multidimensional behaviors. Syst. Control Lett. **54**, 375–380 (2005)
[48] Seiler, W.: On the arbitrariness of the general solution of an involutive partial differential equation. J. Math. Phys. **35**, 486–498 (1994)

[49] Seiler, W.: Indices and solvability for general systems of differential equations. In: V. Ghanza, E. Mayr, E. Vorozhtsov (eds.) Computer Algebra in Scientific Computing — CASC '99, pp. 365–385. Springer-Verlag, Berlin (1999)
[50] Seiler, W.: A combinatorial approach to involution and δ-regularity I: Involutive bases in polynomial algebras of solvable type. Appl. Alg. Eng. Comm. Comp. **20**, 207–259 (2009)
[51] Seiler, W.: A combinatorial approach to involution and δ-regularity II: Structure analysis of polynomial modules with Pommaret bases. Appl. Alg. Eng. Comm. Comp. **20**, 261–338 (2009)
[52] Seiler, W.: Involution — The Formal Theory of Differential Equations and its Applications in Computer Algebra. Algorithms and Computation in Mathematics 24. Springer-Verlag, Berlin (2009)
[53] Willems, J.: Paradigms and puzzles in the theory of dynamical systems. IEEE Trans. Autom. Control **36**, 259–294 (1991)
[54] Wood, J., Rogers, E., Owens, D.: A formal theory of matrix primeness. Math. Control Signals Syst. **11**, 40–78 (1998)
[55] Zerz, E.: Extension modules in behavioral linear systems theory. Multidimensional Syst. Signal Process. **12**, 309–327 (2001)
[56] Zerz, E.: Multidimensional behaviours: an algebraic approach to control theory for PDE. Int. J. Control **77**, 812–820 (2004)
[57] Zerz, E.: An algebraic analysis approach to linear time-varying systems. IMA J. Math. Control Information **23**, 113–126 (2006)
[58] Zerz, E.: Discrete multidimensional systems over $\mathbb{Z}_n$. Syst. Control Lett. **56**, 702–708 (2007)
[59] Zerz, E., Rocha, P.: Controllability and extendibility of continuous multidimensional behaviors. Multidimensional Syst. Signal Process. **17**, 97–106 (2006)
[60] Zerz, E., Seiler, W., Hausdorf, M.: On the inverse syzygy problem. Comm. Alg. **38**, 2037–2047 (2010)

Index