Werner M. Seiler

Involution

The formal theory of differential equations
and its applications in computer algebra
and numerical analysis

involution: anything internally complex or intricate

Oxford Advanced Learner's Dictionary of Current English

inuoluere, ML *involvere*, to envelop, roll up, wrap up, whence 'to *involve*'; pp *inuolutus*, ML *involutus*, yields both the tech adj, Geom n, *involute* and the L *inuolutio*, ML o/s *involution* whence E *involution*

Eric Partridge: *Origins: an etymological dictionary of modern English*

Preface

> *Mathematics is being lazy. Mathematics is letting the principles do the work for you so that you do not have to do the work for yourself.*
>
> *George Pólya*

The theory of differential equations is one of the largest fields within mathematics and probably any graduate in mathematics has attended at least one course on differential equations. But differential equations are also of fundamental importance in most applied sciences; whenever a continuous process is modelled mathematically, chances are high that differential equations appear. So it does not surprise that many textbooks exist on both ordinary and partial differential equations. But the huge majority of these books makes an implicit assumption on the structure of the equations: either one deals with scalar equations or with *normal* systems, i.e. with systems in Cauchy–Kovalevskaya form. The main topic of the present work is what happens, if this assumption is dropped.

This is not just an academic exercise; non-normal systems are ubiquitous in applications. Classical examples include the incompressible Navier–Stokes equations of fluid dynamics, Maxwell's equations of electrodynamics, the Yang–Mills equations of the fundamental gauge theories in modern particle physics or Einstein's equations of general relativity. But also the simulation and control of multibody systems, electrical circuits or chemical reactions lead to non-normal systems of ordinary differential equations, often called differential algebraic equations. In fact, most of the differential equations encountered by engineers and scientists are probably not normal.

In view of this great importance of non-normal systems, the lack of literature on their general theory is all the more surprising. Of course, much has been written on specific systems like the Navier–Stokes equations but the existence of general approaches to non-normal systems seems to be hardly known, although some of them were developed about a century ago!

We will study the so-called *formal theory of differential equations*. It merges ideas from the Janet–Riquier Theory of partial differential equations and the Cartan–Kähler Theory of exterior differential systems. This combines geometric and algebraic structures into a powerful framework useful not only for theoretical analysis but also for concrete applications. Compared with other books on the formal theory we put a stronger emphasis on computations. This will become most evident in the last chapter where concrete algorithms for the completion to involution are discussed. But also in the other chapters we try to be as constructive as possible.

The adjective "formal" has at least two different meanings. A simple one is that the origin of the theory was the problem of determining formal power series solutions for general systems of differential equations, i. e. of proving the formal integrability. In a broader sense, the word "formal" refers to the fact that we nowhere use solutions; all analysis is based solely on manipulations of the equations themselves.

As the title clearly indicates, our topic will be the notion of involution. Roughly speaking, the message is that whatever one intends to do with a differential equation, one should first render it involutive: *involution is the central principle in the theory of under- or overdetermined systems.*

Unfortunately, the term "involutive" is used in different approaches with somewhat different meanings. In fact, even within this book we will use it sometimes in a broader and sometimes in a narrower meaning. In the algebraic theory we introduce a fairly general notion of involutive bases. But within the geometric theory only the special case of Pommaret bases appears. In the case of differential equations, involution is often confused with formal integrability, i. e. with the mere absence of integrability conditions. But involution requires more and we will stress at several occasions the great importance of this "more".

The book contains six chapters and three appendices. The first chapter is concerned with the geometry behind the formal theory. It introduces the jet bundle formalism as the geometric foundation of differential equations. In fact, it contains two different introductions of jet bundles. The "pedestrian" one is based on the usual Taylor series approach and puts more emphasis on local coordinates. The "intrinsic" one uses an abstract construction stressing the affine structure of jet bundles. It does not often appear in the literature but turns out to be rather natural for the formal theory.

Then differential equations are intrinsically defined as fibred submanifolds of jet bundles and the basic geometric operations with them, prolongation and projection, are introduced. This leads to an intrinsic picture of integrability conditions and the notion of formal integrability. Finally, the geometric symbol of a differential equation is defined and related to the principal symbol well-known from standard partial differential equation theory.

While formal integrability is a very natural and intuitive concept, for many purposes it does not suffice. The first problem is already that no constructive criterion for formal integrability is known so far. But we will also see at several places that it has further shortcomings; for example even in the analytic category an existence and uniqueness theorem can be proven only for involutive equations.

So the second chapter introduces our main theme: involution. But it does so in a purely algebraic framework which, at first sight, is not at all related to differential equations. Involution is introduced for the Abelian monoid $\mathbb{N}_0^n$ and then extended to a fairly general class of algebras: the polynomial algebras of solvable type. This rather abstract formalism has the advantage that it does not only allow us to extract the simple idea underlying involution in a very clear form, it may also be applied without modifications in many different situations: polynomial ideals, rings of differential operators, universal enveloping algebras of Lie algebras, ...

We show that the fundamental idea behind the theory is the determination of combinatorial decompositions of a polynomial module. Involutive bases unite these decompositions with Gröbner bases, the central algorithmic tool in commutative algebra. Although fairly obvious, this point of view seems to be new. For us, the most important type of involutive bases are Pommaret bases, as these will later reappear in the theory of differential equations. They define a special kind of decomposition, namely Rees decompositions, existing only in generic coordinates. We present a new and very cheap criterion for δ-singularity solving the termination problem in completion algorithms.

The literature on involutive bases is almost exclusively concerned with their efficient construction. They are considered mainly because they offer an interesting alternative to Buchberger's algorithm for the determination of Gröbner bases. But the combinatorial decompositions induced by them contain much information about the structure of the ideal they generate. A trivial application is the determination of the Hilbert function. We will show that in addition one can straightforwardly read off a Pommaret basis the depth, the projective dimension and the Castelnuovo–Mumford regularity of their ideal. They also provide a simple realisation of Hironaka's criterion for Cohen-Macaulay modules. Most of these results follow from an involutive version of Schreyer's syzygy theorem presented here for the first time.

In the third chapter, the relation of the developed algebraic theory to differential equations is exhibited by showing that in local coordinates the (geometric) symbol defines at each point of the differential equation a polynomial module. Thus we finally arrive at the definition of an involutive differential equation where the symbol equations yield a Pommaret basis of this module. An important structural property of such an equation is the existence of the Cartan normal form for local representations.

As a first simple application of involution, we rigorously define with the help of the principal symbol under- and overdetermined equations. Such a definition makes sense only for involutive equations and it turns out that the classical counting rules comparing the number of equations and unknowns may be misleading. Rather surprisingly, we have not been able to find a completely sound definition in the literature. Finally, we discuss the differential version of the syzygy resolution studied in the last chapter, the Janet sequence. In particular, we solve the inverse problem which is of considerable importance in control theory.

The fourth chapter contains some "combinatorial games": abstract measures for the size of the formal solution space like the Hilbert polynomial or the Cartan characters. One may argue about their usefulness, but some classical notions like the number of degrees of freedom of a physical system are in fact based on such "games". We introduce the concept of a differential relation between two differential equations generalising Bäcklund transformations and study the induced relation between the respective Hilbert polynomials. For applications in physics, it is often of interest to formally subtract the effect of gauge symmetries. Within the framework of the formal theory, we present a new and fairly simple pseudogroup approach. Einstein introduced for similar purposes the strength of a differential equation. We relate it to the formal theory and show that it contains less information than the full set of Cartan characters.

As an important application of involution we prove in the fifth chapter some existence and uniqueness results. The jet bundle formalism may be understood as a geometric version of Taylor series, so it is not surprising that the simplest results are obtained for analytic solutions. We recall the Cauchy–Kovalevskaya Theorem, as the most general (with respect to the structure of the equation) existence and uniqueness result for analytic solutions of analytic normal systems.

One could try to directly extend its proof to involutive systems. Indeed, this approach is taken in the Janet–Riquier Theory leading to Riquier's theorem. We will use another approach leading to the Cartan–Kähler Theorem. Our proof is based on considering a sequence of normal systems to which the Cauchy–Kovalevskaya Theorem is applied; thus power series appear only indirectly in contrast to the proof of Riquier's theorem. It clearly demonstrates why formal integrability is not sufficient for the analysis of differential equations but the crucial tool of the Cartan normal form of an involutive equation is necessary.

Because of the use of the Cauchy–Kovalevskaya Theorem, the Cartan–Kähler Theorem inherits the restriction to the analytic category and this seriously limits its practical value. However, our proof provides us with a powerful technique for obtaining stronger results, if more is known about the structure of the equation. In such situations it is sometimes possible to substitute the Cauchy–Kovalevskaya Theorem by other results from the extensive literature on existence and uniqueness for normal systems. As a first example we generalise the Holmgren Theorem to involutive systems; despite its triviality this generalisation seems to be new. We also analyse the class of what we call weakly overdetermined systems; this class includes in particular Maxwell's equations. Here the above mentioned technique is used in order to obtain results on smooth solutions of hyperbolic systems with elliptic constraints.

The last chapter is mainly concerned with computational aspects. It studies the problem of completing an arbitrary differential equation to an involutive one and provides concrete algorithms that can be (and are) implemented in a computer algebra system. For ordinary differential equations, completion is a rather simple operation and the basic algorithm has been rediscovered many times in different fields. For partial differential equations, the Cartan–Kuranishi Theorem provides a very general geometric completion strategy which, however, has yet to be turned into an algorithm.

On the algebraic side we present a simple but nevertheless surprisingly efficient algorithm for the construction of involutive bases (some explanations for its efficiency follow from our structure theory of polynomial modules). For linear systems of partial differential equations, it is possible to combine it with the Cartan–Kuranishi strategy. The result is a new algebraic algorithm that is able to deliver intrinsic geometric results. Compared with earlier realisations of the Cartan–Kuranishi completion it is much faster, as due to a clever book-keeping much less prolongations must be computed.

Finally, as a completely new application of involution, we turn to the numerical treatment of non-normal differential equations. As already mentioned, in the case of ordinary differential equations one often speaks of differential algebraic equations. In their numerical analysis the notion of an index plays a fundamental role; roughly speaking an index measures how far away a differential algebraic equation is from a normal ordinary differential equation. This can be translated into how much trouble one has to expect in a numerical integration.

The literature on differential algebraic equations abounds with index definitions. We show that the formal theory does not only help us to get a better understanding of the meaning of the various index concepts, it even leads naturally to formulations that apply equally well to partial differential equations. In fact, we obtain the first index concepts for partial differential equations that do not require integral transforms, semi-discretisations or any other technique of reduction to a differential algebraic equation.

For the special case of linear systems with constant coefficients we furthermore study semi-discretisations transforming overdetermined partial differential equations into differential algebraic equations. Probably one can nowhere see the importance of involution better than in this application. One might expect that if the original partial differential equation does not possess any hidden integrability conditions, then the same should be true for the arising ordinary differential equations. However, this is not the case: we must require that the partial differential equation is involutive! Thus these abstract combinatorial decompositions of polynomial modules acquire suddenly a very concrete meaning as in semi-discretisations obstructions to involution turn into integrability conditions.

As one can see, we cover rather diverse topics. Somebody reading the whole text needs a certain familiarity with different branches of mathematics: jet bundles and differential geometry, commutative and computer algebra, and finally some numerical analysis. Not all readers will have this background or actually be interested in all these topics. This has been taken into account in a two-fold way.

Three appendices collect some basic results from analysis, algebra and differential geometry that are used in the main text. They try to make the text to a reasonable degree self-contained and fix some notations. But of course they cannot substitute a real introduction into the subjects involved. For this reason, references to standard textbooks are given.

Many chapters and sections are rather independent of each other. Depending on the interests of a reader many parts may be safely ignored. For example, somebody only interested in the algebraic theory of involutive bases needs only Chapters 2 and 6. On the other hand, somebody who does not like commutative algebra may ignore these two chapters. For the benefit of readers not so familiar with differential geometry most geometric constructions are described both in an intrinsic manner and in local coordinates. Furthermore, many concrete examples are included to illustrate the theoretical results.

Each chapter ends with a brief section called *Notes*. It usually gives some pointers to the literature and to alternative approaches. I have also tried to trace a bit the history of the main ideas. But as I am not a historian of science, the results should be more taken as rough indications than as a definite answer. Furthermore, in several places I rely on other sources and I have not always checked the correctness of the references. A valuable source of information on the state of the art of the theory of differential equations at the end of the 19th century is the article of Weber [192] in the *Enzyklopädie der Mathematischen Wissenschaften.*

I am indebted to Ernst Binz and the members of his group at Universität Mannheim for many stimulating discussions. This holds in particular for Marcus Hausdorf, my collaborator on most of the topics covered here. He has implemented large parts of the *MuPAD* code used in concrete computations. I would also like to acknowledge the continuing hospitality of Jacques Calmet at Universität Karlsruhe and his support over many years.

Most of my work in the last years has been financially supported by Deutsche Forschungsgemeinschaft under various grants. Some additional financial support was obtained by INTAS (grant 99-01222).

But my deepest gratitude goes to my wife Marion, patiently bearing the live with a scientist. Only her love and support has made it all possible.

Karlsruhe/Mannheim, November 2001 *Werner M. Seiler*

Contents

1. Formal Geometry of Differential Equations

He who can digest a second or third flux-ion, a second or third difference, need not, we think, be squeamish about any point of divinity.

George Berkeley

In this chapter we lay the geometric foundations of the so-called *formal theory of differential equations*. The name "formal theory" stems from the fact that, at least indirectly, much of it is based on the analysis of formal power series solutions, i.e. one ignores the question whether the series actually converge. Another interpretation of the name is that one tries to extract as much information as possible on the solution space by purely formal operations like algebraic manipulations of the equations or their differentiations without actually solving the given equation.

The basic tool for a geometric approach to differential equations is the jet bundle formalism. The first two sections of this chapter give an introduction to it; first in a more pedestrian way considering jets as a geometric approach to power series, then in a more abstract, intrinsic way stressing certain structural properties which will later be of importance.

The third section starts with the geometric definition of differential equations as fibred submanifolds of jet bundles and introduces the two basic operations of prolongation to higher order and projection to lower order. This leads almost immediately to the notion of formal integrability, the topic of the next section. From a more computational point of view this is related to the problem of constructing order by order formal power series solutions for a given differential equation. For a formally integrable equation such a construction is always possible, as no obstructions in the form of integrability conditions exist.

The fifth section introduces the geometric and the principal symbol of a differential equation and related notions like characteristics. At this point it does not yet become apparent why the symbol is so important, but in later chapters we will see that not only the concept of involution is based on it but that it is also of fundamental importance in computations.

In the final section we present some classical differential equations of mathematical physics. In the subsequent chapters, they will be repeatedly used as examples in order to demonstrate that the theory can be used in realistic applications and not only in toy problems.

1.1 A Pedestrian Approach to Jet Bundles

Let $\mathcal{X}$ and $\mathcal{U}$ be two (finite-dimensional) manifolds, for example two open subsets of $\mathbb{R}^n$ and $\mathbb{R}^m$, respectively. Let furthermore $\phi : \mathcal{X} \to \mathcal{U}$ be a smooth map between these two manifolds. If we take some charts on $\mathcal{X}$ and $\mathcal{U}$ with local coordinates $\mathbf{x} = (x^1, \dots, x^n)$ and $\mathbf{u} = (u^1, \dots, u^m)$, respectively, we can locally write ϕ in the form $u^\alpha = \phi^\alpha(x^1, \dots, x^n)$ for $1 \leq \alpha \leq m$. In the context of differential equations, the coordinates $\mathbf{x}$ are often called *independent variables*, whereas the coordinates $\mathbf{u}$ represent the *dependent variables*.

According to Taylor's Theorem we can expand the components ϕ^α of the map ϕ in the neighbourhood of an arbitrary but fixed point $\mathbf{x}_0 \in \mathcal{X}$

$$\phi^\alpha(\mathbf{x}) = \sum_{0 \leq |\mu| \leq q} \frac{p^\alpha_\mu}{\mu!}(\mathbf{x} - \mathbf{x}_0)^\mu + R^\alpha_q(\mathbf{x}, \mathbf{x}_0) \,. \tag{1.1}$$

Here $q \geq 0$ is an arbitrary integer and we use the standard multi index notation of multivariate analysis (see Appendix A.1). It is well-known that the coefficients $p^\alpha_\mu \in \mathbb{R}$ of the expansion (1.1) are the derivatives of the functions ϕ^α evaluated at $\mathbf{x}_0$

$$p^\alpha_\mu = \frac{\partial^{|\mu|}\phi^\alpha}{\partial x^\mu}(\mathbf{x}_0) \tag{1.2}$$

and that in the Lagrange representation the remainder term takes the form

$$R^\alpha_q(\mathbf{x}, \mathbf{x}_0) = \sum_{|\mu|=q+1} \frac{\partial^{q+1}\phi^\alpha}{\partial x^\mu}\big(\mathbf{x}_0 + \theta(\mathbf{x} - \mathbf{x}_0)\big)\frac{(\mathbf{x} - \mathbf{x}_0)^\mu}{\mu!} \tag{1.3}$$

for some $\theta \in]0, 1[$.

We introduce an equivalence relation on the space of all maps $\mathcal{X} \to \mathcal{U}$. The maps ϕ and ψ are equivalent to order q at $\mathbf{x}_0 \in \mathcal{X}$, written $\phi \sim_q \psi$, if their expansions up to order q at $\mathbf{x}_0$ are identical. Geometrically, this means that the graphs of the two maps have at $\mathbf{x}_0$ a contact of order q. Thus if ϕ is given by (1.1) and ψ by a similar expansion with coefficients $\bar{p}^\alpha_\mu$, then $\phi \sim_q \psi$ implies that $p^\alpha_\mu = \bar{p}^\alpha_\mu$ for $1 \le \alpha \le m$ and $0 \le |\mu| \le q$. A *q-jet* at $\mathbf{x}_0$ is defined as an equivalence class

$$[\phi]^{(q)}_{\mathbf{x}_0} = \left\{ \psi \in C^\infty(\mathcal{X}, \mathcal{U}) \mid \psi \sim_q \phi \text{ at } \mathbf{x}_0 \right\} . \tag{1.4}$$

Loosely speaking, the jet bundle of order q is the space of all power series truncated at order q, i.e. a point in it is such an equivalence class $[\phi]^{(q)}_{\mathbf{x}_0}$. Note that we always consider only a finite part of the series; thus it does not matter whether the series actually converge and we do not have to restrict to analytic functions.

Definition 1.1.1. *The* jet bundle of order q *over the manifolds $\mathcal{X}$ and $\mathcal{U}$ is the space $J_q(\mathcal{X}, \mathcal{U})$ of all q-jets $[\phi]^{(q)}_{\mathbf{x}}$ with $\mathbf{x} \in \mathcal{X}$.*

This definition is very similar to the introduction of the tangent bundle in Appendix C.1. A point in it was also defined as an equivalence class of functions (more precisely, of curves). As for the tangent bundle, it is not difficult to show that $J_q(\mathcal{X}, \mathcal{U})$ is a manifold, too. In fact, it is obvious what one may use as local coordinates: given the coordinates $\mathbf{x}$ on $\mathcal{X}$ and $\mathbf{u}$ on $\mathcal{U}$, we take as coordinates for the q-jet $[\phi]^{(q)}_{\mathbf{x}_0}$ the tuple (x^i_0, p^α_μ) with $0 \le |\mu| \le q$. The jet bundle of order 0 is the product $\mathcal{X} \times \mathcal{U}$, as a 0-jet tells us only the result $\phi(\mathbf{x}_0)$ of evaluating any map ϕ contained in it at the point $\mathbf{x}_0$.

In concrete calculations one often ignores that points in $J_q(\mathcal{X}, \mathcal{U})$ are actually equivalence classes of functions. One simply thinks of $J_q(\mathcal{X}, \mathcal{U})$ as a space whose coordinates are besides the independent variables $\mathbf{x}$ and dependent variables $\mathbf{u}$ all derivatives p^α_μ with $|\mu| \le q$.

$J_q(\mathcal{X}, \mathcal{U})$ is not just any manifold but possesses many internal structures. In particular, jet bundles form a natural hierarchy. If $r > q$, we have an obvious projection map $\pi^r_q : J_r(\mathcal{X}, \mathcal{U}) \to J_q(\mathcal{X}, \mathcal{U})$ defined by "forgetting" the higher-order terms of the power series. We can project even further by forgetting everything but the expansion point; this yields maps $\pi^q : J_q(\mathcal{X}, \mathcal{U}) \to \mathcal{X}$. All these projection maps are surjective submersions, so that $J_q(\mathcal{X}, \mathcal{U})$ may be endowed in many different ways with the structure of a fibred manifold (cf. Appendix C.2).

The dimension of a jet bundle is easily calculated by some elementary combinatorics. As $\dim \mathcal{X} = n$, $\dim \mathcal{U} = m$, we get $\dim J_0(\mathcal{X}, \mathcal{U}) = n + m$. If we go from $J_{q-1}(\mathcal{X}, \mathcal{U})$ to $J_q(\mathcal{X}, \mathcal{U})$, we add one new coordinate for each derivative of order q of m functions of n variables. Thus, by (A.4a),

$$\dim J_q(\mathcal{X}, \mathcal{U}) - \dim J_{q-1}(\mathcal{X}, \mathcal{U}) = m \binom{n+q-1}{q} , \tag{1.5}$$

which, by (A.4c), implies that

$$\dim J_q(\mathcal{X},\mathcal{U}) = n + m\binom{n+q}{q}\,. \tag{1.6}$$

In practice, one considers usually only the fibre dimension, i. e. the number of variables p^α_μ with $0 \le |\mu| \le q$, and omits the summand n stemming from the base space $\mathcal{X}$.

Example 1.1.2. Assume that $\mathcal{X}$ is two-dimensional with coordinates (x^1, x^2) and $\mathcal{U}$ one-dimensional with coordinate u^1. The jet bundle $J_0(\mathcal{X},\mathcal{U}) = \mathcal{X}\times\mathcal{U}$ is then three-dimensional with coordinates $(x^1, x^2; u^1)$. The first-order jet bundle $J_1(\mathcal{X},\mathcal{U})$ is five-dimensional with coordinates $(x^1, x^2; u^1; p^1_{[1,0]}, p^1_{[0,1]})$ and for the eight-dimensional second-order jet bundle $J_2(\mathcal{X},\mathcal{U})$ we use the coordinates $(x^1, x^2; u^1; p^1_{[1,0]}, p^1_{[0,1]}, p^1_{[2,0]}, p^1_{[1,1]}, p^1_{[0,2]})$ and so on. The fibre dimension is always two less.

Obviously, the multi index notation is rather cumbersome in such concrete examples. For this reason, we restrict its use to theoretical parts and use more natural variable names in examples. Thus we may take variables (x, y) for $\mathcal{X}$ and u for $\mathcal{U}$. The derivatives in $J_2(\mathcal{X},\mathcal{U})$ are then denoted by $(x, y; u; u_x, u_y, u_{xx}, u_{xy}, u_{yy})$ which brings us back to the standard notations in differential equations. ◁

Any smooth map $\phi : \mathcal{X} \to \mathcal{U}$ can be *prolonged* to a smooth map

$$j_q\phi : \begin{cases} \mathcal{X} \longrightarrow J_q(\mathcal{X},\mathcal{U}) \\ \mathbf{x} \longmapsto [\phi]^{(q)}_{\mathbf{x}} = \left(x^i, \phi^\alpha(\mathbf{x}), \frac{\partial^{|\mu|}\phi^\alpha}{\partial x^\mu}(\mathbf{x})\right) \end{cases}. \tag{1.7}$$

Of course, not all maps $\mathcal{X} \to J_q(\mathcal{X},\mathcal{U})$ are of this special form, but these prolonged maps are particularly important. In local coordinates, a general map $\mathcal{X} \to J_q(\mathcal{X},\mathcal{U})$ has the form $\mathbf{x} \mapsto \bigl(x^i, \phi^\alpha(\mathbf{x}), \phi^\alpha_\mu(\mathbf{x})\bigr)$ without any relationship between the functions ϕ^α_μ and ϕ^α.

Those maps that come from a prolongation can be characterised with the help of the *contact structure* of the jet bundle $J_q(\mathcal{X},\mathcal{U})$. The name refers to the fact that the equivalence relation underlying our definition of the jet bundle identifies maps whose graphs have a contact of order q. The structure is often written in terms of differential one-forms on $J_q(\mathcal{X},\mathcal{U})$. Given a one-form $\omega \in \Omega^1(J_q(\mathcal{X},\mathcal{U}))$, we can compute for an arbitrary n-dimensional fibred submanifold $\mathcal{N} \subset J_q(\mathcal{X},\mathcal{U})$, i. e. satisfying $\pi^q\mathcal{N} = \mathcal{X}$, the pull-back $\iota^*\omega$ where $\iota : \mathcal{N} \to J_q(\mathcal{X},\mathcal{U})$ is the natural inclusion map. A one-form on $J_q(\mathcal{X},\mathcal{U})$ is called a *contact form*, if this pull-back vanishes, if and only if $\mathcal{N}$ is the image of a prolonged map $j_q\phi$. The contact forms span the *contact codistribution* $\mathcal{C}^0_q \subset T^*J_q\mathcal{E}$.

Locally, $\mathcal{C}_q^0$ is generated by the following one-forms[1]

$$\omega_\mu^\alpha = dp_\mu^\alpha - p_{\mu+1_i}^\alpha dx^i \tag{1.8}$$

with $1 \leq \alpha \leq m$ and $0 \leq |\mu| < q$. Thus $\dim \mathcal{C}_q^0 = \dim J_{q-1}\mathcal{E}$ (the fibre dimension!). It is easy to show that the forms ω_μ^α are indeed contact forms and generate $\mathcal{C}_q^0$ [86]. Because of the transversality condition $\pi^q\mathcal{N} = \mathcal{X}$, we may use the coordinates $\mathbf{x}$ as parameters for $\mathcal{N}$ and describe it in the form $(\mathbf{x}, \mathbf{p}) = (\mathbf{x}, \psi(\mathbf{x}))$ with some function ψ. Hence pulling-back the contact forms by the natural inclusion map $\iota : \mathcal{N} \hookrightarrow J_q(\mathcal{X}, \mathcal{U})$ yields

$$(\iota^*\omega_\mu^\alpha)(\mathbf{x}) = \left(\frac{\partial \psi_\mu^\alpha}{\partial x^i}(\mathbf{x}) - \psi_{\mu+1_i}^\alpha(\mathbf{x})\right) dx^i \,. \tag{1.9}$$

The pulled-back form vanishes, if and only if the components of ψ satisfy the relations $\psi_\mu^\alpha = \partial^{|\mu|}\psi_0^\alpha/\partial x^\mu$. But this is only the case, if $\psi = j_q\psi_0$.

An alternative description of the contact structure uses vector fields instead of differential forms: the *contact distribution* $\mathcal{C}_q \subset TJ_q\mathcal{E}$ consists of all vectors that are annihilated by the contact one-forms. It follows from (1.8) that $\dim \mathcal{C}_q = \dim J_q\mathcal{E} - \dim J_{q-1}\mathcal{E} + \dim \mathcal{X}$. $\mathcal{C}_q$ is locally generated by the following two sets of vector fields:

$$C_i = \partial_{x^i} + p_{\mu+1_i}^\alpha \partial_{p_\mu^\alpha} \,, \qquad 1 \leq i \leq n \tag{1.10a}$$

(summation over all $1 \leq \alpha \leq m$ and all multi indices μ with $0 \leq |\mu| < q$ in the second term) and

$$C_\alpha^\mu = \partial_{p_\mu^\alpha} \,, \qquad 1 \leq \alpha \leq m \,, \quad |\mu| = q \,. \tag{1.10b}$$

Now we may characterise those fibred submanifolds $\mathcal{N} \subset J_q\mathcal{E}$ that are the image of a prolonged map $j_q\phi$ by the fact that their tangent spaces are subspaces of $\mathcal{C}_q$.

For another kind of prolongation we need the *formal* or *total derivatives*. They take as argument a function $\Phi : J_q(\mathcal{X}, \mathcal{U}) \to \mathbb{R}$ and return a function $J_{q+1}(\mathcal{X}, \mathcal{U}) \to \mathbb{R}$. In local coordinates, the formal derivative D_i with respect to the independent variable x^i is defined as

$$D_i\Phi(\mathbf{x}, \mathbf{p}) = \frac{\partial \Phi}{\partial x^i}(\mathbf{x}, \mathbf{p}) + \frac{\partial \Phi}{\partial p_\mu^\alpha}(\mathbf{x}, \mathbf{p})\, p_{\mu+1_i}^\alpha \,. \tag{1.11}$$

[1] Here we use on the right hand side the *summation convention* that a summation is understood over repeated indices, i.e. in this case over i. The range of the summation should be clear from the context; here it is obviously $1 \leq i \leq n$. In the sequel we will mostly omit the sums in formulae.

Note that the function $D_i\Phi$ is always *quasi-linear*, i.e. linear in the derivatives of maximal order: as Φ depends only on the derivatives p^α_μ with $|\mu| \leq q$, derivatives of order $q+1$ appear in (1.11) only in the terms $p^\alpha_{\mu+1_i}$ with $|\mu| = q$.

Eq. (1.11) represents the well-known *chain rule* of calculus within the jet bundle formalism. Indeed, suppose that we consider the function Φ restricted to the graph of a map $j_q\psi : \mathcal{X} \to J_q(\mathcal{X}, \mathcal{U})$. Then we can substitute the variables p^α_μ by $\partial^{|\mu|}\psi^\alpha/\partial x^\mu$ and obtain a function $\tilde{\Phi} : \mathcal{X} \to \mathcal{U}$. The partial derivative $\partial\tilde{\Phi}/\partial x_i$ must be computed by the chain rule. Exactly the same result is reached, if we perform the substitutions in the function $D_i\Phi$.

Example 1.1.3. Let (x, t) be coordinates on $\mathcal{X}$ and u on $\mathcal{U}$. We consider the function $\Phi(x, t, u, u_x, u_t) = (u_t)^2 - xu_x$ defined on $J_1(\mathcal{X}, \mathcal{U})$. Its formal derivatives with respect to t and x, respectively, are given by the following two functions defined on $J_2(\mathcal{X}, \mathcal{U})$

$$D_x\Phi(x, t, u, u_x, u_t, u_{xx}, u_{xt}, u_{tt}) = 2u_tu_{xt} - xu_{xx} - u_x \,, \tag{1.12a}$$

$$D_t\Phi(x, t, u, u_x, u_t, u_{xx}, u_{xt}, u_{tt}) = 2u_tu_{tt} - xu_{xt} \,. \tag{1.12b}$$

These are indeed quasi-linear functions, i.e. linear in all second-order derivatives, although Φ is quadratic in u_t.

If we substitute u by some function $\psi(x, t)$ in Φ, we obtain a function $\tilde{\Phi}(x, t) = \Phi\big(x, t, \psi(x, t), \psi_x(x, t), \psi_t(x, t)\big)$. Differentiating this function with respect to x yields according to the chain rule

$$\frac{\partial\tilde{\Phi}}{\partial x}(x, t) = 2\psi_t(x, t)\psi_{xt}(x, t) - x\psi_{xx}(x, t) - \psi_x(x, t) \tag{1.13}$$

and correspondingly for $\partial\tilde{\Phi}/\partial t$. The back substitutions $\psi \to u$, $\psi_x \to u_x$, $\psi_{xx} \to u_{xx}$ etc. lead exactly to the above formal derivatives. $\lhd$

We have mentioned above that jet bundles form a natural hierarchy. The projections $\pi^q_{q-1} : J_q(\mathcal{X}, \mathcal{U}) \to J_{q-1}(\mathcal{X}, \mathcal{U})$ have an additional structural property that will become rather important for us in the sequel.

Proposition 1.1.4. *The jet bundle $J_q(\mathcal{X}, \mathcal{U})$ of order q is affine over the jet bundle $J_{q-1}(\mathcal{X}, \mathcal{U})$ of order $q - 1$.*

Before we prove this proposition, we discuss what it means. Recall that a function is affine, if it is of the form $x \mapsto ax + b$ for constants a, b. Proposition 1.1.4 claims that the highest-order derivatives p^α_μ with $|\mu| = q$ transform under a change of coordinates in $\mathcal{X}$ and $\mathcal{U}$ as affine functions with coefficients a, b depending on the remaining coordinates p^α_μ with $|\mu| < q$. A precise definition of an affine bundle is given in Appendix C.2.

Proof. We study the effect of changes of coordinate $\bar{\mathbf{x}} = \bar{\mathbf{x}}(\mathbf{x})$ and $\bar{\mathbf{u}} = \bar{\mathbf{u}}(\mathbf{u})$ in the manifolds $\mathcal{X}$ and $\mathcal{U}$, respectively. They induce a change of coordinates $\bar{\mathbf{p}} = \bar{\mathbf{p}}(\mathbf{x}, \mathbf{p})$ in the jet bundle $J_q(\mathcal{X}, \mathcal{U})$. It can be determined either by using the chain rule or, equivalently, by requiring that the contact forms (1.8) remain invariant. In either case the result is that in repeated index notation

$$\bar{p}^{\alpha}_{j_1 \cdots j_q} = \left(\frac{\partial \bar{u}^{\alpha}}{\partial u^{\beta}} \frac{\partial x^{i_1}}{\partial \bar{x}^{j_1}} \cdots \frac{\partial x^{i_q}}{\partial \bar{x}^{j_q}} \right) p^{\beta}_{i_1 \cdots i_q} + \cdots \tag{1.14}$$

where the dots represent a complicated expression which, however, does not depend on any derivative p^{α}_{μ} with $|\mu| = q$ and where $\partial \mathbf{x}/\partial \bar{\mathbf{x}}$ represents the inverse of the Jacobian $\partial \bar{\mathbf{x}}/\partial \mathbf{x}$. But this implies that (1.14) is indeed affine in the derivatives of order q as claimed. □

1.2 An Intrinsic Approach to Jet Bundles

We describe now a direct intrinsic construction of the jet bundles based on Proposition 1.1.4. This construction does not seem to be widely known in the literature, but it is very convenient for our purposes. Strictly speaking, we do not construct the jet bundles as introduced in the previous section but a bundle canonically diffeomorphic to it. Thus we identify it with the jet bundle. We also slightly generalise: instead of jets of maps $\mathcal{X} \to \mathcal{U}$ we consider jets of sections of a fibred manifold $\pi : \mathcal{E} \to \mathcal{X}$ (see Appendix C.2 for some basic material on fibred manifolds). Note that this affects only global properties; locally there is no real difference. The situation of the last section is recovered, if we take the trivial fibration $\mathrm{pr}_1 : \mathcal{E} = \mathcal{X} \times \mathcal{U} \to \mathcal{X}$ with the canonical projection on the first factor.

Definition 1.2.1. *Let $\pi : \mathcal{E} \to \mathcal{X}$ be a fibred manifold. The* first-order jet bundle *over $\mathcal{E}$ is the affine bundle $\pi^1_0 : J_1\mathcal{E} \to \mathcal{E}$ whose fibre at the point $\xi \in \mathcal{E}$ with $\pi(\xi) = x \in \mathcal{X}$ is given by the affine space*

$$(J_1\mathcal{E})_{\xi} = \mathcal{A}^{(1)}_{\xi} = \left\{ \gamma \in T^*_x\mathcal{X} \otimes T_{\xi}\mathcal{E} \mid T_{\xi}\pi \circ \gamma = \mathrm{id}_{T_x\mathcal{X}} \right\} \tag{1.15}$$

modelled on the vector space

$$\mathcal{V}^{(1)}_{\xi} = T^*_x\mathcal{X} \otimes V_{\xi}\mathcal{E} \, . \tag{1.16}$$

If we take two arbitrary maps $\gamma_1, \gamma_2 \in \mathcal{A}^{(1)}_{\xi}$, then their difference satisfies $T_{\xi}\pi \circ (\gamma_1 - \gamma_2) = \mathrm{id}_{T_x\mathcal{X}} - \mathrm{id}_{T_x\mathcal{X}} = 0$. Hence $\gamma_1 - \gamma_2$ maps any vector in $T_x\mathcal{X}$ into a vertical vector and $\gamma_1 - \gamma_2 \in \mathcal{V}^{(1)}_{\xi}$. So $\mathcal{A}^{(1)}_{\xi}$ is indeed an affine space modelled on the vector space $\mathcal{V}^{(1)}_{\xi}$. Any element of $\mathcal{A}^{(1)}_{\xi}$ has the form $\gamma(v) = \big(v, B(v)\big)$ with some linear map B.

The affine bundle $J_1\mathcal{E}$ is modelled on the vector bundle $\pi^*(T^*\mathcal{X})\otimes V\mathcal{E} \longrightarrow \mathcal{E}$. Here the first factor is the pull-back of $T^*\mathcal{X} \longrightarrow \mathcal{X}$ over $\pi : \mathcal{E} \longrightarrow \mathcal{X}$. As the pull-back notation is rather cumbersome, we will write this bundle as $T^*\mathcal{X} \underset{\mathcal{E}}{\otimes} V\mathcal{E}$. The $\mathcal{E}$ under the tensor product sign signals that both factors are to be considered as bundles over $\mathcal{E}$ using, if necessary, appropriate pull-backs which are usually obvious from the context.

At first sight, Definition 1.2.1 seems to be completely unrelated to our previous Definition 1.1.1. So we show next that points in $J_1\mathcal{E}$ may be interpreted as equivalence classes of sections $\sigma \in \Gamma_{loc}(\mathcal{X}, \mathcal{E})$. For such a section we define its (first) *prolongation* $j_1\sigma \in \Gamma_{loc}(\mathcal{X}, J_1\mathcal{E})$ by

$$j_1\sigma : \begin{cases} \mathcal{X} & \longrightarrow & J_1\mathcal{E} \\ x & \longmapsto & (\sigma(x), T_x\sigma) \end{cases} . \tag{1.17}$$

The tangent map $T_x\sigma$ indeed belongs to the space $\mathcal{A}^{(1)}_\xi$ with $\xi = \sigma(x)$. As σ is a map from $\mathcal{X}$ to $\mathcal{E}$, its tangent map $T_x\sigma$ is a linear map $T_x\mathcal{X} \longrightarrow T_\xi\mathcal{E}$ and thus an element of $T^*_x\mathcal{X} \otimes T_\xi\mathcal{E}$. Furthermore, by definition of a section, σ satisfies $\pi \circ \sigma = \mathrm{id}_\mathcal{X}$ and hence by the chain rule $T_\xi\pi \circ T_x\sigma = \mathrm{id}_{T_x\mathcal{X}}$. Conversely, to any $\gamma \in \mathcal{A}^{(1)}_\xi$ many sections $\sigma \in \Gamma_{loc}(\mathcal{X}, \mathcal{E})$ exist such that $\sigma(x) = \xi$ and $T_x\sigma = \gamma$. The bijection

$$\gamma \in \mathcal{A}^{(1)}_\xi \longleftrightarrow [\sigma] = \left\{\sigma \in \Gamma_{loc}(\mathcal{X}, \mathcal{E}) \mid \sigma(x) = \xi,\ T_x\sigma = \gamma\right\} \tag{1.18}$$

gives us thus a canonical identification between the jet bundles introduced in the last section and those of Definition 1.2.1. Hence we will not distinguish between them in the sequel.

In adapted coordinates $(\mathbf{x}, \mathbf{u})$ on $\mathcal{E}$, we may locally write the section σ as $\sigma(\mathbf{x}) = \big(\mathbf{x}, s(\mathbf{x})\big)$. For the tangent map $T_x\sigma$ we obtain then the local expression $(\mathbf{x}, \dot{\mathbf{x}}) \mapsto \big(\sigma(\mathbf{x}), d\sigma(\mathbf{x}) \cdot \dot{\mathbf{x}}\big)$ with the matrix

$$d\sigma(\mathbf{x}) = \begin{pmatrix} \frac{\partial s^1}{\partial x^1}(\mathbf{x}) & \cdots & \frac{\partial s^1}{\partial x^n}(\mathbf{x}) \\ \vdots & & \vdots \\ \frac{\partial s^m}{\partial x^1}(\mathbf{x}) & \cdots & \frac{\partial s^m}{\partial x^n}(\mathbf{x}) \\ 1 & 0 \cdots 0 & 0 \\ & \ddots & \\ 0 & 0 \cdots 0 & 1 \end{pmatrix} . \tag{1.19}$$

It consists of the Jacobian of the function s evaluated at the fixed point $x \in \mathcal{X}$ stacked over the $n \times n$ identity matrix. Conversely, the matrix of any linear map contained in $\mathcal{A}^{(1)}_\xi$ has the form (1.19), i.e. an arbitrary $m \times n$ matrix stacked over the $n \times n$ identity matrix.

Remark 1.2.2. A point $(\xi, \gamma_\xi) \in J_1\mathcal{E}$ induces a decomposition of the tangent space $T_\xi\mathcal{E}$. By definition, the fibre component γ_ξ represents an element of $A^{(1)}_\xi$, i.e. a linear map $T_x\mathcal{X} \to T_\xi\mathcal{E}$ (where again $x = \pi(\xi)$). Because of the condition $T_\xi\pi \circ \gamma_\xi = \mathrm{id}_{T_x\mathcal{X}}$, the intersection $\operatorname{im}\gamma_\xi \cap V_\xi\mathcal{E}$ contains only the zero vector and $T_\xi\mathcal{E} = \operatorname{im}\gamma_\xi \oplus V_\xi\mathcal{E}$. This implies that a (global) section γ of the bundle $\pi^1_0 : J_1\mathcal{E} \to \mathcal{E}$ with $\gamma(\xi) = (\xi, \gamma_\xi)$ may be interpreted as a *connection* (see Appendix C.2) on $\mathcal{E}$ where the horizontal space $H_\xi\mathcal{E}$ at the point $\xi \in \mathcal{E}$ is given by $\operatorname{im}\gamma_\xi$.

In fact, *any* connection on $\mathcal{E}$ can be generated that way. More precisely, there exists a bijection between the space of all connections on $\mathcal{E}$ and $\Gamma(\mathcal{E}, J_1\mathcal{E})$, the space of all global sections $\mathcal{E} \to J_1\mathcal{E}$. Given a section $\gamma \in \Gamma(\mathcal{E}, J_1\mathcal{E})$, we define a connection via its *horizontal lift* $A : \mathcal{E} \underset{\mathcal{X}}{\times} T\mathcal{X} \to T\mathcal{E}$ by $A(\xi, v_x) = \big(\xi, \gamma_\xi(v_x)\big)$. Conversely, given a horizontal lift A, we define a section $\gamma \in \Gamma(\mathcal{E}, J_1\mathcal{E})$ by setting $\gamma_\xi(v_x) = A(\xi, v_x)$. ◁

The geometry of the jet bundle is to a large extent determined by its *contact structure*. In the previous section we introduced it in terms of the contact forms locally spanned by the one-forms ω^α_μ defined by (1.8). Their definition is obviously not intrinsic, although our characterisation of contact forms via pull-backs is. In this section we take another approach to the contact structure via the *contact map*. Using Definition 1.2.1, it becomes a simple evaluation map:

$$\Gamma_1 : \begin{cases} J_1\mathcal{E} \underset{\mathcal{X}}{\times} T\mathcal{X} \longrightarrow T\mathcal{E} \\ (\xi, \gamma_\xi, v_x) \longmapsto \big(\xi, \gamma_\xi(v_x)\big) \end{cases} . \tag{1.20}$$

Note the close relation of the contact map to the horizontal lift A introduced in the remark above: given a global section $\gamma : \mathcal{E} \to J_1\mathcal{E}$ we can write $A(\xi, v_x) = \Gamma_1\big(\gamma(\xi), v_x\big)$. The *complementary contact map* is defined by

$$\theta_1 : \begin{cases} J_1\mathcal{E} \underset{\mathcal{E}}{\times} T\mathcal{E} \longrightarrow V\mathcal{E} \\ (\xi, \gamma_\xi, w_\xi) \longmapsto w_\xi - \Gamma_1\big(\xi, \gamma_\xi, T_\xi\pi(w_\xi)\big) \end{cases} . \tag{1.21}$$

Like Γ_1 is related to the horizontal lift A, the complementary contact map θ_1 is related to the vertical projector P^V of the connection induced by a section $\gamma : \mathcal{E} \to J_1\mathcal{E}$, namely $P^V(\xi, w_\xi) = \theta_1\big(\gamma(\xi), w_\xi\big)$.

As the map Γ_1 is linear in its last argument, we may interpret it alternatively as a map $J_1\mathcal{E} \to T^*\mathcal{X} \underset{\mathcal{E}}{\otimes} T\mathcal{E}$. Then its local coordinate form is $\Gamma_1 : (x^i, u^\alpha, p^\alpha_i) \mapsto \big(x^i, u^\alpha; dx^i \otimes (\partial_{x^i} + p^\alpha_i \partial_{u^\alpha})\big)$. It is no accident that the vector fields in the parentheses look exactly like the contact fields (1.10) in the case $q = 1$; after an appropriate pull-back they are indeed identical. Similarly, we interpret θ_1 as a map $J_1\mathcal{E} \to T^*\mathcal{E} \underset{\mathcal{E}}{\otimes} V\mathcal{E}$ with the local coordinate form $\theta_1 : (x^i, u^\alpha, p_i^\alpha) \mapsto \big(x^i, u^\alpha; (du^\alpha - p_i^\alpha dx^i) \otimes \partial_{u^\alpha}\big)$. Now we recover (after some appropriate pull-back) in the parentheses the contact forms (1.8).

The jet bundles of higher order permit a similar direct construction, although it becomes somewhat less elegant, as the characterisation of the appearing affine spaces is more cumbersome. We proceed inductively and define $J_q\mathcal{E}$ as an affine bundle over $J_{q-1}\mathcal{E}$.

Definition 1.2.3. *Let $\pi : \mathcal{E} \to \mathcal{X}$ be a fibred manifold and $J_{q-1}\mathcal{E}$ the jet bundle of order $q-1$. The* qth order jet bundle *over $\mathcal{E}$ is the affine bundle $\pi^q_{q-1} : J_q\mathcal{E} \to J_{q-1}\mathcal{E}$ whose fibre at the point $\rho \in J_{q-1}\mathcal{E}$ with $\pi_0^{q-1}(\rho) = \xi \in \mathcal{E}$ and $\pi(\xi) = x \in \mathcal{X}$ is given by the affine space*

$$
\begin{aligned}
(J_q\mathcal{E})_\rho = \mathcal{A}_\xi^{(q)} = \Big\{\gamma \in & (T_x^*\mathcal{X})^q \otimes T_\xi\mathcal{E} \mid \\
& \gamma(v_1, \dots, v_q) = \big(v_q, B_q(v_1, \dots, v_q)\big) \\
& \text{with } B_q \text{ totally symmetric}\Big\}
\end{aligned}
\tag{1.22}
$$

modelled on the vector space

$$
\mathcal{V}_\xi^{(q)} = S_q(T_x^*\mathcal{X}) \otimes V_\xi\mathcal{E}
\tag{1.23}
$$

where S_q denotes the q-fold symmetric product.

The elements of $\mathcal{A}_\xi^{(q)}$ may also be considered as maps $T_x\mathcal{X} \to \mathcal{A}_\xi^{(q-1)}$. It follows by the same argument as in the case of the first-order jet bundle that $\mathcal{A}_\xi^{(q)}$ is indeed an affine space. If we take the difference of two elements $\gamma_1, \gamma_2 \in \mathcal{A}_\xi^{(q)}$, we obtain a linear symmetric map whose value is a vertical vector, i.e. an element of $\mathcal{V}_\xi^{(q)}$. The affine bundle $\pi^q_{q-1} : J_q\mathcal{E} \to J_{q-1}\mathcal{E}$ is thus modelled on the vector bundle $S_q(T^*\mathcal{X}) \underset{J_{q-1}\mathcal{E}}{\otimes} V\mathcal{E}$ (writing out the involved pull-backs, this is the bundle $(\pi^{q-1})^* S_q(T^*\mathcal{X}) \otimes (\pi_0^{q-1})^* V\mathcal{E} \to J_{q-1}\mathcal{E}$). It follows from (1.5) (or (A.4a), respectively) that the fibre dimension of this vector bundle is

$$
\dim\big(S_q(T^*\mathcal{X}) \underset{J_{q-1}\mathcal{E}}{\otimes} V\mathcal{E}\big) = m\binom{n+q-1}{q} .
\tag{1.24}
$$

Remark 1.2.4. As in the first-order case any global section $J_{q-1}\mathcal{E} \to J_q\mathcal{E}$ defines a connection on $J_{q-1}\mathcal{E}$. We do not give here any details how the splitting of the tangent bundle into a horizontal and a vertical subbundle is performed. This will become obvious below, when we study the relation to iterated jet bundles. ◁

We still have to show that the thus constructed bundle may indeed be identified with the jet bundle. In the case of the first-order jet bundle this was straightforward, as the elements of $\mathcal{A}^{(1)}_{\xi}$ are of the form $T_x\sigma$ for sections $\sigma : \mathcal{X} \to \mathcal{E}$. For the higher-order jet bundles it is similarly possible to identify the elements of $\mathcal{A}^{(q)}_{\xi}$ with higher-order tangent maps, i. e. tangent maps of tangent maps etc. As this becomes notationally very messy, we show it explicitly only for the second-order jet bundle.

Consider a section $\sigma : \mathcal{X} \to \mathcal{E}$ with $\sigma(x) = \xi = \bigl(x, s(x)\bigr)$. Then $T_x\sigma$ is a map $T_x\mathcal{X} \to T_\xi\mathcal{E}$. Let $w = T_x\sigma(v)$ for some vector $v \in T_x\mathcal{X}$; we have seen above that we may write $w = \bigl(v, (\partial s/\partial x)v\bigr)$. The second-order tangent map is a linear map $T_{(x,v)}(T_x\sigma) : T_{(x,v)}(T_x\mathcal{X}) \to T_{(\xi,w)}(T_\xi\mathcal{E})$. A vector $z \in T_{(x,v)}(T_x\mathcal{X})$ can be split into two parts $z = (z_1, z_2)$ with z_1 being “tangent in x-direction” and z_2 in “v-direction”. Then

$$T_{(x,v)}(T_x\sigma)(z) = \Bigl(z_1, \frac{\partial s}{\partial x}v, z_2, \frac{\partial^2 s}{\partial x^2}(v, z_1) + \frac{\partial s}{\partial x}z_2\Bigr) . \tag{1.25}$$

Obviously, this map is completely determined by the Jacobian $\partial s/\partial x$ and the Hessian $\partial^2 s/\partial x^2$. The Jacobian is already encoded in the point $\rho \in J_1\mathcal{E}$ where we consider the fibre of $J_2\mathcal{E}$. The Hessian evaluated at a fixed point $x \in \mathcal{X}$ defines a linear symmetric map B_2 as in Definition 1.2.3. Conversely, to every linear symmetric map B_2 many sections σ exist such that their Hessian at the single point $x \in \mathcal{X}$ coincides with B_2. So we obtain again a bijection between a point in $J_q\mathcal{E}$ and an equivalence class of local sections of $\mathcal{E}$.

As each jet bundle $J_q\mathcal{E}$ is also a fibred manifold over the base space $\mathcal{X}$, we may construct iterated jet bundles $J_r(J_q\mathcal{E})$. Another way to view $J_{q+1}\mathcal{E}$ is then as a submanifold of $J_1(J_q\mathcal{E})$. The mapping $\iota_{q,1} : J_{q+1}\mathcal{E} \to J_1(J_q\mathcal{E})$ describing this identification is defined intrinsically as the prolongation of the identity map on $J_q\mathcal{E}$, i. e. as the unique mapping satisfying for any section $\sigma \in \Gamma_{\text{loc}}(\mathcal{X}, \mathcal{E})$ the condition $\iota_{q,1} \circ j_{q+1}\sigma = j_1(j_q\sigma)$.

As by Definition 1.2.1 the fibre components of elements of $J_1(J_q\mathcal{E})$ may be considered as maps $\gamma : T\mathcal{X} \to TJ_q\mathcal{E}$ satisfying $T\pi^q \circ \gamma = \mathrm{id}_{T\mathcal{X}}$, the same may be done with the elements of $\mathcal{A}^{(q)}$. More precisely, $\mathcal{A}^{(q)}$ is isomorphic to an affine subspace of the space of all such maps γ which is characterised by a symmetry condition.

If we take on $J_q\mathcal{E}$ our usual local coordinates (x^i, p^α_μ), then local coordinates on $J_1(J_q\mathcal{E})$ are $(x^i, p^\alpha_\mu; p^\alpha_{\mu,i})$, i. e. all derivatives p^α_μ are considered as dependent variables and all their first-order derivatives $p^\alpha_{\mu,i}$ are introduced. The submanifold of $J_1(J_q\mathcal{E})$ with which we identify $J_{q+1}\mathcal{E}$ is described by the local equations $p^\alpha_{\mu,i} = p^\alpha_{\nu,j}$ whenever $\mu + 1_i = \nu + 1_j$. Obviously, these equations express that partial derivatives commute.

Similarly, we introduce mappings $\iota_{q,r} : J_{q+r}\mathcal{E} \to J_r(J_q\mathcal{E})$ by the requirement $\iota_{q,r} \circ j_{q+r}\sigma = j_r(j_q\sigma)$ for any section $\sigma \in \Gamma_{\text{loc}}(\mathcal{X}, \mathcal{E})$. Local coordinates on $J_r(J_q\mathcal{E})$ are of the form $p^\alpha_{\mu,\nu}$ with $0 \leq |\mu| \leq q$ and $0 \leq |\nu| \leq r$. The image of $J_{q+r}\mathcal{E}$ in $J_r(J_q\mathcal{E})$ is then described by the equations $p^{\alpha}_{\mu,\nu} = p^{\alpha}_{\bar\mu,\bar\nu}$ whenever $\mu+\nu = \bar\mu+\bar\nu$.

More generally, we define the prolongation of any map $\Phi : J_q\mathcal{E} \to \mathcal{E}'$ where $\mathcal{E}'$ is another fibred manifold over $\mathcal{X}$ and Φ is fibred over the identity. The r-fold prolongation $\rho_r\Phi$ is a map $J_{q+r}\mathcal{E} \to J_r\mathcal{E}'$ such that the diagram

$$\begin{tikzcd}
J_{q+r}\mathcal{E} \arrow[r, "\pi^{q+r}_r"] \arrow[rrr, bend left, "\rho_r\Phi"] & J_q\mathcal{E} \arrow[r, "\Phi"] & \mathcal{E}' & J_r\mathcal{E}' \arrow[l, "\pi^r_0"'] \\
& \mathcal{X} \arrow[ul, "j_{q+r}\sigma"] \arrow[u, "j_q\sigma"] \arrow[r, equal] & \mathcal{X} \arrow[u, "\sigma'"'] \arrow[ur, "j_r\sigma'"'] &
\end{tikzcd} \tag{1.26}$$

commutes for any local section $\sigma \in \Gamma_{loc}(\mathcal{X},\mathcal{E})$, i.e. the prolonged map satisfies $j_r(\Phi \circ j_q\sigma) = (\rho_r\Phi) \circ (j_{q+r}\sigma)$.

If we use adapted coordinates $(\mathbf{x},\mathbf{u})$ on $\mathcal{E}$ and $(\mathbf{x},\mathbf{v})$ on $\mathcal{E}'$, so that locally the map Φ is of the form $v^\tau = \Phi^\tau(\mathbf{x},\mathbf{p})$, then the prolongation $\rho_r\Phi$ has the coordinate form $q^\tau_\sigma = D_\sigma\Phi^\tau(\mathbf{x},\mathbf{p})$ with $0 \le |\sigma| \le r$ (we denote the jet coordinates of $J_r\mathcal{E}'$ by q^τ_σ). Here $D_\sigma = D_1^{\sigma_1}\cdots D_n^{\sigma_n}$ with D_i the formal derivative defined by (1.11). Thus (1.26) is just an abstract representation of the chain rule.

The *contact structure* of the higher-order jet bundle may be introduced in a similar manner to the first-order case. The q-contact map is a map $\Gamma_q : J_q\mathcal{E} \times_{\mathcal{X}} T\mathcal{X} \to TJ_{q-1}\mathcal{E}$ such that the diagram

$$\begin{tikzcd}
J_q\mathcal{E} \times_{\mathcal{X}} T\mathcal{X} \arrow[rr, "\Gamma_q"] & & TJ_{q-1}\mathcal{E} \\
& T\mathcal{X} \arrow[ul, "((j_q\sigma)\circ\tau_{\mathcal{X}})\times \mathrm{id}_{T\mathcal{X}}"] \arrow[ur, "Tj_{q-1}\sigma"'] &
\end{tikzcd} \tag{1.27}$$

commutes for any section $\sigma \in \Gamma_{loc}(\mathcal{X},\mathcal{E})$. This condition defines it uniquely. It is easy to see that Γ_q is a linear fibred morphism over π^q_{q-1} which can be expressed by the fact that the diagram

$$\begin{tikzcd}
J_q\mathcal{E} \times_{\mathcal{X}} T\mathcal{X} \arrow[r, "\Gamma_q"] \arrow[d, "\mathrm{pr}_1"'] & TJ_{q-1}\mathcal{E} \arrow[d, "\tau_{J_{q-1}\mathcal{E}}"] \\
J_q\mathcal{E} \arrow[r, "\pi^q_{q-1}"'] & J_{q-1}\mathcal{E}
\end{tikzcd} \tag{1.28}$$

commutes. Here pr_1 is the projection on the first factor and τ the tangent bundle projection. The linearity implies that we may consider Γ_q as a map

$J_q\mathcal{E} \to T^*\mathcal{X} \underset{J_{q-1}\mathcal{E}}{\otimes} TJ_{q-1}\mathcal{E}$ which takes in adapted coordinates the concise form $\Gamma_q : (x^i, u^\alpha, p^\alpha_\mu) \mapsto \bigl(x^i, u^\alpha, p^\alpha_\mu; dx^i \otimes (\partial_{x^i} + p^\alpha_{\mu+1_i}\partial_{p^\alpha_\mu})\bigr)$. Note that on the right hand side only multi indices with $0 \leq |\mu| < q$ appear whereas on the left hand side $|\mu| = q$ is admitted, too. As in the first-order case, it is not difficult to recognise the contact fields (1.10) in the parentheses.

The *complementary contact map* of the qth order jet bundle is the map $\theta_q : J_q\mathcal{E} \underset{J_{q-1}\mathcal{E}}{\times} TJ_{q-1}\mathcal{E} \to VJ_{q-1}\mathcal{E}$ defined by $\theta_q(\rho, w) = w - \Gamma_q(\rho, T_{\bar\rho}\pi^{q-1}(w))$ where $\bar\rho = \pi^q_{q-1}\rho$. It takes in adapted coordinates on the jet bundle $J_q\mathcal{E}$ the form $\theta_q : (x^i, u^\alpha, p^\alpha_\mu) \mapsto \bigl(x^i, u^\alpha, p^\alpha_\mu; (dp^\alpha_\mu - p^\alpha_{\mu+1_i}dx^i) \otimes \partial_{p^\alpha_\mu}\bigr)$ where now the contact forms (1.8) reappear. Given a global section $J_{q-1}\mathcal{E} \to J_q\mathcal{E}$ defining a connection on $J_{q-1}\mathcal{E}$ we get the same relation between Γ_q and the horizontal lift on one hand and θ_q and the vertical projector on the other hand as in the first-order case.

1.3 Differential Equations

With the concept of jet bundles at our hands we give a geometric definition of (systems of) differential equations. It does not require coordinates or local equations but defines a differential equation as an intrinsic geometric object, namely a fibred submanifold.

Definition 1.3.1. *A (non-linear)* differential equation *of order* q *is a fibred submanifold* $\mathcal{R}_q$ *of the jet bundle* $J_q\mathcal{E}$*.*

Locally, i.e. within some open subset $\mathcal{V}_q \subseteq J_q\mathcal{E}$, we may always assume that a fibred map

$$\Phi : \begin{cases} \mathcal{V}_q \subseteq J_q\mathcal{E} \longrightarrow \mathcal{E}' \\ (\mathbf{x}, \mathbf{p}) \longmapsto \Phi(\mathbf{x}, \mathbf{p}) \end{cases} \tag{1.29}$$

exists taking values in a vector bundle $\pi' : \mathcal{E}' \to \mathcal{X}$ such that $\mathcal{R}_q \cap \mathcal{V}_q$ is the zero set of Φ. Thus locally the differential equation $\mathcal{R}_q$ can be described by a system of equations $\Phi^\tau(\mathbf{x}, \mathbf{p}) = 0$ with $1 \leq \tau \leq t = \dim\mathcal{E}'$ which brings us back to the traditional picture.

In the sequel we will mostly assume that there exists a *global* morphism $\Phi : J_q\mathcal{E} \to \mathcal{E}'$ such that $\mathcal{R}_q$ is its zero set, i.e. that our differential equation is defined by the *differential operator* $\Delta[\Phi] : \Gamma_{loc}(\mathcal{E}) \to \Gamma_{loc}(\mathcal{E}')$ with $\Delta[\Phi](\sigma) = \Phi \circ j_q\sigma$. As we have mainly local applications in mind, this assumption is not a serious restriction but simplifies the notation at some places. If $\pi : \mathcal{E} \to \mathcal{X}$ is also a vector bundle, we define *linear differential equations* by the condition that Φ is a vector bundle morphism. Note that the linearity refers only to the derivatives; the dependency on the independent variables may be arbitrary.

Definition 1.3.1 does *not* distinguish between scalar equations and systems. In fact, we will use in the sequel the following convention. When we speak of a *differential equation*, we always think of it as such a geometric object $\mathcal{R}_q$. When we speak of a *system*, we think of a local representation of $\mathcal{R}_q$ given by a map of the form (1.29). Of course, such a system consists again of equations $\Phi(x,p) = 0$, so that it is somewhat difficult to strictly maintain this distinction ...

Definition 1.3.1 requires $\mathcal{R}_q$ to be a *fibred* submanifold. This condition implies that $\pi^q(\mathcal{R}_q) = \mathcal{X}$. If it was violated, the independent variables would not be really independent, as the differential equation would impose relations between them. More precisely, some local representations of $\mathcal{R}_q$ would contain equations of the form $\Phi(\mathbf{x}) = 0$ which obviously should not happen.

Remark 1.3.2. First-order *ordinary* differential equations $\mathcal{R}_1 \subset J_1\mathcal{E}$ possess a special geometric structure. If $\mathcal{R}_1$ is the image of a section $\gamma : \mathcal{E} \to J_1\mathcal{E}$, we may interpret it according to Remark 1.2.2 as a connection on $\mathcal{E}$. Locally, such equations correspond to systems of the form $\dot{\mathbf{u}} = \boldsymbol{\phi}(t,\mathbf{u})$, i.e. we are dealing with a pure differential equation without algebraic constraints. If constraints are present, they determine a fibred submanifold $\mathcal{C} \subset \mathcal{E}$ and we may consider equations defined by sections $\gamma : \mathcal{C} \to J_1\mathcal{E}$. They correspond to connections on $\mathcal{C}$. Obviously, this can straightforwardly be extended to an identification of qth-order ordinary differential equations with connections on $J_{q-1}\mathcal{E}$.

In the case of *partial* differential equations, the images of global sections $\gamma : \mathcal{E} \to J_1\mathcal{E}$ yield *equations of finite type*, also called *maximally overdetermined* equations (see Remark 1.5.2). If such an equation is of first order, it can locally be represented by a system of the form $\mathbf{p} = \boldsymbol{\phi}(\mathbf{x},\mathbf{u})$ where the vector $\mathbf{p}$ contains all first-order derivatives. Again the inclusion of constraints, i.e. of additional algebraic equations defining a fibred submanifold $\mathcal{C} \subset \mathcal{E}$, is trivial. $\triangleleft$

Two natural geometric operations with differential equations exist: *prolongation* and *projection*. The former one "lifts" an equation to higher order; the latter one lowers the order. Both are of central importance for all what we will do with differential equations.

Prolonging a differential equation is easy to perform effectively, once a local representation is given. Its intrinsic geometric description is a bit more involved. We need again the embeddings $\iota_{q,r} : J_{q+r}\mathcal{E} \hookrightarrow J_r(J_q\mathcal{E})$. As $\mathcal{R}_q$ is a fibred manifold, too, we can construct jet bundles over it, for example $J_r\mathcal{R}_q$. It may be consider in a natural way as a submanifold of $J_rJ_q\mathcal{E}$. We denote the corresponding embedding by $\bar{\iota}_{q,r}$. Then we define the prolongation

$$\mathcal{R}_{q+r} = \iota_{q,r}^{-1}\big(\bar{\iota}_{q,r}(J_r\mathcal{R}_q) \cap \iota_{q,r}(J_{q+r}\mathcal{E})\big) \subseteq J_{q+r}\mathcal{E} \ . \tag{1.30}$$

In words, this somewhat awkward construction has a fairly simple explanation. We consider both $J_r\mathcal{R}_q$ and $J_{q+r}\mathcal{E}$ as submanifolds of $J_r(J_q\mathcal{E})$ using the corresponding embeddings. Then we take their intersection and consider the result as a submanifold of $J_{q+r}\mathcal{E}$. As already discussed in the last section, the intersection enforces the commutativity of partial differentiations.

If our differential equation $\mathcal{R}_q$ comes from a differential operator, its prolongation can be expressed via the prolongation of the corresponding global map Φ. We get the following commuting diagram

$$\begin{tikzcd} 0 \arrow[r] & \mathcal{R}_{q+r} \arrow[r] \arrow[d, "\pi^{q+r}_q"'] & J_{q+r}\mathcal{E} \arrow[r, "\rho_r\Phi"] \arrow[d, "\pi^{q+r}_q"'] & J_r\mathcal{E}' \arrow[d, "(\pi')^r_0"] \\ 0 \arrow[r] & \mathcal{R}_q \arrow[r] & J_q\mathcal{E} \arrow[r, "\Phi"] & \mathcal{E}' \end{tikzcd} \tag{1.31}$$

and the prolonged equation $\mathcal{R}_{q+r}$ is the zero set of the prolonged map $\rho_r\Phi$.

In local coordinates, the prolongation is computed with the help of the formal derivative (1.11). If $\Phi(\mathbf{x},\mathbf{p}) = 0$ is a local representation of $\mathcal{R}_q$, then $\mathcal{R}_{q+r}$ is described locally by the system

$$\mathcal{R}_{q+r} : \begin{cases} \Phi^\tau(\mathbf{x},\mathbf{p}) &= 0 \ , \qquad 1 \leq \tau \leq t \ , \\ D_j\Phi^\tau(\mathbf{x},\mathbf{p}) &= 0 \ , \qquad 1 \leq j \leq n \ , \\ \quad \vdots & \\ D_\nu\Phi^\tau(\mathbf{x},\mathbf{p}) &= 0 \ , \qquad |\nu| = r \ . \end{cases} \tag{1.32}$$

Indeed, these are just the coordinate expressions of $\rho_r\Phi$.

In contrast, projections are easily described geometrically but difficult to perform effectively in local coordinates. Geometrically we simply apply the natural projections $\pi^q_{q-r} : J_q\mathcal{E} \to J_{q-r}\mathcal{E}$ and define for $r \leq q$

$$\mathcal{R}^{(r)}_{q-r} = \pi^q_{q-r}(\mathcal{R}_q) \ . \tag{1.33}$$

To obtain a local representation of the differential equation $\mathcal{R}^{(r)}_{q-r} \subseteq J_{q-r}\mathcal{E}$, one takes a local representation of $\mathcal{R}_q$, say $\Phi^\tau(x,p) = 0$, and tries to eliminate by purely algebraic operations the derivatives of order greater than $q-r$ in as many equations as possible. If we cannot construct any equation depending only on lower-order derivatives, then $\mathcal{R}^{(r)}_{q-r} = J_{q-r}\mathcal{E}$. Otherwise it is a proper submanifold locally described by all such equations.

We will see later many examples of such computations. Fortunately, we mostly project differential equations generated by prolongations. In this case, the elimination of the higher-order derivatives requires only linear algebra. A slight complication arises from the fact that the prolongation or the projection of a given differential equation $\mathcal{R}_q$ is not necessarily a submanifold anymore. For projections this is rather obvious and the following example shows a counter example for a prolongation.

Example 1.3.3. We consider the *non-linear* second-order equation

$$\mathcal{R}_2 : \begin{cases} u_{yy} - \frac{1}{2}(u_{xx})^2 = 0 , \\ u_{xy} - u_{xx} \qquad = 0 . \end{cases} \tag{1.34}$$

Adding the x- and the y-derivative of the second equation and subtracting the x-derivative of the first one yields $(u_{xx} - 1)u_{xxx} = 0$. Thus $\mathcal{R}_3$ lies in the union of the two hyperplanes $u_{xx} = 1$ and $u_{xxx} = 0$. Such a union is *not* a submanifold, as in the neighbourhood of the intersection it is not homeomorphic to an open subset of some $\mathbb{R}^k$. We must distinguish two cases in which the differential equation might show quite different behaviour; in particular, the solution spaces will generally differ. ◁

For simplicity, we will always assume that all performed projections and prolongations yield again submanifolds. This means that we consider only *regular* equations. Regularity can always be achieved by restricting to some submanifold of $\mathcal{R}_q$ but is difficult to check effectively.

One could be tempted to think that prolongation and projection are some kind of "inverse" operations: if one first prolongs a differential equation $\mathcal{R}_q$ to $\mathcal{R}_{q+r}$ for some $r > 0$ and subsequently projects back to $J_q\mathcal{E}$ with π^{q+r}_q, one might naively expect that the obtained equation $\mathcal{R}^{(r)}_q$ is identical with the original equation $\mathcal{R}_q$. However, this is in general not correct, as *integrability conditions* may arise: we only get that always $\mathcal{R}^{(r)}_q \subseteq \mathcal{R}_q$.

Example 1.3.4. A trivial example may explain this phenomenon and also demonstrate explicitly the operations of prolongation and projection. We consider the first-order equation $\mathcal{R}_1$ defined by the linear system

$$\mathcal{R}_1 : \begin{cases} u_z + y u_x = 0 , \\ u_y \qquad\quad = 0 . \end{cases} \tag{1.35}$$

In order to obtain the system for the prolonged equation $\mathcal{R}_2$ we must formally differentiate all equations in (1.35) with respect to all independent variables x, y, and z. The arising equations are added to the original system and we get

$$\mathcal{R}_2 : \begin{cases} u_z + y u_x = 0 , & u_y = 0 , \\ u_{xz} + y u_{xx} = 0 , & u_{xy} = 0 , \\ u_{yz} + y u_{xy} + u_x = 0 , & u_{yy} = 0 , \\ u_{zz} + y u_{xz} = 0 , & u_{yz} = 0 . \end{cases} \tag{1.36}$$

In order to compute the projection to $J_1\mathcal{E}$ of the prolongation $\mathcal{R}_2$ we must try to obtain by algebraic manipulations as many first-order equations as possible. Since we are dealing with a linear system, this is not very difficult. The first two equations are obviously first order, as they stem from the original system. In the first equation of the third row we can eliminate both second-order derivatives using equations of the second column. No other eliminations are possible. So we find the integrability condition $u_x = 0$ and a local representation of the projected equation $\mathcal{R}_1^{(1)}$ is given by

$$\mathcal{R}_1^{(1)} : \begin{cases} u_z = 0 , \\ u_y = 0 , \\ u_x = 0 . \end{cases} \tag{1.37}$$

It is easy to see that further prolongations and subsequent projections do not yield any additional conditions and thus $\mathcal{R}_1^{(1)} = \mathcal{R}_1^{(2)} = \mathcal{R}_1^{(3)} = \cdots$. ◁

It is important to note that in Example 1.3.4 it is not possible to construct the integrability condition by purely algebraic operations without differentiations. This property may serve as an informal definition of integrability conditions: they are equations algebraically independent of the original system that can be generated using formally differentiated equations. Obviously, any sufficiently smooth solution of the system *automatically* satisfies the integrability conditions; thus they do not represent additional restrictions to the solution space but they are equations that have been "hidden" in the structure of the system.

In Example 1.3.4 we see the classical form of an integrability condition. It arises as a generalised cross-derivative: a linear combination of differentiated equations that yields a lower-order equation. Obviously, cross-derivatives make sense only for partial differential equations. However, a second mechanism for the generation of integrability conditions exists which also works for ordinary differential equations: if the original system contains equations of differing orders, the differentiation of a lower-order equation may generate an integrability condition.

Example 1.3.5. We consider the trivial second-order equation $\mathcal{R}_2$ described in local coordinates by the linear system

$$\mathcal{R}_2 : \begin{cases} u_{yy} = 0 , \\ u_x = 0 . \end{cases} \tag{1.38}$$

Prolongation to third order yields a system that contains the second-order equations $u_{xy} = u_{yy} = 0$. Thus the projection to $J_2\mathcal{E}$ is not surjective and we must add these two integrability conditions to the system (1.38) in order to obtain a local representation of $\mathcal{R}_2^{(1)}$. $\lhd$

Geometrically, it makes no sense to speak of individual integrability conditions or to distinguish between cross-derivatives and the prolongation of lower-order equations. Instead both kinds of integrability conditions have the same effect, namely that for some $r > 0$ the differential equation $\mathcal{R}_q^{(r)}$ is a proper subset of $\mathcal{R}_q$. More precisely, it is a submanifold of lower dimension. In Example 1.3.4 $\mathcal{R}_1$ is two-dimensional (the fibre dimension of $J_1\mathcal{E}$ is four and we have two independent equations), whereas $\dim \mathcal{R}_1^{(1)} = 1$.

Finally, we need of course the notion of a solution. Classically, solutions are functions that entered into a given system let it vanish identically. In our geometric framework, functions are substituted by sections.

Definition 1.3.6. *Let* $\mathcal{R}_q \subseteq J_q\mathcal{E}$ *be a differential equation of order* q. *A* (local) solution *is a smooth section* $\sigma \in \Gamma_{loc}(\mathcal{E})$ *such that* $\operatorname{im} j_q\sigma \subseteq \mathcal{R}_q$.

In local coordinates, such a section $\sigma : \mathcal{V} \subseteq \mathcal{X} \to \mathcal{E}$ can be written as a mapping $\mathbf{x} \mapsto (\mathbf{x}, \mathbf{s}(\mathbf{x}))$. Its prolongation $j_q\sigma : \mathcal{V} \to J_q\mathcal{E}$ has locally the form $\mathbf{x} \mapsto (\mathbf{x}, \mathbf{s}(\mathbf{x}), \partial^{|\mu|}\mathbf{s}(\mathbf{x})/\partial x^{\mu})$. If the differential equation $\mathcal{R}_q$ is locally represented by the system $\Phi(\mathbf{x}, \mathbf{p}) = 0$, the condition $\operatorname{im} j_q\sigma \subseteq \mathcal{R}_q$ requires that $\Phi(\mathbf{x}, \partial^{|\mu|}\mathbf{s}(\mathbf{x})/\partial x^{\mu}) = 0$ for all $\mathbf{x} \in \mathcal{V}$, i.e. the (smooth) function $\mathbf{s}(\mathbf{x})$ defined on $\mathcal{V}$ is a solution in the conventional sense.

Finally, we compare our definition of a differential equation with the situation in algebraic geometry. There a variety is the zero set of some polynomials, i.e. every point on the variety represents a solution of the algebraic equations defined by the polynomials. Let the differential equation $\mathcal{R}_q \subseteq J_q\mathcal{E}$ be locally represented by the system $\Phi(\mathbf{x}, \mathbf{p}) = 0$. It may be understood as a system of *algebraic* equations on $J_q\mathcal{E}$ and the points on $\mathcal{R}_q$ are the solutions of these equations. As in algebraic geometry, the functions Φ do not really matter but only the ideal in $\mathcal{F}(J_q\mathcal{E})$ generated by them.

Recall that a point of the jet bundle $J_q\mathcal{E}$ was defined as an equivalence class of sections. Let the class $[\sigma]_{\mathbf{x}}^{(q)}$ be a point in $J_q\mathcal{E}$ and $\mathcal{V} \subseteq \mathcal{X}$ a neighbourhood of $\mathbf{x}$ such that σ is defined on $\mathcal{V}$. If the section σ is a solution of the differential equation $\mathcal{R}_q$ in the sense of Definition 1.3.6, the points $[\sigma]_{\mathbf{x}}^{(q)}$ (for all $\mathbf{x} \in \mathcal{V}$) must lie on the submanifold $\mathcal{R}_q$. However, the converse does not necessarily hold and the equivalence classes $[\sigma]_{\mathbf{x}}^{(q)}$ in general contain many sections that are *not* solutions of $\mathcal{R}_q$. This is easy to understand, as the equivalence relation $\sim_q$ takes into account only the values of the derivatives of σ up to order q at a *single* point $\mathbf{x}$.

1.4 Formal Integrability

We have seen that in general integrability conditions arise, if we prolong and project a differential equation. We now show that such conditions are obstructions for the construction of power series solutions. Obviously, such a computation can only be performed in local coordinates. Thus we assume that we are given a differential equation $\mathcal{R}_q$ described in some neighbourhood of the point $\mathbf{x}_0 \in \mathcal{X}$ by the system $\Phi(\mathbf{x}, \mathbf{p}) = 0$ and make the usual ansatz for a power series solution

$$u^\alpha(x) = \sum_{|\mu|=0}^{\infty} \frac{a^\alpha_\mu}{\mu!}(\mathbf{x} - \mathbf{x}_0)^\mu \tag{1.39}$$

with real coefficients a^α_μ. Note that we are only dealing with formal series, as we do not discuss their convergence (but see Chapter 5).

Of course, without initial or boundary conditions the equation possesses in general many solutions and we cannot determine uniquely all coefficients in (1.39). But we can at least exhibit some relations between them. We enter the ansatz into the system describing $\mathcal{R}_q$ and evaluate the arising expressions at the point $\mathbf{x} = \mathbf{x}_0$. One easily checks that this yields the following *algebraic* equations for the coefficients a^α_μ with $|\mu| \leq q$:

$$\Phi^\tau(x^i_0, a^\alpha_\mu) = 0\ , \qquad 1 \leq \tau \leq t\ , \tag{1.40}$$

i.e. in the functions Φ^τ we only substitute the derivatives p^α_μ by the coefficients a^α_μ of the ansatz (1.39). In general, (1.40) represent a non-linear system of equations. Hence its solution space might have a very complicated structure. For example, it could consist of disjoint parts of different dimensions. Each of these must be treated separately.

Now we turn our attention to the prolonged equation $\mathcal{R}_{q+1}$. According to (1.32) it is described in a neighbourhood of $\mathbf{x}_0$ by the original equations $\Phi(\mathbf{x}, \mathbf{p}) = 0$ and in addition by the equations $D_i\Phi(\mathbf{x}, \mathbf{p}) = 0$ for $1 \leq i \leq n$. Again we enter our ansatz into this system and evaluate at $\mathbf{x} = \mathbf{x}_0$. The first part of the arising algebraic system for the coefficients a^α_μ with $|\mu| \leq q + 1$ has already been handled in the previous step. The only new equations are

$$D_j\Phi^\tau(x^i_0, a^\alpha_\mu) = 0\ , \qquad \begin{cases} 1 \leq \tau \leq t\ , \\ 1 \leq j \leq n\ . \end{cases} \tag{1.41}$$

If we assume that all relations between the coefficients up to order q are already taken into account with (1.40), then (1.41) is an inhomogeneous *linear* system for the coefficients of order $q + 1$. Its matrix and right hand side depends on the coefficients of lower order. The linearity of the system stems from the quasi-linearity of the formal derivative (1.11).

In this manner we proceed order by order. At order $q+r$ we get from the local representation of $\mathcal{R}_{q+r}$ as additional equations for the coefficients a^α_μ with $|\mu| \leq q+r$

$$D_\nu \Phi^\tau(x^i_0, a^\alpha_\mu) = 0\ , \qquad \begin{cases} 1 \leq \tau \leq t\ , \\ |\nu| = r\ . \end{cases} \tag{1.42}$$

Again we may consider it as an inhomogeneous linear system for the coefficients of order $q+r$.

In the end we obtain a system of infinitely many equations for the infinitely many coefficients of the ansatz (1.39). Each solution of this system defines a formal power series solution. Except (1.40) all systems are linear and their solution spaces have a very simple structure. But if we choose different solutions of (1.40), the solution spaces of the linear systems (1.42) may change their dimension, as their matrices depend on the coefficients of lower order. This corresponds to the existence of so-called *singular solutions* briefly discussed in Section 4.1.

Each of these infinitely many equations can be solved for a different coefficient a^α_μ of our ansatz. These coefficients are called *principal*, the remaining ones are the *parametric coefficients*. Of course, this distinction is not uniquely defined; in general, there are many ways to solve the equations for some coefficients. The parametric coefficients may be chosen arbitrarily; they parametrise the formal solution space. The principal coefficients are determined by the differential equation $\mathcal{R}_q$ and its prolongations.

In a concrete computation we can perform only a finite number of prolongations. Thus we must stop the above construction at some order $s \geq q$. Of course, we expect that with the help of the so far found equations for the coefficients a^α_μ with $0 \leq |\mu| \leq s$ we obtain at least a correct truncation of power series solutions. However, this is only true, if no integrability conditions of order less than or equal to s are hidden in the system describing $\mathcal{R}_s$, as otherwise some further conditions on the coefficients a^α_μ with $0 \leq |\mu| \leq s$ exist which we have not taken into account.

Example 1.4.1. In Example 1.3.4 $\mathcal{R}_1^{(1)} \neq \mathcal{R}_1$. Entering a power series ansatz into (1.35) yields only that we must have $a_{[0,0,1]} + y_0 a_{[1,0,0]} = 0$ which can be satisfied in many ways. Using (1.37) we get that in fact $a_{[0,0,1]} = a_{[1,0,0]} = 0$ for any power series solution of $\mathcal{R}_1$. $\triangleleft$

As this example demonstrates, one should always remember that integrability conditions are not additional restrictions of the solution space but that any solution of the original system *automatically* satisfies them.

Definition 1.4.2. *The differential equation* $\mathcal{R}_q$ *is called* formally integrable*, if for all* $r \geq 0$ *the equality* $\mathcal{R}^{(1)}_{q+r} = \mathcal{R}_{q+r}$ *holds.*

Thus an equation is formally integrable, if at no order of prolongation integrability conditions arise. For such equations the above method for the construction of formal power series solutions works without problems which explains the name "formal integrability". Furthermore, we can be sure that we do not overlook any hidden conditions on the lower order coefficients, if we build only a truncated series. In Example 1.3.4, the equation $\mathcal{R}_1$ is obviously not formally integrable, but one easily shows that $\mathcal{R}^{(1)}_1$ is.

Example 1.4.3. In Remark 1.3.2 we introduced differential equations of finite type and saw that we may interpret them as connections. Now we exploit this observation for the integration of such equations. For simplicity of notation, we restrict to first-order equations $\mathcal{R}_1 \subset J_1\mathcal{E}$ which are the image of a section $\gamma : \mathcal{E} \to J_1\mathcal{E}$. Locally, we choose a representation of the form

$$\mathcal{R}_1 \ : \{\ p^\alpha_i = \phi^\alpha_i(\mathbf{x}, \mathbf{u})\ , \qquad 1 \leq i \leq n\ . \tag{1.43}$$

A natural question is when such an equation is formally integrable. As (1.43) contains no equations of lower order, the only possibility for integrability conditions are cross-derivatives

$$0 = p^\alpha_{ij} - p^\alpha_{ji} = D_i\phi^\alpha_j - D_j\phi^\alpha_i = \Delta^\alpha_{ij}\ . \tag{1.44}$$

Evaluation of these on the submanifold $\mathcal{R}_1$ yields the integrability conditions

$$\Delta^\alpha_{ij}\Big|_{\mathcal{R}_1} = \frac{\partial\phi^\alpha_j}{\partial x^i} + \frac{\partial\phi^\alpha_j}{\partial u^\beta}\phi^\beta_i - \frac{\partial\phi^\alpha_i}{\partial x^j} + \frac{\partial\phi^\alpha_i}{\partial u^\beta}\phi^\beta_j = 0\ . \tag{1.45}$$

Obviously, the satisfaction of these equations is a necessary condition for the formal integrability of $\mathcal{R}_1$. In our particular case of a maximally overdetermined equation, it is not difficult to see that this is also a sufficient condition: any cross-derivative in a prolonged equation $\mathcal{R}_{1+r}$ is a prolongation of one of the $\Delta^\alpha_{ij}|_{\mathcal{R}_1}$. We do not show this here in detail, as we will prove a much more general result in Section 3.2.

A more geometric way to view these integrability conditions starts with the introduction of the vector fields $X_i \in \mathfrak{X}(\mathcal{E})$ locally defined by

$$X_i\Big|_{(\mathbf{x},\mathbf{u})} = \partial_{x^i} + \phi^\alpha_i(\mathbf{x}, \mathbf{u})\partial_{u^\alpha} \qquad 1 \leq i \leq n\ . \tag{1.46}$$

They span the horizontal distribution of the connection on $\mathcal{E}$ defined by the section γ. Indeed, if we interpret as in Section 1.2 the fibre component of each point $(\mathbf{x}, \mathbf{u}, \mathbf{p}) \in \mathcal{R}_1$ as a linear map $T_{\mathbf{x}}\mathcal{X} \to T_{(\mathbf{x},\mathbf{u})}\mathcal{E}$, then in our local coordinates its matrix is given by stacking the $m \times n$ matrix defined by ϕ_i^α over the $n \times n$ identity matrix and the horizontal lift of the basis vector ∂_{x^i} yields the vector X_i.

The integrability conditions (1.45) are equivalent to the vanishing of all Lie brackets $[X_i, X_j]$. In other words, $\mathcal{R}_1$ is formally integrable, if and only if the section γ defines a *flat* connection with an involutive horizontal distribution. By the Frobenius Theorem (see Appendix C.3), any involutive distribution is integrable and the n-dimensional horizontal distribution $\mathcal{H} = \langle X_1, \dots, X_n\rangle \subset T\mathcal{E}$ possesses n-dimensional integral manifolds $\mathcal{N} \subset \mathcal{E}$.

We claim that these integral manifolds are images of solutions of $\mathcal{R}_1$. Indeed, let $\sigma : \mathcal{U} \subseteq \mathcal{X} \to \mathcal{E}$ be a (local) solution defined in some open set $\mathcal{U}$. In our case this implies that $\gamma(\sigma(x)) = j_1\sigma(x)$ for all $x \in \mathcal{U}$. According to our intrinsic approach to jet bundles, we can write $j_1\sigma(x) = \big(\sigma(x), T_x\sigma\big)$ and obtain that $\operatorname{im} T_x\sigma = \mathcal{H}_{\sigma(x)}$. But this equality entails that $\operatorname{im}\sigma$ is an integral manifold of the horizontal distribution $\mathcal{H}$.

Conversely, let $\mathcal{N} \subset \mathcal{E}$ be an n-dimensional integral manifold of $\mathcal{H}$. Obviously, it is transversal to the fibration $\pi : \mathcal{E} \to \mathcal{X}$ and hence the image of a (local) section $\sigma : \mathcal{U} \subseteq \mathcal{X} \to \mathcal{E}$. By definition of an integral manifold, this implies that $\operatorname{im} T_x\sigma = \mathcal{H}_{\sigma(x)}$ or, equivalently, $\gamma(\sigma(x)) = j_1\sigma(x)$ for all $x \in \mathcal{U}$. Thus $\operatorname{im} j_1\sigma \subset \mathcal{R}_1$ and σ is a solution.

This result has also practical implications: the concrete determination of integral manifolds of an involutive distribution requires only the solution of *ordinary* differential equations, although we are dealing with a partial differential equation. This was already shown by Lie (see [107]) interpreting the vector fields X_i as generators of Lie point symmetries [129] of $\mathcal{R}_1$. Indeed, let an n-dimensional integral manifold $\mathcal{N}$ of $\mathcal{H}$ be locally described by the m equations $\psi^\alpha(\mathbf{x}, \mathbf{u}) = 0$. As the vector fields X_i span the tangent space of $\mathcal{N}$, we must have $d\psi^\alpha(X_i) = X_i\psi^\alpha = 0$ (cf. Appendix C.3). Thus it suffices to determine all solutions of the overdetermined system $X_i v = 0$ for one unknown $v(\mathbf{x}, \mathbf{u})$. Such systems are studied in Example 3.2.5 where it is shown that their solution requires only the integration of ordinary differential equations. $\triangleleft$

1.5 (Geometric) Symbol and Principal Symbol

In our introduction of jet bundles we have stressed several times the fact that $J_q\mathcal{E}$ is an affine bundle over $J_{q-1}\mathcal{E}$ modelled on the vector space $S_q(T^*\mathcal{X}) \otimes V\mathcal{E}$; in fact our intrinsic approach to jet bundles was built on it. The reason for this emphasis becomes apparent now. It allows us to introduce a very useful geometric object, namely the symbol of a differential equation. Many properties of a differential equation can be read off its symbol and it effectively reduces some problems to questions in linear algebra.

Definition 1.5.1. *Let $\mathcal{R}_q \subseteq J_q\mathcal{E}$ be a differential equation. The* (geometric) symbol $\mathcal{M}_q$ *of $\mathcal{R}_q$ is the family of vector spaces over $\mathcal{R}_q$ with the vector space at the point $\rho \in \mathcal{R}_q$ given by*

$$(\mathcal{M}_q)_\rho = V_\rho^{(q)}\mathcal{R}_q \subseteq V_\rho^{(q)}J_q\mathcal{E}\,. \tag{1.47}$$

Here $V_\rho^{(q)}J_q\mathcal{E} = \ker T_\rho\pi^q_{q-1}$ is the vertical bundle with respect to the fibration $\pi^q_{q-1} : J_q\mathcal{E} \to J_{q-1}\mathcal{E}$.

Thus the geometric symbol is the vertical part of the tangent space of $\mathcal{R}_q$ with respect to the fibration π^q_{q-1} (here we identify $T\mathcal{R}_q$ with a subspace of $TJ_q\mathcal{E}$). But as mentioned above, for this fibration the vertical part is isomorphic to $S_q(T^*\mathcal{X}) \underset{J_{q-1}\mathcal{E}}{\otimes} V\mathcal{E}$, so that we may consider $\mathcal{M}_q$ as a subspace of the latter vector bundle. If the differential equation $\mathcal{R}_q$ is globally described by the map $\Phi : J_q\mathcal{E} \to \mathcal{E}'$ with a vector bundle $\pi' : \mathcal{E}' \to \mathcal{X}$, then we introduce the *symbol map* $\sigma : S_q(T^*\mathcal{X}) \underset{J_{q-1}\mathcal{E}}{\otimes} V\mathcal{E} \to T\mathcal{E}'$ given by $\sigma = T\Phi\big|_{V^{(q)}J_q\mathcal{E}}$ and define $\mathcal{M}_q = \ker\sigma$.

Locally, this leads to the following picture. Let $(\mathbf{x}, \mathbf{p})$ be local coordinates on $J_q\mathcal{E}$ in a neighbourhood of ρ. We must first determine $T_\rho\mathcal{R}_q$ as a subspace of $T_\rho J_q\mathcal{E}$. Let $(\mathbf{x}, \mathbf{p}; \dot{\mathbf{x}}, \dot{\mathbf{p}})$ be the induced coordinates on $T_\rho J_q\mathcal{E}$; then every vector $X \in T_\rho J_q\mathcal{E}$ can be written

$$X = \dot{x}^i\partial_{x^i} + \dot{p}^\alpha_\mu\partial_{p^\alpha_\mu}\,. \tag{1.48}$$

Assuming that a local representation of the submanifold $\mathcal{R}_q$ is given by the equations $\Phi^\tau(\mathbf{x}, \mathbf{p}) = 0$ with $\tau = 1, \dots, p$, its tangent space $T_\rho\mathcal{R}_q$ consists of all vectors $X \in T_\rho J_q\mathcal{E}$ such that $d\Phi^\tau(X) = 0$ (see Remark C.3.5) which yields the linear system

$$\sum_{i=1}^{n} \frac{\partial\Phi^\tau}{\partial x^i}(\mathbf{x}, \mathbf{p})\,\dot{x}^i + \sum_{\substack{1\leq\alpha\leq m\\ 0\leq|\mu|\leq q}} \frac{\partial\Phi^\tau}{\partial p^\alpha_\mu}(\mathbf{x}, \mathbf{p})\,\dot{p}^\alpha_\mu = 0\,. \tag{1.49}$$

The symbol is by definition the vertical part of this tangent space. This means that we are only interested in those solutions of (1.49) where $\dot{\mathbf{x}} = 0$ and $\dot{p}^\alpha_\mu = 0$ whenever $|\mu| < q$. Thus $\mathcal{M}_q$ can be locally described as the solution space of the following system of linear equations

$$(\mathcal{M}_q)_\rho : \left\{ \sum_{\substack{1\leq\alpha\leq m\\ |\mu|=q}} \frac{\partial\Phi^\tau}{\partial p^\alpha_\mu}(\mathbf{x}, \mathbf{p})\,\dot{p}^\alpha_\mu = 0\,, \quad \tau = 1, \dots, p\,. \right. \tag{1.50}$$

This is a system with real coefficients, as the derivatives $\partial\Phi^\tau/\partial p^\alpha_\mu$ are evaluated at the point $\rho \in \mathcal{R}_q$. If there are n independent and m dependent variables, the coefficient matrix has p rows (labelled by τ) and $m\binom{n+q-1}{n-1}$ columns (labelled by α and μ). We call it *symbol matrix* denoted by $M_q(\rho)$. It is also the matrix of the symbol map σ in local coordinates.

Obviously, $\dim(\mathcal{M}_q)_\rho$ (or equivalently $\operatorname{rank} M_q(\rho)$) might vary with ρ. For this reason, we speak in Definition 1.5.1 only of a family of vector spaces and not of a vector bundle. Only if the dimension remains constant over $\mathcal{R}_q$, the symbol $\mathcal{M}_q$ is a vector subbundle of $S_q(T^*\mathcal{X}) \underset{J_{q-1}\mathcal{E}}{\otimes} V\mathcal{E}$. In the sequel we will always assume that this is the case if not explicitly stated otherwise. In particular, $\mathcal{M}_q$ is a vector bundle, if $\mathcal{R}_q$ is globally defined as zero set of a fibred map $\Phi : J_q\mathcal{E} \to \mathcal{E}'$.

The symbol is most easily understood for linear differential equations. Loosely speaking, the geometric symbol is then simply the highest order or principal part of the system. For non-linear systems we perform a brute force linearisation at the point ρ in order to obtain $(\mathcal{M}_q)_\rho$.

Remark 1.5.2. A particular class of differential equations are those where the dimension of the symbol is zero. These equations are called *maximally overdetermined* or of *finite type*. The latter name stems from the fact that, as we will see later, the formal solution space of such an equation is always finite-dimensional. $\lhd$

If we are interested in the symbol $\mathcal{M}_{q+r} \subseteq S_{q+r}(T^*\mathcal{X}) \underset{J_{q+r-1}\mathcal{E}}{\otimes} V\mathcal{E}$ of the prolonged equation $\mathcal{R}_{q+r} \subseteq J_{q+r}\mathcal{E}$, there is no need to explicitly compute a local representation of $\mathcal{R}_{q+r}$ in order to obtain local equations for its symbol. We can directly derive them from a local representation of $\mathcal{R}_q$. This follows easily from the coordinate expression (1.11) of the formal derivative.

By (1.50) and (1.32), we need only the partial derivatives $\partial D_i\Phi^\tau/\partial p^\alpha_\nu$ with $1 \leq i \leq n$ and $|\nu| = q+1$, i.e. the highest-order part of the formal derivative $D_i\Phi^\tau$, for determining the symbol $\mathcal{M}_{q+1}$. According to (1.11) it is given by $\partial D_i\Phi^\tau/\partial p^\alpha_\nu = \partial\Phi^\tau/\partial p^\alpha_{\nu-1_i}$ (if $\nu_i = 0$, the derivative vanishes). Thus a local representation of the first prolongation $\mathcal{M}_{q+1}$ is

$$(\mathcal{M}_{q+1})_\rho : \left\{ \sum_{\substack{1\leq\alpha\leq m \\ 0\leq|\mu|\leq q}} \frac{\partial\Phi^\tau}{\partial p^\alpha_\mu}(\mathbf{x},\mathbf{p})\, \dot{p}^\alpha_{\mu+1_i} = 0 \,, \quad \begin{array}{l} \tau = 1,\dots,p\,, \\ i = 1,\dots,n\,. \end{array} \right. \tag{1.51}$$

At first sight one might be tempted to think that the linear systems (1.50) and (1.51) have the same coefficient matrices. This is of course not true. The matrix M_{q+1} of (1.51) has np rows and $m\binom{n+q}{n-1}$ columns and is thus much larger. Because of the particular way in which the index i appears in (1.51), one may say that the matrix M_{q+1} emerges from M_q by a kind of combinatorial blowing up. For each row in M_q we obtain n rows in M_{q+1} and similarly for the columns. These n rows contain exactly the same entries but distributed in different patterns.

Example 1.5.3. We consider a scalar second-order equation in two independent variables x, y with the local representation

$$\mathcal{R}_2 : \{ \Phi(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy}) = 0 \ . \tag{1.52}$$

If we define the three functions $a(\rho) = \partial\Phi/\partial u_{xx}(\rho)$, $b(\rho) = \partial\Phi/\partial u_{xy}(\rho)$ and $c(\rho) = \partial\Phi/\partial u_{yy}(\rho)$, each $\mathcal{R}_2 \to \mathbb{R}$, then the matrix of the symbol $\mathcal{M}_2$ at a point $\rho \in \mathcal{R}_2$ contains only one row

$$M_2(\rho) = \begin{pmatrix} a(\rho) & b(\rho) & c(\rho) \end{pmatrix} \ . \tag{1.53}$$

If no point $\rho \in \mathcal{R}_2$ exists where all three functions a, b, c vanish simultaneously, $\mathcal{M}_2$ is a two-dimensional vector bundle.

For the prolonged symbol $\mathcal{M}_3$ at a point $\rho' \in \mathcal{R}_3$ such that $\pi^3_2(\rho') = \rho$ we obtain the matrix

$$M_3(\rho') = \begin{pmatrix} a(\rho) & b(\rho) & c(\rho) & 0 \\ 0 & a(\rho) & b(\rho) & c(\rho) \end{pmatrix} \ . \tag{1.54}$$

Here the columns correspond to the unknowns $\dot{u}_{xxx}$, $\dot{u}_{xxy}$, $\dot{u}_{xyy}$ and $\dot{u}_{yyy}$. If $\mathcal{M}_2$ is a vector bundle, then $\mathcal{M}_3$ is too. $\triangleleft$

The considerations above are straightforwardly extended by iteration to higher prolongations and we obtain as local representation

$$(\mathcal{M}_{q+r})_\rho : \left\{ \sum_{\substack{1\leq\alpha\leq m \\ 0\leq|\mu|\leq q}} \frac{\partial\Phi^\tau}{\partial p^\alpha_\mu}(\mathbf{x}, \mathbf{p})\, \dot{p}^\alpha_{\mu+\nu} = 0 \ , \quad \begin{array}{l} \tau = 1, \dots, p \ , \\ |\nu| = r \ . \end{array} \right. \tag{1.55}$$

The size of the corresponding symbol matrices M_{q+r} is growing rapidly. Because of (1.24), one easily calculates that M_{q+r} has $p\binom{n+r-1}{n-1}$ rows and $m\binom{n+q+r-1}{n-1}$ columns.

The matrix of the symbol and in particular of its prolongations indirectly already appeared in Section 1.4 during the construction of a formal power series solution (1.39). The rank of the symbol matrix M_q tells us how many coefficients of order q are determined by (1.40) (nothing is said about the coefficients of lower order!). The matrices M_{q+r} of the prolonged symbols appear, if we interpret (1.42) as an inhomogeneous linear system for the coefficients of order $q+r$: its matrix is precisely M_{q+r}.

In Section 1.4 we saw the this simple algorithm for the construction of formal power series solutions works only, if we are dealing with a formally integrable equation. The symbol aids us in deciding whether or not a given equation is formally integrable. It indicates whether during the prolongation of a differential equation integrability conditions may have occurred and it allows us to construct them using only linear algebra. The proof of the next theorem shows this explicitly.

Theorem 1.5.4. *If the symbol $\mathcal{M}_{q+1}$ is a vector bundle, then*

$$\dim \mathcal{R}_q^{(1)} = \dim \mathcal{R}_{q+1} - \dim \mathcal{M}_{q+1}\,. \tag{1.56}$$

Proof. The dimension of a submanifold is determined by counting how many independent equations are needed to describe it locally, i. e. by the rank of a Jacobian. We assume that $\mathcal{R}_q$ has the local representation $\Phi^\tau(\mathbf{x}, \mathbf{p}) = 0$. The part of the Jacobian of $\mathcal{R}_{q+1}$ relevant for the determination of the fibre dimension can be divided into four blocks:

$$\left(\begin{array}{|c|c|}\hline \dfrac{\partial D_i\Phi^\tau}{\partial p^\alpha_\mu}\,,\ |\mu| = q+1 & \dfrac{\partial D_i\Phi^\tau}{\partial p^\alpha_\mu}\,,\ 0 \le |\mu| \le q \\ \hline 0 & \dfrac{\partial \Phi^\tau}{\partial p^\alpha_\mu}\,,\ 0 \le |\mu| \le q \\ \hline \end{array}\right)\,. \tag{1.57}$$

The lower part stems from the local representation of $\mathcal{R}_q$; its rank is $\operatorname{codim}\mathcal{R}_q$. The upper part comes from the prolonged equation; the left block is the symbol matrix M_{q+1}. We must distinguish two cases. If M_{q+1} has maximal rank, then $\dim\mathcal{R}_{q+1} = \dim\mathcal{R}_q + \dim\mathcal{M}_{q+1}$ and thus $\mathcal{R}_q^{(1)} = \mathcal{R}_q$.

If M_{q+1} has not maximal rank, one can construct by linear operations a row with only zeros in the left block. The question is how this row looks like in the right block. If it is independent of the rows in the lower part, we have found an integrability condition. It is constructed by performing exactly the same linear operations on the full equations $D_i\Phi^\tau = 0$. Otherwise, we have only obtained a redundant description of $\mathcal{R}_{q+1}$. This leads to *identities* or *compatibility conditions*, if we add a right hand side to the differential equation. Comparing ranks yields the claim for one point, but since $\mathcal{M}_{q+1}$ is a vector bundle, it holds everywhere. □

In the classical theory of partial differential equations a different notion of symbol is used which should not be confused with the geometric symbol introduced above: the classical symbol is *not* an intrinsic object. Our notion of symbol is closely related to what is traditionally called the principal symbol which is an intrinsically defined object.

Assume we are given a one-form $\chi \in T^*\mathcal{X}$. It induces for every $q > 0$ a map $\iota_{\chi,q} : V\mathcal{E} \to S_q(T^*\mathcal{X}) \otimes V\mathcal{E}$ defined by $\iota_{\chi,q}(\nu) = \chi^q \otimes \nu$ where χ^q denotes the q-fold symmetric product of χ. In local coordinates, we write $\chi = \chi_i dx^i$ and obtain $\iota_{\chi,q} : \nu^\alpha \partial_{u^\alpha} \mapsto \chi_\mu \nu^\alpha dx^\mu \otimes \partial_{u^\alpha}$ where μ runs over all multi indices of length q and $\chi_\mu = \chi_1^{\mu_1} \cdots \chi_n^{\mu_n}$. Now we may define the principal symbol in terms of the symbol map σ.

Definition 1.5.5. *Let $\chi \in T^*\mathcal{X}$ be a one-form over $\mathcal{X}$ and $\mathcal{R}_q \subseteq J_q\mathcal{E}$ a differential equation globally described by the map $\Phi : J_q\mathcal{E} \to \mathcal{E}'$. The* principal symbol *of $\mathcal{R}_q$ is the linear map $\tau_\chi : V\mathcal{E} \to T\mathcal{E}'$ defined by $\tau_\chi = \sigma \circ \iota_{\chi,q}$.*

The one-form χ plays the role of a parameter in the principal symbol. For many purposes, it is of great importance how the properties of the linear map τ_χ change with χ. Locally, we can associate a matrix $T[\chi]$ with τ_χ:

$$T^\tau_\alpha[\chi] = \sum_{|\mu|=q} \frac{\partial \Phi^\tau}{\partial p^\alpha_\mu} \chi^\mu \, . \tag{1.58}$$

If $\dim \mathcal{E} = m$ and $\dim \mathcal{E}' = p$, it has p rows and m columns. Its entries are homogeneous polynomials of degree q in the coefficients of χ. We may think of $T[\chi]$ as a kind of contraction of the symbol matrix M_q of $\mathcal{R}_q$. Both matrices have the same number of rows. The column with index α of $T[\chi]$ is a linear combination of all columns in M_q corresponding to a variable $\dot{p}^\alpha_\mu$ with the coefficients given by χ_μ.

Example 1.5.6. Consider again the equation $\mathcal{R}_2$ described by (1.52). Its symbol matrix is the 1×3 matrix shown in (1.53). The matrix of its principal symbol is 1×1 with the single entry $a(\rho)\chi_x^2 + b(\rho)\chi_x\chi_y + c(\rho)\chi_y^2$. $\lhd$

Definition 1.5.7. *Let $\mathcal{R}_q$ be an involutive differential equation. We call a non-zero one-form $\chi \in T^*\mathcal{X}$* characteristic *at a point $\rho \in \mathcal{R}_q$, if the principal symbol τ_χ is not injective at ρ. The set of all characteristic one-forms at ρ is the* characteristic variety *C_ρ.*

C_ρ is indeed a variety in the strict algebraic meaning of the word. The condition that τ_χ is not injective at a point ρ can be expressed as the vanishing of certain subdeterminants of the matrix $T[\chi]$, i. e. as the vanishing of (homogeneous) polynomials in the coefficients χ_i. For *semi-linear* equations, where the symbol space $(\mathcal{M}_q)_\rho$ depends only on $x = \pi^q(\rho) \in \mathcal{X}$, one can define characteristic one-forms over points $x \in \mathcal{X}$.

Example 1.5.8. In introductory textbooks on differential equations one usually does not consider characteristic one-forms but *characteristic surfaces*. We show now for linear first-order systems the relation between the two approaches. Consider

$$\mathcal{R}_1 : \left\{ \sum_{\alpha=1}^{m} \sum_{i=1}^{n} A_\alpha^{\tau i}(\mathbf{x}) p_i^\alpha + \sum_{\alpha=1}^{m} B_\alpha^\tau(\mathbf{x}) + C^\tau(x) = 0 \;, \qquad 1 \leq \tau \leq p \right. \tag{1.59}$$

and the $(n-1)$-dimensional hypersurface $\Sigma \subset \mathcal{X}$ described by the equation

$$\Sigma = \left\{ x \in \mathcal{X} \mid x^n = \phi(x^1, \dots, x^{n-1}) \right\} \,. \tag{1.60}$$

Such a surface is called characteristic, if prescribing the value of $\mathbf{u}$ on Σ does not suffice to compute all first-order derivatives on Σ. Obviously, this condition concerns only the derivatives $(\partial\mathbf{u}/\partial x^n)\big|_\Sigma$; all other ones follow easily from our "initial values".

Thus assume that $\mathbf{u}\big|_\Sigma = \mathbf{f}(x^1, \dots, x^{n-1})$. Using the chain rule we obtain

$$\left.\frac{\partial\mathbf{u}}{\partial x^i}\right|_\Sigma + \frac{\partial\phi}{\partial x^i}\left.\frac{\partial\mathbf{u}}{\partial x^n}\right|_\Sigma = \frac{\partial\mathbf{f}}{\partial x^i} \;, \qquad 1 \leq i < n \,. \tag{1.61}$$

Substituting these relations into (1.59) and restricting to Σ yields a linear system for the derivatives $(\partial\mathbf{u}/\partial x^n)\big|_\Sigma$ whose matrix T is

$$T_\alpha^\tau = A_\alpha^{\tau n} - \sum_{i=1}^{n-1} A_\alpha^{\tau i} \frac{\partial\phi}{\partial x^i} \,. \tag{1.62}$$

If $\operatorname{rank} T < m$, i.e. the associated linear map is not injective, then it is *not* possible to determine all derivatives on the hypersurface Σ and it is characteristic. The matrix of the principal symbol of (1.59) is $T_\alpha^\tau[\chi] = A_\alpha^{\tau i}\chi_i$ which coincides with (1.62) for the one-form $\chi = dx^n - (\partial\phi/\partial x^i)dx^i$. Hence the hypersurface Σ is characteristic, if and only if this one-form is characteristic. More abstractly, we may say that the conormal space $N^*\Sigma$ (see Appendix C.3) is spanned by the one-form χ. ◁

1.6 Some Examples

In this section we briefly discuss some classical differential equations which will be used as examples at many places later on. They stem from physics and are of great importance in applications. Most of them are already formally integrable, however, we have not yet the necessary tools to prove this.

Example 1.6.1. Maxwell's equations form the foundations of electrodynamics. They describe the evolution of the electric and magnetic fields and fluxes. Their most general (differential) form is

$$\mathbf{D}_t = \nabla \times \mathbf{H} - \mathbf{J}_e \,, \qquad \mathbf{B}_t = -\nabla \times \mathbf{E} - \mathbf{J}_m \,, \tag{1.63a}$$

$$0 = \nabla \cdot \mathbf{D} - \rho_e \,, \qquad 0 = \nabla \cdot \mathbf{B} - \rho_m \,. \tag{1.63b}$$

Here ∇ denotes the gradient with respect to the spatial variables only; $\cdot$ and $\times$ represent the scalar and the vector product, respectively. $\mathbf{E}$ and $\mathbf{H}$ are the electric and magnetic field, respectively, $\mathbf{D}$ and $\mathbf{B}$ the corresponding fluxes, $\mathbf{J}_e$ and $\mathbf{J}_m$ are currents and finally, ρ_e and ρ_m charge densities. In physics, $\mathbf{J}_m$ and ρ_m are usually set to zero, as nobody has ever observed experimentally magnetic charges or currents. But there is no mathematical reason for not adding such terms (at the level of (1.64); in gauge theory the situation changes).

(1.63) is yet underdetermined. In a simple medium, we have the constitutive relations $\mathbf{D} = \epsilon\mathbf{E}$, $\mathbf{B} = \mu\mathbf{H}$ and $\mathbf{J}_e = \sigma\mathbf{E}$ where ϵ, μ and σ are the permittivity, the permeability and the conductivity, respectively. If the medium is isotropic and homogeneous, these quantities are constant scalars. In general, they are tensor valued functions of space-time. We mainly study Maxwell's equations in vacuum. There $\sigma = 0$ and $\epsilon\mu = 1/c$ where c is the speed of light. In natural units $c = 1$ and we obtain the following system:

$$\mathbf{E}_t = \nabla \times \mathbf{B} \,, \qquad \mathbf{B}_t = -\nabla \times \mathbf{E} \,, \tag{1.64a}$$

$$0 = \nabla \cdot \mathbf{E} \,, \qquad 0 = \nabla \cdot \mathbf{B} \,. \tag{1.64b}$$

The equations in (1.64a) describe the time evolution of the fields. The equations in (1.64b) are often called the *Gauß* or *divergence laws*; they render (1.64) overdetermined, as we have got eight equations for six field components. Because of the well-known identity $\nabla \cdot (\nabla \times \mathbf{X}) = 0$ for any vector field $\mathbf{X}$, we obtain no integrability conditions, if we first prolong and then project back (1.64). Note that the same is true for the general form (1.63), if we ignore the currents and charges.

This identity also implies another interesting property. Let us take only the evolutionary part (1.64a); it forms a hyperbolic (in t-direction) system in Cauchy–Kovalevskaya form. So we may consider the initial value problem for it. Assume that we prescribe for $t = 0$ fields $\mathbf{E}_0(\mathbf{x})$ and $\mathbf{B}_0(\mathbf{x})$ satisfying the divergence laws (1.64b). Then the solution $\mathbf{E}(\mathbf{x}, t)$, $\mathbf{B}(\mathbf{x}, t)$ of (1.64a) for these initial data satisfy (1.64b) for all times t, i. e. it is a solution of the full Maxwell equations (1.64). In Section 5.4 we will see that this is no accident but in fact a characteristic and very important property of involutive equations.

If we include the currents J_e and J_m and the charge densities ρ_e and ρ_m, we see that they must satisfy the *continuity equations*

$$(\rho_e)_t + \nabla \cdot J_e = 0 , \qquad (\rho_m)_t + \nabla \cdot J_m = 0 . \tag{1.65}$$

Otherwise the equations become inconsistent and do not admit any solutions. These are examples of *compatibility conditions.* ◁

Example 1.6.2. The fundamental equations governing the flow of fluids and gas are the *Navier–Stokes equations.* We are particularly interested in *incompressible flows.* The flow lives in some domain $\Omega \subseteq \mathbb{R}^3$; for simplicity we assume it to be open. Our base manifold $\mathcal{X}$ is then given by the product $\Omega \times \mathbb{R}$ where the factor $\mathbb{R}$ represents time. The dependent variables are the velocity of the flow and the pressure. Thus we may use as total space $\mathcal{E} = T\Omega \times \mathbb{R}^2$ with local coordinates $(\mathbf{x}, t, \mathbf{u}, p)$: $\mathbf{x}$ represent a point in Ω, t the time, $\mathbf{u}$ are coordinates of a tangent vector in $T_{\mathbf{x}}\Omega$ and p is the pressure.

The incompressible Navier–Stokes equations are

$$\mathbf{u}_t + (\mathbf{u} \cdot \nabla)\mathbf{u} = \nu\Delta\mathbf{u} - \nabla p , \tag{1.66a}$$

$$\nabla \cdot \mathbf{u} = 0 . \tag{1.66b}$$

Here we use again notations from vector analysis: ∇ is the gradient with respect to the spatial variables $\mathbf{x}$ and the scalar product $\nabla \cdot \mathbf{u}$ corresponds to the divergence of $\mathbf{u}$. The left hand side of (1.66a) represents the convective derivative of $\mathbf{u}$ expressing that our coordinate system moves with the flow; in components it is given by $\partial u^i/\partial t + u^j \partial u^i/\partial x^j$.

The coefficient ν is the *viscosity* of the flow. If it is zero, we obtain *Euler's equations* of inviscid flow. They can be described in a much more geometric way as a flow on the group of volume preserving diffeomorphisms [9]. With respect to the formal analysis, the Euler and the Navier–Stokes equations behave very similar.

(1.66) describes a submanifold $\mathcal{R}_2 \subset J_2\mathcal{E}$. As the incompressibility constraint (1.66b) is of first order, $\mathcal{R}_2$ is *not* formally integrable; trivial integrability conditions are given by the prolongations of (1.66b) to second order. Thus $\mathcal{R}_2^{(1)}$ is described by (1.66) plus the equations

$$\nabla \cdot \mathbf{u}_t = 0 , \qquad \nabla(\nabla \cdot \mathbf{u}) = 0 . \tag{1.67}$$

$\mathcal{R}_2^{(1)}$ is still not formally integrable, as we find in the next prolongation a non-trivial integrability condition: if we take the divergence of (1.66a), we can eliminate the third-order term $\Delta(\nabla \cdot \mathbf{u})$ by the divergence of the second equation in (1.67). There remains the second-order equation

$$\Delta p = -\nabla \cdot \big((\mathbf{u} \cdot \nabla)\mathbf{u}\big) , \tag{1.68}$$

which may be interpreted as a Poisson equation for the pressure. It plays an important role in the numerical integration of the Navier–Stokes equations. One can check that the equation $\mathcal{R}_2^{(2)}$ defined by (1.66), (1.67), and (1.68) is formally integrable. $\triangleleft$

The following two examples are actually of a very geometric nature. Nevertheless, we describe them only in local coordinates. There are two reasons for this. First of all, we do not want to introduce all the necessary differential geometry. Secondly, we will take at many places a more computationally point of view. In order to be able to demonstrate many of our computational techniques with the help of these examples, we must write them down in local coordinates.

Example 1.6.3. In modern elementary particle physics three of the four fundamental interactions are described by *Yang–Mills theories.* The proper framework for them are principal fibre bundles, but we take here a simplified approach. Underlying the theory is the space-time, an n-dimensional manifold equipped with a Lorentzian metric. Again for simplicity we restrict here to the case of a flat space-time with the Minkowski metric η^{ij} defined by $\eta^{ii} = 1$ for $1 \leq i < n$, $\eta^{nn} = -1$ and $\eta^{ij} = 0$ in all other cases.

Assume we are given a d-dimensional Lie group $\mathcal{G}$. The basic field of the corresponding Yang–Mills theory, the *vector potential*, is a one-form A over space-time taking values in the Lie algebra $\mathfrak{g}$ of $\mathcal{G}$. If a basis of $\mathfrak{g}$ is given by the elements T_a with $1 \leq a \leq d$, such a one-form has the local form $A = A_i^a T_a dx^i$. Thus we are dealing with a fibre bundle over space-time whose fibre dimension is dn.

Let the structure constants of the Lie algebra $\mathfrak{g}$ be C^c_{ab}, i.e. we have $[T_a, T_b] = C^c_{ab} T_c$. Then the exterior derivative of A yields the *field strength* two-form $F = dA = F^a_{ij} T_a dx^i \wedge dx^j$ where the coefficients are given by

$$F^a_{ij} = \partial_{x^i} A^a_j - \partial_{x^j} A^a_i + C^a_{bc} A^b_i A^c_j \, . \tag{1.69}$$

They allow us to write the field equations in a concise form:

$$\Phi^a_\ell \left(A^b_i, \partial_{x^j} A^b_i, \partial_{x^j x^k} A^b_i\right) = \eta^{ij} \left(\partial_{x^i} F^a_{j\ell} + C^a_{bc} A^b_i F^c_{j\ell}\right) = 0 \, . \tag{1.70}$$

This is a semi-linear system of dn second-order equations. The non-linearity is introduced by the third term in the definition of the field strengths which is quadratic in the vector potentials. It disappears, if the Lie algebra $\mathfrak{g}$ is Abelian, as in this case all structure constants are zero.

In a rather lengthy computation one can show that the partial differential equation $\mathcal{R}_2$ defined locally by (1.70) is formally integrable. The key is the *Noether identity*

$$\eta^{ij}\left(D_i\Phi^a_j + C^a_{bc}A^b_i\Phi^c_j\right) \equiv 0 \tag{1.71}$$

stemming from the properties of the structure constants C^a_{bc} due to the Jacobi identity. It implies that $\mathcal{R}_2^{(1)} = \mathcal{R}_2$ and, as we will see later, this suffices to prove formal integrability.

The Yang–Mills equations for the Lie group $U(1)$ are closely related to Maxwell's equations. As we are dealing with a one-dimensional group, we skip the upper index on the coefficients of the potential A and the field strength F. The coefficients of latter form thus in four dimensions a 4×4 antisymmetric matrix related to the electric and magnetic field:

$$F = \begin{pmatrix} 0 & B^{(z)} & -B^{(y)} & -E^{(x)} \\ -B^{(z)} & 0 & B^{(x)} & -E^{(y)} \\ B^{(y)} & -B^{(x)} & 0 & -E^{(z)} \\ E^{(x)} & E^{(y)} & E^{(z)} & 0 \end{pmatrix} . \tag{1.72}$$

In other words, if we build the three-dimensional vector $\mathbf{A}$ consisting of the first three coefficients of A, then $\mathbf{B} = \nabla \times \mathbf{A}$ and $\mathbf{E} = -\partial_t\mathbf{A} - \nabla \cdot A_4$. Note that these relations automatically imply half of the equations in (1.64), namely $\nabla \cdot \mathbf{B} = 0$ and $\mathbf{B}_t = -\nabla \times \mathbf{E}$. The remaining four equations are equivalent to the four Yang–Mills equations (1.70). ◁

Example 1.6.4. General relativity describes the fourth fundamental interaction, gravity; a cornerstone of the theory are *Einstein's equations.* We consider only the vacuum case without any matter fields but in arbitrary dimensions.[2] Space-time is modelled as an n-dimensional Lorentzian manifold, i.e. as a manifold $\mathcal{M}$ equipped with a metric g_{ij} that has the same signature as the Minkowski metric η_{ij}. This manifold $\mathcal{M}$ is our base space $\mathcal{X}$. The dependent variables are the components g_{ij} of the metric. Thus $\mathcal{E}$ is the vector bundle of symmetric bilinear forms; its fibre dimension is $m = n(n+1)/2$.

The vacuum Einstein equations express the condition that the Ricci tensor R_{ij} of the Levi–Civita connection for the metric g_{ij} vanishes:

$$R_{ij}[g_{k\ell}] = 0 . \tag{1.73}$$

As the Ricci tensor is symmetric, this is a system of $n(n+1)/2$ equations so that we have as many equations as unknown functions. (1.73) defines a submanifold $\mathcal{R}_2 \subset J_2\mathcal{E}$, since the Ricci tensor depends on the second-order derivatives of the metric.

Writing out the components R_{ij} of the Ricci tensor in terms of the metric and its derivatives leads to rather complicated non-linear expressions which

[2] Strictly speaking, we assume that the dimension is at least four; for topological reasons two and three dimensions are special cases where different equations have to be used.

can be found in any textbook on either Riemannian geometry or general relativity. We omit them here and use instead a little trick for simplifying the coordinate form of (1.73): we take *locally geodesic coordinates*. Such coordinates exist for every point $x_0 \in \mathcal{M}$ and are characterised by two properties: (i) the metric becomes the Minkowski metric at the point x_0, $g_{ij}(x_0) = \eta_{ij}$, and (ii) the Christoffel symbols vanish at x_0. At x_0, (1.73) takes in these coordinates the following simple linear form

$$\begin{aligned} R_{ij}(x_0) = \frac{1}{2}\eta^{k\ell}\Big[&\partial_{x^i x^k} g_{j\ell}(x_0) + \partial_{x^j x^\ell} g_{ik}(x_0) - \\ &\partial_{x^i x^j} g_{k\ell}(x_0) - \partial_{x^k x^\ell} g_{ij}(x_0)\Big] = 0\ . \end{aligned} \tag{1.74}$$

Note that the defining properties of a locally geodesic coordinate system hold only *at the single point* x_0 and generally not even in some neighbourhood of it. Thus (1.74) should not be considered as a differential equation but only as a relation between the fibre coordinates of $J_2\mathcal{E}$ in the fibre over the point $x_0 \in \mathcal{M}$. In particular, we cannot expect that local representations of prolongations $\mathcal{R}_{2+r} \subset J_{2+r}\mathcal{E}$ are obtained by formally differentiating (1.74). However, this holds for the first prolongation of $\mathcal{R}_2$, as $\partial_{x^k} g_{ij}(x_0) = 0$ because of the metricity of the Levi–Civita connection. In further prolongations, lower-order terms appear. Fortunately, it turns out that $\mathcal{R}_3$ suffices for the formal analysis of the Einstein equations.

Because of the *contracted Bianchi identity* (a formal analogue to the Noether identity (1.71) of the Yang–Mills equations)

$$g^{ki}\partial_{x^k} R_{ij} - \partial_{x^j}(g^{ki}R_{ki}) = \Gamma^k_{ij} g^{i\ell} R_{k\ell} - \Gamma^\ell_{\ell k} g^{ik} R_{ij} \tag{1.75}$$

(here g^{ki} is the inverse to g_{ij}, i.e. $g^{ki}g_{ij} = \delta^k_j$), we find that $\mathcal{R}_2^{(1)} = \mathcal{R}_2$. We will see later that this suffices to prove the formal integrability of Einstein's equations. $\triangleleft$

1.7 Notes

Implicitly, jet bundles have been used for a long time. They were formally introduced by Ehresmann in a series of papers [44, 45] (see also [43]). They have become a standard tool for the geometric analysis of differential equations. By now, many approaches to jet bundles exist with various degrees of abstraction and sophistication. A standard reference is the book by Saunders [155]; useful introductions are also contained in [61, 129].

The pedestrian approach based on power series we presented in Section 1.1 is completely sufficient for all computational purposes. It can also be described in a coordinate free manner. This requires only an intrinsic formulation of the equivalence relation $\sim_q$. We say $\phi \sim_1 \psi$ at $x_0 \in \mathcal{X}$, if the tangent maps satisfy $T_{x_0}\phi = T_{x_0}\psi$. Then we extend recursively: $\phi \sim_q \psi$ at $x_0 \in \mathcal{X}$, if $T_{x_0}\phi \sim_{q-1} T_{x_0}\psi$ at all points $v \in T_{x_0}\mathcal{X}$. This formulation is detailed in [61].

Another approach is discussed in [132, Ch. IV]. It requires that $\mathcal{U}$ is a real vector space, which is the case in most applications. We define for any point $x_0 \in \mathcal{X}$ the ideal $\mathcal{I}_{x_0}(\mathcal{X})$ of all smooth functions $\phi : \mathcal{X} \to \mathbb{R}$ with $\phi(x_0) = 0$. Then we introduce the spaces

$$\mathcal{Z}^{(q)}_{x_0}(\mathcal{X},\mathcal{U}) = [\mathcal{I}_{x_0}(\mathcal{X})]^q \cdot C^\infty(\mathcal{X},\mathcal{U}) . \tag{1.76}$$

All functions in $\mathcal{Z}^{(q)}_{x_0}(\mathcal{X},\mathcal{U})$ possess at x_0 an at least q-fold zero and thus all their derivatives up to order $q-1$ vanish at this point. Finally, we identify q-jets at x_0 with the elements of the quotient space $C^\infty(\mathcal{X},\mathcal{U})/\mathcal{Z}^{(q+1)}_{x_0}(\mathcal{X},\mathcal{U})$.

The intrinsic approach to (first-order) jet bundle presented in Section 1.2 does not seem to be very popular. It is rarely found in the literature, although it has a number of advantages for studying the geometry of jet bundles. In recent times, it has gained some importance in the context of multisymplectic geometry, as the affine dual of $J_1\mathcal{E}$ carries a natural multisymplectic structure [62]. Our treatment of the contact structure is taken from Modugno [125].

We have seen that there are some analogies between our geometric view of differential equations and the situation in algebraic geometry. However, there are also some apparent differences. In particular, not every section contained in an equivalence class $[\sigma]^{(q)}_x \in \mathcal{R}_q$ is a solution of the differential equation. If one wants such a one-to-one relation between points and solutions, one must take *all* prolongations of $\mathcal{R}_q$ into account. This leads to the introduction of the jet bundle $J_\infty\mathcal{E}$ of infinite order a point of which defines a unique formal power series.

The proper definition of $J_\infty\mathcal{E}$ requires some mathematical care, as it is an infinite-dimensional space and it is not obvious how to equip it with a topological and a differentiable structure. This is achieved using the notions of projective and inductive limits in directed systems of sets [17]. As the jet bundles of different orders are related by the natural projections π^r_q, we may define

$$J_\infty\mathcal{E} = \operatorname*{proj\,lim}_{q\to\infty} J_q\mathcal{E} . \tag{1.77}$$

In order to introduce locally analytic functions on $J_\infty\mathcal{E}$ we exploit the natural inclusions $C^\omega(J_q\mathcal{E}) \subset C^\omega(J_r\mathcal{E})$ for $r > q$ and set

$$C^\omega(J_\infty\mathcal{E}) = \operatorname*{ind\,lim}_{q\to\infty} C^\omega(J_q\mathcal{E}) . \tag{1.78}$$

Note that the definition of an inductive limit implies that any function in $C^\omega(J_\infty\mathcal{E})$ is an element of at least one space $C^\omega(J_q\mathcal{E})$. Thus although these functions are defined on an infinite-dimensional manifold, they always depend only on a finite number of variables. In a similar, though slightly more complicated manner, one constructs the tangent and cotangent bundle of $J_\infty\mathcal{E}$ [17].

Then we introduce the infinite prolongation $\mathcal{R}_\infty \subseteq J_\infty\mathcal{E}$ of a differential equation. As each point on $\mathcal{R}_\infty$ represents a formal power series solution of the equation, the submanifold $\mathcal{R}_\infty$ may be identified with the formal solution space. Thus at the formal level we are in the same situation as in algebraic geometry. This point of view is much used by the school of Vinogradov (see e.g. [5, 38, 96, 181]) who calls such a submanifold $\mathcal{R}_\infty \subseteq J_\infty\mathcal{E}$ a *differential variety* or short *diffiety*. For our purposes we do not need this concept.

An interesting aspect of $J_\infty\mathcal{E}$ is that the contact codistribution $\mathcal{C}^0_\infty$ is involutive. In a jet bundle $J_q\mathcal{E}$ of finite order q, this can never be the case, as one can easily see from the basic contact forms ω^α_μ defined in (1.8). Indeed, if we take the exterior derivative of a form ω^α_μ, we obtain $d\omega^\alpha_\mu = -dp^\alpha_{\mu+1_i} \wedge dx^i$. If $|\mu| < q-1$, we may rewrite this as $d\omega^\alpha_\mu = -\omega^\alpha_{\mu+1_i} \wedge dx^i$ and the derivative vanishes modulo $\mathcal{C}^0_q$. But if $|\mu|$ takes the maximal value $q-1$, this is not possible and the exterior derivative does not vanish. Hence $\mathcal{C}^0_q$ is not involutive for any finite order q. In $\mathcal{C}^0_\infty$ no maximal value for $|\mu|$ exists and thus it is involutive.

Similarly, the contact distribution $\mathcal{C}_\infty$ is involutive. It consists of the n vector fields C_i given by (1.10a). Note that they are now defined by a series. Their application to a function yields its formal derivatives.

2. Involution I: Algebraic Theory

By the help of God and with His precious assistance, I say that Algebra is a scientific art.

Omar Khayyam

The first chapter was mainly concerned with the geometry behind the formal theory. For many problems a purely geometric approach is not sufficient and some additional algebraic ingredients are needed. These lead us to the title concept of this book: involution. We use in this chapter a rather abstract presentation that at first sight has no relations at all to differential equations but is rooted in commutative algebra. Only in the next chapter we will see how these algebraic ingredients appear in the analysis of the (geometric) symbol of a differential equation.

In the first section we introduce the notion of an involutive division, a restriction of the usual divisibility relation of power products. As we are interested in applications beyond the commutative case, we introduce the theory not for power products but more abstractly for the monoid of multi indices. Similarly, we proceed not just to the usual polynomial ring but to a more general class of not necessarily commutative algebras: the polynomial algebras of solvable types. They are introduced with many examples like rings of linear differential or difference operators or universal enveloping algebras of Lie algebras in Section 2.2.

In the following section we introduce involutive bases for arbitrary algebras of solvable type. It turns out that they may be understood as a special kind of Gröbner bases with additional combinatorial properties. The question of their existence and construction is postponed to Chapter 6; here we are only concerned with some elementary consequences of the definition. The fourth section studies in more details Pommaret bases which are for us the most important involutive bases. In particular, the problem of δ-regularity requires some scrutiny.

In Section 2.5 we motivate our approach to involution by discussing combinatorial decompositions of polynomial modules. It is a trivial fact that any involutive basis allows us to identify the submodule it generates with the direct sum of some free modules. Arbitrary polynomial modules may be presented as the quotient of a free module by one of its submodule. Any Gröbner basis of this submodule induces a Stanley decomposition, i. e. a vector space isomorphism between the module and a direct sum of free modules. A trivial application is the straightforward determination of the Hilbert function of the module. Pommaret bases lead to a special kind of decompositions, namely Rees decompositions, and thus provide us with a simple realisation of Hironaka's criterion for Cohen-Macaulay modules.

In the final section we study as a further application of involutive bases syzygies and free resolutions. Again Pommaret bases possess special properties. Given a Pommaret basis of a module, it is straightforward to construct a Pommaret basis of its syzygy module. By iteration, we easily obtain the full syzygy resolution and a stronger form of Hilbert's Syzygy Theorem. Finally, we show that the degree of a Pommaret basis is nothing but the Castelnuovo–Mumford regularity.

This chapter requires some familiarity with basic concepts in commutative algebra, in particular with Gröbner bases. Our introduction of involutive bases is closely modelled on a classical approach to Gröbner bases. A brief review of the most important points of this approach in the form we need is contained in Appendix B.

2.1 Involutive Divisions

As we will use involution in different contexts, we need a fairly general and flexible approach. For this reason we do not define it in terms of power products or differential operators or something similar but in terms of multi indices (introduced in Appendix A.1). We study the Abelian monoid $(\mathbb{N}_0^n, +)$ with the addition defined componentwise. The multi indices may be identified in a natural way with the vertices of an n-dimensional integer lattice, so that we can easily visualise subsets of $\mathbb{N}_0^n$.

For a multi index $\nu \in \mathbb{N}_0^n$ we introduce its *cone* $\mathcal{C}(\nu) = \nu + \mathbb{N}_0^n$, i. e. the set of all multi indices that can be reached from ν by adding another multi index. We say that ν *divides* μ, written $\nu \mid \mu$, if $\mu \in \mathcal{C}(\nu)$. It might appear strange to speak here of a "division", as the operation in $\mathbb{N}_0^n$ is an addition, but if we think of the power products x^ν and x^μ, then $\mu \in \mathcal{C}(\nu)$ is equivalent to $x^\nu \mid x^\mu$. Given a finite subset $\mathcal{N} \subset \mathbb{N}_0^n$, we define its *span* as the monoid ideal generated by $\mathcal{N}$:

$$\langle \mathcal{N} \rangle = \bigcup_{\nu \in \mathcal{N}} \mathcal{C}(\nu) \,. \tag{2.1}$$

Again this is easy to interpret for power products. The set $\{x^\nu \mid \nu \in \mathcal{N}\}$ generates a monomial ideal $\mathcal{J}$. We have $\mu \in \langle \mathcal{N} \rangle$, if and only if $x^\mu \in \mathcal{J}$.

The basic idea of an involutive division is to introduce a restriction of the cone of a multi index, the involutive cone: it is only allowed to add multi indices certain entries of which vanish. This is equivalent to a restriction of the above defined divisibility relation. The final goal will be having a *disjoint* union in (2.1) by using only these involutive cones on the right hand side. This will naturally lead to the combinatorial decompositions discussed in Section 2.5.

In order to finally give the definition of an involutive division, we need one more notation: let $N \subseteq \{1, \dots, n\}$ be an arbitrary subset of the set of the first n integers; then we write $\mathbb{N}_N^n = \{\nu \in \mathbb{N}_0^n \mid \nu_j = 0,\ \forall j \notin N\}$ for the set of all multi indices where the only entries who may be non-zero are those whose indices are contained in N.

Definition 2.1.1. *An* involutive division L *is defined on* $(\mathbb{N}_0^n, +)$*, if a subset* $N_{L,\mathcal{N}}(\nu) \subseteq \{1, \dots, n\}$ *of* multiplicative indices *is associated to every multi index* ν *in a finite subset* $\mathcal{N} \subset \mathbb{N}_0^n$ *such that the following two conditions on the* involutive cones $\mathcal{C}_{L,\mathcal{N}}(\nu) = \nu + \mathbb{N}^n_{N_{L,\mathcal{N}}(\nu)} \subseteq \mathbb{N}_0^n$ *of the multi indices* $\nu \in \mathcal{N}$ *are satisfied.*

1. *If there exist two elements* $\mu, \nu \in \mathcal{N}$ *with* $\mathcal{C}_{L,\mathcal{N}}(\mu) \cap \mathcal{C}_{L,\mathcal{N}}(\nu) \neq \emptyset$*, either* $\mathcal{C}_{L,\mathcal{N}}(\mu) \subseteq \mathcal{C}_{L,\mathcal{N}}(\nu)$ *or* $\mathcal{C}_{L,\mathcal{N}}(\nu) \subseteq \mathcal{C}_{L,\mathcal{N}}(\mu)$ *holds.*
2. *If* $\mathcal{N}' \subset \mathcal{N}$*, then* $N_{L,\mathcal{N}}(\nu) \subseteq N_{L,\mathcal{N}'}(\nu)$ *for all* $\nu \in \mathcal{N}'$*.*

An arbitrary multi index $\mu \in \mathbb{N}_0^n$ *is* involutively divisible *by* $\nu \in \mathcal{N}$*, written* $\nu \mid_{L,\mathcal{N}} \mu$*, if* $\mu \in \mathcal{C}_{L,\mathcal{N}}(\nu)$*.*

Before we discuss the precise meaning of this definition and in particular of the two conditions contained in it, we should stress the following important point: as indicated by the notation, involutive divisibility is always defined with respect to both an involutive division L and a fixed finite set $\mathcal{N} \subset \mathbb{N}_0^n$; only an element of $\mathcal{N}$ can be an involutive divisor. Obviously, involutive divisibility $\nu \mid_{L,\mathcal{N}} \mu$ implies ordinary divisibility $\nu \mid \mu$.

The involutive cone $\mathcal{C}_{L,\mathcal{N}}(\nu)$ of an element $\nu \in \mathcal{N}$ is a subset of the full cone $\mathcal{C}(\nu)$. We are not allowed to add arbitrary multi indices to ν but may increase only certain entries of ν determined by the multiplicative indices. The first condition in the above definition says that involutive cones can intersect only trivially: if two intersect, one must be a subset of the other.

The *non-multiplicative indices* form the complement of $N_{L,\mathcal{N}}(\nu)$ in N and are denoted by $\bar{N}_{L,\mathcal{N}}(\nu)$. If we remove some elements from the set $\mathcal{N}$ and determine the multiplicative indices of the remaining elements with respect to the subset $\mathcal{N}'$, we will obtain in general a different result than before. The second condition for an involutive division says now that while it may happen that a non-multiplicative index becomes multiplicative for some $\nu \in \mathcal{N}'$, the converse cannot happen.

Example 2.1.2. An important example of an involutive division is the *Janet division* J. In order to define it, we must introduce certain subsets of the given set $\mathcal{N} \subset \mathbb{N}_0^n$:

$$(d_k, \dots, d_n) = \left\{ \nu \in \mathcal{N} \mid \nu_i = d_i\,,\ k \leq i \leq n \right\}\,. \tag{2.2}$$

The index n is multiplicative for $\nu \in \mathcal{N}$, if $\nu_n = \max_{\mu\in\mathcal{N}}\{\mu_n\}$, and $k < n$ is multiplicative for $\nu \in (d_{k+1}, \dots, d_n)$, if $\nu_k = \max_{\mu\in(d_{k+1},\dots,d_n)}\{\mu_k\}$.

Let us consider for $n = 2$ the set $\mathcal{N} = \left\{\nu^{(1)} = [2,0], \nu^{(2)} = [0,2]\right\}$. As $\nu_2^{(2)} > \nu_2^{(1)}$, the index 2 is multiplicative only for $\nu^{(2)}$. In order to decide whether 1 is multiplicative, we must look at the subsets (0) and (2), respectively. As each of these subsets consists of only one element, the index 1 is multiplicative for both multi indices and we obtain that $N_{J,\mathcal{N}}(\nu^{(1)}) = \{1\}$ and $N_{J,\mathcal{N}}(\nu^{(2)}) = \{1,2\}$. ◁

Lemma 2.1.3. *The Janet division is an involutive division.*

Proof. The second condition of Definition 2.1.1 is easy to verify. If k is a multiplicative index for $\nu \in \mathcal{N}$, then ν_k is maximal in a certain subset $(d_{k+1}, \dots, d_n) \subseteq \mathcal{N}$. If we remove some elements of the set $\mathcal{N}$, this maximality is unchanged. Thus all multiplicative indices remain multiplicative. It may happen that some non-multiplicative indices become multiplicative, but this is allowed.

Concerning the first condition we will prove an even stronger result: if $\mathcal{C}_{J,\mathcal{N}}(\mu) \cap \mathcal{C}_{J,\mathcal{N}}(\nu) \neq \emptyset$ for two multi indices $\mu, \nu \in \mathcal{N}$, then μ and ν must in fact be identical. So assume there exists a multi index $\rho \in \mathcal{C}_{J,\mathcal{N}}(\mu) \cap \mathcal{C}_{J,\mathcal{N}}(\nu)$. Then there must exist two multi indices $\bar{\mu} \in \mathbb{N}^n_{N_{J,\mathcal{N}}(\mu)}$ and $\bar{\nu} \in \mathbb{N}^n_{N_{J,\mathcal{N}}(\nu)}$ such that $\rho = \mu + \bar{\mu} = \nu + \bar{\nu}$. We look first at ρ_n. If $\bar{\mu}_n > 0$ and $\bar{\nu}_n > 0$, n is multiplicative for both μ and ν. But by the definition of the Janet division this requires $\mu_n = \nu_n$. It is not possible that $\bar{\mu}_n > 0$ and $\bar{\nu}_n = 0$, as then $\mu_n < \nu_n$ and nevertheless $n \in N_{J,\mathcal{N}}(\mu)$ contradicting the definition of the Janet division. The same holds of course for $\bar{\mu}_n = 0$ and $\bar{\nu}_n > 0$. Thus we may conclude that $\mu_n = \nu_n$.

We proceed now to ρ_{n-1}. By the considerations above we know already that $\mu, \nu \in (\mu_n)$. But this implies that we can repeat the same argument for the next entry. For $n-1$ to be a multiplicative index for μ, we must have $\mu_{n-1} \geq \nu_{n-1}$ and vice versa. Hence $\mu_{n-1} = \nu_{n-1}$ and $\mu, \nu \in (\mu_{n-1}, \mu_n)$. Iterating this argument, we get $\mu = \nu$. □

We have stressed above that involutive divisibility is always defined with respect to a given set $\mathcal{N}$. However, there exist some involutive divisions where the multiplicative variables are independent of $\mathcal{N}$. Such divisions are very convenient in concrete computations and they have got a special name.

Definition 2.1.4. *An involutive division* L *is* globally defined*, if the assignment of the multiplicative indices is independent of the set* $\mathcal{N}$*; in this case we write simply* $N_L(\nu)$*.*

Example 2.1.5. An involutive division that will become very important for us in the sequel is the *Pommaret division* P. It assigns the multiplicative indices according to the following simple rule: if $1 \leq k \leq n$ is the smallest index such that $\nu_k > 0$ for some multi index $\nu \in \mathbb{N}_0^n \setminus \{[0,\dots,0]\}$, then we call k the *class* of ν, written $\operatorname{cls}\nu$, and set $N_P(\nu) = \{1,\dots,k\}$. Finally, we define $N_P([0,\dots,0]) = \{1,\dots,n\}$. Thus P is globally defined. ◁

Lemma 2.1.6. *The Pommaret division is an involutive division.*

Proof. The second condition of Definition 2.1.1 obviously holds for *any* globally defined division, as the set $\mathcal{N}$ does not influence the multiplicative indices. For the first condition assume that there is a multi index $\rho \in \mathcal{C}_P(\mu) \cap \mathcal{C}_P(\nu)$. As in the proof of Lemma 2.1.3 we write $\rho = \mu + \bar{\mu} = \nu + \bar{\nu}$ with multi indices $\bar{\mu} \in \mathbb{N}^n_{N_P(\mu)}$ and $\bar{\nu} \in \mathbb{N}^n_{N_P(\nu)}$. Without loss of generality, we may assume that $\operatorname{cls}\mu \leq \operatorname{cls}\nu = k$. By definition of the Pommaret division this implies $\rho_i = \mu_i = \nu_i$ for all $i > k$. We must distinguish two cases. If $\operatorname{cls}\mu = k$, too, then either $\nu_k \leq \mu_k$ and consequently $\mu \in \mathcal{C}_P(\nu)$ or $\nu_k > \mu_k$ and $\nu \in \mathcal{C}_P(\mu)$. If $\operatorname{cls}\mu < k$ and $\nu_k \leq \mu_k$, then again $\mu \in \mathcal{C}_P(\nu)$. This finishes the proof, as it is not possible that $\operatorname{cls}\mu < k$ and $\nu_k > \mu_k$: because of $\rho_k \geq \nu_k$ this would require that k is a multiplicative index for μ contradicting $\operatorname{cls}\mu < k$. □

Above we introduced the span of a set $\mathcal{N} \subset \mathbb{N}_0^n$ as the union of the cones of its elements. Given an involutive division it appears natural to consider also the union of the involutive cones. Obviously, this yields in general only a subset of the monoid ideal $\langle\mathcal{N}\rangle$.

Definition 2.1.7. *The* involutive span *of a finite set* $\mathcal{N} \subset \mathbb{N}_0^n$ *is*

$$\langle\mathcal{N}\rangle_L = \bigcup_{\nu \in \mathcal{N}} \mathcal{C}_{\mathcal{N},L}(\nu)\ . \tag{2.3}$$

The set $\mathcal{N}$ *is called* weakly involutive *for the division* L *or a* weak involutive basis *of the monoid ideal* $\langle\mathcal{N}\rangle$*, if* $\langle\mathcal{N}\rangle_L = \langle\mathcal{N}\rangle$*. The set* $\mathcal{N}$ *is a* strong involutive basis *or for short an* involutive basis*, if the union in* (2.3) *is disjoint, i. e. the intersections of the involutive cones are empty. We call any finite set* $\mathcal{N} \subseteq \bar{\mathcal{N}} \subset \mathbb{N}_0^n$ *such that* $\langle\bar{\mathcal{N}}\rangle_L = \langle\mathcal{N}\rangle$ *a* (weak) involutive completion *of* $\mathcal{N}$*. An* obstruction to involution *for the set* $\mathcal{N}$ *is a multi index* $\nu \in \langle\mathcal{N}\rangle \setminus \langle\mathcal{N}\rangle_L$*.*

Remark 2.1.8. An obvious necessary condition for a strong involutive basis is that no multi indices $\mu, \nu \in \mathcal{N}$ exist such that $\mu \mid_{L,\mathcal{N}} \nu$. Sets with this property are called *involutively autoreduced.* ◁

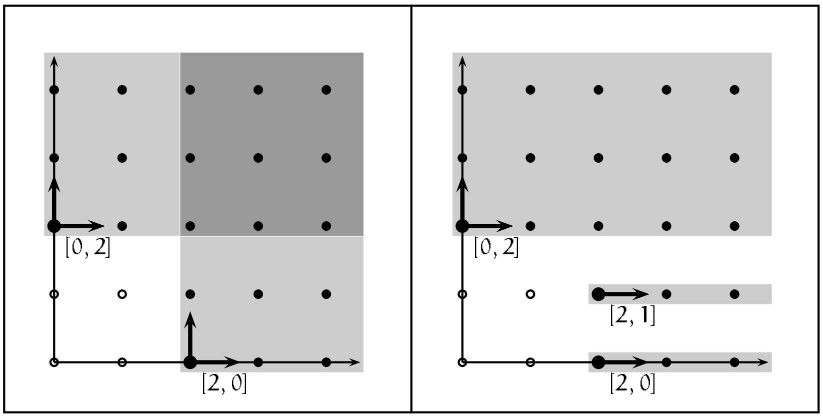

Fig. 2.1. *Left:* intersecting cones. *Right:* involutive cones.

Example 2.1.9. Figure 2.1 demonstrates the geometric interpretation of involutive divisions for $n = 2$. In both diagrams one can see the monoid ideal generated by the set $\mathcal{N} = \{[0,2],[2,0]\}$; the vertices belonging to it are marked by dark points. The arrows represent the multiplicative indices, i. e. the "allowed directions". The full cones of the two elements of $\mathcal{N}$ intersect in the darkly shaded area. The right diagram shows an involutive basis of $\langle\mathcal{N}\rangle$ for both the Janet and the Pommaret division. We must add to $\mathcal{N}$ the multi index $[2,1]$ and both for it and for $[2,0]$ only the index 1 is multiplicative. One can clearly see how $\langle\mathcal{N}\rangle$ is decomposed into three disjoint involutive cones: one of dimension 2, two of dimension 1. ◁

We are particularly interested in *strong* involutive bases, as only these lead to direct sums. The following result shows that in the "monomial" case any weak involutive basis can be reduced to a strong one by simply eliminating some elements.

Proposition 2.1.10. *If $\mathcal{N}$ is a weakly involutive set, a subset $\mathcal{N}' \subseteq \mathcal{N}$ exists such that $\mathcal{N}'$ is a strong involutive basis of $\langle\mathcal{N}\rangle$.*

Proof. The proof represents a nice motivation for the two conditions in Definition 2.1.1 of an involutive division. If $\mathcal{N}$ is not yet a strong involutive basis, the union in (2.3) cannot be disjoint. Thus there must exist involutive cones which intersect. By the first condition, this implies that some cones are contained in other ones; no other form of intersection is possible. If we eliminate the corresponding elements of $\mathcal{N}$, we get a subset $\mathcal{N}' \subset \mathcal{N}$ which has by the second condition the same involutive span, as the remaining elements may only gain additional multiplicative indices. Thus after a finite number of such eliminations we arrive at a strong involutive basis. $\square$

Remark 2.1.11. For the Janet division any weakly involutive set is also strongly involutive. This is a simple consequence of the proof of Lemma 2.1.3 where we showed that with respect to the Janet division two involutive cones only intersect, if they are identical. $\triangleleft$

The explicit construction of (strong) involutive completions for a given set $\mathcal{N} \subset \mathbb{N}_0^n$ will be discussed in detail in Section 6.4. For the moment we only note that we cannot expect that for an arbitrary set $\mathcal{N}$ and an arbitrary involutive division L an involutive basis $\mathcal{N}'$ of $\langle\mathcal{N}\rangle$ exists.

Example 2.1.12. We take the Pommaret division and the set $\mathcal{N} = \{[1,1]\}$. As $\operatorname{cls}[1,1] = 1$, we get $N_P([1,1]) = \{1\}$. So $\mathcal{C}_P([1,1]) \subset \mathcal{C}([1,1])$. But any multi index contained in $\langle\mathcal{N}\rangle$ has also class 1. Hence there does not exist a *finite* involutive basis of $\langle\mathcal{N}\rangle$ for the Pommaret division. We can generate it involutively only with the infinite set $\{[1,k] \mid k \in \mathbb{N}\}$. $\triangleleft$

Definition 2.1.13. *An involutive division L is called* Noetherian, *if any finite subset $\mathcal{N} \subset \mathbb{N}_0^n$ has a finite involutive completion with respect to L.*

Lemma 2.1.14. *The Janet division is Noetherian.*

Proof. Let $\mathcal{N} \subset \mathbb{N}_0^n$ be an arbitrary finite set. We will explicitly construct a Janet basis for $\langle\mathcal{N}\rangle$. Let μ be the multi index defined by $\mu_i = \max_{\nu\in\mathcal{N}} \nu_i$. We claim that the set

$$\bar{\mathcal{N}} = \{\bar{\nu} \in \mathbb{N}_0^n \cap \langle\mathcal{N}\rangle \mid \mu \in \mathcal{C}(\bar{\nu})\} \tag{2.4}$$

is an involutive completion of $\mathcal{N}$ with respect to the Janet division. Obviously, $\mathcal{N} \subseteq \bar{\mathcal{N}}$ and $\bar{\mathcal{N}} \subset \langle\mathcal{N}\rangle$. Let $\rho \in \langle\mathcal{N}\rangle$ be an arbitrary element. If $\rho \in \bar{\mathcal{N}}$, then trivially $\rho \in \langle\bar{\mathcal{N}}\rangle_J$. Otherwise set $I = \{i \mid \rho_i > \mu_i\}$ and define the multi index $\bar{\rho}$ by $\bar{\rho}_i = \rho_i$ for $i \notin I$ and $\bar{\rho}_i = \mu_i$ for $i \in I$, i.e. $\bar{\rho}_i = \min\{\rho_i, \mu_i\}$. By construction of the set $\bar{\mathcal{N}}$ and the definition of μ we have that $\bar{\rho} \in \bar{\mathcal{N}}$ and that $I \subseteq N_{J,\bar{\mathcal{N}}}(\bar{\rho})$. But this implies that $\rho \in \mathcal{C}_{J,\bar{\mathcal{N}}}(\bar{\rho})$ and thus $\bar{\mathcal{N}}$ is a finite Janet basis for $\langle\mathcal{N}\rangle$. □

2.2 Polynomial Algebras of Solvable Type

We introduced in the last section involutive bases in $\mathbb{N}_0^n$. We could interpret the multi indices as monomials and proceed to define involutive bases for polynomial ideals. But we are also interested in other rings containing linear differential or difference operators. As the basic ideas remain unchanged in many different situations, we need an algebraic structure within which all interesting cases can be handled in a unified manner. Slightly generalising a concept introduced by Kandry-Rody and Weispfenning [88], we use *polynomial algebras of solvable type.*

Let $\mathcal{P} = \Bbbk[x^1, \dots, x^n]$ be a polynomial ring over the field $\Bbbk$ of characteristic zero. $\mathcal{P}$ is a commutative ring with unit with respect to the usual multiplication. We equip the vector space $\mathcal{P}$ with alternative multiplications, in particular with non-commutative ones. We allow that both the variables x^i do not commute any more and that they operate on the coefficients. The usual multiplication is denoted either by a dot $\cdot$ or by no symbol at all. Alternative multiplications $\mathcal{P} \times \mathcal{P} \to \mathcal{P}$ are always denoted by $f \star g$.

Like Gröbner bases, involutive bases are defined with respect to a *term order*. It selects in each polynomial $f \in \mathcal{P}$ a *leading term* $\mathrm{lt}_\prec f = x^\mu$ with *leading exponent* $\mathrm{le}_\prec f = \mu$. Based on these leading exponents we can associate to each finite set $\mathcal{F}$ of polynomials a finite set $\mathrm{le}_\prec \mathcal{F} \subset \mathbb{N}_0^n$ to which we can apply the theory developed in the last section. But in order for this approach to make sense, we need a kind of compatibility between the multiplication $\star$ and the chosen term order.

Definition 2.2.1. $(\mathcal{P}, \star, \prec)$ *is a* polynomial algebra of solvable type *for the term order* $\prec$*, if the multiplication* $\star : \mathcal{P} \times \mathcal{P} \to \mathcal{P}$ *satisfies the following three axioms.*

(i) $(\mathcal{P}, \star)$ *is a ring with unit* 1.
(ii) $\forall a \in \Bbbk,\ f \in \mathcal{P}:\ a \star f = af$.
(iii) $\forall f, g \in \mathcal{P}:\ \mathrm{le}_\prec(f \star g) = \mathrm{le}_\prec f + \mathrm{le}_\prec g$.

Condition (i) ensures that the arithmetics in $(\mathcal{P}, \star, \prec)$ obeys the usual associative and distributive laws. Because of Condition (ii), $(\mathcal{P}, \star, \prec)$ is still a left $\mathbb{k}$-module. We do not require that it is a right $\mathbb{k}$-module, as this would exclude the possibility that the variables x^i operate on the coefficients. Condition (iii) relates the leading exponent of a product to the leading exponents of the factors. If $f, g \in \mathcal{P}$ are two arbitrary polynomials, a field element $c \in \mathbb{k}$ and a polynomial $h \in \mathcal{P}$ with $\mathrm{le}_\prec h \prec \mathrm{le}_\prec(f \cdot g)$ exist such that

$$f \star g = c\,(f \cdot g) + h\,. \tag{2.5}$$

We say that the multiplication $\star$ *respects the term order* $\prec$.

Since $\star$ is associative and distributive, the multiplication is fixed, as soon as the following data are given. We need for $1 \leq i, j \leq n$ constants $c_{ij} \in \mathbb{k}\setminus\{0\}$, polynomials $h_{ij} \in \mathcal{P}$ and maps $c_i : \mathbb{k} \to \mathbb{k}$, $h_i : \mathbb{k} \to \mathcal{P}$ such that

$$x^i \star a = c_i(a)x^i + h_i(a)\,, \quad \forall a \in \mathbb{k}\,, \tag{2.6a}$$

$$x^i \star x^j = c_{ij}x^i x^j + h_{ij}\,. \tag{2.6b}$$

Besides the conditions on the leading exponents of the polynomials h_{ij} and $h_i(a)$ imposed by (iii), each map c_i must be an injective field endomorphism and each map h_i must satisfy $h_i(a+b) = h_i(a)+h_i(b)$ and a kind of Leibniz rule $h_i(ab) = c_i(a)h_i(b) + h_i(a)b$.

Because of (iii) we can define Gröbner bases for ideals in algebras of solvable type. In case we endow $\mathcal{P}$ with a non-commutative multiplication, we must in principle distinguish left, right and two-sided ideals. However, we will exclusively work with left ideals and thus do not make this distinction. Recall that the left ideal generated by a finite subset $\mathcal{F} \subset \mathcal{P}$ is given by

$$\langle \mathcal{F} \rangle = \sum_{f \in \mathcal{F}} \mathcal{P} \star f\,. \tag{2.7}$$

Definition 2.2.2. *Let $(\mathcal{P}, \star, \prec)$ be a polynomial algebra of solvable type and $\mathcal{I} \subseteq \mathcal{P}$ a left ideal. A finite set $\mathcal{G} \subset \mathcal{P}$ is a* Gröbner basis *of the ideal $\mathcal{I}$, if $\langle \mathrm{le}_\prec \mathcal{G} \rangle = \mathrm{le}_\prec \mathcal{I}$.*

For the ordinary multiplication this definition reduces to the classical Definition B.2.3 of Gröbner bases. The decisive point, explaining the conditions imposed in Definition 2.2.1, is that *normal forms* with respect to a finite set $\mathcal{F} \subset \mathcal{P}$ can be computed in algebras of solvable type in precisely the same way as in the ordinary polynomial ring. Assume we are given a polynomial $f \in \mathcal{P}$ such that $\mathrm{le}_\prec g \mid \mathrm{le}_\prec f$ for some $g \in \mathcal{G}$ and set $\mu = \mathrm{le}_\prec f - \mathrm{le}_\prec g$. If we consider $g_\mu = x^\mu \star g$, then by (iii) $\mathrm{le}_\prec g_\mu = \mathrm{le}_\prec f$. Setting $d = \mathrm{lc}_\prec f / \mathrm{lc}_\prec g_\mu$, we find by (ii) that $\mathrm{lt}_\prec(f - d g_\mu) \prec \mathrm{lt}_\prec f$. Hence we may use the simple Algorithm B.1 for computing normal form; in particular, it always terminates by the same argument as in the ordinary case. Note that in general $d \neq \mathrm{lc}_\prec f / \mathrm{lc}_\prec g$, if $c \neq 1$

in (2.5), and that normal form computations are typically more expensive due to the appearance of the additional polynomial h in (2.5).

The Gröbner basis theory of Appendix B.2 can be extended straightforwardly to algebras of solvable type, as most proofs are based on the computation of normal forms. The remaining arguments mostly take place in $\mathbb{N}_0^n$ and thus can be applied without changes. In particular, we get the following result crucial for the termination of Buchberger's Algorithm B.2.

Proposition 2.2.3. *Let $(\mathcal{P}, \star, \prec)$ be a polynomial algebra of solvable type. Then $\mathcal{P}$ is a (left) Noetherian ring.*

Proof. For the ordinary multiplication this is Hilbert's Basis Theorem. There exist many different proofs of it; we present here one that can handle our more general situation. Let $\mathcal{I} \subseteq \mathcal{P}$ be a (left) ideal. By Dickson's Lemma the monoid ideal $\operatorname{le}_\prec \mathcal{I}$ has a finite basis. So we can choose a finite set $\mathcal{G} \subset \mathcal{I}$ such that $\operatorname{le}_\prec \mathcal{G}$ is a finite basis of $\operatorname{le}_\prec \mathcal{I}$. It follows from a simple normal form argument that $\mathcal{G}$ generates $\mathcal{I}$. $\square$

Note that we have in fact proven more than required by the proposition. The set $\mathcal{G}$ is obviously a Gröbner basis of $\mathcal{I}$, so that we have shown the existence of a Gröbner basis for arbitrary ideals in a polynomial algebra of solvable type. We stop here and do not give more details on extending the Gröbner basis theory to these algebras, as we will use in the next section a completely different approach leading to involutive bases. For the remainder of this section we consider a number of important examples of algebras of solvable type in order to make this abstract concept better understandable. They demonstrate the wide range of applicability of these algebras. Further examples can be found in [88] (and references therein).

Example 2.2.4. Obviously, the most trivial example of an algebra of solvable type is the usual polynomial ring, i.e. we set $f \star g = fg$. Its multiplication respects any term order, as by definition a term order is compatible with the monoid structure of the terms.

In contrast to Kandry-Rody and Weispfenning [88], we allow that the terms operate on the coefficients. The most important example for us are rings of *linear differential operators*. We take for the coefficients some function field $\mathbb{F}$, say the rational functions $\mathbb{Q}(x^1, \dots, x^n)$, and consider the polynomial ring $\mathcal{D} = \mathbb{F}[\partial_1, \dots, \partial_n]$ where $\partial_i = \partial/\partial x^i$. More generally, we may take any differential field $\mathbb{F}$ with commuting derivations ∂_i. The product of two monomial operators $a(x)\partial^\mu$ and $b(x)\partial^\nu$ is given by

$$a(x)\partial^\mu \star b(x)\partial^\nu = \sum_{\sigma+\rho=\mu} \binom{\mu}{\sigma} a(x) \frac{\partial^{|\sigma|} b(x)}{\partial x^\sigma} \partial^{\rho+\nu} \tag{2.8}$$

where $\binom{\mu}{\sigma}$ is a shorthand for $\prod_{i=1}^{n}\binom{\mu_i}{\sigma_i}$. By the properties of a term order

$$\mathrm{le}_{\prec}\big(a(x)\partial^{\mu} \star b(x)\partial^{\nu}\big) = \mu + \nu = \mathrm{le}_{\prec}\big(a(x)\partial^{\mu}\big) + \mathrm{le}_{\prec}\big(b(x)\partial^{\nu}\big)\ , \tag{2.9}$$

as any term $\partial^{\rho+\nu}$ appearing on the right hand side of (2.8) divides $\partial^{\mu+\nu}$ and thus $\partial^{\rho+\nu} \preceq \partial^{\mu+\nu}$ for any term order $\prec$. Hence $\mathcal{D}$ is a polynomial algebra of solvable type for all term orders. ◁

This example can be considerably extended to a rather general class of non-commutative polynomials originally introduced by Ore [130]; our exposition follows [26].

Example 2.2.5. Let $\mathbb{F}$ be any field of characteristic zero and $\sigma : \mathbb{F} \to \mathbb{F}$ an injective endomorphism. A *pseudo-derivation* with respect to σ is a map $\delta : \mathbb{F} \to \mathbb{F}$ such that (i) $\delta(f+g) = \delta(f)+\delta(g)$ and (ii) $\delta(f\cdot g) = \sigma(f)\cdot\delta(g)+\delta(f)\cdot g$ for all $f, g \in \mathbb{F}$. If $\sigma = \mathrm{id}_{\mathbb{F}}$, the identity map, (ii) is the standard Leibniz rule for derivations. If $\sigma \neq \mathrm{id}_{\mathbb{F}}$, one can show that there exists an $h \in \mathbb{F}$ such that $\delta = h(\sigma - \mathrm{id}_{\mathbb{F}})$. And conversely, if $\delta \neq 0$, there exists an $h \in \mathbb{F}$ such that $\sigma = h\delta + \mathrm{id}_{\mathbb{F}}$. Ore called $\sigma(f)$ the *conjugate* and $\delta(f)$ the *derivative* of f.

Given σ and δ, the ring $\mathbb{F}[\partial;\sigma,\delta]$ of univariate *Ore polynomials* consists of all formal polynomials in ∂ with coefficients in $\mathbb{F}$, i.e. of expressions of the form $\theta = \sum_{i=0}^{q} f_i\partial^i$ with $f_i \in \mathbb{F}$ and $q \in \mathbb{N}_0$. The addition is defined as usual. The "variable" ∂ operates on an element $f \in \mathbb{F}$ according to the rule

$$\partial \star f = \sigma(f)\partial + \delta(f) \tag{2.10}$$

which is extended associatively and distributively to define the multiplication in $\mathbb{F}[\partial;\sigma,\delta]$: given two elements $\theta_1, \theta_2 \in \mathbb{F}[\partial;\sigma,\delta]$, we can transform the product $\theta_1 \star \theta_2$ to the above normal form by repeatedly applying (2.10). The product of two linear polynomials evaluates then to

$$\begin{aligned}(f_1 + f_2\partial) \star (g_1 + g_2\partial) = f_1 g_1 &+ f_2\delta(g_1) + \\ &\big[f_1 g_2 + f_2\sigma(g_1) + f_2\delta(g_2)\big]\partial + \\ &f_2\sigma(g_2)\partial^2\ .\end{aligned} \tag{2.11}$$

The condition that σ is injective ensures that $\deg(\theta_1 \star \theta_2) = \deg\theta_1 + \deg\theta_2$. We call $\mathbb{F}[\partial;\sigma,\delta]$ the *Ore algebra* generated by σ and δ.

A simple familiar example is given by $\mathbb{F} = \mathbb{Q}(x)$, $\delta = \frac{d}{dx}$ and $\sigma = \mathrm{id}_{\mathbb{F}}$ yielding linear ordinary differential operators with rational functions as coefficients. Similarly, we obtain *recurrence* and *difference operators*. We set $\mathbb{F} = \mathbb{C}(n)$, the field of sequences with complex elements, and take for σ the shift operator, i.e. the automorphism mapping s_n to s_{n+1}. Then $\Delta = \sigma - \mathrm{id}_{\mathbb{F}}$ is a pseudo-derivation. $\mathbb{F}[E;\sigma,0]$ consists of linear ordinary recurrence operators, $\mathbb{F}[E;\sigma,\Delta]$ of linear ordinary difference operators.

Note that so far it is not yet clear why we call the elements of these classical examples "operators". In fact, the elements of any Ore algebra $\mathbb{F}[\partial;\sigma,\delta]$ can be interpreted as operators on vector spaces over $\mathbb{F}$. Let $\mathcal{V}$ be such a vector space and L an $\mathbb{F}$*-pseudo-linear map* $L : \mathcal{V} \to \mathcal{V}$, i.e. $L(u+v) = L(u) + L(v)$ and $L(fu) = \sigma(f)L(u) + \delta(f)u$ for all $f \in \mathbb{F}$ and $u, v \in \mathcal{V}$. Then we introduce an action $\mathcal{L} : \mathbb{F} \times \mathcal{V} \to \mathcal{V}$ mapping $(\theta = \sum f_i\partial^i, u)$ to $L_\theta u = \sum f_i L^i(u)$. For $\mathcal{V} = \mathbb{F}^m$ a natural choice for L is $L(v_1,\dots,v_m) = \bigl(\delta(v_1),\dots,\delta(v_m)\bigr)$. If $\delta = 0$ as for the recurrence operators, we may also take $L(v_1,\dots,v_m) = \bigl(\sigma(v_1),\dots,\sigma(v_m)\bigr)$. With these definitions the above Ore algebras contain linear operators in the familiar sense.

For multivariate Ore polynomials we take a set $\Sigma = \{\sigma_1,\dots,\sigma_n\}$ of injective $\mathbb{F}$-endomorphisms and a set $\Delta = \{\delta_1,\dots,\delta_n\}$ 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 of the form (2.10). If we require that all the maps σ_i, δ_j commute with each other, i.e. $\sigma_i \circ \sigma_j = \sigma_j \circ \sigma_i$, $\delta_i \circ \delta_j = \delta_j \circ \delta_i$ and $\sigma_i \circ \delta_j = \delta_j \circ \sigma_i$ for all $i \neq j$, one easily checks that $\partial_i \star \partial_j = \partial_j \star \partial_i$, i.e. the "variables" ∂_i commute.

Setting $\mathcal{D} = \{\partial_1,\dots,\partial_n\}$, we denote by $\mathbb{F}[\mathcal{D};\Sigma,\Delta]$ the ring of multivariate Ore polynomials. Because of the commutativity of the ∂_i we can 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 $\bigl(\mathbb{F}[\mathcal{D};\Sigma,\Delta],\star\bigr)$ is an algebra of solvable type for any term order goes completely analogously to the case of linear differential operators treated explicitly in Example 2.2.4 because of the great similarity of (2.10) and (2.8) (for $\mu = 1_j$ and $\nu = 0$). $\lhd$

Ore algebras are the special case where the maps h_i in (2.6a) take their values in $\Bbbk$ and not in $\mathcal{P}$ and where the variables commute. We consider now an example with non-commutative variables. By the same construction principle, factorisation of a free tensor algebra by a suitable quadratic ideal, one can generate many interesting algebras like Clifford algebras.

Example 2.2.6. Let $\mathfrak{A}$ be a finite-dimensional *Lie algebra* over a field $\Bbbk$. As usual we denote the product in $\mathfrak{A}$ by brackets: $[X,Y]$. Readers unfamiliar with Lie algebras should think of a subalgebra of $\mathfrak{gl}(m,\Bbbk)$, the algebra of all $m \times m$ matrices with entries from $\Bbbk$. The bracket is then the well-known commutator $[X,Y] = XY - YX$. Note that the multiplication in our algebra is the commutator and *not* the ordinary matrix multiplication!

As an algebra over a field, $\mathfrak{A}$ is in particular a vector space and we can choose a basis $\{X_1,\dots,X_n\}$. The structure of the Lie algebra is defined by the commutation relations of the generators X_i

$$[X_i, X_j] = C_{ij}^k X_k \tag{2.12}$$

where the $C_{ij}^k \in \Bbbk$ are the *structure constants* of the algebra $\mathfrak{A}$.

We construct now the *universal enveloping algebra* of $\mathfrak{A}$, an important tool in the representation theory of Lie algebras (see e. g. [185, Chapt. 3]). We start with the tensor algebra $\mathfrak{T}(\mathfrak{A})$ over $\mathfrak{A}$; a basis of it is given by all formal tensor products $X_{i_1} \otimes X_{i_2} \otimes \cdots \otimes X_{i_q}$ with $q \in \mathbb{N}_0$. As $X_i \otimes X_j \neq X_j \otimes X_i$, we cannot identify such "terms" with multi indices in $\mathbb{N}_0^n$. So we factor by the two-sided ideal generated by the quadratic elements $X_i \otimes X_j - X_j \otimes X_i - [X_i, X_j]$ in order to obtain the universal enveloping algebra $\mathfrak{U}(\mathfrak{A})$.

One can show that $\mathfrak{U}(\mathfrak{A})$ is an associative algebra and has the following universal property. Let $\pi : \mathfrak{A} \to \mathfrak{U}(\mathfrak{A})$ be the canonical linear map defined by $X_i \mapsto [X_i]$, i.e. each generator X_i is mapped on its equivalence class in the factor algebra $\mathfrak{U}(\mathfrak{A})$. The map π satisfies $\pi([X, Y]) = [\pi(X), \pi(Y)]$ for all $X, Y \in \mathfrak{A}$. Let $\mathfrak{B}$ be another associative algebra with a linear map $\phi : \mathfrak{A} \to \mathfrak{B}$ such that $\phi([X, Y]) = [\phi(X), \phi(Y)]$ for all $X, Y \in \mathfrak{A}$. Then there exists a unique homomorphism $\tilde{\phi} : \mathfrak{U}(\mathfrak{A}) \to \mathfrak{B}$ such that $\phi = \tilde{\phi} \circ \pi$.

In each equivalence class there is exactly one representative containing only terms of the form $X_{i_1} \otimes X_{i_2} \otimes \cdots \otimes X_{i_q}$ where now the indices form an ascending sequence $i_1 \leq i_2 \leq \cdots \leq i_q$. Given an arbitrary element $t \in \mathfrak{T}(\mathfrak{A})$ this representative can be constructed by substituting each product $X_i \otimes X_j$ with $i > j$ by $X_j \otimes X_i + [X_i, X_j]$ until no more such products appear. This leads to a normal form which can be identified with an element of $\Bbbk[X_1, \dots, X_n]$. Thus as vector spaces $\mathfrak{U}(\mathfrak{A})$ and $\Bbbk[X_1, \dots, X_n]$ are isomorphic (this is the *Poincaré–Birkhoff–Witt Theorem*).

As the right hand side of (2.12) is of lower degree than the left hand side, $\mathfrak{U}(\mathfrak{A})$ is a polynomial algebra of solvable type for any degree compatible term order. Kandry-Rody and Weispfenning [88] showed that if $\mathfrak{A}$ is a *solvable* Lie algebra, one can find bases $\{X_1, \dots, X_n\}$ such that $\mathfrak{U}(\mathfrak{A})$ is also of solvable type with respect to the lexicographic order.

As a concrete example we study the three-dimensional Lie algebra $\mathfrak{so}(3)$. Its commutation relations are $[X_i, X_j] = \epsilon_{ijk} X_k$ where ϵ_{ijk} is the totally antisymmetric tensor and we obtain for the multiplication in $\mathfrak{U}(\mathfrak{so}(3))$

$$\begin{aligned}
x^1 \star x^2 &= x^1 x^2\,, & x^2 \star x^1 &= x^1 x^2 - x^3\,, \\
x^1 \star x^3 &= x^1 x^3\,, & x^3 \star x^1 &= x^1 x^3 + x^2\,, \\
x^2 \star x^3 &= x^2 x^3\,, & x^3 \star x^2 &= x^2 x^3 - x^1\,.
\end{aligned} \tag{2.13}$$

It is obvious that while the multiplication respects any degree compatible term order it does not respect the lexicographic order. ◁

Example 2.2.7. Non-commuting variables also occur in the *Weyl algebra* consisting of linear differential operators with polynomial coefficients. We may consider it as a polynomial ring $\mathcal{W}_n = \Bbbk[x^1, \dots, x^n, \partial_1, \dots, \partial_n]$ where the unknowns satisfy the commutation rules

$$[x^i, x^j] = 0\,, \quad [\partial_i, \partial_j] = 0\,, \quad [x^i, \partial_j] = \delta^i_j\,. \tag{2.14}$$

In fact, adding the "variable" $\mathbb{1}$ commuting which all other ones and writing the right hand side of (2.14) as $\delta^i_j \mathbb{1}$, we see that (2.14) defines a $(2n+1)$-dimensional Lie algebra, the *Heisenberg algebra*, which forms the basis of quantum mechanics. The Weyl algebra is a factor algebra of its enveloping algebra; we must factor in addition by the relations $\mathbb{1} \cdot x^i = x^i$ and $\mathbb{1} \cdot \partial_i = \partial_i$.

Modules over the Weyl algebra, often called $\mathcal{D}$*-modules*, are used in many different fields, in particular in singularity theory and in the representation theory of algebraic groups. An introduction into their theory may be found in [39]; more advanced material is contained in [19, 22]. Gröbner bases in the Weyl algebra are discussed in [154].

With respect to the solvability the Weyl algebra behaves like a ring of differential operators with polynomial coefficients. By the same argument as in Example 2.2.4 one can show that it is an algebra of solvable type for any term order. The difference to the there studied rings of differential operators lies in the different definition of terms. A term in the Weyl algebra has the form $x^\mu \partial^\nu$ where $\mu, \nu \in \mathbb{N}_0^n$. Thus the leading exponent of an element of $\mathcal{W}_n$ is a multi index with $2n$ entries. Terms in $\mathcal{D}$ are of the form ∂^μ with $\mu \in \mathbb{N}_0^n$ so that exponents have only n entries. $\lhd$

2.3 Involutive Bases

We proceed to define involutive bases for left ideals of polynomial algebras of solvable type. Our approach closely follows the introduction of Gröbner bases in Appendix B.2. In principle, we could at once consider submodules of free modules over such an algebra. But this only complicates the notation. So we start with the ideal case and the extension to submodules goes as for Gröbner bases.

Let $(\mathcal{P}, \star, \prec)$ be an algebra of solvable type for some term order $\prec$ and $\mathcal{F} \subset \mathcal{P}$ a finite subset. Furthermore, we choose on $\mathbb{N}_0^n$ an involutive division L. With the help of the leading exponents in $\operatorname{le}_\prec \mathcal{F}$, we may assign each element $f \in \mathcal{F}$ a set of *multiplicative variables*

$$X_{L,\mathcal{F},\prec}(f) = \left\{x^i \mid i \in N_{L,\operatorname{le}_\prec \mathcal{F}}(\operatorname{le}_\prec f)\right\}\,. \tag{2.15}$$

This leads naturally to the following extension of Definition 2.1.7 of the involutive span.

Definition 2.3.1. *Let $\mathcal{F} \subset \mathcal{P}$ be a finite set. Its* involutive span *for a given involutive division* L *and the term order* $\prec$ *is*

$$\langle \mathcal{F} \rangle_{L,\prec} = \sum_{f \in \mathcal{F}} \Bbbk[X_{L,\mathcal{F},\prec}(f)] \star f\,. \tag{2.16}$$

The definition of an obstruction to involution is more involved than in the monomial case. We need two central notions in the theory of Gröbner bases: reduction and normal form. We introduce variants adapted to the context of involutive bases.

Definition 2.3.2. *Let $\mathcal{F} \subset \mathcal{P}$ be a finite set and* L *an involutive division. A polynomial $g \in \mathcal{P}$ is* involutively reducible *with respect to $\mathcal{F}$, if it contains a term $x^\mu \in \operatorname{supp} g$ such that $\mathrm{le}_\prec f \,|_{L,\mathrm{le}_\prec \mathcal{F}}\, \mu$ for some $f \in \mathcal{F}$. The polynomial g is in* involutive normal form *with respect to $\mathcal{F}$, if it is not involutively reducible. The set $\mathcal{F}$ is* involutively autoreduced*, if no polynomial $f \in \mathcal{F}$ contains a term $x^\mu \in \operatorname{supp} f$ such that another polynomial $f' \in \mathcal{F} \setminus \{f\}$ exists with $\mathrm{le}_\prec f' \,|_{L,\mathrm{le}_\prec \mathcal{F}}\, \mu$.*

Remark 2.3.3. The definition of an involutively autoreduced set *cannot* be formulated more concisely by saying that each $f \in \mathcal{F}$ is in involutive normal form with respect to $\mathcal{F} \setminus \{f\}$. If we are not dealing with a global division, the removal of f from $\mathcal{F}$ will generally change the assignment of the multiplicative indices and thus affect the involutive divisibility. $\triangleleft$

Involutive reducibility is obviously a restriction of ordinary reducibility. An *obstruction to involution* is a polynomial $g \in \langle \mathcal{F} \rangle \setminus \langle \mathcal{F} \rangle_{L,\prec}$ possessing a standard representation (B.9) with respect to $\mathcal{F}$; i.e. $g = \sum_{f \in \mathcal{F}} g_f \star f$ for polynomials g_f satisfying $\mathrm{le}_\prec(g_f \star f) \preceq \mathrm{le}_\prec g$. We will later see that these elements make the difference between an involutive and a Gröbner basis.

In a differential equations context, elements of $\langle \mathcal{F} \rangle$ which do not possess a standard representation with respect to $\mathcal{F}$ are called *integrability conditions*. Note that this notion depends on the chosen term order $\prec$; only for degree compatible term orders it coincides with the informal definition of integrability conditions used in Section 1.4. By Proposition B.2.7, $\mathcal{F}$ is a Gröbner basis, if and only if no "integrability conditions" exist.

Example 2.3.4. Consider the set $\mathcal{F} = \{f_1, f_2, f_3\} \subset \Bbbk[x, y, z]$ with the polynomials $f_1 = z^2 - xy$, $f_2 = yz - x$ and $f_3 = y^2 - z$. For any degree compatible term order, the leading terms of f_2 and f_3 are unique. For f_1 we have two possibilities: if we use the degree lexicographic order[1] (i.e. for $x \prec y \prec z$), it is z^2, for the degree inverse lexicographic order (i.e. for $x \succ y \succ z$) the

[1] All used term orders are introduced in Example B.2.1.

leading term is xy. In the first case, neither "integrability conditions" nor obstructions to involution for the Janet division exist, as $\langle\mathcal{F}\rangle_{J,\prec_{\text{deglex}}} = \langle\mathcal{F}\rangle$. In the second case, $f_4 = z^3 - x^2 = zf_3 + xf_1$ is an "integrability condition". Adding it to $\mathcal{F}$ yields a Gröbner basis $\mathcal{G}$ of $\langle\mathcal{F}\rangle$, as one can easily check using Proposition B.2.12. But now $f_5 = zf_4$ is an obstruction to involution of $\mathcal{G}$, as it is not involutively reducible with respect to the Janet division. $\lhd$

It often suffices, if one does not consider all terms in $\operatorname{supp} g$ but only the leading term $\operatorname{lt}_\prec g$: the polynomial g is *involutively head reducible*, if $\operatorname{le}_\prec f\,|_{L,\operatorname{le}_\prec\mathcal{F}}\operatorname{le}_\prec g$ for some $f \in \mathcal{F}$. Similarly, the set $\mathcal{F}$ is *involutively head autoreduced*, if no leading exponent of an element $f \in \mathcal{F}$ is involutively divisible by the leading exponent of another element $f' \in \mathcal{F} \setminus \{f\}$.

Example 2.3.5. In the case of the Janet division, a set $\mathcal{F}$ is involutively head autoreduced as soon as all elements have different leading terms. Indeed, we have seen in the proof of Lemma 2.1.3 that with respect to the Janet division two involutive cones only intersect, if they are identical. $\lhd$

As involutive reducibility is a restriction of ordinary reducibility, involutive normal forms can be determined with trivial adaptions of the familiar algorithms like for example Algorithm B.1. The termination follows by the same argument as usual, namely that $\prec$ is a well-order. If g' is an involutive normal form of $g \in \mathcal{P}$ with respect to $\mathcal{F}$ for the division L, then we write $g' = \operatorname{NF}_{\mathcal{H},L,\prec}(g)$, although involutive normal forms are in general not unique (like ordinary normal forms – see Example B.2.6). Depending on the order in which reductions are applied different results are obtained.

The ordinary normal form is unique, if and only if it is computed with respect to a Gröbner basis (Proposition B.2.7). In fact, this property is often used to define Gröbner bases. For the involutive normal form, the situation is somewhat different.

Lemma 2.3.6. *If the finite set $\mathcal{F} \subset \mathcal{P}$ is involutively head autoreduced with respect to the involutive division L, the sum in (2.16) is direct.*

Proof. Let f_1, f_2 be two distinct elements of $\mathcal{F}$ and $X_i = X_{L,\mathcal{F},\prec}(f_i)$ their sets of multiplicative variables. Assume two polynomials $P_i \in \Bbbk[X_i]$ exist such that $P_1 \star f_1 = P_2 \star f_2$ implying in particular that $\operatorname{le}_\prec(P_1 \star f_1) = \operatorname{le}_\prec(P_2 \star f_2)$. As the multiplication $\star$ respects the term order $\prec$, this can only happen, if $\mathcal{C}_{L,\operatorname{le}_\prec\mathcal{F}}(\operatorname{le}_\prec f_1) \cap \mathcal{C}_{L,\operatorname{le}_\prec\mathcal{F}}(\operatorname{le}_\prec f_2) \neq \emptyset$. Thus one of the involutive cones is completely contained in the other one and either $\operatorname{le}_\prec f_1\,|_{L,\operatorname{le}_\prec\mathcal{F}}\operatorname{le}_\prec f_2$ or $\operatorname{le}_\prec f_2\,|_{L,\operatorname{le}_\prec\mathcal{F}}\operatorname{le}_\prec f_1$ contradicting that $\mathcal{F}$ is involutively head autoreduced. $\square$

Proposition 2.3.7. *If the finite set $\mathcal{F} \subset \mathcal{P}$ is involutively head autoreduced, each polynomial $g \in \mathcal{P}$ has a unique involutive normal form $\operatorname{NF}_{\mathcal{F},L,\prec}(g)$.*

Proof. If 0 is an involutive normal form of g, then obviously $g \in \langle\mathcal{F}\rangle_{L,\prec}$. Conversely, assume that $g \in \langle\mathcal{F}\rangle_{L,\prec}$, i.e. the polynomial g can be written in the form $g = \sum_{f\in\mathcal{F}} P_f \star f$ with $P_f \in \Bbbk[X_{L,\mathcal{F},\prec}(f)]$. As $\mathcal{F}$ is involutively head autoreduced, the leading terms of the summands can never cancel (see the proof of Lemma 2.3.6). Thus $\mathrm{lt}_\prec g = \mathrm{lt}_\prec(P_f \star f)$ for some $f \in \mathcal{F}$ and any polynomial $g \in \langle\mathcal{F}\rangle_{L,\prec}$ is involutively head reducible with respect to $\mathcal{F}$. Each reduction step in an involutive normal form algorithm leads to a new polynomial $g' \in \langle\mathcal{F}\rangle_{L,\prec}$ with $\mathrm{lt}_\prec g' \preceq \mathrm{lt}_\prec g$. If the leading term is reduced, we even get $\mathrm{lt}_\prec g' \prec \mathrm{lt}_\prec g$. As each terminating normal form algorithm must sooner or later reduce the leading term, we eventually obtain 0 as unique involutive normal form of any $g \in \langle\mathcal{F}\rangle_{L,\prec}$.

Let g_1 and g_2 be two involutive normal forms of the polynomial g. Obviously, $g_1 - g_2 \in \langle\mathcal{F}\rangle_{L,\prec}$. By definition of a normal form both $\operatorname{supp} g_1$ and $\operatorname{supp} g_2$ do not contain any term involutively reducible with respect to $\mathcal{F}$ and the same holds for $\operatorname{supp}(g_1 - g_2) \subseteq \operatorname{supp} g_1 \cup \operatorname{supp} g_2$. Hence $g_1 - g_2$ is also in involutive normal form and by our considerations above we must have $g_1 - g_2 = 0$. □

Finally, we define involutive bases. As in the monomial case, we distinguish weak and strong bases. Although a strong involutive basis always induces a direct sum decomposition of the ideal $\mathcal{I}$ by Lemma 2.3.6, we cannot use this for the definition but must require a stronger property.

Definition 2.3.8. *Let $\mathcal{I} \subseteq \mathcal{P}$ be a (left) ideal. A* weak involutive basis *of $\mathcal{I}$ with respect to L and $\prec$ is a finite set $\mathcal{H}$ with $\langle\mathcal{H}\rangle_{L,\prec} = \mathcal{I}$. The set $\mathcal{H}$ is a* strong involutive basis *or shortly* involutive basis*, if it is in addition involutively head autoreduced.*

Example 2.3.9. We continue with Example 2.3.4. The set $\mathcal{F}$ is a Janet basis, i.e. an involutive basis with respect to the Janet division, of the ideal $\mathcal{I} = \langle\mathcal{F}\rangle$ for the term order $\prec_{\text{deglex}}$, although we have not yet the necessary tools to prove this fact. If we use $\prec_{\text{deginvlex}}$ as term order, a Janet basis is given by $\mathcal{F}' = \{f_1, f_2, f_3, f_4, f_5\}$; in fact, this is the smallest Janet basis of $\mathcal{I}$ for this term order, as it is not possible to remove an element. By Example 2.3.5, both $\mathcal{F}$ and $\mathcal{F}'$ are involutively head autoreduced for the respective term order. Note that for $\prec_{\text{deginvlex}}$ the basis is not only larger but also contains polynomials of higher degree. ◁

We postpone the question of existence and actual construction of involutive bases to Chapter 6. There we will also provide effective tools for checking whether a given set is an involutive basis. For the remainder of this section we concentrate instead on some important properties of these bases and especially on their relation to Gröbner bases. We start with an involutive form of

Proposition B.2.7. It is in so far stronger, as the involutive standard representation (2.17) is *unique* in contrast to the ordinary one (cf. Example B.2.8).

Proposition 2.3.10. *Let $\mathcal{I} \subseteq \mathcal{P}$ be a non-zero ideal and $\mathcal{H} \subset \mathcal{P}$ a finite set involutively head autoreduced with respect to the involutive division* L. *Then the following statements are equivalent.*

(i) *The set $\mathcal{H}$ is an involutive basis for the involutive division* L *and the term order $\prec$.*
(ii) *Every polynomial $f \in \mathcal{I}$ can be uniquely written in the form*

$$f = \sum_{h \in \mathcal{H}} P_h \star h \tag{2.17}$$

where the coefficients $P_h \in \Bbbk[X_{L,\mathcal{H},\prec}(h)]$ satisfy $\operatorname{lt}_\prec(P_h \star h) \preceq \operatorname{lt}_\prec f$ for all polynomials $h \in \mathcal{H}$.

Proof. As $\mathcal{H}$ is involutively head autoreduced, involutive normal forms are unique and $\mathrm{NF}_{\mathcal{H},L,\prec}(f) = 0$ for any $f \in \mathcal{I}$. Applying the division algorithm corresponding to the involutive form of Algorithm B.1 yields a representation in the form (2.17) satisfying the conditions on the leading terms. Its uniqueness follows from Lemma 2.3.6. Thus (i) implies (ii).

The converse is simple, too. Obviously, (ii) implies that $\langle \mathcal{H} \rangle_{L,\prec} = \mathcal{I}$. As by assumption $\mathcal{H}$ is furthermore involutively head autoreduced, it is an involutive basis of $\mathcal{I}$. □

Any monomial basis of a monomial ideal is trivially a Gröbner basis. Thus we can rephrase the definition of Gröbner bases as follows: a finite set $\mathcal{G}$ is a Gröbner basis of the ideal $\mathcal{I}$, if and only if $\operatorname{lt}_\prec \mathcal{G}$ is a Gröbner basis of $\operatorname{lt}_\prec \mathcal{I}$. Involutive bases satisfy a similar characterisation theorem.

Theorem 2.3.11. *Let $\mathcal{I} \subseteq \mathcal{P}$ be an ideal and $\mathcal{H} \subset \mathcal{P}$ a finite set. If $\operatorname{le}_\prec \mathcal{H}$ is a (weak) involutive basis of $\operatorname{le}_\prec \mathcal{I}$ for a given involutive division* L, *then $\mathcal{H}$ is a (weak) involutive basis of $\mathcal{I}$. If $\mathcal{H}$ is a strong involutive basis, the converse holds, too.*

Proof. For the first assertion take an arbitrary polynomial $f \in \mathcal{I}$. According to our assumption, its leading exponent $\operatorname{le}_\prec f$ lies in the involutive cone $\mathcal{C}_{L,\operatorname{le}_\prec \mathcal{H}}(h)$ of some element $h \in \mathcal{H}$. Let $\mu = \operatorname{le}_\prec f - \operatorname{le}_\prec h$ and set $f_1 = f - cx^\mu \star h$ where the coefficient $c \in \Bbbk$ is chosen such that the leading terms cancel. Obviously, $f_1 \in \mathcal{I}$ and $\operatorname{lt}_\prec f_1 \prec \operatorname{lt}_\prec f$. Thus we can iterate the procedure obtaining a sequence of polynomials $f_i \in \mathcal{I}$. After a finite number of steps we must reach $f_N = 0$, as the leading terms are always decreasing and by assumption *any* polynomial $f \in \mathcal{I}$ is involutively head reducible with respect to $\mathcal{H}$. But this implies that $f \in \langle \mathcal{H} \rangle_{L,\prec}$.

Thus we have proved that if $\mathrm{le}_\prec\mathcal{H}$ is a weak involutive basis of $\mathrm{le}_\prec\mathcal{I}$, then $\mathcal{H}$ is a weak involutive basis of $\mathcal{I}$. If $\mathrm{le}_\prec\mathcal{H}$ is a strong involutive basis of $\mathrm{le}_\prec\mathcal{I}$, then $\mathcal{H}$ is obviously involutively head autoreduced so that in this case it is in fact a strong involutive basis of $\mathcal{I}$.

For the converse we invoke Proposition 2.3.10. If the set $\mathcal{H}$ is a strong involutive basis, any polynomial f has a standard representation $f = \sum_{h\in\mathcal{H}} P_h\star h$ with $P_h \in \Bbbk[X_{L,\mathcal{H},\prec}(h)]$ and $\mathrm{lt}_\prec(P_h \star h) \preceq \mathrm{lt}_\prec f$ for all $h \in \mathcal{H}$. Thus we must have $\mathrm{le}_\prec f = \mathrm{le}_\prec(P_h\star h)$ for some $h \in \mathcal{H}$ and $\mathrm{le}_\prec\mathcal{H}$ is at least a weak involutive basis of the monoid ideal $\mathrm{le}_\prec\mathcal{I}$. But as $\mathcal{H}$ is involutively head autoreduced, it is in fact a strong involutive basis. □

Corollary 2.3.12. *Let $\mathcal{H}$ be an involutive basis of the ideal $\mathcal{I} \subseteq \mathcal{P}$. Then $\mathcal{H}$ is a Gröbner basis of $\mathcal{I}$. Furthermore, the involutive and the ordinary normal form of any polynomial $f \in \mathcal{P}$ are identical.*

Proof. The first assertion follows at once from Theorem 2.3.11. As $\mathrm{le}_\prec\mathcal{H}$ is even an involutive basis of $\mathrm{le}_\prec\mathcal{I}$, it satisfies of course $\langle\mathrm{le}_\prec\mathcal{H}\rangle = \mathrm{le}_\prec\mathcal{I}$ so that $\mathcal{H}$ is a Gröbner basis of $\mathcal{I}$ by Definition 2.2.2.

The second assertion is a trivial consequence of the first. If $\mathcal{H}$ is an involutive basis, it is involutively head autoreduced. Thus by Proposition 2.3.7 involutive normal forms with respect to $\mathcal{H}$ are unique. As $\mathcal{H}$ is also a Gröbner basis, ordinary normal forms are unique, too. Finally, any polynomial in ordinary normal form is trivially in involutive normal form. Hence the two normal forms must coincide. □

Remark 2.3.13. If $\mathcal{G}$ is a Gröbner basis of the ideal $\mathcal{I}$, then $\mathcal{I}$ cannot contain "integrability conditions" in the sense defined above, as any element of $\mathcal{I}$ has a standard representation by Proposition B.2.7. But this does not imply that for a given division L the ideal $\mathcal{I}$ is free of obstructions to involution. In order to obtain at least a weak involutive basis, we must add further elements of $\mathcal{I}$ to $\mathcal{G}$ until $\langle\mathrm{le}_\prec\mathcal{G}\rangle_L = \mathrm{le}_\prec\mathcal{I}$. Obviously, this observation allows us to reduce the construction of polynomial involutive bases to the monomial case. But we will see in Section 6.4 that better possibilities exist.

It follows from these considerations that in general involutive bases are not reduced Gröbner bases, as we already observed in Example 2.3.9. For $\prec_{\mathrm{deglex}}$ the set $\mathcal{F}$ was simultaneously a Janet basis and a reduced Gröbner basis. But for $\prec_{\mathrm{deginvlex}}$ the reduced Gröbner basis is $\mathcal{F} \cup \{f_4\}$, whereas a Janet basis requires in addition the polynomial f_5. We will see in the next sections that this "redundancy" in involutive bases is the key for their use in the structure analysis of polynomial ideals and modules. ◁

Finally, we discuss the generalisation of Proposition 2.1.10 to the polynomial case. It turns out that only under additional assumptions, it is easily possible to obtain a strong involutive basis from a weak one.

Proposition 2.3.14. *Let $\mathcal{I} \subseteq \mathcal{P}$ be an ideal and $\mathcal{H} \subset \mathcal{P}$ a finite set such that $\mathrm{le}_\prec \mathcal{H}$ is a weak involutive basis of $\mathrm{le}_\prec \mathcal{I}$ for the involutive division L. Then there exists a subset $\mathcal{H}' \subseteq \mathcal{H}$ that is a strong involutive basis of $\mathcal{I}$.*

Proof. If the set $\mathcal{H}$ is involutively head autoreduced, it is already a strong involutive basis and we are done. Otherwise $\mathcal{H}$ contains polynomials h_1, h_2 such that $\mathrm{le}_\prec h_1 \mid_{L,\mathrm{le}_\prec \mathcal{H}} \mathrm{le}_\prec h_2$. Consider $\mathcal{H}' = \mathcal{H} \setminus \{h_2\}$. By the same arguments as in the proof of Proposition 2.1.10 one easily shows that $\mathrm{le}_\prec \mathcal{H}' = \mathrm{le}_\prec \mathcal{H} \setminus \{\mathrm{le}_\prec h_2\}$ is still a weak involutive basis of $\mathrm{le}_\prec \mathcal{I}$ and thus by Theorem 2.3.11 the set $\mathcal{H}'$ is still a weak involutive basis of $\mathcal{I}$. We can repeat such eliminations, until we reach a strong involutive basis. □

Example 2.3.15. In order to demonstrate why we really need in Proposition 2.3.14 the additional assumption on the leading exponents of the basis, we consider in the ordinary polynomial ring $\Bbbk[x, y]$ the ideal $\mathcal{I}$ generated by the two polynomials $f_1 = y^2$ and $f_2 = y^2 + x^2$. If we order the variables as $x^1 = x$ and $x^2 = y$, then $\mathcal{H} = \{f_1, f_2\}$ is a weak Janet basis of $\mathcal{I}$, as all variables are multiplicative for each generator. However, $\mathrm{le}_\prec \mathcal{H} = \{[0, 2]\}$ does *not* generate $\mathrm{le}_\prec \mathcal{I}$, as obviously $[2, 0] \in \mathrm{le}_\prec \mathcal{I} \setminus \langle\{[0, 2]\}\rangle$.

Thus the assumption of Proposition 2.3.14 is not satisfied. Indeed, elimination of one of the generators leads to a set which no longer generates $\mathcal{I}$. One could hope that an involutive autoreduction of $\mathcal{H}$ might lead to a strong involutive basis, but this is not the case. We obtain the set $\mathcal{H}' = \{x^2, y^2\}$ which is not even a weak involutive basis of $\mathcal{I}$, as y is not multiplicative for x^2 and $x^2 y \notin \langle \mathcal{H}' \rangle_{J,\prec}$. ◁

2.4 Pommaret Bases and δ-Regularity

We have seen in Example 2.1.12 that not every monoid ideal in $\mathbb{N}_0^n$ possesses a finite Pommaret basis: the Pommaret division is not Noetherian. This phenomenon is known under the name *δ-regularity.* We study it in this section for (ordinary) polynomial ideals; the extension to differential equations is straightforward. But before, we prove two technical lemmata.

Lemma 2.4.1. *Let $\mathcal{H}$ be a homogeneous Pommaret basis of the homogeneous ideal $\mathcal{I} \subseteq \mathcal{P}$. If $q \geq \max_{h \in \mathcal{H}} \deg h$, then a Pommaret basis of the truncation $\mathcal{I}_{\geq q} = \bigoplus_{p \geq q} \mathcal{I}_p$ is given by*

$$\mathcal{H}_q = \left\{ x^\mu h \mid h \in \mathcal{H},\ |\mu| + \deg h = q,\ \mu_j = 0\ \forall j > \operatorname{cls} h \right\} . \tag{2.18}$$

Conversely, if $\mathcal{I}_{\geq q}$ possesses a Pommaret basis, then so does $\mathcal{I}$.

Proof. By the conditions in (2.18), any polynomial $h \in \mathcal{H}$ is multiplied by terms x^μ containing only variables multiplicative for it. Thus trivially $\operatorname{cls}(x^\mu h) = \operatorname{cls}\mu$. Furthermore, $\mathcal{H}_q$ is involutively head autoreduced, as $\mathcal{H}$ is. Now let $f \in \mathcal{I}_{\geq q}$ be an arbitrary homogeneous polynomial. As $\mathcal{H}$ is a Pommaret basis of $\mathcal{I}$, it has a standard representation $f = \sum_{h\in\mathcal{H}} P_h h$ with $P_h \in \Bbbk[x^1,\dots,x^{\operatorname{cls} h}]$. Hence f can be written as a linear combination of polynomials $x^\nu h$ where $|\nu| = \deg f - \deg h \geq q - \deg h$ and where x^ν contains only multiplicative variables. We decompose $\nu = \mu + \rho$ with $|\mu| = q - \deg h$ and $\rho_j = 0$ for all $j > \operatorname{cls}\mu$. Thus $x^\nu h = x^\rho(x^\mu h)$ with $x^\mu h \in \mathcal{H}_q$ and x^ρ contains only variables multiplicative for it. But this trivially implies the existence of a standard representation $f = \sum_{h'\in\mathcal{H}_q} P_{h'} h'$ with $P_{h'} \in \Bbbk[x^1,\dots,x^{\operatorname{cls} h'}]$ and thus $\mathcal{H}_q$ is a Pommaret basis of $\mathcal{I}_{\geq q}$.

The converse is also very simple. Let $\mathcal{H}_q$ be a Pommaret basis of $\mathcal{I}_{\geq q}$ and $\mathcal{H}_p$ bases of $\mathcal{I}_p$ as a vector space for $0 \leq p < q$. If we set $\mathcal{H} = \bigcup_{p=0}^{q} \mathcal{H}_q$, then $\operatorname{le}_\prec \mathcal{H}$ is obviously a weak Pommaret basis of $\operatorname{le}_\prec \mathcal{I}$ and by Proposition 2.3.14 an involutive head autoreduction yields a strong basis. $\square$

We call the number $q = \max_{h\in\mathcal{H}} \deg h$ the *degree* of the basis $\mathcal{H}$ and write $\deg \mathcal{H} = q$.

Lemma 2.4.2. *With the same notations as in Lemma 2.4.1, let $\mathcal{N} = \operatorname{le}_\prec \mathcal{H}_q$. If $\nu \in \mathcal{N}$ with $\operatorname{cls}\nu = k$, then $\nu - 1_k + 1_j \in \mathcal{N}$ for all $j > k$.*

Proof. j is non-multiplicative for ν. As $\mathcal{N}$ is an involutive basis of $\operatorname{le}_\prec \mathcal{I}_{\geq q}$, it must contain a multi index μ with $\mu \mid_P \nu + 1_j$. Obviously, $\operatorname{cls}(\nu + 1_j) = k$ and thus $\operatorname{cls}\mu \geq k$. Because of $|\mu| = |\nu|$, the only possibility is $\mu = \nu + 1_j - 1_k$. $\square$

The basic observation is that δ-regularity is a question of the used coordinates. With the help of a linear coordinate transformation, any ideal can be mapped into another one that has a finite Pommaret basis. The first problem is to detect that one is using "bad" coordinates and the second problem to find "good" ones. A number of different solutions exist; most of them tend to be computationally rather expensive. We present here a new one; its main advantage is that it is very cheap.

We start with an algebraic definition of δ-regularity. Let as always $\mathcal{P} = \Bbbk[x^1,\dots,x^n]$ be the polynomial ring and $\prec$ a term order. A linear change of variables $\tilde{\mathbf{x}} = A\mathbf{x}$ with a regular matrix $A \in \Bbbk^{n\times n}$ transforms any polynomial in $\mathcal{P}$ into a polynomial of the same degree in $\tilde{\mathcal{P}} = \Bbbk[\tilde{x}^1,\dots,\tilde{x}^n]$. We use on $\tilde{\mathcal{P}}$ the same term order as on $\mathcal{P}$; i.e. if we ordered, say, degree reverse lexicographically on $\mathcal{P}$, we order again degree reverse lexicographically on $\tilde{\mathcal{P}}$. In other words, we are actually using an order defined on $\mathbb{N}_0^n$; performing a variable transformation only changes the multi indices appearing as exponents.

Let $\mathcal{F} \subset \mathcal{P}$ be a finite subset which is involutively head autoreduced with respect to an involutive division L. We call the total number of multiplicative variables of its elements its *involutive size* and denote it by

$$|\mathcal{F}|_{L,\prec} = \sum_{f\in\mathcal{F}} |X_{L,\prec,\mathcal{F}}(f)| . \tag{2.19}$$

Applying the above transformation yields a set $\tilde{\mathcal{F}} \subset \tilde{\mathcal{P}}$. In general, $\tilde{\mathcal{F}}$ is no longer involutively head autoreduced, but we can easily perform an involutive head autoreduction obtaining a set $\tilde{\mathcal{F}}^{\triangle}$. The leading exponents of $\tilde{\mathcal{F}}^{\triangle}$ may be very different from those of $\mathcal{F}$ and thus $|\mathcal{F}|_{L,\prec}$ may differ from $|\tilde{\mathcal{F}}^{\triangle}|_{L,\prec}$.

Definition 2.4.3. *Let the finite set $\mathcal{F} \subset \mathcal{P}$ be involutively head autoreduced with respect to the Pommaret division. The coordinates $\mathbf{x}$ are* δ-regular *for $\mathcal{F}$, if after any linear change of coordinates $\tilde{\mathbf{x}} = A\mathbf{x}$ the inequality $|\mathcal{F}|_{P,\prec} \geq |\tilde{\mathcal{F}}^{\triangle}|_{P,\prec}$ holds.*

Example 2.4.4. Let us reconsider Example 2.1.12 in terms of polynomials. Thus we take the set $\mathcal{F} = \{xy\} \subset \Bbbk[x, y]$. Independent of how we order the variables, the class of xy is 1. After the change of coordinates $x = \tilde{x} + \tilde{y}$ and $y = \tilde{y}$, we obtain the set $\tilde{\mathcal{F}} = \{\tilde{y}^2 + \tilde{x}\tilde{y}\} \subset \Bbbk[\tilde{x}, \tilde{y}]$. If we use the degree reverse lexicographic order, the leading term is $\tilde{y}^2$ which is of class 2. Thus it is possible to enlarge the involutive size of $\mathcal{F}$ and the original coordinates are not δ-regular. The new coordinates obviously are, as a higher class than 2 is not possible. ◁

Proposition 2.4.5. *Let $\mathcal{H}$ be a Pommaret basis of an ideal $\mathcal{I} \subseteq \mathcal{P}$. Then the coordinates $\mathbf{x}$ are δ-regular for $\mathcal{H}$.*

Proof. By Theorem 2.3.11, $\mathrm{le}_\prec\mathcal{H}$ is a Pommaret basis of $\mathrm{le}_\prec\mathcal{I}$. Thus it suffices to consider homogeneous ideals and, by Lemma 2.4.1, we may further reduce to the case that all generators are of the same degree q. The involutive size of $\mathcal{H}_q$ is then the vector space dimension of $\mathcal{I}_{q+1}$. This value is independent of the chosen coordinates and no involutively head autoreduced set may lead to a larger involutive size. Thus our coordinates are δ-regular. □

Thus δ-regularity is necessary for the existence of Pommaret bases and for the termination of the completion algorithms discussed in Chapter 6. From an algorithmic point of view, δ-singular coordinates are "bad". Fortunately, there are not many of them. Choosing an arbitrary coordinate system as reference, we may identify every coordinate system with the regular matrix $A \in \Bbbk^{n\times n}$ defining the transformation from our reference system to it.

Proposition 2.4.6. *Those coordinate systems that are δ-singular for a given finite involutively head autoreduced set $\mathcal{F} \subset \mathcal{P}$ form a Zariski closed set in $\Bbbk^{n\times n}$.*

Proof. Let us first perform a coordinate transformation with an undetermined matrix A, i.e. we are treating its entries as parameters. This obviously leads to a δ-regular coordinate system, as each polynomial in $\tilde{\mathcal{F}}^{\triangle}$ will get its maximally possible class. δ-singular coordinates are defined by the vanishing of certain (leading) coefficients. These coefficients are polynomials in the entries of A. Thus the set of all δ-singular coordinate systems can be described as the zero set of an ideal. □

By definition of the Zariski topology, this means that hardly any coordinate system is δ-singular. More precisely, they form a subset of dimension less than n^2 in an n^2-dimensional space. Thus if we randomly choose a coordinate system, it is δ-regular with probability 1! Nevertheless, δ-singular coordinate systems exist for most sets $\mathcal{F}$; exceptions are sets like $\mathcal{F} = \{x^2+y^2\}$ where one easily shows that any coordinate system is δ-regular. We will see below in the proof of Theorem 2.4.10 that it suffices for the construction of δ-regular coordinates to consider only transformations $\mathbf{x} \mapsto A\mathbf{x}$ where the matrix A satisfies $A_{ii} = 1$ and $A_{ij} = 0$ for $j > i$.

Our goal is to find a simple criterion for δ-singular coordinates. The basic idea is to compare the multiplicative variables assigned by the Pommaret and by the Janet division, respectively. The definitions of these two divisions are very different. Somewhat surprisingly, they yield very similar results.

Proposition 2.4.7. *Let the finite set $\mathcal{N} \subset \mathbb{N}_0^n$ be involutively autoreduced with respect to the Pommaret division. Then for all $\nu \in \mathcal{N}$ the inclusion $N_P(\nu) \subseteq N_{J,\mathcal{N}}(\nu)$ holds.*

Proof. Let $\nu \in \mathcal{N}$ with $\operatorname{cls}\nu = k$. Thus $\nu = [0, \dots, 0, \nu_k, \dots, \nu_n]$ with $\nu_k > 0$. We must show that the indices $1, \dots, k$ are multiplicative for ν with respect to the Janet division. In order to decide whether $k \in N_{J,\mathcal{N}}(\nu)$, we study the set $(\nu_{k+1}, \dots, \nu_n) \subseteq \mathcal{N}$. If it contained a multi index μ with $\mu_k > \nu_k$, the set $\mathcal{N}$ would not be involutively autoreduced with respect to the Pommaret division, as $\nu \mid_P \mu$. Thus k is multiplicative for the Janet division, too. The same argument can be applied for $k-1, \dots, 1$ where sets of the form $(0, \dots, 0, \nu_k, \dots, \nu_n)$ with an increasing number of zeros are considered. Hence $N_P(\nu) \subseteq N_{J,\mathcal{N}}(\nu)$. □

For later use, we note the following simple corollaries which further study the relationship between the Janet and the Pommaret division. Recall that by Remark 2.1.11 any set $\mathcal{N} \subset \mathbb{N}_0^n$ is involutively autoreduced with respect to the Janet division. We show first that by an involutive autoreduction of $\mathcal{N}$ with respect to the *Pommaret* division its *Janet* span can become only larger but not smaller.

Corollary 2.4.8. *Let $\mathcal{N} \subset \mathbb{N}_0^n$ be an arbitrary finite set of multi indices and $\mathcal{N}_P = \mathcal{N} \setminus \{\nu \in \mathcal{N} \mid \exists \mu \in \mathcal{N} : \mu \,|_P\, \nu\}$, i. e. we eliminate all multi indices possessing a Pommaret divisor in $\mathcal{N}$. Then $\langle\mathcal{N}\rangle_J \subseteq \langle\mathcal{N}_P\rangle_J$.*

Proof. If $\mu^{(1)} \,|_P\, \mu^{(2)}$ and $\mu^{(2)} \,|_P\, \nu$, then trivially $\mu^{(1)} \,|_P\, \nu$. Thus for each eliminated multi index $\nu \in \mathcal{N} \setminus \mathcal{N}_P$ another multi index $\mu \in \mathcal{N}_P$ exists with $\mu \,|_P\, \nu$. Let $\operatorname{cls}\mu = k$. By the proposition above $\{1, \dots, k\} \subseteq N_{J,\mathcal{N}_P}(\mu)$. Assume that an index $j > k$ exists with $j \in N_{J,\mathcal{N}}(\nu)$. By definition of the Pommaret division, $\mu_i = \nu_i$ for all $i > k$. Thus $\mu \in (\nu_{j+1}, \dots, \nu_n)$ and $j \in N_{J,\mathcal{N}}(\mu)$. As by the second condition on an involutive division $N_{J,\mathcal{N}}(\mu) \subseteq N_{J,\mathcal{N}_P}(\mu)$ for all $\mu \in \mathcal{N}_P$, we conclude that $j \in N_{J,\mathcal{N}_P}(\mu)$ and $\mathcal{C}_{J,\mathcal{N}}(\nu) \subset \mathcal{C}_{J,\mathcal{N}_P}(\mu)$. But this immediately implies $\langle\mathcal{N}\rangle_J \subseteq \langle\mathcal{N}_P\rangle_J$. □

The second corollary implies that any Pommaret basis is simultaneously a Janet basis. Thus if $\mathcal{H}$ is a Pommaret basis, then $X_{P,\prec}(h) = X_{J,\mathcal{H},\prec}(h)$ for all polynomials $h \in \mathcal{H}$.

Corollary 2.4.9. *Let the finite set $\mathcal{H} \subset \mathcal{P}$ be involutive with respect to the Pommaret division (and some term order). Then $\mathcal{H}$ is also involutive with respect to the Janet division.*

Proof. By the proposition above, it is obvious that the set $\mathcal{H}$ is at least weakly involutive with respect to the Janet division. For the Janet division any weakly involutive set is strongly involutive, as soon as no two elements have the same leading terms (recall Example 2.3.5). But as $\mathcal{H}$ is a Pommaret basis, this cannot happen. □

We show now that for δ-regular coordinate systems the inclusions in Proposition 2.4.7 must be equalities. In other words, if a variable x^ℓ exists which is multiplicative for an element of $\mathcal{F}$ with respect to the Janet division but non-multiplicative with respect to the Pommaret division, then the used coordinates are δ-singular. Our proof is constructive in the sense that it shows us how to find coordinates leading to a larger involutive span.

Theorem 2.4.10. *Let the finite set $\mathcal{F} \subset \mathcal{P}$ be involutively head autoreduced with respect to the Pommaret division and a class respecting term order. If $|\mathcal{F}|_{J,\prec} > |\mathcal{F}|_{P,\prec}$, the coordinates $\mathbf{x}$ are δ-singular for $\mathcal{F}$.*

Proof. By the proposition above, we have $X_P(f) \subseteq X_{J,\mathcal{F}}(f)$ for all $f \in \mathcal{F}$. Assume that for a polynomial $h \in \mathcal{F}$ the strict inclusion $X_P(h) \subset X_{J,\mathcal{F}}(h)$ holds. Thus a variable $x^\ell \in X_{J,\mathcal{F}}(h)$ with $\ell > k = \operatorname{cls} h$ exists. We perform the linear change of variables $x^i = \tilde{x}^i$ for $i \neq k$ and $x^k = \tilde{x}^k + a\tilde{x}^\ell$ with a yet arbitrary parameter $a \in \Bbbk \setminus \{0\}$. This induces the following transformation of the terms $x^\mu \in \mathbb{T}$:

$$x^\mu = \sum_{j=0}^{\mu_k} \binom{\mu_k}{j} a^j \tilde{x}^{\mu - j_k + j_\ell} . \tag{2.20}$$

Let $\mathrm{le}_\prec h = \mu$. Thus $\mu = [0, \dots, 0, \mu_k, \dots, \mu_n]$ with $\mu_k > 0$. We define $\nu = \mu - (\mu_k)_k + (\mu_k)_\ell$; obviously, $\mathrm{cls}\,\nu > k$ by construction. Applying our transformation to the polynomial h leads to a polynomial $\tilde{h}$ containing the term $\tilde{x}^\nu$. Note that ν cannot be an element of $\mathrm{le}_\prec \mathcal{F}$. Indeed, if it was, it would be an element of the same set $(\mu_{\ell+1}, \dots, \mu_n)$ as μ. But this contradicts our assumption that ℓ is multiplicative for the multi index μ with respect to the Janet division, as $\nu_\ell > \mu_\ell$.

Applying the transformation to all polynomials $f \in \mathcal{F}$ yields the set $\tilde{\mathcal{F}}$ on which we perform an involutive head autoreduction in order to obtain the set $\tilde{\mathcal{F}}^\triangle$. We are interested in its involutive size with respect to the Pommaret division. We can choose the parameter a such that after the transformation each polynomial $\tilde{f} \in \tilde{\mathcal{F}}$ has at least the same class as the corresponding polynomial $f \in \mathcal{F}$, as our term order respects classes. This is a simple consequence of (2.20): only for a finite number of values of a cancellations of terms may occur. It follows from the definition of the Pommaret division that if $\mathrm{le}_\prec f_2 \,|_P\, \mathrm{le}_\prec f_1$, then $\mathrm{cls}\,\mathrm{le}_\prec f_2 \geq \mathrm{cls}\,\mathrm{le}_\prec f_1$. Hence even after the involutive head autoreduction the involutive size of $\tilde{\mathcal{F}}^\triangle$ cannot be smaller than that of $\mathcal{F}$.

Now consider again the polynomial h. The leading term of the transformed polynomial $\tilde{h}$ must be greater than or equal to $\tilde{x}^\nu$. Thus its class is greater than k. This remains true even after an involutive head autoreduction with all those polynomials $\tilde{f} \in \tilde{\mathcal{F}}$ that are of class greater than k, as $x^\nu \notin \mathrm{lt}_\prec \mathcal{F}$. Hence the only possibility to obtain a leading term of class less than or equal to k consists of an involutive reduction with respect to a polynomial $\tilde{f} \in \tilde{\mathcal{F}}$ with $\mathrm{cls}\, f \leq k$. But this implies that $\mathrm{cls}\,\mathrm{le}_\prec \tilde{f} > k$. So we may conclude that after the transformation we have at least one polynomial more whose class is greater than k. So the coordinates $\mathbf{x}$ cannot be δ-regular. □

Corollary 2.4.11. *If the coordinates $\mathbf{x}$ are δ-regular for the Pommaret head autoreduced set $\mathcal{F}$, then $\langle \mathcal{F} \rangle_{J,\prec} = \langle \mathcal{F} \rangle_{P,\prec}$ for any class respecting term order $\prec$.*

It is important to note that this corollary provides us only with a necessary but *not* with a sufficient criterion for δ-regularity. In other words, even if the Janet and the Pommaret size are equal for a given set $\mathcal{F} \subset \mathcal{P}$, this does not imply that the used coordinates are δ-regular.

Example 2.4.12. Let $\mathcal{F} = \{\underline{z^2} + yz - x^2,\ \underline{xz} + xy - z,\ \underline{yz} + y^2 + x^2 - x\}$. The underlined terms are the leaders with respect to the degree reverse lexicographic order. One easily checks that the Janet and the Pommaret division yield the same multiplicative variables. If we perform the transformation $\tilde{x} = z$, $\tilde{y} = y + z$ and $\tilde{z} = x$, then after an autoreduction we obtain the set $\tilde{\mathcal{F}}^\triangle = \{\underline{\tilde{z}^2} - \tilde{x}\tilde{y},\ \underline{\tilde{y}\tilde{z}} - \tilde{x},\ \underline{\tilde{y}^2} - \tilde{z}\}$. Again the Janet and the Pommaret division yield the same multiplicative variables, but $|\tilde{\mathcal{F}}^\triangle|_{P,\prec} > |\mathcal{F}|_{P,\prec}$. Thus the coordinates (x, y, z) are not δ-regular for $\mathcal{F}$. $\lhd$

While we have defined δ-regularity as a problem of the Pommaret division, this example shows that in fact other divisions "feel" it, too. Usually they yield in δ-singular coordinates a lower number of multiplicative variables and thus larger involutive bases. δ-regularity is a typical phenomenon in commutative algebra. Many results are valid only in "generic" coordinates, i. e. for coordinates in a Zariski open set.

Besides being necessary for the mere existence of a finite Pommaret basis, a second application of δ-regular coordinates is the construction of $\mathcal{I}$-regular sequences for a homogeneous ideal $\mathcal{I} \subseteq \mathcal{P}$.

Proposition 2.4.13. *Let $\mathcal{I} \subseteq \mathcal{P}$ be a homogeneous ideal and $\mathcal{H}$ a homogeneous Pommaret basis for it with respect to the degree reverse lexicographic order. Let $d = \min_{h\in\mathcal{H}} \operatorname{cls} h$. Then $(x^1, \dots, x^d)$ is a maximal $\mathcal{I}$-regular sequence and* $\operatorname{depth} \mathcal{I} = d$.

Proof. A Pommaret basis $\mathcal{H}$ induces a decomposition of $\mathcal{I}$ of the form

$$\mathcal{I} = \bigoplus_{h\in\mathcal{H}} \Bbbk[x^1, \dots, x^{\operatorname{cls} h}] \cdot h\,. \tag{2.21}$$

If $d = \min_{h\in\mathcal{H}} \operatorname{cls} h$ denotes the minimal class of a generator in $\mathcal{H}$, then (2.21) trivially implies that the sequence $(x^1, \dots, x^d)$ is $\mathcal{I}$-regular.

Let us try to extend this sequence by a variable x^k with $k > d$. We set $\mathcal{H}_d = \{h \in \mathcal{H} \mid \operatorname{cls} h = d\}$ and choose an element $\bar{h} \in \mathcal{H}_d$ of maximal degree. As we use the degree reverse lexicographic order, $\bar{h} \in \langle x^1, \dots, x^d\rangle$ by Lemma B.2.2. Consider the product $x^k\bar{h}$; by construction x^k is non-multiplicative for $\bar{h}$. For each $h \in \mathcal{H}$ a polynomial $P_h \in \Bbbk[x^1, \dots, x^{\operatorname{cls} h}]$ exists such that $x^k\bar{h} = \sum_{h\in\mathcal{H}} P_h h$. No polynomial h with $\operatorname{cls} h > d$ lies in $\langle x^1, \dots, x^d\rangle$ (obviously $\operatorname{lt}_\prec h \notin \langle x^1, \dots, x^d\rangle$). As the leading terms cannot cancel in the sum, $P_h \in \langle x^1, \dots, x^d\rangle$ for all $h \in \mathcal{H} \setminus \mathcal{H}_d$. Thus $x^k\bar{h} = \sum_{h\in\mathcal{H}_d} c_h h + g$ with $c_h \in \Bbbk$ and $g \in \langle x^1, \dots, x^d\rangle\mathcal{I}$. As $\mathcal{I}$ is a homogeneous ideal and as the degree of $\bar{h}$ is maximal in $\mathcal{H}_d$, all constants c_h must vanish.

It is not possible that $\bar{h} \in \langle x^1, \dots, x^d\rangle\mathcal{I}$, as otherwise $\bar{h}$ would be involutively head reducible by some other element of $\mathcal{H}$. Hence we have shown that any variable x^k with $k > d$ is a zero divisor in $\mathcal{I}/\langle x^1, \dots, x^d\rangle\mathcal{I}$ and the $\mathcal{I}$-regular sequence $(x^1, \dots, x^d)$ cannot be extended by any x^k with $k > d$.

Finally, assume that $y^1, \dots, y^{d+1} \in \mathcal{P}_1$ form an $\mathcal{I}$-regular sequence of length $d+1$. We extend them to a basis $\{y^1, \dots, y^n\}$ of $\mathcal{P}_1$ and perform a coordinate transformation $\boldsymbol{x} \mapsto \boldsymbol{y}$. Our basis $\mathcal{H}$ transforms into the set $\mathcal{H}_{\boldsymbol{y}}$ and after an involutive head autoreduction we obtain the set $\mathcal{H}_{\boldsymbol{y}}^{\triangle}$. In general, the coordinates $\boldsymbol{y}$ are not δ-regular for $\mathcal{H}_{\boldsymbol{y}}^{\triangle}$. But there exist coordinates $\tilde{\boldsymbol{y}}$ of the form $\tilde{y}^k = y^k + \sum_{i=1}^{k-1} a_i^k y^i$ with $a_i^k \in \Bbbk$ such that if we transform $\mathcal{H}$ to them and perform afterwards an involutive head autoreduction, they are δ-regular for the obtained set $\tilde{\mathcal{H}}_{\boldsymbol{y}}^{\triangle}$.

This implies that $\tilde{\mathcal{H}}_{\boldsymbol{y}}^{\triangle}$ is a Pommaret basis of the ideal it generates.[2] Thus $\min_{\tilde{h} \in \tilde{\mathcal{H}}_y^{\triangle}} \operatorname{cls} \tilde{h} = d$ and by the same argument as above $\tilde{y}^{d+1}$ is a zero divisor in $\mathcal{I}/\langle \tilde{y}^1, \dots, \tilde{y}^d \rangle \mathcal{I}$. Because of the special form of the transformation from $\boldsymbol{y} \mapsto \tilde{\boldsymbol{y}}$, we have $\langle \tilde{y}^1, \dots, \tilde{y}^d \rangle = \langle y^1, \dots, y^d \rangle$ and y^{d+1} must be a zero divisor in $\mathcal{I}/\langle y^1, \dots, y^d \rangle \mathcal{I}$. But this contradicts the assumption that $(y^1, \dots, y^{d+1})$ is an $\mathcal{I}$-regular sequence and $\operatorname{depth} \mathcal{I} = d$. □

Remark 2.4.14. One may wonder to what extent this result really requires the Pommaret division. Given an arbitrary involutive basis $\mathcal{H}$ of $\mathcal{I}$, we may introduce the set $X_{\mathcal{I}} = \bigcap_{h \in \mathcal{H}} X_{L,\mathcal{H},\prec}(h)$; obviously, for a Pommaret basis $\mathcal{X}_{\mathcal{I}} = \{x^1, \dots, x^d\}$ with $d = \operatorname{depth} \mathcal{I}$. Again it is trivial to conclude from the induced decomposition that any sequence formed by elements of $\mathcal{X}_{\mathcal{I}}$ is $\mathcal{I}$-regular. But in general we cannot claim that these are *maximal* $\mathcal{I}$-regular sequences. Thus only a lower bound for the depth is obtained.

As a simple example we consider the ideal $\mathcal{I}$ generated by $f_1 = z^2 - xy$, $f_2 = yz - x$ and $f_3 = y^2 - z$. If we set $x^1 = x$, $x^2 = y$ and $x^3 = z$, then the set $\mathcal{F} = \{f_1, f_2, f_3\}$ is a Pommaret basis of $\mathcal{I}$ with respect to the degree reverse lexicographic order. By Proposition 2.4.13, (x, y) is a maximal $\mathcal{I}$-regular sequence and $\operatorname{depth} \mathcal{I} = 2$.

If we set $x^1 = z$, $x^2 = y$ and $x^3 = x$ (i.e. we invert the ordering of the variables), then no finite Pommaret basis exists; these coordinates are not δ-regular. In order to obtain a Janet basis $\mathcal{F}_J$ of $\mathcal{I}$ (for the degree reverse lexicographic order with respect to the new ordering of the variables), we must enlarge $\mathcal{F}$ by $f_4 = z^3 - x^2$ and $f_5 = z^3x - x^3$. We find $X_{\mathcal{I}} = \{z\}$, as

$$\begin{aligned} &X_{J,\mathcal{F}_J,\prec_{\text{deglex}}}(f_1) = \{z, y, x\}, \quad X_{J,\mathcal{F}_J,\prec_{\text{deglex}}}(f_2) = \{z\}, \\ &X_{J,\mathcal{F}_J,\prec_{\text{deglex}}}(f_3) = \{z, y\}, \qquad X_{J,\mathcal{F}_J,\prec_{\text{deglex}}}(f_4) = \{z\}, \\ &X_{J,\mathcal{F}_J,\prec_{\text{deglex}}}(f_5) = \{z, x\}. \end{aligned} \tag{2.22}$$

The minimal number of multiplicative variables for a generator is 1 which is less than $\operatorname{depth} \mathcal{I}$. Thus no obvious way exists to deduce $\operatorname{depth} \mathcal{I}$ or a maximal

[2] The involutive sizes of $\mathcal{H}$ and $\tilde{\mathcal{H}}_y^{\triangle}$ are equal. Thus local involution of $\mathcal{H}$ implies local involution of $\tilde{\mathcal{H}}_y^{\triangle}$ and for the Pommaret division local involution is equivalent to involution (cf. Proposition 6.4.5 and Lemma 6.3.5).

$\mathcal{J}$-regular sequence. In particular, it is not possible to extend (z) to a longer $\mathcal{J}$-regular sequence: x is a zero divisor for $\mathcal{J}/z\mathcal{J}$, as $x(xf_2 - f_4) \in z\mathcal{J}$; similarly y, as $y(yf_2 - f_1) \in z\mathcal{J}$; and even for an arbitrary linear combination $ax + by$ with $a, b \in \Bbbk$ we find $(ax + by)\big[b(yf_2 - f_1) - a(xf_2 - f_4)\big] \in z\mathcal{J}$. $\triangleleft$

2.5 Applications I: Combinatorial Decompositions

In the proof of Proposition 2.4.13 we could already see the power of the direct sum decompositions induced by (strong) involutive bases. In this section we want to study this aspect in more details. All results apply to arbitrary finitely generated polynomial modules. But for notational simplicity, we restrict to graded $\Bbbk$-algebras $\mathcal{A} = \mathcal{P}/\mathcal{J}$ for some homogeneous ideal $\mathcal{J} \subseteq \mathcal{P}$ where $\mathcal{P} = \Bbbk[x^1, \dots, x^n]$ is the ordinary polynomial ring. If we speak of a basis of the ideal $\mathcal{J}$, we always assume that it is homogeneous, too.

Recall that the main motivation of Buchberger for the introduction of Gröbner bases was to be able to compute effectively in such factor spaces. Indeed given a Gröbner basis $\mathcal{G}$ of the ideal $\mathcal{J}$, the normal form with respect to $\mathcal{G}$ distinguishes a unique representative in each equivalence class (cf. Remark B.2.9). Our goal in this and the next section is to show that Pommaret bases yield in addition a lot of structural information about the algebra $\mathcal{A}$. More precisely, we want to compute fundamental invariants like the Hilbert polynomial (which immediately yields the Krull dimension and the multiplicity), the depth or the Castelnuovo–Mumford regularity. Our basic tools are combinatorial decompositions of $\mathcal{A}$ into a direct sum of polynomial rings.

Definition 2.5.1. *A* Stanley decomposition *of the graded $\Bbbk$-algebra $\mathcal{A} = \mathcal{P}/\mathcal{J}$ consists of an isomorphism as $\Bbbk$-vector spaces*

$$\mathcal{A} \cong \bigoplus_{t \in \mathcal{T}} \Bbbk[X_t] \cdot t \tag{2.23}$$

with a finite set $\mathcal{T} \subset \mathbb{T}$ and sets $X_t \subseteq \{x^1, \dots, x^n\}$.

We call the elements of X_t again multiplicative variables. As a first, almost trivial application of such decompositions we determine the Hilbert series and the (Krull) dimension.

Proposition 2.5.2. *Let the graded algebra $\mathcal{A}$ possess the Stanley decomposition* (2.23)*. Then its Hilbert series is*

$$\mathcal{H}_{\mathcal{A}}(\lambda) = \sum_{t \in \mathcal{T}} \frac{\lambda^{q_t}}{(1 - \lambda)^{k_t}} \tag{2.24}$$

where $q_t = \deg t$ *and* $k_t = |X_t|$. *Thus the (Krull) dimension of* $\mathcal{A}$ *is given by* $D = \max_{t\in\mathcal{T}}\{k_t\}$ *and the multiplicity (or degree) by the number of terms* $t \in \mathcal{T}$ *with* $k_t = D$.

Proof. The Hilbert series of a polynomial ring $\Bbbk[y^1,\dots,y^d]$ is given by $\mathcal{H}(\lambda) = 1/(1-\lambda)^d$. Indeed, the Taylor coefficient of order q of this function is $\binom{d+q-1}{q}$, the number of monomials of degree q in the ring by (A.4a). If the generator t is of degree q_t, we must shift all coefficients by q_t which is easily achieved by a multiplication with λ^{q_t}. Finally, we may simply add the contributions of the individual generators, as the right hand side of (2.23) is a direct sum. Thus we arrive at (2.24). The statements about the dimension and the multiplicity are simple corollaries. □

Vasconcelos [186, p. 23] calls Stanley decompositions *"an approach that is not greatly useful computationally but it is often nice theoretically"*. One reason for this assessment is surely that so far the main algorithms for their construction work only for monomial ideals $\mathcal{I}$. Thus one must first compute a Gröbner basis of $\mathcal{I}$ for some term order $\prec$ and then, exploiting that as vector spaces $\mathcal{P}/\mathcal{I} \cong \mathcal{P}/\mathrm{lt}_\prec\mathcal{I}$, one determines a Stanley decomposition. Its existence is guaranteed by the following result.

Proposition 2.5.3. *Let* $\mathcal{I} \subseteq \mathbb{N}_0^n$ *be a monoid ideal and* $\bar{\mathcal{I}} = \mathbb{N}_0^n \setminus \mathcal{I}$ *its complementary set. There exists a finite set* $\bar{\mathcal{N}} \subset \bar{\mathcal{I}}$ *and for each multi index* $\nu \in \bar{\mathcal{N}}$ *a set* $\emptyset \subseteq N_\nu \subseteq \{1,\dots,n\}$ *such that*

$$\bar{\mathcal{I}} = \bigcup_{\nu\in\bar{\mathcal{N}}} \left(\nu + \mathbb{N}^n_{N_\nu}\right) \tag{2.25}$$

and $(\nu + \mathbb{N}^n_{N_\nu}) \cap (\mu + \mathbb{N}^n_{N_\mu}) = \emptyset$ *for all* $\mu, \nu \in \bar{\mathcal{N}}$.

Proof. The assertion is trivial for the zero ideal, so we exclude this case and proceed by an induction on the number n. The case $n = 1$ is trivial, as any monoid ideal in $\mathbb{N}_0$ is of the form $\mathcal{I} = \langle [q] \rangle$ for some $q \in \mathbb{N}_0$ and hence $\bar{\mathcal{I}} = \{[0],[1],\dots,[q-1]\}$ which is obviously of the form (2.25) with $N_\nu = \emptyset$ for all ν.

Suppose now that the assertion is true for $n-1$. We want to show that it holds for n, too. But before we must introduce some notations. Primed letters always refer to multi indices $\nu' \in \mathbb{N}_0^{n-1}$; unprimed ones to multi indices $\nu \in \mathbb{N}_0^n$. For a given $\nu \in \mathbb{N}_0^n$, we define $\nu' = [\nu_1,\dots,\nu_{n-1}] \in \mathbb{N}_0^{n-1}$, i.e. we simply drop the last entry. Finally, we write $[\nu', q]$ for the multi index $[\nu_1,\dots,\nu_{n-1},q] \in \mathbb{N}_0^n$.

Given a monoid ideal $\mathcal{I} \subseteq \mathbb{N}_0^n$, we define for all $q \in \mathbb{N}_0$ the monoid ideals $\mathcal{I}'_q = \left\{\nu' \in \mathbb{N}_0^{n-1} \mid [\nu', q] \in \mathcal{I}\right\}$ in $\mathbb{N}_0^{n-1}$. Thus we "slice" $\mathcal{I}$ by intersecting it with the "hyperplanes" $\nu_n = Const$. Obviously, for $q < r$ we have $\mathcal{J}'_q \subseteq \mathcal{J}'_r$ so that these ideals form an ascending chain. As $\mathbb{N}_0^{n-1}$ is a Noetherian monoid, an integer $q_0 \in \mathbb{N}$ exists such that $\mathcal{J}'_q = \mathcal{J}'_{q_0}$ for all $q \geq q_0$.

According to our inductive hypothesis, the complements $\bar{\mathcal{J}}'_q$ of each of these monoid ideals for $0 \leq q \leq q_0$ can be disjointly decomposed in the form

$$\bar{\mathcal{J}}'_q = \bigcup_{\nu' \in \bar{\mathcal{N}}_q^{(n-1)}} \left(\nu' + \mathbb{N}^{n-1}_{N_{\nu'}}\right) \tag{2.26}$$

with some subsets $\bar{\mathcal{N}}_q^{(n-1)} \subset \mathbb{N}_0^{n-1}$ and $N_{\nu'} \subseteq \{1, \dots, n-1\}$. We set now $\bar{\mathcal{N}}_q^{(n)} = \left\{[\nu', q] \mid \nu' \in \bar{\mathcal{N}}_q^{(n-1)}\right\}$ and $\bar{\mathcal{N}} = \bigcup_{q=0}^{q_0} \bar{\mathcal{N}}_q^{(n)}$. Furthermore, we choose $N_{[\nu',q]} = N_{\nu'}$ for $q < q_0$ and $N_{[\nu',q]} = N_{\nu'} \cup \{n\}$ for $q = q_0$.

Setting $\mathcal{U} = \bigcup_{\nu \in \bar{\mathcal{N}}} (\nu + \mathbb{N}^n_{N_\nu})$, the union is obviously disjoint. We claim that $\mathcal{U} = \bar{\mathcal{J}}$ which implies our proposition. Let $\mu \in \mathcal{U}$, then by construction of the set $\bar{\mathcal{N}}$ we have that $\mu' \in \bar{\mathcal{J}}'_q$ for some $0 \leq q \leq q_0$ and hence by definition of the ideals $\mathcal{J}'_q$, the multi index μ cannot be an element of $\mathcal{J}$. Thus $\mathcal{U} \subseteq \bar{\mathcal{J}}$. Conversely, $\mu \in \bar{\mathcal{J}}$ implies $\mu' \in \bar{\mathcal{J}}'_{\mu_n}$. If $\mu_n \geq q_0$, then $\mathcal{J}'_{\mu_n} = \mathcal{J}'_{q_0}$ and a multi index $\nu' \in \bar{\mathcal{N}}_{q_0}^{(n-1)}$ exists such that $\mu' \in \nu' + \mathbb{N}^{n-1}_{N_{\nu'}}$. Thus $\mu \in \nu + \mathbb{N}^n_\nu$ with $\nu = [\nu', q_0]$. If $\mu_n < q_0$, then a multi index $\nu' \in \bar{\mathcal{N}}_{\mu_n}^{(n-1)}$ exists such that $\mu' \in \nu' + \mathbb{N}^{n-1}_{N_{\nu'}}$ and in this case $\mu \in \nu + \mathbb{N}^n_\nu$ with $\nu = [\nu', \mu_n]$. Thus we have $\bar{\mathcal{J}} \subseteq \mathcal{U}$ and consequently $\mathcal{U} = \bar{\mathcal{J}}$ as claimed. □

One must stress that the decomposition (2.25) is not unique and different decompositions may use sets $\bar{\mathcal{N}}$ of different sizes. The elements in the sets N_ν are again called *multiplicative indices*. The proof above is not completely constructive, as the number q_0 was defined by a Noetherian argument. But it is not difficult to see that we may take $q_0 = \max_{\nu \in \mathcal{N}} \nu_n$ where $\mathcal{N}$ is the minimal basis of the monoid ideal $\mathcal{J}$. Now one can straightforwardly transform our proof into a recursive algorithm for the construction of Stanley decompositions. If a Janet basis of $\mathcal{J}$ is known, an alternative algorithm exists.

Proposition 2.5.4. *Let $\mathcal{N}_J$ be a Janet basis of the monoid ideal $\mathcal{J} \subseteq \mathbb{N}_0^n$. Then the set $\bar{\mathcal{N}} \subset \mathbb{N}_0^n$ in (2.25) may be chosen such that for all $\nu \in \bar{\mathcal{N}}$ the equality $N_\nu = N_{J,\mathcal{N}_J \cup \{\nu\}}(\nu)$ holds.*

Proof. We explicitly construct such a set $\bar{\mathcal{N}}$. Let $q_n = \max_{\mu \in \mathcal{N}_J} \mu_n$. Then we put into $\bar{\mathcal{N}}$ all multi indices $\nu = [0, \dots, 0, q]$ such that $0 \leq q < q_n$ and such that $\mathcal{N}_J$ does not contain any multi index μ with $\mu_n = q$. We set $N_\nu = \{1, \dots, n-1\}$ which obviously is just $N_{J,\mathcal{N}_J \cup \{\nu\}}(\nu)$ according to the definition of the Janet division given in Example 2.1.2.

Let $(d_k, \dots, d_n) \subseteq \mathcal{N}_J$ be one of the subsets (2.2) considered in the assignment of the multiplicative variables to the elements of $\mathcal{N}_J$ according to the definition of the Janet division and set $q_{k-1} = \max_{\mu\in(d_k,\dots,d_n)} \mu_{k-1}$. We enlarge the set $\bar{\mathcal{N}}$ by all those multi indices $\nu = [0,\dots,0,q,d_k,\dots,d_n]$ where $0 \leq q < q_{k-1}$ and where $(d_k,\dots,d_n)$ does not contain any multi index μ with $\mu_{k-1} = q$. The corresponding sets N_ν of multiplicative indices contain the indices $1,\dots,k-1$ and all those indices greater than k that are multiplicative for the elements of $(d_k,\dots,d_n)$. Again it is easy to verify that this entails $N_\nu = N_{J,\mathcal{N}_J\cup\{\nu\}}(\nu)$.

We claim that the thus constructed set $\bar{\mathcal{N}}$ (together with the sets N_ν) defines a disjoint decomposition (2.25) of the complementary set $\bar{\mathcal{I}}$. Like in the proof of the previous proposition we proceed by an induction on n and use again the notations introduced there. The case $n = 1$ is trivial. Suppose that our assertion is true for $n-1$. Let $\rho \in \mathbb{N}_0^n$ be an arbitrary multi index and let $q_1 < q_2 < \cdots < q_r$ represent all values that occur for the entry μ_n in the multi indices $\mu \in \mathcal{N}_J$. We must distinguish three cases depending on the value of ρ_n: (i) $\rho_n < q_r$ but $\rho_n \notin \{q_1,\dots,q_{r-1}\}$, (ii) $\rho_n \in \{q_1,\dots,q_{r-1}\}$, and (iii) $\rho_n \geq q_r$.

In the first case, $\bar{\mathcal{N}}$ contains the multi index $\nu = [0,\dots,0,\rho_n]$ and, as $N_\nu = \{1,\dots,n-1\}$, obviously $\rho \in \nu + \mathbb{N}^n_{N_\nu}$. It is easy to see that ρ cannot be an element of $\mathcal{I}$ or of the cone $\bar{\nu} + \mathbb{N}^n_{N_{\bar{\nu}}}$ of any other multi index $\bar{\nu} \in \bar{\mathcal{N}}$.

In the second case, we "slice" the sets $\mathcal{N}_J$ and $\bar{\mathcal{N}}$ at degree ρ_n and obtain thus the two sets $\mathcal{N}'_J = \{\mu' \in \mathbb{N}_0^{n-1} \mid \mu \in \mathcal{N}_J,\ \mu_n = \rho_n\} \subset \mathbb{N}_0^{n-1}$ and $\bar{\mathcal{N}}' = \{\nu' \in \mathbb{N}_0^{n-1} \mid \nu \in \bar{\mathcal{N}},\ \nu_n = \rho_n\} \subset \mathbb{N}_0^{n-1}$, respectively. If we compute the sets $N_{J,\mathcal{N}'_J}(\mu')$ of multiplicative indices for the elements $\mu' \in \mathcal{N}'_J$, it is straightforward to verify that they are just $N_{J,\mathcal{N}_J}(\mu)$, as μ_n is not maximal and $\mu \in (\rho_n)$, if and only if $\mu' \in \mathcal{N}'_J$. Furthermore, $\mathcal{N}'_J$ is a Janet basis of the monoid ideal $\mathcal{I}'_{\rho_n}$. If we apply the procedure above to this ideal in $\mathbb{N}_0^{n-1}$, we obtain $\bar{\mathcal{N}}'$ as complementary basis and the sets of multiplicative variables remain unchanged.

By our inductive hypothesis, $\bar{\mathcal{N}}'$ defines a disjoint decomposition of the sought form of the set $\bar{\mathcal{I}}'_{\rho_n}$. If $\rho \in \bar{\mathcal{I}}$, then $\rho' \in \bar{\mathcal{I}}'_{\rho_n}$ and a multi index $\nu' \in \bar{\mathcal{N}}'$ exists such that $\rho' \in \nu' + \mathbb{N}^{n-1}_{N_{\nu'}}$. By construction, $\nu = [\nu',\rho_n] \in \bar{\mathcal{N}}$ and $\rho \in \nu + \mathbb{N}^n_{N_\nu}$. If $\rho \in \mathcal{I}$, then $\rho' \in \mathcal{I}'_{\rho_n}$, and it is not possible to find a multi index $\nu' \in \bar{\mathcal{N}}'$ such that $\rho' \in \nu' + \mathbb{N}^{n-1}_{N_{\nu'}}$. Hence $\rho \notin \nu + \mathbb{N}^n_{N_\nu}$ for any $\nu \in \bar{\mathcal{N}}$.

The third case is very similar to the second one. For the definition of $\mathcal{N}'_J$ and $\bar{\mathcal{N}}'$ we consider only multi indices where the value of the last entry is q_r. Note that for all multi indices $\nu \in \bar{\mathcal{N}}$ that contribute to $\bar{\mathcal{N}}'$ the index n is multiplicative. If $\rho \in \bar{\mathcal{I}}$, then $\rho' \in \bar{\mathcal{I}}'_{\rho_n}$ and a multi index $\nu' \in \bar{\mathcal{N}}'$ exists such that $\rho' \in \nu' + \mathbb{N}^{n-1}_{N_{\nu'}}$ and we conclude as above that $\rho \in \nu + \mathbb{N}^n_{N_\nu}$ for $\nu = [\nu', q_r] \in \bar{\mathcal{N}}$. Again it is obvious that for a multi index $\rho \in \mathcal{I}$, it is not possible to find such a $\nu \in \bar{\mathcal{N}}$. □

We have formulated this proposition only for the case that $\mathcal{N}_J$ is a Janet basis, but our proof yields an even stronger result. Let $\mathcal{N}_J$ be an *arbitrary* subset of $\mathbb{N}_0^n$ and construct the corresponding set $\bar{\mathcal{N}}$. Then we may substitute everywhere in the proof $\mathcal{I}$ by the involutive span $\langle\mathcal{N}_J\rangle_J$ and still obtain that any multi index $\rho \in \mathbb{N}_0^n$ either lies in $\langle\mathcal{N}_J\rangle_J$ or in exactly one of the cones defined by $\bar{\mathcal{N}}$ and the sets N_ν.

Example 2.5.5. Consider the simple example $\mathcal{N}_J = \{[1,1]\}$. Our construction yields $\nu^{(1)} = [0,0]$ with $N_{\nu^{(1)}} = \{1\}$ and $\nu^{(2)} = [0,1]$ with $N_{\nu^{(2)}} = \{2\}$. Obviously, the corresponding cones disjointly decompose the complementary set $\bar{\mathcal{I}} = \{[i,j] \in \mathbb{N}_0^2 \mid ij = 0\}$. ◁

According to Corollary 2.4.9, we may apply the same construction to Pommaret bases. But as the Pommaret division is globally defined, it is very simple to give an alternative decomposition.

Proposition 2.5.6. *Let $\mathcal{N}_P \subset \mathbb{N}_0^n$ be a Pommaret basis of the monoid ideal $\mathcal{I} \subseteq \mathbb{N}_0^n$ and $q = \max_{\nu\in\mathcal{N}_P} |\nu|$. If we introduce $\bar{\mathcal{N}}_0 = \{\nu \in \bar{\mathcal{I}} \mid |\nu| < q\}$ and $\bar{\mathcal{N}}_1 = \{\nu \in \bar{\mathcal{I}} \mid |\nu| = q\}$, then*

$$\bar{\mathcal{I}} = \bar{\mathcal{N}}_0 \cup \bigcup_{\nu\in\bar{\mathcal{N}}_1} \mathcal{C}_P(\nu)\,. \tag{2.27}$$

Proof. The definition of the Pommaret division implies the identity

$$(\mathbb{N}_0^n)_{\geq q} = \bigcup_{|\nu|=q} \mathcal{C}_P(\nu) \tag{2.28}$$

from which the proposition follows trivially. Here $(\mathbb{N}_0^n)_{\geq q}$ denotes the set of all multi indices of length greater than or equal to q. By the definition of an involutive division, the union on the right hand side is disjoint. □

Example 2.5.7. The decomposition (2.27) is usually redundant. Considering for $\bar{\mathcal{N}}_1$ only multi indices of length q makes the formulation much easier but it is not optimal. Consider the trivial example $\mathcal{N}_P = \{[0,1]\}$. According to Proposition 2.5.6 we should set $\bar{\mathcal{N}}_0 = \{[0,0]\}$ and $\bar{\mathcal{N}}_1 = \{[1,0]\}$. But obviously $\bar{\mathcal{I}} = [0,0] + \mathbb{N}^2_{\{1\}}$. Applying the construction used in the proof of Proposition 2.5.4 yields directly this more compact form. ◁

A very interesting special type of Stanley decompositions are *Rees decompositions.* Here the sets X_t are of the form $X_t = \{x^1, \dots, x^{\operatorname{cls} t}\}$ where the number $\operatorname{cls} t$ is called the *class*[3] of the generator t. It is no coincidence that we use here the same terminology as in the Pommaret division. By Proposition 2.5.6, a Pommaret basis of the ideal $\mathcal{I} \subseteq \mathcal{P}$ immediately induces a Rees

[3] Some authors also use the term *level.*

decomposition of the algebra $\mathcal{P}/\mathcal{I}$. If $t = x^\mu$ with $\mu \in \bar{\mathcal{N}}_1$, then indeed its class is just the usual one of its multi index. As an application, we show that given a Pommaret basis of $\mathcal{I}$, we can easily read off the dimension and the depth of $\mathcal{A} = \mathcal{P}/\mathcal{I}$.

Proposition 2.5.8. *Let $\mathcal{H}$ be a Pommaret basis of degree q of the homogeneous ideal $\mathcal{I} \subseteq \mathcal{P}$. The dimension of the algebra $\mathcal{A} = \mathcal{P}/\mathcal{I}$ is given by $D = \min\{i \mid \langle \mathcal{H}, x^1, \dots, x^i\rangle_q = \mathcal{P}_q\}$.*

Proof. The Hilbert polynomials of $\mathcal{A}$ and the truncation $\mathcal{A}_{\geq q}$ coincide. Thus it suffice to consider the latter algebra. By Lemma 2.4.1, a Pommaret basis of $\mathcal{I}_{\geq q}$ is given by the set $\mathcal{H}_q$ determined in (2.18). If D is the smallest number such that $\langle \mathcal{H}_q, x^1, \dots, x^D\rangle_q = \mathcal{P}_q$, then all multi indices $\nu \in \mathbb{N}_0^n$ with $|\nu| = q$ and $\operatorname{cls}\nu > D$ lie in $\operatorname{le}_\prec \mathcal{H}_q$ but a multi index $\mu \in \mathbb{N}_0^n$ exists such that $|\mu| = q$, $\operatorname{cls}\mu = D$ and $\mu \neq \operatorname{le}_\prec \mathcal{H}_q$. By Proposition 2.5.6 this entails that μ is a generator of class D of the decomposition (2.27) and that the decomposition does not contain a generator of higher class. But this trivially implies that $\dim \mathcal{A} = D$. $\square$

In a terminology apparently introduced by Gröbner [65, Section 131], a subset $X_\mathcal{I} \subseteq \{x^1, \dots, x^n\}$ is *independent modulo* $\mathcal{I}$, if $\mathcal{I} \cap \mathbb{k}[X_\mathcal{I}] = \{0\}$ and $\dim \mathcal{I}$ is the maximal size of an independent set. This approach to the dimension of an ideal has been taken up again by Kredel and Weispfenning [97] using Gröbner bases (see also [15, Sections 6.3 and 9.3]).

Corollary 2.5.9. *Under the assumptions of Proposition 2.5.8, $\{x^1, \dots, x^D\}$ is a maximal independent set modulo the ideal $\mathcal{I}$.*

Proof. Let $d = \min_{h \in \mathcal{H}} \operatorname{cls} h$. Then $\{x^1, \dots, x^{d-1}\}$ is trivially an independent set. Indeed, since $\operatorname{lt}_\prec \mathcal{H} \subset \mathbb{k}[x^d, \dots, x^n]$ and $\operatorname{lt}_\prec \mathcal{I} = \langle \operatorname{lt}_\prec \mathcal{H}\rangle$, no non-zero element of $\mathcal{I}$ has a leading term in $\mathbb{k}[x^1, \dots, x^{d-1}]$. Obviously, $D \geq d$.

As above, it suffices to consider the truncation $\mathcal{I}_{\geq q}$. Assume a non-zero polynomial $f \in \mathcal{I}_{\geq q} \cap \mathbb{k}[x^1, \dots, x^D]$ exists. As $\mathcal{H}_q$ is a Pommaret basis of the ideal $\mathcal{I}_{\geq q}$, it must contain a polynomial h such that $\operatorname{le}_\prec h \mid_P \operatorname{le}_\prec f$; this implies $\operatorname{lt}_\prec h \in \mathbb{k}[x^d, \dots, x^D]$. Repeated applications of Lemma 2.4.2 to $\operatorname{le}_\prec h$ with $j = D$ yields that $\nu = q_D \in \operatorname{le}_\prec \mathcal{H}_q$. Further applications of Lemma 2.4.2 to ν with $k = D$ show that in fact any multi index of length q and class D is contained in $\operatorname{le}_\prec \mathcal{H}_q$. But thus already $\langle \mathcal{H}, x^1, \dots, x^{D-1}\rangle_q = \mathcal{P}_q$ in contradiction to the characterisation of D in Proposition 2.5.8. $\square$

As our proof is solely based on an analysis of the leading terms, we showed in fact that $\{x^1, \dots, x^D\}$ is a *strongly* independent set modulo $\mathcal{I}$ and the term order $\prec$, i.e. that $\mathbb{k}[x^1, \dots, x^D] \cap \operatorname{lt}_\prec \mathcal{I} = \{0\}$ [15, Definition 9.22].

Proposition 2.5.10. *Let $\mathcal{H}$ be a Pommaret basis of the homogeneous ideal $\mathcal{I} \subseteq \mathcal{P}$ for the degree reverse lexicographic order and $d = \min_{h\in\mathcal{H}} \operatorname{cls} h$. The depth of $\mathcal{A} = \mathcal{P}/\mathcal{I}$ is $d-1$.*

Proof. We claim that $(x^1, \dots, x^{d-1})$ is a maximal $\mathcal{A}$-regular sequence. Let $f \in \mathcal{P}$ be an arbitrary polynomial. If $x^j f \in \mathcal{I}$ for an index $1 \leq j < d$, then we must have $f \in \mathcal{I}$, as no $h \in \mathcal{H}$ is divisible by x^j because of the condition on the classes. Indeed, in the involutive standard representation $x^j f = \sum_{h\in\mathcal{H}} P_h h$ all polynomials P_h must lie in $\langle x^j\rangle$ and dividing by x^j we obtain the standard representation of f. Hence none of the variables in $(x^1, \dots, x^{d-1})$ is a zero divisor for $\mathcal{A}$. The same argument shows that $x^j f \in \langle x^1, \dots, x^k\rangle\mathcal{I}$ for an index $k < j < d$ implies that $f \in \langle x^1, \dots, x^k\rangle\mathcal{I}$ and none of the variables in $(x^{k+1}, \dots, x^{d-1})$ is a zero divisor in $\mathcal{A}/\langle x^1, \dots, x^k\rangle\mathcal{A}$, as we may identify $\langle x^1, \dots, x^k\rangle\mathcal{A} \cong \langle x^1, \dots, x^k\rangle/\langle x^1, \dots, x^k\rangle\mathcal{I}$. Hence $(x^1, \dots, x^{d-1})$ is an $\mathcal{A}$-regular sequence.

Now we show that it is not possible to extend this sequence by any x^k with $k \geq d$. We identify $\mathcal{A}/\langle x^1, \dots, x^{d-1}\rangle\mathcal{A}$ with $\Bbbk[x^d, \dots, x^n]/\bar{\mathcal{I}}$ where the homogeneous ideal $\bar{\mathcal{I}} = \{f \in \Bbbk[x^d, \dots, x^n] \mid \exists g \in \langle x^1, \dots, x^{d-1}\rangle : f + g \in \mathcal{I}\}$ is isomorphic to $\mathcal{I}/\langle x^1, \dots, x^{d-1}\rangle$. If we drop in each element of $\mathcal{H}$ all terms contained in $\langle x^1, \dots, x^{d-1}\rangle$, we obtain a basis $\bar{\mathcal{H}}$ of $\bar{\mathcal{I}}$. As this operation does not affect the leading terms of the generators, $\bar{\mathcal{H}}$ is still a Pommaret basis.

Let $\bar{h}_d \in \bar{\mathcal{H}}$ be a polynomial of class d and of maximal degree among all generators of this class. As the degree reverse lexicographic order respects classes, it is of the form $\bar{h}_d = x^d \hat{h}_d$. It is not possible that $\hat{h}_d \in \bar{\mathcal{I}}$, as otherwise $\bar{\mathcal{H}}$ would not be involutively head autoreduced. Thus x^d is a zero divisor in $\mathcal{A}/\langle x^1, \dots, x^{d-1}\rangle\mathcal{A}$, as $x^d\hat{h}_d \in \bar{\mathcal{I}}$.

Now consider x^k with $k > d$. Obviously, it is a non-multiplicative variable for $\bar{h}_d$. Thus for each generator $\bar{h} \in \bar{\mathcal{H}}$ a polynomial $P_{\bar{h}} \in \Bbbk[x^d, \dots, x^{\operatorname{cls}\bar{h}}]$ exists such that $x^k\bar{h}_d = \sum_{\bar{h}\in\bar{\mathcal{H}}} P_{\bar{h}}\bar{h}$. As in the proof of Proposition 2.4.13, one easily deduces from the fact that the leading terms cannot cancel that $P_{\bar{h}} \in \langle x^d\rangle$ for all $\bar{h}$ with $\operatorname{cls}\bar{h} > d$. If $\operatorname{cls}\bar{h} = d$, then $P_{\bar{h}} \in \Bbbk[x^d]$. As $\bar{h}_d$ is of maximal degree among all generators of class d and $\bar{\mathcal{I}}$ is a homogeneous ideal, we must have $P_{\bar{h}} \in \langle x^d\rangle$, too. Thus we may divide by x^d and obtain $x^k\hat{h}_d = \sum_{\bar{h}\in\bar{\mathcal{H}}}(P_{\bar{h}}/x^d)\bar{h} \in \bar{\mathcal{I}}$. But this implies that x^k is a zero divisor in $\mathcal{A}/\langle x^1, \dots, x^{d-1}\rangle\mathcal{A}$.

Finally, assume that $y^1, \dots, y^d \in \mathcal{P}_1$ form an $\mathcal{A}$-regular sequence. We extend to a basis $\{y^1, \dots, y^n\}$ of $\mathcal{P}_1$ and perform a change of coordinates $\mathbf{x} \mapsto \mathbf{y}$. This transforms $\mathcal{H}$ into $\tilde{\mathcal{H}}$ and after an involutive head autoreduction we obtain the set $\tilde{\mathcal{H}}^{\triangle}$. Let $c = \min_{\tilde{h}\in\tilde{\mathcal{H}}^{\triangle}} \operatorname{cls}\tilde{h}$. One proves in the same manner as above that y^c is a zero divisor in $\mathcal{A}/\langle y^1, \dots, y^{c-1}\rangle\mathcal{A}$ and thus that $c > d$. This implies that the original coordinates $\mathbf{x}$ are not δ-regular for $\mathcal{H}$ contradicting Proposition 2.4.5. Hence $\operatorname{depth}\mathcal{A} = d - 1$. $\square$

As a trivial corollary we obtain that for any graded algebra $\mathcal{A} = \mathcal{P}/\mathcal{I}$ the inequality $\operatorname{depth}\mathcal{A} \leq \dim\mathcal{A}$ holds. In the limit case $\operatorname{depth}\mathcal{A} = \dim\mathcal{A}$, the algebra is by definition *Cohen–Macaulay*. Thus we obtain the following characterisation of algebras of this type.

Theorem 2.5.11. *Let $\mathcal{H}$ be a Pommaret basis of degree q of the homogeneous ideal $\mathcal{I} \subseteq \mathcal{P}$ with respect to the degree reverse lexicographic order and $d = \min_{h\in\mathcal{H}} \operatorname{cls} h$. The algebra $\mathcal{A} = \mathcal{P}/\mathcal{I}$ is Cohen–Macaulay, if and only if $\langle \mathcal{H}, x^1, \dots, x^{d-1}\rangle_q = \mathcal{P}_q$.*

Proof. By the results above, $\operatorname{depth}\mathcal{A} = \dim\mathcal{A} = d - 1$. □

Corollary 2.5.12 (Hironaka criterion). *$\mathcal{A} = \mathcal{P}/\mathcal{I}$ is a Cohen–Macaulay algebra, if and only if a Rees decomposition of $\mathcal{A}$ exists where all generators have the same class.*

Proof. One direction is trivial. If such a special decomposition exists with d the common class of all generators, then obviously both the dimension and the depth of $\mathcal{A}$ is d and thus $\mathcal{A}$ is Cohen–Macaulay.

For the converse, let us assume that $\mathcal{A}$ is Cohen–Macaulay and that $\dim\mathcal{A} = \operatorname{depth}\mathcal{A} = d$. Let $\mathcal{H}$ be a Pommaret basis of $\mathcal{I}$ with respect to the degree reverse lexicographic order. We will show later that in suitably chosen variables $\mathbf{x}$ such a basis always exists (Corollary 6.4.7). By our results above, this implies that $\min_{h\in\mathcal{H}} \operatorname{cls} h = d + 1$. We introduce the set $\bar{\mathcal{N}} = \{\nu \in \mathbb{N}_0^n \setminus \langle \operatorname{le}_\prec \mathcal{H}\rangle \mid \operatorname{cls}\nu > d\}$ (recall that by convention we defined $\operatorname{cls}[0,\dots,0] = n$ so that $[0,\dots,0] \in \bar{\mathcal{N}}$ whenever $\mathcal{I} \neq \mathcal{P}$). $\bar{\mathcal{N}}$ is finite, as all its elements satisfy $|\nu| < \deg\mathcal{H}$, and we claim that

$$\mathcal{A} \cong \bigoplus_{\nu\in\bar{\mathcal{N}}} \mathbb{k}[x^1,\dots,x^d]\cdot x^\nu \,. \tag{2.29}$$

In fact, (2.29) is just the decomposition obtained by applying the algorithm implicitly contained in the proof of Proposition 2.5.4. Consider an element $\nu \in \bar{\mathcal{N}}$; obviously, it is of the form $\nu = [0,\dots,0,\nu_{d+1},\dots,\nu_n]$ with $\sum_{i=d+1}^n \nu_i < q = \deg\mathcal{H}$. If we set $q' = q - \sum_{i=d+2}^n \nu_i$, then by Theorem 2.5.11 the multi index $[0,\dots,0,q',\nu_{d+2},\dots,\nu_n]$ lies in the monoid ideal $\langle \operatorname{le}_\prec\mathcal{H}\rangle$. But this implies the existence of a multi index $\nu' \in \operatorname{le}_\prec\mathcal{H}$ with $\nu' = [0,\dots,0,\nu'_{d+1},\nu_{d+2},\dots,\nu_n]$ with $\nu_{d+1} < \nu'_{d+1} \leq q'$. Hence the set $(\nu_{d+2},\dots,\nu_n)$ is considered in the assignment of multiplicative variables to the elements of $\mathcal{H}$ with respect to the Janet division and it consists only of the multi index ν', as $\mathcal{H}$ is involutively head autoreduced (with respect to the Pommaret division). But this implies that our algorithm chooses ν as an element of $\bar{\mathcal{N}}$ and assigns to it the multiplicative variables $x^1,\dots,x^d$.

The algorithm cannot lead to a larger set $\bar{\mathcal{N}}$, as any further multi index would be of class less than or equal to d and thus be contained in $\Bbbk[x^1,\dots,x^d]\cdot 1$. But we know that the sets are disjoint, so that this cannot happen and we get the decomposition (2.29). □

Example 2.5.13. Consider the ideal $\mathcal{I}=\langle z^3, yz^2-xz^2, y^2-xy\rangle \subset \Bbbk[x,y,z]$. A Pommaret basis of $\mathcal{I}$ for the degree reverse lexicographic order is given by $\mathcal{H}=\{z^3, yz^2-xz^2, y^2z-xyz, y^2-xy\}$ and thus both the depth and the dimension of $\mathcal{P}/\mathcal{I}$ is 1. Following the proof of Proposition 2.5.4 we obtain $\mathcal{T}=\{1,y,z,yz,z^2\}$ and the Hironaka decomposition is

$$\mathcal{P}/\mathcal{I}\cong \Bbbk[x]\oplus\Bbbk[x]\cdot y\oplus\Bbbk[x]\cdot z\oplus\Bbbk[x]\cdot yz\oplus\Bbbk[x]\cdot z^2\,. \tag{2.30}$$

The term z^2 comes from the set $(2)=\{[0,1,2]\}$. The terms yz and z arise from $(1)=\{[0,2,1]\}$ and finally 1 and y stem from $(0)=\{[0,2,0]\}$. The set $(3)=\{[0,0,3]\}$ does not contribute to the decomposition. ◁

2.6 Applications II: Syzygies and Free Resolutions

Appendix B.3 shows that Gröbner bases are very useful in syzygy theory; the Schreyer Theorem B.3.3 gives us even directly a Gröbner basis of the first syzygy module. In this section we want to consider the use of involutive bases for this problem. In particular, Pommaret bases allow us to make statements about the whole syzygy resolution. We consider only the ordinary polynomial ring $\mathcal{P}=\Bbbk[x^1,\dots,x^n]$ and free modules $\mathcal{P}^m$ over it.

In this section we cannot avoid the explicit use of modules and we use the vector notation introduced in the Appendices B.2 and B.3. Let $\mathcal{H}\subset\mathcal{P}^m$ be a finite set, $\prec$ a term order on $\mathbb{T}_m$ and L an involutive division on $\mathbb{N}_0^n$. We divide $\mathcal{H}$ into m disjoint sets $\mathcal{H}_\alpha=\{\mathbf{h}\in\mathcal{H}\mid \operatorname{lt}_\prec\mathbf{h}=t\mathbf{e}_\alpha,\ t\in\mathbb{T}\}$. This leads naturally to m sets $\mathcal{N}_\alpha=\{\mu\in\mathbb{N}_0^n\mid x^\mu\mathbf{e}_\alpha\in\operatorname{lt}_\prec\mathcal{H}_\alpha\}$. If $\mathbf{h}\in\mathcal{H}_\alpha$, we assign the multiplicative variables $X_{L,\mathcal{H},\prec}(\mathbf{h})=\{x^i\mid i\in N_{L,\mathcal{N}_\alpha}(\operatorname{le}_\prec\mathbf{h})\}$. The involutive span $\langle\mathcal{H}\rangle_{L,\prec}$ is defined by an obvious generalisation of (2.16).

Let $\mathcal{M}\subseteq\mathcal{P}^m$ be a submodule and $\mathcal{H}$ an involutive basis of $\mathcal{M}$, i.e. $\langle\mathcal{H}\rangle_L=\mathcal{M}$ and $\mathcal{H}$ is involutively head autoreduced. Take an arbitrary element $\mathbf{h}_\alpha\in\mathcal{H}$ and choose a non-multiplicative variable $x^k\in\bar{X}_{L,\mathcal{H},\prec}(\mathbf{h}_\alpha)$. By Proposition 2.3.10, we can determine for each $\mathbf{h}_\beta\in\mathcal{H}$ with an involutive normal form algorithm a unique polynomial $P_\beta^{(\alpha;k)}\in\Bbbk[X_{L,\mathcal{H},\prec}(\mathbf{h}_\beta)]$ such that $x^k\mathbf{h}_\alpha=\sum_{\beta=1}^s P_\beta^{(\alpha;k)}\mathbf{h}_\beta$. The corresponding syzygy is

$$\mathbf{S}_{\alpha;k}=x^k\mathbf{e}_\alpha-\sum_{\beta=1}^s P_\beta^{(\alpha;k)}\mathbf{e}_\beta\in\mathcal{P}^s\,. \tag{2.31}$$

We claim now that the set of all these syzygies $\mathbf{S}_{\alpha;k}$ generates the syzygy module $\mathrm{Syz}(\mathcal{H})$. Before the proof, we need a simple technical lemma.

Lemma 2.6.1. *Let* $\mathbf{S} = \sum_{\beta=1}^{s} S_\beta \boldsymbol{e}_\beta$ *be an arbitrary syzygy in* $\mathrm{Syz}(\mathcal{H})$*. Then* $S_\beta \in \Bbbk[X_{L,\mathcal{H},\prec}(\mathbf{h}_\beta)]$ *for all* $1 \leq \beta \leq s$*, if and only if* $\mathbf{S} = 0$*.*

Proof. By definition of a syzygy, $\sum_{\beta=1}^{s} S_\beta \mathbf{h}_\beta = 0$. As $\mathcal{H}$ is involutively head autoreduced, each element $\mathbf{f} \in \langle\mathcal{H}\rangle$ possesses a unique involutive normal form by Proposition 2.3.7. In particular, this holds for $0 \in \langle\mathcal{H}\rangle$. Thus either $S_\beta \notin \Bbbk[X_{L,\mathcal{H},\prec}(\mathbf{h}_\beta)]$ for at least one β or $\mathbf{S} = 0$. □

We recall from Appendix B.3 the definition of the term order $\prec_{\mathcal{F}}$ on $\mathbb{T}_s$ induced by an arbitrary finite set $\mathcal{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_s\} \subset \mathcal{P}^m$ and an arbitrary term order $\prec$ on $\mathbb{T}_m$. Given two terms $\mathbf{s} = s\boldsymbol{e}_\sigma$ and $\mathbf{t} = t\boldsymbol{e}_\tau$, we set $\mathbf{s} \prec_{\mathcal{F}} \mathbf{t}$, if either $\mathrm{lt}_\prec(s\mathbf{f}_\sigma) \prec \mathrm{lt}_\prec(t\mathbf{f}_\tau)$ or $\mathrm{lt}_\prec(s\mathbf{f}_\sigma) = \mathrm{lt}_\prec(t\mathbf{f}_\tau)$ and $\tau < \sigma$.

Proposition 2.6.2. *The set* $\mathcal{H}_{\mathrm{Syz}} = \left\{\mathbf{S}_{\alpha;k} \mid 1 \leq \alpha \leq s;\ x^k \in \bar{X}_{L,\mathcal{H},\prec}(\mathbf{h}_\alpha)\right\}$ *is a generating set of the syzygy module* $\mathrm{Syz}(\mathcal{H})$*.*

Proof. This result is a simple corollary to Schreyer's Theorem and Buchberger's second criterion in the form of Lemma B.3.4. Without loss of generality, we assume that $\mathcal{H}$ is a monic basis, i. e. all leading coefficients are 1. Let $\mathbf{S}_{\alpha;k} \in \mathcal{H}_{\mathrm{Syz}}$. By Proposition 2.3.10 the polynomials $P_\beta^{(\alpha;k)}$ in (2.31) satisfy $\mathrm{lt}_\prec(P_\beta^{(\alpha;k)}\mathbf{h}_\beta) \preceq \mathrm{lt}_\prec(x^k\mathbf{h}_\alpha)$. By Lemma 2.3.6 there exists one and only one index $\bar\beta$ such that $\mathrm{lt}_\prec(P_{\bar\beta}^{(\alpha;k)}\mathbf{h}_{\bar\beta}) = \mathrm{lt}_\prec(x^k\mathbf{h}_\alpha)$.

Set $\mathbf{t}_\alpha = \mathrm{lt}_\prec\mathbf{h}_\alpha$ and $\mathbf{t}_{\alpha\beta} = \mathrm{lcm}(\mathbf{t}_\alpha, \mathbf{t}_\beta)$. As an involutive basis is simultaneously a Gröbner basis, we may write all S-polynomials in the form $\mathbf{S}_\prec(\mathbf{h}_\alpha, \mathbf{h}_\beta) = \sum_{\gamma=1}^{s} f_{\alpha\beta\gamma}\mathbf{h}_\gamma$ with some polynomials $f_{\alpha\beta\gamma} \in \mathcal{P}$. In particular, we can choose these polynomials such that $\mathrm{lt}_\prec\big(\mathbf{S}_\prec(\mathbf{h}_\alpha, \mathbf{h}_\beta)\big) \succeq \mathrm{lt}_\prec(f_{\alpha\beta\gamma}\mathbf{h}_\gamma)$ for $1 \leq \gamma \leq s$. As in Appendix B.3, we introduce for $\alpha \neq \beta$ the syzygies

$$\mathbf{S}_{\alpha\beta} = \frac{\mathbf{t}_{\alpha\beta}}{\mathbf{t}_\alpha}\boldsymbol{e}_\alpha - \frac{\mathbf{t}_{\alpha\beta}}{\mathbf{t}_\beta}\boldsymbol{e}_\beta - \mathbf{f}_{\alpha\beta}\,. \tag{2.32}$$

The set $\mathcal{H}_{\mathrm{Schreyer}} = \{\mathbf{S}_{\alpha\beta} \mid 1 \leq \alpha < \beta \leq s\}$ of all these syzygies generates the syzygy module $\mathrm{Syz}(\mathcal{H})$ by the Schreyer Theorem B.3.3. It is easy to see that $\mathbf{S}_{\alpha;k} = \mathbf{S}_{\alpha\bar\beta}$. Thus $\mathcal{H}_{\mathrm{Syz}} \subseteq \mathcal{H}_{\mathrm{Schreyer}}$.

Let $\mathbf{S}_{\alpha\beta} \in \mathcal{H}_{\mathrm{Schreyer}} \setminus \mathcal{H}_{\mathrm{Syz}}$ be an arbitrary syzygy. We prove first that $\mathcal{H}_{\mathrm{Schreyer}} \setminus \{\mathbf{S}_{\alpha\beta}\}$ is still a generating set of $\mathrm{Syz}(\mathcal{H})$. Any syzygy in $\mathcal{H}_{\mathrm{Schreyer}}$ has the form $\mathbf{S}_{\alpha\beta} = x^\mu\boldsymbol{e}_\alpha - x^\nu\boldsymbol{e}_\beta + \mathbf{R}_{\alpha\beta}$. By construction, one of the monomials x^μ and x^ν must contain a non-multiplicative variable x^k for $\mathbf{h}_\alpha$ or $\mathbf{h}_\beta$, respectively. Without loss of generality, we assume that $x^k \in \bar{X}_{L,\mathcal{H},\prec}(\mathbf{h}_\alpha)$ and $\mu_k > 0$. This implies that $\mathcal{H}_{\mathrm{Syz}}$ contains the syzygy $\mathbf{S}_{\alpha;k}$. As shown above, a unique index $\gamma \neq \beta$ exists such that $\mathbf{S}_{\alpha;k} = \mathbf{S}_{\alpha\gamma}$.

Let $\mathbf{S}_{\alpha\gamma} = x^k \mathbf{e}_\alpha - x^\rho \mathbf{e}_\gamma + \mathbf{R}_{\alpha\gamma}$. By construction, $x^\rho \mathbf{t}_\gamma = x^k \mathbf{t}_\alpha$ divides $x^\mu \mathbf{t}_\alpha = \mathbf{t}_{\alpha\beta}$. Thus $\mathbf{t}_\gamma \mid \mathbf{t}_{\alpha\beta}$ and by Lemma B.3.4 the set $\mathcal{H}_{\text{Schreyer}} \setminus \{\mathbf{S}_{\alpha\beta}\}$ still generates $\text{Syz}(\mathcal{H})$. If we try to iterate this argument, we encounter the following problem. In order to be able to eliminate $\mathbf{S}_{\alpha\beta}$ we need both $\mathbf{S}_{\alpha\gamma}$ and $\mathbf{S}_{\beta\gamma}$ in the remaining set. For $\mathbf{S}_{\alpha\gamma} \in \mathcal{H}_{\text{Syz}}$, this is always guaranteed. But we know nothing about $\mathbf{S}_{\beta\gamma}$ and, if it is not an element of $\mathcal{H}_{\text{Syz}}$, it could have been removed in an earlier iteration.

We claim that with respect to the term order $\prec_{\mathcal{H}}$ the term $\text{lt}_{\prec_{\mathcal{H}}} \mathbf{S}_{\alpha\beta}$ is greater than both $\text{lt}_{\prec_{\mathcal{H}}} \mathbf{S}_{\alpha\gamma}$ and $\text{lt}_{\prec_{\mathcal{H}}} \mathbf{S}_{\beta\gamma}$. Without loss of generality, we may assume for simplicity that $\alpha < \beta < \gamma$, as the syzygies $\mathbf{S}_{\alpha\beta}$ and $\mathbf{S}_{\beta\alpha}$ differ only by a sign. Thus $\text{lt}_{\prec_{\mathcal{H}}} \mathbf{S}_{\alpha\beta} = \frac{\mathbf{t}_{\alpha\beta}}{\mathbf{t}_\alpha} \mathbf{e}_\alpha$ and similarly for $\mathbf{S}_{\alpha\gamma}$ and $\mathbf{S}_{\beta\gamma}$. Furthermore, $\mathbf{t}_\gamma \mid \mathbf{t}_{\alpha\beta}$ trivially implies $\mathbf{t}_{\alpha\gamma} \mid \mathbf{t}_{\alpha\beta}$ and hence $\mathbf{t}_{\alpha\gamma} \prec \mathbf{t}_{\alpha\beta}$ for any term order $\prec$. Obviously, the same holds for $\mathbf{t}_{\beta\gamma}$. Now a straightforward application of the definition of the term order $\prec_{\mathcal{H}}$ proves our claim.

Thus if we always remove a syzygy $\mathbf{S}_{\alpha\beta} \in \mathcal{H}_{\text{Schreyer}} \setminus \mathcal{H}_{\text{Syz}}$ whose leading term is maximal with respect to the term order $\prec_{\mathcal{H}}$, it can never happen that the syzygy $\mathbf{S}_{\beta\gamma}$ required for the application of Lemma B.3.4 has already been eliminated earlier and $\mathcal{H}_{\text{Syz}}$ is indeed a generating set of $\text{Syz}(\mathcal{H})$. □

This result is weaker than Schreyer's Theorem, as it only yields a generating set of the syzygy module and not a Gröbner basis for it. However, if $\mathcal{H}$ is not any involutive basis but a Pommaret basis, then $\mathcal{H}_{\text{Syz}}$ is again a Pommaret basis for $\text{Syz}(\mathcal{H})$ with respect to $\prec_{\mathcal{H}}$.

Theorem 2.6.3. *Let $\mathcal{H}$ be a Pommaret basis of $\mathcal{M}$. If we order its elements according to their class, i. e. if we assume that $\operatorname{cls} \mathbf{h}_\alpha < \operatorname{cls} \mathbf{h}_\beta$ implies $\alpha < \beta$, then $\mathcal{H}_{\text{Syz}}$ is a Pommaret basis of $\text{Syz}(\mathcal{H})$ with respect to the term order $\prec_{\mathcal{H}}$.*

Proof. Our ordering of $\mathcal{H}$ implies that $\text{lt}_{\prec_{\mathcal{H}}} \mathbf{S}_{\alpha;k} = x^k \mathbf{e}_\alpha$. Indeed, as mentioned above $\text{lt}_\prec(P_\beta^{(\alpha;k)} \mathbf{h}_\beta) \preceq \text{lt}_\prec(x^k \mathbf{h}_\alpha)$ for all β and only one index $\bar\beta$ exists with $\text{lt}_\prec(P_{\bar\beta}^{(\alpha;k)} \mathbf{h}_{\bar\beta}) = \text{lt}_\prec(x^k \mathbf{h}_\alpha)$. But as $P_{\bar\beta}^{(\alpha;k)} \in \Bbbk[\mathcal{X}_{P,\prec}(\mathbf{h}_{\bar\beta})]$ this equality implies that $\operatorname{cls} \mathbf{h}_{\bar\beta} > \operatorname{cls} \mathbf{h}_\alpha$. Hence $\bar\beta > \alpha$ and according to the definition of the term order $\prec_{\mathcal{H}}$ we find $\text{lt}_{\prec_{\mathcal{H}}} \mathbf{S}_{\alpha;k} = x^k \mathbf{e}_\alpha$.

In order to show that $\mathcal{H}_{\text{Syz}}$ is a Pommaret basis, we must express every non-multiplicative product $x^\ell \mathbf{S}_{\alpha;k}$ with $\ell > k$ as a linear combination of multiplicative products. As x^ℓ is obviously non-multiplicative for $\mathbf{h}_\alpha$, the set $\mathcal{H}_{\text{Syz}}$ also contains the syzygy $\mathbf{S}_{\alpha;\ell}$ and x^k is multiplicative for it. Let us study the difference

$$x^\ell \mathbf{S}_{\alpha;k} - x^k \mathbf{S}_{\alpha;\ell} = \sum_{\beta=1}^{s} \left(x^\ell P_\beta^{(\alpha;k)} - x^k P_\beta^{(\alpha;\ell)} \right) \mathbf{e}_\beta \,. \tag{2.33}$$

If all coefficients on the right hand side are multiplicative, we know by Lemma 2.6.1 that it must actually be zero and we are done.

While the polynomials $P_\beta^{(\alpha;k)}$ and $P_\beta^{(\alpha;\ell)}$ depend only on multiplicative variables, it is possible that either x^ℓ or x^k is non-multiplicative for one or more $\mathbf{h}_\beta$ where $P_\beta^{(\alpha;k)}$ or $P_\beta^{(\alpha;\ell)}$, respectively, is not zero. If this is the case, we choose the pair (β, k) or (β, ℓ) for that $\mathrm{lt}_{\prec_\mathcal{H}}(x^\ell P_\beta^{(\alpha;k)}\mathbf{e}_\beta)$ (or $\mathrm{lt}_{\prec_\mathcal{H}}(x^k P_\beta^{(\alpha;\ell)}\mathbf{e}_\beta)$, respectively) is maximal with respect to $\prec_\mathcal{H}$. Without loss of generality we assume that this pair is of the form (β, k); the following argument remains true for pairs of the form (β, ℓ).

As x^k is assumed to be non-multiplicative for $\mathbf{h}_\beta$, the set $\mathcal{H}_{\mathrm{Syz}}$ contains the syzygy $\mathbf{S}_{\beta;k}$ and we can eliminate $\mathrm{lt}_{\prec_\mathcal{H}}(x^k P_\beta^{(\alpha;\ell)}\mathbf{e}_\beta)$ by subtracting $P_\beta^{(\alpha;\ell)}\mathbf{S}_{\beta;k}$. Note that the polynomial $P_\beta^{(\alpha;\ell)}$ depends only on variables x^i with $i \leq \operatorname{cls}\mathbf{h}_\beta$. By assumption $\operatorname{cls}\mathbf{h}_\beta < k$; hence this is a multiplicative product. All terms introduced by this operation are less than $\mathrm{lt}_{\prec_\mathcal{H}}(x^k P_\beta^{(\alpha;\ell)}\mathbf{e}_\beta)$. It follows that continuing this way we can eliminate all terms with non-multiplicative coefficients in a finite number of steps. As soon as this has happened, it follows from Lemma 2.6.1 that the remainder is zero and we have expressed $x^\ell\mathbf{S}_{\alpha;k}$ as a linear combination of multiplicative products. □

Note that this proof makes extensive use of the definition of the Pommaret division and cannot easily be extended to other divisions. In particular, it is decisive that we can trivially determine first the leading terms and then the multiplicative variables of each syzygy $\mathbf{S}_{\alpha;k}$. As a simple corollary we obtain the following stronger version of Hilbert's Syzygy Theorem B.3.6.

Corollary 2.6.4. *Let $\mathcal{H}$ be a Pommaret basis of the polynomial module $\mathcal{M}$. Denote by $\beta_0^{(k)}$ the number of generators $\mathbf{h} \in \mathcal{H}$ such that $\operatorname{cls}\operatorname{le}_\prec\mathbf{h} = k$ and set $d = \min\{k \mid \beta_0^{(k)} > 0\}$. Then $\mathcal{M}$ possesses a free resolution*

$$0 \longrightarrow \mathcal{P}^{t_{n-d}} \longrightarrow \cdots \longrightarrow \mathcal{P}^{t_1} \longrightarrow \mathcal{M} \longrightarrow 0 \tag{2.34}$$

where the ranks of the free modules are given by

$$t_i = \sum_{k=1}^{n-i} \binom{n-k}{i} \beta_0^{(k)} . \tag{2.35}$$

Proof. By Theorem 2.6.3, $\mathcal{H}_{\mathrm{Syz}}$ is a Pommaret basis of $\mathrm{Syz}(\mathcal{H})$. Applying Theorem 2.6.3 again, we can construct a Pommaret basis of the second syzygy module $\mathrm{Syz}^2(\mathcal{H})$ and so on. It follows from the definition (2.31) of the generators $\mathbf{S}_{\alpha;k}$ that $\operatorname{cls}\mathbf{S}_{\alpha;k} = k > \operatorname{cls}\mathbf{h}_\alpha$. If d is the minimal class appearing in $\mathcal{H}$, the minimal class in $\mathcal{H}_{\mathrm{Syz}}$ is $d+1$. This yields the length of the resolution (2.34), as a Pommaret basis with $d = n$ generates a free module.

The ranks of the modules follow from a rather straightforward combinatorial calculation. Let $\beta_i^{(k)}$ denote the number of generators of class k of the ith syzygy module $\mathrm{Syz}^i(\mathcal{H})$. By definition of the generators $\mathbf{S}_{\alpha;k}$, we find $\beta_i^{(k)} = \sum_{j=1}^{k-1} \beta_{i-1}^{(j)}$, as each generator of class less than k in the Pommaret basis of $\mathrm{Syz}^{i-1}(\mathcal{H})$ contributes one generator of class k in the basis of $\mathrm{Syz}^i(\mathcal{H})$. A simple induction allows us to express the $\beta_i^{(k)}$ in terms of the $\beta_0^{(k)}$:

$$\beta_i^{(k)} = \sum_{j=1}^{k-i} \binom{k-j-1}{i-1} \beta_0^{(j)} . \tag{2.36}$$

The ranks of the modules in (2.34) are given by $t_i = \sum_{k=1}^{n} \beta_i^{(k)}$; entering (2.36) yields via a classical identity for binomial coefficients (2.35). □

A natural question is when the resolution (2.34) is *minimal* in the sense of Definition B.3.7, as in this case t_i are the Betti numbers of the module $\mathcal{M}$. In general, (2.34) is not minimal, as the generators $\mathbf{S}_{\alpha;k}$ may contain constant coefficients. But (2.34) is at least a resolution of minimal length, i.e. $\operatorname{proj\,dim} \mathcal{M} = n - d$. Indeed, by Proposition 2.4.13, $\operatorname{depth} \mathcal{M} = d$ and the Auslander–Buchsbaum formula (graded form [47, Exercise 19.8]) for a $\mathcal{P}$-module asserts that $\operatorname{proj\,dim} \mathcal{M} + \operatorname{depth} \mathcal{M} = n$.

In the case of monomial modules, Eliahou and Kervaire [49] identified a class of modules for which (2.34) is a minimal resolution: stable modules. Their result is readily derived using Pommaret bases; in fact, implicitly they used such bases.

Definition 2.6.5. *A (possibly infinite) set $\mathcal{N} \subseteq \mathbb{N}_0^n$ is* stable, *if for each multi index $\nu \in \mathcal{N}$ all multi indices $\nu - 1_k + 1_j$ with $k = \operatorname{cls} \nu < j \leq n$ are also elements of $\mathcal{N}$. A monomial submodule $\mathcal{M} \subseteq \mathcal{P}^m$ is called* stable, *if each of the sets $\mathcal{N}_\alpha = \{\mu \mid x^\mu \mathbf{e}_\alpha \in \mathcal{M}\} \subset \mathbb{N}_0^n$ with $1 \leq \alpha \leq m$ is stable.*

Any monomial submodule has a unique minimal basis. For stable submodules it must coincide with its Pommaret basis. This represents a very simple and effective characterisation of stable submodules.

Proposition 2.6.6. *Let $\mathcal{M} \subseteq \mathcal{P}^m$ be a monomial submodule. $\mathcal{M}$ is stable, if and only if its minimal basis $\mathcal{H}$ is also a Pommaret basis.*

Proof. Let us assume first that $\mathcal{M}$ is stable; we have to show that $\langle \mathcal{H} \rangle_P = \mathcal{M}$. For every term $\mathbf{s} \in \mathcal{M}$ a unique term $\mathbf{t}_1 \in \mathcal{H}$ exists such that $\mathbf{s} = s_1 \mathbf{t}_1$ for some term $s_1 \in \mathbb{T}$. If $s_1 \in \Bbbk[X_P(\mathbf{t}_1)]$, we are done. Otherwise there exists an index $j > k = \operatorname{cls} \mathbf{t}_1$ such that $x^j \mid s_1$ and we rewrite

$$\mathbf{s} = \left(\frac{x^k}{x^j} s_1\right) \left(\frac{x^j}{x^k} \mathbf{t}_1\right) . \tag{2.37}$$

Since, by assumption, $\mathcal{M}$ is stable, $(x^j/x^k)\mathbf{t}_1 \in \mathcal{M}$. Thus a term $\mathbf{t}_2 \in \mathcal{H}$ exists such that $(x^j/x^k)\mathbf{t}_1 = s_2\mathbf{t}_2$ for some term $s_2 \in \mathbb{T}$. We are done, if $\mathbf{t}_2 \mid_P \mathbf{s}$. Otherwise we iterate this construction. By the continuity of the Pommaret division, this cannot go on infinitely, i.e. after a finite number of steps we must reach a term $\mathbf{t}_N \in \mathcal{H}$ such that $\mathbf{t}_N \mid_P \mathbf{s}$ and thus $\mathbf{s} \in \langle\mathcal{H}\rangle_P$.

For the converse, assume that the minimal basis $\mathcal{H}$ is a Pommaret basis. Then a unique $\mathbf{t} \in \mathcal{H}$ exists for each $\mathbf{s} \in \mathcal{M}$ such that $\mathbf{t} \mid_P \mathbf{s}$. We must show that with $k = \operatorname{cls}\mathbf{s} \leq \operatorname{cls}\mathbf{t}$ for all $i > k$ the terms $(x^i/x^k)\mathbf{s}$ are also elements of $\mathcal{M}$. We distinguish two cases. If $\mathbf{s} = \mathbf{t}$, a $\bar{\mathbf{t}} \in \mathcal{H}$ exists with $\bar{\mathbf{t}} \mid_P (x^i\mathbf{t})$. As $\mathcal{H}$ is a minimal basis, it cannot be that $\bar{\mathbf{t}} = x^i\mathbf{t}$. Instead we must have that $\bar{\mathbf{t}} \mid (x^i/x^k)\mathbf{t} = (x^i/x^k)\mathbf{s}$ and we are done. If $\mathbf{s} \neq \mathbf{t}$, we write $\mathbf{s} = s\mathbf{t}$ with $s \in \mathbb{T}$. If $k < \operatorname{cls}\mathbf{t}$, then $x^k \mid s$ which implies that we can divide by x^k and thus $(x^i/x^k)\mathbf{s} \in \mathcal{M}$. Otherwise, $\operatorname{cls}\mathbf{s} = \operatorname{cls}\mathbf{t}$ and we know from the first case that $(x^i/x^k)\mathbf{t} \in \mathcal{M}$. But $(x^i/x^k)\mathbf{s} = (x^i/x^k)s\mathbf{t}$. $\square$

Example 2.6.7. It follows from Lemma 2.4.2, that if $\mathcal{H}$ is a Pommaret basis of the submodule $\mathcal{M} \subseteq \mathcal{P}^m$ of degree q, then $(\operatorname{lt}_{\prec}\mathcal{M})_{\geq q}$ is a stable monomial submodule, as $\operatorname{lt}_{\prec}\mathcal{H}_q$ is obviously its minimal basis and simultaneously a Pommaret basis. $\triangleleft$

Theorem 2.6.8. *Let $\mathcal{M} \subseteq \mathcal{P}^m$ be a monomial module. Then the syzygy resolution* (2.34) *is minimal, if and only if $\mathcal{M}$ is stable.*

Proof. The proof is rather simple but notationally messy. We denote the elements of the Pommaret basis of $\operatorname{Syz}^i(\mathcal{H})$ by $\mathbf{S}_{\alpha;k_1,\dots,k_i}$ where the inequalities $k_i > \cdots > k_1 > \operatorname{cls}\mathbf{h}_\alpha$ hold. Similarly, we denote the standard basis of the free module $\mathcal{P}^{t_i}$ by $\mathbf{e}_{\alpha;k_1,\dots,k_{i-1}}$ with the same inequalities for the indices. The ith syzygies have the form

$$\mathbf{S}_{\alpha;k_1,\dots,k_i} = x^{k_i}\mathbf{e}_{\alpha;k_1,\dots,k_{i-1}} - \sum P^{(\alpha;k_1,\dots,k_i)}_{\beta;\ell_1,\dots,\ell_{i-1}}\mathbf{e}_{\beta;\ell_1,\dots,\ell_{i-1}} \tag{2.38}$$

with polynomials $P^{(\alpha;k_1,\dots,k_i)}_{\beta;\ell_1,\dots,\ell_{i-1}} \in \mathbb{k}[x^1,\dots,x^{\ell_{i-1}}]$ and where the summation is over all $1 \leq \beta \leq p$ and $\ell_{i-1} > \cdots > \ell_1 > \operatorname{cls}\mathbf{h}_\beta$. By definition, the resolution is minimal, if and only if all polynomials $P^{(\alpha;k_1,\dots,k_i)}_{\beta;\ell_1,\dots,\ell_{i-1}}$ are elements of the ideal $\mathcal{P}_+ = \langle x^1,\dots,x^n\rangle$.

We proceed by an induction over the resolution. If $\mathcal{M}$ is stable, by Proposition 2.6.6 the Pommaret basis $\mathcal{H}$ is the minimal basis of $\mathcal{M}$. To each element $\mathbf{h}_\alpha \in \mathcal{H}$ and each non-multiplicative index $k > \operatorname{cls}\mathbf{h}_\alpha$ there exists precisely one $\mathbf{h}_\beta \in \mathcal{H}$ such that $x^k\mathbf{h}_\alpha = m\mathbf{h}_\beta$ where the term $m \in \mathbb{T}$ contains only multiplicative variables of $\mathbf{h}_\beta$. As $\mathcal{H}$ is the minimal basis of $\mathcal{M}$, it is not possible that $m = 1$, as otherwise $\mathbf{h}_\alpha \mid \mathbf{h}_\beta$. This proves our claim for the case $i = 1$, as $P^{(\alpha;k)}_\beta \in \mathcal{P}_+$ for all possible values of α, β and k.

Assume now that $P^{(\alpha;k_1,\dots,k_i)}_{\beta;\ell_1,\dots,\ell_{i-1}} \in \mathcal{P}_+$ for all possible values of the indices. For the induction we must show that all coefficients $P^{(\alpha;k_1,\dots,k_{i+1})}_{\beta;\ell_1,\dots,\ell_i} \in \mathcal{P}_+$. For this purpose, we rewrite

$$x^{k_{i+1}}\mathbf{S}_{\alpha;k_1,\dots,k_i} = x^{k_i}(x^{k_{i+1}}\mathbf{e}_{\alpha;k_1,\dots,k_{i-1}}) - \sum P^{(\alpha;k_1,\dots,k_i)}_{\beta;\ell_1,\dots,\ell_{i-1}}(x^{k_{i+1}}\mathbf{e}_{\beta;\ell_1,\dots,\ell_{i-1}}) \tag{2.39}$$

and proceed as in the proof of Theorem 2.6.3: we compute a normal form with respect to the Pommaret basis of $\mathrm{Syz}^{i+1}(\mathcal{H})$. We solve each syzygy (2.38) for its initial term, the first term on the right hand side, and substitute the result in (2.39) whenever there is a non-multiplicative product of a basis vector $\mathbf{e}_{\beta;\ell_1,\dots,\ell_{i-1}}$. As all $P^{(\alpha;k_1,\dots,k_i)}_{\beta;\ell_1,\dots,\ell_{i-1}}$ lie in $\mathcal{P}_+$, such a substitution can never decrease the degree of a coefficient. Thus $P^{(\alpha;k_1,\dots,k_{i+1})}_{\beta;\ell_1,\dots,\ell_i} \in \mathcal{P}_+$ for all possible values of the indices.

This proves that for a stable module the resolution (2.34) is minimal. Assume that $\mathcal{M}$ is not stable. By Proposition 2.6.6 the Pommaret basis $\mathcal{H}$ cannot be the minimal basis. In this case, $\mathcal{H}$ contains elements $\mathbf{h}_\alpha$, $\mathbf{h}_\beta$ with $\mathbf{h}_\beta = x^k\mathbf{h}_\alpha$ for some $x^k \in \bar{X}_P(\mathbf{h}_\alpha)$. So the syzygy $\mathbf{S}_{\alpha;k} = x^k\mathbf{e}_\alpha - \mathbf{e}_\beta$ contains a constant coefficient and the resolution is not minimal. □

We return to the general, polynomial, case but restrict for notational simplicity again to ideals instead of submodules. In many situations it is of interest to obtain an estimate on the degree of an ideal basis, in particular for the complexity analysis of completion algorithms. Up to date, no simple answer is known to this question. Somewhat surprisingly, the stronger problem of bounding not only the degree of a basis of $\mathcal{I}$ but also of its syzygies can be treated effectively.

Definition 2.6.9. *Let $\mathcal{I} \subseteq \mathcal{P}$ be a homogeneous ideal. $\mathcal{I}$ is* q-regular, *if the jth syzygy module is generated by elements of degree less than or equal to $q + j$. The* Castelnuovo–Mumford regularity $\operatorname{reg}\mathcal{I}$ *of the ideal $\mathcal{I}$ is the least q for which $\mathcal{I}$ is q-regular.*

We show in the remainder of this section that $\operatorname{reg}\mathcal{I}$ can be trivially read off a Pommaret basis of $\mathcal{I}$. Our basic tool is the following theorem combining results from Bayer and Stillman [14] and Eisenbud and Goto [48], respectively. These works also contain proofs of it.

Theorem 2.6.10. *The following statements are equivalent.*

(i) *The homogeneous module $\mathcal{I} \subseteq \mathcal{P}$ is q-regular.*
(ii) *The truncation $\mathcal{I}_{\geq q}$ admits a linear free resolution, i.e. $\mathcal{I}_{\geq q}$ is generated by elements of degree q and all maps in the resolution are linear.*

(iii) *For some* $0 \leq d \leq n$ *elements* $y^1, \dots, y^d \in \mathcal{P}_1$ *exist such that*

$$\left(\langle \mathcal{I}, y^1, \dots, y^{j-1} \rangle : y^j\right)_q = \left(\langle \mathcal{I}, y^1, \dots, y^{j-1} \rangle\right)_q\,, \quad 1 \leq j \leq d\,, \tag{2.40a}$$

$$\left(\langle \mathcal{I}, y^1, \dots, y^d \rangle\right)_q = \mathcal{P}_q\,. \tag{2.40b}$$

Recall that the truncation is defined by $\mathcal{I}_{\geq q} = \bigoplus_{p \geq q} \mathcal{I}_p$. The notation $\mathcal{I} : g$ refers to the quotient of the ideal $\mathcal{I}$ by the principal ideal generated by $g \in \mathcal{P}$, i.e. $\mathcal{I} : g = \{f \in \mathcal{P} \mid gf \in \mathcal{I}\}$.

Theorem 2.6.11. *Let $\mathcal{I} \subseteq \mathcal{P}$ be a homogeneous ideal. The Castelnuovo–Mumford regularity of $\mathcal{I}$ is q, if and only if $\mathcal{I}$ has (in suitable coordinates) a Pommaret basis of degree q with respect to the degree reverse lexicographic order.*

Proof. One direction is trivial. If $\mathcal{I}$ has a Pommaret basis $\mathcal{H}$ of degree q, then the set $\mathcal{H}_q$ defined by (2.18) is a Pommaret basis of the truncation $\mathcal{I}_{\geq q}$ according to Lemma 2.4.1. Now it follows easily from Theorem 2.6.3 that $\mathcal{I}_{\geq q}$ possesses a linear free resolution, as all syzygies of $\mathcal{H}_q$ are homogeneous of degree 1. Thus the ideal $\mathcal{I}$ is q-regular by Part (ii) of Theorem 2.6.10.

For the converse we apply Part (iii) of Theorem 2.6.10. We use coordinates $\mathbf{x}$ such that $x^i = y^i$ for $1 \leq i \leq d$. By Part (ii) of Theorem 2.6.10, $\mathcal{I}_{\geq q}$ can be generated by elements of degree q. Let $\mathcal{H}_q$ be a basis of $\mathcal{I}_q$ as a vector space in triangular form, i.e. $\mathrm{lt}_\prec h_1 \neq \mathrm{lt}_\prec h_2$ for all $h_1, h_2 \in \mathcal{H}_q$. We claim that $\mathcal{H}_q$ is a Pommaret basis of $\mathcal{I}_{\geq q}$.

Let us write $\mathcal{H}_q = \{h_{k,\ell} \mid 1 \leq k \leq n,\ 1 \leq \ell \leq \ell_k\}$ where $\operatorname{cls} h_{k,\ell} = k$. This implies that a minimal basis of the ideal $\langle \mathcal{I}_{\geq q}, x^1, \dots, x^j \rangle$ is given by $\{h \in \mathcal{H}_q \mid \operatorname{cls} h > j\} \cup \{x^1, \dots, x^j\}$. We show now that

$$\mathcal{H}_{q+1} = \{x^j h_{k,\ell} \mid 1 \leq j \leq k,\ 1 \leq k \leq n,\ 1 \leq \ell \leq \ell_k\} \tag{2.41}$$

is a basis of $\mathcal{I}_{q+1}$ as a vector space. We will prove in Chapter 6 that this implies that $\mathcal{H}_q$ is a Pommaret basis of $\mathcal{I}_{\geq q}$ ($\mathcal{H}_q$ is then locally involutive, cf. Proposition 6.4.5 and Lemma 6.3.5).

Let $f \in \mathcal{I}_{q+1}$ and $\operatorname{cls} f = j$. By the properties of the degree reverse lexicographic order this implies that $f = x^j \hat{f} + g$ with $\hat{f} \in \left(\mathbb{k}[x^j, \dots, x^n] \setminus \{0\}\right)_q$ and $g \in \left(\langle x^1, \dots, x^{j-1} \rangle\right)_{q+1}$ (cf. Lemma B.2.2). We distinguish two cases. By (2.40b) we have $\left(\langle \mathcal{I}, x^1, \dots, x^d \rangle\right)_q = \mathcal{P}_q$. Thus if $j > d$, we may write $\hat{f} = \sum_{k=d+1}^{n} \sum_{\ell=1}^{\ell_k} c_{k,\ell} h_{k,\ell} + \hat{g}$ with $c_{k,\ell} \in \mathbb{k}$ and $\hat{g} \in \left(\langle x^1, \dots, x^d \rangle\right)_q$. We set $\hat{f}_0 = \sum_{k=j}^{n} \sum_{\ell=1}^{\ell_k} c_{k,\ell} h_{k,\ell}$ and $\hat{f}_1 = \sum_{k=d+1}^{j-1} \sum_{\ell=1}^{\ell_k} c_{k,\ell} h_{k,\ell} + \hat{g}$. Obviously, $\hat{f} \in \left(\langle \mathcal{I}, x^1, \dots, x^{j-1} \rangle : x^j\right)_q$. Thus if we have $j \leq d$, then (2.40a) implies that $\hat{f} \in \left(\langle \mathcal{I}, x^1, \dots, x^{j-1} \rangle\right)_q$. Hence in this case we may write $\hat{f} = \hat{f}_0 + \hat{f}_1$ with $\hat{f}_0 = \sum_{k=j}^{n} \sum_{\ell=1}^{\ell_k} c_{k,\ell} h_{k,\ell}$ and $\hat{f}_1 \in \left(\langle x^1, \dots, x^{j-1} \rangle\right)_q$.

It is trivial that $\langle\mathcal{H}_{q+1}\rangle \subseteq \mathcal{I}_{q+1}$ (here we mean the span over $\Bbbk$ and not over $\mathcal{P}$). We show by an induction over j that $\mathcal{I}_{q+1} \subseteq \langle\mathcal{H}_{q+1}\rangle$. If $j = 1$, then $f = x^1\hat{f}$ with $\hat{f} \in \mathcal{I}_q$. Thus $f \in \langle\mathcal{H}_{q+1}\rangle$. If $j > 1$, we write $f = f_0 + f_1$ with $f_0 = x^j\hat{f}_0$ and $f_1 = x^j\hat{f}_1 + g$ where $\hat{f}_0$ and $\hat{f}_1$ have been defined above. By construction, $f_0 \in \langle\mathcal{H}_{q+1}\rangle$, as x^j is multiplicative for all generators contained in $\hat{f}_0$, and $f_1 \in \mathcal{I}_{q+1}$ with $\operatorname{cls} f_1 < j$. According to our inductive hypothesis this implies that $f_1 \in \langle\mathcal{H}_{q+1}\rangle$, too. Hence $\langle\mathcal{H}_{q+1}\rangle = \mathcal{I}_{q+1}$. □

Remark 2.6.12. Again it is a natural question whether this result holds only for Pommaret bases. And again we may simply take the example already used in Remark 2.4.14. In δ-regular coordinates, the ideal $\mathcal{I}$ considered there possesses a Pommaret basis of degree 2. Hence the Castelnuovo–Mumford regularity $\operatorname{reg}\mathcal{I}$ is 2. But we saw that a simple permutation of the variables yields a Janet basis of degree 4 for the same term order. Thus we may conclude that arbitrary involutive bases contain no information on the Castelnuovo–Mumford regularity. ◁

2.7 Notes

Combinatorial decompositions of polynomial modules and involutive bases are implicitly contained in the *Janet–Riquier Theory* of linear systems of differential equations developed in the late 19th, early 20th century by a number of authors (see e.g. [83, 84, 150, 177, 180]). In this theory, the term *passive* introduced by Méray and Riquier [122] is more common than *involutive*; the latter one seems to go back to Lie. The idea of assigning multiplicative variables is apparently due to Janet. In this context, one should mention that the distinction into parametric and principal derivatives used in Chapter 1 goes back to Riquier.

The results of this theory remained more or less unknown in commutative algebra until fairly recently. Rees [145] introduced much later the special decomposition now carrying his name. The concept was generalised by Stanley [169, 170] who was interested in a simple way to determine Hilbert functions; Proposition 2.5.2 is due to him. In recent years, a number of authors have picked up these ideas and applied them in various fields; see e.g. [12, 18, 54, 172, 173] and references therein.

Concrete algorithms for their construction have been proposed by Sturmfels and White [173]. They presented a simple recursive algorithm to determine a Stanley decomposition for arbitrary monomial modules which is essentially equivalent to our proof of Proposition 2.5.3 (a related algorithm was designed by Reid [147] in the context of initial value problems for overdetermined systems of differential equations). Thus its extension to polynomial modules $\mathcal{P}^m/\mathcal{N}$ requires that one must first compute a Gröbner basis of $\mathcal{N}$ for some term order $\prec$ in order to obtain a basis of $\operatorname{lt}_\prec\mathcal{N}$. They further developed a special method for determining Rees decompositions. However, it is rather expensive, as it needs repeated computations of Gröbner bases. Our approach via Pommaret bases is more direct and efficient.

Proposition 2.5.3 on the decomposition of the complementary set is a classical result in commutative algebra [40, pp. 417–418] and is sometimes used to prove the existence of the Hilbert polynomial. We slightly modified the proof in order to obtain a disjoint decomposition. Gerdt [55, Lemma 24] proposed a version of this result that relates the form of the cones to an involutive division. However, his proof is not correct, as the there constructed set $\mathcal{U}$ (in the notation of our proof of Proposition 2.5.3) may also contain elements of $\mathcal{I}$. Indeed, it is trivial to provide examples for this effect using the Janet division. Proposition 2.5.4 was already proven by Janet [84, Section 15].

Involutive divisions were introduced by Gerdt and Blinkov [56]. They also defined all examples of divisions appearing in this chapter. One should note a slight difference in conventions. We always analyse multi indices from the right, i. e. we start with the last entry μ_n and not with the first entry μ_1. Thus for the Janet division we first check whether n is multiplicative and then continue with the lower values;[4] for the Pommaret division multi indices of the form $[0, \dots, 0, \mu_n]$ possess the most multiplicative indices. In [56], it is exactly the other way round: they start the analysis of a multi index with μ_1 and not μ_n; for the Janet division they first check whether 1 is multiplicative and for the Pommaret division multi indices $[\mu_1, 0, \dots, 0]$ possess the most multiplicative indices. Obviously, the two conventions become identical, if one simply inverts the order of the variables.

A weaker notion of an involutive division was introduced by Apel [8]. One could say that he uses a "local" approach whereas ours is "global". He defines an involutive division only with respect to a *single* set $\mathcal{N} \subseteq \mathbb{N}_0^n$ and hence does not need the second condition in Definition 2.1.1 relating the set of multiplicative indices for subsets. We will discuss the effect of this in more detail at the end of Chapter 6 in the context of completion algorithms.

The term "algebra of solvable type" was coined by Kandry-Rody and Weispfenning [88], when they tried to introduce Gröbner bases for non-commutative rings. Their definition is more restrictive than ours, as it does not allow that the terms operate on the coefficients and requires a stronger form of compatibility between the multiplication $\star$ and the term order $\prec$. It automatically implies that $\star$ respects $\prec$. For our purposes this is the decisive property and thus we have used it instead of the more technical axioms in [88] for Definition 2.2.1.

In the application sections we restricted to the ordinary polynomial ring $\mathcal{P}$. In fact, many of the results presented there do not hold in arbitrary polynomial algebras of solvable type. For example, in all our proofs concerning the depth of a module we implicitly assumed that if $f \in \langle x^i \rangle$, then each term in $\operatorname{supp} f$ is divisible by x^i. In case of the universal enveloping algebra of $\mathfrak{so}(3)$ this is obviously not the case according to (2.13). In fact, we find even that $\langle x^1, x^2 \rangle = \langle x^1, x^2, x^3 \rangle$.

The definition of a Stanley decomposition for a module over a polynomial algebra of solvable type is identical to the definition in the case of the ordinary polynomial ring, as it is based on an isomorphism as vector spaces. Thus the multiplication does not matter and it is important to note that on the right hand side of (2.23) the $\cdot$ is *not* replaced by a $\star$. If the multiplication $\star$ preserves the natural grading of the polynomial ring, Proposition 2.5.2 remains valid.

[4] In fact, this is the way the division was introduced by Janet [84, pp. 16–17].

Our syzygy theory can be generalised under certain conditions. In particular, the proof of Theorem 2.6.3 depends mainly on normal form arguments that work in any polynomial algebra of solvable type. We must require that $\star$ respects term orders of the form $\prec_{\mathcal{H}}$ in order for the theorem to make sense. In addition, we must take care of the arguments on multiplicative variables. In arbitrary polynomial algebras of solvable type $(ax^i) \star (bx^j) = cx^ix^j + h$ where in general the polynomial h may depend on any variables x^k. We need that if x^i and x^j are both multiplicative for some other polynomial, then the product $(ax^i) \star (bx^j)$ should also be multiplicative. This implies that h may depend only on variables with $k \leq \max\{i, j\}$. All these conditions are for example satisfied by rings of differential operators.

The notion of an involutive basis is also due to Gerdt and Blinkov [56]; the special case of a Pommaret basis was already earlier introduced by Zharkov and Blinkov [195]. But as already mentioned, such bases are implicitly contained in many articles, even in rather old ones. The main credit of the three mentioned authors is that, inspired by the work of Janet, they were the first to give a rigorous algebraic formulation of the theory and to relate it to the modern theory of Gröbner bases. In particular, they introduced involutive reductions and normal forms, two concepts missing in the classical Janet–Riquier Theory.

One may wonder why we have introduced weak involutive bases. This concept is indeed very weak: a weak involutive basis does not define a Stanley decomposition, it is not necessarily a Gröbner basis and so on. All interesting properties hold only for strong bases; indeed, we will exclusively work with strong bases. One simple reason for the introduction of weak bases is that in our approach to involutive bases their definition appears very naturally. In some proofs it is convenient to show first that one has a weak basis and later that it can be reduced to a strong one.

In a recent preprint [73] on involutive bases in the Weyl algebra, weak bases played a more prominent role. There bases with respect to a larger class of orders, so-called *multiplicative monomial orders*, are studied. In general, these orders are not well-orders and normal form computations do not always terminate. It was not possible to show the existence of strong involutive bases for such orders but a constructive proof of the existence of weak involutive bases which are simultaneously Gröbner bases could be given. This work showed that normal form arguments are even more important for involutive bases than for Gröbner bases.

δ-regularity is often considered as a purely technical nuisance. It is of course unpleasant that for some ideals no finite Pommaret basis exists. In particular, the completion algorithms discussed in Chapter 6 do not terminate in this case. However, δ-singularity is often related to important properties of the ideal considered. For example, in applications to differential equations it is related to characteristics; if the hyperplane $x^n = 0$ is a characteristic surface, then the coordinates are δ-singular (see Section 3.3).

Several solutions to the problem of δ-regularity have been proposed in the literature; mainly in the context of differential equations. Most of them are based on the proof of Proposition 2.4.6. Pommaret [139] advocates the brute force method of computing in a generic coordinate system. As it requires the introduction of n^2 additional parameters which are considered as algebraically independent transcendentals over $\Bbbk$, this approach is only of theoretical value. Even for small examples computations in generic coordinates very quickly blow up. An advantage is that one obtains an explicit characterisation of all δ-singular coordinate systems which is sometimes of interest (see e.g. [173, Proposition 5.3])

[159] contains an improvement on this approach. If k is the lowest class present in the set $\mathcal{F}$, it requires only the introduction of $n(n-k)$ parameters. Furthermore, it is not necessary to explicitly perform the coordinate transformation. As the ranks of several parametric matrices must be determined, this solution is still computationally rather expensive. It also yields all genericity conditions.

Another possibility consists of relying on Proposition 2.4.6 and performing a random coordinate transformation. This approach is, for example, used by Hartley and Tucker [69] in their implementation of the Cartan–Kähler theory (and also by Sturmfels and White [173] in their above mentioned method for the construction of Rees decompositions). As we have seen, it leads with probability 1 to a δ-regular coordinate system. Besides the possibility that we still might get a δ-singular system, the main disadvantage of this approach is that it destroys any sparsity that might be present in the original set $\mathcal{F}$. In particular, in applications to differential equations the elements of $\mathcal{F}$ are very often sparse, i.e. not every derivative p^α_μ up to a given length q of the multi indices μ does actually appear. Of course, this is a great advantage in any concrete computations which one would like to preserve.

The approach presented in Section 2.4 was developed in [71] only very recently. It has two interesting aspects for computations. On one hand it provides us with a very cheap method to detect termination problems in the completion; this will be discussed in more detail in Section 6.4. On the other hand, the proof indicates a concrete coordinate transformation that brings us closer to δ-regularity. It preserves comparatively much of the sparsity of $\mathcal{F}$, as it only affects terms containing x^ℓ.

We have seen that (strong) involutive bases are a special form of Gröbner bases. Very often they are reduced to this one aspect. Much of the (not very extensive) literature on involutive bases treats them solely as an alternative approach to compute Gröbner bases. It is completely overlooked that the combinatorial decompositions defined by involutive bases reveal a lot of structural information not contained in arbitrary Gröbner bases. This holds in particular for Pommaret bases where e.g. Proposition 2.4.13 and Corollary 2.6.4 yield the depth and the projective dimension together with a simple proof of Hilbert's Syzygy Theorem.

The characterisation of stable monomial modules in terms of Pommaret bases is due to Mall [111] (but see also the remark after Lemma 1.2 in [49]). Eliahou and Kervaire [49] provide an explicit representation of their minimal free resolution. It is based on some manipulations of a Koszul complex. Unfortunately, it is not possible to extend their results to the polynomial case. The main reason is very simple. Assume we are given a stable monomial module with basis $\{\mathbf{t}_\alpha\}$. If we multiply the generator $\mathbf{t}_\alpha$ with a term s, there exists a unique generator $\mathbf{t}_\beta$ such that $s\mathbf{t}_\alpha \in \mathcal{C}_P(\mathbf{t}_\beta)$. The estimate $\operatorname{cls} \mathbf{t}_\beta \geq \operatorname{cls} \mathbf{t}_\alpha$ is completely trivial but essential for the construction of the differential of Eliahou and Kervaire. In the polynomial case with a basis $\{\mathbf{h}_\alpha\} \subset \mathcal{P}^m$, the standard representation of a product $s\mathbf{h}_\alpha$ consists of a sum involving in general many generators $\mathbf{h}_\beta$. In particular, it is no longer possible to provide a lower bound for the classes of the $\mathbf{h}_\beta$. This prevents the construction of a complex along the lines of Eliahou and Kervaire.

Of course, based on the Schreyer Theorem, any Gröbner basis may be used to construct a syzygy resolution. And in many textbooks this resolution is used for a proof of Hilbert's Syzygy Theorem. However, the special structures induced by Pommaret basis does not only yield a stronger estimate on the length of the minimal resolution (in fact, we have seen that it always yields a resolution of minimal length) but even a closed formula for the ranks of the modules contained in the resolution. At the moment, it is still an open question to what extent this result may be exploited for obtaining tight estimates on the Betti numbers or even for computing them directly.

A number of different approaches to the Castelnuovo–Mumford regularity exists. A few notes on the history of this concept can be found in [47, pp. 514–515]. Often it is defined using local cohomology. In view of Theorem 2.6.11, this should be closely related to the definition of involution via the Spencer cohomology. To our knowledge, this relation has not yet been worked out. Our approach exploits the more constructive criteria provided by [14, 48].

Among other applications, the Castelnuovo–Mumford regularity $\operatorname{reg}\mathcal{I}$ is a useful measure for the complexity of Gröbner basis computations [13]. Roughly speaking, only for ideals $\mathcal{I}$ with a not too high value of $\operatorname{reg}\mathcal{I}$ one has a good chance to determine a Gröbner basis in a reasonable time. Bayer and Stillman [14] proved that in generic coordinates it is not possible to find a Gröbner basis of degree less than $\operatorname{reg}\mathcal{I}$. This estimate is sharp, as it is realised by bases with respect to the degree reverse lexicographic order. As by Theorem 2.6.11, any Pommaret basis for the degree reverse lexicographic order has this very degree, we obtain as a trivial corollary an estimate how much larger a Pommaret basis may become in comparison to the corresponding reduced Gröbner basis.

3. Involution II: Differential Theory

I consider that I understand an equation when I can predict the properties of its solutions, without actually solving it.

Paul A.M. Dirac

In the previous chapter we introduced involutive bases for polynomial algebras. Now we relate them to the formal theory of differential equations. The key is the (geometric) symbol $\mathcal{M}_q$. In the first section we show that in local coordinates it defines at each point of the differential equation $\mathcal{R}_q$ a homogeneous submodule of a free polynomial module. This permits us to apply our algebraic theory and leads naturally to the notion of an involutive symbol. For easier reference we repeat some results of the previous chapter in the language of differential equations.

In the following section we finally define involutive differential equations and study some of their elementary properties. In particular, we show that, in contrast to formal integrability, involution can be checked by a finite criterion. Furthermore, we introduce the Cartan normal form of an involutive equation, a useful tool in many situations.

As a first application of involution we revisit in the third section the principal symbol and give a rigorous definition of under- and overdetermined equations. It turns out that the classical counting rules are not always valid. Furthermore, we indicate how the principal symbol allows us a partial solution of the problem of δ-regularity.

The last section extends some of the results of Section 2.6 on syzygies of polynomials to linear differential operators. Syzygies correspond to compatibility conditions and the syzygy resolution is dual to the Janet sequence. We study the inverse problem of compatibility and provide an algorithmic solution of it for systems with constant coefficients.

3.1 Involutive Symbols

In Section 1.5 we introduced the (geometric) symbol $\mathcal{M}_q$ of a differential equation $\mathcal{R}_q$. Now we apply the algebraic methods developed in the last chapter to analyse $\mathcal{M}_q$. This analysis is pointwise and takes place in local coordinates. Thus the question arises how intrinsic the results are. In particular, we are using Pommaret bases where we have already seen in Section 2.4 that they only exist in δ-regular coordinates. But there we also showed that generic coordinates are δ-regular and we gave a simple criterion for detecting δ-singularity. Thus we assume throughout this section that we are dealing with δ-regular coordinates.

Let $(\mathbf{x}, \mathbf{p})$ be local coordinates in $J_q\mathcal{E}$ in a neighbourhood of the point $\rho \in \mathcal{R}_q$. The symbol space $(\mathcal{M}_q)_\rho$ consists of all solutions of the real linear system (1.50). If we denote $x = \pi^q(\rho) \in \mathcal{X}$ and $\xi = \pi^q_0(\rho) \in \mathcal{E}$, then $(\mathcal{M}_q)_\rho$ is a subspace of $S_q(T^*_x\mathcal{X}) \otimes V_\xi\mathcal{E}$. Thus the linear equations in (1.50) may be considered as elements of the dual space $S_q(T_x\mathcal{X}) \otimes V^*_\xi\mathcal{E}$. A basis of this space is given by $\{\partial_{x^\mu} \otimes du^\alpha \mid 1 \le \alpha \le m, |\mu| = q\}$ where $\partial_{x^\mu} = (\partial_{x^1})^{\mu_1} \cdots (\partial_{x^n})^{\mu_n}$. Our linear system describes then the kernel of the forms

$$\lambda^\tau = \sum_{\substack{1\le\alpha\le m\\ |\mu|=q}} \frac{\partial\Phi^\tau}{\partial p^\alpha_\mu}(\mathbf{x}, \mathbf{p})\, \partial_{x^\mu} \otimes du^\alpha\,, \quad \tau = 1, \dots, p\,. \tag{3.1}$$

Instead of considering only $S_q(T_x\mathcal{X})$, we take the full symmetric algebra $S(T_x\mathcal{X})$ which is isomorphic to the polynomial ring $\mathcal{P} = \mathbb{k}[x^1, \dots, x^n]$ (see Appendix B.1). So we may identify $S(T_x\mathcal{X}) \otimes V^*_\xi\mathcal{E}$ with a free $\mathcal{P}$-module and the forms λ^τ generate a homogeneous submodule $\mathcal{M}^\mathcal{P}$; comparing with (1.55) we see that the kernel of all the elements of the component $\mathcal{M}^\mathcal{P}_{q+r}$ is the prolonged symbol space $(\mathcal{M}_{q+r})_\rho$.

Definition 3.1.1. *The symbol $\mathcal{M}_q$ of the differential equation $\mathcal{R}_q$ is called* involutive *at the point $\rho \in \mathcal{R}_q$, if local coordinates $\mathbf{x}$ exist on a neighbourhood of $\pi^q(\rho)$ such that an involutive head autoreduction of the set $\Lambda = \{\lambda^\tau \mid 1 \le \tau \le p\}$ yields a Pommaret basis of the symbol module $\mathcal{M}^\mathcal{P}$ for a class respecting term order.*

Recall that essentially only one class respecting term order exists: the degree reverse lexicographic order (see Lemma B.2.2). We require here the use of such an order for a simple reason: the involutive size of the autoreduced set $\Lambda^\triangle$ is maximal for a class respecting term order. Thus, if we do not get a Pommaret basis for such an order, we will not get a Pommaret basis for any other term order.

Obviously, it is well possible for non-linear equations that the symbol is involutive at some points $\rho \in \mathcal{R}_q$ but not at other points. Thus involution is a *local* property. For simplicity, we will assume throughout this book, that the properties of $\mathcal{M}_q$ do not vary over the manifold $\mathcal{R}_q$. In the remainder of this section, we translate this definition into the form usually found in books on the formal theory. Thus to a certain extent we just repeat material of the last chapter in a different language.

In practice, the test for involution of a symbol is performed as follows. We take the matrix M_q and sort its columns according to a class respecting term order. In fact, it suffices, if we take care that a column corresponding to an unknown $\dot{p}^\alpha_\mu$ (representing coordinates in $S_q(T^*_x\mathcal{X}) \otimes V_\xi\mathcal{E}$) is always to the left of a column corresponding to the unknown $\dot{p}^\beta_\nu$, if $\operatorname{cls}\mu > \operatorname{cls}\nu$. Then we compute a row echelon form $M^\triangle_q$ of M_q using only row operations. This is equivalent to an involutive head autoreduction of the set Λ. The rows of $M^\triangle_q$ define a new set $\Lambda^\triangle \subset S_q(T_x\mathcal{X}) \otimes V^*_\xi\mathcal{E}$. We are interested in which column the first non-vanishing entry of each row sits; the corresponding unknown $\dot{p}^\alpha_\mu$ is called the *leader* of the row.

Definition 3.1.2. *Let $\beta^{(k)}_q$ be the number of leaders that are of class k. These numbers are the* indices *of the symbol $\mathcal{M}_q$.*

The problem of δ-regularity concerns this definition. The class of a derivative is not invariant under coordinate transformations. In different coordinate systems we may thus obtain different values for the indices. δ-regular coordinates are distinguished by the fact that the sum $\sum_{k=1}^n k\beta^{(k)}_q$ takes its maximal value, as it gives the involutive size of $\Lambda^\triangle$ with respect to the Pommaret division and a class respecting term order.

If we recall the construction of formal power series solutions in Section 1.4, then the computation of the row echelon form corresponds to a special choice of the principal derivatives: the derivative p^α_μ is principal, if and only if $\dot{p}^\alpha_\mu$ is a leader. Thus the index $\beta^{(k)}_q$ determines the number of principal derivatives of order q and class k.

Proposition 3.1.3. *The symbol $\mathcal{M}_q$ with the indices $\beta^{(k)}_q$ is involutive, if and only if the matrix M_{q+1} of the prolonged symbol $\mathcal{M}_{q+1}$ satisfies*

$$\operatorname{rank} M_{q+1} = \sum_{k=1}^{n} k\beta^{(k)}_q . \tag{3.2}$$

Proof. One direction is a trivial consequence of the definition of a Pommaret basis. If $\Lambda^\triangle$ is such a basis, all elements of $\mathcal{M}^{\mathcal{P}}$ of degree $q+1$ are obtained via multiplicative prolongations and their number is given by the right hand side of (3.2). We postpone the proof of the converse to Chapter 6 where the concept of *local involution* will be introduced and where we will show that for the Pommaret division it is equivalent to involution. If the equality (3.2) holds, the basis $\Lambda^\triangle$ is locally involutive. $\square$

Let us consider this result at the level of the differential equation and not only the symbol. In the neighbourhood of a point $\rho \in \mathcal{R}_q$ we solve each equation in a local representation for a derivative. We can do this in such a way that each equation is solved for a different derivative and that this derivative is the leading derivative of the equation with respect to a class respecting term order. If the class of a leading derivative is k, we assign the multiplicative variables $x^1, \dots, x^k$ to the corresponding equation. If we do the same with the prolonged equation $\mathcal{R}_{q+1}$ in the neighbourhood of a point in $(\pi^{q+1}_q)^{-1}(\rho)$, then all equations of order $q+1$ in the local representation of $\mathcal{R}_{q+1}$ are algebraically dependent of the formal derivatives of the equations of order q in the representation of $\mathcal{R}_q$. If the symbol is involutive, we do not need all formal derivatives of these equations.

Proposition 3.1.4. *The symbol $\mathcal{M}_q$ of the differential equation $\mathcal{R}_q$ is involutive, if and only if all equations of order $q+1$ in the local representation of the prolonged equation $\mathcal{R}_{q+1}$ are algebraically dependent of the formal derivatives of the equations of order q in $\mathcal{R}_q$ with respect to their multiplicative variables only.*

Proof. Obviously, we may prepare our local representation of $\mathcal{R}_q$ in such a way that applying (1.50) yields the symbol directly in row echelon form. All equations obtained by prolonging each equation only with respect to its multiplicative variables are independent, as they possess distinct leading derivatives. Since there are $\beta^{(k)}_q$ equations of order q and class k in our local representation of $\mathcal{R}_q$, this yields at least $\sum_{k=1}^{n} k\beta^{(k)}_q$ independent equations of order $q+1$ in the local representation of $\mathcal{R}_{q+1}$. By assumption, $\operatorname{rank} \mathcal{M}_{q+1}$ equals this number. Hence the local representation of $\mathcal{R}_{q+1}$ cannot contain more independent equations of order $q+1$. Note that it may well contain further independent equations, but these must be of lower order, i.e. they are integrability conditions (cf. the proof of Theorem 1.5.4). □

Example 3.1.5. We reconsider Example 2.1.9 in form of the second-order equation $\mathcal{R}_2$ defined by the system $u_{xx} = u_{yy} = 0$. Its symbol is

$$\mathcal{M}_2 : \begin{cases} \dot{u}_{yy} = 0\,, \\ \dot{u}_{xx} = 0\,. \end{cases} \tag{3.3}$$

If we order $x^1 = x$ and $x^2 = y$, we see that the first equation is of class 2 and the second one of class 1. The symbol matrix is already in row echelon form and for the second equation y is a non-multiplicative variable.

For the indices of $\mathcal{M}_2$ we find $\beta^{(2)}_2 = \beta^{(1)}_2 = 1$. The prolonged symbol $\mathcal{M}_3$ is defined by the equations

$$\mathcal{M}_3 : \begin{cases} \dot{u}_{yyy} = 0\,, & \dot{u}_{xyy} = 0\,, \\ \dot{u}_{xxy} = 0\,, & \dot{u}_{xxx} = 0\,. \end{cases} \tag{3.4}$$

Thus we get that $\operatorname{rank} M_3 = 4 \neq \beta_2^{(1)} + 2\beta_2^{(2)} = 3$ and according to our definition $\mathcal{M}_2$ is *not* involutive. In contrast, it is easy to see that $\mathcal{M}_3$ is involutive, as $\dim \mathcal{M}_3 = 0$: a vanishing symbol is always involutive (exercise!). We will show later in Chapter 6 that this is not accidental; any symbol becomes involutive after a finite number of prolongations. Here already one prolongation suffices.

At the level of the differential equation we get the following picture. Prolonging the equation $u_{xx} = 0$ with respect to it non-multiplicative variable yields the equation $u_{xxy} = 0$. It is easy to see that this equation is algebraically independent of all multiplicative prolongations. Note that the equation $u_{xxy} = 0$ is *not* an integrability condition; one easily checks that $\mathcal{R}_q^{(1)} = \mathcal{R}_q$ for all $q \geq 2$. As in the last chapter we call such an equation an *obstruction to involution*. $\triangleleft$

The leaders of an involutive symbol matrix in row echelon form cannot take arbitrary values. The following result is a reformulation of Lemma 2.4.2 showing that the leading exponents define a stable monoid ideal.

Lemma 3.1.6. *Let $\mathcal{M}_q$ be an involutive symbol and $M_q^{\triangle}$ its matrix in row echelon form where the columns have been ordered according to a class respecting term order. If $\dot{p}_\mu^\alpha$ is a leader of $M_q^{\triangle}$ of class k, then $\dot{p}^\alpha_{\mu-1_k+1_j}$ is also a leader for all $j > k$.*

Example 3.1.7. We apply our criterion to the analysis of the symbol of the Einstein equations (1.73). In arbitrary coordinates it is rather unpleasant to write down the defining equations. So we exploit that involution is a local property: if we want to check involution at a point in the fibre over $x_0 \in \mathcal{X}$, we may use the locally geodesic coordinates mentioned in Example 1.6.4.

The form (1.74) permits an easy construction of the symbol at any point in the fibre over x_0. As no lower-order terms are present, it makes no difference whether we consider the symbol equations or the differential equations. Our first task is to determine the leaders. We distinguish four cases.

$i \neq n$ and $j \neq n$: The leader is $\partial_{x^n x^n} g_{ij}$.
$i = n$ and $j < n-1$: The leader is $\partial_{x^n x^{n-1}} g_{n-1,j}$.
$i = n$ and $j = n-1$: The leader is $\partial_{x^n x^{n-1}} g_{11}$.
$i = n$ and $j = n$: The leader is $\partial_{x^{n-1} x^{n-1}} g_{11}$.

In the last case several substitutions of other equations are necessary for obtaining an autoreduced form. The other cases are rather obvious. One easily verifies that the leaders satisfy Lemma 3.1.6.

The indices of the Einstein equations are therefore

$$\beta_2^{(n)} = \frac{n(n-1)}{2}\,, \quad \beta_2^{(n-1)} = n\,, \quad \beta_2^{(n-k)} = 0 \text{ for } 1 < k < n\,. \tag{3.5}$$

We mentioned already in Example 1.6.4 that for the first prolongation one may act as if (1.74) was indeed a differential equation and not just some relations valid only in the fibre over x_0. So the prolonged symbol is again easily constructed and the rank of its matrix evaluated to $\frac{1}{2}n^2(n+1)-n$. But this is just $n\beta_2^{(n)}+(n-1)\beta_2^{(n-1)}$ and, by Proposition 3.1.3, the symbol of the Einstein equations is involutive. Note that this implies that our coordinates are δ-regular, as in δ-singular coordinates the criterion of Proposition 3.1.3 always fails. ◁

For *first-order* symbols this lemma implies that the indices form an ascending sequence. For higher-order symbols this is in general not the case; a trivial counter example is the symbol of the second-order equation defined by $u_{yy} = u_{xy} = u_{xx} = 0$ where $\beta_2^{(1)} = 2 > \beta_2^{(2)} = 1$.

Corollary 3.1.8. *Let $\mathcal{M}_1$ be the involutive symbol of a first order equation in n independent and m dependent variables. Then*

$$0 \le \beta_1^{(1)} \le \beta_1^{(2)} \le \cdots \le \beta_1^{(n)} \le m\,. \tag{3.6}$$

Proof. This is an immediate consequence of Lemma 3.1.6. If $\dot{p}_k^\alpha$ is a leader of class k, then for any $j > k$ we have a leader $\dot{p}_j^\alpha$ of class j. Hence $\beta_1^{(k)} \le \beta_1^{(j)}$ for $k < j$. □

Example 3.1.9. Let us analyse the symbol of Maxwell's equations (1.64). We order the independent variables as (x, y, z, t) so that time derivatives have the highest class. With this ordering (1.64) yields the symbol matrix directly in row echelon form. In fact, as no lower-order terms are present in the equations, it makes no real difference whether we study the symbol or the differential equation. Thus we do not introduce any extra notations for the variables in the symbol equations.

The rows of maximal class, i. e. of class 4, correspond to the six evolution equations in (1.64a); the remaining two constraint equations (1.64b) are of class 3, as each can by solved for a z-derivative. Thus we obtain for the indices

$$\beta_1^{(4)} = 6\,, \quad \beta_1^{(3)} = 2\,, \quad \beta_1^{(2)} = \beta_1^{(1)} = 0\,. \tag{3.7}$$

They form indeed an ascending sequence of the form (3.6). Only for the z-components of the fields $\mathbf{E}$ and $\mathbf{B}$ several leaders appear, namely $E_z^{(z)}$ and $E_t^{(z)}$ (and correspondingly for $B^{(z)}$). This agrees with Lemma 3.1.6.

In order to check whether the symbol is involutive, we must determine what happens in the non-multiplicative prolongations. In our case, this means that we must differentiate the constraint equations (1.64b) with respect to t. We find

$$\partial_t(\nabla \cdot \mathbf{E}) = \nabla \cdot \mathbf{E}_t = \nabla \cdot (\nabla \times \mathbf{B}) = 0 \tag{3.8}$$

where in the first step we used multiplicative prolongations of equations in (1.64a) and the last step is a well-known result in vector analysis. The same result is obtained with the equations for the magnetic field $\mathbf{B}$. Thus we can rewrite any non-multiplicative prolongation as a linear combination of multiplicative ones and the symbol is involutive. $\lhd$

Finally, we prove the following two somewhat technical results. Every prolongation of an involutive symbol is again involutive. In principle, this is obvious from the properties of a Pommaret basis; however, we give here a direct computational proof. The second result concerns a kind of commutativity of projection and prolongation for equations with an involutive symbol. Both results will be needed in the next section for proving a simple criterion for involutive differential equations.

Proposition 3.1.10. *Let the symbol $\mathcal{M}_q$ of the differential equation $\mathcal{R}_q$ be involutive.*

(i) *The prolonged symbol $\mathcal{M}_{q+1}$ is involutive, too.*
(ii) *The equality $(\mathcal{R}_q^{(1)})_{+1} = \mathcal{R}_{q+1}^{(1)}$ holds.*

Proof. We start with Part (i); Part (ii) will be a corollary. Without loss of generality, we assume that the symbol matrix M_q is in row echelon form with the columns ordered according to a class respecting term order. Let $\dot{p}^\alpha_\mu$ be the leaders of M_q. By Proposition 3.1.4, the leaders of M_{q+1} are $\dot{p}^\alpha_{\mu+1_\ell}$ with $1 \leq \ell \leq \operatorname{cls}\mu$. The class of such a leader is ℓ.

To prove Part (i), we must show that prolonging a row in M_{q+1} with respect to a non-multiplicative variable leads to a linearly dependent equation. Thus let $k = \operatorname{cls}\mu$ and let $\ell \leq k$. Prolonging the rows of M_{q+1} with respect to all independent variables leads to rows with leaders of the form $\dot{p}^\alpha_{\mu+1_\ell+1_j}$ for $1 \leq j \leq n$. Three cases must be distinguished:

$j \leq \ell$: These are the prolongations with respect to multiplicative variables. All obtained rows are linearly independent, as their leaders are different.

$\ell < j \leq k$: Since $j \leq k$, the variable x^j is multiplicative for the row in M_q with the leader $\dot{p}^\alpha_\mu$. Thus we can first prolong this row with respect to x^j and then the result with respect to x^ℓ and obtain the same result. As now each prolongation is only with respect to a multiplicative variable, the resulting row is already contained in the rows of the previous case.

$j > k$: Let us denote the row in M_{q+2} by r. As in the previous case we interchange the prolongations, thus prolong first with respect to x^j and then with respect to x^ℓ to obtain r. Since $j > k$, $\dot{p}^\alpha_{\mu-1_k+1_j}$ is by Lemma 3.1.6 also a leader in M_q. Its class is greater than or equal k; thus x^k is a multiplicative variable for the corresponding row. By prolonging it with respect to x^k we see that $\dot{p}^\alpha_{\mu+1_j}$ is among the leaders of M_{q+1}. Since x^ℓ is multiplicative for this equation, we have found a row contained in the first case with leader $\dot{p}^\alpha_{\mu+1_\ell+1_j}$.
If we compare the two rows with leader $\dot{p}^\alpha_{\mu+1_\ell+1_j}$, we see that they are both prolongations with respect to x^ℓ of rows with the leader $\dot{p}^\alpha_{\mu+1_j}$. One of these was obtained by prolonging a row of M_q with respect to a non-multiplicative variable. Since $\mathcal{M}_q$ is involutive, this row is not contained in the set of independent rows we use for $\mathcal{M}_{q+1}$ but is a linear combination of those. Thus r can be written as x^ℓ-prolongation of this linear combination. If x^ℓ is a non-multiplicative variable for some of the rows involved, we apply the same analysis again. Since the considered leaders become smaller and smaller, this process eventually terminates and then we have expressed r as a linear combination of rows contained in the first case.

Hence we have shown that all independent rows in M_{q+2} can be obtained by prolonging the rows of M_{q+1} only with respect to their multiplicative variables. But this means nothing else than that $\mathcal{M}_{q+1}$ is involutive.

Part (ii) may look trivial at first sight, but it is not. On the left hand side only the integrability conditions that arise during the projection from order $q+1$ to order q are taken into account; then one prolongs again to order $q+1$. On the right hand side, however, we project from order $q+2$ to order $q+1$. Hence we have to show that all integrability conditions which occur during this projection are prolongations of integrability conditions of the projection of order $q+1$ to q.

In order to achieve this we make basically the same considerations as in the proof of Part (i) but now with the full equations instead of only with their highest-order part. Let $\Phi^\tau = 0$ be an equation of class k in a local description of $\mathcal{R}_q$. We have to analyse all equations of the form $D_\ell D_j \Phi^\tau = 0$. We distinguish two cases assuming always $\ell \leq j$:

$j \leq k$: In this case $D_\ell D_j \Phi^\tau = 0$ contains only prolongations with respect to multiplicative variables. By Proposition 3.1.4, these are the only independent equations of order $q+2$ in our local description of $\mathcal{R}_{q+2}$.

$k < j$: This implies that $D_j \Phi^\tau = 0$ was a prolongation with respect to a non-multiplicative variable. By Proposition 3.1.4 we can express it as a linear combination of prolongations of equations of $\mathcal{R}_q$ with respect to multiplicative variables and perhaps some integrability conditions. Following again the argument of the third case above, we find that $D_\ell D_j \Phi^\tau = 0$ is a linear combination of the equations contained in the first case and of prolonged integrability conditions.

Note that it is not necessary here to distinguish between $\ell \leq k$ and $\ell > k$, as we are allowed to use all prolongations to construct $(\mathcal{R}_q^{(1)})_{+1}$. □

Part (ii) says that for a differential equation with involutive symbol we obtain the same, if we either first prolong it twice and then project once or alternately prolong, project and prolong. Denoting prolongations by one order by ρ and projections by one order by π, this may concisely be written as $\rho \circ \pi \circ \rho = \rho^2 \circ \pi$. Note that it is crucial that we start on both sides with a prolongation; we may not omit the leading ρ.

3.2 Involutive Differential Equations

Based on the results in the last section, we finally introduce the notion of an involutive differential equation. It combines the geometric concept of formal integrability with the combinatorial ideas underlying involutive bases. We will see in the subsequent chapters that this indeed makes sense. In fact, it will turn out that involution provides a very powerful framework for analysing many different aspects of differential equations.

Definition 3.2.1. *The differential equation $\mathcal{R}_q$ is called* involutive*, if it is formally integrable and if its symbol $\mathcal{M}_q$ is involutive.*

In the literature, the term "involution" is often used in a rather imprecise manner. In particular, involution is sometimes taken as a synonym for formal integrability. While Definition 3.2.1 obviously implies that an involutive equation is also formally integrable, the converse is generally not true: involution is a stronger concept than formal integrability.

Example 3.2.2. The probably simplest example of a differential equation which is formally integrable but *not* involutive is the following second-order equation in two independent and one dependent variable:

$$\mathcal{R}_2 : \begin{cases} u_{yy} = 0 \,, \\ u_{xx} = 0 \,. \end{cases} \tag{3.9}$$

We have already shown in Example 3.1.5 that its symbol $\mathcal{M}_2$ is not involutive. But $\mathcal{R}_2$ is trivially formally integrable, as all right hand sides are zero. Since the prolonged symbol $\mathcal{M}_3$ is involutive, $\mathcal{R}_3$ is an involutive equation. ◁

The first important point about involution is that it is much easier to verify than formal integrability. Recall that Definition 1.4.2 requires to check infinitely many conditions, namely that no integrability conditions arise at any order of prolongation. Up to now, no finite criterion for formal integrability is known. But although formal integrability is a necessary condition for involution, the latter one can be checked in a finite manner.

Theorem 3.2.3. *$\mathcal{R}_q$ is an involutive differential equation, if and only if its symbol $\mathcal{M}_q$ is involutive and $\mathcal{R}_q^{(1)} = \mathcal{R}_q$.*

Proof. The necessity of the two given conditions is obvious from Definition 3.2.1, as they are weaker than the conditions there. The interesting part is their sufficiency and represents largely a simple corollary to Proposition 3.1.10 which can be applied since the symbol $\mathcal{M}_q$ is assumed to be involutive. By Part (ii) of it and our assumption on $\mathcal{R}_q^{(1)}$, we obtain $\mathcal{R}_{q+1}^{(1)} = (\mathcal{R}_q^{(1)})_{+1} = \mathcal{R}_{q+1}$. Furthermore, Part (i) implies that all prolonged symbols $\mathcal{M}_{q+r}$ are involutive, too, so that we can use this argument inductively at any order. Hence $\mathcal{R}_{q+r}^{(1)} = \mathcal{R}_{q+r}$ for all $r \geq 0$ and $\mathcal{R}_q$ is formally integrable. □

Now we understand the meaning of the at first sight rather technical Proposition 3.1.10. If the symbol $\mathcal{M}_q$ is involutive and at some order of prolongation integrability conditions are hidden, then at least some of these conditions must arise already during the *first* prolongation (here we meet again the idea of local involution to be discussed in Section 6.3); the remaining ones are prolongations of these. This is obviously a very strong result and forms the basis of any concrete computation with involutive equations. In particular, the Cartan–Kuranishi completion, which we will discuss in Section 6.2, depends crucially on it.

Example 3.2.4. Theorem 3.2.3 permits us to conclude that the Einstein equations are involutive. We showed already in Example 1.6.4 that $\mathcal{R}_2^{(1)} = \mathcal{R}_2$ and in Example 3.1.7 that the symbol $\mathcal{M}_2$ is involutive. Thus all conditions of the theorem are satisfied. ◁

Involutive differential equations possess local representations of a special form: the *Cartan normal form.* In order to avoid excessive notation we present it only for a first-order equation $\mathcal{R}_1$. This poses no real restriction, as every differential equation can be transformed into an equivalent first-order one (see Appendix A.3). The first two main points about the Cartan normal form are that it yields the symbol matrix immediately in row echelon form and that each equation in the system is solved for its principal derivative. Thus it looks as follows:

$$p_n^\alpha = \phi_n^\alpha(x^i, u^\beta, p_j^\gamma, p_n^\delta)\,, \qquad \begin{cases} 1 \le \alpha \le \beta_1^{(n)}\,, \\ 1 \le j < n\,, \\ \beta_1^{(n)} < \delta \le m\,, \end{cases} \tag{3.10a}$$

$$p_{n-1}^\alpha = \phi_{n-1}^\alpha(x^i, u^\beta, p_j^\gamma, p_{n-1}^\delta)\,, \qquad \begin{cases} 1 \le \alpha \le \beta_1^{(n-1)}\,, \\ 1 \le j < n-1\,, \\ \beta_1^{(n-1)} < \delta \le m\,, \end{cases} \tag{3.10b}$$

$$\vdots$$

$$p_1^\alpha = \phi_1^\alpha(x^i, u^\beta, p_1^\delta)\,, \qquad \begin{cases} 1 \le \alpha \le \beta_1^{(1)}\,, \\ \beta_1^{(1)} < \delta \le m\,, \end{cases} \tag{3.10c}$$

$$u^\alpha = \phi^\alpha(x^i, u^\beta)\,, \qquad \begin{cases} 1 \le \alpha \le \beta_0 \le m;, \\ \beta_0 < \beta \le m\,. \end{cases} \tag{3.10d}$$

The subsystem in the first line comprises all equations of class n, the one in the second line all of class $n-1$ and so on. The derivatives on the right hand side have always a class lower than or equal to the one on the left hand side. The subsystem in the last line collects all algebraic constraints; their number is denoted by β_0. The subsystems may be empty from a certain class on, as the indices satisfy $0 \le \beta_1^{(1)} \le \cdots \le \beta_1^{(n)} \le m$ (Corollary 3.1.8).

These purely structural aspects of the normal form (3.10) do not yet capture that our differential equation is involutive; they only express that we have chosen a local representation in triangular form. Any differential equation can be brought into such a form. The most important point about the Cartan normal form is that involution implies a number of relations between prolongations of the individual equations in (3.10). First of all, involution of the symbol $\mathcal{M}_1$ requires by Proposition 3.1.4 that any non-multiplicative prolongation can be written as a linear combination of multiplicative ones. Thus functions $A_{ij}^\beta(\mathbf{x}, \mathbf{u}, \mathbf{p})$, $B_i^\beta(\mathbf{x}, \mathbf{u}, \mathbf{p})$, and $C^\beta(\mathbf{x}, \mathbf{u})$ must exist such that whenever $1 \le \ell < k \le n$

$$\begin{aligned} D_k(p_\ell^\alpha - \phi_\ell^\alpha) = \sum_{i=1}^{k} \sum_{\beta=1}^{\beta_1^{(i)}} \left\{ \sum_{j=1}^{i} A_{ij}^\beta D_j(p_i^\beta - \phi_i^\beta) + B_i^\beta(p_i^\beta - \phi_i^\beta) \right\} + \\ + \sum_{\beta=1}^{\beta_0} C^\beta(u^\beta - \phi^\beta)\,, \end{aligned} \tag{3.11}$$

as x^k is non-multiplicative for all equations solved for an x^ℓ-derivative. It follows easily from an analysis of the appearing derivatives that no equations of class higher than k are needed on the right hand side. Strictly speaking, involution of the symbol requires only that in (3.11) the second-order derivatives cancel. The functions B_i^β, and C^β take already care that (3.11) does not lead to integrability conditions.

Formal integrability of $\mathcal{R}_1$ requires furthermore that the prolongation of the algebraic equations in (3.10d) does not lead to new equations. This implies the existence of functions $\bar{C}^\beta(\mathbf{x}, \mathbf{u})$ such that

$$\phi_k^\alpha - \sum_{\beta=\beta_0+1}^{m} \frac{\partial \phi^\alpha}{\partial u^\beta}\phi_k^\beta = \frac{\partial \phi^\alpha}{\partial x^k} + \sum_{\beta=1}^{\beta_0} \bar{C}^\beta \left(u^\beta - \phi^\beta\right) . \tag{3.12}$$

Note that involution of (3.10) requires not only that $\beta_0 \leq \beta_1^{(1)}$ but in fact that even on the right hand side of (3.10d) always $\beta \leq \beta_1^{(1)}$. Otherwise, some derivatives in the prolonged equations could not be eliminated by the differential equations in (3.10).

Example 3.2.5. As a larger example we study linear homogeneous first-order equations in one dependent variable. This is a very classical problem which found much interest in the late nineteenth century. Most textbooks on differential equations in the early twentieth century discussed it at length (see e. g. [32]). It leads to what is nowadays called the *Frobenius Theorem* in differential geometry (see Appendix C.3). In local coordinates the differential equations we are considering take the form

$$\mathcal{R}_1 \; : \; \left\{ \sum_{i=1}^{n} a_\tau^i(\mathbf{x}) p_i = 0 \, , \qquad 1 \leq \tau \leq t \, . \right. \tag{3.13}$$

Note that no terms of order zero appear.

We want to study under what conditions $\mathcal{R}_1$ is involutive. In order to facilitate the analysis of the symbol, we transform the system into a triangular form. By taking appropriate linear combinations we solve for some derivatives. We assume that these are those of maximal class; otherwise we change the independent variables. So we arrive at a system of the form

$$\mathcal{R}_1 \; : \; \left\{ p_{n+1-\tau} + \sum_{i=1}^{n-t} \bar{a}_\tau^i(\mathbf{x}) p_i = 0 \, , \qquad 1 \leq \tau \leq t \, . \right. \tag{3.14}$$

If (3.13) is involutive, this is its Cartan normal form.

Now we must check whether relations of the form (3.11) hold; so we consider the non-multiplicative prolongations of the equations in (3.14). For the equation with index τ these are the prolongations with respect to the variables $x^{n+1-\sigma}$ for $1 \leq \sigma < \tau$. They lead to second-order equations

$$p_{n+1-\sigma,n+1-\tau} + \sum_{i=1}^{n-t} \bar{a}^i_\tau(\mathbf{x}) p_{i,n+1-\sigma} + \sum_{i=1}^{n-t} \frac{\partial \bar{a}^i_\tau(\mathbf{x})}{\partial x^{n+1-\sigma}} p_i = 0\,. \tag{3.15}$$

In order for $\mathcal{R}_1$ to be involutive, we must be able to express (3.15) as a linear combination of the remaining prolongations. The first term can be cancelled by subtracting the prolongation of the equation with index σ with respect to $x^{n+1-\tau}$: it is obtained by interchanging σ and τ in (3.15).

Next we study the remaining second-order derivatives in the first sum in (3.15) (and in the equation with σ and τ interchanged, respectively). Each of them can be cancelled individually by subtracting (respectively adding) the prolongation of the equation with index τ (respectively σ) with respect to x^i multiplied by $\bar{a}^i_\tau(\mathbf{x})$ (respectively $\bar{a}^i_\sigma(\mathbf{x})$). One easily checks that hereby all appearing derivatives p_{ij} with $1 \le i, j \le n-t$ cancel.

Thus we are able to eliminate all second-order derivatives implying that the symbol $\mathcal{M}_1$ is involutive. But if we perform all these eliminations, we obtain as integrability condition for each pair $1 \le \sigma < \tau \le t$ a first-order differential equation:

$$\begin{aligned}\sum_{i=1}^{n-t} \Biggl[& \frac{\partial \bar{a}^i_\tau(\mathbf{x})}{\partial x^{n+1-\sigma}} - \frac{\partial \bar{a}^i_\sigma(\mathbf{x})}{\partial x^{n+1-\tau}} + \\ & \sum_{j=1}^{n-t} \bar{a}^j_\sigma(\mathbf{x}) \frac{\partial \bar{a}^i_\tau(\mathbf{x})}{\partial x^j} - \sum_{j=1}^{n-t} \bar{a}^j_\tau(\mathbf{x}) \frac{\partial \bar{a}^i_\sigma(\mathbf{x})}{\partial x^j} \Biggr] p_i = 0\,.\end{aligned} \tag{3.16}$$

As it does not contain any derivative p_i with $i > n-t$, it cannot be expressed as a linear combination of equations contained in (3.14): $\mathcal{R}_1$ is formally integrable and thus involutive, if and only if in (3.16) the coefficients in the bracket vanishes.

For a more geometric version of this result, we consider the following set of vector fields

$$\bar{X}_\tau = \partial_{x^{n+1-\tau}} + \sum_{i=1}^{n-t} \bar{a}^i_\tau(\mathbf{x}) \partial_{x^i}\,, \qquad 1 \le \tau \le t\,. \tag{3.17}$$

They allow us to express (3.14) succinctly in the form $\bar{X}_\tau u = 0$. Comparing with (C.6) it is not difficult to see that the above conditions for involution are equivalent to the vanishing of the Lie brackets of the fields $\bar{X}_\tau$:

$$\bigl[\bar{X}_\tau, \bar{X}_\sigma\bigr] = 0\,, \qquad 1 \le \sigma < \tau \le t\,. \tag{3.18}$$

Thus we are in the same situation as in Example 1.4.3 where also the integrability conditions could be expressed as the vanishing of Lie brackets.

We may write the original system (3.13) in terms of the vector fields

$$X_\tau = \sum_{i=1}^{n} a^i_\tau(\mathbf{x})\partial_{x^i}\ , \qquad 1 \le \tau \le t\ . \tag{3.19}$$

They are linear combinations of the vector fields $\bar{X}_\tau$ – but with $\mathfrak{X}(M)$ considered as $\mathcal{F}(M)$-module (see Appendix C.3). The Lie bracket (C.6) is bilinear only with respect to linear combinations with real coefficients. Thus we cannot conclude that for an involutive system (3.13) the vector fields X_τ must commute. But it is easy to see that the commutator of two of the vector fields must be an $\mathcal{F}(M)$-linear combination of the X_τ, i.e. functions $C^\rho_{\sigma\tau}(\mathbf{x})$ must exist such that

$$[X_\sigma, X_\tau] = \sum_{\rho=1}^{t} C^\rho_{\sigma\tau}(\mathbf{x})X_\rho\ , \tag{3.20}$$

i.e. the distribution $\mathcal{D}$ spanned by these vector fields is closed under the Lie bracket. Thus the equation $\mathcal{R}_1$ is involutive, if and only if the distribution $\mathcal{D}$ is involutive.

In the classical language, a system (3.13) such that the vector fields X_τ satisfy (3.20) is called *complete*; this terminology goes back at least to Clebsch. If it is in the triangular form (3.14), one speaks of a *Jacobian system*.

The results above may be partially extended to non-linear systems using the so-called *Jacobi bracket* [32]. If F and G are two arbitrary functions defined on the first-order jet bundle $J_1\mathcal{E}$ where the fibre dimension of $\mathcal{E}$ is one, then the Jacobi bracket is given in local coordinates by

$$[F, G] = \frac{\partial F}{\partial p_i}\left(\frac{\partial G}{\partial x^i} + p_i\frac{\partial G}{\partial u}\right) - \frac{\partial G}{\partial p_i}\left(\frac{\partial F}{\partial x^i} + p_i\frac{\partial F}{\partial u}\right)\ . \tag{3.21}$$

It possesses similar properties as the commutator of two vector fields but it obeys a different Jacobi identity. Using the formal derivative (1.11) it is more concisely written as $[F, G] = F_{p_i}D_iG - G_{p_i}D_iF$. In a straightforward extension of the computations above one shows that the non-linear first-order system $\Phi^\tau(\mathbf{x}, u, \mathbf{p}) = 0$ is involutive, if and only if all Jacobi brackets $[\Phi^\sigma, \Phi^\tau]$ vanish modulo the system.

Finally, let us note that any complete system can be integrated solving only *ordinary* differential equations. If it consists of t equations, then it possesses $n - t$ functionally independent solutions $\psi_k(\mathbf{x})$ with $1 \le k \le n - t$ and its general solution is given by $\Psi(\psi_1(\mathbf{x}), \dots, \psi_{n-t}(\mathbf{x}))$ where Ψ is an arbitrary function.

Indeed, if $t = 1$, this is a simple consequence of the classical method of characteristics discussed in any introductory textbook on partial differential equations. For $t > 1$ we use an extension of it due to Mayer [117]. For simplicity, we formulate it in local coordinates and assume without loss of generality that we are dealing with the Jacobian system defined by the vector fields (3.17). Consider the one-forms

$$\omega^k = dx^k - \sum_{\tau=n-t+1}^{n} \bar{a}^k_\tau(\mathbf{x})dx^\tau\,, \qquad 1 \le k \le n-t\,. \tag{3.22}$$

They generate a Pfaffian system which is the annihilator $\mathcal{D}^0$ of the distribution $\mathcal{D}$ spanned by vector fields (3.17).

We choose a point $x_0 \in \mathcal{X}$ and introduce t parameters λ^τ and one new variable s by setting $x^\tau(s) = x^\tau_0 + \lambda^\tau s$ for $n-t < \tau \le n$. This defines a curve on a t-dimensional coordinate plane through x_0. On it $dx^\tau = \lambda ds$ and the Pfaffian system (3.22) becomes a system of $n-t$ ordinary differential equations

$$\frac{dx^k}{ds} = \sum_{\tau=n-t+1}^{n} \lambda^\tau \bar{a}^k_\tau(x^\ell, x^\sigma_0 + \lambda^\sigma s)\,, \qquad \begin{cases} 1\le k,\ell \le n-t\,, \\ n-t<\sigma\le n\,. \end{cases} \tag{3.23}$$

Its solutions are of the form $x^k(s) = \phi^k(s, x_0, \lambda^\sigma)$. If $u(\mathbf{x})$ is a solution of our Jacobian system, then one easily shows that $(d/ds)u\big(x^i(s)\big) = 0$. As in the classical theory of characteristics, solutions of the original system are constant along the solution of the (generalised) characteristic system (3.23). The parameters λ^τ can be eliminated by setting $\lambda^\tau = (x^\tau - x^\tau_0)/s$. Then as claimed above, $n-t$ functionally independent functions ψ_k exist such that $(d/ds)\psi_k\big(x^i(s)\big) = 0$.

This relation of the Pfaffian system (3.22) to the method of characteristics is not surprising, as the forms $\omega^k\big|_{x_0}$ generate the characteristic variety C_{x_0} of our equation at the point $x_0 \in \mathcal{X}$. ◁

The final proposition in this section should be rather obvious to any reader who has understood the meaning of involution. We mention it only for historical reasons, as it generalises an old result of Finzi [51] (cf. [129]) on systems with the same number of equations and dependent variables.

Proposition 3.2.6. *A differential equation $\mathcal{R}_q$ possesses identities or integrability conditions, if and only if either $\beta^{(n-1)}_q > 0$ or $\mathcal{R}^{(1)}_{q-1} \subset J_{q-1}\mathcal{E}$.*

Proof. The second case, $\mathcal{R}^{(1)}_{q-1} \subset J_{q-1}\mathcal{E}$, is trivial. It means that any local representation of $\mathcal{R}_q$ contains not only equations of order q but also of lower order. Obviously, this implies the existence of either identities or integrability conditions arising from the prolongation of these lower order equations. Thus we exclude this possibility and concentrate on the first case.

If $\beta^{(n-1)}_q > 0$, any local representation of $\mathcal{R}_q$ contains equations with non-multiplicative variables, as only equations of class n do not possess any.

Thus according to Proposition 3.1.4 the system must have either identities or integrability conditions. For the converse we note that $\beta_q^{(n-1)} = 0$ implies that all indices $\beta_q^{(k)}$ with $k < n$ must vanish; this can be shown analogously to Corollary 3.1.8. Thus there exists a local representation of $\mathcal{R}_q$ where all equations are of class n. Hence all their prolongations are algebraically independent and neither identities nor integrability conditions can arise from cross-derivatives. □

3.3 The Principal Symbol Revisited

In this section we prove a result on the relation of the highest index $\beta_q^{(n)}$ and the principal symbol of a differential equation. In principle, this result can be extended to give a fully constructive solution of the problem of δ-regularity (see [159]). However, this solution is computationally rather expensive; the approach presented in Section 2.4 is simpler and cheaper. So instead we use the mentioned relation only for a rigorous definition of under- and overdetermined equation.

Theorem 3.3.1. *Let $\mathcal{R}_q$ be a differential equation in n independent variables with principal symbol τ_χ. In different coordinate systems we may obtain different values for the index $\beta_q^{(n)}$ of the geometric symbol $\mathcal{M}_q$ at the point $\rho \in \mathcal{R}_q$ with $\pi(\rho) = x$; let $\tilde{\beta}_q^{(n)}$ be the maximum of these values. Then at the point ρ*

$$\max_{\chi \in T_x^*\mathcal{X}} \operatorname{rank} \tau_\chi = \tilde{\beta}_q^{(n)} . \tag{3.24}$$

Proof. By the definition of $\tilde{\beta}_q^{(n)}$, we find in a neighbourhood of ρ a local description of $\mathcal{R}_q$ where the first $\tilde{\beta}_q^{(n)}$ equations $\Phi^\tau(\mathbf{x}, \mathbf{p}) = 0$ satisfy

$$\frac{\partial \Phi^\tau}{\partial p^\alpha_{[0,\dots,0,q]}} = \delta^\tau_\alpha . \tag{3.25}$$

Setting $\chi = \chi_i dx^i\big|_x$ has the following effect on the matrix of the principal symbol τ_χ at the point ρ: the diagonal elements of the top left $\tilde{\beta}_q^{(n)} \times \tilde{\beta}_q^{(n)}$ submatrix of $T[\chi]$ are of the form $(\chi_n)^q + \dots$ and the monomial $(\chi_n)^q$ appears nowhere else in the matrix. If we choose $\chi = dx^n\big|_x$, all entries vanish with the exception of these diagonal elements which become one. Hence the rank of τ_χ at ρ is $\tilde{\beta}_q^{(n)}$ for this particular choice of χ.

Assume that for some $\chi \in T_x^*\mathcal{X}$ the rank of τ_χ was greater than $\tilde{\beta}_q^{(n)}$. We choose in a neighbourhood of x a new coordinate system $\tilde{\mathbf{x}}$ given by $\tilde{\mathbf{x}} = \boldsymbol{\psi}(\mathbf{x})$

where the function ψ^n satisfies $d\psi^n\big|_x = \chi$. Obviously, such coordinates always exist, as this is only a condition at a single point. For the dependency on the derivatives of class n we obtain after this transformation

$$\frac{\partial\tilde{\Phi}^\tau}{\partial\tilde{p}^\alpha_{[0,\dots,0,q]}} = \frac{\partial\Phi^\tau}{\partial p^\alpha_\nu}\prod_{i=1}^{n}\left(\frac{\partial\psi^n}{\partial x^i}\right)^{\nu_i} . \tag{3.26}$$

At the point x we can evaluate the derivatives of ψ^n and the right hand side becomes the matrix $T[\chi]$ of the principal symbol τ_χ. By assumption, its rank is greater than $\tilde{\beta}^{(n)}_q$. Thus we have found a coordinate system where more than $\tilde{\beta}^{(n)}_q$ equations are of class n contradicting the definition of $\tilde{\beta}^{(n)}_q$. □

Recall that by definition of δ-regular coordinates, they must always yield the value $\tilde{\beta}^{(n)}_q$ for the index $\beta^{(n)}_q$. Thus for such coordinates the one-form $\chi = dx^n$ must lead to a principal symbol of maximal rank. One-forms that lead to a lower rank are sometimes called *systatic*.

Corollary 3.3.2. *If a one-form χ exists such that the principal symbol τ_χ of the differential equation $\mathcal{R}_q$ is surjective, then $\mathcal{R}_q$ is involutive.*

Proof. The principal symbol τ_χ can only be surjective, if the number t of (independent) equations in a local representation of $\mathcal{R}_q$ is less than or equal to the number m of dependent variables and if all these equations are of order q. The rank of τ_χ is then t and by Theorem 3.3.1 all equations are of class n. But this implies that $\mathcal{M}_q$ is involutive and that no integrability conditions arise; hence $\mathcal{R}_q$ is involutive. □

Example 3.3.3. Despite the computational complexity of this approach, Theorem 3.3.1 is sometimes quite useful for checking the δ-regularity of a coordinate system. We demonstrate this for the Yang–Mills equations (1.70). Their principal part is linear and given by $\eta^{ij}(\partial_{x^ix^j}A^a_k - \partial_{x^ix^k}A^a_j)$. Thus we obtain in these coordinates $\beta^{(n)}_2 = d(n-1)$, as for all equations with an index $k \neq n$ we may take $\partial_{x^nx^n}A^a_k$ as principal derivative of class n. In the d remaining equations with $k = n$ we choose $\partial_{x^{n-1}x^n}A^a_{n-1}$ as principal derivative of class $n-1$. Thus $\beta^{(n-1)}_2 = d$ and all other indices vanish.

The question is whether some of the equations we classified as being of class $n-1$ could be "promoted" to class n by a coordinate transformation, i.e. whether our coordinates are δ-regular. As only $\beta^{(n)}_2$ and $\beta^{(n-1)}_2$ are greater than zero, we are done once we know whether $\beta^{(n)}_2 = \tilde{\beta}^{(n)}_2$ and by Theorem 3.3.1 this can be decided with the help of the principal symbol. For the entries of its matrix we obtain

$$T^{a\ell}_{bk}[\chi] = \delta^a_b\eta^{ij}\chi_i\left(\delta^\ell_k\chi_j - \delta^\ell_j\chi_k\right) \tag{3.27}$$

with $1 \leq a, b \leq d$ and $1 \leq k, \ell \leq n$. The indices a, k label the rows (i.e. the equations in the system) and the indices b, ℓ the columns (i.e. the fields).

It is easy to see that these entries satisfy for any one-form χ the d identities $\chi_\ell T^{a\ell}_{bk}[\chi] = 0$, as then the expression in the parentheses vanishes. This implies that the rank of the matrix $T[\chi]$ can never be larger than $d(n-1)$. But from our analysis above, we know already that for $\chi = dx^n$ we obtain this rank. Thus this is indeed the maximal rank and our coordinates are δ-regular.

The special form of these identities implies immediately that the symbol $\mathcal{M}_2$ of the Yang–Mills equations is involutive, as a multiplication by χ corresponds to a differentiation. The principal part of the prolongation $\mathcal{R}_3$ has the form $\eta^{ij}(\partial_{x^i x^j x^\ell} A^a_k - \partial_{x^i x^k x^\ell} A^a_j)$ and thus consists of dn^2 equations. Contraction with $\eta^{k\ell}$ yields obviously zero and we have at least d dependencies between rows of the symbol matrix M_3 (corresponding to the d identities above) and its maximal rank is $d(n^2-1)$. On the other hand, the number of multiplicative variables of the symbol $\mathcal{M}_2$ is a lower bound for this rank and in our case $n\beta_2^{(n)} + (n-1)\beta_2^{(n-1)} = d(n^2-1)$. Hence $\mathcal{M}_2$ is involutive by Proposition 3.1.3. We determined already in Example 1.6.3 that $\mathcal{R}_2^{(1)} = \mathcal{R}_2$. Hence the Yang–Mills equations are involutive.

Note that the concrete form of neither the metric η^{ij} nor the structure coefficients C^a_{bc} plays a role for this result! In particular, it is independent of the signature of the metric and remains valid in curved space-times. In the analysis of the symbol the structure constants do not appear at all; the formal integrability requires only that they satisfy the Jacobi identity. $\lhd$

Example 3.3.4. In the same manner one may analyse the Einstein equations and give an alternative proof that their symbol is involutive. In locally geodesic coordinates the matrix of the principal symbol has the entries

$$\begin{aligned} T^{ij}_{k\ell}[\chi] = \chi_m \Big[&\eta^{jm}(\delta^i_\ell \chi_k + \delta^i_k \chi_\ell) + \eta^{im}(\delta^j_\ell \chi_k + \delta^j_k \chi_\ell) - \\ &2\eta^{ij}\chi_k\chi_\ell - \eta^{mn}\chi_m\chi_n(\delta^i_k\delta^j_\ell + \delta^i_\ell\delta^j_k)\Big]\,. \end{aligned} \tag{3.28}$$

Due to the symmetries of the metric and the Ricci tensor, T is a square matrix with $n(n+1)/2$ rows and columns; the former ones are labelled by the indices k, ℓ, the latter ones by i, j. One easily verifies the identities

$$\eta^{m\ell}\Big(T^{ij}_{km}[\chi]\chi_\ell - \frac{1}{2}T^{ij}_{m\ell}[\chi]\chi_k\Big) = 0\,. \tag{3.29}$$

They imply that the maximal rank of T cannot exceed $n(n-1)/2$. But we know already from Example 3.1.7 that in locally geodesic coordinates the index $\beta_2^{(n)}$ has this value. Thus no further identities exist and the coordinates are δ-regular.

We may again interpret the identities (3.29) as relations between the rows of the prolonged symbol matrix and a count similar to the previous example proves involution of the symbol. Note that (3.29) represents the principal symbol of the contracted Bianchi identity. Thus our calculations in locally geodesic coordinates yield an indirect proof of them. ◁

It is really surprising that one has difficulties to find in the literature rigorous definitions for such basic notions as under- and overdeterminacy. Often one uses the simple counting familiar from linear algebra: how many equations and how many unknown functions are there. Unfortunately, this may easily lead to wrong results, as the next example shows. The main point is that here a definition and a criterion are confused. In linear algebra, a linear system of equations is called underdetermined, if some components of the solution vector are not restricted by the system, i. e. may be chosen arbitrarily. It is trivial to show that this happens, if and only if the system comprises less independent equations than unknowns. For differential equations the situation is more complicated, as the "only if" is no longer true.

Example 3.3.5. Consider the $U(1)$-Yang–Mills equations on a two-dimensional space-time (i. e. electrodynamics in one space dimension):

$$\mathcal{R}_2 : \begin{cases} u_{tt} - v_{xt} = 0 \, , \\ u_{xt} - v_{xx} = 0 \, . \end{cases} \tag{3.30}$$

Obviously, we have a system with two equations and two unknown functions $u(x,t)$ and $v(x,t)$ (corresponding to A_2 and A_1). Thus one might expect that this should be a well-determined system; however, (3.30) is underdetermined! This can be seen as follows. One easily checks that (3.30) is invariant under the transformation[1] $u \to u + \partial_x \Lambda$ and $v \to v + \partial_t \Lambda$ where Λ is an arbitrary function of the independent variable x and t. Thus we may choose either u or v arbitrarily and still find a solution of (3.30). This is precisely the behaviour we expect of an under- and not of a well-determined system.

In fact, all Yang–Mills equations in any dimension of space-time are underdetermined, as they remain invariant under the transformation

$$\bar{A}_i^a = A_i^a + \partial_{x^i} \Lambda^a + C^a_{bc} A_i^b \Lambda^c \tag{3.31}$$

where the number of arbitrary functions $\Lambda^a(\mathbf{x})$ equals the dimension of the underlying Lie group $\mathcal{G}$. In differential forms language, $\bar{\mathbf{A}} = \mathbf{A} + d\boldsymbol{\Lambda}$ where the function $\boldsymbol{\Lambda}$ takes its values in the Lie algebra $\mathfrak{g}$ of $\mathcal{G}$. ◁

[1] This is a simple example of a *gauge symmetry*; more on this topic is contained in Section 4.3.

With the help of the principal symbol we can give rigorous definitions of under-, over-, and well-determined equations. They reduce the problem to linear algebra. In concrete computations with local coordinates the classification requires only the analysis of a parametrised matrix, essentially the determination of its rank.

Definition 3.3.6. *An involutive differential equation* $\mathcal{R}_q$ *is* underdetermined *at a point* $\rho \in \mathcal{R}_q$*, if no one-form* $\chi \in T^*_{\pi^q(\rho)}\mathcal{X}$ *exists such that the principal symbol* τ_χ *of* $\mathcal{R}_q$ *is injective at* ρ*. If a one-form* χ *exists such that* τ_χ *is bijective at* ρ*, the differential equation* $\mathcal{R}_q$ *is* well-determined *or* normal*. In any other case* $\mathcal{R}_q$ *is* overdetermined*.*

These definitions are rather abstract but have simple interpretations in terms of the highest index $\beta_q^{(n)}$ of the symbol $\mathcal{M}_q$ (of course computed in δ-regular coordinates). We will see in Section 4.2 that they indeed capture the intuitive idea of an underdetermined equation.

Proposition 3.3.7. *An involutive differential equation* $\mathcal{R}_q$ *in* n *independent and* m *dependent variables is underdetermined, if and only if* $\beta_q^{(n)} < m$*. It is normal, if and only if* $\beta_q^{(n)} = m$ *and all other indices vanish.*

Proof. By definition, $\mathcal{R}_q$ is underdetermined, if and only if no one-form χ exists such that $\operatorname{rank} \tau_\chi = m$. By Theorem 3.3.1 this happens if and only if $\beta_q^{(n)} < m$. Similarly, we get the value of $\beta_q^{(n)}$ for a normal equation. It follows from the bijectivity of the principal symbol that in this case all other indices vanish, as a local representation cannot contain any further equations. Indeed, they would be of lower order and thus contribute zero rows to the matrix of the principal symbol. □

Based on this result it is very easy to characterise the local form of normal equations. In the case of an ordinary differential equation, normality excludes the presence of lower-order equations. Thus a first-order normal ordinary differential equation possesses local representations of the form $\dot{\mathbf{u}} = \mathbf{f}(x, \mathbf{u})$, i. e. each equation is solved for a derivative and we have one equation for each derivative. In the case of a partial differential equation we rename x^n as t. Then a normal first-order equation takes locally the *Cauchy–Kovalevskaya form* $\mathbf{u}_t = \mathbf{f}(\mathbf{x}, t, \mathbf{u}, \mathbf{u}_\mathbf{x})$: each equation is solved for a t-derivative and we have one equation for each t-derivative. Thus, as expected, normality requires more than just the same number of equations and unknown functions; a distinguished independent variable t must exist such that we can write the system in the above special form.

Recalling Definition 1.5.7 of a characteristic one-form, we see that for an underdetermined equation any one-form $\chi \in T^*\mathcal{X}$ is characteristic. For any other equation characteristic and systatic are equivalent and we may say that a necessary condition for a δ-regular coordinate system is that dx^n is not characteristic. Indeed, if δ-singular coordinates appear in applications, this is mostly due to the use of characteristic coordinates. A trivial example is the wave equation in the form $u_{xy} = 0$; this is just the differential equation version of Example 2.1.12.

Example 3.3.8. We continue with Example 3.3.5. The matrix of the principal symbol is

$$T[\chi] = \begin{pmatrix} (\chi_1)^2 & -\chi_1\chi_2 \\ \chi_1\chi_2 & -(\chi_2)^2 \end{pmatrix} . \tag{3.32}$$

If we multiply the first row by χ_2 and the second by χ_1, we obtain the same. Thus an upper bound for the rank of T is 1 and it takes indeed this value for every non-vanishing one-form χ. This yields $\tilde{\beta}_2^{(2)} = 1$ and it is not possible to find a coordinate system in which both equations of (3.30) are of class 2. The principal symbol is never injective and every one-form is characteristic. Hence $\mathcal{R}_2$ is indeed underdetermined, although there are as many equations as dependent variables. ◁

Note that Definition 3.3.6 applies only to *involutive* equations. Otherwise, no reasonable definitions of under- or overdeterminacy can be given, as hidden integrability conditions may change the character of the equation. We must be sure that we know all equations that possibly affect the principal symbol are exhibited.

Example 3.3.9. We augment (3.30) by the first-order equation $v_t - u_x = 0$ (this corresponds to the so-called *Lorentz gauge*). The arising equation $\bar{\mathcal{R}}_2$ is no longer formally integrable. If we nevertheless applied Definition 3.3.6, it would still classify the equation as underdetermined: the addition of a lower-order equation only adds a row of zeros to the matrix of the principal symbol. A local representation of $\bar{\mathcal{R}}_2^{(1)}$ is

$$\bar{\mathcal{R}}_2^{(1)} : \begin{cases} u_{tt} - u_{xx} = 0 \,, \quad u_{xt} - v_{xx} = 0 \,, \\ v_{tt} - v_{xx} = 0 \,, \quad v_{xt} - u_{xx} = 0 \,, \\ \qquad\qquad v_t - u_x = 0 \,. \end{cases} \tag{3.33}$$

It is not difficult to show that this equation is involutive. The matrix of its principal symbol is

$$T[\chi] = \begin{pmatrix} (\chi_1)^2 - (\chi_2)^2 & 0 \\ \chi_1\chi_2 & -(\chi_2)^2 \\ 0 & (\chi_1)^2 - (\chi_2)^2 \\ -(\chi_2)^2 & \chi_1\chi_2 \\ 0 & 0 \end{pmatrix} \tag{3.34}$$

and it is obviously injective for any one-form χ with $\chi_1 \neq \chi_2$. Thus the principal symbol is never bijective and we are dealing with an overdetermined equation. In fact, the basic idea behind such a *gauge fixing* consists of reducing the arbitrariness in the solutions of a differential equation. $\triangleleft$

3.4 Applications: The Janet Sequence

Let $\pi_0 : \mathcal{E}_0 \to \mathcal{X}$ and $\pi_1 : \mathcal{E}_1 \to \mathcal{X}$ be two fibred manifolds over the same base space $\mathcal{X}$. A differential operator Δ_0 of order q is a map $\Gamma_{loc}(\mathcal{E}_0) \to \Gamma_{loc}(\mathcal{E}_1)$ that can be "factored" in the form $\Delta_0 = \Phi_0 \circ j_q$ with a fibred morphism $\Phi_0 : J_q\mathcal{E}_0 \to \mathcal{E}_1$ (cf. Section 1.3). If $\mathcal{E}_0$ and $\mathcal{E}_1$ are vector bundles over $\mathcal{X}$ and Φ_0 is a vector bundle morphism, then Δ_0 is a linear differential operator. In this section we are exclusively concerned with linear operators.

Let $\pi_2 : \mathcal{E}_2 \to \mathcal{X}$ be a further vector bundle and $\Delta_1 : \Gamma_{loc}(\mathcal{E}_1) \to \Gamma_{loc}(\mathcal{E}_2)$ a linear differential operator. It represents the *compatibility conditions* of the operator Δ_0, if $\operatorname{im}\Delta_0 = \ker\Delta_1$. In other words, the inhomogeneous linear system $\Delta_0\sigma_0 = \sigma_1$ for a given right hand side $\sigma_1 \in \Gamma_{loc}(\mathcal{E}_1)$ has a solution, if and only if $\Delta_1\sigma_1 = 0$. Obviously, we may iterate this construction and search for the compatibility conditions of Δ_1 and so on. Denoting by $\Theta \subseteq \Gamma_{loc}(\mathcal{E}_0)$ the solution space of Δ_0, this leads to an exact sequence

$$0 \longrightarrow \Theta \longrightarrow \Gamma_{loc}(\mathcal{E}_0) \xrightarrow{\Delta_0} \Gamma_{loc}(\mathcal{E}_1) \xrightarrow{\Delta_1} \Gamma_{loc}(\mathcal{E}_2) \longrightarrow \cdots, \tag{3.35}$$

the *Janet sequence* of the linear operator Δ_0. By the definition of exactness, the equality $\operatorname{im}\Delta_i = \ker\Delta_{i+1}$ holds at every step. The sequence terminates as soon as a surjective operator is reached.

Example 3.4.1. A simple but classical example of a Janet sequence is formed by the basic operators of vector analysis: grad, curl, and div. Let us start with the differential equation $\operatorname{grad} u = \nabla u = 0$. If we add a right hand side, we obtain the system

$$u_x = v^1, \quad u_y = v^2, \quad u_z = v^3. \tag{3.36}$$

It is trivial to check that this an involutive system. It possesses solutions, if and only if the three functions v^α appearing on the right hand side satisfy the differential equation

$$v^2_z - v^3_y = 0, \quad v^3_x - v^1_z = 0, \quad v^1_y - v^2_x = 0, \tag{3.37}$$

in other words, if $\operatorname{curl}\mathbf{v} = \nabla \times \mathbf{v} = 0$. This corresponds to the well-known identity $\operatorname{curl} \circ \operatorname{grad} = 0$.

Now we may add a right hand side to (3.37), i. e. we consider $\operatorname{curl} \boldsymbol{v} = \boldsymbol{w}$. This equation is solvable, if and only if $\boldsymbol{w}$ satisfies

$$w_x^1 + w_y^2 + w_z^3 = 0 \,. \tag{3.38}$$

Of course, this means nothing but $\operatorname{div} \boldsymbol{w} = \boldsymbol{\nabla} \cdot \boldsymbol{w} = 0$ and we recover the next classical identity $\operatorname{div} \circ \operatorname{curl} = 0$.

We may represent these identities in the following exact sequence

$$\begin{gathered} 0 \longrightarrow \Gamma_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{R}) \xrightarrow{\mathrm{grad}} \Gamma_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{R}^3) \xrightarrow{\mathrm{curl}} \\ \Gamma_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{R}^3) \xrightarrow{\mathrm{div}} \Gamma_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{R}) \longrightarrow 0 \,. \end{gathered} \tag{3.39}$$

Readers familiar with differential geometry know of course that this is nothing but a special case of the *de Rham complex*. Let $\mathcal{M}$ be an n-dimensional manifold. We introduced in Appendix C.3 the vector bundles $\Omega^k(\mathcal{M})$ of differential k-forms. Together with the exterior derivative d as differential, they form the complex

$$0 \longrightarrow \Omega^0(M) \xrightarrow{d} \Omega^1(M) \xrightarrow{d} \cdots \xrightarrow{d} \Omega^n(M) \longrightarrow 0 \,. \tag{3.40}$$

If we consider global forms, it is in general not exact and its cohomology contains important topological information on the manifold $\mathcal{M}$. Locally, (3.40) is exact by the Poincaré Lemma (see Remark C.3.6). $\triangleleft$

Above we have defined differential operators between spaces of smooth sections. Of course, one could consider instead, say, distributional solutions. In principle, the exactness of the Janet sequence depends on the chosen function spaces and its verification requires the corresponding existence theory. Even for linear operators, this is a rather hard subject. In Section 5.4 we will discuss the existence and uniqueness of analytic solutions for arbitrary involutive equations. But already the extension to smooth solutions is only possible under additional assumptions on the structure of the equation (an example of such an extension will be given in Section 5.6). For this reason, we consider the *formal* Janet sequence, i. e. we are only concerned with formal power series solutions.

This restriction permits us to take a purely algebraic approach to the problem of compatibility. The basic idea is to ignore the function spaces on which our differential operators act and to consider instead free modules over the ring of linear differential operators. We must use *right* modules, as otherwise we do not obtain module homomorphisms. This inverts the arrows in the sequence (3.35) and we obtain a free resolution in the form (B.17).

Formal series are local objects, so we go into a coordinate chart where the linear differential operator $\Delta_0 : \Gamma_{\mathrm{loc}}(\mathcal{E}_0) \to \Gamma_{\mathrm{loc}}(\mathcal{E}_1)$ takes the form

$$v^\beta(\mathbf{x}) = \sum_{\alpha=1}^{m_0} \sum_{0\leq|\mu|\leq q} A^\beta_{\alpha\mu}(\mathbf{x}) D^\mu u^\alpha(\mathbf{x}) . \tag{3.41}$$

Here m_i denotes the fibre dimension of the vector bundle $\mathcal{E}_i$. We assume that the coefficients $A^\beta_{\alpha\mu}$ are elements of some function field $\mathbb{F}$ and introduce as in Example 2.2.4 the ring $\mathcal{D} = \mathbb{F}[\partial_1, \dots, \partial_n]$ with $\partial_i = \partial/\partial x^i$. We associate with Δ_0 the (left) $\mathcal{D}$-module homomorphism

$$D_0 : \begin{cases} \mathcal{D}^{m_1} & \longrightarrow \quad \mathcal{D}^{m_0} \\ (P_\beta) & \longmapsto \quad \Bigl(\sum_{\beta=1}^{m_1} \sum_{0\leq|\mu|\leq q} P_\beta \cdot A^\beta_{\alpha\mu}\partial^\mu\Bigr) \end{cases} . \tag{3.42}$$

Note that we act on the vector $\mathbf{P}$ from the right; thus we should think of it as a row vector. If we represent the homomorphism D_0 by the $m_1 \times m_0$ matrix $(\hat{D}_0)^\beta_\alpha = \sum_{0\leq|\mu|\leq q} A^\beta_{\alpha\mu}\partial^\mu$, we may write compactly $D_0(\mathbf{P}) = \mathbf{P}\cdot\hat{D}_0$. It is essential that we multiply from the right, as otherwise we would not obtain a left $\mathcal{D}$-module homomorphism (obviously $\hat{D}_0 \cdot (Q\mathbf{P}) \neq Q(\hat{D}_0 \cdot \mathbf{P})$ but $(Q\mathbf{P}) \cdot \hat{D}_0 = Q(\mathbf{P}\cdot\hat{D}_0)$ for arbitrary operators $Q \in \mathcal{D}$).

Lemma 3.4.2. *The sequence*

$$\cdots \longrightarrow \mathcal{D}^{m_2} \xrightarrow{D_1} \mathcal{D}^{m_1} \xrightarrow{D_0} \mathcal{D}^{m_0} \longrightarrow \mathcal{M} \longrightarrow 0 \tag{3.43}$$

with $\mathcal{M} = \operatorname{coker} D_0 = \mathcal{D}^{m_0}/\operatorname{im} D_0$ *is exact, if and only if the Janet sequence* (3.35) *is formally exact.*

Proof. We have $(D_i \circ D_{i+1})(\mathbf{P}) = \mathbf{P}\cdot\hat{D}_{i+1}\cdot\hat{D}_i = 0$ for all $\mathbf{P} \in \mathcal{D}^{m_{i+2}}$, if and only if $\Delta_{i+1}\circ\Delta_i = 0$. Furthermore, $D_i(\mathbf{P}) = \mathbf{P}\cdot\hat{D}_i = 0$ implies that $\mathbf{P}$ induces a compatibility condition of Δ_i and vice versa. All such conditions are dependent of Δ_{i+1}, if and only if $\mathbf{P} = \mathbf{Q}\cdot\hat{D}_{i+1}$ for some $\mathbf{Q} \in \mathcal{D}^{m_{i+2}}$. □

Thus determining the formal Janet sequence of Δ_0 is equivalent to computing the free resolution (3.43) of $\mathcal{M} = \operatorname{coker} D_0$ which amounts to syzygy computations in free $\mathcal{D}$-modules. As discussed in the Notes of the last chapter, the proof of Theorem 2.6.3 remains valid, if we substitute the commutative ring $\mathcal{P}$ by the non-commutative ring $\mathcal{D}$. Now we may reformulate it in the following form.

Theorem 3.4.3. *Assume that the rows of the matrix* $\hat{D}_0$ *form a Pommaret basis of the submodule* $\operatorname{im} D_0 \subseteq \mathcal{D}^{m_0}$ *for a class respecting term order and let* $n-d$ *be the lowest class of a row. Then the Janet sequence* (3.35) *terminates with the operator* Δ_d.

Proof. By Theorem 2.6.3 the free resolution (3.43) terminates under our assumptions with D_d. The claim follows now from Lemma 3.4.2. □

Example 3.4.4. We return to Example 3.4.1 above. Here $\Delta_0 = \text{grad}$ and thus $\hat{D}_0 = (\partial_x\ \partial_y\ \partial_z)^t$. Obviously, the rows of the matrix form a Pommaret basis and we have $d = 2$. Thus the Janet sequence (3.39) has the maximally possible length for an operator in $n = 3$ independent variables and stops only with the operator $\Delta_2 = \text{div}$. ◁

Remark 3.4.5. Let $\Delta_0 = \Phi_0 \circ j_q$ be a differential operator. We associate with it the submanifold $\mathcal{R}_q \subseteq J_q\mathcal{E}_0$ defined by the zeros of Φ_0. One may think that under the assumptions of Theorem 3.4.3 the differential equation $\mathcal{R}_q$ is involutive, but generally this is not the case. However, if q' is the minimal order of a row in $\hat{D}_0$, then $\mathcal{R}_q^{(q-q')}$ is involutive. This can be shown in the same manner as Lemma 2.4.1: we simply add the multiplicative prolongations of all lower-order equations. In the language of Section 6.5, the rows of $\hat{D}_0$ form a "skeleton" of the differential equation $\mathcal{R}_q^{(q-q')}$. ◁

For the remainder of the section we turn to the inverse problem: given a linear differential operator Δ_1, does another differential operator Δ_0 exist such that $\operatorname{im}\Delta_0 = \ker\Delta_1$? If yes, then we call Δ_0 a *parametrisation* of Δ_1. Somewhat surprisingly, this problem permits a constructive solution.

We present here only the theory for systems with constant coefficients. In this case $\mathcal{D}$ is commutative, i.e. the ordinary polynomial ring, and we may write the maps D_i as *left* multiplication of *column* vectors with the *transposed* matrix $\hat{D}_i^t$. Now our problem can be formulated as follows: given a module homomorphism $\beta : \mathcal{D}^q \to \mathcal{D}^r$, we would like to find a further module homomorphism $\gamma : \mathcal{D}^r \to \mathcal{D}^s$ such that $\operatorname{im}\beta = \ker\gamma$. We develop an algorithm that decides the existence of such a map and, if the answer is affirmative, explicitly determines it.

The basic observation is that a syzygy computation "goes in the wrong direction": it determines a map $\alpha : \mathcal{D}^p \to \mathcal{D}^q$ with $\operatorname{im}\alpha = \ker\beta$. Thus we must revert the directions of the arrows. This is achieved by taking the *dual*; in our commutative situation this means applying the Hom functor. For notational simplicity, we denote the module $\operatorname{Hom}_{\mathcal{D}}(\mathcal{D}^q, \mathcal{D})$ of all $\mathcal{D}$-linear maps from $\mathcal{D}^q$ to $\mathcal{D}$ by $(\mathcal{D}^q)^*$. It is trivially isomorphic to $\mathcal{D}^q$. Now β induces the dual map $\beta^* : (\mathcal{D}^r)^* \to (\mathcal{D}^q)^*$ defined by $\beta(\phi) = \phi \circ \beta$. If B is a matrix representation of the map β (with respect to some bases of the free modules $\mathcal{D}^q$ and $\mathcal{D}^r$), then a matrix representation of β^* (with respect to the dual bases) is simply the transposed matrix B^t.

The algorithm proceeds in five steps:

1. Compute β^* (as mentioned above, at the level of the matrices, this simply means to transpose the matrix of β).

2. Determine a map $\gamma^* : (\mathcal{D}^s)^* \to (\mathcal{D}^r)^*$ such that $\operatorname{im}\gamma^* = \ker\beta^*$. This corresponds to a syzygy computation.
3. Compute γ (transpose again).
4. Determine a map $\beta_c : \mathcal{D}^{q_c} \to \mathcal{D}^r$ such that $\operatorname{im}\beta_c = \ker\gamma$. This requires another syzygy computation.
5. Check whether $\operatorname{im}\beta_c = \operatorname{im}\beta$. If yes, then γ is a map with the required properties. Otherwise, no such map exists.

All maps and modules appearing in the course of the algorithm are sketched in the following diagram:

$$\begin{tikzcd}
\mathcal{D}^{q_c} \arrow[dr, "\beta_c"] & & \\
\mathcal{D}^{q} \arrow[r, "\beta"'] \arrow[d, dotted, leftrightarrow] & \mathcal{D}^{r} \arrow[r, "\gamma"] \arrow[d, dotted, leftrightarrow] & \mathcal{D}^{s} \arrow[d, dotted, leftrightarrow] \\
(\mathcal{D}^{q})^* & (\mathcal{D}^{r})^* \arrow[l, "\beta^*"'] & (\mathcal{D}^{s})^* \arrow[l, "\gamma^*"']
\end{tikzcd} \tag{3.44}$$

Let B and C be the matrices of the maps β and γ, respectively. Then the condition in the second step implies that $B^tC^t = 0$ and thus $CB = 0$ or equivalently $\operatorname{im}\beta \subseteq \ker\gamma$. But although $\operatorname{im}\gamma^* = \ker\beta^*$, we cannot conclude that $\operatorname{im}\beta = \ker\gamma$ (the functor $\operatorname{Hom}_{\mathcal{D}}(\cdot, \mathcal{D})$ is not exact, as free $\mathcal{D}$-modules are not injective [131, Section 2.4]).

Example 3.4.6. A special case arises in our algorithm, if the map β^* is injective. Then $s = 0$, as no syzygies exist, and the algorithm stops already in the second step. In this situation the map β obviously does not define the compatibility conditions of any other system.

As a concrete example we consider the *Poincaré–Steklov problem* in three dimensions. Given a vector field $\mathbf{v}$ and a function w, it looks for a vector field $\mathbf{u}$ satisfying $\nabla \times \mathbf{u} = \mathbf{v}$ and $\nabla \cdot \mathbf{u} = w$. These equations appear for instance in electrostatics (if we study only time-independent solutions, Maxwell's equations decouple and the electrical field $\mathbf{E}$ solves a Poincaré–Steklov problem) but also in the theory of the Navier–Stokes equations.

One can show that only a single compatibility condition exists for this problem: $\nabla \cdot \mathbf{v} = 0$. If we study the inverse problem via our algorithm, the matrix of the map β is

$$B = \begin{pmatrix} 0 & \partial_z & -\partial_y & \partial_x \\ -\partial_z & 0 & \partial_x & \partial_y \\ \partial_y & -\partial_x & 0 & \partial_z \end{pmatrix} . \tag{3.45}$$

It is easy to see that the rows of this matrix are linearly independent; thus the map β^* (with matrix B^t) is injective. At the level of the differential operators this means that the *adjoint* Poincaré–Steklov problem is underdetermined, as it comprises less equations than unknowns. $\triangleleft$

We still have to prove the correctness of our algorithm. One direction is trivial: if $\operatorname{im}\beta = \operatorname{im}\beta_c$, then γ is a map with the wanted properties. The difficult part is to show that no such map exists, if $\operatorname{im}\beta \neq \operatorname{im}\beta_c$. Recall that a *torsion element* of a $\mathcal{D}$-module $\mathcal{M}$ is an $\mathbf{m} \in \mathcal{M}$ such that for at least one $p \in \mathcal{D} \setminus \{0\}$ the product $p\mathbf{m}$ is zero; a module without torsion elements is *torsion free.* Now the correctness is implied by the following result.

Theorem 3.4.7. *Let $\beta : \mathcal{D}^q \to \mathcal{D}^r$ be a module homomorphism and the map $\beta_c : \mathcal{D}^q \to \mathcal{D}^r$ be constructed by the algorithm above. Then the following three statements are equivalent.*

(i) *The $\mathcal{D}$-module $\operatorname{coker}\beta = \mathcal{D}^r/\operatorname{im}\beta$ is torsion free.*
(ii) *A module homomorphism $\bar{\gamma} : \mathcal{D}^r \to \mathcal{D}^{\bar{s}}$ exists with $\operatorname{im}\beta = \ker\bar{\gamma}$.*
(iii) *The equality $\operatorname{im}\beta = \operatorname{im}\beta_c$ holds.*

Proof. (iii) trivially implies (ii); we simply choose $\bar{\gamma} = \gamma$. Assume that (ii) holds and consider the module homomorphism $\phi : \operatorname{coker}\beta \to \mathcal{D}^{\bar{s}}$ given by $\phi([\mathbf{f}]) = \bar{\gamma}(\mathbf{f})$ where $[\mathbf{f}]$ denotes the equivalence class of $\mathbf{f} \in \mathcal{D}^r$ in the factor space $\mathcal{D}^r/\operatorname{im}\beta$. The map ϕ is well-defined as $\operatorname{im}\beta = \ker\bar{\gamma}$ and obviously injective. Thus $\operatorname{coker}\beta$ is isomorphic to a submodule of a free module implying that it is torsion free and (ii) implies (i).

There remains to show that (i) implies (iii). We given an indirect proof. If $\mathcal{F}$ is the quotient field of $\mathcal{D}$, i.e. the field of rational functions in $x^1, \dots, x^n$, then β trivially extends to a vector space homomorphism $\tilde{\beta} : \mathcal{F}^q \to \mathcal{F}^r$ and similarly for the maps β_c and γ. We claim that $\operatorname{im}\tilde{\beta} = \operatorname{im}\tilde{\beta}_c$. Indeed, by construction, $\operatorname{im}\gamma^* = \ker\beta^*$ and thus also $\operatorname{im}\tilde{\gamma}^* = \ker\tilde{\beta}^*$; as $\mathcal{F}$ is a field, the functor $\operatorname{Hom}_{\mathcal{F}}(\cdot, \mathcal{F})$ is exact[2] and $\operatorname{im}\tilde{\beta} = \ker\tilde{\gamma} = \operatorname{im}\tilde{\beta}_c$.

Let us now assume that (iii) does *not* hold. This means $\operatorname{im}\beta \subset \operatorname{im}\beta_c$, as by construction $\operatorname{im}\beta \subseteq \operatorname{im}\beta_c$. Consider an element $\mathbf{f} \in \operatorname{im}\beta_c \setminus \operatorname{im}\beta$. We may identify it with an element of $\operatorname{im}\tilde{\beta}_c$. Since $\operatorname{im}\tilde{\beta} = \operatorname{im}\tilde{\beta}_c$, rational functions $a_\alpha \in \mathcal{F}$ exist such that $\mathbf{f} = \sum_{\alpha=1}^{q} a_\alpha \tilde{\beta}(\mathbf{e}_\alpha)$ (here $\mathbf{e}_\alpha$ denotes again the elements of the standard basis of a free module). We can choose a polynomial $h \in \mathcal{D}$ such that $h_\alpha = h a_\alpha \in \mathcal{D}$. This implies the existence of a syzygy over the ring $\mathcal{D}$,

$$h\mathbf{f} = \sum_{\alpha=1}^{q} h_\alpha \beta(\mathbf{e}_\alpha) \,, \tag{3.46}$$

[2] For arbitrary rings $\mathcal{R}$ the functor $\operatorname{Hom}_{\mathcal{R}}(\cdot, \mathcal{R})$ is only left exact; full exactness requires that $\mathcal{R}$ is a field [102, pp. 97–98].

so that $f \notin \operatorname{im}\beta$ but $hf \in \operatorname{im}\beta$. Thus $[f]$ is a torsion element in $\operatorname{coker}\beta$ and (i) cannot hold. $\square$

Example 3.4.8. Our algorithm permits us to derive the relation between the $U(1)$ Yang–Mills equations and the Maxwell equations discussed in Example 1.6.3. For this purpose we take the four equations $\nabla \cdot \mathbf{B} = 0$ and $\mathbf{B}_t = -\nabla \times \mathbf{E}$. Rewriting them as a linear differential operator yields

$$\Delta_0 = \begin{pmatrix} \partial_t & 0 & 0 & 0 & -\partial_z & \partial_y \\ 0 & \partial_t & 0 & \partial_z & 0 & -\partial_x \\ 0 & 0 & \partial_t & -\partial_y & \partial_x & 0 \\ \partial_x & \partial_y & \partial_z & 0 & 0 & 0 \end{pmatrix} . \tag{3.47}$$

As β is defined by the transposed matrix, this is simultaneously a matrix representation of the map β^* in (3.44). Following our algorithm yields in a straightforward computation the parametrisation

$$\Delta_1 = \begin{pmatrix} 0 & -\partial_z & \partial_y & 0 \\ \partial_z & 0 & -\partial_x & 0 \\ -\partial_y & \partial_x & 0 & 0 \\ -\partial_t & 0 & 0 & -\partial_x \\ 0 & -\partial_t & 0 & -\partial_y \\ 0 & 0 & -\partial_t & -\partial_z \end{pmatrix} . \tag{3.48}$$

(This matrix also represents the map γ^*). This result does not come as a surprise, as Δ_1 obviously encodes the familiar relations $\mathbf{B} = \nabla \times \mathbf{A}$ and $\mathbf{E} = -\nabla \cdot A_4 - \partial_t \mathbf{A}$. Thus we have rediscovered the vector potential.

It is interesting to note that if we omit in Δ_0 one of the first three rows, we still obtain the same operator Δ_1. But now $\operatorname{im}\beta \neq \operatorname{im}\beta_c$ and β_c is given by the original operator Δ_0. Thus the compatibility analysis automatically recovers the dropped equation. $\triangleleft$

Example 3.4.9. As a further quite famous example we consider the question whether the linearised Einstein equations are the compatibility conditions of some other system. It was posed by Wheeler about 1970 and much later solved by Pommaret [141].

The linearisation of the Einstein equations (1.73) around the Minkowski metric, i.e. the ansatz[3] $g_{ij} = \eta_{ij} + h_{ij}$, yields the following linear system

$$\eta^{k\ell}\left[\partial_{x^i x^j} h_{k\ell} + \partial_{x^k x^\ell} h_{ij} - \partial_{x^k x^i} h_{\ell j} - \partial_{x^k x^j} h_{\ell i}\right] = 0 . \tag{3.49}$$

[3] Recall from Example 1.6.3 that $\eta^{ii} = 1$ for $1 \leq i < n$, $\eta^{nn} = -1$ and $\eta^{ij} = 0$ in all other cases. η_{ij} are the entries of the inverse matrix.

Because of the symmetry we may restrict to $1 \leq i \leq j \leq n$ so that these are $n(n+1)/2$ equations for the same number of unknowns but the system is still underdetermined. Thus for $n = 4$ we have 10 equations.

For space reasons we omit to write down the matrix corresponding to the map β; its entries follow straightforwardly from (3.49). The matrix representation of the map γ is

$$C = \begin{pmatrix} 2\partial_x & 0 & 0 & 0 & \partial_y & 0 & 0 & \partial_z & 0 & \partial_t \\ 0 & 2\partial_y & 0 & 0 & \partial_x & \partial_z & 0 & 0 & \partial_t & 0 \\ 0 & 0 & 2\partial_z & 0 & 0 & \partial_y & \partial_t & \partial_x & 0 & 0 \\ 0 & 0 & 0 & -2\partial_t & 0 & 0 & -\partial_z & 0 & -\partial_y & -\partial_x \end{pmatrix} . \tag{3.50}$$

Computing its syzygies yields a system of 20 independent equations. Thus $\operatorname{im}\beta \neq \operatorname{im}\beta_c$ and (3.49) does not admit a parametrisation.

One can obtain further interesting results in differential geometry; for lack of space we refer again to [141]. We mention here only that the differential operator corresponding to γ^* has a simple geometric interpretation. We may write it as $h_{ij} = \eta_{kj}\partial_{x^i}L^k + \eta_{ik}\partial_{x^j}L^k$ or $h = \mathcal{L}_L\eta$, i.e. h is the Lie derivative of the Minkowski metric with respect to the vector field $L = L^k\partial_{x^k}$. $\lhd$

3.5 Notes

We remarked already at the end of the last chapter that many of the ideas around involution have their origin in the theory of differential equations (although we presented the theory first in algebraic form). But the Janet–Riquier theory mentioned there was not the only approach to deal with general systems of differential equations. A completely intrinsic theory based on a representation by an exterior ideal of differential forms was developed by Cartan [33] and Kähler [87]; a modern presentation of their work is contained in [27]. The Cartan–Kähler Theory is in particular very natural for many problems in differential geometry. A dual version was developed by Vessiot [189] (see also [171])

We used a combinatorial approach to involution; in the formal theory it is traditionally introduced as a cohomological concept. We only briefly sketch this theory here; more details can be found in [27, 42, 113, 139] (and references therein). For notational simplicity, we omit throughout the explicit reference to a base point, although all constructions are to be understood pointwise.

Consider the map $\delta : S_{r+1}(T^*\mathcal{X}) \to T^*\mathcal{X} \otimes S_r(T^*\mathcal{X})$ defined by the composition of the natural inclusion $S_{r+1}(T^*\mathcal{X}) \hookrightarrow T^*\mathcal{X} \otimes \bigotimes^r T^*\mathcal{X}$ with the canonical projection $T^*\mathcal{X} \otimes \bigotimes^r T^*\mathcal{X} \to T^*\mathcal{X} \otimes S_r(T^*\mathcal{X})$. By wedging from the left with $\Lambda_s(T^*\mathcal{X})$ and tensoring from the right with $V\mathcal{E}$ we can extend the map δ to another map $\Lambda_s(T^*\mathcal{X}) \otimes S_{r+1}(T^*\mathcal{X}) \otimes V\mathcal{E} \to \Lambda_{s+1}(T^*\mathcal{X}) \otimes S_r(T^*\mathcal{X}) \otimes V\mathcal{E}$. This leads to the so-called *δ-sequences*

$$\begin{aligned} 0 \longrightarrow S_r(T^*\mathcal{X}) \otimes V\mathcal{E} \longrightarrow T^*\mathcal{X} \otimes S_{r-1}(T^*\mathcal{X}) \otimes V\mathcal{E} \longrightarrow \cdots \\ \longrightarrow \Lambda_s(T^*\mathcal{X}) \otimes S_{r-s}(T^*\mathcal{X}) \otimes V\mathcal{E} \longrightarrow \cdots \\ \longrightarrow \Lambda_n(T^*\mathcal{X}) \otimes S_{r-n}(T^*\mathcal{X}) \otimes V\mathcal{E} \longrightarrow 0 \end{aligned} \tag{3.51}$$

where we set $S_i(T^*\mathcal{X}) = 0$ for $i < 0$. The *formal Poincaré Lemma* states that these sequences are exact for all $r > 0$.

Setting $\mathcal{M}_i = 0$ for $i < 0$ and $\mathcal{M}_i = S_i(T^*\mathcal{X}) \otimes V\mathcal{E}$ for $0 \leq i < q$, the δ-sequence (3.51) can be restricted to a sequence

$$0 \longrightarrow \mathcal{M}_{q+r} \longrightarrow T^*\mathcal{X} \otimes \mathcal{M}_{q+r-1} \longrightarrow \cdots \longrightarrow \Lambda_n(T^*)\mathcal{X} \otimes \mathcal{M}_{q+r-n} \longrightarrow 0 \tag{3.52}$$

which is still a complex but in general no longer exact. Its (bigraded) cohomology is called the *Spencer cohomology* of the symbol $\mathcal{M}_q$. We denote by $H^{s,r}(\mathcal{M}_q)$ the cohomology group at $\Lambda_s(T^*\mathcal{X}) \otimes \mathcal{M}_r$ and define the symbol $\mathcal{M}_q$ to be involutive, if the groups $H^{s,q+r}(\mathcal{M}_q)$ vanish for all $0 \leq s \leq n$ and all $r \geq 0$.

This approach can be connected to ours as follows. Let $\{e_1, \dots, e_n\}$ be an ordered basis of $T^*\mathcal{X}$. For a given symbol $\mathcal{M}_q$ we define for $1 \leq k \leq n$ the subspaces

$$\mathcal{M}_{q,k} = \{\rho \in \mathcal{M}_q \mid \rho(e_i, v_1, \dots, v_{q-1}) = 0\,,\ \forall 1 \leq i \leq k\,, \forall v_1, \dots, v_{q-1} \in T^*\mathcal{X}\} \tag{3.53}$$

and $\mathcal{M}_{q,0} = \mathcal{M}_q$. We call the basis *quasi-regular* (for $\mathcal{M}_q$), if

$$\dim \mathcal{M}_{q+1} = \sum_{k=0}^{n-1} \dim \mathcal{M}_{q,k}\,. \tag{3.54}$$

One can show that a symbol is involutive, if and only if quasi-regular bases exist for it. In fact, coordinates $(x^1, \dots, x^n)$ on $\mathcal{X}$ are δ-regular, if and only if the corresponding basis $\{\partial/\partial x^1, \dots, \partial/\partial x^n\}$ of $T^*\mathcal{X}$ is quasi-regular. The linear equations defining $\mathcal{M}_{q,k}$ correspond to the symbol equations of class greater than or equal to k and thus (3.54) is equivalent to our criterion (3.2).

Our definition of an underdetermined equation is more precise than the one given by Pommaret [139, Definition 6.6]. He defines an equation to be underdetermined, if a one-form χ exists such that the principal symbol τ_χ is surjective. While this condition covers the elementary case that any local representation comprises less equations than unknown functions, it fails for Example 3.3.5: the rank of the matrix $T[\chi]$ given by (3.32) is at most one, thus τ_χ can never be surjective.

Our approach generalises the definitions of Olver [129, Definition 2.86] for systems with the same number of equations and unknown functions. They are in so far not satisfactory, as he does not require the equation to be involutive. Thus it may happen that the addition of an integrability condition renders an underdetermined equation overdetermined or vice versa. It should be stressed again that this classification makes sense only for involutive equations!

The satisfaction of the compatibility conditions of a linear differential equation is an obvious necessary condition for the existence of solutions for the inhomogeneous equation. Whether it is also sufficient depends of course on in what kind of solutions one is interested. For the simple case of formal solutions the sufficiency is obvious. The results of Section 5.4 will imply the sufficiency for analytic solutions of analytic equations. In the case of constant coefficients stronger results have been obtained in [133] (and references therein).

The inverse problem of syzygies has not found much interest in commutative algebra; we are not aware of any work discussing it in the context of polynomials. The situation is different for differential equations where it is of considerable interest for applications in control theory [128, 140, 194]. A central concept here is the *controllability* of a system. Roughly it means that we can "steer" it from any point to any other point by a suitable choice of the controls.

Within the behavioural approach [138] a continuous system (or *behaviour*) is modelled by the solution space of a differential equation. In contrast to the traditional approach, no distinction into input, output and state variables is made. Somewhat surprisingly it turns out that controllability is equivalent to the existence of a parametrisation, i. e. a solution to the inverse problem.

Apparently, the presented algorithm for the construction of a parametrisation is due to Oberst [128]; our proof of Theorem 3.4.7 is a mixture of results found in [142] and [194, Section 1.2]. Pommaret [140] showed that the theory remains valid for linear systems with variable coefficients, if we use the formally adjoint operator instead of the transposed (the meaning of the $*$ in (3.44) must be changed). The extension of our proof to this more general situation will appear elsewhere.

Finally, we mention that the described algorithm for the solution of the inverse problem can be easily extended to compute the extension modules $\mathrm{Ext}^i_{\mathcal{D}}(\mathcal{M}, \mathcal{D})$. In fact, as the proof of Theorem 3.4.7 shows, $\mathrm{Ext}^1_{\mathcal{D}}(\mathcal{M}, \mathcal{D})$ is generated precisely by the constructed torsion elements. For the higher extension modules we first follow a few steps the free resolution (3.43) and then apply the same algorithm [143].

4. The Size of the Formal Solution Space

Knowing what is big and what is small is more important than being able to solve partial differential equations.

Stan Ulam

The results in this chapter should be used with some care. We introduce measures for the size of the formal solution space of a differential equation. First of all, one cannot stress strongly enough that this only concerns *formal* solutions, i. e. formal power series satisfying the given equation. In contrast to Chapter 5, we do not bother here about the convergence of these series and thus the results do not imply any statement on the existence of strong solutions. So one might say that we perform in this chapter only combinatorial games. Nevertheless, there are situations where the results can be quite useful. This should be evident from the fact that scientists like Cartan and Einstein actively participated in these "games".

In the first section we briefly discuss the concept of a "general" solution. This word is often used but hardly anybody cares to give a rigorous definition. We do not give one either, as this turns out to be rather difficult and of doubtful value. So we only discuss some of the arising problems in order to motivate why we afterwards restrict to the formal solution space where the situation is much simpler.

Section 4.2 introduces the classical tools of the formal theory for measuring the size of the formal solution space: the Cartan characters and the Hilbert function. We exhibit the connection between them and show how they can be explicitly determined for an involutive equation. In applications to physics, one is interested in subtracting the effect of gauge symmetries; this is easily done using a pseudogroup approach. In the following section we analyse how the Cartan characters and the Hilbert function change under certain operations on the differential equation. This leads to the rather general concept of a differential relation.

The final section of this chapter is more of a "historical" nature. We show how Einstein's approach based on the so-called strength of a differential equation fits into our framework. It turns out that the machinery of formal theory is more powerful and flexible than this approach, as the strength corresponds to just one of the Cartan characters.

4.1 General Solutions

The notion of a "general solution" or a "general integral" is notoriously difficult. Although it is intuitively clear what it is supposed to mean, it is rather difficult to give a rigorous definition. From the point of view of the functional analytic approach that has dominated the theory of partial differential equations for many decades, this is not so surprising as obviously the answer decisively depends on the considered function spaces. The classical idea of a parameter dependent expression which yields all solutions through suitable specialisations of the parameters becomes rather meaningless, if one is interested in solutions living in distribution spaces.

Example 4.1.1. If one admits distributions as solutions, already for ordinary differential equations the classical counting rules fail. As a trivial example we may consider the following first-order equation $xu' = 0$. The classical point of view is that the solution space of a scalar linear first-order equation is a one-dimensional vector space. In our case a basis of this space is obviously given by $u(x) = 1$. However, because of the singularity of the coefficient at $x = 0$, the equation admits further solutions in a distributional sense, for example $u(x) = H(x)$ where $H(x)$ denotes the Heaviside or step function defined by $H(x) = 0$ for $x < 0$ and $H(x) = 1$ for $x \geq 0$ [149]. Thus we have an at least two-dimensional solution space! $\triangleleft$

Another problem with the concept of a general solution is the existence of so-called *singular solutions*. They appear only in non-linear equations. The word "singular" does not refer to singularities in the solution but to the fact that such solutions cannot be obtained from a general solution; typically they arise as envelopes of families of solutions.

Example 4.1.2. Again it suffices to consider ordinary differential equations. The general solution of the following first-order system in two unknown functions u and v [53, p. 159]

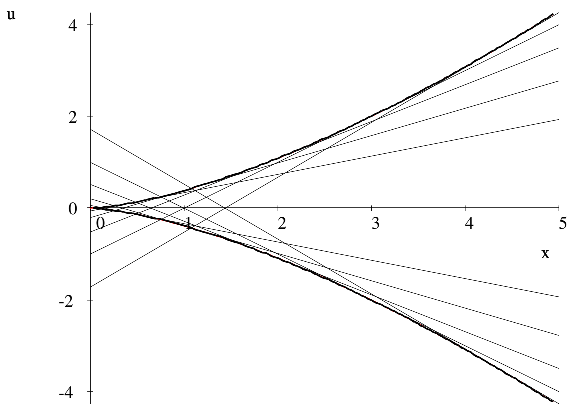

Fig. 4.1. Singular and general solution for (4.2)

$$(u')^2 + xu' + v' - u = 0\,, \tag{4.1a}$$
$$u'v' + xv' - v = 0 \tag{4.1b}$$

is the two-parameter family $u(x) = \alpha x + \alpha^2 + \beta$ and $v(x) = \beta x + \alpha\beta$. However, there exists a further one-parameter family of solutions, namely $u(x) = -\frac{1}{4}(x+c)^2 + c^2$ and $v(x) = -\frac{1}{4}c(x-c)^2$. Obviously, it is impossible to choose the parameters α, β in such a way that these solutions are part of the general solution.

A similar example for a single equation is [52, p. 242]

$$(u')^3 - xu' + u = 0\,. \tag{4.2}$$

Its general solution is the one-parameter family $u(x) = ax - a^3$. The two singular solutions are given by the quadratic equation $u^2 = \frac{4}{27}x^3$; they envelope the general solution (Figure 4.1). $\lhd$

In both examples above the origin of the problem is that the rank of the symbol matrix is not constant. In Example 4.1.1 the rank drops at $x = 0$, i.e. at a certain point of the base manifold $\mathcal{X}$. While this does not affect the classical solution space, it allows for additional distributional solutions. In Example 4.1.2, the determinant of the symbol matrix of the system (4.1) is $D = 2(u')^2 + 3xu' - v' + x^2$. Thus the rank of the symbol matrix drops whenever $D = 0$, i.e. on a submanifold of $J_1\mathcal{E}$, as this represents again a differential equation. The "general solution" consists of all those solutions for which $D \neq 0$. The "singular solutions" are the remaining ones; they arise as solutions of the overdetermined system obtained by adding $D = 0$ to (4.1).

Within our geometric approach to differential equations one may use a slightly different ansatz that avoids some of these problems, at least for smooth solutions. The downside is that we can only define *a* and not *the* general solution. Furthermore, the definition is not very constructive. The basic idea is very simple and rather old. Consider the inverse problem: we are given an expression containing some arbitrary constants and functions and we wish to determine differential equations for which this expression is a solution. For the "smallest" such differential equation our expression is a general solution. Following Hermann [76] we can make this idea more precise in our geometric framework.

Definition 4.1.3. *A* general solution *of the differential equation $\mathcal{R}_q \subseteq J_q\mathcal{E}$ is a subset $\Theta \subseteq \Gamma_{loc}(\mathcal{E})$ such that the following two conditions are satisfied.*

(i) *For all $r \geq 0$ we have $j_{q+r}(\Theta) \subseteq \mathcal{R}_{q+r}$.*
(ii) *For no $r \geq 0$ a submanifold $\bar{\mathcal{R}}_{q+r} \subset \mathcal{R}_{q+r}$ of lower dimension exists such that $j_{q+r}(\Theta) \subseteq \bar{\mathcal{R}}_{q+r}$.*

Θ is a local general solution *at a point $x \in \mathcal{X}$, if all sections $\sigma \in \Theta$ are defined in neighbourhood of x.*

A natural question is how one may describe such subsets $\Theta \subseteq \Gamma_{loc}(\mathcal{E})$. The intuitive idea is that it stems from an expression parametrised by some arbitrary constants and functions. Again Hermann [76] showed how this may be formulated in a geometric manner. Let $\mathcal{Y}_\lambda$ be manifolds and and $\mathcal{V}_\lambda$ vector spaces for $1 \leq \lambda \leq \ell$ and let $\mathcal{W}$ be a further manifolds. All manifolds and spaces are assumed to be finite-dimensional. Then a (local) *closed form representation* of a (local) general solution Θ is a smooth map

$$s : C^\infty(\mathcal{Y}_1, \mathcal{V}_1) \times \cdots \times C^\infty(\mathcal{Y}_\ell, \mathcal{V}_\ell) \times \mathcal{W} \to \Gamma_{loc}(\mathcal{E}) \tag{4.3}$$

such that $\operatorname{im} s = \Theta$. $\dim \mathcal{W}$ corresponds to the number of arbitrary constants in the general solution. Each factor $C^\infty(\mathcal{Y}_\lambda, \mathcal{V}_\lambda)$ represents $\dim \mathcal{V}_\lambda$ functions depending on $\dim \mathcal{Y}_\lambda$ arguments.

Example 4.1.4. We apply these ideas to the wave equation $u_{tt} - u_{xx} = 0$. As it is a linear equation, we do not have to bother about singular solutions. Furthermore, it is simple to derive closed form expressions for its general solution. Using the method of characteristics we obtain the classical superposition formula

$$s_1 : \begin{cases} C^\infty(\mathbb{R}, \mathbb{R}^2) & \longrightarrow & \Gamma_{loc}(\mathbb{R}^2 \times \mathbb{R}) \\ \binom{f}{g} & \longmapsto & f(x-t) + g(x+t) \end{cases} . \tag{4.4}$$

Alternatively, we take the D'Alembert solution of the initial value problem

$$s_2 : \begin{cases} C^\infty(\mathbb{R}, \mathbb{R}^2) & \longrightarrow & \Gamma_{\text{loc}}(\mathbb{R}^2 \times \mathbb{R}) \\ \binom{f}{g} & \longmapsto & \frac{1}{2}\left[f(x-t) + f(x+t)\right] + \frac{1}{2}\int_{x-t}^{x+t} g(\tau)d\tau \end{cases} . \tag{4.5}$$

In both cases, the right hand side is to be interpreted as a section of the trivial fibre bundle $\pi : \mathbb{R}^2 \times \mathbb{R} \to \mathbb{R}^2$. ◁

4.2 Cartan Characters and Hilbert Function

We are interested in measuring the size of the formal solution space. In particular, we would like to be able to compare two differential equations. A simple possibility would be to count the parametric derivatives. However, for most partial differential equations the result would be infinity which does not help us much in comparing different equations. Intuitively it is clear that although most partial differential equations have an infinite-dimensional solution space, there are differences in the size. Let us for example assume that we could find a closed form representation like (4.3). Then for some equations more arbitrary functions or manifolds $\mathcal{Y}_\lambda$ of higher dimensions are needed than for others.

The basic trick to capture these ideas in a rigorous manner consists of exploiting the natural grading of derivatives. Instead of simply counting all parametric derivatives, we count them order by order. Obviously, this yields at each order a finite number. Just comparing these numbers for one fixed order does not give us much information, but studying the asymptotic growth turns out to be very useful.

In Section 1.5 we introduced the indices $\beta_q^{(k)}$ of the symbol $\mathcal{M}_q$ of a differential equation $\mathcal{R}_q$. If the equation $\mathcal{R}_q$ is involutive, $\beta_q^{(k)}$ gives the number of principal derivatives of class k and order q. Now we introduce similar numbers $\alpha_q^{(k)}$ for counting the remaining derivatives of class k and order q which are of course parametric.

Definition 4.2.1. *Let $\mathcal{R}_q$ be an involutive differential equation in n independent and m dependent variables with indices $\beta_q^{(k)}$. The* Cartan characters *$\alpha_q^{(k)}$ of $\mathcal{R}_q$ are*

$$\alpha_q^{(k)} = m\binom{q+n-k-1}{q-1} - \beta_q^{(k)}, \qquad 1 \leq k \leq n . \tag{4.6}$$

By (A.4b), the binomial coefficient on the right hand side of (4.6) counts the total number of derivatives of order q and class k. The Cartan characters of an involutive equation form a descending sequence. Recall that the corresponding property for the indices, Corollary 3.1.8, holds only for first-order equations.

Proposition 4.2.2. *The Cartan characters $\alpha_q^{(k)}$ of an involutive differential equation $\mathcal{R}_q$ in m dependent variables satisfy*

$$\alpha_q^{(1)} \geq \alpha_q^{(2)} \geq \cdots \geq \alpha_q^{(n)} \geq 0 \,. \tag{4.7}$$

Proof. For $q = 1$ this follows immediately from Corollary 3.1.8, as in this case $\alpha_1^{(k)} = m - \beta_1^{(k)}$. For a higher-order equation we exploit the fact (shown in Appendix A.3), that it can be transformed into an equivalent involutive first-order equation with exactly the same Cartan characters. □

One of the key properties of an involutive symbol is that it is possible to compute the indices and the Cartan characters of its prolongations without explicitly determining their equations. This is a simple consequence of our combinatorial interpretation of involutive bases.

Proposition 4.2.3. *Let $\mathcal{M}_q$ be an involutive symbol with Cartan characters $\alpha_q^{(k)}$ and indices $\beta_q^{(k)}$. Then the Cartan characters and indices of its prolongations $\mathcal{M}_{q+r}$ are given by*

$$\alpha_{q+r}^{(k)} = \sum_{i=k}^{n} \binom{r+i-k-1}{r-1} \alpha_q^{(i)} \,, \tag{4.8a}$$

$$\beta_{q+r}^{(k)} = \sum_{i=k}^{n} \binom{r+i-k-1}{r-1} \beta_q^{(i)} \,. \tag{4.8b}$$

Proof. It suffices to prove (4.8b), as (4.8a) is then an immediate consequence of (4.6). We use an induction over r. Prolonging an equation of class k or higher with respect to x^k yields an equation of class k. By Proposition 3.1.4 these are the only independent equations of order $q+1$ in the prolongation, as $\mathcal{M}_q$ is involutive. Thus $\beta_{q+1}^{(k)} = \beta_q^{(k)} + \cdots + \beta_q^{(n)}$ and we have proven (4.8b) for $r = 1$.

Let us assume that (4.8b) holds for $r-1$. Since according to Proposition 3.1.10 the prolongation of an involutive symbol is again involutive, we can apply the formula for $r = 1$ to get $\beta_{q+r}^{(k)} = \beta_{q+r-1}^{(k)} + \cdots + \beta_{q+r-1}^{(n)}$. The inductive hypothesis yields

$$\beta_{q+r}^{(k)} = \sum_{i=k}^{n} \sum_{l=k}^{i} \binom{r+l-k-2}{r-2} \beta_q^{(i)} = \sum_{i=k}^{n} \binom{r+i-k-1}{r-1} \beta_q^{(i)} \,. \tag{4.9}$$

The last equality follows from an identity for binomial coefficients. □

Definition 4.2.4. *The* Hilbert function $h(r)$ *of a differential equation* $\mathcal{R}_q$ *gives the number of parametric derivatives of order* r.

The term Hilbert function stems from commutative algebra where such a function is used to measure the size of graded modules (see Appendix B.1); in differential algebra sometimes the term *differential dimension polynomial* is preferred [94]. Here it measures the size of the formal solution space which is parametrised by the parametric coefficients, as we have seen in Section 1.4. For an *involutive* differential equation $\mathcal{R}_q$ it is straightforward to determine the Hilbert function. We have

$$h(r) = \begin{cases} \dim \mathcal{R}_0^{(q)} & \text{for } r = 0\ , \\ \dim \mathcal{R}_r^{(q-r)} - \dim \mathcal{R}_{r-1}^{(q-r+1)} & \text{for } 0 < r \leq q\ , \\ \dim \mathcal{R}_r - \dim \mathcal{R}_{r-1} & \text{for } r > q\ . \end{cases} \tag{4.10}$$

However, this form is not very satisfying. In fact, as in commutative algebra, the Hilbert function $h(r)$ becomes a polynomial for sufficiently large values of r. More precisely, it is polynomial at least for $r \geq q$ and it is possible to derive a closed form expression for this polynomial.

Proposition 4.2.5. *The Hilbert polynomial* $H(r)$ *of the involutive differential equation* $\mathcal{R}_q$ *in* n *independent variables with Cartan characters* $\alpha_q^{(k)}$ *can be written in the form*

$$H(q+r) = \sum_{i=0}^{n-1} \left(\sum_{k=i}^{n-1} \frac{\alpha_q^{(k+1)}}{k!} s_{k-i}^{(k)}(0) \right) r^i \tag{4.11}$$

where the coefficients $s_k^{(i)}(0)$ *are the modified Stirling numbers introduced in Appendix A.4.*

Conversely, if the Hilbert polynomial of the differential equation $\mathcal{R}_q$ *has the form* $H(q+r) = \sum_{i=0}^{n-1} h_i r^i$, *then its Cartan characters can be computed recursively starting with* $k = n$ *by*

$$\alpha_q^{(k)} = (k-1)!\, h_{k-1} - \sum_{i=k+1}^{n} \frac{(k-1)!}{(i-1)!} s_{i-k}^{(i-1)}(0) \alpha_q^{(i)}\ . \tag{4.12}$$

Proof. By Theorem 1.5.4, $H(r) = \dim \mathcal{M}_r$ for $r \geq q$. It follows trivially from the definition of the indices and the Cartan characters, respectively, that $\dim \mathcal{M}_{q+r} = \sum_{i=1}^{n} \alpha_{q+r}^{(i)}$. Together with (4.8a) this yields

$$\begin{aligned} H(q+r) &= \sum_{i=1}^{n} \sum_{k=i}^{n} \binom{r+k-i-1}{r-1} \alpha_q^{(k)} \\ &= \sum_{k=1}^{n} \sum_{i=1}^{k} \binom{r+k-i-1}{r-1} \alpha_q^{(k)} = \sum_{k=1}^{n} \binom{r+k-1}{r} \alpha_q^{(k)}\ . \end{aligned} \tag{4.13}$$

Here we used in the last step the identity (A.36). Now we can enter the defining relation (A.31) of the modified Stirling numbers and obtain after a simple index shift (4.11). It defines obviously a one-to-one correspondence between the coefficients h_i of the Hilbert polynomial and the Cartan characters $\alpha_q^{(k)}$. A straightforward inversion yields (4.12). □

Of course, this result represents nothing but a reformulation of Proposition 2.5.2 for the special case of a Rees decomposition.

The Cartan characters (or the Hilbert function) measure the size of the formal solution space in a rather abstract manner. But in some sense they are the best measures available, as they do not require any assumptions about the form of the general solution. The main caveat is that the Cartan characters are local quantities defined only at a point $\rho \in \mathcal{R}_q$. In particular, if singular solutions exist, the Cartan characters will vary over $\mathcal{R}_q$. As the question of global solutions is beyond the scope of this book anyway, we can live with this restriction.

For the remainder of this section we want to relate the Cartan characters to numbers that can be more easily interpreted, namely to the number of arbitrary functions appearing in a closed form solution like (4.3). However, such an interpretation can be given only under some restrictive assumptions that are often *not* satisfied in applications. The main assumption is that the map s is purely algebraic, i. e. does not involve any differentiations or integrations. Thus we exclude cases as the D'Alembert formula (4.5). More precisely, we require that our solutions can be written in the form

$$\mathbf{u}(\mathbf{x}) = \mathbf{\Psi}\Big(\dots, F_\lambda\big(\phi_1(\mathbf{x}), \dots, \phi_k(\mathbf{x})\big), \dots\Big) . \tag{4.14}$$

Here F_λ represents the arbitrary functions. The one shown depends on k arguments which are defined by the fixed functions ϕ_i (these functions define the manifold $\mathcal{Y}_\lambda$). Note that no explicit dependency on the independent variables $\mathbf{x}$ is permitted; they appear only via the functions ϕ_i.

Example 4.2.6. The nonlinear equation $uu_{xy} - u_x u_y = 0$ admits such an algebraic representation of its general solution. It contains two arbitrary functions F_1 and F_2 each depending on one argument and is given by $u(x,y) = \Psi\big(F_1(\phi_1(x,y)), F_2(\phi_2(x,y))\big)$ with $\Psi(a,b) = ab$, $\phi_1(x,y) = x$ and $\phi_2(x,y) = y$. Thus our differential equation describes all functions of x and y that can be separated multiplicatively: $u(x,y) = F_1(x)F_2(y)$. ◁

Our goal is to determine the number f_k of arbitrary functions in (4.14) that depend on k arguments for $1 \leq k \leq n$. It turns out that if such a representation exists, these numbers are unique. Following [114], we introduce the following short hand

$$\begin{bmatrix} k \\ r \end{bmatrix} = \binom{k+r-1}{k-1} . \tag{4.15}$$

Then a smooth function of k arguments possesses $\left[\begin{smallmatrix} k \\ r \end{smallmatrix}\right]$ Taylor coefficients of order r (cf. (1.24)).

We expand (4.14) in a Taylor series. In order to avoid degenerate cases, we choose the expansion point $x_0 \in \mathcal{X}$ such that if $\partial\phi_j/\partial x^i \not\equiv 0$ for some $1 \le i \le n$, then $(\partial\phi_j/\partial x^i)(x_0) \neq 0$. The right hand side of (4.14) depends thus on

$$T(r) = \sum_{k=1}^{n} f_k \begin{bmatrix} k \\ r \end{bmatrix} \tag{4.16}$$

arbitrary Taylor coefficients of order r. If (4.14) is indeed the general solution of the differential equation $\mathcal{R}_q$, these coefficients must be in a one-to-one correspondence with the parametric derivatives of $\mathcal{R}_q$. At least for orders $r \ge q$, it is straightforward to give a necessary condition for the existence of such a correspondence. We write both the Hilbert polynomial $H(r)$ of $\mathcal{R}_q$ and $T(r)$ as explicit polynomials in r and equate the coefficients. This yields a triangular system for the numbers f_k which is easily solved by back-substitution. We phrase the result in form of a proposition.

Proposition 4.2.7. *If the involutive differential equation $\mathcal{R}_q$ with Cartan characters $\alpha_q^{(k)}$ possesses an algebraic representation of its general solution in the form* (4.14)*, then this representation contains f_k arbitrary functions depending on k arguments where the numbers f_k are determined by the recursion relation*

$$f_n = \alpha_q^{(n)} , \tag{4.17a}$$

$$f_k = \alpha_q^{(k)} + \sum_{i=k+1}^{n} \frac{(k-1)!}{(n-1)!} \left(s_{i-k}^{(i-1)}(0)\alpha_q^{(i)} - s_{i-k}^{(i-1)}(q) f_i \right) . \tag{4.17b}$$

Such an algebraic representation can exist only, if the solution of this recursion relation contains only non-negative integers.[1]

Now we see why Definition 3.3.6 of an underdetermined equation makes sense. By Proposition 3.3.7, we get for the highest index of an underdetermined equation $\beta_q^{(n)} < m$ and thus $\alpha_q^{(n)} > 0$. Hence any algebraic representation of the general solution of such an equation depends on at least

[1] One can show by explicit computations that even for fairly large values of n, the recursion relation (4.17) always yields integer solutions. Thus it is tempting to conjecture that this holds for all n. However, one can also show that this fact is highly non-trivial, as the modified Stirling number $s_{i-k}^{(i-1)}(q)$ is in general not divisible by the denominator $(n-1)!$.

one arbitrary function of n arguments. Obviously, we may simply choose $\mathbf{x}$ for these arguments and at least one component of $\mathbf{u}$ remains completely unrestricted by the differential equation.

Example 4.2.8. We consider again Maxwell's equations (1.64). We know from Example 1.6.1 that they are formally integrable and from Example 3.1.9 that their symbol is involutive. We determined already that $\beta_1^{(4)} = 6$, $\beta_1^{(3)} = 2$ and all other indices vanish. This yields the Cartan characters

$$\alpha_1^{(4)} = 0\,, \quad \alpha_1^{(3)} = 4\,, \quad \alpha_1^{(2)} = \alpha_1^{(1)} = 6\,. \tag{4.18}$$

The Hilbert polynomial is quadratic, $H(1+r) = 2r^2 + 12r + 16$, and for the numbers of arbitrary functions in an algebraic representation of the general solution we compute from (4.17)

$$f_4 = 0\,, \quad f_3 = 4\,, \quad f_2 = 2\,, \quad f_1 = 0\,. \tag{4.19}$$

In this particular case, these numbers are easy to interpret. Maxwell's equations are obviously not underdetermined. For each of the x- and y-components of the fields we have one evolution equation; thus we may prescribe in an initial value problem arbitrary values on a three-dimensional hypersurface. For each of the z-components we have two equations; as we will discuss in more detail in Chapter 5, this implies that, at least formally, we may prescribe initial data for them only on a two-dimensional surface. $\triangleleft$

For first-order equations the situation becomes very simple and we can solve (4.17) in closed form. We can even show that the solution consists always of non-negative integers.

Corollary 4.2.9. *For an involutive first-order equation $\mathcal{R}_1$ the solution of (4.17) is given by*

$$f_n = \alpha_1^{(n)} = m - \beta_1^{(n)}\,, \tag{4.20a}$$

$$f_k = \alpha_1^{(k)} - \alpha_1^{(k+1)} = \beta_1^{(k+1)} - \beta_1^{(k)}\,, \quad 1 \leq k < n\,. \tag{4.20b}$$

It contains only non-negative integers.

Proof. The proof is by a descending induction. The case $k = n$ is obvious. If we assume that (4.20b) holds for $k+1$, we get for k that

$$\begin{aligned} f_k = {} & \alpha_1^{(k)} - \alpha_1^{(k+1)} + \\ & (k-1)! \sum_{i=k+2}^{n-1} i!\,\alpha_1^{(i+1)} \left(i s_{i-k}^{(i-1)}(1) + s_{i-k+1}^{(i)}(0) - s_{i-k+1}^{(i)}(1) \right)\,. \end{aligned} \tag{4.21}$$

By Lemma A.4.3 the expression in the bracket vanishes for all values of i and k; this yields (4.20b) for k. It is obvious that all f_k are integers and Proposition 4.2.2 implies that they are non-negative. $\square$

One might be tempted to think that the general solutions of most partial differential equations should possess such algebraic representation. In fact, for higher-order equations it is a rather rare phenomenon. In particular, in the important case of normal equations no algebraic representations exist.

Example 4.2.10. Consider a normal differential equation $\mathcal{R}_q$ of order $q > 1$ in n independent and m dependent variables. By Proposition 3.3.7 the index $\beta_q^{(n-1)} = m$ and all other indices vanish. Thus its Cartan characters are

$$\alpha_q^{(n)} = 0 \,, \qquad \alpha_q^{(k)} = m\binom{q+k-1}{q-1} \,, \quad 1 \leq k < n \,. \tag{4.22}$$

Evaluating (4.17) yields $f_n = 0$, $f_{n-1} = mq$ and $f_{n-2} = mq(1-q)/2$. Thus for $q > 1$ and $n > 2$ we always get a negative value for f_{n-2}. $\lhd$

This simple example shows that the notion of an algebraic representation is only of limited use. One can consider more general representations involving differentiations and integrations of the arbitrary functions [160], but one obtains in general no unique values for the numbers f_k. Furthermore, a number of technical assumptions are required. So the value of such a formal counting of arbitrary functions is somewhat dubious. We have presented it here only because it helps to get a better understanding of how the Cartan characters and the Hilbert polynomial measure the size of the formal solution space. As already mentioned above, these abstract tools do not depend on any representation of the general solution and thus can always be applied without problems.

It is no coincidence that for first-order equation one always obtains reasonable values for the numbers f_k. As we will see in Chapter 5, these numbers are related to a concrete initial value problem. Solving this problem with a power series ansatz yields then indeed at each order a one-to-one correspondence between the Taylor coefficients of the initial data and the parametric coefficients of the solution.

For higher-order equations the situation becomes more difficult, as in general it is no longer so easy to relate the orders of the coefficients of the data and of the solution.[2] Only one thing is clear: the asymptotic growth of the number of parametric coefficients is dominated by the arbitrary functions with the highest number of arguments. The number k_0 of their arguments determines the degree of the Hilbert polynomial, $\deg H = k_0 - 1$ and is equal to the highest index k_0 such that $\alpha_q^{(k_0)} > 0$; it is sometimes called the *Cartan genus* of the differential equation and corresponds to the (Krull) dimension of an ideal in algebraic geometry.

[2] Of course, a simple solution is to rewrite the equation as an equivalent first-order equation as described in Appendix A.3 and then to analyse this equation.

The value $\alpha_q^{(k_0)}$ is known as *index of generality* [34, pp. 42–43]. It is not difficult to see that any representation of the general solution depends on precisely $\alpha_q^{(k_0)}$ arbitrary functions of k_0 arguments. In field theories $\alpha_q^{(k_0)}$ gives the number of degrees of freedom [164]; in algebraic geometry it corresponds to the multiplicity of an ideal.

4.3 Differential Relations and Gauge Symmetries

One can act with various kinds of transformations on a differential equation. In the simplest case, the transformation is just a change of variables in a local representation of the equation. Obviously, this does not affect the Hilbert function or the Cartan characters. However, more "brutal" transformations may well change these measures for the size of the formal solution space. A simple example is the reduction to a first-order equation as discussed in Appendix A.3; depending on the way the reduction is performed, the Cartan characters may change.

In this section we provide a rather general framework for such transformations: differential relations. These are slightly more complicated objects than differential equations; in a certain sense they contain two equations in one. Using this concept we can derive an equation for the Hilbert functions of differentially related differential equations.

Definition 4.3.1. *Let $\pi : \mathcal{E} \to \mathcal{X}$ and $\bar\pi : \bar{\mathcal{E}} \to \mathcal{X}$ be two fibred manifolds over the same base space $\mathcal{X}$. A* differential relation *of order $r, \bar r$ is a submanifold $\mathcal{T}_{r,\bar r} \subseteq J_r\mathcal{E} \times J_{\bar r}\bar{\mathcal{E}}$ such that $\mathcal{T}_{r,\bar r}$ is a fibred submanifold with respect to the two fibrations induced by π and $\bar\pi$, namely $\pi^r \times \mathrm{id} : J_r\mathcal{E} \times J_{\bar r}\bar{\mathcal{E}} \to \mathcal{X} \times J_{\bar r}\bar{\mathcal{E}}$ and $\mathrm{id} \times \bar\pi^{\bar r} : J_r\mathcal{E} \times J_{\bar r}\bar{\mathcal{E}} \to J_r\mathcal{E} \times \mathcal{X}$.*

The defining properties of a differential relation allow us to consider for a given section $\sigma : \mathcal{X} \to \mathcal{E}$ the submanifold $\mathcal{T}_{\bar r}[\sigma] = \mathcal{T}_{r,\bar r} \cap \big(\mathrm{im}\, j_r\sigma \times J_{\bar r}\bar{\mathcal{E}}\big)$ as a differential equation in $J_{\bar r}\bar{\mathcal{E}}$ and similarly for any section $\bar\sigma : \mathcal{X} \to \bar{\mathcal{E}}$ the submanifold $\mathcal{T}_r[\bar\sigma] = \mathcal{T}_{r,\bar r} \cap \big(J_r\mathcal{E} \times \mathrm{im}\, j_{\bar r}\bar\sigma\big)$ as a differential equation in $J_r\mathcal{E}$.

Definition 4.3.2. *Let $\mathcal{R}_q \subseteq J_q\mathcal{E}$ and $\bar{\mathcal{R}}_{\bar q} \subseteq J_{\bar q}\bar{\mathcal{E}}$ be two differential equations. They are* differentially related *via $\mathcal{T}_{r,\bar r} \subseteq J_r\mathcal{E} \times J_{\bar r}\bar{\mathcal{E}}$, if the following condition is satisfied: if the section σ is a solution of $\mathcal{R}_q$, then every solution of $\mathcal{T}_{\bar r}[\sigma]$ solves $\bar{\mathcal{R}}_q$, and conversely if $\bar\sigma$ is a solution of $\bar{\mathcal{R}}_q$, then every solution of $\mathcal{T}_r[\bar\sigma]$ solves $\mathcal{R}_q$.*

Example 4.3.3. A classical instance of a differential relation are *Bäcklund transformations* [137, 153]. Two additional conditions are imposed on them: (i) $\mathcal{T}_s[\bar\sigma]$ and $\mathcal{T}_{\bar s}[\sigma]$ are solvable, if and only if σ and $\bar\sigma$ are solutions of $\mathcal{R}_q$ and

$\bar{\mathcal{R}}_{\bar{q}}$, respectively; (ii) both $\mathcal{T}_s[\bar{\sigma}]$ and $\mathcal{T}_{\bar{s}}[\sigma]$ are equations of finite type. These two conditions allow for a geometric interpretation of Bäcklund transformations as so-called zero-curvature representations based on Remark 1.3.2; we do not pursue here this point of view but refer to Brandt [24].

Bäcklund [11] introduced these transformations for the sine-Gordon equation $u_{xt} = \sin u$. Here we even find an auto-Bäcklund transformation mapping the equation into itself. It has the following form:

$$\left(\frac{u+v}{2}\right)_x = a\sin\left(\frac{u-v}{2}\right), \qquad \left(\frac{u-v}{2}\right)_t = \frac{1}{a}\sin\left(\frac{u+v}{2}\right) \tag{4.23}$$

where $a \neq 0$ is an arbitrary real parameter. Obviously, given v this is an equation of finite type for u and vice versa. In order to check whether these finite type equations are formally integrable, we differentiate the first equation with respect to t and the second one with respect to x and enter (4.23) into the result. With the help of some elementary trigonometric identities we find that it is necessary and sufficient for the formal integrability that both u and v satisfy the sine-Gordon equation.

Bäcklund transformations play an important role in the theory of soliton equations. In particular, they are related to the inverse scattering transform and allow for the iterative construction of N-soliton solutions [1, Section 3.1]. For example, we consider the transformation (4.23) for $u(x,t) \equiv 0$ which leads to $v_x = -2a\sin(v/2)$ and $v_t = -(2/a)\sin(v/2)$. Combining these two equations yields the simple linear equation $v_x - a^2 v_t = 0$ whose general solution is trivially given by $v(x,t) = F(t + a^2 x)$ with a yet arbitrary function F. Entering it into the first equation of our system shows that F must satisfy the ordinary differential equation $aF' = -2\sin(F/2)$. This is a separable equation and an elementary integration gives $F(\tau) = 4\arctan\bigl(\exp(c - \tau/a)\bigr)$ where c is an arbitrary integration constant. Thus, starting from the trivial solution $u(x,t) \equiv 0$, we have constructed the single kink solution

$$v(x,t) = 4\arctan\bigl(\exp(ax + \frac{t}{a} + c)\bigr) \tag{4.24}$$

of the sine-Gordon equation. Iteration yields solutions with more kinks. $\triangleleft$

If two differential equations are differentially related, we can associate to each solution of one of them some solutions of the other one. But in general this is not a one-to-one correspondence. In the case of the Bäcklund transformation (4.23), $v(x,t)$ solves an equation of finite type; and each solution $u(x,t)$ of the sine-Gordon equation leads via the transformation to a one-parameter family of new solutions. Our goal is now to compare the Hilbert functions of two such equations. For simplicity we will assume that all differential equations involved are involutive.

Let $h(s)$ and $\bar{h}(s)$ be the Hilbert functions of the differential equations $\mathcal{R}_q$ and $\bar{\mathcal{R}}_{\bar{q}}$, respectively. We introduce two different Hilbert functions, $g(s)$ and $\bar{g}(s)$, for the differential relation $\mathcal{T}_{r,\bar{r}}$: for the former one we consider a differential equation $\mathcal{T}_r[\bar{\sigma}]$ and for the second one an equation $\mathcal{T}_{\bar{r}}[\sigma]$. We assume here that the Hilbert functions of these equations do not depend on the sections σ and $\bar{\sigma}$, respectively.

Proposition 4.3.4. *The above defined Hilbert functions satisfy for all arguments* $s \geq \max\{r, \bar{r}\}$

$$h(s) - g(s) = \bar{h}(s + \bar{r} - r) - \bar{g}(s + \bar{r} - r)\ . \tag{4.25}$$

Proof. The formal power series solution of $\mathcal{R}_q$ contains $h(s)$ parametric coefficients of order s. Via the differential relation $\mathcal{T}_{r,\bar{r}}$, a family of solutions of $\bar{\mathcal{R}}_{\bar{q}}$ depending on $\bar{g}(s)$ parametric coefficients of order s corresponds to each solution σ of $\mathcal{R}_q$. However, each of these solutions of $\bar{\mathcal{R}}_{\bar{q}}$ can be obtained not only by starting with σ but also by starting with any solution in a whole family parameterised at order s by $g(s)$ coefficients (obtained by applying $\mathcal{T}_{r,\bar{r}}$ "backwards"). Since for $s \geq \max\{r, \bar{r}\}$ a coefficient of σ of order s corresponds via $\mathcal{T}_{r,\bar{r}}$ to coefficients of order $s + \bar{r} - r$ of $\bar{\sigma}$, the formal power series solution of $\bar{\mathcal{R}}_{\bar{q}}$ contains $\bar{h}(s + \bar{r} - r) = h(s) - g(s) + \bar{g}(s + \bar{r} - r)$ parametric coefficients of order $s + \bar{r} - r$. □

For lower values of s, the equality (4.25) does not need to hold. This is due to the fact that in general we cannot say that the coefficients of σ of order $s \leq \max\{r, \bar{r}\}$ are related to coefficients of $\bar{\sigma}$ of order $s + \bar{r} - r$. We only know that the coefficients of σ *up to order* r correspond to the coefficients of $\bar{\sigma}$ *up to order* $\bar{r}$. Only when the relation $\mathcal{T}_{r,\bar{r}}$ has a special form, we may be able to make stronger statements. In particular, if both equations $\mathcal{T}_r[\bar{\sigma}]$ and $\mathcal{T}_{\bar{r}}[\sigma]$ are of Cauchy–Kovalevskaya form, the relation (4.25) holds for the full Hilbert functions.

Example 4.3.5. We exhibited the relation between Maxwell's equations and the $U(1)$ Yang–Mills equations in Example 1.6.3. Maxwell's equations use as dependent variables the electric and magnetic fields $\mathbf{E}$ and $\mathbf{B}$ which may be combined into the field strength tensor F given by (1.72); the Yang–Mills equations the vector potential A. As already mentioned, in the language of differential forms F may be considered as the exterior derivative of A. At the level of the components the two sets of fields are related by

$$F_{ij} = \frac{\partial A_i}{\partial x^j} - \frac{\partial A_j}{\partial x^i}\ . \tag{4.26}$$

This is nothing else but a differential relation $\mathcal{T}_{1,0}$. For a given potential A the field strength F is uniquely determined; hence the corresponding Hilbert function is simply $\bar{g}(s) = 0$. For given field strength F, (4.26) represents a first-order differential equation in the potential A; its Hilbert function is easily determined to

$$g(0) = 4\,, \quad g(1) = 10\,, \quad g(2+s) = 20 + \frac{37}{3}s + \frac{5}{2}s^2 + \frac{1}{6}s^3\,. \tag{4.27}$$

The Hilbert functions of the two differential equations are given by

$$h(0) = 4\,, \quad h(1) = 16\,, \quad h(2+s) = 36 + \frac{73}{3}s + \frac{9}{2}s^2 + \frac{1}{6}s^3 \tag{4.28}$$

for the second-order Yang–Mills equations and

$$\bar{h}(0) = 6\,, \quad \bar{h}(1+s) = 16 + 12s + 2s^2 \tag{4.29}$$

for the first-order Maxwell equations. As predicted by our theory, these Hilbert functions satisfy the equality (4.25) for all values $s \geq 0$. ◁

A somewhat similar situation arises in the formal analysis of gauge theories. Such theories are of great importance in physics; today all fundamental interactions are modelled by field theories invariant under gauge symmetries. In fact, this is the guiding principle in modern elementary particle physics. In this section we use a very broad definition of the term "gauge symmetry" that is much more general than the one appearing in elementary particle physics. In particular, it turns out that when a symmetry is to be promoted to a gauge symmetry depends generally on the physical interpretation and cannot be decided on the basis of purely mathematical considerations.

We assume that we are given a differential equation $\mathcal{R}_q \subseteq J_q\mathcal{E}$, the "field equations". We do not distinguish whether these are ordinary or partial differential equations and we do not assume any special structure of $\mathcal{R}_q$ like that they are Lagrangian or Hamiltonian field equations. Our considerations are valid for any kind of formalism to set up the field equations.

For us symmetries are diffeomorphisms $\mathcal{E} \to \mathcal{E}$. Thus we may treat them as global sections of the trivial bundle $\mathrm{pr}_1 : \mathcal{E} \times \mathcal{E} \to \mathcal{E}$ and form the jet bundles $J_r(\mathcal{E} \times \mathcal{E})$ with $r \geq 0$ over it. However, not every section of this bundle corresponds to a diffeomorphism. We define the open subbundles $I_r(\mathcal{E} \times \mathcal{E})$ consisting only of the jets of those local sections that are invertible. Now (local) symmetry transformations correspond to solutions of differential equations in $I_r(\mathcal{E} \times \mathcal{E})$; this is the basic idea behind the theory of Lie pseudogroups. For simplicity of notation, we identify local sections $\mathcal{U} \subseteq \mathcal{E} \to \mathcal{E} \times \mathcal{E}$ with maps $\mathcal{U} \to \mathcal{E}$.

Definition 4.3.6. *A* Lie pseudogroup *is a differential equation* $\mathcal{G}_r \subseteq I_r(\mathcal{E} \times \mathcal{E})$ *(often called a* finite Lie equation*) whose solutions satisfy the following two conditions.*

(i) *If the two sections* $\gamma_1 : \mathcal{U}_1 \subseteq \mathcal{E} \to \mathcal{E}$ *and* $\gamma_2 : \mathcal{U}_2 \subseteq \mathcal{E} \to \mathcal{E}$ *are both local solutions of* $\mathcal{G}_r$ *with* $\gamma_1(\mathcal{U}_1) \cap \mathcal{U}_2 \neq \emptyset$, *then the section* $\gamma_2 \circ \gamma_1$ *defined on some subset* $\mathcal{U}_3 \subseteq \mathcal{U}_1$ *with* $\gamma_1(\mathcal{U}_3) \subseteq \mathcal{U}_2$ *is also a local solution of* $\mathcal{G}_r$.

(ii) *If the section* $\gamma : \mathcal{U} \subseteq \mathcal{E} \to \mathcal{E}$ *is a local solution of* $\mathcal{G}_r$, *then the section* $\gamma^{-1} : \gamma(\mathcal{U}) \to \mathcal{U}$ *is also a local solution of* $\mathcal{G}_r$.

Obviously, this definition just rephrases the usual axioms of a local group: composition and inversion of group elements yield (where defined) again group elements. The associativity follows trivially from the properties of the composition of maps.

In our case, $\mathcal{E}$ is not an arbitrary manifold but the total space of a fibration $\pi : \mathcal{E} \to \mathcal{X}$. A *gauge transformation* γ preserves the fibration; in other words, it is a fibred morphism and induces a map $\bar{\gamma} : \mathcal{X} \to \mathcal{X}$ such that $\pi \circ \gamma = \bar{\gamma} \circ \pi$. For simplicity, we assume that $\bar{\gamma}$ is the identity map on $\mathcal{X}$. Such a gauge transformation is a *symmetry* of the differential equation $\mathcal{R}_q$, if it maps solutions into solutions. This means, if $\sigma : \mathcal{X} \to \mathcal{E}$ is a solution of $\mathcal{R}_q$, then $\gamma \circ \sigma$ is again a solution.

This is the above mentioned very general mathematical definition of a gauge symmetry; it is just a fibre preserving symmetry. In physical applications, gauge symmetries have a more narrow meaning which, however, incorporates already physical interpretations. Thus no mathematical criterion exists when a fibre preserving symmetry should be called a gauge symmetry in the physical sense. In physics, gauge symmetries represent a certain redundancy in the mathematical description of a theory. Solutions related by a gauge symmetry are identified, as one says that they correspond to the same physical state and hence cannot be distinguished by measurements.

These considerations imply that a physicist is not interested in the size of the full solution space but only in the size modulo the action of the gauge group, i.e. in the size of the "state space". We determine now this "gauge corrected" size from the Hilbert functions of the field equation $\mathcal{R}_q$ and the Lie pseudogroup $\mathcal{G}_r$.

Proposition 4.3.7. *Let* $h(s)$ *be the Hilbert function of the involutive differential equation* $\mathcal{R}_q$ *and* $g(s)$ *the Hilbert function of the involutive Lie pseudogroup* $\mathcal{G}_r$. *Then the size of the state space is measured by the gauge correct Hilbert function*

$$\bar{h}(s) = h(s) - g(s) . \tag{4.30}$$

Proof. We proceed as in the proof of Proposition 4.3.4. By definition of the Hilbert function, the formal power series solution of $\mathcal{R}_q$ depends on $h(s)$ parametric coefficients of order s. Using the gauge invariance $g(s)$ of these coefficients can be given arbitrary values. Thus $\bar{h}(s)$ parametric coefficients remain to distinguish physical states. □

Example 4.3.8. Let us consider again the $U(1)$ Yang–Mills equations. For them Maxwell's equations represent a gauge-invariant form, so that we can explicitly compare the result of a gauge correction with the values obtained directly from Maxwell's equations.

As already mentioned in Example 3.3.5, the $U(1)$ Yang–Mills equations are invariant under the gauge transformation $\bar{A}_i = A_i + \partial_{x^i}\Lambda$ (or $\bar{A} = A + d\Lambda$ in differential forms language) with an arbitrary function Λ. These transformations build the solution space of the following Lie equation

$$\mathcal{G}_1 : \left\{ \frac{\partial \bar{A}_i}{\partial x^j} - \frac{\partial \bar{A}_j}{\partial x^i} = 0 , \quad \frac{\partial \bar{A}_i}{\partial A_j} = \delta_i^j . \right. \tag{4.31}$$

$\mathcal{G}_1$ is involutive and in a four-dimensional space-time its Hilbert function is ($\dim I_1(\mathcal{E} \times \mathcal{E}) = 36$ here and (4.31) consists of 22 first-order equations and of no algebraic equations)

$$g(0) = 4 , \quad g(1) = 10 , \quad g(2 + s) = 20 + \frac{37}{3}s + \frac{5}{2}s^2 + \frac{1}{6}s^3 . \tag{4.32}$$

According to Proposition 4.3.7 we must subtract it from the Hilbert function (4.28) of the $U(1)$ Yang–Mills equations. Thus the gauge corrected Hilbert function is

$$\bar{h}(0) = 0 \quad \bar{h}(1) = 6 , \quad \bar{h}(2 + s) = 16 + 12s + 2s^2 . \tag{4.33}$$

Up to a shift of the argument by one which is necessary because of the lower order of Maxwell's equations, we find indeed the Hilbert function (4.29) of Maxwell's equations. It is no coincidence that (4.32) is identical with (4.27): the gauge symmetry is of course sitting in the differential relation $\mathcal{T}_{1,0}$; if A and $\bar{A}$ are connected by a gauge transformation, then they yields the same field strength F.

The gauge corrected Hilbert function corresponds to the following gauge corrected Cartan characters for the Yang–Mills equations:

$$\bar{\alpha}_2^{(4)} = 0 , \quad \bar{\alpha}_2^{(3)} = 4 , \quad \bar{\alpha}_2^{(2)} = \bar{\alpha}_2^{(1)} = 6 . \tag{4.34}$$

Of course, they are identical with the characters (4.18) of Maxwell's equations. Since the highest character vanishes after the gauge correction, the underdeterminacy of the Yang–Mills equations is solely due to the gauge freedom. In fact, we expect this for any reasonable physical theory.

The Hilbert function of an arbitrary Yang–Mills theory over a four-dimensional space-time is simply d times (4.28), respectively (4.33) after the gauge correction, where d is the dimension of the gauge group. This is not surprising, since it is determined by the symbol where the structure constants of the Lie algebra do not appear. The principal part of an arbitrary Yang–Mills theory consists of d copies of the principal part of the $U(1)$ Yang–Mills theory; there is no coupling between terms belonging to different generators of the Lie algebra. ◁

Gauge symmetries can be "fixed" by adding additional equations such that the augmented differential equation is no longer invariant. Obviously, this reduces the solution space. However, the gauge fixing conditions much be chosen so that we still have at least one solution in each orbit of the gauge group. Otherwise we have not removed gauge freedom but parts of the state space. If we are able to completely fix the gauge, then we obtain a differential equation such that each of these solutions lies on a different orbit. Thus its solution space can be bijectively mapped to the state space. The Hilbert function of this equation is then $\bar{h}(s)$. Using (4.12) we can also determine in advance its Cartan characters.

Example 4.3.9. The Lorentz gauge of Example 3.3.9 can be extended to general Yang–Mills theories. In does not eliminate all arbitrariness due to the gauge group, but it manages at least to render the originally underdetermined equation overdetermined. The gauge fixing condition is $\eta^{\mu\nu}\partial_\mu A^a_\nu = 0$. As it is of lower order, the augmented equation is not formally integrable but it becomes involutive after one prolongation with subsequent projection. It is straightforward to determine its Cartan characters:

$$\tilde{\alpha}_2^{(4)} = 0\,, \quad \tilde{\alpha}_2^{(3)} = 4d\,, \quad \tilde{\alpha}_2^{(2)} = 12d\,, \quad \tilde{\alpha}_2^{(1)} = 16d\,. \tag{4.35}$$

Comparing with (4.34) we find that while the two higher characters have the same value as the gauge corrected ones the two lower ones are larger. Thus the gauge fixed system still contains a residual gauge freedom. ◁

4.4 Einstein's Strength

Einstein [46] introduced an alternative measure for the size of the formal solution space of a differential equation $\mathcal{R}_q$: the *strength*. We show now how his approach fits into our framework. Let $\mathcal{R}_q$ be an involutive differential equation in n independent variables. The starting point is the function

$$Z(r) = h(r)/\left[{n \atop r}\right] \tag{4.36}$$

where h is the Hilbert function of $\mathcal{R}_q$. For values $r \geq q$ the function Z is rational in r: its denominator is of degree $n-1$; the numerator is then the Hilbert polynomial (4.11) whose degree is at most $n-1$. The function Z describes the growth of the number of parametric coefficients relative to the growth of number of Taylor coefficients of an analytic function of n variables. We expand it in powers of $1/r$:

$$Z(q+r) = Z_q^{(0)} + Z_q^{(1)}/r + O(1/r^2)\,. \tag{4.37}$$

Definition 4.4.1. *$Z_q^{(0)}$ is the* compatibility coefficient*, $Z_q^{(1)}$ the* strength[3] *of the differential equation $\mathcal{R}_q$. An equation with $Z_q^{(0)} = 0$ is called* absolutely compatible*.*

As the next proposition shows (but this is also clear from the interpretation of the function Z), an equation which is not absolutely compatible is necessarily underdetermined.

Proposition 4.4.2. *The compatibility coefficient and the strength of an involutive differential equation $\mathcal{R}_q$ in n independent variables with Cartan characters $\alpha_q^{(k)}$ are*

$$Z_q^{(0)} = \alpha_q^{(n)}\,, \tag{4.38a}$$

$$Z_q^{(1)} = (n-1)\Big(\frac{1}{2}n\,\alpha_q^{(n)} + \alpha_q^{(n-1)}\Big)\,. \tag{4.38b}$$

Proof. Proposition 4.2.5 provides the explicit form (4.11) of the Hilbert polynomial in terms of the Cartan characters. Together with (A.31) this yields readily the compatibility coefficient $Z_q^{(0)}$ by taking the limit $r \to \infty$. Subtracting the result from (4.37), multiplying by r and taking again the limit $r \to \infty$ gives the formula for the strength $Z_q^{(1)}$. □

In the case of gauge theories, Einstein used the gauge corrected Hilbert function for the definition of Z. Then the gauge corrected Cartan characters appear in Proposition 4.4.2. As we mentioned in the last section, we always expect for a reasonable physical theory that all underdeterminacy stems from the gauge freedom. After the gauge correction, the field equations of such a theory are always absolutely compatible and we get $Z_q^{(1)} = (n-1)\bar{\alpha}_q^{(n-1)}$. Thus up to a constant factor $n-1$ depending solely on the dimension of space-time, the strength coincides with Cartan's index of generality (we assume here that $\bar{\alpha}_q^{(n-1)} > 0$ which is the case for all classical examples of field theories).

[3] Actually, Einstein called $Z_q^{(1)}$ the *coefficient of freedom*; he only titled his whole discussion of the subject with the word strength. Later, other authors started to call $Z_q^{(1)}$ strength, although this is quite contradictory: the stronger the differential equation is in the sense of Einstein, the smaller is its strength!

In any case, we may conclude that Einstein's approach provides us with less information than a full involution analysis of the symbol. The compatibility coefficient and the strength correspond only to the two highest Cartan characters; the lower characters are ignored. One could probably recover them by determining further coefficients in the expansion (4.37). But it is clear that analysing the symbol yields the characters much faster and easier.

4.5 Notes

The notions of "general" and "singular" integrals have disappeared from modern textbooks on differential equations. Older books discuss them often at considerable length; see e. g. [52, Chapters VIII and XIII]. Our approach is based on ideas of Ampère [6] who was apparently the first to propose the definition that a general solution uniquely determine the differential equation and follows a formulation given by Hermann [76]. Within differential algebra, i. e. for differential equations with at most polynomial non-linearities, a rigorous definition of general and singular solutions exists. Apparently Ritt [151] was the first to introduce it; more recent work in this direction is due to Hubert [80, 81].

The here used notion of a differential relation was introduced recently in [164]. It generalises the classical concept of a Bäcklund transformation and allows for rather rough transformations of a differential equation. The gauge correction of Cartan characters is also fairly new. [160] presented a first approach based on an explicit representation of the gauge transformation; the approach via Lie pseudogroups is again from [164]. This article also discusses the use of the gauge corrected Cartan characters for counting the degrees of freedom of a gauge theory. The formal analysis of gauge fixings conditions was first proposed in [162]. More on symmetries of differential equations can be found in the textbooks [20, 129]

The Notes of Chapter 3 discussed the definition of involution via the Spencer cohomology. In this approach the Cartan characters arise as follows: assume that we are given a quasi-regular basis of $T^*\mathcal{X}$ for the symbol $\mathcal{M}_q$ and consider the subspaces $\mathcal{M}_{q,k}$ defined by (3.53); the Cartan characters are then the differences $\alpha_q^{(k)} = \dim \mathcal{M}_{q,k-1} - \dim \mathcal{M}_{q,k}$.

Einstein showed considerable interest in formal measures for the size of the solution space. His reason [46, p. 133] was the following:

> Given certain field variables and a system of field equations for them, the latter will not in general determine the field completely. There still remain certain free data for a solution of the field equation. The smaller the number of free data consistent with the system of field equations, the "stronger" is the system. It is clear that in the absence of any other viewpoint from which to select the equations, one will prefer the "stronger" system to a less strong one.

Around 1930 he discussed this question in his correspondence with Cartan [34]. Cartan tried to explain Einstein his theory of involutive systems; in fact, one letter is one of very few places where Cartan explained his theory in terms of partial differential equations and not of exterior differential systems. But for some reasons Einstein apparently never used this approach.

Many years later, he developed his own notion, the strength, in an appendix of his book on the meaning of relativity [46, pp. 133–139]. This appendix on a generalised theory of gravitation was introduced in the third edition and completely rewritten for the fifth edition under the title "Relativistic Theory of the Non-Symmetric Field". The second version starts with discussing the strength, as

> This discussion is of intrinsic interest quite apart from the particular theory presented here. For a deeper understanding of our problem, however, it is indispensable.

Of course, Einstein did not use the Hilbert function for defining the strength. He gave a much more cumbersome recipe for determining it. In order to obtain the numerator of (4.36), he counted the number of field equations, the number of identities satisfied by them, the number of identities of these identities and so on. Obviously, this process is prone to errors, in particular when applied to an equation which is not involutive. Indeed, one can find several examples in the literature where wrong results for the strength were computed. In contrast, the determination of the Cartan characters via the symbol is trivial.

The strength did not spark much interest. Only few authors applied it to various field theories [29, 78, 114–116, 127, 158, 167]. Penney [135] used it for some metaphysical speculations on the dimensionality of the world. He conjectured that the world is four-dimensional, because only in four dimensions the Maxwell and the Einstein equations possess the same strength. In fact, already Einstein [46, p. 139] remarked:

> It is surprising that the gravitational equations for empty space determine their field just as strongly as do Maxwell's equations in the case of the electromagnetic field.

However, if one computes the full set of (gauge corrected) Cartan characters for both theories, one sees that even in four dimensions only the highest non-vanishing one is identical; the other ones and thus the size of the formal solution spaces of the two theories are different [160]. Furthermore, there is no compelling physical reason why the two theories should have formal solution spaces of the same size.

Sué [174] was apparently the first who tried to relate Einstein's and Cartan's approach, especially for gauge theories. But he compared a strength which was corrected for the gauge freedom with Cartan characters which were not corrected, so some of his conclusions are misleading. The relation (4.38) between strength and compatibility index on one side and the two highest Cartan characters on the other side was first derived in [160].

5. Existence and Uniqueness of Solutions

In order to solve this differential equation you look at it till a solution occurs to you.

George Pólya

A fundamental issue in the theory of differential equations is to prove the existence and uniqueness of solutions. Before we discuss partial differential equations, we analyse in Section 5.1 the situation for general systems of ordinary differential equations, often called differential algebraic equations. For involutive equations it is here straightforward to extend the classical existence and uniqueness theorem.

As the formal theory is closely connected to the study of power series solutions, it is relatively easy to obtain results on *analytic* solutions of partial differential equations. In Section 5.2 we recall the Cauchy–Kovalevskaya Theorem for normal systems. After introducing in Section 5.3 the concept of a formally well-posed initial value problem, we generalise in the following section this famous theorem to arbitrary involutive systems, the Cartan–Kähler Theorem. We see here again why formal integrability is *not* sufficient for analysing partial differential equations but that one indeed needs involution.

From an application point of view the restriction to analytic solutions is very severe. This is the price we have to pay for the Cartan–Kähler Theorem requiring essentially no other assumptions than involution. At such a level of generality no stronger results can be expected. However, the technique of proof we use allows us in certain cases to go beyond analytic solutions. As a first simple example we extend in Section 5.5 the classical uniqueness theorem of Holmgren for C^1 solutions to arbitrary linear involutive systems.

The final section studies what we call "weakly overdetermined" systems and derives under certain structural assumptions some existence and uniqueness results for smooth solutions. This class includes in particular Maxwell's equations. Again we exploit the technique of the proof of the Cartan–Kähler Theorem to “lift” classical results on normal systems to overdetermined ones.

As we present in this chapter only local results, we will work throughout in local coordinates and therefore speak of systems.

5.1 Ordinary Differential Equations

In order to simplify the discussion of the existence and uniqueness of solutions of an involutive first-order ordinary differential equation, we specialise the Cartan normal form (3.10) to the case of only one independent variable. One speaks of a *semi-explicit form*, as it entails a split into algebraic and differential equations and the differential equations are solved for some derivatives. As there is no need to distinguish between derivatives of different class, we divide the vector $\mathbf{u}$ of the unknown functions into two parts $\mathbf{v}$ and $\mathbf{w}$ such that no derivatives $\mathbf{w}'$ appear. Now (3.10) takes the form

$$\mathbf{v}' = \boldsymbol{\phi}(x, \mathbf{v}, \mathbf{w}) \,, \tag{5.1a}$$

$$0 = \boldsymbol{\psi}(x, \mathbf{v}, \mathbf{w}) \,. \tag{5.1b}$$

Involution imposes conditions on the functions $\boldsymbol{\phi}$ and $\boldsymbol{\psi}$. The conditions (3.11) are empty in our case, as they stem from cross-derivatives which are not possible for ordinary differential equations. The conditions (3.12) become

$$\frac{\partial \boldsymbol{\psi}}{\partial \mathbf{w}} = \mathbf{A}\boldsymbol{\psi} \,, \qquad \frac{\partial \boldsymbol{\psi}}{\partial x} + \left(\frac{\partial \boldsymbol{\psi}}{\partial \mathbf{v}}\right)^t \boldsymbol{\phi} = B\boldsymbol{\psi} \,. \tag{5.2}$$

Here $\mathbf{A}(x, \mathbf{v}, \mathbf{w})$ is a vector of matrix valued functions and $B(x, \mathbf{v}, \mathbf{w})$ a matrix of appropriate sizes. If the first condition was violated, the prolongation of (5.1b) would lead to new differential equations. It implies that the unknown functions $\mathbf{w}$ may be considered as parameters. If the differential equation $\mathcal{R}_1$ is not underdetermined, the vector $\mathbf{w}$ is empty. The second condition ensures that the prolongation yields no new algebraic equations.

Proposition 5.1.1. *Assume that the system* (5.1) *is involutive and not underdetermined. Then the function* $\boldsymbol{\psi}$ *is a weak invariant of the normal ordinary differential equation* (5.1a)*.*

Proof. This is a trivial reformulation of (5.2). □

“Weak” invariant refers to the fact that the left hand sides of (5.2) vanish only on the constraint manifold $\mathcal{C} \subset \mathcal{E}$ defined by the equations $\boldsymbol{\psi} = 0$. For a strong invariant one requires that they vanish on the full manifold $\mathcal{E}$. Another way to formulate Proposition 5.1.1 is to say that $\mathcal{C}$ is an *invariant manifold* of the ordinary differential equation (5.1a), i. e. if we choose for (5.1a) initial data on $\mathcal{C}$, then the full solution will stay on $\mathcal{C}$.

Remark 5.1.2. The fact that the constraints yield generally only weak invariants has some implications for the numerical solution of differential algebraic equations. While $\mathcal{C}$ is an invariant manifold for any exact solution of (5.1a), this is not necessarily the case for approximate solutions. If we apply some standard numerical method to (5.1a), we cannot expect that the obtained approximation lies on $\mathcal{C}$.

This problem is known as *drift* off the constraint manifold. How severe it becomes, depends on the stability properties of $\mathcal{C}$, i. e. on the behaviour of (5.1a) in its neighbourhood. If $\mathcal{C}$ is attractive, numerical errors are automatically damped; if $\mathcal{C}$ is repulsive, they are amplified and the computations typically break down rapidly.

As we are always free to modify $\boldsymbol{\phi}$ by adding arbitrary linear combinations of $\boldsymbol{\psi}$, it is a natural thought to choose $\boldsymbol{\phi}$ such that $\mathcal{C}$ becomes attractive for (5.1a). This procedure is known as *stabilisation* [10]. Unfortunately, it is rather difficult to find a stabilisation for an arbitrary ordinary differential equation and it tends to make the evaluation of (5.1a) rather expensive. Thus in practice one prefers other approaches like projections on the constraint manifold or special methods for differential algebraic equations [25, 68]. $\lhd$

Proposition 5.1.1 makes the existence and uniqueness theory for involutive ordinary differential equations almost trivial. We can reduce the problem to a normal ordinary differential equation living on some submanifold $\mathcal{C} \subset \mathcal{E}$ and apply the standard existence and uniqueness theorem. For simplicity, we assume that $\mathcal{X} \subseteq \mathbb{R}$ which is not a very restrictive assumption, as any connected one-dimensional manifold is either a submanifold of $\mathbb{R}$ or $\mathbb{S}^1$.

Theorem 5.1.3. *Let $\mathcal{X} \subseteq \mathbb{R}$ and $[a, b] \subset \mathcal{X}$ a closed interval. Assume that $\boldsymbol{\phi}$ is continuous on $\pi^{-1}([a, b]) \cap \mathcal{C}$ and furthermore satisfies there a Lipschitz condition in $\mathbf{u} = (\mathbf{v}, \mathbf{w})$. Then the initial value problem for (5.1) with the initial condition $\mathbf{u}(a) = \mathbf{u}_0 \in \mathcal{C}$ possesses a unique solution $\mathbf{u}(x)$ on the whole interval $[a, b]$.*

Of course, one should note that we obtain such a trivial theory only because of our standard assumption that we are dealing with a *regular* differential equation. This entails that the ranks of certain Jacobians are constant and that we may indeed rewrite any local representation in the form (5.1). Many ordinary differential equations appearing in applications do not satisfy this assumption and their existence and uniqueness is harder. As we are mainly interested in partial differential equations, we do not give any details but refer to the literature, e. g. [144].

For a more geometric approach to an existence and uniqueness theory, we may exploit the contact structure of the jet bundle $J_1\mathcal{E}$. The contact distribution $\mathcal{C}_1 \subset TJ_1\mathcal{E}$ is $(m+1)$-dimensional and locally generated by the vector fields $\{\partial_{p_1^\alpha},\ p_1^\alpha\partial_{u^\alpha}+\partial_x\}$ (this is (1.10) specialised to $q = n = 1$). For a given ordinary differential equation $\mathcal{R}_1 \subseteq J_1\mathcal{E}$ we introduce on it the distribution $\mathcal{D} = T\mathcal{R}_1 \cap \mathcal{C}_1$. As for ordinary differential equations $\dim \mathcal{X} = 1$, we may consider a section $\sigma : \mathcal{X} \to \mathcal{E}$ as a curve.

Lemma 5.1.4. *The section $\sigma \in \Gamma_{\mathrm{loc}}(\mathcal{E})$ is a solution of $\mathcal{R}_1$, if and only if at every point x of its domain of definition the tangent vector of the prolonged section $j_1\sigma \in \Gamma_{\mathrm{loc}}(J_1\mathcal{E})$ lies in the vector space $\mathcal{D}_{j_1\sigma(x)}$.*

Proof. By the definition of a solution, the curve $j_1\sigma$ lies completely in $\mathcal{R}_1$ and thus the curve $(j_1\sigma)'$ in $T\mathcal{R}_1$. On the other hand, by the definition of the contact structure, the tangent space of any prolonged section $j_1\sigma$ is a subspace of the contact distribution $\mathcal{C}_1$. □

As a corollary of this lemma, we may easily characterise geometrically the existence of unique solutions for initial value problems. The differential equation $\mathcal{R}_1$ possesses a unique local solution passing through a prescribed point $\xi \in \mathcal{R}_0^{(1)} \subseteq \mathcal{E}$, if and only if the distribution $\mathcal{D}$ is one-dimensional everywhere on $\mathcal{R}_1$. Obviously, a necessary condition for unique solutions is that $\mathcal{R}_1$ is not underdetermined. In this case, the manifold $\mathcal{R}_1$ contains exactly one point ρ such that $\pi_0^1(\rho) = \xi$. A one-dimensional distribution is trivially involutive and thus integrable. Hence a unique integral curve to $\mathcal{D}$ exists passing through ρ. By Lemma 5.1.4 this integral curve is the prolongation of a solution of $\mathcal{R}_1$.

Theorem 5.1.5. *Let the regular ordinary differential equation $\mathcal{R}_1$ be involutive and not underdetermined. Then the distribution $\mathcal{D}$ is one-dimensional and the initial value problem for $\mathcal{R}_1$ is uniquely solvable.*

Proof. As the equation $\mathcal{R}_1$ is assumed to be regular and not underdetermined, we can locally transform it into the semi-explicit form $\mathbf{p} = \boldsymbol{\phi}(x, \mathbf{u})$, $\boldsymbol{\psi}(x, \mathbf{u}) = 0$ where $\mathbf{p}$ denotes the vector of all first-order derivatives. Any vector in $T\mathcal{R}_1$ must thus annihilate the following one-forms

$$dp_1^\alpha - \frac{\partial\phi^\alpha}{\partial u^\beta}du^\beta - \frac{\partial\phi^\alpha}{\partial x}dx\,, \qquad \frac{\partial\psi^\gamma}{\partial u^\beta}du^\beta + \frac{\partial\psi^\gamma}{\partial x}dx\,. \tag{5.3}$$

The distribution $\mathcal{D}$ is generated by those vectors $X \in \mathcal{C}_1$ that satisfy this condition. The ansatz $X = a^\alpha\partial_{p_1^\alpha} + b(p_1^\alpha\partial_{u^\alpha} + \partial_x)$ yields the following linear system for the coefficients a^α, b:

$$a^\alpha = b\left(\frac{\partial\phi^\alpha}{\partial u^\beta}p_1^\beta + \frac{\partial\phi^\alpha}{\partial x}\right)\,, \qquad 0 = b\left(\frac{\partial\psi^\gamma}{\partial u^\beta}p_1^\beta + \frac{\partial\psi^\gamma}{\partial x}\right)\,. \tag{5.4}$$

As these equations have to be evaluated *on* $\mathcal{R}_1$, the second set is empty for an involutive equation by (5.2). Thus the linear system (5.4) has a one-dimensional solution space. □

For notational simplicity, we have formulated this theorem and its derivation for first-order equations. But obviously, all the considerations extend straightforwardly to higher-order equations; there is not need to rewrite them as first-order equations, as it is usually done. For a regular, involutive and not underdetermined equation $\mathcal{R}_q \subset J_q\mathcal{E}$ we may always introduce the one-dimensional distribution $\mathcal{D} = T\mathcal{R}_q \cap \mathcal{C}_q$ the integral curves of which are the prolongations of solutions of $\mathcal{R}_q$.

5.2 The Cauchy–Kovalevskaya Theorem

With respect to the class of differential equations to which it can be applied, the Cauchy–Kovalevskaya Theorem represents probably the most general existence and unique result we know for the initial value problem for normal equations. Its only structural restriction is that we are dealing with a non-characteristic problem. However, the downside of this generality is that it applies only to *analytic* differential equations, *analytic* initial data and makes only statements about *analytic* solutions. These restrictions are very severe, in fact, too severe for most applications.

Nevertheless, the Cauchy–Kovalevskaya Theorem is the natural continuation of the work we have done so far. In Section 1.4 we discussed the construction of formal power series solutions. Now we show that under certain conditions these series converge. We formulate the theorem only for quasi-linear first-order equations. As detailed in Appendix A.3, any non-linear first-order differential equation can be reduced to a quasi-linear one via a prolongation and the introduction of additional dependent variables. One easily verifies that this operation maps a normal first-order equation again into a normal first-order equation.

It follows from the discussion of Proposition 3.3.7 that a normal quasi-linear first-order system may always be brought into the form

$$\mathbf{u}_{x^n} = \sum_{i=1}^{n-1} C_i(\mathbf{x},\mathbf{u})\,\mathbf{u}_{x^i} + \mathbf{c}(\mathbf{x},\mathbf{u}) \tag{5.5a}$$

where the C_i are $m \times m$ matrices with entries $c^\alpha_{i\beta}$ and $\mathbf{c}$ is an m-dimensional vector with components c_α. We consider the pure value problem

$$\mathbf{u}(x^1, \dots, x^{n-1}, 0) = 0 \,. \tag{5.5b}$$

The extension to general (analytic) initial data $\mathbf{f}$ is trivial; one only has to perform the coordinate transformation $\mathbf{v} = \mathbf{u} - \mathbf{f}$.

Note that the structure of (5.5a) implies that the hyperplane $x^n = 0$ on which the initial data are prescribed is non-characteristic. In physical applications the "evolution parameter" x^n represents usually the time t.

Theorem 5.2.1 (Cauchy–Kovalevskaya). *Let the functions C_i and $\mathbf{c}$ be real-analytic at the origin. The initial value problem* (5.5) *possesses a unique solution $\mathbf{u}(\mathbf{x})$ that is real-analytic at the origin.*

Proof. As (5.5a) is a normal equation, it is automatically formally integrable. Thus we can apply the order by order construction of a formal power series solution discussed in Section 1.4. Because of the special structure of (5.5) it is easy to distinguish the principal and the parametric coefficients in the series: any coefficient a^α_μ of the series for u^α with $\mu_n > 0$ is principal; all others are parametric.

The initial conditions (5.5b) determine uniquely all parametric coefficients of the series. It is not difficult to see that for our special initial data any Taylor coefficient a^α_μ can be represented as a polynomial with positive coefficients in the Taylor coefficients of the functions $c^\alpha_{i\beta}$ and c_α expanded at the origin.

Thus the initial value problem (5.5) possesses a unique formal power series solution. There remains to show that it converges in a neighbourhood of the origin. As the finitely many functions $c^\alpha_{i\beta}$ and c_α are analytic at the origin, constants M and r exist such that they are all contained in $\mathcal{C}_{M,r}(0)$ (see Appendix A.2). Consider the initial value problem

$$\mathbf{U}_{x^n} = \sum_{i=1}^{n-1} \bar{C}_i(\mathbf{x}, \mathbf{U})\, \mathbf{U}_{x^i} + \bar{\mathbf{c}}(\mathbf{x}, \mathbf{U}) \,, \tag{5.6a}$$

$$\mathbf{U}(x^1, \dots, x^{n-1}, 0) = 0 \tag{5.6b}$$

where all entries $\bar{c}^\alpha_{i\beta}$ of the matrices $\bar{C}_i$ and components $\bar{c}_\alpha$ of the vector $\bar{\mathbf{c}}$ are given by the same function

$$\phi(\mathbf{x}, \mathbf{U}) = \frac{Mr}{r - (x^1 + \cdots + x^n + U^1 + \cdots + U^m)} \tag{5.7}$$

and thus majorise the functions in (5.5a): $\bar{c}^\alpha_{i\beta} \succeq c^\alpha_{i\beta}$ and $\bar{c}_\alpha \succeq c_\alpha$.

The formal power series solution $\mathbf{U}$ of (5.6) majorises the formal power series solution $\mathbf{u}$ of (5.5a) (recall, the Taylor coefficients a^α_μ are determined by polynomials with *positive* coefficients). Hence our proof is completed, if we can show that the series $\mathbf{U}$ converges.

Entering (5.7) into (5.6) results in the differential equation

$$U_{x^n} = \frac{Mr}{r-(x^1+\cdots+x^n+U^1+\cdots+U^m)}\left(1+\sum_{i=1}^{n-1}\sum_{\alpha=1}^{m}U^\alpha_{x^i}\right) \tag{5.8}$$

which can be solved in closed form. With the help of the simple ansatz $U^\alpha(\mathbf{x}) = V(x^1+\cdots+x^{n-1}, x^n)$ for $1 \leq \alpha \leq m$ we can reduce (5.8) to a scalar equation for the single unknown $V(\sigma,\tau)$. It solution, computed for example by the method of characteristics, is given by

$$V(\sigma,\tau) = \frac{1}{nm}\left(r-\sigma-\sqrt{(r-\sigma)^2-2nmMr\tau}\right) . \tag{5.9}$$

This function is analytic at $\sigma = \tau = 0$ and (5.8) possesses a unique solution which is analytic at the origin. Thus the above constructed power series solution of (5.6) converges in a neighbourhood of the origin and one easily checks that it also satisfies the initial condition. □

It should be stressed that the uniqueness of the solution is asserted only in the category of functions that are real-analytic at the origin. The Cauchy–Kovalevskaya Theorem does not exclude the existence of further solutions that are not real-analytic. We will see in Section 5.5 that for linear systems a stronger uniqueness statement is possible.

Another important point is that Theorem 5.2.1 gives a purely local result; nothing is asserted about the domain of definition of the solution. A lower bound for its radius of convergence is given by the radius of convergence of the function V defined by (5.9). We do not give a precise formula, but obviously the bound depends only on the number of independent and dependent variables, respectively, n and m and on the numbers M, r determined by the coefficients of the equations.

The assumption of analyticity is crucial in the Cauchy–Kovalevskaya Theorem. In general, it is not possible to relax it, say to smooth functions, as the following classical example of Lewy [106] demonstrates (see also [85]). He proved that there exists a real-valued function $f \in C^\infty(\mathbb{R}^3)$ such that the following linear equation for a complex-valued unknown $u(x, y, z)$

$$u_x + iu_y - 2i(x+iy)u_z = f(x,y,z) \tag{5.10}$$

does not possess any solutions in $C^1(\Omega)$ for any open set $\Omega \subset \mathbb{R}^3$.

The way we have stated the Cauchy–Kovalevskaya Theorem (which is the usual way in textbooks), it gives *sufficient* conditions for the existence of a unique analytic solution. An obvious – but rarely posed – question is whether these are also *necessary*. Kovalevskaya gave in her thesis [95] an affirmative answer, however, without proof. The essential question is whether it is necessary to require that on the right hand side of (5.5a) the derivatives with respect to x^i for $i < n$ are also of first order.

All other conditions are obviously necessary. If the initial conditions are posed on a characteristic surface, it is not possible to construct a unique power series solution; this is the essence of the definition of characteristic. If x^n is non-characteristic and the first order system cannot be brought into the form (5.5a), it is either under- or overdetermined. In the first case we do not get a unique solution, in the latter one the initial data cannot be specified arbitrarily but must satisfy compatibility conditions. The necessity of the analyticity of the involved functions is trivial.

Example 5.2.2. Kovalevskaya [95, p. 22] provided a simple counter example to her theorem, if the right hand side of (5.5a) contains higher-order derivatives. Consider the following initial value problem for the heat equation:

$$u_t = u_{xx} \,, \tag{5.11a}$$

$$u(x,0) = 1/(1-x) \,. \tag{5.11b}$$

It is straightforward to construct the unique formal power series solution of this problem:

$$u(x,t) = \sum_{i,j=0}^{\infty} \frac{(2i)!}{i!j!} x^j t^i \,. \tag{5.12}$$

However, one can show that this series converges for no $t \neq 0$. Indeed, it is well-known that the solution of the pure initial value problem for the heat equation is given by convolution with the heat kernel

$$u(x,t) = \frac{1}{\sqrt{4\pi t}} \int_{-\infty}^{\infty} \exp\left[-\frac{(x-\xi)^2}{4t}\right] f(\xi)\, d\xi \tag{5.13}$$

for $u(x,0) = f(x)$. Obviously, the integral does not converge for arbitrary analytic initial data f, but we must impose on the function f some decay conditions for $|x| \to \infty$. $\triangleleft$

Of course, such a counter example does not suffice to prove the necessity of the conditions in the Cauchy–Kovalevskaya Theorem. A rigorous proof was achieved comparatively recently by Mizohata [124] for the case of linear systems. However, the techniques used are beyond the scope of this book.

5.3 Formally Well-Posed Initial Value Problems

In the previous section we discussed the initial value problem for normal differential equation. There it was obvious what initial conditions one should impose. For a non-normal equation this question becomes less trivial. In this section we consider the problem of choosing the right form and number of initial conditions on the level of formal power series, i.e. we are only considering formal power series solutions and the initial data take the form of formal power series, too.

Definition 5.3.1. *An initial value problem for a differential equation $\mathcal{R}_q$ is* formally well-posed, *if it possesses a unique formal power series solution for arbitrary formal power series as initial data.*

Thus in a formally well-posed initial value problem the Taylor coefficients of the initial data are in a one-to-one correspondence with the parametric derivatives of the differential equation. It is easy to see that the initial value problem considered in the Cauchy–Kovalevskaya Theorem 5.2.1 is formally well-posed. Here all derivatives that are only with respect to $x^1, \dots, x^{n-1}$ are parametric. As the initial data prescribed consist of functions depending on these variables, each such derivative is uniquely determined by the corresponding coefficient of the initial data.

Remark 5.3.2. The concept of well-posedness plays an important role in the theory of differential equations. According to the classical definition of Hadamard a problem is well-posed, if (i) it has a solution, (ii) the solution is unique and (iii) it depends continuously on the initial data. Of course, this is only a rough description that must be made rigorous by specifying function spaces and topologies. In Definition 5.3.1 we gave a precise formulation of the first two points by restricting to formal power series, but we completely omitted the third point.

This is a general problem with the Cauchy–Kovalevskaya Theorem. It permits the study of initial value problem that are not well-posed in the classical sense, as a famous example of Hadamard [66, p. 33] shows. Consider the following initial value problem for the Laplace equation:

$$u_{tt} + u_{xx} = 0 \,, \tag{5.14a}$$

$$u(x,0) = 0 \,, \tag{5.14b}$$

$$u_t(x,0) = \sin(nx)/n \,. \tag{5.14c}$$

Its (unique) solution is given by $u(x,t) = \sinh(nt)\sin(nx)/n^2$. If we consider the limit $n \to \infty$, we see that the initial data vanish but that for any $t > 0$ the solution blows up. As $u(x,t) \equiv 0$ is the unique solution for vanishing initial data, solutions of (5.14) cannot depend continuously on the initial data: arbitrarily small changes of the initial data may lead to arbitrarily large changes in the solution.

This is a typical effect in elliptic equations and the reason why one usually studies boundary value problems for such equations. We will ignore in the sequel the question of continuous dependency on the initial data and concentrate solely on the existence and uniqueness of (formal) solutions. $\triangleleft$

Our goal in this section is to determine in a systematic way formally well-posed initial value problems for involutive systems. We start with linear systems and for simplicity we restrict to the case of one dependent variable; in the general case one must simply perform the same analysis for each dependent variable separately. Taking up again the notations of Example 2.2.4, we study an ideal $\mathcal{I} \subseteq \mathcal{D}$ where $\mathcal{D} = \mathbb{F}[\partial_1, \dots, \partial_n]$ denotes the ring of linear differential operators with coefficients in some function field $\mathbb{F}$.

Example 5.3.3. Before we go into technical details, we demonstrate our basic strategy for the construction of formally well-posed initial value problems at a simple example in one dependent and two independent variables where the right initial conditions are easily found by inspection. Assume that we are given an involutive equation of the form

$$\mathcal{R}_2 : \begin{cases} u_{yy} = au + bu_x + cu_y \\ u_{xy} = du + eu_x + fu_y \end{cases} \tag{5.15}$$

where the coefficients are functions of x and y. Thus principal derivatives are u_{yy}, u_{xy} and all their derivatives. Conversely, parametric derivatives are u, u_y and all pure x-derivatives.

In a formally well-posed initial value problem all parametric coefficients must be prescribed by initial conditions. As all pure x-derivatives are parametric, it is obvious that the initial data must comprise a function of x. Thus we take as one initial condition $u(x, 0) = f(x)$. This determines all parametric coefficients in our system with the exception of the one corresponding to u_y. Hence we impose as second initial condition $u_y(0, 0) = g$. This time it is not possible to prescribe an arbitrary function but only a constant.

It is a typical feature of non-normal systems that even in the non-characteristic initial value problem one cannot prescribe all initial data on one hypersurface as in the Cauchy–Kovalevskaya Theorem. In the case of an underdetermined system the data would not suffice to determine all parametric coefficients, as some unknown functions are completely arbitrary. For an overdetermined system one could not prescribe arbitrary initial data, as they would have to satisfy some consistency conditions. $\triangleleft$

As a first step towards the construction of formally well-posed initial value problems, we need an algebraic approach to the distinction into principal and parametric derivatives. This is easily achieved with the help of a term order and leads to concepts very familiar from Chapter 2.

Definition 5.3.4. *Let $\mathcal{I} \subseteq \mathcal{D}$ be a (left) ideal. For a given term order $\prec$, we call the derivatives ∂^μ with $\mu \in \Delta_\prec(\mathcal{I}) = \operatorname{le}_\prec \mathcal{I}$* principal. *All remaining derivatives are* parametric *and their exponents form the* complementary set $\Pi_\prec(\mathcal{I}) = \mathbb{N}_0^n \setminus \Delta_\prec(\mathcal{I})$.

Comparing with the introduction of principal and parametric derivatives (or coefficients, respectively) in Section 1.4, we see that there at each order of prolongation the principal derivatives could be chosen completely independent of the choices at the previous orders. Now we select the principal derivatives in a systematic manner based on a term order.

While $\Delta_\prec(\mathcal{I})$ is a monoid ideal, the set $\Pi_\prec(\mathcal{I})$ does in general not possess any algebraic structure. Given an involutive division L and an involutive basis $\mathcal{H}$ of the ideal $\mathcal{I}$ with respect to L and $\prec$, we immediately obtain a disjoint decomposition of $\Delta_\prec(\mathcal{I})$ via the involutive cones of the multi indices in $\operatorname{le}_\prec \mathcal{H}$. The basic idea underlying the construction of formally well-posed initial value problems is to construct a similar disjoint decomposition of $\Pi_\prec(\mathcal{I})$. But this question was already successfully treated in Section 2.5 and we exploit the results for obtaining formally well-posed initial value problems.

Given an ideal $\mathcal{I} \subseteq \mathcal{D}$, we decompose the set $\Pi_\prec(\mathcal{I})$ in the form (2.25). We choose an arbitrary but fixed point $\mathbf{x}_0 \in \mathcal{X}$. For subsets $N \subseteq \{1, \dots, n\}$ we introduce two notations: $\mathbf{y} = N(\mathbf{x})$ where $y^i = x^i$, if $i \in N$, and $y^i = x_0^i$, if $i \notin N$ and $\mathbf{x}_N = (x^{i_1}, \dots, x^{i_k})$, if $N = \{i_1, \dots, i_k\}$. In other words, $N(\mathbf{x})$ denotes a projection of the vector $\mathbf{x}$ where those components of it that do not correspond to indices contained in N obtain the values of the entries of $\mathbf{x}_0$ and $\mathbf{x}_N$ simply collects those components of $\mathbf{x}$ that correspond to the elements of N. The decomposition (2.25) is defined by a set $\bar{\mathcal{N}} \subset \Pi_\prec(\mathcal{I})$ and associated sets $N_\nu \subseteq \{1, \dots, n\}$ for each $\nu \in \bar{\mathcal{N}}$. Based on these data, we prescribe in the neighbourhood of $\mathbf{x}_0$ the following initial conditions for $\mathcal{I}$:

$$D_\nu u\bigl(N_\nu(\mathbf{x})\bigr) = f_\nu(\mathbf{x}_{N_\nu}) \,, \qquad \forall \nu \in \bar{\mathcal{N}} \,. \tag{5.16}$$

Here, the right hand sides f_ν are arbitrary formal power series in their arguments. If $N_\nu = \emptyset$, then f_ν is simply an arbitrary constant.

Theorem 5.3.5. *The initial value problem for $\mathcal{I}$ defined by the initial conditions* (5.16) *is formally well-posed.*

Proof. We expand the solution of $\mathcal{I}$ into a formal power series about $\mathbf{x}_0$. By definition, its parametric coefficients are in a one-to-one correspondence with the elements of $\Pi_\prec(\mathcal{I})$. Let $\mu \in \Pi_\prec(\mathcal{I})$ be the multi index corresponding to a given parametric coefficient. Then exactly one multi index $\nu \in \bar{\mathcal{N}}$ exists such that $\mu = \nu + \rho$ with $\rho \in \mathbb{N}_{N_\nu}^n$. Our initial value problem prescribes for this coefficient the unique value $D_\rho f_\nu(\mathbf{x}_{0,N})$. This implies that we obtain for arbitrary initial data a unique formal power series solution and, by definition, our problem is formally well-posed. □

Thus we see that the construction of a formally well-posed initial value problem for an involutive system $\mathcal{F} \subset \mathcal{D}$ is in fact equivalent to the determination of a Stanley decomposition of the $\mathcal{D}$-module $\mathcal{D}/\langle\mathcal{F}\rangle$. Both operations reduce to the same problem, namely to find a decomposition of the form (2.25) for the complementary set of $\mathrm{le}_{\prec}\mathcal{F}$.

Example 5.3.6. Given a system in Cauchy–Kovalevskaya form, we must analyse for each dependent variable the monoid ideal $\langle\{[0,\dots,0,1]\}\rangle$. Its complement is obviously given by $[0,\dots,0] + \mathbb{N}^n_{\{1,\dots,n-1\}}$. Thus we are led to the initial value problem

$$u^\alpha(x^1,\dots,x^{n-1},x^n_0) = f^\alpha(x^1,\dots,x^{n-1})\ , \qquad 1 \leq \alpha \leq m \tag{5.17}$$

which is used in the Cauchy–Kovalevskaya Theorem.

As a simple example of a system not in Cauchy–Kovalevskaya form we consider the wave equation in characteristic coordinates $u_{xy} = 0$. The corresponding monoid ideal $\mathcal{J} = \langle\{[1,1]\}\rangle$ was already studied in Example 2.5.5. We obtain the classical characteristic initial value problem

$$u(x,y_0) = f(x)\ , \qquad u_y(x_0,y) = g(y)\ . \tag{5.18}$$

The second condition is often substituted by $u(x_0,y) = g(y)$, but then we must impose the consistency condition $f(0) = g(0)$ so that the initial value problem does not possess a solution for arbitrary initial data and thus is not formally well-posed. Our formulation allows us to choose g arbitrarily. ◁

As mentioned in Section 2.5, a decomposition (2.25) of the complementary set $\bar{\mathcal{J}}$ can be determined algorithmically. Concrete algorithms follow from the proofs of Proposition 2.5.3 or, in the important special case of Janet bases, of Proposition 2.5.4. A related algorithm for the construction of formally well-posed initial value problems was proposed by Reid [147].

Remark 5.3.7. In Section 4.1 we introduced closed form representations of general solutions in which manifolds $\mathcal{Y}_\ell$ appears. Our initial value problems take "planes" for these manifolds, namely, $\mathcal{Y}_\nu = \{\mathbf{x} \in \mathcal{X} \mid x^i = x^i_0\ \forall i \notin N_\nu\}$. If we construct them with the help of Pommaret bases, these submanifolds form a flag $\{\mathbf{x}_0\} = \mathcal{Y}_0 \subset \mathcal{Y}_1 \subset \cdots \subset \mathcal{Y}_n = \mathcal{X}$ where $\mathcal{Y}_k = \{\mathbf{x} \in \mathcal{X} \mid x^i = x^i_0\ \forall i > k\}$. For other decompositions this is not necessarily the case, as we have seen above for characteristic initial value problems. ◁

Finally, we study the extension of these results to systems with arbitrary non-linearities. The only available notion of involution for such systems is the intrinsic Definition 3.2.1 within our geometric approach. Thus we start with an involutive equation $\mathcal{R}_q \subseteq J_q\mathcal{E}$ and construct in local coordinates an initial value problem in the neighbourhood of a point $\rho \in \mathcal{R}_q$. We prepare again a local representation in Cartan normal form: each equation is solved for a different derivative which is the largest derivative in the equation with respect to the degree reverse lexicographic order.

Consider again the symbol module $\mathcal{M}^{\mathcal{P}}$; obviously only the equations of order q contribute to it. As we assume that $\mathcal{R}_q$ is involutive, these equations induce by Definition 3.1.1 of an involutive symbol a Pommaret basis of it and their leading derivatives yield a Pommaret basis of $\mathrm{le}_{\prec}\mathcal{M}^{\mathcal{P}}$. Let p_μ be the leading derivative of an equation whose order is less than q. As $\mathcal{R}_q$ is formally integrable, we may choose our local representation such that $p_{\mu+1_i}$ for any $1 \leq i \leq n$ is the leading derivative of some equation, too.

Let $\Delta_{\prec}(\mathcal{R}_q) \subseteq \mathbb{N}_0^n$ be the monoid ideal generated by the multi indices of the leading derivatives;[1] obviously, $\bigl(\Delta_{\prec}(\mathcal{R}_q)\bigr)_{\geq q} = \mathrm{le}_{\prec}\mathcal{M}^{\mathcal{P}}$. Thus our generating set is a weak Pommaret basis of $\Delta_{\prec}(\mathcal{R}_q)$ and, by Proposition 2.1.10, an involutive autoreduction yields a strong Pommaret basis. But now we are in the same situation as above for linear systems: we have a monoid ideal defined by the principal derivatives and must decompose its complement in the form (2.25). Then we impose the initial conditions (5.16) and show as in Theorem 5.3.5 that this yields a formally well-posed initial value problem in the neighbourhood of ρ.

As a corollary to these considerations, we may characterise all the involutive equations $\mathcal{R}_q$ the symbol module $\mathcal{M}^{\mathcal{P}}$ of which is Cohen–Macaulay. It turns out that this property is in a certain sense equivalent to normality, thus these equations have a particularly simple structure.

Proposition 5.3.8. *Let the symbol module $\mathcal{M}^{\mathcal{P}}$ of the involutive differential equation $\mathcal{R}_q$ be Cohen–Macaulay and let d be its depth. Then $\mathcal{R}_q$ is equivalent to a normal equation in d independent variables in the sense that a bijection between their formal solution spaces exists.*

Proof. Let us first consider the case that $d = 1$. Then $\mathcal{R}_q$ is an equation of finite type, as obviously $\mathcal{M}_q$ vanishes. In Example 1.4.3 we have shown that the integration of such an equation may be reduced to the integration of an involutive distribution which in turn may be reduced to solving a normal system of ordinary differential equations (Example 3.2.5).

If $d > 1$, then a local representation of $\mathcal{R}_q$ cannot contain any equation of order less than q, in other words $\mathcal{R}_{q-1}^{(1)} = J_{q-1}\mathcal{E}$. Indeed, the depth of $\mathcal{M}^{\mathcal{P}}$

[1] Despite our notation this monoid ideal is of course not intrinsically defined but depends on the chosen local representation.

corresponds to the lowest class present in the symbol. If a local representation contains lower-order equations, it must also contain all their formal derivatives, as $\mathcal{R}_q$ is formally integrable. Thus equations of class 1 appear in the symbol and its depth is 1.

Now our proposition is a direct consequence of (our proof of) Hironaka's criterion for Cohen–Macaulay modules (Corollary 2.5.12). It implies that a formally well-posed initial value problem for $\mathcal{R}_q$ can be defined by prescribing all derivatives with respect to $x^d, \dots, x^n$ up to order $q-1$ as functions of $x^1, \dots, x^{d-1}$. But these are precisely the initial data for a normal equation in d variables. □

At the end of the next section we will describe the equivalent normal equation and the bijection between the formal solution spaces in more detail. Note that if $\mathcal{R}_q$ is already a normal equation, then $\mathcal{M}^{\mathcal{P}}$ is even a free module.

Example 5.3.9. Assume that the involutive differential equation $\mathcal{R}_2$ in three independent and one dependent variables has a local representation of the following form:

$$\mathcal{R}_2 : \begin{cases} u_{zz} = \phi_1(x, y, z, u, u_x, u_y, u_z, u_{xx}, u_{xy}, u_{xz}) , \\ u_{yz} = \phi_2(x, y, z, u, u_x, u_y, u_z, u_{xx}, u_{xy}, u_{xz}) , \\ u_{yy} = \phi_3(x, y, z, u, u_x, u_y, u_z, u_{xx}, u_{xy}, u_{xz}) . \end{cases} \tag{5.19}$$

It is not difficult to see that its symbol module $\mathcal{M}^{\mathcal{P}}$ is Cohen–Macaulay and $\operatorname{depth} \mathcal{M}^{\mathcal{P}} = 2$. Applying the algorithm of the proof of Proposition 2.5.4 we find that a formally well-posed initial value problem is given by

$$u_z(x, 0, 0) = f(x) , \qquad u_y(x, 0, 0) = g(x) , \qquad u(x, 0, 0) = g(x) . \tag{5.20}$$

Obviously, the same initial data are prescribed for a normal system in two independent and three dependent variables. For the independent variables we may choose the $\mathcal{M}^{\mathcal{P}}$-regular sequence (x, y). ◁

5.4 The Cartan–Kähler Theorem

The Cartan–Kähler Theorem extends the Cauchy–Kovalevskaya Theorem to arbitrary involutive systems. This implies that it inherits all its restrictions. So again the practical use of this theorem is rather limited. However, the proof is quite enlightening. It does not only highlight the importance of involution in contrast to mere formal integrability; it also shows a fairly general methodology to extend results on normal systems to involutive ones.

For simplicity, we formulate it for an involutive first-order system in Cartan normal form without any algebraic equations. Thus we assume that we are given an involutive system of the following form (cf. (3.10)):

$$p_n^\alpha = \phi_n^\alpha(x^i, u^\beta, p_j^\gamma, p_n^\delta)\,, \qquad \begin{cases} 1 \le \alpha \le \beta_1^{(n)}\,, \\ 1 \le j < n\,, \\ \beta_1^{(n)} < \delta \le m\,, \end{cases} \tag{5.21a}$$

$$p_{n-1}^\alpha = \phi_{n-1}^\alpha(x^i, u^\beta, p_j^\gamma, p_{n-1}^\delta)\,, \qquad \begin{cases} 1 \le \alpha \le \beta_1^{(n-1)}\,, \\ 1 \le j < n-1\,, \\ \beta_1^{(n-1)} < \delta \le m\,, \end{cases} \tag{5.21b}$$

$$\vdots$$

$$p_1^\alpha = \phi_1^\alpha(x^i, u^\beta, p_1^\delta)\,, \qquad \begin{cases} 1 \le \alpha \le \beta_1^{(1)}\,, \\ \beta_1^{(1)} < \delta \le m\,. \end{cases} \tag{5.21c}$$

In order to obtain a unique solution, we prescribe the following initial conditions to the system (5.21)

$$u^\alpha(x^1, \dots, x^n) = f^\alpha(x^1, \dots, x^n)\,, \qquad \beta_1^{(n)} < \alpha \le m\,, \tag{5.22a}$$

$$u^\alpha(x^1, \dots, x^{n-1}, 0) = f^\alpha(x^1, \dots, x^{n-1})\,, \qquad \beta_1^{(n-1)} < \alpha \le \beta_1^{(n)}\,, \tag{5.22b}$$

$$\vdots$$

$$u^\alpha(x^1, 0, \dots, 0) = f^\alpha(x^1)\,, \qquad \beta_1^{(1)} < \alpha \le \beta_1^{(2)}\,, \tag{5.22c}$$

$$u^\alpha(0, \dots, 0) = f^\alpha\,, \qquad 1 \le \alpha \le \beta_1^{(1)}\,. \tag{5.22d}$$

This is a formally well-posed initial value problem. The leading derivatives of the equations in (5.21) define a Pommaret basis (hence by Corollary 2.4.9 also a Janet basis) of monoid module generated by their multi indices. The initial conditions (5.22) are obtained from (5.16), if we decompose this module according to the proof of Proposition 2.5.4.

Again one clearly sees that the system is underdetermined, if $\beta_1^{(n)} < m$, as in this case some of the dependent variables u^α are not restricted at all by the differential equations and must therefore be completely prescribed by the initial conditions. One might say that these variables have more the character of parameters than of unknowns.

Theorem 5.4.1 (Cartan–Kähler). *Let the functions ϕ_k^α and f^α be real-analytic at the origin and let the system* (5.21) *be involutive. Then it possesses one and only one solution that is analytic at the origin and satisfies the initial conditions* (5.22).

We first prove a lemma which is at the very heart of the proof of the Cartan–Kähler Theorem we will give. It is also of independent interest, as it shows an important property of involutive systems.

Lemma 5.4.2. *Under the assumptions of Theorem 5.4.1, consider the subsystem consisting of the equations of class* k *in* (5.21),

$$p_k^\alpha = \phi_k^\alpha(x^i, u^\beta, p_j^\gamma, p_k^\delta)\,, \qquad \begin{cases} 1\le\alpha\le\beta_1^{(k)}\,,\\ 1\le j\le k\,,\\ \beta_1^{(k)}<\delta\le m\,,\end{cases} \tag{5.23}$$

and set $x^i = 0$ *for* $i > k$ *in it. Let* $\mathbf{U}(x^1,\dots,x^k)$ *be an analytic solution of the thus restricted system* (5.23) *for the initial conditions*

$$\begin{aligned} u^\alpha(x^1,\dots,x^{k-1},0) &= f^\alpha(x^1,\dots,x^{k-1})\,, & 1\le\alpha\le\beta_1^{(k)}\,,\\ u^\alpha(x^1,\dots,x^{k-1},x^k) &= f^\alpha(x^1,\dots,x^{k-1},x^k,0,\dots,0)\,, & \beta_1^{(k)}<\alpha\le m\,. \end{aligned}$$

(Thus the functions u^α *with* $\alpha > \beta_1^{(k)}$ *appear only as parameters). If the initial data* $\mathbf{f}(\mathbf{x})$ *are such that for* $x^k = \cdots = x^n = 0$ *they satisfy all equations in* (5.21) *of class lower than* k*, then the solution* $\mathbf{U}(x^1,\dots,x^k)$ *satisfies these equations also for* $x^k \neq 0$.

Proof. We must analyse what happens, if we enter the solution $\mathbf{U}(x^1,\dots,x^k)$ into the equations of lower class. So we introduce the residuals stemming from the equations of class ℓ

$$\begin{aligned}\Delta_\ell^\alpha(x^1,\dots,x^k) =& \frac{\partial U^\alpha}{\partial x^\ell}(x^1,\dots,x^k) -\\ &\phi_\ell^\alpha\Big(x^1,\dots,x^k,0,\dots,0,\mathbf{U}(x^1,\dots,x^k),\frac{\partial\mathbf{U}}{\partial\mathbf{x}}(x^1,\dots,x^k)\Big)\end{aligned}$$

for all classes $\ell < k$ and all indices $1\le\alpha\le\beta_1^{(\ell)}$. We exploit the fact that the system (5.21) is involutive in order to derive differential equations for the functions Δ_ℓ^α. Adapting (3.11) to the special case we have got here, we see that involution of (5.21) is equivalent to the existence of functions $A_{ij}^\beta(\mathbf{x},\mathbf{u},\mathbf{p})$ and $B_i^\beta(\mathbf{x},\mathbf{u},\mathbf{p})$ such that whenever $1\le\ell<k\le n$

$$D_k(p_\ell^\alpha-\phi_\ell^\alpha) = \sum_{i=1}^{k}\sum_{\beta=1}^{\beta_1^{(i)}}\left\{\sum_{j=1}^{i} A_{ij}^\beta D_j(p_i^\beta-\phi_i^\beta) + B_i^\beta(p_i^\beta-\phi_i^\beta)\right\}. \tag{5.24}$$

Entering $\mathbf{U}(x^1,\dots,x^k)$ into these relations and setting $x^i = 0$ for $i > k$ implies that the residuals Δ_ℓ^α must satisfy the linear system

$$\frac{\partial \Delta_\ell^\alpha}{\partial x^k} = \sum_{i=1}^{k-1} \sum_{\beta=1}^{\beta_1^{(i)}} \left\{ \sum_{j=1}^{i} \bar{A}_{ij}^\beta \frac{\partial \Delta_i^\beta}{\partial x^j} + \bar{B}_i^\beta \Delta_i^\beta \right\}, \quad 1 \le \alpha \le \beta_1^{(k)}, \tag{5.25}$$

where

$$\bar{A}_{ij}^\beta(x^1,\dots,x^k) = A_{ij}^\beta\left(x^1,\dots,x^k,0,\dots,0,\mathbf{U}(x^1,\dots,x^k),\frac{\partial \mathbf{U}}{\partial \mathbf{x}}(x^1,\dots,x^k)\right),$$
$$\bar{B}_i^\beta(x^1,\dots,x^k) = B_i^\beta\left(x^1,\dots,x^k,0,\dots,0,\mathbf{U}(x^1,\dots,x^k),\frac{\partial \mathbf{U}}{\partial \mathbf{x}}(x^1,\dots,x^k)\right)$$

are analytic functions, as A_{ij}^β, B_i^β and $\mathbf{U}$ are all analytic and the analytic functions form a monoid under composition. On the right hand side of (5.25) the terms with $i = k$ have disappeared, as by assumption $\mathbf{U}$ is a solution of (5.23). This makes (5.25) an analytic *normal* system in the independent variables $x^1,\dots,x^k$. Obviously, $\Delta_\ell^\alpha \equiv 0$ is an analytic solution of it. By the Cauchy–Kovalevskaya Theorem 5.2.1 this solution is unique. Thus the solution $\mathbf{U}$ also satisfies all the equations of lower class. □

Lemma 5.4.2 expresses a kind of conservation property of involutive equations: if the equations of lower class are satisfied for $x^k = 0$, then they are satisfied at all values of x^k for which a solution exists. For the special case of Maxwell's equations (1.64), this is a very classical result. As already mentioned in Example 1.6.1, if a solution of the evolution part of Maxwell's equations satisfies the Gauss laws at the time $t = 0$, it satisfies them at all times. Using a classical identity from vector analysis, this is proven in almost any textbook on electrodynamics. But as we have seen, it is in fact a very general property of involutive systems.

We may also consider Lemma 5.4.2 as a generalisation of Proposition 5.1.1 to partial differential equations. It states that the subsystems of lower class are a kind of "weak invariants" for the "flow" generated by the subsystem of class k. In fact, this becomes a very useful point of view within the formalism of semigroups [134, 149] where partial differential equations are considered as ordinary differential equations on infinite-dimensional spaces.

Proof (of the Cartan–Kähler Theorem). We follow the proof given by Pommaret [139]. The solution is constructed step by step; at each step a normal system appears to which we apply the Cauchy–Kovalevskaya Theorem 5.2.1. Finally, we use repeatedly Lemma 5.4.2.

We start with the subsystem (5.21c) comprising all equations of class 1:

$$p_1^\alpha = \phi_1^\alpha(x^i, u^\beta, p_1^\gamma), \qquad \begin{cases} 1 \le \alpha \le \beta_1^{(1)}, \\ \beta_1^{(1)} < \gamma \le m. \end{cases} \tag{5.26a}$$

If we set $x^i = 0$ for $i > 1$, the initial conditions (5.22) allow us to substitute for u^β and p_1^β with $\beta > \beta_1^{(1)}$ the functions $\bar{f}^\beta(x^1) = f^\beta(x^1, 0, \dots, 0)$ and their x^1-derivatives. Now (5.26a) becomes a normal system (even of ordinary differential equations) in the independent variable x^1 and we may apply the Cauchy–Kovalevskaya Theorem for the last set of initial conditions in (5.22)

$$u^\alpha(0, \dots, 0) = f^\alpha \in \mathbb{R}\,, \qquad 1 \le \alpha \le \beta_1^{(1)}\,. \tag{5.26b}$$

It guarantees the existence of a unique analytic solution $U_1^\alpha(x^1)$ (where now $1 \le \alpha \le \beta_1^{(1)}$) of the initial value problem (5.26).

In the next step, we consider the equations of class 2 in (5.21)

$$p_2^\alpha = \phi_2^\alpha(x^i, u^\beta, p_1^\gamma, p_2^\delta)\,, \qquad \begin{cases} 1 \le \alpha \le \beta_1^{(2)}\,, \\ \beta_1^{(2)} < \delta \le m\,. \end{cases} \tag{5.27a}$$

This time we set $x^i = 0$ only for $i > 2$. Then we substitute u^β, p_1^β and p_2^β with $\beta > \beta_1^{(2)}$ by the functions $\bar{f}^\beta(x^1, x^2) = f^\beta(x^1, x^2, 0, \dots, 0)$ and their derivatives with respect to x^1 and x^2. This makes (5.27a) a normal system in the two independent variables x^1 and x^2. We prescribe the initial conditions

$$u^\alpha(x^1, 0) = \begin{cases} f^\alpha(x^1) & \text{for } \beta_1^{(1)} < \alpha \le \beta_1^{(2)}\,, \\ U_1^\alpha(x^1) & \text{for } 1 \le \alpha \le \beta_1^{(1)}\,. \end{cases} \tag{5.27b}$$

Thus for those components u^α we considered already in the first step, we take as initial data the solution $U_1^\alpha(x^1)$ obtained above. For the remaining components we use the initial data in (5.22).

We may again apply the Cauchy–Kovalevskaya Theorem, as all appearing functions are analytic. It guarantees us the existence of a unique analytic solution $U_2^\alpha(x^1, x^2)$ with $1 \le \alpha \le \beta_1^{(2)}$ of the initial value problem (5.27). However, it is not obvious that these functions also satisfy (5.26a) with $x^i = 0$ for $i > 2$; we only know this, if in addition $x^2 = 0$. But this is precisely the situation covered by Lemma 5.4.2: it asserts that the functions $U_2^\alpha(x^1, x^2)$ also satisfy the equations of lower class, i. e. in our case (5.26a), for all values of the variable x^2.

We iterate this process. In step k we consider the equations of class k, i. e. (5.23). Setting $x^i = 0$ for $i > k$, we substitute u^β and p_j^β with $\beta > \beta_1^{(k)}$ by the functions $\bar{f}^\beta(x^1, \dots, x^k) = f^\beta(x^1, \dots, x^k, 0, \dots, 0)$ and their derivatives. Then (5.23) becomes a normal system in the independent variables $x^1, \dots, x^k$. As initial conditions we use

$$u^\alpha(x^1, \dots, x^{k-1}, 0) = \begin{cases} f^\alpha(x^1, \dots, x^{k-1}) & \text{for } \beta_1^{(k-1)} < \alpha \le \beta_1^{(k)}\,, \\ U_{k-1}^\alpha(x^1, \dots, x^{k-1}) & \text{for } 1 \le \alpha \le \beta_1^{(k-1)}\,. \end{cases} \tag{5.28}$$

Here $U^\alpha_{k-1}(x^1,\dots,x^{k-1})$ represents the solution of the previous step. By the Cauchy–Kovalevskaya Theorem the initial value problem (5.23, 5.28) possesses a unique analytic solution $U^\alpha_k(x^1,\dots,x^k)$ with $1 \leq \alpha \leq \beta_1^{(k)}$ which by Lemma 5.4.2 also satisfies all equations of lower class.

The solution $U^\alpha_n(x^1,\dots,x^n)$ with $1 \leq \alpha \leq \beta_1^{(n)}$ obtained in step n is then the sought for analytic solution of the complete original initial value problem (5.21, 5.22), as the remaining components of $\mathbf{u}$ are given by the initial conditions. The uniqueness of the solution follows from the construction. □

One may wonder whether one really needs the assumption that the system (5.21) is involutive; perhaps formal integrability suffices? In fact, the step by step construction used in the proof above is possible for *any* system in triangular form. But it is unclear whether the thus constructed solution also satisfies all equations of lower class. Our proof of Lemma 5.4.2 depends crucially on the involution of the system and it does not hold for arbitrary non-involutive systems, as the following example demonstrates.

Example 5.4.3. We consider the first-order equation

$$\mathcal{R}_1 : \begin{cases} v_t = w_x , \\ w_t = 0 , \\ v_x = 0 . \end{cases} \tag{5.29}$$

It arises when we rewrite the second-order equation (3.9) as a first-order equation by setting $w = u_t$ and $v = u_x$ (compare Appendix A.3). Thus we already know that $\mathcal{R}_1$ is formally integrable but not involutive. Indeed, the non-multiplicative prolongation $D_t(v_x)$ cannot be reduced to zero and we obtain the obstruction to involution $w_{xx} = 0$.

The crucial step in the proof above applied to this example would be to show that any solution of the normal subsystem comprising the first two equation of (5.29) that satisfies the third equation for $t = 0$ also satisfies it for all values of t. If (5.29) was involutive, this would indeed be the case by Lemma 5.4.2. But as we are only dealing with a formally integrable equation, it is easy to construct a counter example.

Consider the functions

$$\begin{aligned} v(t,x) &= \alpha'(x)t + c , \\ w(t,x) &= \alpha(x) \end{aligned} \tag{5.30}$$

where $c \in \mathbb{R}$ is an arbitrary constant and $\alpha(x)$ an arbitrary real-analytic function. These functions solve the first two equations of (5.29). Furthermore, for $t = 0$, they also solve the last equation. But obviously this is not the case for any other value of t, if we choose $\alpha(x)$ such that $\alpha''(x) \neq 0$. Only for

$\alpha(x) = ax + b$ with constants $a, b \in \mathbb{R}$ we obtain a solution of $\mathcal{R}_1$, as only for this choice of α the equation $w_{xx} = 0$ is satisfied. $\triangleleft$

In fact, this example can easily be generalised to a proof that involution is not only sufficient for Lemma 5.4.2 to be applicable at each step of the proof of Theorem 5.4.1 but also necessary. Formal integrability is trivially necessary; thus we restrict to systems which are formally integrable but not involutive. Without loss of generality, we may further assume that in the Cartan normal form (5.21) the subsystem of class n is normal and that an obstruction to involution arises from the prolongation of an equation of lower class with respect to x^n, as this is precisely the situation in which we apply Lemma 5.4.2 in the proof of Theorem 5.4.1.

As the subsystem of class n is normal, we can eliminate any x^n-derivative in the obstruction to involution. Thus it may be considered as a differential equation in $x^1, \dots, x^{n-1}$ which, by construction, is independent of the equations of lower class present in our system. In our example above this was the equation $w_{xx} = 0$. According to the assumptions of Lemma 5.4.2, we choose initial data for $x^n = 0$ satisfying all equations of lower class. But we can always find initial data that do *not* satisfy our obstruction to involution. If we solve the subsystem of class n for such initial data, for $x^n \neq 0$ the solution will *not* satisfy that equation of lower class that generated the obstruction to involution, as its x^n-prolongation does not vanish.

Thus, if (5.21) is only formally integrable but not involutive, then it does not possess a solution for arbitrary choices of the initial data in (5.22). Hence involution is a necessary condition in Theorem 5.4.1, as otherwise our initial value problem is not even formally well-posed. Furthermore, these considerations show that it is a characteristic property of involutive systems that Lemma 5.4.2 holds for $1 \leq k \leq n$.

The Cartan–Kähler Theorem inherits all the restrictions and disadvantages of the Cauchy–Kovalevskaya Theorem, as its proof is based on the latter. For arbitrary differential equations there seems to be no way to avoid this, as no other existence and uniqueness theorem for arbitrary normal equations exists. But if more is known about the structure of the subsystems considered in the proof above, one might be able to use other existence and uniqueness theorems and to obtain stronger results still following the basic strategy of our proof.

If we want to go beyond analytic solutions, a crucial step will be the extension of Lemma 5.4.2. We used the Cauchy–Kovalevskaya Theorem in its proof. But in fact all we need is to prove the uniqueness of the zero solution of (5.25). Note that this is always a *linear* system, even if (5.21) is non-linear. Thus one may apply the Holmgren Theorem (see the next section) instead of the Cauchy–Kovalevskaya Theorem and obtains that the zero solution is unique even among all C^1 solutions. However, this still requires that we are dealing with analytic equations. In order to be able to relax that condition one needs more information about the structure of the system (5.21). A simple example will be given in Section 5.6.

Finally, we continue the discussion of involutive equations with a Cohen–Macaulay symbol module started at the end of the last section. Assume first that we are dealing with a first-order equation and denote again by d the depth of the symbol module. In our proof of the Cartan–Kähler Theorem we considered the subsystems of constant class. If the symbol module is Cohen–Macaulay, each of them becomes normal, as soon as we set $x^{k+1}, \dots, x^n$ to constant values where k is the class of the subsystem. No further substitutions are necessary, as all initial conditions comprise only functions of $x^1, \dots, x^{d-1}$.

In the first step of the proof we study the subsystem of class d, as no equations of lower class are present in the Cartan normal form. Its initial conditions are identical with the initial conditions for the full equation $\mathcal{R}_q$. Thus we may take it for the normal system referred to in Proposition 5.3.8. The bijection between the formal solution spaces of the full system and this normal system is implicitly contained in the proof of the Cartan–Kähler Theorem. We extend the solution $\mathbf{U}(x^1, \dots, x^d)$ of the subsystem to the solution $\mathbf{u}(x^1, \dots, x^n)$ of the full system following step by step the proof. At each step the initial conditions are just the solutions of the previous step; as the symbol module is Cohen–Macaulay, no additional conditions appear. In other words, in (5.28) no functions f^α appear, as $\beta_1^{(k-1)} = \beta_1^{(k)}$. Thus the extension is uniquely determined by the subsystems of higher class; their solutions define the bijection.

The discussion of higher-order equations is similar but somewhat messy. The basic idea is to rewrite as a first-order equation. But all arising equations of class less than d are imposed only as initial conditions. In order to avoid excessive notations, we simply continue Example 5.3.9 instead of providing the general theory.

Example 5.4.4. Consider the local representation (5.19) of the involutive differential equation $\mathcal{R}_2$. It is already in Cartan normal form and splits into two subsystems of class 2 and 3, respectively. The leading derivatives of the subsystem of class 2 are u_{yy} and u_{yz}. This suggests to consider it as a normal first-order system for u_y and u_z. But unfortunately, the right hand sides are not yet of the right form. Thus we introduce the new dependent variables $r = u_x$, $s = u_y$ and $t = u_z$ and augment our system by the equations $u_y = s$ and $r_y = s_x$. Substituting $z = 0$ yields the normal system

$$\begin{aligned} t_y &= \phi_2(x, y, 0, u, r, s, t, r_x, s_x, t_x)\;, \\ s_y &= \phi_3(x, y, 0, u, r, s, t, r_x, s_x, t_x)\;, \\ r_y &= s_x\;, \quad u_y = s\;. \end{aligned} \tag{5.31a}$$

We prescribe the following initial conditions:

$$t(x,0) = f(x)\;,\; s(x,0) = g(x)\;,\; r(x,0) = h'(x)\;,\; u(x,0) = h(x)\;. \tag{5.31b}$$

This is just a trivial reformulation of (5.20). The initial data for r and u are not independent but imply $r = u_x$ for $y = z = 0$. By Lemma 5.4.2 the solution of the initial value problem also satisfies this relation.

As in the proof of the Cartan–Kähler Theorem we continue with the equations of class 3. In our case there is only one with leading derivative u_{zz}. We rewrite again as a first-order system

$$\begin{aligned} t_z &= \phi_1(x, y, z, u, r, s, t, r_x, s_x, t_x)\;, \\ s_z &= t_y\;, \quad r_z = t_x\;, \quad u_z = t\;. \end{aligned} \tag{5.32a}$$

As initial conditions we take the solution (T, S, R, U) of (5.31):

$$\begin{aligned} t(x,y,0) &= T(x,y)\;, \quad & s(x,y,0) &= S(x,y)\;, \\ r(x,y,0) &= R(x,y)\;, \quad & u(x,y,0) &= U(x,y)\;. \end{aligned} \tag{5.32b}$$

Note that the initial conditions are completely fixed; all arbitrariness is contained in (5.31b). It is obvious that the u-component of the solution of the initial value problem (5.32) solves the original system (5.19). Conversely, any solution of (5.19) induces one of (5.32). Thus we conclude that a one-to-one correspondence exists between the (formal) solutions of (5.19) and of the normal system (5.31a). The bijection is provided by the definition of the variables r, s, t and by the solution of (5.32a).

Basically, we have just rewritten the original second-order equation $\mathcal{R}_2$ as a first-order equation according to the method outlined in Appendix A.3. The main difference is that the equation $u_x = r$ does not appear in any system we consider. Indeed, it is easy to check that if we straightforwardly transform the original equation to first order, then the arising equation will *not* have a Cohen–Macaulay symbol module. The reason is precisely the appearance of the equation $u_x = r$. We have hidden it in the initial conditions (5.31b) and Lemma 5.4.2 ensures that it holds for all values of (x, y, z). $\lhd$

5.5 The Holmgren Theorem

For linear systems it is possible to eliminate some of the restrictions of the Cauchy–Kovalevskaya and the Cartan–Kähler Theorem, respectively. Namely we can prove the uniqueness of continuous solutions of analytic equations. For normal systems, this is known as Holmgren's Theorem. Based on the proof of the Cartan–Kähler Theorem 5.4.1 presented in the last section, it can be easily extended to arbitrary involutive systems.

The basic idea is very simple; it is a bit more difficult to make it rigorous. We consider the following *normal* linear homogeneous system with analytic coefficients

$$A_i^{\alpha\beta}(\mathbf{x})p_i^\alpha + B^{\alpha\beta}(\mathbf{x})u^\alpha = 0\,, \qquad 1 \le \beta \le m \tag{5.33}$$

in a "lens-shaped" domain $\Omega \subset \mathbb{R}^n$ whose boundary consists of two non-characteristic hypersurfaces R and S (see Figure 5.1). We furthermore assume that S is analytic. The claim is that the initial value problem for (5.33) with $\mathbf{u}$ prescribed on R has at most one solution in $C^1(\overline{\Omega})$.

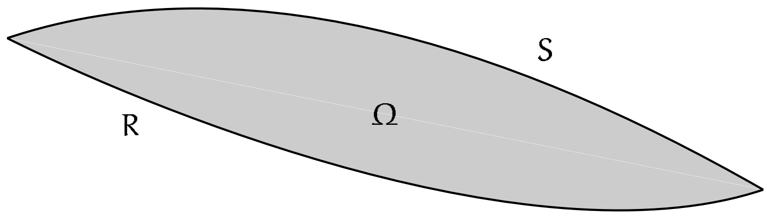

Fig. 5.1. A "lens-shaped" domain

An informal proof can be given as follow. We impose the initial conditions $\mathbf{u}|_R \equiv 0$. Then we add the equations in (5.33) with some yet undetermined functions $\bar{\mathbf{u}} \in C^1(\overline{\Omega})$ as coefficients and integrate over the domain Ω. A partial integration yields

$$\begin{aligned} 0 &= \int_\Omega \left(\bar{u}^\beta A_i^{\alpha\beta}p_i^\alpha + \bar{u}^\beta B^{\alpha\beta}u^\alpha\right)dx \\ &= \int_\Omega \left(\left[-\bar{p}_i^\beta A_i^{\alpha\beta}u^\alpha + \bar{u}^\beta \frac{\partial A_i^{\alpha\beta}}{\partial x^i}u^\alpha\right] + \bar{u}^\beta B^{\alpha\beta}u^\alpha\right)dx + \\ &\quad \int_{\partial\Omega} \bar{u}^\beta A_i^{\alpha\beta}u^\alpha n_i\, dA \end{aligned} \tag{5.34}$$

where $\mathbf{n}$ is the outer unit normal to the hypersurface $\partial\Omega = R \cup S$ and $\bar{p}_i^\beta$ is the x^i-derivative of $\bar{u}^\beta$.

Let us assume that $\bar{\mathbf{u}}$ is a solution of the adjoint system

$$-A_i^{\alpha\beta}\bar{p}_i^\beta + \left[\frac{\partial A_i^{\alpha\beta}}{\partial x^i} + B^{\alpha\beta}\right]\bar{u}^\beta = 0\,, \quad 1 \leq \alpha \leq m \tag{5.35}$$

satisfying the initial conditions $\bar{\mathbf{u}}\big|_S = \mathbf{f}$ for some arbitrary continuous function $\mathbf{f}$ defined on S. Then (5.34) yields

$$\int_S f^\beta (A_i^{\alpha\beta} n_i) u^\alpha \, dA = 0\,. \tag{5.36}$$

Since we assumed that S is an analytic non-characteristic hypersurface, the term $\sum_{i=1}^n A_i^{\alpha\beta} n_i$ vanishes nowhere on S. Thus (5.36) can only hold for *all* continuous functions $\mathbf{f}$, if $\mathbf{u}\big|_S \equiv 0$.

The problem is to show that we always find such functions $\bar{\mathbf{u}}$. As (5.33) is an analytic system, (5.35) is analytic, too. For analytic functions $\mathbf{f}$ we can thus apply the Cauchy–Kovalevskaya Theorem. By the Weierstraß Approximation Theorem any continuous function on a compact subset of $\mathbb{R}^n$ can be uniformly approximated by polynomials, so that it indeed suffices to treat analytic initial data, namely polynomials. However, the Cauchy–Kovalevskaya Theorem is a purely local result and we do not know whether the solution of the initial value problem for (5.35) exists in the whole domain Ω.

This forces us to proceed in the following rather technical fashion. We replace the single surface S by a one-parameter family S_λ that sweeps through the whole domain Ω which is now assumed to be of a somewhat special form. Let $\Omega' \subset \mathbb{R}^n$ be a bounded domain such that the coefficients of (5.33) are analytic in it. Let $\Psi(x)$ be an analytic function defined in Ω' such that $\nabla\Psi$ does not vanish anywhere. We set $Z = \{x \in \Omega' \mid x^n = 0\}$ and assume that it is an non-empty and non-characteristic hypersurface.

Now we introduce for $\lambda \in [a, b]$ the parametrised family of hypersurfaces $S_\lambda = \{x \in \Omega' \mid \Psi(x) = \lambda, x^n \geq 0\}$. We make the following assumptions (see Figure 5.2):

1. All $\lambda \in [a, b]$ are regular values of Ψ.
2. $\overline{\Omega} = \bigcup_{\lambda \in [a,b]} S_\lambda$.
3. $S_a \subset Z$ consists of a single point.
4. For each $\lambda \in (a, b]$ the hypersurfaces S_λ are non-characteristic and intersect Z transversally.

Theorem 5.5.1 (Holmgren). *Let* $\mathbf{u} \in C^1(\overline{\Omega})$ *be a solution of* (5.33) *such that* $\mathbf{u} \equiv 0$ *on* $Z \cap \partial\Omega$*. Then* $\mathbf{u} \equiv 0$ *in* $\overline{\Omega}$*.*

Proof. We set $\Lambda = \{\lambda \in [a, b] \mid \mathbf{u}\big|_{S_\lambda} \equiv 0\}$. The theorem is equivalent to $\Lambda = [a, b]$. By definition Λ is closed; we will show that it is also open in $[a, b]$, as then $\Lambda = [a, b]$.

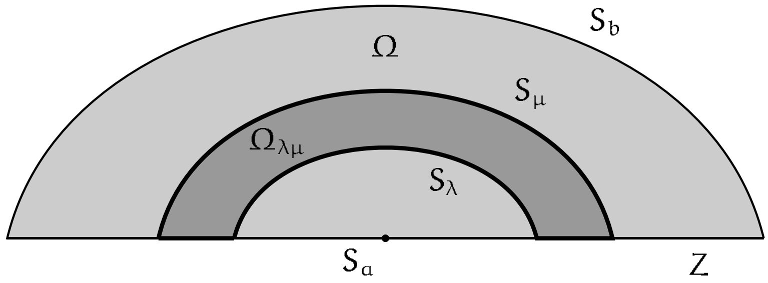

Fig. 5.2. A parametrised family of hypersurfaces

The functions $A_i^{\alpha\beta}$, $B^{\alpha\beta}$, Ψ are analytic and $\overline{\Omega}$ is compact; thus we can find values M, r independent of $\boldsymbol{x}$ such that all of them are in $\mathcal{C}_{M,r}(\boldsymbol{x})$ at each point $\boldsymbol{x} \in \overline{\Omega}$. If we solve the initial value problem for (5.33) with initial data in $\mathcal{C}_{M,r}$ prescribed on the hypersurface S_λ for $\lambda \in (a, b]$, by the Cauchy–Kovalevskaya Theorem the unique analytic solution will exist in an ϵ-neighbourhood of S_λ where ϵ does not depend on λ.

Polynomials are analytic functions and thus elements of some class $\mathcal{C}_{\bar{M},\bar{r}}$. As polynomials possess only a finite number of non-vanishing derivatives, we can choose an arbitrarily large value for $\bar{r}$ provided we take $\bar{M}$ large enough. As (5.33) is a linear system, we can multiply any solution by a constant and obtain again a solution; this does not change the domain of definition of the solution. By multiplying a polynomial in $\mathcal{C}_{\bar{M},r}$ with $M/\bar{M}$ we obtain a polynomial in $\mathcal{C}_{M,r}$. Thus we can conclude that for any polynomial prescribed as initial data on S_λ the unique analytic solution of (5.35) exist in the whole ϵ-neighbourhood of S_λ.

Let $\lambda \in [a, b]$ and $\epsilon > 0$ be given. In the neighbourhood of some point $\boldsymbol{y} \in S_\lambda$ we can apply the Implicit Function Theorem to solve the parametric equation $\Psi(\boldsymbol{x}) = \mu$ for some coordinate x^i. This yields a function $x^i = \psi(x^1, \dots, x^{i-1}, x^{i+1}, \dots, x^n; \mu)$ which is continuous in μ. So we can choose a $\delta(\boldsymbol{y}) > 0$ such that for all $\mu \in [a, b]$ with $|\lambda - \mu| < \delta(\boldsymbol{y})$ there exists a point $\boldsymbol{z} \in S_\mu$ with $\|\boldsymbol{z} - \boldsymbol{y}\| < \epsilon$. As S_λ is compact, there exists a constant δ such that $0 < \delta \leq \delta(\boldsymbol{y})$ for all $\boldsymbol{y} \in S_\lambda$. But this implies that for all $\mu \in [a, b]$ with $|\lambda - \mu| < \delta$ the hypersurface S_μ lies completely in an ϵ-neighbourhood of the hypersurface S_λ.

Now take a $\lambda \in \Lambda$ and a $\mu \in [a, b]$ such that $\mu \neq \lambda$ and $|\lambda - \mu| < \delta$. Then we can apply the argument above to the domain $\Omega_{\lambda\mu}$ bounded by the surfaces S_λ, S_μ and Z (recall that above we did not require that R is analytic, thus we may take S_λ plus parts of Z for the surface R). This time we can guarantee that the solution of the adjoint system (5.35) exists in the whole domain. Thus $\mathbf{u}|_{S_\mu} \equiv 0$ and $\mu \in \Lambda$. But this implies that Λ is open in $[a, b]$, as obviously $a \in \Lambda$ and thus Λ is not empty. □

This is not the strongest possible formulation of the Holmgren Theorem. Using, say, density arguments one can extend the uniqueness to more general spaces; Taylor [176] gives for example a version valid in Sobolev spaces. Note that we are still forced to require that the system is analytic, as we apply the Cauchy–Kovalevskaya Theorem at an intermediate step. But this is much less severe than the restriction to analytic solution which has now been removed.

An interesting aspect for us is that in the proof of Lemma 5.4.2 the decisive normal system (5.25) is always linear. Thus provided that its coefficients $\bar{A}^\beta_{ij}(\mathbf{x})$ and $\bar{B}^\beta_i(\mathbf{x})$ are analytic (for example, if A^β_{ij} and B^β_i in (5.24) do not depend on $\mathbf{u}$ and $\mathbf{p}$), we can prove Lemma 5.4.2 for much larger classes of solutions, as we can replace the use of the Cauchy–Kovalevskaya Theorem by the Holmgren Theorem.

Finally, we proceed to extend the Holmgren Theorem to arbitrary involutive systems. If the system (5.21) is linear, there exist analytic functions $A^{\alpha\delta}_k(\mathbf{x})$, $B^{\alpha\gamma}_{kj}(\mathbf{x})$, $C^{\alpha\beta}_k(\mathbf{x})$, $D^\alpha_k(\mathbf{x})$ such that

$$\begin{aligned}\phi^\alpha_k(\mathbf{x}, \mathbf{u}, p^\gamma_j, p^\delta_k) = &\sum_{\delta=\beta^{(k)}_1+1}^{m} A^{\alpha\delta}_k(\mathbf{x}) p^\delta_k + \sum_{\gamma=1}^{m}\sum_{j=1}^{k-1} B^{\alpha\gamma}_{kj}(\mathbf{x}) p^\gamma_j + \\ &+ \sum_{\beta=1}^{m} C^{\alpha\beta}_k(\mathbf{x}) u^\beta + D^\alpha_k(\mathbf{x})\,.\end{aligned} \tag{5.37}$$

Theorem 5.5.2. *The initial value problem* (5.21), (5.22) *with the right hand sides given by* (5.37) *possesses at most one solution in* $C^1(\overline{\Omega})$.

Proof. As for a normal system, it suffices to prove that in the case of a homogeneous system the only solution for vanishing initial data is the zero solution. Thus we study what happens in our proof of the Cartan–Kähler Theorem in this situation. In the first step we consider the normal system (5.26a) with vanishing initial data. Theorem 5.5.1 implies that the unique solution in $C^1(\overline{\Omega} \cap \{x^2 = \cdots = x^n = 0\})$ is $U^\alpha_1(x^1) \equiv 0$. Thus in the next step we have again vanishing initial data and can again apply Theorem 5.5.1. In this manner we can go through all steps and obtain finally that $\mathbf{u} \equiv 0$ is the unique solution in $C^1(\overline{\Omega})$. □

5.6 Applications: Weakly Overdetermined Systems

In the last section we only obtained a uniqueness result; in this section we want to prove the existence and uniqueness of *smooth* solutions for a special class of linear systems.

Definition 5.6.1. *An involutive differential equation $\mathcal{R}_q$ in n independent variables is called* weakly overdetermined, *if $\beta_q^{(n-1)} > 0$ and $\beta_q^{(k)} = 0$ for $k < n-1$, i. e. if any local representation in δ-regular coordinates contains only equations of class n and $n-1$.*

Note that such a weakly overdetermined equation might well be underdetermined in the sense of Definition 3.3.6. In the sequel, we are only interested in equations that can be interpreted in some sense as evolution equations. We slightly change our notation and denote the independent variables by $(x^1, \dots, x^n, t)$, i. e. we have $n+1$ variables and write x^{n+1} as t. We study now systems with variable coefficients of the following form:

$$E\mathbf{u}_t = A_i \mathbf{u}_{x_i} + B\mathbf{u} \,, \tag{5.38a}$$

$$0 = C_i \mathbf{u}_{x_i} + D\mathbf{u} \,. \tag{5.38b}$$

Here $\mathbf{u}$ is as usual the m-dimensional vector of dependent variables. The matrices $E(\mathbf{x}, t)$, $A_i(\mathbf{x}, t)$ and $B(\mathbf{x}, t)$ have q rows and m columns; the matrices $C_i(\mathbf{x}, t)$ and $D(\mathbf{x}, t)$ have p rows. We make three assumptions that are crucial for our analysis.

(a) The hypersurface $t = 0$ is non-characteristic.
(b) rank $E(\mathbf{x}, t) \equiv q$.
(c) For at least one i we have rank $C_i(\mathbf{x}, t) \equiv p$. Without loss of generality, we assume that this is the case for C_n.

All three assumptions together imply that we are dealing with a weakly overdetermined system in a δ-regular coordinate system. Note that (a) excludes parabolic systems. (b) implies that (5.38a) contains all equations of class $n+1$; (c) ensures that all equations in (5.38b) are of class n.

Example 5.6.2. An important example of a weakly overdetermined system are *Maxwell's equations*. It is well-known that their evolution part (1.64a), corresponding to (5.38a) form a hyperbolic system, whereas the constraint equations (1.64b), corresponding to (5.38b), are elliptic. $\triangleleft$

Our first task is to derive the conditions under which (5.38) represents an involutive system. This is a straightforward computation.

Lemma 5.6.3. *The system* (5.38) *is involutive, if and only if there exist* $p \times q$ *matrices* $M_i(\mathbf{x}, t)$, $N(\mathbf{x}, t)$ *and* $p \times p$ *matrices* $H_i(\mathbf{x}, t)$, $K(\mathbf{x}, t)$ *such that for all values* $1 \leq i, j \leq n$

$$M_i E = C_i \ , \tag{5.39a}$$

$$H_i C_j + H_j C_i = M_i A_j + M_j A_i \ , \tag{5.39b}$$

$$NE = D - M_k E_{x^k} \ , \tag{5.39c}$$

$$H_i D + K C_i + H_k C_{i,x^k} = M_i B + N A_i + M_k A_{i,x^k} + C_{i,t} \ , \tag{5.39d}$$

$$KD + H_k D_{x^k} = NB + M_k B_{x^k} + D_t \ . \tag{5.39e}$$

Proof. As the system is weakly overdetermined, non-multiplicative variables appear only in (5.38b), namely t is non-multiplicative for each equation. Thus we must try to express the t-prolongation of (5.38b) as linear combination of the original equations and their multiplicative prolongations. The matrices M_i, N, H_i, K are nothing but the coefficients of this linear combination. More precisely, a simple calculation shows that the formal equality

$$\partial_t(5.38b) = M_i \partial_{x^i}(5.38a) + H_i \partial_{x^i}(5.38b) + N(5.38a) + K(5.38b) \tag{5.40}$$

holds, if and only if the matrices satisfy the above conditions. □

The first two conditions are necessary for an involutive symbol; the remaining three conditions ensure the formal integrability of the system. Because of the rank assumptions (b) and (c), it is not difficult to see that if such matrices M_i, N, H_i, K exists, they are uniquely determined by (5.39). For later use, we derive the compatibility conditions of the linear system (5.38) under the assumption that it is involutive. For this purpose, we add on the right hand side of (5.38a) a "perturbation" $\boldsymbol{\delta}$ and on the right hand side of (5.38b) a "perturbation" $-\boldsymbol{\epsilon}$.

Lemma 5.6.4. *The perturbed system possesses a formal solution, if and only if the perturbations satisfy the compatibility condition*

$$\boldsymbol{\epsilon}_t - H_i \boldsymbol{\epsilon}_{x_i} - K\boldsymbol{\epsilon} = M_i \boldsymbol{\delta}_{x_i} + N\boldsymbol{\delta} \ . \tag{5.41}$$

Proof. (5.41) follows immediately from (5.40). □

In order to obtain an existence and uniqueness theorem for smooth solutions, we need to make some further assumptions on the structure of (5.38). First of all, we want to avoid underdetermined situations. So we assume that $q = m$. This implies together with Assumption (b) that E is a regular matrix and without loss of generality, we take $E(\mathbf{x}, t) = \mathbb{1}_m$, the $m \times m$ identity matrix. Then it follows from (5.39a) that $M_i = C_i$.

Recall the following classical definitions. (5.38a) with $E = \mathbb{1}_m$ is called *hyperbolic* in t-direction, if for any vector $\xi = (\xi^1, \dots, \xi^n) \in \mathbb{R}^n$ the matrix $A_\xi(\mathbf{x}, t) = \xi^i A_i(\mathbf{x}, t)$ has at any point $(\mathbf{x}, t)$ only real eigenvalues and an eigenbasis. It is *strongly hyperbolic*, if there exists for any $A_\xi(\mathbf{x}, t)$ a symmetric, positive definite matrix $P_\xi(\mathbf{x}, t)$, a *symmetriser*, depending smoothly on $\mathbf{x}$, t and ξ such that $P_\xi A_\xi - A_\xi^t P_\xi = 0$. (5.38b) is *elliptic*, if the matrix $C_\xi(\mathbf{x}, t) = \xi^i C_i(\mathbf{x}, t)$ defines for all vectors $0 \neq \xi \in \mathbb{R}^n$ a surjective mapping. This condition implies that *all* matrices C_i must possess everywhere maximal rank: $\operatorname{rank} C_i(\mathbf{x}, t) \equiv p$.

A classical result in the theory of strongly hyperbolic systems is the following existence and uniqueness theorem (cf. [98, Theorem 6.2.2]) which will form the basis of our existence and uniqueness theorem for (5.38). It provides in addition an estimate on the norm of the solution. In the case of a symmetric hyperbolic problem, i.e. $A = A^t$ and we do not need a symmetriser, we take the usual Sobolev norms $\|\mathbf{u}(\mathbf{x})\|_{H^p}$ defined as square root of the sum of the squares of the L^2 norm of $\mathbf{u}(\mathbf{x})$ and all its derivatives up to order p (note we consider here only functions of the spatial variables $\mathbf{x}$).

In the general case we need a kind of weighted Sobolev norm where the "weight" consists of a pseudodifferential operator constructed out of the symmetriser $P_\xi(\mathbf{x}, t)$. More precisely, we introduce the operator

$$\hat{P}(t)v(\mathbf{x}) = \int P_{\xi/|\xi|}(\mathbf{x}, t) e^{2\pi i \xi \cdot \mathbf{x}} \hat{v}(\xi)\, d\xi \tag{5.42}$$

acting on functions depending only on the spatial variables. Here $\hat{v}$ is as usually the Fourier transform of v. Then we define the inner product

$$\big(\mathbf{v}(\mathbf{x}), \mathbf{w}(\mathbf{x})\big)_{P(t)} = \int \mathbf{v}(\mathbf{x}) \cdot \big(P(t)\mathbf{w}(\mathbf{x})\big)\, d\mathbf{x}\,. \tag{5.43}$$

The norm $\|\mathbf{u}(\mathbf{x})\|_{P(t),H^p}$ is then a Sobolev norm where the underlying L^2 norm has been substituted by the norm induced by the inner product (5.43). There are some subtleties involved in the definition of this norm like the domain over which the integrals are taken; in general, one needs a partition of unity on this domain etc. We refer to [98] and references therein.

Theorem 5.6.5. *Let the system*

$$\mathbf{u}_t = A_i(\mathbf{x}, t)\mathbf{u}_{x_i} + B(\mathbf{x}, t)\mathbf{u} + \mathbf{F}(\mathbf{x}, t) \tag{5.44}$$

have smooth coefficients A_i, B, $\mathbf{F}$ and be strongly hyperbolic with symmetriser P_ξ. Then it possesses a unique smooth solution satisfying periodic boundary conditions and the smooth initial conditions

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{f}(\mathbf{x})\,. \tag{5.45}$$

This solution satisfies at any time $t \in [0, T]$ *the following estimate*

$$\|\mathbf{u}(\cdot, t)\|_{P(t),H^p} \leq K_p \left[\|\mathbf{f}\|_{P(t),H^p} + \int_0^t \|\mathbf{F}(\cdot, \tau)\|_{P(t),H^p}\, d\tau\right] \tag{5.46}$$

where the constant K_p *depends only on the coefficients of the system, their derivatives up to order* p *and on* T.

Our goal is to extend this theorem to weakly overdetermined systems where (5.38a) is strongly hyperbolic and (5.38b) is elliptic. If we could apply Lemma 5.4.2, this would be straightforward. However, we are not dealing with analytic equations. Thus we must show that the lemma still holds in our more general situation. This can be done with the help of the compatibility condition (5.41) inheriting some properties of (5.38a).

Lemma 5.6.6. *Let* (5.38a) *with* $E = \mathbb{1}_m$ *be hyperbolic in* t*-direction and* (5.38b) *elliptic. Then the compatibility condition* (5.41) *— considered as a system for* $\boldsymbol{\epsilon}$ *only — is hyperbolic in* t*-direction, too.*

Proof. As we assume that the system (5.38) is involutive, the matrices A_i, $M_i = C_i$, H_i satisfy the relations (5.39b). This implies (with obvious notation) that $H_\xi C_\xi = C_\xi A_\xi$ for any vector $\xi \in \mathbb{R}^n$. Let $\mathbf{v}$ be an eigenvector of A_ξ for the (real) eigenvalue λ. As $H_\xi C_\xi \mathbf{v} = C_\xi A_\xi \mathbf{v} = \lambda C_\xi \mathbf{v}$, we see that $C_\xi \mathbf{v}$ is an eigenvector of H_ξ for the eigenvalue λ.

According to our assumptions, all matrices A_ξ possess only real eigenvalues and an eigenbasis and all matrices C_ξ define surjective mappings. Thus there exists an eigenbasis of H_ξ consisting of vectors of the form $C_\xi \mathbf{v}$ where $\mathbf{v}$ is an eigenvector of A_ξ. This implies furthermore that all eigenvalues of H_ξ are real, as they are also eigenvalues of A_ξ. □

Lemma 5.6.7. *Let the assumptions of Lemma 5.6.6 be satisfied and let furthermore* (5.38a) *be strongly hyperbolic with symmetriser* $P_\xi(\mathbf{x}, t)$. *Then the compatibility condition* (5.41) *is strongly hyperbolic, too. As symmetriser we may use the matrix* $Q_\xi(\mathbf{x}, t)$ *defined by*

$$Q_\xi^{-1} = C_\xi P_\xi^{-1} C_\xi^t\,. \tag{5.47}$$

Proof. We must show that Q_ξ satisfies the three properties of a symmetriser. Obviously, it inherits the symmetry of P_ξ. For the positive definiteness we note that $\mathbf{v}^t Q_\xi^{-1} \mathbf{v} = (C_\xi^t \mathbf{v})^t P_\xi^{-1} (C_\xi^t \mathbf{v})$. As C_ξ defines a surjective mapping, its transpose C_ξ^t defines an injective mapping. Thus the expression above vanishes only for $\mathbf{v} = 0$. This also implies that Q_ξ^{-1} is indeed invertible. Finally, using again (5.39b)

$$\begin{aligned} H_\xi Q_\xi^{-1} - Q_\xi^{-1} H_\xi^t &= H_\xi C_\xi P_\xi^{-1} C_\xi^t - C_\xi P_\xi^{-1} C_\xi^t H_\xi^t \\ &= C_\xi (A_\xi P_\xi^{-1} - P_\xi^{-1} A_\xi^t) C_\xi^t . \end{aligned} \tag{5.48}$$

But this last expression vanishes, as P_ξ is a symmetriser for A_ξ. □

Based on these technical lemmata, it is straightforward to extend Theorem 5.6.5 to weakly overdetermined equations. The decisive point is that this theorem implies a uniqueness result for the compatibility condition (5.41).

Theorem 5.6.8. *Let* (5.38a) *with* $E = \mathbb{1}_m$ *be strongly hyperbolic and* (5.38b) *elliptic. If the initial data in* (5.45) *satisfies for* $t = 0$ *the equations* (5.38b)*, then the initial value problem* (5.38, 5.45) *with periodic boundary conditions possesses a unique smooth solution.*

Proof. Let us first ignore the constraints (5.38b). Then Theorem 5.6.5 guarantees the existence and uniqueness of a smooth solution. Entering this solution into the equations (5.38b) yields residuals $\boldsymbol{\epsilon}$. These residuals must satisfy the compatibility condition (5.41) with $\boldsymbol{\delta} = 0$. As by assumption the initial data satisfy (5.38b) at $t = 0$, our initial conditions for (5.41) are $\boldsymbol{\epsilon}(\mathbf{x}, 0) = 0$. Obviously, $\boldsymbol{\epsilon} \equiv 0$ is one solution of this initial value problem. According to Lemma 5.6.7, we are dealing with a strongly hyperbolic system, so that we can again apply Theorem 5.6.5. It guarantees that the zero solution is in fact the only solution. Hence there exists a unique smooth solution of the overdetermined problem (5.38, 5.45). □

Our method of proving this result has as side effect some implications for the numerical solution of a weakly overdetermined systems. A simple approach consists of considering the equations of lower class (5.38b) only as constraints on the initial data and otherwise ignoring them, i. e. one simply solves the initial value problem for (5.38a) with initial data satisfying (5.38b).

For exact solutions this approach is fine, as we have seen above. For numerical solutions the situation is somewhat different. In general, we must expect a drift off the constraints, i. e. after some time the numerical solution ceases to satisfy the equations of lower class (5.38b) (we noted this problem already for ordinary differential equations in Remark 5.1.2). For us this is easy to understand. A numerical solution is only an approximation of an exact solution, hence there appears a non-vanishing residual $\boldsymbol{\delta}$ leading to a non-vanishing right hand side in the compatibility condition (5.41) which determines the evolution of the residual $\boldsymbol{\epsilon}$ in the equations (5.38b). So the size of the drift can be estimated by (5.46):

$$\|\boldsymbol{\epsilon}(\cdot, t)\|_{P(t), H^p} \leq K_p \int_0^t \|M_i \boldsymbol{\delta}_{x_i}(\cdot, \tau) + N\boldsymbol{\delta}(\cdot, \tau)\|_{P(t), H^p} \, d\tau \tag{5.49}$$

The problem is aggravated by the fact that the right hand side of the estimate depends not only on the residuals $\boldsymbol{\delta}$ but also on their spatial derivatives. While any reasonable numerical method takes care of controlling the size of $\boldsymbol{\delta}$, it is hardly possible to control the size of the derivatives.

Example 5.6.9. The approach outlined above is fairly popular for numerically solving Maxwell's equations. Here the compatibility conditions are the continuity equations (1.65). It follows from their derivation that for Maxwell's equations the estimate (5.46) depends only on the divergence of $\boldsymbol{\delta}$ and not on $\boldsymbol{\delta}$ itself. Thus a good numerical method for Maxwell's equations should be constructed such that the divergence of the residual vanishes. ◁

5.7 Notes

The geometric approach to the initial value problem for ordinary differential equations presented in Section 5.1 is not only of theoretical interest. Tuomela and Arponen [183] use it for the numerical integration of overdetermined equations. One motivation is that the prolonged curve $j_1\sigma$ in $J_1\mathcal{E}$ is often smoother than its projection σ in $\mathcal{E}$. In fact, one can show that many apparent singularities disappear this way [182]. While our proof of Theorem 5.1.5 was based on a local representation in semi-explicit form, this is not necessary for computations. Even if a fully non-linear representation is given, the distribution $\mathcal{D}$ always arises as the solution of a linear system. The only advantage of the semi-linear form is that it yields this system in the special form (5.4); but for numerical computations this is not relevant.

A similar geometric theory exists for partial differential equations and yields an alternative approach to the Cartan–Kähler Theorem. It is known under the name *Vessiot theory*, as it was introduced by Vessiot [189] as a dual form of the Cartan–Kähler theory based on differential forms. A formulation in modern language was given by Stormark [171]; for the special case of hyperbolic equations in the plane see also [187]. The theory becomes considerably more involved for partial differential equations, as the distribution $\mathcal{D}$ itself does not directly yield solutions. One has to search for transverse involutive subdistributions of dimension n.

The proof of the Cauchy–Kovalevskaya Theorem we have used is the one found in most textbooks (see e. g. [85, 149]) and essentially goes back to Goursat [63]. Cauchy gave a first proof in 1842 in a series of notes to the French academy of science [35–37]; later Kovalevskaya studied the problem in her thesis [95]. Note that Cauchy's method of majorants proves the convergence of the formal power series solutions indirectly by comparison with a known series. A direct convergence proof turns out to be surprisingly tricky; Shinbrot and Welland [166] provided a relatively simple one.

Several alternative proofs exist which do not use power series but functional analytic methods. The basic idea is to convert the initial value problem into a fixed point equation in some Banach space; a nice example for this technique can be found in [190]. These approaches have the advantage that they can also handle more general situations leading to "abstract" Cauchy–Kovalevskaya Theorems.

The Janet–Riquier Theory (see the Notes of Chapter 2) also provides a generalisation of the Cauchy–Kovalevskaya Theorem: *Riquier's Theorem*. It considers a formally well-posed initial value problem constructed with the help of an involutive basis. In contrast to our proof of the Cartan–Kähler Theorem, one tries to prove directly the convergence of formal power series. However, so far one could show the convergence only for degree compatible term orders. Thus we meet again the condition that each equation should be solved for a derivative of maximal order. We noted this already for the Cauchy–Kovalevskaya Theorem where it was even a necessary condition. Whether it is also necessary for Riquier's Theorem seems to be an open question.

Our proof of Holmgren's Theorem follows [149]. The basic trick using the adjoint system goes back to Holmgren [79]. Its proper global formulation with the one-parameter family S_λ of analytic surfaces is due to John [85]. The extension to arbitrary involutive systems seems to be new despite its simplicity; it was first presented in [159].

The notion of a weakly overdetermined system was introduced in [165]. That article studies only systems with constant coefficients where the situation is simpler. In this case one does not need the assumption that the constraints are elliptic, as one may simply apply the Holmgren Theorem for proving that the equations of lower class are satisfied.

6. Completion to Involution

I wish to God these calculations had been executed by steam.

Charles Babbage

We have seen in the previous chapters that involutive equations possess many attractive features. So the question naturally arises what one should do, if one faces a non-involutive equation. The answer is that one should *complete* it to an equivalent involutive one. We will show that this is, in principle, always possible and discuss algorithms to perform this task. However, algorithms in the strict sense of computer science can be obtained only for differential equations with at most polynomial non-linearities.

Even in the cases where a completion algorithm exists, it tends to be rather expensive and tedious. This is not surprising, if one recalls the one-to-one correspondence between linear partial differential equations in one dependent variable with constant coefficients and polynomials (we simply identify the derivative $\partial^{|\mu|}u/\partial x^{\mu}$ with the term x^{μ}). Completing a system of such equations to involution is equivalent to determining a Gröbner basis for the corresponding polynomial ideal. But this operation is known to possess a doubly exponential worst case complexity (see Appendix B.2). As a consequence, computer algebra is an important tool for concrete computations with general systems of differential equations. Most of our calculations were performed with *MuPAD*.[1] This computer algebra system contains implementations of many of the algorithms discussed here [16].

We start with the fairly simple case of ordinary differential equations. Here completion takes the very intuitive geometric form of imposing tangency conditions. Because of its simplicity it has been discovered and rediscovered

[1] See the web site `www.mupad.de` for further information.

many times in different fields. For partial differential equations, completion becomes considerably more involved. We describe in Section 6.2 a general geometric method as a constructive proof of the Cartan–Kuranishi Theorem.

Rendering the Cartan–Kuranishi completion algorithmic requires some work. We first go back to the algebraic theory introduced in Chapter 2 and develop in Section 6.4 an algorithm for determining involutive bases. Unfortunately, the termination of this algorithm (which is equivalent to the existence of a basis) is a non-trivial question and requires the introduction of some technicalities which are the topic of the preceeding section.

Section 6.5 combines for linear differential equations the geometric and the algebraic approach into a hybrid algorithm providing a concrete realisation of the Cartan–Kuranishi method (and thus yielding intrinsic results) based on the more efficient algebraic algorithms.

Finally, we discuss the application of the completion theory in the numerical analysis of general systems of differential equations. In Section 6.6, we show how some of the index concepts proposed for differential algebraic equations can be intrinsically formulated within the formal theory. This leads to definitions that can be directly applied to partial differential equations, too. Section 6.7 studies the relation between the index of a partial differential equation and the differential algebraic equations obtained from it by a semi-discretisation. Somewhat surprisingly, it turns out that obstructions to involution become integrability conditions so that again involution and not mere formal integrability is relevant.

6.1 Ordinary Differential Equations

We first restrict our attention to *autonomous* ordinary differential equations where the independent variable x does not occur explicitly. This is only for the simplicity of the presentation and to make the theory more intuitive. The general case will be treated at the end of the section.

Autonomous systems are traditionally described within the tangent bundle (see Appendix C.1). So let $\mathcal{U}$ be an m-dimensional manifold. The right hand side of an ordinary differential equations in the solved form

$$\mathbf{u}' = \boldsymbol{\phi}(\mathbf{u}) \tag{6.1}$$

is interpreted as a vector field $X \in \mathfrak{X}(\mathcal{U})$. Systems of the more general form

$$\boldsymbol{\Phi}\left(\mathbf{u}, \mathbf{u}'\right) = 0\ , \tag{6.2}$$

where the number of equations may differ from $\dim \mathcal{U}$, appear in many applications and are known in the literature under different names. A classical one is *implicit equation*; in numerical analysis the name *differential algebraic equation* has become popular; in engineering applications one often speaks of a *descriptor form* in contrast to the *state space form* (6.1). Geometrically, such a system corresponds (under some regularity assumptions) to a submanifold $\mathcal{S}^{(0)} \subseteq T\mathcal{U}$.

Assuming that $\mathbf{\Phi}$ is a smooth function and that the rank of the Jacobian $\partial\mathbf{\Phi}/\partial\mathbf{u}'$ is constant, we transform (6.2) into *semi-explicit form*: we split the vector $\mathbf{u}$ into two parts $\mathbf{v}$ and $\mathbf{w}$ (in general of different dimensions) such that the derivatives $\mathbf{w}'$ do not appear in (6.2); by the Implicit Function Theorem we can solve for the derivative $\mathbf{v}'$ and obtain

$$\mathbf{v}' = \boldsymbol{\phi}(\mathbf{v}, \mathbf{w}) \,, \tag{6.3a}$$

$$0 = \boldsymbol{\psi}(\mathbf{v}, \mathbf{w}) \,. \tag{6.3b}$$

This also explains the name *differential algebraic equation*; in semi-explicit form such a system comprises a mixture of differential and algebraic equations. The algebraic equations are often called *constraints*. Note that so far we do not assume that we are dealing with an involutive equation. Hence in contrast to Section 5.1, we may *not* call (6.3) Cartan normal form; the necessary conditions (5.2) are generally not satisfied.

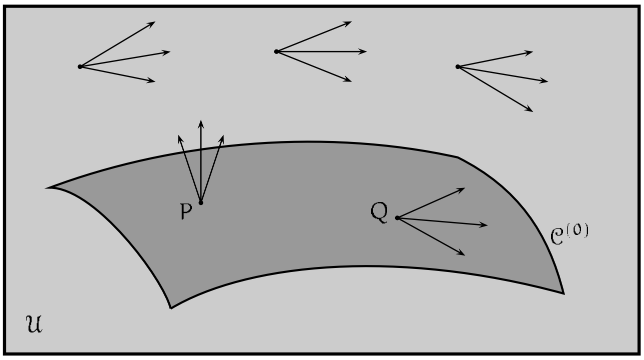

Fig. 6.1. Vector field pencil and submanifold

Thus we are not given a simple vector field on a manifold but a whole *pencil* of vector fields on $\mathcal{U}$ and a *submanifold* $\mathcal{C}^{(0)} = \tau_{\mathcal{U}}(\mathcal{S}^{(0)}) \subseteq \mathcal{U}$, the *constraint manifold* obtained via the tangent bundle projection (see Figure 6.1). In the semi-explicit form, the submanifold is described by the constraint equations (6.3b) and we have a pencil $\mathcal{P}^{(0)}$ of vector fields, as (6.3a) does not uniquely determine all components of a vector field (the $\mathbf{w}$ components remain unknown).

Example 6.1.1. As a trivial example we may consider the following linear system in semi-explicit form defined over $\mathcal{U} = \mathbb{R}^3$ with $\mathbf{u} = (u, v, w)$:

$$v' = u - 2\,, \quad w' = -v\,, \quad v = 3\,. \tag{6.4}$$

Here the constraint manifold $\mathcal{C}^{(0)}$ is the plane $v = 3$ and the pencil $\mathcal{P}^{(0)}$ contains all vector fields $X_a \in \mathfrak{X}(\mathcal{U})$ of the form $X_a(\mathbf{u}) = \bigl(a(\mathbf{u}), u - 2, -v\bigr)^t$ where $a \in \mathcal{F}(\mathcal{U})$ is an arbitrary function. $\lhd$

A solution of the differential algebraic equation must define an integral curve to one of the vector fields in the pencil $\mathcal{P}^{(0)}$ and this curve must lie in the constraint manifold $\mathcal{C}^{(0)}$. But this is only possible, if the vector field is tangential to $\mathcal{C}^{(0)}$. As shown in Figure 6.1, two possibilities arise. At some points $P \in \mathcal{C}^{(0)}$ all vectors contained in $\mathcal{P}^{(0)}$ are transversal to $\mathcal{C}^{(0)}$; such points can never lie on a curve defined by a solution of the differential algebraic equation and must be eliminated. This leads to a smaller constraint manifold $\mathcal{C}^{(1)} \subseteq \mathcal{C}^{(0)}$. At the remaining points Q the pencil $\mathcal{P}^{(0)}$ contains at least one vector tangential to $\mathcal{C}^{(0)}$. Here all the transversal vectors must be eliminated of the pencil and we obtain a shrunk pencil $\mathcal{P}^{(1)}$.

If $\mathcal{C}^{(1)}$ is a proper submanifold of $\mathcal{C}^{(0)}$, we must again check the tangency, as vectors tangential to $\mathcal{C}^{(0)}$ are not necessarily tangential to $\mathcal{C}^{(1)}$. Thus we obtain a decreasing sequence of submanifolds $\mathcal{C}^{(0)} \supset \mathcal{C}^{(1)} \supset \cdots \supset \mathcal{C}^{(k)}$ and accompanying vector field pencils $\mathcal{P}^{(0)} \supseteq \mathcal{P}^{(1)} \supseteq \cdots \supseteq \mathcal{P}^{(k)}$. The process stops as soon as $\mathcal{C}^{(i)} = \mathcal{C}^{(i+1)}$; depending on whether or not $\mathcal{P}^{(i)} = \mathcal{P}^{(i+1)}$ we have $k = i$ or $k = i + 1$. At this point we have obtained a submanifold $\mathcal{C}^{(k)} \subseteq \mathcal{U}$ and a pencil $\mathcal{P}^{(k)}$ of vector fields tangential to it.

In a more constructive manner we describe this process by two sequences of submanifolds $\mathcal{C}^{(i)} \subseteq \mathcal{U}$ and $\mathcal{S}^{(i)} \subseteq T\mathcal{U}$. The constraint manifold $\mathcal{C}^{(i)}$ is always obtained via the tangent bundle projection: $\mathcal{C}^{(i)} = \tau_{\mathcal{U}}(\mathcal{S}^{(i)})$. The new differential equation $\mathcal{S}^{(i)}$ arises as the intersection of the old one with the tangent space of its constraint manifold: $\mathcal{S}^{(i+1)} = \mathcal{S}^{(i)} \cap T\mathcal{C}^{(i)}$. The pencils $\mathcal{P}^{(i)}$ are uniquely defined only on the constraint manifolds $\mathcal{C}^{(i)}$; there they simply span the submanifolds $\mathcal{S}^{(i)}$. Off the submanifolds $\mathcal{C}^{(i)}$ the vector fields contained in $\mathcal{P}^{(i)}$ may be continued arbitrarily.

In most applications the final pencil consists of a single vector field (see Figure 6.2); otherwise we are dealing with an underdetermined system. Such a field is called an *underlying vector field.* Note again that it is uniquely determined only *on* the constraint manifold $\mathcal{C}^{(k)}$. If we add to an underlying vector field any field vanishing on $\mathcal{C}^{(k)}$, we obtain another underlying vector field. It is well possible that $\mathcal{C}^{(k)}$ is an empty submanifold. This simply means that the original system was *inconsistent*; it does not possess any solution, as the conditions it poses cannot be fulfilled simultaneously.

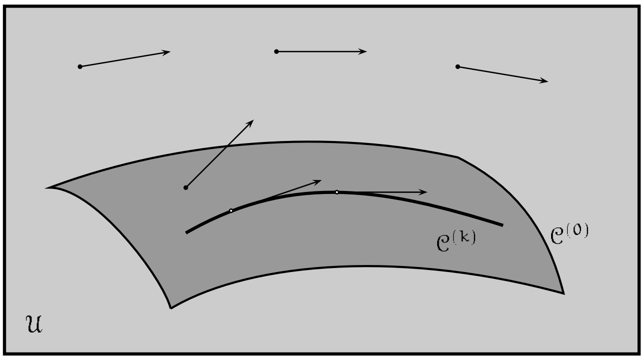

Fig. 6.2. Completed system

Proposition 5.1.1 may now be interpreted as follows: the final constraint manifold $\mathcal{C}^{(k)}$ is an *invariant manifold* for any vector field contained in the final pencil $\mathcal{P}^{(k)}$. Indeed, in dynamical systems theory an invariant manifold for a vector field has the property that any solution for initial data on it stays within the manifold for all times. And this property was our guiding principle in the discussion above.

Example 6.1.2. We continue Example 6.1.1. As the constraint manifold $\mathcal{C}^{(0)}$ is parallel to the u-w plane, a vector field is tangential to it, if and only if its v-component vanishes on $\mathcal{C}^{(0)}$. As the parameter a does not appear in this component, the tangency cannot be achieved by shrinking the pencil. Instead we have to eliminate all points $P \in \mathcal{C}^{(0)}$ which are not simultaneously on the plane $u = 2$. Thus the manifold $\mathcal{C}^{(1)}$ is the straight line $u = 2$, $v = 3$.

As the constraint manifold has become smaller, we have to check the tangency again. Vector fields tangential to $\mathcal{C}^{(1)}$ must have a u-component vanishing on $\mathcal{C}^{(1)}$. This can be achieved by shrinking the pencil to all vector fields X_a where the function a vanishes on $\mathcal{C}^{(1)}$. The constraint manifold remains unchanged, so that the completion stops here.

A simple underlying vector field is $X_1(\mathbf{u}) = (0, 0, -3)^t$. However, the field $X_2(\mathbf{u}) = (2 - u, v - 3, u - v - 2)^t$ is also a valid choice, as X_1 and X_2 coincide on the constraint manifold $\mathcal{C}_1$. $\triangleleft$

The extension of these results to *non-autonomous systems* requires essentially not more than replacing everywhere the tangent by the jet bundle. Thus we assume now that we are given a fibred manifold $\pi : \mathcal{E} \to \mathcal{X}$ over a one-dimensional base space $\mathcal{X}$; for simplicity we take $\mathcal{X} = \mathbb{R}$. Our starting point is a differential equation $\mathcal{R}_1 \subseteq J_1\mathcal{E}$. Instead of the sequence $\mathcal{S}^{(i)}$ we consider the sequence $\mathcal{R}_1^{(i)} \subseteq J_1\mathcal{E}$ constructed by $\mathcal{R}_1^{(i+1)} = \pi_1^2(\mathcal{R}_2^{(i)})$. The completion stops as soon as $\mathcal{R}_1^{(i+1)} = \mathcal{R}_1^{(i)}$.

At first sight, this looks very different compared to the completion process for autonomous systems. The definition of $\mathcal{R}_1^{(i+1)}$ requires as first step a prolongation to second order and then we project back to first order. In the autonomous case, we reversed the order of the operations: first we projected to the constraint manifold, then we "prolonged" by considering its tangent space. But for *ordinary* differential equations we have the following result.

Proposition 6.1.3. *Let $\mathcal{R}_q \subseteq J_q\mathcal{E}$ be an ordinary differential equation. Then $\mathcal{R}_q^{(1)} = \mathcal{R}_q \cap J_1\mathcal{R}_{q-1}^{(1)}$ (here we consider $J_1\mathcal{R}_{q-1}^{(1)}$ as a submanifold of $J_q\mathcal{E}$).*

Proof. For notational simplicity we give the proof for the case of a first-order equation: $q = 1$. The extension to the general case is trivial. Let a local representation of the differential equation $\mathcal{R}_1$ be

$$\mathcal{R}_1 : \begin{cases} \mathbf{v}' = \boldsymbol{\phi}(x, \mathbf{v}, \mathbf{w}) \,, \\ 0 = \boldsymbol{\psi}(x, \mathbf{v}, \mathbf{w}) \,. \end{cases} \tag{6.5}$$

This is the obvious generalisation of the semi-explicit form (6.3) to non-autonomous systems. We need a local description of $\mathcal{R}_1^{(1)}$. Because of the semi-explicit form it is easy to see that the prolongation of the equations containing $\mathbf{v}'$ leads to second-order equations and each equation is solved for a different derivative. Thus no cancellations of second-order derivatives are possible and these equations drop out when we project back to first order.

A local representation of $\mathcal{R}_1^{(1)}$ is obtained by adding to the equations already contained in (6.5) the equations

$$\frac{\partial \boldsymbol{\psi}}{\partial x} + \frac{\partial \boldsymbol{\psi}}{\partial \mathbf{v}}\mathbf{v}' + \frac{\partial \boldsymbol{\psi}}{\partial \mathbf{w}}\mathbf{w}' = 0 \tag{6.6}$$

obtained by prolonging the algebraic equations in (6.5). Of course, in general this is a redundant description, as these new equations are not necessarily independent of the original ones. But this is of no importance here.

It is trivial to see that the same equations also define a local representation of $\mathcal{R}_1 \cap J_1\mathcal{R}_0^{(1)}$. The constraint manifold $\mathcal{R}_0^{(1)}$ is locally described by the algebraic equations in (6.5) and we must simply add their prolongations (6.6) to the original equations for the intersection. □

Hence we may reverse the order of prolongation and projection for ordinary differential equations. There remains to show that the final differential equation $\mathcal{R}_1^{(k)}$ obtained in our completion process is indeed an involutive equation. Obviously, any smooth solution of the original equation $\mathcal{R}_1$ is also a solution of $\mathcal{R}_1^{(k)}$, as it automatically satisfies all the integrability conditions added during the completion.

Corollary 6.1.4. *The final differential equation $\mathcal{R}_1^{(k)}$ is involutive.*

Proof. As we are dealing with ordinary differential equations, involution is equivalent to formal integrability. By the previous proposition and our construction it is obvious that $\mathcal{R}_1^{(k)} = \mathcal{R}_1^{(k+\ell)}$ for any $\ell > 0$. But this means that it is a formally integrable equation. □

A closer analysis of (6.6) clarifies the two different effects described in the autonomous case: shrinking of either the constraint manifold or the vector field pencil. The derivatives $\mathbf{v}'$ in (6.6) can be eliminated using the differential equations in (6.5). Thus the rank of the Jacobian $\partial\boldsymbol{\psi}/\partial\mathbf{w}$ determines how many new differential equations, i.e. integrability conditions of order one, appear. Such new differential equations correspond to a shrinking of the pencil. If the Jacobian does not possess maximal rank, we can derive some algebraic equations from (6.6). If these are independent of the constraints in (6.5), we have found further integrability conditions leading to a smaller constraint manifold.

Obviously, the completion process always terminates, as the proof of Proposition 6.1.3 shows that $\dim \mathcal{R}_1^{(i+1)} < \dim \mathcal{R}_1^{(i)}$ until the process stops (we have additional equations). As $J_1\mathcal{E}$ is a finite-dimensional manifold, we cannot have an infinite decreasing sequence. One should note that the assumption of a regular equation is crucial for this proof. It is equivalent to the assumption of constant rank of all appearing Jacobians which is necessary to allow us at each step the distinction into algebraic and differential equations.

6.2 The Cartan–Kuranishi Completion

In the previous section we mainly worked in the tangent bundle, as this makes the completion process for (autonomous) ordinary differential equations very transparent and intuitive. Now we derive in an intrinsic manner a rather general completion method that can – in principle – handle any *regular* system of partial differential equations. However, as we will see, some of the required operations cannot always be performed effectively.

Recall that a differential equation $\mathcal{R}_q$ is regular, if all its prolongations and subsequent projections lead to fibred submanifolds. Furthermore, we assume that the indices of all symbols are constant. Obviously, it is not easy to check in advance whether a given differential equation satisfies this condition. Usually, it can only be verified during the completion. A trivial exception are linear equations.

Theorem 6.2.1 (Cartan–Kuranishi). *Let $\mathcal{R}_q \subseteq J_q\mathcal{E}$ be a regular differential equation. Then two integers $r, s \geq 0$ exist such that $\mathcal{R}_{q+r}^{(s)}$ is involutive.*

Proof. We give a constructive proof leading directly to the promised general completion method. It is based on Algorithm 6.1. If we can show that it always terminates with an involutive equation of the form $\mathcal{R}^{(s)}_{q+r}$, we are done. Thus there are two points to prove: the correctness and the termination.

Algorithm 6.1 Cartan–Kuranishi Completion

Input: regular differential equation $\mathcal{R}_q \subseteq J_q\mathcal{E}$
Output: equivalent involutive equation $\mathcal{R}^{(s)}_{q+r}$

$r \leftarrow 0$; $s \leftarrow 0$
repeat
    while $\mathcal{M}^{(s)}_{q+r}$ not involutive **do**
        $r \leftarrow r + 1$
    end_while
    $intConds \leftarrow \mathcal{R}^{(s+1)}_{q+r} \subsetneq \mathcal{R}^{(s)}_{q+r}$
    if $intConds$ **then**
        $s \leftarrow s + 1$
    end_if
until $\neg intConds$
return $\mathcal{R}^{(s)}_{q+r}$

The *correctness* is simple. Assume that the algorithm terminates. At the end of the inner loop the symbol $\mathcal{M}^{(s)}_{q+r}$ is involutive. If *intConds* is false, then $\mathcal{R}^{(s+1)}_{q+r} = \mathcal{R}^{(s)}_{q+r}$. Thus all conditions of Theorem 3.2.3 are satisfied and the equation $\mathcal{R}^{(s)}_{q+r}$ is involutive.

For the *termination* we must consider the two loops in Algorithm 6.1. In local coordinates at a given point $\rho \in \mathcal{R}^{(s)}_{q+r}$ the inner one corresponds by definition of an involutive symbol to the determination of a Pommaret basis. We will prove later in this chapter that such a basis always exists and thus the loop always terminates provided our coordinates are δ-regular (Theorem 6.4.6 together with the results of Section 2.4).

The termination of the outer loop follows from a Noetherian argument. At the end of the sth iteration of the outer loop we arrive at the equation $\mathcal{R}^{(s)}_{q+r_s}$ where r_s is the total number of iterations through the inner loop performed so far. If $\rho \in \mathcal{R}^{(s)}_{q+r_s}$ is an arbitrary fixed point, then for $0 \leq \bar{s} \leq s$ the points $\rho^{(\bar{s})} = \pi^{q+r_s}_{q+r_{\bar{s}}}(\rho)$ lie in $\mathcal{R}^{(\bar{s})}_{q+r_{\bar{s}}}$. As discussed in Section 3.1, the equations defining the symbol space $(\mathcal{M}^{(\bar{s})}_{q+r_{\bar{s}}})_{\rho^{(\bar{s})}}$ generate the symbol module[2] $\mathcal{M}^{\mathcal{P}}_{\bar{s}}$, a submodule of a free polynomial module over $\mathcal{P} = \Bbbk[x^1, \dots, x^n]$. Obviously, we have $\mathcal{M}^{\mathcal{P}}_0 \subseteq \langle \mathcal{M}^{\mathcal{P}}_0, \mathcal{M}^{\mathcal{P}}_1 \rangle \subseteq \cdots \subseteq \langle \mathcal{M}^{\mathcal{P}}_0, \mathcal{M}^{\mathcal{P}}_1, \dots, \mathcal{M}^{\mathcal{P}}_s \rangle$; in fact, each integrability condition added in the outer loop yields a new module generator.

[2] Here the subscript does not refer to a homogeneous component but labels different modules.

Thus we have constructed an ascending chain of submodules of a free polynomial module. As polynomial modules are Noetherian, this chain must become stationary for some value of s. But this implies that in the iteration of the outer loop corresponding to this value *no* integrability condition was found. Hence the algorithm terminates at this iteration. $\square$

If we apply Algorithm 6.1 to a concrete differential equation, we alternate prolongations and projections, thus, strictly speaking, we work with differential equations $((((\mathcal{R}_{q+r_1})^{(1)})_{+r_2})^{(1)})\cdots)_{+r_s}$. However, we only project equations with an involutive symbol. Thus, by Part (ii) of Proposition 3.1.10, these equations coincide with $\mathcal{R}^{(s)}_{q+r_1+\cdots+r_s}$.

The involutive equation $\mathcal{R}^{(s)}_{q+r}$ is *equivalent* to the original equation $\mathcal{R}_q$ in the sense that both equations possess the same formal solution space. In fact, both equations have even the same smooth solutions. Differences may show up only for solutions of $\mathcal{R}_q$ contained in $C^k(\Omega)$ with $q \leq k < q + r$. They cannot be strong solutions of $\mathcal{R}^{(s)}_{q+r}$, as they are lacking the necessary differentiability. But we may expect that they are at least weak solutions.

Example 6.2.2. We demonstrate the working of Algorithm 6.1 for a classical linear second-order equation due to Janet:

$$\mathcal{R}_2 : \begin{cases} u_{zz} + y u_{xx} = 0 \, , \\ u_{yy} = 0 \, . \end{cases} \tag{6.7}$$

Thus we have one dependent variable and three independent variable. Obviously, the system is overdetermined.

According to our algorithm we prolong until the symbol becomes involutive. One readily checks that the indices of the symbol $\mathcal{M}_2$ are $\beta_2^{(3)} = \beta_2^{(2)} = 1$ and $\beta_2^{(1)} = 0$. The rank of the symbol matrix M_3 is 6 (the non-multiplicative prolongation $u_{yyz} = 0$ is necessary to generate $\mathcal{M}_3$), thus $\mathcal{M}_2$ is not involutive and we must prolong. In a lengthy computation one shows that $\mathcal{M}_3$ is involutive and the inner loop terminates with $r = 1$. Now we must check for integrability conditions. One finds $\dim \mathcal{R}_3^{(1)} = 11 < \dim \mathcal{R}_3 = 12$. Thus $\mathcal{R}_3$ is not involutive, as one condition is hidden. Indeed,

$$\partial_{yy}(u_{zz} + y u_{xx}) - \partial_{zz} u_{yy} - y \partial_{xx} u_{yy} = u_{xxy} = 0 \, . \tag{6.8}$$

Thus at the end of the first iteration of the outer loop $r = s = 1$ and we continue with the analysis of $\mathcal{R}_3^{(1)}$. The calculations are now becoming more and more tedious, as the number of equations needed in local representations of either the differential equations or their symbols rapidly grows. So we can only sketch the progress of the computation.

$\mathcal{M}_3^{(1)}$ is again not involutive, since setting up $\mathcal{M}_4^{(1)}$ requires the non-multiplicative prolongation $u_{xxyz} = 0$. The symbol $\mathcal{M}_4^{(1)}$ is involutive and at the end of the inner loop $r = 2$. But again we find an integrability condition,

$$\partial_{xxy}(u_{zz} + yu_{xx}) - y\partial_{xx}u_{xxy} - \partial_{zz}u_{xxy} = u_{xxxx} = 0 \,, \tag{6.9}$$

and $\mathcal{R}_4^{(1)}$ is not involutive. Thus now $s = 2$ and we continue with $\mathcal{R}_4^{(2)}$.

Again, the symbol $\mathcal{M}_4^{(2)}$ is not involutive because of the non-multiplicative prolongation $u_{xxxxz} = 0$. Thus we prolong once more in order to obtain $\mathcal{R}_5^{(2)}$. Finally, the algorithm terminates with $r = 3$ and $s = 2$, as the symbol $\mathcal{M}_5^{(2)}$ is involutive and no integrability conditions appear during the prolongation so that $\mathcal{R}_5^{(2)}$ is involutive. Thus, $\mathcal{R}_4^{(2)}$ is another example of an equation which is formally integrable but *not* involutive. The problem is that we can detect this only in hindsight, after the completion has terminated.

In fact, the involution of $\mathcal{M}_5^{(2)}$ is trivial, as it vanishes. Thus we are dealing here with an equation of finite type. Its solution space is twelve-dimensional as $\dim \mathcal{R}_5^{(2)} = 12$. Indeed, a direct integration of the completed equation yields the solution

$$\begin{aligned} u(x,y,z) = &\frac{1}{3}a_1x^3z - a_2xyz^2 + a_3x^3 + a_2x^2z + a_4xyz - a_1yz^2 + \\ &a_5x^2 + a_6xy + a_7xz + a_8yz + a_9x + a_{10}y + a_{11}z + a_{12} \,. \end{aligned} \tag{6.10}$$

As one can see the straightforward application of Algorithm 6.1 becomes quite tedious even for such small equations. In our particular example one needs a total of 132 prolongations in order to get local representations for all appearing equations. Furthermore, one must compute row echelon forms of symbol matrices and Jacobians; the largest is an 86×84 matrix. We will show later that by a clever organisation of the computations one can completely avoid the determination of row echelon forms and reduce the number of prolongations by about 80%. ◁

Algorithm 6.1 does not really represent an algorithm but only a "method", i. e. it is not clear whether all steps can be performed effectively. This concerns in particular the test for inequality in Line /6/. To decide whether or not a symbol is involutive requires basically only some linear algebra and is thus fully algorithmic. However, even there it may happen for non-linear systems that not all necessary operations with the coefficients of the symbol can be performed effectively.

The test in Line /6/ requires to decide whether certain equations are algebraically independent. Depending on the type of the equation this can be highly non-trivial. For *linear* equations it is again only a matter of some linear algebra. For *polynomial* equations it can – at least in principle – be done using Gröbner bases (see Appendix B.2), although the computations suffer potentially from a very high complexity. For arbitrary non-linear equations no general criterion is known. Depending on the class of functions considered the situation may become even worse. While analytic functions are either independent or not, smooth functions may be independent in some domains and dependent in others.

Finally, we compare Algorithm 6.1 with the geometric completion procedure for ordinary differential equations introduced in the last section. As any ordinary differential equation has an involutive symbol, the inner loop becomes superfluous. The algorithm reduces to a simple *prolong–project*, *prolong–project*, ... But we have seen in the proof of Proposition 6.1.3 that it is possible to swap the operations and use *project–prolong* cycles. The result is the same, but the modified algorithm is computationally much more efficient, as much less equations must be prolonged. But one should stress again that this modification is allowed only for *ordinary* differential equations, as only there one may omit the analysis of the symbol.

6.3 Continuous and Constructive Divisions

In Chapter 2 we introduced involutive bases. But we postponed the question how they can actually be constructed. Unfortunately, for arbitrary involutive division no satisfying solution is known so far. One can follow a brute force approach, namely performing a breadth first search through the tree of all possible completions. Obviously, it terminates only, if a finite basis exists. But for divisions satisfying some additional properties one can design a fairly efficient completion algorithm.

The first problem in constructing an involutive completion of a finite subset $\mathcal{N} \subset \mathbb{N}_0^n$ for a division L is to check whether $\mathcal{N}$ is already involutive. The trouble is that we do not know a priori where obstructions to involution might lie. As these multi indices must somehow be related to the non-multiplicative indices of the elements of $\mathcal{N}$, the multi indices $\nu + 1_j$ with $\nu \in \mathcal{N}$ and $j \in \bar{N}_{L,\mathcal{N}}(\nu)$ are a natural first guess.

Definition 6.3.1. *The finite set $\mathcal{N} \subset \mathbb{N}_0^n$ is* locally involutive *for the involutive division L, if $\nu + 1_j \in \langle\mathcal{N}\rangle_L$ for every non-multiplicative index $j \in \bar{N}_{L,\mathcal{N}}(\nu)$ of every multi index $\nu \in \mathcal{N}$.*

While involution obviously implies local involution, the converse does not hold, as the following rather bizarre example of a globally defined division due to Gerdt and Blinkov [56] demonstrates.

Example 6.3.2. We consider $\mathbb{N}^3$. The involutive division L is defined by the following assignment of multiplicative indices. With the exception of four multi indices all elements of $\mathbb{N}^3$ do not have any multiplicative indices. These four exceptions are: $N_L([0,0,0]) = \{1,2,3\}$, $N_L([1,0,0]) = \{1,3\}$, $N_L([0,1,0]) = \{1,2\}$, and $N_L([0,0,1]) = \{2,3\}$. It is straightforward to show that this indeed defines an involutive division.

Now we take the set $\mathcal{N} = \{[1,0,0],[0,1,0],[0,0,1]\}$. It is not involutive, as the multi index $[1,1,1] \in \langle\mathcal{N}\rangle$ is not contained in the involutive span $\langle\mathcal{N}\rangle_L$. But it is locally involutive, as the three multi indices $[1,1,0]$, $[0,1,1]$ and $[1,0,1]$ obtained by taking the elements of $\mathcal{N}$ and adding their respective non-multiplicative index are contained in $\langle\mathcal{N}\rangle_L$. $\triangleleft$

Definition 6.3.3. *Let* L *be an involutive division and* $\mathcal{N} \subset \mathbb{N}_0^n$ *a finite set. Let furthermore* $(\nu^{(1)},\dots,\nu^{(t)})$ *be a finite sequence of elements of* $\mathcal{N}$ *where every multi index* $\nu^{(k)}$ *with* $k < t$ *has a non-multiplicative index* $j_k \in \bar{N}_{L,\mathcal{N}}(\nu^{(k)})$ *such that* $\nu^{(k+1)} \mid_{L,\mathcal{N}} \nu^{(k)} + 1_{j_k}$. *The division* L *is called* continuous, *if any such sequence consists only of distinct elements, i. e. if* $\nu^{(k)} \neq \nu^{(\ell)}$ *for all* $k \neq \ell$.

It is easy to see that the division of Example 6.3.2 is not continuous. And indeed for this special class of divisions local involution always implies involution. This makes these divisions very convenient for concrete computations: if obstructions to involution exist, some of them must be of the form $\nu + 1_j$ with $\nu \in \mathcal{N}$ and index $j \in \bar{N}_{L,\mathcal{N}}(\nu)$ and all others are "multiples" of these.

Proposition 6.3.4. *For a continuous division* L, *any locally involutive set* $\mathcal{N} \subset \mathbb{N}_0^n$ *is involutive.*

Proof. Let the set Σ contain those obstructions to involution that are of minimal length. We claim that for a continuous division L all multi indices $\sigma \in \Sigma$ are of the form $\nu + 1_j$ with $\nu \in \mathcal{N}$ and $j \in \bar{N}_{L,\mathcal{N}}(\nu)$. This immediately implies our proposition: since for a locally involutive set all such multi indices are contained in $\langle\mathcal{N}\rangle_L$, we must have $\Sigma = \emptyset$ and thus $\langle\mathcal{N}\rangle = \langle\mathcal{N}\rangle_L$.

In order to prove our claim, we choose a $\sigma \in \Sigma$ for which no $\nu \in \mathcal{N}$ exists with $\sigma = \nu + 1_j$. We collect in $\mathcal{N}_\sigma$ all divisors $\nu \in \mathcal{N}$ of σ of maximal length. Let $\nu^{(1)}$ be an element of $\mathcal{N}_\sigma$; by assumption the multi index $\mu^{(1)} = \sigma - \nu^{(1)}$ satisfies $|\mu^{(1)}| > 1$ and at least one non-multiplicative index $j_1 \in \bar{N}_{L,\mathcal{N}}(\nu^{(1)})$ exists with $\mu^{(1)}_{j_1} > 0$. By the definition of Σ we have $\nu^{(1)} + 1_{j_1} \in \langle\mathcal{N}\rangle_L$. Thus a multi index $\nu^{(2)} \in \mathcal{N}$ exists with $\nu^{(2)} \mid_{L,\mathcal{N}} \nu^{(1)} + 1_{j_1}$. This implies $\nu^{(2)} \mid \sigma$ and we set $\mu^{(2)} = \sigma - \nu^{(2)}$. By the definition of the set $\mathcal{N}_\sigma$ we have $|\nu^{(2)}| \leq |\nu^{(1)}|$. Hence $\nu^{(2)} + 1_j \in \langle\mathcal{N}\rangle_L$ for all j.

Choose a non-multiplicative index $j_2 \in \bar{N}_{L,\mathcal{N}}(\nu^{(2)})$ with $\mu^{(2)}_{j_2} > 0$. Such an index exists as otherwise $\sigma \in \langle\mathcal{N}\rangle_L$. By the same arguments as above, a multi index $\nu^{(3)} \in \mathcal{N}$ exists with $\nu^{(3)} |_{L,\mathcal{N}} \nu^{(2)} + 1_{j_2}$ and $|\nu^{(3)}| \leq |\nu^{(2)}|$. We can iterate this process and produce an infinite sequence $(\nu^{(1)}, \nu^{(2)}, \dots)$ where each multi index $\nu^{(i)} \in \mathcal{N}$ and $\nu^{(i+1)} |_{L,\mathcal{N}} \nu^{(i)} + 1_{j_i}$ with $j_i \in \bar{N}_{L,\mathcal{N}}(\nu^{(i)})$. As $\mathcal{N}$ is a finite set, the elements of the sequence cannot be all different. This contradicts our assumption that L is a continuous division: by taking a sufficiently large part of this sequence we obtain a finite sequence with all properties mentioned in Definition 6.3.3 but containing some identical elements. Hence a multi index $\nu \in \mathcal{N}$ must exist such that $\sigma = \nu + 1_j$. □

Continuity is not only sufficient for the equivalence of local involution and involution but also useful for proving other properties. The main point is that it excludes the existence of infinite reduction chains. The two involutive divisions introduced in Examples 2.1.2 and 2.1.5 are both continuous.

Lemma 6.3.5. *The Janet and the Pommaret division are continuous.*

Proof. Let $\mathcal{N} \subseteq \mathbb{N}_0^n$ be a finite set and $(\nu^{(1)}, \dots, \nu^{(t)})$ a finite sequence where $\nu^{(i+1)} |_{L,\mathcal{N}} \nu^{(i)} + 1_j$ with $j \in \bar{N}_{L,\mathcal{N}}(\nu^{(i)})$ for $1 \leq i < t$.

We claim that for $L = J$, the Janet division, $\nu^{(i+1)} \succ_{\text{lex}} \nu^{(i)}$ implying that the sequence is strictly ascending and cannot contain any identical entries. Set $k = \max\{i \mid \mu_i \neq \nu_i\}$. Then $j \leq k$, as otherwise $j \in N_{J,\mathcal{N}}(\nu^{(i+1)})$ entails $j \in N_{J,\mathcal{N}}(\nu^{(i)})$ contradicting our assumption that j is non-multiplicative for the multi index $\nu^{(i)}$. But $j < k$ is also not possible, as then $\nu_k^{(i+1)} < \nu_k^{(i)}$ and so k cannot be multiplicative for $\nu^{(i+1)}$. There remains as only possibility $j = k$. In this case we must have $\nu_j^{(i+1)} = \nu_j^{(i)} + 1$ (otherwise j could not be multiplicative for $\nu^{(i+1)}$). Thus we conclude that $\nu^{(i+1)} \succ_{\text{lex}} \nu^{(i)}$ and the Janet division is continuous.

The proof for the case $L = P$, the Pommaret division, is slightly more subtle.[3] The condition $j \in \bar{N}_P(\nu^{(i)})$ implies that $\operatorname{cls}(\nu^{(i)} + 1_j) = \operatorname{cls}\nu^{(i)}$ and if $\nu^{(i+1)} |_P \nu^{(i)} + 1_j$, then $\operatorname{cls}\nu^{(i+1)} \geq \operatorname{cls}\nu^{(i)}$, i.e. the class of the elements of the sequence is monotonously increasing. If $\operatorname{cls}\nu^{(i+1)} = \operatorname{cls}\nu^{(i)} = k$, then the involutive divisibility requires that $\nu_k^{(i+1)} \leq \nu_k^{(i)}$, i.e. among the elements of the sequence of the same class the corresponding entry is monotonously decreasing. And if finally $\nu_k^{(i+1)} = \nu_k^{(i)}$, then we must have $\nu^{(i+1)} = \nu^{(i)} + 1_j$, i.e. the length of the elements is strictly increasing. Hence all elements of the sequence are different and the Pommaret division is continuous. □

[3] It is tempting to tackle the lemma for the Pommaret division in the same manner as for the Janet division using $\prec_{\text{revlex}}$ instead of $\prec_{\text{lex}}$; in fact, such a "proof" is contained in [56]. Unfortunately, it is not correct, as $\prec_{\text{revlex}}$ is not a term order (see Example B.2.1): if $\nu^{(i+1)} = \nu^{(i)} + 1_j$, then $\nu^{(i+1)} \prec_{\text{revlex}} \nu^{(i)}$ although the latter multi index is a divisor of the former one! Thus the sequences considered in Definition 6.3.3 are in general not strictly ascending with respect to $\prec_{\text{revlex}}$.

Definition 6.3.6. *Let $\mathcal{N} \subset \mathbb{N}_0^n$ be a finite set of multi indices and choose a multi index $\nu \in \mathcal{N}$ and a non-multiplicative index $j \in \bar{N}_{L,\mathcal{N}}(\nu)$ such that*

(i) $\nu + 1_j \notin \langle \mathcal{N} \rangle_L$.
(ii) *If there exists $\mu \in \mathcal{N}$ and $k \in \bar{N}_{L,\mathcal{N}}(\nu)$ such that $\mu + 1_k \mid \nu + 1_j$ but $\mu + 1_k \neq \nu + 1_j$, then $\mu + 1_k \in \langle \mathcal{N} \rangle_L$.*

The division L is constructive, *if for any such set $\mathcal{N}$ and any such multi index $\nu + 1_j$ no multi index $\rho \in \langle \mathcal{N} \rangle_L$ with $\nu + 1_j \in \mathcal{C}_{L,\mathcal{N}\cup\{\rho\}}(\rho)$ exists.*

In words, constructivity can roughly be explained as follows. The conditions imposed on ν and j ensure a kind of minimality: no proper divisor of $\nu + 1_j$ is of the form $\mu + 1_k$ for a $\mu \in \mathcal{N}$ and not contained in the involutive span $\langle \mathcal{N} \rangle_L$. The conclusion implies that it is useless to add multi indices to $\mathcal{N}$ that lie in some involutive cone, as none of them can be an involutive divisor of $\nu + 1_j$. An efficient completion algorithm for a constructive division should consider only non-multiplicative indices. Because of the restriction to multi indices of the form $\nu + 1_j$, constructivity is really important only for divisions which are also continuous.

Lemma 6.3.7. *Any globally defined division (and thus the Pommaret division) is constructive. The Janet division is constructive, too.*

Proof. For a globally defined division the proof is very simple. For any multi index $\rho \in \langle \mathcal{N} \rangle_L$ there exists a multi index $\mu \in \mathcal{N}$ such that $\rho \in \mathcal{C}_L(\mu)$. As for a globally defined division the multiplicative indices are independent of the reference set, we must have by the definition of an involutive division that $\mathcal{C}_L(\rho) \subseteq \mathcal{C}_L(\mu)$. Hence adding such a multi index to $\mathcal{N}$ cannot change the involutive span and if $\nu + 1_j \notin \langle \mathcal{N} \rangle_L$, then also $\nu + 1_j \notin \langle \mathcal{N} \cup \{\rho\} \rangle_L$. But this implies the constructivity of L.

The proof of the constructivity of the Janet division is more involved. The basic idea is to show that if it was not constructive, it also could not be continuous. Let $\mathcal{N}$, ν, j be as described in Definition 6.3.6. Assume for a contradiction that a multi index $\rho \in \langle \mathcal{N} \rangle_J$ exists with $\nu + 1_j \in \mathcal{C}_{J,\mathcal{N}\cup\{\rho\}}(\rho)$. We write $\rho = \nu^{(1)} + \mu$ for a multi index $\nu^{(1)} \in \mathcal{N}$ with $\rho \in \mathcal{C}_{J,\mathcal{N}}(\nu^{(1)})$. As $\nu + 1_j \notin \langle \mathcal{N} \rangle_J$, we must have $|\mu| > 0$. Set $\lambda = \nu + 1_j - \rho$ and let m, l be the maximal indices such that $\mu_m > 0$ and $\lambda_l > 0$, respectively.

We claim that $j > \max\{m, l\}$. Indeed, if $j \leq m$, then $\nu_m^{(1)} < \nu_m$ and, by definition of the Janet division, this implies that $m \notin N_{J,\mathcal{N}}(\nu^{(1)})$, a contradiction. Similarly, we cannot have $j < l$, as then $l \notin N_{J,\mathcal{N}\cup\{\rho\}}(\rho)$. Finally, $j = l$ is not possible. As we know already that $j > m$, we have in this case that $\rho_i = \nu_i^{(1)} = \nu_i$ for all $i > j$ and $\rho_j \leq \nu_j$. Hence $j \in \bar{N}_{J,\mathcal{N}\cup\{\rho\}}(\nu)$ and this implies furthermore $j \in \bar{N}_{J,\mathcal{N}\cup\{\rho\}}(\rho)$, a contradiction.

Now we are going to construct a sequence as in Definition 6.3.3 of a constructive division. Choose an index j_1 with $\lambda_{j_1} > 0$ and $j_1 \in \bar{N}_{J,\mathcal{N}}(\nu^{(1)})$. Such an index exists, as otherwise $\nu + 1_j \in \mathcal{C}_{J,\mathcal{N}}(\nu^{(1)}) \subset \langle \mathcal{N} \rangle_J$. So we may write $\nu + 1_j = (\nu^{(1)} + 1_{j_1}) + \mu + \lambda - 1_{j_1}$. Because of $|\mu| > 0$, the multi index $\nu^{(1)} + 1_{j_1}$ is a proper divisor of $\nu + 1_j$ and according to our assumptions a $\nu^{(2)} \in \mathcal{N}$ exists with $\nu^{(1)} + 1_{j_1} \in \mathcal{C}_{J,\mathcal{N}}(\nu^{(2)})$.

By the same arguments as above an index $j_2 \in \bar{N}_{J,\mathcal{N}}(\nu^{(2)})$ must exist with $(\mu + \lambda - 1_{j_1})_{j_2} > 0$ and a further multi index $\nu^{(3)} \in \mathcal{N}$ such that $\nu^{(2)} + 1_{j_2} \in \mathcal{C}_{J,\mathcal{N}}(\nu^{(3)})$. Thus we can iterate and produce an infinite sequence $(\nu^{(1)}, \nu^{(2)}, \dots)$ where $\nu^{(i+1)} \,|_{J,\mathcal{N}}\, \nu^{(i)} + 1_{j_i}$ with $j_i \in \bar{N}_{J,\mathcal{N}}(\nu^{(i)})$. By the continuity of the Janet division all members of the sequence must be different. But every multi index $\nu^{(i)}$ is a divisor of $\nu + 1_j$, so obviously only finitely many of them can be different. Thus the sequence must terminate which only happens, if $\nu + 1_j \in \mathcal{C}_{J,\mathcal{N}}(\nu^{(i)})$ for some i contradicting our assumptions. □

6.4 Computation of Involutive Bases

We present now an algorithm for determining involutive bases of left ideals in a polynomial algebra of solvable type $(\mathcal{P}, \star, \prec)$. Note that this includes simultaneously as special cases the completion of polynomial ideals and of linear systems of differential equations. As mentioned above, for arbitrary involutive divisions, nobody has so far been able to find a reasonable algorithm. But if we assume that the division is continuous and constructive, then a very simple algorithm exists. As in Chapter 2 we start with the "monomial" case, i. e. we consider the problem how we can complete a given finite set $\mathcal{N} \subset \mathbb{N}_0^n$ of multi indices to an involutive set with respect to some involutive division L, and extend later to finite polynomial sets $\mathcal{F} \subset \mathcal{P}$.

Algorithm 6.2 Completion in $(\mathbb{N}_0^n, +)$

Input: a finite set $\mathcal{N} \subset \mathbb{N}_0^n$, an involutive division L, a term order $\prec$
Output: an involutive completion $\bar{\mathcal{N}}$ of $\mathcal{N}$

/1/ $\bar{\mathcal{N}} \leftarrow \mathcal{N}$
/2/ **repeat**
/3/ $\quad \mathcal{S} \leftarrow \{\nu + 1_j \mid \nu \in \bar{\mathcal{N}},\ j \in \bar{N}_{L,\bar{\mathcal{N}}}(\nu),\ \nu + 1_j \notin \langle \bar{\mathcal{N}} \rangle_L\}$
/4/ $\quad \bar{\mathcal{N}} \leftarrow \bar{\mathcal{N}} \cup \{\min_\prec \mathcal{S}\}$
/5/ **until** $\mathcal{S} = \emptyset$
/6/ **return** $\bar{\mathcal{N}}$

The strategy behind Algorithm 6.2 is fairly natural given the results of the last section. In Line /3/ it collects in a set $\mathcal{S}$ all "minimal" obstructions to involution. For a continuous division L, the set $\mathcal{N}$ is involutive, if and only if $\mathcal{S} = \emptyset$. Furthermore, for a constructive division L it does not make sense to add elements of $\langle\mathcal{N}\rangle_L$ to $\mathcal{N}$ in order to complete. Thus we add in Line /4/ an element of $\mathcal{S}$. This element is chosen with the help of a term order.

Proposition 6.4.1. *Let the finite set $\mathcal{N} \subset \mathbb{N}_0^n$ possess a finite involutive completion with respect to the continuous and constructive division* L. *Then Algorithm 6.2 terminates with an involutive completion $\bar{\mathcal{N}}$ of $\mathcal{N}$.*

Proof. If Algorithm 6.2 terminates, its correctness is obvious. The criterion for its termination, $\mathcal{S} = \emptyset$, is equivalent to local involution of $\bar{\mathcal{N}}$. By Proposition 6.3.4, local involution implies for a continuous division involution. Thus the result $\bar{\mathcal{N}}$ is an involutive completion of $\mathcal{N}$, as by construction $\mathcal{N} \subseteq \bar{\mathcal{N}} \subset \langle\mathcal{N}\rangle$.

If the input set $\mathcal{N}$ is already involutive, Algorithm 6.2 leaves it unchanged and thus obviously terminates. Let us assume that $\bar{\mathcal{N}}$ is not yet involutive. In the first iteration of the `repeat` loop a multi index of the form $\nu + 1_j$ is added to $\mathcal{N}$. It is not contained in $\langle\mathcal{N}\rangle_L$ and minimal with respect to $\prec$ among all such non-multiplicative "multiples" of multi indices in $\mathcal{N}$. If $\mathcal{N}_L$ is an arbitrary involutive completion of $\mathcal{N}$, it must contain a multi index $\mu \notin \mathcal{N}$ such that $\mu\,|_{L,\mathcal{N}_L}\,\nu + 1_j$. We claim that $\mu = \nu + 1_j$.

Assume that $\mu \neq \nu + 1_j$. Since $\mathcal{N}_L \subset \langle\mathcal{N}\rangle$, μ must lie in the cone of some multi index $\nu^{(1)} \in \mathcal{N}$. We will show that, because of the continuity of L, $\mu \in \langle\mathcal{N}\rangle_L$ contradicting the constructivity of L. If $\nu^{(1)}\,|_{L,\mathcal{N}}\,\mu$, we are done. Otherwise we write $\mu = \nu^{(1)} + \rho^{(1)}$ for some multi index $\rho^{(1)} \in \mathbb{N}_0^n$. A non-multiplicative index $j_1 \in \bar{N}_{L,\mathcal{N}}(\nu^{(1)})$ with $\rho^{(1)}_{j_1} > 0$ must exist. Consider the multi index $\nu^{(1)} + 1_{j_1}$. Because of $\nu^{(1)} + 1_{j_1} \mid \mu$ and $\mu \mid \nu + 1_j$ the inequality $\nu^{(1)} + 1_{j_1} \prec (\nu + 1_j)$ holds. As $\nu + 1_j$ is the minimal element of $\mathcal{S}$ with respect to $\prec$, we must have $\nu^{(1)} + 1_{j_1} \in \langle\mathcal{N}\rangle_L$. Thus a multi index $\nu^{(2)} \in \mathcal{N}$ exists such that $\nu^{(2)}\,|_{L,\mathcal{N}}\,\nu^{(1)} + 1_{j_1}$.

By iteration of this argument we obtain a sequence $(\nu^{(1)}, \nu^{(2)}, \dots)$ where each element $\nu^{(i)}$ is a divisor of μ and where $\nu^{(i+1)}\,|_{L,\mathcal{N}}\,\nu^{(i)} + 1_{j_i}$ with a non-multiplicative index $j_i \in \bar{N}_{L,\mathcal{N}}(\nu^{(i)})$. This sequence cannot become infinite for a continuous division, as μ has only finitely many different divisors and all the multi indices $\nu^{(i)}$ in arbitrary finite pieces of the sequence must be different. But the sequence can only stop, if some $\nu^{(i)} \in \mathcal{N}$ exists such that $\nu^{(i)}\,|_{L,\mathcal{N}}\,\mu$ and hence $\mu \in \langle\mathcal{N}\rangle_L$.

Thus *any* involutive completion $\mathcal{N}_L$ of the set $\mathcal{N}$ must contain the multi index $\nu + 1_j$. In the next iteration of the `repeat` loop, Algorithm 6.2 treats the enlarged set $\mathcal{N}_1 = \mathcal{N} \cup \{\nu + 1_j\}$. As any involutive completion $\mathcal{N}_L$ of $\mathcal{N}$ is also an involutive completion of $\mathcal{N}_1$, we may apply the same argument again.

As a completion $\mathcal{N}_L$ is by definition a finite set, we must reach after a finite number k of iterations an involutive basis $\mathcal{N}_k$ of $\langle\mathcal{N}\rangle$. $\square$

Note the crucial difference between this result and Theorem B.2.13 on the termination of the Buchberger algorithm for the construction of Gröbner bases. In the latter case, we can show the termination for arbitrary input, i.e. the theorem provides a constructive proof for the existence of such a basis. Here we are only able to prove the termination under the assumption that a finite involutive basis exists; the existence has to be shown separately. For example, Lemma 2.1.14 guarantees us that any monoid ideal possesses a finite Janet basis.

An interesting aspect of the proof above is that it shows that our completion algorithm is largely independent of the term order $\prec$. It affects only the order in which multi indices are added but not which or how many multi indices are added during the completion. We may even use different term orders at each step, as for the proof it is only important that we always add a multi index that is minimal in $\mathcal{S}$ with respect to *some* term order.

Corollary 6.4.2. *If Algorithm 6.2 terminates, its output $\tilde{\mathcal{N}}$ is independent of the chosen term order $\prec$. This holds even, if the term order is changed in every iteration of the* `repeat` *loop. Furthermore, if $\mathcal{N}_L$ is any involutive completion of $\mathcal{N}$ with respect to the division L, then $\tilde{\mathcal{N}} \subseteq \mathcal{N}_L$.*

Proof. Consider the set $\mathcal{L}(\mathcal{N})$ of all involutive completions of $\mathcal{N}$ with respect to the division L and define

$$\tilde{\mathcal{N}} = \bigcap_{\mathcal{N}_L \in \mathcal{L}(\mathcal{N})} \mathcal{N}_L\,. \tag{6.11}$$

We claim that Algorithm 6.2 determines this set $\tilde{\mathcal{N}}$ independent of the used term order. Obviously, this implies our corollary.

In our proof of Proposition 6.4.1, we have shown that all the multi indices added in the course of Algorithm 6.2 are contained in *any* involutive completion of $\mathcal{N}$. Thus all these multi indices are elements of $\tilde{\mathcal{N}}$. As our algorithm terminates with an involutive completion, its output is an element of $\mathcal{L}(\mathcal{N})$, too. Hence the output must be $\tilde{\mathcal{N}}$. $\square$

Recall that for arbitrary (monomial or polynomial) ideals a basis $\mathcal{F}$ is called *minimal*, if it is not possible to remove an element of $\mathcal{F}$ without losing the property that we have a basis. A similar notion can naturally be introduced for involutive bases.

Definition 6.4.3. *Let $\mathcal{I} \subseteq \mathbb{N}_0^n$ be a monoid ideal and L an involutive division. An involutive basis $\mathcal{N}$ of $\mathcal{I}$ with respect to L is called* minimal*, if any other involutive basis $\mathcal{N}'$ of $\mathcal{I}$ with respect to L satisfies $\mathcal{N}' \subseteq \mathcal{N}$.*

Any monomial ideal has a *unique* minimal monomial basis: take an arbitrary monomial basis and eliminate all terms having a divisor in the basis. Obviously, these eliminations do not change the span and the result is a minimal basis. Corollary 6.4.2 implies similarly that if a monoid ideal $\mathcal{I} \subseteq \mathbb{N}_0^n$ has a finite involutive basis for a given division L, then a unique minimal involutive basis exists. It can easily be constructed by taking the unique minimal basis of $\mathcal{I}$ in the usual sense and applying Algorithm 6.2 to it.

A trivial method to compute an involutive basis for an ideal $\mathcal{I}$ in a polynomial algebra $(\mathcal{P}, \star, \prec)$ of solvable type follows immediately from Remark 2.3.13: we determine first a Gröbner basis $\mathcal{G}$ of $\mathcal{I}$ and then with Algorithm 6.2 the involutive completion of the monomial set $\mathrm{lt}_\prec \mathcal{G}$. In fact, a similar approach is proposed by Sturmfels and White [173] for the construction of Stanley decompositions. However, we prefer to extend the ideas behind Algorithm 6.2 to a direct completion algorithm for polynomial ideals, as we believe that this is more efficient. As a first step, we need subalgorithms for two important tasks: *involutive normal forms* and *involutive head autoreductions*.

The design of an algorithm $\texttt{NormalForm}_{L,\prec}(g, \mathcal{H})$ determining an involutive normal form of the polynomial g with respect to the finite set $\mathcal{H} \subset \mathcal{P}$ is trivial. We may use the standard Algorithm B.1 for normal forms in the Gröbner theory (see p. 254), if we replace in Lines /1/, /5/ and /7/ of it the ordinary divisibility relation by its involutive version. Obviously, this does not affect the termination. Actually, for our purposes it is not necessary to compute a full normal form; we may stop as soon as we have obtained a polynomial that is not involutively head reducible.

A simple realisation of an algorithm $\texttt{InvHeadAutoReduce}_{L,\prec}(\mathcal{F})$ for an involutive head autoreduction is given by Algorithm 6.3. Its correctness is obvious and the termination follows again from the fact that $\prec$ is a well-order, as after each reduction step one element of $\mathcal{H}$ has a smaller leading term than before. For a full autoreduction instead of only a head autoreduction, one must consider in Line /2/ the full support of h and not only the leading term. In this case it is tempting to replace Lines /5-7/ by $\bar{h} \leftarrow \texttt{NormalForm}_{L,\prec}(h, \bar{\mathcal{H}})$. However, this is correct only for global divisions, as then the multiplicative variables are determined with respect to the set $\bar{\mathcal{H}}$ and not $\mathcal{H}$ (cf. Remark 2.3.3).

Note that in the determination of the coefficient $c \in \Bbbk$ in Line /6/ we must divide by the leading coefficient of the product $x^\mu \star f$. In general, this differs from the leading coefficient of f, as one can see from (2.5). However, for the two cases of greatest interest for us, ordinary polynomials and linear differential operators, we may take $\mathrm{lc}_\prec f$ instead.

Algorithm 6.3 Involutive Head Autoreduction in $(\mathcal{P}, \star, \prec)$

Input: a finite set $\mathcal{F} \subset \mathcal{P}$, an involutive division L
Output: an involutively head autoreduced set $\mathcal{H}$ with $\langle\mathcal{H}\rangle = \langle\mathcal{F}\rangle$

$\mathcal{H} \leftarrow \mathcal{F}$
while $\exists h \in \mathcal{H}, f \in \mathcal{H} \setminus \{h\} : \mathrm{le}_\prec f \mid_{L,\mathrm{le}_\prec \mathcal{H}} \mathrm{le}_\prec h$ **do**
  choose such a pair (h, f)
  $\bar{\mathcal{H}} \leftarrow \mathcal{H} \setminus \{h\}$
  $\mu \leftarrow \mathrm{le}_\prec h - \mathrm{le}_\prec f$
  $c \leftarrow \mathrm{lc}_\prec h / \mathrm{lc}_\prec (x^\mu \star f)$
  $\bar{h} \leftarrow h - c x^\mu \star f$
  if $\bar{h} = 0$ **then**
    $\mathcal{H} \leftarrow \bar{\mathcal{H}}$
  else
    $\mathcal{H} \leftarrow \bar{\mathcal{H}} \cup \{\bar{h}\}$
  end_if
end_while
return $\mathcal{H}$

Based on these two subalgorithms, we propose now Algorithm 6.4 for the computation of involutive bases in $\mathcal{P}$. It follows the same strategy as the monomial algorithm. We multiply each element of the set by its non-multiplicative variables. Then we look whether the result is already contained in the involutive span of the set; if not, it is added to the set. The check whether an element lies in the involutive span is performed via an involutive normal form computation; for each element of the span the involutive normal form is zero. In order to obtain reasonable multiplicative variables, we must take care that our set is always involutively head autoreduced.

The way in which we select in Line /5/ the next polynomial $\bar{g}$ to be treated is more or less identical with the so-called normal selection strategy in the theory of Gröbner bases (see Appendix B.2). There this strategy is known to work very well for degree compatible term orders but not so well for other orders like the purely lexicographic one. Here we will see below that the normal selection strategy is important for proving the termination of the algorithm. It is currently unclear to what extent other strategies may be used for the selection of $\bar{g}$.

In order to prove the correctness and the termination of Algorithm 6.4 we need a generalisation of the notion of local involution and of Proposition 6.3.4 to polynomial sets.

Definition 6.4.4. *The finite set $\mathcal{F} \subset \mathcal{P}$ is* locally involutive *for the division L, if $x^j \star f \in \langle\mathcal{F}\rangle_{L,\prec}$ for any non-multiplicative variable $x^j \in \bar{X}_{L,\mathcal{F},\prec}(f)$ of any polynomial $f \in \mathcal{F}$.*

Algorithm 6.4 Completion in $(\mathcal{P}, \star, \prec)$

Input: a finite set $\mathcal{F} \subset \mathcal{P}$, an involutive division L
Output: an involutive basis $\mathcal{H}$ of $\mathcal{I} = \langle \mathcal{F} \rangle$ with respect to L and $\prec$

$\mathcal{H} \leftarrow \texttt{InvHeadAutoReduce}_{L,\prec}(\mathcal{F})$
loop
  $\mathcal{S} \leftarrow \{x^j \star h \mid h \in \mathcal{H},\ x^j \in \bar{X}_{L,\mathcal{H},\prec}(h),\ x^j \star h \notin \langle \mathcal{H} \rangle_{L,\prec}\}$
  repeat
    $\bar{g} \leftarrow \min_{\prec} \mathcal{S}$
    $\mathcal{S} \leftarrow \mathcal{S} \setminus \{\bar{g}\}$
    $g \leftarrow \texttt{NormalForm}_{L,\prec}(\bar{g}, \mathcal{H})$
  until $(g \neq 0)$ or $(\mathcal{S} = \emptyset)$
  if $g \neq 0$ **then**
    $\mathcal{H} \leftarrow \texttt{InvHeadAutoReduce}_{L,\prec}(\mathcal{H} \cup \{g\})$
  else
    return $\mathcal{H}$
  end_if
end_loop

Proposition 6.4.5. *Given a continuous division* L, *any involutively head autoreduced set* $\mathcal{F} \subset \mathcal{P}$ *that is locally involutive is also involutive.*

Proof. Local involution may be expressed as follows. Choose for an arbitrary polynomial $g \in \mathcal{F}$ a non-multiplicative variable $x^j \in \bar{X}_{L,\mathcal{F}}(g)$. Then for each $f \in \mathcal{F}$ a polynomial $P_f \in \mathbb{k}[X_{L,\mathcal{F},\prec}(f)]$ exists such that

$$x^j \star g = \sum_{f \in \mathcal{F}} P_f \star f \,. \tag{6.12}$$

As by assumption, $\mathcal{F}$ is involutively head autoreduced, (6.12) implies that the set $\operatorname{le}_\prec \mathcal{F} \subset \mathbb{N}_0^n$ is locally involutive and thus by the continuity of L involutive (Proposition 6.3.4). By Theorem 2.3.11, $\mathcal{F}$ is thus an involutive basis of $\mathcal{I} = \langle \mathcal{F} \rangle$, if we can show that $\langle \operatorname{le}_\prec \mathcal{F} \rangle = \operatorname{le}_\prec \mathcal{I}$.

We claim that for any term x^μ and for any polynomial $f_1 \in \mathcal{F}$ a polynomial $h \in \mathcal{F}$ exists such that $\operatorname{le}_\prec(x^\mu \star f_1) \in \mathcal{C}_{L,\operatorname{le}_\prec \mathcal{F}}(h)$. If this is true, we may take arbitrary linear combinations (with coefficients in $\mathbb{k}$) of such products and the leading terms cannot cancel, as $\operatorname{le}_\prec \mathcal{F}$ is a strong involutive basis of its span. But this implies immediately that the leading exponent of any polynomial in $\mathcal{I}$ lies in $\langle \operatorname{le}_\prec \mathcal{F} \rangle$ and the proposition holds.

If the term x^μ contains only variables that are multiplicative for $\operatorname{le}_\prec f_1$, nothing is to be shown. Otherwise we choose an index $j_1 \in \bar{N}_{L,\operatorname{le}_\prec \mathcal{F}}(\operatorname{le}_\prec f_1)$ such that $\mu_{j_1} > 0$. By (6.12), for each $f \in \mathcal{F}$ a polynomial $P_f^{(1)} \in \mathbb{k}[X_{L,\mathcal{F},\prec}(f)]$ exists such that $x^{j_1} \star f_1 = \sum_{f \in \mathcal{F}} P_f^{(1)} \star f$. Let the leading term on the right hand side be $\operatorname{lt}_\prec(P_{f_2}^{(1)} \star f_2)$. If the term $x^{\mu - 1_{j_1}}$ contains only variables that are multiplicative for $\operatorname{le}_\prec f_2$, we are done.

Otherwise we choose a non-multiplicative index $j_2 \in \bar{N}_{L,\mathrm{le}_\prec \mathcal{F}}(\mathrm{le}_\prec f_2)$ such that $(\mu - 1_{j_1})_{j_2} > 0$. Applying again (6.12), for each $f \in \mathcal{F}$ a polynomial $P_f^{(2)} \in \mathbb{k}[X_{L,\mathcal{F},\prec}(f)]$ exists such that $x^{j_2} \star f_2 = \sum_{f\in\mathcal{F}} P_f^{(2)} \star f$. Let the leading term on the right hand side be $\mathrm{lt}_\prec(P_{f_3}^{(2)} \star f_3)$ and introduce the multi index $\nu = \mathrm{le}_\prec(x^{j_1} \star f_1) - \mathrm{le}_\prec f_2$. Then $\mathrm{le}_\prec(x^\mu \star f_1) = \mathrm{le}_\prec(x^{\mu+\nu-1_{j_1}-1_{j_2}} \star f_3)$, as the multiplication $\star$ in a polynomial algebra of solvable type respects the term order $\prec$. If all variables on the right hand side are multiplicative, we are done.

Otherwise we iterate: we choose an index j_3 and decompose $x^{j_3} \star f_3$ with (6.12) into multiplicative products obtaining a new polynomial f_4 and so on. This process yields a sequence $\big(\nu^{(1)} = \mathrm{le}_\prec f_1, \nu^{(2)} = \mathrm{le}_\prec f_2, \dots\big)$ where all multi indices $\nu^{(k)}$ are elements of the finite set $\mathrm{le}_\prec \mathcal{F}$ and where to each $\nu^{(k)}$ a non-multiplicative index $j_k \in \bar{N}_{L,\mathrm{le}_\prec \mathcal{F}}(\nu^{(k)})$ exists such that $\nu^{(k+1)} |_{L,\mathrm{le}_\prec \mathcal{F}} \nu^{(k)} + 1_{j_k}$. By the definition of a continuous division, this sequence cannot become infinite and the above described process must terminate. But this implies, as claimed, the existence of a polynomial $h \in \mathcal{F}$ such that $\mathrm{le}_\prec(x^\mu \star f_1) \in \mathcal{C}_{L,\mathrm{le}_\prec \mathcal{F}}(h)$. □

As in the proof of Proposition 6.4.1, local involution of $\mathcal{H}$ is obviously equivalent to the termination condition $\mathcal{S} = \emptyset$ of the `repeat` loop in Algorithm 6.4. Thus we are now in the position to prove the following result.

Theorem 6.4.6. *Let* L *be a continuous and constructive involutive division and* $(\mathcal{P}, \star, \prec)$ *a polynomial algebra of solvable type. Assume that the (left) ideal* $\mathcal{I} \subseteq \mathcal{P}$ *is such that* $\mathrm{le}_\prec \mathcal{I}$ *possesses a finite involutive completion. Then Algorithm 6.4 terminates for any finite generating set* $\mathcal{F}$ *of* $\mathcal{I}$ *with an involutive basis.*

Proof. We begin by proving the *correctness* of the algorithm under the assumption that it terminates. The relation $\mathcal{I} = \langle \mathcal{H} \rangle$ remains valid throughout, although $\mathcal{H}$ changes. But the only changes are the addition of further elements of $\mathcal{I}$ and involutive head autoreductions; both operations do not affect the ideal generated by $\mathcal{H}$. When the algorithms terminates, we have $\mathcal{S} = \emptyset$ and thus the output $\mathcal{H}$ is locally involutive and by Proposition 6.4.5 involutive.

There remains the problem of *termination*. Algorithm 6.4 produces a sequence $(\mathcal{H}_1, \mathcal{H}_2, \dots)$ with $\langle \mathcal{H}_i \rangle = \mathcal{I}$. The set $\mathcal{H}_{i+1}$ is determined from $\mathcal{H}_i$ in Line /10/. We distinguish two cases, namely whether or not during the computation of the involutive normal form in Line /7/ the leading exponent changes. If $\mathrm{le}_\prec \bar{g} = \mathrm{le}_\prec g$, then $\langle \mathrm{le}_\prec \mathcal{H}_i \rangle = \langle \mathrm{le}_\prec \mathcal{H}_{i+1} \rangle$, as $\mathrm{le}_\prec g = \mathrm{le}_\prec h + 1_j$ for some $h \in \mathcal{H}_i$. Otherwise we claim that $\langle \mathrm{le}_\prec \mathcal{H}_i \rangle \subsetneq \langle \mathrm{le}_\prec \mathcal{H}_{i+1} \rangle$.

By construction, the polynomial g is in involutive normal form with respect to the set $\mathcal{H}_i$ implying that $\mathrm{le}_\prec g \in \langle \mathrm{le}_\prec \mathcal{H}_i \rangle \setminus \langle \mathrm{le}_\prec \mathcal{H}_i \rangle_L$. If we had $\langle \mathrm{le}_\prec \mathcal{H}_i \rangle = \langle \mathrm{le}_\prec \mathcal{H}_{i+1} \rangle$, a polynomial $h \in \mathcal{H}_i$ would exist such that

$\mathrm{le}_\prec g = \mathrm{le}_\prec h + \mu$ where the multi index μ has a non-vanishing entry μ_j for at least one non-multiplicative index $j \in \bar{N}_{L,\mathrm{le}_\prec \mathcal{H}_i}(h)$. This implies that $\mathrm{le}_\prec h + 1_j \preceq \mathrm{le}_\prec g \prec \mathrm{le}_\prec \bar{g}$. But we choose the polynomial $\bar{g}$ in Line /5/ such that its leading exponent is minimal among all non-multiplicative products $x^k \star h$ with $h \in \mathcal{H}_i$; hence $\mathrm{le}_\prec \bar{g} \preceq \mathrm{le}_\prec h + 1_j$. As this is a contradiction, we must have $\langle \mathrm{le}_\prec \mathcal{H}_i \rangle \subsetneq \langle \mathrm{le}_\prec \mathcal{H}_{i+1} \rangle$.

So the outer infinite loop generates an ascending chain of monoid ideals $\langle \mathrm{le}_\prec \mathcal{H}_1 \rangle \subseteq \langle \mathrm{le}_\prec \mathcal{H}_2 \rangle \subseteq \cdots \subseteq \mathrm{le}_\prec \mathcal{I}$. As $\mathbb{N}_0^n$ is Noetherian, the chain must become stationary at some index N. It follows from the considerations above that in all iterations of the outer loop after the Nth one $\mathrm{le}_\prec \bar{g} = \mathrm{le}_\prec g$ in Line /7/. At this stage Algorithm 6.4 reduces to an involutive completion of the set $\mathrm{le}_\prec \mathcal{H}_N$ using Algorithm 6.2. Indeed, in Line /5/ we choose the polynomial $\bar{g}$ such that $\mathrm{le}_\prec \bar{g}$ is the same multi index as Algorithm 6.2 adds in Line /4/. By Proposition 6.4.1, the latter algorithm terminates under the made assumptions and hence Algorithm 6.4 terminates, too. □

Corollary 6.4.7. *Let* L *be a Noetherian division. Then every ideal* $\mathcal{I} \subseteq \mathcal{P}$ *possesses a finite involutive basis with respect to the division* L*. Furthermore, in suitably chosen variables* $\mathbf{x}$ *every ideal has a finite Pommaret basis, if the term order* $\prec$ *is class respecting.*

Proof. The first part is trivial, as for a Noetherian division by definition any monoid ideal possesses a finite involutive basis. Thus the conditions of Theorem 6.4.6 are always satisfied. For the second part we invoke the results of Section 2.4 on δ-regular coordinate systems. As a first step we show that every ideal has a *Pommaret* head autoreduced *Janet* basis. Indeed, let us apply Algorithm 6.4 for the Janet division with a slight modification: in the Lines /2/ and /10/ we perform the involutive head autoreductions with respect to the Pommaret division. It is obvious that if the algorithm terminates, the result is a basis with the wanted properties.

The Janet division is Noetherian by Lemma 2.1.14. Thus without our modification the termination is obvious. With respect to the Janet division every set of multi indices is involutively autoreduced. Hence a Janet head autoreduction only takes care that no two elements of a set have the same leading exponents (recall Example 2.3.5). But in Line /10/ we add a polynomial that is in involutive normal form so that no involutive head reductions are possible. As the Pommaret head autoreduction may only lead to a larger monoid ideal $\mathrm{le}_\prec \mathcal{H}_i$, the Noetherian argument in the proof of Theorem 6.4.6 remains valid after our modification.

Once the ascending ideal chain $\langle \mathrm{le}_\prec \mathcal{H}_1 \rangle \subseteq \langle \mathrm{le}_\prec \mathcal{H}_2 \rangle \subseteq \cdots$ has become stationary, Algorithm 6.4 reduces to the "monomial" Algorithm 6.2. According to Corollary 2.4.8, the Pommaret head autoreductions may only increase the

Janet spans $\langle \mathrm{le}_{\prec}\mathcal{H}_i\rangle_J$. Thus the termination of Algorithm 6.2 is not affected by our modification and the algorithm terminates for arbitrary input.

By Proposition 2.4.6, the coordinate systems that are δ-singular for a given finite set $\mathcal{F} \subset \mathcal{P}$ form a Zariski closed set. As our modified algorithm terminates, it treats only a finite number of sets $\mathcal{H}_i$ and the coordinate systems that are δ-singular for at least one of them still form a Zariski closed set. Thus generic coordinates are δ-regular for all sets $\mathcal{H}_i$ and by Corollary 2.4.11 their Janet and their Pommaret spans coincide for a class respecting term order. But this implies that the result of the modified algorithm is not only a Janet but also a Pommaret basis. □

Theorem 2.4.10 provides us with an effective mean to detect that the completion to a Pommaret basis does *not* terminate. We follow Algorithm 6.4 with the Pommaret division (and a class respecting term order) but check after each iteration whether $|\mathcal{H}_i|_{J,\prec} = |\mathcal{H}_i|_{P,\prec}$. If no finite Pommaret basis exists in the given coordinates, sooner or later the Janet size is greater than the Pommaret size. Indeed, by the considerations above the completion with respect to the Janet division terminates. Thus either the result is simultaneously a Pommaret basis or our criterion must detect a variable which is multiplicative only for the Janet but not for the Pommaret division.

Example 6.4.8. Let us apply Algorithm 6.4 to the Pommaret completion of the set $\mathcal{F} = \{\underline{z^2} + yz - x^2,\ \underline{xz} + xy - z,\ \underline{yz} + y^2 + x^2 - x\}$ (with respect to the degree reverse lexicographic order). We have seen in Example 2.4.12 that the coordinates are not δ-regular, although the Janet and the Pommaret span of $\mathcal{F}$ coincide. According to our algorithm we must first analyse the polynomial $y(xz + xy - z)$. Its involutive normal form with respect to $\mathcal{F}$ is $-\underline{x^3}+y^2+2x^2-x$. If we determine the multiplicative variables for the enlarged set, they do not change for the old elements. For the new polynomial the Janet division yields $\{x, y\}$. But y is obviously not multiplicative for the Pommaret division. Thus our criterion tells us that the Pommaret completion will not terminate. Indeed it is easy to see that no matter how often we multiply the new polynomial by y, it will never become involutively head reducible and no finite Pommaret basis exists.

The Janet completion (with Pommaret autoreductions) yields for this example the following involutive basis

$$\begin{aligned}\mathcal{F}_J = \{&\underline{z^2} - y^2 - 2x^2 + x,\ \underline{xz} + xy - z,\ \underline{yz} + y^2 + x^2 - x,\\ &\underline{y^2} - x,\ \underline{x^3} - 2x^2 - x + y,\ \underline{x^2y} - xy + z\}\ .\end{aligned} \tag{6.13}$$

Without the Pommaret autoreductions the basis is even larger. In Example 2.4.12 we showed that the transformation $\tilde{x} = z$, $\tilde{y} = y + z$ and $\tilde{z} = x$ yields after an autoreduction the set $\tilde{\mathcal{F}}^{\triangle} = \{\underline{\tilde{z}^2} - \tilde{x}\tilde{y},\ \underline{\tilde{y}\tilde{z}} - \tilde{x},\ \underline{\tilde{y}^2} - \tilde{z}\}$. One easily checks that it is a Pommaret and thus also a Janet basis. We see again that other divisions also "feel" δ-singularity in the sense that in such coordinates they typically lead to larger bases of higher degree. For efficiency reasons one should therefore always try to work in δ-regular coordinates. $\triangleleft$

Our proof of Theorem 6.4.6 has an interesting consequence. Assume that the term order $\prec$ is of type ω, i. e. for any two multi indices μ, ν with $\mu \prec \nu$ only finitely many multi indices $\rho^{(i)}$ exist with $\mu \prec \rho^{(1)} \prec \rho^{(2)} \prec \cdots \prec \nu$. Then even if our algorithm does *not* terminate, it determines in a finite number of steps a Gröbner basis of the ideal $\mathcal{I}$.

Proposition 6.4.9. *Let the term order $\prec$ be of type ω. Then Algorithm 6.4 determines for any input $\mathcal{F}$ in a finite number of steps a Gröbner basis of the ideal $\mathcal{I} = \langle\mathcal{F}\rangle$.*

Proof. In the proof of Theorem 6.4.6 we introduced the set $\mathcal{H}_N$ such that $\langle\mathcal{H}_{N+\ell}\rangle = \langle\mathcal{H}_N\rangle$ for all $\ell > 0$. We claim that $\mathcal{H}_N$ is a Gröbner basis of $\mathcal{I}$.

Let $f \in \mathcal{I}$ be an arbitrary element of the ideal. As $\mathcal{H}_N$ is a basis of $\mathcal{I}$, we find for each $h \in \mathcal{H}_N$ a polynomial $g_h \in \mathcal{P}$ such that

$$f = \sum_{h \in \mathcal{H}_N} g_h \star h\,. \tag{6.14}$$

By Proposition B.2.7, $\mathcal{H}_N$ is a Gröbner basis, if and only if we can choose the coefficients g_h such that $\mathrm{lt}_\prec(g_h \star h) \preceq \mathrm{lt}_\prec f$. Assume that for f no such standard representation exists and let $\mu = \max_{h\in\mathcal{H}_N}\{\mathrm{le}_\prec g_h + \mathrm{le}_\prec h\}$. If we denote by $\bar{\mathcal{H}}_N$ the set of all polynomials $\bar{h} \in \mathcal{H}_N$ for which $\mathrm{le}_\prec g_{\bar{h}} + \mathrm{le}_\prec \bar{h} = \mu$, then we must have a non-trivial syzygy $\sum_{\bar{h}\in\bar{\mathcal{H}}_N}(\mathrm{le}_\prec g_{\bar{h}} + \mathrm{le}_\prec \bar{h}) = 0$. By Lemma 2.6.1, at least one $\bar{h} \in \bar{\mathcal{H}}_N$ exists such that some non-multiplicative variable $x^j \in \bar{X}_{L,\mathcal{H}_N}(\bar{h})$ divides $\mathrm{lt}_\prec g_{\bar{h}}$.

As $\prec$ is of type ω, after a finite number of steps the non-multiplicative product $x^j \star \bar{h}$ is analysed in Algorithm 6.4. Thus for some $n_1 \geq 0$ the set $\mathcal{H}_{N+n_1}$ contains an element $\bar{h}'$ with $\mathrm{le}_\prec \bar{h}' = \mathrm{le}_\prec(x^j \star \bar{h})$. Let $\mu = \mathrm{le}_\prec g_{\bar{h}}$, $x^{\mu-1_j} \star x^j = cx^\mu + r_1$ and $\bar{h}' = dx^j \star \bar{h} + r_2$. Then we may rewrite

$$g_{\bar{h}} \star \bar{h} = \frac{\mathrm{lc}_\prec g_{\bar{h}}}{cd}\bigl[x^{\mu-1_j} \star (\bar{h}' - r_2) - dr_1 \star \bar{h}\bigr] + \bigl(g_{\bar{h}} - \mathrm{lm}_\prec g_{\bar{h}}\bigr) \star \bar{h}\,. \tag{6.15}$$

As $\bar{h}'$ was determined via an involutive normal form computation applied to $x^j \star \bar{h}$ and as we know that at this stage of the algorithm the leading term does not change during the computation, the leading term on the right hand side of (6.15) is $\mathrm{lt}_\prec(x^{\mu-1_j} \star \bar{h}')$. If the term $x^{\mu-1_j}$ contains a non-multiplicative variable $x^k \in \bar{X}_{L,\mathcal{H}_{N+n_1}}(\bar{h}')$, we repeat the argument obtaining an element $\bar{h}'' \in \mathcal{H}_{N+n_1+n_2}$ such that $\mathrm{le}_\prec \bar{h}'' = \mathrm{le}_\prec(x^k \star \bar{h}')$.

Obviously, this process terminates after a finite number of steps — even if we do it for each $\bar{h} \in \bar{\mathcal{H}}_N$. Thus after $N + \ell$ iterations (for some finite number $\ell \geq 0$) we obtain a set $\mathcal{H}_{N+\ell}$ such that, after applying all the found relations (6.15), f can be expressed in the form $f = \sum_{h \in \mathcal{H}_{N+\ell}} \tilde{g}_h \star h$ where still $\mu = \max_{h \in \mathcal{H}_{N+\ell}} \{\mathrm{le}_\prec \tilde{g}_h + \mathrm{le}_\prec h\}$. Denote again by $\bar{\mathcal{H}}_{N+\ell} \subseteq \mathcal{H}_{N+\ell}$ the set of all those polynomials $\bar{h}$ for which this maximum is achieved.

By construction, no term $\mathrm{lt}_\prec \tilde{g}_{\bar{h}}$ with $\bar{h} \in \bar{\mathcal{H}}_{N+\ell}$ contains a variable that is non-multiplicative for $\bar{h}$. Hence, by Lemma 2.6.1, it is not possible that $\sum_{\bar{h} \in \bar{\mathcal{H}}_{N+\ell}} (\mathrm{le}_\prec \tilde{g}_{\bar{h}} + \mathrm{le}_\prec \bar{h}) = 0$. But this contradicts $\mu \succ \mathrm{lt}_\prec f$. Thus each polynomial $f \in \mathcal{P}$ possesses a standard representation already with respect to $\mathcal{H}_N$ and this set is a Gröbner basis. □

Remark 6.4.10. All degree compatible term orders are of type ω. Purely lexicographic orders are not of type ω. Indeed, if $x \prec_{\mathrm{lex}} y$, then the infinitely many terms x^k are all less than y with respect to $\prec_{\mathrm{lex}}$. ◁

The notion of a minimal involutive basis may be extended to ideals in $(\mathcal{P}, \star, \prec)$ in close analogy to Definition B.2.10 of a minimal Gröbner basis.

Definition 6.4.11. *Let $\mathcal{I} \subseteq \mathcal{P}$ be an ideal and L an involutive division. An involutive basis $\mathcal{H}$ of $\mathcal{I}$ with respect to L is* minimal*, if $\mathrm{lt}_\prec \mathcal{H}$ is a minimal involutive basis of $\mathrm{lt}_\prec \mathcal{I}$ for L.*

Uniqueness requires now two additional assumptions. First of all, our definition of an involutive basis requires only that it is involutively head autoreduced; for uniqueness we obviously need a full involutive autoreduction. Secondly, we must normalise the leading coefficients to one, i. e. we must take a *monic* basis.

Proposition 6.4.12. *Let $\mathcal{I} \subseteq \mathcal{P}$ be an ideal and L an involutive division. Then $\mathcal{I}$ has at most one monic, involutively autoreduced, minimal involutive basis with respect to L.*

Proof. Assume that $\mathcal{H}_1$ and $\mathcal{H}_2$ are two different monic, involutively autoreduced, minimal involutive bases of $\mathcal{I}$ with respect to L and $\prec$. By definition of a minimal involutive bases, this implies that $\mathrm{lt}_\prec \mathcal{H}_1 = \mathrm{lt}_\prec \mathcal{H}_2$. As $\mathcal{H}_1$ and $\mathcal{H}_2$ are not identical, we must have two polynomials $h_1 \in \mathcal{H}_1$ and $h_2 \in \mathcal{H}_2$ such that $\mathrm{lt}_\prec h_1 = \mathrm{lt}_\prec h_2$ but $h_1 \neq h_2$. Now consider the polynomial $h = h_1 - h_2 \in \mathcal{I}$. Its leading term must lie in the involutive span of $\mathrm{lt}_\prec \mathcal{H}_1 = \mathrm{lt}_\prec \mathcal{H}_2$. On the other hand, $\mathrm{lt}_\prec h \in \operatorname{supp} h_1 \cup \operatorname{supp} h_2$. But this implies that either $\mathcal{H}_1$ or $\mathcal{H}_2$ is not involutively autoreduced. □

In contrast to the monomial case, one does not automatically obtain a minimal involutive basis by making some minor modifications of Algorithm 6.4. In particular, it does not suffice to apply it to a minimal basis in the ordinary sense. Gerdt and Blinkov [57] presented an algorithm that always returns a minimal involutive basis provided a finite involutive basis exists. While it still follows the same basic strategy of study all products with non-multiplicative variables, it requires a more complicated organisation of the algorithm. We omit here the details.

6.5 Linear Differential Equations

We stressed in Section 6.2 that the Cartan–Kuranishi completion is more a "method" than an algorithm, as it is not clearly specified how the required operations should be performed. We now develop for linear differential equations an effective algebraic realisation of it based on Algorithm 6.4 for involutive bases. A naive realisation of the Cartan–Kuranishi completion is immediate: one explicitly determines local representations for all differential equations $\mathcal{R}^{(s)}_{q+r}$ and symbols $\mathcal{M}^{(s)}_{q+r}$ appearing in the course of the algorithm. For linear or polynomial equations it is – at least in principle – straightforward to perform all required operations algorithmically using Gröbner bases.

Obviously, such a brute force approach is not very efficient; in general, many unnecessary prolongations (which are usually rather expensive to compute) are performed and one has to deal with rather large matrices (recall Example 6.2.2). Pure algebra in form of involutive bases leads to fairly fast algorithms for linear equations, but all geometric information is lost. For example, it is not possible to determine the number of performed prolongations and projections. We combine the geometric and the algebraic approach into a "hybrid" algorithm, so that we maintain the efficiency of pure algebra and still obtain the full geometric information.

In the sequel we assume that the coefficients of our system stem from a function field $\mathbb{k}$ where all required operations (arithmetics and differentiations) can be performed effectively and that contains the coordinates $\mathbf{x}$. Thus a typical instance of $\mathbb{k}$ is a finite algebraic extension of the field $\mathbb{Q}(x^1,\dots,x^n)$ of rational functions with rational coefficients.

We assume that the linear system is given in the form

$$\Phi^\tau(x,p) = A^{\tau\mu}_\alpha(x)p^\alpha_\mu = 0\,, \qquad \tau = 1,\dots,t \tag{6.16}$$

where $0 \leq |\mu| \leq q$. Thus we are only considering homogeneous equations. But of course one could add right hand sides as discussed in Section 3.4. For the remainder of this section, we concentrate on the left hand side and call it a *row*. Thus a linear system is for us a set $\mathcal{F}$ of rows; one may also say that we are operating on a matrix representation of the system.

The derivatives are ordered according to the degree reverse lexicographic order $\prec_{\text{degrevlex}}$; the leading term of a row $f \in \mathcal{F}$ is called *leading derivative* and denoted by[4] $\operatorname{ld} f$. As usual, multiplicative and non-multiplicative variables are assigned to a row according to its leading derivative. Reductions correspond now to elementary row operations and an autoreduction yields a triangular form of the system. We introduce the following notation for the result of a reduction. Let $f = a(x) \operatorname{ld} f + \sum_{p^\alpha_\mu \prec \operatorname{ld} f} A^\mu_\alpha(x) p^\alpha_\mu$ and $g = b(x) \operatorname{ld} g + \sum_{p^\alpha_\mu \prec \operatorname{ld} g} B^\mu_\alpha(x) p^\alpha_\mu$ be two rows of the system $\mathcal{F}$ with $\operatorname{ld} f = \operatorname{ld} g$; then

$$\operatorname{rem}(f, g) = \sum_{p^\alpha_\mu \prec \operatorname{ld} f} \left(A^\mu_\alpha(x) - \frac{a(x)}{b(x)} B^\mu_\alpha(x) \right) p^\alpha_\mu \,. \tag{6.17}$$

Note that this is a purely algebraic operation; no differentiations are allowed!

In the naive realisation of the Cartan–Kuranishi completion it is necessary to differentiate every equation in the system with respect to every independent variables in order to obtain a representation of the prolonged equation. We use the Pommaret division to distinguish multiplicative and non-multiplicative variables. Our goal is to avoid as much as possible to perform differentiations with respect to multiplicative variables. As in Algorithm 6.4 they should appear only in involutive normal form computations. In order to achieve this goal and still be able to decide involution of the appearing symbols and differential equations, we need a careful book-keeping. The necessary tools are introduced in the next definition. It might appear somewhat strange at first sight, but it will prove its usefulness in the sequel.

Definition 6.5.1. *Let* f *denote a row from a linear system* $\mathcal{F}$*. We can make* f *into an* indexed row*, either a* single indexed row $f_{(k)}$ *or a* double indexed row *(or* phantom row*)* $f_{(k,\ell)}$ *by assigning to it one respectively two non-negative integers* k *and* ℓ *where* $0 \leq k < \ell$ *in the second case. We call* k *the* initial level *and* ℓ *the* phantom level *of the indexed row. For a* global level $\lambda \geq k$*, we define the* truncated involutive cone $\bar{\mathcal{C}}_\lambda$ *of an indexed row to be the following set of rows:*

$$\bar{\mathcal{C}}_\lambda(f_{(k)}) = \left\{ D_\mu f \mid 0 \leq |\mu| \leq (\lambda - k);\ \forall i > \operatorname{cls} f : \mu_i = 0 \right\} , \tag{6.18a}$$

$$\bar{\mathcal{C}}_\lambda(f_{(k,\ell)}) = \left\{ D_\mu f \mid (\lambda - \ell) < |\mu| \leq (\lambda - k);\ \forall i > \operatorname{cls} f : \mu_i = 0 \right\} . \tag{6.18b}$$

The truncated involutive cone of an indexed row determined by the numbers λ, k and possibly ℓ contains all multiplicative prolongations whose orders satisfy the given conditions. With the help of this notion, we introduce the basic object of our approach, the skeleton of a linear system.

[4] As we have now fixed one specific term order, we omit for notational simplicity the subscript $\prec$.

Definition 6.5.2. *A* skeleton $\mathcal{S}_\lambda$ *with global level* λ *is the union of a set* $\mathcal{U}_\lambda = \{f^\sigma_{(k_\sigma)}\}$ *of single indexed rows and a set* $\mathcal{W}_\lambda = \{f^\sigma_{(k_\sigma, l_\sigma)}\}$ *of double indexed rows such that for all* σ *the inequalities* $k_\sigma \leq \lambda$ *and* $k_\sigma < l_\sigma \leq \lambda$, *respectively, hold. It defines the linear system*

$$\bar{\mathcal{S}}_\lambda = \left\{\bar{\mathcal{C}}_\lambda(f_{(k)}) \mid f_{(k)} \in \mathcal{U}_\lambda\right\} \cup \left\{\bar{\mathcal{C}}_\lambda(f_{(k,\ell)}) \mid f_{(k,\ell)} \in \mathcal{W}_\lambda\right\} . \tag{6.19}$$

Thus in order to obtain the linear system $\bar{\mathcal{S}}_\lambda$ corresponding to a given skeleton $\mathcal{S}_\lambda$, we must simply add all permitted multiplicative prolongations of the contained rows and drop the indices.

In each iteration of the Cartan–Kuranishi completion, the system is *prolonged* one order. Since we are interested in autoreduced systems, we must afterwards perform a *triangulation*. We consider now how these two operations are carried out efficiently on the level of skeletons. We may assume that the system $\mathcal{F}$ to be completed is already given in triangular, i.e. (head) autoreduced, form. We turn $\mathcal{F}$ into a skeleton $\mathcal{S}_0$ by simply setting the initial level k_σ of every row f^σ to 0. The global level λ is always equal to the number of the current iteration step; if a new row enters the skeleton in this step, its initial level becomes λ. Double indexed rows arise, if during the triangulation process a row is eliminated which has already produced some multiplicative prolongations. In this case, the phantom level is set to the current global level. Thus the completion algorithm produces a sequence of skeletons

$$\mathcal{S}_0 = \mathcal{S}_0^\triangle \longrightarrow \mathcal{S}_1 \longrightarrow \mathcal{S}_1^\triangle \longrightarrow \mathcal{S}_2 \longrightarrow \mathcal{S}_2^\triangle \longrightarrow \mathcal{S}_3 \longrightarrow \cdots \tag{6.20}$$

where the $^\triangle$ indicates a skeleton in triangular form.

Algorithm 6.5 Prolongation of a Skeleton

Input: skeleton $\mathcal{S}_\lambda = \mathcal{U}_\lambda \cup \mathcal{W}_\lambda$, number n of independent variables
Output: prolonged skeleton $\mathcal{S}_{\lambda+1} = \mathcal{U}_{\lambda+1} \cup \mathcal{W}_{\lambda+1}$
$\mathcal{U}_{\lambda+1} \leftarrow \mathcal{U}_\lambda$
$\mathcal{W}_{\lambda+1} \leftarrow \mathcal{W}_\lambda$
for all $f^\sigma_{(\lambda)} \in \mathcal{U}_\lambda$ **do**
  for $\operatorname{cls} f^\sigma_{(\lambda)} \leq i \leq n$ **do**
    $\mathcal{U}_{\lambda+1} \leftarrow \mathcal{U}_{\lambda+1} \cup \left\{(D_i f^\sigma)_{(\lambda+1)}\right\}$
  end_for
end_for
return $\mathcal{U}_{\lambda+1} \cup \mathcal{W}_{\lambda+1}$

Algorithm 6.5 merely computes the non-multiplicative prolongations of those single indexed rows that have been newly added in the previous iteration. Due to the increase of the global level multiplicative prolongations of *all* rows from the skeleton are automatically present in the linear system it defines. No rows are explicitly removed; the new global level causes the elimination of the least-order multiplicative prolongations of all phantom rows in the system. We will show below that these operations yield the desired skeleton for the prolonged system.

Algorithm 6.6 Triangulation of a Skeleton

Input: skeleton $\mathcal{S}_\lambda = \mathcal{U}_\lambda \cup \mathcal{W}_\lambda$
Output: equivalent skeleton $\mathcal{S}^{\triangle}_\lambda = \mathcal{U}^{\triangle}_\lambda \cup \mathcal{W}^{\triangle}_\lambda$ such that $\bar{\mathcal{S}}^{\triangle}_\lambda$ is in triangular form
/1/ $\mathcal{U}^{\triangle}_\lambda \leftarrow \emptyset$
/2/ $\mathcal{W}^{\triangle}_\lambda \leftarrow \mathcal{W}_\lambda$
/3/ **while** $\mathcal{U}_\lambda \neq \emptyset$ **do**
/4/ $\quad f^{\sigma}_{(k_\sigma)} \leftarrow \max_{\prec} \mathcal{U}_\lambda$
/5/ $\quad \bar{f} \leftarrow f^{\sigma}$
/6/ $\quad \mathcal{U}_\lambda \leftarrow \mathcal{U}_\lambda \setminus \{f^{\sigma}_{(k_\sigma)}\}$
/7/ $\quad$ **while** $\exists f^{\tau} \in \mathcal{S}_\lambda : \operatorname{ld} \bar{f} \in \bar{\mathcal{C}}_\lambda(\operatorname{ld} f^{\tau})$ **do**
/8/ $\qquad p^{\alpha}_{\mu} \leftarrow \operatorname{ld} \bar{f}$
/9/ $\qquad p^{\alpha}_{\nu} \leftarrow \operatorname{ld} f^{\tau}$
/10/ $\qquad g \leftarrow D_{\mu-\nu} f^{\tau}$
/11/ $\qquad \bar{f} \leftarrow \operatorname{rem}(\bar{f}, g)$
/12/ $\quad$ **end_while**
/13/ $\quad$ **if** $\bar{f} = f$ **then**
/14/ $\qquad \mathcal{U}^{\triangle}_\lambda \leftarrow \mathcal{U}^{\triangle}_\lambda \cup \{f^{\sigma}_{(k_\sigma)}\}$
/15/ $\quad$ **else**
/16/ $\qquad$ **if** $\bar{f} \neq 0$ **then**
/17/ $\qquad\quad \mathcal{U}^{\triangle}_\lambda \leftarrow \mathcal{U}^{\triangle}_\lambda \cup \{\bar{f}_{(\lambda)}\}$
/18/ $\qquad$ **end_if**
/19/ $\qquad$ **if** $k_\sigma < \lambda$ **then**
/20/ $\qquad\quad \mathcal{W}^{\triangle}_\lambda \leftarrow \mathcal{W}^{\triangle}_\lambda \cup \{f^{\sigma}_{(k_\sigma,\lambda)}\}$
/21/ $\qquad$ **end_if**
/22/ $\quad$ **end_if**
/23/ **end_while**
/24/ **return** $\mathcal{U}^{\triangle}_\lambda \cup \mathcal{W}^{\triangle}_\lambda$

The triangulation operation described in Algorithm 6.6 is slightly more complicated. Nevertheless, it is very similar to the autoreduction mentioned in Section 6.4. It suffices to reduce only the single indexed rows, as the multiplicative prolongations of the phantom rows obviously have different leading derivatives. In the outer `while` loop, we take a row f from $\mathcal{U}_\lambda$ with maximal leading derivative and try to reduce it modulo the rows contained in the system $\bar{\mathcal{S}}_\lambda$. Hence we check in Line /7/ whether a reduction with a row lying in one of the truncated involutive cones is possible. In the inner `while` loop, the actual reduction is performed. If no changes occur, the row is written back into the skeleton in Line /14/; otherwise the reduced row is added with initial level λ in Line /17/. If already multiplicative prolongations of the eliminated row have been computed, its removal would cause them to vanish as well. Thus we must keep track of their existence: in Line /20/ we turn the row into a phantom row, assuring that in future iterations its multiplicative prolongations are still taken into account.

The correctness of these algorithms is equivalent to showing that the system $\bar{\mathcal{S}}_\lambda^\triangle$ defined by the skeleton $\mathcal{S}_\lambda^\triangle$ yields a local representation for the differential equation $\mathcal{R}_{q+\lambda}$, if the initial differential equation was $\mathcal{R}_q$ (represented by the skeleton $\mathcal{S}_0$). As the main step towards this result we consider the relation between the prolonged rows of $\bar{\mathcal{S}}_\lambda^\triangle$ and rows in $\mathcal{S}_{\lambda+1}$.

Proposition 6.5.3. *Consider the sequence* (6.20) *derived from a skeleton* $\mathcal{S}_0$ *by repeatedly applying Algorithms 6.5 and 6.6. For* $\lambda \geq 0$ *let* g *be a linear combination of rows of the system* $\bar{\mathcal{S}}_\lambda^\triangle$*. Then each prolongation of* g *by one order can be written as a linear combination of rows of the system* $\bar{\mathcal{S}}_{\lambda+1}$*.*

Proof. We first slightly modify Algorithm 6.6 by replacing Line /20/ with $\mathcal{U}_\lambda^\triangle \leftarrow \mathcal{U}_\lambda^\triangle \cup \{f^\sigma_{(k_\sigma)}\}$. This implies that no phantom rows occur. So we ignore, for the moment, all references to the set $\mathcal{W}_\lambda$. Any row of which multiplicative prolongations have been computed is contained in all subsequent skeletons. Of course, one thereby loses the triangular form of the systems $\bar{\mathcal{S}}_\lambda^\triangle$.

Let the skeleton $\mathcal{S}_\lambda^\triangle$ consist of the single indexed rows $\{f^\sigma_{(k_\sigma)}\}$ and let

$$g = \sum_{l=1}^{s} a_l(x) D_{\mu_l} f^{\sigma_l} \tag{6.21}$$

be a finite linear combination of differential consequences of rows in the system $\bar{\mathcal{S}}_\lambda^\triangle$. The rows in the system $\bar{\mathcal{S}}_\lambda$ are multiplicative prolongations of rows in the corresponding skeleton $\mathcal{S}_\lambda$. We formally differentiate (6.21) with respect to an arbitrary independent variable x^j obtaining

$$D_j g = \sum_{l=1}^{s} \left[a_l(x) D_{\mu_l+1_j} f^{\sigma_l} + \frac{\partial a_l(x)}{\partial x^j} D_{\mu_l} f^{\sigma_l}) \right] . \tag{6.22}$$

We must show that all the rows of $\bar{\mathcal{S}}_\lambda^\triangle$ *and* their prolongations by one order are elements of $\bar{\mathcal{S}}_{\lambda+1}$. By the remark above, this is obvious for the unprolonged rows. For the other rows, we proceed by an induction on the global level λ. For $\lambda = 0$, Algorithm 6.5 computes all non-multiplicative prolongations (since $k_\sigma = \lambda = 0$ for each row $f^\sigma_{(k_\sigma)}$). The multiplicative prolongations are taken care of by the incrementation of the global level λ. Now assume $\lambda > 0$. Let $h = D_{\mu+1_j} f$ be a prolongation with $D_\mu f \in \bar{\mathcal{S}}_\lambda^\triangle$ and

$f_{(k)} \in \mathcal{S}^{\triangle}_{\lambda}$. We prove that $h \in \bar{\mathcal{S}}_{\lambda+1}$ by a Noetherian argument. The terminal cases are:

- If x^j is a *multiplicative* variable for $f_{(k)}$, we are done: all differentiations in $D_{\mu+1_j}$ are multiplicative for f and, as the global level is raised by the prolongation, $h \in \bar{\mathcal{S}}_{\lambda+1}$.
- If $|\mu| = 0$ and x^j is a *non-multiplicative* variable for $f_{(k)}$, then the prolongation is explicitly computed and therefore contained in the next system.

In all other cases, we set $c = \operatorname{cls} \mu + 1_j$ and rewrite $h = D_c(D_{\mu+1_j-1_c} f)$. Because of $|\mu| \leq (\lambda - k)$, the total derivative in the parentheses is an element of $\bar{\mathcal{S}}_{\lambda'}$ for some $\lambda' \leq \lambda$. Thus it is also contained in $\bar{\mathcal{S}}_{\lambda}$ and we may apply the induction hypothesis in order to write h in the form

$$h = D_c\Big[\sum_{l=1}^{s'} b_l(x) D_{\nu_l} f^{\sigma_l}\Big] . \tag{6.23}$$

Observe that for the summand h' possessing the highest leading derivative (its multi index is $\mu+1_j-1_c$) the variable x^c is by construction multiplicative. Hence $D_c h' \in \bar{\mathcal{S}}_{\lambda+1}$ as desired. With each of the remaining summands, we repeat the above manipulations. As their leading exponents are all less than $\mu + 1_j - 1_c$, this process must terminate after a finite number of steps. This proves the statement above under the made assumption.

Finally, we include the phantom rows, i.e. we change back Line /20/ of Algorithm 6.6. It follows from (6.22) that if a row is dependent of the other rows in $\bar{\mathcal{S}}_{\lambda}$ and therefore eliminated, all its multiplicative prolongations by one order reduce to zero modulo $\bar{\mathcal{S}}_{\lambda+1}$. Therefore it does not matter for the system (however, it does matter for its triangular form) whether we carry out these reductions or not. Non-multiplicative prolongations of phantom rows never need to be computed, as for $f_{(k,\ell)}$ always $k < \ell \leq \lambda$ holds. □

Once this somewhat messy proof is finished, we obtain as corollary the for us crucial theorem on the relation between the sequence (6.20) obtained by alternately prolonging and triangulising the initial skeleton $\mathcal{S}_0$ and the prolongations of the linear differential equation $\mathcal{R}_q$ corresponding to $\mathcal{S}_0$.

Theorem 6.5.4. *Let the system $\mathcal{F}$ be a local representation for the linear differential equation $\mathcal{R}_q$. If we turn $\mathcal{F}$ into a skeleton $\mathcal{S}_0$ as described above and form the sequence* (6.20)*, then for all $\lambda \geq 0$ the system $\bar{\mathcal{S}}^{\triangle}_{\lambda}$ is triangulised and represents the equation $\mathcal{R}_{q+\lambda}$.*

Proof. We denote by $\mathcal{F}_0 = \mathcal{F}, \mathcal{F}_1, \mathcal{F}_2, \dots$ the prolongations of the system $\mathcal{F}$. Each $\mathcal{F}_i$ is a local representation for the prolonged differential equation $\mathcal{R}_{q+1}$. If we set $\mathcal{F}_0 = \bar{\mathcal{S}}^{\triangle}_0$, we have also equality of the linear hulls: $\langle \mathcal{F}_0 \rangle = \langle \bar{\mathcal{S}}^{\triangle}_0 \rangle$.

By repeatedly applying Proposition 6.5.3, we get for $\lambda \geq 0$: $\langle \mathcal{F}_\lambda \rangle = \langle \bar{\mathcal{S}}_\lambda \rangle$. As we perform only elementary row operations during Algorithm 6.6, it follows that $\bar{\mathcal{S}}_\lambda^\triangle$ is a representation of $\mathcal{R}_{q+\lambda}$.

It remains to show is that the systems $\bar{\mathcal{S}}_\lambda^\triangle$ are really in triangular form. First of all, the rows in $\mathcal{U}_\lambda^\triangle$ obviously possess different leading derivatives. Now suppose that there are two rows $f^1, f^2 \in \bar{\mathcal{S}}_\lambda^\triangle$ with $\operatorname{ld} D_{\mu_1} f^1 = \operatorname{ld} D_{\mu_2} f^2$. Let $f^1_{(k_1,\ell_1)}$ and $f^2_{(k_2,\ell_2)}$ be the corresponding rows in the skeleton $\mathcal{S}_\lambda^\triangle$ from which they have been derived by multiplicative prolongations. If either f^1 or f^2 is a single indexed row, we simply ignore all references made to the phantom level.

By definition of an involutive division, if two involutive cones intersect, one is entirely contained in the other one. Thus without loss of generality, we assume $\operatorname{ld} f^2 = \operatorname{ld} D_\nu f^1$ for some multiplicative prolongation D_ν. This implies that in step $k_1 + |\nu|$ the row f^2 was reduced and became a phantom row. In the next $|\mu_2|$ steps, all multiplicative prolongations of f^2 up to this order are reduced. From $|\nu| + |\mu_2| = |\mu_1| \leq (\lambda - k_1)$ it follows that $D_{\mu_2} f^2 \notin \bar{\mathcal{S}}_\lambda^\triangle$, and we have arrived at a contradiction. □

By working with skeletons instead of full systems, we need to compute the multiplicative prolongations only when they are required for a reduction. In each step of the algorithm, we prolong the skeleton by exactly one order and compute a triangular form. However, as we want to realise the Cartan–Kuranishi completion, we have to relate these ideas to the geometric theory.

We have already seen that the skeleton $\mathcal{S}_\lambda^\triangle$ yields a representation for the differential equation $\mathcal{R}_{q+\lambda}$. In fact, it contains even more information. Projecting into jet bundles of lower order, one can extract from $\mathcal{S}_\lambda^\triangle$ representations for all differential equations $\mathcal{R}^{(s)}_{q+r}$ with $\lambda = r + s$:

$$\mathcal{S}_\lambda^\triangle : \begin{cases} \mathcal{R}_{q+\lambda} \xrightarrow{\pi^{q+\lambda}_{q+\lambda-1}} \mathcal{R}^{(1)}_{q+\lambda-1} \xrightarrow{\pi^{q+\lambda-1}_{q+\lambda-2}} \mathcal{R}^{(2)}_{q+\lambda-2} \xrightarrow{\pi^{q+\lambda-2}_{q+\lambda-3}} \cdots \end{cases} . \tag{6.24}$$

Because of the triangular form of $\bar{\mathcal{S}}_\lambda^\triangle$, this amounts to dropping the rows of appropriate orders, although we will never do this explicitly. Instead, we identify the differential equation $\mathcal{R}^{(s)}_{q+r}$ currently considered by the values of the parameters r and s.

If we juxtapose all the sequences of the form (6.24), we arrive at the grid pictured in Figure 6.3. Starting with $\mathcal{R}_q$, the Cartan–Kuranishi algorithm moves along the horizontal lines searching for an equation with involutive symbol. If one is found, it is checked for integrability conditions by comparing it with the equation one row below and one column to the right. If the two equations differ, the process is continued in the row below, i.e. after adding the integrability conditions. Recall that in general $\bigl(\mathcal{R}^{(s)}_{q+r}\bigr)_{+1} \neq \mathcal{R}^{(s)}_{q+r+1}$;

$$
\begin{tikzcd}
\mathcal{S}_0^{\triangle} \arrow[r, Rightarrow] \arrow[d, dash] & \mathcal{S}_1^{\triangle} \arrow[r, Rightarrow] \arrow[d, dash] & \mathcal{S}_2^{\triangle} \arrow[r, Rightarrow] \arrow[d, dash] & \cdots \\
\mathcal{R}_q \arrow[r] \arrow[d, dash] \arrow[dr, equal, "?"] & \mathcal{R}_{q+1} \arrow[r] \arrow[d, dash] \arrow[dr, equal, "?"] & \mathcal{R}_{q+2} \arrow[r] \arrow[d, dash] & \cdots \\
\mathcal{R}_{q-1}^{(1)} \arrow[r, dashed] \arrow[d, dash] \arrow[dr, equal, "?"] & \mathcal{R}_q^{(1)} \arrow[r, dashed] \arrow[d, dash] \arrow[dr, equal, "?"] & \mathcal{R}_{q+1}^{(1)} \arrow[r, dashed] \arrow[d, dash] & \cdots \\
\mathcal{R}_{q-2}^{(2)} \arrow[r, dashed] \arrow[d, dash] & \mathcal{R}_{q-1}^{(2)} \arrow[r, dashed] \arrow[d, dash] & \mathcal{R}_q^{(2)} \arrow[r, dashed] \arrow[d, dash] & \cdots \\
\vdots & \vdots & \vdots &
\end{tikzcd}
$$

Fig. 6.3. Skeletons and the corresponding equations

to ensure equality we need an involutive symbol (cf. Part (ii) of Proposition 3.1.10). So two tasks remain to be considered:

1. provide a method for deciding which of the differential equations represented by a skeleton have involutive symbols;
2. determine whether $\mathcal{R}_{q+r}^{(s)}$ and $\mathcal{R}_{q+r}^{(s+1)}$ are equal.

The first one is solved by the following result.

Proposition 6.5.5. *Let $\mathcal{S}_\lambda^{\triangle}$ be a skeleton for the differential equation $\mathcal{R}_{q+\lambda}$ in δ-regular coordinates and let t denote the maximal order of a single indexed row in $\mathcal{S}_{\lambda+1}^{\triangle}$. The symbol $\mathcal{M}_{q+r}^{(s)}$ is involutive for all values r, s such that $r+s=\lambda$ and $q+r \geq t$.*

Proof. We choose r, s with $r+s=\lambda$. A representation for $\mathcal{R}_{q+r}^{(s)}$ can be obtained from $\bar{\mathcal{S}}_\lambda^{\triangle}$ by dropping all rows of order greater than $q+r$. We prolong this system by one order and compute a triangular form. If no non-multiplicative prolongations of order $q+r+1$ remain, Proposition 3.1.4 can be invoked to prove that the symbol $\mathcal{M}_{q+r}^{(s)}$ is involutive. By assumption, even after the prolongation and subsequent triangulation of the full skeleton no single indexed rows of this or a greater order exist in $\mathcal{S}_{\lambda+1}^{\triangle}$. As an independent non-multiplicative prolongation would result in such a row, the involution of the examined symbol can be read off from this skeleton. □

A useful side effect of this result is that it allows us to carry out additional projections while working inside a given skeleton: we can trace a vertical line in Figure 6.3 belonging to some skeleton $\mathcal{S}_\lambda^\triangle$ downwards as long as the symbols of the respective differential equations remain involutive and continue with the last such equation. This is shown in Figure 6.4 in comparison to the strategy followed in the generic version of the Cartan–Kuranishi completion (boxed equations possess involutive symbols).

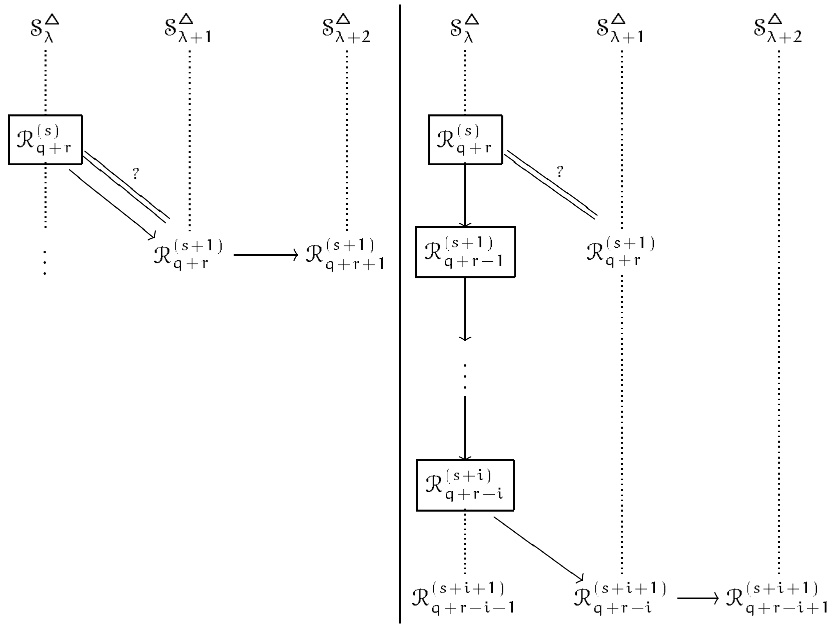

Fig. 6.4. Analysis of symbols and systems (left: classical Cartan–Kuranishi algorithm; right: new algorithm).

For the second task mentioned above, we note that it suffices to consider the dimensions of the involved manifolds, as one is always a submanifold of the other. As the triangulation guarantees algebraic independence of the rows in a skeleton, the determination of dimensions reduces to a simple counting of equations.

Definition 6.5.6. *Let $\mathcal{S}_\lambda^\triangle$ be a triangulised skeleton. We denote by $\#\mathcal{S}_{\lambda,t}^\triangle$ the number of rows of order t in the system $\bar{\mathcal{S}}_\lambda^\triangle$. The* rank vector *of the skeleton is the list*

$$\mathbf{r}_{\mathcal{S}_\lambda^\triangle} = [\#\mathcal{S}_{\lambda,0}^\triangle, \dots, \#\mathcal{S}_{\lambda,q+\lambda}^\triangle] \tag{6.25}$$

where q is the order of the original skeleton $\mathcal{S}_0$.

The numbers $\#\mathcal{S}^{\triangle}_{\lambda,t}$ yield straightforwardly the dimensions of all differential equations and symbols represented by $\mathcal{S}^{\triangle}_{\lambda}$. For $\lambda = r + s$

$$\dim \mathcal{R}^{(s)}_{q+r} = \dim J_{q+r}\mathcal{E} - \sum_{i=0}^{q+r} \#\mathcal{S}^{\triangle}_{\lambda,i}\,, \tag{6.26a}$$

$$\dim \mathcal{M}^{(s)}_{q+r} = \dim S_q(T^*\mathcal{X}) \underset{J_{q-1}\mathcal{E}}{\otimes} V\mathcal{E} - \#\mathcal{S}^{\triangle}_{\lambda,q+r}\,. \tag{6.26b}$$

The two differential equations $\mathcal{R}^{(s)}_{q+r}$ and $\mathcal{R}^{(s+1)}_{q+r}$ are now identical, i.e. no integrability conditions appear during the next prolongation, if the rank vectors $\mathbf{r}_{\mathcal{S}^{\triangle}_{\lambda}}$ and $\mathbf{r}_{\mathcal{S}^{\triangle}_{\lambda+1}}$ coincide in their first $q + r$ entries. The higher entries play no role.

It is straightforward to modify the Algorithms 6.5 and 6.6 so that they update the rank vector. Computing it for the initial skeleton is trivial. During the prolongation from $\mathcal{S}_\lambda$ to $\mathcal{S}_{\lambda+1}$ the following corrections are necessary for each row $f \in \mathcal{S}_\lambda$:

- If $f_{(k)}$ is a single indexed row of order t with level $k = \lambda$, prolongations with respect to all independent variables are computed. Hence the entry for order $t + 1$ is increased by n, the number of independent variables.
- If the initial level of $f_{(k)}$ with order t and class c is less than λ, $\binom{\lambda-k+c}{c-1}$ new rows of order $t + \lambda - k + 1$ enter the system.
- If $f_{(k,\ell)}$ is a double indexed row of order t and class c, there are $\binom{\lambda-k+c}{c-1}$ new rows of order $t + \lambda - k + 1$, and $\binom{\lambda-\ell+c}{c-1}$ rows of order $t + (\lambda - \ell) + 1$ are removed.

For a triangulation, the corrections are even simpler: if a row is reduced to a new row of lower order, the entry in the rank vector for the old order has to be decreased and the one for the new order has to be increased by one.

Combining all these ideas and results we finally arrive at the promised hybrid Algorithm 6.7. In contrast to the Cartan–Kuranishi Algorithm 6.1, it consists of only one loop in which the skeleton is prolonged and triangulised (Lines /8/ and /9/) using the above described algorithms. The analysis of the symbol based on Proposition 6.5.5 takes place in Line /10/. A simple comparison of rank vectors in Line /13/ signals whether integrability conditions have appeared. Note that we compare the entries of the vectors only up to the currently relevant order. The `while` loop in Lines /18–21/ realises the idea sketched in Figure 6.4 and projects as far as possible.

Strictly speaking, this is *not* a realisation of the Cartan–Kuranishi completion because of the inner `while` loop performing additional projections. A strict realisation would allow only one iteration through this loop. In practice, these additional projections are, however, a benefit. Of course, one is

Algorithm 6.7 A Hybrid Completion Algorithm for Linear Systems

Input: skeleton $\mathcal{S}_0^{\triangle}$ of the differential equation $\mathcal{R}_q$
Output: skeleton $\mathcal{S}_\lambda^{\triangle}$ and integers $r, s \in \mathbb{N}$ such that $\mathcal{R}_{q+r}^{(s)}$ is involutive

$r \leftarrow 0$
$s \leftarrow 0$
$\lambda \leftarrow 0$
systemInvolutive $\leftarrow$ *FALSE*
repeat
  $\bar{\mathbf{r}} \leftarrow \mathbf{r}_{\mathcal{S}_\lambda^{\triangle}}$
  $\lambda \leftarrow \lambda + 1$
  $\mathcal{S}_\lambda \leftarrow \mathtt{ProlongSkeleton}(\mathcal{S}_{\lambda-1}^{\triangle})$
  $\mathcal{S}_\lambda^{\triangle} \leftarrow \mathtt{TriangulizeSkeleton}(\mathcal{S}_\lambda)$
  if $\exists f_{(k)} \in \mathcal{U}_\lambda^{\triangle} : \operatorname{ord} f_{(k)} = q + r + 1$ **then** { *symbol not involutive* }
    $r \leftarrow r + 1$
  else { $\mathcal{M}_{q+r}^{(s)}$ *involutive; check formal integrability of* $\mathcal{R}_{q+r}^{(s)}$ }
    if $\forall 0 \leq i \leq q + r : \mathbf{r}_{\mathcal{S}_\lambda^{\triangle}}[i] = \bar{\mathbf{r}}[i]$ **then**
      systemInvolutive $\leftarrow$ *TRUE*
    else
      $r \leftarrow r + 1$
    end_if
    while $\neg\bigl(\exists f_{(k)} \in \mathcal{U}_\lambda^{\triangle} : \operatorname{ord} f_{(k)} = q + r\bigr)$ **do** { *project as far as possible* }
      $r \leftarrow r - 1$
      $s \leftarrow s + 1$
    end_while
  end_if
until *systemInvolutive*
return $(r, s, \mathcal{S}_\lambda^{\triangle})$

interested in those values r, s for which $\mathcal{R}_{q+r}^{(s)}$ is involutive *and* the sum $r+s$ is minimal. Our new algorithm delivers these values.

Theorem 6.5.7. *If the coordinates* $\mathbf{x}$ *are* δ*-regular for all appearing symbols, Algorithm 6.7 terminates with an involutive system.*

Proof. This follows immediately from the Cartan–Kuranishi Theorem 6.2.1. The only problem for the termination is that at same stage we might fail to recognise that the current symbol is involutive. In this case, Algorithm 6.7 would prolong infinitely often. But if the used coordinate system is δ-regular for each appearing symbol, involution of the symbols is easily checked by Proposition 6.5.5. □

Theorem 2.4.10 allows us to verify the condition of Theorem 6.5.7. If the used coordinates are not δ-regular for one of the symbols, then at some stage of the completion a larger Janet span appears. Thus with the help of this simple criterion, we can effectively avoid an infinite computation.

Given the triangulised skeleton $\mathcal{S}_\lambda^\triangle$ of a differential equation $\mathcal{R}_q$, we refine the above considerations about the dimensions in order to determine the indices $\beta_q^{(i)}$ of its symbol $\mathcal{M}_q$. A row $f_{(k,\ell)} \in \mathcal{S}_\lambda^\triangle$ of order t and class c obviously contributes to them, if and only if (i) $\lambda - k + t \geq q$ and (ii) $\ell = 0$ or $\lambda - \ell + t < q$. Its contribution $\beta_q^{(i)}(f_{(k,\ell)})$ is then given by

$$\beta_q^{(i)}(f_{(k,\ell)}) = \begin{cases} B(c-i+1, q-t, 1) & \text{for } 1 \leq i \leq c\,, \\ 0 & \text{for } c < i \leq n \end{cases} \tag{6.27}$$

where $B(n,q,c) = \binom{n-c+q-1}{q-1}$ denotes the number of multi indices of length n, order q and class c. The Cartan characters $\alpha_q^{(i)}$ of the equation are then computed from (4.6).

Example 6.5.8. We demonstrate the working of Algorithm 6.7 for the Janet Example 6.2.2. Our algorithm starts with the skeleton

$$\mathcal{S}_0 = \mathcal{S}_0^\triangle : \begin{cases} (u_{zz} + yu_{xx})_{(0)}\,, \\ (u_{yy})_{(0)}\,. \end{cases} \tag{6.28}$$

The initial rank vector is $\mathbf{r}_{\mathcal{S}_0^\triangle} = [0,0,2]$ and the algorithm proceeds as follows.

1. *Iteration*: There is only one non-multiplicative prolongation: u_{yyz}. It cannot be reduced and is added to the skeleton with initial level 1. Since $\mathcal{S}_1^\triangle$ contains a third-order row, the symbol $\mathcal{M}_3$ is not involutive.
2. *Iteration*: u_{yyzz} is the only new non-multiplicative prolongation and it reduces to u_{xxy}. As no row of order 4 remains in the skeleton, $\mathcal{M}_4$ is involutive. We must now compare the rank vectors up to order 3: $\bar{\mathcal{S}}_1^\triangle$ contains 6 equations of order 3, whereas $\bar{\mathcal{S}}_2^\triangle$ contains 7. Thus one integrability condition has appeared (which is of course u_{xxy}) and the algorithm continues with the equation $\mathcal{R}_3^{(1)}$.
3. *Iteration*: The computed non-multiplicative prolongations are u_{xxyy} and u_{xxyz}. Only the first row can be reduced to zero, while the second one obstructs the involution of $\mathcal{M}_3^{(1)}$.
4. *Iteration*: Again there are two non-multiplicative prolongations: u_{xxyyz} can be reduced to zero; u_{xxyyz} is replaced by u_{xxxx}. Due to the lack of an equation of order 5 in $\mathcal{S}_4^\triangle$, the symbol $\mathcal{M}_4^{(1)}$ is involutive. The rank vectors that must be compared are: $\mathbf{r}_{\mathcal{S}_3^\triangle} = [0,0,2,7,13,17]$ and $\mathbf{r}_{\mathcal{S}_4^\triangle} = [0,0,2,7,14,19,24]$. Only the first five components are taken into account, but the difference between the entries 13 and 14 signals the appearance of another integrability condition (u_{xxxx}). The next equation to be examined is $\mathcal{R}_4^{(2)}$.

5. *Iteration*: Of the two non-multiplicative prolongations, u_{xxxxy} is reduced to zero, whereas u_{xxxxz} represents an obstruction to involution for the symbol $\mathcal{M}_4^{(2)}$. The current skeleton is now

$$\mathcal{S}_5^{\triangle} : \begin{cases} (u_{xxxxz})_{(5)}\,, & (u_{yyz})_{(1)}\,, \\ (u_{xxxx})_{(4)}\,, & (u_{zz} + yu_{xx})_{(0)}\,, \\ (u_{xxyz})_{(3)}\,, & (u_{yy})_{(0)}\,, \\ (u_{xxy})_{(2)} \end{cases} \tag{6.29}$$

with rank vector $\mathbf{r}_{\mathcal{S}_5^{\triangle}} = [0, 0, 2, 7, 14, 21, 26, 32]$.

6. *Iteration*: As both non-multiplicative prolongations u_{xxxxyz} and u_{xxxxzz} can be reduced to zero, the symbol $\mathcal{M}_5^{(2)}$ is involutive. The comparison of the rank vectors shows that no new integrability conditions have appeared. Thus the algorithm stops and outputs the skeleton $\mathcal{S}_6^{\triangle}$, which is identical with $\mathcal{S}_5^{\triangle}$ given above, for the differential equation $\mathcal{R}_5^{(2)}$. A local representation is obtained by prolonging all equations in the skeleton multiplicatively up to fifth order.

Thus the path of the Cartan–Kuranishi completion can be summarised in the following sequence of differential equations:

$$\mathcal{R}_2 \to \mathcal{R}_3 \to \mathcal{R}_3^{(1)} \to \mathcal{R}_4^{(1)} \to \mathcal{R}_4^{(2)} \to \mathcal{R}_5^{(2)}\,. \tag{6.30}$$

In order to demonstrate the superior efficiency compared to a direct implementation of the Cartan–Kuranishi completion we do some statistics. The new algorithm computes 10 non-multiplicative and 18 multiplicative prolongations (for the reductions); 5 of the latter ones are used several times. In contrast, a full determination of all systems requires the prolongation of 132 equations — almost five times more. In addition, the naive approach requires to determine row echelon forms for a number of matrices (Jacobians and symbols), the largest being an 86×84 matrix. $\triangleleft$

One may wonder why we need rank vectors, as all integrability conditions in the Janet example stem from non-multiplicative prolongations that do not reduce to zero. This is, however, only one type of integrability condition; there exists a second type for systems comprising equations of different orders. As a trivial example we consider the equation $\mathcal{R}_2$ locally represented by $u_{yy} = u_{xy} = u_x = 0$. Obviously, the symbol $\mathcal{M}_2$ is involutive. But $\mathcal{R}_2$ is not formally integrable: prolongation of the lower order equation $u_x = 0$ with respect to the *multiplicative* variable x yields the additional equation $u_{xx} = 0$ required to represent $\mathcal{R}_2^{(1)}$.

In our algorithm the skeletons $\mathcal{S}_0^{\triangle}$ and $\mathcal{S}_1^{\triangle}$ contain exactly the same rows. But their rank vectors differ: $\mathbf{r}_{\mathcal{S}_0^{\triangle}} = [0, 1, 2]$ and $\mathbf{r}_{\mathcal{S}_1^{\triangle}} = [0, 1, 3, 3]$. The increase in the third entry is due to the integrability condition $u_{xx} = 0$. Our algorithm never explicitly determines this type of integrability conditions; they are solely visible through changes in the rank vector.

In the Janet example, phantom rows do not occur and the projections are always only by one order. In the next example, stemming from Lie symmetry theory [20, 129], both phenomena are present.

Example 6.5.9. We consider the determining system for the generators of Lie point symmetries of the heat equation. It consists of nine equations in three dependent variables η, ξ, τ and three independent variables t, x, u.

$$\mathcal{R}_2 : \begin{cases} \tau_u = 0\,, & \tau_{ux} + \xi_u = 0\,, & \eta_{uu} - \xi_{ux} = 0\,, \\ \tau_x = 0\,, & \tau_{xx} + 2\xi_x - \tau_t = 0\,, & \eta_{xx} - \eta_t = 0\,, \\ \tau_{uu} = 0\,, & \eta_{ux} - \frac{1}{2}\xi_{xx} + \frac{1}{2}\xi_t = 0\,, & \xi_{uu} = 0\,. \end{cases} \tag{6.31}$$

As above, we obtain $\mathcal{S}_0^{\triangle}$ by dropping the right hand side of each equation and assigning an initial level of 0. During the triangulation of $\mathcal{S}_1$, the row $\tau_{xx} + 2\xi_x - \tau_t$ can be reduced by the (multiplicative) prolongation with respect to x of the row τ_x. This yields the row $\xi_x - \frac{1}{2}\tau_t$, which is added to $\mathcal{S}_1^{\triangle}$. However, the system it defines also contains multiplicative prolongations of $\tau_{xx} + 2\xi_x - \tau_t$. Thus dropping the row completely would lose information and lead to wrong values for the dimensions. Instead, its phantom level is set to 1. In the λth iteration, all rows obtained by prolonging λ times the row $\tau_{xx} + 2\xi_x - \tau_t$ are still present in $\bar{\mathcal{S}}_\lambda$ and available for reducing other rows.

The algorithm continues to produce the following sequence of equations (note the double projection in the third iteration step):

$$\mathcal{R}_2 \to \mathcal{R}_3 \to \mathcal{R}_3^{(1)} \to \mathcal{R}_2^{(3)} \to \mathcal{R}_2^{(4)} \to \mathcal{R}_3^{(4)}\,. \tag{6.32}$$

The final skeleton is

$$\mathcal{S}_5^{\triangle} : \begin{cases} \tau_{ttt} = 0\,, & \eta_{xx} - \eta_t = 0\,, & \xi_u = 0\,, \\ \eta_{uu} - \xi_{ux} = 0\,, & \xi_{tt} = 0\,, & \xi_x - \frac{1}{2}\tau_t = 0\,, \\ \eta_{ut} + \frac{1}{2}\xi_{xt} = 0\,, & \tau_u = 0\,, & \tau_x = 0\,, \\ \eta_{ux} - \frac{1}{2}\xi_{xx} + \frac{1}{2}\xi_t = 0\,. \end{cases} \tag{6.33}$$

Since only head reductions have been performed during the algorithm, there exist further simplifications. ◁

6.6 Applications I: Differentiation and Perturbation Indices

In the numerical theory of differential algebraic equations (see e.g. [25, 68]) the concept of an *index*[5] plays an important role. It serves as an indicator

[5] Not to be confused with the indices $\beta_q^{(k)}$ of the symbol!

for the difficulties one has to expect, if one tries to numerically solve the given differential equation. The higher the index of an equation, the more problems arise in its integration. Typical problems are that classical methods exhibit suddenly a lower order when applied to non-normal systems or that constraints are violated, the so-called drift off the constraint manifold.

The literature abounds with definitions of indices; a partial survey may be found in [30]. A recent trend is the extension of results on differential algebraic equations to non-normal systems of partial differential equations. One speaks then often of *partial differential algebraic equations*, but this terminology is somewhat misleading. In many important cases like the Maxwell or Navier–Stokes equations, such systems do not contain any algebraic equations; the non-normality is due to the presence of equations of lower class not of lower order. We will therefore avoid this term.

In this section we show that many of the proposed index concepts can be well understood within the framework of the completion to involution. In particular, we provide intrinsic formulations of them applicable to partial differential equations, too. In contrast to most proposals in the literature, our definitions are directly based on the partial differential equations and do not require the use of integral transformations, semi-discretisations or similar means of reduction to differential algebraic equations.

We distinguish two main classes of indices. *Differentiation indices* count the number of prolongations needed until the equation satisfies some condition. In the case of "the" differentiation index, the condition is that the equation is not underdetermined. *Perturbation indices* are based on estimates on the difference between solutions of the original equation and of a perturbation of it. Essentially, the question is up to which order derivatives of the perturbation enter the estimate.

The most widely used index for differential algebraic equations is "the" differentiation index usually defined via the so-called derivative arrays. The rth array consists of a local representation of the prolonged equation $\mathcal{R}_{q+r}$. The classical approach to "the" differentiation index of a first-order ordinary differential equation $\mathcal{R}_1$ defines it as the smallest number ν such that all derivatives $\dot{\mathbf{u}}$ can be determined as functions of the independent variable t and the dependent variables $\mathbf{u}$. In the language of Section 6.1 we may say that then the differential part of the system yields an underlying equation. In an intrinsic manner we define it as follows.

Definition 6.6.1. *The* determinacy index *of a differential equation* $\mathcal{R}_q \subseteq J_q\mathcal{E}$ *in* n *independent and* m *dependent variables is the smallest integer* $\nu_d = r + s$ *such that* $\beta_{q+r}^{(n)} = m$ *for the differential equation* $\mathcal{R}_{q+r}^{(s)}$.

In order to avoid writing "the" differentiation index all the time, we have coined the new term determinacy index, as the defining property of this index is that $\mathcal{R}^{(s)}_{q+r}$ is not an underdetermined equation. ν_d has no finite value for an underdetermined equation, as then by Proposition 3.3.7 $\beta^{(n)}_{q+r} < m$ for any value of r and s. Note that, as promised, it makes no difference whether we analyse an ordinary or a partial differential equation.

Remark 6.6.2. The straightforward translation of the classical definition of the determinacy index into our geometric language would be to require that the symbol $\mathcal{M}^{(s)}_{q+r}$ vanishes, i.e. has dimension zero. For ordinary differential equations this is equivalent to Definition 6.6.1, as here $\dim \mathcal{M}^{(s)}_{q+r} = m - \beta^{(1)}_{q+r}$. However, for partial differential equations this approach yields a finite value only for equations of finite type, as only for these all derivatives of highest order can be expressed as functions of the remaining ones. Obviously, such a definition is not very useful in practice. $\lhd$

The determinacy index cannot be used as a basis of an existence theory for solutions. It is automatically determined during the completion of the original equation $\mathcal{R}_q$, but does not require the full completion. Thus it is well possible that the equation is in fact inconsistent. We will see below in our discussion of the perturbation index, why the determinacy index is nevertheless a useful concept.

From the point of view of the completion theory, Definition 6.6.1 appears to be a bit strange: why stop in the middle of the completion? Thus a more natural index concept is simply to count the number of prolongations needed until we reach an involutive equation. In the case of differential algebraic equations this approach is in fact equivalent to the *strangeness index* introduced by Kunkel and Mehrmann [100, 101] and to the geometric index definitions in [144, 146, 175].

Definition 6.6.3. *The* involution index *of the differential equation* $\mathcal{R}_q$ *is the smallest integer* $\nu_i = r + s$ *such that* $\mathcal{R}^{(s)}_{q+r}$ *is involutive.*

Of course, one could similarly define an *integrability index*, as the smallest value $r+s$ such that $\mathcal{R}^{(s)}_{q+r}$ is formally integrable. But we will see later that involution is more important in this context. Obviously, we have $\nu_d \leq \nu_i$

The perturbation index was first introduced by Hairer et al. [67] for differential algebraic equations and later extended to partial differential equations by Campbell and Marszalek [31]. It is an analytic concept and thus we formulate it only for systems. We assume that $\mathcal{X} \subset \mathbb{R}^n$ is a compact domain and $\mathcal{E} = \mathbb{R}^m \times \mathbb{R}$. We choose a vector norm $|\cdot|$ on $\mathbb{R}^m$ and a norm $\|\cdot\|$ on some function space $\mathcal{F}$ which is usually either the maximum norm $\|\cdot\|_{L^\infty}$ or the uniform norm $\|\cdot\|_{L^1}$. In addition, we define on $\mathcal{F}$ for each integer $k > 0$ a kind of Sobolev norm

$$\|\mathbf{f}\|_k = \sum_{0 \leq |\mu| \leq k} \left\| \frac{\partial^{|\mu|}\mathbf{f}}{\partial x^\mu} \right\| , \tag{6.34}$$

i.e. we sum the norms of $\mathbf{f}$ and all its partial derivatives up to order k. Of course, the function space $\mathcal{F}$ must be chosen such that these norms make sense on it.

Partial differential equations are usually accompanied by initial or boundary conditions. In order to accommodate for this, we take the following simple approach. Let $\mathcal{X}' \subseteq \partial\mathcal{X}$ be a subdomain of the boundary of $\mathcal{X}$ and introduce on the restriction $\mathcal{F}'$ of $\mathcal{F}$ to $\mathcal{X}'$ similar norms denoted by $\|\cdot\|'_k$. The conditions are then written in the form $\mathbf{\Psi}(\mathbf{x}, \mathbf{p})\big|_{\mathcal{X}'} = 0$. This comprises most kinds of initial or boundary conditions found in applications. The highest derivative in $\mathbf{\Psi}$ determines the order ℓ of the conditions.

We consider a general system of partial differential equations

$$\mathbf{\Phi}(\mathbf{x}, \mathbf{p}) = 0 \tag{6.35}$$

where $\mathbf{\Phi}$ takes its values in $\mathbb{R}^t$. As usually, we do not require that $t = m$ or that the Jacobian $\partial\mathbf{\Phi}/\partial\mathbf{p}$ has any special properties. We assume that we are given a smooth solution $\mathbf{u}(\mathbf{x})$ of it defined on the whole domain $\mathcal{X}$ and satisfying our initial or boundary conditions on $\mathcal{X}'$. We want to compare it with solutions of the perturbed equation

$$\mathbf{\Phi}(\mathbf{x}, \mathbf{p}) = \mathbf{\delta}(\mathbf{x}) \tag{6.36}$$

with a smooth right hand side $\mathbf{\delta}$.

Definition 6.6.4. *Let* $\mathbf{u}(\mathbf{x})$ *be a smooth solution of* (6.35) *satisfying some initial or boundary conditions of order* ℓ *on* $\mathcal{X}'$. *The system* (6.35) *has* perturbation index ν_p *along this solution, if* ν_p *is the smallest integer such that for any solution* $\tilde{\mathbf{u}}(\mathbf{x})$ *of the perturbed equation* (6.36) *at every point* $\mathbf{x} \in \mathcal{X}$ *the estimate*

$$|\mathbf{u}(\mathbf{x}) - \tilde{\mathbf{u}}(\mathbf{x})| \leq C\big(\|\mathbf{f} - \tilde{\mathbf{f}}\|'_\ell + \|\mathbf{\delta}\|_{\nu_p - 1}\big) \tag{6.37}$$

holds, whenever the right hand side is sufficiently small. Here $\mathbf{f}$ *and* $\tilde{\mathbf{f}}$ *represent the restrictions of the solutions* $\mathbf{u}$ *and* $\tilde{\mathbf{u}}$, *respectively, to* $\mathcal{X}'$. *The constant* C *may depend only on the domains* $\mathcal{X}$, $\mathcal{X}'$ *and on the function* $\mathbf{\Phi}$.

In the case of initial value problems for differential algebraic equations this definition coincides with the one given by Hairer et al. [67]. The term

$\|\mathbf{f}-\tilde{\mathbf{f}}\|'_k$ takes the simple form $|\mathbf{u}(\mathbf{x}_0)-\tilde{\mathbf{u}}(\mathbf{x}_0)|$ where $\mathcal{X}'=\{\mathbf{x}_0\}$. For partial differential equations our definition is almost identical with the one given by Campbell and Marszalek [31]; the only difference is that we do not include the order ℓ of the initial or boundary conditions in the index.

Whereas it is not so clear why differentiation indices should indicate the difficulty of the numerical integration of (6.35), this is rather obvious for the perturbation index. If we take for $\tilde{\mathbf{u}}(\mathbf{x})$ an approximate solution of (6.35), we may interpret $\boldsymbol{\delta}(\mathbf{x})$ as the residual obtained by entering it into the equation. (6.37) tells us that for an equation with $\nu_p > 1$ it does not suffice to keep this residual as small as possible, since also some of its derivatives enter the estimate. Strictly speaking, this implies that the considered initial or boundary value problem is ill-posed in the sense of Hadamard!

Unfortunately, it is rather difficult to derive estimates of the form (6.37), especially for partial differential equations. On the other hand, it is comparatively straightforward to compute the determinacy or the involution index. Thus it is a natural thought to bound the perturbation index by one of these. However, Campbell and Gear [30] showed that, in general, no relation between ν_d, ν_i and ν_p exists. This is not a surprise, as ν_p is based on the perturbed equation (6.36) and the perturbation may affect the completion for non-linear equations.

Example 6.6.5. Consider the following simple quasi-linear system of ordinary differential equations for two unknowns u and v:

$$vv'+u=0\,,\qquad v=0\,. \tag{6.38}$$

Obviously it is equivalent to $u=v=0$. Thus its determinacy and involution indices are $\nu_d=\nu_i=1$, respectively, as after one step we obtain an underlying equation $u'=v'=0$.

For the perturbed system

$$vv'+u=\delta\,,\qquad v=\epsilon\,, \tag{6.39}$$

the situation is different, as we cannot eliminate the derivative v' by algebraic manipulations. The first completion step yields $v'=\epsilon'$ and only now the first equation of (6.39) becomes equivalent to $u=\delta-\epsilon\epsilon'$. The second completion step leads to $u'=\delta'-(\epsilon\epsilon')'$ and thus an underlying equation. The indices of the perturbed system are $\tilde{\nu}_d=\tilde{\nu}_i=2$. From the solution of the system we easily read off that also $\nu_p=2$, as u depends on ϵ'.

As shown in [30], this example can be extended. Using an m-dimensional vector of unknowns $\mathbf{u}$, one constructs systems such that for the unperturbed system still $\nu_d=\nu_i=1$ but for the perturbed one $\tilde{\nu}_d=\tilde{\nu}_i=m$ and also $\nu_p=m$. Thus the difference between the indices of the unperturbed and the perturbed system may become arbitrarily large. $\triangleleft$

Definition 6.6.6. *Let $\mathbf{\Phi}(\mathbf{x}, \mathbf{p}) = 0$ be a system of differential equations. Its* perturbed determinacy index $\tilde{\nu}_d$ *and* perturbed involution index $\tilde{\nu}_i$, *respectively, are defined as the indices of the perturbed system $\mathbf{\Phi}(\mathbf{x}, \mathbf{p}) = \boldsymbol{\delta}(\mathbf{x})$.*

Note that these perturbed indices are not intrinsically defined for differential equations $\mathcal{R}_q$ but only for systems. The reason lies simply in the fact that our approach to perturbing the equation is based on a fixed local representation. This approach also reflects the situation appearing in concrete numerical computations.

Remark 6.6.7. Definition 6.6.6 implicitly assumes that the completion process of the perturbed equation is independent of the precise form of $\boldsymbol{\delta}$. In most applications this is probably the case. Otherwise the perturbed indices are computed with the generic branch of the completion, i. e. under the assumption that the perturbations do *not* satisfy any special conditions. $\lhd$

We show now that the perturbation index may be estimated by the perturbed determinacy index. We do this first for the case of ordinary differential equations, where we can obtain very general results. For partial differential equations we need to make a number of assumptions in order to prove a rigorous theorem.

Before we get to our main theorem, we need a proposition which is a simple consequence of the *Gronwall Lemma.* Recall that this lemma states that if a continuous function $\psi : [a, b] \to \mathbb{R}$ satisfies $\psi(x) \leq A + B\int_a^x \psi(\xi)d\xi$ for some constants $A, B \geq 0$, then the estimate $\psi(x) \leq Ae^{B(x-a)}$ holds for all $x \in [a, b]$. A proof of it is contained in almost any textbook on ordinary differential equations (see e. g. [188]). Based on it we can compute the perturbation index ν_p for normal systems of ordinary differential equations.

Proposition 6.6.8. *Let $\mathbf{\Phi}(x, \mathbf{u}, \mathbf{u}') = 0$ be an implicit ordinary differential equation such that the Jacobian $\partial\mathbf{\Phi}/\partial\mathbf{u}'$ has a bounded inverse on the interval $\mathcal{X} = [0, X]$. Furthermore, let the function $\mathbf{\Phi}$ be Lipschitz continuous in $\mathbf{u}$ with constant L_0 and let $\tilde{\mathbf{u}}(x)$ be a solution of the perturbed equation $\mathbf{\Phi}(x, \tilde{\mathbf{u}}, \tilde{\mathbf{u}}') = \boldsymbol{\delta}(x)$. Then for all $x \in [0, X]$*

$$|\mathbf{u}(x) - \tilde{\mathbf{u}}(x)| \leq C\Big(|\mathbf{u}(0) - \tilde{\mathbf{u}}(0)| + \int_0^x |\boldsymbol{\delta}(\xi)|d\xi\Big) \tag{6.40}$$

where the constant C depends only on $\mathbf{\Phi}$ and the length X of the interval.

Proof. We solve the original and the perturbed equation for the derivatives $\mathbf{u}'$ and $\tilde{\mathbf{u}}'$, respectively, which is always possible under the made assumptions. This yields the equations $\tilde{\mathbf{u}}' = \boldsymbol{\phi}(x, \tilde{\mathbf{u}}, \boldsymbol{\delta})$ and $\mathbf{u}' = \boldsymbol{\phi}(x, \mathbf{u}, 0)$. The function $\boldsymbol{\phi}$ satisfies a Lipschitz condition in $\tilde{\mathbf{u}}$ and $\boldsymbol{\delta}$ with a constant L_1 depending on

L_0 and the bound on $(\partial\mathbf{\Phi}/\partial\mathbf{u}')^{-1}$. Now we integrate both equations from 0 to x, subtract them and take the norm:

$$\begin{aligned}|\mathbf{u}(x)-\tilde{\mathbf{u}}(x)| \leq |\mathbf{u}(0)-\tilde{\mathbf{u}}(0)| &+ \\ &\Big|\int_0^x \big[\boldsymbol{\phi}\big(\xi,\mathbf{u}(\xi),0\big)-\boldsymbol{\phi}\big(\xi,\tilde{\mathbf{u}}(\xi),\boldsymbol{\delta}(\xi)\big)\big]\,d\xi\Big|\,.\end{aligned} \tag{6.41}$$

Interchanging the integration and the norm, we exploit the Lipschitz continuity of $\boldsymbol{\phi}$ to obtain

$$|\mathbf{u}(x)-\tilde{\mathbf{u}}(x)| \leq |\mathbf{u}(0)-\tilde{\mathbf{u}}(0)| + L_1\int_0^x \big(|\mathbf{u}(\xi)-\tilde{\mathbf{u}}(\xi)| + |\boldsymbol{\delta}(\xi)|\big)d\xi\,. \tag{6.42}$$

Obviously, $\int_0^x |\boldsymbol{\delta}(\xi)|d\xi$ is a monotonically increasing function of x on the interval $[0,X]$ taking its maximal value at $x=X$. Applying Gronwall's Lemma with $\psi(x) = |\mathbf{u}(x)-\tilde{\mathbf{u}}(x)|$ and $A = |\mathbf{u}(0)-\tilde{\mathbf{u}}(0)| + L_1\int_0^X |\boldsymbol{\delta}(\xi)|d\xi$, $B = L_1$ yields immediately the claimed estimate with $C = \max(1, L_1)\exp(L_1X)$. □

Thus a fully implicit system of ordinary differential equations has generally the perturbation index $\nu_p = 1$, as the right hand side of (6.40) is the L^1 norm of $\boldsymbol{\delta}$. But note that in order to derive (6.42) we had to interchange in (6.41) the integration and the norm. If we start with an *explicit* system, so that the perturbed form is $\tilde{\mathbf{u}}' = \mathbf{f}(x,\tilde{\mathbf{u}}) + \boldsymbol{\delta}$, such an interchange is not necessary and we obtain on the right hand side of (6.40) a term of the form $|\int_0^x \boldsymbol{\delta}(\xi)d\xi|$. Interpreting $\|\boldsymbol{\delta}\|_{-1}$ as the (maximum) norm of $\int_0^x \boldsymbol{\delta}(\xi)d\xi$, we see that such a system has in fact the perturbation index $\nu_p = 0$.

Theorem 6.6.9. *Let* $\mathbf{\Phi}(x,\mathbf{u},\mathbf{u}') = 0$ *be a differential algebraic equation. Then*

$$\tilde{\nu}_d \leq \nu_p \leq \tilde{\nu}_d + 1\,. \tag{6.43}$$

Proof. The essential step of the proof is to show that the completion of the perturbed form of the given differential algebraic equation leads to an underlying equation of the form

$$\mathbf{\Phi}^c(x,\tilde{\mathbf{u}},\tilde{\mathbf{u}}',\boldsymbol{\delta},\boldsymbol{\delta}',\dots,\boldsymbol{\delta}^{(\tilde{\nu}_d)}) = 0 \tag{6.44}$$

with $\partial\mathbf{\Phi}^c/\partial\boldsymbol{\delta}^{(\tilde{\nu}_d)} \neq 0$. Then we apply Proposition 6.6.8 and get an estimate of the form (6.37) with $\nu_p - 1 = \tilde{\nu}_d$. The lower bound is obtained by taking into account the remark above: if the integration can be left within the norm, it kills one derivative. ν_p cannot be less than $\tilde{\nu}_d$, as the solution of (6.44) depends at least on $\boldsymbol{\delta}^{(\tilde{\nu}_d-1)}$, and by definition $\tilde{\nu}_d$ is the smallest number of differentiations leading to an underlying equation.

Before the completion we transform the perturbed differential algebraic equation into semi-explicit form; this is a purely algebraic process which does not introduce any derivatives. As outlined in Section 6.1, the completion process takes then the following simple form: the algebraic equations are differentiated and it is checked whether this yields any new equations; the completion stops as soon as no new *algebraic* equations arise.

In the first step the algebraic part of the differential algebraic equation is differentiated. If this leads to any new equations, they contain a first-order derivative of a perturbation. If some of the new equations are algebraic, they are differentiated again. Thus any new equations arising out of them contains a second-order derivative of a perturbation and so on. The perturbed determinacy index $\tilde{\nu}_d$ counts the last completion step where new *differential* equations appear, i. e. equations that are part of the underlying equation $\mathbf{\Phi}^c = 0$. Thus these equations depend effectively on derivatives of order $\tilde{\nu}_d$ of the perturbations and $\mathbf{\Phi}^c$ is of the form claimed above. □

Example 6.6.10. We consider again the system of Example 6.6.5. As already mentioned, here $\nu_p = \tilde{\nu}_d$ in agreement with our theorem. We are at the lower bound, as the underlying equation is of the *explicit* form $u' = \delta' - (\epsilon\epsilon')'$, $v' = \epsilon'$ and we can thus perform the integration within the norm. ◁

Now we want to extend these results to partial differential equations. We cannot expect to obtain a theory with the same level of generality, as estimates on solutions of partial differential equations are much more difficult to find than for ordinary differential equations. So we will make a rigorous statement only for a rather narrow class of equations. Looking at the proof of Theorem 6.6.9, we see that it depends essentially on the concept of an underlying equation and that this is the reason why the determinacy and not the involution index is appearing.

So far we have a notion of an underlying equation, or more precisely of an underlying vector field, only for ordinary differential equations. Thus we need first a more general definition. The basic idea is that an underlying equation is always a *normal* equation certain solutions of which coincide with solutions of our given differential equation.

Definition 6.6.11. *Let $\mathcal{R}_q \subseteq J_q\mathcal{E}$ be an arbitrary differential equation. A normal differential equation $\bar{\mathcal{R}}_{q+r} \subset J_{q+r}\mathcal{E}$ such that any (local) solution $\sigma : \mathcal{X} \rightarrow \mathcal{E}$ of $\mathcal{R}_q$ also solves $\bar{\mathcal{R}}_{q+r}$ is an* underlying equation *for $\mathcal{R}_q$.*

Let $\mathcal{R}^{(s)}_{q+r}$ be the involutive completion of $\mathcal{R}_q$. We may take a local representation of it in Cartan normal form. Assuming that $\mathcal{R}^{(s)}_{q+r}$ is not underdetermined, the subsystem of class n defines a normal equation satisfying the conditions of Definition 6.6.11. Thus, as for ordinary differential equations, the completion provides us with a simple method for the construction of a underlying equation. Again, we do not need a full completion; the determinacy index ν_d equals the number of prolongations necessary to find the underlying equation.

Theorem 6.6.12. *Let $\mathcal{R}_q$ be a linear equation and assume that the right hand side of its underlying equation $\bar{\mathcal{R}}_{q+r}$ obtained via a completion can be written in the form $A\mathbf{u}$ where A is the generator of a strongly continuous semigroup acting on an appropriately chosen Banach space. Then $\tilde{\nu}_d \leq \nu_p \leq \tilde{\nu}_d + 1$.*

Proof. As the basic idea behind semigroup theory [134, 149] is to treat normal partial differential equations as ordinary differential equations on some infinite-dimensional Banach space, we try to follow the same strategy as in the proof of Theorem 6.6.9.

Let us assume for simplicity that $q + r = 1$. We slightly change our notation and write t instead of x^n. Derivatives with respect to x^n are denoted by a dot. Adding a perturbation δ, the underlying equation has then the form $\dot{\mathbf{u}} = A\mathbf{u} + \delta$ where $\mathbf{u}$ and δ are considered as functions of t taking values in the chosen Banach space. It is well-known that the solution $\tilde{\mathbf{u}}$ of this equation is given by

$$\tilde{\mathbf{u}}(t) = S(t)\tilde{\mathbf{u}}(0) + \int_0^t S(t-\tau)\delta(\tau)\,d\tau\,. \tag{6.45}$$

Here $S(t)$ is the semigroup generated by A and $\tilde{\mathbf{u}}(0)$ an arbitrary element of the domain of A.

We are interested in an estimate of the difference between $\tilde{\mathbf{u}}(t)$ and the solution $\mathbf{u}(t)$ obtained for $\delta \equiv 0$. If $\|\cdot\|$ denotes the norm of the used Banach space, it follows immediately from (6.45) that

$$\|\tilde{\mathbf{u}}(t) - \mathbf{u}(t)\| \leq C\Big(\|\tilde{\mathbf{u}}(0) - \mathbf{u}(0)\| + \int_0^t \|\delta(\tau)\|\,d\tau\Big) \tag{6.46}$$

for some positive constant C, as the norms of the operators $S(t)$ of a strongly continuous semigroup are bounded on a finite interval.

If the perturbed determinacy index of $\mathcal{R}_q$ is $\tilde{\nu}_d$, the perturbation δ of the underlying equation depends on derivatives up to order $\tilde{\nu}_d$ of the perturbation of the original system. Thus we get $\nu_p \leq \tilde{\nu}_d + 1$. By exactly the same reasoning as in the case of ordinary differential equations, it may happen that the semigroup $S(t)$ has a particularly simple structure such that we obtain (6.46) with the integration within the norm. In this case $\nu_p = \tilde{\nu}_d$; obviously no lower value is possible. □

6.7 Applications II: Semi-Discretisations

In this section we study *semi-discretisations* of partial differential equations. For simplicity, we restrict to first-order linear homogeneous systems with constant coefficients. We assume that we are given a partial differential equation in $n+1$ independent variables. Then we discretise by some finite difference method the spatial derivatives with respect to n of these variables. This yields a differential algebraic equation in the remaining independent variable. We are interested in the relation between the involution indices of the original partial differential equation and the obtained differential algebraic equation.

Thus consider the first-order linear homogeneous system

$$\sum_{i=1}^{n+1} M_i \mathbf{u}_{x^i} + N\mathbf{u} = 0 \,. \tag{6.47}$$

The matrices M_i and N are here completely arbitrary (we are *not* assuming that we are dealing with a weakly overdetermined system). Instead of (6.47) we complete to involution the *perturbed* system

$$\sum_{i=1}^{n+1} M_i \mathbf{u}_{x^i} + N\mathbf{u} = \boldsymbol{\gamma}(\mathbf{x}) \,. \tag{6.48}$$

As we are dealing with a linear constant coefficients system, the perturbation does not affect the completion. However, it serves as a convenient mean of "book-keeping" which prolongations are needed during the completion.

Definition 6.7.1. *The* involution index ν_ℓ in direction x_ℓ *of the system* (6.47) *is the maximal number of differentiations with respect to* x^ℓ *appearing in a* $\boldsymbol{\gamma}$*-derivative contained in the involutive completion of the perturbed system* (6.48).

For the semi-discretisation we proceed as follows. Assume that x^ℓ is the independent variable "surviving" the discretisation; in other words, afterwards we are dealing with an ordinary differential equation containing only x^ℓ-derivatives. In order to simplify the notation, we rewrite (6.47). We denote x^ℓ by t and renumber the remaining independent variables as x^i with $1 \leq i \leq n$. Then we solve as many equations as possible for a t-derivative, as we consider these as the derivatives of highest class. This yields a system of the form

$$E\mathbf{u}_t = \sum_{i=1}^{n} A_i \mathbf{u}_{x_i} + B\mathbf{u} + \boldsymbol{\delta} \,, \tag{6.49a}$$

$$0 = \sum_{i=1}^{n} C_i \mathbf{u}_{x_i} + D\mathbf{u} + \boldsymbol{\epsilon} \,. \tag{6.49b}$$

Here we assume, as for a weakly overdetermined system, that the matrix E is of maximal rank, i. e. every equation in the subsystem (6.49a) really depends on a t-derivative. And we assume that the coordinates $(\mathbf{x}, t)$ are δ-regular. However, we do not pose any restrictions on the ranks of the matrices C_i.

In (6.49) we have introduced perturbations $\boldsymbol{\delta}$ and $\boldsymbol{\epsilon}$ which are related to the perturbations $\boldsymbol{\gamma}$ in (6.48) by linear transformations with constant coefficients. As we are only interested in the number of differentiations applied to them during the completion, such a transformation has no effect.

We discretise the "spatial" derivatives, i. e. the derivatives with respect to the x^i, on a grid where the points $\mathbf{x}_\mathbf{k}$ are labelled by integer vectors $\mathbf{k} = [k_1, \dots, k_n]$. $\mathbf{u}_\mathbf{k}(t)$ denotes the value of the function $\mathbf{u}$ at the point $\mathbf{x}_\mathbf{k}$ at time t. We approximate in (6.49) the spatial derivative $\mathbf{u}_{x_i}(\mathbf{x}_\mathbf{k}, t)$ by the finite difference

$$\delta_i \mathbf{u}_\mathbf{k}(t) = \sum_{\ell_i = -a_i}^{b_i} \alpha^{(i)}_{\ell_i} \mathbf{u}_{[k_1, \dots, k_i + \ell_i, \dots, k_n]}(t) \tag{6.50}$$

with some real coefficients $\alpha^{(i)}_{\ell_i}$. Thus different discretisations are allowed for different values of i, but $\mathbf{u}_{x_i}$ is everywhere discretised in the same way. Entering the approximations (6.50) of the spatial derivatives into (6.49) yields the differential algebraic equation

$$E\dot{\mathbf{u}}_\mathbf{k} = \sum_{i=1}^{n} A_i \delta_i \mathbf{u}_\mathbf{k} + B\mathbf{u}_\mathbf{k} \,, \tag{6.51a}$$

$$0 = \sum_{i=1}^{n} C_i \delta_i \mathbf{u}_\mathbf{k} + D\mathbf{u}_\mathbf{k} \,. \tag{6.51b}$$

Theorem 6.7.2. *Let the coordinates* $(\mathbf{x}, t)$ *be* δ*-regular for* (6.47)*. The involution index of the differential algebraic equation* (6.51) *obtained in the above described semi-discretisation of* (6.47) *with respect to* x^ℓ *is* ν_ℓ*.*

Proof. We must compare what happens during the completion of (6.49) and (6.51), respectively. Determining the index ν_ℓ requires keeping track of when some the perturbations are differentiated with respect to t. The index of the differential algebraic equation tells us how many prolongations of the algebraic equations in (6.51b) lead to integrability conditions. Thus we must show that whenever a perturbation in (6.49) is differentiated with respect to t an integrability condition appears in (6.51) and vice versa.

For the partial differential system (6.49) the differentiations with respect to t lead to the second-order equations

$$\mathsf{E}\mathbf{u}_{tt} = \sum_{i=1}^{n} \mathsf{A}_i \mathbf{u}_{x_i t} + \mathsf{B}\mathbf{u}_t + \boldsymbol{\delta}_t \,, \tag{6.52a}$$

$$0 = \sum_{i=1}^{n} \mathsf{C}_i \mathbf{u}_{x_i t} + \mathsf{D}\mathbf{u}_t + \boldsymbol{\epsilon}_t \,. \tag{6.52b}$$

As t is non-multiplicative for all equations in (6.49b), the equations (6.52b) are decisive for our analysis. The spatial differentiations of the partial differential system (6.49) yield for $1 \leq j \leq n$

$$\mathsf{E}\mathbf{u}_{x_j t} = \sum_{i=1}^{n} \mathsf{A}_i \mathbf{u}_{x_i x_j} + \mathsf{B}\mathbf{u}_{x_j} + \boldsymbol{\delta}_{x_j} \,, \tag{6.53a}$$

$$0 = \sum_{i=1}^{n} \mathsf{C}_i \mathbf{u}_{x_i x_j} + \mathsf{D}\mathbf{u}_{x_j} + \boldsymbol{\epsilon}_{x_j} \,. \tag{6.53b}$$

All equations in (6.53a) are the result of differentiations with respect to multiplicative variables. In (6.53b) we cannot make a similar statement, as we have not transformed our original system into a full triangular form and thus do not know which variables are multiplicative for which equation. However, this information is not needed later.

Similarly, we differentiate the differential algebraic equation (6.51)

$$\mathsf{E}\ddot{\mathbf{u}}_{\mathbf{k}} = \sum_{i=1}^{n} \mathsf{A}_i \delta_i \dot{\mathbf{u}}_{\mathbf{k}} + \mathsf{B}\dot{\mathbf{u}}_{\mathbf{k}} \,, \tag{6.54a}$$

$$0 = \sum_{i=1}^{n} \mathsf{C}_i \delta_i \dot{\mathbf{u}}_{\mathbf{k}} + \mathsf{D}\dot{\mathbf{u}}_{\mathbf{k}} \,. \tag{6.54b}$$

In addition, we write down – in formal analogy to the equations (6.53) obtained by spatial differentiations – the following equations obtained by applying the difference operators δ_j for $1 \leq j \leq n$ to (6.51)

$$\mathsf{E}\delta_j \dot{\mathbf{u}}_{\mathbf{k}} = \sum_{i=1}^{n} \mathsf{A}_i \delta_i \delta_j \mathbf{u}_{\mathbf{k}} + \mathsf{B}\delta_j \mathbf{u}_{\mathbf{k}} \,, \tag{6.55a}$$

$$0 = \sum_{i=1}^{n} \mathsf{C}_i \delta_i \delta_j \mathbf{u}_{\mathbf{k}} + \mathsf{D}\delta_j \mathbf{u}_{\mathbf{k}} \,. \tag{6.55b}$$

They are obviously just linear combinations of equations in (6.54); however, their addition simplifies the subsequent analysis.

Given all these equations, we analyse where either integrability conditions or obstructions to involution might occur. We distinguish two cases. Case I analyses (6.52b) and (6.54b), respectively. New equations arise whenever it is not possible to express all equations in (6.52b) (in (6.54b), respectively) as linear combinations of the equations in (6.53a) (in (6.55a), respectively). Obviously, the number of new equations is the same in both systems in view of their identical structure and those for the differential algebraic equation are semi-discretisations of those for the original partial differential system. For the partial differential system each of them contains a term $\boldsymbol{\epsilon}_t$. In the case of the differential algebraic system all new equations are integrability conditions; for the partial differential system some of them might represent obstructions to involution, namely those of second order.

Case II occurs only for the partial differential system: integrability conditions or obstructions to involution could arise *within* the subsystem (6.53b). No simple criterion exists for this, as we have not transformed to a full triangular form. But this is not important for us, as all these new equations contain only spatial derivatives of the perturbations and thus do not affect our theorem. There are no corresponding equations for the differential algebraic equation, as (6.55) is algebraically dependent of the original system.

If Case I does not lead to any new equations, no hidden constraints exist in the differential algebraic equation. Thus it is involutive and of index 0. The partial differential system is not necessarily involutive, as Case II might still lead to new equations. Our theorem would be wrong, if differentiating these with respect to the non-multiplicative variable t yielded again new equations. However, this cannot happen: by interchanging the order of differentiations one sees that any equation arising this way may also be generated by combining spatial derivatives of equations stemming from Case I.[6] This implies that no additional differentiations with respect to t of the perturbations are performed – no matter how many further iterations within Case II are necessary. Hence the index ν_ℓ vanishes in accordance with our claim.

If Case I yields new equations, both the involution index of the differential algebraic equation and the index ν_ℓ are at least 1. We may then continue in precisely the same manner. For the partial differential system we take all equations in (6.52) and (6.53) and first differentiate them with respect to t. Then we differentiate them with respect to the spatial variables x^k. Similarly, we take the equations in (6.54) and (6.55) and differentiate them with respect to t. Finally, we consider what we obtain by applying the difference operators δ_k to them.

The analysis of these third-order systems proceeds as above. We distinguish the same two cases: Case I involves time derivatives and thus adds *one* differentiation with respect to t to the perturbations; Case II concerns only the partial differential system and does not lead to any t-differentiations.

[6] This is exactly the same argument as in the proof of Proposition 3.1.10 where all calculations involved are performed in full detail.

New equations arising in Case II do not lead at later stages of the completion to any additional t-differentiations of the perturbations, as we may always interchange the order of differentiations and thus all the differentiations with respect to t have already been studied at an earlier stage in Case I.

In this manner we proceed to higher-order systems until no new equations appear in Case I. Because of our assumption that the coordinates are δ-regular, this happens after a finite number of steps. It is a clear that at stage k new equations arise only out of the new equations of stage $k-1$, so that those of Case I in stage k contain derivatives of the perturbations with precisely k differentiations with respect to t. Thus the involution index of the differential algebraic equation (6.51) is indeed equal to the maximal number of differentiations with respect to t applied to the perturbations $\boldsymbol{\epsilon}$. But this is just the definition of the involution index ν_ℓ in direction x^ℓ. □

This result is somewhat surprising. While one surely expects that integrability conditions of the original partial differential equation induce integrability conditions in the differential algebraic equation obtained by semi-discretisation, Theorem 6.7.2 says that obstructions to involution also turn into integrability conditions upon semi-discretisation. Thus even if the original partial differential equation is formally integrable, the differential algebraic equation might contain integrability conditions.

Example 6.7.3. Consider the following linear system for two unknown functions $v(x,t)$, $w(x,t)$:

$$\begin{pmatrix} v \\ w \end{pmatrix}_t = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix}_x, \qquad 0 = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} v \\ w \end{pmatrix}_x \tag{6.56}$$

This is the same system as in Example 5.4.3 written in matrix notation. We know that it is formally integrable but not involutive, as differentiating the constraint equation with respect to the non-multiplicative variable t yields (after a simplification) a new second-order equation as obstruction to involution: $w_{xx} = 0$.

A semi-discretisation with backward differences for the spatial derivatives leads to the differential algebraic equation

$$\Delta x \begin{pmatrix} \dot v_n \\ \dot w_n \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} v_n - v_{n-1} \\ w_n - w_{n-1} \end{pmatrix}, \quad 0 = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} v_n - v_{n-1} \\ w_n - w_{n-1} \end{pmatrix}. \tag{6.57}$$

As predicted by Theorem 6.7.2, it is not involutive, as it has the integrability condition $w_n - 2w_{n-1} + w_{n-2} = 0$ obviously representing an approximation of the obstruction to involution $w_{xx} = 0$ by centred differences. ◁

Corollary 6.7.4. *Let* ν_i *be the involution index of* (6.47) *and assume that all one-forms* $dx^\ell = 0$ *are non-characteristic. Then any differential algebraic equation arising by a semi-discretisation of* (6.47) *has an involution index less than or equal to* ν_i.

Proof. We trevally have $\nu_\ell \leq \nu_i$ for all ℓ. □

For the weakly overdetermined systems introduced in Section 5.6, Theorem 6.7.2 can be strengthened to the statement that the differential algebraic equation obtained by a semi-discretisation with respect to t is formally integrable, *if and only if* the original system of partial differential equations is involutive [165]. And this result does not only hold for semi-discretisations by finite difference methods but also for a spectral approach. Using periodic boundary condition we may expand the solution into a spatial Fourier series with time-dependent coefficients. These coefficients satisfy then a differential algebraic equation which is formally integrable, *if and only if* the original system was involutive.

6.8 Notes

We described the completion process for ordinary differential equations first in terms of the tangent bundle, as in this form it has been rediscovered many times; see e.g. [121, 126, 144, 146, 175]. In some of these works the authors also treat non-autonomous systems within the tangent bundle. This requires some slightly awkward constructions. Usually, the time variable is included in the configuration space, so that one deals with the manifold $\mathcal{U} \times \mathbb{R}$. Then one defines the "restricted" tangent bundle as the set of all those vectors in $T(\mathcal{U} \times \mathbb{R})$ which have a 1 in the last component. The jet bundle offers a much more natural framework for such systems and the restricted tangent bundle is easily recovered with the contact map. An intrinsic proof of the equivalence of the tangent and the jet bundle approach to completion was given in [50].

The notions of continuous and constructive divisions were introduced by Gerdt and Blinkov [56]. Obviously, these are rather technical concepts and especially constructivity is often difficult to prove. They are designed exactly in such a way that our proof of Proposition 6.4.1 works. Combined they ensure that no infinite chains can appear in Algorithm 6.2. Despite its complicated definition, it is obvious that constructivity is a necessary property for the correctness of our algorithms. Otherwise we would be forced to consider also multiplicative indices. For continuity the situation is less clear. While it is sufficient for Proposition 6.3.4 to hold, it is not known whether it is also necessary.

The basic ideas of Algorithm 6.4 are already contained in the Janet–Riquier Theory of differential equations (see the Notes of Chapter 2). Our formulation is essentially due to Blinkov, Gerdt and Zharkov [56, 57, 195] (compare also the notes at the end of Chapter 2). In recent years, they have spent, together with further collaborators, much effort on making the algorithm more efficient and on providing fast implementations. [57] contains a more sophisticated version of Algorithm 6.4 which has in particular for Janet bases some advantages.

As already mentioned in the Notes of Chapter 2, Apel [8] introduced his own concept of an involutive division. His divisions are defined with respect to only one subset $\mathcal{N} \subseteq \mathbb{N}_0^n$. As during the completion the set $\mathcal{H}$ is changed in every step, we always need a new division for the new set $\operatorname{le}_\prec \mathcal{H}$ and Apel discusses some heuristic strategies for choosing the divisions.[7] Whether this additional freedom can be exploited to achieve a higher efficiency is an open question. A definite advantage of this approach is that it is almost trivial to prove the existence of an involutive basis for arbitrary ideals.

However, one should note the following difference. In Apel's approach it is not even possible to formulate the task to determine an involutive basis for a prescribed involutive division, as we do not know the relevant set $\operatorname{le}_\prec \mathcal{H}$, before the completion is over. Thus this approach is only useful, if the goal is to somehow get some Stanley decomposition (or if the involutive theory is only used as an alternative for Buchberger's algorithm to compute Gröbner bases); the role of the involutive division is mainly that of a technical trick to improve efficiency. Up to now, this is not really a disadvantage, as not much is known about special properties of Stanley decompositions for different divisions (with the exception of Pommaret bases!). But one might expect that in the future the involutive division will gain a similar importance as the term order for a Gröbner basis and that for different tasks one is interested in different involutive divisions.

Our algorithm can in principle be applied to nonlinear equations, too, provided the nonlinearities are only polynomial. However, the reduction process is then more subtle; one must resort to *pseudo-reductions* (see e. g. [123]). This makes it much harder to interpret the outcome. In particular, it is possible that one leaves the differential ideal spanned by the original system, as equations must be multiplied with differential polynomials. In the language of differential algebra one obtains a *coherent autoreduced set* and our algorithm may be considered as an alternative to the classical *Kolchin–Ritt algorithm* [93, 151].

A special situation arises for quasi-linear equations. Since they are linear in the derivatives of highest order, all reductions can be performed effectively as linear operations. However, we cannot expect the equation to remain quasi-linear during the completion, as the integrability conditions may be fully non-linear. Provided all appearing equations are quasi-linear, Algorithm 6.4 can be used without problems and produces the correct result. In this case, also no problems with the regularity of the equation appear. A simple example are the Navier–Stokes equations.

The hybrid algorithm of Section 6.5 combining geometric and algebraic ideas was developed in the master's thesis of Hausdorf [70] (see also [71]). Within this thesis, the algorithm was also implemented in the computer algebra system *MuPAD*.

[7] One should note that Apel admits the case $\mathcal{N} = \mathbb{N}_0^n$, i. e. divisions defined on the set of all multi indices. For such divisions, there is no need to change the divisions within a computation and in fact one can formulate the classical involutive divisions like the Pommaret and the Janet division in his framework. In this case his theory largely coincides with ours.

Compared with earlier works [157, 161] using a naive realisation of the Cartan–Kuranishi completion it is much more efficient. The results of Section 2.4 on detecting δ-singular coordinate systems can also be incorporated into the algorithm; for details we refer to [71, Section 5]. For the application to non-linear equations, the same comments as above for the purely algebraic case apply.

The brute force realisation of the Cartan–Kuranishi completion can, in principle, handle equations with polynomial non-linearities fully algorithmically, as all necessary operations can be performed effectively using (algebraic) Gröbner bases. However, this approach requires the determination of a Gröbner basis for the local representation of each appearing differential equation. As the number of variables and polynomials in these bases are rapidly growing with each prolongation, this seems to be feasible only for fairly small examples becoming involutive very soon.

The completion approaches discussed in this chapter are not the only ones; a number of alternatives exists. Gröbner bases are by now a standard tool in commutative algebra and algebraic geometry. Several algorithms for their construction has been designed and their improvement is a very active field of research. More on these bases can be found in Appendix B.2.

As any involutive basis is also a Gröbner basis, it appears natural to compare Algorithm 6.4 with Buchberger's Algorithm B.2. A direct implementation of the latter one is very inefficient; only the use of optimisations like the two Buchberger criteria (Propositions B.2.15 and B.2.16) for detecting unnecessary reductions makes it usable. In contrast, a direct implementation of Algorithm 6.4 yields already reasonable results, although improvements are possible. Some are discussed in [58, 59] and references therein. The second article also contains tables with the results of extensive benchmarks showing that the involutive approach is highly competitive as long as the involutive basis is not considerably larger than the Gröbner basis.

Some simple heuristic explanations of these observations are possible. Algorithm 6.4 implicitly also computes S-polynomials but less than a naive implementation of the Buchberger algorithm. Both Buchberger criteria are not completely but to a large extent automatically contained in the strategy of multiplying only by non-multiplicative variables. For the first criterion this is obvious; for the second one it follows from Proposition 2.6.2. Some further arguments are given in [8].

This strategy also has certain similarities with the so-called *Hilbert driven Buchberger algorithm* introduced by Traverso [179]. In this variant of the Buchberger algorithm one assumes that one knows already the Hilbert function of the ideal and this knowledge is used to avoid further unnecessary reductions. In concrete computations, this idea turns out to be quite effective. Unfortunately, the determination of the Hilbert function requires usually a Gröbner basis. Thus the main application of this technique consists of transforming a Gröbner basis with respect to one term order into one with respect to another term order.

In this context, our algorithm can be understood as follows. Given a basis of the ideal, we make a conjecture about its Hilbert function based on the assignment of multiplicative variables. The multiplication of the elements by their non-multiplicative variables checks whether the conjectured function is already the correct Hilbert function. The algorithm stops, as soon as this is the case. Note that in contrast to the Hilbert driven Buchberger algorithm no a priori knowledge of the Hilbert function is required.

Proposition 6.4.9 says that even if Algorithm 6.4 does not terminate, it yields in a finite number of steps a Gröbner basis for term orders of type ω. In this form, this result is only of theoretical interest, as it is unclear how one should detect the Gröbner basis. Using standard criteria would destroy all potential advantages of our algorithm. For the special case of Pommaret bases, Apel [7] found a simple criterion, so that a variant of Algorithm 6.4 can be used for the construction of Gröbner bases independent of the existence of a finite involutive basis.

In the case of differential equations one would like for many applications to have a simple algorithm to determine all integrability conditions. We have seen in Example 6.2.2 that they significantly simplify the explicit solution of the differential equation. Unfortunately, so far nobody has been able to find a finite criterion for formal integrability. Thus all approaches construct more than just the integrability conditions and they differ in what this "more" is. Most of them have been studied in the context of Lie symmetry theory [20, 129] for the treatment of the determining systems, which are typically large overdetermined linear systems of partial differential equations. Hence a number of references to theoretical works and implementations can be found in the review articles of Hereman [74, 75].

We mention here only three approaches popular in computer algebra. The probably first implementation of the *Janet–Riquier Theory* in a computer algebra system is due to Topunov [178]. Reid [147] developed a *standard form*, also inspired by the Janet–Riquier Theory, and implemented it in MAPLE. The approach is fully algorithmic only for linear equations, but this suffices for applications in symmetry theory. In the more recent work [148] he speaks of a "reduced involutive form" which is not identical with our notion of involution. Mansfield [112] introduced *differential Gröbner bases* by combining ideas from differential algebra and classical Gröbner basis theory; she also provides a MAPLE implementation. One should, however, be careful with the name. There does not yet exist a generally accepted definition of how Gröbner bases should be generalised to differential equations. Mansfield's bases cannot solve the differential ideal membership problem and thus one may argue whether they should be called Gröbner bases.

This last approach is rooted in *differential algebra*, a branch of algebra studying rings and fields with an additional structure, namely derivations, that models in an algebraic manner the basic properties of differentiation. Some classical references are [89, 93, 151]. As an example how this theory may be used in the context of completion, see also [23].

A high worst case complexity is a common problem for all these approaches. As already mentioned, this is not very surprising, as most of them comprise Gröbner basis computations as a special case. Thus the complexity results obtained for Gröbner bases (see Appendix B.2) represent a lower bound for the completion in the differential case. Besides this fairly simple observation not much is known about the complexity of all the mentioned completion algorithms.

The relation between the formal theory and the index concepts in the numerical theory of differential algebraic equations was first noted by Le Vey [104, 105] and Piirilä and Tuomela [136]. Tuomela and Arponen [183, 184] discuss some concrete numerical methods based on these ideas. The indices were generalised to partial differential equations in [163]. These works concentrated on differentiation indices; the estimation of perturbation indices by perturbed differentiation indices was first done in [72] based on earlier works by Campbell and Gear [30].

The definition of indices for partial differential equations is still an active field of research. Most definitions in the literature (see e. g. [31, 108, 109]) are based on first reducing to only one independent variable by semi-discretisations, integral transforms or some other methods. Then the index of the arising differential algebraic equation is used. Obviously, the result depends in general on the used reduction technique. In contrast, our index definitions can be directly applied to the given partial differential equation.

It is of great interest to relate these two approaches, as such reduction methods often form the basis of concrete numerical methods. If the index computed directly from the partial differential equation differs from the index of the differential algebraic equation obtained by the chosen numerical method, this could indicate that the method destroys some structural properties of the original differential equation. But of course this requires that the "right" indices are used in both cases. Currently not much is known about such comparisons. The results we presented in Section 6.7 on the semi-discretisation of linear systems with constant coefficients are the first ones in this direction and stem from [165].

An obvious question is whether similar results hold for linear systems with *variable* coefficients. This is not the case, as differential and difference operators obey different Leibniz rules. The proof of Theorem 6.7.2 depends decisively on the possibility to mimic the operations on the partial differential equation on the differential algebraic equation. This is not possible for systems with variable coefficients, as the evaluation points of the coefficients have to be taken into account.

Hence for arbitrary discretisations the indices of the partial differential equation and of the differential algebraic equation might differ considerably. However, it is conceivable that for *special* discretisations something similar to Theorem 6.7.2 holds. In fact, this should be considered as a very good and desirable property. In recent times, some researchers have studied so-called *mimetic* discretisations (see for example [82]). While their motivation is not an involution analysis, their approach should lead to similar results.

A. Analysis

The first section of this appendix fixes some notations for multi indices that are widely used in the main text. Except the distinction into multi and "repeated" indices these are standard notations in multivariate analysis. The next section recalls those properties of analytic functions that are needed in the proof of the Cauchy–Kovalevskaya Theorem. Then some elementary operations with differential equations like reduction to first order or quasi-linearisation are discussed within the context of the formal theory. Finally, a few facts about the modified Stirling numbers that appear in manipulations of Cartan characters and Hilbert functions are presented.

A.1 Multi Indices

We use mostly the standard multi index notation of multivariate analysis. If we are dealing with n variables $x^1, \dots, x^n$, a *multi index* is an ordered n-tuple $\mu = [\mu_1, \dots, \mu_n] \in \mathbb{N}_0^n$ of integers (we always use square brackets to denote multi indices). The sum of its entries is often called its *length* written $|\mu| = \mu_1 + \cdots + \mu_n$. Its factorial is defined by $\mu! = \mu_1! \cdots \mu_n!$. The expression $(x - y)^\mu$ is a short hand for $(x^1 - y^1)^{\mu_1} \cdots (x^n - y^n)^{\mu_n}$. For a function $\phi(x^1, \dots, x^n)$ we write

$$\frac{\partial^{|\mu|}\phi}{\partial x^\mu} = \frac{\partial^{|\mu|}\phi}{\partial (x^1)^{\mu_1} \cdots \partial (x^n)^{\mu_n}} \,. \tag{A.1}$$

For convenience, we define for the special multi index $0 = [0, \dots, 0]$ that $x^0 = 1$ and $\partial^{|0|}\phi/\partial x^0 = \phi$. Obviously, $0! = 1$ and $|0| = 0$. Other special multi indices are

$$\ell_i = [0, \dots, 0, \ell, 0, \dots, 0] \tag{A.2}$$

where $\ell \in \mathbb{N}$ is the ith entry and all other entries vanish.

The addition of multi indices is defined componentwise, i.e.

$$\mu + \nu = [\mu_1 + \nu_1, \dots, \mu_n + \nu_n] \,. \tag{A.3}$$

With this addition the set $\mathbb{N}^n$ of all multi indices acquires the structure of an Abelian monoid. If we want to increase by one the ith entry of a multi index μ, we can now simply write $\mu + 1_i$ using (A.2).

If k is the smallest value such that $\mu_k \neq 0$, we call it the *class* of the multi index μ: $\operatorname{cls}\mu = k$. In (A.1) this means that there are no differentiations with respect to $x^1, \dots, x^{\operatorname{cls}\mu - 1}$. In principle, the class of the multi index $[0,\dots,0]$ is undefined, but in many situations it is convenient to set $\operatorname{cls}[0,\dots,0] = n$.

Some elementary combinatorics leads to the following results:

$$\left|\left\{\mu \in \mathbb{N}_0^n \mid |\mu| = q\right\}\right| = \binom{n+q-1}{q}\,, \tag{A.4a}$$

$$\left|\left\{\mu \in \mathbb{N}_0^n \mid |\mu| = q \wedge \operatorname{cls}\mu = k\right\}\right| = \binom{n-k+q-1}{q-1}\,, \tag{A.4b}$$

$$\left|\left\{\mu \in \mathbb{N}_0^n \mid 0 \leq |\mu| \leq q\right\}\right| = \binom{n+q}{q}\,. \tag{A.4c}$$

A multi index μ of length q and class k is obtained by first setting μ_k to a value $1 \leq q' \leq q$ and then adding an arbitrary multi index in $\mathbb{N}_0^{n-k}$ of length $q - q'$. Thus the second line arises from summing up the first one from 0 to $q-1$ (the possible values of $q - q'$) after replacing n by $n-k$ and using a classical identity for binomial coefficients. In a similar manner, the third line stems from summing up the first one from 0 to q and applying the same identity.

On a few occasions we use an alternative multi index notation, which we call for lack of a better name *repeated index* notation. We distinguish it from multi indices by parentheses instead of brackets. A repeated index ρ of length q always contains exactly q entries with values between 1 and n: $\rho = (\rho_1, \dots, \rho_q)$. The analogue to (A.1) is

$$\frac{\partial^q \phi}{\partial x^\rho} = \frac{\partial^q \phi}{\partial x^{\rho_1} \cdots \partial x^{\rho_q}}\,. \tag{A.5}$$

As we always assume that partial derivatives commute, there is obviously no need to order the entries of a repeated index: any permutation of the entries leaves the index unchanged.

A.2 Real-Analytic Functions

Much of the formal theory is based on formal power series. If such a series converges, it defines an analytic function. In this appendix we collect some basic facts about these functions. Throughout we consider real-valued functions defined in some domain $\Omega \subset \mathbb{R}^n$.

Definition A.2.1. *The function* $f : \Omega \to \mathbb{R}$ *is called* real-analytic *at the point* $x \in \Omega$*, if there exists a neighbourhood* $x \in U \subseteq \Omega$ *such that for all* $y \in U$ *the function* f *can be represented by a power series*

$$f(y) = \sum_{|\mu|=0}^{\infty} a_\mu (y-x)^\mu \,. \tag{A.6}$$

f *is real-analytic in the whole domain* Ω*, if it is real-analytic at every point* $x \in \Omega$*. The set of all functions that are real-analytic in the domain* Ω *is denoted by* $C^\omega(\Omega)$*.*

Obviously, any analytic function is smooth: $C^\omega(\Omega) \subset C^\infty(\Omega)$. The converse is, however, not true. Consider the function $f : \mathbb{R} \to \mathbb{R}$ defined by

$$f : x \mapsto \begin{cases} 0 & x \leq 0 \,, \\ e^{-\frac{1}{x^2}} & x > 0 \,. \end{cases} \tag{A.7}$$

It is easy to see that f is smooth on $\mathbb{R}$. But it is not analytic at the origin. Analyticity is a rather strong property, as the following proposition implies that if we know an analytic function and all its derivatives at one point of Ω, we know it in all Ω. This is sometimes called *unique continuation.*

Proposition A.2.2. *Let* f *and* g *be two functions that are real-analytic in the domain* Ω*. If there exists a point* $x \in \Omega$ *such that*

$$\frac{\partial^{|\mu|} f}{\partial x^\mu}(x) = \frac{\partial^{|\mu|} g}{\partial x^\mu}(x) \qquad \forall \mu \in \mathbb{N}_0^n \,, \tag{A.8}$$

then $f \equiv g$ *in* Ω*.*

Proof. We use a simple topological argument. Consider the set

$$\Sigma = \left\{ y \in \Omega \mid \frac{\partial^{|\mu|} f}{\partial x^\mu}(y) = \frac{\partial^{|\mu|} g}{\partial x^\mu}(y) \quad \forall \mu \in \mathbb{N}_0^n \right\} . \tag{A.9}$$

Σ is relatively closed in Ω, as it is defined by equations. On the other hand, Σ must be open, because at any point $x \in S$ the two functions f and g possess identical power series and therefore must be identical in an open neighbourhood. As, by definition, a domain is a connected open set, we must have $\Sigma = \Omega$. □

Another important and characteristic property of analytic functions is the existence of a bound on their derivatives. For a smooth function f defined in a neighbourhood of an arbitrary but fixed point $x \in \Omega$, we say that $f \in C_{M,r}(x)$, if its derivatives at x satisfy *Cauchy's estimate*

$$\left| \frac{\partial^{|\mu|} f}{\partial x^\mu}(x) \right| \leq M \frac{|\mu|!}{r^{|\mu|}} \qquad \forall \mu \in \mathbb{N}^n \,. \tag{A.10}$$

Proposition A.2.3. *Let* $f \in C^\infty(\Omega)$. *Then* $f \in C^\omega(\Omega)$, *if and only if for every compact subset* $\Sigma \subset \Omega$ *there exist two numbers* M *and* r *such that* $f \in \mathcal{C}_{M,r}(x)$ *for all* $x \in \Sigma$.

Proof. Let us first assume that f is analytic. Then it is not difficult to show that for all $x \in \Omega$ its derivatives satisfy in a neighbourhood $\mathcal{U}(x)$ the estimate (A.10) with constants $M(x)$ and $r(x)$ (see e.g. [149, Section 2.2.1]). As any compact subset Σ can be covered by a finite number of such neighbourhoods, we can take the maximum of the $M(x)$ and the minimum of the $r(x)$ and obtain the assertion.

For the converse, take a point $x \in \Omega$ and choose as compact set Σ a closed ball of radius R with its centre at x (obviously, the radius R must be chosen small enough so that $\Sigma \subset \Omega$). Then, by assumption, there exist numbers M and r such that $f \in \mathcal{C}_{M,r}(x)$. Consider an arbitrary point $y \in \Sigma$ with $\delta = \sum_{i=1}^{n} |x_i - y_i| < \min(r, R)$ and introduce for $0 \leq \tau \leq 1$ the function $F(\tau) = f\bigl(x + \tau(y - x)\bigr)$. We apply Taylor's Theorem to $F(\tau)$ and obtain

$$f(y) = F(1) = \sum_{k=0}^{q-1} \frac{1}{k!}\frac{d^kF}{d\tau^k}(1) + \frac{1}{q!}\frac{d^qF}{d\tau^q}(\bar{\tau}) \tag{A.11}$$

for some value $0 \leq \bar{\tau} \leq 1$. A straightforward computation yields for any order $q \geq 0$ the following estimate of the remainder term

$$\Bigl|\frac{1}{q!}\frac{d^qF}{d\tau^q}(\bar{\tau})\Bigr| \leq M\left(\frac{\delta}{r}\right)^q . \tag{A.12}$$

As by construction $\delta < r$, it tends to zero and we find that f can be represented in a neighbourhood of x by a convergent series of the form (A.6). □

Finally, we recall that an analytic function f is *majorised* in a neighbourhood of the point x by another analytic function g, written $f \preceq g$, if the inequality

$$\Bigl|\frac{\partial^{|\mu|}f}{\partial x^\mu}(x)\Bigr| \leq \Bigl|\frac{\partial^{|\mu|}g}{\partial x^\mu}(x)\Bigr| \tag{A.13}$$

holds for all multi indices $\mu \in \mathbb{N}_0^n$. As a simple example we note that any function in $\mathcal{C}_{M,r}(x)$ is majorised by the function

$$g(x) = \frac{Mr}{r - (x_1 + \cdots + x_n)} , \tag{A.14}$$

as the Taylor coefficients of g at the origin are just given by the right hand side of Cauchy's estimate (A.10).

A.3 Elementary Operations with Differential Equations

In this section we study three elementary operations with differential equations: rewriting a higher-order equation as a first-order one, quasi-linearisation and transformation to one dependent variable. In particular, we determine what happens, if we apply the operation to an involutive equation.

Reduction to first order. It is a classical operation to rewrite a differential equation of order q as a first-order equation through the introduction of additional dependent variables representing the derivatives of the original dependent variables. Geometrically, this is based on the identification of $J_q\mathcal{E}$ with a submanifold of $J_1(J_{q-1}\mathcal{E})$ based on the immersion $\iota_{q,1}$ (see Section 1.2). We define intrinsically $\tilde{\mathcal{R}}_1 = \iota_q(\mathcal{R}_q) \subseteq J_1(J_{q-1}\mathcal{E})$.

In local coordinates we get the following picture. Let $(x^i, u^\alpha, p^\alpha_\mu)$ be our usual coordinates on $J_q\mathcal{E}$; on $J_1(J_{q-1}\mathcal{E})$ we use $(x^i, \tilde{u}^\alpha_\nu, \tilde{p}^\alpha_{\nu,i})$ with $0 \leq |\nu| < q$. Then $\iota_{q,1}(x^i, u^\alpha, p^\alpha_\mu) = (x^i, p^\alpha_\nu, p^\alpha_{\nu+1_i})$. If the differential equation $\mathcal{R}_q$ is locally represented by the system $\Phi^\tau(x^i, u^\alpha, p^\alpha_\mu) = 0$, a local representation of $\tilde{\mathcal{R}}_1$ is given by

$$\tilde{\mathcal{R}}_1 : \begin{cases} \tilde{\Phi}^\tau(x^i, \tilde{u}^\alpha_\mu, \tilde{p}^\alpha_{\mu,i}) = 0 \,, & \\ \quad \tilde{p}^\alpha_{\mu,i} = \tilde{u}^\alpha_{\mu+1_i} \,, & |\mu| < q \,, \\ \tilde{p}^\alpha_{\mu+1_i,j} = \tilde{p}^\alpha_{\mu+1_j,i} \,, & |\mu| = q-1 \,. \end{cases} \tag{A.15}$$

The second and third line represent the equations defining the submanifold $\iota_{q,1}(J_q\mathcal{E}) \subset J_1(J_{q-1}\mathcal{E})$. The function $\tilde{\Phi}^\tau$ is not uniquely defined, as in general several possibilities exist to express a derivative p^α_μ by one of the new coordinates. One way which allows us to compute easily the indices of the symbol $\tilde{\mathcal{M}}_1$ is to use the mapping

$$p^\alpha_\mu \longmapsto \begin{cases} \tilde{u}^\alpha_\mu & \text{for } |\mu| \leq q-1 \,, \\ \tilde{p}^\alpha_{\mu-1_k,k} & \text{for } |\mu| = q, \ \operatorname{cls}\mu = k \,. \end{cases} \tag{A.16}$$

Proposition A.3.1. *Let $\tilde{\alpha}^{(k)}_1$ be the Cartan characters of $\tilde{\mathcal{R}}_1$ and $\alpha^{(k)}_q$ those of $\mathcal{R}_q$. Then $\tilde{\alpha}^{(k)}_1 = \alpha^{(k)}_q$ for $1 \leq k \leq n$ and there exists a one-to-one correspondence between the solutions*[1] *of $\tilde{\mathcal{R}}_1$ and $\mathcal{R}_q$. The equation $\mathcal{R}_q$ is involutive, if and only if $\tilde{\mathcal{R}}_1$ is involutive.*

Proof. Due to the substitution rule (A.16) the first line of (A.15) yields $\beta^{(k)}_q$ equations of class k, the same value as in $\mathcal{R}_q$. The second line represents

[1] As usual we are only considering either smooth or formal solutions.

$\dim J_{q-2}\mathcal{E}$ equations of class k and a little combinatorial calculation shows that the last line leads to additionally $\dim\left(S_{q-1}(T^*\mathcal{X})\otimes V\mathcal{E}\right)-\binom{q+n-k-1}{q-1}$ equations of class k. Adding these values we get

$$\tilde{\beta}_1^{(k)}=\dim J_{q-2}\mathcal{E}+\dim\left(S_{q-1}(T^*\mathcal{X})\otimes V\mathcal{E}\right)-\alpha_q^{(k)} \tag{A.17}$$

and thus for the Cartan characters

$$\tilde{\alpha}_1^{(k)}=\dim J_{q-1}\mathcal{E}-\tilde{\beta}_1^{(k)}=\alpha_q^{(k)}\,. \tag{A.18}$$

If $\mathbf{x}$ forms a δ-regular coordinate system for $\mathcal{R}_q$, then it does so for $\tilde{\mathcal{R}}_1$, too, and vice versa. The additional equations in (A.15) are in a form that leads to maximal values for the number of equations of class n and so on. Thus only the equations $\tilde{\Phi}^\tau=0$ could make trouble. But again due to our substitution rule (A.16) we see that if we get maximal values for the indices of $\mathcal{M}_q$ we get them also for $\tilde{\mathcal{M}}_1$ and vice versa.

These considerations imply furthermore that involution of the symbol of one of the equations is equivalent to involution of the symbol of the other one. It is easy to see that the second and third line in (A.15) alone form an involutive system and that there arise no integrability conditions between them and the equations $\tilde{\Phi}^\tau=0$. But if the equations $\Phi^\tau=0$ generate integrability conditions, then the same happens with the equations $\tilde{\Phi}^\tau=0$ and vice versa.

The relation between the solution spaces of $\mathcal{R}_q$ and $\tilde{\mathcal{R}}_1$, respectively, is obvious. If a section $\sigma:\mathcal{X}\to\mathcal{E}$ is a solution of $\mathcal{R}_q$, then $j_{q-1}\sigma:\mathcal{X}\to J_{q-1}\mathcal{E}$ is a solution of $\tilde{\mathcal{R}}_1$. Conversely, the second and third line of (A.15) ensure that every solution $\tilde{\sigma}:\mathcal{X}\to J_{q-1}\mathcal{E}$ of $\tilde{\mathcal{R}}_1$ is a section of the form $\tilde{\sigma}=j_{q-1}\sigma$ with $\sigma:\mathcal{X}\to\mathcal{E}$ a solution of $\mathcal{R}_q$. □

Quasi-Linearisation. The above mentioned method for a reduction to first order is not the only possible one. We may apply it not to $\mathcal{R}_q$ itself but to its prolongation $\mathcal{R}_{q+1}$. Recall that a prolongation leads always to a quasi-linear equation. This implies that the arising first-order equation $\bar{\mathcal{R}}_1\subset J_1J_q\mathcal{E}$ is quasi-linear, too. Note that the original equations are now transformed into algebraic constraints.

Proposition A.3.2. *Let $\bar{\alpha}_1^{(k)}$ be the Cartan characters of $\bar{\mathcal{R}}_1$ and $\alpha_q^{(k)}$ those of $\mathcal{R}_q$. Then $\bar{\alpha}_1^{(k)}=\sum_{i=k}^n\alpha_q^{(i)}$ for $1\le k\le n$ and there exists a one-to-one correspondence between the solutions of $\bar{\mathcal{R}}_1$ and $\mathcal{R}_q$. If $\mathcal{R}_q$ is involutive, then $\bar{\mathcal{R}}_1$ is involutive, too. If $\bar{\mathcal{R}}_1$ is involutive, then $\mathcal{R}_q$ is formally integrable and $\mathcal{R}_{q+1}$ is involutive.*

Proof. Most of the assertions follow immediately from Proposition A.3.1. According to Proposition 4.2.3, the Cartan characters of $\mathcal{R}_{q+1}$ and $\mathcal{R}_q$ are related by $\alpha_{q+1}^{(k)} = \sum_{i=k}^{n} \alpha_q^{(i)}$. If $\mathcal{R}_q$ is involutive, then $\mathcal{R}_{q+1}$ and hence $\bar{\mathcal{R}}_1$ are involutive, too.

For the final assertion, we first note that if $\bar{\mathcal{R}}_1$ is involutive, then by our previous results $\mathcal{R}_{q+1}$ is involutive. Applying our transformation to $\mathcal{R}_q$ yields a differential equation $\bar{\mathcal{R}}_0 \supseteq \bar{\mathcal{R}}_0^{(1)}$. As the transformation of $\mathcal{R}_{q+1}$ yields $\bar{\mathcal{R}}_1$ and as, because of the involution of $\bar{\mathcal{R}}_1$, the equality $(\bar{\mathcal{R}}_0^{(1)})_{+1} = \bar{\mathcal{R}}_1$ holds, we must in fact have $\bar{\mathcal{R}}_0 = \bar{\mathcal{R}}_0^{(1)}$. But this implies that $\mathcal{R}_q^{(1)} = \mathcal{R}_q$. Hence $\mathcal{R}_q$ is at least formally integrable. □

Example A.3.3. We cannot conclude that involution of $\bar{\mathcal{R}}_1$ entails involution of $\mathcal{R}_q$, as its construction uses a prolongation. Consider again the equation of Example 3.2.2

$$\mathcal{R}_2 : \left\{ u_{yy} = 0 , \quad u_{xx} = 0 . \right. \tag{A.19}$$

In order to make the notation less cumbersome, we introduce the following variable names in $J_2\mathcal{E}$: $r = u_{xx}$, $s = u_{xy}$, $t = u_{yy}$, $v = u_x$ and $w = u_y$. The first order equation $\bar{\mathcal{R}}_1 \subset J_1(J_2\mathcal{E})$ has a local representation of the form

$$\bar{\mathcal{R}}_1 : \begin{cases} t_y = 0 , & t_x = 0 , \\ s_x = 0 , & r_x = 0 , \\ r = 0 , & t = 0 , \\ r_y = s_x , & s_y = t_x , \\ v_y = s , & v_x = r , \\ w_y = t , & w_x = s , \\ u_y = w , & u_x = v . \end{cases} \tag{A.20}$$

The first six equations represent $\mathcal{R}_3$ in the new coordinates; the remaining equations define the submanifold $\iota_{3,1}(J_3\mathcal{E}) \subset J_1(J_2\mathcal{E})$.

One easily checks that $\bar{\mathcal{R}}_1$ is involutive. We know from Example 3.2.2 that $\mathcal{R}_2$ is only formally integrable but not involutive whereas $\mathcal{R}_3$ is involutive. ◁

Transformation to one dependent variable. It is always possible to transform a differential equation $\mathcal{R}_q$ in m dependent variables u^α and n independent variables x^i into one with only one dependent variable v. However, we must introduce m additional independent variables y^α; the order is raised by one and the relation between the transformed and the original equation is more complicated than in the above reduction to first order.

The basic idea is very simple: we define

$$v = \sum_{\alpha=1}^{m} y^\alpha u^\alpha . \tag{A.21}$$

Then for any multi-index $\mu \in \mathbb{N}_0^n$ the identity

$$\frac{\partial^{|\mu|} u^\alpha}{\partial x^\mu} = \frac{\partial^{|\mu|+1} v}{\partial y^\alpha \partial x^\mu} \tag{A.22}$$

holds and the differential equation $\mathcal{R}_q$ can be rewritten as an equation of order $q+1$ in the single dependent variable v. The definition (A.21) implies that we must add the equations

$$\frac{\partial^2 v}{\partial y^\alpha \partial y^\beta} = 0\,, \qquad 1 \leq \alpha \leq \beta \leq m \tag{A.23}$$

plus all their prolongations up to order $q+1$. This determines a differential equation $\bar{\mathcal{R}}_{q+1}$.

The equations (A.23) imply that any solution of the transformed equation $\bar{\mathcal{R}}_{q+1}$ is of the form

$$v(\mathbf{x}, \mathbf{y}) = \sum_{\alpha=1}^{m} y^\alpha u^\alpha(\mathbf{x}) + \Lambda(\mathbf{x}) \tag{A.24}$$

where $\mathbf{u}(\mathbf{x})$ is a solution of the original equation $\mathcal{R}_q$ and $\Lambda(\mathbf{x})$ is an arbitrary function. Obviously, this arbitrary function is of no interest for us; thus we may consider it as a *gauge symmetry* of $\bar{\mathcal{R}}_{q+1}$.

Proposition A.3.4. *Let $\bar{\alpha}_{q+1}^{(k)}$ for $1 \leq k \leq n+m$ be the Cartan characters of $\bar{\mathcal{R}}_{q+1}$ where we set $x^k = y^\alpha$ for $k = n+\alpha$ and $\alpha_q^{(k)}$ for $1 \leq k \leq n$ the Cartan characters of $\mathcal{R}_q$. Then*

$$\bar{\alpha}_{q+1}^{(k)} = \begin{cases} 0 & \text{for } n < k \leq n+m\,, \\ \alpha_q^{(k)} + \binom{q+n-k}{q} & \text{for } 1 \leq k \leq n\,. \end{cases} \tag{A.25}$$

$\bar{\mathcal{R}}_{q+1}$ is involutive, if and only if $\mathcal{R}_q$ is involutive.

Proof. Several ways exist to prove this proposition. One may proceed head-on and determine the indices of $\bar{\mathcal{R}}_{q+1}$; they are given by

$$\bar{\beta}_{q+1}^{(k)} = \begin{cases} \binom{q+n+m-k}{q} & \text{for } n < k \leq n+m\,, \\ \beta_q^{(k)} + \frac{m(m+1)}{2} \sum_{i=k}^{n+m} \binom{q+n+m-i-2}{q-2} & \text{for } 1 \leq k \leq n\,. \end{cases} \tag{A.26}$$

The binomial coefficients count the contribution of (A.23) and their prolongations. This yields (A.25) after some computations involving non-trivial identities of binomial coefficients.

A simpler and more direct way exploits (A.24). It shows that the solution of $\bar{\mathcal{R}}_{q+1}$ does not contain any arbitrary functions of the new variables $\mathbf{y}$. Thus no parametric derivatives exist that are pure $\mathbf{y}$-derivatives and $\bar{\alpha}^{(k)}_{q+1} = 0$ for $k > n$. Concerning the $\mathbf{x}$-dependency we see that we recover all arbitrariness of $\mathcal{R}_q$ and have in addition the gauge symmetry. It contributes $\binom{q+n-k}{q}$ parametric derivatives of order $q+1$ and class k for $1 \leq k \leq n$. Hence $\bar{\alpha}^{(k)}_{q+1} = \alpha^{(k)}_q + \binom{q+n-k}{q}$ for $1 \leq k \leq n$.

The assertion about involution is trivial. (A.23) represents obviously an involutive equation; if it was not for the $\mathbf{x}$-dependency, it would be a finite type equation. Cross-derivatives between the original and these added equations vanish trivially. Thus the only source for possible integrability conditions or obstructions to involution is the original equation and there the $\mathbf{y}$-derivatives play no role at all. □

Using Hilbert polynomials, it is much easier to express (A.25). Let $\bar{H}(s)$ be the Hilbert polynomial of $\bar{\mathcal{R}}_{q+1}$, $H(s)$ the one of $\mathcal{R}_q$ and $G(s)$ the one of the Lie pseudogroup defined by the Lie equation

$$\mathcal{G}_1 : \begin{cases} \dfrac{\partial \bar{v}}{\partial v} = 1 \,, \quad \dfrac{\partial \bar{v}}{\partial y^\alpha} = 0 \,, \qquad 1 \leq \alpha \leq m \ . \end{cases} \tag{A.27}$$

It describes the gauge symmetry $\bar{v} = v + \Lambda(\mathbf{x})$ of $\bar{\mathcal{R}}_{q+1}$. Then we get

$$\bar{H}(s) = H(s-1) + G(s) \tag{A.28}$$

with $G(s) = \binom{q+n-1}{q}$.

A.4 Modified Stirling Numbers

The modified Stirling numbers appear in some manipulations of the Cartan characters and the Hilbert polynomial. Thus they are of a combinatorial origin. We introduce them in a fashion similar to the classical Stirling numbers of the first kind [64]. Let us denote by $x^{\overline{n}} = x(x+1)\cdots(x+n-1)$ the nth rising factorial powers (sometimes also called Pochhammer's symbol). Then we can define the Stirling numbers of the first kind[2] $S^{(k)}_n$ for $n \geq k$ by the identity

$$x^{\overline{n}} = \sum_{k=0}^{n} S^{(k)}_n x^k \ . \tag{A.29}$$

The modified Stirling numbers $s^{(n)}_k(q)$ arise, if we write $x = q + r + 1$ for some non-negative integer q in (A.29) and then express the right hand side as a polynomial in r. Table A.1 on page 242 contains some explicit values.

[2] We follow here the notation of Abramowitz and Stegun [3] for the Stirling numbers of the first kind. Graham et al. [64] propose the notation $\left[{n \atop k}\right]$ which we, however, use already in a different context.

Definition A.4.1. *The* modified Stirling numbers $s_k^{(n)}(q)$ *are defined for all non-negative integers* n, k, q *with* $n \geq k$ *by the identity*

$$(q+r+1)^{\overline{n}} = \sum_{k=0}^{n} s_{n-k}^{(n)}(q) r^k \tag{A.30}$$

(note the different position of the scripts compared with (A.29) *and the use of* $n-k$ *instead of* k*).*

We are mainly interested in the fact that the modified Stirling numbers appear in the expansion of binomial coefficients as polynomials. Dividing (A.30) by $n!$ yields immediately

$$\binom{q+r+n}{q+r} = \frac{1}{n!} \sum_{k=0}^{n} s_{n-k}^{(n)}(q)\, r^k \,. \tag{A.31}$$

This also explains why we have used $q+r+1$ in (A.30) instead of the perhaps more natural $q+r$. Using the identity $S_{n+1}^{(k+1)} = \sum_{i=0}^{k} \binom{i}{k} S_n^{(i)}$ [64] one can show that $s_k^{(n)}(0) = S_{n+1}^{(n-k+1)}$.

With the help of the elementary symmetric polynomials one can easily derive a closed form expression for them. The elementary symmetric polynomial $\sigma_k^{(n)}$ is a polynomial of degree k in n variables $x^1, \dots, x^n$ that remains invariant under arbitrary permutations of the variables [40, Chapt. 7]. Thus $\sigma_1^{(n)} = x^1 + \cdots + x^n$, $\sigma_2^{(n)} = x^1x^2 + x^1x^3 + \cdots + x^2x^3 + \cdots + x^{n-1}x^n$ and so on until $\sigma_n^{(n)} = x^1x^2 \cdots x^n$. Now we can write

$$s_k^{(n)}(q) = \begin{cases} 1 & \text{for } k = 0\,, \\ \sigma_k^{(n)}(q+1, q+2, \dots, q+n) & \text{for } 0 < k \leq n\,. \end{cases} \tag{A.32}$$

The modified Stirling numbers inherit from the elementary symmetric polynomials the recursion relation

$$s_k^{(n)}(q) = s_k^{(n-1)}(q) + (q+n)\, s_{k-1}^{(n-1)}(q)\,. \tag{A.33}$$

One can use this relation to construct explicit expressions for $s_k^{(n)}(q)$. They become fairly lengthy for larger k.

Lemma A.4.2.

$$s_n^{(n)}(q) = (q+n)!/q!\,, \tag{A.34a}$$

$$s_1^{(n)}(q) = \frac{1}{2} n(n+2q+1)\,, \tag{A.34b}$$

$$s_2^{(n)}(q) = \frac{1}{24} n(n-1)\Big(3n^2 + 12qn + 5n + 12q^2 + 12q + 2\Big)\,. \tag{A.34c}$$

Proof. (A.34a) and (A.34b) follow immediately from Definition A.4.1. For (A.34c) we apply (A.33)

$$
\begin{aligned}
s_2^{(n)}(q) &= s_2^{(n-1)}(q) + (q+n)\, s_1^{(n-1)}(q) \\
&= \cdots = \sum_{l=2}^{n-1} (q+l+1)\, s_1^{(l)}(q) + s_2^{(2)}(q) \\
&= \frac{1}{2} \sum_{l=1}^{n-1} (q+l+1) \Big[l^2 + (2q+1) l \Big] .
\end{aligned}
$$

Evaluation of this sum yields the claim. □

One can also deduce recursion formulae relating the modified Stirling numbers for different values of q or express $s_k^{(n)}(q)$ directly through $s_l^{(m)}(0)$ with $0 \leq m \leq n$. Like most properties of these numbers they stem from a well-known formula for binomial coefficients.

Lemma A.4.3.

$$
s_{n-k+1}^{(n)}(q) = n\, s_{n-k}^{(n-1)}(q) + s_{n-k+1}^{(n)}(q-1) . \tag{A.35}
$$

Proof. The basis of this lemma is the following identity coming directly from the Pascal triangle

$$
\binom{q+r+n}{q+r} = \sum_{k=0}^{n} \binom{q+r+k-1}{q+r-1} . \tag{A.36}
$$

Plugging in (A.31) yields after a few simple index manipulations yet another recursion relation for the modified Stirling numbers

$$
s_k^{(n)}(q) = \sum_{j=0}^{k} \frac{n!}{(n-j)!} s_{k-j}^{(n-j)}(q-1) . \tag{A.37}
$$

Applying this relation in (A.35) proves at once the lemma. □

In order to express $s_k^{(n)}(q)$ through the modified Stirling numbers with $q = 0$, we must generalise (A.37). Each binomial coefficient $\binom{r+k}{k}$ represents by (A.31) a polynomial in r of degree k. Thus every polynomial in r of degree n can be written as a linear combination of $\binom{r}{0}, \binom{r+1}{1}, \dots, \binom{r+n}{n}$. Especially there exist numbers $d_k^{(n)}(q)$ such that

$$
\binom{q+r+n}{q+r} = \sum_{k=0}^{n} d_k^{(n)}(q) \binom{r+k}{k} . \tag{A.38}
$$

Applying repeatedly (A.36) one can express these coefficients as nested sums

$$d_k^{(n)}(q) = \sum_{k_1=k}^{n} \sum_{k_2=k_1}^{n} \cdots \sum_{k_{q-1}=k_{q-2}}^{n} 1 . \tag{A.39}$$

For lower values of q these sums are easily evaluated. We omit, however, explicit expressions. Now we obtain similarly as above

$$s_k^{(n)}(q) = \sum_{j=0}^{k} \frac{n!}{(n-j)!} d_k^{(n)}(q) s_{k-j}^{(n-j)}(0) . \tag{A.40}$$

Finally, we list in Table A.1 some values of the modified Stirling numbers; the rows correspond to different values of n, the columns of k. We remark that many such integer sequences can be found in the encyclopedia [168] of which also an extended electronic version exists on the web at `www.research.att.com/~njas/sequences`. It contains a number of other generalisations of the classical Stirling numbers (with references).

	q=0				q=1				q=2				q=3			
$n\backslash k$	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
1	1				2				3				4			
2	3	2			5	6			7	12			9	20		
3	6	11	6		9	26	24		12	47	60		15	74	120	
4	10	35	50	24	14	71	154	120	18	119	342	360	22	179	638	840

Table A.1. Some values of the modified Stirling numbers

B. Algebra

A proper understanding of involution in the form introduced in Chapter 2 requires certain ideas from commutative algebra which are collected in this appendix. Some classical references for all mentioned topics are [40, 41, 47]; older textbooks on commutative algebra are usually less constructive. The introductory textbook [99] puts strong emphasis on computational issues (using the computer algebra system CoCoA).

The first section quickly reviews some basic algebraic structures like monoids, rings, algebras etc. Much of the material is elementary. Special emphasis is put on modules; they are central for the algebraic analysis of the symbol of a differential equation. In particular, we introduce the Hilbert function and polynomial, as similar concepts appear for differential equations in Section 4.2. A simple determination of the Hilbert function may be considered as one motivation for the introduction of involutive bases.

The second section gives an introduction into the theory of Gröbner bases of polynomial ideals and modules. These bases represent a fundamental tool in commutative algebra and algebraic geometry. Like most computations in a Euclidian vector space become significantly easier, if one chooses an orthonormal basis, almost any computation with ideals profits from the knowledge of Gröbner bases. Standard references are [4, 15, 40]; most books on computer algebra also contain at least an introduction.

For us Gröbner bases are so important, as much of the theory of involutive bases is modelled on them and an involutive basis is a Gröbner basis with additional combinatorial properties. We recall the famous Buchberger algorithm for determining a Gröbner basis in order to allow for a comparison with the completion algorithm derived in Section 6.4. However, deeper questions of the optimisation of the algorithm are not discussed.

Similarly, for lack of space applications of Gröbner bases are omitted; many can be found in above references. They include in particular the problem of determining points on the variety defined by an ideal, in other words solving systems of polynomial equations. Other applications are the union or intersection of ideals, computations with cosets, elimination problems, ...

The third section recalls some aspects of syzygies and free resolutions. It discusses the theorem of Schreyer on a Gröbner basis of the syzygy module and the construction of the syzygy resolution.

B.1 Some Basic Algebraic Structures

Definition B.1.1. *A set $\mathcal{S}$ with an operation $\cdot : \mathcal{S} \times \mathcal{S} \to \mathcal{S}$ is a* semigroup, *if $\cdot$ is associative, i. e. if $s_1 \cdot (s_2 \cdot s_3) = (s_1 \cdot s_2) \cdot s_3$ for all $s_1, s_2, s_3 \in \mathcal{S}$. The semigroup is* commutative *or* Abelian, *if $s_1 \cdot s_2 = s_2 \cdot s_1$ for all $s_1, s_2 \in \mathcal{S}$. A semigroup $\mathcal{S}$ is a* monoid, *if there exists a neutral element $e \in \mathcal{S}$ such that $e \cdot s = s \cdot e = s$ for all $s \in \mathcal{S}$. A monoid $\mathcal{S}$ is a* group, *if every $s \in \mathcal{S}$ has an inverse $s^{-1} \in \mathcal{S}$ such that $s \cdot s^{-1} = e$.*

The for us most important example of an Abelian monoid is given by the multi indices $(\mathbb{N}_0^n, +)$ discussed in Appendix A.1. Its neutral element is $[0, \dots, 0]$. The integers $\mathbb{Z}$ are both a monoid and a group: a monoid with respect to multiplication and a group with respect to addition; the respective neutral elements are 1 and 0. A non-commutative monoid $(C^\infty(\mathbb{R}), \circ)$ is defined by the smooth functions $\mathbb{R} \to \mathbb{R}$ with the composition as operation and the identity as neutral element. The invertible functions form a group.

Let $(\mathcal{S}, \cdot)$ be a monoid and $\mathcal{J} \subseteq \mathcal{S}$ a subset such that $\mathcal{S} \cdot \mathcal{J} \subseteq \mathcal{J}$, i.e. if we "multiply" an element of $\mathcal{J}$ by an arbitrary element of $\mathcal{S}$, we always obtain again an element of $\mathcal{J}$. Then we call $\mathcal{J}$ a *(left) monoid ideal*. In case that the monoid is not Abelian we may analogously define right monoid ideals by the condition $\mathcal{J} \cdot \mathcal{S} \subseteq \mathcal{J}$ or double sided monoid ideals by $\mathcal{S} \cdot \mathcal{J} \cdot \mathcal{S} \subseteq \mathcal{J}$.

Definition B.1.2. *A* (left) monoid module *over a monoid $\mathcal{S}$ with neutral element e is a set $\mathcal{M}$ together with an operation $\star : \mathcal{S} \times \mathcal{M} \to \mathcal{M}$ so that*

- $\forall m \in \mathcal{M} : \ e \star m = m$;
- $\forall s_1, s_2 \in \mathcal{S}, \forall m \in \mathcal{M} : \ (s_1 \cdot s_2) \star m = s_1 \star (s_2 \star m)$.

A trivial example of a monoid module over $\mathcal{S}$ is a monoid ideal $\mathcal{J} \subseteq \mathcal{S}$ with $\star = \cdot$. If $\mathcal{M}$ is a monoid module, a subset $\mathcal{G} \subset \mathcal{M}$ is a *generating set* of $\mathcal{M}$, if $\mathcal{M}$ is the smallest monoid module containing $\mathcal{G}$. In this case, any element of $\mathcal{M}$ can be represented in the form $s \star g$ with some $s \in \mathcal{S}$ and $g \in \mathcal{G}$. We write $\mathcal{M} = \langle \mathcal{G} \rangle$. The monoid module $\mathcal{M}$ is *finitely generated*, if it possesses a finite generating set.

The monoid $(\mathcal{S}, \cdot)$ is *Noetherian*, if every monoid ideal $\mathcal{J} \subseteq \mathcal{S}$ is finitely generated. Equivalently, every ascending chain of monoid ideals $\mathcal{J}_1 \subseteq \mathcal{J}_2 \subseteq \mathcal{J}_3 \subseteq \cdots \subseteq \mathcal{S}$ becomes stationary, i.e. an index $N \geq 1$ exists such that $\mathcal{J}_N = \mathcal{J}_{N+1} = \mathcal{J}_{N+2} = \cdots$. The monoid $(\mathbb{N}_0^n, +)$ is Noetherian. This can be shown by an induction on n. If we identify the multi index $\mu \in \mathbb{N}_0^n$ with the term x^μ, this result is often called *Dickson's Lemma* (it is usually formulated differently but is equivalent to this statement).

Definition B.1.3. *A* ring *is a set* $\mathcal{R}$ *with two operations* $+, \cdot : \mathcal{R} \times \mathcal{R} \to \mathcal{R}$ *such that* $(\mathcal{R}, +)$ *is an Abelian group, the neutral element of which we denote by* 0*, and* $(\mathcal{R}, \cdot)$ *is a semigroup. Furthermore, the following distributive laws must hold:*

- $\forall r_1, r_2, r_3 \in \mathcal{R}: \ r_1 \cdot (r_2 + r_3) = r_1 \cdot r_2 + r_1 \cdot r_3$;
- $\forall r_1, r_2, r_3 \in \mathcal{R}: \ (r_1 + r_2) \cdot r_3 = r_1 \cdot r_3 + r_2 \cdot r_3$.

$\mathcal{R}$ *is a* ring with unit*, if* $(\mathcal{R}, \cdot)$ *is a monoid whose neutral element is usually denoted by* 1.

A classical example of a ring with unit are the integers $\mathbb{Z}$. If $\mathcal{R}$ is any commutative ring with unit, the set $\mathcal{P} = \mathcal{R}[x^1, \dots, x^n]$ of all polynomials in the variables[1] $x^1, \dots, x^n$ with coefficients in $\mathcal{R}$ is again a ring with unit. Examples of non-commutative rings are rings of linear differential operators with variable coefficients. We deal almost exclusively with rings with unit; in fact in many books the definition of a ring requires the existence of a unit.

Definition B.1.4. *A* field $\Bbbk$ *is a ring with unit such that* $(\Bbbk \setminus \{0\}, \cdot)$ *is an Abelian group with neutral element* 1.

The rational numbers $\mathbb{Q}$ are the quotient field of the integers $\mathbb{Z}$. Similarly, the rational functions $\mathcal{R}(x^1, \dots, x^n)$ are the quotient field of the polynomial ring $\mathcal{P}$. The analytic functions do not form a field but only a ring, as the quotient of two analytic functions is no longer analytic at the zeros of the denominator. But the *meromorphic functions*, i.e. the functions that are everywhere analytic except on a discrete set of points where they have a pole, form a field. We consider exclusively fields of *characteristic zero*; any such field contains the rational numbers $\mathbb{Q}$ as a subfield.

Definition B.1.5. *Let* $\mathcal{R}$ *be a ring with unit. An Abelian group* $(\mathcal{M}, +)$ *is called a* (left) $\mathcal{R}$-module*, if a scalar multiplication* $\cdot : \mathcal{R} \times \mathcal{M} \to \mathcal{M}$ *exists with the following properties:*

- $\forall m \in \mathcal{M}: \ 1 \cdot m = m$;
- $\forall r_1, r_2 \in \mathcal{R}, \forall m \in \mathcal{M}: \ r_1 \cdot (r_2 \cdot m) = (r_1 r_2) \cdot m$;
- $\forall r \in \mathcal{R}, \forall m_1, m_2 \in \mathcal{M}: \ r \cdot (m_1 + m_2) = r \cdot m_1 + r \cdot m_2$;
- $\forall r_1, r_2 \in \mathcal{R}, \forall m \in \mathcal{M}: \ (r_1 + r_2) \cdot m = r_1 \cdot m + r_2 \cdot m$.

[1] We follow here (as everywhere in this text) the convention in differential geometry and use superscripts to index variables.

This is an obvious extension of the definition of a monoid module. Much of the literature deals only with modules over *commutative* rings. We have not made this restriction, as we want to be able to consider modules over, say, rings of differential operators. One can analogously define *right* modules with a scalar multiplication $\cdot : \mathcal{M} \times \mathcal{R} \to \mathcal{M}$. As we exclusively deal with left modules, we usually omit the "left" and simply speak of modules.

Example B.1.6. The simplest and most familiar examples of modules are *vector spaces*. Indeed, if the ring $\mathcal{R}$ is even a field, then an $\mathcal{R}$-module is nothing but a vector space over $\mathcal{R}$.

Vector spaces are a special case of *free modules*. An $\mathcal{R}$-module $\mathcal{M}$ is called free, if it is of the form $\mathcal{R}^m$ for some positive integer m. Thus the elements of $\mathcal{M}$ can be identified with m-dimensional vectors whose entries are elements of $\mathcal{R}$. Addition and scalar multiplication are defined componentwise and obviously satisfy the conditions above.

Another way to define a free $\mathcal{R}$-module is to take a set $\mathcal{G}$ of *generators* and to consider all formal linear combinations of elements of $\mathcal{G}$ with coefficients in $\mathcal{R}$. An element of the arising module $\mathcal{M}_\mathcal{G}$ has the form $m = \sum_{g \in \mathcal{G}} r_g g$ where only finitely many of the coefficients $r_g \in \mathcal{R}$ do not vanish. Again, addition and scalar multiplication are defined componentwise. ◁

Let $\mathcal{G} \subset \mathcal{M}$ be a subset of $\mathcal{M}$. Then $\mathcal{G}$ is a *generating set* of $\mathcal{M}$, if any element $m \in \mathcal{M}$ can be written as a linear combination $m = \sum_{g \in \mathcal{G}} r_g g$ where only finitely many of coefficients $r_g \in \mathcal{R}$ do not vanish. Note that this representation is in general not unique! A module is called *finitely generated*, if it possesses a finite generating set. A set $\mathcal{G}$ is called *linearly independent*, if $\sum_{g \in \mathcal{G}} r_g g = 0$ implies that all coefficients r_g vanish. A linearly independent generating set is a *basis* of the module $\mathcal{M}$; only in this case the above representation of the elements of $\mathcal{M}$ as linear combinations is unique. Note, however, that most modules do not possess a basis! In fact, a module has a basis, if and only if it is free.

Remark B.1.7. These notions are familiar from linear algebra. One should, however, note that many facts which are true for vector spaces do not hold, in general, for modules. Assume, for example, that we are given a set of elements that are linearly dependent. In a vector space this implies that some of the elements can be expressed as linear combinations of the other ones. In a module this is not necessarily true. Take the free $\mathbb{Z}$-module $\mathbb{Z}^2$ and the two elements $m_1 = \binom{2}{4}$ and $m_2 = \binom{3}{6}$. Obviously, they are linearly dependent, as $3m_1 - 2m_2 = 0$. But over $\mathbb{Z}$ none is a multiple of the other one. ◁

Let $\mathcal{M}$ and $\mathcal{N}$ be two $\mathcal{R}$-modules. A map $\phi : \mathcal{M} \to \mathcal{N}$ is a *module homomorphism*, if it respects the module structures of $\mathcal{M}$ and $\mathcal{N}$, i.e. if for all module elements $m_1, m_2 \in \mathcal{M}$ and for all ring elements $r \in \mathcal{R}$ the equality $\phi(m_1 + r \cdot m_2) = \phi(m_1) + r \cdot \phi(m_2)$ hold. As usual, a bijective homomorphism is called an *isomorphism* and two isomorphic modules can be identified to some extent.

Any finitely generated module is isomorphic to a factor module of a free module; this is called a *presentation by generators and relations*. In principle, the construction of such a presentation for a given $\mathcal{R}$-module $\mathcal{M}$ is straightforward. One takes a finite generating set $\mathcal{G} = \{g_1, \dots, g_m\}$ of $\mathcal{M}$ and the free $\mathcal{R}$-module $\mathcal{R}^m$. If $\mathcal{G}$ is a basis of $\mathcal{M}$, then $\mathcal{M} \cong \mathcal{R}^m$. Otherwise, we define the map

$$\phi : \begin{cases} \mathcal{R}^m & \longrightarrow \quad \mathcal{M} \\ (r_1, \dots, r_m) & \longmapsto \quad \sum_{i=1}^m r_i g_i \end{cases} \tag{B.1}$$

and set $\mathcal{N} = \ker \phi$. Then $\mathcal{M} \cong \mathcal{R}^m/\mathcal{N}$. In practice, it may be difficult to effectively construct a generating set $\mathcal{G}$ and its relations module $\mathcal{N}$ for an arbitrary module $\mathcal{M}$. Fortunately, most of the modules we are dealing with automatically arise in a presentation by generators and relations, so that we do not bother about this problem.

A finite sequence $(x_1, \dots, x_k)$ of elements of the underlying ring $\mathcal{R}$ is called $\mathcal{M}$-*regular* for the $\mathcal{R}$-module $\mathcal{M}$, if x_1 is not a zero divisor of $\mathcal{M}$ and x_i is not a zero divisor for $\mathcal{M}/\langle x_1, \dots, x_{i-1}\rangle\mathcal{M}$ for $1 < i \leq k$. The maximal length of an $\mathcal{M}$-regular sequence is the *depth* d of the module $\mathcal{M}$. It represents a measure for the size of $\mathcal{M}$.

Definition B.1.8. *Let $\mathcal{R}$ be a ring with unit. An* $\mathcal{R}$-algebra $\mathcal{A}$ *is a (left) $\mathcal{R}$-module together with an $\mathcal{R}$-bilinear multiplication* $\cdot : \mathcal{A} \times \mathcal{A} \to \mathcal{A}$.

The ring $\mathcal{P} = \mathcal{R}[x^1, \dots, x^n]$ of polynomials with coefficients in an Abelian ring $\mathcal{R}$ is a simple example of an Abelian $\mathcal{R}$-algebra. More generally, let $\mathcal{M}$ be a finitely generated free $\mathcal{R}$-module. We introduce the *symmetric algebra* $S\mathcal{M}$ as follows.[2] Let $\{g_1, \dots, g_n\}$ be a basis of $\mathcal{M}$, then $S\mathcal{M}$ is the infinite-dimensional free $\mathcal{R}$-module generated by the basis $\{g^\mu = g_1^{\mu_1} \cdots g_n^{\mu_n} \mid \mu \in \mathbb{N}_0^n\}$. Thus $S\mathcal{M}$ is isomorphic to the polynomial ring $\mathcal{P}$ via the identification $x^i \leftrightarrow g_i$. However, it is important to note that the definition of $S\mathcal{M}$ is independent of the chosen basis of $\mathcal{M}$, in contrast to this isomorphism which is obviously not canonical. Linear differential operators with variable coefficients form a non-commutative algebra over their coefficient ring.

The multiplication in an algebra is not necessarily associative. A very important example of a non-associative algebra is a *Lie algebra*. Here the

[2] A more satisfying definition of the symmetric algebra is as a quotient space of the full tensor algebra over $\mathcal{M}$. However, for our purposes this "pedestrian" approach via a basis suffices.

multiplication is usually written in form of a bracket, $a_1 \cdot a_2 = [a_1, a_2]$, and it must satisfy in addition to the distributive laws the following two axioms: (i) skew-symmetry, $[a_1, a_2] = -[a_2, a_1]$, and (ii) the *Jacobi identity*, $[a_1, [a_2, a_3]] + [a_2, [a_3, a_1]] + [a_3, [a_1, a_2]] = 0$. Note that an associative algebra is always a ring.

Subgroups, subrings, submodules etc. are defined in the obvious way: a subset that is closed under the given operations, i. e. the addition respectively the multiplication of two elements of the subset yields again an element of the subset. Thus in the case of a subring $\mathcal{I}$ one requires for example that $\mathcal{I}+\mathcal{I} \subseteq \mathcal{I}$ and $\mathcal{I} \cdot \mathcal{I} \subseteq \mathcal{I}$. There exists a special class of subrings that are particularly important, namely those which are even a monoid ideal with respect to the ring multiplication.

Definition B.1.9. *Let $\mathcal{R}$ be an arbitrary ring. A subring $\mathcal{I} \subseteq \mathcal{R}$ is called a* (left) ideal*, if $\mathcal{R} \cdot \mathcal{I} \subseteq \mathcal{I}$.*

In a *(left) Noetherian ring* $\mathcal{R}$ any (left) ideal $\mathcal{I} \subseteq \mathcal{R}$ is finitely generated or, equivalently, any ascending chain of (left) ideals $\mathcal{I}_1 \subseteq \mathcal{I}_2 \subseteq \mathcal{I}_3 \subseteq \cdots \subseteq \mathcal{R}$ becomes eventually stationary, i. e. there exists an $n \in \mathbb{N}$ such that $\mathcal{I}_{n+k} = \mathcal{I}_n$ for all k. By the *Hilbert Basis Theorem* the polynomial ring $\mathcal{R}[x^1, \cdots, x^n]$ is Noetherian, if the coefficient ring $\mathcal{R}$ is Noetherian.[3]

Definition B.1.10. *The ring $\mathcal{R}$ is* (non-negatively) graded*, if it has as additive Abelian group $(\mathcal{R}, +)$ a direct sum decomposition $\mathcal{R} = \bigoplus_{i \in \mathbb{N}_0} \mathcal{R}_i$ such that $\mathcal{R}_i \mathcal{R}_j \subseteq \mathcal{R}_{i+j}$. A* graded $\mathcal{R}$-module $\mathcal{M}$ *is an $\mathcal{R}$-module with a direct sum decomposition $\mathcal{M} = \bigoplus_{i \in \mathbb{N}_0} \mathcal{M}_i$ such that $\mathcal{R}_i \cdot \mathcal{M}_j \subseteq \mathcal{M}_{i+j}$. The elements of $\mathcal{R}_i$ or $\mathcal{M}_i$, respectively, are* homogeneous *of degree i. For a given degree q the* truncation $\mathcal{M}_{\geq q}$ *is defined as the module $\bigoplus_{p \geq q} \mathcal{M}_p$, i. e. we throw away all elements of degree less than q.*

If $\mathcal{R}$ is a graded ring, $\mathcal{R}_0$ is a subring and each component $\mathcal{R}_i$ is an $\mathcal{R}_0$-module. The graded ring we are mostly interested in is the polynomial ring $\mathcal{P} = \Bbbk[x^1, \dots, x^n]$ where a natural grading is given by the total degree. Here $\mathcal{P}_0 = \Bbbk$ is even a field and the components $\mathcal{P}_i$ consisting of all homogeneous polynomials of degree i are vector spaces over $\Bbbk$. A submodule $\mathcal{N} \subseteq \mathcal{M}$ is graded, if it can be decomposed in the form $\mathcal{N} = \bigoplus_{i \in \mathbb{N}_0} (\mathcal{N} \cap \mathcal{M}_i)$. A graded submodule can always be generated by homogeneous elements.

Definition B.1.11. *Let $\mathcal{R}$ be a graded ring with $\mathcal{R}_0$ a field and $\mathcal{M}$ a finitely generated graded $\mathcal{R}$-module. Then the homogeneous components $\mathcal{M}_r$ are*

[3] The proof of Hilbert's Basis Theorem requires the *axiom of choice* (cf. [15, Section 4.1]). If you cannot accept this axiom, you should stop reading now, as most of our algebraic results depend on it!

finite-dimensional vector spaces over $\mathcal{R}_0$. *The* Hilbert function *of* $\mathcal{M}$ *is the numerical function*[4] *defined by*

$$h_{\mathcal{M}}(r) = \dim \mathcal{M}_r \,. \tag{B.2}$$

The Hilbert function represents a measure for the size of the graded module $\mathcal{M}$. As a vector space over $\mathcal{R}_0$ the whole module is generally infinite-dimensional, so that this dimension is not a useful measure. Instead we must use a more complicated object like the Hilbert function. Robbiano [152] gives a nice introduction into the theory of Hilbert functions (and numerical functions in general).

Let the ring $\mathcal{R}$ be generated over $\mathcal{R}_0$ by n elements of degree 1; this is for example the case, if $\mathcal{R} = \mathcal{P}$. It is a classical result in commutative algebra that at some degree r_0 the Hilbert function becomes polynomial, i. e. there exists a polynomial $H_{\mathcal{M}}(r)$, the *Hilbert polynomial* of $\mathcal{M}$, such that for all integers $r \geq r_0$ the identity $H_{\mathcal{M}}(r) = h_{\mathcal{M}}(r)$ holds. One can show that $\deg h_{\mathcal{M}} = D \leq n$. The number D is the *(Krull) dimension* of the module $\mathcal{M}$. If the coefficient of r^D in $H_{\mathcal{M}}(r)$ is a_D, then the *multiplicity* of $\mathcal{M}$ is $a_D D!$.

Above we introduced the depth d as an alternative measure for the size of the module $\mathcal{M}$. One can show that always $D \geq d$. Modules such that $D = d$ are called *Cohen–Macaulay*.

Sometimes one prefers instead of the Hilbert function the *Hilbert series*.[5] It is the generating function of $h_{\mathcal{M}}(r)$, i. e. the univariate series in some variable λ defined by

$$\mathcal{H}_{\mathcal{M}}(\lambda) = \sum_{r=0}^{\infty} h_{\mathcal{M}}(r)\lambda^r \,. \tag{B.3}$$

One can show (see Proposition 2.5.2) that this series can always be summed in closed form yielding a rational function $\mathcal{H}_{\mathcal{M}}(\lambda) = f(\lambda)/(1-\lambda)^n$ where $f(\lambda)$ is a polynomial with integer coefficients. For the free module $\mathcal{M} = \mathcal{R}^m$ we get $\mathcal{H}_{\mathcal{M}}(\lambda) = m/(1-\lambda)^n$. The dimension D of $\mathcal{M}$ arises now as the order of the pole of $\mathcal{H}_{\mathcal{M}}$ at $\lambda = 1$.

These notions were originally introduced by Hilbert [77] for the case of polynomial ideals.[6] Here one must distinguish between the *affine* and the *projective* Hilbert function. Let $\mathcal{I}$ be a homogeneous ideal in the polynomial ring $\mathcal{P}$, i. e. an ideal generated by homogeneous elements. Then the quotient $\mathcal{P}/\mathcal{I}$ is a graded $\mathcal{P}$-module and we define the projective Hilbert function $h_{\mathcal{I}}$

[4] A *numerical function* $f(r)$ is a function returning for every integer argument $r \in \mathbb{Z}$ an integer value.

[5] The names *Poincaré* or *Hilbert-Poincaré series* are also used frequently.

[6] Hilbert used the term *characteristic function*.

as the Hilbert function of the graded module $\mathcal{P}/\mathcal{I}$. A particular case of a homogeneous ideal is a *monomial ideal*, i.e. an ideal with a basis consisting only of monomials. Then $h_{\mathcal{I}}(r)$ counts the terms of degree r that are *not* contained in $\mathcal{I}$. As we will see in Remark B.2.9, the general case can always be reduced to the monomial one with the help of Gröbner bases.

For non-homogeneous ideals $\mathcal{I}$, the quotient $\mathcal{P}/\mathcal{I}$ has no obvious grading. There exist two different ways to define nevertheless a Hilbert function which both lead to the same result. We may homogenise the ideal $\mathcal{I}$: we introduce an additional new variable x^0 and consider the ideal

$$\mathcal{I}^h = \langle f^h \mid f \in \mathcal{I} \rangle \subseteq \mathbb{k}[x^0, x^1, \dots, x^n] \tag{B.4}$$

where the homogeneous polynomial f^h is obtained from f by multiplying each term $x^\mu \in \operatorname{supp} f$ by $(x^0)^{\deg f - |\mu|}$. Then we define the affine Hilbert function of $\mathcal{I}$ as the projective Hilbert function of $\mathcal{I}^h$: $h^a_{\mathcal{I}} = h_{\mathcal{I}^h}$. Alternatively, we may use the following approach. Denote by $\mathcal{P}_{\leq r}$ the vector space of all polynomials f with $\deg f \leq r$ and set $\mathcal{I}_{\leq r} = \mathcal{I} \cap \mathcal{P}_{\leq r}$. Now we define the affine Hilbert function by $h^a_{\mathcal{I}}(r) = \dim \mathcal{P}_{\leq r} - \dim \mathcal{I}_{\leq r}$. For a homogeneous ideal we have the obvious relation $h_{\mathcal{I}}(r) = h^a_{\mathcal{I}}(r) - h^a_{\mathcal{I}}(r-1)$.

The degree of the affine Hilbert polynomial $H^a_{\mathcal{I}}(r)$ is the *(Krull) dimension* D of the ideal $\mathcal{I}$. For an algebraically closed coefficient field $\mathbb{k}$, the affine variety defined by the ideal $\mathcal{I}$, i.e. the set of points in $\mathbb{k}^n$ that are common zeros of all polynomials in $\mathcal{I}$, has also the dimension D.

The Hilbert polynomial $H_{\mathcal{I}}(r)$ is an example of a *numerical polynomial*, i.e. $H_{\mathcal{I}} : \mathbb{Z} \to \mathbb{Z}$. These polynomials form a $\mathbb{Z}$-module. One can show with a simple argument from interpolation theory that any numerical polynomial of degree d can be written in the form

$$H_{\mathcal{I}}(r) = \sum_{i=0}^{d} a_i r^i = \sum_{i=0}^{d} b_i \binom{r}{i} . \tag{B.5}$$

The transformation between the coefficients a_i and b_i is a combinatorial operation leading to the modified Stirling numbers discussed in Appendix A.4. The appearance of binomial coefficients in (B.5) is not surprising, as according to (A.4a) the affine Hilbert polynomial of the monomial ideal in $\mathcal{P}$ generated by the single term x^μ is $H^a(r) = \binom{n+r-|\mu|}{n}$.

B.2 Gröbner Bases for Polynomial Ideals and Modules

Let $\Bbbk$ be a field of characteristic zero and $\mathcal{P} = \Bbbk[x^1, \dots, x^n]$ a polynomial ring. We denote by $\mathbb{T} = \{x^\mu \mid \mu \in \mathbb{N}_0^n\}$ the set of all *terms*.[7] $(\mathbb{T}, \cdot)$ where $\cdot$ is the usual multiplication defines an Abelian monoid which is obviously isomorphic to the monoid $(\mathbb{N}_0^n, +)$ of multi indices introduced in Appendix A.1 via the identification $\mu \leftrightarrow x^\mu$.

A *term order* $\prec$ is a total order on $\mathbb{T}$ compatible with the monoid structure. This means that it satisfies the following two conditions: (i) $1 \preceq t$ for all terms $t \in \mathbb{T}$ and (ii) $s \preceq t$ implies $r \cdot s \preceq r \cdot t$ for all terms $r, s, t \in \mathbb{T}$. These two conditions ensure that $\prec$ is a *well-order*, i.e. that a strictly decreasing sequence $t_1 \succ t_2 \succ \cdots$ cannot be infinite. If a term order fulfils in addition the condition that $s \prec t$ whenever $\deg s < \deg t$, it is called *degree compatible*. As $\mathbb{T}$ and $\mathbb{N}_0^n$ are isomorphic as monoids, we can also speak of term orders on $\mathbb{N}_0^n$. In fact, most term orders are defined via multi indices.

Example B.2.1. Some important term orders are the following ones. The *lexicographic* order is defined by $x^\mu \prec_{\text{lex}} x^\nu$, if the last non-vanishing entry of $\mu - \nu$ is negative. Thus $(x^2)^2 x^3 \prec_{\text{lex}} x^1 (x^3)^2$. With respect to the *reverse lexicographic* order $x^\mu \prec_{\text{revlex}} x^\nu$, if the first non-vanishing entry of $\mu - \nu$ is positive. Now we have $x^1(x^3)^2 \prec_{\text{revlex}} (x^2)^2 x^3$. However, $\prec_{\text{revlex}}$ is *not* a term order, as it violates the first condition: $x^1 \prec_{\text{revlex}} 1$. The reverse lexicographic order should not be confused with the *inverse lexicographic* order $\prec_{\text{invlex}}$ which arises from $\prec_{\text{lex}}$ by inverting the order of the variables, i.e. $x^\mu \prec_{\text{invlex}} x^\nu$, if the first non-vanishing entry of $\mu - \nu$ is negative.

Degree compatible versions exist of all these orders. $x^\mu \prec_{\text{deglex}} x^\nu$, if $|\mu| < |\nu|$ or if $|\mu| = |\nu|$ and $x^\mu \prec_{\text{lex}} x^\nu$. Similarly, $x^\mu \prec_{\text{degrevlex}} x^\nu$, if $|\mu| < |\nu|$ or if $|\mu| = |\nu|$ and $x^\mu \prec_{\text{revlex}} x^\nu$. Note that $\prec_{\text{degrevlex}}$ *is* a term order, in fact a very important one!

Here we have defined the orders *inverse* to the usual convention in most texts on Gröbner bases. The classical forms arises, if one inverts the order of the variables: $x^1, \dots, x^n \to x^n, \dots, x^1$. Our version fits better to the conventions used in differential equations theory, in particular to our definition of the class of a multi index in Appendix A.1 (see below). $\triangleleft$

Any polynomial $f \in \mathcal{P}$ can be written in the form $f = \sum_{t \in \mathbb{T}} c_t t$ where only finitely many of the coefficients $c_t \in \Bbbk$ are different from 0. The set

$$\operatorname{supp} f = \{t \in \mathbb{T} \mid c_t \neq 0\} \tag{B.6}$$

[7] There is a notorious dispute in the commutative algebra community whether the elements of $\mathbb{T}$ should be called terms or monomials. We call them terms; a monomial is the product of a term and a coefficient.

is the *support* of f. For a given term order $\prec$, $\operatorname{supp} f$ contains a largest term, the *leading term* $\operatorname{lt}_\prec f$ of the polynomial f. Its coefficient is the *leading coefficient* $\operatorname{lc}_\prec f$ and if $\operatorname{lt}_\prec f = x^\mu$, we call μ the *leading exponent* $\operatorname{le}_\prec f$. Finally, the *leading monomial* is defined by $\operatorname{lm}_\prec f = \operatorname{lc}_\prec f \cdot \operatorname{lt}_\prec f$. If $\mathcal{F}$ is a finite set of polynomials, we write $\operatorname{lt}_\prec \mathcal{F}$ for the set $\{\operatorname{lt}_\prec f \mid f \in \mathcal{F}\}$.

Let $\mathcal{F} = \{f_1, \dots, f_r\}$ be a finite set of polynomials; it generates the following ideal in the ring $\mathcal{P}$

$$\langle \mathcal{F} \rangle = \Big\{ h \in \mathcal{P} \mid h = \sum_{i=1}^{r} g_i f_i \;\; g_i \in \mathcal{P} \Big\} \,. \tag{B.7}$$

If $\mathcal{I} = \langle \mathcal{F} \rangle$ for some ideal $\mathcal{I} \subseteq \mathcal{P}$, then $\mathcal{F}$ is called a *basis*[8] of $\mathcal{I}$. By *Hilbert's Basis Theorem* any ideal in $\mathcal{P}$ has a finite basis (which is of course not unique), as any field is trivially a Noetherian ring.

Lemma B.2.2. *Let for a term order $\prec$ the condition $\operatorname{lt}_\prec f \in \Bbbk[x^1, \dots, x^k]$ be equivalent to $f \in \Bbbk[x^1, \dots, x^k]$ for any polynomial $f \in \mathcal{P}$. Then $\prec$ is the lexicographic order $\prec_{\text{lex}}$. If the condition $\operatorname{lt}_\prec f \in \langle x^1, \dots, x^k \rangle$ is equivalent to $f \in \langle x^1, \dots, x^k \rangle$ for any homogeneous polynomial $f \in \mathcal{P}$, then $\prec$ is the degree reverse lexicographic order $\prec_{\text{degrevlex}}$.*

Proof. We leave the case of the lexicographic order as an exercise; it goes completely analogously to the for us more important case of the degree reverse lexicographic order. Let x^μ and x^ν be two arbitrary monomials of the same degree such that the first non-vanishing entry of $\mu - \nu$ is $\mu_k - \nu_k$. Without loss of generality, we assume that $\mu_k > \nu_k$. Set $\rho = [\nu_1, \dots, \nu_k, 0, \dots, 0]$ and consider the multi indices $\bar{\mu} = \mu - \rho$ and $\bar{\nu} = \nu - \rho$. Obviously, $x^{\bar{\mu}} \in \langle x^1, \dots, x^k \rangle$ whereas $x^{\bar{\nu}} \notin \langle x^1, \dots, x^k \rangle$. Considering the homogeneous polynomial $f = x^{\bar{\mu}} + x^{\bar{\nu}}$, the assumption of the lemma implies that the leading term of f must be $x^{\bar{\nu}}$. By the monotonicity of term orders, we conclude that $x^\mu \prec x^\nu$. But by definition, we also have $x^\mu \prec_{\text{degrevlex}} x^\nu$. □

We say that a term order *respects classes*, if for multi indices μ, ν of the same length $\operatorname{cls} \mu < \operatorname{cls} \nu$ implies $x^\mu \prec x^\nu$. It is now easy to see that by Lemma B.2.2 only one class respecting term order exists: the degree reverse lexicographic order.

If $\mathcal{I}$ is any ideal in $\mathcal{P}$, its *leading ideal* $\operatorname{lt}_\prec \mathcal{I}$ for a given term order $\prec$ is the monomial ideal generated by the leading terms of all polynomials in $\mathcal{I}$

$$\operatorname{lt}_\prec \mathcal{I} = \langle \{\operatorname{lt}_\prec h \mid h \in \mathcal{I}\} \rangle \,. \tag{B.8}$$

[8] Obviously, every polynomial ideal is a module over the polynomial ring. One should note that strictly speaking these "ideal bases" are only generating sets and not bases, as a proper ideal never defines a free module. But this abuse of language has become standard in commutative algebra.

Definition B.2.3. *Let $\mathcal{I} \subseteq \mathcal{P}$ be a non-zero ideal. A finite set $\mathcal{G} \subset \mathcal{P}$ is a* Gröbner basis[9] *of $\mathcal{I}$ for the term order $\prec$, if $\mathrm{lt}_\prec \mathcal{I} = \langle \mathrm{lt}_\prec \mathcal{G} \rangle$.*

Example B.2.4. Consider the set $\mathcal{F} = \{f_1 = \underline{xy} - y,\ f_2 = \underline{y^2} - x\} \subset \Bbbk[x,y]$ and let $\prec$ be a degree compatible order. The underlined terms are the leading terms of the elements of $\mathcal{F}$. Set $\mathcal{I} = \langle \mathcal{F} \rangle$; then $\mathcal{F}$ is *not* a Gröbner basis of $\mathcal{I}$ with respect to $\prec$. The polynomial $f_3 = \underline{x^2} - x = y \cdot f_1 + (1-x) \cdot f_2$ is obviously an element of $\mathcal{I}$, but its leading term x^2 is not contained in the monomial ideal generated by $\mathrm{lt}_\prec \mathcal{F} = \{xy, y^2\}$. We have not yet a way to prove it, but the augmented set $\mathcal{G} = \{f_1, f_2, f_3\}$ is a Gröbner basis of $\mathcal{I}$. $\lhd$

Note that Definition B.2.3 does not require $\mathcal{G} \subset \mathcal{I}$; but we will later see that this is indeed always the case. A set $\mathcal{G}$ may well be a Gröbner basis of an ideal for one term order but not for another one; the chosen term order has a great influence on the basis. We sometimes simply say that a set $\mathcal{G}$ is a Gröbner basis. This means that it is a Gröbner basis of the ideal $\langle \mathcal{G} \rangle$ for some term order $\prec$ which is obvious from the context.

While Definition B.2.3 is probably the simplest and most direct way to define a Gröbner basis, it does not make obvious why Gröbner bases are so important or how one can check whether a given set $\mathcal{G}$ is a Gröbner basis. Furthermore, it is not clear whether Gröbner bases exist at all. In order to answer such questions we need a few additional concepts.

Definition B.2.5. *For a given term order $\prec$, the polynomial $g \in \mathcal{P}$ is* reducible *with respect to the finite subset $\mathcal{F} \subset \mathcal{P}$, if there exists a term $t \in \operatorname{supp} g$ and a polynomial $f \in \mathcal{F}$ such that $\mathrm{lt}_\prec f \mid t$. A polynomial $g' \in \mathcal{P}$ that is not reducible with respect to $\mathcal{F}$ and such that $g' - g \in \langle \mathcal{F} \rangle$ is called a* normal form *of g with respect to F: $g' = \mathrm{NF}_{\mathcal{F},\prec}(g)$.*

Algorithm B.1 allows us to determine a normal form of a polynomial g with respect to a set $\mathcal{F}$. If g is not in normal form with respect to $\mathcal{F}$, we take the largest term $t \in \operatorname{supp} g$ divisible by the leading term of an element $f \in \mathcal{F}$ and compute the difference $g' = g - (c_t t / \mathrm{lm}_\prec f) f$; obviously, this eliminates the monomial $c_t t$ in g. If g' is still reducible, we iterate. As after each step the largest divisible term in $\operatorname{supp} g'$ becomes smaller and $\prec$ is a well-order, the algorithm terminates after a finite number of steps.

The normal form of a polynomial g with respect to an arbitrary set $\mathcal{F}$ is not unique. This can already be seen from Algorithm B.1. In Line /5/ there might exist several elements $f \in \mathcal{F}$ such that $\mathrm{lt}_\prec f \mid t$ and any of them can be used for eliminating the monomial $c_t t$. For different choices we must expect to obtain different normal forms.

[9] The terminology was introduced by Buchberger in honour of his thesis supervisor Gröbner. Some authors prefer to speak of *standard bases*.

Algorithm B.1 A simple normal form algorithm

Input: finite set $\mathcal{F} \subset \mathcal{P}$, polynomial $g \in \mathcal{P}$, term order $\prec$
Output: a normal form $g' = \mathrm{NF}_{\mathcal{F},\prec}(g)$

```
/1/ S ← {t ∈ supp g | ∃f ∈ F : lt≺f | t}
/2/ g′ ← g
/3/ while S ≠ ∅ do
/4/     t = max≺ S
/5/     choose f ∈ F such that lt≺f | t
/6/     g′ ← g′ − (c_t t/lm≺f) f
/7/     S ← {t ∈ supp g′ | ∃f ∈ F : lt≺f | t}
/8/ end_while
/9/ return g′
```

Example B.2.6. We continue with the set $\mathcal{F} = \{f_1, f_2\}$ of Example B.2.4 and a degree compatible term order. If we reduce the polynomial $\underline{xy^2} - x$ first with respect to f_1, we obtain the normal form $\underline{x^2} - x$. If we start with a reduction with respect to f_2, the normal form 0 arises after one further reduction step. $\triangleleft$

Closely related to the normal form algorithm is the so-called *division algorithm* which is used in many textbooks for the introduction of Gröbner bases. By a slight modification of Algorithm B.1, namely recording the individual reduction steps, one can determine a representation $g = \sum_{f \in \mathcal{F}} h_f f + g'$ with coefficients $h_f \in \mathcal{P}$. In Line /2/ we initialise $h_f \leftarrow 0$ for every $f \in \mathcal{F}$; in Line /6/ we compute in addition $h_f \leftarrow h_f + c_t t/\mathrm{lm}_\prec f$ for the f chosen in the previous line. This may be considered as a multivariate generalisation of polynomial division and g' is sometimes called the *remainder* of the reduction. By construction, the "quotients" h_f satisfy $\mathrm{lt}_\prec(h_f f) \preceq \mathrm{lt}_\prec g$.

Proposition B.2.7. *Let $\mathcal{I} \subseteq \mathcal{P}$ be a non-zero ideal and $\mathcal{G} \subset \mathcal{P}$ a finite set. Then the following statements are equivalent.*

(i) *The set $\mathcal{G}$ is a Gröbner basis of $\mathcal{I}$ for the term order $\prec$.*
(ii) *Every polynomial $f \in \mathcal{P}$ possesses a unique normal form with respect to the set $\mathcal{G}$.*
(iii) *Every polynomial $f \in \mathcal{I}$ can be written in the form*

$$f = \sum_{g \in \mathcal{G}} h_g g \tag{B.9}$$

where the coefficients $h_g \in \mathcal{P}$ satisfy $\mathrm{lt}_\prec(h_g g) \preceq \mathrm{lt}_\prec f$ for all $g \in \mathcal{G}$.

Proof. Assume that $\mathcal{G}$ is a Gröbner basis of $\mathcal{I}$ with respect to the given term order $\prec$. Let f_1 and f_2 be two normal forms of an arbitrary polynomial

$f \in \mathcal{P}$; then obviously $f_1 - f_2 \in \mathcal{J}$. The leading term of $f_1 - f_2$ is an element of $\operatorname{supp} f_1 \cup \operatorname{supp} f_2$. By definition of a normal form this set cannot contain any term that is divisible by the leading term of an element of $\mathcal{G}$. On the other hand, by definition of a Gröbner basis, $\operatorname{lt}_\prec(f_1 - f_2)$ must be divisible by the leading term of an element of $\mathcal{G}$. This contradiction can only be resolved, if $f_1 - f_2 = 0$ and thus (i) implies (ii).

If every polynomial in $\mathcal{P}$ possesses a unique normal form, then the normal form of all polynomials $f \in \mathcal{J}$ is zero, as zero is trivially a normal form of any element of the ideal. If we apply Algorithm B.1 in its extended form, i.e. the division algorithm, to such a polynomial f, we obtain a representation of the form (B.9). As already remarked above, the coefficients h_g satisfy by construction the required condition. Thus (iii) follows from (ii).

Finally, we assume that (iii) holds. The condition on the coefficients obviously implies that for any $f \in \mathcal{J}$ the leading term $\operatorname{lt}_\prec f$ is divisible by $\operatorname{lt}_\prec g$ for some $g \in \mathcal{G}$. Thus $\operatorname{lt}_\prec \mathcal{J} = \langle \operatorname{lt}_\prec \mathcal{G} \rangle$ and consequently (i) is true. □

Example B.2.8. The right hand side of (B.9) is called a *standard representation* of f. It is important to note that it is *not* unique. While the proposition above asserts that all possible normal form algorithms return the same result, they may achieve it in different ways and thus the associated division algorithms yield different standard representations. Consider again the Gröbner basis $\mathcal{G} = \{f_1, f_2, f_3\}$ of Example B.2.4. The polynomial $f = xy^2 - x \in \langle \mathcal{G} \rangle$ has two different standard representations: $f = yf_1 + f_2 = xf_2 + f_3$. ◁

Remark B.2.9. Proposition B.2.7 implies that any Gröbner basis $\mathcal{G}$ of an ideal $\mathcal{J}$ provides us with a method to determine unique representatives for the cosets in $\mathcal{P}/\mathcal{J}$, as all polynomials in a coset $[f]$ have the same normal form with respect to $\mathcal{G}$. The normal form even defines an isomorphism between $\mathcal{P}/\mathcal{J}$ considered as a vector space over $\mathbb{k}$ and the vector space generated by all terms *not* contained in $\operatorname{lt}_\prec \mathcal{J} = \langle \operatorname{lt}_\prec \mathcal{G} \rangle$, as any normal form contains only such terms. This fact was already observed by Macaulay [110] who concluded that every ideal $\mathcal{J}$ has the same Hilbert function as the *monomial* ideal $\operatorname{lt}_\prec \mathcal{J}$.[10]

An effective mean to perform arithmetics with cosets was the original motivation for Buchberger to introduce Gröbner bases in his thesis [28]. It also allows us to decide algorithmically the *ideal membership problem*: given a Gröbner basis $\mathcal{G}$, we have $f \in \mathcal{J}$, if and only if $\operatorname{NF}_{\mathcal{G},\prec}(f) = 0$. ◁

Note that by the definition of a basis, it is obvious that all elements of $\mathcal{J}$ can be written in the form (B.9) (compare with (B.7)). The particular property of a Gröbner basis is the additional condition on the coefficients.

[10] In the case of the *affine* Hilbert function, i.e. for non-homogeneous ideals, this holds only for degree compatible term orders.

Definition B.2.10. *Let $\mathcal{G}$ be a Gröbner basis with respect to the term order $\prec$. The basis $\mathcal{G}$ is* minimal, *if it does not contain two distinct elements $g, g' \in \mathcal{G}$ such that $\operatorname{lt}_\prec g \mid \operatorname{lt}_\prec g'$. The basis $\mathcal{G}$ is called* reduced, *if every element $g \in \mathcal{G}$ is in normal form with respect to the set $\mathcal{G}\setminus\{g\}$ and if $\operatorname{lc}_\prec g = 1$ for all $g \in \mathcal{G}$.*

Obviously, any reduced Gröbner basis is minimal. The terminology "minimal basis" has the following origin. For an ideal $\mathcal{I}$ one calls a basis $\mathcal{F}$ minimal, if one cannot remove an element of $\mathcal{F}$ without losing the property that $\mathcal{F}$ is a basis of $\mathcal{I}$. A general ideal has many minimal bases (even with different numbers of elements); but every *monomial* ideal possesses a unique minimal *monomial* basis. If $\mathcal{G}$ is a minimal Gröbner basis of $\mathcal{I}$ with respect to the term order $\prec$, then the set $\operatorname{lt}_\prec \mathcal{G}$ is the unique minimal monomial basis of $\operatorname{lt}_\prec \mathcal{I}$.

Theorem B.2.11. *Let $\mathcal{I} \subseteq \mathcal{P}$ be an arbitrary non-zero ideal. Then $\mathcal{I}$ possesses a unique reduced Gröbner basis for any term order $\prec$.*

All the characterisations of Gröbner bases given so far are not very constructive, as they do not tell us how to find a Gröbner basis for a given ideal. This will change now with the help of the *S-polynomial*[11] $S_\prec(f, g)$ of two polynomials f, g. It is defined as

$$S_\prec(f,g) = \frac{\operatorname{lcm}(\operatorname{lm}_\prec f, \operatorname{lm}_\prec g)}{\operatorname{lm}_\prec f} \cdot f - \frac{\operatorname{lcm}(\operatorname{lm}_\prec f, \operatorname{lm}_\prec g)}{\operatorname{lm}_\prec g} \cdot g \tag{B.10}$$

where lcm denotes as usual the least common multiple. Note that the factors are chosen such that the leading monomials cancel in the subtraction.

Proposition B.2.12. *A finite set $\mathcal{G}$ is a Gröbner basis of the ideal $\langle \mathcal{G} \rangle$, if and only if for every pair $f, g \in \mathcal{G}$ the S-polynomial $S_\prec(f, g)$ reduces to zero with respect to $\mathcal{G}$, i. e. $\operatorname{NF}_{\mathcal{G},\prec}(S_\prec(f, g)) = 0$.*

Based on this result, we describe a concrete algorithm for the construction of Gröbner bases: the *Buchberger algorithm* (see Algorithm B.2). Given a finite set $\mathcal{G}$, we form the set $\mathcal{S}$ of all pairs $\{g_1, g_2\}$ of different elements of $\mathcal{G}$. Then we take one pair and compute the corresponding S-polynomial. If its normal form $\bar{g}$ with respect to $\mathcal{G}$ vanishes, we choose the next pair. Otherwise we add $\bar{g}$ to $\mathcal{G}$ and start anew. The algorithm stops with a Gröbner basis, when all S-polynomials reduce to zero.[12]

---

[11] The S stands for *syzygy*; see Appendix B.3.

[12] Normal form computations may also be considered in the more abstract framework of *term rewriting*. Buchberger's algorithm is then an instance of a famous critical pair completion procedure, namely the *Knuth–Bendix algorithm* [90]. More on this relation can be found in [193].

Algorithm B.2 Buchberger algorithm for determining Gröbner bases

Input: finite set $\mathcal{F} \subset \mathcal{P}$, term order $\prec$
Output: Gröbner basis $\mathcal{G}$ of $\langle \mathcal{F} \rangle$

/1/ $\mathcal{G} \leftarrow \mathcal{F}$
/2/ $\mathcal{S} \leftarrow \{\{g_1, g_2\} \mid g_1, g_2 \in \mathcal{G}, g_1 \neq g_2\}$
/3/ **while** $\mathcal{S} \neq \emptyset$ **do**
/4/   choose $\{g_1, g_2\} \in \mathcal{S}$
/5/ $\mathcal{S} \leftarrow \mathcal{S} \setminus \{\{g_1, g_2\}\}$
/6/ $\bar{g} = \mathrm{NF}_{\mathcal{G},\prec}(S_\prec(g_1, g_2))$
/7/   **if** $\bar{g} \neq 0$ **then**
/8/ $\mathcal{S} \leftarrow \mathcal{S} \cup \{\{\bar{g}, g\} \mid g \in \mathcal{G}\}$
/9/ $\mathcal{G} \leftarrow \mathcal{G} \cup \{\bar{g}\}$
/10/   **end_if**
/11/ **end_while**
/12/ **return** $\mathcal{G}$

Theorem B.2.13. *Algorithm B.2 terminates for any finite set $\mathcal{F} \subset \mathcal{P}$ and any term order $\prec$ with a Gröbner basis of the ideal $\langle \mathcal{F} \rangle$.*

Proof. The main point is to show the termination of the algorithm. If it terminates, the outcome is clearly a Gröbner basis of $\langle \mathcal{F} \rangle$: we only add elements of this ideal so that the span does not change and the termination criterion is that all S-polynomials reduce to zero which by Proposition B.2.12 characterises Gröbner bases.

The termination follows from a simple Noetherian argument. The algorithm determines a sequence $\mathcal{G}_1 \subsetneq \mathcal{G}_2 \subsetneq \mathcal{G}_3 \subsetneq \cdots$ of finite sets. Consider the new element $\bar{g}$ added in Line /9/ to $\mathcal{G}_i$ in order to obtain $\mathcal{G}_{i+1}$. It results from a normal form computation with respect to $\mathcal{G}_i$ in Line /6/. This implies that $\mathrm{lt}_\prec \bar{g} \notin \langle \mathrm{lt}_\prec \mathcal{G}_i \rangle$ and thus we obtain a strictly ascending chain of monomial ideals: $\langle \mathrm{lt}_\prec \mathcal{G}_1 \rangle \subsetneq \langle \mathrm{lt}_\prec \mathcal{G}_2 \rangle \subsetneq \langle \mathrm{lt}_\prec \mathcal{G}_3 \rangle \subsetneq \cdots$. By Dickson's Lemma it is not possible that such a chain becomes infinite. Hence after a finite number of steps we must have $\mathcal{G}_i = \mathcal{G}_{i+1}$ and the algorithm terminates with a Gröbner basis of the ideal $\langle \mathcal{F} \rangle$. $\square$

Example B.2.14. We apply Buchberger's algorithm to the set $\mathcal{F} = \{f_1, f_2\}$ of Example B.2.4 and some degree compatible term order $\prec$. The S-polynomial of the two generators is $S_\prec(f_1, f_2) = x^2 - xy$ and a normal form of it with respect to $\mathcal{F}$ is given by $f_3 = \underline{x^2} - x$. As one can easily check by explicitly computing the normal forms of the two new S-polynomials $S_\prec(f_1, f_3)$ and $S_\prec(f_2, f_3)$, the set $\mathcal{G} = \{f_1, f_2, f_3\}$ is a Gröbner basis (even a reduced one) of the ideal $\langle \mathcal{F} \rangle$ for any degree compatible term order. $\triangleleft$

Note that it is *not* recommended to use the algorithm in the described form for concrete computations! The form above makes the basic ideas fairly transparent, but it is not efficient. Furthermore, Algorithm B.2 does not determine a reduced Gröbner basis; this would require additional normal form computations each time a new element has been added to the basis.

The choice of the pairs $\{g_1, g_2\}$ in Line /4/ is important for the efficency. A number of strategies have been developed for it. The classical *normal selection strategy* chooses that pair for which the term $\mathrm{lcm}(\mathrm{lt}_\prec g_1, \mathrm{lt}_\prec g_2)$ is minimal with respect to the used term order $\prec$. It works very well for degree compatible term orders. Another popular approach is the "sugar cube" strategy [60] which is based on a simulated homogenisation of the input. In contrast to the normal selection strategy, this approach works well also for term orders like the purely lexicographic one.

Another crucial observation is that the algorithm spends most of its time computing normal forms. Buchberger found two criteria for predicting that certain S-polynomials reduce to zero. Their use often leads to drastic gains in efficiency, as the corresponding pairs $\{g_1, g_2\}$ can be thrown away without any computation.

Proposition B.2.15 (Buchberger's first criterion). *Let* $g_1, g_2 \in \mathcal{P}$ *be two polynomials such that* $\gcd(\mathrm{lt}_\prec g_1, \mathrm{lt}_\prec g_2) = 1$. *Then* $\mathrm{NF}(S_\prec(g_1, g_2), \{g_1, g_2\}) = 0$, *i. e. the S-polynomial* $S_\prec(g_1, g_2)$ *reduces to zero modulo the set* $\{g_1, g_2\}$.

Whereas the first criterion gives an immediate reducibility, the second one predicts that a certain S-polynomial will become reducible at some stage of the algorithm, possibly much later. Thus it is a considerably deeper result.

Proposition B.2.16 (Buchberger's second criterion). *Let* $\mathcal{G} \subset \mathcal{P}$ *be a finite set and* $g_1, g_2, h \in \mathcal{G}$ *such that* $\mathrm{lt}_\prec h \mid \mathrm{lcm}(\mathrm{lt}_\prec g_1, \mathrm{lt}_\prec g_2)$. *Then it suffices to treat in Algorithm B.2 the S-polynomials* $S_\prec(g_1, h)$ *and* $S_\prec(g_2, h)$; *the S-polynomial* $S_\prec(g_1, g_2)$ *may be ignored.*

Various ways exist to incorporate these criteria in Algorithm B.2, but we omit here a discussion. The proof of the first criterion is not very difficult; it follows from a direct computation. The second one is based on syzygy theory and we will briefly mention the underlying theory in the next section.

Optimisations of the Buchberger algorithm (and the design of alternative algorithms for the construction of Gröbner bases) is still a major research topic in computer algebra and we must refer to the literature (cited in the above mentioned textbooks). The complexity of computing Gröbner bases and deciding the ideal membership problem was thoroughly analysed by Mayr and collaborators in a number of publications; a survey with many references is given in [119]. A by now classical result is that the ideal membership problem for polynomials with rational coefficients is in the worst case exponential in space [118, 120]. The same holds for determining a reduced Gröbner basis.

One should stress that this complexity result is less a question of designing clever algorithms but ideals exist where the *size* of the basis is so huge, as Mayr and Meyer [120] demonstrated with an explicit example. Thus simply storing the basis requires already exponential space. At first sight, this seems to imply that it is hopeless to compute Gröbner bases. But one should always keep in mind that all these results concern the *worst case* complexity. There are some indications that many examples appearing in applications have a considerably lower complexity, as they are of a more geometric origin whereas all the known worst case examples are of a combinatorial origin. Some discussions of this and related issues can be found in [13].

All the results in this section apply to polynomial rings over a *field* $\Bbbk$. This restriction is important; we have nowhere payed particular attention to the coefficients, as we could always assume that we can divide by them, i. e. divisibility of monomials was solely a matter of the terms involved. This is no longer the case, if we consider polynomials with coefficients out of some ring. There exists a theory of Gröbner bases for this more general situation (see e. g. [4, Chapt. 4]), but things are getting more complicated.

Gröbner bases are defined for submodules of free modules in the same manner as for ideals; one only needs a bit more notation. We represent the elements of the free module $\mathcal{P}^m$ as vectors $\mathbf{f} = (f_1, \dots, f_m)$ with $f_\alpha \in \mathcal{P}$. The unit vectors $\mathbf{e}_\alpha$ with $e_{\alpha\beta} = \delta_{\alpha\beta}$ and $1 \leq \alpha \leq m$ form the standard basis of the module $\mathcal{P}^m$; thus $\mathbf{f} = f_1\mathbf{e}_1 + \cdots + f_m\mathbf{e}_m$.

First, we must extend the concept of a term order. A term $\mathbf{t}$ is now a vector of the form $\mathbf{t} = t\mathbf{e}_\alpha$ for some α and with $t \in \mathbb{T}$ a term in $\mathcal{P}$. We denote the set of all terms by $\mathbb{T}_m$; it is a monoid module over $\mathbb{T}$. By a slight abuse of notation, we will sometimes "divide" terms: if $\mathbf{t} = r\mathbf{s}$, then we set $\mathbf{t}/\mathbf{s} = r\mathbf{e}_\alpha$; thus the result is again a vector. Terms belonging to different unit vectors cannot be divided. A *term order* $\prec$ on $\mathbb{T}_m$ must satisfy the following two conditions: if $\mathbf{s}, \mathbf{t} \in \mathbb{T}_m$ are two arbitrary terms, then (i) $\mathbf{t} \preceq r\mathbf{t}$ and (ii) $\mathbf{s} \prec \mathbf{t}$ implies $r\mathbf{s} \prec r\mathbf{t}$ for all $r \in \mathbb{T}$. Obviously, these are simple generalisations of the conditions for a term order on $\mathbb{T}$ and assert that $\prec$ is compatible with the structure of $\mathbb{T}_m$ as a monoid module over $\mathbb{T}$.

In practice one uses on $\mathbb{T}_m$ mostly term orders derived from orders on $\mathbb{T}$. Two natural ways to "lift" an order are *TOP* and *POT*. These abbreviations stand for *Term Over Position* and *Position Over Term*, respectively. Let $\prec$ be a term order on $\mathbb{T}$. Then $s\mathbf{e}_\alpha \prec_{\text{TOP}} t\mathbf{e}_\beta$, if either $s \prec t$ or $s = t$ and $\alpha < \beta$, i. e. we rank the comparison of the terms s, t over the comparison of their position in the vector. We have $s\mathbf{e}_\alpha \prec_{\text{POT}} t\mathbf{e}_\beta$, if either $\alpha < \beta$ or

$\alpha = \beta$ and $s \prec t$, i.e. the position in the vector is more important than the terms. It is easy to show that both $\prec_{TOP}$ and $\prec_{POT}$ are term orders on $\mathbb{T}_m$.

The notion of a class respecting term order can trivially be extended to orders on $\mathbb{T}_m$. In a similar manner to the proof of Lemma B.2.2 one shows that the only class respecting term orders are the TOP lifts of the degree reverse lexicographic order for arbitrary labellings of the unit vectors $\mathbf{e}_\alpha$.

Given this definition of term orders on $\mathbb{T}_m$, we can straightforwardly extend all the notions introduced in the last section to submodules $\mathcal{M} \subseteq \mathcal{P}^m$. The leading term, coefficient, and monomial are defined as before; again we simply write $\operatorname{lt}_\prec \mathcal{M}$ for the monomial submodule generated by the leading terms of the elements of $\mathcal{M}$. Now we can easily generalise Definition B.2.3.

Definition B.2.17. *Let $\mathcal{M}$ be a submodule of the free module $\mathcal{P}^m$. A finite set $\mathcal{G} \subset \mathcal{P}^m$ is called a* Gröbner basis *of $\mathcal{M}$ for a given term order $\prec$, if $\operatorname{lt}_\prec \mathcal{M} = \langle \operatorname{lt}_\prec \mathcal{G} \rangle$.*

We may continue to use Algorithm B.1 for computing normal forms. In order to apply Buchberger's Algorithm B.2, we still have to generalise the S-polynomials. First, we introduce the least common multiple of two terms $\mathbf{s} = s\mathbf{e}_\alpha$ and $\mathbf{t} = t\mathbf{e}_\beta$ in $\mathbb{T}_m$: it is zero, if $\alpha \neq \beta$, and $\operatorname{lcm}(s,t)\mathbf{e}_\alpha$ otherwise. Then we define the S-"polynomial" of two elements $\mathbf{f}, \mathbf{g} \in \mathcal{P}^m$ by simply copying (B.10):

$$\mathbf{S}_\prec(\mathbf{f},\mathbf{g}) = \frac{\operatorname{lcm}(\operatorname{lm}_\prec \mathbf{f}, \operatorname{lm}_\prec \mathbf{g})}{\operatorname{lm}_\prec \mathbf{f}} \cdot \mathbf{f} - \frac{\operatorname{lcm}(\operatorname{lm}_\prec \mathbf{f}, \operatorname{lm}_\prec \mathbf{g})}{\operatorname{lm}_\prec \mathbf{g}} \cdot \mathbf{g}\,. \tag{B.11}$$

Now the results and algorithms of Appendix B.2 can be straightforwardly extended to the module case. In particular, it turns out that all mentioned propositions and theorems remain valid.

B.3 Syzygies and Free Resolutions

Like many other things in commutative algebra, the term *syzygy* goes back to Hilbert. It comes from the Greek word for yoke and indeed a syzygy defines a "yoke" connecting some module elements.

Definition B.3.1. *Let $\mathcal{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_s\}$ be a finite set of elements of the free module $\mathcal{P}^m$ where again $\mathcal{P} = \Bbbk[x^1, \dots, x^n]$. A* syzygy *of $\mathcal{F}$ is an element $\mathbf{h} = (h_1, \dots, h_s) \in \mathcal{P}^s$ such that*

$$\sum_{\tau=1}^{s} h_\tau \mathbf{f}_\tau = 0\,. \tag{B.12}$$

The syzygies of $\mathcal{F}$ form a submodule of $\mathcal{P}^s$, the syzygy module $\operatorname{Syz}(\mathcal{F})$.

The components of a syzygy can be understood as solutions of a linear system of equations over the ring $\mathcal{P}$. Its $m \times s$ matrix is easily obtained, if we write the module elements $\mathbf{f}_\tau$ as m-dimensional column vectors. In other words, $\mathrm{Syz}(\mathcal{F})$ is the kernel of the linear map

$$\phi_{\mathcal{F}} : \begin{cases} \mathcal{P}^s & \longrightarrow & \mathcal{P}^m \\ (h_1, \dots, h_s) & \longmapsto & h_1\mathbf{f}_1 + \cdots + h_s\mathbf{f}_s \end{cases} . \tag{B.13}$$

Thus $\mathrm{Syz}(\mathcal{F}) = \{0\}$, if and only if $\mathcal{F}$ is the basis of a free module.

A generating set of the module $\mathrm{Syz}(\mathcal{F})$ can be easily determined, if $\mathcal{F}$ is a Gröbner basis of $\langle\mathcal{F}\rangle$ (for some term order $\prec$). For an arbitrary set $\mathcal{F}$ one first computes a Gröbner basis $\mathcal{G}$ of the ideal $\langle\mathcal{F}\rangle$ and then determines $\mathrm{Syz}(\mathcal{F})$ from $\mathrm{Syz}(\mathcal{G})$ (see e.g. [4]). So let $\mathcal{G} = \{\mathbf{g}_1, \dots, \mathbf{g}_s\}$ be a Gröbner basis for some term order $\prec$; without loss of generality we may assume that $\mathcal{G}$ is reduced. Furthermore, let $\mathbf{t}_\sigma = \mathrm{lt}_\prec \mathbf{g}_\sigma$ and $\mathbf{t}_{\sigma\tau} = \mathrm{lcm}(\mathbf{t}_\sigma, \mathbf{t}_\tau)$. It follows from Proposition B.2.7 (iii) that we can write

$$\mathbf{S}_\prec(\mathbf{g}_\rho, \mathbf{g}_\sigma) = \sum_{\tau=1}^{s} h_{\rho\sigma\tau}\mathbf{g}_\tau \tag{B.14}$$

where the leading terms of the polynomials $h_{\rho\sigma\tau} \in \mathcal{P}$ satisfy $\mathrm{lt}_\prec \mathbf{S}_\prec(\mathbf{g}_\rho, \mathbf{g}_\sigma) = \max_\prec \{\mathrm{lt}_\prec(h_{\rho\sigma\tau}\mathbf{g}_\tau) \mid 1 \leq \tau \leq s\}$. Now we define for $\rho \neq \sigma$

$$\mathbf{S}_{\rho\sigma} = \frac{\mathbf{t}_{\rho\sigma}}{\mathbf{t}_\rho}\boldsymbol{e}_\rho - \frac{\mathbf{t}_{\rho\sigma}}{\mathbf{t}_\sigma}\boldsymbol{e}_\sigma - \mathbf{h}_{\rho\sigma} \in \mathcal{P}^s \tag{B.15}$$

with $\mathbf{h}_{\rho\sigma} = (h_{\rho\sigma 1}, \dots, h_{\rho\sigma s})$. By the definition of the S-"polynomial" and (B.14), this is a syzygy in $\mathrm{Syz}(\mathcal{G})$.

Theorem B.3.2. *Let $\mathcal{G}$ be a Gröbner basis. The syzygy module* $\mathrm{Syz}(\mathcal{G})$ *is generated by the set* $\mathcal{S}_{\mathcal{G}} = \{\mathbf{S}_{\rho\sigma} \mid 1 \leq \rho < \sigma \leq s\}$.

Thus we automatically obtain a generating set of the module $\mathrm{Syz}(\mathcal{G})$, if we compute the Gröbner basis $\mathcal{G}$ with Buchberger's algorithm and use the extended form of Algorithm B.1, the division algorithm discussed after Example B.2.6. It delivers us the coefficients $h_{\tau\sigma\rho}$ of the representation (B.14).

An interesting result given by Schreyer [156] in his diploma thesis is that this generating set is even a Gröbner basis for a certain term order. Let $\mathcal{F} = \{\mathbf{f}_1, \dots, \mathbf{f}_s\}$ be a subset of $\mathcal{P}^m$ and $\prec$ a term order on $\mathbb{T}_m$. We define an order $\prec_{\mathcal{F}}$ on $\mathbb{T}_s$ by setting $s\boldsymbol{e}_\sigma \prec_{\mathcal{F}} t\boldsymbol{e}_\tau$ if either $\mathrm{lt}_\prec(s\mathbf{f}_\sigma) \prec \mathrm{lt}_\prec(t\mathbf{f}_\tau)$ or $\mathrm{lt}_\prec(s\mathbf{f}_\sigma) = \mathrm{lt}_\prec(t\mathbf{f}_\tau)$ and $\tau < \sigma$ (note that we reverse the order of the indices). One can show that for arbitrary sets $\mathcal{F}$ (not necessarily Gröbner bases) $\prec_{\mathcal{F}}$ is a term order on $\mathbb{T}_s$.

Theorem B.3.3 (Schreyer). *The set $\mathcal{S}_\mathcal{G}$ is a Gröbner basis of the syzygy module* $\mathrm{Syz}(\mathcal{G})$ *with respect to the term order $\prec_\mathcal{G}$.*

Syzygy theory is also underlying Buchberger's second criterion (Proposition B.2.16). Let $\mathcal{T} = \{t_1, \dots, t_s\} \subset \mathbb{T}$ be a finite set of terms. As above we set $t_{\rho\sigma} = \mathrm{lcm}(t_\rho, t_\sigma)$ and $\mathbf{S}_{\rho\sigma} = \frac{t_{\rho\sigma}}{t_\rho}\mathbf{e}_\rho - \frac{t_{\rho\sigma}}{t_\sigma}\mathbf{e}_\sigma$. As we are only dealing with terms, no polynomials $h_{\rho\sigma}$ arise. Setting $t_{\rho\sigma\tau} = \mathrm{lcm}(t_\rho, t_\sigma, t_\tau)$, it is a simple exercise to verify by direct computation that

$$\frac{t_{\rho\sigma\tau}}{t_{\rho\sigma}}\mathbf{S}_{\rho\sigma} + \frac{t_{\rho\sigma\tau}}{t_{\sigma\tau}}\mathbf{S}_{\sigma\tau} + \frac{t_{\rho\sigma\tau}}{t_{\tau\rho}}\mathbf{S}_{\tau\rho} = 0\,. \tag{B.16}$$

An obvious consequence is that if $t_\tau \mid t_{\rho\sigma}$, then $\mathbf{S}_{\rho\sigma} \in \langle \mathbf{S}_{\sigma\tau}, \mathbf{S}_{\tau\rho}\rangle$. More generally, we may formulate the following statement.

Lemma B.3.4. *Let $\mathcal{S} \subseteq \mathcal{S}_\mathcal{T} = \{\mathbf{S}_{\rho\sigma} \mid 1 \leq \rho < \sigma \leq s\}$ be a generating set of* $\mathrm{Syz}(\mathcal{T})$*. Assume that the three pairwise distinct indices ρ, σ, τ are such that*[13] $\mathbf{S}_{\rho\sigma}, \mathbf{S}_{\sigma\tau}, \mathbf{S}_{\rho\tau} \in \mathcal{S}$ *and $t_\tau \mid t_{\rho\sigma}$. Then the smaller set $\mathcal{S} \setminus \{\mathbf{S}_{\rho\sigma}\}$ still generates* $\mathrm{Syz}(\mathcal{T})$*.*

Buchberger's second criterion is now a simple corollary to a generalisation of Proposition B.2.12. If $\mathcal{G} \subset \mathcal{P}$ is a finite set, we take $\mathcal{T} = \mathrm{lt}_\prec \mathcal{G}$. Given a subset $\mathcal{S} \subseteq \mathcal{S}_\mathcal{T}$ that generates $\mathrm{Syz}(\mathcal{T})$, one can show that it suffices for $\mathcal{G}$ being a Gröbner basis, if all S-polynomials $S_\prec(g_\rho, g_\sigma)$ with $\mathbf{S}_{\rho\sigma} \in \mathcal{S}$ reduce to zero modulo $\mathcal{G}$.

Definition B.3.5. *Let $\mathcal{R}$ be a ring and $\mathcal{M}$ an $\mathcal{R}$-module. An exact sequence*

$$\cdots \longrightarrow \mathcal{R}^{s_j} \xrightarrow{\psi_j} \mathcal{R}^{s_{j-1}} \longrightarrow \cdots \longrightarrow \mathcal{R}^{s_1} \xrightarrow{\psi_1} \mathcal{R}^{s_0} \xrightarrow{\psi_0} \mathcal{M} \longrightarrow 0\,, \tag{B.17}$$

i. e. a sequence where $\mathrm{im}\,\psi_{j+1} = \ker\psi_j$ for all $j \geq 0$ and $\mathrm{im}\,\psi_0 = \mathcal{M}$, is called a free resolution *of $\mathcal{M}$. If there exists an index ℓ such that $s_\ell \neq 0$ but $s_j = 0$ for all $j > \ell$, the resolution is* finite *and its* length *is $\ell + 1$.*

For a finitely generated $\mathcal{R}$-module $\mathcal{M}$ we can construct such a free resolution step by step using presentations by generators and relations. Assume we know such a presentation for $\mathcal{M}$, i. e. $\mathcal{M} \cong \mathcal{R}^{s_0}/\mathcal{M}_0$. This is equivalent to the short exact sequence

$$0 \longrightarrow \mathcal{M}_0 \xrightarrow{\iota_0} \mathcal{R}^{s_0} \xrightarrow{\phi_0} \mathcal{M} \longrightarrow 0 \tag{B.18}$$

where ι_0 is the inclusion map and ϕ_0 the map defined by (B.1). As $\mathcal{M}_0$ is again finitely generated, we can find a presentation $\mathcal{M}_0 \cong \mathcal{R}^{s_1}/\mathcal{M}_1$ leading to

[13] If $\rho > \sigma$, then we understand that $\mathbf{S}_{\sigma\rho} \in \mathcal{S}$ etc.

an analogous short exact sequence with maps ι_1 and ϕ_1. If we define $\psi_0 = \phi_0$ and $\psi_1 = \iota_0 \circ \phi_1$, we obtain the exact sequence

$$0 \longrightarrow \mathcal{M}_1 \xrightarrow{\iota_1} \mathcal{R}^{s_1} \xrightarrow{\psi_1} \mathcal{R}^{s_0} \xrightarrow{\psi_0} \mathcal{M} \longrightarrow 0\,. \tag{B.19}$$

We can iterate this construction taking a presentation of $\mathcal{M}_1$ leading to a finitely generated module $\mathcal{M}_2$ and so on. Each presentation leads to a short exact sequence of the form (B.18) with associated maps ι_i and ϕ_i. Defining $\psi_i = \iota_{i-1} \circ \phi_i$ we obtain the following picture where the diagonals come from the presentations and the central row defines a free resolution of $\mathcal{M}$:

$$\begin{tikzcd}
 & & 0 \arrow[dr] & & 0 & & & & \\
 & & & \mathcal{M}_1 \arrow[ur] \arrow[dr] & & & & & \\
\cdots \arrow[rr, "\psi_3"] \arrow[dr] & & \mathcal{R}^{s_2} \arrow[ur] \arrow[rr, "\psi_2"] & & \mathcal{R}^{s_1} \arrow[rr, "\psi_1"] \arrow[dr] & & \mathcal{R}^{s_0} \arrow[r, "\psi_0"] & \mathcal{M} \arrow[r] & 0\,. \\
 & \mathcal{M}_2 \arrow[ur] \arrow[dr] & & & & M_0 \arrow[ur] \arrow[dr] & & & \\
0 \arrow[ur] & & 0 & & 0 \arrow[ur] & & 0 & &
\end{tikzcd} \tag{B.20}$$

We are mainly interested in the case that $\mathcal{R} = \mathcal{P}$. Here explicit presentations are easily obtained. Let the finite set $\mathcal{F}$ generate the $\mathcal{P}$-module $\mathcal{M}$ and take $\mathrm{Syz}(\mathcal{F})$ as $\mathcal{M}_0$. If $\mathcal{F}$ is a Gröbner basis of $\mathcal{M}$, the Schreyer Theorem B.3.3 provides us with a Gröbner basis of $\mathrm{Syz}(\mathcal{F})$, so that we can iterate the construction obtaining a *syzygy resolution* of $\mathcal{M}$. Hilbert's famous Syzygy Theorem states that any finitely generated polynomial module has a finite free resolution.

Theorem B.3.6 (Hilbert). *Let* $\mathcal{P} = \Bbbk[x^1, \dots, x^n]$. *Every finitely generated* $\mathcal{P}$-*module has a finite free resolution whose length is at most* n.

Free resolutions are not unique. But if we impose an additional condition, then we obtain a resolution that is unique up to isomorphism. Denote by $\mathcal{P}_+$ the ideal $\langle x^1, \dots, x^n \rangle$.

Definition B.3.7. *A free resolution of a* $\mathcal{P}$-*module is called* minimal, *if the maps* $\psi_j : \mathcal{P}^{s_j} \to \mathcal{P}^{s_{j-1}}$ *satisfy* $\operatorname{im} \psi_j \subseteq \mathcal{P}_+ \cdot \mathcal{P}^{s_{j-1}}$. *In this case, the numbers* s_j *are the* Betti numbers *of* $\mathcal{M}$ *and the length of the resolution is the* projective dimension $\operatorname{proj\,dim} \mathcal{M}$.

C. Differential Geometry

As we are concerned with a geometric approach to differential equations, we need some basic notions in differential geometry which are collected in this appendix. We start with a section introducing (differentiable) manifolds and some spaces associated with them: the tangent and the cotangent bundle. The following section extends to a slightly more complicated concept, fibred manifolds, and also briefly introduces bundles, although we hardly need their additional structures. The final section studies in some more detail vector fields and differential forms and the basic operations with them.

Standard references on differential geometry are [2, 21, 91, 92, 103, 191], but they go much further than we need. For most of our purposes a working knowledge of some basic notions around manifolds is sufficient. Much of the material we need can also be found in a very accessible form in [129].

C.1 Manifolds

The jet bundle formalism is based on fibred manifolds, but before we study these we briefly recall the definition of an ordinary manifold. Loosely speaking, it is something that locally looks like an $\mathbb{R}^m$. When we speak of a manifold, we will always mean a finite-dimensional, smooth (C^∞) manifold.

Definition C.1.1. *An m-dimensional* manifold *is a set $\mathcal{M}$ with a countable family $\mathcal{A}$ of* coordinate charts *$(\mathcal{U}, \phi_\mathcal{U})$ where $\mathcal{U} \subseteq \mathcal{M}$ is a subset and $\phi_\mathcal{U} : \mathcal{U} \to \Omega_\mathcal{U}$ a map onto a connected open subset $\Omega_\mathcal{U} \subseteq \mathbb{R}^m$ such that:*

1. *The charts in $\mathcal{A}$ cover the whole manifold: $\mathcal{M} = \bigcup_{(\mathcal{U},\phi_\mathcal{U})\in\mathcal{A}} \mathcal{U}$.*
2. *If two charts $(\mathcal{U}, \phi_\mathcal{U}), (\mathcal{V}, \phi_\mathcal{V}) \in \mathcal{A}$ overlap, the* transition function
$$\psi_{\mathcal{U}\mathcal{V}} = \phi_\mathcal{V} \circ \phi_\mathcal{U}^{-1} : \phi_\mathcal{U}(\mathcal{U} \cap \mathcal{V}) \subseteq \mathbb{R}^m \longrightarrow \phi_\mathcal{V}(\mathcal{U} \cap \mathcal{V}) \subseteq \mathbb{R}^m \tag{C.1}$$
is a smooth function.
3. *If $p_1 \in \mathcal{U}_1$ and $p_2 \in \mathcal{U}_2$ are two distinct points of $\mathcal{M}$, there exist open neighbourhoods $\Omega_1 \subseteq \Omega_{\mathcal{U}_1}$ of $\phi_{\mathcal{U}_1}(p_1)$ and $\Omega_2 \subseteq \Omega_{\mathcal{U}_2}$ of $\phi_{\mathcal{U}_2}(p_2)$ such that $\phi_{\mathcal{U}_1}^{-1}(\Omega_1) \cap \phi_{\mathcal{U}_2}^{-1}(\Omega_2) = \emptyset$.*

$\mathcal{A}$ is called an *atlas* and provides the set $\mathcal{M}$ with a *differentiable structure.* The charts endow the manifold $\mathcal{M}$ with a *topology*: the open sets are the inverse images $\phi_{\mathcal{U}}^{-1}(\Omega)$ of open sets $\Omega \subseteq \Omega_{\mathcal{U}}$ in $\mathbb{R}^m$. The third point in the definition above states that $\mathcal{M}$ with this topology is a Hausdorff space. These topological aspects are not so important for us, as we are mainly working locally, i. e. we stay within a chart $(\mathcal{U}, \phi_{\mathcal{U}}) \in \mathcal{A}$ where we can map everything into $\mathbb{R}^m$ and then calculate in coordinates as usual. Typically, we do not explicitly mention the chart but simply say *"let $\mathbf{x} = (x^1, \dots, x^m)$ be local coordinates on $\mathcal{M}$"*. This means that we go into a chart $(\mathcal{U}, \phi_{\mathcal{U}}) \in \mathcal{A}$ and identify each point $p \in \mathcal{U}$ with the vector $\mathbf{x} = \phi_{\mathcal{U}}(p) \in \mathbb{R}^m$.

One can also define analytic or C^k or topological manifolds; in this case one requires that the transition functions $\psi_{\mathcal{U}\mathcal{V}}$ describing changes of coordinates are of class C^ω or C^k or just homeomorphisms, respectively. Infinite-dimensional manifolds are usually obtained by modelling them on some Banach space instead of $\mathbb{R}^m$.

Example C.1.2. The simplest example of a manifold is a connected open subset $\Omega \subseteq \mathbb{R}^m$. In this case $\mathcal{A}$ contains only one chart, namely $(\Omega, \mathrm{id}_\Omega)$.

A slightly more complicated example of an m-dimensional manifold is the surface S^m of the unit sphere in $\mathbb{R}^{m+1}$. Here it is not possible to cover the whole manifold with a single chart, as S^m is not homeomorphic to $\mathbb{R}^m$; one needs at least two charts using the stereographic projection. $\triangleleft$

The space $C^\infty(\mathcal{M}, \mathbb{R})$ of all smooth real-valued functions defined on the manifold $\mathcal{M}$ is denoted by $\mathcal{F}(\mathcal{M})$. It is a ring, as we can add and multiply such functions simply by adding and multiplying their values. A function $f \in \mathcal{F}(\mathcal{M})$ induces in a chart $(\mathcal{U}, \phi_{\mathcal{U}})$ a function $f_{\mathcal{U}} = f \circ \phi_{\mathcal{U}} : \Omega_{\mathcal{U}} \to \mathbb{R}$. We call f differentiable in $\mathcal{U}$, if $f_{\mathcal{U}}$ is differentiable. It is easy to see that differentiability at a point $p \in \mathcal{M}$ is independent of the used chart, as the transition functions are smooth.

A familiar object from elementary differential geometry is the tangent plane of a surface. Its generalisation to an arbitrary manifold $\mathcal{M}$ leads to the tangent bundle $T\mathcal{M}$. We start by introducing the tangent space $T_p\mathcal{M}$ at an arbitrary but fixed point $p \in \mathcal{M}$. There exist several possibilities; we use a rather geometric one which also stresses the close relation between tangent and jet bundles.

A curve on the manifold $\mathcal{M}$ is a differentiable map $\gamma : \mathcal{I} \subseteq \mathbb{R} \to \mathcal{M}$ from some real interval into $\mathcal{M}$. Without loss of generality we assume that $0 \in \mathcal{I}$. For the construction of $T_p\mathcal{M}$ we take all curves with $\gamma(0) = p$. Two such curves γ_1, γ_2 are *tangent* at p, if in some coordinate chart $(\mathcal{U}, \phi_{\mathcal{U}})$ with $p \in \mathcal{U}$ the tangent vectors $(\phi_{\mathcal{U}} \circ \gamma_1)'(0)$ and $(\phi_{\mathcal{U}} \circ \gamma_2)'(0)$ coincide (note that $\phi_{\mathcal{U}} \circ \gamma_i$ is a curve in $\mathbb{R}^m$ where we know tangent vectors from calculus). One can show that this definition is independent of the used chart.

Let $[\gamma]_p$ be the set of all curves $\bar{\gamma}$ with $\bar{\gamma}(0) = p$ which are tangent to γ at p. An element of $T_p\mathcal{M}$ is such an equivalence class of curves. It is not difficult to see that $T_p\mathcal{M}$ is a real vector space of the same dimension as the manifold $\mathcal{M}$, as a coordinate chart maps the elements of $T_p\mathcal{M}$ bijectively on vectors in $\mathbb{R}^m$.

Definition C.1.3. *Let $p \in \mathcal{M}$ be a point on the manifold $\mathcal{M}$. The* tangent bundle $T\mathcal{M}$ *is the union of all tangent spaces*

$$T\mathcal{M} = \bigcup_{p \in \mathcal{M}} \{p\} \times T_p\mathcal{M} \,. \tag{C.2}$$

The tangent bundle projection *is the map $\tau_{\mathcal{M}} : T\mathcal{M} \to \mathcal{M}$ defined by $\tau_{\mathcal{M}}(\{p\} \times [\gamma]_p) = p$ for any $[\gamma]_p \in T_p\mathcal{M}$.*

Let $\mathcal{M}$ and $\mathcal{N}$ be two manifolds and $f : \mathcal{M} \to N$ an arbitrary smooth map. It induces a map $Tf : T\mathcal{M} \to T\mathcal{N}$, the *tangent map*, which can be defined intrinsically as follows. Let $\gamma_{\mathcal{M}} : \mathcal{I} \subseteq \mathbb{R} \to \mathcal{M}$ be an arbitrary curve on $\mathcal{M}$. Together with f it defines a curve $\gamma_{\mathcal{N}} : \mathcal{I} \to \mathcal{N}$ on the manifold $\mathcal{N}$ by $\gamma_{\mathcal{N}} = f \circ \gamma_{\mathcal{M}}$. By definition, a point in $T_p\mathcal{M}$ is an equivalence class of such curves $\gamma_{\mathcal{M}}$ with $\gamma_{\mathcal{M}}(0) = p$. Obviously, all curves in this class are mapped on curves $\gamma_{\mathcal{N}}$ with $\gamma_{\mathcal{N}}(0) = f(p)$ and it is not difficult to see that in fact the class $[\gamma_{\mathcal{M}}]_p$ is mapped into the class $[\gamma_{\mathcal{N}}]_{f(p)}$. Thus we obtain a *linear* map $T_pf : T_p\mathcal{M} \to T_{f(p)}\mathcal{N}$. Putting together all these maps for different p yields the map Tf.

The tangent bundle $T\mathcal{M}$ is again a manifold, as the differentiable structure of $\mathcal{M}$ induces a differentiable structure on $T\mathcal{M}$. Namely, let $\mathcal{A}$ be an atlas of $\mathcal{M}$; then $T\mathcal{A} = \{(T\mathcal{U}, T\phi_{\mathcal{U}}) \mid (\mathcal{U}, \phi_{\mathcal{U}}) \in \mathcal{A}\}$ is an atlas of $T\mathcal{M}$. We have $\dim T\mathcal{M} = 2\dim\mathcal{M}$. Let $(\mathcal{U}, \phi_{\mathcal{U}})$ be a coordinate chart with $p \in \mathcal{U}$ and $\phi_{\mathcal{U}}(p) = \mathbf{x} = (x^1, \dots, x^m)$. Consider for $1 \leq i \leq m$ the curves $\gamma_i : \mathcal{I} \to \mathcal{M}$ with $(\phi_{\mathcal{U}} \circ \gamma_i)(t) = \mathbf{x} + t \cdot \mathbf{e}_i$ where $\mathbf{e}_i$ is the ith unit vector in $\mathbb{R}^m$. We denote the vector $[\gamma_i]_p$ by $\partial_{x^i}\big|_p$. These vectors form a basis of $T_p\mathcal{M}$. One often denotes coordinates with respect to this basis by $\dot{\mathbf{x}} = (\dot{x}^1, \dots, \dot{x}^m)$, so that local coordinates on $T\mathcal{M}$ have the form $(\mathbf{x}, \dot{\mathbf{x}}) \in \mathbb{R}^m \times \mathbb{R}^m$.

Locally, the tangent map Tf corresponds to the linearisation of f via its Jacobian (sometimes denoted df). Let us introduce coordinates $(\mathbf{x}, \dot{\mathbf{x}})$ on $T\mathcal{M}$ (with some chart $(\mathcal{U}, \phi_{\mathcal{U}})$ on $\mathcal{M}$) and $(\mathbf{y}, \dot{\mathbf{y}})$ on $T\mathcal{N}$ (with some chart $(\mathcal{V}, \psi_{\mathcal{V}})$ on $\mathcal{N}$). We write $\mathbf{y} = \bar{f}(\mathbf{x})$, where $\bar{f} = \phi_{\mathcal{U}}^{-1} \circ f \circ \psi_{\mathcal{V}} : \mathcal{U} \to \mathcal{V}$ is the coordinate representation of f. Then Tf has in these coordinates the following form:

$$\mathsf{T}f : \begin{cases} \mathsf{T}\mathcal{U} & \longrightarrow & \mathsf{T}\mathcal{V} \\ (\mathbf{x}, \dot{\mathbf{x}}) & \longmapsto & \left(\bar{f}(\mathbf{x}), \dfrac{\partial \bar{f}}{\partial \mathbf{x}}(\mathbf{x})\dot{\mathbf{x}}\right) \end{cases} . \tag{C.3}$$

The *rank* of the map f at a point $p \in \mathcal{U}$ is defined as the rank of the Jacobian $\partial\bar{f}/\partial\mathbf{x}$ at the point $\phi_{\mathcal{U}}(p)$. One can show that this definition is independent of the chosen coordinate charts. If $\mathcal{K}$ is a further manifold and $g : \mathcal{N} \to \mathcal{K}$ a further map, the familiar *chain rule* can be expressed as $\mathsf{T}(g \circ f) = \mathsf{T}g \circ \mathsf{T}f$.

Mostly, we are somewhat sloppy with our notations and do not distinguish between the map f and its coordinate representation $\bar{f}$. Thus we simply write $\mathbf{y} = f(\mathbf{x})$ omitting the coordinate maps $\phi_{\mathcal{U}}$ and $\psi_{\mathcal{V}}$.

Let $f : \mathcal{M} \to \mathcal{N}$ be a smooth map. We assume for simplicity that rank f is constant on the manifold $\mathcal{M}$. Maps with the maximal possible rank (which is obviously the minimum of $\dim\mathcal{M}$ and $\dim\mathcal{N}$) are of particular importance and have got special names: if $\dim\mathcal{M} \leq \dim\mathcal{N}$, one speaks of an *immersion*; for $\dim\mathcal{M} \geq \dim\mathcal{N}$ of a *submersion*. Locally, immersions look like linear injections, submersions like projections (see [61, Section I.2] for a precise form of this statement). If f is a bijection and f^{-1} is also smooth, then f is called a *diffeomorphism* and the two manifolds $\mathcal{M}$ and $\mathcal{N}$ are *diffeomorphic*. Obviously, this requires that $\mathsf{T}f$ is an isomorphism with $(\mathsf{T}f)^{-1} = \mathsf{T}(f^{-1})$ and thus $\dim\mathcal{M} = \dim\mathcal{N}$.

An n-dimensional *submanifold* of a manifold $\mathcal{M}$ is a subset $\mathcal{N} \subseteq \mathcal{M}$ where for every point $p \in \mathcal{N}$ there exists a coordinate chart $\mathcal{U} \subseteq \mathcal{M}$ with a coordinate map $\phi_{\mathcal{U}} : \mathcal{U} \to \mathbb{R}^n \times \mathbb{R}^{m-n}$ such that $\phi_{\mathcal{U}}(\mathcal{N} \cap \mathcal{U}) = \phi_{\mathcal{U}}(\mathcal{U}) \cap (\mathbb{R}^n \times \{0\})$. This implies that a submanifold is in particular a manifold, as it inherits a differentiable structure from $\mathcal{M}$. The natural inclusion map $\iota : \mathcal{N} \hookrightarrow \mathcal{M}$ mapping each point on itself is an injective immersion. The *codimension* of the submanifold $\mathcal{N}$ in $\mathcal{M}$ is defined as $\operatorname{codim}\mathcal{N} = \dim\mathcal{M} - \dim\mathcal{N}$.

If $\mathcal{N}_1$ and $\mathcal{N}_2$ are two submanifolds of $\mathcal{M}$, we say that they are *transversal*, if for all $p \in \mathcal{N}_1 \cap \mathcal{N}_2$ the tangent spaces satisfy $\mathsf{T}_p(\mathcal{N}_1 \cap \mathcal{N}_2) = \mathsf{T}_p\mathcal{N}_1 \cap \mathsf{T}_p\mathcal{N}_2$. This condition can be interpreted as follows. Let γ_i be curves on $\mathcal{N}_i$ with $\gamma_1(0) = p = \gamma_2(0)$. The two curves can be tangent at p, if and only if there exists an interval $\mathcal{I} \subset \mathbb{R}$ such that $\gamma_i(\mathcal{I}) \subseteq \mathcal{N}_1 \cap \mathcal{N}_2$.

Definition C.1.4. *The dual space of the tangent space* $\mathsf{T}_p\mathcal{M}$*, denoted by* $\mathsf{T}_p^*\mathcal{M}$*, is called the* cotangent space *of the manifold* $\mathcal{M}$ *at the point* $p \in \mathcal{M}$*. Thus its elements are linear maps (functionals)* $\lambda_p : \mathsf{T}_p\mathcal{M} \to \mathbb{R}$*. The* cotangent bundle $\mathsf{T}^*\mathcal{M}$ *is the union of all cotangent spaces*

$$\mathsf{T}^*\mathcal{M} = \bigcup_{p \in \mathcal{M}} \{p\} \times \mathsf{T}_p^*\mathcal{M} . \tag{C.4}$$

The cotangent bundle projection *is the map* $\tau_{\mathcal{M}}^* : \mathsf{T}\mathcal{M} \to \mathcal{M}$ *defined by* $\tau_{\mathcal{M}}^*(\{p\} \times \lambda_p) = p$ *for any* $\lambda_p \in \mathsf{T}_p^*\mathcal{M}$*.*

Above we have mentioned that the vectors $\partial_{x^i}\big|_p$ form a basis of $T_p\mathcal{M}$. The elements of the dual basis in $T_p^*\mathcal{M}$ are denoted by $dx^i\big|_p$, i.e. we have the relations $dx^i\big|_p(\partial_{x^j}\big|_p) = \delta^i_j$.

C.2 Fibred Manifolds and Bundles

Fibred manifolds can be seen as a generalisation of the Cartesian product of two manifolds. As we will see, in the neighbourhood of a point a fibred manifold is in fact always of this form.

Definition C.2.1. *A* fibred manifold *is a triple* $(\mathcal{E}, \pi, \mathcal{B})$ *where* $\mathcal{E}$ *and* $\mathcal{B}$ *are manifolds and* $\pi : \mathcal{E} \to \mathcal{B}$ *is a surjective submersion.* $\mathcal{E}$ *is called the* total space *of the fibred manifold,* $\mathcal{B}$ *the* base space *and* π *the* projection*. The subset* $\pi^{-1}(b) \subset \mathcal{E}$ *is the* fibre *over the point* $b \in \mathcal{B}$ *and written* $\mathcal{E}_b$*.*

We usually speak simply of the fibred manifold $\mathcal{E}$, if the base space $\mathcal{B}$ and the projection π are obvious from the context. If $\dim \mathcal{E} = n + m$ and $\dim \mathcal{B} = n$, then m is called the *fibre dimension* of the fibred manifold.

Example C.2.2. The simplest example of a fibred manifold is a *trivial fibred manifold*. Its total space is of the form $\mathcal{E} = \mathcal{B} \times \mathcal{F}$ for some manifold $\mathcal{F}$ and the map π is the projection pr_1 on the first factor. In this case, each fibre can be identified with the manifold $\mathcal{F}$. ◁

Locally, any fibred manifold looks like a trivial fibred manifold. With the help of the Implicit Function Theorem one can show that, because of the properties of a submersion, to each point $\xi \in \mathcal{E}$ a neighbourhood $\mathcal{U}_\xi \subset \mathcal{E}$, a manifold $\mathcal{F}_\xi$ and a diffeomorphism $\phi_\xi : \mathcal{U}_\xi \to \pi(\mathcal{U}_\xi) \times \mathcal{F}_\xi$ exist such that the diagram

$$\begin{tikzcd} \mathcal{U}_\xi \arrow[r, "\phi_\xi"] \arrow[d, "\pi|_{\mathcal{U}_\xi}"'] & \pi(\mathcal{U}_\xi) \times \mathcal{F}_\xi \arrow[d, "pr_1"] \\ \pi(\mathcal{U}_\xi) \arrow[r, "id"'] & \pi(\mathcal{U}_\xi) \end{tikzcd} \tag{C.5}$$

commutes. This means that $pr_1(\phi_\xi(\zeta)) = \pi(\zeta)$ for any point $\zeta \in \mathcal{U}_\xi$. Thus any fibred manifold can be covered with subsets $\mathcal{U}_\xi$ which are diffeomorphic to products $\pi(\mathcal{U}_\xi) \times \mathcal{F}_\xi$. It is a natural thought to exploit this for the introduction of special coordinate charts on $\mathcal{E}$, so-called *adapted coordinates*. Namely, we first map with ϕ_ξ on $\pi(\mathcal{U}_\xi) \times \mathcal{F}_\xi$ and then use charts on $\pi(\mathcal{U}_\xi)$ and $\mathcal{F}_\xi$, respectively. The arising coordinate maps are of the form $\mathcal{U}_\xi \to \mathbb{R}^n \times \mathbb{R}^m$

and if two points $\xi_1, \xi_2 \in \mathcal{U}_\xi$ lie in the same fibre, i.e. $\pi(\xi_1) = \pi(\xi_2)$, then their first n coordinates are identical. We mostly denote adapted coordinates by $(\mathbf{x}, \mathbf{u}) = (x^1, \dots, x^n, u^1, \dots, u^m)$.

A submanifold $\mathcal{F} \subseteq \mathcal{E}$ of the total space is a *fibred submanifold*, if the restriction $\pi|_{\mathcal{F}} : \mathcal{F} \to \mathcal{B}$ is still a surjective submersion. In other words, $(\mathcal{F}, \pi|_{\mathcal{F}}, \mathcal{B})$ must again be a fibred manifold.

Fibre bundles are a special kind of fibred manifolds that look even stronger like a trivial bundle. Here we require that we may choose the subsets $\mathcal{U}_\xi$ in the form $\pi^{-1}(\mathcal{W}_b)$ where $\mathcal{W}_b$ is a neighbourhood in $\mathcal{B}$ of the point $b = \pi(\xi)$, i.e. U_ξ consists of complete fibres. A covering of $\mathcal{E}$ with such subsets is a *local trivialisation* of the bundle. One can show that to every local trivialisation a manifold $\mathcal{F}$ exists such that all the manifolds $\mathcal{F}_\xi$ are diffeomorphic to it. $\mathcal{F}$ is called the *typical fibre* of the bundle.

Special kinds of bundles arise, if the typical fibre $\mathcal{F}$ has additional structures. For us important are *vector bundles* where $\mathcal{F}$ is a vector space $\mathcal{V}$ and *affine bundles* where $\mathcal{F}$ is an affine space $\mathcal{A}$. Recall that an affine space $\mathcal{A}$ is modelled on a vector space $\mathcal{V}$: $\mathcal{A}$ is a set on which $\mathcal{V}$ acts free and transitively as additive group; in other words, we identify the "difference" $\alpha - \beta$ of any two points $\alpha, \beta \in \mathcal{A}$ with a vector $v \in \mathcal{V}$. As we can do this over every point of the base space, a vector bundle is associated with every affine bundle.

Example C.2.3. Simple examples of bundles are both tangent and cotangent bundle of a manifold $\mathcal{M}$. Their typical fibre is the vector space $\mathbb{R}^m$ where $m = \dim \mathcal{M}$; thus they are even vector bundles. The coordinates $(\mathbf{x}, \dot{\mathbf{x}})$ on the tangent bundle $T\mathcal{M}$ are adapted. ◁

This special form of the sets $\mathcal{U}_\xi$ seems to be a rather weak condition. In fact, it has a number of consequences. For example, the differentiable structure of $\mathcal{E}$ is completely determined by those of $\mathcal{B}$ and $\mathcal{F}$. Bundles also carry an additional structure, namely a group acting on the fibres. As we do not need this, we stop here and refer to the literature for more details.

Definition C.2.4. *A* (local) section σ *of the fibred manifold* $\pi : \mathcal{E} \to \mathcal{B}$ *is a map* $\sigma : \mathcal{U} \subseteq \mathcal{B} \to \mathcal{E}$ *such that* $\pi \circ \sigma = \mathrm{id}_{\mathcal{U}}$. *The space of all local sections is denoted by* $\Gamma_{\mathrm{loc}}(\mathcal{B}, \mathcal{E})$ *or, if the base space is obvious from the context, shortly by* $\Gamma_{\mathrm{loc}}(\mathcal{E})$.

In adapted coordinates $(\mathbf{x}, \mathbf{u})$ on the fibred manifold $\mathcal{E}$, a section σ has the local form $\sigma(\mathbf{x}) = (\mathbf{x}, s(\mathbf{x}))$, as the condition $\pi \circ \sigma = \mathrm{id}_{\mathcal{U}}$ implies that every point $b \in \mathcal{U}$ is mapped to a point in the fibre $\mathcal{E}_b$ above it. The notion of a section can be understood as a generalisation of the graph of a function. Indeed, if we are dealing with a trivial fibred manifold $\mathcal{E} = \mathcal{B} \times \mathcal{F}$, every function $s : \mathcal{B} \to \mathcal{F}$ defines a section $\sigma(b) = (b, s(b))$ and the set $\sigma(\mathcal{B}) \subset \mathcal{B}\times\mathcal{F}$ is the graph of s.

Let $(\mathcal{E}, \pi, \mathcal{B})$ and $(\mathcal{F}, \rho, \mathcal{B})$ be two fibred manifolds over the same base space. Their *fibred product* is the fibred manifold $(\mathcal{E} \underset{\mathcal{B}}{\times} \mathcal{F}, \pi \underset{\mathcal{B}}{\times} \rho, \mathcal{B})$ where the total space $\mathcal{E} \underset{\mathcal{B}}{\times} \mathcal{F}$ consists of all pairs $(\xi, \zeta) \in \mathcal{E} \times \mathcal{F}$ with $\pi(\xi) = \rho(\zeta)$ and the projection is defined by $(\pi \underset{\mathcal{B}}{\times} \rho)(\xi, \zeta) = \pi(\xi) = \rho(\zeta)$. It is straightforward to show that this indeed yields again a fibred manifold.

The total space $\mathcal{E} \underset{\mathcal{B}}{\times} \mathcal{F}$ may also be considered as a fibred manifold over either $\mathcal{E}$ or $\mathcal{F}$, respectively, with obvious projections on the first or second factor, respectively. In this case one speaks of a *pull-back* and uses the notation $\rho^*(\pi) : \rho^*(\mathcal{E}) \to \mathcal{F}$ and $\pi^*(\rho) : \pi^*(\mathcal{F}) \to \mathcal{E}$, respectively.

Let $(\mathcal{E}, \pi, \mathcal{B})$ and $(\mathcal{F}, \rho, \mathcal{C})$ be two arbitrary fibred manifolds. A smooth map $F : \mathcal{E} \to \mathcal{F}$ is a *fibred morphism*, if there exists a map $f : \mathcal{B} \to \mathcal{C}$ (which is in fact uniquely determined and smooth) such that $\rho \circ F = f \circ \pi$. It maps fibres into fibres: if $\xi_1, \xi_2 \in \mathcal{E}_b$, then $F(\xi_1), F(\xi_2) \in \mathcal{F}_{f(b)}$. An example of a fibred morphism is the tangent map $Tf : T\mathcal{M} \to T\mathcal{N}$ of a map $f : \mathcal{M} \to \mathcal{N}$.

The projection π introduces in a natural way the notion of a vertical vector. Vertical vectors are tangential to the fibres. As π is a map between the two manifolds $\mathcal{E}$ and $\mathcal{B}$, it induces a linear map $T\pi$ between the tangent spaces $T\mathcal{E}$ and $T\mathcal{B}$ locally given by (C.3).

Definition C.2.5. *The* vertical bundle *of the fibred manifold* $\pi : \mathcal{E} \to \mathcal{B}$ *is the subbundle* $V\mathcal{E} = \ker T\pi \subset T\mathcal{E}$.

In adapted coordinates $(\mathbf{x}, \mathbf{u})$ on the manifold $\mathcal{E}$, vertical vectors at a point $\xi \in \mathcal{E}$ are of the form $v = v^\alpha \partial_{u^\alpha}\big|_\xi$, i.e. all their components in the directions $\partial_{x^i}\big|_\xi$ corresponding to the coordinates in the base space $\mathcal{B}$ vanish.

Example C.2.6. In the case of the trivial fibred manifold $\mathcal{E} = \mathcal{B} \times \mathcal{F}$ we have the decomposition $T\mathcal{E} = T\mathcal{B} \times T\mathcal{F}$ and we can identify $V\mathcal{E} = \mathcal{E} \times T\mathcal{F}$. $\lhd$

If vertical vectors exists, one expects also horizontal vectors. However, this requires an additional structure on the fibred manifold $\mathcal{E}$, namely a *connection*, as it is not possible to give an intrinsic definition of a horizontal vector for arbitrary fibred manifolds. The simple idea that in adapted coordinates such vectors should have the form $v = v^i \partial_{x^i}\big|_\xi$ does not work, as this form does not remain invariant under changes to other adapted coordinates.

One way to define a connection is to choose (in a smooth manner) at each point $\xi \in \mathcal{E}$ a subspace $H_\xi\mathcal{E} \subset T_\xi\mathcal{E}$, a *horizontal space*, such that $T_\xi\mathcal{E} = H_\xi\mathcal{E} \oplus V_\xi\mathcal{E}$; i.e. each vector $v_\xi \in T_\xi\mathcal{E}$ can be decomposed in a sum $v_\xi = v_\xi^H + v_\xi^V$ with $v_\xi^H \in H_\xi\mathcal{E}$ and $v_\xi^V \in V_\xi\mathcal{E}$. This induces two natural maps: the *horizontal projector* $P^H : T\mathcal{E} \to H\mathcal{E}$ with $P^H(\xi, v_\xi) = (\xi, v_\xi^H)$ and the *vertical projector* $P^V : T\mathcal{E} \to V\mathcal{E}$ with $P^V(\xi, v_\xi) = (\xi, v_\xi^V)$. Conversely, each of these projectors uniquely defines the connection.

Alternatively, one may define a connection with a *horizontal lift*. This is a map $A : \mathcal{E} \times_{\mathcal{B}} T\mathcal{B} \to T\mathcal{E}$ which is linear in the second factor and a right inverse to the map $\rho = \tau_{\mathcal{E}} \times T\pi : T\mathcal{E} \to \mathcal{E} \times_{\mathcal{B}} T\mathcal{B}$, i.e. $\rho \circ A = \mathrm{id}_{\mathcal{E}\times_{\mathcal{B}} T\mathcal{B}}$. In adapted coordinates ρ maps a point $\big((\mathbf{x},\mathbf{u});(\dot{\mathbf{x}},\dot{\mathbf{u}})\big) \in T\mathcal{E}$ into the point $\big((\mathbf{x},\mathbf{u});(\mathbf{x},\dot{\mathbf{x}})\big) \in \mathcal{E} \times_{\mathcal{B}} T\mathcal{B}$. This implies that A must be locally of the form $\big((\mathbf{x},\mathbf{u});(\mathbf{x},\dot{\mathbf{x}})\big) \mapsto \big((\mathbf{x},\mathbf{u});(\dot{\mathbf{x}},\bar{A}(\mathbf{x},\mathbf{u})\dot{\mathbf{x}})\big)$ where $\bar{A}(\mathbf{x},\mathbf{u})$ is an $m \times n$ matrix. The horizontal spaces of the connection defined by the map A are given by $H_\xi\mathcal{E} = A(\xi, T_{\pi(\xi)}\mathcal{B})$. This explains the name horizontal lift: A takes a vector in $T\mathcal{B}$ and lifts it to a horizontal vector in $T\mathcal{E}$. Each connection uniquely determines such a horizontal lift and conversely each horizontal lift defines uniquely a connection.

C.3 Vector Fields and Differential Forms

In this section we want to discuss in some more detail sections of the tangent and cotangent bundle of a manifold. They are fundamental objects in differential geometry.

Definition C.3.1. *Sections of the bundle* $T\mathcal{M}$ *are called* vector fields*. We write* $\mathfrak{X}(\mathcal{M})$ *for* $\Gamma_{\mathrm{loc}}(T\mathcal{M})$*.*

$\mathfrak{X}(\mathcal{M})$ possesses some algebraic structures (see Appendix B.1). In particular, it is an $\mathcal{F}(\mathcal{M})$-module: we can add vector fields and multiply them by functions in $\mathcal{F}(\mathcal{M})$. In local coordinates $(x^1, \dots, x^m)$, the module $\mathfrak{X}(\mathcal{M})$ is generated by the fields $\partial_{x^i} : p \mapsto \partial_{x^i}\big|_p$. Thus every vector field can be written locally as $X = v^i\partial_{x^i}$ with some functions $v^i \in \mathcal{F}(\mathcal{M})$.

The notation ∂_{x^i} already indicates that vector fields can also be interpreted as linear first-order differential operators acting on $\mathcal{F}(\mathcal{M})$. This is a more algebraic alternative to our geometric introduction of tangent vectors in the last section: a vector field is an $\mathbb{R}$-linear map $X : \mathcal{F}(\mathcal{M}) \to \mathcal{F}(\mathcal{M})$ that satisfies the Leibniz rule $X(fg) = X(f)g + fX(g)$ (such maps are also called *derivations*). The relation between the two approaches is provided by the tangent map. If $f \in \mathcal{F}(\mathcal{M})$, then $Tf : T\mathcal{M} \to T\mathbb{R} = \mathbb{R} \oplus \mathbb{R}$ and for any vector field $X \in \mathfrak{X}(\mathcal{M})$ we define the function $X(f) \in \mathcal{F}(\mathcal{M})$ by $X(f)(p) = T_pf(X_p)$ for any point $p \in \mathcal{M}$. In local coordinates, the vector field ∂_{x^i} has then the meaning one expects: $\partial_{x^i}(f) = \partial f/\partial x^i$.

With this interpretation, $\mathfrak{X}(\mathcal{M})$ can be easily given the structure of a real Lie algebra. As multiplication we take the *Lie bracket*: $[X, Y] = X \circ Y - Y \circ X$

for any two fields $X, Y \in \mathfrak{X}(\mathcal{M})$. It is easy to verify that this yields again a derivation on $\mathcal{F}(\mathcal{M})$, i.e. a vector field in $\mathfrak{X}(\mathcal{M})$. If the two vector fields are given in local coordinates by $X = v^i \partial_{x^i}$ and $Y = w^i \partial_{x^i}$, then their Lie bracket is the vector field

$$[X, Y] = \left(v^j \frac{\partial w^i}{\partial x^j} - w^j \frac{\partial v^i}{\partial x^j} \right) \partial_{x^i} . \tag{C.6}$$

Because of these two linear structures, module over $\mathcal{F}(\mathcal{M})$ and Lie algebra over $\mathbb{R}$, one must be careful when speaking about linear combinations of vector fields. In order to avoid misunderstandings, one should always state where the coefficients live, if it is not evident from the context.

A d-dimensional *distribution* $\mathcal{D}$ on a manifold $\mathcal{M}$ assigns in a smooth manner to each point $p \in \mathcal{M}$ a vector subspace $\mathcal{D}_p \subseteq T_p\mathcal{M}$ with $\dim \mathcal{D}_p = d$. Typically, it is given as the span of some linearly independent vector fields $X^1, \dots, X^d$, i.e. $\mathcal{D}_p = \langle X^1_p, \dots, X^d_p \rangle$. A distribution is *involutive*, if it is closed under the Lie bracket. This means, if X and Y are two vector fields in $\mathcal{D}$ (thus $X_p, Y_p \in \mathcal{D}_p$ for all $p \in \mathcal{M}$), then $[X, Y]$ also lives in $\mathcal{D}$. The distribution is called *integrable*, if every point $p \in \mathcal{M}$ lies on a *maximal integral manifold* of $\mathcal{D}$, i.e. on a submanifold $\mathcal{N} \subseteq \mathcal{M}$ of dimension d such that $\mathcal{D}\big|_{\mathcal{N}} = T\mathcal{N}$. The famous *Frobenius Theorem* states that a distribution is integrable, if and only if it is involutive.

A simple example of a distribution is the vertical bundle of a fibred manifold. It is not difficult to see that it is even an involutive distribution; the fibres are integral manifolds of maximal dimension. If there is a connection defined on the fibred manifold, its horizontal spaces also form a distribution. In contrast to the vertical bundle, it is in general not involutive. Connections with an involutive horizontal distribution are called *flat*.

Definition C.3.2. *Sections of the bundle* $T^*\mathcal{M}$ *are called* differential (one-) forms. *We write* $\Omega^1(\mathcal{M})$ *for* $\Gamma_{\mathrm{loc}}(T^*\mathcal{M})$.

Just like $\mathfrak{X}(\mathcal{M})$, the bundle $\Omega^1(\mathcal{M})$ is an $\mathcal{F}(\mathcal{M})$-module. In a chart $\mathcal{U}$ with local coordinates $x^1, \dots, x^n$, a module basis is given by the one-forms $dx^i : p \mapsto dx^i\big|_p$. Obviously, this is the dual basis to $\{\partial_{x^i}\}$.

A d-dimensional *codistribution* $\mathcal{D}^*$ on a manifold $\mathcal{M}$ maps each point $p \in \mathcal{M}$ smoothly to a vector subspace $\mathcal{D}^*_p \subseteq T^*_p\mathcal{M}$ with $\dim \mathcal{D}^*_p = d$. This notion is dual to distributions. If $\mathcal{D} \subseteq T\mathcal{M}$ is a distribution, its *annihilator* $\mathcal{D}^0 = \{\omega_p \in T^*_p\mathcal{M} \mid p \in \mathcal{M},\ \omega_p(\mathcal{D}_p) = 0\}$ is a codistribution (of dimension $\dim \mathcal{M} - \dim \mathcal{D} = \operatorname{codim} \mathcal{D}$). If a codistribution is spanned by some one-forms $\omega^1, \dots, \omega^d$, these are classically called a *Pfaffian system*.

A codistribution $\mathcal{D}^*$ is called *involutive*, if the exterior derivative of any form $\omega \in \mathcal{D}^*$ vanishes modulo $\mathcal{D}^*$, i.e. we can write $d\omega = \sum_{i=1}^{r} \eta_i \wedge \omega_i$

with $\omega_i \in \mathcal{D}^*$. With the help of (C.15) one easily checks that this condition is dual to the condition for an involutive distribution: $\mathcal{D}$ is involutive, if and only if $\mathcal{D}^0$ is involutive. An *integral manifold* of a codistribution $\mathcal{D}^*$ is a submanifold $\mathcal{N} \subseteq \mathcal{M}$ such that the pull-back $\iota^*(\mathcal{D}^*)$ with respect to the inclusion map $\iota : \mathcal{N} \hookrightarrow \mathcal{M}$ vanishes: $\iota^*\omega = 0$ for all $\omega \in \mathcal{D}^*$. The dual version of the *Frobenius Theorem* asserts that a codistribution is involutive, if and only if it is *integrable*, i. e. $\mathcal{D}^*$ possesses at any point $p \in \mathcal{M}$ integral manifolds of dimension d.

If there exist one-forms, one expects that there are also two-forms etc. Indeed, like over any finite-dimensional vector space, we can build at each point $p \in \mathcal{M}$ the *exterior algebra* $\Lambda(T_p^*\mathcal{M})$. It is a graded algebra (see Appendix B.1) with $\Lambda^0(T_p^*\mathcal{M}) = \mathbb{R}$ and $\Lambda^1(T_p^*\mathcal{M}) = T_p^*\mathcal{M}$. The forms of higher degree can be obtained with the help of the *wedge product*: $\wedge : \Lambda^k(T_p^*\mathcal{M}) \times \Lambda^\ell(T_p^*\mathcal{M}) \to \Lambda^{k+\ell}(T_p^*\mathcal{M})$. It is associative and distributive but non-commutative. Instead we have for $\omega \in \Lambda^k(T_p^*\mathcal{M})$ and $\eta \in \Lambda^\ell(T_p^*\mathcal{M})$ that $\omega \wedge \eta = (-1)^{k\ell}\eta \wedge \omega$. The precise definition is obtained by anti-symmetrisation of the tensor product and requires the choice of a combinatorial factor. We will use the following convention

$$(\omega \wedge \eta)(v_1, \dots, v_{k+\ell}) = \\ \frac{1}{k!\ell!} \sum_{\pi \in S_{k+\ell}} \epsilon(\pi)\omega(v_{\pi(1)}, \dots, v_{\pi(k)})\eta(v_{\pi(k+1)}, \dots, v_{\pi(k+\ell)}) \,. \quad \text{(C.7)}$$

Here $S_{k+\ell}$ is the symmetric group containing all permutations of the numbers $1, \dots, k+\ell$ and $\epsilon(\pi) = \pm 1$ depending on whether we are dealing with an even or an odd permutation. Some authors use a factor $1/(k+\ell)!$, but we will stick to the convention above.

We can also form vector bundles with the spaces $\Lambda^k(T_p^*\mathcal{M})$. Their local sections are contained in $\Omega^k(\mathcal{M})$ and are called *differential* k-*forms*.Note that zero-forms are just (local) functions $\mathcal{M} \to \mathbb{R}$. The direct sum of all these vector bundles yields the bundle $\Omega(\mathcal{M})$. In some local coordinates on the manifold $\mathcal{M}$, a basis of $\Omega^k(\mathcal{M})$ is given by the forms $dx^{i_1} \wedge \cdots \wedge dx^{i_k}$ with $i_1 < i_2 < \cdots < i_k$. However, typically on writes a k-form locally as $\omega = \omega_{i_1 \dots i_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k}$ where the smooth functions $\omega_{i_1 \dots i_k}$ are totally antisymmetric in their indices.

If $f : \mathcal{N} \to \mathcal{M}$ is a smooth map, we can *pull-back* forms. If $\omega \in \Omega^k(\mathcal{M})$, we define via the tangent map Tf a form $f^*\omega \in \Omega^k(\mathcal{N})$ by setting

$$(f^*\omega)_p(v_1, \dots, v_k) = \omega_{f(p)}\big(T_pf(v_1), \dots, T_pf(v_k)\big) \,. \quad \text{(C.8)}$$

Note that this can in particular be applied to the inclusion map ι, if $\mathcal{N}$ is a submanifold of $\mathcal{M}$. In local coordinates we get the following picture. Let

$x^1, \dots, x^n$ be coordinates in some chart $\mathcal{U} \subseteq \mathcal{N}$ and $y^1, \dots, y^m$ coordinates in a chart $\mathcal{V} \subseteq \mathcal{M}$ where the charts are chosen such that $f(\mathcal{U}) \subseteq \mathcal{V}$ and we can write $y^j = f^j(\mathbf{x})$. If $\omega = \omega_{j_1 \dots j_k} dy^{j_1} \wedge \cdots \wedge dy^{j_k}$ is a k-form defined on $\mathcal{V}$, then its pull-back takes on $\mathcal{U}$ the local form

$$f^* \omega = \frac{\partial f^{j_1}}{\partial x^{i_1}} \cdots \frac{\partial f^{j_k}}{\partial x^{i_k}} \omega_{j_1 \dots j_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k} \,. \tag{C.9}$$

Given an algebra $\mathcal{A}$, a *derivation* on it is a map $\delta : \mathcal{A} \to \mathcal{A}$ that is linear with respect to the vector space structure and that satisfies a (possibly graded) *Leibniz rule* $\delta(fg) = \delta(f)g \pm f\delta(g)$. There exist several important derivations on $\Omega(\mathcal{M})$.

Definition C.3.3. *The* interior derivative *with respect to a given vector field* $X \in \mathfrak{X}(\mathcal{M})$ *is a map* $\iota_X : \Omega^{k+1}(\mathcal{M}) \to \Omega^k(\mathcal{M})$ *defined by*

$$(\iota_X \omega)(X_1, \dots, X_k) = \omega(X, X_1, \dots, X_k) \tag{C.10}$$

for all forms $\omega \in \Omega^{k+1}(\mathcal{M})$ *and vector fields* $X_1, \dots, X_k \in \mathfrak{X}(\mathcal{M})$.

Thus the interior derivative reduces the form degree. Its graded Leibniz rule is easily derived from the properties of the wedge product of forms: $\iota_X(\omega_1 \wedge \omega_2) = (\iota_X \omega_1) \wedge \omega_2 + (-1)^{k_1} \omega_1 \wedge (\iota_X \omega_2)$, if $\omega_i \in \Omega^{k_i}(\mathcal{M})$. In local coordinates, the interior derivative corresponds to what is classically called a "contraction": if $X = v^i \partial_{x^i}$ and $\omega = \omega_{i_1 \cdots i_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k}$, then

$$\iota_X \omega = \omega_{i_1 i_2 \cdots i_k} v^{i_1} dx^{i_2} \wedge \cdots \wedge dx^{i_k} \,. \tag{C.11}$$

The next derivative increases the form degree, but it is less easy to introduce. We take here an algebraic approach and define it via some characteristic properties. One can show that there exists one and only one map satisfying all these properties.

Definition C.3.4. *The* exterior derivative *of differential forms is the unique map* $d : \Omega^k(\mathcal{M}) \to \Omega^{k+1}(\mathcal{M})$ *satisfying the following properties.*

- d *is an* $\mathbb{R}$*-linear map: for all* $a_1, a_2 \in \mathbb{R}$ *and* $\omega_1, \omega_2 \in \Omega^k(\mathcal{M})$

$$d(a_1 \omega_1 + a_2 \omega_2) = a_1 d\omega_1 + a_2 d\omega_2 \,. \tag{C.12}$$

- d *satisfies the Leibniz rule: for all* $\omega_1 \in \Omega^{k_1}(\mathcal{M})$ *and* $\omega_2 \in \Omega^{k_2}(\mathcal{M})$

$$d(\omega_1 \wedge \omega_2) = (d\omega_1) \wedge \omega_2 + (-1)^{k_1} \omega_1 \wedge d\omega_2 \,. \tag{C.13}$$

- d *is nilpotent:* $d \circ d = 0$.

- *If* $f : M \to \mathbb{R}$ *is a zero-form, the one-form* df *is locally given by*

$$df(x) = \frac{\partial f}{\partial x^i}(x)dx^i\ . \tag{C.14}$$

Given a k-form ω, an explicit coordinate-free expression for $d\omega$ is

$$\begin{aligned} d\omega(X_1, \dots, X_{k+1}) = {} & \sum_{i=1}^{k+1} (-1)^{i+1} X_i\big(\omega(X_1, \dots, \widehat{X}_i, \dots, X_{k+1})\big) + \\ & + \sum_{i<j} \omega\big([X_i, X_j], X_1, \dots, \widehat{X}_i, \dots, \widehat{X}_j, \dots, X_{k+1}\big) \end{aligned} \tag{C.15}$$

where the hat indicates that the corresponding vector is omitted. In particular, for a function $f \in \mathcal{F}(\mathcal{M})$ we obtain $df(X) = X(f)$; thus the one-form df is just the fibre component of the tangent map Tf. An important property of the exterior derivative is that it commutes with the pull-back: if $f : \mathcal{M} \to \mathcal{N}$ is a smooth map, then $d \circ f^* = f^* \circ d$. In local coordinates, one obtains for $\omega = \omega_{i_1 \cdots i_k} dx^{i_1} \wedge \cdots \wedge dx^{i_k}$ that

$$d\omega = \frac{\partial \omega_{i_1 \cdots i_k}}{\partial x^i} dx^i \wedge dx^{i_1} \wedge \cdots \wedge dx^{i_k}\ . \tag{C.16}$$

Remark C.3.5. If $f : \mathcal{M} \to \mathbb{R}$ is a smooth function on $\mathcal{M}$ such that zero is a regular value of it, then it is a classical result in differential geometry that its zero set $\mathcal{N} = \{p \in \mathcal{M} \mid f(p) = 0\}$ is a submanifold of $\mathcal{M}$. The tangent spaces $T_p\mathcal{N}$ may be identified with vector subspaces of $T_p\mathcal{M}$. These vector subspaces are easily determined with the help of the one-form $df \in \Omega^1(\mathcal{M})$: $T_p\mathcal{N} = \{v \in T_p\mathcal{M} \mid df_p(v) = 0\}$. As zero is a regular value, by definition df_p does not vanish for any point $p \in \mathcal{N}$ so that this condition is always non-trivial.

This follows directly from our geometric definition of tangent vectors as equivalence class of curves. If $\gamma : \mathcal{I} \to \mathcal{N}$ is a curve with $\gamma(0) = p$, then obviously $f(\gamma(t)) \equiv 0$. Differentiating this relation with respect to t yields $df(\dot\gamma(t)) \equiv 0$. Evaluation at $t = 0$ yields our result.

More generally, if the submanifold $\mathcal{N}$ is the common zero set of the functions $f^1, \dots, f^r$, then any vector in $T_p\mathcal{N}$ satisfies $df^1_p(v) = \cdots = df^r_p(v) = 0$ and vice versa. The vector subspace $N^*_p\mathcal{N} \subset T^*_p\mathcal{N}$ spanned by $df^1_p, \dots, df^r_p$ is sometimes called the *conormal space* of $\mathcal{N}$ at the point p. ◁

Remark C.3.6. We call a k-form $\omega \in \Omega^k(\mathcal{M})$ *exact*, if a $(k-1)$-form $\eta \in \Omega^{k-1}(\mathcal{M})$ exists such that $\omega = d\eta$. A form ω such that $d\omega = 0$ is *closed*. Obviously, any exact form is closed, as $d^2 = 0$. The *Poincaré Lemma* asserts that locally (more precisely, in star-shaped domains), the converse also holds: every closed form is exact. Globally, however, this is in general not true. Those closed forms that are not exact generate the *de Rham cohomology* of the manifold $\mathcal{M}$, a topological invariant. ◁

The last derivative is again with respect to a vector field and can be defined for arbitrary tensor fields, although we will introduce it only for differential forms. The general definition uses a limit process taking place along the flow lines of X. Our definition is then a theorem and (C.17) usually referred to as *Cartan's magic formula.*

Definition C.3.7. *The* Lie derivative *with respect to a vector field* $X \in \mathfrak{X}(\mathcal{M})$ *is the map* $\mathcal{L}_X : \Omega^k(\mathcal{M}) \to \Omega^k(\mathcal{M})$ *defined by*

$$\mathcal{L}_X = \iota_X \circ d + d \circ \iota_X \,. \tag{C.17}$$

For zero-forms, i.e. functions, this yields $\mathcal{L}_X f = Xf$. For a k-form ω we obtain the following coordinate-free expression

$$\begin{aligned}\mathcal{L}_X\omega(X_1,\dots,X_k) = X\bigl(\omega(X_1,\dots,X_k)\bigr) &+ \\ + \sum_{i=1}^{k} \omega(X_1,\dots,[X_i,X],\dots,X_k) \,.&\end{aligned} \tag{C.18}$$

Locally, we find for $X = v^i\partial_{x^i}$ and $\omega = \omega_{i_1\cdots i_k}dx^{i_1}\wedge\cdots\wedge dx^{i_k}$ that

$$\begin{aligned}\mathcal{L}_X\omega = \biggl(&v^i\frac{\partial\omega_{i_1\cdots i_k}}{\partial x^i} + \\ &\frac{\partial v^i}{\partial x^{i_1}}\omega_{i i_2\cdots i_k} + \cdots + \frac{\partial v^i}{\partial x^{i_k}}\omega_{i_1\cdots i_{k-1} i}\biggr)dx^{i_1}\wedge\cdots\wedge dx^{i_k} \,.\end{aligned} \tag{C.19}$$

The Lie derivative commutes with the exterior derivative, $d\circ\mathcal{L}_X = \mathcal{L}_X\circ d$, and satisfies for arbitrary differential forms $\omega_1, \omega_2 \in \Omega(\mathcal{M})$ the *ungraded* Leibniz rule $\mathcal{L}_X(\omega_1\wedge\omega_2) = \mathcal{L}_X(\omega_1)\wedge\omega_2 + \omega_1\wedge\mathcal{L}_X(\omega_2)$.

References

[1] M.J. Ablowitz and H. Segur. *Solitons and the Inverse Scattering Transform*. SIAM Studies in Applied Mathematics. SIAM, Philadelphia, 1981.

[2] R. Abraham, J.E. Marsden, and T. Ratiu. *Manifolds, Tensor Analysis, and Applications*. Applied Mathematical Sciences 75. Springer-Verlag, New York, 1988.

[3] M. Abramowitz and I. Stegun, editors. *Handbook of Mathematical Functions*. Dover, New York, 1965.

[4] W.W. Adams and P. Loustaunau. *An Introduction to Gröbner Bases*. Graduate Studies in Mathematics 3. AMS, Providence, 1994.

[5] D.V. Alekseevskij, A.M. Vinogradov, and V.V. Lychagin. Basic ideas and concepts of differential geometry. In R.V. Gamkrelidze, editor, *Geometry I*, Encyclopaedia of Mathematical Sciences 28, pages 1–264. Springer-Verlag, Berlin, 1991.

[6] A.M. Ampère. Considérations Générales sur les Intégrales des Équations aux Différentielles Partielles. *J. École Polytéchnique*, 10:549–611, 1815.

[7] J. Apel. The computation of Gröbner bases using an alternative algorithm. In M. Bronstein, J. Grabmeier, and V. Weispfenning, editors, *Symbolic Rewriting Techniques*, Progress in Computer Science and Applied Logic 15, pages 35–45. Birkhäuser, Basel, 1998.

[8] J. Apel. Theory of involutive divisions and an application to Hilbert function computations. *J. Symb. Comput.*, 25:683–704, 1998.

[9] V.I. Arnold. *Mathematical Methods of Classical Mechanics*. Graduate Texts in Mathematics 60. Springer-Verlag, New York, 1978.

[10] U.M. Ascher, H. Chin, and S. Reich. Stabilization of DAEs and invariant manifolds. *Num. Math.*, 67:131–149, 1994.

[11] A.V. Bäcklund. Zur Theorie der partiellen Differentialgleichungen erster Ordnung. *Math. Ann.*, 17:285–328, 1880.

[12] K. Baclawski and A.M. Garsia. Combinatorial decompositions of rings. *Adv. Math.*, 39:155–184, 1981.

[13] D. Bayer and D. Mumford. What can be computed in algebraic geometry? In D. Eisenbud and L. Robbiano, editors, *Computational Algebraic Geometry and Commutative Algebra*, Symposia Mathematica 34, pages 1–48. Cambridge University Press, Cambridge, 1993.

[14] D. Bayer and M. Stillman. A criterion for detecting $\mathfrak{m}$-regularity. *Invent. Math.*, 87:1–11, 1987.

[15] Th. Becker and V. Weispfenning. *Gröbner Bases.* Springer-Verlag, New York, 1993.

[16] J. Belanger, M. Hausdorf, and W.M. Seiler. A MuPAD library for differential equations. In V.G. Ghanza, E.W. Mayr, and E.V. Vorozhtsov, editors, *Computer Algebra in Scientific Computing — CASC 2001*, pages 25–42. Springer-Verlag, Berlin, 2001.

[17] I.N. Bernstein and B.I. Rosenfeld. Homogeneous spaces of infinite Lie algebras and characteristic classes of foliations. *Uspekhi Mat. Nauk*, 28:103–138, 1973.

[18] L.J. Billera, R. Cushman, and J.A. Sanders. The Stanley decomposition of the harmonic oscillator. *Indagat. Math.*, 50:375–394, 1988.

[19] J.E. Björk. *Analytical* $\mathcal{D}$*-Modules and Applications.* Mathematics and Its Applications 247. Kluwer, Dordrecht, 1993.

[20] G.W. Bluman and S. Kumei. *Symmetries and Differential Equations.* Applied Mathematical Sciences 81. Springer-Verlag, New York, 1989.

[21] W.M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Pure and Applied Mathematics 120. Academic Press, Orlando, 1986.

[22] A. Borel, editor. *Algebraic* $\mathcal{D}$*-Modules.* Perspectives in Mathematics 2. Academic Press, Boston, 1987.

[23] F. Boulier, D. Lazard, F. Ollivier, and M. Petitot. Representation for the radical of a finitely generated differential ideal. In A.H.M. Levelt, editor, *Proc. ISSAC '95*, pages 158–166. ACM Press, New York, 1995.

[24] F. Brandt. Bäcklund transformations and zero-curvature representations of systems of partial differential equations. *J. Math. Phys.*, 35: 2463–2484, 1994.

[25] K.E. Brenan, S.L. Campbell, and L.R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations.* Classics in Applied Mathematics 14. SIAM, Philadelphia, 1996.

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

[27] R.L. Bryant, S.S. Chern, R.B. Gardner, H.L. Goldschmidt, and P.A. Griffiths. *Exterior Differential Systems.* Mathematical Sciences Research Institute Publications 18. Springer-Verlag, New York, 1991.

[28] B. Buchberger. *Ein Algorithmus zum Auffinden der Basiselemente des Restklassenringes nach einem nulldimensionalen Polynomideal.* PhD thesis, Universität Innsbruck, 1965.

[29] R. Burman. On the strengths of field equations. *Czech. J. Phys. B*, 27:113–116, 1977.

[30] S.L. Campbell and C.W. Gear. The index of general nonlinear DAEs. *Numer. Math.*, 72:173–196, 1995.

[31] S.L. Campbell and W. Marszalek. The index of an infinite dimensional implicit system. *Math. Model. Syst.*, 1:1–25, 1996.

[32] C. Carathéodory. *Variationsrechnung und Partielle Differentialgleichungen erster Ordnung, Band I.* Teubner, Leipzig, 1956.

[33] E. Cartan. *Les Systèmes Différentielles Extérieurs et leurs Applications Géométriques.* Hermann, Paris, 1945.

[34] E. Cartan and A. Einstein. *Lettres sur la Parallélisme Absolue 1929–1932,* edited by R. Debever. Palais des Académies, Bruxelles, 1979.

[35] A. Cauchy. Mémoire sur l'application du calcul des limites à l'intégration d'un système d' équations aux dérivées partielles. In *Œuvres Complètes d'Augustin Cauchy, Série I, Tome VII*, pages 33–49. Gauthier-Villars, 1892. (Original: Compt. Rend. Acad. Sci., t. XV (1842), p. 85).

[36] A. Cauchy. Mémoire sur l'emploi du calcul des limites dans l'intégration des équations aux dérivées partielles. In *Œuvres Complètes d'Augustin Cauchy, Série I, Tome VII*, pages 17–33. Gauthier-Villars, 1892. (Original: Compt. Rend. Acad. Sci., t. XV (1842), p. 44).

[37] A. Cauchy. Mémoire sur les systèmes d'équations aux dérivées partielles d'ordres quelconques, et sur leur réduction à des systèmes d'équations linéaires du premier ordre. In *Œuvres Complètes d'Augustin Cauchy, Série I, Tome VII*, pages 52–58. Gauthier-Villars, 1892. (Original: Compt. Rend. Acad. Sci., t. XV (1842), p. 131).

[38] J. Chrastina. *The Formal Theory of Differential Equations.* Folia Mathematica 6. Masaryk University, Brno, 1998.

[39] S.C. Coutinho. *A Primer of Algebraic* D*-Modules.* London Mathematical Society Student Texts 33. Cambridge University Press, 1995.

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

[41] D. Cox, J. Little, and D. O'Shea. *Using Algebraic Geometry.* Graduate Texts in Mathematics 185. Springer-Verlag, New York, 1998.

[42] M. Dubois-Violette. The theory of overdetermined linear systems and its applications to non-linear field equations. *J. Geom. Phys.*, 1:139–172, 1984.

[43] C. Ehresmann. Introduction à la théorie des structures infinitésimals es des pseudo-groupes de Lie. In *Géométrie Différentielle*, Colloq. Intern. Centre Nat. Rech. Sci. 52, pages 97–110. CNRS, Paris, 193.

[44] C. Ehresmann. Les prolongements d'une variété différentiable. *Comp. Rend. Acad. Sci.*, 233:598–600, 777–779, 1081–1083, 1951.

[45] C. Ehresmann. Les prolongements d'une variété différentiable. *Comp. Rend. Acad. Sci.*, 234:1028–1030, 1424–1425, 1951.

[46] A. Einstein. *The Meaning of Relativity.* Princeton University Press, Princeton, 5th edition, 1956.

[47] D. Eisenbud. *Commutative Algebra with a View Toward Algebraic Geometry.* Graduate Texts in Mathematics 150. Springer-Verlag, New York, 1995.

[48] D. Eisenbud and S. Goto. Linear free resolutions and minimal multiplicity. *J. Alg.*, 88:89–133, 1984.

[49] S. Eliahou and M. Kervaire. Minimal resolutions of some monomial ideals. *J. Algebra*, 129:1–25, 1990.

[50] D. Fesser, D. Saller, and W.M. Seiler. A differential equations approach to Hamiltonian systems. Preprint Universität Mannheim, 2001.

[51] A. Finzi. Sur les Systèmes d'Équations aux Derivées Partielles qui, comme les Systèmes Normaux, Comportent Autant d'Équations que de Fonctions Inconnues. *Proc. Kon. Neder. Akad. v. Wetenschappen*, 50:136–142, 143–150, 288–297, 351–356, 1947.

[52] A.R. Forsyth. *Theory of Differential Equations II: Ordinary Equations, Not Linear.* Cambridge University Press, Cambridge, 1900.

[53] A.R. Forsyth. *Theory of Differential Equations III: Ordinary Equations, Not Linear.* Cambridge University Press, Cambridge, 1900.

[54] K. Gatermann. *Computer Algebra Methods for Equivariant Dynamical Systems.* Lecture Notes in Mathematics 1728. Springer-Verlag, Berlin, 2000.

[55] V.P. Gerdt. Completion of linear differential systems to involution. In V.G. Ghanza, E.W. Mayr, and E.V. Vorozhtsov, editors, *Computer Algebra in Scientific Computing — CASC 1999*, pages 115–137. Springer-Verlag, Berlin, 1999.

[56] V.P. Gerdt and Yu.A. Blinkov. Involutive bases of polynomial ideals. *Math. Comp. Simul.*, 45:519–542, 1998.

[57] V.P. Gerdt and Yu.A. Blinkov. Minimal involutive bases. *Math. Comp. Simul.*, 45:543–560, 1998.

[58] V.P. Gerdt, Yu.A. Blinkov, and D.A. Yanovich. Construction of Janet bases I: Monomial bases. In V.G. Ghanza, E.W. Mayr, and E.V. Vorozhtsov, editors, *Computer Algebra in Scientific Computing — CASC 2001*, pages 233–247. Springer-Verlag, Berlin, 2001.

[59] V.P. Gerdt, Yu.A. Blinkov, and D.A. Yanovich. Construction of Janet bases II: Polynomial bases. In V.G. Ghanza, E.W. Mayr, and E.V. Vorozhtsov, editors, *Computer Algebra in Scientific Computing — CASC 2001*, pages 249–263. Springer-Verlag, Berlin, 2001.

[60] A. Giovani, T. Mora, G. Niesi, L. Robbiano, and C. Traverso. "One sugar cube, please" or selection strategies in the Buchberger algorithm. In S.M. Watt, editor, *Proc. ISSAC '91*, pages 49–54. ACM Press, New York, 1991.

[61] M. Golubitsky and V. Guillemin. *Stable Mappings and Their Singularities*. Graduate Texts in Mathematics 14. Springer-Verlag, New York, 1973.

[62] M.J. Gotay, J. Isenberg, and J.E. Marsden. Momentum maps and classical relativistic fields I: Covariant field theory. Preprint physics/9801019, 1998.

[63] E. Goursat. *Cours d'Analyse Mathématique*, volume II. Gauthier-Villars, Paris, 1942.

[64] R.L. Graham, D.E. Knuth, and O. Patashnik. *Concrete Mathematics*. Addison-Wesley, Reading, 1989.

[65] W. Gröbner. *Moderne Algebraische Geometrie*. Springer-Verlag, Wien, 1949.

[66] J. Hadamard. *Lectures on Cauchy's Problem in Linear Partial Differential Equations*. Dover, New York, 1952.

[67] E. Hairer, C. Lubich, and M. Roche. *The Numerical Solution of Differential-Algebraic Equations by Runge-Kutta Methods*. Lecture Notes in Mathematics 1409. Springer-Verlag, Berlin, 1989.

[68] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II*. Springer Series in Computational Mathematics 14. Springer-Verlag, Berlin, 1996.

[69] D.H. Hartley and R.W. Tucker. A constructive implementation of the Cartan-Kähler theory of exterior differential systems. *J. Symb. Comp.*, 12:655–667, 1991.

[70] M. Hausdorf. Geometrisch-Algebraische Vervollständigung Allgemeiner Systeme von Differentialgleichungen. Master's thesis, Fakultät für Informatik, Universität Karlsruhe, 2000.

[71] M. Hausdorf and W.M. Seiler. An efficient algebraic algorithm for the geometric completion to involution. Preprint Universität Mannheim, 2000.

[72] M. Hausdorf and W.M. Seiler. Perturbation versus differentiation indices. In V.G. Ghanza, E.W. Mayr, and E.V. Vorozhtsov, editors, *Computer Algebra in Scientific Computing — CASC 2001*, pages 323–337. Springer-Verlag, Berlin, 2001.

[73] M. Hausdorf, W.M. Seiler, and R. Steinwandt. Involutive bases in the Weyl algebra. Preprint Universität Mannheim, 2001.

[74] W. Hereman. Review of symbolic software for the computation of Lie symmetries of differential equations. *Euromath Bull.*, 2:45–82, 1994.

[75] W. Hereman. Symbolic software for Lie symmetry analysis. In N.H. Ibragimov, editor, *CRC Handbook of Lie Group Analysis of Differential Equations, Volume 3: New Trends in Theoretical Development and Computational Methods*, chapter 13. CRC Press, Boca Raton, Florida, 1995.

[76] R. Hermann. *The Geometry of Non-Linear Differential Equations, Bäcklund Transformations, and Solitons, Part A*. Interdisciplinary Mathematics XII. Math Sci Press, Brookline, 1976.

[77] D. Hilbert. Über die Theorie der algebraischen Formen. *Math. Ann.*, 36:473–534, 1890.

[78] C. Hoenselaers. The strength of a system of differential equations. *Prog. Theor. Phys.*, 58:1185–1190, 1977.

[79] E. Holmgren. Über Systeme von linearen, partiellen Differentialgleichungen. *Ofversigt af kongl. Vetenskapsakad. Förh.*, 58:91–103, 1901.

[80] E. Hubert. The general solution of an ordinary differential equation. In Y.N. Lakshman, editor, *Proc. ISSAC '96*, pages 189–195. ACM Press, New York, 1996.

[81] E. Hubert. Detecting degenerate behaviors in first order algebraic differential equations. *Theor. Comp. Sci.*, 187:7–25, 1997.

[82] J.M. Hyman and M. Shashkov. Mimetic discretizations for Maxwell's equations. *J. Comp. Phys.*, 151:881–909, 1999.

[83] M. Janet. Sur les Systèmes d'Équations aux Dérivées Partielles. *J. Math. Pure Appl.*, 3:65–151, 1920.

[84] M. Janet. *Leçons sur les Systèmes d'Équations aux Dérivées Partielles.* Cahiers Scientifiques, Fascicule IV. Gauthier-Villars, Paris, 1929.

[85] F. John. *Partial Differential Equations.* Applied Mathematical Sciences 1. Springer-Verlag, New York, 1982.

[86] H.H. Johnson. Classical differential invariants and applications to partial differential equations. *Math. Ann.*, 148:308–329, 1962.

[87] E. Kähler. *Einführung in die Theorie der Systeme von Differentialgleichungen.* Teubner, Leipzig, 1934.

[88] A. Kandry-Rody and V. Weispfenning. Non-commutative Gröbner bases in algebras of solvable type. *J. Symb. Comp.*, 9:1–26, 1990.

[89] I. Kaplansky. *An Introduction to Differential Algebra.* Hermann, Paris, 1957.

[90] D.E. Knuth and P.B. Bendix. Simple word problems in universal algebras. In J. Leech, editor, *Computational Problems in Abstract Algebra*, pages 263–297. Pergamon, Oxford, 1970.

[91] S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry*, volume 1. Interscience Publisher, New York, 1963.

[92] S. Kobayashi and K. Nomizu. *Foundations of Differential Geometry*, volume 2. Interscience Publisher, New York, 1963.

[93] E.R. Kolchin. *Differential Algebra and Algebraic Groups*. Academic Press, New York, 1973.

[94] M.V. Kondratieva, A.B. Levin, A.V. Mikhalev, and E.V. Pankratiev. *Differential and Difference Dimension Polynomials*. Mathematics and its Applications 461. Kluwer, Dordrecht, 1999.

[95] S. Kowalevsky. Zur Theorie der partiellen Differentialgleichungen. *J. Reine Angew. Math.*, 80:1–32, 1875.

[96] I.S. Krasilshik, V.V. Lychagin, and A.M. Vinogradov. *Geometry of Jet Spaces and Nonlinear Partial Differential Equations*. Gordon & Breach, New York, 1986.

[97] H. Kredel and V. Weispfenning. Computing dimension and independent sets for polynomial ideals. *J. Symb. Comp.*, 6:231–247, 1988.

[98] H.-O. Kreiss and J. Lorenz. *Initial-Boundary Value Problems and the Navier-Stokes Equations*. Pure and Applied Mathematics 136. Academic Press, Boston, 1989.

[99] M. Kreuzer and L. Robbiano. *Computational Commutative Algebra 1*. Springer-Verlag, Berlin, 2000.

[100] P. Kunkel and V. Mehrmann. Canonical forms for linear differential-algebraic equations with variable coefficients. *J. Comp. Appl. Math.*, 56:225–251, 1994.

[101] P. Kunkel and V. Mehrmann. A new look at pencils of matrix valued functions. *Lin. Alg. Appl.*, 212/213:215–248, 1994.

[102] S. Lang. *Algebra*. Addison-Wesley, Reading, 1984.

[103] S. Lang. *Fundamentals of Differential Geometry*. Graduate Texts in Mathematics 191. Springer-Verlag, New York, 1999.

[104] G. Le Vey. Differential algebraic equations: A new look at the index. Rapport de Recherche 2239, INRIA Rennes, 1994.

[105] G. Le Vey. Some remarks on solvability and various indices for implicit differential equations. *Num. Algo.*, 19:127–145, 1998.

[106] H. Lewy. An example of a smooth linear partial differential equation without solution. *Ann. Math.*, 66:155–158, 1957.

[107] S. Lie and F. Engel. *Theorie der Transformationsgruppen*, volume 1. Teubner, Leipzig, 1888.

[108] W. Lucht and K. Strehmel. Discretization based indices for semilinear partial differential algebraic equations. *Appl. Num. Math.*, 28:371–386, 1998.

[109] W. Lucht, K. Strehmel, and C. Eichler-Liebenow. Indexes and special discretization methods for linear partial differential algebraic equations. *BIT*, 39:484–512, 1999.

[110] F.S. Macaulay. Some properties of enumeration in the theory of modular systems. *Proc. London Math. Soc.*, 26:531–555, 1927.

[111] D. Mall. On the relation between Gröbner and Pommaret bases. *Appl. Alg. Eng. Comm. Comp.*, 9:117–123, 1998.

[112] E.L. Mansfield. *Differential Gröbner Bases*. PhD thesis, Macquarie University, Sydney, 1991.

[113] E.L. Mansfield. A simple criterion for involutivity. *J. London Math. Soc.*, 54:323–345, 1996.

[114] K.H. Mariwalla. Applications of the concept of strength of a system of partial differential equations. *J. Math. Phys.*, 15:468–473, 1974.

[115] N.F.J. Matthews. On the strength of Maxwell's equations. *J. Math. Phys.*, 28:810–814, 1987.

[116] N.F.J. Matthews. The strength of Einstein's equations. *Gen. Rel. Grav.*, 24:17–33, 1992.

[117] A. Mayer. Über unbeschränkt integrable Systeme von linearen totalen Differentialgleichungen. *Math. Ann.*, 5:448–470, 1872.

[118] E.W. Mayr. Membership in polynomial ideals over $\mathbb{Q}$ is exponential space complete. In B. Monien and R. Cori, editors, *Theoretical Aspects of Computer Science — STACS '89*, Lecture Notes in Computer Science 349, pages 400–406. Springer-Verlag, 1989.

[119] E.W. Mayr. Some complexity results for polynomial ideals. *J. Compl.*, 13:303–325, 1997.

[120] E.W. Mayr and A.R. Meyer. The complexity of the word problems for commutative semigroups and polynomial ideals. *Adv. Math.*, 46: 305–329, 1982.

[121] G. Mendella, G. Marmo, and W.M. Tulczyjew. Integrability of implicit differential equations. *J. Phys. A: Math. Gen.*, 28:149–163, 1995.

[122] C. Méray and C. Riquier. Sur la convergence des développements des intégrales ordinaires d'un système d'équations différentielles partielles. *Ann. Sci. Ec. Norm. Sup.*, 7:23–88, 1890.

[123] B. Mishra. *Algorithmic Algebra*. Texts and Monographs in Computer Science. Springer-Verlag, New York, 1993.

[124] S. Mizohata. On the Cauchy-Kowalewski theorem. In L. Nachbin, editor, *Mathematical Analysis and Applications, Part B*, Advances

in Mathematics Supplementary Studies 7B, pages 617–652. Academic Press, New York, 1981.

[125] M. Modugno. Covariant quantum mechanics. Unpublished manuscript, Dept. of Mathematics, University of Florence, 1999.

[126] M.C. Muñoz-Lecanda and N. Román-Roy. Implicit quasilinear differential systems: A geometrical approach. *Electr. J. Diff. Eqs.*, 1999. Paper No. 10.

[127] G.L. Murphy. On the strength of Einstein's unified field equations. *Int. J. Theor. Phys.*, 18:323–327, 1979.

[128] U. Oberst. Multidimensional constant linear systems. *Acta Appl. Math.*, 20:1–175, 1990.

[129] P.J. Olver. *Applications of Lie Groups to Differential Equations.* Graduate Texts in Mathematics 107. Springer-Verlag, New York, 1986.

[130] O. Ore. Theory of non-commutative polynomials. *Ann. Math.*, 34: 480–508, 1933.

[131] M.S. Osborne. *Basic Homological Algebra.* Graduate Texts in Mathematics 196. Springer-Verlag, New York, 2000.

[132] R.S. Palais. *Seminar on the Atiyah-Singer Index Theorem.* Annals of Mathematical Studies 57. Princeton University Press, Princeton, 1965.

[133] V.P. Palamodov. *Linear Differential Operators with Constant Coefficients.* Grundlehren der mathematischen Wissenschaften 168. Springer-Verlag, Berlin, 1970.

[134] A. Pazy. *Semigroups of Linear Operators and Applications to Partial Differential Equations.* Applied Mathematical Sciences 44. Springer-Verlag, New York, 1983.

[135] R. Penney. On the dimensionality of the real world. *J. Math. Phys.*, 6:1607–1611, 1965.

[136] O.-P. Piirilä and J. Tuomela. Differential-algebraic systems and formal integrability. Research Report A326, Helsinki University of Technology, Institute of Mathematics, 1993.

[137] F.A.E. Pirani, D.C. Robinson, and W.F. Shadwick. *Local Jet Bundle Formulation of Bäcklund Transformations.* Mathematical Physics Studies 1. Reidel, Dordrecht, 1979.

[138] J.W. Polderman and J.C. Willems. *Introduction to Mathematical Systems Theory.* Texts in Applied Mathematics 26. Springer-Verlag, New York, 1998.

[139] J.F. Pommaret. *Systems of Partial Differential Equations and Lie Pseudogroups.* Gordon & Breach, London, 1978.

[140] J.F. Pommaret. *Partial Differential Equations and Group Theory.* Kluwer, Dordrecht, 1994.

[141] J.F. Pommaret. Einstein equations do not admit a generic potential. In J. Janyska, I. Kolár, and J. Slovák, editors, *Proc. 6^{th} International Conference on Differential Geometry and Applications*. Dept. of Mathematics, Masaryk University, Brno, Czech Republic, 1995. (electronically available at `www.emis.de/proceedings/6ICDGA`).

[142] J.F. Pommaret and A. Quadrat. Generalized Bezout identity. *Appl. Alg. Eng. Comm. Comp.*, 9:91–116, 1998.

[143] J.F. Pommaret and A. Quadrat. A formal approach to optimal control. Rapport de Recherche 1999-166, CERMICS, École Nationale des Ponts et Chaussées, Marne-la-Vallée, 1999.

[144] P.J. Rabier and W.C. Rheinboldt. A geometric treatment of implicit differential algebraic equations. *J. Diff. Eq.*, 109:110–146, 1994.

[145] D. Rees. A basis theorem for polynomial modules. *Proc. Cambridge Phil. Soc.*, 52:12–16, 1956.

[146] S. Reich. On an existence and uniqueness theory for nonlinear differential-algebraic equations. *Circ. Sys. Sig. Proc.*, 10:343–359, 1991.

[147] G.J. Reid. Algorithms for reducing a system of PDEs to standard form, determining the dimension of its solution space and calculating its Taylor series solution. *Eur. J. Appl. Math.*, 2:293–318, 1991.

[148] G.J. Reid, A.D. Wittkopf, and A. Boulton. Reduction of systems of nonlinear partial differential equations to simplified involutive forms. *Eur. J. Appl. Math.*, 7:635–666, 1996.

[149] M. Renardy and R.C. Rogers. *An Introduction to Partial Differential Equations*. Texts in Applied Mathematics 13. Springer-Verlag, New York, 1993.

[150] C. Riquier. *Les Systèmes d'Équations aux Derivées Partielles*. Gauthier-Villars, Paris, 1910.

[151] J.F. Ritt. *Differential Algebra*. Dover, New York, 1966. (Original: AMS Colloquium Publications, Vol. XXXIII, 1950).

[152] L. Robbiano. Introduction to the theory of Hilbert functions. *Queen's Papers Pure Appl. Math.*, 85:B1–B26, 1990. Queen's Curves Seminar Vol. VII.

[153] C. Rogers and W.F. Shadwick. *Bäcklund Transformations and Their Applications*. Mathematics in Science and Engineering 161. Academic Press, New York, 1982.

[154] M. Saito, B. Sturmfels, and N. Takayama. *Gröbner Deformations of Hypergeometric Differential Equations*. Algorithms and Computation in Mathematics 6. Springer-Verlag, Berlin, 2000.

[155] D.J. Saunders. *The Geometry of Jet Bundles.* London Mathematical Society Lecture Notes Series 142. Cambridge University Press, Cambridge, 1989.

[156] F.O. Schreyer. Die Berechnung von Syzygien mit dem verallgemeinerten Weyerstraßschen Divisionssatz. Master's thesis, Fakultät für Mathematik, Universität Hamburg, 1980.

[157] J. Schü, W.M. Seiler, and J. Calmet. Algorithmic methods for Lie pseudogroups. In N. Ibragimov, M. Torrisi, and A. Valenti, editors, *Proc. Modern Group Analysis: Advanced Analytical and Computational Methods in Mathematical Physics*, pages 337–344. Kluwer, Dordrecht, 1993.

[158] B.F. Schutz. On the strength of a system of partial differential equations. *J. Math. Phys.*, 16:855–856, 1975.

[159] W.M. Seiler. *Analysis and Application of the Formal Theory of Partial Differential Equations.* PhD thesis, School of Physics and Materials, Lancaster University, 1994.

[160] W.M. Seiler. On the arbitrariness of the general solution of an involutive partial differential equation. *J. Math. Phys.*, 35:486–498, 1994.

[161] W.M. Seiler. Applying AXIOM to partial differential equations. Internal Report 95–17, Universität Karlsruhe, Fakultät für Informatik, 1995.

[162] W.M. Seiler. Arbitrariness of the general solution and symmetries. *Acta Appl. Math.*, 41:311–322, 1995.

[163] W.M. Seiler. Indices and solvability for general systems of differential equations. In V.G. Ghanza, E.W. Mayr, and E.V. Vorozhtsov, editors, *Computer Algebra in Scientific Computing — CASC 1999*, pages 365–385. Springer-Verlag, Berlin, 1999.

[164] W.M. Seiler. Involution and constrained dynamics III: Intrinsic degrees of freedom count. *Tech. Mech.*, 20:137–146, 2000.

[165] W.M. Seiler. Completion to involution and semi-discretisations. *Appl. Num. Math.*, to appear, 2001.

[166] M. Shinbrot and R.R Welland. The Cauchy-Kowalewskaya theorem. *J. Math. Anal. Appl.*, 55:757–772, 1976.

[167] S.T.C. Siklos. Counting solutions of Einstein's equations. *Class. Quant. Grav.*, 13:1931–1948, 1996.

[168] N.J.A. Sloane and S. Plouffe. *The Encyclopedia of Integer Sequences.* Academic Press, San Diego, 1995.

[169] R.P. Stanley. Hilbert functions of graded algebras. *Adv. Math.*, 28: 57–83, 1978.

[170] R.P. Stanley. Linear diophantine equations and local cohomology. *Invent. Math.*, 68:175–193, 1982.

[171] O. Stormark. *Lie's Structural Approach to PDE Systems*. Encyclopedia of Mathematics and its Applications 80. Cambridge University Press, Cambridge, 2000.

[172] B. Sturmfels. *Algorithms in Invariant Theory*. Texts and Monographs in Symbolic Computation. Springer-Verlag, Wien, 1993.

[173] B. Sturmfels and N. White. Computing combinatorial decompositions of rings. *Combinatorica*, 11:275–293, 1991.

[174] M. Sué. Involutive systems of differential equations: Einstein's strength versus Cartan's degré d'arbitraire. *J. Math. Phys.*, 32:392–399, 1991.

[175] A. Szatkowski. Geometric characterization of singular differential algebraic equations. *Int. J. Sys. Sci.*, 23:167–186, 1992.

[176] M.E. Taylor. *Partial Differential Equations*. Texts in Applied Mathematics 23. Springer-Verlag, New York, 1996.

[177] J.M. Thomas. *Differential Systems*. AMS, New York, 1937.

[178] V.L. Topunov. Reducing systems of linear differential equations to a passive form. *Acta Appl. Math.*, 16:191–206, 1989.

[179] C. Traverso. Hilbert functions and the Buchberger algorithm. *J. Symb. Comput*, 22:355–376, 1996.

[180] A. Tresse. Sur les invariants différentiels des groupes continus de transformations. *Acta Math.*, 18:1–88, 1894.

[181] T. Tsujishita. Formal geometry of systems of differential equations. *Sugaku Exp.*, 3:25–73, 1990.

[182] J. Tuomela. On singular points of quasilinear differential and differential-algebraic equations. *BIT*, 37:968–977, 1997.

[183] J. Tuomela and T. Arponen. On the numerical solution of involutive ordinary differential systems. *IMA J. Num. Anal.*, 20:561–599, 2000.

[184] J. Tuomela and T. Arponen. On the numerical solution of involutive ordinary differential systems: Higher order methods. *BIT*, 41:599–628, 2001.

[185] V.S. Varadarajan. *Lie Groups, Lie Algebras, and Their Representations*. Graduate Texts in Mathematics 102. Springer-Verlag, New York, 1984.

[186] W.V. Vasconcelos. *Computational Methods in Commutative Algebra and Algebraic Geometry*. Algorithms and Computations in Mathematics 2. Springer-Verlag, Berlin, 1997.

[187] P.J. Vassiliou. Vessiot structure for manifolds of (p, q)-hyperbolic type: Darboux integrability and symmetry. *Trans. Amer. Math. Soc.*, 353: 1705–1739, 2001.

[188] F. Verhulst. *Nonlinear Differential Equations and Dynamical Systems*. Springer-Verlag, Berlin, 2^{nd} edition, 1996.

[189] E. Vessiot. Sur une théorie nouvelle des problèmes généraux d'intégration. *Bull. Soc. Math. Fr.*, 52:336–395, 1924.

[190] W. Walter. An elementary proof of the Cauchy-Kowalevsky theorem. *Amer. Math. Monthly*, 92:115–126, 1985.

[191] F. Warner. *Foundations of Differentiable Manifolds and Lie Groups*. Scott, Foreman, Glenview, London, 1971.

[192] E. von Weber. Partielle Differentialgleichungen. In *Enzyklopädie der mathematischen Wissenschaften, Vol. II, Part 1.1*, chapter A5, pages 294–399. Teubner, Leipzig, 1900.

[193] F. Winkler. Knuth-Bendix procedure and Buchberger algorithm — a synthesis. In G.H. Gonnet, editor, *Proc. ISSAC '89*, pages 55–67. ACM Press, New York, 1989.

[194] E. Zerz. *Topics in Multidimensional Linear Systems Theory*. Lecture Notes in Control and Information Sciences 256. Springer-Verlag, London, 2000.

[195] A.Yu. Zharkov and Yu.A. Blinkov. Involution approach to solving systems of algebraic equations. In G. Jacob, N.E. Oussous, and S. Steinberg, editors, *Proc. Int. IMACS Symp. Symbolic Computation*, pages 11–17, Lille, 1993.

Author Index

Glossary

$|_{L,\mathcal{N}}$, 39
$\left[\begin{smallmatrix}k\\r\end{smallmatrix}\right]$, 125
$\langle\mathcal{F}\rangle$, 252
$\langle\mathcal{F}\rangle_{L,\prec}$, 51
$\langle\mathcal{N}\rangle_L$, 41
$\prec_{\mathcal{F}}$, 261
$\prec_{\text{deglex}}$, 251
$\prec_{\text{degrevlex}}$, 251
$\prec_{\text{invlex}}$, 251
$\prec_{\text{lex}}$, 251
$\prec_{\text{revlex}}$, 251

$\alpha_q^{(k)}$, 121

$\beta_q^{(k)}$, 87

$\mathcal{C}_{L,\mathcal{N}}(\nu)$, 39
$\mathcal{C}_{M,r}(x)$, 233
$\mathcal{C}_q$, 5
$\mathcal{C}_q^0$, 4
$\operatorname{cls}\mu$, 232
$\operatorname{codim}\mathcal{N}$, 268
$\operatorname{coker}\beta$, 111
$C^\omega(\Omega)$, 233

$\mathcal{D}$, 46
$d\omega$, 275
$\Delta_\prec(\mathcal{I})$, 149

e_α, 259
$\mathcal{E}_b$, 269
$\mathcal{E}\times_{\mathcal{B}}\mathcal{F}$, 271
$\operatorname{Ext}^i_{\mathcal{D}}(\mathcal{M},\mathcal{D})$., 115

$\mathcal{F}(\mathcal{M})$, 266
$f\star g$, 44

$\Gamma_{loc}(\mathcal{E})$, 270
Γ_q, 9, 12

$\mathcal{H}_{\mathcal{M}}(\lambda)$, 249
$H_{\mathcal{M}}(r)$, 249
$h_{\mathcal{M}}(r)$, 249
Hom, 109

$\mathcal{I}:g$, 79
$\mathcal{I}^h$, 250
$\iota_X\omega$, 275

$j_1\sigma$, 8
$J_q\mathcal{E}$, 7, 10
$J_q(\mathcal{X},\mathcal{U})$, 3

$\operatorname{lc}_\prec f$, 252
$\operatorname{ld} f$, 199
$\operatorname{le}_\prec f$, 252
$\operatorname{lm}_\prec f$, 252
$\operatorname{lt}_\prec f$, 252
$\operatorname{lt}_\prec \mathcal{I}$, 252
$\mathcal{L}_X\omega$, 277

$\mathcal{M}_{\geq q}$, 248
$M_q(\rho)$, 24
$\mathcal{M}_q$, 23

$NF_{\mathcal{F},\prec}(g)$, 253
$N_{L,\mathcal{N}}(\nu)$, 39
$\bar{N}_{L,\mathcal{N}}(\nu)$, 40
$\mathbb{N}^n_N$, 39
$N^*_p\mathcal{N}$, 276
ν_ℓ, 220
ν_d, 212
$\tilde{\nu}_d$, 216
ν_i, 213

Index