HABILITATION THESIS

Algebraic and algorithmic approaches to analysis: Integro-differential equations, positive steady states, and wavelets

Georg Regensburger

---

cumulative treatise submitted
in partial fulfillment of the requirements
for the venia docendi in mathematics

Johannes Kepler University
Linz, Austria
August 2018

Contents

Introduction

This habilitation thesis is designed in cumulative form and is composed of 16 peer-reviewed articles in journals, conference proceedings, and collections, and one invited paper.

A unifying theme in my research is the combination of algebraic and geometric methods with symbolic computation to treat problems coming from analysis and applications. I am interested in aspects ranging from purely algebraic over constructive and algorithmic considerations to the development of software for computer algebra systems. The habilitation thesis collects papers on three topics of my research related to analysis. A main focus over the last years was on developing algebraic and algorithmic methods for integro-differential equations and boundary (value) problems. The first eleven papers (**I**–**XI**) cover the main developments and applications in this area. An overview on the literature on symbolic and algebraic methods for integro-differential operators is given in [7]. The second area (papers **XII**–**XIV**) is on positive steady states of dynamical systems arising from chemical reaction networks. In terms of the corresponding (generalized) polynomial equations (with real exponents), our results guarantee uniqueness and existence of positive solutions for all positive parameters. The last three papers (**XV**–**XVII**) are earlier work on parametrizing filter coefficients for orthonormal wavelets and applications of such parametrized families.

Within each topic, the papers are in chronological order. In the following overview, numbered labels refer to some related publications (not included in the habilitation thesis) at the end of the introduction.

Integro-differential equations

Boundary problems play a dominant role in applied mathematics since in practical problems differential equations usually come along with boundary conditions. Despite this fact, they have not been considered much from an algebraic and symbolic computation perspective. For differential equations per se, there has been a lot of research in (computer) algebra to develop algorithmic methods for simplifying and solving (systems of) differential equations.

A long-term research goal with our co-authors is to develop an algebraic foundation and algorithmic framework for solving, transforming, and simplifying (systems of) integro-differential equations and boundary problems, complementing numerical methods. For boundary problems with linear ordinary differential equations (LODEs), we introduced with Markus Rosenkranz the structure of integro-differential algebras that combines a differential algebra with suitable notions of integration and evaluation. The ring of integro-differential operators associated with an ordinary integro-differential algebra allows us to express and compute with boundary problems (differential operator plus boundary conditions) as well as solution operators (Green's operators) in one algebraic structure; see **Article I** and **Article II**. For a construction of integro-differential operators with polynomial coefficients using skew-polynomials, see **Article IV**. In **Article I**, we also develop an approach to factor a boundary problem into "smaller" boundary problems along a given factorization of the

corresponding differential operator. A prototype implementation of the corresponding algorithms for solving and factoring boundary problems in the Theorema system of Bruno Buchberger is described in **Article V**; see also the survey paper [2] for further details.

In **Article III**, we study Green's operators and the factorization of linear boundary problems from a purely linear algebra perspective. This setting applies also to (systems of) linear partial differential equations differential equations (PDEs). Some first steps for making this approach algorithmic for PDEs with constant coefficients and an implementation are discussed in **V** and in the co-supervised PhD thesis of Loredana Tec.

In a collaboration with Hansjörg Albrecher and Corina Constantinescu (former Financial Mathematics group at RICAM Linz now respectively University of Lausanne and Liverpool) and others, we adapt and apply our approach for factoring boundary problems to study ruin probabilities and related quantities in renewal risk models. The starting point was the observation that for certain probability distributions the integral equations for the quantities of interest can be transformed into boundary problems. Lifting the factorization of the differential operator to the corresponding boundary problems gives new explicit and asymptotic expressions for the so-called Gerber-Shiu function in terms of the penalty function. In **Article VI**, we study the classical case with constant premium rates leading to differential equations with constant coefficients. However, it is clear that it will often be more realistic to let premium amounts depend on the current surplus level of the insurance portfolio. In **Article VII**, we investigate these more general models based on linear boundary problems (on the half bounded interval from zero to infinity) with variable coefficients and the corresponding factorization of Green's operators.

We also investigated the generalization of results and methods for regular boundary problems (having a unique solution for every right-hand side) to singular boundary problems, which appear in several applications. We discuss how to compute generalized Green's operators for LODEs and present an implementation of integro-differential operators and the corresponding algorithms for boundary problems in the computer algebra system Maple in [1]. In **Article X**, we develop linear algebra results needed for generalizing the composition and factorization of boundary problems to singular ones. We consider generalized inverses of linear operators and study the question when their product in reverse order is again a generalized inverse. This problem (reverse order law) is well-studied in the literature for various kinds of generalized inverses, especially for matrices. Motivated by our application to boundary problems, we use implicit representation of subspaces via "boundary conditions" from the dual space. This approach gives necessary and sufficient conditions for the reverse order law to hold and a new representation of the product of generalized inverses on arbitrary vector spaces. We discuss algorithmic aspects and an implementation for linear ordinary differential equations in [4]. For further details, we refer also to the co-supervised PhD thesis of Anja Korporal.

In the frame of my Schrödinger Fellowship at INRIA Saclay–Île-de-France, we investigated with Alban Quadrat (INRIA Lille–Nord Europe) algebraic and algorithmic properties of ordinary integro-differential operators with polynomial coefficients. Differential operators with polynomial coefficients (Weyl algebras) provide a rich algebraic structure with a wealth of results and algorithmic methods. Adding an integral operator, many new phenomena appear, including zero devisors and non-finitely generated ideals. For an algorithmic approach to linear integro-differential equations, it turned out that computing polynomial solutions is a fundamental task. Combining ideas for

computing polynomial solutions for linear differential equations and homological algebra, we introduce in **Article VIII** a class of algorithmic Fredholm operators on the polynomial ring (rational indicial maps) including integro-differential operators. Based on these results, we give a constructive proof that the right annihilator of an integro-differential operator (with evaluations) is finitely generated. For initial value problems, an involution on the algebra of integro-differential operators allows us to compute also left annihilators, which can be interpreted as compatibility conditions. We give a first implementation of the corresponding algorithms based on our `IntDiffOp(operations)` package [3]. See also the extended version [9] of **VIII**, which includes a self-contained introduction to ordinary integro-differential operators with polynomial coefficients and several evaluations.

In **Article IX**, we study algebraic aspects of integro-differential algebras and their relation to so-called differential Rota-Baxter algebras. We generalize this concept to that of integro-differential algebras with weight. Based on free commutative Rota-Baxter algebras, we investigate the construction of free integro-differential algebras with weight generated by a regular differential algebra. The explicit construction is not only interesting from an algebraic point of view but is also an important step for algorithmic extensions of differential algebras to integro-differential algebras. It is also related to the universal algebra construction of integro-differential polynomials in **Article II** and [2]. Algorithmic methods for integrating fractions of differential polynomials are described in [5].

Skew polynomials (Ore extensions and Ore algebras) are a well-established algebraic and algorithmic setting for studying many common operators like differential and difference operators. However, integro-differential operators over an arbitrary integro-differential algebra, for example, do not fit this structure. In **Article XI**, we propose a general algorithmic approach to noncommutative operator algebras generated by additive operators using quotients of tensor rings. For a constructive approach, these quotients are defined by confluent tensor reduction systems, which are a basis-free analog of noncommutative Gröbner bases. See [6] for the corresponding MATHEMATICA package `TenRes`. The tensor approach also allows to model integro-differential operators with matrix coefficients, where constants are not commutative. Using tensor reduction systems, we construct normal forms for the ring of integro-differential operators with linear substitutions having matrix coefficients; see [10] for an application to linear differential time-delay systems.

Positive steady states

In **Article XII**, a joint work with Stefan Müller (University of Vienna), we propose the notion of generalized mass-action systems that can serve as a more realistic model for reaction networks in intracellular environments; classical mass-action systems capture chemical reaction networks in homogeneous and dilute solutions. In addition to the complexes of a network and the related stoichiometric subspace, we introduce corresponding kinetic complexes, which represent the exponents in the rate functions and determine the kinetic-order subspace. We show that several results of chemical reaction network theory developed by Feinberg and coauthors can be extended to the case of generalized mass-action kinetics.

Our main result gives conditions for the existence of a unique positive steady state for arbitrary initial conditions and independent of rate constants in this generalized setting. We also give necessary and sufficient conditions for multistationarity, which is an important property in many applications, for example, in connection with cell differentiation. The conditions are formulated in

terms of sign vectors (oriented matroids) of the stoichiometric and kinetic-order subspace and face lattices of related cones. In terms of the corresponding (generalized) polynomial equations, our results guarantee uniqueness and existence of positive solutions for all positive parameters. In the invited paper **Article XIII**, we focus on a constructive characterization of positive solutions of these generalized polynomial equations with real and symbolic exponents. The algorithmic methods are implemented in the Maple package `GMAK`. We discuss dynamical properties of planar generalized mass-action systems in the recent papers [8] and [11].

In **Article XIV**, a collaboration with Anne Shiu (Texas A&M University), Alicia Dickenstein (University of Buenos Aires), Carsten Conradi (HTW Berlin), Elisenda Feliu (University of Copenhagen), and Stefan Müller, we characterize the injectivity of families of generalized polynomial maps on the positive orthant in terms of sign vectors. Our work relates to and extends existing injectivity conditions expressed in terms of determinants. As one application, we give criteria for the uniqueness of steady states in chemical reaction networks with power-law kinetics. In the context of real algebraic geometry, our results allow a first partial multivariate generalization of the classical Descartes' rule, which bounds the number of positive real roots of a univariate real polynomial in terms of the number of sign variations of its coefficients. In the recent preprint [12], we give an effective characterization of the bijectivity of families of generalized polynomial maps.

Wavelets

In **Article XV** with Otmar Scherzer (University of Vienna) and in **Article XVI**, we use Gröbner bases for constructing parametrizations of filter coefficients of scaling functions and compactly supported orthonormal wavelets with several vanishing moments. The discrete moments of the filter coefficients are used as parameters and we take advantage of relations between these moments. Applications of parametrized filter coefficients include compression of signals and images and the construction of most regular or least asymmetric wavelets; see **Article XVII**. Using our parametrizations, we can also reduce the question of the existence of rational filter coefficients to finding rational points on algebraic curves. For example, we prove that there do not exist orthonormal rational filter coefficients of length six with at least two vanishing moments.

Related publications

[1] Anja Korporal, Georg Regensburger, and Markus Rosenkranz. Regular and singular boundary problems in Maple. In V. Gerdt, W. Koepf, E. Mayr, and E. Vorozhtsov, editors, *Computer Algebra in Scientific Computing. Proceedings of the 13th International Workshop (CASC 2011)*, volume 6885 of *Lecture Notes in Comput. Sci.*, pages 280–293, Berlin/Heidelberg, 2011. Springer.

[2] Markus Rosenkranz, Georg Regensburger, Loredana Tec, and Bruno Buchberger. Symbolic analysis for boundary problems: From rewriting to parametrized Gröbner bases. In U. Langer and P. Paule, editors, *Numerical and Symbolic Scientific Computing: Progress and Prospects*, Texts and Monographs in Symbolic Computation, pages 273–331. SpringerWienNew York, Vienna, 2012.

[3] Anja Korporal, Georg Regensburger, and Markus Rosenkranz. Symbolic computation for ordinary boundary problems in Maple. *ACM Commun. Comput. Algebra*, 46:154–156, 2012. Software presentation at ISSAC 2012.

[4] Anja Korporal and Georg Regensburger. Composing and factoring generalized Green's operators and ordinary boundary problems. In M. Barkatou, T. Cluzeau, G. Regensburger, and M. Rosenkranz, editors, *AADIOS 2012*, volume 8372 of *Lecture Notes in Comput. Sci.*, pages 116–134, Berlin/Heidelberg, 2014. Springer.

[5] François Boulier, François Lemaire, Joseph Lallemand, Georg Regensburger, and Markus Rosenkranz. Additive normal forms and integration of differential fractions. *J. Symbolic Comput.*, 77:16–38, 2016.

[6] Jamal Hossein Poor, Clemens G. Raab, and Georg Regensburger. Normal forms for operators via Gröbner bases in tensor algebras. In Gert-Martin Greuel, Thorsten Koch, Peter Paule, and Andrew Sommese, editors, *Mathematical Software – ICMS 2016*, volume 9725 of *Lecture Notes in Comput. Sci.*, pages 505–513. Springer International Publishing, 2016.

[7] Georg Regensburger. Symbolic computation with integro-differential operators. In Markus Rosenkranz, editor, *ISSAC '16: Proceedings of the 41th international symposium on Symbolic and algebraic computation*, pages 17–18, New York, NY, USA, 2016. ACM. Tutorial, extended abstract.

[8] Balázs Boros, Josef Hofbauer, Stefan Müller, and Georg Regensburger. The center problem for the Lotka reactions with generalized mass-action kinetics. *Qual. Theory Dyn. Syst.*, 17:403–410, 2018.

[9] Alban Quadrat and Georg Regensburger. Computing polynomial solutions and annihilators of integro-differential operators with polynomial coefficients. In *Algebraic Methods and Symbolic-Numeric Computation in Systems Theory*, Advances in Delays and Dynamics (ADD), 26 pages. Springer, 2018. To appear.

[10] Thomas Cluzeau, Jamal Hossein Poor, Alban Quadrat, Clemens G. Raab, and Georg Regensburger. Symbolic computation for integro-differential-time-delay operators with matrix coefficients. In *14th IFAC Workshop on Time Delay Systems*, IFAC-PapersOnLine, 6 pages, 2018. To appear.

[11] Balázs Boros, Josef Hofbauer, Stefan Müller, and Georg Regensburger. Planar S-systems: Global stability and the center problem. *Discrete Contin. Dyn. Syst.*, 26 pages, 2018. Under revision.
arXiv:1707.02104 [math.DS].

[12] Stefan Müller, Josef Hofbauer, and Georg Regensburger. On the bijectivity of families of exponential/generalized polynomial maps. 26 pages, 2018. Submitted.
arXiv:1804.01851 [math.AG].

Available online at www.sciencedirect.com
ScienceDirect
Journal of Symbolic Computation 43 (2008) 515–544
Journal of Symbolic Computation
www.elsevier.com/locate/jsc

Solving and factoring boundary problems for linear ordinary differential equations in differential algebras☆

Markus Rosenkranz, Georg Regensburger

Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria

Received 8 May 2007; accepted 26 November 2007
Available online 4 December 2007

Abstract

We present a new approach for expressing and solving boundary problems for linear ordinary differential equations in the language of differential algebras. Starting from an algebra with a derivation and integration operator, we construct an algebra of linear integro-differential operators that is expressive enough for specifying regular boundary problems with arbitrary Stieltjes boundary conditions as well as their solution operators.

On the basis of these structures, we define a new multiplication on regular boundary problems in such a way that the resulting Green's operator is the reverse composition of the constituent Green's operators. We provide also a method for lifting any factorization of the underlying differential operator to the level of boundary problems. Since this method only needs the computation of initial value problems, it can be used as an effective alternative for computing Green's operators in the case where one knows how to factor the given differential operators.

Keywords: Linear boundary value problems; Ordinary differential equations; Green's operators; Factorization; Differential algebra; Noncommutative Gröbner bases

1. Introduction

In this paper, we develop a new approach for handling *boundary problems* in the language of differential algebras, restricting ourselves to the case of linear boundary problems for

☆ This work was supported by the Austrian Science Fund (FWF) under the SFB grant F1322.
E-mail addresses: Markus.Rosenkranz@oeaw.ac.at (M. Rosenkranz), Georg.Regensburger@oeaw.ac.at (G. Regensburger).
URLs: http://www.ricam.oeaw.ac.at (M. Rosenkranz), http://www.ricam.oeaw.ac.at (G. Regensburger).

0747-7171/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jsc.2007.11.007

ordinary differential equations. (We reserve the traditional term "boundary value problem" for the particular type of boundary problems that have only point evaluations, i.e. point conditions in the terminology of Section 5.) The algebraic language that we build up allows us

- to state boundary problems in a natural algebraic language,
- to express their solution operators in the same language,
- to compute the solution operators from a fundamental system,
- to multiply boundary problems corresponding to the solution operators,
- to lift factorizations of differential operators to boundary problems.

The present paper extends the ideas from Rosenkranz (2005) and Rosenkranz et al. (2003) in several aspects: Boundary problems can now be formulated and solved in any differential algebra that meets some natural conditions (Theorem 26), the case of variable coefficients is fully included, and a new monoid structure on boundary problems provides an elegant description and an alternative computation method for the corresponding solution operators.

For developing an appropriate notion of a boundary problem in a given differential algebra, it will be useful to have a look at the *classical setting* of Stakgold (1979, p. 203) dealing with a two-point boundary problem on a finite interval $[a, b]$. Disregarding weak solutions and ill-posed problems for simplicity, the general idea is that a differential equation

$$u^{(n)}(x) + c_{n-1}(x)\, u^{(n-1)}(x) + \cdots + c_1(x)\, u'(x) + c_0(x)\, u(x) = f(x) \tag{1}$$

with coefficient functions $c_{n-1}, \ldots, c_1, c_0 \in C^\infty[a, b]$ and forcing function $f \in C^\infty[a, b]$ is supplemented with additional conditions that determine the solution $u \in C^\infty[a, b]$ uniquely. In certain cases, these may be initial conditions, but in general one has to deal with constraints that combine the values and derivatives of u at both endpoints a and b. In the context of a linear differential equation like (1), it is natural to restrict oneself to linear conditions of the form

$$p_{n-1}\, u^{(n-1)}(a) + \cdots + p_0\, u(a) + q_{n-1}\, u^{(n-1)}(b) + \cdots + q_0\, u(b) = e, \tag{2}$$

where the p_i, q_i and e are given complex numbers. For obvious reasons, boundary conditions of the form (2) are known as two-point boundary conditions; note that they include initial conditions as the special case where all the q_i vanish. In order to obtain a regular boundary problem, one imposes n suitable linear boundary conditions (2) on a given linear differential equation (1). Since all differential equations, operators and conditions will be linear in this paper, we will from now on drop the attribute "linear".

Classical boundary problems (1), (2) have a rich structure. First of all, it is clear that one can decompose the solution of (1), (2) into a solution of the semi-inhomogeneous problem (obtained from (2) by setting all $e = 0$) and a solution of the semi-homogeneous problem (obtained from (1) by setting $f = 0$). Since we assume that fundamental systems are available, the latter problem reduces to linear algebra, and we can concentrate on the semi-inhomogeneous problem. Thus we assume from now on homogeneous boundary conditions.

A second crucial observation is that the solution u depends linearly on the forcing function f. In fact, the assumption of a regular boundary problem means (Definition 25) that there is a unique u for every given f, so there is a solution operator $G\colon C^\infty[a, b] \to C^\infty[a, b]$ with $u = Gf$. This so-called *Green's operator* G is linear.

Taking advantage of the linear structure, it is possible to compute the Green's operator G rather than a particular solution u belonging to a fixed forcing function f. We may view this as solving the parametrized differential equation (1) together with boundary conditions (2). There

is also a practical reason why it is useful to have the Green's operator: The forcing function f is often more likely to change (e.g. as the "source term" in heat conduction), while the shape of the differential equation (its left-hand side) and the boundary conditions remain fixed. In the classical setting, the Green's operator $G\colon C^\infty[a,b] \to C^\infty[a,b]$ can be represented in the form of an integral operator

$$Gf(x) = \int_a^b g(x,\xi)\, f(\xi)\, \mathrm{d}\xi$$

with a uniquely determined *Green's function* $g \in C^{n-2}[a,b]^2$. So once g is found, one can compute each desired solution u in a single integration.

Now let us describe our strategy of rebuilding this scenario in a (moderately general) *differential algebra*. In the place of $C^\infty[a,b]$, we take a differential algebra $\mathcal{F}$ as our starting point. Obviously, a differential equation (1) is then given by

$$Tu = f \tag{3}$$

with a differential operator $T \in \mathcal{F}[\partial]$, and one has to find the solution $u \in \mathcal{F}$ in terms of a given forcing function $f \in \mathcal{F}$. (In order to gain flexibility, we will actually consider differential operators $T \in \mathcal{F}_0[\partial]$ for a suitable subalgebra $\mathcal{F}_0 \leq \mathcal{F}$; see Definition 18.) Boundary conditions can be given by

$$\beta_1 u = \cdots = \beta_n u = 0 \tag{4}$$

for suitable functionals $\beta_1, \ldots, \beta_n \in \mathcal{F}^*$, where $\mathcal{F}^*$ denotes the dual space of $\mathcal{F}$. We will allow rather general boundary conditions of the so-called Stieltjes type (see Definition 14), including not only two-point conditions like (2) but also global conditions involving integrals.

At this point, we would like to make a general remark on *point evaluation* in differential algebra. This is a topic not often considered (within the given algebraic setting), despite its undisputed importance in the applications. The problem is that the elements of a differential algebra (or differential ring or differential field) are abstractions of functions that are not meant to be "evaluated". Robinson (1961) has addressed this discrepancy by introducing what he called localized differential rings. Working in the much wider scope of polynomial differential equations, he has developed a solvability criterion for initial value problems. To our knowledge, his ideas have not found much resonance. For a more practical perspective on initial value problems for differential–algebraic equations, see the recent survey by Pritchard and Sit (in press), containing a method for determining admissible initial conditions. Our own approach is to consider boundary conditions in their natural context: as functionals of the aforementioned type.

This is why we require a differential algebra—they provide a vector space structure together with the structure of a differential ring. In fact, we need more than that (Section 2): Since we want to express the Green's operator of a boundary problem (3), (4), we need a linear operator $\int$ denoting *integration*, just like ∂ is used for differentiation. We stipulate that $\int$ is a section (right inverse) of ∂, meaning that

$$\partial \int = 1.$$

Further analysis will make it clear that we must also require $\int$ to satisfy a version of the Baxter axiom, an algebraic formulation of integration by parts. As we shall see, this necessarily excludes differential fields from the admissible differential algebras $\mathcal{F}$. We are thus led to the following

crucial observation (Proposition 6): Despite their extremely useful role e.g. in the Galois theory of linear differential equations (van der Put and Singer, 2003), differential fields are inadequate for treating initial/boundary conditions along with the differential equations. In some sense, this result is to be expected: Point evaluations correspond to maximal ideals, which are not available in fields.

We call the resulting structure $(\mathcal{F}, \partial, \int)$ an integro-differential algebra. They induce a natural algebra of *integro-differential operators* $\mathcal{F}[\partial, \int]$, just like $(\mathcal{F}, \partial)$ alone induces the algebra of differential operators $\mathcal{F}[\partial]$. We introduce a suitable rewrite system (Baader and Nipkow, 1998) for these operators (Section 3), enabling their convenient symbolic manipulation. Our rewrite system is both Noetherian and confluent (Proposition 13), and the corresponding normal forms have a natural description (Proposition 17). The advantage of the $\mathcal{F}[\partial, \int]$ language is that it provides a uniform frame for stating initial/boundary problems as well as deriving and expressing their Green's operators.

The departure from differential fields has the consequence that inhomogeneous differential equations cannot be reduced to homogeneous ones in the way explained by van der Put and Singer (2003, Exercise 1.14.1). Hence we have to resort to an algebraic version of the familiar method of "variation of the constant" for solving even *initial value problems* (Section 4), and this necessitates a condition on solutions of inhomogeneous first-order differential equations. It essentially requires that exponential solutions exist and behave as normal: they have a reciprocal.

For treating *boundary problems* (3), (4) in a convenient fashion, we specify them as pairs:

$$(T, \mathcal{B}) \quad \text{with} \quad T \in \mathcal{F}_0[\partial] \quad \text{and} \quad \mathcal{B} = [\beta_1, \ldots, \beta_n] \leq \mathcal{F}^*.$$

Using this setup, we will show (Section 5) that they have a Green's operator that can be expressed in $\mathcal{F}[\partial, \int]$, and we sketch how one can compute it. For a concrete implementation in the classical C^∞ setting, see the previous article (Rosenkranz, 2005). Generalizing the idea of a boundary problem as "a surjective linear map with linear functionals as side conditions", we have also developed an abstract treatment for general vector spaces in our forthcoming paper Regensburger and Rosenkranz (in press). This approach allows us to apply the ideas of Sections 6 and 7, e.g. to linear partial differential equations or systems of linear ordinary differential equations.

The algebraic treatment of boundary problems applied in this paper not only allows for a symbolic solution, it is also a natural setting for exposing an important structure connecting boundary problems amongst themselves (Section 6): It turns out that the composition structure of Green's operators is reflected in a *monoid structure* on the boundary problems, arising as a semi-direct product of $\mathcal{F}_0[\partial]$ and the additive structure of subspaces in $\mathcal{F}^*$.

Finally (Section 7), we will show how to factor a given boundary problem $(T, \mathcal{B})$ into smaller ones. While *factorization* of linear ordinary differential operators is an important topic in symbolic computation (Grigoriev, 1990; van der Put and Singer, 2003; Schwarz, 1989; Tsarev, 1996), it neglects the presence of boundary conditions (possibly addressed in a post-processing step). We will show how every factorization of the differential operator T gives rise to various factorizations of $(T, \mathcal{B})$, whose full classification is stated. In order to lift a factorization of T to the level of boundary problems, one only needs to solve an initial value problem. Hence one may employ factorization as a tool for computing the Green's operator G. In the extreme case of splitting T into linear factors, one obtains G as a composition of first-order Green's operators, which can be computed easily. (In practical examples, one will often be content with a partial factorization.)

Some remarks on *notation*. We write $\mathbb{N}$ for the set of all natural numbers including zero. The variable n ranges over $\mathbb{N}$. All algebras are assumed to be commutative with identity. The zero-

dimensional subspace of any vector space will be denoted by $O = \{0\}$. We write $[f_1, \ldots, f_n]$ for the subspace generated by the vectors $f_1, \ldots, f_n$ of some vector space $\mathcal{F}$. For subsets $\mathcal{A} \subseteq \mathcal{F}$ and $\mathcal{B} \subseteq \mathcal{F}^*$, the so-called orthogonal is defined as

$$\mathcal{A}^\perp = \{\varphi \in \mathcal{F}^* \mid \forall_{f\in\mathcal{A}}\, \varphi(f) = 0\} \leq \mathcal{F}^*,$$
$$\mathcal{B}^\perp = \{f \in \mathcal{F} \mid \forall_{\varphi\in\mathcal{B}}\, \varphi(f) = 0\} \leq \mathcal{F};$$

see Section 5 for more details.

2. Integration in differential algebras

Let $(\mathcal{F}, \partial)$ be a *differential algebra* over a field K, so $\partial\colon \mathcal{F} \to \mathcal{F}$ is a K-linear map fulfilling the Leibniz rule $\partial(fg) = f\,\partial(g) + g\,\partial(f)$. For convenience, we may assume $K \leq \mathcal{F}$, and we write f' as shorthand for $\partial(f)$. Furthermore, we will assume that K has characteristic zero (even though some definitions and results would make sense in positive characteristic), except when stated otherwise. Then we may also assume $\mathbb{Q} \leq K$, so that $\mathcal{F}$ is what is sometimes called a Ritt algebra (Kaplansky, 1957, p. 12).

The algebra of (formal) *differential operators* over the differential algebra $\mathcal{F}$ is denoted by $\mathcal{F}[\partial]$, as e.g. in van der Put and Singer (2003). Addition in $\mathcal{F}[\partial]$ is obvious, while multiplication is determined by the rule $\partial f = f\partial + f'$. Each $T \in \mathcal{F}[\partial]$ acts on $\mathcal{F}$ as an (actual) differential operator $T\colon \mathcal{F} \to \mathcal{F}$. The identity operator of $\mathcal{F}[\partial]$ is denoted by $\partial^0 = 1$ just like the unit element $1 \in \mathcal{F}$; it will be clear from the context which is meant.

Our goal is to solve inhomogeneous differential equations by using Green's operators. The simplest such equation is $u' = f$, and its solution operators $\int\colon f \mapsto u$ are exactly the sections of the differential operator ∂. A derivation need not have any sections; e.g. in the algebra of univariate differential polynomials, the indeterminate cannot be a derivative. But if it does, their description follows from linear algebra.

Proposition 1. *Every section $\int\colon \mathcal{F} \to \mathcal{F}$ of the derivation $\partial\colon \mathcal{F} \to \mathcal{F}$ corresponds to a unique projector $P\colon \mathcal{F} \to \mathcal{F}$ with $P = 1 - \int\partial$, and to a unique direct sum $\mathcal{F} = \mathcal{C} \dotplus \mathcal{I}$ with $\mathcal{C} = \operatorname{Ker}(\partial) = \operatorname{Im}(P)$ and $\mathcal{I} = \operatorname{Im}(\int) = \operatorname{Ker}(P)$.*
If $\int$ is any fixed section of ∂, every projector P with $\operatorname{Im}(P) = \operatorname{Ker}(\partial)$ induces a section $(1-P)\int$, and every section of ∂ arises uniquely in this way.

Proof. See Nashed and Votruba (1976, p. 17) or Regensburger and Rosenkranz (in press). □

We refer to the elements of $\mathcal{I} = \operatorname{Im}(\int)$ as the *initialized* functions (with respect to $\int$), while those of $\mathcal{C} = \operatorname{Ker}(\partial)$ are usually known as the *constants* (with respect to ∂). In the prototypical case of $\mathcal{F} = C^\infty(\mathbb{R})$, the initialized functions are those that can be written as $F(x) = \int_\alpha^x f(\xi)\,\mathrm{d}\xi$ for an integrand $f \in C^\infty(\mathbb{R})$ and an initialization point $\alpha \in \mathbb{R}$; hence F is exactly that antiderivative of f that fulfills the initial condition $F(\alpha) = 0$.

For solving inhomogeneous differential equations $Tu = f$ of higher order, one must expect to iterate the section $\int$. In general, this could lead to "nested integrals" of arbitrary complexity. But we know from the classical C^∞ setting (see Section 1) that the Green's operator G can always be expressed using a single integration, with the so-called *Green's function* g as its integral kernel. The essential role of Green's functions is to resolve nested integrals, whereas the passage from an operator $G\colon C^\infty[a,b] \to C^\infty[a,b]$ to a function $g \in C^{n-2}[a,b]^2$ is immaterial from our viewpoint.

In order to capture this behavior, we need an identity for resolving nested integrals (eventually leading to the $\int f \int$ rule in Table 1). Such an identity is given by the so-called *Baxter axiom* (of weight zero), asserting

$$(\int f)(\int g) = \int(f \int g) + \int(g \int f) \tag{5}$$

for all $f, g \in \mathcal{F}$; see Guo (2002), Baxter (1960) and Rota (1969) for more details. One sees immediately that (5) is an algebraic version of integration by parts, rewritten in such a way that it need not refer to any derivation. A Baxter algebra $(\mathcal{F}, \int)$ is then a K-algebra $\mathcal{F}$ with a K-linear operation $\int$ fulfilling the Baxter axiom (5).

If $\int$ is again a section of a derivation ∂ on $\mathcal{F}$, we note an important consequence of (5). Writing x as an abbreviation for $\int 1$, we obtain $x^2/2 = \int\int 1$ and inductively $x^n/n! = \int \cdots \int 1$ with n iterates of $\int$. Hence the powers $u = x^k$ with $k < n$ are solutions of $u^{(n)} = 0$, and one checks immediately that they are all linearly independent. This means that $\mathrm{Ker}(\partial^n)$ contains $[1, x, \ldots, x^{n-1}]$ as an n-dimensional subspace. So we see that $\mathcal{F}$ contains (an isomorphic copy of) the *polynomial ring* $K[x]$ and is thus infinite dimensional. Note that $K[x] \leq \mathcal{F}$ is simultaneously a differential algebra under ∂ and a Baxter algebra under $\int$, so $(K[x], \partial, \int)$ is an integro-differential algebra in the sense of Definition 4.

What we shall actually need is the *differential Baxter axiom*, requiring

$$\int fg = f \int g - \int(f' \int g) \tag{6}$$

for all $f, g \in \mathcal{F}$. Note that this is what most people do when they actually apply integration by parts (eventually leading to the $\int f \partial$ rule in Table 1), but (6) cannot be stated in pure Baxter algebras. The variant (5) follows immediately by substituting $\int f$ for f in (6), and often the two versions are actually equivalent (especially in the cases relevant for us—see after Definition 8). For seeing that in general (6) is stronger than (5), we need a somewhat artificial construction (Example 3). In fact, we can easily characterize what makes the *differential Baxter axiom* stronger than the pure one.

Lemma 2. *A section $\int$ of ∂ fulfills the differential Baxter axiom* (6) *iff it fulfills the pure Baxter axiom* (5) *and the homogeneity condition $\int cf = c \int f$ for all $c \in \mathcal{C}$ and $f \in \mathcal{F}$.*

Proof. Assume $\int$ fulfills (6). Then $\int$ also fulfills (5) as observed above, while substituting a constant $c \in \mathcal{C}$ for f in (6) gives homogeneity. Conversely, assume that $\int$ fulfills (5) and the homogeneity condition. The latter hypothesis means that (6) is satisfied if $f \in \mathcal{C}$. Now consider $f \in \mathcal{I}$ so that $\int f' = f$. Substituting f' for f in (5), we see that (6) is also satisfied for these $f \in \mathcal{I}$. But then the general case of $f \in \mathcal{F}$ follows via the direct sum $\mathcal{F} = \mathcal{C} \dotplus \mathcal{I}$. □

Example 3. Let K be a field of characteristic zero. Then $(R[x], \partial)$ with $R = K[y]/y^4$ and $\partial f = f_x$ is a differential algebra over K. Defining

$$\int f = \int_0^x f(\xi, y)\, \mathrm{d}\xi + f(0,0)\, y^2, \tag{7}$$

we obtain a K-linear map $\int\colon R[x] \to R[x]$. Since the second term vanishes under ∂, we see immediately that $\int$ is a section of ∂. For verifying the Baxter axiom (5), let us write $\int\!\!\!\!\!-$ for the ordinary integral in (7) and compute

$$(\int f)(\int g) = (\int\!\!\!\!\!- f)(\int\!\!\!\!\!- g) + y^2 \int\!\!\!\!\!- \big(g(0,0)\, f + f(0,0)\, g\big) + f(0,0)\, g(0,0)\, y^4,$$
$$\int(f \int g) = \int f(\int\!\!\!\!\!- g + g(0,0)\, y^2) = \int\!\!\!\!\!- (f \int\!\!\!\!\!- g) + y^2 \int\!\!\!\!\!- \big(g(0,0)\, f\big).$$

Since $y^4 \equiv 0$ and the ordinary integral $\int$ fulfills the Baxter axiom (5), this implies immediately that $\int$ does also. However, it does not fulfill the stronger axiom (6), because the homogeneity condition is violated: Observe that $\operatorname{Ker}(\partial) = R$, so in particular we should have $\int y \cdot 1 = y \cdot \int 1$. But one checks immediately that the left-hand side yields xy, while the right-hand side yields $xy + y^3$.

For excluding cases like the preceding example, we will insist that "integral operators" must satisfy the *differential Baxter axiom*.

Definition 4. Let $(\mathcal{F}, \partial)$ be a differential algebra. A section $\int$ of ∂ is called an *integral* if it satisfies the differential Baxter axiom (6). In this case, we call $(\mathcal{F}, \partial, \int)$ an *integro-differential algebra*.

Example 5. As an *example*, detailed in Rosenkranz (2005, p. 176), take $\mathcal{F} = C^\infty[a, b]$ with its usual derivation ∂ and integral operators

$$\int^*\colon f \mapsto \int_a^x f(\xi)\,\mathrm{d}\xi \quad \text{and} \quad \int_*\colon f \mapsto \int_x^b f(\xi)\,\mathrm{d}\xi.$$

Then both $(\mathcal{F}, \partial, \int^*)$ and $(\mathcal{F}, \partial, -\int_*)$ are integro-differential algebras. By contrast, the operator

$$f \mapsto \int_a^b \int_\tau^x f(\xi)\,\mathrm{d}\xi\,\mathrm{d}\tau,$$

used for regularizing an ill-posed problem in Rosenkranz (2005, p. 192), is just a section for ∂, but not an integral.

Using Proposition 1, we can characterize integrals by their projectors and direct sums. In the above example, we observe that the projectors $f \mapsto f(a)$ and $f \mapsto f(b)$, corresponding respectively to the integrals $\int^*$ and $-\int_*$, are *multiplicative*, whereas the projector $\int_a^b$ for the third operator is not. This behavior is the key to their characterization.

Proposition 6. *A section $\int\colon \mathcal{F} \to \mathcal{F}$ of the derivation $\partial\colon \mathcal{F} \to \mathcal{F}$ is an integral iff its projector $P\colon \mathcal{F} \to \mathcal{F}$ is multiplicative iff $\mathcal{I} = \operatorname{Im}(\int)$ is an ideal.*

Proof. Assume first that $\int$ is an integral for ∂, let $P = 1 - \int\partial$ be its projector and $\mathcal{F} = \mathcal{C} \dotplus \mathcal{I}$ the corresponding direct sum with $\mathcal{C} = \operatorname{Ker}(\partial)$ and $\mathcal{I} = \operatorname{Im}(\int)$, according to Proposition 1. We must prove $P(fg) = P(f)\,P(g)$ for all $f, g \in \mathcal{F}$. Substituting g' for g in (6), we obtain

$$\begin{aligned} 0 &= \textstyle\int fg' - f\int g' + \int(f'\int g') = \int fg' - f(g - Pg) + \int(f'(g - Pg)) \\ &= \textstyle\int fg' + \int f'g - fg + f\,Pg - (\int f')\,Pg, \end{aligned}$$

where we have used the homogeneity of $\int$ in the last step. But then

$$P(fg) = fg - \textstyle\int(f'g + fg') = (f - \int f')\,Pg = Pf\,Pg,$$

as claimed. Assume conversely that P is multiplicative, and take $f, G \in \mathcal{F}$ arbitrary. Expanding the definition of P and using the Leibniz law gives

$$P(fG) = (1 - \textstyle\int\partial)fG = fG - \int f'G - \int fG'$$

and

$$Pf\,PG = (f - \textstyle\int f')(G - \int G') = fG - G\int f' - f\int G' + (\int f')(\int G');$$

equating the two expressions, we obtain

$$(\int f')(\int G') + \int f'G + \int fG' = G\int f' + f\int G',$$

which yields indeed (6) by specializing to $G = \int g$.

Let us now prove that $\mathcal{I}$ is an ideal under the assumption that P is multiplicative. Since P is a projector along $\mathcal{I}$, we have $PG = 0$ iff $G \in \mathcal{I}$. Hence for all $f \in \mathcal{F}$ and $G \in \mathcal{I}$ we have $P(fG) = Pf\ PG = 0$, and $fG \in \mathcal{I}$ as claimed. Finally, we assume that $\mathcal{I}$ is an ideal and prove that P is multiplicative. Taking $f, g \in \mathcal{F}$ arbitrary, we set $f_0 = Pf \in \mathcal{C}$ and $g_0 = Pg \in \mathcal{C}$. Then $f_1 = f - f_0 \in \mathcal{I}$ and likewise $g_1 = g - g_0 \in \mathcal{I}$, so we obtain

$$P(fg) = P(f_0g_0) + P(f_0g_1) + P(f_1g_0) + P(f_1g_1) = f_0g_0 = Pf\ Pg$$

since all of f_0g_1, f_1g_0, $f_1g_1 \in \mathcal{I}$ vanish under P, while $f_0g_0 \in \mathcal{C}$ is fixed by P. □

For the operators $\int^*$ and $\int_*$ in Example 5, the *Baxter axiom* is of course known to hold. In the following example, where this is not obvious, we can take advantage of Proposition 6.

Example 7. Consider $\mathcal{F} = C^\infty(\mathbb{R}^2)$ with the derivation $\partial u = u_x + u_y$. Finding sections for ∂ means solving the partial differential equation $u_x + u_y = f$. Its general solution is given by

$$u(x, y) = \int_\alpha^x f(t, t - x + y)\,\mathrm{d}t + g(y - x),$$

where $g \in C^\infty(\mathbb{R})$ and $\alpha \in \mathbb{R}$ are arbitrary. In order to ensure a linear section, one has to choose $g = 0$, arriving at

$$\int f = \int_\alpha^x f(t, t - x + y)\,\mathrm{d}t.$$

Using a change of variables, one may verify that $\int$ satisfies the Baxter axiom (5), so $(\mathcal{F}, \int)$ is a Baxter algebra. We see also that $\mathcal{C} = \mathrm{Ker}(\partial)$ is given by the functions $(x, y) \mapsto g(x - y)$ with arbitrary $g \in C^\infty(\mathbb{R})$, while $\mathcal{I} = \mathrm{Im}(\int)$ consists of the functions $f \in \mathcal{F}$ satisfying $f(\alpha, y) = 0$ for all $y \in \mathbb{R}$. The projector $P\colon \mathcal{F} \to \mathcal{F}$ maps a function f to the function $(x, y) \mapsto f(\alpha, \alpha - x + y)$. Since the homogeneity condition is obviously satisfied, we conclude that $(\mathcal{F}, \partial, \int)$ is an integro-differential algebra. But with Proposition 6, we could have derived this result immediately since P is multiplicative and $\mathcal{I}$ an ideal.

As we see from the above example, the space of constants for an integro-differential algebra may be infinite dimensional. Since we want to treat boundary problems for *ordinary differential equations*, we will exclude these cases. Note that in the following definition our terminology deviates from that of Kolchin (1973, p. 58), which simply requires having a single derivation. So in Kolchin's sense, the differential algebra of Example 7 would be addressed as "ordinary".

Definition 8. A differential algebra $(\mathcal{F}, \partial)$ is called *ordinary* if $\dim \mathrm{Ker}(\partial) = 1$.

Having an ordinary differential algebra $\mathcal{F}$ has several important *consequences*. First of all, it is clear that we have $K = \mathcal{C}$, so $\mathcal{F}$ is an algebra over its own field of constants. But then a section is automatically homogeneous over $\mathcal{C}$, so the pure Baxter axiom (5) and its differential version (6) coincide. Furthermore, we obtain the familiar relation

$$\mathrm{Ker}(\partial^n) = [1, x, \ldots, x^{n-1}], \tag{8}$$

which can be seen thus: As mentioned above, the Baxter axiom implies the inclusion $\supseteq$. Equality follows from $\dim \operatorname{Ker}(\partial^n) = n$, which is a consequence of the identity

$$\operatorname{Ker}(T^2) = G\ \operatorname{Ker}(T) \dotplus \operatorname{Ker}(T)$$

in Regensburger and Rosenkranz (in press), generally valid for epimorphisms T and sections G of T.

One knows from linear algebra that a projector P onto a one-dimensional subspace $[w]$ of a K-vector space V can be written as $P(v) = \varphi(v)\, w$, where φ is a unique functional with $\varphi(w) = 1$. If V is moreover a K-algebra, a projector onto $K = [1]$ is canonically described by the functional φ with normalization $\varphi(1) = 1$. Hence in an ordinary differential algebra, the projectors corresponding (via Proposition 1) to sections of the derivation can be regarded as *normalized functionals*.

In an ordinary *integro*-differential algebra $(\mathcal{F}, \partial, \int)$, the normalized functional corresponding to the integral $\int$ is moreover *multiplicative* by Proposition 6. Since this will be a crucial ingredient for our later development, it deserves a special name.

Definition 9. Let $(\mathcal{F}, \partial, \int)$ be an ordinary integro-differential algebra. Then we call the multiplicative functional $\mathtt{E} = 1 - \int\partial$ its *evaluation*.

The terminology stems from the *standard model* described in Example 5, where $\mathtt{E}$ is a point evaluation. For boundary problems on a finite interval, it is natural to treat both endpoints specially, leading to a pair of evaluations and integrals. This is the situation described in Rosenkranz (2005, p. 182) by the concept of "analytic algebra".

Example 10. An *analytic algebra* $(\mathcal{F}, \partial, \int^*, \int_*)$ is equivalent to a pair of ordinary integro-differential algebras $(\mathcal{F}, \partial, \int^*)$ and $(\mathcal{F}, \partial, -\int_*)$. Writing as in the above reference $f \mapsto f^{\leftarrow}$ and $f \mapsto f^{\rightarrow}$ for the evaluations of respectively $\int^*$ and $\int_*$, one finds that

$$(\textstyle\int^* f)^{\rightarrow} = \int^* f + \int_* f = (\int_* f)^{\leftarrow}.$$

This relation implies (after some calculation) that $\int_*$ is the adjoint of $\int^*$, with respect to the inner product $\langle|\rangle\colon \mathcal{F} \times \mathcal{F} \to \mathcal{C}$ given by

$$\langle f\,|\,g\rangle = (\textstyle\int^* + \int_*)\, fg.$$

In the standard model $\mathcal{F} = C^\infty[a, b]$, we have $f^{\leftarrow} = f(a)$ and $f^{\rightarrow} = f(b)$, yielding the L^2 inner product $\langle f\,|\,g\rangle = \int_a^b f(x)g(x)\,\mathrm{d}x$.

The multiplicative functionals on an algebra are known as its *characters* (note that all characters are normalized). We write $\mathcal{M}(\mathcal{F})$ for the vector space of all characters on an ordinary integro-differential algebra $(\mathcal{F}, \partial, \int)$. The evaluation of $\mathcal{F}$ is a distinguished character $\mathtt{E} \in \mathcal{M}(\mathcal{F})$ whose kernel $\mathcal{I}$ is an ideal with $\mathcal{F} = K \dotplus \mathcal{I}$ according to Proposition 6.

One calls a K-algebra *augmented* if there exists a character on it. Its kernel $\mathcal{I}$ is then known as an *augmentation ideal* and forms a direct summand of K; see Cohn (2003, p. 132). Augmentation ideals are always maximal ideals (generalizing the $C^\infty[0, 1]$ case) since the direct sum $\mathcal{F} = K \dotplus \mathcal{I}$ induces a ring isomorphism $\mathcal{F}/\mathcal{I} \cong K$. Reformulating Proposition 6, we obtain now a characterization of integrals in ordinary differential algebras.

Corollary 11. *In an ordinary differential algebra $(\mathcal{F}, \partial)$, a section $\int$ of ∂ is an integral iff its normalized functional is a character iff $\mathcal{I} = \operatorname{Im}(\int)$ is an augmentation ideal.*

Note that the augmentation ideal $\mathcal{I}$ corresponding to an integral is in general not a differential ideal of $\mathcal{F}$. We can see this e.g. for $\int^*$ in Example 5, where $\mathcal{I}$ consists of all $f \in \mathcal{F}$ with $f(0) = 0$, so that $\mathcal{I}$ is *not differentially closed* since $(x \mapsto x) \in \mathcal{I}$ but $(x \mapsto 1) \notin \mathcal{I}$.

We have now gathered the main ingredients needed for treating boundary problems, namely integro-differential algebras. Similar structures are introduced under the name *Rota–Baxter algebras* in the recent preprint by Guo and Keigher (2007), which came to our attention only after completing this article. The situation considered there is more general in four respects: The algebras are over unital commutative rings rather than fields, they may be noncommutative, they may have nonzero weight, and they satisfy the pure Baxter axiom (5) rather than the differential version (6). Their interest stems mainly from combinatorial investigations of tree-like structures, where the weight is usually nonzero.

3. Integro-differential operators

From here onwards, let $(\mathcal{F}, \partial, \int)$ be an ordinary integro-differential algebra over a field K with evaluation E. We introduce now an algebra of operators on $\mathcal{F}$ using rewrite systems (Baader and Nipkow, 1998) in the spirit of Bergman (1978). The *integro-differential operators* $\mathcal{F}[\partial, \int]$ are defined as the K-algebra generated by the symbols ∂ and $\int$, the "functions" $f \in \mathcal{F}$ and the multiplicative "functionals" $\varphi \in \mathcal{M}(\mathcal{F})$, modulo the rewrite rules given in Table 1. We will use the variables f, g for elements of $\mathcal{F}$ and the variables φ, ψ, χ for elements of $\mathcal{M}(\mathcal{F})$. Every integro-differential operator can be written as a sum of "monomials", every monomial as a coefficient times a "term".

In the rules of Table 1 as well as in the rest of this paper, we use the notation $U \cdot f$ for the *action* of U on a function f, where U is an element of the free algebra in the above generators. It is an easy matter to check that the rewrite rules of Table 1 are fulfilled in $(\mathcal{F}, \partial, \int)$, so we may regard $\cdot$ as an action of $\mathcal{F}[\partial, \int]$ on $\mathcal{F}$. In particular, $f \cdot g$ now denotes the product of functions $f, g \in \mathcal{F}$.

We remark that Table 1 is to be understood as including *implicit rules* for $\int\int$, $\int\partial$ and $\int\varphi$ by substituting $f = 1$ in the rules for $\int f \int$, $\int f \partial$ and $\int f \varphi$, respectively. Moreover, one obtains the *derived rule* $\mathrm{E}\int = 0$ from the definition of the evaluation E. Note that $\mathcal{F}[\partial]$ is a subalgebra of $\mathcal{F}[\partial, \int]$ with the same induced action on $\mathcal{F}$.

Example 12. The *analytic polynomials* of Rosenkranz (2005, p. 176) are also an important special case of integro-differential operators (the restriction to $K = \mathbb{C}$ imposed there is not essential). They are constructed on top of an analytic algebra $(\mathcal{F}, \partial, \int^*, \int_*)$ with evaluations $f \mapsto f^{\leftarrow}$ and $f \mapsto f^{\rightarrow}$, as explained in Example 10. As usual, we can express one integral using the other, yielding either $-\int_* = (1 - {}^{\rightarrow})\int^*$ or $-\int^* = (1 - {}^{\leftarrow})\int_*$. Choosing randomly the first alternative, we work with the integro-differential algebra $(\mathcal{F}, \partial, \int^*)$. Up to notational details, the analytic polynomials over $(\mathcal{F}, \partial, \int^*, \int_*)$ are then the subalgebra of $\mathcal{F}[\partial, \int^*]$ generated by the operators

$$\begin{aligned} D &= \partial, & L &= {}^{\leftarrow}, \\ A &= \textstyle\int^*, & R &= {}^{\rightarrow}, \\ B &= \textstyle\int_*, = (1 - {}^{\leftarrow})\int^* & \lceil f \rceil &= f, \end{aligned}$$

using the same names as in the cited article. We use also the abbreviation $F = A + B$ for the operator of definite integration.

Table 1
Rewrite rules for integro-differential operators

fg	$\to$	$f \cdot g$	∂f	$\to$	$\partial \cdot f + f\partial$	$\int f \int$	$\to$	$(\int \cdot f)\int - \int(\int \cdot f)$
$\varphi\psi$	$\to$	ψ	$\partial\varphi$	$\to$	0	$\int f \partial$	$\to$	$f - \int(\partial \cdot f) - (\mathtt{E} \cdot f)\,\mathtt{E}$
φf	$\to$	$(\varphi \cdot f)\,\varphi$	$\partial\int$	$\to$	1	$\int f \varphi$	$\to$	$(\int \cdot f)\,\varphi$

Note that for analytic polynomials, the multiplication operators $\lceil f \rceil$ are restricted to *basis elements* $f \in \mathcal{F}$; similar restrictions could be made here. The point is that a system of normal forms on $\mathcal{F}[\partial, \int]$ presupposes a canonical simplifier on the free algebra generated by ∂ and $\int$, the functions $f \in \mathcal{F}$ and the functionals $\varphi \in \mathcal{M}(\mathcal{F})$. Expansion with respect to fixed bases of $\mathcal{F}$ and $\mathcal{M}(\mathcal{F})$ provides such a canonical simplifier, but there may also be others. In Rosenkranz (2005), we have implemented a ground simplifier via such a basis expansion (where $\mathcal{F}$ was given by the exponential polynomials). In the present paper, we take the viewpoint that the free algebra is equipped with *some* canonical simplifier (the "ground simplifier"), and the confluence result of the following proposition has to be understood relative to such a ground simplifier.

Proposition 13. *The rewrite system of Table 1 is Noetherian and confluent.*

Proof. By the Diamond Lemma 1.2 from Bergman (1978), it suffices to ensure the following two facts: First we must construct a partial well-order $>$ on the word monoid in the generators of $\mathcal{F}[\partial, \int]$ such that $>$ is compatible with the monoid structure and the rewrite system in Table 1. Second we have to prove that all ambiguities of the rewrite system are resolvable. For defining the partial well-order, we put $\partial > f$ for all functions f and extend this to words by the graded lexicographic construction. The resulting partial order is clearly well-founded (since it is on the generators) and compatible with the monoid structure (by its grading). It is also compatible with the rewrite system because all rules reduce the word length except for the Leibniz rule, which is compatible because $\partial > f$.

For proving that the ambiguities of Table 1 are resolvable, note first that we have no inclusion ambiguities while there are exactly 14 overlap ambiguities. For overlapping rules $ww_1 \to p_1$ and $w_2 w \to p_2$ to be resolvable, their S-polynomial $p_2 w_1 - w_2 p_1$ must reduce to zero. This is indeed the case, as one can check by an easy calculation (using also the axioms of integro-differential algebras for $\mathcal{F}$). As a representative example, let us reassure ourselves that the S-polynomial from the rules for $ww_1 = \int f \partial$ and $w_2 w = \int g \int$ does indeed reduce to

$$
\begin{aligned}
&(\textstyle\int \cdot g)\int f\partial - \int(\int \cdot g)\, f\partial - \int gf + \int g \int f' + \int g\,(\mathtt{E} \cdot g)\,\mathtt{E} \\
&\quad = (\textstyle\int \cdot g) f - (\int \cdot g)\int f' - (\int \cdot g)\,(\mathtt{E} \cdot f)\,\mathtt{E} - (\int \cdot g) f + \int \partial \cdot ((\int \cdot g) \cdot f) \\
&\qquad + (\mathtt{E} \cdot ((\textstyle\int \cdot g) \cdot f)\,\mathtt{E} - \int(g \cdot f) + (\int \cdot g)\int f' - \int(\int \cdot g) f' + (\mathtt{E} \cdot f)(\int \cdot g)\,\mathtt{E} \\
&\quad = \textstyle\int \partial \cdot ((\int \cdot g) \cdot f) + (\mathtt{E} \cdot ((\int \cdot g) \cdot f)\,\mathtt{E} - \int(g \cdot f) - \int(\int \cdot g) f' \\
&\quad = \textstyle\int(g \cdot f) + \int(\int \cdot g) f' + 0 - \int(g \cdot f) - \int(\int \cdot g) f' \\
&\quad = 0,
\end{aligned}
$$

as it should. □

In other words, the polynomials given by the difference between the left-hand and right-hand sides of Table 1 form a *two-sided noncommutative Gröbner basis*. For the theory of Gröbner bases, we refer the reader to Buchberger (1965, 1970, 1998), for its noncommutative extension to Mora (1986), Mora (1994) and Ufnarovski (1998).

Comparing the analytic polynomials in Rosenkranz (2005, p. 183) with the rewrite system of Table 1, we would like to emphasize the gain in *simplicity and economy*: Despite their higher generality, the integro-differential operators of $\mathcal{F}[\partial, \int]$ require just 9 instead of 36 identities! Consequently, their confluence proof (resolving 14 overlaps) can still be produced by hand, while the automatically generated confluence proof for the analytic polynomials (resolving 233 overlaps) contains 2000 lines; see Rosenkranz (2005, p. 184f) for a small fragment of it.

Having a Noetherian and confluent rewrite system, every integro-differential operator has a *unique normal form* (Baader and Nipkow, 1998, p. 12). In order to describe these normal forms explicitly, it is useful to single out a particular portion of the operators that will also turn out to play a distinguished role in specifying boundary conditions (see Section 5).

Definition 14. The elements of the right ideal

$$\mathcal{S}(\mathcal{F}) = \mathcal{M}(\mathcal{F})\,\mathcal{F}[\partial, \textstyle\int]$$

are called *Stieltjes boundary conditions* over $\mathcal{F}$; if there is no danger of ambiguity, we will henceforth just speak of "boundary conditions".

We will now describe the *normal forms* in $\mathcal{F}[\partial, \int]$, starting with a simple observation on reducibility (in general not describing normal forms), which is afterwards used for characterizing the normal forms of boundary conditions.

Lemma 15. *Every integro-differential operator in $\mathcal{F}[\partial, \int]$ can be reduced to a linear combination of monomials $f\varphi\int g\psi\partial^i$, where $i \geq 0$ and each of $f, \varphi, \int, g, \psi$ may also be absent.*

Proof. Call a monomial consisting only of functions and functionals "algebraic". Using the left column of Table 1, it is immediately clear that all such monomials can be reduced to f or φ or $f\varphi$. Now let w be an arbitrary monomial in the generators of $\mathcal{F}[\partial, \int]$. By using the middle column of Table 1, we may assume that all occurrences of ∂ are moved to the right, so that all monomials have the form $w = w_1 \cdots w_n \partial^i$ with $i \geq 0$ and each of $w_1, \ldots, w_n$ either a function, a functional or $\int$. We may further assume that there is at most one occurrence of $\int$ among the $w_1, \ldots, w_n$. Otherwise the monomials $w_1 \cdots w_n$ contain $\int \tilde{w} \int$, where each $\tilde{w} = f\varphi$ is an algebraic monomial. But then we can reduce

$$\textstyle\int \tilde{w} \int = (\int f\varphi)\int = (\int \cdot\, f)\varphi\int$$

by using the corresponding rule of Table 1. Applying these rules repeatedly, we arrive at algebraic monomials left and right of $\int$ (or just a single algebraic monomial if $\int$ is absent). □

Proposition 16. *Every boundary condition of $\mathcal{S}(\mathcal{F})$ has the normal form*

$$\sum_{\varphi \in \mathcal{M}(\mathcal{F})} \left(\sum_{i \in \mathbb{N}} a_{\varphi,i}\, \varphi\partial^i + \varphi \textstyle\int f_\varphi \right)$$

with $a_{\varphi,i} \in K$ and $f_\varphi \in \mathcal{F}$ almost all zero.

Proof. By Lemma 15, every boundary condition of $\mathcal{S}(\mathcal{F})$ is a linear combination of monomials having the form

$$w = \chi f\varphi \textstyle\int g\psi\partial^i \quad \text{or} \quad w = \chi f\varphi\partial^i \tag{9}$$

where each of f, g, φ, ψ may also be missing. Using the left column of Table 1, the prefix $\chi f\varphi$ can be reduced to a scalar multiple of a functional, so we may as well assume that f and φ are not

present; this finishes the right-hand case of (9). For the remaining case $w = \chi \int g \psi \partial^i$, assume first that ψ is present. Then we have

$$\chi\,(\textstyle\int g\psi) = \chi\,(\int \cdot\, g)\,\psi = (\chi \int \cdot\, g)\,\chi\psi = (\chi \int \cdot\, g)\,\psi,$$

so w is again a scalar multiple of $\psi\partial^i$, and we are done. Finally, assume we have $w = \chi \int g \partial^i$. If $i = 0$, this is already a normal form. Otherwise we obtain

$$w = \chi\,(\textstyle\int g\partial)\,\partial^{i-1} = (\chi \cdot g)\,\chi\partial^{i-1} - \chi \int g'\partial^{i-1} - (\mathtt{E} \cdot g)\,\mathtt{E}\partial^{i-1},$$

where the first and the last summand are in the required normal form, while the middle summand is to be reduced recursively, eventually leading to a middle term in normal form $\pm\chi \int g'\partial^0 = \pm\chi \int g'$. □

The Stieltjes boundary conditions have the additional benefit of allowing a simple description of the normal forms for all integro-differential operators. Just as we obtain the *differential operators* $\mathcal{F}[\partial] \subset \mathcal{F}[\partial, \int]$ with their usual normal forms, we write also $\mathcal{F}[\int] \subset \mathcal{F}[\partial, \int]$ for the subalgebra of *integral operators*, generated by the functions and $\int$ modulo the Baxter rule (uppermost in the right column of Table 1). Using Lemma 15, it is clear that the normal forms of integral operators are linear combinations of $f \int g$ with $f, g \in \mathcal{F}$.

Finally, we write $\mathcal{F}[\mathtt{E}]$ for the left $\mathcal{F}$-submodule generated by $\mathcal{S}(\mathcal{F})$ and call them *Stieltjes boundary operators* (briefly "boundary operators"). Note that $\mathcal{F}[\mathtt{E}]$ includes $\mathcal{S}(\mathcal{F})$ as well as all finite dimensional projectors P along Stieltjes boundary conditions. The latter can all be described as follows: If $u_1, \ldots, u_n \in \mathcal{F}$ and $\beta_1, \ldots, \beta_n \in \mathcal{S}(\mathcal{F})$ are biorthogonal in the sense that $\beta_i(u_j) = \delta_{ij}$, then

$$P = \sum_{i=1}^{n} u_i\,\beta_i, \tag{10}$$

is the projector onto $[u_1, \ldots, u_n]$ along $[\beta_1, \ldots, \beta_n]^\perp$; see for example Köthe (1969, p. 71) and Regensburger and Rosenkranz (in press, Prop. 2). From the representation (10) it is immediately clear that $P \in \mathcal{F}[\mathtt{E}]$. All elements of $\mathcal{F}[\mathtt{E}]$ have the normal form (10), except that the (u_j) need not be biorthogonal to the (β_i).

It turns out now that every monomial of an integro-differential operator is either a *differential operator* or an *integral operator* or a *boundary operator*.

Proposition 17. *Up to ordering the summands, every normal form of $\mathcal{F}[\partial, \int]$ with respect to the rewrite system of Table 1 can be written uniquely as a sum $T + G + B$ having the following normal-form summands: a differential operator $T \in \mathcal{F}[\partial]$, an integral operator $G \in \mathcal{F}[\int]$, and a boundary operator $B \in \mathcal{F}[\mathtt{E}]$.*

Proof. Inspection of Table 1 confirms that all integro-differential operators having the described sum representation $T + G + P$ are indeed in normal form. Let us now prove that every integro-differential operator of $\mathcal{F}[\partial, \int]$ has such a representation. It is sufficient to consider its monomials w. If w starts with a functional, we obtain a boundary condition by Proposition 16; so assume this is not the case. From Lemma 15 we know that

$$w = f\varphi \textstyle\int g\psi\partial^i \quad \text{or} \quad w = f\varphi\partial^i,$$

where each of φ, g, ψ may be absent. But $w \in \mathcal{F}[\mathtt{E}]$ unless φ is absent, so we may actually assume

$$w = f \int g \psi \partial^i \quad \text{or} \quad w = f \partial^i.$$

The right-hand case yields $w \in \mathcal{F}[\partial]$. If ψ is present in the other case, we may reduce $\int g\psi$ to $(\int \cdot\, g)\,\psi$, and we obtain again $w \in \mathcal{F}[\mathtt{E}]$. Hence we are left with $w = f \int g \partial^i$, and we may assume $i > 0$ since otherwise we have $w \in \mathcal{F}[\int]$ immediately. But then we can reduce

$$\begin{aligned} w &= f\,(\textstyle\int g \partial)\,\partial^{i-1} = f\Big(g - \int (\partial \cdot g) - (\mathtt{E} \cdot g)\,\mathtt{E}\Big)\partial^{i-1} \\ &= (fg)\,\partial^{i-1} - f \textstyle\int (\partial \cdot g)\,\partial^{i-1} - (\mathtt{E} \cdot g)\, f\, \mathtt{E}\, \partial^{i-1}, \end{aligned}$$

where the first term is obviously in $\mathcal{F}[\partial]$ and the last one in $\mathcal{F}[\mathtt{E}]$. The middle term may be reduced recursively until the exponent of ∂ has dropped to zero, leading to a term in $\mathcal{F}[\int]$. □

4. Initial value problems

Up to now we have not discussed the existence of solutions for differential equations, except for two particularly simple cases: the homogeneous differential equation $u^{(n)} = 0$ whose solution space is given by $[1, x, \ldots, x^{n-1}]$ as stated in (8), and the inhomogeneous equation $u' = f$ with $\int f$ as particular solution. In order to have some finer control on which differential equations we want to have solutions, we will allow specifying the *coefficients* of the pertinent linear differential operators. (In differential Galois theory, one usually works with differential fields, where one can study extensions in a much more convenient manner. As we have seen above, though, this route is not accessible for us here.)

Definition 18. A differential subalgebra $\mathcal{F}_0 \leq \mathcal{F}$ is called *saturated* for a differential algebra $\mathcal{F}$ if $\dim \operatorname{Ker}(T) = n$ for every monic $T \in \mathcal{F}_0[\partial]$ with $\deg T = n$ and if all nonzero solutions u of $u' = au$, with $a \in \mathcal{F}_0$, are invertible in $\mathcal{F}$. In this context, we call $\mathcal{F}$ the *ground algebra* and $\mathcal{F}_0$ the *coefficient algebra*. If $\mathcal{F}_0$ coincides with $\mathcal{F}$, we simply speak of a saturated integro-differential algebra.

Some remarks on this definition are in order. First of all, we point out that we need $\mathcal{F}_0$ to be differentially closed such that we can multiply within $\mathcal{F}_0[\partial]$, which will be needed for multiplying boundary problems in Section 6. The first condition on solvability ensures that *homogeneous equations* $Tu = 0$ have a fundamental system with the appropriate number of solutions, while the second condition means that *exponentials* behave as usual. Note also that $\mathcal{F}$ is an ordinary differential algebra as soon as it possesses a saturated coefficient algebra.

Not every integro-differential algebra has a saturated coefficient algebra; e.g. the polynomial algebra $(K[x], \partial, \int)$ does not. We do not know any useful criteria for settling this question. However, there are several important *examples* of integro-differential algebras with saturated coefficient algebras:

Example 19. The prototypical example is furnished by $C^\infty[a, b]$ where $[a, b]$ is a finite interval of $\mathbb{R}$. As a coefficient algebra, one may take either $C^\infty[a, b]$ itself or any differential subalgebra like $\mathbb{R}$ or $\mathbb{C}$ or $\mathbb{C}[x]$. Similarly, one may take analytic functions $C^\omega[a, b]$ and its differential subalgebras. Less demanding but practically important, the exponential polynomials, as defined in Rosenkranz (2005, p. 176), can be taken as a ground algebra with $\mathbb{C}$ as a coefficient algebra.

Example 20. For any field K of characteristic 0, the formal power series $K[[z]]$ are a saturated integro-differential algebra, with derivation and integration defined as usual. This may also be inferred from the next example by the isomorphism described there.

Example 21. Let K be an arbitrary field (note that we are explicitly including the case of positive characteristic in this example). Then the algebra $H(K)$ Hurwitz series (Keigher, 1997) over K is defined as the K-vector space of infinite K-sequences with the multiplication defined as

$$(a_n) \cdot (b_n) = \left(\sum_{i=0}^{n} \binom{n}{i} a_i b_{n-i} \right)_n$$

for all $(a_n), (b_n) \in H(K)$. If one introduces derivation and integration through

$$\partial\,(a_0, a_1, a_2, \dots) = (a_1, a_2, \dots),$$
$$\textstyle\int (a_0, a_1, \dots) = (0, a_0, a_1, \dots),$$

the Hurwitz series form an integro-differential algebra $(H(K), \partial, \int)$, as explained by Keigher and Pritchard (2000) and Guo (2002).

Note that as an additive group, $H(K)$ coincides with the formal power series $K[[z]]$, but its multiplicative structure differs: We have an isomorphism

$$\sum_{n=0}^{\infty} a_n\, z^n \;\mapsto\; (n!\, a_n)$$

from $K[[z]]$ to $H(K)$ if and only if K has characteristic zero. The point is that one can integrate every element of $H(K)$, whereas the formal power series z^{p-1} does not have an antiderivative in $K[[z]]$ if K has characteristic p.

Defining the exponential function $\exp = (1, 1, 1, \dots)$, we obtain immediately $\partial \exp = \exp$. One can introduce a composition $f \circ g$ for $f, g \in H(K)$ whenever g has vanishing constant term, and the usual chain rule is satisfied for this composition (Keigher and Pritchard, 2000). Then the first-order homogeneous equation $u' = au$ with $a \in H(K)$ is solved by

$$u = c\, \exp \circ (\textstyle\int a),$$

which is easily seen to be invertible in $H(K)$. By Corollary 4.3 in Keigher and Pritchard (2000), we know also that all monic homogeneous differential equations of order n have an n-dimensional kernel. Hence $H(K)$ is a saturated integro-differential algebra.

Throughout the rest of this paper, we assume that $(\mathcal{F}, \partial, \int)$ is an integro-differential algebra with a saturated coefficient algebra $\mathcal{F}_0$. As before, we write $\mathtt{E}$ for its evaluation. Having integrals, it is natural to expect that we can also solve *inhomogeneous equations*. As we shall see now, it is always possible to find a particular solution, but we can be more specific than that.

We formulate the *initial value problem* for a monic differential operator $T \in \mathcal{F}_0[\partial]$ and character $\eta \in \mathcal{M}(\mathcal{F})$ as follows: Given a forcing function $f \in \mathcal{F}$, find $u \in \mathcal{F}$ such that

$$\begin{aligned} Tu &= f \\ \eta u = \eta u' &= \cdots = \eta u^{(n-1)} = 0, \end{aligned} \tag{11}$$

where $\deg T = n$. Problems of this kind can be solved uniquely.

Proposition 22. *For every monic $T \in \mathcal{F}_0[\partial]$ and $\eta \in \mathcal{M}(\mathcal{F})$, the initial value problem of the form (11) has a unique solution $u \in \mathcal{F}$ for given $f \in \mathcal{F}$.*

Proof. We can use the usual technique of reformulating (11) as a system of linear first-order differential equations with companion matrix $A \in \mathcal{F}_0^{n\times n}$; then we apply the familiar variation-of-constants formula, as described e.g. by Coddington and Levinson (1955, p. 74). To this end, we pick a fundamental system $u_1, \dots, u_n \in \mathcal{F}$ for T and compute the Wronskian matrix

$$W = \begin{pmatrix} u_1 & \dots & u_n \\ u_1' & \dots & u_n' \\ \vdots & \ddots & \vdots \\ u_1^{(n-1)} & \dots & u_n^{(n-1)} \end{pmatrix}.$$

Observe that $d = \det W$ satisfies the first-order differential equation $d' = ad$, where a is the trace of $A \in \mathcal{F}_0$; see for example Exercise 1.14.5 in van der Put and Singer (2003), but note that we do not need a differential field. Since $\mathcal{F}_0$ is saturated for $\mathcal{F}$, the determinant d must be invertible and hence W a regular matrix.

By Proposition 1 and Corollary 11, the operator $\int\!\!\!\!\!-\, = (1-\eta)\int$ is the integral having the evaluation $\eta = 1 - \int\!\!\!\!\!-\,\partial$. We extend the action of the operators $\int\!\!\!\!\!-, \partial, \eta$ componentwise to $\mathcal{F}^n$. Setting now

$$\hat{u} = (W \int\!\!\!\!\!-\, W^{-1})\, \hat{f}$$

with $\hat{f} = (0, \dots, 0, f)^\top \in \mathcal{F}^n$, one may readily check that $\hat{u} \in \mathcal{F}^n$ is a solution of the first-order system $\hat{u}' = A\hat{u} + \hat{f}$ with initial condition $\eta\hat{u} = 0$. Writing u for the first component of $\hat{u}$, we have a solution of (11).

For proving uniqueness, assume u is a solution of (11) for $f = 0$; we must show $u = 0$. We may expand $u = c_1u_1 + \cdots + c_nu_n$ in terms of the fundamental system $u_1, \dots, u_n$ with suitable coefficients $c_1, \dots, c_n \in K$. Then the initial conditions of (11) may be summarized by $\eta(Wc) = 0$ with the coefficient vector $c = (c_1, \dots, c_n)^\top \in K^n$. But $\eta(Wc) = \eta(W)c$ because η is linear, and $\det \eta(W) = \eta(\det W)$ because it is moreover multiplicative. Since $\det W \in \mathcal{F}$ is invertible, this implies that $\eta(W) \in K^{n\times n}$ is regular, so $c = \eta(W)^{-1}0 = 0$ and $u = 0$. □

As mentioned after Example 10, every integro-differential algebra $(\mathcal{F}, \partial, \int)$ comes with a distinguished character: the evaluation $\eta = \mathrm{E}$. Hence we may speak of the initial value problem associated with a monic $T \in \mathcal{F}_0[\partial]$. If $u \in \mathcal{F}$ is the unique solution to such an initial value problem with forcing function f, we obtain an operator $T^\blacklozenge\colon \mathcal{F} \to \mathcal{F}$ with $u = T^\blacklozenge f$, which we shall call the *fundamental right inverse* for T. The notation and terminology are in accordance with Rosenkranz (2005), where the evaluation $\mathrm{E}\colon C^\infty[a,b] \to C^\infty[a,b]$ is given by $u \mapsto u(a)$. We observe also that $T^\blacklozenge$ is a particular case of a Green's operator.

Proposition 23. *For every monic $T \in \mathcal{F}_0[\partial]$, the fundamental right inverse can be realized as an integro-differential operator $T^\blacklozenge \in \mathcal{F}[\partial, \int]$.*

Proof. Inspecting the proof of Proposition 22, one can see that u may in fact be obtained from f by the *operation of an integro-differential operator* from $\mathcal{F}[\partial, \int]$. This holds in particular for the initial value problem with $\eta = \mathrm{E}$. □

5. Boundary problems

The main purpose of $\mathcal{F}[\partial, \int]$ is to provide a unified language for expressing *boundary problems* as well as their *solutions*. As explained in Section 1, a boundary problem of order

n is typically formulated as follows: Given a forcing function $f \in \mathcal{F}$, we have to find $u \in \mathcal{F}$ such that

$$\begin{aligned} Tu &= f, \\ \beta_1 u = \cdots &= \beta_n u = 0, \end{aligned} \tag{12}$$

for a monic differential operator $T \in \mathcal{F}_0[\partial]$ with $\deg T = n$ and boundary conditions $\beta_1, \ldots, \beta_n \in \mathcal{F}^*$. Clearly we have $T \in \mathcal{F}[\partial, \int]$, but also $\beta_1, \ldots, \beta_n \in \mathcal{F}[\partial, \int]$ if we restrict ourselves to the (relatively large) class of Stieltjes boundary conditions (Definition 14). The solution is usually expressed as $u = Gf$, where $G\colon \mathcal{F} \to \mathcal{F}$ is the so-called Green's operator of the boundary problem (12). As we shall see in Theorem 26, the Green's operator G can also be expressed as the action of an element in $\mathcal{F}[\partial, \int]$.

We think of the boundary conditions $\beta_1, \ldots, \beta_n \in \mathcal{F}^*$ of (12) as specifying a *space of admissible functions*

$$\mathcal{A} = \{\beta_1, \ldots, \beta_n\}^\perp \leq \mathcal{F}.$$

Obviously we may replace the boundary conditions $\beta_1, \ldots, \beta_n \in \mathcal{F}^*$ by other boundary conditions $\tilde{\beta}_1, \ldots, \tilde{\beta}_n \in \mathcal{F}^*$ such that $\tilde{\beta}_i = c_{i1}\beta_1 + \cdots + c_{in}\beta_n$ for a regular matrix $(c_{ij}) \in K^{n\times n}$, leading to the same space of admissible functions $\mathcal{A} = \{\tilde{\beta}_1, \ldots, \tilde{\beta}_n\}^\perp$. This means that the admissible functions may be described invariantly as $\mathcal{A} = \mathcal{B}^\perp$ in terms of $\mathcal{B} = [\beta_1, \ldots, \beta_n] = [\tilde{\beta}_1, \ldots, \tilde{\beta}_n]$. Such a finite dimensional subspace $\mathcal{B} \leq \mathcal{F}^*$ will be called a *space of boundary conditions*.

The operators $\ldots^\perp$ on $\mathcal{F}$ and $\mathcal{F}^*$ create an order-reversing *lattice isomorphism* (a fortiori a Galois connection) between the modular lattices of finite codimensional subspaces of $\mathcal{F}$ and finite dimensional subspaces of $\mathcal{F}^*$. Specifically, we have

$$\mathcal{B}^\perp = \{u \in \mathcal{F} \mid \forall_{\beta\in\mathcal{B}}\ \beta(u) = 0\}$$

for the space of functions satisfying the boundary conditions in $\mathcal{B}$ and

$$\mathcal{A}^\perp = \{\beta \in \mathcal{F}^* \mid \forall_{u\in\mathcal{A}}\ \beta(u) = 0\}$$

for the space of boundary conditions satisfied by the functions in $\mathcal{A}$. The lattice isomorphism provides crucial relations for treating boundary problems (Section 6), specifically

$$(\mathcal{B}_1 \cap \mathcal{B}_2)^\perp = \mathcal{B}_1^\perp + \mathcal{B}_2^\perp \quad \text{and} \quad (\mathcal{B}_1 + \mathcal{B}_2)^\perp = \mathcal{B}_1^\perp \cap \mathcal{B}_2^\perp \tag{13}$$

for finite dimensional subspaces $\mathcal{B}_1, \mathcal{B}_2 \leq \mathcal{F}^*$ and

$$\mathcal{K} \dotplus \mathcal{B}^\perp = \mathcal{F} \ \Leftrightarrow\ \mathcal{K}^\perp \dotplus \mathcal{B} = \mathcal{F}^* \tag{14}$$

for finite dimensional subspaces $\mathcal{K} \leq \mathcal{F}$ and finite codimensional subspaces $\mathcal{B} \leq \mathcal{F}^*$. We are thus in a similar situation to in algebraic geometry, where affine varieties correspond to subspaces of $\mathcal{F}$ while radical ideals correspond to subspaces of $\mathcal{F}^*$. (Our forthcoming article Regensburger and Rosenkranz (in press) provides an abstract approach along these lines.)

For our present purposes, however, we are interested in an *algorithmic treatment* of boundary conditions and their associated spaces of admissible functions. As indicated above, this can be achieved by working with Stieltjes boundary conditions—they are wide enough for practical applications while allowing convenient implementation of the operations expressed in the above identities. Our notion of Stieltjes boundary conditions is naturally motivated by the classical setting obtained by setting $\mathcal{F} = C^\infty[a, b]$ in Example 12.

In a traditional boundary problem (Stakgold, 1979, p. 203), one prescribes only a so-called two-point boundary condition

$$\beta u = \sum_{i=0}^{n-1} a_i\, u^{(i)}(a) + b_i\, u^{(i)}(b)$$

with $a_0, \ldots, a_{n-1}, b_0, \ldots, b_{n-1} \in \mathbb{C}$. Obviously, we may view

$$\beta = \sum_{i=0}^{n-1} a_i\, L D^i + b_i\, R D^i$$

as an element of $\mathcal{F}[\partial, \int]$ since $L, R \in \mathcal{M}(\mathcal{F})$. In a general integro-differential algebra $\mathcal{F}$, we define a *point condition* as a linear combination of conditions having the form $\varphi\partial^i$ with $\varphi \in \mathcal{M}(\mathcal{F})$.

In the literature Brown and Krall (1974, 1977), one also considers boundary conditions of the form

$$\beta u = \sum_{i=0}^{n-1} a_i\, u^{(i)}(a) + b_i\, u^{(i)}(b) + \int_a^b f(\xi)\, u(\xi)\, \mathrm{d}\xi$$

under the name "Stieltjes boundary conditions". Here the sum part gives a point condition as before, while the integral kernel $f \in \mathcal{F}$ is used for prescribing an *integral condition*. Note that such boundary conditions are in the normal form described by Proposition 16, which is the reason for the terminology in Definition 14. We call a Stieltjes boundary condition global if $f \neq 0$.

There are at least three *reasons* for considering Stieltjes boundary conditions: First of all, they are interesting in themselves because certain boundary problems are naturally expressed in terms of global side conditions (for example, specifying the heat radiated through the boundary). This in also true for regularizing ill-posed problems and computing their generalized Green's function (Rosenkranz, 2005, p. 191). A second reason for introducing Stieltjes boundary conditions will become manifest in Section 7: Factoring a boundary problem leads to factor problems with global conditions, even for a problem having only point conditions (see Example 28). Finally, a third advantage of Stieltjes boundary conditions is that they have a natural algebraic characterization by Definition 14.

We write $\mathfrak{B}_n$ for the set of all subspaces $\mathcal{B} = [\beta_1, \ldots, \beta_n] \leq \mathcal{F}^*$ generated by n linearly independent Stieltjes boundary conditions $\beta_1, \ldots, \beta_n \in \mathcal{S}(\mathcal{F})$; note that $[] = O$ is the only element of $\mathfrak{B}_0$. Then $\mathfrak{B} = \bigcup_n \mathfrak{B}_n$ is closed under the operation $+$ of constructing the sum of vector spaces, thus yielding an abelian monoid $(\mathfrak{B}, +)$, which we call the *monoid of boundary conditions*. Specifically, the sum of an m-dimensional and an n-dimensional space of boundary conditions gives

$$[\beta_1, \ldots, \beta_m] + [\tilde{\beta}_1, \ldots, \tilde{\beta}_n] = [\beta_1, \ldots, \beta_m, \tilde{\beta}_1, \ldots, \tilde{\beta}_n] = [\gamma_1, \ldots, \gamma_k],$$

with dimension $k \leq m + n$. In order to compute linearly independent boundary conditions $\gamma_1, \ldots, \gamma_k$, we can apply the following evident strategy.

Proposition 24. *There is an algorithm for computing a basis $\beta_1, \ldots, \beta_n \in \mathcal{S}(\mathcal{F})$ for an arbitrary $\mathcal{B} \in \mathfrak{B}$ given by generators $\gamma_1, \ldots, \gamma_m \in \mathcal{S}(\mathcal{F})$.*

Proof. Expand each of $\gamma_1, \ldots, \gamma_m$ in the K-basis of normal-form monomials as given by Proposition 16. Although the number of such basis elements is infinite, the expansions of

$\gamma_1, \dots, \gamma_m$ will only use finitely many of them, say, $m_1, \dots, m_r$. This yields an $m \times r$ matrix (a_{ij}) over K such that $\gamma_i = a_{i1}m_1 + \cdots + a_{ir}m_r$ for all $i \in \{1, \dots, m\}$. Reducing the matrix (a_{ij}) to row echelon and discarding the zero rows leads to the desired K-basis $\beta_1, \dots, \beta_n$ of $\mathcal{B}$. □

Let us write $\mathfrak{D}_n$ for the set of all monic $T \in \mathcal{F}_0[\partial]$ with $\deg T = n$, setting $\mathfrak{D} = \bigcup_n \mathfrak{D}_n$. In this paper, we will only be concerned with boundary problems (12) that are *regular* in the sense that they have a unique solution u for each forcing function f. Below we reformulate the condition of regularity directly in terms of the differential operator and the space of boundary conditions.

Definition 25. A *boundary problem* of order n is a pair $(T, \mathcal{B})$ with $T \in \mathfrak{D}_n$ and $\mathcal{B} \in \mathfrak{B}_n$; it is called *regular* if $\operatorname{Ker}(T) \dotplus \mathcal{B}^\perp = \mathcal{F}$. We write $\mathfrak{P}_n$ for the set of all regular boundary problems of order n, setting $\mathfrak{P} = \bigcup_n \mathfrak{P}_n$.

As explained in Regensburger and Rosenkranz (in press), the requirement of the direct sum is equivalent to $\operatorname{Ker}(T) \cap \mathcal{B}^\perp = O$ and also to $\operatorname{Ker}(T) + \mathcal{B}^\perp = \mathcal{F}$ since we have insisted on $\deg T = \dim \mathcal{B}$ in our current setting. It is moreover equivalent to regularity in the sense discussed above and to the following *algorithmic criterion*: If $u_1, \dots, u_n$ is any basis of $\operatorname{Ker}(T)$ and $\beta_1, \dots, \beta_n$ any basis of $\mathcal{B}$, the problem $(T, \mathcal{B})$ is regular iff

$$\begin{pmatrix} \beta_1(u_1) & \cdots & \beta_1(u_n) \\ \vdots & \ddots & \vdots \\ \beta_n(u_1) & \cdots & \beta_n(u_n) \end{pmatrix} \tag{15}$$

is regular in $K^{n\times n}$. This test may be found in Kamke (1967, p. 184) for the special case of two-point boundary conditions, but it generalizes even to the abstract setting described in Regensburger and Rosenkranz (in press). Since in this paper we consider only regular boundary problems, we will suppress the attribute "regular".

Note that we do not require well-posedness. Following Hadamard, a *well-posed* problem (Engl et al., 1996, p. 86) must be regular as well as stable (meaning that the solution u depends continuously on the data f). Our approach is purely algebraic, so we do not care about stability (which would first of all require a topology on $\mathcal{F}$). For example, the following boundary problem in $\mathcal{F} = C^\infty[0, 1]$ is regular but not well-posed, at least not when in the common setting of the Banach space $(\mathcal{F}, \|\cdot\|_\infty)$: Given f, find u such that $u' - u = f$ and $u''(0) = 0$. In this case, the solution exists and is unique; in fact, it is given by $u(x) = \int_0^x f(\xi)\,\mathrm{d}\xi - (f(0) + f'(0))\,\mathrm{e}^x$, so the Green's operator is $\mathrm{e}^x - \mathrm{e}^x L - \mathrm{e}^x LD$. Incidentally, this example illustrates another unusual feature of our setting—we do not restrict the derivatives in the boundary conditions to orders below the order of the differential equation (even though it will often be reasonable to make such a restriction).

The *Green's operator* G of a boundary problem $(T, \mathcal{B})$ is specified by the two requirements

$$TG = 1 \quad \text{and} \quad \operatorname{Im}(G) = \mathcal{B}^\perp.$$

If $\deg T = n$, the space of boundary conditions $\mathcal{B}$ can be described by n basis elements $\beta_1, \dots, \beta_n$, and we can rewrite this in the traditional form (12). Then the Green's operator G is given by the mapping $f \mapsto u$. Since every boundary problem $(T, \mathcal{B})$ has a unique Green's operator G in this sense, we can introduce the notation $(T, \mathcal{B})^{-1}$ for it.

In Rosenkranz (2005), we have explained how to compute from a fundamental system for T the Green's operator of a two-point boundary problem $(T, \mathcal{B})$ for the analytic algebra $C^\infty[a, b]$

Table 2
Outline for computing Green's operators

Input: $(T, \mathcal{B}) \in \mathfrak{P}$ with bases $\{u_j\}$ of $\operatorname{Ker}(T)$ and $\{\beta_i\}$ of $\mathcal{B}$
Output: $G \in \mathcal{F}[\partial, \int]$ such that $G = (T, \mathcal{B})^{-1}$
Determine $T^{\blacklozenge} \in \mathcal{F}[\partial, \int]$ as in Proposition 23, using $\{u_j\}$
Determine projector $P \in \mathcal{F}[\mathtt{E}]$ as in (16), using $\{u_j\}$ and $\{\beta_i\}$
Compute $G = (1 - P)\, T^{\blacklozenge}$ in $\mathcal{F}[\partial, \int]$

of Example 12. This result generalizes to our present setting; see Table 2 for an *outline of the computation* and Example 33 a sample problem (Green's operator for the left factor).

Theorem 26. *Every boundary problem $(T, \mathcal{B}) \in \mathfrak{P}$ has a Green's operator that can be written as an integro-differential operator $G \in \mathcal{F}[\partial, \int]$.*

Proof. The decomposition method explained in Rosenkranz (2005) is also valid in our case; based on the algebraic generalized inverse (Nashed and Votruba, 1976; Engl and Nashed, 1981), it even carries over to the general setting described in Regensburger and Rosenkranz (in press). Thus we have

$$G = (1 - P)\, T^{\blacklozenge},$$

where P is the projector onto $\operatorname{Ker}(T)$ along $\mathcal{B}^{\perp}$, and $T^{\blacklozenge}$ is the fundamental right inverse of T. From Proposition 23 we know that $T^{\blacklozenge} \in \mathcal{F}[\partial, \int]$. (In fact, we could take any right inverse of T, but $T^{\blacklozenge}$ is a canonical choice.) For computing the projector in the form (10), we choose a fundamental system $u_1, \ldots, u_n$ for T. If $\mathcal{B}$ is given by a basis $\beta_1, \ldots, \beta_n \in \mathcal{S}(\mathcal{F})$, we can change to a new basis $\tilde{\beta}_1, \ldots, \tilde{\beta}_n$ that is biorthogonal to $u_1, \ldots, u_n$ by setting

$$(\tilde{\beta}_1, \ldots, \tilde{\beta}_n)^{\top} = B^{-1}(\beta_1, \ldots, \beta_n)^{\top},$$

where B is the matrix (15). Then

$$P = \sum_{i=1}^{n} u_i \tilde{\beta}_i \in \mathcal{F}[\mathtt{E}] \subseteq \mathcal{F}[\partial, \int] \tag{16}$$

is the desired projector, and we have $G = (1 - P)\, T^{\blacklozenge} \in \mathcal{F}[\partial, \int]$. □

The *factorization method* described in Section 7 provides an alternative approach to computing Green's operators. The crucial point will be that multiplying boundary problems corresponds to composing their Green's operators in reverse order (see Proposition 27). In the case of differential operators with constant coefficients, one can express any Green's operator as a product of first-order Green's operators, which can be described by a simple formula.

6. Multiplying boundary problems

Using actions, a *semi-direct product* may be defined for monoids just as for groups; the resulting structure is again a monoid (Cohn, 1982, p. 277). Unlike for groups, one has to distinguish semi-direct products (for left actions) and reverse semi-direct products (for right actions); see Eilenberg (1976) and also Regensburger and Rosenkranz (in press).

We define a right action as follows. Every integro-differential operator $U \in \mathcal{F}[\partial, \int]$ acts on $\mathfrak{B}$ as

$$\mathcal{B} \cdot U = \{\beta \circ U \mid \beta \in \mathcal{B}\};$$

if $\mathcal{B}$ is generated by n conditions $\gamma_1, \ldots, \gamma_n$, this gives

$$[\gamma_1, \ldots, \gamma_n] \cdot U = [\gamma_1 \circ U, \ldots, \gamma_n \circ U].$$

For a differential operators $T \in \mathcal{F}_0[\partial]$, a basis $\beta_1, \ldots, \beta_n$ of $\mathcal{B}$ is transformed into a basis $\beta_1 \circ T, \ldots, \beta_n \circ T$ of $\mathcal{B} \cdot T$ since T has a right inverse like $T^\blacklozenge$.

The resulting reverse semi-direct product $\mathfrak{D} \ltimes \mathfrak{B} = (\mathfrak{D} \times \mathfrak{B}, \cdot)$ then has the *multiplication* defined by

$$(T_1, \mathcal{B}_1) \cdot (T_2, \mathcal{B}_2) = (T_1 T_2, \mathcal{B}_1 \cdot T_2 + \mathcal{B}_2). \tag{17}$$

The neutral element under this multiplication is given by the degenerate boundary problem $(1, O)$, which is regular by definition. (Written out in the classical notation, this is the following "problem": Given $f \in \mathcal{F}$, find $u \in \mathcal{F}$ such that $u = f$ without further boundary conditions!)

As mentioned after Theorem 26, we can compute Green's operators from the constituent Green's operators in a factorization, and in Section 7 we will present a method for producing such factorizations from a factorization of the differential operator. But of course this presupposes that the product of boundary problems *corresponds to the composition* of their Green's operators in reverse order. Let us write $\mathfrak{G}$ for the monoid generated by all Green's operators for boundary problems in $\mathfrak{P}$.

Proposition 27. *The boundary problems $\mathfrak{P} \subseteq \mathfrak{D} \times \mathfrak{B}$ form a submonoid of $\mathfrak{D} \ltimes_\Phi \mathfrak{B}$, and the transformation $(T, \mathcal{B}) \mapsto (T, \mathcal{B})^{-1}$ is an anti-isomorphism from $\mathfrak{P}$ to $\mathfrak{G}$. In other words, every Green's operator corresponds to exactly one boundary problem, and we have*

$$(\mathcal{P}_1 \mathcal{P}_2)^{-1} = \mathcal{P}_2^{-1} \mathcal{P}_1^{-1}$$

for all $\mathcal{P}_1, \mathcal{P}_2 \in \mathfrak{P}$.

Proof. From the remark above we know already $(1, O) \in \mathfrak{P}$. By the definition (17) of the multiplication, we have

$$(T_1, \mathcal{B}_1)(T_2, \mathcal{B}_2) = (T_1 T_2, \mathcal{B}_1 \cdot T_2 + \mathcal{B}_2).$$

We first prove that the right-hand boundary problem is regular and that its Green's operator is given by $G_2 G_1$. Clearly we have

$$TG = (T_1 T_2)(G_2 G_1) = T_1 (T_2 G_2) G_1 = T_1 G_1 = 1,$$

so $G_2 G_1$ is a section of $T_1 T_2$. Hence $\operatorname{Ker}(T_1 T_2) \dotplus \operatorname{Im}(G_1 G_2) = \mathcal{F}$, and it remains to show

$$\operatorname{Im}(G_2 G_1) = (\mathcal{B}_1 \cdot T_2 + \mathcal{B}_2)^\perp.$$

Consider first $u = G_2 G_1 f$. We have $\beta(u) = 0$ for all $\beta \in \mathcal{B}_2$ since $\operatorname{Im}(G_2) = \mathcal{B}_2^\perp$, and $\beta(T_2 u) = \beta(G_1 f) = 0$ for all $\beta \in \mathcal{B}_1$ since $\operatorname{Im}(G_1) = \mathcal{B}_1^\perp$, so $u \in (\mathcal{B}_1 \cdot T_2 + \mathcal{B}_2)^\perp$. Conversely, assume $u \in (\mathcal{B}_1 \cdot T_2 + \mathcal{B}_2)^\perp$. Then $u \in (\mathcal{B}_1 \cdot T_2)^\perp$ and $u \in \mathcal{B}_2^\perp$ by (13). The latter condition means $u = G_2 v$ for some v, while the former condition implies $v \in \mathcal{B}_1^\perp$; hence $v = G_1 f$ and $u = G_2 G_1 f$ for some f.

Now for the uniqueness of the Green's operators. Consider two boundary problems $(T, \mathcal{B}), (\tilde{T}, \tilde{\mathcal{B}}) \in \mathfrak{P}$ with the same Green's operator G. Then we obtain from $TG = 1$ and $\tilde{T}G = 1$ that $(T - \tilde{T})G = 0$, so $T - \tilde{T}$ vanishes on the infinite dimensional space $\mathrm{Im}(G) \leq \mathcal{F}$. Assume now $T \neq \tilde{T}$ for a contradiction. Then $T - \tilde{T}$ is a nonzero differential operator over a saturated coefficient algebra $\mathcal{F}_0$, so it has a finite dimensional kernel and cannot vanish on all of $\mathrm{Im}(G)$. Hence we have indeed $T = \tilde{T}$. Finally, we have also $\mathcal{B}^\perp = \mathrm{Im}(G) = \tilde{\mathcal{B}}^\perp$ and therefore $\mathcal{B} = \tilde{\mathcal{B}}$. □

Let us carry out a *simple multiplication* in the monoid $(\mathfrak{P}, \cdot)$, working with the analytic polynomials of Example 12 over the ground algebra $\mathcal{F} = C^\infty[0, 1]$.

Example 28. We claim that

$$(D, [F]) \cdot (D, [L]) = (D^2, [L, R]). \tag{18}$$

Indeed, we have $[F] \cdot D = [FD] = [AD + BD] = [(1 - L) + (-1 + R)] = [R - L]$ and $[F] \cdot D + [L] = [L, R]$, so (18) follows. Written in classical notation, we have multiplied the boundary problems

$$\boxed{\begin{array}{l} u' = f \\ \int_0^1 u(\xi)\,\mathrm{d}\xi = 0 \end{array}} \cdot \boxed{\begin{array}{l} u' = f \\ u(0) = 0 \end{array}} = \boxed{\begin{array}{l} u'' = f \\ u(0) = u(1) = 0 \end{array}}.$$

We see at this point that global conditions are necessary for the converse process: If we want to factor the boundary problem (see Section 7) on the right-hand side, we cannot have two-point boundary conditions in the left factor since it is unique (Proposition 31).

7. Factoring boundary problems

In this section we will study how to split boundary problems into smaller ones. In fact, it turns out that *every* factorization of a differential operator can be "lifted" to the level of boundary problems (Theorem 32).

Definition 29. A boundary problem $(T_2, \mathcal{B}_2) \in \mathfrak{P}$ is called a *right factor* of a boundary problem $(T, \mathcal{B}) \in \mathfrak{P}$ if T_2 is a right factor of T and $\mathcal{B}_2$ a subspace of $\mathcal{B}$.

Proposition 30. *Let $(T, \mathcal{B}) \in \mathfrak{P}$ be a boundary problem and $T = T_1 T_2$ a factorization of its differential operator. Then $(T, \mathcal{B})$ has a right factor $(T_2, \mathcal{B}_2) \in \mathfrak{P}$.*

Proof. Set $n = \deg T_1$ and $m = \deg T_2$. Choose a basis

$$u_1, \ldots, u_m, u_{m+1}, \ldots, u_{m+n} \in \mathcal{F}$$

of $\mathrm{Ker}(T)$ such that $u_1, \ldots, u_m$ is a basis of $\mathrm{Ker}(T_2)$, and choose any basis

$$\beta_1, \ldots, \beta_{m+n} \in \mathcal{S}(\mathcal{F})$$

of $\mathcal{B}$. Since $(T, \mathcal{B})$ is a regular problem, the matrix

$$B = \begin{pmatrix} \beta_1(u_1) & \ldots & \beta_1(u_m) & \beta_1(u_{m+1}) & \ldots & \beta_1(u_{m+n}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \beta_{m+n}(u_1) & \ldots & \beta_{m+n}(u_m) & \beta_{m+n}(u_{m+1}) & \ldots & \beta_{m+n}(u_{m+n}) \end{pmatrix}$$

is regular. Hence we may use row operations to obtain a matrix with a regular upper left $m \times m$ block and zeros below. (We could reduce the matrix B to row echelon form, but this is more than we need at this point.) These operations are realized by left-multiplying B with a suitable matrix $P \in \mathrm{GL}(K, m+n)$ such that the upper left is transformed into a regular matrix

$$B_2 = \begin{pmatrix} \tilde{\beta}_1(u_1) & \dots & \tilde{\beta}_1(u_m) \\ \vdots & \ddots & \vdots \\ \tilde{\beta}_m(u_1) & \dots & \tilde{\beta}_m(u_m) \end{pmatrix}$$

with new boundary conditions

$$\tilde{\beta}_i = \sum_{j=1}^{m+n} P_{ij}\beta_j \qquad (i = 1, \dots, m).$$

But the regularity of B_2 means that $(T_2, \mathcal{B}_2) \in \mathfrak{P}$ with $\mathcal{B}_2 = [\tilde{\beta}_1, \dots, \tilde{\beta}_m] \leq \mathcal{B}$. □

A refined analysis of Proposition 30 leads to a full *classification of all right factors* $(T_2, \mathcal{B}_2) \in \mathfrak{P}$ of a given boundary problem $(T, \mathcal{B}) \in \mathfrak{P}$; see Regensburger and Rosenkranz (in press) for the detailed statement and proof in an abstract setting. The bottom line is that there is a bijection between right factors of $(T, \mathcal{B})$ and direct summands of $\mathrm{Ker}(T_2)$ in $\mathrm{Ker}(T)$. In detail, every right factor $(T_2, \mathcal{B}_2)$ corresponds to $\mathcal{L}_2 = \mathcal{B}_2^\perp \cap \mathrm{Ker}(T)$, while every direct summand $\mathcal{L}_2$ corresponds to $(T_2, \mathcal{B}_2)$ with $\mathcal{B}_2 = \mathcal{B} \cap \mathcal{L}_2^\perp$. One can also show that $(T_2, \mathcal{B}_2)$ is regular iff

$$\mathrm{Ker}(T_2)^\perp \cap \mathcal{B} \dotplus \mathcal{B}_2 = \mathcal{B}, \tag{19}$$

using the preservation of direct sums (14).

When referring to $\mathcal{P}_2 = (T_2, \mathcal{B}_2)$ as a right factor of $\mathcal{P} = (T, \mathcal{B})$, we are actually anticipating that there is also a *left factor* $\mathcal{P}_1 = (T_1, \mathcal{B}_1)$ such that their product yields $\mathcal{P}$. This is indeed the case, as we will see in Proposition 31. But what is immediately clear is that if $\mathcal{P}_1$ exists, it is uniquely determined by $\mathcal{P}$ alone. Indeed, we know from Proposition 27 that $G = G_2 G_1$, where G, G_1, G_2 denote the Green's operators respectively of $\mathcal{P}, \mathcal{P}_1, \mathcal{P}_2$. But this implies $G_1 = T_2 G$ and hence $\mathcal{B}_1 = \mathrm{Im}(T_2 G)^\perp$.

Apart from the existence question, the disturbing feature of the relation $\mathcal{B}_1 = \mathrm{Im}(T_2 G)^\perp$ is that it presupposes knowledge of the Green's operator G. This defeats the plan of using factorization for determining the Green's operator from those of its factors. The next proposition remedies this flaw: it turns out that all we need is an *arbitrary right inverse* H_2 of the differential operator T_2. Of course we take $H_2 = G_2$, but this still needs the computation of a Green's operator (albeit of a smaller size). A more reasonable choice is $H_2 = T_2^\blacklozenge$, thus reducing the task of computing Green's operators to initial value problems. (The fundamental right inverse is a canonical choice here, but in specific settings it may be algorithmically advantageous to choose other right inverses of T_2.)

Proposition 31. *Given $(T, \mathcal{B}) \in \mathfrak{P}$ with $T = T_1 T_2$, there is a unique $(T_1, \mathcal{B}_1) \in \mathfrak{P}$ such that every right factor $(T_2, \mathcal{B}_2) \in \mathfrak{P}$ of $(T, \mathcal{B})$ satisfies $(T, \mathcal{B}) = (T_1, \mathcal{B}_1) \cdot (T_2, \mathcal{B}_2)$. Moreover, we have*

$$\mathcal{B}_1 = (\mathrm{Ker}(T_2)^\perp \cap \mathcal{B}) \cdot H_2,$$

where H_2 is any right inverse of T_2 and $\mathcal{B}_1 = \mathcal{B} \cdot G_2$ where G_2 is the Green's operator of any right factor $(T_2, \mathcal{B}_2)$.

Proof. We have already seen that if $(T_1, \mathcal{B}_1)$ exists, it is unique with $\mathcal{B}_1 = \mathrm{Im}(T_2G)^\perp$. Since T_2G is a right inverse of T_1, we have also $\mathrm{Ker}(T_1) \dotplus \mathrm{Im}(T_2G) = \mathcal{F}$. But this means $(T_1, \mathcal{B}_1) \in \mathfrak{P}$ if we can just ensure that $\mathcal{B}_1$ has a basis of Stieltjes boundary conditions. And this follows immediately once we have proved that

$$\mathrm{Im}(T_2G)^\perp = (\mathrm{Ker}(T_2)^\perp \cap \mathcal{B}) \cdot H_2 \tag{20}$$

since when $\mathcal{B}$ is generated by Stieltjes boundary conditions, its intersection with $\mathrm{Ker}(T_2)^\perp$ is generated by certain linear combinations of them, while right-multiplication by H_2 still yields Stieltjes boundary conditions by the definition of $\mathcal{S}(\mathcal{F})$.

For proving (20), assume first $\beta(T_2Gu) = 0$ for all $u \in \mathcal{F}$. Setting $\tilde{\beta} = \beta \circ T_2$, we have $\beta = \tilde{\beta} \circ H_2$, and it suffices to show $\tilde{\beta} \in \mathrm{Ker}(T_2)^\perp$ and $\tilde{\beta} \in \mathcal{B} = \mathrm{Im}(G)^\perp$. But the former is immediate from the definition of $\tilde{\beta}$, and the latter follows since $\tilde{\beta}(Gu) = \beta(T_2Gu) = 0$ by hypothesis. Conversely, let us assume $\tilde{\beta} \in \mathrm{Ker}(T_2)^\perp \cap \mathcal{B}$ and show $\tilde{\beta} \circ H_2 \in \mathrm{Im}(T_2G)^\perp$. Indeed, we have

$$(\tilde{\beta} \circ H_2)(T_2Gu) = \tilde{\beta}(H_2T_2Gu) = \tilde{\beta}\big(Gu\big) - \tilde{\beta}\big((1 - H_2T_2)\, Gu\big) = 0$$

because the left summand vanishes by the hypothesis $\tilde{\beta} \in \mathcal{B} = \mathrm{Im}(G)^\perp$ and the right summand by the hypothesis $\tilde{\beta} \in \mathrm{Ker}(T_2)^\perp$ and the fact that $1 - H_2T_2$ is a projector onto $\mathrm{Ker}(T_2)$.

Next let us prove the product $(T, \mathcal{B}) = (T_1, \mathcal{B}_1) \cdot (T_2, \mathcal{B}_2)$. Using (20), it suffices to ensure the relation

$$(\mathrm{Ker}(T_2)^\perp \cap \mathcal{B}) \cdot H_2T_2 = \mathrm{Ker}(T_2)^\perp \cap \mathcal{B} \tag{21}$$

since the regularity of $(T_2, \mathcal{B}_2)$ is equivalent to $\mathrm{Ker}(T_2)^\perp \cap \mathcal{B} \dotplus \mathcal{B}_2 = \mathcal{B}$ by (19). For proving (21), we apply the stronger result that $\beta \mapsto \beta \circ H_2T_2$ leaves $\mathrm{Ker}(T_2)^\perp \cap \mathcal{B}$ pointwise invariant, which follows from the fact that $1 - H_2T_2$ is a projector onto $\mathrm{Ker}(T_2)$.

Finally, we prove $\mathcal{B}_1 = \mathcal{B} \cdot G_2$. Substituting G_2 for H_2 in the generic representation of $\mathcal{B}_1$, we show

$$\mathcal{B} \cdot G_2 = (\mathrm{Ker}(T_2)^\perp \cap \mathcal{B}) \cdot G_2.$$

Since $(T_2, \mathcal{B}_2)$ is regular, we can substitute $\mathrm{Ker}(T_2)^\perp \cap \mathcal{B} \dotplus \mathcal{B}_2$ for $\mathcal{B}$ in (19) in the left-hand side, and it remains to show that $\mathcal{B}_2 \cdot G_2 = 0$. But this follows from $\mathrm{Im}(G_2) = \mathcal{B}_2^\perp$. □

The constructive method for computing $\mathcal{B}_1 = (\mathrm{Ker}(T_2)^\perp \cap \mathcal{B}) \cdot H_2$ is the same as in the proof of Proposition 30. Using the row-operation matrix $P \in \mathrm{GL}(K, m + n)$ constructed there (the original version creating zeros only in the lower left block), we compute the new boundary conditions

$$\tilde{\beta}_i = \sum_{j=1}^{m+n} P_{ij}\beta_j \quad (i = m + 1, \ldots, m + n)$$

to obtain a basis $\tilde{\beta}_{m+1} \circ H_2, \ldots, \tilde{\beta}_{m+n} \circ H_2$ of $\mathcal{B}_1$.

Putting together Proposition 30 and Proposition 31, we have now established the following *Factorization Theorem for Boundary Problems*.

Theorem 32. *Given a boundary problem $(T, \mathcal{B}) \in \mathfrak{P}$, every factorization $T = T_1T_2$ of the differential operator can be lifted to a factorization $(T, \mathcal{B}) = (T_1, \mathcal{B}_1) \cdot (T_2, \mathcal{B}_2)$ of the boundary problem with $(T_1, \mathcal{B}_1), (T_2, \mathcal{B}_2) \in \mathfrak{P}$ and $\mathcal{B}_2 \leq \mathcal{B}$.*

We conclude this section with an example of a fourth-order boundary problem arising in mechanics; see Kamke (1967, p. 525).

Example 33. Using the language of analytic polynomials (see Example 12), we consider the boundary problem $\mathcal{P} = (D^4 + 4, [L, R, LD, RD])$, in traditional formulation

$$\boxed{\begin{aligned} &u'''' + 4u = f, \\ &u(0) = u(1) = u'(0) = u'(1) = 0. \end{aligned}}$$

We employ the natural factorization $D^4 + 4 = (D^2 - 2i)(D^2 + 2i)$. Using the basis functions $u_{\pm\pm} = \mathrm{e}^{\pm 1 \pm i}$ for the kernel of $D^4 + 4$, we choose the boundary conditions for the right factor $D^2 + 2i$ in such a way that its Green's operator G_2 has a convenient formulation (this is not necessary in principle but keeps expressions shorter). By the generic second-order formula from Stakgold (1979, p. 195), also derived in Rosenkranz (2005, p. 196), we are led to the right factor $\mathcal{P}_2 = (D^2 + 2i, [(i-1)L - LD, (1-i)R - RD])$ or

$$\boxed{\begin{aligned} &u'' + 2i\,u = f, \\ &(i-1)\,u(0) - u'(0) = (1-i)\,u(1) - u'(1) = 0 \end{aligned}}$$

in traditional formulation.

Boundary problem $\mathcal{P}_2$ can now be solved easily by the generic second-order formula. Alternatively, one could also apply the algorithm from Table 2 or a factorization into first-order problems as explained at the end of Section 5. In any case, one arrives at the Green's operator

$$G_2 = \frac{1+i}{4}\left(\lceil u_{+-}\rceil A \lceil u_{-+}\rceil + \lceil u_{-+}\rceil B \lceil u_{+-}\rceil\right),$$

acting on a function $f \in C^\infty[0,1]$ according to

$$G_2 f(x) = \frac{1+i}{4}\left(\int_0^x \mathrm{e}^{(1-i)(x-\xi)} f(\xi)\,\mathrm{d}\xi + \int_x^1 \mathrm{e}^{(i-1)(x-\xi)} f(\xi)\,\mathrm{d}\xi\right).$$

We use the Green's operator G_2 of boundary problem $\mathcal{P}_2$ for determining the boundary conditions of the (unique!) left factor $\mathcal{P}_1$ in the factorization $\mathcal{P} = \mathcal{P}_1\mathcal{P}_2$ according to Proposition 31. One may easily verify that $\mathcal{P}_1 = (D^2 - 2i, [F\lceil u_{+-}\rceil, F\lceil u_{-+}\rceil])$ or

$$\boxed{\begin{aligned} &u'' - 2i\,u = f \\ &\textstyle\int_0^1 \mathrm{e}^{(1-i)\xi} f(\xi)\,\mathrm{d}\xi = \int_0^1 \mathrm{e}^{(i-1)\xi} f(\xi)\,\mathrm{d}\xi = 0 \end{aligned}} \tag{22}$$

in traditional formulation.

Since this is not a two-point boundary problem, let us go through the algorithm of Table 2 in detail. The first step is to determine the fundamental right inverse of $D^2 - 2i$. A straightforward computation yields

$$H_1 = \frac{i-1}{4}\left(\lceil u_{--}\rceil\, A\, \lceil u_{++}\rceil - \lceil u_{++}\rceil\, A\, \lceil u_{--}\rceil\right).$$

Next we compute the projector P onto $\mathrm{Ker}(D^2 - 2i)$ along $[F\lceil u_{+-}\rceil, F\lceil u_{-+}\rceil]^\perp$. Using the representation (16), we compute a basis $(\hat{u}_{+-}, \hat{u}_{-+})$ biorthogonal to $(F\lceil u_{+-}\rceil, F\lceil u_{-+}\rceil)$, obtaining $P = \lceil \hat{u}_{+-}\rceil\, F\, \lceil u_{+-}\rceil + \lceil \hat{u}_{-+}\rceil\, F\, \lceil u_{-+}\rceil$. Carrying out the computation (which involves four definite integrals and inverting a 2×2 matrix) leads to

Table 3
Coefficients for G_1

	u_{--}	u_{-+}	u_{+-}	u_{++}
a_{--}	$(1+i)(\mathrm{e}^2-\mathrm{e}^{2i})$	$2i(1-\mathrm{e}^2)$	$2(\mathrm{e}^{2i}-1)$	$(1-i)(2-\mathrm{e}^2-\mathrm{e}^{2i})$
b_{--}	$(1+i)(\mathrm{e}^2-\mathrm{e}^{2i})$	$2i(1-\mathrm{e}^2)$	$2(\mathrm{e}^{2i}-1)$	$(1-i)(\mathrm{e}^{-2}+\mathrm{e}^{-2i}-2)$
a_{++}	$(1-i)(\mathrm{e}^{-2}+\mathrm{e}^{-2i}-2)$	$2(1-\mathrm{e}^{-2i})$	$2i(\mathrm{e}^{-2}-1)$	$(1+i)(\mathrm{e}^{-2i}-\mathrm{e}^{-2})$
b_{++}	$(1-i)(2-\mathrm{e}^{2i}-\mathrm{e}^2)$	$2(1-\mathrm{e}^{-2i})$	$2i(\mathrm{e}^{-2}-1)$	$(1+i)(\mathrm{e}^{-2i}-\mathrm{e}^{-2})$

Table 4
Coefficients for G

	u_{--}	u_{-+}	u_{+-}	u_{++}
a_{--}	$i(\mathrm{e}^{2i}-\mathrm{e}^2)$	$(1-i)(1-\mathrm{e}^2)$	$(1+i)(1-\mathrm{e}^{2i})$	$\mathrm{e}^2+\mathrm{e}^{2i}-2$
b_{--}	$i(\mathrm{e}^{2i}-\mathrm{e}^2)$	$(1-i)(1-\mathrm{e}^2)$	$(1+i)(1-\mathrm{e}^{2i})$	$2-\mathrm{e}^{-2}-\mathrm{e}^{-2i}$
a_{-+}	$(1-i)(1-\mathrm{e}^2)$	$\mathrm{e}^2-\mathrm{e}^{-2i}$	$i(2-\mathrm{e}^2-\mathrm{e}^{-2i})$	$(1+i)(\mathrm{e}^{-2i}-1)$
b_{-+}	$(1-i)(1-\mathrm{e}^2)$	$\mathrm{e}^2-\mathrm{e}^{-2i}$	$i(\mathrm{e}^{-2}+\mathrm{e}^{2i}-2)$	$(1+i)(\mathrm{e}^{-2i}-1)$
a_{+-}	$(1+i)(1-\mathrm{e}^{2i})$	$i(\mathrm{e}^{-2}+\mathrm{e}^{2i}-2)$	$\mathrm{e}^{2i}-\mathrm{e}^{-2}$	$(1-i)(\mathrm{e}^{-2}-1)$
b_{+-}	$(1+i)(1-\mathrm{e}^{2i})$	$i(2-\mathrm{e}^2-\mathrm{e}^{-2i})$	$\mathrm{e}^{2i}-\mathrm{e}^{-2}$	$(1-i)(\mathrm{e}^{-2}-1)$
a_{++}	$2-\mathrm{e}^{-2}-\mathrm{e}^{-2i}$	$(1+i)(\mathrm{e}^{-2i}-1)$	$(1-i)(\mathrm{e}^{-2}-1)$	$i(\mathrm{e}^{-2}-\mathrm{e}^{-2i})$
b_{++}	$\mathrm{e}^2+\mathrm{e}^{2i}-2$	$(1+i)(\mathrm{e}^{-2i}-1)$	$(1-i)(\mathrm{e}^{-2}-1)$	$i(\mathrm{e}^{-2}-\mathrm{e}^{-2i})$

$$\hat{u}_{+-} = \frac{(\mathrm{e}^2-1)\,u_{--} - (\mathrm{e}^{-2i}-1)i\,u_{++}}{\Delta},$$
$$\hat{u}_{-+} = \frac{(\mathrm{e}^{2i}-1)i\,u_{--} - (\mathrm{e}^{-2}-1)\,u_{++}}{\Delta},$$

where $\Delta = \cos 2 + \cosh 2 - 2$. Then we compute the Green's operator of boundary problem $\mathcal{P}_1$ as $G_1 = (1-P)\,H_1$. Using the normalization engine for analytic polynomials described in Rosenkranz (2005), we arrive at

$$G_1 = \frac{1}{8\Delta}\Big(\lceil u_{--}\rceil\, A\, \lceil a_{--}\rceil + \lceil u_{--}\rceil\, B\, \lceil b_{--}\rceil + \lceil u_{++}\rceil\, A\, \lceil a_{++}\rceil + \lceil u_{++}\rceil\, B\, \lceil b_{++}\rceil\Big),$$

where each of $a_{--}, b_{--}, a_{++}, b_{++}$ is a linear combination $u_{--}, u_{-+}, u_{+-}, u_{++}$ as indicated in Table 3.

According to Proposition 27, the Green's operator G of the full boundary problem $\mathcal{P}$ is given by G_2G_1. Its explicit form, obtained by noncommutative multiplication and subsequent normalization, is given here for reference; often one might prefer the factored representation in terms of G_2 and G_1. We have

$$G = \frac{1+i}{32\Delta}\Big(\lceil u_{--}\rceil\, A\, \lceil a_{--}\rceil + \cdots + \lceil u_{++}\rceil\, B\, \lceil b_{++}\rceil\Big),$$

similar to G_1 in structure, but now with four additional summands coming from u_{-+} and u_{+-}. The eight functions $a_{--}, \ldots, a_{++}$ are again linear combinations of the type before, with coefficients given in Table 4.

8. Conclusion

Factoring a differential equation reduces the order and thus aids in solving the given equation. Since differential equations usually come together with boundary conditions, they must be incorporated in an additional step (typically viewed as external to differential algebra). The theory presented in this paper extends the factorization techniques for linear ordinary differential equations in such a way that the boundary conditions become an integral part, leading to an algorithmic machinery for *factoring and solving boundary problems* over integro-differential algebras. The implementation of these algorithms will be described in a subsequent paper.

Let us now discuss some possibilities for *extending* our approach into various directions: partial differential equations, systems of linear ordinary differential equations, difference equations, polynomial boundary conditions, semilinear boundary problems, dual pairings and duality theory, analytical aspects, and localization.

In this paper, we have restricted ourselves to ordinary differential equations (and thus to ordinary integro-differential algebras in the sense of Definition 8). This is convenient since – relative to given fundamental systems – it allows us to compute Green's operators in closed form. But the concept of multiplying (and hence factoring) boundary problems, as defined in (17), may be transferred to a *more general setting* that allows for infinitely many "boundary conditions"; see Regensburger and Rosenkranz (in press).

It can in particular be applied to *linear partial differential equations*, where one can exploit suitable results about factoring linear partial differential operators (Grigoriev and Schwarz, 2007, 2005, 2004; Tsarev, 1998). As a prototype (Regensburger and Rosenkranz, in press), we have factored the one-dimensional inhomogeneous wave equation on a bounded interval into two first-order "boundary problems". Along these lines, we plan to develop symbolic algorithms for first-order partial differential equations (typical factor problems) in non-trivial geometries. Since factorization will normally end up with (symbolically) irreducible boundary problems, it becomes more important to address stability issues: Well-posed boundary problems should be factored into well-posed blocks (Engl et al., 1996), if possible.

Going into a different direction, one can also apply our methodology of multiplying and factoring boundary problems to *systems* of linear ordinary differential equations. We expect that the solution theory (now using "Green's matrices" instead of Green's functions) as well as the algorithms will essentially carry over to this setting.

Everything considered in this paper was directed towards the continuous case of linear differential equations, but we expect the discrete case of linear *difference equations* to be tractable in principle by the same methods, except for the well-known complications arising from a skew Leibniz rule and a Baxter axiom with weight unity instead of zero; see Example 1.6 in Guo (2002). As pointed out in Section 2, the concept of integro-differential algebras generalizes naturally to this situation (Guo and Keigher, 2007).

By contrast, the restriction to linear differential equations seems to be quite rigid: we do not see how to translate our ideas to nonlinear differential equations. What could be considered, though, is the case of linear differential equations with *polynomial boundary conditions*, a case that is also of interest in applications. (A classical example is given by the heat equation with radiation on the boundary, described by the Stefan–Boltzmann law: The normal derivative of the temperature is proportional to its fourth power.) Although the solution operator of such a problem is necessarily nonlinear, we hope that one can adapt some of our ideas by handling the boundary conditions through ideals instead of linear subspaces.

In this article, we have worked with the (algebraic) dual of the vector space structure of the underlying differential algebra. We think that our approach could in principle be transferred to a setting with a *dual pairing* instead of the canonical bilinear form; this would include important topological vector spaces like C^k and L^p. Of course, this requires a modification of the composition structure, leading to a category rather than a monoid of boundary problems as pointed out in Regensburger and Rosenkranz (in press). The advantage might be that one gains topological insights relating various operators (like the differential and Green's operators) and spaces (like images and kernels).

Speaking of duality, one should also mention that the usual *duality theory* of linear boundary problems (Coddington and Levinson, 1955, Chapter 11) can be transferred to "classical" Stieltjes boundary conditions (on real- or complex-valued functions); see for example Brown (1975). The idea is that every boundary problem should have a dual or "adjoint" problem whose solution operator is the "transpose" of the original problem. The adjoint problem is often useful for characterizing certain aspects of a given primal problem (e.g. the solvability criterion for the Fredholm alternative).

We have not yet exploited the factored representation of Green's operators for *characterizing Green's functions* (possibly restricted to the well-posed case to avoid distributions). This may be done from two different perspectives: From an algebraic viewpoint, one might proceed in a manner similar to the Galois theory of linear ordinary differential equations; from an analytic viewpoint, the singular value decomposition would be of interest.

Finally, we mention that we have also treated singular boundary problems, where one needs a modified Green's function/operator as in the example from Section 3.5 in Rosenkranz (2005). This leads to a *localization* in the algebra of Green's operators—differential operators appear as the "reciprocals" of suitable integral operators. In this manner, one obtains a noncommutative generalization of the Mikusiński calculus that allows a symbolic treatment of boundary problems just like the ordinary Mikusiński calculus does for initial value problems (Mikusiński, 1959). These ideas will be discussed in a future paper.

Acknowledgements

We would like to thank our project leaders Bruno Buchberger and Heinz W. Engl for their continuous support, critical comments and helpful suggestions. Moreover, we want to express our gratitude to the referees, in particular the first referee, for their thorough reading and for many valuable suggestions.

References

Baader, F., Nipkow, T., 1998. Term Rewriting and All That. Cambridge University Press, Cambridge.

Baxter, G., 1960. An analytic problem whose solution follows from a simple algebraic identity. Pacific J. Math. 10, 731–742.

Bergman, G.M., 1978. The diamond lemma for ring theory. Adv. Math. 29 (2), 179–218.

Brown, R.C., 1975. Duality theory for *n*th order differential operators under Stieltjes boundary conditions. SIAM J. Math. Anal. 6 (5), 882–900.

Brown, R.C., Krall, A.M., 1974. Ordinary differential operators under Stieltjes boundary conditions. Trans. Amer. Math. Soc. 198, 73–92.

Brown, R.C., Krall, A.M., 1977. *n*-th order ordinary differential systems under Stieltjes boundary conditions. Czechoslovak Math. J. 27 (1), 119–131.

Buchberger, B., 1965, An algorithm for finding the bases elements of the residue class ring modulo a zero dimensional polynomial ideal (German). Ph.D. Thesis. Univ. of Innsbruck. English translation published in J. Symbolic Comput., 41 (3–4): 475–511, 2006.

Buchberger, B., 1970. Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems. Aequationes Math. 4, 374–383. English translation: An algorithmical criterion for the solvability of a system of algebraic equations. In Buchberger and Winkler (1998).

Buchberger, B., 1998. Introduction to Gröbner Bases. pp. 3–31, In Buchberger and Winkler (1998).

Buchberger, B., Winkler, F. (Eds.), 1998. Gröbner bases and applications. London Mathematical Society Lecture Note Series, vol. 251. Cambridge University Press, Cambridge. Papers from the Conference on 33 Years of Gröbner Bases held at the University of Linz, Linz, February 2–4, 1998.

Coddington, E.A., Levinson, N., 1955. Theory of Ordinary Differential Equations. McGraw-Hill Book Company, Inc., New York, Toronto, London.

Cohn, P.M., 1982. Algebra, 2nd ed. vol. 1. John Wiley & Sons, Chichester.

Cohn, P.M., 2003. Basic Algebra: Groups, Rings and Fields. Springer, London.

Eilenberg, S., 1976. Automata, languages, and machines (Volume B). In: Pure and Applied Mathematics, vol. 59. Academic Press, New York.

Engl, H.W., Hanke, M., Neubauer, A., 1996. Regularization of inverse problems. In: Mathematics and its Applications, vol. 375. Kluwer Academic Publishers Group, Dordrecht.

Engl, H.W., Nashed, M.Z., 1981. New extremal characterizations of generalized inverses of linear operators. J. Math. Anal. Appl. 82 (2), 566–586.

Grigoriev, D., Schwarz, F., 2004. Factoring and solving linear partial differential equations. Computing 73 (2), 179–197.

Grigoriev, D., Schwarz, F., 2005. Generalized Loewy-decomposition of D-modules. In: Kauers, M. (Ed.), ISSAC '05: Proceedings of the 2005 International Symposium on Symbolic and Algebraic Computation. ACM Press, New York, NY, USA, pp. 163–170.

Grigoriev, D., Schwarz, F., 2007. Loewy- and primary decompositions of D-modules. Adv. in Appl. Math. 38, 526–541.

Grigoriev, D.Y., 1990. Complexity of factoring and calculating the GCD of linear ordinary differential operators. J. Symbolic Comput. 10 (1), 7–37.

Guo, L., 2002. Baxter algebras and differential algebras. In: Differential Algebra and Related Topics (Newark, NJ, 2000). World Sci. Publ., River Edge, NJ, pp. 281–305.

Guo, L., Keigher, W., 2007, On differential Rota–Baxter algebras, arXiv:math/0703780v1 [math.RA].

Kamke, E., 1967. Differentialgleichungen. Lösungsmethoden und Lösungen. Teil I: Gewöhnliche Differentialgleichungen, 8th ed. In: Mathematik und ihre Anwendungen in Physik und Technik A, vol. 18. Akademische Verlagsgesellschaft, Leipzig.

Kaplansky, I., 1957. An Introduction to Differential Algebra. Actualités Sci. Ind., No. 1251 = Publ. Inst. Math. Univ. Nancago, No. 5. Hermann, Paris.

Keigher, W.F., 1997. On the ring of Hurwitz series. Comm. Algebra 25 (6), 1845–1859.

Keigher, W.F., Pritchard, F.L., 2000. Hurwitz series as formal functions. J. Pure Appl. Algebra 146 (3), 291–304.

Kolchin, E., 1973. Differential algebra and algebraic groups. In: Pure and Applied Mathematics, vol. 54. Academic Press, New York, London.

Köthe, G., 1969. Topological vector spaces (Volume I). In: Die Grundlehren der mathematischen Wissenschaften, vol. 159. Springer, New York.

Mikusiński, J., 1959. Operational calculus. In: International Series of Monographs on Pure and Applied Mathematics, vol.8. Pergamon Press, New York.

Mora, F., 1986. Groebner bases for non-commutative polynomial rings. In: AAECC-3: Proceedings of the 3rd International Conference on Algebraic Algorithms and Error-Correcting Codes. Springer-Verlag, London, UK, pp. 353–362.

Mora, T., 1994. An introduction to commutative and noncommutative Gröbner bases. Theoret. Comput. Sci. 134 (1), 131–173. Second International Colloquium on Words, Languages and Combinatorics (Kyoto, 1992).

Nashed, M.Z., Votruba, G.F., 1976. A unified operator theory of generalized inverses. In: Nashed, M.Z. (Ed.), Generalized Inverses and Applications (Proc. Sem., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1973). Academic Press, New York, pp. 1–109.

Pritchard, F.L., Sit, W.Y., 2007. On initial value problems for ordinary differential–algebraic equations. In: Rosenkranz, M., Wang, D. (Eds.), Gröbner Bases in Symbolic Analysis. Proceedings of the Special Semester on Gröbner Bases and Related Methods. In: Radon Series Comp. Appl. Math., vol. 2. Walter de Gruyter, Berlin, pp. 283–340.

Regensburger, G., Rosenkranz, M., 2007, An algebraic foundation for factoring linear boundary problems. Ann. Mat. Pura Appl. (4) (in press).

Robinson, A., 1961. Local differential algebra. Trans. Amer. Math. Soc. 97, 427–456.

Rosenkranz, M., 2005. A new symbolic method for solving linear two-point boundary value problems on the level of operators. J. Symbolic Comput. 39 (2), 171–199.

Rosenkranz, M., Buchberger, B., Engl, H.W., 2003. Solving linear boundary value problems via non-commutative Gröbner bases. Appl. Anal. 82, 655–675.

Rota, G.-C., 1969. Baxter algebras and combinatorial identities (I, II). Bull. Amer. Math. Soc. 75, 325–334.

Schwarz, F., 1989. A factorization algorithm for linear ordinary differential equations. In: ISSAC '89: Proceedings of the ACM-SIGSAM 1989 International Symposium on Symbolic and Algebraic Computation. ACM Press, New York, NY, USA, pp. 17–25.

Stakgold, I., 1979. Green's Functions and Boundary Value Problems. John Wiley & Sons, New York.

Tsarev, S.P., 1996. An algorithm for complete enumeration of all factorizations of a linear ordinary differential operator. In: ISSAC '96: Proceedings of the 1996 International Symposium on Symbolic and Algebraic Computation. ACM Press, New York, NY, USA, pp. 226–231.

Tsarev, S.P., 1998. Factorization of linear partial differential operators and Darboux integrability of nonlinear PDEs. SIGSAM Bull. 32 (4), 21–28.

Ufnarovski, V., 1998. Introduction to Noncommutative Gröbner Bases Theory. pp. 259–280, In Buchberger and Winkler (1998).

van der Put, M., Singer, M.F., 2003. Galois theory of linear differential equations. In: Grundlehren der Mathematischen Wissenschaften, vol. 328. Springer, Berlin.

Integro-Differential Polynomials and Operators

Markus Rosenkranz and Georg Regensburger *
Johann Radon Institute for Computational and Applied Mathematics (RICAM)
Austrian Academy of Sciences
Altenbergerstraße 69
A-4040 Linz, Austria
{markus.rosenkranz,georg.regensburger}@oeaw.ac.at

ABSTRACT

We propose two algebraic structures for treating integral operators in conjunction with derivations: The algebra of integro-differential polynomials describes nonlinear integral and differential operators together with initial values. The algebra of integro-differential operators can be used to solve boundary problems for linear ordinary differential equations. In both cases, we describe canonical/normal forms with algorithmic simplifiers.

Categories and Subject Descriptors

I.1.1 [**Symbolic and Algebraic Manipulation**]: Expressions and Their Representation—*simplification of expressions*; I.1.2 [**Symbolic and Algebraic Manipulation**]: Algorithms—*algebraic algorithms*

General Terms

Theory, Algorithms

Keywords

Integral operators, integro-differential algebras, noncommutative Gröbner bases, Green's operators, linear boundary value problems

1. INTRODUCTION

While differential operators are studied extensively in symbolic computation, this cannot be asserted about *integral operators*. In the former case, one uses two fundamental structures for transferring analysis to algebra: "differential operators" and "differential polynomials"; both of these can act on suitable function spaces (the former linearly and the latter nonlinearly). In this paper, we propose two analogous algebraic structures for treating integral operators (along with differential operators): "integro-differential operators" and "integro-differential polynomials"; both of these are extensions of the corresponding differential structures. In Section 2, we review some notions about integro-differential algebras needed for these constructions.

The *integro-differential polynomials* are introduced here for the first time. Their construction is explained in Section 3, the computational approach in Section 4. While modeling nonlinear integral operators, the most important use of the integro-differential polynomial ring $\mathcal{F}\{u\}$ over a given integro-differential algebra $\mathcal{F}$ is probably to describe extensions of integro-differential algebras in a constructive fashion. In practice, one can start with the integro-differential algebra $\mathcal{F}_0$ of exponential polynomials, adjoin a solution of differential equations (with initial values) by passing to a quotient $\mathcal{F}_1$ of $\mathcal{F}_0\{u\}$, and iterate this procedure.

The notion of *integro-differential operators* has been introduced in [23], where it is used for multiplying and factoring BVPs (= linear boundary value problems for ordinary differential equations). In fact, one of the main applications of integro-differential operators is that they describe the differential equation, boundary conditions and solution operator (Green's operator) of a BVP in uniform language. In [23] we have constructed a monoid on BVPs isomorphic to the compositional structure of their Green's operators, studied in [21] from an abstract viewpoint. In this paper, we will review their construction and main properties in Section 5 and focus on computational aspects in Section 6.

For both integro-differential polynomials and operators, the crucial instrument for an algorithmic treatment is of course the usage of standard representatives, but they arise in fairly different contexts: In the former case, where we prefer to speak of *canonical forms*, we employ tools from universal algebra to build a canonical simplifier for the appropriate polynomial concept. In the latter case, where we shall use word *normal forms*, our approach is to construct a confluent rewrite system (equivalently: a noncommutative Gröbner basis).

2. INTEGRO-DIFFERENTIAL ALGEBRAS

Our starting point is a commutative *differential algebra* $(\mathcal{F}, \partial)$ over a field K, so $\partial\colon \mathcal{F} \to \mathcal{F}$ is a K-linear map satisfying the Leibniz rule

$$\partial(fg) = f\,\partial(g) + g\,\partial(f). \tag{1}$$

For convenience, we may assume $K \leq \mathcal{F}$, and we write f' as a shorthand for $\partial(f)$. Furthermore, we will assume that K has characteristic zero and $\mathbb{Q} \leq K$, hence $\mathcal{F}$ is what is

*This work was supported by the Austrian Science Fund (FWF) under the SFB grant F1322.

ISSAC'08, July 20–23, 2008, Hagenberg, Austria.

sometimes called a Ritt algebra [14, p. 12]. The algebra of differential operators over $\mathcal{F}$ is denoted by $\mathcal{F}[\partial]$ as in [27].

For inhomogeneous differential equations $Tu = f$ with $T \in \mathcal{F}[\partial]$, the solution operators (mapping $f \in \mathcal{F}$ to $u \in \mathcal{F}$) are integral operators. The simplest equation is $u' = f$, and its solution operators $\int$ are exactly the *sections* (i.e. K-linear right inverses) of the differential operator ∂ so that

$$\partial\textstyle\int = 1. \tag{2}$$

Note that derivations need not have sections (for example in the algebra of univariate differential polynomials, the indeterminate cannot be a derivative).

The *characterization* of sections follows from Linear Algebra, see [19, p. 17] or [21]: Every section $\int\colon \mathcal{F} \to \mathcal{F}$ of the derivation $\partial\colon \mathcal{F} \to \mathcal{F}$ corresponds to a unique projector $P\colon \mathcal{F} \to \mathcal{F}$ with

$$P = 1 - \textstyle\int\partial \tag{3}$$

and to a unique direct sum decomposition $\mathcal{F} = \mathcal{C} \dotplus \mathcal{I}$ of K-vector spaces with

$$\mathcal{C} = \operatorname{Ker}(\partial) = \operatorname{Im}(P) \quad \text{and} \quad \mathcal{I} = \operatorname{Im}(\textstyle\int) = \operatorname{Ker}(P).$$

Moreover, if $\int$ is any fixed section of ∂, every projector P with $\operatorname{Im}(P) = \operatorname{Ker}(\partial)$ induces a section $(1-P)\int$, and every section of ∂ arises uniquely in this way.

We refer to the elements of $\mathcal{I} = \operatorname{Im}(\int)$ as *initialized* (with respect to $\int$), while those of $\mathcal{C} = \operatorname{Ker}(\partial)$ are usually called the *constants* (with respect to ∂).

As a standard example, we take $\mathcal{F} = C^\infty[a,b]$ where differentiability in the endpoints is understood in the sense of one-sided derivatives. The initialized functions are those that can be written as $F(x) = \int_\alpha^x f(\xi)\,d\xi$ for $f \in C^\infty[a,b]$ and an initialization point $\alpha \in [a,b]$; hence F is the unique antiderivative of f that fulfills the initial condition $F(\alpha) = 0$.

For solving inhomogeneous differential equations of higher order, one must expect to iterate the section $\int$. While this would in general lead to nested integrals, we know from the classical C^∞ setting that the Green's operator can always be expressed via the Green's function [26] by a single integration. To capture this behavior, we need an identity for resolving nested integrals (eventually leading to the rewrite rule for $\int f \int$ in Table 1). Such an identity is given by the so-called *Baxter axiom* (of weight zero), asserting

$$\textstyle\int f \cdot \int g = \int f \int g + \int g \int f \tag{4}$$

for all $f, g \in \mathcal{F}$. Note that we apply the following convention in this paper: An integral like $\int f \int g$ should be interpreted as $\int(f \int g)$, unless we use $\cdot$ as on the left-hand side above.

Obviously (4) is an algebraic version of integration by parts, written in a way that does not involve the derivation. (For the integro-differential polynomials, the role of the Baxter axiom is more subtle: From the left to right, it "flattens" products of nested integrals; in the other direction, it is used for "integrating out" coefficient functions—see Section 4.) A weight-zero *Baxter algebra* $(\mathcal{F}, \int)$ is then a K-algebra $\mathcal{F}$ with a K-linear operation $\int$ fulfilling the Baxter axiom (4); we refer to [11, 2, 24] for more details.

What we shall actually use is the *differential Baxter axiom*, which requires

$$\textstyle\int fg = f\int g - \int f' \int g \tag{5}$$

for all $f, g \in \mathcal{F}$. Note that this is what most people do when they actually apply integration by parts. Variant (4) follows immediately by substituting $\int f$ for f in (5), and often both versions are equivalent (see after Definition 11).

We can also characterize what makes the differential Baxter axiom stronger than the pure one: A section $\int$ of ∂ fulfills the differential Baxter axiom (5) iff it fulfills the pure Baxter axiom (4) and the *homogeneity condition*

$$\textstyle\int cf = c\int f \tag{6}$$

for all $c \in \mathcal{C}$ and $f \in \mathcal{F}$. In fact, (6) implies that $\int\colon \mathcal{F} \to \mathcal{F}$ is $\mathcal{C}$-linear and not only K-linear.

We refer to [23] for the proof of the equivalence and for an example of a differential algebra with a section that satisfies the pure Baxter axiom but not its differential form. To exclude such cases we will insist that *integral operators* must satisfy the differential Baxter axiom.

Definition 1. *Let $\mathcal{F}$ be a differential algebra over a field K. A section $\int$ of ∂ is called an* integral *if it satisfies the differential Baxter axiom* (5)*. In this case, we call $(\mathcal{F}, \partial, \int)$ an* integro-differential algebra.

As an example, take $\mathcal{F} = C^\infty[a,b]$ with its usual derivation ∂ and integral operators

$$\textstyle\int^*\colon f \mapsto \int_a^x f(\xi)\,d\xi \quad \text{and} \quad \int_*\colon f \mapsto \int_x^b f(\xi)\,d\xi.$$

Then both $(\mathcal{F}, \partial, \int^*)$ and $(\mathcal{F}, \partial, -\int_*)$ are integro-differential algebras. By contrast, the operator

$$f \mapsto \int_a^b \int_\tau^x f(\xi)\,d\xi\,d\tau,$$

is just a section for ∂, but not an integral.

In the above example, the projectors $P^\star\colon f \mapsto f(a)$ and $P_\star\colon f \mapsto f(b)$ corresponding to the respective integral operators $\int^*$ and $\int_*$ are *multiplicative* (see (7) below), whereas the projector $\int_a^b$ for the third operator is not. This is true in general—we can characterize integrals by their projectors or images as detailed in [23]: A section $\int\colon \mathcal{F} \to \mathcal{F}$ of the derivation $\partial\colon \mathcal{F} \to \mathcal{F}$ is an integral iff $\mathcal{I} = \operatorname{Im}(\int)$ is an ideal of $\mathcal{F}$ iff $P = 1 - \int\partial$ is multiplicative, meaning

$$P(fg) = P(f)P(g) \tag{7}$$

for all $f, g \in \mathcal{F}$. Using the homogeneity condition (6), this implies also

$$\textstyle\int fg' = fg - \int f'g - P(f)P(g) \tag{8}$$

as an equivalent formulation (corresponding to the rewrite rule for $\int f\partial$ in Table 1) of the differential Baxter axiom (5).

Similar structures are introduced under the name *differential Rota-Baxter algebras* in the recent article [12]. A crucial difference is that they only require the section axiom (2) for connecting derivation and integral, but not the differential Baxter axiom (5). They construct free objects in more general categories where the algebras are over unital commutative rings rather than fields, they may be noncommuative, and the weight can be an arbitrary scalar.

3. THE ALGEBRA OF INTEGRO-DIFFERENTIAL POLYNOMIALS

In this section, we introduce the algebra of *integro-differential polynomials* obtained by adjoining one indeterminate function to an integro-differential algebra. This is a special

case of the general construction of polynomials in universal algebra. See for example [1] for the basic notions in universal algebra that we use in the following and [7, 13, 16] for details on polynomials in universal algebra.

The idea of the construction is as follows. Let $\mathcal{V}$ be a variety defined by a set E of identities or "laws" over a signature Σ. Let A be a fixed "coefficient domain" from the variety $\mathcal{V}$, and let X be a set of "variables" or "indeterminates". Then all terms in the signature Σ with constants (henceforth called "coefficients") in A and variables in X represent the same polynomial if their equality can be derived in finitely many steps from the identities in E and the operations in A. The set of all such terms $\mathcal{T}_\Sigma(A \cup X)$ modulo this congruence $\equiv$ is an algebra in $\mathcal{V}$, called the *polynomial algebra* (for $\mathcal{V}$) in X over A, denoted by $A_\mathcal{V}[X]$.

The polynomial algebra $A_\mathcal{V}[X]$ contains A as a subalgebra, and $A \cup X$ is a generating set. As in the case of polynomials for commutative rings, we have the *substitution homomorphism* in general polynomial algebras. Let B be an algebra in $\mathcal{V}$. Then given a homomorphism $\varphi_1 \colon A \to B$ and a map $\varphi_2 \colon X \to B$, there exists a unique homomorphism

$$\varphi \colon A_\mathcal{V}[X] \to B$$

such that $\varphi(a) = \varphi_1(a)$ for all $a \in A$ and $\varphi(x) = \varphi_2(x)$ for all $x \in X$.

In order to compute with polynomials one can use an effective canonical simplifier [7], that is, a computable map

$$\sigma \colon \mathcal{T}_\Sigma(A \cup X) \to \mathcal{T}_\Sigma(A \cup X)$$

such that

$$\sigma(T) \equiv T \quad \text{and} \quad S \equiv T \Rightarrow \sigma(S) = \sigma(T)$$

for all terms $S, T \in \mathcal{T}_\Sigma(A \cup X)$. The representatives in $\mathcal{R} := \mathrm{Im}(\sigma)$ are called *canonical forms*. Canonical simplifiers correspond uniquely to so-called systems of canonical forms, i.e. a set of terms

$$\mathcal{R} \subseteq \mathcal{T}_\Sigma(A \cup X)$$

such that for every $T \in \mathcal{T}_\Sigma(A \cup X)$ one can compute a canonical form $R \in \mathcal{R}$ with $T \equiv R$ and such that $R \neq \tilde{R} \Rightarrow R \not\equiv \tilde{R}$ for $R, \tilde{R} \in \mathcal{R}$. In other words, for every polynomial in $A_\mathcal{V}[X]$ represented by a term T one can compute a term $R \in \mathcal{R}$ representing the same polynomial, with different terms in $\mathcal{R}$ representing different polynomials, see [16, p. 23].

As a well-known example take the *polynomial ring* $R[x]$ in one indeterminate x over a commutative ring R. The set of all terms of the form $a_n x^n + \ldots + a_0$ with coefficients $a_i \in R$ and $a_n \neq 0$ together with 0 is a system of canonical forms for $R[x]$. One usually defines the polynomial ring directly in terms of these canonical forms. Polynomials for groups, bounded lattices and Boolean algebras are discussed in [16] along with systems of canonical forms.

Let us now consider the variety $\mathcal{V}$ of integro-differential algebras. Its *signature* Σ contains (besides the ring operations): the derivation ∂, the integral $\int$, the family of unary "scalar multiplications" $(\cdot\lambda)_{\lambda \in K}$; for convenience we also include the projection P. The *identities* E are (besides those of a K-algebra and the K-linearity of the operators ∂, $\int$, P): the Leibniz rule (1), the section axiom (2), the definition of the projection (3), and the differential Baxter axiom (5).

Definition 2. *Let $\mathcal{F}$ be an integro-differential algebra. Then $\mathcal{F}_\mathcal{V}[u]$ is called the algebra of* integro-differential polynomials *in u over $\mathcal{F}$ and denoted it by $\mathcal{F}\{u\}$ in analogy to differential polynomials.*

We will also use the following *identities* following from E and describing the basic interactions between the operations in $\mathcal{F}$: the pure Baxter axiom (4), the multiplicativity of the projection (7), the identities

$$P^2 = P, \quad \partial P = 0, \quad P\textstyle\int = 0, \quad \int P(f)g = P(f)\int g, \qquad (9)$$

and the variant (8) of the differential Baxter axiom connecting all three operations. Moreover, we use also the shuffle identity [25, 20] obtained from iterating the Baxter axiom

$$\textstyle\int f_1 \int \ldots \int f_m \cdot \int g_1 \int \ldots \int g_n = \sum \int h_1 \int \ldots \int h_{m+n}, \qquad (10)$$

where the sum ranges over all shuffles of $(f_1, \ldots, f_m)$ and $(g_1, \ldots, g_n)$. By construction of the polynomial algebra, all these identities hold also for $\mathcal{F}\{u\}$.

We will use f, g for denoting coefficients in $\mathcal{F}$ and V for terms in $\mathcal{T}_\Sigma(\mathcal{F} \cup \{u\})$. As for differential polynomials, we write u_n for the nth derivative of u. We use the *multi-index notation*

$$u^\beta = \prod_{i=0}^{\infty} u_i^{\beta_i}$$

for a sequence β in $\mathbb{N}$ with only finitely many nonzero entries. The order of a differential monomial u^β is the highest derivative appearing in u^β or $-\infty$ if $\beta = 0$. Moreover, we write $V(0)$ for $P(V)$ and

$$u(0)^\alpha = \prod_{i=0}^{\infty} u_i(0)^{\alpha_i}.$$

for a multi-index α.

4. CANONICAL FORMS FOR INTEGRO-DIFFERENTIAL POLYNOMIALS

Our goal is to find a system of canonical forms for integro-differential polynomials. As a first step, we describe a system of terms that is sufficient for *representing every polynomial*, but not in a unique (canonical) way.

Lemma 3. *Every polynomial in $\mathcal{F}\{u\}$ can be represented by a finite sum of terms of the form*

$$f u(0)^\alpha u^\beta \textstyle\int f_1 u^{\gamma_1} \int \ldots \int f_n u^{\gamma_n}, \qquad (11)$$

where each multi-index as well as n may be zero.

Proof. By induction on the structure of terms, using the identities of integro-differential algebras and the above mentioned consequences (except the differential variants of the Baxter axiom). □

Note that for terms only involving the derivation, (11) gives already the usual canonical form for differential polynomials. With the aid of Lemma 3, we can now determine the *constants* in $\mathcal{F}\{u\}$.

Proposition 4. *Every constant in $\mathcal{F}\{u\}$ can be represented as a finite sum $\sum_\alpha c_\alpha u(0)^\alpha$ with constants c_α in $\mathcal{F}$.*

PROOF. By the identity $\int\partial = 1 - P$, a term V represents a constant in $\mathcal{F}\{u\}$ iff $P(V) \equiv V$. Since V is congruent to a finite sum of terms of the form (11) and since $\mathrm{Im}(P) = \mathcal{C}$, the identities for P imply that V is congruent to a finite sum of terms of the form $c_\alpha u(0)^\alpha$. □

It is immediately clear that terms of the form (11) cannot be canonical forms for general integro-differential polynomials since for example $f\int gu$ and $\lambda^{-1}f\int\lambda gu$ with $\lambda \in K$ represent the same polynomial. This can be solved by *choosing a basis* $\mathcal{B}$ for $\mathcal{F}$ containing 1.

A second problem for canonical forms comes from the fact that we can integrate certain differential polynomials using *integration by parts* (8). For example, the terms $\int fu'$ and $fu - \int f'u - f(0)u(0)$ represent the same polynomial. More generally, we have the following identity.

LEMMA 5. *We have*

$$\int V u_k^{\beta_k} u_{k+1} \equiv \frac{1}{\beta_k + 1}\left(V u_k^{\beta_k+1} - \int V' u_k^{\beta_k+1} - V(0)u_k(0)^{\beta_k+1}\right) \quad (12)$$

where $k, \beta_k \geq 0$.

PROOF. Using (8) and the Leibniz rule, we see that

$$\int V u_k^{\beta_k} u_{k+1} = \int (V u_k^{\beta_k})(u_k)' \equiv V u_k^{\beta_k+1} - \int V' u_k^{\beta_k+1} - \beta_k \int V u_k^{\beta_k} u_{k+1} - V(0)u_k(0)^{\beta_k+1},$$

and the equation follows. □

In particular, if $V = f u_0^{\beta_0} \dots u_{k-1}^{\beta_{k-1}}$, then V' and hence also the right-hand side of (12) contains only differential monomials with order at most k. So if the highest derivative in the differential monomial u^β of order $k+1$ appears linearly, the term $\int fu^\beta$ is congruent to a sum of terms involving only differential monomials of order at most k. This motivates the following *classification of differential monomials*; confer also [4, 10].

DEFINITION 6. *A monomial* (11)*, with* u^β *having order* k*, is said to have depth* n *and order* k*. It is called* quasiconstant *if* $\beta = 0$*,* quasilinear *if* $k > 0$ *and the highest derivative appears linearly; otherwise it is called* functional*.*

DEFINITION 7. *We write* $\mathcal{R}$ *for the set of all* K*-linear combinations of terms of the form*

$$b u(0)^\alpha u^\beta \int b_1 u^{\gamma_1} \int \dots \int b_n u^{\gamma_n}, \quad (13)$$

where $b, b_1, \dots, b_n \in \mathcal{B}$*, the multi-indices* α, β *as well as* n *may be zero, and* $u^{\gamma_1}, \dots, u^{\gamma_n}$ *are functional.*

As we will see, $\mathcal{R}$ forms a *system of canonical forms* for $\mathcal{F}\{u\}$. The easier part of this claim is that every polynomial has such a representation.

PROPOSITION 8. *Every polynomial in* $\mathcal{F}\{u\}$ *can be represented by a term in* $\mathcal{R}$*.*

PROOF. Using basis expansions and the K-linearity of the integral, we can represent with Lemma 3 every polynomial in $\mathcal{F}\{u\}$ as a K-linear combination of terms of the form

$$b u(0)^\alpha u^\beta \int b_1 u^{\gamma_1} \int \dots \int b_n u^{\gamma_n}, \quad (14)$$

where the multi-indices and n can also be zero.

With basis expansions and the identity

$$\int f \int V \equiv \int f \cdot \int V - \int V \int f,$$

coming from the pure Baxter axiom (4), we can achieve that every multi-index γ_k in (14) is nonzero (induction on depth). Using Lemma 5, one sees that a term $\int b_1 u^{\gamma_1}$ is congruent to a sum of terms involving only integral terms with functional differential monomials (induction on order). Finally one shows (induction on depth and order) that this also holds for terms of the form

$$\int b_1 u^{\gamma_1} \int \dots \int b_n u^{\gamma_n}.$$

The proposition then follows by basis expansions and the K-linearity of the integral. □

It remains to show that each term in $\mathcal{R}$ represents a different polynomial. To this end, let $\langle\mathcal{R}\rangle$ be the *free vector space* over the set of terms (13). In order to distinguish the basis vectors of $\langle\mathcal{R}\rangle$ from the corresponding terms in $\mathcal{R}$, we denote them by

$$\langle b u(0)^\alpha u^\beta \int b_1 u^{\gamma_1} \int \dots \int b_n u^{\gamma_n}\rangle. \quad (15)$$

If $b, b_1, \dots, b_n$ are no basis vectors, (15) is to be understood as an abbreviation for the corresponding basis expansion. We equip the free vector space $\langle\mathcal{R}\rangle$ with the structure of an integro-differential algebra. The operations are defined on the basis vectors mimicking the corresponding operations in $\mathcal{T}_\Sigma(\mathcal{F} \cup \{u\})$, and reducing to congruent terms in $\mathcal{R}$.

The *multiplication* in $\langle\mathcal{R}\rangle$ is introduced in stages. Let J and $\tilde{J}$ range over pure integral terms $\int b_1 u^{\gamma_1} \int \dots \int b_n u^{\gamma_n}$, including 1 for $n = 0$. The product of a term $\langle b u(0)^\alpha u^\beta\rangle$ with a general term $\langle \tilde{b} u(0)^{\tilde{\alpha}} u^{\tilde{\beta}} \tilde{J}\rangle$ is defined as

$$\langle b\tilde{b}\, u(0)^{\alpha+\tilde{\alpha}} u^{\beta+\tilde{\beta}} \tilde{J}\rangle.$$

Corresponding to the shuffle identity (10), we define the product $\langle \int b u^\gamma J\rangle \langle \int \tilde{b} u^{\tilde{\gamma}} \tilde{J}\rangle$ of pure integrals recursively as

$$\langle \int b u^\gamma\rangle \star \langle J\rangle \langle \int \tilde{b} u^{\tilde{\gamma}} \tilde{J}\rangle + \langle \int \tilde{b} u^{\tilde{\gamma}}\rangle \star \langle \tilde{J}\rangle \langle \int b u^\gamma J\rangle,$$

where $\star$ denotes the operation of nesting integrals (multiplication binds stronger than $\star$); the base case is given by the neutral element 1. With this product, the pure integral terms form a subalgebra isomorphic to the shuffle algebra so that $\cdot$ is associative and commutative. Finally, the product of two general basis vectors $\langle b u(0)^\alpha u^\beta J\rangle$ and $\langle \tilde{b} u(0)^{\tilde{\alpha}} u^{\tilde{\beta}} \tilde{J}\rangle$ is given by multiplying $\langle b u(0)^\alpha u^\beta\rangle \langle \tilde{b} u(0)^{\tilde{\alpha}} u^{\tilde{\beta}}\rangle$ with $\langle J\rangle\langle\tilde{J}\rangle$.

The *derivation* of basis vector is defined through the Leibniz rule, using also the identities $\partial P = 0$, $\partial\int = 1$ and basis expansions.

The *integral* of a basis vector is defined recursively (first by depth and then by order), based on the classification of Definition 6. In the quasiconstant case, we define

$$\int \langle b u(0)^\alpha J\rangle = \langle \int b\rangle \langle u(0)^\alpha J\rangle - \int \langle J'\rangle \int \langle b u(0)^\alpha\rangle,$$

where J' is J with the integral removed (zero for $J = 1$). For a quasilinear basis vector

$$\langle b u(0)^\alpha V u_k^{\beta_k} u_{k+1} J\rangle \quad \text{with} \quad V = u_0^{\beta_0} \dots u_{k-1}^{\beta_{k-1}},$$

we set $s = \beta_k + 1$ and define the integral by

$$s\int \langle b u(0)^\alpha V u_k^{\beta_k} u_{k+1} J\rangle = \langle b u(0)^\alpha V u_k^s J\rangle - \langle u(0)^\alpha\rangle \int \langle bVJ\rangle' \langle u_k^s\rangle - \langle bV u^\alpha u_k^s J\rangle(0);$$

the third summand is absent unless $J = 1$.

In the functional case, we use

$$\int\langle bu(0)^\alpha u^\beta J\rangle = \langle u(0)^\alpha \int bu^\beta J\rangle,$$

as a definition for the integral.

For showing that $\langle\mathcal{R}\rangle$ is an *integro-differential algebra*, we have to verify the axioms: First of all we see that it is a commutative K-algebra by our previous remark about the shuffle product. The Leibniz rule and the section axiom follow immediately from the definition. The only difficult task is to prove the differential Baxter axiom. An easy calculation shows that

$$\int\langle u(0)^\alpha\rangle\langle R\rangle = \langle u(0)^\alpha\rangle\int\langle R\rangle.$$

Proposition 4 then implies that $\int$ is homogeneous over the constants in $\langle\mathcal{R}\rangle$. By the observation before (6), it suffices therefore to verify the pure Baxter axiom. The proof is lengthy (using inductions over depth and order, with case distinctions according to the definition of the integral) and will be presented in a subsequent publication.

PROPOSITION 9. *With the operations defined as above, $\langle\mathcal{R}\rangle$ is an integro-differential algebra.*

The integro-differential algebra $\langle\mathcal{R}\rangle$ provides the key for showing that all terms in $\mathcal{R}$ represent different polynomials of $\mathcal{F}\{u\}$.

THEOREM 10. *The terms in $\mathcal{R}$ constitute a system of canonical forms for $\mathcal{F}\{u\}$, provided that basis expansion in $\mathcal{F}$ is computable.*

PROOF. Since $\langle\mathcal{R}\rangle$ is an integro-differential algebra, there exists a unique substitution homomorphism

$$\varphi\colon \mathcal{F}\{u\} \to \langle\mathcal{R}\rangle$$

such that $\varphi(f) = \langle f\rangle$ for all $f \in \mathcal{F}$ and $\varphi(u) = \langle u\rangle$. Let

$$\pi\colon \mathcal{R} \to \mathcal{F}\{u\}$$

denote the restriction of the canonical epimorphism associated with $\equiv$. Then $\varphi\circ\pi$ is injective since it maps $R \in \mathcal{R}$ to $\langle R\rangle \in \langle\mathcal{R}\rangle$ and surjective by Proposition 8. We conclude that π is also bijective, so $\mathcal{R}$ is indeed a system of canonical forms. □

5. THE ALGEBRA OF INTEGRO-DIFFERENTIAL OPERATORS

As explained in the Introduction, one important application of integro-differential polynomials is the *adjunction* of new elements to an initially given integro-differential algebra $\mathcal{F}$; this issue will be broached in a future paper. If $\mathcal{F}$ is ordinary (see Definition 11 below), we can thus ensure that a given homogeneous differential equation $Tu = 0$ with monic $T \in \mathcal{F}[\partial]$ is dimensionally adequate, meaning $\dim_K \operatorname{Ker}(T) = \deg T$. This is the prerequisite for finding the Green's operator of the corresponding inhomogeneous equation $Tu = f$; see [23] for a detailed description of the solution method. Its groundwork consists of adding and multiplying integro-differential operators, and this is what we shall consider here.

Before giving the construction of integro-differential operators, we will explicitly restrict ourselves to *ordinary differential equations* in the following sense. Note that in the following definition our terminology deviates from [15, p. 58], where it only refers to having a single derivation.

DEFINITION 11. *A differential algebra $\mathcal{F}$ over a field K is called* ordinary *if $\dim_K \operatorname{Ker}(\partial) = 1$. An integro-differential algebra $(\mathcal{F}, \partial, \int)$ is called ordinary if $(\mathcal{F}, \partial)$ is ordinary.*

As a consequence, the solution space of a homogeneous differential equation $Tu = 0$ with monic $T \in \mathcal{F}[\partial]$ is now finite-dimensional, so we can indeed enforce *dimensional adequacy* by adjunction. (The notion of saturated integro-differential algebra [23] postulates dimensional adequacy for every monic $T \in \mathcal{F}[\partial]$.)

Clearly we have $K = \mathcal{C}$ in an ordinary differential algebra $\mathcal{F}$, which is thus an algebra over its own field of constants. But then a section is automatically homogeneous over $\mathcal{C}$, so the pure Baxter axiom (4) and its differential version (5) coincide. Moreover, one knows from Linear Algebra that a *projector* P onto a one-dimensional subspace $[w]$ of a K-vector space V can be written as $P(v) = \varphi(v)\,w$ with a functional φ that can be made unique by the normalization $\varphi(w) = 1$. If V is a K-algebra, a projector onto $K = [1]$ is canonically described by the functional φ with $\varphi(1) = 1$. This holds in particular in an ordinary differential algebra, where the projectors (3) corresponding to sections of the derivation can be regarded as normalized functionals.

In an ordinary integro-differential algebra $\mathcal{F}$, the normalized functional corresponding to the integral $\int$ is moreover multiplicative, as explained at the end of Section 2. We call this multiplicative functional

$$\mathtt{E} = 1 - \int\partial \tag{16}$$

its *evaluation*. The terminology stems from the standard example $\mathcal{F} = C^\infty[a,b]$, where $\mathtt{E}$ is a point evaluation (see below Definition 1). The multiplicative functionals on an algebra are known as its *characters* (note that all characters are normalized). We write $\mathcal{M}(\mathcal{F})$ for the vector space of all characters on an ordinary integro-differential algebra $\mathcal{F}$, including the evaluation $\mathtt{E}$ as a distinguished character.

Let $\mathcal{F}$ be a *fixed ordinary integro-differential algebra* over a field K with evaluation $\mathtt{E}$. The variables f, g are used for elements of $\mathcal{F}$, the variables φ, ψ for elements of $\mathcal{M}(\mathcal{F})$. We introduce now an algebra of operators on $\mathcal{F}$ using rewrite systems [1] in the spirit of [3].

fg	$\to$	$f\cdot g$	∂f	$\to$	$\partial\cdot f + f\partial$
$\varphi\psi$	$\to$	ψ	$\partial\varphi$	$\to$	0
φf	$\to$	$(\varphi\cdot f)\,\varphi$	$\partial\int$	$\to$	1
$\int f\int$	$\to$	$(\int\cdot f)\int - \int(\int\cdot f)$			
$\int f\partial$	$\to$	$f - \int(\partial\cdot f) - (\mathtt{E}\cdot f)\,\mathtt{E}$			
$\int f\varphi$	$\to$	$(\int\cdot f)\,\varphi$			

Table 1: Rewrite System for $\mathcal{F}[\partial,\int]$

DEFINITION 12. *The* integro-differential operators $\mathcal{F}[\partial,\int]$ *are defined as the K-algebra generated by the symbols ∂ and $\int$, the "functions" $f \in \mathcal{F}$ and the multiplicative "functionals" $\varphi \in \mathcal{M}(\mathcal{F})$, modulo the rewrite system of Table 1.*

In the rules of Table 1, we use the notation $U\cdot f$ for the *action* of U on an element $f \in \mathcal{F}$, where U is an element of the free algebra in the above generators. It is an easy matter

to check that the rewrite rules of Table 1 are fulfilled in $\mathcal{F}$, so we may lift $\cdot$ to an action of $\mathcal{F}[\partial, \int]$ on $\mathcal{F}$. In particular, $f \cdot g$ now denotes the product in $\mathcal{F}$.

We remark that Table 1 is to be understood as including *implicit rules* for $\int\int$, $\int\partial$ and $\int\varphi$ by substituting $f = 1$ in the rules for $\int f\int$, $\int f\partial$ and $\int f\varphi$, respectively. Moreover, one obtains the *derived rule* $\mathtt{E}\int = 0$ from the definition of the evaluation $\mathtt{E}$. Note also that $\mathcal{F}[\partial] \subseteq \mathcal{F}[\partial, \int]$, with the same induced action on $\mathcal{F}$.

THEOREM 13. *The rewrite system for $\mathcal{F}[\partial, \int]$ in Table 1 is convergent.*

In other words, the polynomials given by the difference between the left-hand and right-hand sides of Table 1 form a two-sided noncommutative *Gröbner basis*. The proof is given in [23]. For the theory of Gröbner bases, we refer to [5, 6], for its noncommutative extension to [17, 18].

6. NORMAL FORMS FOR INTEGRO-DIFFERENTIAL OPERATORS

Having a convergent rewrite system, every integro-differential operator has a *unique normal form* [1, p. 12]. To compute such normal forms we also need a canonical simplifier on the free algebra generated by ∂ and $\int$, the functions $f \in \mathcal{F}$ and the functionals $\varphi \in \mathcal{M}(\mathcal{F})$; one possibility is by basis expansion in $\mathcal{F}$. Here we summerize the description of the normal forms on $\mathcal{F}[\partial, \int]$ obtained in [23].

We first consider operators in the right ideal

$$\mathcal{S}(\mathcal{F}) = \mathcal{M}(\mathcal{F})\,\mathcal{F}[\partial, \int],$$

which we call *Stieltjes boundary conditions* over $\mathcal{F}$ or "boundary conditions" for short. Every such boundary condition has the normal form

$$\sum_{\varphi\in\mathcal{M}(\mathcal{F})} \left(\sum_{i\in\mathbb{N}} a_{\varphi,i}\,\varphi\partial^i + \varphi\int f_\varphi \right)$$

with $a_{\varphi,i} \in K$ and $f_\varphi \in \mathcal{F}$ almost all zero. We write $\mathcal{F}[\mathtt{E}]$ for the left $\mathcal{F}$-submodule generated by $\mathcal{S}(\mathcal{F})$ and call them *Stieltjes boundary operators* or "boundary operators" for short.

With the rule for ∂f of Table 1 it is clear that the *differential operators* $\mathcal{F}[\partial] \subset \mathcal{F}[\partial, \int]$ have their usual normal forms. Analogously, we write $\mathcal{F}[\int] \subset \mathcal{F}[\partial, \int]$ for the subalgebra of *integral operators*, generated by the functions and $\int$ modulo the rule for $\int f\int$ of Table 1; one sees immediately that that their normal forms are linear combinations of $f\int g$ with $f, g \in \mathcal{F}$.

THEOREM 14. *Up to ordering the summands, every normal form of $\mathcal{F}[\partial, \int]$ with respect to the rewrite system of Table 1 can be written uniquely as a sum $T + G + B$ with $T \in \mathcal{F}[\partial]$ and $G \in \mathcal{F}[\int]$ and $B \in \mathcal{F}[\mathtt{E}]$.*

We can use integro-differential operators for specifying and solving boundary problems. Since space is limited, we can only state the main results here; for details and complete proofs, we must again refer to [23]. We formulate the *boundary problem* for a monic differential operator $T \in \mathcal{F}[\partial]$ with $\deg T = n$ and Stieltjes boundary conditions $\beta_1, \ldots, \beta_n \in \mathcal{S}(\mathcal{F})$ as follows.

Given a forcing function $f \in \mathcal{F}$, find $u \in \mathcal{F}$ such that

$$\begin{aligned} Tu &= f, \\ \beta_1 u = \ldots &= \beta_n u = 0. \end{aligned} \tag{17}$$

We call the boundary problem regular if there is a unique $u \in \mathcal{F}$ for every $f \in \mathcal{F}$; this implies in particular that $\beta_1, \ldots, \beta_n$ are linearly independent over K.

The first step in solving (17) is to consider the corresponding *initial value problem* based on a character $\eta \in \mathcal{M}(\mathcal{F})$, where one replaces the boundary conditions $\beta_1, \ldots, \beta_n$ by $\eta, \eta\partial, \ldots, \eta\partial^{n-1}$. Note that one may in particular choose $\eta = \mathtt{E}$, evaluating in the initialization point. The main idea of solving initial value problems is of course an adaption of the familiar variation-of-constants formula (see for example in [9, p. 74] for systems and [9, p. 87] for scalar differential equations).

PROPOSITION 15. *Let $T \in \mathcal{F}[\partial]$ be a monic differential operator with $\deg T = n$ such that $Tu = 0$ has a fundamental system of solutions $u_1, \ldots, u_n \in \mathcal{F}$. If W is its Wronskian matrix and $d = \det W$ is invertible in $\mathcal{F}$, the initial value problem $Tu = f$ based on $\eta \in \mathcal{M}(\mathcal{F})$ has the unique solution*

$$u = \sum_{i=1}^{n} u_i(1-\eta)\textstyle\int d^{-1}\, d_i f \tag{18}$$

for every forcing function $f \in \mathcal{F}$. Here $d_i = \det W_i$, where W_i is the matrix obtained from W by replacing the ith column by the nth unit vector.

PROOF. We can use the usual technique of reformulating $Tu = f$ as a system of linear first-order differential equations with companion matrix $A \in \mathcal{F}^{n\times n}$. The integral operator

$$\textstyle\int\!\!\!\!\!\!- = (1-\eta)\int$$

is a section of ∂ with corresponding projector $1 - \int\!\!\!\!\!\!-\,\partial = \eta$. Since η is multiplicative, we know from Section 2 that $\int\!\!\!\!\!\!-$ is an integral. We extend the action of the operators $\int\!\!\!\!\!\!-, \partial, \eta$ componentwise to $\mathcal{F}^n$. Setting now

$$\hat{u} = W \textstyle\int\!\!\!\!\!\!- W^{-1} \hat{f}$$

with $\hat{f} = (0, \ldots, 0, f)^\top \in \mathcal{F}^n$, one may readily check that $\hat{u} \in \mathcal{F}^n$ is a solution of the first-order system $\hat{u}' = A\hat{u} + \hat{f}$ with initial condition $\eta\hat{u} = 0$. Writing u for the first component of $\hat{u}$, we have a solution of the initial value problem $Tu = f$ based on $\eta \in \mathcal{M}(\mathcal{F})$. Using Cramer's rule to compute the nth column of W^{-1}, we see that

$$W^{-1}\,\hat{f} = d^{-1} f\,(d_1, \ldots, d_n)^\top,$$

and (18) follows since the first row of W is $(u_1, \ldots, u_n)$.

For proving uniqueness, assume $Tu = 0$ along with the initial conditions $\eta u = \ldots = \eta u^{(n-1)} = 0$. Let

$$u = c_1 u_1 + \ldots + c_n u_n$$

with coefficients in K. Then the initial conditions yield $\eta(Wc) = 0$ with $c = (c_1, \ldots, c_n)^\top \in K^n$. But $\eta(Wc) = \eta(W)c$ because η is linear, and $\det\eta(W) = \eta(\det W)$ because it is moreover multiplicative. Since $\det W \in \mathcal{F}$ is invertible, this implies that $\eta(W) \in K^{n\times n}$ is regular, so $c = \eta(W)^{-1}0 = 0$ and $u = 0$. □

The above proposition hinges on two conditions: The first has already been discussed and can be satisfied by adjunction. The second condition needs an *invertible Wronskian d*.

This could also be enforced by a suitable localization of $\mathcal{F}$, as for Picard-Vessiot rings [27, p. 12]. But in many applications, this condition will come out naturally: The Wronskian d is always an exponential over $\mathcal{F}$ since it satisfies the differential equation $d' = ad$, where a is the trace of the companion matrix A.

Since every integro-differential algebra $\mathcal{F}$ comes with the evaluation $\eta = \mathtt{E}$ as a distinguished character, we can speak of the initial value problem associated with a monic $T \in \mathcal{F}[\partial]$. Then the map $T^\blacklozenge : \mathcal{F} \to \mathcal{F}$ described by the assignment $f \mapsto u$ in (18) simplifies to

$$T^\blacklozenge = \sum_{i=1}^{n} u_i \textstyle\int d^{-1}\, d_i. \tag{19}$$

We call $T^\blacklozenge \in \mathcal{F}[\partial, \int]$ the *fundamental right inverse* of T. Note that (19) can be further simplified if T has constant coefficients; see [22].

The next step in solving (17) is to compute the *projector* onto $\operatorname{Ker}(T) = [u_1, \ldots, u_n]$ along

$$[\beta_1, \ldots, \beta_n]^\perp = \{u \in \mathcal{F} \mid \beta_1 u = \ldots = \beta_n u = 0\},$$

which can be achieved as follows: Change from the basis $\beta_1, \ldots, \beta_n$ of $[\beta_1, \ldots, \beta_n]$ to a new basis $\tilde{\beta}_1, \ldots, \tilde{\beta}_n$ over K biorthogonal to $u_1, \ldots, u_n$ in the sense that $\tilde{\beta}_i(u_j) = \delta_{ij}$. Then the projector can be determined as

$$P = u_1\tilde{\beta}_1 + \ldots + u_n\tilde{\beta}_n \in \mathcal{F}[\mathtt{E}].$$

See [23] for further details.

We can now put everything together for determining the *Green's operator* $G\colon f \mapsto u$ of (17). The point is that $T^\blacklozenge$ solves the initial value problem, while $1 - P$ "translates" the initial conditions $\mathtt{E}, \mathtt{E}\partial, \ldots, \mathtt{E}\partial^{n-1}$ to the required boundary conditions $\beta_1, \ldots, \beta_n$.

THEOREM 16. *Let $T \in \mathcal{F}[\partial]$ be monic with $\deg T = n$ and $\beta_1, \ldots, \beta_n \in \mathcal{S}(\mathcal{F})$ such that the boundary problem (17) is regular. If the conditions of Proposition 15 are satisfied, the Green's operator of (17) is given by*

$$G = (1 - P)\, T^\blacklozenge,$$

where P is the projector onto $\operatorname{Ker}(T)$ along $[\beta_1, \ldots, \beta_n]^\perp$.

PROOF. Let $u_1, \ldots, u_n$ be a fundamental system for T. We have $TG = TT^\blacklozenge - PT^\blacklozenge = 1 - 0$ since P annihilates $u_1, \ldots, u_n$. Thus $u = Gf$ satisfies the differential equation $Tu = f$ of (17).

For ensuring the boundary conditions of (17), we prove $\beta_i G = 0$ for $i = 1, \ldots, n$. But we have even $\beta_i(1 - P) = 0$ because $1 - P$ projects onto $[\beta_1, \ldots, \beta_n]^\perp$. □

In analysis, the Green's operator G is usually written as an integral operator with the bivariate *Green's function* as its kernel. As remarked in Section 2, this is the effect of the Baxter axiom. Hence the abstract version of a Green's function is the Green's operator $G \in \mathcal{F}[\partial, \int]$ written in its normal form. In the classical $C^\infty[a, b]$ setting, there is indeed a straight-forward correspondence between normal forms and Green's functions [22].

7. CONCLUSION

We have presented two algorithmic tools for studying integration from an *algebraic operator perspective*. The integro-differential polynomials, introduced for the first time in this paper, enjoy a rich structure that deserves further analysis. Specifically, their quotient algebras are relevant in view of adjunctions (see at the beginning of Section 5). Unlike the integro-differential polynomials, the integro-differential operators are an algebraic model of linear operators, based on the (noncommutative) compositional structure. Their normal forms are much easier to describe since one can fall back on Gröbner basis methods. We can benefit from both in the study of differential equations—particularly when considered with boundary conditions.

Acknowledgments

We would like to thank our project leaders *Bruno Buchberger* and *Heinz W. Engl* for their continuous support, critical comments and helpful suggestions.

8. REFERENCES

[1] F. Baader and T. Nipkow. *Term rewriting and all that*. Cambridge University Press, Cambridge, 1998.

[2] G. Baxter. An analytic problem whose solution follows from a simple algebraic identity. *Pacific J. Math.*, 10:731–742, 1960.

[3] G. M. Bergman. The diamond lemma for ring theory. *Adv. in Math.*, 29(2):178–218, 1978.

[4] A. H. Bilge. A REDUCE program for the integration of differential polynomials. *Comput. Phys. Comm.*, 71(3):263–268, 1992.

[5] B. Buchberger. *An algorithm for finding the bases elements of the residue class ring modulo a zero dimensional polynomial ideal (German)*. PhD thesis, Univ. of Innsbruck, 1965. English translation published in *J. Symbolic Comput.*, 41(3-4):475–511, 2006.

[6] B. Buchberger. Introduction to Gröbner bases. 1998. In [8], pp. 3–31.

[7] B. Buchberger and R. Loos. Algebraic simplification. In *Computer algebra*, pages 11–43. Springer, Vienna, 1983.

[8] B. Buchberger and F. Winkler, editors. *Gröbner bases and applications*, volume 251 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1998. Papers from the Conference on 33 Years of Gröbner Bases held at the University of Linz, Linz, February 2–4, 1998.

[9] E. A. Coddington and N. Levinson. *Theory of ordinary differential equations*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1955.

[10] I. M. Gelfand and L. A. Dikiĭ. Fractional powers of operators, and Hamiltonian systems. *Funkcional. Anal. i Priložen.*, 10(4):13–29, 1976. English translation: Functional Anal. Appl. 10 (1976), no. 4, 259–273 (1977).

[11] L. Guo. Baxter algebras and differential algebras. In *Differential algebra and related topics (Newark, NJ, 2000)*, pages 281–305. World Sci. Publ., River Edge, NJ, 2002.

[12] L. Guo and W. Keigher. On differential Rota-Baxter algebras. *J. Pure Appl. Algebra*, 212(3):522–540, 2008.

[13] H. Hule. Polynome über universalen Algebren. *Monatsh. Math.*, 73:329–340, 1969.

[14] I. Kaplansky. *An introduction to differential algebra*. Publ. Inst. Math. Univ. Nancago, No. 5. Hermann, Paris, 1957.

[15] E. Kolchin. *Differential algebra and algebraic groups*, volume 54 of *Pure and Applied Mathematics*. Academic Press, New York-London, 1973.

[16] H. Lausch and W. Nöbauer. *Algebra of polynomials*, volume 5 of *North-Holland Mathematical Library*. North-Holland Publishing Co., Amsterdam, 1973.

[17] F. Mora. Groebner bases for non-commutative polynomial rings. In *AAECC-3: Proceedings of the 3rd International Conference on Algebraic Algorithms and Error-Correcting Codes*, pages 353–362, London, UK, 1986. Springer-Verlag.

[18] T. Mora. An introduction to commutative and noncommutative Gröbner bases. *Theoret. Comput. Sci.*, 134(1):131–173, 1994. Second International Colloquium on Words, Languages and Combinatorics (Kyoto, 1992).

[19] M. Z. Nashed and G. F. Votruba. A unified operator theory of generalized inverses. In M. Z. Nashed, editor, *Generalized inverses and applications (Proc. Sem., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1973)*, pages 1–109. Academic Press, New York, 1976.

[20] R. Ree. Lie elements and an algebra associated with shuffles. *Ann. of Math. (2)*, 68:210–220, 1958.

[21] G. Regensburger and M. Rosenkranz. An algebraic foundation for factoring linear boundary problems. *Ann. Mat. Pura Appl. (4)*, 2008. DOI:10.1007/s10231-008-0068-3.

[22] M. Rosenkranz. A new symbolic method for solving linear two-point boundary value problems on the level of operators. *J. Symbolic Comput.*, 39(2):171–199, 2005.

[23] M. Rosenkranz and G. Regensburger. Solving and factoring boundary problems for linear ordinary differential equations in differential algebras. *J. Symbolic Comput.*, 2007. DOI:10.1016/j.jsc.2007.11.007.

[24] G.-C. Rota. Baxter algebras and combinatorial identities (I, II). *Bull. Amer. Math. Soc.*, 75:325–334, 1969.

[25] G.-C. Rota. Ten mathematics problems I will never solve. *Mitt. Dtsch. Math.-Ver.*, (2):45–52, 1998.

[26] I. Stakgold. *Green's functions and boundary value problems*. John Wiley & Sons, New York, 1979.

[27] M. van der Put and M. F. Singer. *Galois theory of linear differential equations*, volume 328 of *Grundlehren der Mathematischen Wissenschaften*. Springer, Berlin, 2003.

Annali di Matematica (2009) 188:123–151
DOI 10.1007/s10231-008-0068-3

An algebraic foundation for factoring linear boundary problems

Georg Regensburger · Markus Rosenkranz

Received: 31 May 2007 / Revised: 12 October 2007 / Published online: 26 March 2008

Abstract Motivated by boundary problems for linear differential equations, we define an abstract boundary problem as a pair consisting of a surjective linear map ("differential operator") and an orthogonally closed subspace of the dual space ("boundary conditions"). Defining the composition of boundary problems corresponding to their Green's operators in reverse order, we characterize and construct all factorizations of a boundary problem from a given factorization of the defining operator. For the case of ordinary differential equations, the main results can be made algorithmic. We conclude with a factorization of a boundary problem for the wave equation.

Keywords Linear boundary value problems · Factorization · Green's operators

Mathematics Subject Classification (2000) 47A68 · 34B05 · 35G15

1 Introduction

To motivate our algebraic setting and terminology, we begin with two illustrative examples for boundary problems, one for ordinary and one for partial differential equations. The goal is to determine the operator mapping the right-hand side ("forcing function") of the differential equation to its solution, subject to the given boundary conditions. It is known as *Green's operator* [26], since it is the integral operator induced by the *Green's function*. This name

This work was supported by the Austrian Science Fund (FWF) under the SFB grant F1322.

G. Regensburger ( ) · M. Rosenkranz
Johann Radon Institute for Computational and Applied Mathematics (RICAM),
Austrian Academy of Sciences, Altenbergerstraße 69, 4040 Linz, Austria
e-mail: georg.regensburger@oeaw.ac.at

M. Rosenkranz
e-mail: markus.rosenkranz@oeaw.ac.at

was introduced by Neumann [16] and Riemann [18, Sect. 23] in honor of the mathematician Green (1793–1841), who invented the concept in [8, p. 12].

The first example is a classical *two-point boundary value problem* on a finite interval; see for example Stakgold [23]. Writing V for the complex vector space $C^\infty[0, 1]$, we consider the following problem: given $f \in V$, find $u \in V$ such that

$$\boxed{\begin{aligned} u'' &= f, \\ u(0) &= u(1) = 0. \end{aligned}} \tag{1.1}$$

Let $D\colon V \to V$ denote the usual derivation and $L, R \in V^*$ the two linear functionals $L\colon f \mapsto f(0)$ and $R\colon f \mapsto f(1)$. Note that u is annihilated by any linear combination of these two functionals so that problem (1.1) can be described by $(D^2, [L, R])$, where $[L, R]$ is the subspace of the dual space generated by L and R.

Based on an operator approach first presented in [20], a symbolic method for computing Green's operators for regular two-point boundary problems with constant coefficients was given in [19]. We describe a *symbolic framework* treating boundary problems for arbitrary linear ordinary differential equations in [21]. A crucial step is the computation of normal forms using a suitable noncommutative Gröbner basis that reflects the essential interactions between certain basic operators. Gröbner bases were introduced by Buchberger in [2,3].

As a second example consider the following boundary problem for the *wave equation* on the domain $\Omega = \mathbb{R} \times \mathbb{R}_{\geq 0}$, now writing V for $C^\infty(\Omega)$: Given $f \in V$, find $u \in V$ such that

$$\boxed{\begin{aligned} u_{tt} - u_{xx} &= f, \\ u(x, 0) = u_t(x, 0) &= 0. \end{aligned}} \tag{1.2}$$

Note that we use the terms "boundary condition/problem" in the general sense of linear conditions. (Usually one calls the above problem an initial value problem; for a genuine boundary problem we refer to the end of the paper. We prefer the term "boundary problem" to the more common expression "boundary value problem" since the latter would suggest that boundary conditions are always point evaluations, while we will also need integral conditions.)

The boundary conditions in (1.2) can be expressed by the infinite family of linear functionals $L_x\colon u \mapsto u(x, 0)$, $M_x\colon u \mapsto u_t(x, 0)$ with $x \in \mathbb{R}$, so we can represent the boundary problem by $(\partial_t^2 - \partial_x^2, [L_x, M_x]_{x\in\mathbb{R}})$. The space $[\ldots]$ here denotes the *orthogonal closure* (see Appendix A.1 for details) of the subspace generated by the boundary conditions: Since u is annihilated by the L_x and M_x, it is also annihilated by all functionals in $[L_x, M_x]$, for example the functionals $u \mapsto \int_0^x u(\eta, 0)\, d\eta$ for $x \in \mathbb{R}$.

Abstracting from the above examples, we define a *boundary problem* as a pair consisting of a surjective linear map and an orthogonally closed subspace of the dual space. Every finite-dimensional vector space of the dual is orthogonally closed (like the boundary conditions in the first example), but we need the notion of orthogonal closure to deal with infinite dimensional vector spaces (as in the second example) if we are to remain in an algebraic setting.

It would be interesting to extend our results such that additional *topological assumptions* on the vector spaces and operators are taken into account. For example, it should be possible to use a dual pairing [13] instead of a vector space and its algebraic dual. For an approach along these lines, see Wyler [26], dealing with generalized Green's operators.

One motivation for us was that understanding algebraic aspects of boundary problems is important for treating boundary problems by *symbolic computation*, where one usually

considers manipulations of the operators that are independent of the spaces they act on. Since the surjective linear map may also be a matrix differential operator, this approach can be extended to boundary problems for systems of linear differential equations.

In the abstract setting, computing the *Green's operator* of a boundary problem means determining the right inverse of the defining operator corresponding to the kernel complement given by the space of boundary conditions. Going back from a Green's operator to its boundary problem can be interpreted as solving a suitably defined dual boundary problem.

The crucial step in our approach consists in the passage from a single problem to a *compositional structure on boundary problems*, defined in such a way that it corresponds to the composition of the Green's operators in reverse order. As we will see, the computation of Green's operators can then be seen as an anti-isomorphism between boundary problems and dual boundary problems.

Our main result in this paper is the description of *factorizations* in this compositional structure: given a boundary problem, we characterize and construct all possible factorizations along a given factorization of the defining operator. By the above anti-isomorphism, this also yields a method for factoring Green's operators.

In the setting of *differential equations*, factoring boundary problems allows us to split a problem of higher order into subproblems of lower order, provided we can factor the differential operator. For the latter, we can exploit algorithms and results about factoring ordinary [11,17,22,24] and partial differential operators [9,10,25]. The factor problems can then be dealt with by symbolic, numerical or hybrid methods. For numerical or hybrid methods one has to consider stability issues [6]: a given well-posed problem should be factored such that the lower-order problems are well-posed.

The paper is organized as follows: in Sect. 2, we introduce abstract boundary problems and dual boundary problems. The composition of boundary problems with the above anti-isomorphism is described in Sect. 3. We consider the question of factoring boundary problems in Sect. 4. For endomorphisms, we give in Sect. 5 an interpretation of the composition as a semidirect product of monoids. In Sect. 6, we focus on operators with finite dimensional kernel, where all the main constructions can be made algorithmic. This includes in particular boundary problems for ordinary differential equations, treated from a symbolic computation perspective in [21]. We conclude in Sect. 7 with computing factorizations and Green's operators for (1.1) and (1.2).

In the appendix, we recall and develop various auxiliary results from linear algebra. In Appendix A.1 we treat the duality between subspaces of a vector space and orthogonally closed subspaces of its dual. The relation between orthogonality and the transpose is discussed in Appendix A.2. Left and right inverses are covered in Appendix A.3, the dimension arguments needed for finitely many boundary conditions in Appendix A.4.

2 Boundary problems and Green's operators

A *boundary problem* is given by a pair $(T, \mathcal{F})$, where $T : V \to W$ is a surjective linear map between vector spaces V, W and $\mathcal{F} \subseteq V^*$ an orthogonally closed subspace of *boundary conditions*. We say that $u \in V$ is a solution of $(T, \mathcal{F})$ for a given $w \in W$, if

$$Tu = w \quad \text{and} \quad f(u) = 0 \quad \text{for all } f \in \mathcal{F}$$

or equivalently $u \in \mathcal{F}^\perp$. A boundary problem $(T, \mathcal{F})$ is *regular* if $\mathcal{F}^\perp$ is a complement of $K = \operatorname{Ker} T$ so that $V = K \dotplus \mathcal{F}^\perp$. Then there exists a unique right inverse $G : W \to V$ of T with $\operatorname{Im} G = \mathcal{F}^\perp$, see Appendix A.3. We call G the *Green's operator* for the boundary

problem $(T, \mathcal{F})$. Since $TGw = w$ and $Gw \in \mathcal{F}^\perp$, we see that the Green's operator maps every right-hand side $w \in W$ to its unique solution $u = Gw \in V$. Hence we say that G solves the boundary problem $(T, \mathcal{F})$, and we use the notation

$$G = (T, \mathcal{F})^{-1}.$$

Conversely, if there exists a right inverse G of T for a boundary problem $(T, \mathcal{F})$ such that $\operatorname{Im} G = \mathcal{F}^\perp$, it is regular by (A.17). Since orthogonality preserves direct sums, we see that $(T, \mathcal{F})$ is regular iff

$$V^* = \mathcal{F} \dotplus K^\perp. \tag{2.1}$$

By Proposition A.6, we have

$$\operatorname{Ker} G^* = (\operatorname{Im} G)^\perp = \mathcal{F}^{\perp\perp} = \mathcal{F} \quad \text{and} \quad \operatorname{Im} T^* = (\operatorname{Ker} T)^\perp = K^\perp \tag{2.2}$$

for a regular boundary problem $(T, \mathcal{F})$. Given any right inverse $\tilde{G}$ of T, we know with Lemma A.8 that the Green's operator for a regular boundary problem $(T, \mathcal{F})$ is given by

$$G = (1 - P)\tilde{G}, \tag{2.3}$$

where P is the projection with $\operatorname{Im} P = K$ and $\operatorname{Ker} P = \mathcal{F}^\perp$.

If T is invertible, then $(T, 0)$ is the only regular boundary problem for T, and its Green's operator is $(T, 0)^{-1} = T^{-1}$. In particular, we have

$$(1, 0)^{-1} = 1 \tag{2.4}$$

for the identity operator.

A *dual boundary problem* is given by a pair (K, G), where $G\colon W \to V$ is an injective linear map and $K \subseteq V$ a subspace of *dual boundary conditions*. We say that $g \in V^*$ is a solution of (K, G) for a given $h \in W^*$ if

$$G^* g = h \quad \text{and} \quad g(v) = 0 \quad \text{for all } v \in K$$

or equivalently $g \in K^\perp$. A dual boundary problem (K, G) is *regular* if K is a complement of $I = \operatorname{Im} G$ so that $V = K \dotplus I$. Then there exists a unique left inverse $T\colon V \to W$ of G with $\operatorname{Ker} T = K$, see Appendix A.3. We call T the dual Green's operator for the *dual boundary problem* (K, G). Since $G^*T^* = 1$ and $\operatorname{Im} T^* = K^\perp$ by Proposition A.6, we see that $G^*T^*h = h$ and $T^*h \in K^\perp$, and so T^* maps every right-hand side $h \in W^*$ to its unique solution $g = T^*h$. Hence we say that T solves the dual boundary problem (K, G), and we use the notation

$$T = (K, G)^{-1}.$$

Conversely, if there exists a left inverse T of G for a dual boundary problem (K, G) such that $\operatorname{Ker} T = K$, it is regular by (A.17). Given any left inverse $\tilde{T}$ of G, we know with Lemma A.10 that the dual Green's operator for a regular dual boundary problem (K, G) is given by $T = \tilde{T}(1 - P)$, where P is the projection with $\operatorname{Im} P = K$ and $\operatorname{Ker} P = I$.

If G is invertible, then $(0, G)$ is the only regular dual boundary problem with G and its dual Green's operator is $(0, G)^{-1} = G^{-1}$. In particular, we have

$$(0, 1)^{-1} = 1 \tag{2.5}$$

for the identity operator.

For fixed vector spaces V and W we denote the set of all regular (dual) boundary problems respectively by

$$R = \{(T, \mathcal{F}) \mid T\colon V \to W,\ (T, \mathcal{F}) \text{ regular}\}$$

and

$$R^* = \{(K, G) \mid G\colon W \to V,\ (K, G) \text{ regular}\}.$$

We can interpret the bijection (A.20) between left and right inverses in terms of boundary and dual boundary problems. The main part is always solving a (dual) regular boundary problem, that is, computing its (dual) Green's operator. Note that for boundary problem we specify a complement of the kernel by an orthogonally closed subspace of the dual space.

Proposition 2.1 *The map*

$$\begin{aligned} R &\to R^* \\ (T, \mathcal{F}) &\mapsto \left(\operatorname{Ker} T, (T, \mathcal{F})^{-1}\right) \end{aligned}$$

is a bijection between the sets of regular (dual) boundary problems, and

$$\begin{aligned} R^* &\to R \\ (K, G) &\mapsto \left((K, G)^{-1}, (\operatorname{Im} G)^{\perp}\right). \end{aligned}$$

is its inverse.

Proof Clear with Proposition A.11. □

3 Composing boundary problems

Let $(T_1, \mathcal{F}_1)$ and $(T_2, \mathcal{F}_2)$ be boundary problems with $T_1\colon V \to W$ and $T_2\colon U \to V$. We define the *composition* of $(T_1, \mathcal{F}_1)$ and $(T_2, \mathcal{F}_2)$ by

$$(T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2) = (T_1 T_2, T_2^*(\mathcal{F}_1) + \mathcal{F}_2). \tag{3.1}$$

Proposition 3.1 *The composition of boundary problems is again a boundary problem.*

Proof The composition of surjective maps is surjective. We must show that $T_2^*(\mathcal{F}_1) + \mathcal{F}_2$ is an orthogonally closed subspace of U^*. But from Corollary A.5 we know that the transpose maps orthogonally closed subspaces to orthogonally closed subspaces and from Proposition A.3 that the sum of two orthogonally closed subspaces is orthogonally closed.

□

The composition of boundary problems is associative. Moreover, we have

$$(1_V, 0) \circ (T, \mathcal{F}) = (T, \mathcal{F}) \quad \text{and} \quad (T, \mathcal{F}) \circ (1_W, 0) = (T, \mathcal{F})$$

with $T\colon V \to W$ and 0 the zero-dimensional vector space. So all boundary problems of vector spaces over a fixed field form a category with objects the vector spaces and morphisms the boundary problems.

The next proposition tells us that the composition of boundary problems preserves regularity, and the corresponding Green's operator is the composition of Green's operators in reverse order. Hence the regular boundary problems form a subcategory of the category of all boundary problems. We denote the *category of regular boundary problems* by $\mathcal{R}$.

Proposition 3.2 *Let $(T_1, \mathcal{F}_1)$ and $(T_2, \mathcal{F}_2)$ be regular boundary problems with Green's operators G_1 and G_2. Then the composition*

$$(T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2) = (T, \mathcal{F})$$

is regular with Green's operator $G_2 G_1$ so that

$$((T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2))^{-1} = (T_2, \mathcal{F}_2)^{-1} \circ (T_1, \mathcal{F}_1)^{-1}.$$

Moreover, the sum

$$\mathcal{F} = T_2^*(\mathcal{F}_1) \dotplus \mathcal{F}_2 \tag{3.2}$$

is direct.

Proof We have

$$T_1 T_2 G_2 G_1 = T_1 1 G_1 = T_1 G_1 = 1$$

so that $G_2 G_1$ is a right inverse of $T_1 T_2$. Since $\operatorname{Ker} G_1^* = \mathcal{F}_1$ and $\operatorname{Ker} G_2^* = \mathcal{F}_2$ by (2.2), we have with Proposition A.6 and (A.21)

$$(\operatorname{Im} G_2 G_1)^\perp = \operatorname{Ker} (G_2 G_1)^* = \operatorname{Ker} G_1^* G_2^* = T_2^*(\mathcal{F}_1) \dotplus \mathcal{F}_2.$$

The proposition now follows by the characterization of regular boundary problems through Green's operators. □

Note that with (A.15) and (A.5) we see that

$$T_2^*(\mathcal{F}_1^{\perp\perp}) + \mathcal{F}_2^{\perp\perp} = (T_2^*(\mathcal{F}_1) + \mathcal{F}_2)^{\perp\perp}$$

for arbitrary (not necessarily orthogonally closed) subspaces $\mathcal{F}_1$ and $\mathcal{F}_2$. If the boundary conditions are given by the orthogonal closure of arbitrary subspaces $\mathcal{F}_1$ and $\mathcal{F}_2$, the composition of two boundary problems is equal to

$$\left(T_1, \mathcal{F}_1^{\perp\perp}\right) \circ \left(T_2, \mathcal{F}_2^{\perp\perp}\right) = \left(T_1 T_2, (T_2^*(\mathcal{F}_1) + \mathcal{F}_2)^{\perp\perp}\right). \tag{3.3}$$

We will use this observation for boundary problems with partial differential equations in Sect. 7.

Let now (K_2, G_2) and (K_1, G_1) be dual boundary problems with $G_2 \colon V \to U$ and $G_1 \colon W \to V$. We define the *composition* of (K_2, G_2) and (K_1, G_1) by

$$(K_2, G_2) \circ (K_1, G_1) = (K_2 + G_2(K_1), G_2 G_1). \tag{3.4}$$

Obviously, the composition is again a dual boundary problem. It is associative, and we have

$$(0, 1_W) \circ (K, G) = (K, G) \quad \text{and} \quad (K, G) \circ (0, 1_V) = (K, G)$$

with $G \colon W \to V$. So all dual boundary problems of vector spaces over a fixed field form a category.

As we will see, also for dual boundary problems the composition of two regular problems is again regular. Hence the regular dual boundary problems form a subcategory of the category of all dual boundary problems. We denote the *category of regular dual boundary problems* by $\mathcal{R}^*$.

Proposition 3.3 *Let (K_2, G_2) and (K_1, G_1) be regular dual boundary problems with dual Green's operators T_2 and T_1. Then the composition*

$$(K_2, G_2) \circ (K_1, G_1) = (K, G)$$

is regular with dual Green's operator T_1T_2 so that

$$((K_2, G_2) \circ (K_1, G_1))^{-1} = (K_1, G_1)^{-1} \circ (K_2, G_2)^{-1}.$$

Moreover, the sum $K = K_2 \dotplus G_2(K_1)$ is direct.

Proof We have

$$T_1T_2G_2G_1 = T_1 1 G_1 = T_1G_1 = 1$$

so that T_1T_2 is a left inverse of G_2G_1. By (A.21), we have

$$\operatorname{Ker}(T_1T_2) = G_2(K_1) \dotplus K_2$$

with $K_1 = \operatorname{Ker} T_1$ and $K_2 = \operatorname{Ker} T_2$. The proposition follows now by the characterization of regular dual boundary problems through dual Green's operators. □

Summing up, we see that solving regular (dual) boundary problems gives an anti-isomorphism between the categories of regular (dual) boundary problems, justifying our terminology for dual boundary problems.

Theorem 3.4 *The contravariant functor*

$$\begin{aligned} F &: \mathcal{R} \to \mathcal{R}^* \\ (T, \mathcal{F}) &\mapsto \left(\operatorname{Ker} T, (T, \mathcal{F})^{-1}\right) \end{aligned}$$

is an anti-isomorphism between the categories of regular (dual) boundary problems, and

$$\begin{aligned} F^* &: \mathcal{R}^* \to \mathcal{R} \\ (K, G) &\mapsto \left((K, G)^{-1}, (\operatorname{Im} G)^{\perp}\right). \end{aligned}$$

is its inverse.

Proof By (2.4) and (2.5), we have $F(1) = 1$ as well as $F^*(1) = 1$. Hence F and F^* are contravariant functors by Propositions 3.2 and 3.3. Finally, $FF^* = 1$ and $F^*F = 1$ by Proposition 2.1. □

4 Factoring boundary problems

Let $(T, \mathcal{F})$ be a boundary problem with $T : U \to W$ and assume that we have a factorization

$$(T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2) = (T, \mathcal{F}) \tag{4.1}$$

into boundary problems with $T_1 : V \to W$ and $T_2 : U \to V$. By definition (3.1), this means that we have a factorization

$$T = T_1T_2$$

for the defining operators and a sum

$$\mathcal{F} = T_2^*(\mathcal{F}_1) + \mathcal{F}_2$$

for the boundary conditions. In this section, we characterize all possible factorizations of a boundary problem into two boundary problems. In particular, we show that if $(T, \mathcal{F})$ is regular and a factorization $T = T_1 T_2$ is fixed, there exists a unique regular left factor $(T_1, \mathcal{F}_1)$, and we describe all right factors $(T_2, \mathcal{F}_2)$.

Given a factorization $T = T_1 T_2$ with surjective linear maps T_1 and T_2, we construct all corresponding factorizations into (regular) boundary problems. The boundary conditions for the factor problems can be described in terms of the boundary conditions $\mathcal{F}$ and the factorization $T = T_1 T_2$. More precisely, we need $K_2 = \operatorname{Ker} T_2$ and an arbitrary right inverse of T_2, which we denote in this section by H_2. We begin without any assumption on the regularity.

Lemma 4.1 *Let $(T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2) = (T, \mathcal{F})$. Then*

$$T_2^*(\mathcal{F}_1) \subseteq \mathcal{F} \cap K_2^\perp \tag{4.2}$$

and

$$T_2^* H_2^*(\tilde{\mathcal{F}}_1) = \tilde{\mathcal{F}}_1 \tag{4.3}$$

for any $\tilde{\mathcal{F}}_1 \subseteq K_2^\perp$.

Proof Note that $\operatorname{Im} T_2^* = K_2^\perp$ by Proposition A.6 and $T_2^*(\mathcal{F}_1) \subseteq T_2^*(\mathcal{F}_1) + \mathcal{F}_2 = \mathcal{F}$. For the second equation observe that $T_2^* H_2^*$ is a projection with $\operatorname{Im} T_2^* H_2^* = \operatorname{Im} T_2^* = K_2^\perp$ by (A.16). □

Proposition 4.2 *Let $T = T_1 T_2$ be a factorization with surjective linear maps T_1 and T_2. Let*

$$\tilde{\mathcal{F}}_1 \subseteq \mathcal{F} \cap K_2^\perp \quad \text{and} \quad \mathcal{F}_2 \subseteq \mathcal{F}$$

be orthogonally closed subspaces such that $\mathcal{F} = \tilde{\mathcal{F}}_1 + \mathcal{F}_2$, and $\mathcal{F}_1 = H_2^*(\tilde{\mathcal{F}}_1)$. Then

$$(T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2) = (T, \mathcal{F})$$

is a factorization of $(T, \mathcal{F})$.

Proof By Corollary A.5, we know that $\mathcal{F}_1 = H_2^*(\tilde{\mathcal{F}}_1)$ is orthogonally closed, and so $(T_1, \mathcal{F}_1)$ is a boundary problem. Using (4.3), we observe

$$(T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2) = (T_1 T_2, T_2^* H_2^*(\tilde{\mathcal{F}}_1) + \mathcal{F}_2) = (T, \tilde{\mathcal{F}}_1 + \mathcal{F}_2) = (T, \mathcal{F}),$$

and the proposition is proved. □

Let now $(T, \mathcal{F})$ be regular with Green's operator G, and assume that we have a factorization $T = T_1 T_2$ with T_1 and T_2 surjective. Then $T_2 G$ is a right inverse of T_1 since

$$T_1 T_2 G = T G = 1.$$

So $(T_1, (\operatorname{Im} T_2 G)^\perp)$ is a regular boundary problem. We can describe its boundary conditions without G only in terms of $\mathcal{F}$ and T_2 with a right inverse H_2.

Lemma 4.3 *Let $(T, \mathcal{F})$ be regular with Green's operator G and let $T = T_1 T_2$ be a factorization with surjective linear maps T_1 and T_2. Then*

$$(\operatorname{Im} T_2 G)^\perp = H_2^*(\mathcal{F} \cap K_2^\perp),$$

and $(T_1, H_2^*(\mathcal{F} \cap K_2^\perp))$ is regular with Green's operator $T_2 G$.

Proof Using Proposition A.6 and (A.22), we obtain

$$(\operatorname{Im} T_2 G)^\perp = \operatorname{Ker} (T_2 G)^* = \operatorname{Ker} G^* T_2^* = H_2^*(\operatorname{Ker} G^* \cap \operatorname{Im} T_2^*).$$

From (2.2) we know that $\operatorname{Ker} G^* = \mathcal{F}$ and $\operatorname{Im} T_2^* = K_2^\perp$. □

The following theorem tells us that given a regular boundary problem $(T, \mathcal{F})$ and a factorization $T = T_1 T_2$, there is a unique regular left factor described by the previous lemma.

Theorem 4.4 *Let $(T, \mathcal{F})$ be regular and $T = T_1 T_2$ a factorization with surjective linear maps T_1 and T_2. Then*

$$(T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2) = (T, \mathcal{F})$$

is a factorization with $(T_1, \mathcal{F}_1)$ regular iff

$$\mathcal{F}_1 = H_2^*(\mathcal{F} \cap K_2^\perp)$$

and $\mathcal{F}_2 \subseteq \mathcal{F}$ is an orthogonally closed subspace such that

$$\mathcal{F} = (\mathcal{F} \cap K_2^\perp) + \mathcal{F}_2.$$

Moreover, if $(T_1, \mathcal{F}_1)$ is regular, its Green's operator is $T_2 G$.

Proof Let $(T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2) = (T, \mathcal{F})$ with $(T, \mathcal{F})$ and $(T_1, \mathcal{F}_1)$ regular. Writing $\bar{\mathcal{F}}_1 = H_2^*(\mathcal{F} \cap K_2^\perp)$, we see with Equation (4.2) that $\mathcal{F}_1 \subseteq \bar{\mathcal{F}}_1$. Since $(T_1, \mathcal{F}_1)$ is regular by assumption and $(T_1, \bar{\mathcal{F}}_1)$ by the previous lemma, we have

$$\mathcal{F}_1 \dotplus K_1^\perp = \bar{\mathcal{F}}_1 \dotplus K_1^\perp = V^*$$

by (2.1), so that $\mathcal{F}_1$ and $\bar{\mathcal{F}}_1$ have a common complement. Using modularity, we see that

$$\mathcal{F}_1 = \mathcal{F}_1 + (K_1^\perp \cap \bar{\mathcal{F}}_1) = (\mathcal{F}_1 + K_1^\perp) \cap \bar{\mathcal{F}}_1 = \bar{\mathcal{F}}_1 = H_2^*(\mathcal{F} \cap K_2^\perp).$$

By (4.3), we have $T_2^*(\mathcal{F}_1) = T_2^* H_2^*(\mathcal{F} \cap K_2^\perp) = \mathcal{F} \cap K_2^\perp$, and so

$$\mathcal{F} = (\mathcal{F} \cap K_2^\perp) + \mathcal{F}_2.$$

Conversely, we know by the previous lemma that $(T_1, H_2^*(\mathcal{F} \cap K_2^\perp))$ is regular, and $(T_1, H_2^*(\mathcal{F} \cap K_2^\perp)) \circ (T_2, \mathcal{F}_2) = (T, \mathcal{F})$ by Proposition 4.2. □

Finally, assume that *all* boundary problems in the factorization (4.1) are regular with corresponding Green's operators G, G_1 and G_2. Then we have the factorizations

$$T = T_1 T_2 \quad \text{and} \quad G = G_2 G_1,$$

by Proposition 3.2, and a direct sum of the boundary conditions

$$\mathcal{F} = T_2^*(\mathcal{F}_1) \dotplus \mathcal{F}_2$$

by (3.2). Since $T_2 G = T_2 G_2 G_1 = G_1$, we know from Lemma 4.3 that $\mathcal{F}_1 = H_2^*(\mathcal{F} \cap K_2^\perp)$. By (4.3), we obtain $T_2^*(\mathcal{F}_1) = \mathcal{F} \cap K_2^\perp$ so that

$$\mathcal{F} = (\mathcal{F} \cap K_2^\perp) \dotplus \mathcal{F}_2.$$

We write $\bar{\mathbb{P}}(V^*)$ for the lattice of orthogonally closed subspaces of V^*; see Appendix A.1 in the appendix. With the following proposition relating complements, subspaces and orthogonality, we can characterize all regular problems $(T_2, \mathcal{F}_2)$ with $\mathcal{F}_2 \subseteq \mathcal{F}$.

Proposition 4.5 *Let $K_2 \subseteq K \subseteq V$ be subspaces and $\mathcal{F} \subseteq V^*$ an orthogonally closed subspace such that*

$$V = K \dotplus \mathcal{F}^\perp.$$

Then we have a bijection

$$\begin{aligned}\{\mathcal{F}_2 \in \bar{\mathbb{P}}(V^*) \mid \mathcal{F}_2 \subseteq \mathcal{F} \text{ and } V = K_2 \dotplus \mathcal{F}_2^\perp\} \\ \cong \{V_2 \in \mathbb{P}(V) \mid K = V_2 \dotplus K_2\}\end{aligned}$$

given by

$$\mathcal{F}_2 \mapsto \mathcal{F}_2^\perp \cap K \quad \textit{and} \quad V_2 \mapsto \mathcal{F} \cap V_2^\perp. \tag{4.4}$$

Moreover,

$$V = K_2 \dotplus \mathcal{F}_2^\perp \quad \textit{iff} \quad \mathcal{F} = (\mathcal{F} \cap K_2^\perp) \dotplus \mathcal{F}_2,$$

for orthogonally closed subspaces $\mathcal{F}_2 \subseteq \mathcal{F}$.

Proof Let $\mathcal{F}_2 \subseteq \mathcal{F}$ be orthogonally closed such that $V = K_2 \dotplus \mathcal{F}_2^\perp$. We obtain

$$K = V \cap K = (K_2 + \mathcal{F}_2^\perp) \cap K = K_2 + (\mathcal{F}_2^\perp \cap K),$$

and the sum is direct since $K_2 \cap \mathcal{F}_2^\perp = 0$, so $\mathcal{F}_2^\perp \cap K$ is a complement of K_2 in K. Since $\mathcal{F} \cap K^\perp = 0$, we have

$$\mathcal{F} \cap (\mathcal{F}_2^\perp \cap K)^\perp = \mathcal{F} \cap (\mathcal{F}_2 + K^\perp) = \mathcal{F}_2 + (\mathcal{F} \cap K^\perp) = \mathcal{F}_2.$$

Conversely, let V_2 be a subspace such that $K = V_2 \dotplus K_2$. Since $V = K \dotplus \mathcal{F}^\perp$ and $(\mathcal{F} \cap V_2^\perp)^\perp = \mathcal{F}^\perp + V_2$, we have

$$V = K + \mathcal{F}^\perp = K_2 \dotplus (\mathcal{F}^\perp + V_2) = K_2 \dotplus (\mathcal{F} \cap V_2^\perp)^\perp.$$

Moreover, note that

$$(\mathcal{F} \cap V_2^\perp)^\perp \cap K = (V_2 + \mathcal{F}^\perp) \cap K = V_2 + (\mathcal{F}^\perp \cap K) = V_2$$

since $\mathcal{F}^\perp \cap K = 0$.

Now let $\mathcal{F}_2 \subseteq \mathcal{F}$ be orthogonally closed such that $V = K_2 \dotplus \mathcal{F}_2^\perp$. Let $V_2 = \mathcal{F}_2^\perp \cap K$. Then we know from above that $K = V_2 \dotplus K_2$, so

$$V = K \dotplus \mathcal{F}^\perp = V_2 \dotplus K_2 \dotplus \mathcal{F}^\perp.$$

Since orthogonality preserves direct sums, we obtain

$$V^* = (\mathcal{F} \cap K_2^\perp) \dotplus V_2^\perp.$$

So we have

$$\mathcal{F} = \mathcal{F} \cap V^* = \mathcal{F} \cap ((\mathcal{F} \cap K_2^\perp) + V_2^\perp) = (\mathcal{F} \cap K_2^\perp) + (\mathcal{F} \cap V_2^\perp),$$

and the sum is direct since $(\mathcal{F} \cap K_2^\perp) \cap V_2^\perp = 0$. Since we also know from above that $\mathcal{F} \cap V_2^\perp = \mathcal{F}_2$, the first part of the equivalence is proved.

Conversely, let $\mathcal{F}_2$ be an orthogonally closed subspace such that

$$\mathcal{F} = (\mathcal{F} \cap K_2^\perp) \dotplus \mathcal{F}_2.$$

Then $(\mathcal{F} \cap K_2^\perp) \cap \mathcal{F}_2 = 0$ and hence by passing to the orthogonal

$$V = K_2 + \mathcal{F}^\perp + \mathcal{F}_2^\perp = K_2 + \mathcal{F}_2^\perp,$$

the latter since $\mathcal{F}_2^\perp \supseteq \mathcal{F}^\perp$. Moreover, note that

$$\mathcal{F}^\perp = (\mathcal{F} \cap K_2^\perp)^\perp \cap \mathcal{F}_2^\perp = (\mathcal{F}^\perp + K_2) \cap \mathcal{F}_2^\perp = \mathcal{F}^\perp + (K_2 \cap \mathcal{F}_2^\perp).$$

Since $K \cap \mathcal{F}^\perp = 0$, we obtain

$$0 = K \cap (\mathcal{F}^\perp + (K_2 \cap \mathcal{F}_2^\perp)) = (K \cap \mathcal{F}^\perp) + (K_2 \cap \mathcal{F}_2^\perp) = K_2 \cap \mathcal{F}_2^\perp.$$

Hence $V = K_2 \dotplus \mathcal{F}_2^\perp$, and the proposition is proved. □

Corollary 4.6 *Let $(T, \mathcal{F})$ be regular and T_2 surjective with* $\operatorname{Ker} T_2 \subseteq \operatorname{Ker} T$. *Then* (4.4) *defines a bijection between*

$$\{\mathcal{F}_2 \subseteq \mathcal{F} \mid (T_2, \mathcal{F}_2) \text{ regular}\}$$

and complements of $\operatorname{Ker} T_2$ *in* $\operatorname{Ker} T$. *Moreover, $(T_2, \mathcal{F}_2)$ is regular iff $\mathcal{F}_2$ is an orthogonally closed complement of $(\mathcal{F} \cap K_2^\perp)$ in $\mathcal{F}$.*

The following corollary allows us to compute the boundary conditions for the unique regular left factor if we have the Green's operator for a regular right factor.

Corollary 4.7 *Let $(T, \mathcal{F})$ be regular and T_2 surjective with* $\operatorname{Ker} T_2 \subseteq \operatorname{Ker} T$. *Then*

$$G_2^*(\mathcal{F}) = G_2^*(\mathcal{F} \cap K_2^\perp)$$

if G_2 is the Green's operator for $(T_2, \mathcal{F}_2)$ regular with $\mathcal{F}_2 \subseteq \mathcal{F}$.

Proof If $G_2 = (T_2, \mathcal{F}_2)^{-1}$ with $\mathcal{F}_2 \subseteq \mathcal{F}$, then

$$\mathcal{F} = (\mathcal{F} \cap K_2^\perp) \dotplus \mathcal{F}_2,$$

by the previous corollary. Since $\operatorname{Ker} G_2^* = \mathcal{F}_2$ by (2.2), this implies $G_2^*(\mathcal{F}) = G_2^*(\mathcal{F} \cap K_2^\perp)$. □

Summing up, we can now characterize and construct all possible factorizations of a regular boundary problem into two regular boundary problems given a factorization of the defining operator.

Theorem 4.8 *Let $(T, \mathcal{F})$ be regular and $T = T_1 T_2$ a factorization with surjective linear maps T_1 and T_2. Then*

$$(T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2) = (T, \mathcal{F})$$

is a factorization with $(T_2, \mathcal{F}_2)$ regular iff

$$\mathcal{F}_1 = H_2^*\big(\mathcal{F} \cap K_2^\perp\big)$$

and $\mathcal{F}_2 \subseteq \mathcal{F}$ is an orthogonally closed subspace such that

$$\mathcal{F} = (\mathcal{F} \cap K_2^\perp) \dotplus \mathcal{F}_2.$$

In particular, the left factor $(T_1, \mathcal{F}_1)$ is necessarily regular.

Proof Let $(T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2) = (T, \mathcal{F})$ with $(T, \mathcal{F})$ and $(T_2, \mathcal{F}_2)$ regular. Let G_2 be the Green's operator for $(T_2, \mathcal{F}_2)$. Since $\operatorname{Ker} G_2^* = \mathcal{F}_2$ by (2.2) and $\mathcal{F} = T_2^*(\mathcal{F}_1) + \mathcal{F}_2$, we obtain $G_2^*(\mathcal{F}) = \mathcal{F}_1$. With the previous corollary this yields

$$\mathcal{F}_1 = G_2^*(\mathcal{F} \cap K_2^\perp),$$

and so $(T_1, \mathcal{F}_1)$ is regular by Lemma 4.3. The theorem follows with Corollary 4.6 and Theorem 4.4. □

5 A monoid of boundary problems

In this section, we consider boundary problems with *endomorphisms*; this case is also the basis for the symbolic computation treatment in [21]. Having endomorphisms, the composition of boundary problems (3.1) and dual boundary problems (3.4) coincides with the multiplication in a reverse semidirect product of suitably defined monoids and actions. Moreover, the contravariant functors from Theorem 3.4 between regular (dual) boundary problems specialize to anti-isomorphisms between the submonoids of regular (dual) boundary problems.

Given a monoid action, one can define the semidirect product of monoids just as for groups. In contrast to groups, one must distinguish between left and right actions and accordingly define the multiplication for semidirect products.

We recall the definitions. Let M and N be monoids. Following a convention introduced by Eilenberg [5], which also fits perfectly with our application, we write the product in M additively (without assuming commutativity in general). Given a left action of N on M, denoted by $n \cdot m$, and specified by a homomorphism $\varphi\colon N \to \operatorname{End} M$, the *semidirect product* $M \rtimes_\varphi N$ is the set $M \times N$ with the multiplication "from the left"

$$(m_1, n_1)(m_2, n_2) = (m_1 + n_1 \cdot m_2, n_1 n_2) = (m_1 + \varphi_{n_1}(m_2), n_1 n_2).$$

One verifies that this multiplication is associative with identity $(0, 1)$, so the semidirect product $M \rtimes_\varphi N$ is indeed a monoid.

Analogously, given a right action of N on M, denoted by $m \cdot n$, and specified by an anti-homomorphism $\varphi\colon N \to \operatorname{End} M$, the *reverse semidirect product* $N \ltimes_\varphi M$ is the set $N \times M$ with the multiplication "from the right"

$$(n_1, m_1)(n_2, m_2) = (n_1 n_2, m_1 \cdot n_2 + m_2) = (n_1 n_2, \varphi_{n_2}(m_1) + m_2).$$

Again $N \ltimes_\varphi M$ is a monoid with identity $(1, 0)$.

Let now V be a vector space and $L(V)$ the monoid of endomorphisms with respect to composition. The subspace lattice of V is denoted by $\mathbb{P}(V)$, and $L(V)$ acts on it from the left by $A \cdot V_1 = A(V_1)$, so we have a homomorphism $\varphi\colon L(V) \to \operatorname{End} \mathbb{P}(V)$ with $\varphi_A(V_1) = A(V_1)$. The multiplication in the semidirect product $\mathbb{P}(V) \rtimes_\varphi L(V)$ is

$$(V_1, A_1)(V_2, A_2) = (V_1 + A_1(V_2), A_1 A_2),$$

which is exactly the definition (3.4) of the composition of dual boundary problems. Writing H for the submonoid of all injective endomorphisms, we see that the semidirect product $\mathbb{P}(V) \rtimes_\varphi H$ is the *monoid of dual boundary problems*. The regular dual boundary problems form a submonoid

$$R^* = \{(K, G) \in \mathbb{P}(V) \times H \mid (K, G) \text{ regular}\}$$

since the composition of two regular dual boundary problems is regular by Proposition 3.3.

We now discuss the situation for boundary problems. By Proposition A.3, the sum of two orthogonally closed subspaces is orthogonally closed, so $\bar{\mathbb{P}}(V^*)$ is an additive monoid. We know from Corollary A.5 that the transpose maps orthogonally closed subspaces to orthogonally closed subspaces. Hence $L(V)$ acts on $\bar{\mathbb{P}}(V^*)$ from the right via the transpose $\mathcal{F} \cdot A = A^*(\mathcal{F})$, and we have the anti-homomorphism $\varphi\colon L(V) \to \operatorname{End} \bar{\mathbb{P}}(V^*)$ with $\varphi_A(\mathcal{F}) = A^*(\mathcal{F})$. The multiplication in the reverse semidirect product $L(V) \ltimes_\varphi \bar{\mathbb{P}}(V^*)$ is

$$(A_1, \mathcal{F}_1)(A_2, \mathcal{F}_2) = (A_1 A_2, A_2^*(\mathcal{F}_1) + \mathcal{F}_2),$$

which is the definition (3.1) of the composition of boundary problems. Writing S for the submonoid of all surjective endomorphisms, the reverse semidirect product $S \ltimes_\varphi \bar{\mathbb{P}}(V^*)$ is the *monoid of boundary problems*. The regular boundary problems form a submonoid

$$R = \{(T, \mathcal{F}) \in S \times \bar{\mathbb{P}}(V^*) \mid (T, \mathcal{F}) \text{ regular}\}$$

since the composition of two regular boundary problems is regular by Proposition 3.2.

Solving regular (dual) boundary problems gives an anti-isomorphism between the monoids of regular (dual) boundary problems. More precisely, we have the following result as a special case of Theorem 3.4.

Proposition 5.1 *The map*

$$\begin{aligned} R &\to R^* \\ (T, \mathcal{F}) &\mapsto \left(\operatorname{Ker} T, (T, \mathcal{F})^{-1}\right) \end{aligned}$$

is an anti-isomorphism between the monoids of regular (dual) boundary problems, and

$$\begin{aligned} R^* &\to R \\ (K, G) &\mapsto \left((K, G)^{-1}, (\operatorname{Im} G)^\perp\right). \end{aligned}$$

is its inverse.

Given a submonoid S_1 of all surjective endomorpisms S, we can consider the monoid of boundary problems $S_1 \ltimes \bar{\mathbb{P}}(V^*)$ with linear maps in S_1. We can also restrict the boundary conditions to a submonoid F of $\bar{\mathbb{P}}(V^*)$ if F is closed under S_1 in the sense that

$$T^*(\mathcal{F}) \in F \quad \text{for all } T \in S_1 \text{ and } \mathcal{F} \in F,$$

so that S_1 acts on F. In all such cases, the regular boundary problems form a submonoid. As an example, take the submonoid of surjective endomorphisms with finite dimensional kernel with finite dimensional subspaces of boundary conditions.

Analogously, we can consider submonoids of all injective endomorphisms and restrict the dual boundary conditions to suitable submonoids of $\mathbb{P}(V)$. The corresponding dual problems for the previous example are injective endomorphisms with finite codimensional image with finite dimensional subspaces as dual boundary conditions.

Note that with the results from Sect. 4, given a factorization in S_1, we can construct all factorizations of a (regular) boundary problem into (regular) boundary problems with arbitrary boundary conditions. If we restrict the boundary conditions to a submonoid F, we have to check whether the constructed boundary conditions are again in F.

6 Finitely many boundary conditions

In this section, we specialize some results and discuss algorithmic aspects for boundary problems where the corresponding linear maps have finite dimensional kernels and the spaces of boundary conditions are finite dimensional. Note that this includes boundary value problems for (systems of) ordinary differential equations and systems of partial differential equations with finite dimensional solution space.

More precisely, we consider boundary problems $(T, \mathcal{F})$ where $T\colon V \to W$,

$$\dim K < \infty \quad \text{and} \quad \mathcal{F} = [f_1, \ldots, f_n]$$

with $K = \operatorname{Ker} T$. We can rewrite the condition that $u \in V$ is a solution of the boundary problem $(T, \mathcal{F})$ for a given $w \in W$ in the following traditional form

$$\boxed{\begin{aligned} &Tu = w, \\ &f_1(u) = \cdots = f_n(u) = 0. \end{aligned}}$$

By Corollary A.17, a necessary condition for the regularity of $(T, \mathcal{F})$ is

$$\dim \operatorname{Ker} T = \dim \mathcal{F},$$

meaning that we have the "correct" number of boundary conditions. Moreover, we get the following algorithmic regularity test for boundary problems (to be found in Kamke [12, p. 184] for the special case of two-point boundary conditions).

Proposition 6.1 *A boundary problem $(T, \mathcal{F})$ with $\dim \operatorname{Ker} T = \dim \mathcal{F}$ is regular iff the matrix*

$$\begin{pmatrix} f_1(u_1) & \cdots & f_1(u_n) \\ \vdots & \ddots & \vdots \\ f_n(u_1) & \cdots & f_n(u_n) \end{pmatrix}$$

is regular, where the f_i and u_j are any basis of respectively $\mathcal{F}$ and $\operatorname{Ker} T$.

Let T be a fixed surjective linear map. By (2.3), given any right inverse $\tilde{G}$ of T, the Green's operator for a regular boundary problem $(T, \mathcal{F})$ is given by $G = (1 - P)\tilde{G}$, where P is the projection with $\operatorname{Im} P = K$ and $\operatorname{Ker} P = \mathcal{F}^\perp$. If T has a finite dimensional kernel with basis $u_1, \ldots, u_n$, we can easily describe the projection P in terms of a basis $f_1, \ldots, f_n$ of $\mathcal{F}$. Since the matrix $B = (f_i(u_j))$ is regular by the previous proposition, we can define

$$(\tilde{f}_1, \ldots, \tilde{f}_n)^t = B^{-1}(f_1, \ldots, f_n)^t.$$

Then the $(\tilde{f}_i)$ and (u_j) are biorthogonal, and $P\colon V \to V$ defined by

$$v \mapsto \sum_{i=1}^{n} \langle v, \tilde{f}_i \rangle u_i$$

is the projection with $\operatorname{Im} P = K$ and $\operatorname{Ker} P = \mathcal{F}^\perp$ by Lemma A.1.

Given a factorization $T = T_1 T_2$ and a right inverse H_2 of T_2, we know from Theorem 4.8 how to construct all possible factorizations of a regular boundary problem $(T, \mathcal{F})$ into two regular problems. The boundary conditions for the left factor $(T_1, \mathcal{F}_1)$ are uniquely given by

$$\mathcal{F}_1 = H_2^*(\mathcal{F} \cap K_2^\perp),$$

and all regular boundary problems $(T_2, \mathcal{F}_2)$ correspond to direct sums

$$\mathcal{F} = (\mathcal{F} \cap K_2^\perp) \dotplus \mathcal{F}_2.$$

In the following, we discuss how all such factorizations can be computed by linear algebra if T has a finite dimensional kernel.

Let $(T, \mathcal{F})$ be regular, $K = \operatorname{Ker} T$, $K_2 = \operatorname{Ker} T_2$, and $f_1, \ldots, f_{m+n}$ a basis of $\mathcal{F}$. Choose a basis

$$u_1, \ldots, u_m, u_{m+1}, \ldots, u_{m+n}$$

of K such that $u_1, \ldots, u_m$ is basis of K_2, and let

$$B = \begin{pmatrix} f_1(u_1) & \ldots & f_1(u_m) & f_1(u_{m+1}) & \ldots & f_1(u_{m+n}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ f_{m+n}(u_1) & \ldots & f_{m+n}(u_m) & f_{m+n}(u_{m+1}) & \ldots & f_{m+n}(u_{m+n}) \end{pmatrix}. \tag{6.1}$$

Since B is regular, we can perform row operations corresponding to a regular matrix P such that

$$P\,B = \begin{pmatrix} B_2 & C \\ 0 & D \end{pmatrix} \tag{6.2}$$

is a block matrix, where B_2 is a regular $m \times m$ matrix. Let

$$(\tilde{f}_1, \ldots, \tilde{f}_m, \tilde{f}_{m+1}, \ldots, \tilde{f}_{m+n})^t = P\,(f_1, \ldots, f_{m+n})^t, \tag{6.3}$$

that is,

$$\tilde{f}_i = \sum_{j=1}^{m+n} P_{ij} f_j,$$

and $\mathcal{F}_2 = [\tilde{f}_1, \ldots, \tilde{f}_m]$. Then we have obviously $[\tilde{f}_{m+1}, \ldots, \tilde{f}_{m+n}] \subseteq \mathcal{F} \cap K_2^\perp$ and since $\dim(\mathcal{F} \cap K_2^\perp) = \operatorname{codim}(\mathcal{F}^\perp + K_2) = n$, they are equal. So

$$\mathcal{F} = (\mathcal{F} \cap K_2^\perp) \dotplus \mathcal{F}_2$$

is a direct sum. Conversely, it is clear that any such direct sum given by bases $\mathcal{F}_2 = [\tilde{f}_1, \ldots, \tilde{f}_m]$ and $\mathcal{F} \cap K_2^\perp = [\tilde{f}_{m+1}, \ldots, \tilde{f}_{m+n}]$ with P as in (6.3) gives a block matrix as in (6.2). By Theorem 4.8, we know that

$$(T, \mathcal{F}) = (T_1, \mathcal{F}_1) \circ (T_2, \mathcal{F}_2)$$

is a factorization into regular boundary problems with

$$\mathcal{F}_1 = [H_2^*(\tilde{f}_{m+1}), \ldots, H_2^*(\tilde{f}_{m+n})] \quad \text{and} \quad \mathcal{F}_2 = [\tilde{f}_1, \ldots, \tilde{f}_m]. \tag{6.4}$$

Note that if H_2 is the Green's operator for a regular right factor $(T_2, \mathcal{F}_2)$ with $\mathcal{F}_2 \subseteq \mathcal{F}$, we have $H_2^*(\mathcal{F}) = H_2^*(\mathcal{F} \cap K_2^\perp)$ by Corollary 4.7. So we can compute the uniquely determined boundary conditions $\mathcal{F}_1$ simply by applying H_2^* to the boundary conditions $\mathcal{F}$; see the examples in Sect. 1.

7 Examples for differential equations

Let us now illustrate our algebraic approach to abstract boundary problems in the concrete setting of differential equations, taking up the examples posed in Sect. 1.

We want to factor the *two-point boundary problem* $(D^2, [L, R])$ of (1.1) into two regular problems with $T_1 = T_2 = D$. The indefinite integral $A = \int_0^x$ is the Green's operator for the regular right factor $(D, [L])$. By Corollary 4.7, the boundary conditions for the unique left factor are

$$A^*[L, R] = [0, RA] = [RA],$$

where $RA = \int_0^1$ is the definite integral. So we obtain the factorization

$$(D, [RA]) \circ (D, [L]) = (D^2, [L, R])$$

or

$$\boxed{\begin{array}{l} u' = f \\ \int_0^1 u(\xi)\, d\xi = 0 \end{array}} \circ \boxed{\begin{array}{l} u' = f \\ u(0) = 0 \end{array}} = \boxed{\begin{array}{l} u'' = f \\ u(0) = u(1) = 0 \end{array}}$$

in the notation from Sect. 1. Note that the boundary condition for the left factor is an integral condition. Such conditions are not considered in the classical setting of two-point boundary problems but are known in the literature as Stieltjes boundary conditions [1]. We check this factorization by multiplying the two boundary problems according to Definition (3.1). Note that

$$(D, [RA]) \circ (D, [L]) = (D^2, [D^*(RA), L])$$

and $D^*(RA) = RAD = \int_0^1 D = L - R$ so that

$$[D^*(RA), L] = [L - R, R] = [L, R],$$

as we expect.

To illustrate the method from the previous section, we factor the boundary problem $(D^2, [LD, R])$. We use again the indefinite integral $A = (D, [L])^{-1}$ as a right inverse of D, but for this boundary problem it is not a Green's operator for a regular right factor since $L \notin [LD, R]$. Hence we cannot simply apply A^* to the boundary conditions as we did before since this would give us two conditions

$$A^*[LD, R] = [LDA, RA] = [L, RA]$$

for a first-order problem. So we have to proceed as described in the previous section. A suitable basis for $\operatorname{Ker} D^2$ is $1, x$. Evaluating the boundary conditions LD, R on $1, x$ yields

$$\begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix},$$

for the matrix B from (6.1). Swapping the first and the second row gives a block triangular matrix as in (6.2). So by (6.4), the boundary condition is given by $A^*(LD) = L$ for the left factor and by R for the right factor, and we obtain the factorization

$$(D, [L]) \circ (D, [R]) = (D^2, [LD, R]).$$

See [21] for a general discussion on solving and factoring boundary problems for ordinary differential equations in an algorithmic context.

As an example of a boundary problem for a partial differential equation, we return to the *wave equation* (1.2) from Sect. 1. We write it as

$$\mathcal{W} = (\partial_t^2 - \partial_x^2, [u(x,0), u_t(x,0)]),$$

where $u(x,0)$ and $u_t(x,0)$ are short for the functionals $u \mapsto u(x,0)$ and $u \mapsto u_t(x,0)$, respectively, and $[\ldots]$ denotes the orthogonal closure of the subspace generated by these functionals with x ranging over $\mathbb{R}$. The Green's operator for $\mathcal{W}$ is given by

$$Gf(x,t) = \frac{1}{2} \int_0^t \int_{x-(t-\tau)}^{x+(t-\tau)} f(\xi,\tau)\,\mathrm{d}\xi\,\mathrm{d}\tau, \tag{7.1}$$

as can be found in the literature [23, p. 485]. We show that one can determine G by constructing a factorization of $\mathcal{W}$ along the factorization

$$\partial_t^2 - \partial_x^2 = (\partial_t - \partial_x)(\partial_t + \partial_x).$$

A regular right factor is given by

$$\mathcal{W}_2 = (\partial_t + \partial_x, [u(x,0)]).$$

In general, choosing boundary conditions in such a way that they make up a regular boundary problem for a given first-order right factor of a linear partial differential operator amounts to a geometric problem involving the characteristics. The Green's operator for $\mathcal{W}_2$ can easily be computed as

$$G_2 f(x,t) = \int_{x-t}^{x} f(\xi, \xi - x + t)\,\mathrm{d}\xi$$

and can be used for finding the boundary conditions for the uniquely determined left factor

$$\mathcal{W}_1 = (\partial_t - \partial_x, G_2^*[u(x,0), u_t(x,0)]) = (\partial_t - \partial_x, [u(x,0)])$$

by Corollary 4.7. One can verify the factorization $\mathcal{W} = \mathcal{W}_1 \circ \mathcal{W}_2$, taking into account (3.3). The Green's operator for $\mathcal{W}_1$ is analogously given by

$$G_1 f(x,t) = \int_x^{x+t} f(\xi, x - \xi + t)\,\mathrm{d}\xi,$$

and all we have to do now is to compute the composite

$$G_2 G_1 f(x,t) = \int_{x-t}^{x} \int_{\tau}^{2\tau - x + t} f(\xi, 2\tau - \xi - x + t)\,\mathrm{d}\xi\,\mathrm{d}\tau,$$

which is the Green's operator for $\mathcal{W}$ by Theorem 4.8. Since G and $G_2 G_1$ solve the same regular boundary problem, we know that $G = G_2 G_1$, as one may also verify directly by a change of variables.

The above methodology can also be transferred to the computationally more involved case of the wave equation on the *bounded interval* $[0, 1]$, succinctly expressed in our notation by

$$\mathcal{V} = (\partial_t^2 - \partial_x^2, [u(x,0), u_t(x,0), u(0,t), u(1,t)])$$

with x ranging over $[0, 1]$. In a similar fashion, one can find a factorization $\mathcal{V} = \mathcal{V}_1 \circ \mathcal{V}_2$ with

$$\mathcal{V}_1 = (\partial_t - \partial_x, [u(x,0), \int_{\max(1-t,0)}^{1} u(\xi, \xi + t - 1)\,\mathrm{d}\xi]),$$
$$\mathcal{V}_2 = (\partial_t + \partial_x, [u(x,0), u(0,t)]).$$

Unlike in the unbounded case, the Green's operator for $\mathcal{V}$ involves a finite sum whose upper bound depends on the argument (x, t). These complications are reflected in the Green's operator for the left factor $\mathcal{V}_1$, whose computation leads to a simple functional equation. A systematic investigation of partial differential equations with integral boundary conditions is a subject of future work.

Acknowledgments The authors would like to thank their project leaders Bruno Buchberger and Heinz W. Engl for continuous support, critical comments, and helpful suggestions. The authors also extend their thanks to Ulrich Oberst for pointing out the reference [13] and to Herwig Hauser for advice on an earlier version of the paper.

Appendix

A.1 Orthogonally closed subspaces

In this section, we summarize the results needed for orthogonally closed subspaces of a vector space and its dual. The notation should remind of the analogous well-known results for Hilbert spaces. See for example Conway [4] and Lang [14, pp. 391–394] for the Banach space setting.

First we recall the notion of orthogonality for a bilinear map of modules. Let M and N be left modules over a commutative ring R and $b\colon M \times N \to R$ be a bilinear map. Two vectors $x \in M$ and $y \in N$ are called *orthogonal* with respect to b if $b(x, y) = 0$. Let $X^\perp$ denote the set of all $y \in N$ that are orthogonal to X for a fixed bilinear map b. This is obviously a submodule of N, which we call the *orthogonal* of X. We define orthogonality on the other side in the same way.

It follows directly from the definition that for any subsets $X_1, X_2 \subseteq M$ we have

$$X_1 \subseteq X_2 \Rightarrow X_1^\perp \supseteq X_2^\perp \quad \text{and} \quad X_1 \subseteq X_1^{\perp\perp}. \tag{A.1}$$

These statements hold analogously for subsets of N. Let $\mathbb{P}(M)$ denote the *projective geometry* of a module M, that is, the poset of all submodules (ordered by inclusion). Then the two properties (A.1) for orthogonality imply that we have an order-reversing Galois connection between the projective geometries $\mathbb{P}(M) \rightleftarrows \mathbb{P}(N)$ defined by

$$M_1 \mapsto M_1^\perp \quad \text{and} \quad N_1 \mapsto N_1^\perp. \tag{A.2}$$

Hence we know in particular that $S^\perp = S^{\perp\perp\perp}$ for any submodule S of M or N. Moreover, the map $S \mapsto S^{\perp\perp}$ is a closure operator: an extensive ($S \subseteq S^{\perp\perp}$), order-preserving and idempotent self-map. We call a submodule S *orthogonally closed* if $S = S^{\perp\perp}$. The Galois connection restricted to orthogonally closed submodules is an order-reversing bijection. For further details and references on Galois connections we refer to Erné et al. [7].

We now consider the *canonical bilinear form* $V \times V^* \to k$ of a vector space V over a field k and its dual V^* defined by $(v, f) \mapsto f(v)$ and the induced orthogonality on the subspaces. We use the notation $\langle v, f \rangle$ for $f(v)$.

Let $V_1 \subseteq V$ be a subspace. Using the fact that any basis of a subspace can be extended to a basis for V, we see that for any vector $v \in V$ that is not in V_1 there is a linear form $f \in V^*$ with $f(v_1) = 0$ for all $v_1 \in V_1$ and $f(v) = 1$. It follows immediately that every subspace of V is orthogonally closed with respect to the canonical bilinear form $V \times V^* \to k$. Furthermore, we have a natural isomorphism

$$V_1^\perp \cong (V/V_1)^*.$$

Indeed, each $f \in V_1^\perp$ defines a linear form on V/V_1 since it vanishes on V_1, and it is easy to see that this gives an isomorphism between $V_1^\perp$ and $(V/V_1)^*$. This implies in particular that

$$\dim V_1^\perp = \operatorname{codim} V_1 \quad \text{if } \operatorname{codim} V_1 < \infty.$$

In the following, we consider subspaces of the dual vector space V^*. We first recall some results for biorthogonal systems. Two families $(v_i)_{i\in I}$ of vectors in V and linear forms $(f_i)_{i\in I}$ in V^* are called *biorthogonal* or said to form a *biorthogonal system* if

$$\langle v_i, f_j\rangle = \delta_{ij} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

For a biorthogonal system $(v_i)_{i\in I}$ and $(f_i)_{i\in I}$ we can easily compute the coefficients of a linear combination $v = \sum a_i v_i$ with finitely many $a_i \in k$ nonzero. Applying f_j, we obtain

$$\langle v, f_j\rangle = \sum a_i \langle v_i, f_j\rangle = a_j.$$

Evaluating a linear combination $f = \sum a_j f_j$ at v_i gives analogously

$$\langle v_i, f\rangle = \sum a_j \langle v_i, f_j\rangle = a_i.$$

This implies in particular that the v_i and f_i are linearly independent. Moreover, we can easily compute projections onto finite dimensional vector spaces from a biorthogonal system. One can show the following lemma and proposition for finite biorthogonal systems; for details see Köthe [13, pp. 71–72].

Lemma A.1 *Let $(v_1, \ldots, v_n) \in V$ and $(f_1, \ldots, f_n) \in V^*$ be biorthogonal. Let $V_1 = [v_1, \ldots, v_n]$ and $\mathcal{F}_1 = [f_1, \ldots, f_n]$ be their linear spans. Then $P\colon V \to V$ defined by*

$$v \mapsto \sum_{i=1}^{n} \langle v, f_i\rangle v_i$$

is a projection with $\operatorname{Im} P = V_1$ *and* $\operatorname{Ker} P = \mathcal{F}_1^\perp$ *so that* $V = \mathcal{F}_1^\perp \dotplus V_1$ *and* $\operatorname{codim} \mathcal{F}_1^\perp = n$. *Moreover, for any $f \in \mathcal{F}_1^{\perp\perp}$ we have*

$$f = \sum_{i=1}^{n} \langle v_i, f\rangle f_i,$$

so that $\mathcal{F}_1$ is orthogonally closed.

Proposition A.2 *Let $f_1, \ldots, f_n \in V^*$. Then the f_i are linearly independent iff there exist $v_1, \ldots, v_n \in V$ such that (v_i) and (f_i) are biorthogonal.*

We conclude with the previous lemma that every finite dimensional subspace of V^* is orthogonally closed. But if V is infinite dimensional, there are always linear subspaces, and indeed hyperplanes in V^*, that are not orthogonally closed; see e.g. [13, p. 71]. Nevertheless, since all subspaces of V are orthogonally closed, we have via the Galois connection (A.2) an order-reversing bijection between $\mathbb{P}(V)$ and the poset of all orthogonally closed subspaces of V^*, which we denote by $\bar{\mathbb{P}}(V^*)$.

Recall that the projective geometry $\mathbb{P}(V)$ of any vector space V is a complete complemented modular lattice with the join and meet respectively defined as the sum and intersection of subspaces. Modularity means that

$$V_1 + (V_2 \cap V_3) = (V_1 + V_2) \cap V_3$$

for all $V_1, V_2, V_3 \in \mathbb{P}(V)$ with $V_1 \subseteq V_3$.

Using (A.1) one can show that $\bar{\mathbb{P}}(V^*)$ is a complete lattice with the meet defined as the intersection and the join defined as the orthogonal closure of the sum of subspaces. Hence the Galois connection (A.2) is an order-reversing lattice isomorphism between the complete lattices $\mathbb{P}(V)$ and $\bar{\mathbb{P}}(V^*)$. Therefore $\bar{\mathbb{P}}(V^*)$ is also a complemented modular lattice.

Let $V_1, V_2 \in \mathbb{P}(V)$ and $\mathcal{F}_1, \mathcal{F}_2 \in \bar{\mathbb{P}}(V^*)$. Since the meet in $\bar{\mathbb{P}}(V^*)$ is the set-theoretic intersection, we know that

$$(V_1 + V_2)^\perp = V_1^\perp \cap V_2^\perp \quad \text{and} \quad (\mathcal{F}_1 \cap \mathcal{F}_2)^\perp = \mathcal{F}_1^\perp + \mathcal{F}_2^\perp. \tag{A.3}$$

The sum of infinitely many orthogonally closed subspaces is in general not orthogonally closed when V is infinite dimensional. But using the fact that $\bar{\mathbb{P}}(V^*)$ is a modular lattice, one can show the following proposition [13, p. 72].

Proposition A.3 *The sum of two orthogonally closed subspaces is orthogonally closed.*

Hence we have also

$$(V_1 \cap V_2)^\perp = V_1^\perp + V_2^\perp \quad \text{and} \quad (\mathcal{F}_1 + \mathcal{F}_2)^\perp = \mathcal{F}_1^\perp \cap \mathcal{F}_2^\perp. \tag{A.4}$$

Equations (A.3) and (A.4) imply that orthogonality preserves algebraic complements, that is, for direct sums

$$V = V_1 \dotplus V_2 \quad \text{and} \quad V^* = \mathcal{F}_1 \dotplus \mathcal{F}_2,$$

we have

$$V^* = V_1^\perp \dotplus V_2^\perp \quad \text{and} \quad V = \mathcal{F}_1^\perp \dotplus \mathcal{F}_2^\perp.$$

Every subspace has a complement, hence every orthogonally closed subspace of the dual has an orthogonally closed complement. So if we disregard completeness, the Galois connection (A.2) is an order-reversing lattice isomorphism between the complemented modular lattices $\mathbb{P}(V) \cong \bar{\mathbb{P}}(V^*)$ with join and meet defined as sum and intersection.

Moreover, for arbitrary (not necessarily orthogonally closed) subspaces $\mathcal{F}_1$ and $\mathcal{F}_2$ of V^* we have

$$\mathcal{F}_1^{\perp\perp} + \mathcal{F}_2^{\perp\perp} = (\mathcal{F}_1 + \mathcal{F}_2)^{\perp\perp}. \tag{A.5}$$

Using the fact that taking the double orthogonal is a closure operator, we see namely that $\mathcal{F}_1^{\perp\perp} + \mathcal{F}_2^{\perp\perp} \subseteq (\mathcal{F}_1 + \mathcal{F}_2)^{\perp\perp}$; the reverse inclusion follows since the left hand side of (A.5) is orthogonally closed by Proposition A.3. If $^{\perp\perp}$ were the closure operator of a topology, (A.5) would mean that the sum is continuous and closed.

We have already seen that if $\operatorname{codim} V_1 < \infty$ and $\dim \mathcal{F}_1 < \infty$, then

$$\operatorname{codim} V_1 = \dim V_1^\perp \quad \text{and} \quad \dim \mathcal{F}_1 = \operatorname{codim} \mathcal{F}_1^\perp. \tag{A.6}$$

So we can also consider the restriction of the Galois connection to finite codimensional subspaces of V and finite dimensional subspaces of V^*. This yields an order-reversing lattice isomorphism between modular lattices.

A.2 The transpose

Let V and W be vector spaces over a field k and $A\colon V \to W$ a linear map. We recall some basic properties of the *transpose* or *dual* map $A^*\colon W^* \to V^*$ defined by $h \mapsto h \circ A$. Hence

$$\langle Av, h\rangle_W = \langle v, A^*h\rangle_V \quad \text{for all } v \in V,\ h \in W^* \tag{A.7}$$

with the canonical bilinear forms on W and V, respectively. The map $A \mapsto A^*$ from $L(V, W)$ to $L(W^*, V^*)$ is linear. It is injective since for every nonzero $w \in W$ there exists a linear form $h \in W^*$ with $h(w) \neq 0$. For finite dimensional vector spaces, it is also surjective. We have $(AB)^* = B^*A^*$ for linear maps $A \in L(U, V)$ and $B \in L(V, W)$. Since $1_V{}^* = 1_{V^*}$, this implies that if A is left (respectively right) invertible, A^* is right (respectively left) invertible, so if A is invertible, also A^* is invertible with $(A^*)^{-1} = (A^{-1})^*$. Moreover, the map $A \mapsto A^*$ is an injective k-algebra anti-homomorphism from $L(V)$ to $L(V^*)$.

In the following, we discuss the relations between the image of subspaces under a linear map, its transpose, and orthogonality. From (A.7) it follows immediately that the orthogonal of the image of a subspace $V_1 \subseteq V$ is

$$A(V_1)^\perp = (A^*)^{-1}(V_1^\perp). \tag{A.8}$$

Since $V^\perp = 0$, we have in particular $(\operatorname{Im} A)^\perp = \operatorname{Ker} A^*$. Hence $\operatorname{Ker} A^*$ is orthogonally closed. Taking the orthogonal, we obtain from (A.8)

$$A(V_1) = (A^*)^{-1}(V_1^\perp)^\perp,$$

since every subspace of a vector space is orthogonally closed with respect to the canonical bilinear form. In particular, we have $\operatorname{Im} A = (\operatorname{Ker} A^*)^\perp$. For orthogonally closed subspaces $\mathcal{F}_1 \subseteq V^*$, we obtain

$$A(\mathcal{F}_1^\perp) = (A^*)^{-1}(\mathcal{F}_1)^\perp. \tag{A.9}$$

Now we consider the images under the transpose. Again we see immediately with (A.7) that

$$A^*(\mathcal{H}_1)^\perp = A^{-1}(\mathcal{H}_1^\perp) \tag{A.10}$$

for subspaces $\mathcal{H}_1 \subseteq W^*$. Since $(W^*)^\perp = 0$, we have in particular $(\operatorname{Im} A^*)^\perp = \operatorname{Ker} A$. Taking the orthogonal, we obtain from (A.10)

$$A^*(\mathcal{H}_1) \subseteq A^*(\mathcal{H}_1)^{\perp\perp} = A^{-1}(\mathcal{H}_1^\perp)^\perp. \tag{A.11}$$

Note that in general we have a proper inclusion, as one can see by taking the identity map and a subspace that is not orthogonally closed since the right-hand side is orthogonally closed. But we do have equality for orthogonally closed subspaces. In the Banach space setting, identity (A.13) comes in the context of the Closed Range Theorem [27, p. 205] and holds only for operators with closed range.

Proposition A.4 *We have*

$$A^*(W_1^\perp) = A^{-1}(W_1)^\perp \tag{A.12}$$

for subspaces $W_1 \subseteq W$. In particular,

$$\operatorname{Im} A^* = (\operatorname{Ker} A)^\perp, \tag{A.13}$$

and the image of A^* is orthogonally closed.

Proof With (A.11) and the fact that every subspace a vector space is orthogonally closed with respect to the canonical bilinear form, we know the inclusion $\subseteq$. Conversely, let $f \in A^{-1}(W_1)^\perp$. Then

$$f(v_1) = 0 \quad \text{for all } v_1 \in V \text{ such that } Av_1 \in W_1.$$

So in particular $f(\operatorname{Ker} A) = 0$. We have to find a $h \in W_1^\perp$ such that $f = A^*h$. We define $\tilde{h}\colon \operatorname{Im} A \to K$ by $\tilde{h}(Av) = f(v)$. Then $\tilde{h}$ is well-defined. If $Av_1 = Av_2$, then $v_1 - v_2 \in \operatorname{Ker} A$. Hence $f(v_1) = f(v_2)$ since $f(\operatorname{Ker} A) = 0$. Moreover, note that

$$\tilde{h}(\operatorname{Im} A \cap W_1) = 0.$$

We have to extend $\tilde{h}$ to a linear map $h\colon W \to K$ such that h vanishes on W_1. To this end, let $\tilde{I}_1$ and $\tilde{W}_1$ be complements of $\operatorname{Im} A \cap W_1$ in $\operatorname{Im} A$ and W_1, respectively, so that

$$\operatorname{Im} A = (\operatorname{Im} A \cap W_1) \dotplus \tilde{I}_1 \quad \text{and} \quad W_1 = (\operatorname{Im} A \cap W_1) \dotplus \tilde{W}_1.$$

Then one sees that we have a direct sum

$$\operatorname{Im} A + W_1 = (\operatorname{Im} A \cap W_1) \dotplus \tilde{I}_1 \dotplus \tilde{W}_1.$$

Let $P\colon \operatorname{Im} A + W_1 \to \operatorname{Im} A$ defined by

$$P(\bar{w} + \tilde{w}_1) = \bar{w} \quad \text{where } \bar{w} \in \operatorname{Im} A \text{ and } \tilde{w}_1 \in \tilde{W}_1.$$

Then P is a linear map with $\operatorname{Ker} P = \tilde{W}_1$. We set $h = \tilde{h} \circ P$. Then h is defined on $\operatorname{Im} A + W_1$. We extend h arbitrarily to a linear form on W and denote it again by h. By definition $h = \tilde{h}$ on $\operatorname{Im} A$, and so $f = A^*h$. We have to verify that $h \in W_1^\perp$. Let $w_1 \in W_1$ and

$$w_1 = \bar{w}_1 + \tilde{w}_1 \quad \text{with } \bar{w}_1 \in \operatorname{Im} A \cap W_1 \text{ and } \tilde{w}_1 \in \tilde{W}_1.$$

Then

$$h(w_1) = \tilde{h}(Pw_1) = \tilde{h}(\bar{w}_1) = 0$$

since $\tilde{h}(\operatorname{Im} A \cap W_1) = 0$, and the proposition is proved. □

We know from Appendix A.1 that the Galois connection (A.2) gives an isomorphism between $\mathbb{P}(W)$ and the orthogonally closed subspaces $\bar{\mathbb{P}}(W^*)$. So the previous proposition implies

$$A^*(\mathcal{H}_1) = A^{-1}(\mathcal{H}_1^\perp)^\perp \tag{A.14}$$

for orthogonally closed subspaces $\mathcal{H}_1 \subseteq W^*$. Since the right hand side is orthogonally closed, we obtain the following corollary.

Corollary A.5 *The transpose gives an order-preserving map*

$$\begin{aligned}\bar{\mathbb{P}}(W^*) &\to \bar{\mathbb{P}}(V^*)\\ \mathcal{H}_1 &\mapsto A^*(\mathcal{H}_1)\end{aligned}$$

between orthogonally closed subspaces.

Moreover, using (A.14) and (A.10), we see that

$$A^*\big(\mathcal{H}_1^{\perp\perp}\big) = A^{-1}(\mathcal{H}_1^{\perp})^{\perp} = A^*(\mathcal{H}_1)^{\perp\perp} \tag{A.15}$$

for an arbitrary subspace $\mathcal{H}_1 \subseteq W^*$, which means that A^* is "closed" and "continuous" in the hypothetical topological interpretation mentioned after (A.5).

Finally, we sum up all the identities for the image of subspaces of a linear map and its transpose and orthogonality in the following proposition.

Proposition A.6 *Let V and W be vector spaces over a field k and $A\colon V \to W$ a linear map. Then we have*

$$\begin{aligned}A(V_1)^{\perp} &= (A^*)^{-1}(V_1^{\perp}), \quad A(\mathcal{F}_1^{\perp}) = (A^*)^{-1}(\mathcal{F}_1)^{\perp},\\ A^*(\mathcal{H}_1)^{\perp} &= A^{-1}(\mathcal{H}_1^{\perp}), \quad A^*(W_1^{\perp}) = A^{-1}(W_1)^{\perp},\end{aligned}$$

for subspaces $V_1 \subseteq V$, $\mathcal{H}_1 \subseteq W^*$, $W_1 \subseteq W$ and orthogonally closed subspaces $\mathcal{F}_1 \subseteq V^*$. In particular, we have

$$\begin{aligned}(\operatorname{Im} A)^{\perp} &= \operatorname{Ker} A^*, \quad \operatorname{Im} A = (\operatorname{Ker} A^*)^{\perp},\\ (\operatorname{Im} A^*)^{\perp} &= \operatorname{Ker} A, \quad \operatorname{Im} A^* = (\operatorname{Ker} A)^{\perp},\end{aligned}$$

for the image and kernel of A and A^*.

Proof See (A.8), (A.9), (A.10), and (A.12). □

A.3 Left and right inverses

In this section, we recall and discuss some results for left and right inverses and their relation to projections, complements and inverse images.

Let V and W be vector spaces over a field k. Let $T\colon V \to W$ and $G\colon W \to V$ be linear maps such that $TG = 1$. Then T is surjective and G injective, respectively, and GT is a projection with

$$\operatorname{Ker} GT = \operatorname{Ker} T \quad \text{and} \quad \operatorname{Im} GT = \operatorname{Im} G, \tag{A.16}$$

so that

$$V = \operatorname{Ker} T \dotplus \operatorname{Im} G. \tag{A.17}$$

Conversely, we can begin with a given surjective or injective linear map and a complement of the kernel and image, respectively, and ask if there exists a corresponding right or left inverse. This is a special case of algebraic generalized inverses as in Nashed and Votruba [15]. We discuss the results for both cases.

Let first $T\colon V \to W$ be a surjective linear map with $K = \operatorname{Ker} T$ and I a complement of K in V, so that

$$V = K \dotplus I.$$

Let P be the projection with $\operatorname{Im} P = K$ and $\operatorname{Ker} P = I$. Then by [15, Theorem 1.20] there exists a unique linear map $G\colon W \to V$ with

$$TG = 1, \quad GT = 1 - P, \quad \text{and} \quad GTG = G.$$

Lemma A.7 *The equation $GT = 1 - P$ characterizes G uniquely.*

Proof The third equation above is obviously redundant, and we show that the first follows from the second. We get for $w = Tv$

$$TGw = TGTv = T(v - Pv) = Tv = w$$

since $\operatorname{Im} P = \operatorname{Ker} T$. So $TG = 1$ since T is surjective. □

We can also say that given a complement I of $K = \operatorname{Ker} T$, there exists a unique right inverse G with $\operatorname{Im} G = I$. So we have a bijection

$$\{I \in \mathbb{P}(V) \mid V = K + I\} \cong \{G \in L(W, V) \mid TG = 1\} \tag{A.18}$$

between the set of complements of K in V and the set of right inverses of T. Moreover, all right inverses can be described in terms of a fixed one.

Lemma A.8 *Given any right inverse $\tilde{G}$ of T, the right inverse corresponding to the complement I is given by*

$$G = (1 - P)\tilde{G},$$

where P is the projection with $\operatorname{Im} P = K$ *and* $\operatorname{Ker} P = I$.

Let now $G\colon W \to V$ be an injective linear map with $I = \operatorname{Im} G$ and K a complement of I in V, so that

$$V = K \dotplus I.$$

Let P be the projection with $\operatorname{Im} P = K$ and $\operatorname{Ker} P = I$. Since $\operatorname{Im}(1 - P) = \operatorname{Ker} P = I$, there exists by [15, Theorem 1.20] a unique linear map $T\colon V \to W$ with

$$GT = 1 - P, \quad TG = 1, \quad \text{and} \quad TGT = T.$$

Lemma A.9 *The equation $GT = 1 - P$ characterizes T uniquely.*

Proof Note first that since G is injective $\operatorname{Ker} T = \operatorname{Ker} GT = \operatorname{Ker}(1 - P) = K$. Therefore $TGT = T - TP = T$, which is the third equation above, and hence $TG = (TG)^2$ is a projection. We show that $\operatorname{Ker} TG = 0$, and so TG is the identity. Let $TGw = 0$. Then

$$GTGw = (1 - P)Gw = 0,$$

so that $Gw = PGw$. Since $\operatorname{Ker} P = \operatorname{Im} G$, this implies $Gw = 0$, and thus $w = 0$ because G is injective. □

We can also say that given a complement K of $I = \operatorname{Im} G$, there exists a unique left inverse T with $\operatorname{Ker} T = K$. So we have a bijection

$$\{K \in \mathbb{P}(V) \mid V = K \dotplus I\} \cong \{T \in L(V, W) \mid TG = 1\} \tag{A.19}$$

between the set of complements of I in V and the set of left inverses of G. Analogously as above one can describe all left inverses in terms of a fixed one.

Lemma A.10 *Given any left inverse $\tilde{T}$ of G, the left inverse corresponding to the complement K is given by*

$$T = \tilde{T}(1 - P),$$

where P is the projection with $\operatorname{Im} P = K$ *and* $\operatorname{Ker} P = I$.

Summing up, the bijections (A.18) and (A.19) yield with Lemma A.7 and A.9 the following proposition.

Proposition A.11 *We have a bijection*

$$\begin{aligned}&\{(T, I) \mid T\colon V \to W \text{ surjective, } I \in \mathbb{P}(V) \text{ with } V = \operatorname{Ker} T \dotplus I\}\\ &\cong \{(K, G) \mid G\colon W \to V \text{ injective, } K \in \mathbb{P}(V) \text{ with } V = K \dotplus \operatorname{Im} G\}. \qquad \text{(A.20)}\end{aligned}$$

Given respectively (T, I) or (K, G), we obtain G or T with $TG = 1$ as the solution of

$$GT = 1 - P,$$

where P is the projection with

$$\operatorname{Im} P = \operatorname{Ker} T,\ \operatorname{Ker} P = I \quad \text{and} \quad \operatorname{Im} P = K,\ \operatorname{Ker} P = \operatorname{Im} G,$$

respectively.

The following two propositions describe the inverse image of a composition of an arbitrary and respectively a surjective or injective linear map in terms of one of its right or left inverses.

Proposition A.12 *Let U, V, W be vector spaces over a field k. Let $A \in L(V, W)$ be arbitrary, $T \in L(U, V)$ surjective, G a right inverse of T, and $W_1 \subseteq W$ a subspace. Then we have*

$$(AT)^{-1}(W_1) = G(A^{-1}(W_1)) \dotplus \operatorname{Ker} T$$

for the inverse image of the composite. In particular, we have

$$\operatorname{Ker} AT = G(\operatorname{Ker} A) \dotplus \operatorname{Ker} T \qquad \text{(A.21)}$$

for the kernel of the composite and

$$T^{-1}(V_1) = G(V_1) \dotplus \operatorname{Ker} T$$

for the inverse image.

Proof One inclusion is obvious, since

$$AT(G(A^{-1}(W_1)) + \operatorname{Ker} T) = A(A^{-1}(W_1)) + 0 \subseteq W_1.$$

Conversely, let $u \in (AT)^{-1}(W_1)$. Then $Tu = v$ with $v \in A^{-1}(W_1)$. Hence

$$T(u - Gv) = Tu - v = 0$$

and therefore $u \in G(A^{-1}(W_1)) + \operatorname{Ker}(T)$. The sum is direct by (A.17). □

Proposition A.13 *Let U, V, W be vector spaces over a field k. Let $A \in L(V, W)$ be arbitrary, $G \in L(U, V)$ injective, T a left inverse of G, and $W_1 \subseteq W$ a subspace. Then we have*

$$(AG)^{-1}(W_1) = T(A^{-1}(W_1) \cap \operatorname{Im} G)$$

for the inverse image of the composite. In particular, we have

$$\operatorname{Ker} AG = T(\operatorname{Ker} A \cap \operatorname{Im} G) \tag{A.22}$$

for the kernel of the composite and

$$G^{-1}(V_1) = T(V_1 \cap \operatorname{Im} G)$$

for the inverse image.

Proof Let $v \in A^{-1}(W_1) \cap \operatorname{Im} G$. Since GT is a projection with $\operatorname{Im} GT = \operatorname{Im} G$, see (A.16), we get $AGTv = Av \in W_1$, and one inclusion is proved.

Conversely, let $u \in (AG)^{-1}(W_1)$. Then $Gu = v$ with $v \in A^{-1}(W_1) \cap \operatorname{Im} G$. Hence $TGu = u = Tv$, and therefore $u \in T(A^{-1}(W_1) \cap \operatorname{Im} G)$. □

Observe that for $\dim U = \dim V < \infty$, surjectivity as well as injectivity are of course equivalent to bijectivity, and the propositions are trivial. In particular, if T or G is an endomorphism, the propositions are nontrivial only for an infinite dimensional vector space.

A.4 Dimension and codimension

Recall that for subspaces V_1 and V_2 of a vector space V we have

$$\dim(V_1 + V_2) + \dim(V_1 \cap V_2) = \dim V_1 + \dim V_2$$

and analogously for the codimension

$$\operatorname{codim}(V_1 + V_2) + \operatorname{codim}(V_1 \cap V_2) = \operatorname{codim} V_1 + \operatorname{codim} V_2.$$

Note that if V is finite dimensional, the second equation is a consequence from the first and the equation $\dim V_1 + \operatorname{codim} V_1 = \dim V$. For V finite dimensional, we obtain similarly the equation

$$\operatorname{codim}(V_1 + V_2) + \dim V_1 = \dim(V_1 \cap V_2) + \operatorname{codim} V_2$$

relating the codimension of the sum with the dimension of the intersection of two subspaces. We show that this equation holds for arbitrary vector spaces.

Proposition A.14 *We have*

$$\operatorname{codim}(V_1 + V_2) + \dim V_1 = \dim(V_1 \cap V_2) + \operatorname{codim} V_2$$

for subspaces V_1 and V_2 of a vector space V.

Proof Let $\tilde{V}_1$ and $\tilde{V}_2$ be complements of $V_1 \cap V_2$ in V_1 and V_2, respectively, so that $V_1 = \tilde{V}_1 \dotplus (V_1 \cap V_2)$ and $V_2 = \tilde{V}_2 \dotplus (V_1 \cap V_2)$. Then one sees that we have a direct sum

$$V_1 + V_2 = \tilde{V}_1 \dotplus \tilde{V}_2 \dotplus (V_1 \cap V_2).$$

Let $\tilde{W}$ be a complement of $V_1 + V_2$ in V so that

$$V = (V_1 + V_2) \dotplus \tilde{W} = \tilde{V}_1 \dotplus \tilde{V}_2 \dotplus (V_1 \cap V_2) \dotplus \tilde{W}.$$

Hence $\operatorname{codim}(V_1 + V_2) = \dim \tilde{W}$ and $\operatorname{codim} V_2 = \dim(\tilde{W} + \tilde{V}_1)$. Computing the dimension of the subspace $\tilde{W} \dotplus \tilde{V}_1 \dotplus (V_1 \cap V_2)$ in two different ways, we obtain

$$\begin{aligned}\operatorname{codim}(V_1 + V_2) + \dim V_1 &= \dim \tilde{W} + \dim(\tilde{V}_1 + (V_1 \cap V_2)) \\ &= \dim(V_1 \cap V_2) + \dim(\tilde{W} + \tilde{V}_1) = \dim(V_1 \cap V_2) + \operatorname{codim} V_2,\end{aligned}$$

and the proposition is proved. □

If V_1 is finite dimensional and V_2 finite codimensional, all dimensions and codimensions in the above proposition are finite, and we obtain the following corollaries.

Corollary A.15 *Let V_1 and V_2 be subspaces of a vector space V with $\dim V_1 < \infty$ and $\operatorname{codim} V_2 < \infty$. Then*

$$\operatorname{codim}(V_1 + V_2) - \dim(V_1 \cap V_2) = \operatorname{codim} V_2 - \dim V_1.$$

In particular, we have $\dim(V_1 \cap V_2) = \operatorname{codim}(V_1 + V_2)$ iff $\dim V_1 = \operatorname{codim} V_2$.

Corollary A.16 *Let V_1 and V_2 be subspaces of a vector space V with $\dim V_1 < \infty$ and $\operatorname{codim} V_2 < \infty$. Then $V_1 \dotplus V_2 = V$ iff $V_1 \cap V_2 = 0$ and $\dim V_1 = \operatorname{codim} V_2$ iff $V_1 + V_2 = V$ and $\dim V_1 = \operatorname{codim} V_2$.*

So for testing whether two subspaces V_1 and V_2 with $\dim V_1 = \operatorname{codim} V_2 < \infty$ establish a direct decomposition $V = V_1 \dotplus V_2$, we have to check only one of the two defining conditions $V_1 \cap V_2 = 0$ and $V_1 + V_2 = V$.

The hypothesis that the dimensions are finite is necessary. Let k be a field, $V = k^{\mathbb{N}}$, and consider for example the two subspaces

$$\begin{aligned}V_1 &= \{(0, x_1, 0, x_2, 0, x_3, \ldots) \mid (x_n) \in k^{\mathbb{N}}\} \\ V_2 &= \{(0, 0, x_1, 0, x_2, 0, x_3, \ldots) \mid (x_n) \in k^{\mathbb{N}}\}.\end{aligned}$$

Then $\dim V_1 = \operatorname{codim} V_2 = \dim V = \infty$, $V_1 \cap V_2 = 0$ but $\operatorname{codim}(V_1 + V_2) = 1$.

We use the following corollary in Sect. 6 as a regularity test for boundary problems with finite dimensional kernels and boundary conditions.

Corollary A.17 *Let $V_1 = [v_1, \ldots, v_m]$ be a subspace of a vector space V and $\mathcal{F}_1 = [f_1, \ldots, f_n]$ a subspace of V^* with f_i and v_j linearly independent. Then*

$$V = V_1 \dotplus \mathcal{F}_1^{\perp}$$

is a direct sum iff $m = n$ and the matrix $(f_i(v_j))$ is regular.

Proof By (A.6), $\operatorname{codim} \mathcal{F}_1^{\perp} = \dim \mathcal{F}_1$, so we know from the previous corollary that $V = V_1 \dotplus \mathcal{F}_1^{\perp}$ is a direct sum iff $V_1 \cap \mathcal{F}_1^{\perp} = 0$ and $m = n$. Let $B = (f_i(v_j))$ with columns b_j. Now note that B is singular iff there exists a linear combination $\sum \lambda_j b_j = 0$ with at least one $\lambda_j \neq 0$ iff there exists a nonzero $u = \sum \lambda_j v_j$ in $V_1 \cap \mathcal{F}_1^{\perp}$. □

References

1. Brown, R.C., Krall, A.M.: Ordinary differential operators under Stieltjes boundary conditions. Trans. Am. Math. Soc. **198**, 73–92 (1974)
2. Buchberger, B.: An algorithm for finding the bases elements of the residue class ring modulo a zero dimensional polynomial ideal (German) [English translation published in J. Symbolic Comput., **41**(3-4), 475–511 (2006)]. Ph.D. Thesis, University of Innsbruck (1965)
3. Buchberger, B.: Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems [English translation: an algorithmic criterion for the solvability of a system of algebraic equations. In: Gröbner Bases and Applications, Buchberger, B., Winkler, F. (eds.) London Math. Soc. Lecture Note Ser., vol. 251, pp. 535–545. Cambridge University Press (1998)]. Aequationes Math. **4**, 374–383 (1970)
4. Conway, J.B.: A course in functional analysis. Graduate Texts in Mathematics, vol. 96, 2nd edn. Springer, New York (1990)
5. Eilenberg, S.: Automata, languages, and machines (vol. B). Pure and Applied Mathematics, vol. 59. Academic Press, New York (1976)
6. Engl, H.W., Hanke, M., Neubauer, A.: Regularization of inverse problems. Mathematics and its Applications, vol. 375. Kluwer Academic Publishers Group, Dordrecht (1996)
7. Erné, M., Koslowski, J., Melton, A., Strecker, G.E.: A primer on Galois connections. In: Papers on general topology and applications (Madison, WI, 1991), Annals of New York Academic Science, vol. 704, pp. 103–125. New York Academic Science, New York (1993)
8. Green, G.: An essay on the application of mathematical analysis to the theories of electricity and magnetism. Private, Nottingham (1828). Available at http://gallica.bnf.fr/ark:/12148/bpt6k994483
9. Grigoriev, D., Schwarz, F.: Factoring and solving linear partial differential equations. Computing **73**(2), 179–197 (2004)
10. Grigoriev, D., Schwarz, F.: Loewy- and primary decompositions of D-modules. Adv. App. Math. **38**, 526–541 (2007)
11. Grigoriev, D.Y.: Complexity of factoring and calculating the GCD of linear ordinary differential operators. J. Symbolic Comput. **10**(1), 7–37 (1990)
12. Kamke, E.: Differentialgleichungen. Lösungsmethoden und Lösungen. Teil I: Gewöhnliche Differentialgleichungen. Mathematik und ihre Anwendungen in Physik und Technik A, vol. 18, 8th edn. Akademische Verlagsgesellschaft, Leipzig (1967)
13. Köthe, G.: Topological vector spaces. I. Die Grundlehren der mathematischen Wissenschaften, vol. 159. Springer, New York (1969)
14. Lang, S.: Real and functional analysis. Graduate Texts in Mathematics, vol. 142. Springer, New York (1993)
15. Nashed, M.Z., Votruba, G.F.: A unified operator theory of generalized inverses. In: Generalized Inverses and Applications (Proc. Sem., Math. Res. Center, Univ. Wisconsin, Madison, Wis., 1973), pp. 1–109. Publication of Mathematics Research Center University Wisconsin, No. 32. Academic Press, New York (1976)
16. Neumann, C.: Lösung des allgemeinen Problems über den stationären Temperaturzustand einer homogenen Kugel ohne Hilfe von Reihenentwicklungen, nebst einigen Sätzen zur Theorie der Anziehung. H.W. Schmidt, Halle (1861)
17. van der Put, M., Singer, M.F.: Galois theory of linear differential equations. Grundlehren der Mathematischen Wissenschaften, vol. 328. Springer, Berlin (2003)
18. Riemann, B.: Schwere, Electricität und Magnetismus. Nach den Vorlesungen von Bernhard Riemann bearbeitet. Carl Rümpler, Hannover (1880). Lectures delivered in the summer term of 1861 in Göttingen. Available at http://de.wikisource.org/wiki/Schwere,_Elektricität_und_Magnetismus
19. Rosenkranz, M.: A new symbolic method for solving linear two-point boundary value problems on the level of operators. J. Symbolic Comput. **39**(2), 171–199 (2005)
20. Rosenkranz, M., Buchberger, B., Engl, H.W.: Solving linear boundary value problems via non-commutative Gröbner bases. Appl. Anal. **82**(7), 655–675 (2003)
21. Rosenkranz, M., Regensburger, G.: Solving and factoring boundary problems for linear ordinary differential equations in differential algebras. J. Symbolic Comput. doi:10.1016/j.jsc.2007.11.007 (2007)
22. Schwarz, F.: A factorization algorithm for linear ordinary differential equations. In: ISSAC '89: Proceedings of the ACM-SIGSAM 1989 International Symposium on Symbolic and Algebraic Computation, pp. 17–25. ACM Press, New York (1989)
23. Stakgold, I.: Green's functions and boundary value problems. Wiley, New York (1979)
24. Tsarev, S.P.: An algorithm for complete enumeration of all factorizations of a linear ordinary differential operator. In: ISSAC '96: Proceedings of the 1996 International Symposium on Symbolic and Algebraic Computation, pp. 226–231. ACM Press, New York (1996)

25. Tsarev, S.P.: Factorization of linear partial differential operators and darboux integrability of nonlinear pdes. SIGSAM Bull. **32**(4), 21–28 (1998)
26. Wyler, O.: Green's operators. Ann. Mat. Pura Appl. **66**(4), 252–263 (1964)
27. Yosida, K.: Functional analysis. Grundlehren der Mathematischen Wissenschaften, vol. 123, 6th edn. Springer, Berlin (1980)

A Skew Polynomial Approach to Integro-Differential Operators

Georg Regensburger, Markus Rosenkranz
Johann Radon Institute for Computational and Applied Mathematics (RICAM)
Austrian Academy of Sciences
A-4040 Linz, Austria
Georg.Regensburger@oeaw.ac.at, Markus.Rosenkranz@oeaw.ac.at

Johannes Middeke*
Research Institute for Symbolic Computation (RISC)
Johannes Kepler University
A-4040 Linz, Austria
Johannes.Middeke@risc.uni-linz.ac.at

ABSTRACT

We construct the algebra of integro-differential operators over an ordinary integro-differential algebra directly in terms of normal forms. In the case of polynomial coefficients, we use skew polynomials for defining the integro-differential Weyl algebra as a natural extension of the classical Weyl algebra in one variable. Its normal forms, algebraic properties and its relation to the localization of differential operators are studied. Fixing the integration constant, we regain the integro-differential operators with polynomial coefficients.

Categories and Subject Descriptors

I.1.1 [**Symbolic and Algebraic Manipulation**]: Expressions and Their Representation—*simplification of expressions*; I.1.2 [**Symbolic and Algebraic Manipulation**]: Algorithms—*algebraic algorithms*

General Terms

Theory, Algorithms

Keywords

Integro-differential operators, skew polynomials, Weyl algebra, integro-differential algebra, Baxter algebra.

1. INTRODUCTION

Skew polynomials provide a powerful framework for studying linear differential operators from an algebraic and algorithmic perspective [24, 12, 10]. In this paper, we develop a related approach for ordinary *integro-differential operators*, complementing the development reported in [27].

*The author was supported by the Austrian Science Foundation (FWF) under the project DIFFOP (P20 336–N18).

ISSAC'09, July 28–31, 2009, Seoul, Republic of Korea.

We have introduced the algebra of integro-differential operators in [28] for a symbolic treatment of linear boundary problems following [26]. It is based on integro-differential algebras (Section 2), which bring together the usual derivation structure with a suitable notion of indefinite integration and evaluation. Integro-differential operators are constructed as the corresponding operator algebra. They can be applied for solving boundary problems and for factoring them along a given factorization of the underlying differential equation. A prototype implementation of integro-differential operators in Theorema is presented in [7].

In contrast to our earlier construction, the present treatment of integro-differential operators is directly based on normal forms (Section 3). This is useful for analyzing the algebraic structure and developing algorithms. In this context, polynomial coefficients are of particular interest.

We construct an integro-differential analog of the classical Weyl algebra in one variable—henceforth called the *differential Weyl algebra*—as a skew polynomial ring (Section 4). The integro-differential Weyl algebra has a natural decomposition into the differential Weyl algebra, the *integro Weyl algebra* (Section 5), and the two-sided *evaluation ideal*. Unlike its differential part, the integro-differential Weyl algebra has zero divisors and is neither simple nor Noetherian.

The integro Weyl algebra forms a curious counterpart to the differential Weyl algebra. Following an analogous construction as a skew polynomial ring, the resulting algebra is also a Noetherian integral domain, but otherwise exhibits some striking differences: It is not a simple ring and it lacks a canonical action on the polynomials but it has a natural grading.

Compared to the algebra of integro-differential operators, the integro-differential Weyl algebra has a finer structure, which can be specialized naturally in two different ways, either discarding or fixing the evaluation (Section 6). Factoring out the evaluation ideal leads to a *localization*, where the "integral" is added as a two-sided inverse of the derivation. Factoring out a suitable relation *choosing the integration constant*, we obtain the algebra of integro-differential operators.

Some *notational conventions*: We fix a ground field K of characteristic 0. The inner direct sum of modules is written as $M = M_1 \dotplus M_2$. We use the symbol $\leq$ for algebraic substructures. Unless specified otherwise the variables i, j, k, m, n range over the nonnegative integers.

2. INTEGRO-DIFFERENTIAL ALGEBRAS

In this section, we summarize basic properties of integro-differential algebras from [28]. We recall that $(\mathcal{F}, \partial)$ is a differential K-algebra if $\partial\colon \mathcal{F} \to \mathcal{F}$ is K-linear map satisfying the *Leibniz rule*

$$\partial(fg) = \partial(f)\, g + f\, \partial(g). \tag{1}$$

For convenience, we assume $K \leq \mathcal{F}$ and write f' for $\partial(f)$.

Definition 1. *Let $\mathcal{F}$ be a commutative algebra over a field K. We call $(\mathcal{F}, \partial, \int)$ an* integro-differential algebra *if $(\mathcal{F}, \partial)$ is a differential algebra, $\int\colon \mathcal{F} \to \mathcal{F}$ is a K-linear section of ∂, that is,*

$$\partial \int = 1, \tag{2}$$

and the differential Baxter axiom

$$(\textstyle\int f')(\int g') = (\int f')g + f(\int g') - \int (fg)' \tag{3}$$

holds. Then we call $\int$ an integral operator *for ∂.*

We refer to the elements of $\mathcal{I} = \mathrm{Im}(\int)$ as *initialized*, while those of $\mathcal{C} = \mathrm{Ker}(\partial)$ are usually called *constants*. Since $\int$ is a section of ∂, we have projectors $\int\partial$ and

$$P = 1 - \textstyle\int\partial, \tag{4}$$

and a direct sum $\mathcal{F} = \mathcal{C} \dot{+} \mathcal{I}$ with $\mathcal{C} = \mathrm{Im}(P)$ and $\mathcal{I} = \mathrm{Ker}(P)$. Conversely, for every projector P onto a complement of $\mathcal{C}$ there exists a unique section of ∂ such that (4) holds; see for example [22, p. 17] or [25].

The standard example $\mathcal{F} = C^\infty(\mathbb{R})$ comes from analysis, where ∂ is the usual derivation and $\int$ the integral operator

$$\textstyle\int_c^x\colon f \mapsto \int_c^x f(\xi)\, d\xi.$$

for $c \in \mathbb{R}$. Here (2) is the Fundamental Theorem, while (3) can be verified either directly or by using the characterization of integral operators below. The projector $P\colon f \mapsto f(c)$ corresponds to a point evaluation. For an algorithmic approach to constant coefficient ODE, the subalgebra of exponential polynomials is important.

The polynomial ring $K[x]$ with the usual derivation is similarly seen to form an integro-differential algebra, with integral operator $\int_c^x\colon x^n \mapsto (x^{n+1} - c^{n+1})/(n+1)$ for $c \in K$. The corresponding projector is the evaluation homomorphism determined by $x \mapsto c$; we call c the *constant of integration*.

Substituting respectively $\int f$ for f and $\int g$ for g in (3) and using (1), (2) gives the plain *Baxter axiom* (of weight zero)

$$\textstyle\int f \cdot \int g = \int (f \int g) + \int (g \int f), \tag{5}$$

which is obviously an algebraic version of integration by parts (corresponding to the rewrite rule for $\int f \int$ in Table 1). A *Baxter algebra* $(\mathcal{F}, \int)$ is then a K-algebra $\mathcal{F}$ with a K-linear operation $\int$ fulfilling the Baxter axiom (5); we refer to [14, 2, 29] for more details.

Substituting $\int g$ for g in (3), one obtains with (1), (2) the following one-sided variant of the differential Baxter axiom

$$\textstyle\int fg = f \int g - \int (f' \int g), \tag{6}$$

which we used in [28] for the definition of integro-differential algebras. In the commutative case, both versions of the Baxter axiom are equivalent, but (3) has the advantage that it generalizes to noncommutative algebras over rings and Baxter operators with nonzero weight. Compare to the setting in [15], where a similar structure was introduced independently under the name of differential Rota-Baxter algebras. They only require (2) and the Baxter axiom (5) rather than its differential variant (3).

One can characterize what makes (3) stronger than (5). A section $\int$ of ∂ is an integral operator if and only if it is also $\mathcal{C}$-linear. Moreover, we can characterize the integral operators among sections by requiring the projector in (4) to be multiplicative. Another equivalent formulation of the differential Baxter axiom (corresponding to the usual integration by parts and the identity for $\int f\partial$ in Table 1) is

$$\textstyle\int fg' = fg - \int f'g - P(f)\, P(g), \tag{7}$$

following from $\mathcal{C}$-linearity of $\int$ and multiplicativity of P.

In the rest of the paper, we focus on ordinary differential equations. Thus we call an (integro-)differential algebra *ordinary* if $\dim_K \mathrm{Ker}(\partial) = 1$. Note that this terminology deviates from [17, p. 58], where it only refers to having a single derivation. In an ordinary differential algebra $\mathcal{F}$, we clearly have $K = \mathcal{C}$, so $\mathcal{F}$ is an algebra over its field of constants. A section is then automatically $\mathcal{C}$-linear, so the pure Baxter axiom (5) and its differential version (3) are equivalent.

In this case, the corresponding projector is a character

$$\mathtt{E} = 1 - \textstyle\int\partial \tag{8}$$

since it is multiplicative (by the above characterization of integral operators) and its image is $\mathcal{C} = K$. We write $\mathcal{M}(\mathcal{F})$ for the set of all characters on $(\mathcal{F}, \partial, \int)$, including in particular the *evaluation* $\mathtt{E}$.

3. THE ALGEBRA OF INTEGRO-DIFFERENTIAL OPERATORS

In analogy to differential operators over a differential algebra, it is natural to consider the algebra of linear operators over an integro-differential algebra. In [28] we defined the algebra of integro-differential operators as the quotient of the free algebra in the corresponding operators modulo the parametrized equations in Table 1. We showed that they form an infinite two-sided noncommutative Gröbner basis (or a Noetherian and confluent rewrite system [1]) and determined the corresponding normal forms. (See also [27] for a summary.) For the theory of Gröbner bases, we refer to [5, 6], for its noncommutative extension to [3, 21]. In this section, we want to define the algebra of integro-differential operators directly in terms of their normal forms.

Let $\mathcal{F}$ be an ordinary integro-differential algebra over K. In the following, the variables f, g are used for elements of $\mathcal{F}$ and φ, ψ for characters in $\mathcal{M}(\mathcal{F})$. Moreover, we use $U \bullet f$ for the action of U on f, where U is a combination of $\partial, \int$, functions in $\mathcal{F}$ and characters in $\mathcal{M}(\mathcal{F})$. In particular, we have $\partial \bullet f$ for the derivation, $\int \bullet f$ for the integral operator and $\varphi \bullet f$ for the application of characters, while $g \bullet f$ denotes the product in $\mathcal{F}$.

We remark that Table 1 is to be understood as including implicit equations for $\int\int$, $\int\partial$ and $\int\varphi$ by substituting $f = 1$ in the equations for $\int f \int$, $\int f\partial$ and $\int f\varphi$, respectively. Moreover, one obtains the equation $\mathtt{E}\int = 0$ from the definition of the evaluation $\mathtt{E}$.

For defining the algebra of integro-differential operators in terms of normal forms, we use the fact [28, Prop. 17] that every integro-differential operator can be uniquely written as a sum of a differential, an integral, and a so-called boundary

gf	$=$	$g \bullet f$	∂f	$=$	$f\partial + \partial \bullet f$
$\varphi\psi$	$=$	ψ	$\partial\varphi$	$=$	0
φf	$=$	$(\varphi \bullet f)\,\varphi$	$\partial\int$	$=$	1
$\int f\int$	$=$	$(\int \bullet f)\int - \int(\int \bullet f)$			
$\int f\partial$	$=$	$f - \int(\partial \bullet f) - (\mathtt{E} \bullet f)\,\mathtt{E}$			
$\int f\varphi$	$=$	$(\int \bullet f)\,\varphi$			

Table 1: Relations for Integro-differential Operators

operator. Since all these operators form subalgebras, we first describe them separately, and then the interaction between them. It is clear that the normal forms constitute an algebra isomorphic to the algebra of integro-differential operators in the sense of [28].

Moreover, for simplicity we take the evaluation $\mathtt{E}$ as the only character. For $\mathcal{F} = C^\infty[a,b]$, this amounts to considering only initial conditions, but the approach can be extended by using the normal forms for Stieltjes boundary conditions [28, Def. 14].

We first recall the well-known algebra of *differential operators* $\mathcal{F}[\partial]$ over $\mathcal{F}$. It is defined as sums of terms of the form $f\partial^i$ with the usual addition or, more abstractly, as the free left $\mathcal{F}$-module generated by the ∂^i. The multiplication is determined by viewing $\mathcal{F}$ as a subalgebra of $\mathcal{F}[\partial]$ and by using the equation

$$\partial \cdot f = f\partial + \partial \bullet f \tag{9}$$

coming from the Leibniz rule (1).

Clearly, sums of terms of the form $f\int g$ represent linear integral operators. But they cannot be normal forms since, by linearity, $f\int \lambda g$ and $\lambda f\int g$ with $\lambda \in K$ represent the same operator. This can be solved by choosing a K-basis $\mathcal{B}$ for $\mathcal{F}$. We additionally require $1 \in \mathcal{B}$ so that we can represent integral operators of the form $f\int$. Moreover, we use in the following the convention that $f\int g$ is to be understood as an abbreviation for the corresponding basis expansion if g is not a basis element.

We define the algebra of *integral operators* $\mathcal{F}[\int]$ over $\mathcal{F}$ as sums of terms of the form $f\int b$ with $b \in \mathcal{B}$ (or as the free left $\mathcal{F}$-module generated by the $\int b$). The multiplication is based on the equation

$$\int b \cdot \int = (\int \bullet b)\int - \int(\int \bullet b) \tag{10}$$

corresponding to the Baxter axiom (5). Note that $\mathcal{F}[\int]$ does not contain $\mathcal{F}$; it is an algebra without unit element.

We define the algebra of *boundary operators* $\mathcal{F}[\mathtt{E}]$ as sums of terms of the form $f\mathtt{E}\partial^i$. Their product is determined by

$$\mathtt{E}\partial^i \cdot f\mathtt{E}\partial^j = (\mathtt{E}\partial^i \bullet f)\,\mathtt{E}\partial^j, \tag{11}$$

which is a result of the Leibniz rule and the equations $\partial\mathtt{E} = 0$, $\mathtt{E}f = (\mathtt{E} \bullet f)\mathtt{E}$, $\mathtt{E}^2 = \mathtt{E}$. Also $\mathcal{F}[\mathtt{E}]$ does not contain $\mathcal{F}$.

The additive structure on integro-differential operators is then constructed as the direct sum

$$\mathcal{F}[\partial,\textstyle\int] = \mathcal{F}[\partial] \oplus \mathcal{F}[\int] \oplus \mathcal{F}[\mathtt{E}].$$

We regard the summands as being embedded in $\mathcal{F}[\partial,\int]$.

The multiplication within the summands is given by (9), (10), and (11). It remains to define the multiplication between different summands. To start with, multiplying a differential operator with an integral operator is given by

$$\partial \cdot f\int b = f \bullet b + (\partial \bullet f)\int b,$$

corresponding to (1) and (2). So we have $\mathcal{F}[\partial]\,\mathcal{F}[\int] \subset \mathcal{F}[\partial] + \mathcal{F}[\int]$. The multiplication in the reverse order is based on

$$\int b \cdot f\partial = b \bullet f - \int(\partial b \bullet f) - (\mathtt{E}b \bullet f)\,\mathtt{E},$$

corresponding to the variant of the Baxter axiom (7), so that $\mathcal{F}[\int]\,\mathcal{F}[\partial] \subset \mathcal{F}[\partial] + \mathcal{F}[\int] + \mathcal{F}[\mathtt{E}]$.

The equations for multiplying a boundary operator from either side with a differential or integral operator are

$$\begin{aligned}
\partial^i \cdot f\mathtt{E}\partial^j &= (\partial^i \bullet f)\mathtt{E}\partial^j,\\
\mathtt{E}\partial^i \cdot f\partial^j &= \sum_{k=0}^{i} (\mathtt{E} \bullet f_k)\mathtt{E}\partial^{j+k},\\
\int b \cdot f\mathtt{E}\partial^i &= (\textstyle\int b \bullet f)\mathtt{E}\partial^i,\\
\mathtt{E}\partial^i \cdot f\int b &= \sum_{l=0}^{i-1} (\mathtt{E} \bullet g_l)\mathtt{E}\partial^l,
\end{aligned}$$

where $\partial^i f = \sum_{k=0}^{i} f_k\partial^k$ and $\sum_{k=1}^{i} f_k\partial^{k-1}b = \sum_{l=0}^{i-1} g_l\partial^l$ as differential operators in $\mathcal{F}[\partial]$. Besides the rules used for (11), this involves the rule $\int f\mathtt{E} = (\int \bullet f)\mathtt{E}$. So we have $\mathcal{F}[\partial]\,\mathcal{F}[\mathtt{E}]$, $\mathcal{F}[\mathtt{E}]\,\mathcal{F}[\partial] \subset \mathcal{F}[\mathtt{E}]$ as well as $\mathcal{F}[\int]\,\mathcal{F}[\mathtt{E}]$, $\mathcal{F}[\mathtt{E}]\,\mathcal{F}[\int] \subset \mathcal{F}[\mathtt{E}]$ in these cases.

Since by the above definitions multiplying a boundary operator with any integro-differential operator gives a boundary operator, we see that $\mathcal{F}[\mathtt{E}]$ is the ideal in $\mathcal{F}[\partial,\int]$ generated by the evaluation $\mathtt{E}$, which we call the *evaluation ideal* of $\mathcal{F}[\partial,\int]$. Here and in the following an ideal always means a two-sided ideal. So we have

$$\mathcal{F}[\partial,\textstyle\int] = \mathcal{F}[\partial] \dotplus \mathcal{F}[\int] \dotplus (\mathtt{E}) \tag{12}$$

as a direct sum of $\mathcal{F}$-modules or K-vector spaces.

In the rest of this paper we will deal with the important special case $\mathcal{F} = K[x]$ from a skew polynomial perspective. Using the natural K-basis (x^k) yields a natural K-basis for all normal forms. In this case the above construction can be simplified substantially. We know from the Weyl algebra that the Leibniz rule (9) reduces to $\partial \cdot x = x\partial + 1$ and one can verify (compare Lemma 11) that $\int \cdot \int = x\int - \int x$ suffices to derive (10) for all polynomials. This is the basis for the skew polynomial construction in the following section.

4. THE INTEGRO-DIFFERENTIAL WEYL ALGEBRA

For analyzing rings of formal differential operators it is convenient to view them as skew polynomial rings. Specializing the coefficients to $K[x]$, one is led to the corresponding Weyl algebra. Our goal is to gain a skew polynomial perspective on the above ring $\mathcal{F}[\partial,\int]$ for $\mathcal{F} = K[x]$. In this context, we write ℓ instead of $\int$ to avoid confusion between iterated integrals ℓ^m and integrals with upper bounds $\int^m$.

We recall the construction of skew polynomials [24] [12, p. 276] [10]. Let A be a (noncommutative) ring without zero divisors, ξ an indeterminate, $\sigma\colon A \to A$ an injective endomorphism (also known as "twist") and $\delta\colon A \to A$ a σ-derivation. The skew polynomial ring $A[\xi;\sigma,\delta]$ consists of the elements $a_0 + a_1\xi + \cdots + a_n\xi^n$ with $a_0,\dots,a_n \in A$.

While the addition is defined termwise, the multiplication is determined by the rule

$$\xi a = \sigma(a)\,\xi + \delta(a).$$

It is well-known that $A[\xi;\sigma,\delta]$ is an integral domain since the usual degree equality $\deg fg = \deg f + \deg g$ is valid. We write $A[\xi;\delta]$ for $A[\xi;1,\delta]$.

We concentrate for a moment on the integral operators. One is tempted to take $A = K[x]$ and $\xi = \ell$. But the Baxter axiom requires $\ell x = x\ell - \ell^2$, in violation of the degree requirement. The way out is to reverse the adjunction of x and ℓ, thus picking $A = K[\ell]$ for the coefficient ring and $\xi = x$ for the indeterminate. (In the case of the differential Weyl algebra, the order of adjunction does not matter: This is the point of the well-known automorphism $x \leftrightarrow -\partial$, which does not carry over to its integro counterpart)

We choose a coefficient ring A that includes both ∂ and ℓ so that $A[x;\delta]$ yields in one stroke integro-differential operators that are "almost" isomorphic to $K[x][\partial,\int]$. It turns out that $A[x;\delta]$ has a finer structure than $K[x][\partial,\int]$; their relations will be studied in Section 6.

The coefficient ring A should contain all K-linear combinations of ∂ and ℓ, taking into account that $\partial\ell = 1$. Its derivation δ is set up so as to ensure the relations $\partial x - x\partial = 1$ and $x\ell - \ell x = \ell^2$ when $A[x;\delta]$ is introduced.

Definition 2. *The algebra $K\langle\partial,\ell\rangle$ is the quotient of the free algebra $K\langle D, L\rangle$ modulo the ideal $(DL-1)$. We write ∂ and ℓ for the corresponding residue classes. We define a derivation δ on $K\langle\partial,\ell\rangle$ by $\delta(\partial) = -1$ and $\delta(\ell) = \ell^2$.*

Note that δ is well-defined: Defining first a derivation on the free algebra by $\delta(D) = -1$ and $\delta(L) = L^2$, one sees immediately that $\delta(DL-1) = (DL-1)\,L$, so the passage to the quotient is legitimate.

The algebra $K\langle\partial,\ell\rangle$ is studied by N. Jacobson [16] from the general perspective of one-sided inverses in rings. His results imply immediately that $K\langle\partial,\ell\rangle$ is neither (left or right) Artinian nor (left or right) Noetherian. Extending this approach, L. Gerritzen [13] describes the right modules and derivations on $K\langle\partial,\ell\rangle$; using his classification [13, Prop. 7.1], we have $\delta = -\partial_0$. Some of the following results (without the differential structure—see below) can be found in their papers. Their approach is based on representation theory, while our treatment is based on a more algorithmic normal form perspective.

We shall now establish a decomposition of $K\langle\partial,\ell\rangle$ that is akin to (12). For this goal, observe that the monomials $\ell^i\partial^j$ form a K-basis of $K\langle\partial,\ell\rangle$ since they are normal forms with respect to the Gröbner basis $DL-1$.

In analogy to Equation (8) and [16], we define

$$\mathtt{E} = 1 - \ell\partial \qquad \text{and} \qquad e_{ij} = \ell^i\mathtt{E}\partial^j.$$

The e_{ij} satisfy the multiplication table for matrix units; see for example [16] and [18, Ex. 21.26]. The e_{ij} together with the pure ∂ and ℓ monomials form another K-basis. Indeed, iterating $\ell^{i+1}\partial^{j+1} = -e_{ij} + \ell^i\partial^j$, we obtain

$$\ell^{i+1}\partial^{j+1} = \begin{cases} \ell^{i-j} - \sum_{k=0}^{j} e_{ij} & \text{for } i > j, \\ \partial^{j-i} - \sum_{k=0}^{i} e_{ij} & \text{for } i \le j. \end{cases}$$

Hence ∂^j, ℓ^i, and e_{ij} generate $K\langle\partial,\ell\rangle$ over K. Using the relation $e_{ij} = \ell^i\partial^j - \ell^{i+1}\partial^{j+1}$, one sees that they are also linearly independent because the $\ell^i\partial^j$ are.

We note also that the K-vector space generated by the e_{ij} is the ideal $(\mathtt{E})$ since $\ell e_{ij} = e_{i+1,j}$ and $\partial e_{ij} = e_{i-1,j}$ for $i > 0$ and $\partial e_{0j} = 0$; analogously for multiplication on the right. Confer also [16, 13]. In analogy to $\mathcal{F}[\partial,\int]$, we refer to $(\mathtt{E})$ as the *evaluation ideal* of $K\langle\partial,\ell\rangle$.

Proposition 3. *We have*

$$K\langle\partial,\ell\rangle = K[\partial] \dot{+} K[\ell]\ell \dot{+} (\mathtt{E})$$

as a direct sum of K-vector spaces, where $K[\partial]$ is a differential subring of $K\langle\partial,\ell\rangle$ while $K[\ell]\ell$ is a differential subring without unit and $(\mathtt{E})$ is a δ-ideal.

Proof. We have already seen the decomposition part. All three summands are obviously closed under addition, multiplication and the first one also under derivations. For the second note that $\delta(q) = \frac{dq}{d\ell}\,\ell^2 \in K[\ell]\ell$ for all $q \in K[\ell]\ell$. The third summand is closed under δ since

$$\delta(\mathtt{E}) = -\delta(\ell)\partial - \ell\delta(\partial) = -\ell^2\partial + \ell = \ell\mathtt{E} \in (\mathtt{E}).$$

This completes the proof since δ is a derivation. □

Since $\partial\mathtt{E} = \partial - \partial = 0$ and $\mathtt{E}\ell = \ell - \ell = 0$, we obtain also $\partial^{i+1}e_{ij} = 0$ and $e_{ij}\ell^{j+1} = 0$, so every element of $(\mathtt{E})$ is both a left and and a right zero-divisor. The following minimality property of the evaluation ideal was also noted in [16].

Lemma 4. *Every nonzero ideal in $K\langle\partial,\ell\rangle$ contains $(\mathtt{E})$.*

Proof. Assume I is an ideal and $0 \neq f \in I$. Write now $f = p + q + e$ where $p \in K[\partial]$, $q \in K[\ell]\ell$ and $e \in (\mathtt{E})$ as in Propostion 3. Assume first $p \neq 0$. For a sufficiently high $k \geq 0$ we may assume that $\partial^k f \in K[\partial]$ since $\partial^k e_{ij} = 0$ for $k > i$ while q just gets "shifted" into $K[\partial]$. We may assume $\partial^k f$ is monic. Now let $\mathtt{E}\partial^m$ be the term with highest ∂-power in $\mathtt{E}\partial^k f \in I$. Then $\mathtt{E}\partial^k f\ell^m = \mathtt{E} \in I$ since $\mathtt{E}\partial^n\ell^m = 0$ for $m > n$. If $p = 0$ but $q \neq 0$ we may reason analogously by first looking at $f\ell^k$ for a suitable k.

Therefore, assume now $p = q = 0$ and $e \neq 0$. Let k be maximal such that e_{kj} occurs in e. Then we have $\partial^k f = \partial^k e \in \mathtt{E}K[\partial]\backslash\{0\}$ since all terms e_{ij} with $i < k$ vanish but the terms e_{kj} do not. By the same argument as above $\mathtt{E} \in I$. □

The lattice of differential ideals turns out to be particularly simple.

Proposition 5. *The only proper δ-ideal of $K\langle\partial,\ell\rangle$ is $(\mathtt{E})$.*

Proof. We have already seen in Proposition 3 that $(\mathtt{E})$ is a δ-ideal. Suppose that $I \neq 0$ is another δ-ideal. By Lemma 4 we have $(\mathtt{E}) \subseteq I$. Assume there exists $f = p + q + e \in I\backslash(\mathtt{E})$ where $p \in K[\partial]$, $q \in K[\ell]\ell$ and $e \in (\mathtt{E})$ as above, but with either p or q unequal to zero. Using the same trick as before, we can find $k \geq 0$ such that $\partial^k f \in K[\partial]\backslash\{0\}$. Now, if ∂^m is the leading term of $\partial^k f$, we have $\delta^m(\partial^k f) \in K$ since K has characteristic 0. Hence $I = K\langle\partial,\ell\rangle$. □

We consider for a moment $K[\partial]$, the subring of polynomials in ∂. The derivation δ extends uniquely to the Laurent polynomials $K[\partial,\partial^{-1}]$ if we view them as the localization of $K[\partial]$ by ∂. Intuitively, another way of getting the Laurent polynomials is making ℓ also a left inverse of ∂ in $K\langle\partial,\ell\rangle$. That would mean to set $\mathtt{E} = 1 - \ell\partial = 0$. It turns out that the intuition is right in this case; compare [13, Prop. 2.6] for the algebraic part.

PROPOSITION 6. *The map*

$$K\langle\partial,\ell\rangle/(\mathrm{E}) \xrightarrow{\sim} K[\partial,\partial^{-1}]$$

defined by $\partial + (\mathrm{E}) \mapsto \partial$ *and* $\ell + (\mathrm{E}) \mapsto \partial^{-1}$ *is a differential isomorphism.*

PROOF. The map φ given by $\ell^i\partial^j \mapsto \partial^{j-i}$ is a well-defined K-vector space homomorphism from $K\langle\partial,\ell\rangle$ to $K[\partial,\partial^{-1}]$. We claim that it is also a differential ring homomorphism. Since it is additive, we need to check this just for basis elements of $K\langle\partial,\ell\rangle$: We have $\varphi(\ell^i\partial^j\cdot\ell^k\partial^m) = \varphi(\ell^{i+k-j}\partial^m) = \partial^{j+m-i-k} = \varphi(\ell^i\partial^j)\,\varphi(\ell^k\partial^m)$, assuming $k \geq j$. The computation for $j > k$ is almost the same. We have furthermore $\varphi(\delta(\ell^i\partial^j)) = \varphi(i\ell^{i+1}\partial^j - j\ell^i\partial^{j-1}) = (i-j)\,\partial^{j-i-1} = \delta(\partial^{j-i}) = \delta(\varphi(\ell^i\partial^j))$.

We compute the kernel of φ by considering the basis that corresponds to the decomposition in Proposition 3. The basis vectors ℓ^i and ∂^j are sent to nonzero elements (even basis elements) in $K[\partial,\partial^{-1}]$. On the other hand, we have $\varphi(e_{ij}) = \varphi(\ell^i\partial^j - \ell^{i+1}\partial^{j+1}) = 0$ for all i,j. Hence we conclude $\ker\varphi = (\mathrm{E})$, and by the First Isomorphism Theorem the claim follows. □

Using Proposition 6 and Lemma 4 together with the Third Isomorphism Theorem—see for instance [11, Thm. 1.23]—we see that the ideals of $K\langle\partial,\ell\rangle$ are completely described by the ideals in $K[\partial,\partial^{-1}]$. This is a principal ideal domain by [4, Th. 2.18].

The main purpose of this section is to define the integro-differential analog of the differential Weyl algebra. As noted before Lemma 4, $K\langle\partial,\ell\rangle$ is not an integral domain. One can nevertheless introduce the skew polynomials as before (even with non-injective twists); see [11, Sec. 5.2], [20, Sec. 1.1.2], [18, Ex. 1.9]. Consequently the skew polynomials have zero divisors, and the degree equality must be replaced by the inequality $\deg fg \leq \deg f + \deg g$. The crucial point is that the normal forms are unique as before.

DEFINITION 7. *The* integro-differential Weyl algebra *is given by the skew polynomial ring* $K\langle\partial,\ell\rangle[x;\delta]$ *and is denoted by* $\mathrm{A}_1(\partial,\ell)$.

Any infinite ascending chain $I_1 < I_2 < \cdots$ of left ideals in A yields the infinite ascending chain $RI_1 < RI_2 < \cdots$ of left ideals in $R = A[\xi;\delta]$; similarly for right ideals and for descending chains. Consequently $\mathrm{A}_1(\partial,\ell)$ is also neither (left or right) Artinian nor (left or right) Noetherian. The latter is in stark contrast to the differential Weyl algebra, as the following proposition is.

Over a $\mathbb{Q}$-algebra A, simplicity of skew polynomial rings can be decided by the following practical characterization from [18, Th. 3.15]. The ring $A[\xi;\delta]$ is simple if and only if A has no nontrivial δ-ideals and δ is not an inner derivation. Otherwise, the skew polynomials with coefficients in a δ-ideal of A form an ideal in $A[\xi;\delta]$. Since we have seen in Proposition 3 that (E) is a nontrivial δ-ideal in $K\langle\partial,\ell\rangle$, we can use this criterion to see that the integro-differential Weyl algebra—unlike its differential companion—is not simple.

PROPOSITION 8. *The ring* $\mathrm{A}_1(\partial,\ell)$ *is not simple.*

PROOF. It remains to prove that δ is not an inner derivation. For assume $\delta = [p,\cdot]$ for some $p \in K\langle\partial,\ell\rangle$. Application to ∂ yields $-1 = [p,\partial]$. But $K\langle\partial,\ell\rangle/(\mathrm{E})$ being a commutative ring, every commutator of $K\langle\partial,\ell\rangle$ lies in the ideal (E). Thus we obtain $-1 \in (\mathrm{E})$, in contradiction to Proposition 3. □

5. THE INTEGRO WEYL ALGEBRA

For comparing $\mathrm{A}_1(\partial,\ell)$ with the construction in Section 3, it is useful to investigate the subring of integral operators.

DEFINITION 9. *The subring of* $\mathrm{A}_1(\partial,\ell)$ *consisting of skew polynomials with coefficients in* $K[\ell]$ *is called the* integro Weyl algebra *and denoted by* $\mathrm{A}_1(\ell)$.

Obviously we have $\mathrm{A}_1(\ell) = K[\ell][x;\delta]$, with the derivation δ restricted to $K[\ell]$. In the same fashion, the differential Weyl algebra $\mathrm{A}_1(\partial) = K[\partial][x;\delta]$ is the subring of $\mathrm{A}_1(\partial,\ell)$ consisting of skew polynomials with coefficients in $K[\partial]$.

Note that—unlike its integro-differential companion—the integro Weyl algebra is indeed an integral domain since $K[\ell]$ is. It provides an interesting and natural example of an Ore algebra, which to our knowledge has not been studied in the literature [10, 19].

At first sight, $\mathrm{A}_1(\ell)$ seems to be very similar to $\mathrm{A}_1(\partial)$, but we shall soon realize that appearances are deceptive. To start with, recall that $\mathrm{A}_1(\partial)$ has a canonical action on $K[x]$ in the following sense: If $x \in \mathrm{A}_1(\partial)$ acts by multiplication and $\partial \in \mathrm{A}_1(\partial)$ as a derivation, then $\partial \bullet f = f'$ yields the usual differentiation. The corresponding statement for $\mathrm{A}_1(\ell)$ would require $x \in \mathrm{A}_1(\ell)$ to act by multiplication and $\ell \in \mathrm{A}_1(\ell)$ as a Baxter operator. But this admits any integrals $\ell \bullet f = \int_c^x f$ with arbitrary $c \in K$. We will come back to this in Section 6. Another important difference to the differential case is that $\mathrm{A}_1(\ell)$ comes with a natural grading (by total degree in x and ℓ).

For comparing $\mathrm{A}_1(\ell) \leq \mathrm{A}_1(\partial,\ell)$ with the corresponding summand $K[\textstyle\int] \leq \mathcal{F}[\partial,\int]$, it is necessary to consider different K-bases for $\mathrm{A}_1(\ell)$. The construction of skew polynomials comes with the basis $(\ell^i x^j)$, which we shall call the *left basis* (since the coefficients appear to the left of the indeterminate). It is an easy exercise to determine the transition to the corresponding *right basis* $(x^j\ell^i)$.

LEMMA 10. *We have the identities*

$$x^n\ell^m = \sum_{k=0}^{n} \frac{(-m)^{\underline{k}}\,n^{\underline{k}}}{k!}\,(-1)^k\,\ell^{m+k}x^{n-k}, \tag{13}$$

$$\ell^m x^n = \sum_{k=0}^{n} \frac{(-m)^{\underline{k}}\,n^{\underline{k}}}{k!}\,x^{n-k}\ell^{m+k}, \tag{14}$$

where $n^{\underline{k}} = n(n-1)\ldots(n-k+1)$ *is the falling factorial.*

PROOF. Applying the Leibniz rule in both directions, one shows by induction that

$$x^n f = \sum_{k=0}^{n} \binom{n}{k}\,\delta^k(f)\,x^{n-k},$$

$$fx^n = \sum_{k=0}^{n} \binom{n}{k}\,(-1)^k\,x^{n-k}\,\delta^k(f)$$

for all $f \in K[\ell]$. Setting $f = \ell^m$ and applying $\binom{n}{k} = n^{\underline{k}}/k!$, the claim follows since $\delta^k(\ell^m) = (-1)^k\,(-m)^{\underline{k}}\,\ell^{m+k}$. □

The formulae in Lemma 10 are written in such a way that the similarity to the corresponding formulae for $\mathrm{A}_1(\partial)$ becomes apparent. In fact, Equation (1.4) of [30] coincides with (13) if we allow $m \in \mathbb{Z}$ and identify ℓ with ∂^{-1}. These heuristic observations are made precise in Section 6 by the isomorphism of Proposition 16.

While the left and right bases of $\mathrm{A}_1(\ell)$ are special to the skew polynomial setting, the general ring of integral operators $\mathcal{F}[\int]$ from Section 3 has the K-basis $(\tilde{b}\int b)$. In the present setting, this leads to the *mid basis* $(x^m, x^m\ell x^n)$. As we shall see immediately, its role as a K-basis is justified by the following commutator relation.

LEMMA 11. *We have* $[x^n, \ell] = n\,\ell x^{n-1}\ell$.

PROOF. The case $n = 0$ being trivial, we prove the identity for arbitrary $n + 1$. Substituting $m = 1$ in (13) and multiplying with $(n + 1)\,\ell$ from the left yields

$$\begin{aligned}(n+1)\,\ell x^n\ell &= \sum_{k=0}^{n}(n+1)^{\underline{k+1}}\,\ell^{k+2}x^{n-k}\\ &= -\ell x^{n+1} + \sum_{k=0}^{n+1}(n+1)^{\underline{k}}\,\ell^{k+1}x^{n-k+1};\end{aligned}$$

we conclude by substituting $(n+1, 1)$ for (n, m) in (13). □

COROLLARY 12. *The monomials* (x^m) *and* $(x^m\ell x^n)$ *form a* K*-basis of* $\mathrm{A}_1(\ell)$.

PROOF. In analogy to the differential Weyl algebra, one sees immediately that $\mathrm{A}_1(\ell)$ is isomorphic to the free K-algebra in X and L modulo the ideal generated by $XL - LX - L^2$. Lemma 11 implies that the polynomials

$$LX^nL - (n+1)^{-1}\,[X^{n+1}, L]$$

belong to the ideal. They form a Gröbner basis with respect to the following admissible order [31, p. 268]: Words are compared in L-degree, then in total degree, and finally lexicographically (letters ordered either way). A routine calculation shows that the overlaps LX^nLX^mL are resolvable. The residue classes of the monomials X^n and X^mLX^n form a K-basis of the quotient ring [31, Thm. 7]. □

The transition between the left/right basis and the mid basis is governed by the following formulae.

LEMMA 13. *We have the identities*

$$x^m\ell x^n = \sum_{k=0}^{m}\frac{m!}{k!}\,\ell^{m-k+1}x^{k+n}, \tag{15}$$

$$x^m\ell x^n = \sum_{k=0}^{n}\frac{n!}{k!}\,(-1)^{n-k}x^{m+k}\ell^{n-k+1}, \tag{16}$$

$$\ell^{m+1} = \sum_{k=0}^{m}\frac{(-1)^k}{k!\,(m-k)!}\,x^{m-k}\ell x^k \tag{17}$$

for changing between the left/right and the mid basis.

PROOF. For proving the first formula, it suffices to set $n = 0$. Substituting $(m, 1)$ for (n, m) in (13), one obtains Equation (15) after an index transformation. Analogously, one proves the second formula with $m = 0$ by substituting 1 for m in (14).

We prove the third formula by induction. The base case $m = 0$ is trivial, so assume (17) for $m \geq 0$. Multiplying it with ℓ from the right and using Lemma 11 yields

$$\ell^{m+2} = \sum_{k=0}^{m}\frac{(-1)^k}{(k+1)!\,(m-k)!}\,(x^{m+1}\ell - x^{m-k}\ell x^{k+1}).$$

After expanding the parenthesis and extracting $x^{m+1}\ell$, one is left with the simple binomial sum

$$\sum_{k=0}^{m}\frac{(-1)^k}{(k+1)!\,(m-k)!} = \frac{1}{(m+1)!},$$

so we obtain

$$\begin{aligned}&\frac{1}{(m+1)!}\,x^{m+1}\ell + \sum_{k=1}^{m+1}\frac{(-1)^k}{k!\,(m-k+1)!}\,x^{m-k+1}\ell x^k\\ &\quad= \sum_{k=0}^{m+1}\frac{(-1)^k}{k!\,(m-k+1)!}\,x^{m-k+1}\ell x^k\end{aligned}$$

for ℓ^{m+2}, which is indeed (17) for $m + 1$. □

We note that (17) can be regarded as an algebraic version of the well-known Cauchy formula for repeated integration [23, p. 38].

In view of the transition formulae (15) and (17), one can use the K-basis $(x^m\ell x^n)$ of $\mathrm{A}_1(\ell)\ell$ for setting up a concrete isomorphism (of algebras without unit) to $K[x][\int]$ with its K-basis $(x^m\int x^n)$. Confer Theorem 20 for an analogous statement for the full integro-differential Weyl algebra.

As for $\mathrm{A}_1(\partial, \ell)$, we see that $\mathrm{A}_1(\ell)$ is not a simple ring by the following characterization of the δ-ideals in $K[\ell]$.

LEMMA 14. *An ideal* $I \leq K[\ell]$ *is a nontrivial* δ*-ideal if and only if* $I = (\ell^n)$ *with* $n > 0$.

PROOF. Since $\delta(\ell^n) = n\ell^{n+1}$, ideals generated by ℓ^n are obviously δ-ideals. Conversely, let $I = (q)$ be a nontrivial δ-ideal with $q = \sum_{i=k}^{n} a_i\ell^i \in K[\ell]$ a polynomial of degree $n > 0$ and order k, meaning $a_k \neq 0$. Hence $\delta(q) = rq$ for some $r \in K[\ell]$ so that

$$\delta(q) = \sum_{i=k}^{n} a_i i\,\ell^{i+1} = r\sum_{i=k}^{n} a_i\,\ell^i$$

with $r = b_1\ell + b_0$. Equating the coefficients of ℓ^k and ℓ^{n+1} implies respectively $b_0 = 0$ and $b_1 = n$, the latter since K has characteristic 0. If $k < n$, equating the coefficients of ℓ^{k+1} implies $(n - k)a_k = 0$, in contradiction to our assumption on the characteristic of K. □

PROPOSITION 15. *The ring* $\mathrm{A}_1(\ell)$ *is not simple.*

PROOF. By the previous lemma there are nontrivial δ-ideals in $K[\ell]$. Since δ cannot be an inner derivation, the claim follows as in Proposition 8 from [18, Th. 3.15]. □

6. LOCALIZATION AND EVALUATION

By the construction of $\mathrm{A}_1(\partial, \ell)$, we have set up ℓ as an integral that is a right inverse for ∂. This still leaves some ambiguity for the choice of ℓ, which we will now remove. There are two extreme possibilities: When we require ℓ to be a two-sided inverse, we obtain a localization. On the other hand, we may insist ℓ to be a proper integral by fixing the integration constant; this leads us back to the ring of integro-differential operators $K[x][\partial, \int]$.

Let us start with the localization. Extending the derivation to the Laurent polynomial ring $K[\partial, \partial^{-1}]$ as in Proposition 6, we form the skew polynomial ring $K[\partial, \partial^{-1}][x; \delta]$. Of course, we may also localize $K[\ell]$ to obtain $K[\ell, \ell^{-1}][x; \delta]$ by using an analogous construction. These two rings are

naturally isomorphic, as we will now prove. In the following proofs, we will make use of the universal property of skew polynomial rings [9, Prop. 3.6] [20, §1.2.5] that allows to lift differential homomorphisms from coefficients to skew polynomials.

PROPOSITION 16. *The map*

$$K[\partial, \partial^{-1}][x;\delta] \xrightarrow{\sim} K[\ell, \ell^{-1}][x;\delta]$$

induced by $\partial \mapsto \ell^{-1}$ *is an isomorphism.*

PROOF. The map φ induced by $\partial \mapsto \ell^{-1}$ is a differential homomorphism between $K[\partial, \partial^{-1}]$ and $K[\ell, \ell^{-1}]$ since $\delta(\varphi(\partial)) = \delta(\ell^{-1}) = -\ell^2/\ell^2 = -1 = \varphi(\delta(\partial))$. By the universal property, its extension to $K[\partial, \partial^{-1}][x;\delta]$ is also a homomorphism, and it is clearly bijective. □

The following lemma allows to transfer the skew polynomial structure across quotients as in the commutative case, compare also [9, Prop. 3.15].

LEMMA 17. *Let* (R, δ) *be a differential ring and* $I \leq R$ *a differential ideal. Then*

$$(R/I)[x;\tilde{\delta}] \cong R[x;\delta]/(I)$$

as rings where (I) *denotes the ideal generated by* I *in* $R[x;\delta]$ *and* $\tilde{\delta}$ *is the derivation induced by* δ.

PROOF. The proof is as in the commutative case. First we note that (I) consists exactly of the skew polynomials with coefficients in I. The canonical map $R \to R/I$ is a differential epimorphism and extends therefore to an epimorphism $R[x;\delta] \to (R/I)[x;\tilde{\delta}]$ by the universal property. Its kernel are all skew polynomials whose coefficients are in I. □

The next step is to explore the relation between $\mathrm{A}_1(\partial, \ell)$ and $K[\partial, \partial^{-1}][x;\delta]$. It is very natural—the latter arises from the former by making ℓ also a left inverse of ∂.

THEOREM 18. *We have*

$$\mathrm{A}_1(\partial, \ell)/(\mathtt{E}) \cong K[\partial, \partial^{-1}][x;\delta]$$

as rings.

PROOF. In Proposition 6 we have proved that there exists an isomorphism $\varphi\colon K\langle\partial, \ell\rangle/(\mathtt{E}) \to K[\partial, \partial^{-1}]$. Using again the universal property, there is a corresponding isomorphism $\tilde{\varphi}$ between the skew polynomial rings $(K\langle\partial, \ell\rangle/(\mathtt{E}))[x;\tilde{\delta}]$ and $K[\partial, \partial^{-1}][x;\delta]$, where $\tilde{\delta}$ denote the derivative induced by δ. The claim now follows from Lemma 17. □

We note that the localization can also be applied in the setting of Section 3 by factoring out $(\mathtt{E})$, leading to the isomorphism $\mathcal{F}[\partial, \textstyle\int]/(\mathtt{E}) \cong \mathcal{F}[\partial] \dotplus \mathcal{F}[\textstyle\int]$.

For reconstructing the ring $K[x][\partial, \int]$ of Section 3 from $\mathrm{A}_1(\partial, \ell)$, we need a decomposition analogous to (12). Since the decomposition in Proposition 3 carries over coefficientwise to $\mathrm{A}_1(\partial, \ell)$, we obtain

$$\mathrm{A}_1(\partial, \ell) = \mathrm{A}_1(\partial) \dotplus \mathrm{A}_1(\ell)\ell \dotplus (\mathtt{E}), \tag{18}$$

where $(\mathtt{E})$ is the *evaluation ideal* in $\mathrm{A}_1(\partial, \ell)$. Note that this ideal consists of the skew polynomials with coefficients in $(\mathtt{E}) \subseteq K\langle\partial, \ell\rangle$ as observed before Proposition 8.

The key tool for fixing the integration constant $c \in K$ is the following refinement of the above decomposition. In analogy to the space $K[x][\mathtt{E}]$ introduced in Section 3, we consider the K-vector space $B \leq \mathrm{A}_1(\partial, \ell)$ with basis $(x^k \mathtt{E} \partial^j)$. Note that here and in the following we make use of the right basis $(x^k\partial^i, x^k\ell^i, x^k e_{ij})$ of $\mathrm{A}_1(\partial, \ell)$.

LEMMA 19. *In* $\mathrm{A}_1(\partial, \ell)$, *we have for every* $c \in K$ *the decomposition*

$$(\mathtt{E}) = B \dotplus (\eta),$$

and $(x^k\ell^i\eta\partial^j)$ *is a* K*-basis for* (η), *where* $\eta = \mathtt{E}x - c\mathtt{E}$.

PROOF. One can easily see $\mathtt{E}x = (x - \ell)\mathtt{E}$. This implies $\ell^{i-1}\eta = x\ell^{i-1}\mathtt{E} - i\ell^i\mathtt{E} - c\ell^{i-1}\mathtt{E} \in (\eta)$ and hence

$$x^k e_{ij} + \frac{c}{i} x^k e_{i-1,j} - \frac{1}{i} x^{k+1} e_{i-1,j} \in (\eta)$$

for $i \geq 1$. This allows to replace $x^k e_{ij}$ by terms with smaller powers of ℓ, eventually eliminating all occurrences of ℓ. This means that every element in $(\mathtt{E})$ may be represented as K-linear combination of elements of the form $x^k e_{0j} = x^k \mathtt{E} \partial^j$ and some element in (η).

We write η_{ij} for $\ell^i\eta\partial^j$ and H for the K-vector space generated by $x^k\eta_{ij}$. Obviously H is is a subspace of (η). The product of an element $x^k\eta_{ij}$ by ∂ or ℓ from the right is again in H. By Lemma 10 and by the Leibniz rule we may commute products of the form ℓx^k and ∂x^k, so left multiplication by ℓ and ∂ does not leave H either. Finally, H is also closed under right multiplication by x since $\eta x = (x - \ell)\eta$. Hence H is an ideal, which implies $H = (\eta)$.

For proving directness assume

$$\sum_{m,n} a_{mn} x^k e_{0n} = \sum_{i,j,k} b_{ijk} x^k \eta_{ij} \tag{19}$$

for suitable $a_{mn}, b_{ijk} \in K$. Converting the right-hand side to the basis $(x^k e_{ij})$ by $x^k\eta_{ij} = x^{k+1}e_{ij} - (i+1)x^k e_{i+1,j} - cx^k e_{ij}$ and choosing i maximal, we see that the terms b_{ijk} must all vanish because $(x^k e_{ij})$ is a K-basis and the left-hand side does not contain terms of the form $e_{i+1,j}$. Repeating this for smaller i, it follows that the sum is direct. Using the same argument with 0 as the left-hand side, we conclude that $(x^k\eta_{ij})$ is a K-basis of (η). □

Using the direct sum from Lemma 19, it is now immediate to draw the connection to the ring $K[x][\partial, \textstyle\int]$ of Section 3.

THEOREM 20. *If* $\textstyle\int$ *is an integral operator for the standard derivation* ∂ *on* $K[x]$, *we have*

$$\mathrm{A}_1(\partial, \ell)/(\mathtt{E}x - c\mathtt{E}) \;\cong\; K[x][\partial, \textstyle\int]$$

with $c = \mathtt{E} \bullet x \in K$ *as the constant of integration.*

PROOF. Using Lemma 19 and (18) we see that

$$\mathrm{A}_1(\partial, \ell)/(\eta) = \mathrm{A}_1(\partial) \dotplus \mathrm{A}_1(\ell)\ell \dotplus B.$$

As K-bases we can choose $(x^k\partial^i)$, the mid basis $(x^m\ell x^n)$ and $(x^k\mathtt{E}\partial^j)$, respectively. They map directly to the corresponding basis elements in $K[x][\partial, \textstyle\int]$ detailed in Section 3. This yields a K-linear isomorphism.

For proving that it is also an isomorphism of K-algebras, it suffices to verify that all identities in Table 1 are satisfied. The first six are immediate, for the $\textstyle\int f \int$ rule one uses Lemma 11, for the remaining two rules one can apply the identity $\ell x^k \mathtt{E} \equiv (x^{k+1} - c^{k+1})/(k+1)$ modulo $(\mathtt{E}x - c\mathtt{E})$. □

An alternative proof of Theorem 20 takes the detour via the free algebra $K\langle D, L, X\rangle$. Using its construction, one can show that $\mathrm{A}_1(\partial, \ell)/(\mathtt{E}x - c\mathtt{E})$ is isomorphic to the free algebra modulo the four relations $DL = 1$, $XD = DX - 1$, $XL = LX + L^2$, and $(1 - LD)X = c(1 - LD)$. It remains to prove that these four relations generate the identities of Table 1, which is laborious but straightforward.

7. CONCLUSION

The integro-differential Weyl algebra exhibits an interesting algebraic structure that deserves further study. Encoding integro-differential operators in a skew polynomial setting, it allows to recast our algebraic approach to linear boundary problems in a new language. We hope this will advance the algorithmic treatment of various operations [28], for example the computation of Green's operators and the factorization into lower-order problems.

The current formulation is still very limited in scope. Since we have taken only one character (necessarily the evaluation), boundary problems—both their formulation and their solution—are restricted initial value problems. Adjoining more characters in a skew polynomial setting will be an interesting task.

A more challenging extension concerns the transition from ODE to PDE, analogous to the classical Weyl algebra in several variables. As reported in [25], our algebraic setup (including the factorization) extends to boundary problems for PDE; the task is now to develop an algorithmic framework for relevant classes of such boundary problems. The skew polynomial approach initiated here could provide an appropriate vantage point.

8. REFERENCES

[1] F. Baader and T. Nipkow. *Term rewriting and all that*. Cambridge University Press, Cambridge, 1998.

[2] G. Baxter. An analytic problem whose solution follows from a simple algebraic identity. *Pacific J. Math.*, 10:731–742, 1960.

[3] G. M. Bergman. The diamond lemma for ring theory. *Adv. in Math.*, 29(2):178–218, 1978.

[4] R. Brüske, F. Ischebeck, and F. Vogel. *Kommutative Algebra*. Bibliographisches Institut, Mannheim, 1989.

[5] B. Buchberger. *An algorithm for finding the bases elements of the residue class ring modulo a zero dimensional polynomial ideal (German)*. PhD thesis, Univ. of Innsbruck, 1965.

[6] B. Buchberger. Introduction to Gröbner bases. In *[8]*, pages 3–31. 1998.

[7] B. Buchberger, G. Regensburger, M. Rosenkranz, and L. Tec. General polynomial reduction with Theorema functors: Applications to integro-differential operators and polynomials. *ACM Commun. Comput. Algebra*, 42(3):135–137, 2008. Poster presented at ISSAC'08.

[8] B. Buchberger and F. Winkler, editors. *Gröbner bases and applications*, volume 251 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 1998.

[9] J. Bueso, J. Gómez-Torrecillas, and A. Verschoren. *Algorithmic methods in non-commutative algebra*. Kluwer Academic Publishers, Dordrecht, 2003.

[10] F. Chyzak and B. Salvy. Non-commutative elimination in Ore algebras proves multivariate identities. *J. Symbolic Comput.*, 26(2):187–227, 1998.

[11] P. M. Cohn. *Introduction to ring theory*. Springer Undergraduate Mathematics Series. Springer-Verlag London Ltd., London, 2000.

[12] P. M. Cohn. *Further algebra and applications*. Springer-Verlag London Ltd., London, 2003.

[13] L. Gerritzen. Modules over the algebra of the noncommutative equation $yx = 1$. *Arch. Math. (Basel)*, 75(2):98–112, 2000.

[14] L. Guo. Baxter algebras and differential algebras. In *Differential algebra and related topics (Newark, NJ, 2000)*, pages 281–305. World Sci. Publ., 2002.

[15] L. Guo and W. Keigher. On differential Rota-Baxter algebras. *J. Pure Appl. Algebra*, 212(3):522–540, 2008.

[16] N. Jacobson. Some remarks on one-sided inverses. *Proc. Amer. Math. Soc.*, 1:352–355, 1950.

[17] E. Kolchin. *Differential algebra and algebraic groups*. Academic Press, New York-London, 1973.

[18] T. Y. Lam. *A first course in noncommutative rings*. Springer-Verlag, New York, 1991.

[19] V. Levandovskyy. *Non-commutative Computer Algebra for polynomial algebras: Gröbner bases, applications and implementation*. PhD thesis, Universität Kaiserslauten, 2005.

[20] J. C. McConnell and J. C. Robson. *Noncommutative Noetherian rings*. American Mathematical Society, Providence, RI, revised edition, 2001.

[21] T. Mora. An introduction to commutative and noncommutative Gröbner bases. *Theoret. Comput. Sci.*, 134(1):131–173, 1994.

[22] M. Z. Nashed and G. F. Votruba. A unified operator theory of generalized inverses. In M. Z. Nashed, editor, *Generalized inverses and applications*, pages 1–109. Academic Press, New York, 1976.

[23] K. B. Oldham and J. Spanier. *The fractional calculus*. Academic Press, New York-London, 1974.

[24] Ø. Ore. Theory of non-commutative polynomials. *Ann. Math.*, 34:480–508, 1933.

[25] G. Regensburger and M. Rosenkranz. An algebraic foundation for factoring linear boundary problems. *Ann. Mat. Pura Appl. (4)*, 188(1):123–151, 2009.

[26] M. Rosenkranz. A new symbolic method for solving linear two-point boundary value problems on the level of operators. *J. Symbolic Comput.*, 39(2):171–199, 2005.

[27] M. Rosenkranz and G. Regensburger. Integro-differential polynomials and operators. In D. Jeffrey, editor, *Proceedings of ISSAC '08*, pages 261–268, New York, NY, USA, 2008. ACM.

[28] M. Rosenkranz and G. Regensburger. Solving and factoring boundary problems for linear ordinary differential equations in differential algebras. *J. Symbolic Comput.*, 43(8):515–544, 2008.

[29] G.-C. Rota. Baxter algebras and combinatorial identities. *Bull. Amer. Math. Soc.*, 75:325–334, 1969.

[30] M. Saito, B. Sturmfels, and N. Takayama. *Gröbner deformations of hypergeometric differential equations*. Springer-Verlag, Berlin, 2000.

[31] V. Ufnarovski. Introduction to noncommutative Gröbner bases theory. In *[8]*, pages 259–280. 1998.

A Symbolic Framework for Operations on Linear Boundary Problems

Markus Rosenkranz[1], Georg Regensburger[1],
Loredana Tec[2], and Bruno Buchberger[2]

[1] Johann Radon Institute for Computational and Applied Mathematics,
Austrian Academy of Sciences, Altenberger Str. 69, 4040 Linz, Austria
[2] Research Institute for Symbolic Computation,
Johannes Kepler Universität, 4032 Castle of Hagenberg, Austria

Abstract. We describe a symbolic framework for treating linear boundary problems with a generic implementation in the Theorema system. For ordinary differential equations, the operations implemented include computing Green's operators, composing boundary problems and integro-differential operators, and factoring boundary problems. Based on our factorization approach, we also present some first steps for symbolically computing Green's operators of simple boundary problems for partial differential equations with constant coefficients. After summarizing the theoretical background on abstract boundary problems, we outline an algebraic structure for partial integro-differential operators. Finally, we describe the implementation in Theorema, which relies on functors for building up the computational domains, and we illustrate it with some sample computations including the unbounded wave equation.

Keywords: Linear boundary problem, Green's operator, Integro-Differential Operator, Ordinary Differential Equation, Wave Equation.

1 Introduction

Due to their obvious importance in applications, *boundary problems* play a dominant role in Scientific Computing, but almost exclusively in the numerical segment. It is therefore surprising that they have as yet gained little attention in Symbolic Computation, neither from a theoretical perspective nor in computer algebra systems.

In applications [1, p. 42] one is "concerned not only with solving [the boundary problem] for specific data but also with finding a suitable form for the solution that will exhibit its dependence on the data." In our work, we focus on linear boundary problems (and will henceforth suppress the attribute "linear"). For us, a boundary problem is thus a differential equation with a symbolic right-hand side, supplemented by suitable boundary conditions. Solving it means to determine its *Green's operator*, namely the integral operator that maps the right-hand side to the solution. For a symbolic approach to boundary problems, one has to develop a constructive algebraic theory of integral operators and an algorithmic framework for manipulating boundary conditions.

V.P. Gerdt, E.W. Mayr, and E.V. Vorozhtsov (Eds.): CASC 2009, LNCS 5743, pp. 269–283, 2009.

Such a development was initiated in [2], leading to a symbolic method for computing Green's operators of regular two-point boundary problems with constant coefficients [3]. We extended these results to a *differential algebra setting* in [4], where we also developed a *factorization method* applicable to boundary problems for ordinary differential equations (ODEs). A more *abstract view* on boundary problems and a general factorization theory is described in [5], including in particular partial differential equations (PDEs).

In this paper, we describe a prototype *implementation in Theorema* [6], currently based on a raw interface that will be improved in the future. It provides generic algorithms for various operations on boundary problems and integro-differential operators for ODEs (Section 5), exemplified in (Appendix A): computing Green's operators, composing boundary problems and integro-differential operators, and factoring boundary problems. The computations are realized by a suitable noncommutative Gröbner basis that reflects the essential interactions between certain basic operators. Gröbner bases were introduced by Buchberger in [7]. For an introduction to the theory, we refer to [8], for its noncommutative extension to [9].

Moreover, for *PDEs* we present some first steps for making the abstract setting of [5] algorithmic. We develop an algebraic language for encoding the integro-differential operators appearing as Green's operators of some simple two-dimensional Dirchlet problems for PDEs with constant coefficients (Section 4). Using our generic factorization approach, this allows to find the Green's operator of higher-order boundary problems by composing those of its lower-order factors. This idea is exemplified for the unbounded wave equation with a sample computation (Appendix A).

For the broader audience of Scientific Computing, we summarize the necessary *theoretical background* on abstract boundary problems, omitting all technical details and illustrating it for the case of ODEs (Section 2). After explaining the composition and factorization of boundary problems (Section 3), we outline the algebraic structures used for encoding ordinary as well as partial integro-differential operators (Section 4).

For motivating our algebraic setting of boundary problems, we consider first the *simplest two-point boundary problem*. Writing $\mathcal{F}$ for the real or complex vector space $C^\infty[0,1]$, it reads as follows: Given $f \in \mathcal{F}$, find $u \in \mathcal{F}$ such that

$$\boxed{\begin{aligned} &u'' = f, \\ &u(0) = u(1) = 0. \end{aligned}} \tag{1}$$

Let $D\colon \mathcal{F} \to \mathcal{F}$ denote the usual derivation and L, R the two linear functionals $L\colon f \mapsto f(0)$ and $R\colon f \mapsto f(1)$. Note that u is annihilated by any linear combination of these functionals so that problem (1) can be described by $(D^2, [L, R])$, where $[L, R]$ is the subspace generated by L, R in the dual space $\mathcal{F}^*$.

As a second example, consider the following boundary problem for the *wave equation* on the domain $\Omega = \mathbb{R} \times \mathbb{R}_{\geq 0}$, now writing $\mathcal{F}$ for $C^\infty(\Omega)$: Given $f \in \mathcal{F}$, find $u \in \mathcal{F}$ such that

$$\boxed{\begin{aligned} &u_{tt} - u_{xx} = f, \\ &u(x,0) = u_t(x,0) = 0. \end{aligned}} \tag{2}$$

Note that we use the terms "boundary condition/problem" in the general sense of linear conditions. The boundary conditions in (2) can be expressed by the infinite family of linear functionals $\beta_x \colon u \mapsto u(x,0)$, $\gamma_x \colon u \mapsto u_t(x,0)$ with x ranging over $\mathbb{R}$. So we can represent the boundary problem again by a pair consisting of the differential operator $D_t^2 - D_x^2$ and the (now infinite dimensional) subspace generated by β_x and γ_x in $\mathcal{F}^*$.

For ensuring a unique representation of boundary conditions, we take the *orthogonal closure* of this subspace, which we denote by $[\beta_x, \gamma_x]_{x\in\mathbb{R}}$. This is the space of all linear functionals vanishing on the functions annihilated by β_x, γ_x. Every finite dimensional subspace is orthogonally closed, but here, for example, the functionals $u \mapsto \int_0^x u(\eta,0)\,d\eta$ and $u \mapsto u_x(x,0)$ for arbitrary $x \in \mathbb{R}$ are in the orthogonal closure but not in the space generated by β_x and γ_x. We refer to [10] or [5, App. A.1] for details on the orthogonal closure.

Some *notational conventions*. We use the symbol $\leq$ for algebraic substructures. If $T \colon \mathcal{F} \to \mathcal{G}$ is a linear map and $\mathcal{B} \leq \mathcal{G}^*$, we write $\mathcal{B} \cdot T$ for the subspace $\{\beta \circ T \mid \beta \in \mathcal{B}\} \leq \mathcal{F}^*$. For a subset $\mathcal{B} \subseteq \mathcal{F}^*$ the so-called *orthogonal* is defined as $\mathcal{B}^\perp = \{u \in \mathcal{F} \mid \beta(u) = 0 \text{ for all } \beta \in \mathcal{B}\}$.

2 An Algebraic Formulation of Boundary Problems

In this section, we give a summary of the algebraic setting for boundary problems exposed in [5], see also there for further details and proofs. We illustrate the definitions and statements for ODEs on a compact interval $[a,b] \subseteq \mathbb{R}$. In this setting, most of the statements can be made algorithmic relative to solving homogeneous linear differential equations (and the operations of integration and differentiation).

A *boundary problem* is given by a pair $(T, \mathcal{B})$, where $T \colon \mathcal{F} \to \mathcal{G}$ is a surjective linear map between vector spaces $\mathcal{F}, \mathcal{G}$ and $\mathcal{B} \leq \mathcal{F}^*$ is an orthogonally closed subspace of homogeneous boundary conditions. We say that $u \in \mathcal{F}$ is a solution of $(T, \mathcal{B})$ for a given $f \in \mathcal{G}$ if $Tu = f$ and $u \in \mathcal{B}^\perp$. Note that have restricted ourselves to homogeneous conditions because the general solution is then obtained by adding a "particular solution" satisfying the inhomogeneous conditions. While for ODEs, this amounts to a simple interpolation problem, the treatment of PDEs is more involved.

In the *ODE setting*, $T = D^n + c_{n-1}D^{n-1} + \cdots + c_1 D + c_0$ is a monic differential operator of order n with coefficients $c_i \in \mathcal{G}$. For the spaces $\mathcal{F}, \mathcal{G}$ we could for example choose $\mathcal{F} = \mathcal{G} = C^\infty[a,b]$ or $\mathcal{F} = C^n[a,b]$ and $\mathcal{G} = C[a,b]$, as real or complex vector spaces. The differential operator T is surjective since every inhomogeneous linear differential equation has a solution in $\mathcal{F}$, e.g. given by the formula (3) below. The solution space of the homogeneous equation, $\operatorname{Ker} T$, has dimension n, so we require $\dim \mathcal{B} = n$, and we assume that $\mathcal{B}$ is given by a

basis $\beta_1, \dots, \beta_n$. Then the boundary problem reads as follows: Given $f \in \mathcal{G}$, find $u \in \mathcal{F}$ such that

$$\boxed{\begin{aligned} &Tu = f, \\ &\beta_1(u) = \cdots = \beta_n(u) = 0. \end{aligned}}$$

The boundary conditions can in principle be any linear functionals. In particular, they can be point evaluations of derivatives or also more general boundary conditions of the form $\beta(u) = \sum_{i=0}^{n-1} a_i\, u^{(i)}(a) + b_i\, u^{(i)}(b) + \int_a^b v(\xi)\, u(\xi)\, d\xi$ with $v \in \mathcal{F}$, known in the literature [11] as "Stieltjes boundary conditions". Integral boundary conditions also appear naturally when we factor a boundary problem along a given factorization of the differential operator (Section 3), and they appear in the normal forms of integro-differential operators (Section 4).

A boundary problem $(T, \mathcal{B})$ is *regular* if for each $f \in \mathcal{G}$ there exists exactly one solution u of $(T, \mathcal{B})$. Then we call the linear operator $G\colon \mathcal{G} \to \mathcal{F}$ that maps a right-hand side f to its unique solution $u = Gf$ the *Green's operator* for the boundary problem $(T, \mathcal{B})$, and we say that G solves the boundary problem $(T, \mathcal{B})$. Since $TGf = f$, we see that the Green's operator for a regular boundary problem $(T, \mathcal{B})$ is a right inverse of T, determined by the property $\operatorname{Im} G = \mathcal{B}^\perp$. Therefore we use the notation $G = (T, \mathcal{B})^{-1}$ for the Green's operator.

Regular boundary problems can be characterized as follows. A boundary problem is regular iff $\mathcal{B}^\perp$ is a complement of $\operatorname{Ker} T$ so that $\mathcal{F} = \operatorname{Ker} T \dotplus \mathcal{B}^\perp$ as a direct sum. For ODEs we have the following algorithmic regularity test (compare [12, p. 184] for the special case of two-point boundary conditions): A boundary problem $(T, \mathcal{B})$ for an ODE is regular iff the *evaluation matrix* $B = (\beta_i(u_j))$ is regular, where the β_i and u_j are any basis of respectively $\mathcal{B}$ and $\operatorname{Ker} T$.

Given any right inverse $\tilde{G}$ of a surjective linear map $T\colon \mathcal{F} \to \mathcal{G}$, the Green's operator for a regular boundary problem $(T, \mathcal{B})$ is given by $G = (1 - P)\tilde{G}$, where P is the projector with $\operatorname{Im} P = \operatorname{Ker} T$ and $\operatorname{Ker} P = \mathcal{B}^\perp$. Using this observation, we outline in the following how the Green's operator can be computed in the ODE setting.

Let $(T, \mathcal{B})$ be a regular boundary problem for an ODE of order n with $\mathcal{B} = [\beta_1, \dots, \beta_n]$, and let $u_1, \dots, u_n$ be a fundamental system of solutions. We first compute a right inverse of the differential operator T. This can be done by the usual variation-of-constants formula (see for example [13, p. 87] for continuous functions or [14] in a suitable integro-differential algebra setting): Let $W = W(u_1, \dots, u_n)$ be the Wronskian matrix and $d = \det W$. Moreover, let $d_i = \det W_i$, where W_i is the matrix obtained from W by replacing the ith column by the nth unit vector. Then the solution of the initial value problem $Tu = f$, $u(a) = u'(a) = \cdots = u^{(n-1)}(a) = 0$ is given by

$$u(x) = \sum_{i=1}^{n} u_i(x) \textstyle\int_a^x d_i(\xi)/d(\xi)\, f(\xi)\, d\xi. \tag{3}$$

The integral operator $T^{\blacklozenge}\colon f \mapsto u$ defined by (3) is a right inverse of T, which we also call the *fundamental right inverse*. Computing the projector $P\colon \mathcal{F} \to \mathcal{F}$ with $\operatorname{Im} P = [u_1, \dots, u_n]$ and $\operatorname{Ker} P = [\beta_1, \dots, \beta_n]^\perp$ is a linear algebra problem, see

for example [5, App. A.1]: Let B be the evaluation matrix $B = (\beta_i(u_j))$. Since $(T, \mathcal{B})$ is regular, B is invertible. Set $(\tilde{\beta}_1, \ldots, \tilde{\beta}_n)^t = B^{-1}(\beta_1, \ldots, \beta_n)^t$. Then the projector P is given by $u \mapsto \sum_{i=1}^n \tilde{\beta}_i(u)\, u_i$. Finally, we compute

$$G = (1 - P)T^{\blacklozenge} \tag{4}$$

to obtain the Green's operator for $(T, \mathcal{B})$.

3 Composing and Factoring Boundary Problems

In this section we discuss the composition of boundary problems corresponding to their Green's operators. We also describe how factorizations of a boundary problem along a given factorization of the defining operator can be characterized and constructed. We refer again to [5] for further details. In the following, we assume that all operators are defined on suitable spaces such that the composition is well-defined.

Definition 1. *We define the* composition *of boundary problems* $(T_1, \mathcal{B}_1)$ *and* $(T_2, \mathcal{B}_2)$ *by* $(T_1, \mathcal{B}_1) \circ (T_2, \mathcal{B}_2) = (T_1 T_2, \mathcal{B}_1 \cdot T_2 + \mathcal{B}_2)$.

So the boundary conditions from the first boundary problem are "translated" by the operator from the second problem. The composition of boundary problems is associative but in general not commutative. The next proposition tells us that the composition of boundary problems preserves regularity.

Proposition 1. *Let* $(T_1, \mathcal{B}_1)$ *and* $(T_2, \mathcal{B}_2)$ *be regular boundary problems with Green's operators* G_1 *and* G_2*. Then* $(T_1, \mathcal{B}_1) \circ (T_2, \mathcal{B}_2) = (T, \mathcal{B})$ *is regular with Green's operator* $G_2 G_1$ *so that* $((T_1, \mathcal{B}_1) \circ (T_2, \mathcal{B}_2))^{-1} = (T_2, \mathcal{B}_2)^{-1} \circ (T_1, \mathcal{B}_1)^{-1}$.

The simplest example of composing two boundary (more specifically, initial value) problems for ODEs is the following. Using the notation from the Introduction, one sees that $(D, [L]) \circ (D, [L]) = (D^2, [LD] + [L]) = (D^2, [L, LD])$.

Next we write the wave equation (2) as $\mathcal{P} = (D_t^2 - D_x^2, [u(x,0), u_t(x,0)])$, where $u(x,0)$ and $u_t(x,0)$ are short for the functionals $u \mapsto u(x,0)$ and $u \mapsto u_t(x,0)$, respectively, with x ranging over $\mathbb{R}$, and $[\ldots]$ denotes the orthogonal closure of the subspace generated by these functionals. For boundary problems with PDEs, we usually have to describe the boundary conditions as the orthogonal closure of some subspaces that we can describe in finite terms. As detailed in [5], we can still compute the composition of two such problems since taking the orthogonal closure commutes with the operations needed for computing the boundary conditions for the composite problem (precomposition with a linear operator and sum of subspaces).

Using this observation, we can compute $\mathcal{P}$ as the composition of the two boundary problems $\mathcal{P}_1 = (D_t - D_x, [u(x,0)])$ and $\mathcal{P}_2 = (D_t + D_x, [u(x,0)])$ as follows. By Definition 1, we see that $\mathcal{P}_1 \circ \mathcal{P}_2$ equals

$$(D_t^2 - D_x^2, [u_t(x,0) + u_x(x,0)] + [u(x,0)]) = (D_t^2 - D_x^2, [u(x,0), u_t(x,0)]), \tag{5}$$

where the last equality holds since $u(x,0)=0$ for $x\in\mathbb{R}$ implies also $u_x(x,0)=0$ for $x\in\mathbb{R}$, showing that $u_x(x,0)$ is in the orthogonal closure $[u(x,0)]$.

In the following, we assume that for a boundary problem $(T,\mathcal{B})$ we have a factorization $T=T_1T_2$ of the defining operator with surjective linear maps T_1,T_2. In [5], we characterize and construct all factorizations $(T,\mathcal{B})=(T_1,\mathcal{B}_1)\circ(T_2,\mathcal{B}_2)$ into boundary problems along the given factorization of T. We show in particular that if we factor a regular problem into regular problems, the left factor $(T_1,\mathcal{B}_1)$ is unique, and we can choose for the right factor $(T_2,\mathcal{B}_2)$ any subspace $\mathcal{B}_2\leq\mathcal{B}$ that makes the problem regular. Moreover, if G_2 is the Green's operator for some regular right factor $(T_2,\mathcal{B}_2)$, the boundary conditions for the left factor can be computed by $\mathcal{B}_1=\mathcal{B}\cdot G_2$. Factoring boundary problems for differential equations allows us to split a problem of higher order into subproblems of lower order, provided we can factor the differential operator. For the latter, we can exploit algorithms and results about factoring ordinary [15,16,17] and partial differential operators [18,19].

For ODEs we can factor boundary problems algorithmically as described in [5] and in an integro-differential algebra setting in [4]. There we assume that we are given a fundamental system of the differential operator T and a right inverse of T_2. As we will detail in the next paragraph, we can also compute boundary conditions $\mathcal{B}_2\leq\mathcal{B}$ such that $(T_2,\mathcal{B}_2)$ is a regular right factor, given only a fundamental system of T_2. We can then compute the left factor as explained above. This can be useful in applications, because it still allows us to factor a boundary problem if we can factor the differential operator and compute a fundamental system of only one factor. The remaining lower order problem can then be solved by numerical methods (and we expect that the integral conditions $\mathcal{B}_1=\mathcal{B}\cdot G_2$ may be beneficial since they are stable).

Let now $(T,\mathcal{B})$ be a boundary problem of order $m+n$ with boundary conditions $[\beta_1,\ldots,\beta_{m+n}]$. Let $T=T_1T_2$ be a factorization into factors of respective orders n and m, and let $u_1,\ldots,u_m$ be a fundamental system for T_2. We compute the "partial" $(m+n)\times m$ evaluation matrix $\tilde{B}=\beta_i(u_j)$. Since $(T,\mathcal{B})$ is regular, the full evaluation matrix is regular and hence the columns of $\tilde{B}$ are linearly independent. Therefore computing the reduced row echelon form yields a regular matrix C such that $C\tilde{B}=\left(\begin{smallmatrix}I_m\\0\end{smallmatrix}\right)$, where I_m is the $m\times m$ identity matrix. Let now $(\tilde{\beta}_1,\ldots,\tilde{\beta}_{m+n})^t=C(\beta_1,\ldots,\beta_{m+n})^t$ and $\mathcal{B}_2=[\tilde{\beta}_1,\ldots,\tilde{\beta}_m]$. Then $(T_2,\mathcal{B}_2)$ is a regular right factor since its evaluation matrix is I_m by our construction. See Appendix A for an example.

As a first example, we factor the two-point boundary problem $(D^2,[L,R])$ from the Introduction into two regular problems along the trivial factorization with $T_1=T_2=D$. The indefinite integral $A=\int_0^x$ is the Green's operator for the regular right factor $(D,[L])$. The boundary conditions for the unique left factor are $[LA,RA]=[0,RA]=[RA]$, where $RA=\int_0^1$ is the definite integral. So we obtain $(D,[RA])\circ(D,[L])=(D^2,[L,R])$ or in traditional notation

$$\boxed{\begin{array}{l}u'=f\\ \int_0^1 u(\xi)\,d\xi=0\end{array}}\circ\boxed{\begin{array}{l}u'=f\\ u(0)=0\end{array}}=\boxed{\begin{array}{l}u''=f\\ u(0)=u(1)=0\end{array}}.$$

Note that the boundary condition for the left factor is an integral (Stieltjes) boundary condition.

As an example of a boundary problem for a PDE, we factor the wave equation (2) along the factorization $D_t^2 - D_x^2 = (D_t - D_x)(D_t + D_x)$. In Appendix A, we show that one can use this factorization to determine algorithmically its Green's operator. The boundary problem $\mathcal{P}_2 = (D_t + D_x, [u(x,0)])$ is a regular right factor. In general, choosing boundary conditions in such a way that they make up a regular boundary problem for a given first-order right factor of a linear PDE amounts to a geometric problem involving the characteristics; compare also Section 4. The Green's operator for $\mathcal{P}_2$ is $G_2 f(x,t) = \int_{x-t}^{x} f(\xi, \xi - x + t)\, d\xi$. We can compute the boundary conditions for the left factor by $[u(x,0)\cdot G_2, u_t(x,0)\cdot G_2] = [0, u(x,0)] = [u(x,0)]$ so that $\mathcal{P}_1 = (D_t - D_x, [u(x,0)])$ is the desired left factor. In (5) we have already verified that $\mathcal{P}_1 \circ \mathcal{P}_2 = \mathcal{P}$.

4 Representation of Integro-differential Operators

For representing ordinary boundary problems as well as their Green's operators in a single algebraic structure, we have introduced the algebra of *integro-differential operators* $\mathcal{F}[\partial, \int]$ in [4], see also [14] for a summary. It is based on integro-differential algebras, which bring together the usual derivation structure with a suitable notion of indefinite integration and evaluation. The integro-differential operators are defined as a quotient of the free algebra in the corresponding operators (derivation, integration, evaluation, and multiplication) modulo an infinite parametrized Gröbner basis. See Section 5 for more details and an implementation. Alternatively, integro-differential operators can also be defined directly in terms of normal forms [20].

Let us now turn to the treatment of *partial differential equations*. We are currently forging an adequate notion of integro-differential operators for describing the Green's operators of an interesting class of PDEs, just as $\mathcal{F}[\partial, \int]$ can be used for ODEs. In the remainder of this section we can only give a flavor (and a small test implementation) of how integro-differential operators for PDEs might look like in a simple case that includes the unbounded wave equation (2).

We construct a ring $\mathcal{R}$ of integro-differential operators acting on the function space $\mathcal{F} = C^\infty(\mathbb{R} \times \mathbb{R})$; for simplicity we neglect here the restriction to $\mathbb{R} \times \mathbb{R}_{\geq 0}$. The ring $\mathcal{R}$ is defined as the free $\mathbb{C}$-algebra in the following indeterminates given with their respective action on a function $f(x,t) \in \mathcal{F}$.

Name	Indeterminates	Action
Differential operators	$D_x,\ D_t$	$f_x(x,t),\ f_t(x,t)$
Integral operators	$A_x,\ A_t$	$\int_0^x f(\xi,t)\, d\xi,\ \int_0^t f(x,\tau)\, d\tau$
Evaluation operators	$L_x,\ L_t$	$f(0,t),\ f(x,0)$
Substitution operators	$\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) \in \mathrm{GL}(\mathbb{R}, 2)$	$f(ax+bt, cx+dt)$

Similar to the identities governing $\mathcal{F}[\partial, \int]$, described in [4], various relations among the above operators can now be encoded in a quotient of $\mathcal{R}$. We will only sketch the most important relations, focusing on those that are needed for the sample computations. (In a more complete setup, the indeterminates should also be chosen in a more economical way. For example, it is possible to subsume the evaluations under the substitutions if one allows all affine transformations by adding translations and singular matrices.)

First of all, we can transfer all relations from $\mathcal{F}[\partial, \int]$ that involve D, A and L, once for the corresponding x-operators and once for the corresponding t-operators. Furthermore, each x-operator commutes with each t-operator. For example, we have $D_x A_x = 1$ but $D_x A_t = A_t D_x$. For normalizing such commutative products, we write the x-operators left of the t-operators. Our strategy for normal forms is thus similar to the case of $\mathcal{F}[\partial, \int]$, the only new ingredient being the substitutions: We will move them to the left as much as possible.

Since substitutions operate on the arguments, it is clear that we must reverse their order when multiplying them as elements of $\mathcal{R}$. But the most important relations are those that connect the substitutions with the integro-differential indeterminates: The chain rule governs the interaction with differentiation, the substitution rule with integration. While the former gives rise to the identities

$$D_x M = a\,M D_x + c\,M D_t \quad \text{and} \quad D_t M = b\,M D_x + d\,M D_t$$

for a matrix $M = \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$, the relation between M and integrals is a bit subtler. If M is an upper triangular matrix (so that $c = 0$ and $a \neq 0$), the substitution rule yields

$$A_x M = \tfrac{1}{a}(1 - L_x) M A_x,$$

and if M is a lower triangular matrix (so that $b = 0$ and $d \neq 0$) similarly $A_t M = \frac{1}{d}(1 - L_t) M A_t$.

But there are no such identities for pushing $\left(\begin{smallmatrix} 1 & 0 \\ c & 1 \end{smallmatrix}\right)$ left of A_x or $\left(\begin{smallmatrix} 1 & b \\ 0 & 1 \end{smallmatrix}\right)$ left of A_t; we leave them in their place for the normal forms. For treating the general case, we make use of a variant of the Bruhat decomposition [21, p. 349], writing $M \in \mathrm{GL}(\mathbb{R}, 2)$ as $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 0 \\ c/a & 1 \end{smallmatrix}\right) \left(\begin{smallmatrix} a & b \\ 0 & (ad-bc)/a \end{smallmatrix}\right)$ if $a \neq 0$ and $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} b & 0 \\ d & c \end{smallmatrix}\right) \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ if $a = 0$. Alternatively, we may also use $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & b/d \\ 0 & 1 \end{smallmatrix}\right) \left(\begin{smallmatrix} (ad-bc)/d & 0 \\ c & d \end{smallmatrix}\right)$ if $d \neq 0$ and $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} b & a \\ 0 & c \end{smallmatrix}\right) \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ if $d = 0$. The former decomposition is applied in deriving the rule for A_x, which reads

$$A_x \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \tfrac{1}{a}\,(1 - L_x) \left(\begin{smallmatrix} a & b \\ 0 & (ad-bc)/a \end{smallmatrix}\right) A_x \left(\begin{smallmatrix} 1 & 0 \\ c/a & 1 \end{smallmatrix}\right)$$

if $a \neq 0$ and otherwise $A_x \left(\begin{smallmatrix} 0 & b \\ c & d \end{smallmatrix}\right) = \frac{1}{c}\,(1 - L_x) \left(\begin{smallmatrix} 0 & b \\ c & d \end{smallmatrix}\right) A_t$. Analogously, the latter decomposition yields the rule for A_t as

$$A_t \left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \tfrac{1}{d}\,(1 - L_t) \left(\begin{smallmatrix} (ad-bc)/d & 0 \\ c & d \end{smallmatrix}\right) A_t \left(\begin{smallmatrix} 1 & b/d \\ 0 & 1 \end{smallmatrix}\right)$$

if $d \neq 0$ and otherwise $A_t \left(\begin{smallmatrix} a & b \\ c & 0 \end{smallmatrix}\right) = \frac{1}{b}\,(1 - L_t) \left(\begin{smallmatrix} a & b \\ c & 0 \end{smallmatrix}\right) A_x$.

According to the rules above, an $\mathcal{R}$-operator like $A_x \left(\begin{smallmatrix} 1 & 0 \\ k & 1 \end{smallmatrix}\right)$ is in normal form. Also $A_x \left(\begin{smallmatrix} 1 & 0 \\ k & 1 \end{smallmatrix}\right) A_x$ is a normal form, describing an area integral. For interpreting

it geometrically, it is convenient to postmultiply it with the reverse shear, obtaining thus the integral operator $T_k = \left(\begin{smallmatrix} 1 & 0 \\ -k & 1 \end{smallmatrix}\right) A_x \left(\begin{smallmatrix} 1 & 0 \\ k & 1 \end{smallmatrix}\right) A_x$. One can easily verify that $T_k f(x,t)$ represents the integral of f taken over the triangle with vertices (x,t), $(0,y)$ and $(0,t-kx)$. This is the triangle delimited by the y-axis, the horizontal through (x,y), and the slanted line through (x,t) with slope k. Similar interpretations can be given for products involving A_t.

Finally, we need some rules relating substitutions with evaluations. Here the situation is analogous to the integrals: We can move "most" of the substitutions to the left of an evaluation, but certain shears remain on the right. In detail, we have the rules

$$L_x\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & d \end{smallmatrix}\right) L_x \left(\begin{smallmatrix} 1 & b/d \\ 0 & 1 \end{smallmatrix}\right) \quad \text{if } d \neq 0 \qquad L_x\left(\begin{smallmatrix} a & b \\ c & 0 \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & b \\ 1 & 0 \end{smallmatrix}\right) L_t \quad \text{otherwise}$$

and

$$L_t\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} a & 0 \\ 0 & 1 \end{smallmatrix}\right) L_t \left(\begin{smallmatrix} 1 & 0 \\ c/a & 1 \end{smallmatrix}\right) \quad \text{if } a \neq 0 \qquad L_t\left(\begin{smallmatrix} 0 & b \\ c & d \end{smallmatrix}\right) = \left(\begin{smallmatrix} 0 & 1 \\ c & 0 \end{smallmatrix}\right) L_x \quad \text{otherwise.}$$

As before, certain products remain as normal forms, for example $L_x\left(\begin{smallmatrix} 1 & k \\ 0 & 1 \end{smallmatrix}\right)$. Such an operator acts on a function $f \in \mathcal{F}$ as $f(kt,t)$, collapsing the bivariate function f to the univariate restriction along the diagonal line $x = kt$.

The language of $\mathcal{R}$-operators is not very expressive, but enough for our modest purposes at this point—expressing the boundary problem (2) and computing its Green's operator. Let us first look at the general first-order boundary problem with constant coefficients, prescribing homogeneous Dirichlet conditions on an arbitrary line. Fixing the parameters $a, b, c, k \in \mathbb{R}$, it reads as follows:

$$\boxed{\begin{aligned} a\,u_x + b\,u_t &= f \\ u(kt+c,t) &= 0 \end{aligned}} \tag{6}$$

Here $(a,b)^t$ determines the direction (and speed) of the ground characteristics, while $x = kt + c$ gives the line of boundary values. Of course this excludes the horizontal lines $t = \text{const}$, which would have to be treated separately, in a completely analogous manner. Since (in this paper) we are interested only in regular boundary problems, the characteristics must have a transversal intersection with the line of boundary values. Hence we stipulate that $a - kb \neq 0$. Moreover, we will also assume $a \neq 0$; for otherwise one may switch the x- and t-coordinates. A straightforward computation (or a suitable computer algebra system) gives now

$$u(x,y) = \frac{1}{a} \int_X^x f(\xi, \tfrac{b}{a}(\xi - x) + t)\, d\xi \qquad \text{with} \qquad X = \frac{ac + (at - bx)k}{a - bk}.$$

This solution for the general case can be reduced to $(a,b)^t = (1,0)^t$ and $k = 0$ by first rotating (a,b) into horizontal position, then normalizing it through x-scaling, and finally shearing the line of boundary values into vertical position. This yields the factorization

$$u(x,y) = \left(\begin{smallmatrix} 1/K & -k/K \\ -b/L & a/L \end{smallmatrix}\right) \cdot \textstyle\int_{c/K}^{x} \cdot \left(\begin{smallmatrix} a & kL/K \\ b & L/K \end{smallmatrix}\right) f(x,y), \tag{7}$$

where $K = a - bk$ and $L = a^2 + b^2$. This is almost an $\mathcal{R}$-operator, except that we have only allowed $A_x = \int_0^x$ and its t-analog, so we cannot express $\int_{c/K}^x$ unless we allow more evaluations such that we could write the required integral as $A_x - L_x^{c/K} A_x$, where L_x^ξ acts on a function $g(x, y)$ as $g(\xi, y)$.

While it would be straightforward to incorporate such evaluations by adding suitable relations, it is enough for our purposes to restrict the line of boundaries: We require it to pass through the origin so that $c = 0$. In this case we have of course $\int_{c/K}^x = A_x$, and (7) shows that we can indeed write the Green's operator in the $\mathcal{R}$ language.

5 Implementation in Theorema

As explained in Sections 2 and 4, we compute the Green's operator of a boundary problem for an ODE as an *integro-differential operator*. These operators are realized as noncommutative polynomials (introduced by a generic construct for monoid algebras), taken modulo an infinite parametrized Gröbner basis.

As coefficients we allow either standard polynomials or—more generally—exponential polynomials. Informally speaking, an *exponential polynomial* is a linear combination of terms having the form $x^n e^{\lambda x}$, where n is a natural and λ a complex number. Both the standard and the exponential polynomials can again be generated as an instance of the monoid algebra, respectively using $\mathbb{N}$ and $\mathbb{N} \times \mathbb{C}$ as a term monoid. In this way, we have complete algorithmic control over the coefficient functions (modulo Mathematica's simplifier for constants); see also [22]. Alternatively, we can also take as coefficients all functions representable in Mathematica and let it do the operations on them.

We describe now briefly the representation of integro-differential operators and the implementation of the main algorithms solving, composing and factoring boundary problems. The implementation will soon be available at the website `www.theorema.org`. It is based on Theorema [6], a system designed as an integrated environment for doing mathematics, in particular proving, computing, and solving in various domains of mathematics. Its core language is higher-order predicate logic, containing a natural *programming language* such that algorithms can be coded and verified in a unified formal frame.

We make heavy use of *functors*, introduced and first implemented in Theorema by Buchberger. The general idea—and its use for structuring those domains in which Gröbner bases can be computed—is described in [23,24], where one can also find references to original and early papers by Buchberger on the subject. For a general discussion of functor programming, see also [25].

Functors are a powerful tool for building up *hierarchical domains* in mathematics in a modular and generic way that unites elegance and formal clarity. In Theorema, the notion of a functor is akin to functors in ML, not to be confused with the functors of category theory. From a computational point of view, a Theorema functor is a higher-order function that produces a new domain (carrier and operations) from given domains: operations in the new domain are defined in terms of operations in the underlying domains. Apart from this computational

aspect, functors also have an important reasoning aspect—a functor transports properties of the input domains to properties of the output domain, for example by "conservation theorems".

The `MonoidAlgebra` is the crucial functor that builds up *polynomials*, starting from the base categories of fields with an ordering and ordered monoids. We construct first the free vector space `V` over a field `K` generated by the set of words in an ordered monoid `W` via the functor `FreeVecSpc[K,W]`. Then we extend this domain by introducing a multiplication using the corresponding operations in `K` and `W` as follows.

```
MonoidAlgebra[K, W] = where[V = FreeVecSpc[K, W],
  Functor[P, any[c, d, f, g, ξ, η, m̄, n̄],
    s = ⟨⟩
    ...(* linear operations from V *)
    (* multiplication *)
    ⟨⟩ *_P g = ⟨⟩                                                    ]]
    f *_P ⟨⟩ = ⟨⟩
    ⟨⟨c, ξ⟩, m̄⟩ *_P ⟨⟨d, η⟩, n̄⟩ = ⟨⟨c *_K d, ξ *_W η⟩⟩ +_P ⟨⟨c, ξ⟩⟩ *_P ⟨n̄⟩ +_P ⟨m̄⟩ *_P ⟨⟨d, η⟩, n̄⟩
```

For building up the *integro-differential operators* over an integro-differential algebra $\mathcal{F}$ of coefficient functions, `FreeIntDiffOp[`$\mathcal{F}$`,K]` constructs an instance of the monoid algebra with the word monoid over the infinite alphabet consisting of the letters ∂ and $\int$ along with a basis of $\mathcal{F}$ and all multiplicative characters corresponding to evaluations at points in `K`.

```
Definition["IntDiffOp", any[ℱ, K],
 IntDiffOp[ℱ, K] = where[𝔄 = FreeIntDiffOp[ℱ, K], 𝒢 = GreenSystem[ℱ, K]
     QuotAlg[GBNF[𝔄, 𝒢]]]
]
```

The `GreenSystem` functor contains the encoding of the rewrite system described in Table 1 of [4,14], representing a noncommutative Gröbner basis. The normal forms with respect to total reduction modulo infinite Gröbner bases are introduced in the `GBNF` functor, while the `QuotAlg` functor creates the quotient algebra from the corresponding canonical simplifier.

In Appendix A, we present a few examples of boundary problems for ODEs whose *Green's operators* are computed using (4), which now takes on the following concrete form in Theorema code.

```
GreensOp[F, ℬ] = (1 -_𝔄 Proj[ℬ, F]) *_𝔄 RightInv[F]
                 P                   P
```

Here $\mathcal{B}$ is the vector of boundary conditions and `F` the given fundamental system of solutions.

In a way similar to the integro-differential operators $\mathcal{F}[\partial, \int]$ for ODEs, we have also implemented the integro-differential operators $\mathcal{R}$ for the simple PDE setting outlined in Section 4. Using the same functor hierarchy, we added the corresponding rules for the operators D_x, D_t, A_x, A_t, L_x, L_t and the substitution operators defined by matrices in $\mathrm{GL}(\mathbb{R}, 2)$. Moreover, we implemented the computation of Green's operators for first-order boundary problems (7). With the

factorization (5) we can then compute the Green's operator for the unbounded wave equation (Appendix A).

6 Conclusion

The implementation of our symbolic framework for boundary problems allows us in particular to *solve boundary problems* for *ODEs* from a given fundamental system of the corresponding homogeneous equations. Given a factorization of the differential operator and a fundamental system of one of the factors, we can also *factor boundary problems* into lower order problems. In both cases it would be interesting to investigate the combination with numerical approaches to differential equations and boundary problems. For example, how can we use a fundamental system coming from a numerical algorithm or how can numerical methods be adapted to deal with integral boundary conditions?

The current setting for *PDEs* is of course still very limited and should only be seen as a starting point for future work. But in combination with our factorization approach, we believe that it can be extended to include more complicated problems. For example, the wave equation on the *bounded interval* $[0,1]$, which in our notation reads as $\mathcal{P} = (D_t^2 - D_x^2, [u(x,0), u_t(x,0), u(0,t), u(1,t)])$ with x ranging over $[0,1]$ and t over $\mathbb{R}_{\geq 0}$, can be factored [5] into $\mathcal{P} = \mathcal{P}_1 \circ \mathcal{P}_2$ with

$$\mathcal{P}_1 = (D_t - D_x, [u(x,0), \textstyle\int_{\max(1-t,0)}^{1} u(\xi, \xi + t - 1)\, d\xi])$$

and $\mathcal{P}_2 = (D_t + D_x, [u(x,0), u(0,t)])$. The more complicated structure of the Green's operator for $\mathcal{P}$ (it involves a finite sum with an upper bound depending on its argument) is reflected in the Green's operator for the left factor $\mathcal{P}_1$. Its computation leads in this case to a simple functional equation, but a systematic approach to compute and represent Green's operators for PDEs with *integral boundary conditions* still needs to be developed. In a generalized setting including the bounded wave equation, we would also have to allow for more complicated geometries: as a first step bounded intervals and then also arbitrary convex sets.

References

1. Stakgold, I.: Green's functions and boundary value problems. John Wiley & Sons, New York (1979)
2. Rosenkranz, M., Buchberger, B., Engl, H.W.: Solving linear boundary value problems via non-commutative Gröbner bases. Appl. Anal. 82, 655–675 (2003)
3. Rosenkranz, M.: A new symbolic method for solving linear two-point boundary value problems on the level of operators. J. Symbolic Comput. 39, 171–199 (2005)
4. Rosenkranz, M., Regensburger, G.: Solving and factoring boundary problems for linear ordinary differential equations in differential algebras. J. Symbolic Comput. 43, 515–544 (2008)
5. Regensburger, G., Rosenkranz, M.: An algebraic foundation for factoring linear boundary problems. Ann. Mat. Pura Appl. 188(4), 123–151 (2009)

6. Buchberger, B., Craciun, A., Jebelean, T., Kovacs, L., Kutsia, T., Nakagawa, K., Piroi, F., Popov, N., Robu, J., Rosenkranz, M., Windsteiger, W.: Theorema: Towards computer-aided mathematical theory exploration. J. Appl. Log. 4, 359–652 (2006)
7. Buchberger, B.: An algorithm for finding the bases elements of the residue class ring modulo a zero dimensional polynomial ideal (German). PhD thesis, Univ. of Innsbruck (1965); English translation J. Symbolic Comput. 41(3-4), 475–511 (2006)
8. Buchberger, B.: Introduction to Gröbner bases. In: Buchberger, B., Winkler, F. (eds.) Gröbner bases and applications, Cambridge Univ. Press, Cambridge (1998)
9. Mora, T.: An introduction to commutative and noncommutative Gröbner bases. Theoret. Comput. Sci. 134, 131–173 (1994)
10. Köthe, G.: Topological vector spaces, vol. I. Springer, New York (1969)
11. Brown, R.C., Krall, A.M.: Ordinary differential operators under Stieltjes boundary conditions. Trans. Amer. Math. Soc. 198, 73–92 (1974)
12. Kamke, E.: Differentialgleichungen. Lösungsmethoden und Lösungen. Teil I: Gewöhnliche Differentialgleichungen. Akademische Verlagsgesellschaft, Leipzig (1967)
13. Coddington, E.A., Levinson, N.: Theory of ordinary differential equations. McGraw-Hill Book Company, Inc., New York (1955)
14. Rosenkranz, M., Regensburger, G.: Integro-differential polynomials and operators. In: Jeffrey, D. (ed.) Proceedings of ISSAC 2008, pp. 261–268. ACM, New York (2008)
15. van der Put, M., Singer, M.F.: Galois theory of linear differential equations. Springer, Berlin (2003)
16. Schwarz, F.: A factorization algorithm for linear ordinary differential equations. In: Proceedings of ISSAC 1989, pp. 17–25. ACM, New York (1989)
17. Tsarev, S.P.: An algorithm for complete enumeration of all factorizations of a linear ordinary differential operator. In: Proceedings of ISSAC 1996, pp. 226–231. ACM, New York (1996)
18. Grigoriev, D., Schwarz, F.: Loewy- and primary decompositions of D-modules. Adv. in Appl. Math. 38, 526–541 (2007)
19. Tsarev, S.P.: Factorization of linear partial differential operators and Darboux integrability of nonlinear PDEs. SIGSAM Bull. 32, 21–28 (1998)
20. Regensburger, G., Rosenkranz, M., Middeke, J.: A skew polynomial approach to integro-differential operators. In: Proceedings of ISSAC 2009. ACM, New York (to appear, 2009)
21. Cohn, P.M.: Further algebra and applications. Springer, London (2003)
22. Buchberger, B., Regensburger, G., Rosenkranz, M., Tec, L.: General polynomial reduction with Theorema functors: Applications to integro-differential operators and polynomials. ACM Commun. Comput. Algebra 42, 135–137 (2008)
23. Buchberger, B.: Groebner rings and modules. In: Maruster, S., Buchberger, B., Negru, V., Jebelean, T. (eds.) Proceedings of SYNASC 2001, pp. 22–25 (2001)
24. Buchberger, B.: Groebner bases in Theorema using functors. In: Faugere, J., Wang, D. (eds.) Proceedings of SCC 2008, pp. 1–15. LMIB Beihang University Press (2008)
25. Windsteiger, W.: Building up hierarchical mathematical domains using functors in Theorema. Electr. Notes Theor. Comput. Sci. 23, 401–419 (1999)

A Sample Computations

Let us again consider example (1). By our implementation, we obtain the Green's operator for the boundary problem with the corresponding Green's function. As noted in [3], the Green's function provides a canonical form for the Green's operator. In the following, we use the notation $Au = \int_0^x u(\xi)\,d\xi$, $Bu = \int_x^1 u(\xi)\,d\xi$, $Lu = u(0)$, $Ru = u(1)$, and $A1f(x,t) = \int_0^x f(\xi,t)\,d\xi$.

```
Compute[AsGreen_g[GreensOp_B[D^2, {{{1, {{"⌊⌋", 0}}}}, {{1, {{"⌊⌋", 1}}}}}]]]
```

```
-A x - x B + x A x + x B x
```

```
Compute[GreensFct_gf[GreensOp_B[D^2, {{{1, {{"⌊⌋", 0}}}}, {{1, {{"⌊⌋", 1}}}}}]]]
```

$$\begin{cases} -\xi + x\,\xi & \Leftarrow \quad \xi \le x \\ -x + x\,\xi & \Leftarrow \quad x < \xi \end{cases}$$

As explained in Section 3, we can factor (1) along a factorization of the differential operator, given a fundamental system for the right factor. Here is how we can compute the boundary conditions of the left and right factor problems, respectively.

```
Compute[AsGreen_g[Factorize_Bm[D, D, {{{1, {{"⌊⌋", 0}}}}, {{1, {{"⌊⌋", 1}}}}, {{{1, {}}}}}]]]
```

```
{{A + B}, {L}}
```

We consider as a second example the fourth order boundary problem [4, Ex. 33]:

$$\boxed{\begin{aligned} &u'''' + 4u = f, \\ &u(0) = u(1) = u'(0) = u'(1) = 0. \end{aligned}} \tag{8}$$

Factoring the boundary problem along $D^4 + 4 = (D^2 - 2i)(D^2 + 2i)$, we obtain the following boundary conditions for the factor problems.

```
Compute[AsGreen_g[Factorize_Bm[D^2 - 2 i, D^2 + 2 i,
  {{{1, {{"⌊⌋", 0}}}}, {{1, {{"⌊⌋", 1}}}}, {{1, {{"⌊⌋", 0}, "∂"}}}, {{1, {{"⌊⌋", 1}, "∂"}}},
  {{{1, {{"⌈⌉", {0, -1 + i}}}}}, {{1, {{"⌈⌉", {0, 1 + (-1) i}}}}}}]]]
```

```
{{A E^(Complex[-1,1]) x + B E^(Complex[-1,1]) x, A E^(Complex[1,-1]) x + B E^(Complex[1,-1]) x}, {L, R}}
```

With our implementation we can also compute its Green's operator and verify the solution presented in [4].

The final example for ODEs is a third order boundary problem with exponential coefficients.

$$\boxed{\begin{aligned} &u''' - (e^x + 2)u'' - u' + (e^x + 2)u = f, \\ &u(0) = u(1) = u'(1) = 0. \end{aligned}} \tag{9}$$

Here we use as coefficient algebra all functions representable in Mathematica. The Green's operator is computed as follows.

```
Compute[GreensOp_B[⟨⟨⟨1, mma[e^x]⟩⟩, ⟨⟨1, mma[e^-x]⟩⟩, ⟨⟨-1, mma[e^e^x e^-x]⟩, ⟨1, mma[e^e^x]⟩⟩⟩,
  ⟨⟨⟨1, ⟨⟨"⌊⌋", 0⟩⟩⟩⟩, ⟨⟨1, ⟨⟨"⌊⌋", 1⟩⟩⟩⟩, ⟨⟨1, ⟨⟨"⌊⌋", 1⟩, "∂"⟩⟩⟩⟩]]
```

$$(-1+e)^{-2}\, e^{-1+e}\, e^{e^x} A + (-1+e)^{-2}\, e^{-1+e}\, e^{e^x} B + (-1)\,(-1+e)^{-2}\, e^{-1+e}\, e^{e^{x}+(-1)\,x} A +$$
$$(-1)\,(-1+e)^{-2}\, e^{-1+e}\, e^{e^{x}+(-1)\,x} B + \left(\frac{1}{2} + \frac{1}{2}\,(-1+e)^{-2}\right) e^{-1\,x} A + \frac{1}{2}\,(-1+e)^{-2}\, e^{-1\,x} B +$$
$$\left(\frac{-1}{2}\right)(-1+e)^{-2}\, e^{x} A + \left(\frac{-1}{2}\right)(-1+e)^{-2}\, e^{x} B + (-1+e)^{-2}\, e^{-1+e}\, e^{e^x} A\, e^{-2\,x} +$$
$$(-2)\,(-1+e)^{-2}\, e^{-1+e}\, e^{e^x} A\, e^{-1\,x} + (-1)\, e^{e^x} B\, e^{-1\, e^{x}+(-2)\,x} + (-1+e)^{-2}\, e^{-1+e}\, e^{e^x} B\, e^{-2\,x} +$$
$$(-2)\,(-1+e)^{-2}\, e^{-1+e}\, e^{e^x} B\, e^{-1\,x} + (-1)\,(-1+e)^{-2}\, e^{-1+e}\, e^{e^{x}+(-1)\,x} A\, e^{-2\,x} + 2\,(-1+e)^{-2}\, e^{-1+e}\, e^{e^{x}+(-1)\,x} A\, e^{-1\,x} +$$
$$e^{e^{x}+(-1)\,x} B\, e^{-1\, e^{x}+(-2)\,x} + (-1)\,(-1+e)^{-2}\, e^{-1+e}\, e^{e^{x}+(-1)\,x} B\, e^{-2\,x} + 2\,(-1+e)^{-2}\, e^{-1+e}\, e^{e^{x}+(-1)\,x} B\, e^{-1\,x} +$$
$$\left(1 + \frac{1}{2}\left(-1 + (-1+e)^{-2}\right)\right) e^{-1\,x} A\, e^{-2\,x} + \left(-1 + (-1)\,(-1+e)^{-2}\right) e^{-1\,x} A\, e^{-1\,x} +$$
$$\frac{1}{2}\left(-1 + (-1+e)^{-2}\right) e^{-1\,x} B\, e^{-2\,x} + (-1)\,(-1+e)^{-2}\, e^{-1\,x} B\, e^{-1\,x} + \left(\frac{-1}{2} + \frac{1}{2}\,(-2+e)\,(-1+e)^{-2}\, e\right) e^{x} A\, e^{-2\,x} +$$
$$(-1+e)^{-2}\, e^{x} A\, e^{-1\,x} + \frac{1}{2}\,(-2+e)\,(-1+e)^{-2}\, e\, e^{x} B\, e^{-2\,x} + (-1+e)^{-2}\, e^{x} B\, e^{-1\,x}$$

As a last example, we return to the boundary problem for the *wave equation* (2). With Proposition 1 and using the factorization (5), we can compute the Green's operator for (2) simply by composing the Green's operators of the first-order problems $\mathcal{P}_1 = (D_t - D_x, [u(x,0)])$ and $\mathcal{P}_2 = (D_t + D_x, [u(x,0)])$. Relative to the setting in Section 4, we switch the x- and t-coordinates.

```
Compute[GreensOp_B[1, -1, 0] *_B GreensOp_B[1, 1, 0]]
```

```
⟨⟨1, ⟨⟨mat, ⟨⟨1, 0⟩, ⟨-1, 1⟩⟩⟩, A1, ⟨mat, ⟨⟨1, 0⟩, ⟨2, 1⟩⟩⟩, A1, ⟨mat, ⟨⟨1, 0⟩, ⟨-1, 1⟩⟩⟩⟩⟩⟩
```

Interchanging again t and x, this corresponds in the usual notation to $G_1 f(x,t) = \int_0^t f(\xi, -\xi + x + t)\, d\xi$ and $G_2 f(x,t) = \int_0^t f(\xi, \xi + x - t)\, d\xi$, which yields

$$G_2 G_1 f(x,t) = \int_0^t \int_0^\tau f(\xi, 2\tau - \xi + x - t)\, d\xi\, d\tau$$

for the Green's operator of the unbounded wave equation (2).

Insurance: Mathematics and Economics 46 (2010) 42–51

Contents lists available at ScienceDirect

Insurance: Mathematics and Economics

journal homepage: www.elsevier.com/locate/ime

An algebraic operator approach to the analysis of Gerber–Shiu functions

Hansjörg Albrecher [a,*], Corina Constantinescu [b], Gottlieb Pirsic [b], Georg Regensburger [b], Markus Rosenkranz [b]

[a] Institute of Actuarial Science, Faculty HEC, University of Lausanne, Extranef Building, CH-1015 Lausanne, Switzerland

[b] Johann Radon Institute for Computational and Applied Mathematics (RICAM), Altenbergerstrasse 69, A-4040 Linz, Austria

ARTICLE INFO

Article history:
Received December 2008
Received in revised form
February 2009
Accepted 7 February 2009

ABSTRACT

We introduce an algebraic operator framework to study discounted penalty functions in renewal risk models. For inter-arrival and claim size distributions with rational Laplace transform, the usual integral equation is transformed into a boundary value problem, which is solved by symbolic techniques. The factorization of the differential operator can be lifted to the level of boundary value problems, amounting to iteratively solving first-order problems. This leads to an explicit expression for the Gerber–Shiu function in terms of the penalty function.

1. Introduction

We consider the collective renewal risk model introduced by Sparre Andersen (1957) that describes the amount of free capital $U(t)$ at time t in an insurance portfolio by

$$U(t) = u + ct - \sum_{k=1}^{N(t)} X_k.$$

Here $N(t)$ is a renewal process that counts the number of claims incurred during the time interval $(0, t]$, the constant c is the premium rate and the random variables $(X_k)_{k\geq 0}$ denote the claim sizes that occur at random times $(T_k)_{k\geq 0}$, with $\tau_k = T_k - T_{k-1}$ i.i.d. random variables denoting the k-th interclaim (or inter-arrival) time ($T_0 = 0$). The initial surplus (after the claim at time 0 is paid) is given by $u \geq 0$. Moreover, $(X_k)_{k\geq 0}$ and $(\tau_k)_{k\geq 1}$ are assumed to be independent. Ruin occurs when the surplus process becomes negative for the first time, so the time of ruin is given by

$$T_u = \inf\{t \mid U(t) < 0\}$$

and the ruin probability of a company having initial capital u is given by

$$\psi(u) = P(T_u < \infty \mid U(0) = u).$$

The net profit condition $c\mathbb{E}(T_k) > \mathbb{E}(X_k)$ is imposed to ensure that $\psi(u) < 1$ for all $u \geq 0$.

Denoting by $f(x, y, t \mid u)$ the joint probability density function of the surplus immediately before ruin $U(T_u-)$, the deficit at ruin $|U(T_u)|$ and the time of ruin T_u, we have

$$\int_0^\infty \int_0^\infty \int_0^\infty f(x, y, t \mid u)\mathrm{d}x\mathrm{d}y\mathrm{d}t = \psi(u).$$

Let $w(x, y)$ be a penalty function, nonnegative for $x \geq 0, y \geq 0$. Then for $u \geq 0$, the expected discounted penalty function (also called Gerber–Shiu function) is defined by

$$\begin{aligned} m(u) &= \mathbb{E}\left(\mathrm{e}^{-\delta T_u}\, w(U(T_u-), |U(T_u)|)\, 1_{T_u<\infty} \mid U(0) = u\right) \\ &= \int_0^\infty \int_0^\infty \int_0^\infty \mathrm{e}^{-\delta t}\, w(x, y) f(x, y, t \mid u)\mathrm{d}x\mathrm{d}y\mathrm{d}t, \end{aligned}$$

where $\delta > 0$ is a discount rate.

Since the introduction of this function in the compound Poisson model in the papers of Gerber and Shiu (1997, 1998), there has been a vast literature on its analysis and extensions to more general models. Li and Garrido (2004) and Gerber and Shiu (2005) were the first to investigate the Gerber–Shiu function in renewal models. In this paper we will concentrate on a new method for deriving explicit expressions for $m(u)$ in the case of renewal models. In the renewal context, explicit expressions are usually restricted to models where the claim size distribution and in particular the interclaim distribution are (a subclass of) distributions with rational Laplace transform (which includes Erlang and phase-type distributions as well as mixtures of these); see also Willmot (1999) and Li and Garrido (2005b). Our method is perfectly suitable for this class of distributions.

The established methods for deriving explicit expressions for functions arising in risk theory (e.g. ruin probability, Laplace transform of the time to ruin, Gerber–Shiu function) are either based on defective renewal equations or integral equations

* Corresponding author.
E-mail addresses: hansjoerg.albrecher@unil.ch (H. Albrecher), corina.constantinescu@oeaw.ac.at (C. Constantinescu), gpirsic@gmail.com (G. Pirsic), georg.regensburger@oeaw.ac.at (G. Regensburger), markus.rosenkranz@oeaw.ac.at (M. Rosenkranz).

0167-6687/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.insmatheco.2009.02.002

(Volterra of second kind). Specifically, starting with the defective renewal equation satisfied by the Gerber–Shiu function, Lin and Willmot (2000) propose a solution expressed in terms of the tail of a compound geometric distribution. For particular claim sizes (combinations of exponentials, mixture of Erlangs) they derive explicit analytic solutions for this distribution. In Willmot (2007) this defective renewal equation method is adapted to the analysis of renewal risk models with arbitrary distributions. Another strategy, based on the defective renewal equation, was suggested in the classical compound Poisson model by Drekic et al. (2004). They use *Mathematica* to obtain the moments of the time to ruin, based on the system of defective renewal difference equations derived by Lin and Willmot (2000). In this paper, we introduce an algebraic operator approach with symbolic techniques for deriving explicit expressions for Gerber–Shiu functions. These techniques are easy to implement, and their further analysis can draw on the full potential of current computer algebra systems.

In general renewal models, $m(u)$ can alternatively be expressed as the solution of a Volterra integral equation of the second kind and hence as a Neumann series, see Gerber and Shiu (1998). Under the further assumption that the interclaim times have rational Laplace transform, the integral equation can be transformed into an integro-differential equation (IDE) with suitable boundary conditions. For the solution of the IDE, due to its convolution structure, Laplace transforms are often the key tool to derive explicit solutions; see e.g. Cheng and Tang (2003), Albrecher and Boxma (2005) and Li and Garrido (2005b). Landriault and Willmot (2008) obtain explicit expressions for the Laplace transform that can be inverted back by partial fractions, for arbitrary interclaim times and Coxian claim sizes. However, explicitly inverting the Laplace transform is in general difficult. Li and Garrido (2004) solved the IDE for Erlang(n) [E(n)] (sum of n independent exponential random variables) interclaim times by repeatedly integrating the integro-differential equation satisfied by the Gerber–Shiu function.

In the present paper, we want to advocate an alternative approach to derive explicit expressions for the Gerber–Shiu function in renewal models. For interclaim time distributions with rational Laplace transform—or equivalently if the interclaim density satisfies a linear ordinary differential equations (LODE) with constant coefficients—we first use the systematic approach of Constantinescu (2006) to transform the integral equation for $m(u)$ into an integro-differential equation. If the claim size distribution also has a rational Laplace transform, the IDE can be further reduced to a linear boundary value problem with appropriate boundary conditions (Section 2). Evaluating the IDE and its derivatives at 0 and imposing regularity conditions at ∞, we supplement the differential equation with sufficiently many boundary conditions so that the Gerber–Shiu function is uniquely determined. This program considerably extends the approach of Chen et al. (2007), who derived a LODE for $m(u)$ in a Poisson jump-diffusion process with phase-type jumps and solved it explicitly for penalty functions that depend only on the deficit at ruin.

Having arrived at a linear boundary problem, we employ the symbolic method developed in Rosenkranz (2005) and Rosenkranz and Regensburger (2008) for computing the integral operator (Green's operator) that maps the penalty function to the corresponding Gerber–Shiu function; see Section 3 for a brief description of this approach. Based on an algebraic operator framework, this method uses noncommutative Gröbner bases for transforming integro-differential and boundary operators to normal forms.

Whereas the classical version of this method works only for boundary value problems on compact intervals, we extend the approach to problems on the positive half-line in Section 4. There we consider operators on functions vanishing at infinity, which is the appropriate setup for our purposes.

In Section 5 we present the solution of the boundary value problem in terms of the Green's operator. The method relies on the factorization of the differential operator using the roots of the Lundberg fundamental equation. This factorization is then lifted to the level of boundary value problems: One can iteratively solve a sequence of first-order boundary value problems with appropriate boundary conditions. It turns out that there is a crucial difference between the roots with positive and negative real part and that there are natural links to the so-called Dickson–Hipp operator. Altogether, this approach allows one to compute the Gerber–Shiu function up to quadratures.

In previous papers e.g. Li and Garrido (2004) and Chen et al. (2007), the boundary conditions of the IDE are computed recursively in terms of derivatives of $m(u)$ at zero. In Section 6, we use an integrating factor method with different integration bounds and exploit the Vandermonde-type structure of the resulting matrix for directly deriving an explicit expression for each of these boundary values. This in turn makes it possible to arrive at a fully explicit formula for $m(u)$ in terms of the penalty function. An illustration of our method for E(n) interclaim times with E(m) claim sizes is given in Section 7. The method also covers more general models like the case of renewal risk models perturbed by a Brownian motion treated in Section 8. We conclude in Section 9 by discussing possible extensions of this approach.

2. Reduction to a boundary value problem

Consider T_1 to be the epoch of the first claim. Since ruin cannot occur in the interval $(0, T_1)$, by the standard renewal argument of Feller (1971, p. 183–184) one has

$$\begin{aligned} m(u) &= \mathbb{E}\left(\mathrm{e}^{-\delta T_1} m(u + cT_1 - X_1)\right) \\ &= \int_0^\infty \mathrm{e}^{-\delta t} f_\tau(t) \left(\int_0^{u+ct} m(u + ct - y) \right. \\ &\quad \left. + \int_{u+ct}^\infty w(u + ct, y - u - ct) \right) f_X(y) \mathrm{d}y \mathrm{d}t, \end{aligned} \tag{1}$$

for any claim size density f_X and interclaim time density f_τ. Due to the net profit condition, the model satisfies the regularity condition

$$\lim_{u\to\infty} m(u) = 0. \tag{2}$$

Define the polynomial

$$p_\tau(x) = x^n + a_{n-1}x^{n-1} + \cdots + a_0, \tag{3}$$

where a_j are real numbers for $j = 0, 1, \ldots, n$, and $a_0 \neq 0$. Assume that f_τ satisfies a linear ordinary differential equation with constant coefficients, compactly written in operator notation as

$$p_\tau\left(\frac{\mathrm{d}}{\mathrm{d}t}\right) f_\tau(t) = 0, \tag{4}$$

where $\frac{\mathrm{d}}{\mathrm{d}t}$ is the differentiation operator. For convenience, we consider those LODE representations of f_τ with *almost homogeneous* initial conditions

$$\begin{aligned} &f_\tau^{(k)}(0) = 0 \quad (k = 0, \ldots, n-2), \\ &f_\tau^{(n-1)}(0) = a_0. \end{aligned} \tag{5}$$

The Laplace transform of such a distribution is a rational function that has only a constant as the numerator.

Remark 1. One can express any density which is a convolution of n exponential densities with parameters λ_i in the above way, namely the polynomial (3) is

$$p_\tau(x) = \prod_{i=1}^{n}(x+\lambda_i), \tag{6}$$

with almost homogenous initial conditions (5). In the special case of exponentials with the same parameter λ, this is an Erlang(n) density $f_\tau(t) = \frac{1}{(n-1)!}\lambda^n t^{n-1}\mathrm{e}^{-\lambda t}$, satisfying Eq. (4) with almost homogenous initial conditions (5) and polynomial

$$p_\tau(x) = (x+\lambda)^n. \tag{7}$$

Under assumption (4) one can now use the technique of integration by parts as in Theorem 3 of Constantinescu (2006, Sec. 3.2) to obtain from (1) the integro-differential equation

$$p_\tau^*\left(c\frac{\mathrm{d}}{\mathrm{d}u}-\delta\right)m(u) = a_0\int_0^u m(u-y)\mathrm{d}F_X(y) + a_0\,\omega(u), \tag{8}$$

where the derivatives of m are assumed to exist and to be bounded. Here $\omega(u) = \int_u^\infty w(u, y-u)\mathrm{d}F_X(y)$ and

$$p_\tau^*(x) = (-1)^n x^n + (-1)^{n-1}a_{n-1}x^{n-1} + \cdots + a_0,$$

where $p_\tau^*(\frac{\mathrm{d}}{\mathrm{d}t})$ denotes the adjoint operator of the operator $p_\tau(\frac{\mathrm{d}}{\mathrm{d}t})$ defined through $\left\langle p_\tau(\frac{\mathrm{d}}{\mathrm{d}t})f, g\right\rangle = \left\langle f, p_\tau^*(\frac{\mathrm{d}}{\mathrm{d}t})g\right\rangle$ with $\langle f, g\rangle = \int_0^\infty f(x)\,g(x)\mathrm{d}x$ together with (5). In addition to the model regularity condition (2), we will derive in Section 6 the initial values M_i ($i = 0, \ldots, n-1$) of the IDE (8) through a variation of the classical *integrating factor* method of Gerber and Shiu (1998), obtaining

$$m(0) = M_0, m'(0) = M_1, \ldots, m^{(n-1)}(0) = M_{n-1}. \tag{9}$$

Together with (2), these boundary conditions make the boundary value problem regular.

Remark 2. Note that the same analysis also works for the case in which the boundary conditions are not of homogeneous type (as for instance would be the case for a mixture of Erlangs). In that case the Laplace transform of f_τ has a *polynomial* numerator of lower degree than of the polynomial in the denominator. As a consequence, one obtains further integral terms on the right-hand side of (8), leading to a slightly more cumbersome procedure.

Define the polynomial

$$p_X(x) = x^n + b_{n-1}x^{n-1} + \cdots + b_0. \tag{10}$$

If moreover the claim size density f_X satisfies a LODE with constant coefficients

$$p_X\left(\frac{\mathrm{d}}{\mathrm{d}y}\right)f_X(y) = 0,$$

and (for simplicity) almost homogeneous boundary conditions

$$f_X^{(k)}(0) = 0 \quad (k = 0, \ldots, n-2),$$
$$f_X^{(n-1)}(0) = b_0,$$

then the Gerber–Shiu function satisfies a well-posed boundary value problem, namely the LODE

$$p_X\left(\frac{\mathrm{d}}{\mathrm{d}u}\right)p_\tau^*\left(c\frac{\mathrm{d}}{\mathrm{d}u}-\delta\right)m(u)$$
$$= a_0b_0m(u) + a_0p_X\left(\frac{\mathrm{d}}{\mathrm{d}u}\right)\omega(u) \tag{11}$$

together with boundary conditions (2) and (9). The characteristic equation

$$p_X(s)p_\tau^*(cs-\delta) - a_0b_0 = 0 \tag{12}$$

of (11) is the Lundberg fundamental equation of this model. Since both the claim sizes and the inter-arrival times have rational Laplace transforms, we know by the results in Li and Garrido (2005a) and Landriault and Willmot (2008) that this equation has exactly n roots with positive and m roots with negative real part as long as $\delta > 0$. Note that we exclude the limiting case $\delta = 0$, which is equivalent to having 0 as a solution of the Lundberg equation; see Section 5 for a brief discussion of this case.

3. An algebraic operator approach for boundary value problems

In order to solve the boundary value problem for (11) we will employ the symbolic computation approach developed in Rosenkranz and Regensburger (2008) and Rosenkranz (2005). As this approach is targeted at boundary value problems for LODE in general differential algebras, we have to extract and adapt the parts needed for our present purposes.

As we can restrict ourselves to LODE with constant coefficients, we first consider two-point boundary value problems on a compact interval $[a, b]$: Given a *forcing function* $f(x) \in C[a, b]$, find a solution $g(x) \in C^n[a, b]$ of

$$\begin{aligned}&(D^n + c_{n-1}D^{n-1} + \cdots + c_1D + c_0)\,g = f,\\ &\beta_1(g) = \cdots = \beta_n(g) = 0,\end{aligned} \tag{13}$$

where $D = \frac{\mathrm{d}}{\mathrm{d}x}$, c_i are real numbers and the boundary conditions β_i are linear combinations of $g(a), \ldots, g^{(n-1)}(a)$ and $g(b), \ldots, g^{(n-1)}(b)$.

Note that the boundary conditions in (13) are homogeneous. As one easily sees, the solution for the general case of inhomogeneous boundary conditions is given by the solution of (13) plus the particular solution of the simple boundary value problem with inhomogeneous boundary conditions but $f = 0$.

The boundary value problem (13) is called regular if for every f there exists a unique g or equivalently if the associated homogeneous problem only has the trivial solution. This can be checked by testing whether the matrix formed by evaluating the boundary conditions on a fundamental system is regular; for details see Kamke (1967, p. 184). In this case, there is a well-defined operator $G: C[a, b] \to C^n[a, b]$ mapping $f \mapsto g$, known as the Green's operator of (13). While G is usually represented by its associated Green's function (Stakgold, 2000), the operator formulation is more practical in the present setting.

An essential feature of the symbolic operator calculus is that it allows one to compose two boundary value problems (in particular those of the form (13)) such that the composite Green's operator is given by the composition of the constituent Green's operators. For solving boundary value problems, the other direction is more important: Any factorization of the underlying differential operator can be lifted to a factorization of boundary value problems. Since we are dealing with differential operators with constant coefficients, we can actually achieve a factorization into first-order boundary value problems. For more details on composing and factoring boundary value problems for LODE, we refer again to Rosenkranz and Regensburger (2008). The theory is developed in an abstract algebraic setting, including in principle also boundary value problems for linear partial differential equations, in Regensburger and Rosenkranz (2009).

In the present setting, we can describe the first-order Green's operators as follows. Writing

$$A = \int_a^x, \qquad B = \int_x^b, \quad \text{and} \quad F = \int_a^b = A + B,$$

and

$$A_\sigma = \mathrm{e}^{\sigma x} A \mathrm{e}^{-\sigma x}, \qquad B_\rho = \mathrm{e}^{\rho x} B \mathrm{e}^{-\rho x}, \quad \text{and} \quad F_{\sigma\rho} = \mathrm{e}^{\sigma x} F \mathrm{e}^{-\rho x}$$

for $\rho, \sigma \in \mathbb{C}$, the basic first-order boundary value problems, with respect to each of the end points of the interval, $(D-\sigma)g = f, g(a) = 0$ and $(D-\rho)g = -f, g(b) = 0$, have respectively A_σ and B_ρ as their Green's operators as one can see by the fundamental theorem of calculus. Written as operator identities, this means in particular that

$$\begin{cases} (D-\sigma)A_\sigma = 1, \\ (D-\rho)B_\rho = -1, \end{cases} \tag{14}$$

so A_σ and $-B_\rho$ are right inverses of respectively $D-\sigma$ and $D-\rho$ on $C[a,b]$. By Rosenkranz (2005, Table 1), we obtain furthermore for any $\tilde\rho, \tilde\sigma \in \mathbb{C}$

$$\begin{cases} (\sigma-\tilde\sigma)\, A_\sigma A_{\tilde\sigma} = A_\sigma - A_{\tilde\sigma} \\ (\tilde\rho-\rho)\, B_\rho B_{\tilde\rho} = B_\rho - B_{\tilde\rho} \\ (\rho-\sigma)\, A_\sigma B_\rho = A_\sigma + B_\rho - F_{\sigma\rho} \end{cases} \tag{15}$$

on $C[a,b]$; the first two are called resolvent identities (Yosida, 1995). For the extension to non-compact intervals in Section 4 we mention an alternative, purely algebraic, way to derive (15), namely as a consequence of conditions that will be simpler to establish in the more general case:

Lemma 1. *The identities (15) are algebraic consequences of*

$$\begin{cases} A_\sigma (D-\sigma) A_{\tilde\sigma} = A_{\tilde\sigma} \\ B_\rho (D-\rho) B_{\tilde\rho} = -B_{\tilde\rho} \\ A_\sigma (D-\sigma) B_\rho = B_\rho - F_{\sigma\rho} \end{cases} \tag{16}$$

and the identities (14).

Proof. By (14), we have $A_\sigma = A_\sigma (D-\tilde\sigma) A_{\tilde\sigma} = A_\sigma (\sigma-\tilde\sigma+D-\sigma) A_{\tilde\sigma}$, which equals $(\sigma-\tilde\sigma)\, A_\sigma A_{\tilde\sigma} + A_{\tilde\sigma}$ because of (16); analogously for the other two identities of (15). □

4. Operators on functions vanishing at infinity

In the next section, we need the case $a = 0$ and $b = \infty$. So we consider the Banach algebra $(\mathcal{C}_0, \|\cdot\|_\infty)$ of all continuous functions $f\colon [0,\infty) \to \mathbb{C}$ vanishing at infinity (Conway, 1990, p. 65). The subalgebra of $\mathcal{C}_0$ consisting of n-times continuously differentiable functions is denoted by $\mathcal{C}_0^n$. The following proposition makes precise in how far the situation on $C[a,b]$ carries over to $\mathcal{C}_0$; confer also Butzer and Berens (1967, Prop. 1.3.12) for the case of bounded uniformly continuous functions on $\mathbb{R}$.

Proposition 2. *For $\rho \in \mathbb{C}$ with $\mathrm{Re}(\rho) > 0$, we have continuous integral operators*

$$A_{-\rho}, B_\rho, \mathrm{e}^{-\rho x} A, B\mathrm{e}^{-\rho x} \colon \mathcal{C}_0 \to \mathcal{C}_0^1 \tag{17}$$

with norm bounded by $1/\mathrm{Re}(\rho)$, and the identities (14) and (15) are valid for all $\rho, \tilde\rho, \sigma, \tilde\sigma \in \mathbb{C}$ with $\mathrm{Re}(\rho), \mathrm{Re}(\tilde\rho) > 0$ and $\mathrm{Re}(\sigma), \mathrm{Re}(\tilde\sigma) < 0$.

Proof. Let $\eta = \mathrm{Re}(\rho)$. We first check that the operators (17) map $\mathcal{C}_0$ into $\mathcal{C}_0$. For $A_{-\rho}$ we use that

$$|A_{-\rho} f(x)| \le \mathrm{e}^{-\eta x} \int_0^y \mathrm{e}^{\eta\xi} |f(\xi)| \mathrm{d}\xi + \mathrm{e}^{-\eta x} \int_y^x \mathrm{e}^{\eta\xi} |f(\xi)| \mathrm{d}\xi$$

for all $f \in \mathcal{C}_0$ and $x \ge y \ge 0$. Fixing $\varepsilon > 0$, the first summand is smaller than $\varepsilon/2$ for $x \ge x_0(\varepsilon, y)$ because $\eta > 0$. Since $f \in \mathcal{C}_0$, we have $|f(\xi)| < \varepsilon\eta/2$ for all $\xi \ge y_0(\varepsilon)$, so the second summand is smaller than $\varepsilon/2$ for $x \ge y_0(\varepsilon)$ and $y = y_0(\varepsilon)$. Thus we obtain $|A_{-\rho} f(x)| < \varepsilon$ for all $x \ge \max\{y_0(\varepsilon), x_0(\varepsilon, y_0(\varepsilon))\}$. Using a similar argument as for the second summand, we obtain $B_\rho f \in \mathcal{C}_0$. One immediately checks that $\mathrm{e}^{-\rho x} A$ and $B\mathrm{e}^{-\rho x}$ map even bounded functions into $\mathcal{C}_0$.

Next we verify that the operators are continuous. The norm bound for $A_{-\rho}$ follows from $|A_{-\rho} f(x)| \le \mathrm{e}^{-\eta x} \|f\|_\infty \int_0^x \mathrm{e}^{\eta\xi} \mathrm{d}\xi$ and $\mathrm{e}^{-\eta x} \int_0^x \mathrm{e}^{\eta\xi} \mathrm{d}\xi \le 1/\eta$; similarly for $\mathrm{e}^{-\rho x} A$ and $B\mathrm{e}^{-\rho x}$. For B_ρ we use the representation

$$B_\rho f(x) = \int_0^\infty \mathrm{e}^{-\rho\xi} f(\xi + x) \mathrm{d}\xi \tag{18}$$

and the fact that $\int_0^\infty \mathrm{e}^{-\eta\xi} \mathrm{d}\xi = 1/\eta$.

Now we turn to differentiability and identities (14). For $A_{-\rho}$ this follows immediately from the fundamental theorem of calculus. Using representation (18), the difference quotient $(B_\rho f(x+h) - B_\rho f(x))/h$ is given by

$$\frac{\mathrm{e}^{\rho h} - 1}{h} \int_h^\infty \mathrm{e}^{-\rho\xi} f(\xi + x) \mathrm{d}\xi - \frac{1}{h} \int_0^h \mathrm{e}^{-\rho\xi} f(\xi + x) \mathrm{d}\xi,$$

which converges to $\rho\, B_\rho f(x) - f(x)$ as $h \to 0$. Finally, $\mathrm{e}^{-\rho x} A f$ is differentiable again by the fundamental theorem and $B\mathrm{e}^{-\rho x} f = \mathrm{e}^{-\rho x} B_\rho f$ is differentiable because $B_\rho f$ is by what we have just seen.

It remains to prove the identities (15) and (16); by Lemma 1 it suffices to show the latter. These are an easy consequence of the fact that

$$A_\sigma (D-\sigma) f(x) = f(x) - \mathrm{e}^{\sigma x} f(0) \quad \text{and} \quad B_\rho (D-\rho) f(x) = -f(x)$$

for all $f \in \mathcal{C}_0^1$. The identity for A_σ carries over from the bounded case and is even valid on $C^1[0,\infty)$, the one for B_ρ follows from the representation (18) and integration by parts. □

Remark 3. Note that B_ρ also appears in the literature as the *Dickson–Hipp operator* (Dickson and Hipp, 2001; Li and Garrido, 2004), and the second equation of (15) is also used in these papers. The crucial contribution of the present result is the third equation of (15), i.e. the interaction between the Dickson–Hipp operator B_ρ and its counterpart A_σ.

We write $\mathcal{E}_0 \subset \mathcal{C}_0$ for the subalgebra of exponential polynomials spanned by $x^j \mathrm{e}^{-\rho x}$ with $\mathrm{Re}(\rho) > 0$.

Proposition 3. *The subalgebra $\mathcal{E}_0$ is dense in $\mathcal{C}_0$, and the operators (17) map $\mathcal{E}_0$ into itself.*

Proof. Density follows from the Stone–Weierstrass Theorem for locally compact spaces (Conway, 1990, p. 147). For proving that the operators (17) map $\mathcal{E}_0$ into itself, one uses induction on j and integration by parts. □

Note that—by the same reasoning—the operators A_ρ and B_ρ also map $\mathcal{E}_0$ into itself if $\mathrm{Re}(\rho) = 0$ but they are no longer continuous.

This proposition provides an alternative approach to proving the identities (14) and (15): Since $\mathcal{E}_0$ is dense in $\mathcal{C}_0$ and the operators are continuous, it suffices to prove them for exponential polynomials—this can be done by an elementary computation and induction on j. Density arguments of this type could also be useful for generalizing to larger function spaces like L^p or spaces based on regular variation (Bingham et al., 1987).

5. Solving boundary value problems on the half-line

For computing the Gerber–Shiu function, the method described in Section 2 leads to a boundary value problem on the half-line. In fact, we can rewrite Eq. (11) as

$$Tm = f, \tag{19}$$

with

$$T = p_X\left(\frac{\mathrm{d}}{\mathrm{d}u}\right) p_\tau^*\left(c\frac{\mathrm{d}}{\mathrm{d}u} - \delta\right) - a_0 b_0$$

and $\quad f(u) = a_0\, p_X\left(\frac{\mathrm{d}}{\mathrm{d}u}\right)\omega(u),$

initial values $m^{(i)}(0) = M_i$, and regularity condition $m(\infty) = 0$. As noted earlier (beginning of Section 3), it suffices to consider the corresponding homogeneous boundary conditions and incorporate the boundary values in specific settings afterwards (Sections 7 and 8).

So let us now consider the general boundary value problem on the half-line with homogeneous boundary conditions,

$$\begin{aligned} &Tg = f,\\ &g(0) = \cdots = g^{(m-1)}(0) = 0 \quad \text{and} \quad g \in \mathcal{C}_0, \end{aligned} \tag{20}$$

where the forcing function f is required to vanish at infinity.

We assume that the characteristic equation of T has distinct roots, which we divide into $\rho_1, \ldots, \rho_n$ with positive and $\sigma_1, \ldots, \sigma_m$ with negative real part (for the case of roots with zero real part see the discussion at the end of the section). Thus we have the differential operator $T = T_\rho T_\sigma$ with

$$T_\rho = (D-\rho_1)\cdots(D-\rho_n) \quad \text{and} \quad T_\sigma = (D-\sigma_1)\cdots(D-\sigma_m).$$

Note that in order to have a regular boundary value problem, it is sufficient to prescribe m initial conditions even though the order of T is $m+n$. This is due to the regularity condition $g \in \mathcal{C}_0$: The general solution g of the associated homogeneous differential equation $Tg = 0$ is a linear combination of $\mathrm{e}^{\rho_j x}$ and $\mathrm{e}^{\sigma_i x}$, where all terms with positive roots must vanish and the remaining m coefficients are determined by the m conditions at zero.

The crucial point is that it is possible to factor this boundary value problem along $T = T_\rho T_\sigma$ into the regular boundary value problems

$$\begin{aligned} &T_\sigma g = h,\\ &g(0) = \cdots = g^{(m-1)}(0) = 0 \end{aligned} \quad \text{and} \quad \begin{aligned} &T_\rho h = f,\\ &h \in \mathcal{C}_0 \end{aligned} \tag{21}$$

with forcing function $f \in \mathcal{C}_0$.

Lemma 4. *The boundary value problems (21) have*

$$G_\sigma = A_{\sigma_1}\cdots A_{\sigma_m} = \sum_{i=1}^m a_i A_{\sigma_i} \quad \textit{and}$$

$$G_\rho = (-1)^n B_{\rho_1}\cdots B_{\rho_n} = \sum_{j=1}^n b_j B_{\rho_j}$$

with

$$a_i = \prod_{k=1,k\neq i}^m (\sigma_i - \sigma_k)^{-1} \quad \textit{and}$$

$$b_j = -\prod_{k=1,k\neq j}^n (\rho_j - \rho_k)^{-1}$$

as their Green's operators, so $g = G_\sigma h$ and $h = G_\rho f$, where $\prod_{k=1,k\neq 1}^1 = 1$.

Proof. Let us first prove the identity for G_σ by induction (the case for G_ρ is analogous). The base case $m = 1$ is trivial, so assume the identity for $m-1$. Then (15) yields

$$A_{\sigma_1}\cdots A_{\sigma_{m-1}} A_{\sigma_m} = \sum_{i=1}^{m-1} a_i A_{\sigma_i} - \Big(\sum_{i=1}^{m-1}\prod_{k=1,k\neq i}^m (\sigma_i - \sigma_k)^{-1}\Big) A_{\sigma_m}$$

and we are done since the parenthesis is equal to $-a_m$ by the well-known partial fraction formula.

By Proposition 2, the Green's operators G_ρ and G_σ map $\mathcal{C}_0$ to $\mathcal{C}_0^m$ and $\mathcal{C}_0^n$, respectively, and (14) yields $T_\sigma G_\sigma = 1$ and $T_\rho G_\rho = 1$. It remains to check that $G_\sigma f$ satisfies the initial conditions. For that we prove for all $i < m$ the identity

$$D^i G_\sigma = \sum_{l=0}^i h_{i-l}(\sigma_1, \ldots, \sigma_{l+1})\, A_{\sigma_{l+1}}\cdots A_{\sigma_m}, \tag{22}$$

where h_{i-l} denotes the complete homogeneous symmetric polynomial of degree $i-l$ in the indicated variables (Stanley, 1999, p. 294); the claim then follows because $A_{\sigma_1} f(0), \ldots, A_{\sigma_m} f(0) = 0$. The base case $i = 0$ is trivial, so assume (22) for $i-1$. Using $DA_{\sigma_{l+1}} = 1 + \sigma_{l+1} A_{\sigma_{l+1}}$ from (14), this gives

$$\begin{aligned} D^i G_\sigma &= \sum_{l=0}^{i-1} h_{i-l-1}(\sigma_1, \ldots, \sigma_{l+1})\, DA_{\sigma_{l+1}}\cdots A_{\sigma_m}\\ &= \sum_{l=1}^{i-1}\Big(h_{i-l}(\sigma_1, \ldots, \sigma_l) + \sigma_{l+1} h_{i-l-1}(\sigma_1, \ldots, \sigma_{l+1})\Big)\\ &\quad \times A_{\sigma_{l+1}}\cdots A_{\sigma_m} + \sigma_1^i A_{\sigma_1}\cdots A_{\sigma_m} + A_{\sigma_{i+1}}\cdots A_{\sigma_m} \end{aligned}$$

after a little rearrangement. But the parenthesized factor in the sum simplifies to $h_{i-l}(\sigma_1, \ldots, \sigma_{l+1})$, while the outlying summands also have the right factors $h_{i-0}(\sigma_1) = \sigma_1^i$ and $h_{i-i}(\sigma_1, \ldots, \sigma_{j+1}) = 1$, respectively. □

Theorem 5. *The boundary value problem (20) has the Green's operator*

$$\begin{aligned} G_\sigma G_\rho &= \sum_{i=1}^m\sum_{j=1}^n c_{ij}(A_{\sigma_i} + B_{\rho_j} - F_{\sigma_i\rho_j})\\ &= \sum_{i=1}^m\sum_{j=1}^n c_{ij}\Big(\mathrm{e}^{\sigma_i x} A(\mathrm{e}^{-\sigma_i x} - \mathrm{e}^{-\rho_j x}) + (\mathrm{e}^{\rho_j x} - \mathrm{e}^{\sigma_i x}) B\mathrm{e}^{-\rho_j x}\Big) \end{aligned}$$

where $c_{ij} = a_i b_j\,(\rho_j - \sigma_i)^{-1}$, i.e. $g = G_\sigma G_\rho f$.

Proof. Let $f \in \mathcal{C}_0$. From Proposition 2 we know that $G = G_\sigma G_\rho$ maps f into $\mathcal{C}_0^{m+n}$. By the previous lemma, Gf satisfies the differential equation and the initial conditions. For proving that G has the indicated sum representations, we use again Lemma 4, the identities (15) and the definition of $F_{\sigma_i\rho_j}$. □

If some of the ρ_j have zero real part, the above Green's operator G no longer maps $\mathcal{C}_0$ into itself, so the boundary value problems (20) cannot be expected to have a solution for all $f \in \mathcal{C}_0$. But if $Gf \in \mathcal{C}_0$, it is the unique solution of (20); by the observations after Proposition 3, this is particularly true for $f \in \mathcal{E}_0$.

6. Initial values for E(n) risk processes

The next step for solving the boundary value problem for (19) is to determine the initial values M_i of (9). We consider the case of E(n) distributed interclaim times (under assumption that m has bounded derivatives). Using (7) in the integro-differential equation (8), we obtain

$$\left(-c\frac{\mathrm{d}}{\mathrm{d}u} + (\lambda+\delta)\right)^n m(u) = \lambda^n \int_0^u m(u-y)\mathrm{d}F_X(y) + \lambda^n\omega(u) \tag{23}$$

with the corresponding Lundberg fundamental equation

$$(-cz + (\lambda+\delta))^n - \lambda^n \hat{f}_X(z) = 0, \tag{24}$$

where $\hat{f}_X(z) = \mathbb{E}(e^{-zX})$ is the Laplace transform of $f_X(u)$. Eq. (24) has exactly n solutions ρ_i ($i = 1, \ldots, n$) with positive real part, according to Li and Garrido (2004).

We will use a similar *integrating factors* technique as the one proposed in Gerber and Shiu (1998) and arrive at a system of linear equations in the initial values that we can solve explicitly. A different choice of the integration bounds will simplify some steps compared to a related approach of Li and Garrido (2004). The change of variables and order of integration used in Gerber and Shiu (1998) is then not necessary here. Let us multiply (23) by $e^{-\rho_i u}$ for each $i = 1, \ldots, n$, and then integrate from $u = \infty$ to $u = x$ to arrive at

$$\sum_{j=0}^{n} \binom{n}{j} (-c)^j (\lambda+\delta)^{(n-j)} \int_\infty^x e^{-\rho_i u} m^{(j)}(u)\mathrm{d}u$$
$$= \lambda^n \int_\infty^x e^{-\rho_i u} \int_0^u m(u-y)\mathrm{d}F_X(y)\mathrm{d}u + \lambda^n \int_\infty^x e^{-\rho_i u}\omega(u)\mathrm{d}u.$$

Now we use integration by parts together with

$$\lim_{u\to\infty} e^{-\rho_i u} m^{(j)}(u) = 0 \quad (j = 0, \ldots, n, i = 1, \ldots, n)$$

to obtain

$$\int_\infty^x e^{-\rho_i u} m^{(j)}(u)\mathrm{d}u = \sum_{k=0}^{j-1} e^{-\rho_i x}\rho_i^k m^{(j-k-1)}(x) + \rho_i^j I_i(x),$$

where $I_i(x) = \int_\infty^x e^{-\rho_i u} m(u)\mathrm{d}u$.

Then evaluating each equation at $x = 0$, we note that the left-hand side and the right-hand side terms pertaining to $I_i(0)$ cancel due to (24) evaluated at $z = \rho_i$. Also we see that in the right-hand side the second integral is actually $-\hat{\omega}(\rho_i)$, the Laplace transform of ω evaluated at ρ_i. We obtain a system of n equations in n unknown variables $m^{(k)}(0)$

$$\sum_{j=1}^{n} \binom{n}{j} (-c)^j (\lambda+\delta)^{n-j} \sum_{k=0}^{j-1} \rho_i^k m^{(j-k-1)}(0) = -\lambda^n \hat{\omega}(\rho_i)$$

for $k = 0, \ldots, n-1$. Collecting and rearranging the terms, we get

$$\sum_{k=0}^{n-1} m^{(k)}(0) \underbrace{\sum_{j=0}^{n-k-1} \binom{n}{j} \left(-\frac{\lambda+\delta}{c}\right)^j \rho_i^{(n-k-1)-j}}_{p_{n-k-1}(\rho_i)}$$
$$= -\left(-\frac{\lambda}{c}\right)^n \hat{\omega}(\rho_i), \tag{25}$$

for $i = 1, \ldots, n$. Note that the polynomials

$$p_k(x) = \sum_{j=0}^{k} \binom{n}{k-j} \left(-\frac{\lambda+\delta}{c}\right)^{k-j} x^j \tag{26}$$

appearing in the coefficients of $m^{(n-k-1)}(0)$ are monic of degree k.

We express the system in matrix form $Ax = b$ as

$$\begin{pmatrix} p_0(\rho_1) & \cdots & p_{n-1}(\rho_1) \\ \vdots & \ddots & \vdots \\ p_0(\rho_n) & \cdots & p_{n-1}(\rho_n) \end{pmatrix} \begin{pmatrix} m^{(n-1)}(0) \\ \vdots \\ m^{(0)}(0) \end{pmatrix}$$
$$= -\left(-\frac{\lambda}{c}\right)^n \begin{pmatrix} \hat{\omega}(\rho_1) \\ \vdots \\ \hat{\omega}(\rho_n) \end{pmatrix}.$$

According to Cramer's rule, the solution of this system of equations is of the form

$$m^{(k)}(0) = \frac{\det(B_{n-1-k})}{\det(A)} \quad (k = 0, \ldots, n-1), \tag{27}$$

where B_k is the $n \times n$ matrix obtained from A by replacing the $(k+1)$-th column of A by the right-hand side b.

The following result generalizes the formula for $m(0)$ given in Gerber and Shiu (2005, Eq. 8.1).

Proposition 6. *The k-th derivative of the expected discounted penalty function evaluated at zero has the form*

$$m^{(k)}(0) = (-1)^k \left(\frac{\lambda}{c}\right)^n \sum_{i=1}^{n} \frac{\hat{\omega}(\rho_i)\, S(\rho_i', k)}{\prod_{l=1,\ldots,n;\ l\neq i} (\rho_l - \rho_i)}, \tag{28}$$

for $k = 0, \ldots, n-1$, where $\rho_i' = (\rho_1, \ldots, \rho_{i-1}, \rho_{i+1}, \ldots, \rho_n)$ and

$$S(\rho_i', k) = \sum_{j=0}^{k} \left(-\frac{\lambda+\delta}{c}\right)^j \binom{n-1+j}{j} e_{k-j}(\rho_i'),$$

with e_k the elementary symmetric polynomials of degree k.

Proof. According to Krattenthaler (1999), the determinant of the matrix A is the same as the Vandermonde determinant $V_n = V_n(\rho_1, \ldots, \rho_n)$ so

$$\det(A) = \prod_{1\le i<j\le n} (\rho_j - \rho_i).$$

We will show that the determinant of B_k is the product of a Vandermonde determinant and a linear combination of symmetric polynomials in the ρ_i and $\hat{\omega}(\rho_i)$. Expanding along the $(n-k)$-th column, one gets

$$\det(B_{n-1-k}) = \sum_{i=1}^{n} (-1)^{i+n-k} b_i \det(A_{i,n-k}),$$

where $A_{i,k}$ is the $(n-1)\times(n-1)$ matrix obtained from A by removing the i-th row and the k-th column. By applying Corollary A.2 of the Appendix to the matrix $A_{i,k}$ and observing that

$$q(x) = \left(1 - \frac{\lambda+\delta}{c}x\right)^n - \left(1 + \left(-\frac{\lambda+\delta}{c}x\right)^n\right),$$

we obtain

$$\det(A_{i,n-k}) = V_{n-1}(\rho_i') \sum_{j=0}^{k} d_j e_{k-j}(\rho_i'),$$

where

$$d_j = [x^j] \frac{(-1)^j + \left((1-\frac{\lambda+\delta}{c}x)^n - \left(1 + \left(-\frac{\lambda+\delta}{c}x\right)^n\right)\right)^{j+1}}{(1-\frac{\lambda+\delta}{c}x)^n - \left(-\frac{\lambda+\delta}{c}x\right)^n}$$

and $[x^j]f(x) = f^{(j)}(0)/j!$ denotes the coefficient of x^j of a power series $f(x)$. We will show below that

$$d_j = \left(-\frac{\lambda+\delta}{c}\right)^j \binom{n-1+j}{j}. \tag{29}$$

Inserting the resulting formula for the determinant $A_{i,k}$ into the expansion of $\det(B_k)$ in Cramer's rule, we get

$$m^{(k)}(0) = \left(-\frac{\lambda}{c}\right)^n \sum_{i=1}^{n} (-1)^{i+n-k+1} \hat{\omega}(\rho_i) \frac{V_{n-1}(\rho_i')}{V_n} \sum_{j=0}^{k} d_j e_{k-j}(\rho_i'),$$

which after cancelation of the Vandermode terms leads to the result stated.

It remains to show Eq. (29). From Eq. (34) we get that $(-1)^j d_j = [x^j]\sum_{m=0}^{j}(-q(x))^m$. Since $j < n$ we can safely add terms of order at least n to $q(x)$. We do this and replace $q(x)$ with $(1-\frac{\lambda+\delta}{c}x)^n - 1$.

Inserting the modified $q(x)$ and expanding the expression, we obtain

$$\begin{aligned}\sum_{m=0}^{j}(-q(x))^m &= \sum_{m=0}^{j}(-1)^m\left(\left(1-\frac{\lambda+\delta}{c}x\right)^n-1\right)^m\\ &= \sum_{m=0}^{j}\sum_{l=0}^{m}\binom{m}{l}(-1)^l\sum_{h=0}^{nl}\binom{nl}{h}\left(-\frac{\lambda+\delta}{c}x\right)^h\\ &= \sum_{h=0}^{nm}\left(-\frac{\lambda+\delta}{c}x\right)^h\sum_{m=0}^{j}\sum_{l=0}^{m}(-1)^l\binom{m}{l}\binom{nl}{h},\end{aligned}$$

so that $d_j = \left(\frac{\lambda+\delta}{c}\right)^j\sum_{m=0}^{j}\sum_{l=0}^{m}(-1)^l\binom{m}{l}\binom{nl}{j}$. Rearranging and using the simple binomial identities of Graham et al. (1989, 5.10 and 5.14), we can simplify the double sum to

$$\begin{aligned}&\sum_{l=0}^{j}(-1)^l\binom{nl}{j}\sum_{m=0}^{j}\binom{m}{l} = \sum_{l=0}^{j}(-1)^l\binom{j+1}{l+1}\binom{nl}{j}\\ &= \sum_{l=0}^{j+1}(-1)^{l+1}\binom{j+1}{l}\binom{n(l-1)}{j}+\binom{-n}{j}\\ &= (-1)^j\binom{n-1+j}{j}-\sum_{l=0}^{j+1}(-1)^l\binom{j+1}{l}\binom{n(l-1)}{j}.\end{aligned}$$

Finally, the last sum vanishes due to Graham et al. (1989, 5.42) since it is the $(j+1)$-th difference of $\binom{n(l-1)}{j}$ as a polynomial in l, which is only of degree j. □

Since the Gerber–Shiu function is the unique solution of (19), it has the form

$$m(u) = G_\sigma G_\rho f(u) + m^p(u),$$

where $G_\sigma G_\rho$ is given in Theorem 5 and $m^p(u)$ is the particular solution obtained as a linear combination of the $e^{\sigma_i u}$, with factors determined by the initial values from Proposition 6.

7. Explicit solution for E(n) risk processes with E(m) claims

Let us now specialize the differential equation (11) for the Gerber–Shiu function to the case of Erlang(n, λ) interclaim times and Erlang(m, μ) claim sizes, with discount rate $\delta > 0$. From the previous section we get n boundary conditions. As described in Section 5 one in fact needs m boundary conditions, so we assume $m \le n$ (otherwise, one can derive the remaining conditions by evaluating higher derivatives of the integro-differential equation (23)). We obtain a boundary value problem for the differential equation $Tm = f$ with $D = \frac{\mathrm{d}}{\mathrm{d}u}$, where

$$\begin{aligned}T &= (D+\mu)^m(-c\,D+\lambda+\delta)^n-\lambda^n\mu^m,\\ f(u) &= \frac{\lambda^n\mu^m}{(m-1)!}(D+\mu)^m\int_u^\infty w(u,y-u)\,y^{m-1}e^{-\mu y}\mathrm{d}y \qquad (30)\end{aligned}$$

and boundary conditions (2) and (9). To apply the results from Section 5, we can choose any sufficiently smooth penalty function $w(x, y)$ such that $\lim_{u\to\infty} f(u) = 0$. By Proposition 3 this includes all bivariate exponential polynomials whose terms $x^iy^je^{\alpha x}e^{\beta y}$ satisfy $\alpha < \beta < \mu$.

Since the characteristic equation for T is the Lundberg fundamental equation, we know from the general results mentioned in Section 2 that it has n roots $\rho_1, \ldots, \rho_n$ with positive real part and m roots $\sigma_1, \ldots, \sigma_m$ with negative real part. So we have the factorization

$$T = T_\rho T_\sigma = (D-\rho_1)\cdots(D-\rho_n)(D-\sigma_1)\cdots(D-\sigma_m),$$

and Theorem 5 gives us the Green's operator for the corresponding homogeneous boundary value problem.

Writing $\hat f$ for the Laplace transform of f and using the definition of the corresponding operators, we obtain from Theorem 5 the explicit form of the Gerber–Shiu function

$$\begin{aligned}m(u) = \sum_{i=1}^{m}\sum_{j=1}^{n}c_{ij}\Bigg(&\left(\int_0^u e^{\sigma_i(u-\xi)}+\int_u^\infty e^{\rho_j(u-\xi)}\right)\\ &\times f(\xi)\mathrm{d}\xi-\hat f(\rho_j)e^{\sigma_i u}\Bigg)+m^p(u) \qquad (31)\end{aligned}$$

with

$$c_{ij} = -\prod_{k=1,k\ne i}^{m}(\sigma_i-\sigma_k)^{-1}\prod_{k=1,k\ne j}^{n}(\rho_j-\rho_k)^{-1}(\rho_j-\sigma_i)^{-1}.$$

With the initial values from formula (28) the computation of the particular solution m^p satisfying the inhomogeneous boundary conditions reduces to solving a system of linear equations, obtained from imposing the condition that the particular solution satisfies these given initial conditions, i.e. $(m^p)^{(i)}(0) = M_i$. As remarked in Section 5, formula (31) remains valid for suitable f also in the limiting case $\delta = 0$, which is equivalent to having 0 among the $\rho_1, \ldots, \rho_n$.

So the problem of computing the Gerber–Shiu function for a given penalty function is reduced to quadratures: Since symbolic algorithms for evaluating one-dimensional integrals are very powerful (Bronstein, 2005) and easily accessible in current computer algebra systems, one will often obtain an explicit expression for the Gerber–Shiu function. Otherwise one can resort to standard numerical methods for obtaining approximations.

In the particular case $n = 2, m = 1$ one has

$$\begin{aligned}T &= (D+\mu)(-c\,D+\lambda+\delta)^2-\lambda^2\mu,\\ f(u) &= \lambda^2\mu(D+\mu)\int_u^\infty w(u,y-u)\,e^{-\mu y}\mathrm{d}y.\end{aligned}$$

After calculating the particular solution using the initial value from Proposition 6, we obtain the Gerber–Shiu function in the explicit form

$$\begin{aligned}m(u) = {}&\frac{e^{\sigma u}}{\rho_1-\rho_2}\left(\frac{\hat f(\rho_1)}{\rho_1-\sigma}-\frac{\hat f(\rho_2)}{\rho_2-\sigma}-\left(\frac{\lambda}{c}\right)^2\left(\hat\omega(\rho_1)-\hat\omega(\rho_2)\right)\right)\\ &-\frac{1}{\rho_1-\rho_2}\int_u^\infty\left(\frac{1}{\rho_1-\sigma}e^{\rho_1(u-\xi)}-\frac{1}{\rho_2-\sigma}e^{\rho_2(u-\xi)}\right)f(\xi)\,\mathrm{d}\xi\\ &+\frac{1}{\rho_1-\sigma}\frac{1}{\rho_2-\sigma}\int_0^u e^{\sigma(u-\xi)}f(\xi)\mathrm{d}\xi,\end{aligned}$$

where one should recall that ρ_1, ρ_2 are the positive roots and σ is the negative root of the fundamental Lundberg equation. For example, when $w(x, y) = x^jy^k$ with j and k positive integers, one obtains

$$\begin{aligned}\frac{\Delta\mu^k}{k!\lambda^2}m(u) = {}&-\frac{\rho_2-\sigma}{(\rho_1+\mu)^j}\\ &\times\left(j\,\Gamma\big(j,(\rho_1+\mu)u\big)\,e^{\rho_1 u}+\frac{j!}{c^2}\left(\frac{\rho_1-\sigma}{\rho_1+\mu}-c^2\right)e^{\sigma u}\right)\\ &+\frac{\rho_1-\sigma}{(\rho_2+\mu)^j}\left(j\,\Gamma\big(j,(\rho_2+\mu)u\big)\,e^{\rho_2 u}+\frac{j!}{c^2}\left(\frac{\rho_2-\sigma}{\rho_2+\mu}-c^2\right)e^{\sigma u}\right)\\ &-\frac{\rho_1-\rho_2}{(\sigma+\mu)^j}\left(j\,\Gamma\big(j,(\sigma+\mu)u\big)-j!\right)e^{\sigma u},\end{aligned}$$

where $\Delta = (\rho_1-\rho_2)(\rho_1-\sigma)(\rho_2-\sigma)$ is the square root of the discriminant associated to the fundamental Lundberg equation and $\Gamma(a,x) = \int_x^\infty t^{a-1}e^{-t}\mathrm{d}t$ is the incomplete Gamma function. This formula extends Eq. (3.8) of Cheng and Tang (2003) and similar examples with $n = 2$ from Li and Garrido (2004, 2005b) and Gerber and Shiu (2005).

8. Explicit solution for the classical perturbed risk model

For the case of an Erlang(n, λ) risk model perturbed by a Brownian motion, the Gerber–Shiu function satisfies an integro-differential equation as given in Constantinescu (2006)

$$\left(-\frac{\tilde{\sigma}^2}{2}\frac{\mathrm{d}^2}{\mathrm{d}u^2} - c\frac{\mathrm{d}}{\mathrm{d}u} + \lambda + \delta\right)^n m(u) = \lambda^n \int_0^u m(u-x) f_X(x)\mathrm{d}x + \lambda^n \omega(u), \tag{32}$$

where $\tilde{\sigma}$ is the diffusion coefficient. Since the differential operator of this equation has constant coefficients, the method introduced in this paper applies. As before, for claim distributions with rational Laplace transform, the equation reduces to a LODE. For instance, in the case of E(m, μ) claim sizes, this LODE has the same form $Tm = f$ with $D = \frac{\mathrm{d}}{\mathrm{d}u}$, with

$$T = (D+\mu)^m \left(-\frac{\tilde{\sigma}^2}{2}D^2 - cD + \lambda + \delta\right)^n - \lambda^n \mu^m,$$

and $f(u)$ as in (30) and the appropriate boundary conditions. The characteristic equation for T is again the fundamental Lundberg equation.

Also in this case we can derive explicit expressions for the Gerber–Shiu function. To exemplify, we consider the well-known case of a compound Poisson process perturbed by a Brownian motion with exponential claim sizes, E$(1, \lambda)$–E$(1, \mu)$ in the notation introduced here. Then the LODE is of order three, with

$$T = (D+\mu)\left(-\frac{\tilde{\sigma}^2}{2}D^2 - cD + \lambda + \delta\right) - \lambda\mu$$

and

$$f(u) = \lambda\mu\,(D+\mu)\int_u^\infty w(u, y-u)\,\mathrm{e}^{-\mu y}\mathrm{d}y.$$

The initial value at zero $m(0) = w(0,0)$ is in this case simply the penalty function evaluated at zero. Since according to Li and Garrido (2005a), in the case of a compound Poisson risk model perturbed by a Brownian motion, the Lundberg equation has only one positive solution that we will denote ρ, we can apply the integrating factor technique only once. It yields the linear equation

$$\frac{\tilde{\sigma}^2}{2}m'(0) + \left(\rho\frac{\tilde{\sigma}^2}{2} + c\right)m(0) = \lambda\hat{\omega}(\rho), \tag{33}$$

which we can solve for $m'(0)$. With these initial values, we can compute the particular solution and Eq. (31) leads to

$$\begin{aligned} m(u) = &-\frac{1}{(\rho-\sigma_1)(\rho-\sigma_2)}\int_u^\infty \mathrm{e}^{\rho(u-\xi)} f(\xi)\mathrm{d}\xi \\ &- \frac{\hat{f}(\rho)}{\sigma_2-\sigma_1}\left(\frac{\mathrm{e}^{\sigma_1 u}}{\rho-\sigma_1} - \frac{\mathrm{e}^{\sigma_2 u}}{\rho-\sigma_2}\right) \\ &+ \frac{1}{\sigma_2-\sigma_1}\int_0^u \left(\frac{\mathrm{e}^{\sigma_1(u-\xi)}}{\rho-\sigma_1} - \frac{\mathrm{e}^{\sigma_2(u-\xi)}}{\rho-\sigma_2}\right) f(\xi)\mathrm{d}\xi \\ &+ \frac{1}{\sigma_2-\sigma_1}\left([\sigma_2 m(0) - m'(0)]\mathrm{e}^{\sigma_1 u} + [-\sigma_1 m(0) + m'(0)]\mathrm{e}^{\sigma_2 u}\right) \end{aligned}$$

as an explicit expression for the Gerber–Shiu function. This formula generalizes Eq. (4.6) of Chen et al. (2007) for the case of exponential claim sizes and Example 1 of Li and Garrido (2005a) for exponential inter-arrival times.

9. Conclusion

We have shown that the link between symbolic computation and risk theory can be mutually fruitful and can be utilized to identify fully explicit expressions for the Gerber–Shiu function in general renewal models in terms of the employed penalty function. In the presented approach, Laplace transforms only enter in a very restricted form:

- Only the Laplace transform of the penalty (not of the Gerber–Shiu function) is computed. This has the advantage that one does not need artificial analyticity conditions on m.
- Moreover, the Laplace transform of the penalty is only evaluated at $\rho_1, \ldots, \rho_n$, the positive solutions of the Lundberg equation, for computing the boundary values.
- No inverse Laplace transform is involved. This is in contrast to many previous papers that give explicit formulae for the Laplace transform of the Gerber–Shiu function, which often cannot be inverted in closed form.

In principle, the symbolic method introduced in this paper can be extended to models that include investment as well as to models with interclaim time densities that satisfy ODEs with polynomial coefficients as long as the spectral structure of the Lundberg fundamental equation is still tractable. This will be pursued in future research. The factorization approach for boundary value problems generalizes in principle also to partial differential equations (Regensburger and Rosenkranz, 2009), which in the context of risk theory means that more general models including one more variable could be considered. Finally, the method may be applicable in boundary value problems that occur in other contexts in risk theory.

The formulas developed in this paper can easily be implemented in a computer algebra system, which in turn allows to quickly perform (quantitative and graphical) sensitivity analysis of the corresponding discounting penalty functions with respect to parameter and penalty changes.

Acknowledgements

We would like to thank Jose Garrido for his encouraging feedback at an early stage of this project. We would also like to thank Christian Krattenthaler for valuable hints on simplifying the determinant of Proposition A.1 and Peter Paule and Christoph Koutschan for help with proving the explicit form of d_j in the proof of Proposition 6. Furthermore we thank the anonymous referees for helpful comments to improve the presentation of the manuscript. Hansjörg Albrecher was partly supported by the Austrian Science Fund Project P18392, Gottlieb Pirsic was partly supported by the Austrian Science Fund Project P19004-N18.

Appendix. A generalized vandermonde determinant

For computing the initial values in Proposition 6 we are led to consider the $n \times n$ alternant matrix

$$A = \begin{pmatrix} p_0(x_1) & \cdots & p_{n-1}(x_1) \\ \vdots & \ddots & \vdots \\ p_0(x_n) & \cdots & p_{n-1}(x_n) \end{pmatrix}$$

with polynomials $p_i(x) = a_{i,i}x^i + \cdots + a_{i,0}$ with $a_{i,i} = 1$. In the special case $p_i(x) = x^i$ this is the usual Vandermonde matrix with the determinant V_n in the indeterminates $x_1, \ldots, x_n$, but $\det A = V_n$ holds in general (Krattenthaler, 1999, Prop. 1).

We want to compute the (k, l) minor of A, the determinant of the $(n-1)\times(n-1)$ matrix $A_{k,l}$ obtained by deleting the k-th row and the l-th column. It suffices to consider

$$A_{n,l} = \begin{pmatrix} p_0(x_1) & \cdots & p_{l-1}(x_1) & p_{l+1}(x_1) & \cdots & p_{n-1}(x_1) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ p_0(x_{n-1}) & \cdots & p_{l-1}(x_{n-1}) & p_{l+1}(x_{n-1}) & \cdots & p_{n-1}(x_{n-1}) \end{pmatrix}$$

since

$$A(x_1,\ldots,x_n)_{k,l} = A(x_1,\ldots,x_{k-1},x_{k+1},\ldots,x_n,x_k)_{n,l}.$$

For $p_i(x) = x^i$ it is known (Heineman, 1929) that $\det A_{n,l}/V_{n-1}$ yields the elementary symmetric polynomial e_{n-1-l} in $x_1,\ldots,x_{n-1}$.

Proposition A.1. *We have*

$$\frac{\det A_{n,l}}{V_{n-1}} = e_{n-1-l} + \sum_{j=l+1}^{n-1}\left(\sum_{J}(-1)^{j+l+m}a_{j_1,j_2}a_{j_2,j_3}\cdots a_{j_m,j_{m+1}}\right)e_{n-1-j},$$

where the inner sum ranges over $J = (j_1,\ldots,j_{m+1})$ such that $m \geq 1$ and $j = j_1 > \cdots > j_{m+1} = l$.

Proof. Writing x^j for the column vector $(x_1^j,\ldots,x_{n-1}^j)^{\mathrm{T}}$, the determinant of the matrix

$$A_{n,l} = \left(\sum_{r=0}^{0}a_{0,r}x^r \quad \cdots \quad \sum_{r=0}^{l-1}a_{l-1,r}x^r \quad \sum_{r=0}^{l+1}a_{l+1,r}x^r \quad \cdots \quad \sum_{r=0}^{n-1}a_{n-1,r}x^r\right)$$

is given by multilinearity as

$$\sum_{r_0=0}^{0}\cdots\sum_{r_{l-1}=0}^{l-1}\sum_{r_{l+1}=0}^{l+1}\cdots \times\sum_{r_{n-1}=0}^{n-1}c_r\det(x^{r_0},\ldots,x^{r_{l-1}},x^{r_{l+1}},\ldots,x^{r_{n-1}})$$

with $c_r = a_{0,r_0}\cdots a_{l-1,r_{l-1}}a_{l+1,r_{l+1}}\cdots a_{n-1,r_{n-1}}$. Observe that for the first l indices $r_0,\ldots,r_{l-1}$ there always exist some $i < j < l$ such that $r_i = r_j$ unless $r_i = i$ for all $i < l$. Since the determinant vanishes for the cases $r_i = r_j$ and the p_i are monic (i.e., $a_{i,i} = 1$), the determinant reduces to

$$\sum_{r_{l+1}=l}^{l+1}\cdots\sum_{r_{n-1}=l}^{n-1}a_{l+1,r_{l+1}}\cdots a_{n-1,r_{n-1}} \times\det(x^0,\ldots,x^{l-1},x^{r_{l+1}},\ldots,x^{r_{n-1}})$$

where $r_{l+1},\ldots,r_{n-1}$ can be restricted to mutually distinct indices.

We view the indices as the permutations $r\colon\{l,\ldots,n-1\}\to\{l,\ldots,n-1\}$ satisfying $r_s \leq s$ for $s > l$; note that r_l is determined as the index omitted in $r_{l+1},\ldots,r_{n-1}$. By the monotonicity condition on r, all cycles without l in the cycle representation of r are trivial: If we have a nontrivial cycle $(j_1 \ \ldots \ j_{m+1})$, with $j_{m+1} \neq l$ we are led to the contradiction $j_1 > r_{j_1} = j_2 > \cdots > j_{m+1} > r_{j_{m+1}} = j_1$. Consequently r either possesses only one nontrivial cycle $(j_1 \ \ldots \ j_{m+1})$ with $j_1 > \cdots > j_{m+1} = l$, unless r is the identity. Since the p_i are monic, the factor of the determinant Δ_r occurring in the above sum is given by $a_{j_1,j_2}\cdots a_{j_m,j_{m+1}}$ in the former and by 1 in the latter case.

For finding Δ_r, we use row expansion for computing

$$(-1)^{l+n-1}\Delta_r = \det\begin{pmatrix} x^0 & \cdots & x^{l-1} & x^l & x^{r_{l+1}} & \cdots & x^{r_{n-1}} \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}.$$

This determinant is the result of r acting on the columns of the determinant

$$\det\begin{pmatrix} x^0 & \cdots & x^{j-1} & x^j & x^{j+1} & \cdots & x^{n-1} \\ 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix} = (-1)^{j+n-1}e_{n-1-j}\,V_{n-1}$$

according to the above mentioned result on Vandermonde minors. Since r as a cycle of length $m+1$ has sign $(-1)^m$, this yields

$$\Delta_r = (-1)^{j+l+m}e_{n-1-j}\,V_{n-1},$$

which proves the formula. □

Note that the inner sum in Proposition A.1 can also be interpreted as ranging over all ordered subsets of $\{l,\ldots,j\}$ containing l and j. It can be simplified further in the following special case, which we use in Section 6. We give two representations, one in terms of compositions and the other using generating functions. Here we use the customary notation $[x^i]f(x)$ for the coefficient of x^i in a power series $f(x)$.

Corollary A.2. *If $p_i(x) = a_0x^i + \cdots + a_i$, $a_0 = 1$, the formula in Proposition A.1 simplifies to*

$$\frac{\det A_{n,l}}{V_{n-1}} = e_{n-1-l} + \sum_{j=l+1}^{n-1}(-1)^{j-l} \times\left(\sum_{m\geq 1}(-1)^m\sum_{d_1,\ldots,d_m}a_{d_1}\ldots a_{d_m}\right)e_{n-1-j},$$

where the inner sum ranges over $d_1,\ldots,d_m > 0$ such that $d_1+\cdots+d_m = j-l$. Using generating functions, we have also

$$\frac{\det A_{n,l}}{V_{n-1}} = \sum_{j=0}^{n-1-l}\left([x^j]\,\frac{(-1)^j+q(x)^{j+1}}{1+q(x)}\right)e_{n-1-l-j},$$

where $q(x) = a_1x+\cdots+a_{n-1}x^{n-1}$.

Proof. Applying the above remark to the case $a_{i,j} = a_{i-j}$, the inner sum in Proposition A.1 gives

$$\sum_{j>j_2>\cdots>j_m>l}(-1)^m a_{j-j_2}a_{j_2-j_3}\cdots a_{j_{m-1}-j_m}a_{j_m-l} = \sum_{\substack{d_1,\ldots,d_m>0,\\ \sum_i d_i=j-l}}(-1)^m a_{d_1}\ldots a_{d_m}$$

for $j > l$, since the differences $d_1 = j-j_2$, $d_2 = j_2-j_3,\ldots,d_m = j_m-l$ can take arbitrary nonnegative values, provided they sum up to $j-l$. Now the first formula follows by multiplying with $(-1)^{j+l} = (-1)^{j-l}$.

For the second formula observe that the sum over the compositions of $j-l$ that appears within the bracket of the first formula is equal to the coefficient of x^{j-l} in the product

$$\prod_{i=1}^{m}\left(a_1x+\cdots+a_{n-1}x^{n-1}\right) = q(x)^m,$$

for $m \leq j-l$; for $m > j-l$ the sum over the composition is empty. Note that this even covers the cases $m = 0$, for which the term is zero except for $j-l = 0$, when it becomes one. The stated formula then follows by

$$(-1)^{j-l}\sum_{m=0}^{j-l}[x^{j-l}](-q(x))^m = (-1)^{j-l}[x^{j-l}]\frac{1-(-q(x))^{j-l+1}}{1+q(x)} = [x^{j-l}]\frac{(-1)^{j-l}+q(x)^{j-l+1}}{1+q(x)}. \quad \square \tag{34}$$

As a final remark note that the determinant takes an even simpler form if the p_i are not 'reversed' as they are in the previous corollary.

Corollary A.3. *If $p_i(x) = a_i x^i + \cdots + a_0$, the formula in Proposition A.1 simplifies to*

$$\frac{\det A_{n,l}}{V_{n-1}} = e_{n-1-l} + \sum_{j=l+1}^{n-1} \left(a_l \prod_{k=l+1}^{j-1} (a_k - 1) \right) e_{n-1-j},$$

Proof. The proof proceeds in a similar way as for the previous corollary. Here we have the case $a_{i,j} = a_j$, so the inner sum in Proposition A.1 evaluates to

$$\sum_{\substack{\{j_2,\ldots,j_m\} \subseteq \{l+1,\ldots,j-1\}, \\ j > j_2 > \cdots > j_m > l}} (-1)^m a_{j_2} \cdots a_{j_m} a_{j_{m+1}} = a_l \prod_{k=l+1}^{j-1} (a_k - 1),$$

and the rest follows. □

References

Albrecher, H., Boxma, O.J., 2005. On the discounted penalty function in a Markov-dependent risk model. Insurance: Mathematics & Economics 37 (3), 650–672.

Bingham, N.H., Goldie, C.M., Teugels, J.L., 1987. Regular Variation. In: Encyclopedia of Mathematics and its Applications, vol. 27. Cambridge University Press, Cambridge.

Bronstein, M., 2005. Symbolic Integration. I, 2nd ed. In: Algorithms and Computation in Mathematics, vol. 1. Springer-Verlag, Berlin.

Butzer, P.L., Berens, H., 1967. Semi-groups of Operators and Approximation. In: Die Grundlehren der mathematischen Wissenschaften, Band 145. Springer-Verlag New York Inc., New York.

Chen, Y.-T., Lee, C.-F., Sheu, Y.-C., 2007. An ODE approach for the expected discounted penalty at ruin in a jump-diffusion model. Finance and Stochastics 11 (3), 323–355.

Cheng, Y., Tang, Q., 2003. Moments of the surplus before ruin and the deficit at ruin in the Erlang (2) risk process. North American Actuarial Journal 7 (1), 1–12.

Constantinescu, C., 2006. Renewal risk processes with stochastic returns on investments—A unified approach and analysis of the ruin probabilities. In: Ph.D. Thesis Edition. Valley Library, Corvallis, OR.

Conway, J.B., 1990. A Course in Functional Analysis, 2nd ed. In: Graduate Texts in Mathematics, vol. 96. Springer-Verlag, New York.

Dickson, D.C.M., Hipp, C., 2001. On the time to ruin for Erlang(2) risk processes. Insurance: Mathematics & Economics 29 (3), 333–344.

Drekic, S., Stafford, J.E., Willmot, G.E., 2004. Symbolic calculation of the moments of the time of ruin. Insurance: Mathematics & Economics 34 (1), 109–120.

Feller, W., 1971. An Introduction to Probability Theory and its Applications. Vol. II, 2nd ed. John Wiley & Sons Inc., New York.

Gerber, H.U., Shiu, E.S.W., 1997. The joint distribution of the time of ruin the surplus immediately before ruin and the deficit at ruin. Insurance: Mathematics & Economics 21 (2), 129–137.

Gerber, H.U., Shiu, E.S.W., 1998. On the time value of ruin. North American Actuarial Journal 2 (1), 48–78.

Gerber, H.U., Shiu, E.S.W., 2005. The time value of ruin in a Sparre Andersen model. North American Actuarial Journal 9 (2), 49–84.

Graham, R.L., Knuth, D.E., Patashnik, O., 1989. Concrete Mathematics. Addison-Wesley Publishing Company Advanced Book Program, Reading, MA.

Heineman, E.R., 1929. Generalized Vandermonde determinants. Transactions of the American Mathematical Society 31 (3), 464–476.

Kamke, E., 1967. Differentialgleichungen. Lösungsmethoden und Lösungen. Teil I: Gewöhnliche Differentialgleichungen, 8th ed. In: Mathematik und ihre Anwendungen in Physik und Technik A, vol. 18. Akademische Verlagsgesellschaft, Leipzig.

Krattenthaler, C., 1999. Advanced determinant calculus. Séminaire Lotharingien de Combinatoire 42 (Art. B42q), 67 (electronic), the Andrews Festschrift (Maratea, 1998).

Landriault, D., Willmot, G., 2008. On the Gerber–Shiu discounted penalty function in the Sparre Andersen model with an arbitrary interclaim time distribution. Insurance: Mathematics & Economics 42 (2), 600–608.

Li, S., Garrido, J., 2004. On ruin for the Erlang(n) risk process. Insurance: Mathematics & Economics 34 (3), 391–408.

Li, S., Garrido, J., 2005a. The Gerber–Shiu function in a Sparre Andersen risk process perturbed by diffusion. Scandinavian Actuarial Journal 3, 161–186.

Li, S., Garrido, J., 2005b. On a general class of renewal risk process: Analysis of the Gerber–Shiu function. Advances in Applied Probability 37 (3), 836–856.

Lin, X.S., Willmot, G.E., 2000. The moments of the time of ruin the surplus before ruin and the deficit at ruin. Insurance: Mathematics & Economics 27 (1), 19–44.

Regensburger, G., Rosenkranz, M., 2009. An algebraic foundation for factoring linear boundary problems. Annali di Matematica Pura ed Applicata Series IV 188 (1), 123–151.

Rosenkranz, M., 2005. A new symbolic method for solving linear two-point boundary value problems on the level of operators. Journal of Symbolic Computation 39 (2), 171–199.

Rosenkranz, M., Regensburger, G., 2008. Solving and factoring boundary problems for linear ordinary differential equations in differential algebras. Journal of Symbolic Computation 43 (8), 515–544.

Sparre Andersen, E., 1957. On the collective theory of risk in case of contagion between claims. Bulletin of the Institute of Mathematics and its Applications 12, 275–279.

Stakgold, I., 2000. Boundary Value Problems of Mathematical Physics. Vol. I, II. In: Classics in Applied Mathematics, vol. 29. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, corrected reprint of the 1967–68 original.

Stanley, R.P., 1999. Enumerative Combinatorics. Vol. 2. In: Cambridge Studies in Advanced Mathematics, vol. 62. Cambridge University Press, Cambridge.

Willmot, G.E., 1999. A Laplace transform representation in a class of renewal queueing and risk processes. Journal of Applied Probability 36 (2), 570–584.

Willmot, G.E., 2007. On the discounted penalty function in the renewal risk model with general interclaim times. Insurance: Mathematics & Economics 41 (1), 17–31.

Yosida, K., 1995. Functional analysis. In: Classics in Mathematics. Springer-Verlag, Berlin, reprint of the sixth (1980) edition.

SIAM J. APPL. MATH.
Vol. 73, No. 1, pp. 47–66

EXACT AND ASYMPTOTIC RESULTS FOR INSURANCE RISK MODELS WITH SURPLUS-DEPENDENT PREMIUMS*

HANSJÖRG ALBRECHER†, CORINA CONSTANTINESCU‡, ZBIGNIEW PALMOWSKI§, GEORG REGENSBURGER¶, AND MARKUS ROSENKRANZ‖

Abstract. In this paper we develop a symbolic technique to obtain asymptotic expressions for ruin probabilities and discounted penalty functions in renewal insurance risk models when the premium income depends on the present surplus of the insurance portfolio. The analysis is based on boundary problems for linear ordinary differential equations with variable coefficients. The algebraic structure of the Green's operators allows us to develop an intuitive way of tackling the asymptotic behavior of the solutions, leading to exponential-type expansions and Cramér-type asymptotics. Furthermore, we obtain closed-form solutions for more specific cases of premium functions in the compound Poisson risk model.

Key words. renewal risk models, surplus dependent premiums, boundary value problems, Green's operators, asymptotic expansions

AMS subject classifications. 91B30, 34B27, 34B05

DOI. 10.1137/110852000

1. Introduction.

The study of level crossing events is a classical topic of risk theory and has turned out to be a fruitful area of applied mathematics, as (depending on the model assumptions) often subtle applications of tools from real and complex analysis, functional analysis, asymptotic analysis, and also algebra are needed (see, e.g., [4] for a recent survey).

In classical insurance risk theory, the collective renewal risk model describes the amount of surplus $U(t)$ of an insurance portfolio at time t by

$$U(t) = u + c\,t - \sum_{k=1}^{N(t)} X_k, \tag{1.1}$$

where c represents a constant rate of premium inflow, $N(t)$ is a renewal process that counts the number of claims incurred during the time interval $(0, t]$, and $(X_k)_{k\geq 0}$ is a sequence of independent and identically distributed (i.i.d.) claim sizes with distribution function F_X and density f_X (also independent of the claim arrival process $N(t)$).

*Received by the editors October 17, 2011; accepted for publication (in revised form) August 29, 2012; published electronically January 14, 2013.
http://www.siam.org/journals/siap/73-1/85200.html

†Department of Actuarial Science, Faculty of Business and Economics, University of Lausanne, CH-1015 Lausanne, Switzerland, and Swiss Finance Institute, 8006 Zurich, Switzerland (hansjoerg.albrecher@unil.ch).

‡Institute for Financial and Actuarial Mathematics, Department of Mathematical Sciences, University of Liverpool, Liverpool L69 7ZL, United Kingdom (c.constantinescu@liverpool.ac.uk). This author's research was partially supported by the Swiss National Science Foundation Project 200021-124635/1.

§Mathematical Institute, University of Wroclaw, 50-384 Wroclaw, Poland (zbigniew.palmowski@gmail.com). This author's research was supported by the Ministry of Science and Higher Education grant NCN 2011/01/B/HS4/00982.

¶INRIA Saclay – Île de France, Project DISCO, L2S, Supélec, 91192 Gif-sur-Yvette Cedex, France (georg.regensburger@ricam.oeaw.ac.at). This author's research was supported by the Austrian Science Fund (FWF): J3030-N18.

‖School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, United Kingdom (M.Rosenkranz@kent.ac.uk).

Let $(\tau_k)_{k\geq 0}$ be the i.i.d. sequence of interclaim times. One of the crucial quantities to investigate in this context is the probability that at some point in time the surplus in the portfolio will not be sufficient to cover the claims, which is called the probability of ruin,

$$\psi(u) = P(T_u < \infty \mid U(0) = u),$$

where $U(0) = u \geq 0$ is the initial capital in the portfolio and

$$T_u = \inf\{t \geq 0 : U(t) < 0 \mid U(0) = u\}.$$

A related, more general quantity is the expected discounted penalty function, which penalizes the ruin event for both the deficit at ruin and the surplus before ruin,

$$\Phi(u) = \mathbb{E}\left(e^{-\delta T_u}\, w(U(T_{u-}), |U(T_u)|)\, 1_{T_u<\infty} \mid U(0) = u\right),$$

where $\delta \geq 0$ is a discount rate and the penalty $w(x, y)$ is a bivariate function. (In risk theory literature, $\Phi(u)$ is often referred to as the Gerber–Shiu function; see [8].)

The classical collective risk model is based on the assumption of a constant premium rate c. However, it is clear that it will often be more realistic to let premium amounts depend on the current surplus level. In this case, the risk process (1.1) is replaced by

$$U(t) = u + \int_0^t p(U(s))\, ds - \sum_{k=1}^{N(t)} X_k.$$

Hence, in between jumps (claims) the risk process moves deterministically along the curve $\varphi(u, t)$, which satisfies the partial differential equation

$$\frac{\partial \varphi}{\partial t} = p(u)\frac{\partial \varphi}{\partial u}; \quad \varphi(u, 0) = u.$$

There are only a few situations for which exact expressions for $\psi(u)$ are known for surplus-dependent premiums. One such case is the Cramér–Lundberg risk model (where $N(t)$ is a homogeneous Poisson process with intensity λ), another is the linear premium function $p(u) = c + \varepsilon u$, which has the interpretation of an interest rate ε on the available surplus. In the case of exponential claims, it was already shown by [22] that the probability of ruin then has the form

$$\psi(u) = \frac{\lambda \varepsilon^{\lambda/\varepsilon - 1}}{\mu^{\lambda/\varepsilon} c^{\lambda/\varepsilon} \exp(-\mu c/\varepsilon) + \lambda \varepsilon^{\lambda/\varepsilon - 1} \Gamma(\frac{\mu c}{\varepsilon}, \frac{\lambda}{\varepsilon})}\, \Gamma\big(\tfrac{\mu(c+\varepsilon u)}{\varepsilon}, \tfrac{\lambda}{\varepsilon}\big), \tag{1.2}$$

where $\Gamma(\eta, x) = \int_x^\infty t^{\eta-1} e^{-t} dt$ is the incomplete gamma function (for extensions to finite-time ruin probabilities, see [11, 12, 3]). In fact, for the Cramér–Lundberg risk model with exponential claims and general monotone premium function $p(u)$, one has the explicit expression

$$\psi(u) = \int_u^\infty \frac{\gamma_0 \lambda}{p(x)} \exp\{\lambda q(x) - \mu x\}\, dx, \tag{1.3}$$

where $1/\gamma_0 \equiv 1 + \lambda \int_0^\infty p(x)^{-1} \exp\{\lambda q(x) - \mu x\}\, dx$ and $q(x) \equiv \int_0^x \frac{1}{p(y)}\, dy$ is assumed finite for $x > 0$ (see [23]). Since for surplus-dependent premiums the probabilistic

approach based on random equations does not work, and also the usual analytic methods lead to difficulties because the equations become too complex, it is a challenge to derive explicit solutions beyond the one given above.

In this paper we will employ a method based on boundary problems and Green's operators to derive closed-form solutions and asymptotic properties of $\psi(u)$ and $\Phi(u)$ under more general model assumptions. For that purpose we will employ the algebraic operator approach developed in [2]. However, since that approach was restricted to linear ordinary differential equations (LODEs) with constant coefficients, we will have to extend the theory to tackle the variable-coefficients equations that occur in the present context.

In section 2 we derive the boundary problem for the Gerber–Shiu function $\Phi(u)$ in a renewal risk model with claim and interclaim distributions having rational Laplace transform. For solving it, we employ a new symbolic method, described in section 3. This allows us to construct integral representations for the solution of inhomogeneous LODEs with variable coefficients, for given initial values, under a stability condition. In section 4 we derive a general asymptotic expansion for the discounted penalty function in the renewal model framework. Subsequently, section 5 is dedicated to the more specific case of compound Poisson risk models with exponential claims, for which we have second-order LODEs. More specifically, in section 5.1 we derive exact solutions for a generic premium function $p(u)$. Further, in section 5.2, we consider some interesting particular cases of $p(u)$. In section 5.3 we identify the necessary conditions a premium function should satisfy so that the asymptotic analysis is possible and the assumptions necessary for the asymptotic results in section 4 are validated. We will end by giving concrete examples of such premium functions and their asymptotics.

Throughout the paper we will assume that $U(t) \to \infty$ a.s. This assumption is satisfied, for example, when $p(u) > \mathbb{E}X/\mathbb{E}\tau + \varsigma$ for some $\varsigma > 0$ and sufficiently large u; see, e.g., [4].

2. Deriving the boundary problem. Assume that the distribution of the interclaim time of the renewal process $N(t)$ has rational Laplace transform. For simplicity of notation, we assume further that the rational Laplace transform has a constant numerator. Then its density f_τ satisfies a LODE with constant coefficients

$$\mathcal{L}_\tau\left(\frac{d}{dt}\right) f_\tau(t) = 0 \tag{2.1}$$

and homogeneous initial conditions $f_\tau^{(k)}(t) = 0$ $(k = 0, \dots, n-2)$, where

$$\mathcal{L}_\tau(x) = x^n + \alpha_{n-1}x^{n-1} + \cdots + \alpha_0.$$

Using the method of [6], we can then derive an integro-differential equation for $\Phi(u)$,

$$\mathcal{L}_\tau^*\left(p(u)\frac{d}{du} - \delta\right)\Phi(u) = \alpha_0\left(\int_0^u \Phi(u-y)\,dF_X(y) + \omega(u)\right), \tag{2.2}$$

where $\mathcal{L}_\tau^*$ is the adjoint operator of $\mathcal{L}_\tau$ defined through

$$\mathcal{L}_\tau^*(x) = \mathcal{L}_\tau(-x) = (-x)^n + \alpha_{n-1}(-x)^{n-1} + \cdots + \alpha_0.$$

Assume now that the claim size distribution also has a rational Laplace transform, so that its density f_X satisfies another such LODE,

$$\mathcal{L}_X\left(\frac{d}{dy}\right) f_X(y) = 0 \tag{2.3}$$

with initial conditions $f_X^{(k)}(x) = 0$ $(k = 0, \ldots, m-2)$, where

$$\mathcal{L}_X(x) = x^m + \beta_{m-1}x^{m-1} + \cdots + \beta_0.$$

Then the integro-differential equation (2.2) becomes a LODE with variable coefficients of order $m+n$, namely

$$T\Phi(u) = g(u) \tag{2.4}$$

with differential operator

$$T = \mathcal{L}_X\left(\frac{d}{du}\right)\mathcal{L}_\tau^*\left(p(u)\frac{d}{du} - \delta\right) - \alpha_0\beta_0 \tag{2.5}$$

and right-hand side

$$g(u) = \alpha_0\,\mathcal{L}_X\left(\frac{d}{du}\right)\omega(u),$$

where $\omega(u) \equiv \int_u^\infty w(u, y-u)\,f_X(y)\,dy$. For $\delta = 0$ and $w = 1$, (2.4) reduces to the well-known equation for the probability of ruin.

The equations hold for sufficiently regular functions p. In the special case $p(u) \equiv c$ one recovers the LODE with constant coefficients whose characteristic polynomial is of degree $n+m$ and corresponds to Lundberg's equation. It is known that, for $\delta > 0$, this polynomial has m solutions σ_i, with negative real part, and n solutions ρ_i, with positive real part; see, for example, [14, 13]. In [2], we have derived

$$\Phi(u) = \gamma_1 e^{\sigma_1 u} + \cdots + \gamma_m e^{\sigma_m u} + Gg(u), \tag{2.6}$$

where the γ_i are determined by the initial conditions and

$$Gg(u) \equiv \sum_{i=1}^{m}\sum_{j=1}^{n} c_{ij}\left(\int_0^u e^{\sigma_i(u-\xi)} + \int_u^\infty e^{\rho_j(u-\xi)} - e^{\sigma_i u}\int_0^\infty e^{-\rho_j(\xi)}\right) g(\xi)\,d\xi \tag{2.7}$$

defines Green's operator for the inhomogeneous LODE (2.4) with homogeneous boundary conditions, where

$$c_{ij} = -\prod_{k=1,k\neq i}^{m}(\sigma_i - \sigma_k)^{-1}\prod_{k=1,k\neq j}^{n}(\rho_j - \rho_k)^{-1}\,(\rho_j - \sigma_i)^{-1}.$$

The boundary conditions for (2.4) consist of the initial conditions $\Phi^{(k)}(0)(k = 0, \ldots, m-1)$, determined from the integro-differential equation, and the stability condition $\Phi(\infty) = 0$, provided by the model assumptions.

Analogous to the constant coefficients case, we assume the existence of a fundamental system for (2.4) with m stable solutions $s_i(u)$ and n unstable solutions $r_j(u)$. Here a solution $f(u)$ is called stable if $f(u) \to 0$ and unstable if $f(u) \to \infty$ as $u \to \infty$. We write $t_1, \ldots, t_{m+n}$ for the complete sequence of solutions $s_1, \ldots, s_m, r_1, \ldots, r_n$, and we assume furthermore that the successive Wronskians $w_k \equiv W[t_1, \ldots, t_k]$ for $k = 1, \ldots, m+n$ are all nonzero on the half-line $\mathbb{R}^+ = [0, \infty)$. Under these assumptions, the algebraic operator approach developed for the constant coefficients case [2] will be extended to the surplus-dependent premium case in section 3, and the general solution of (2.4) then has the form

$$\Phi(u) = \gamma_1 s_1(u) + \cdots + \gamma_m s_m(u) + Gg(u),$$

where the γ_i are determined by the initial values and $Gg(u)$ is again the Green's operator for the inhomogeneous LODE (2.4) with homogeneous boundary conditions, but this time with nonconstant $p(u)$. As a consequence, the representation (2.7) is no longer valid, and we will derive a new explicit expression that generalizes it (Theorem 3.4).

Let us complete this section with a remark about how to check that the fundamental system has stable or unstable solutions. Roughly speaking, this amounts to an asymptotic analysis of the solutions of the homogeneous equation. According to [7, Chap. 5], one can identify conditions on $p(u)$ that guarantee the existence of such a fundamental system. These conditions specify the structure of the coefficients, namely, either they converge (sufficiently fast) to constants—in this case one speaks of *almost constant coefficients*—or they diverge to infinity. The canonical form of (2.4) indicates, of course, that the former case applies for our setting here. However, the speed of convergence of the coefficients depends crucially on the premium function $p(u)$. For instance, we will show in Example 5.4 that for $p(u) = c\,e^{\varepsilon/u}$, the LODE with almost constant coefficients converges to the LODE with constant coefficients given in [2].

3. Green's operator approach. In the previous section we have seen that the core task for computing the Gerber–Shiu function $\Phi(u)$ is to determine the Green's operator G for the inhomogeneous LODE (2.4) with homogeneous boundary conditions consisting of the initial conditions $\Phi^{(k)}(0) = 0 (k = 0, \dots, m-1)$ and the stability condition $\Phi(\infty) = 0$. In this section we will present a symbolic method that allows us to construct G for a generic LODE with variable coefficients and homogeneous boundary conditions. In other words, we consider boundary problems of the general type,

$$\begin{cases} T\,\Phi(u) = g(u), \\ \Phi(0) = \Phi'(0) = \cdots = \Phi^{(m-1)}(0) = 0 \quad \text{and} \quad \Phi(\infty) = 0, \end{cases} \tag{3.1}$$

where $T \equiv D^{m+n} + c_{m+n-1}(u)\,D^{m+n-1} + \cdots + c_1(u)\,D + c_0(u)$ is a linear differential operator with variable coefficients (and leading coefficient normalized to unity) and $D \equiv \frac{d}{du}$. Under the conditions described in section 2 the solution of (3.1) is unique and depends linearly on the so-called forcing function $g(u)$. Therefore the assignment $g \mapsto \Phi$ is a linear operator: the Green's operator G of (3.1). The following fact derives immediately from the theory of ordinary differential equations.

THEOREM 3.1. *The Gerber–Shiu function equals*

$$\Phi(u) = \gamma_1 s_1(u) + \cdots + \gamma_m s_m(u) + Gg(u), \tag{3.2}$$

where G is Green's operator for (3.1)*, and the constants γ_i can be identified from the initial conditions.*

For describing our new method of constructing an explicit representation of G, let us recall how this was achieved in [2] for the special case of constant coefficients $c_i(u) \equiv c_i$. We will use the same notation as found there, in particular the basic operators $A = \int_0^u$, $B = \int_u^\infty$, and the definite integral $F = A + B = \int_0^\infty$. Employing the basic operators, the crucial idea was to factor the Green's operator as

$$G = (-1)^n A_{\sigma_1} \cdots A_{\sigma_m} B_{\rho_1} \cdots B_{\rho_n}, \tag{3.3}$$

where the factor operators are defined by $A_\sigma \equiv e^{\sigma x} A e^{-\sigma x}$ and $B_\rho \equiv e^{\rho x} B e^{-\rho x}$ with σ_i and ρ_j as described before. So the strategy was to decompose the problem and tackle the stable exponents with the basic operator A, and the unstable ones with B.

This idea can be carried over to the general case of (3.1). Using the results of [19], any Green's operator can be fully broken down to basic operators if one can factor the differential operator T into first-order factors. Having a fundamental system $t_1, \dots, t_{m+n} = s_1, \dots, s_m, r_1, \dots, r_n$ with successive Wronskians $w_k(u) \neq 0$ ($k = 1, \dots, m+n$) for $u \in \mathbb{R}^+$, such a factorization of T can always be achieved by well-known techniques described, for example, in equation (18) of [15]; see also [17, 24]. Using this factorization, we can break down G in a way similar to (3.3) except that the A_{σ_i} must be replaced by more complicated operators based on A and s_i, and similarly the B_{ρ_j} by suitable operators involving B and r_j. We assume $m, n > 0$ throughout for avoiding degenerate cases.

PROPOSITION 3.2. *The Green's operator of* (3.1) *is given by* $G = G_s G_r$, *where* $G_s = A_{s_1} \cdots A_{s_m}$ *and* $G_r = (-1)^n B_{r_1} \cdots B_{r_n}$ *with*

$$\begin{aligned} A_{t_i} &= A_{s_i} = \tfrac{w_i}{w_{i-1}}\, A\, \tfrac{w_{i-1}}{w_i} && \text{for } 1 \le i \le m, \\ B_{t_j} &= B_{r_{j-m}} = \tfrac{w_j}{w_{j-1}}\, B\, \tfrac{w_{j-1}}{w_j} && \text{for } m+1 \le j \le m+n, \end{aligned}$$

setting $w_0 = 1$ *for convenience.*

Proof. We employ the factorization $T = T_{r_n} \cdots T_{r_1} T_{s_m} \cdots T_{s_1}$, with the first-order operators given by

$$\begin{aligned} T_{t_i} &= \tfrac{w_{i-1}}{w_i}\, D\, \tfrac{w_i}{w_{i-1}} && \text{for } 1 \le i \le m, \\ T_{t_j} &= T_{r_{j-m}} = \tfrac{w_{j-1}}{w_j}\, D\, \tfrac{w_j}{w_{j-1}} && \text{for } m+1 \le j \le m+n. \end{aligned}$$

It is then clear that $G = A_{s_1} \cdots A_{s_m} (-B_{r_1}) \cdots (-B_{r_n})$ is a right inverse of T since both A and $-B$ are right inverses of D. It remains for us to show that $\Phi = Gg$ satisfies the boundary conditions. Differentiating Φ fewer than m times results in an expression whose summands all have the form $h \cdot (A \cdots g)$ for some functions h; evaluating any such summand yields $h(0) \cdot (\int_0^0 \cdots g) = 0$, so the homogeneous initial conditions are indeed satisfied. For showing that the stability condition $\Phi(\infty) = 0$ is also fulfilled we write $\Phi = A_{s_1} \tilde{g}$ with $\tilde{g} \equiv A_{s_2} \cdots A_{s_m} G_r g$. Then $\Phi = s_1 A s_1^{-1} \tilde{g}$ and hence

$$\Phi(\infty) = s_1(\infty) \textstyle\int_0^\infty s_1(u)^{-1} \tilde{g}(u)\, du = 0$$

because $s_1(\infty) = 0$ and the integral is assumed to converge. □

Note that we assume, in the above proof and henceforth, that all forcing functions are chosen so that all infinite integrals have a finite value (this will be the case in all the examples treated here). This is also the reason why the r_j are incorporated in B operators rather than in A operators as for the s_i. Since we want to focus on the symbolic aspects here, we shall not elaborate these points further.

Spelled out in detail, we can now write the Green's operator of (3.1) in the factored form

$$G = \tfrac{w_1}{w_0}\, C_1\, \tfrac{w_0\, w_2}{w_1^2}\, C_2\, \tfrac{w_1\, w_3}{w_2^2}\, C_3 \cdots C_{m+n-1}\, \tfrac{w_{m+n-2}\, w_n}{w_{m+n-1}^2}\, C_{m+n}\, \tfrac{w_{m+n-1}}{w_{m+n}}, \tag{3.4}$$

where C_i is A for $1 \le i \le m$ and $-B$ for $m+1 \le i \le m+n$. Although this already brings us some way towards a closed form for $\Phi(u)$, we would like to collapse the $m+n$ integrals of (3.4) into a single integration, just as we did in [2].

To start with, assume for a moment that we did not have any unstable solutions so that the fundamental system is only $s_1, \dots, s_m$. In that case we must dispense with

the stability condition, imposing only the homogeneous initial conditions in (3.1). The Green's operator consists only of A operators, without any occurrence of B. In this simplified case, how can one collapse the m integral operators $C_1, \dots, C_m = A$ in (3.4) by a linear combination of single integrators (multiplication operators combined with a single A)? The answer is given by the usual variation-of-constants formula, which can be rewritten in our operator notation as follows [16, 20].

PROPOSITION 3.3. *If $s_1, \dots, s_m$ is a fundamental system for the homogeneous equation $T\Phi = 0$, the Green's operator of* (3.1) *is given by*

$$G_s = s_1 A \tfrac{d_{m,1}}{w_m} + \cdots + s_m A \tfrac{d_{m,m}}{w_m}, \tag{3.5}$$

where w_m is the Wronskian determinant of $s_1, \dots, s_m$ and $d_{m,i}$ results from w_m by replacing the ith column by the mth unit vector.

In other words, $\Phi = Gg$ is a particular solution of $T\Phi = g$, made unique by imposing the initial conditions $\Phi(0) = \Phi'(0) = \cdots = \Phi^{(m-1)}(0) = 0$. In our case, the stability condition $\Phi(\infty) = 0$ follows because $s_i(\infty) = 0$ for all $i = 1, \dots, m$. But note that (3.5) is valid for any fundamental system $s_1, \dots, s_m$ of T, yielding a particular solution for the initial value problem (meaning (3.1) without the stability condition).

Let us now turn to the general case, where the fundamental system $t_1, \dots, t_{m+n}$ consists of $m \geq 1$ stable solutions $s_1, \dots, s_m$ and $n \geq 1$ unstable solutions $r_1, \dots, r_n$. In that case the Green's operator has a representation analogous to (3.5) except that we need B operators in addition to A operators *and* we have to include definite integrals F for "balancing" the B against the A operators.

THEOREM 3.4. *Define the constants*

$$\alpha_{i,j} = d_{i,m+j}(0)/w_{m+j-1}(0) \tag{3.6}$$

for $j = 1, \dots, n$ and $i = 1, \dots, m+n$; the functions $a_j = \alpha_{1,j}\, s_1 + \cdots + \alpha_{m,j}\, s_m$ for $j = 1, \dots, n$; and the functions $\tilde{a}_1, \dots, \tilde{a}_n$ by the recursion $\tilde{a}_1 = a_1$, $\tilde{a}_j = a_j - \alpha_{m+1,j}\, \tilde{a}_1 - \cdots - \alpha_{m+j-1,j}\, \tilde{a}_{j-1}$. Then the Green's operator of (3.1) *is given by*

$$G = \sum_{i=1}^{m+n} t_i\, C_i\, \tfrac{d_{i,m+n}}{w_{m+n}} - \sum_{j=1}^{n} \tilde{a}_j\, F\, \tfrac{d_{m+j,m+n}}{w_{m+n}}, \tag{3.7}$$

where C_i is A for $1 \leq i \leq m$ and $-B$ for $m+1 \leq i \leq m+n$.

The proof of this result is given in the appendix, and there is a more explicit way of specifying the sequence of functions $\tilde{a}_1, \dots, \tilde{a}_n$ occurring in Theorem 3.4.

PROPOSITION 3.5. *The functions $\tilde{a}_j$ in Theorem* 3.4 *can be computed by solving the system $T\tilde{a} = a$, where T is the lower triangular matrix with entries*

$$T_{jk} = \begin{cases} \alpha_{m+k,j} & \text{for } j > k, \\ 1 & \text{for } j = k, \\ 0 & \text{otherwise,} \end{cases}$$

while $\tilde{a}$ and a are, respectively, columns with entries $\tilde{a}_1, \dots, \tilde{a}_n$ and $a_1, \dots, a_n$. Hence we have explicitly $\tilde{a}_j = \det T_j / \det T$, where T_j is the matrix resulting from T by replacing its jth column by a.

Proof. We have $\alpha_{m+1,j}\, \tilde{a}_1 + \cdots + \alpha_{m+j-1,j}\, \tilde{a}_{j-1} + \tilde{a}_j = a_j$, for $j > 1$, by the definition of the $\tilde{a}_j$. But this is clearly the jth row of the matrix $T\tilde{a}$, while the recursion base $\tilde{a}_1 = a_1$ provides the first row. The explicit formula is an application of Cramer's rule. □

In either form, the functions $\tilde{a}_1, \ldots, \tilde{a}_n$ can be readily computed from the given fundamental system $s_1, \ldots, s_m, r_1, \ldots, r_n$, and the representation (3.7) provides a closed form for the Green's operator of (3.1).

4. Asymptotic results for the renewal risk model. In what follows, we will write $k(u) \sim l(u)$ if $\lim_{u\to\infty} \frac{k(u)}{l(u)} = 1$ for some functions k and l. Assume that both the interclaim distribution and the claim size distribution have rational Laplace transform, i.e., their densities satisfy the ODEs (2.1) and (2.3), respectively. Assume that the solutions of (2.4) are of the form $t_i(u) \sim u^{\beta_i} e^{y_i u}$, i.e.,

$$t_i(u) \sim \exp\{A\eta_i(u)\}, \quad i = 1, \ldots, n+m, \tag{4.1}$$

with

$$\eta_i(u) \sim y_i + \frac{\beta_i}{u}, \quad i = 1, \ldots, n+m, \tag{4.2}$$

and

$$y_m < \ldots < y_1 \leq 0 < y_{m+1} < \ldots < y_{m+n} \tag{4.3}$$

(so the η_i are not asymptotically equivalent).

Remark 4.1. Note that for the premium functions $p(u) = c + \varepsilon u$, $p(u) = c + \frac{1}{1+\varepsilon u}$ and $p(u) = c \exp \varepsilon/u$, the corresponding t_i fulfill the conditions (4.1)–(4.3). For $m = n = 1$, a more detailed analysis is presented in section 5.3.

Define the constants $h_k = \gamma_k - \sum_{j=1}^{n} \alpha_{jk} F \frac{d_{m+j,m+n}}{w_{m+n}}$ with γ_k appearing in (3.2) and α_{jk} as defined in (3.6). For a permutation φ on $\{1, \ldots, m+n\}$ we define

$$\pi_i = \frac{\sum_{\varphi(i)=n+m} (-1)^{\mathrm{sgn}\varphi} \prod_{k\neq i} y_k^{\varphi(k)}}{\sum_{\varphi} (-1)^{\mathrm{sgn}\varphi} \prod_{k=1}^{n+m} y_k^{\varphi(k)}},$$

where $\mathrm{sgn}\varphi$ denotes the parity of φ.

THEOREM 4.1. *If $g(u) \sim e^{-\nu u}$ for $\nu > -y_1$, then under* (4.1)–(4.3) *the asymptotic expansion*

$$\Phi(u) \approx \sum_{i=1}^{m+1} \vartheta_i(u) \tag{4.4}$$

holds, with $\vartheta_i(u) = h_i s_i(u)$ $(i = 1, \ldots, m)$ and

$$\vartheta_{m+1}(u) \sim \sum_{i=1}^{m+n} \frac{\pi_i}{y_i + \nu} g(u).$$

This is equivalent to saying that $\lim_{u\to\infty} \frac{\Phi(u) - \sum_{i=1}^{k} \vartheta_i(u)}{\vartheta_{k+1}(u)} = 1$, for $k = 1, \ldots, m$.

Proof. Note that by (4.1)–(4.2), $t_i^{(k)}(u) \sim y_i^k e^{A\eta_i(u)}$. Using (3.7) and the Leibniz formula for the determinant, after some calculations one gets expansion (4.4) with $\vartheta_k(u) = l_k t_k(u)$ and

$$\vartheta_{m+1}(u) = \sum_{i=1}^{m+n} \pi_i t_i(u) C_i \frac{g}{t_i}(u).$$

Using l'Hôpital's rule completes the proof. □

5. Compound Poisson risk process with exponential claims. Let us now focus on the case of a compound Poisson model (exponential interclaim times with mean $1/\lambda$) with exponential claim sizes with mean μ and a generic premium function $p(u)$. The differential equation in (2.4) has order two in this case, so we expect to have one stable solution s and one unstable solution r. In fact, here we can relax the notion of an unstable solution, allowing any function where

$$r(\infty) = \lim_{u\to\infty} r(u)$$

exists and is different from zero (so the limit does not necessarily have to be infinity). The reason for this extension is that the basic argument for the ansatz

$$\Phi(u) = \gamma_s s(u) + \gamma_r r(u)$$

carries over: Every solution of (2.4) must be of the form (5) since $r(u), s(u)$ forms a fundamental system. But then the stability condition $\Phi(\infty) = 0$ can only be satisfied if $\gamma_r = 0$ because we require $s(\infty) = 0$. This is why the form (3.2) is still justified in the special case $n = 1$ with $\gamma_1 = \gamma_s$. But note that this argument fails when there are more than two unstable solutions since they can cancel out unless we take some further precautions (e.g., requiring them to be of the same sign).

5.1. Closed-form solutions for generic premium. For a discount factor $\delta > 0$, the expected discounted penalty function satisfies the second-order LODE

$$(D+\mu)\,(-p(u)D + \delta + \lambda)\,\Phi(u) - \lambda\mu\,\Phi(u) = \lambda(D+\mu)\,\omega(u).$$

Expanding the operators, the equation is equivalent to

$$\big(-p(u)D^2 - (\mu\,p(u) + p'(u) - \lambda - \delta)\,D + \delta\mu\big)\,\Phi(u) = \lambda(D+\mu)\,\omega(u).$$

Assuming that $p(u) \neq 0$ for all $u \geq 0$, this is further equivalent to

$$\left(D^2 + \left(\mu + \frac{p'(u)}{p(u)} - \frac{\lambda+\delta}{p(u)}\right) D - \frac{\delta\mu}{p(u)}\right)\Phi(u) = g(u), \tag{5.1}$$

with $g(u) = -\frac{\lambda}{p(u)}(D+\mu)\,\omega(u)$. Furthermore, we assume that $p(u)$ is chosen in such a way that the associated homogeneous solution has a fundamental system s, r with one stable solution s and one unstable solution r with Wronskian $w = w_2 = sr' - s'r$ nonzero on $\mathbb{R}^+$. Then the Green's operator for the boundary problem for the Gerber–Shiu function Φ is given by Theorem 3.4 with $s_1 = s$ and $r_1 = r$, namely

$$Gg(u) = \left(-s(u)\int_0^u \tfrac{r(v)}{w(v)} - r(u)\int_u^\infty \tfrac{s(v)}{w(v)} + \tfrac{r(0)}{s(0)}s(u)\int_0^\infty \tfrac{s(v)}{w(v)}\right) g(v)\,dv. \tag{5.2}$$

For calculating the full expression

$$\Phi(u) = \gamma\, s(u) + Gg(u), \tag{5.3}$$

we have to determine the constant γ. Evaluating the integro-differential equation (2.2) at zero, one obtains

$$-c\,\Phi'(0) + (\lambda+\delta)\,\Phi(0) = \lambda\,\omega(0)$$

and therefore

$$\gamma = \frac{\lambda\,\omega(0) + c\,(Gg)'(0)}{(\lambda+\delta)\,s(0) - c\,s'(0)} = \frac{\lambda\,\omega(0) + c\,\frac{r(0)s'(0)-r'(0)s(0)}{s(0)}\int_0^\infty \frac{s(v)}{w(v)}\,g(v)\,dv}{(\lambda+\delta)\,s(0) - c\,s'(0)} \tag{5.4}$$

for the required constant. For $\delta = 0$, the LODE (5.1) is of first order in Φ', and its associated homogeneous equation has an unstable solution $r(u) = 1$ and a stable solution

$$s(u) = \int_u^\infty \exp\left(-\mu v + \lambda \int_0^v \tfrac{dy}{p(y)}\right) \frac{dv}{p(v)} \tag{5.5}$$

(cf. [4]). For the fundamental system s, r, the Wronskian is just $w = w_2 = -s'$, and the Green's operator (5.2) specializes to

$$Gg(u) = \left(s(u)\int_0^u \tfrac{1}{s'(v)} + \int_u^\infty \tfrac{s(v)}{s'(v)} - \tfrac{s(u)}{s(0)}\int_0^\infty \tfrac{s(v)}{s'(v)}\right) g(v)\,dv$$

while the constant in $\Phi(u) = \gamma\,s(u) + Gg(u)$ is now given by

$$\gamma = \frac{\lambda\omega(0) - p(0)\frac{s'(0)}{s(0)}\int_0^\infty \frac{s(v)}{s'(v)}\,g(v)\,dv}{\lambda s(0) - p(0)s'(0)}.$$

Thus the Gerber–Shiu function can be written generically as

$$\begin{aligned}\Phi(u) = {} & \frac{\lambda\omega(0) - p(0)\frac{s'(0)}{s(0)}\int_0^\infty \frac{s(v)}{s'(v)}\,g(v)\,dv}{\lambda s(0) - p(0)s'(0)}\,s(u) \\ & + \left(s(u)\int_0^u \tfrac{1}{s'(v)} + \int_u^\infty \tfrac{s(v)}{s'(v)} - \tfrac{s(u)}{s(0)}\int_0^\infty \tfrac{s(v)}{s'(v)}\right) g(v)\,dv\end{aligned}$$

in terms of $s(u)$.

Remark 5.1. For $\delta = 0$ and $w = 1$, one has $g = 0$ and $\psi(u) = \gamma\,s(u)$, recovering (1.3) for the ruin probability.

5.2. Closed-form solutions for some particular premium structures.

(A) Linear premium. As discussed in section 1, the linear function $p(u) = c+\varepsilon u$ can be interpreted as describing investments of the surplus into a bond with a fixed interest rate $\varepsilon > 0$; see, for example, [22]. For $\delta > 0$ and $p(u) = c + \varepsilon u$, we can compute a fundamental system for the second-order LODE

$$\left(D^2 + \left(\mu + \tfrac{\varepsilon}{c+\varepsilon u} - \tfrac{\lambda+\delta}{c+\varepsilon u}\right) D - \tfrac{\delta\mu}{c+\varepsilon u}\right)\Phi(u) = -\tfrac{\lambda}{c+\varepsilon u}\,(D+\mu)\,\omega(u)$$

in the form

$$\begin{aligned} s(u) &= U\left(\tfrac{\delta}{\varepsilon}+1,\ \tfrac{\lambda+\delta}{\varepsilon}+1,\ \mu u + \tfrac{\mu c}{\varepsilon}\right)(\varepsilon u + c)^{\frac{\lambda+\delta}{\varepsilon}}\ \exp(-\mu u), \\ r(u) &= M\left(\tfrac{\delta}{\varepsilon}+1,\ \tfrac{\lambda+\delta}{\varepsilon}+1,\ \mu u + \tfrac{\mu c}{\varepsilon}\right)(\varepsilon u + c)^{\frac{\lambda+\delta}{\varepsilon}}\ \exp(-\mu u),\end{aligned} \tag{5.6}$$

where M and U denote the usual Kummer functions as in [1, section 13.1]. For $u \to \infty$, the estimate in [1, section 13.1.8] yields

$$s(u) = K_1\,(\varepsilon u + c)^{\lambda/\varepsilon - 1}\ \exp(-\mu u)\left(1 + \mathcal{O}(\tfrac{1}{\varepsilon u + c})\right) \to 0, \tag{5.7}$$

while the estimate in [1, section 13.1.4] yields

$$r(u) = K_2\,(\varepsilon u+c)^{\delta/\varepsilon}\left(1+\mathcal{O}(\tfrac{1}{\varepsilon u+c})\right) = K_2\,(\varepsilon u+c)^{\delta/\varepsilon} + \mathcal{O}((\varepsilon u+c)^{\delta/\varepsilon-1}) \to \infty, \tag{5.8}$$

where K_1 and K_2 are some constants. Hence s is indeed a stable and r an unstable solution. Using [1, section 13.1.4] one derives the Wronskian

$$w_2 = \tfrac{\Gamma(\frac{\lambda+\delta}{\varepsilon})}{\Gamma(\frac{\delta}{\varepsilon})}\tfrac{\varepsilon(\lambda+\delta)}{\delta}\big(\tfrac{\varepsilon}{\mu}\big)^{(\lambda+\delta)/\varepsilon}(u\varepsilon+c)^{(\lambda+\delta)/\varepsilon-1}\exp(-\mu u+\tfrac{\mu c}{\varepsilon}).$$

Substituting these expressions in (5.2), we end up with

$$\begin{aligned} Gg(u) &= \tfrac{\Gamma(\delta/\varepsilon+1)}{\Gamma((\delta+\lambda)/(1+\varepsilon))}\tfrac{1}{\varepsilon}\left(\tfrac{\mu}{\varepsilon}\right)^{(\lambda+\delta)/\varepsilon}(\varepsilon u+c)^{(\lambda+\delta)/\varepsilon}\exp(-\mu u-\tfrac{\mu c}{\varepsilon}) \\ &\times \left(-U(u)\int_0^u M(v) - M(u)\int_u^\infty U(v) + \tfrac{M(0)}{U(0)}U(u)\int_0^\infty U(v)\right) g(v)\,dv, \end{aligned}$$

where $U(u)$ and $M(u)$ are Kummer functions appearing on the right hand-side of (5.6). This, jointly with (5.4), is sufficient to determine the discounted penalty function in (5.3).

(B) Exponential premium. In general, an exponential premium function leads to intractable results. However, for

$$p(u) = c(1+e^{-u})$$

the probability of ruin can be worked out from the expression in section 5.1:

$$\psi(u) = -\frac{(1+\frac{\lambda}{c})F(\frac{\lambda}{c},\mu;1+\frac{\lambda}{c};e^u+1)(\frac{1}{2}e^u+\frac{1}{2})^{\frac{\lambda}{c}}}{2\mu F(1+\mu,1+\frac{\lambda}{c};2+\frac{\lambda}{c};2)},$$

where $F(a,b;c;z) = {}_2F_1(a,b;c;z)$ stands for the hypergeometric function; see, e.g., [1].

(C) Rational premium. For a basic rational premium like

$$p(u) = c + \frac{1}{1+u},$$

the exact symbolic form for the probability of ruin can be computed up to quadratures, namely

$$\psi(u) = \frac{\lambda(c+1)^{\lambda/c^2}\int_u^\infty e^{-u(c\mu-\lambda)/c}(c+cu+1)^{-(\lambda+c^2)/c^2}(1+u)du}{1+\lambda(c+1)^{\lambda/c^2}\int_0^\infty e^{-u(c\mu-\lambda)/c}(c+cu+1)^{-(\lambda+c^2)/c^2}(1+u)du}.$$

(D) Quadratic premium. For the quadratic function $p(u) = c+u^2$, the probability of ruin can be determined as

$$\psi(u) = \frac{\lambda\int_u^\infty e^{-(-\lambda\arctan(x/\sqrt{c})+\mu x\sqrt{c})/\sqrt{c}}/(c+x^2)\,dx}{1+\lambda\int_0^\infty e^{-(-\lambda\arctan(x/\sqrt{c})+\mu x\sqrt{c})/\sqrt{c}}/(c+x^2)\,dx}.$$

5.3. Asymptotic results for generic premium. Assume the LODE

$$\Phi^{(n)} + c_{n-1}(u)\,\Phi^{(n-1)} + \cdots + c_0(u)\,\Phi = 0 \tag{5.9}$$

has complex coefficients $c_i(u)$ continuous on $\mathbb{R}^+$ and define its characteristic equation as

$$y^n + c_{n-1}(u)\,y^{n-1} + \cdots + c_0(u) = 0. \tag{5.10}$$

Then the asymptotic behavior of solutions of (5.9) for $u \to \infty$ essentially depends on the behavior of the roots $y_1(u),\dots,y_n(u)$ of (5.10) as $u \to \infty$ (see, e.g., [7, section 5.3.1, p. 250]), which will be exploited below.

5.3.1. Probability of ruin. When $\delta = 0$ and $w = 1$, the expected discounted penalty is the probability of ruin. For this quantity we have the following asymptotic estimate (we use the convention $p(\infty) = \lim_{u\to\infty} p(u)$).

PROPOSITION 5.1.

1. *If $p(\infty) = c$, where c is constant, then*

$$\psi(u) \sim \frac{\mu}{\lambda}\gamma \exp\left(-\mu u + \lambda \int_0^u \tfrac{dw}{p(w)}\right), \quad u \to \infty.$$

2. *If $p(\infty) = \infty$, then*

$$\psi(u) \sim \frac{\mu}{\lambda}\gamma \frac{1}{p(u)} \exp\left(-\mu u + \lambda \int_0^u \tfrac{dw}{p(w)}\right), \quad u \to \infty.$$

Proof. Integration by parts in (5.5) gives

$$s(u) = \frac{\mu}{\lambda} \int_u^\infty \exp\left(-\mu v + \lambda \int_0^v \tfrac{dw}{p(w)}\right) dv - \frac{1}{\lambda} \exp\left(-\mu u + \lambda \int_0^u \tfrac{dw}{p(w)}\right)$$

with $s(0) = \frac{\mu}{\lambda}\hat{h}(\mu) - \frac{1}{\lambda}$, $s'(0) = -\frac{1}{p(0)}$, where $\hat{h}$ denotes the Laplace transform of $h(u) = \exp(\lambda \int_0^u \frac{dw}{p(w)})$. Letting $f(u) = \frac{1}{\lambda}\int_u^\infty \exp(-\mu v + \lambda \int_0^v \frac{dw}{p(w)})\, dv$, one gets

$$\psi(u) = \gamma s(u) = \mu f(u) - f'(u) = \mathcal{L}_X^*\left(\frac{d}{du}\right) f(u),$$

with $\gamma = \frac{\lambda}{\lambda s(0) - p(0)s'(0)}$. Note that $f''(u) = \mu f'(u)(1 + \frac{1}{p(u)})$. To prove the first part of the proposition we need to show that $\psi(u) = -\mu \frac{1}{1+c} f'(u)(1 + o(1))$. According to our previous observation, $\frac{\mu f'(u) - f''(u)}{-\mu f''(u)} = \frac{1}{1+p(u)}$, which completes the proof using l'Hôpital rule. The second part can be proved similarly. □

5.3.2. Expected discounted penalty. We consider two cases of premium functions:

P1. the premium function behaves like a constant at infinity,

$$p(\infty) = c, \qquad p'(u) = O\left(\frac{1}{u^2}\right); \tag{5.11}$$

or

P2. the premium function explodes at infinity, $p(\infty) = \infty$ as

$$p(u) = c + \sum_{i=1}^{l} \varepsilon_i u^i, \qquad c > 0. \tag{5.12}$$

The first case is satisfied by the rational and exponential premium functions. The second case is satisfied by the linear and quadratic premium functions (see section 5.2). Consider first the homogeneous equation (5.1) with $g = 0$, i.e., (2.4) with T given in (2.5) with

$$c_0(u) = -\frac{\delta\mu}{p(u)}, \quad c_1(u) = \mu + \frac{p'(u)}{p(u)} - \frac{\lambda + \delta}{p(u)}.$$

After tedious calculations one can check that in the case (5.11) we have

$$\lim_{u\to\infty} \frac{c_1(u)}{\sqrt{c_0(u)}} < \infty$$

so that Conditions 1) and 2) of [7, p. 252] are satisfied. In the second case, (5.12), we have

$$\lim_{u\to\infty} \frac{c_0(u)}{c_1^2(u)} = 0$$

and Conditions 1) and 2') of [7, p. 254] hold. From [7, p. 252] we hence know that for both cases (5.11) and (5.12) the asymptotic behavior of the solution of (5.1) is

$$t_i(u) \sim \exp\left\{\int_0^u (\varrho_i(t) + \varrho_i^{(1)}(t))dt\right\}, \quad i = 1, 2, \tag{5.13}$$

where

$$\varrho_1 = \frac{-\left(\mu + \frac{p'(u)}{p(u)} - \frac{\lambda+\delta}{p(u)}\right) - \sqrt{\left(\mu + \frac{p'(u)}{p(u)} - \frac{\lambda+\delta}{p(u)}\right)^2 + 4\frac{\delta\mu}{p(u)}}}{2}$$

and

$$\varrho_2 = \frac{-\left(\mu + \frac{p'(u)}{p(u)} - \frac{\lambda+\delta}{p(u)}\right) + \sqrt{\left(\mu + \frac{p'(u)}{p(u)} - \frac{\lambda+\delta}{p(u)}\right)^2 + 4\frac{\delta\mu}{p(u)}}}{2}$$

are the negative and positive solutions, respectively, of the characteristic equation

$$x^2 + \left(\mu + \frac{p'(u)}{p(u)} - \frac{\lambda+\delta}{p(u)}\right)x - \frac{\delta\mu}{p(u)} = 0. \tag{5.14}$$

Here $\varrho_1^{(1)}$ and $\varrho_2^{(1))}$ are defined by

$$\varrho_i^{(1)}(u) = -\frac{\varrho_i'(u)}{2\varrho_i(u) + \left(\mu + \frac{p'(u)}{p(u)} - \frac{\lambda+\delta}{p(u)}\right)} = \frac{\varrho_i'(u)}{\sqrt{\left(\mu + \frac{p'(u)}{p(u)} - \frac{\lambda+\delta}{p(u)}\right)^2 + 4\frac{\delta\mu}{p(u)}}}.$$

Remark 5.2. Note that if the premium function $p(u)$ satisfies conditions (5.11) and (5.12), the solutions $t_i(u)$ will be of the asymptotic form (4.1), where $\eta_i = \varrho_i + \varrho_i^{(1)}$. In order to complete the asymptotic analysis of $\Phi(u)$ for large u, recall that the Gerber–Shiu function Φ is given by $\Phi(u) = \gamma s(u) + Gg(u)$, for a normalizing constant γ.

THEOREM 5.2. *Under the assumptions* (5.11) *and* (5.12) *regarding the premium function, the asymptotics of the Gerber–Shiu function are described by*

$$\Phi(u) \sim h_1 s(u) + K_1 g(u),$$

with the exception

$$\Phi(u) \sim h_1 s(u) + K_2 u\, g(u)$$

for the case (5.12) *with* $l = 1$. *Here* $h_1 = \gamma - \int_0^\infty \frac{s(v)}{s'(v)} g(v)\, dv$.

Remark 5.3. Moreover, for the particular examples considered here, the structure of s (and r) is indeed that of the form (4.1)–(4.3) that we had to impose as a condition in the more general framework.

Proof. First note that for $\delta = 0$, $Gg(u) = 0$, and thus $\Phi(u)$ has the same behavior as the probability of ruin $\psi(u) = \gamma s(u)$. Evaluating the expression (5.13) at $\delta = 0$,

$$s(u) \sim e^{-\mu u + \lambda \int_0^u \frac{dv}{p(v)}} \left(\mu p(u) + p'(u) - \lambda\right)^{-1}$$

leads to the classical result regarding the probability of ruin (1.3). For $\delta \neq 0$, one needs the asymptotic behavior of $Gg(u)$, which based on (5.2) can be reduced to analyzing the asymptotic behavior of

$$q(u) = -s(u) \int_0^u \frac{r(v)}{w(v)}\, g(v) dv - r(u) \int_u^\infty \frac{s(v)}{w(v)}\, g(v) dv,$$

since the term $s(u) \int_0^\infty \frac{s(v)}{s'(v)} g(v)\, dv$ behaves as $s(u)$ at infinity. After rewriting

$$q(u) = -\frac{\int_0^u \frac{r(v)}{w(v)}\, g(v) dv}{\frac{1}{s(u)}} - \frac{\int_u^\infty \frac{s(v)}{w(v)}\, g(v) dv}{\frac{1}{r(u)}}$$

and expanding the Wronskian, one can apply l'Hôpital rule and see that as $u \to \infty$ (after some algebra),

$$q(u) \sim \frac{\frac{1}{\frac{r'(u)}{r(u)} - \frac{s'(u)}{s(u)}} g(u)}{\frac{s'(u)}{s(u)}} - \frac{\frac{1}{\frac{s'(u)}{s(u)} - \frac{r'(u)}{r(u)}} g(u)}{\frac{r'(u)}{r(u)}}. \tag{5.15}$$

Using Fedoryuk's asymptotic expressions (5.13) one more time, one can perform the analysis along the two cases introduced here. It easy to check that in the first case, P1, we have

$$s(u) \sim \exp\{-k_1 u\}, \quad r(u) \sim \exp\{-k_2 u\}, \tag{5.16}$$

where

$$k_1 = -\frac{\left(\mu - \frac{\lambda+\delta}{c}\right) + \sqrt{\left(\mu - \frac{\lambda+\delta}{c}\right)^2 + 4\frac{\delta\mu}{c}}}{2}$$

and

$$k_2 = -\frac{\left(\mu - \frac{\lambda+\delta}{c}\right) - \sqrt{\left(\mu - \frac{\lambda+\delta}{c}\right)^2 + 4\frac{\delta\mu}{c}}}{2}.$$

Thus, as $\lim_{u\to\infty} \frac{s'(u)}{s(u)} = k_1$ and $\lim_{u\to\infty} \frac{r'(u)}{r(u)} = k_2$, then

$$h(u) \sim K_1\, g(u), \quad u \to \infty, \tag{5.17}$$

where

$$K_1 = \frac{k_1 + k_2}{k_1 k_2 (k_2 - k_1)} = \frac{\mu - \frac{\lambda+\delta}{c}}{\frac{\delta\mu}{c}\sqrt{\left(\mu - \frac{\lambda+\delta}{c}\right)^2 + 4\frac{\delta\mu}{c}}}.$$

The second case, P2, is more complex, producing more intriguing asymptotics. One can show that in this case

$$s(u) \sim h_1\, u^{\beta}\, e^{-\mu u},$$

with $\beta \in \mathbb{R}$. Note that for $l = 1$, $\varepsilon_1 = \varepsilon$ one recovers the asymptotics (5.7). One can also check that

$$\lim_{u\to\infty} \frac{s'(u)}{s(u)} = -\mu,$$

whereas

$$\frac{r'(u)}{r(u)} \sim \frac{1}{u} \quad \text{for}\, l = 1, \quad \text{and} \quad r(u) \sim 1 \quad \text{for}\ l > 1,$$

producing, respectively,

$$q(u) \sim u g(u) \quad \text{and} \quad q(u) \sim g(u). \qquad \square \tag{5.18}$$

Example 5.3. Consider a compound Poisson risk model with premium functions described by assumptions P1 or P2. Let the penalty be a function of the surplus only, $w(x, y) = e^{-\nu x}$. Since we are in the exponential claims scenario,

$$g(u) = \lambda\mu(D + \mu) \int_u^{\infty} w(u - y) e^{-\mu y}\, dy = \lambda\nu e^{-(\nu+\mu)u}.$$

Thus, for a linear premium function,

$$\Phi(u) = \gamma s(u) + G g(u)\Phi(u) \sim h_1 u^{\beta} e^{-\mu u} + \lambda\nu u e^{-(\nu+\mu)u},$$

with $\beta = \lambda/\varepsilon - 1$, whereas for all the other premium functions in the class considered here,

$$\Phi(u) \sim h_1 u^{\beta} e^{-\mu u} + \lambda\nu e^{-(\nu+\mu)u}, \quad u \to \infty,$$

with $\beta \in \mathbb{R}$.

Example 5.4. When $p(u) = c \exp \varepsilon/u$, one has a differential equation with *almost constant coefficients*,

$$\left(D^2 + \left(\mu - \frac{\varepsilon}{u^2} - \frac{\lambda + \delta}{c} \exp -\varepsilon/u\right) D - \frac{\delta\mu}{c} \exp -\varepsilon/u\right) \Phi(u)$$

$$= -\frac{\lambda}{c} \exp -\varepsilon/u (D + \mu)\omega(u). \tag{5.19}$$

This is an equation of form (5.9), with coefficients satisfying

$$c_k(u) = \alpha_k + a_k(u), \quad k = 1, 2 \tag{5.20}$$

with α_k constant and $\int_1^{\infty} |a_k(u)| du < \infty$. Here $a_1(u) = -\frac{\varepsilon}{u^2} + \frac{\lambda+\delta}{c} a(u)$ and $a_0(u) = \frac{\delta\mu}{c} a(u)$, with $a(u) = \sum_{k=1}^{\infty} \frac{(-1)^n \varepsilon^n}{u^n n!}$ and thus $\int_1^{\infty} |a(u)| du < \infty$, and similarly for a_0 and a_1. From [5, Thm. 8.1, p. 92] (see also [5, Problem 32, p. 105]) we can hence conclude that the homogeneous equation has a fundamental system with asymptotics

$$s(u) = e^{\sigma u}\,(1 + o(1)) \quad \text{and} \quad r(u) = e^{\rho u}\,(1 + o(1))\,,$$

where σ and ρ are solutions of the equation

$$x^2 + \left(\mu - \frac{\lambda+\delta}{c}\right) x - \frac{\delta\mu}{c} = 0,$$

with $\mathrm{Re}(\sigma) < 0$ and $\mathrm{Re}(\rho) > 0$. Note that these solutions coincide with that of the constant premium case. Consequently, one has the same asymptotic behavior as when the premium rate is constant.

Remark 5.4. The above closed-form solutions were worked out for the compound Poisson model. In principle, similar closed-form solutions are possible for more general renewal risk models as discussed in sections 2–4, in which case higher-order differential equations appear (and have a practical meaning). For LODEs with constant coefficients one can often find closed-form solutions for certain functions of interest [2]. For equations with variable coefficients, closed-form solutions can be obtained by our method if explicit fundamental solutions $t_1, \ldots, t_{m+n}$ of the homogeneous equation are available (as for the second-order examples treated in this paper). Typically this will happen for equations with inherent symmetries. But even if this is not the case, one may always consider numerical approximations for the fundamental solutions $t_1, \ldots, t_{m+n}$ and then apply the Green's operator with those approximations inside. Of course this raises the interesting question of how the error propagates, a problem somewhat similar to the asymptotic analysis presented earlier.

6. Conclusion. We have provided a symbolic method and a conceptual framework for studying boundary value problems with variable coefficients as they appear in modeling the surplus level in a portfolio of insurance contracts in classical risk theory. The approach presented allows a detailed analysis of the asymptotic behavior of the solutions of these equations under a set of conditions. For the specific case of the compound Poisson risk model, these conditions were made more explicit in terms of conditions on the form of $p(u)$. Moreover, several new closed-form solutions were established within this framework.

Appendix. Proof of Theorem 3.4. The proof of Theorem 3.4 hinges on the following technical lemma on Wronskian determinants.

LEMMA A.1. *We have* $\left(\frac{d_{i,k+1}}{w_k}\right)' = -d_{i,k}\,\frac{w_{k+1}}{w_k^2}$ *for* $1 \le i \le k < m+n$.

Proof. We have to show $d_{i,k+1}\,w_k' - d_{i,k+1}'\,w_k = d_{i,k}\,w_{k+1}$. We note that all expressions in this formula are certain minors of the Wronskian matrix W for $t_1, \ldots, t_{m+n}$. So let us write $W^{i_1,\ldots,i_l}_{j_1,\ldots,j_l}$ for the minor of W resulting from deleting the columns indexed $i_1, \ldots, i_l$ and the rows indexed $j_1, \ldots, j_l$. Then we have $w_k = W^{k+1}_{k+1}$, $d_{i,k+1} = (-1)^{i+k+1} W^{i}_{k+1}$, $d_{i,k} = (-1)^{i+k} W^{i,k+1}_{k,k+1}$, with the derivatives $d_{i,k+1}' = (-1)^{i+k+1} W^{i}_{k}$ and $w_k' = W^{k+1}_{k}$. For the latter, we use the fact that a Wronskian determinant can be differentiated if one replaces the last row by its derivative; see, for example, [10, p. 118]. Multiplying by $(-1)^{i+k}$, it remains for us to show $W^i_k W^{k+1}_{k+1} - W^i_{k+1} W^{k+1}_k = W^{i,k+1}_{k,k+1} \cdot \det W$. But this is a classical determinant formula of Sylvester; see, for example, [21, p. 1571] or equation (4.49") in [9]. □

The preceding lemma is the key tool for removing the nested integrals in (3.7). For seeing this, note that it can be read backwards as giving the integral of $d_{i,k}\,w_{k+1}/w_k^2$. In conjunction with certain operator identities taken from [18], this allows us to collapse expressions of the form $A \cdots A$ or $B \cdots B$ or, at the interface of the two blocks, $A \cdots B$.

Proof of Theorem 3.4. Note that the case $n = 0$ reduces to Proposition 3.3, so we

may assume $n > 0$ in what follows. We know from Proposition 3.2 that

$$G = (-1)^n A_{s_1} \cdots A_{s_m} B_{r_1} \cdots B_{r_n},$$

using the notation employed there. Based on this factorization, we prove (3.7) by induction on n. In the base case $n = 1$, applying Proposition 3.3 again yields

$$\begin{aligned} G = \; G_s(-B_{r_1}) &= \left(\sum_{i=1}^{m} s_i \, A \, \tfrac{d_{m,i}}{w_m} \right) \tfrac{w_{m+1}}{w_m} \, (-B) \, \tfrac{w_m}{w_{m+1}} \\ &= \sum_{i=1}^{m} s_i \, A \Big(- d_{m,i} \, \tfrac{w_{m+1}}{w_m^2} \Big) \, B \, \tfrac{w_m}{w_{m+1}}, \end{aligned}$$

and Lemma A.1 gives $(d_{m+1,i}/w_m)'$ for the expression in parentheses. Now we employ the identity $AfB = A\,(\int_0^u f) + (\int_0^u f)\,B$, for arbitrary functions f, from [18]. Substituting the expression in parentheses for f, this gives

$$A \, \tfrac{d_{m+1,i}}{w_m} + \tfrac{d_{m+1,i}}{w_m} \, B - \alpha_{1,i} \, F,$$

so we end up with

$$G = \sum_{i=1}^{m} \Big(s_i \, A \, \tfrac{d_{m+1,i}}{w_{m+1}} + s_i \, \tfrac{d_{m+1,i}}{w_m} \, B \tfrac{w_m}{w_{m+1}} - \alpha_{1,i} \, s_i \, F \tfrac{w_m}{w_{m+1}} \Big).$$

In the middle, we factor out $\sum_i s_i \, d_{m+1,i}$, which equals $-r_1 \, d_{m+1,m+1}$ as one sees by replacing the last row in w_{m+1} by the first and then expanding along that last row. But $d_{m+1,m+1} = w_m$, so the middle sum simplifies to $r_1 \, (-B) \, d_{m+1,m+1}/w_{m+1}$ and may thus be incorporated into the first sum. In the third sum of the above expression, we factor out $\sum_i \alpha_{1,i} \, s_i = a_1$. Thus we finally obtain

$$G = \sum_{i=1}^{m+1} t_i \, C_i \, \tfrac{d_{m+1,i}}{w_{m+1}} - \tilde{a}_1 \, F \, \tfrac{d_{m+1,m+1}}{w_{m+1}},$$

which is the desired formula (3.7) for $n = 1$. Now assume (3.7) for n; we prove it for $n + 1$. Using the induction hypothesis we obtain

$$\begin{aligned} G &= (-1)^n A_{s_1} \cdots A_{s_m} B_{r_1} \cdots B_{r_n} (-B_{r_{n+1}}) \\ &= \left(\sum_{i=1}^{m+n} t_i \, C_i \, \tfrac{d_{i,m+n}}{w_{m+n}} - \sum_{j=1}^{n} \tilde{a}_j \, F \, \tfrac{d_{m+j,m+n}}{w_{m+n}} \right) \tfrac{w_{m+n+1}}{w_{m+n}} \, (-B) \, \tfrac{w_{m+n}}{w_{m+n+1}} \\ &= \sum_{i=1}^{m+n} t_i \, C_i \Big(- d_{i,m+n} \, \tfrac{w_{m+n+1}}{w_{m+n}^2} \Big) \, B \, \tfrac{w_{m+n}}{w_{m+n+1}} \\ &\qquad - \sum_{j=1}^{n} \tilde{a}_j \, F \Big(- d_{m+j,m+n} \tfrac{w_{m+n+1}}{w_{m+n}^2} \Big) \, B \, \tfrac{w_{m+n}}{w_{m+n+1}}. \end{aligned}$$

As before, we see that Lemma A.1, with $n + 1$ in place of n, can be applied to the expressions in the two parentheses, yielding $(d_{m+n+1,i}/w_{m+n})'$ for the former and $(d_{m+n+1,m+j}/w_{m+n})'$ for the latter. In addition to the identity for AfB used for the base case, we now also need the related identities $BfB = (\int_u^\infty f)\,B - B\,(\int_u^\infty f)$ and

$FfB = F\,(\int_0^u f)$, also to be found in [18]. When we substitute for f, these identities take on the form

$$\begin{aligned} AfB &= A\,\tfrac{d_{m+n+1,i}}{w_{m+n}} + \tfrac{d_{m+n+1,i}}{w_{m+n}}\,B - \alpha_{n+1,i}\,F, \\ BfB &= B\,\tfrac{d_{m+n+1,i}}{w_{m+n}} - \tfrac{d_{m+n+1,i}}{w_{m+n}}\,B \end{aligned}$$

for the first expression and

$$FfB = F\,d_{m+n+1,m+j}/w_{m+n} - \alpha_{n+1,m+j}\,F$$

for the second. We split the first sum above into the two sums

$$\begin{aligned} &\sum_{i=1}^{m} s_i \left(A\,\tfrac{d_{m+n+1,i}}{w_{m+n}} + \tfrac{d_{m+n+1,i}}{w_{m+n}}\,B - \alpha_{n+1,i}\,F\right) \tfrac{w_{m+n}}{w_{m+n+1}}, \\ &\sum_{j=1}^{n} r_j \left(\tfrac{d_{m+n+1,m+j}}{w_{m+n}}\,B - B\,\tfrac{d_{m+n+1,m+j}}{w_{m+n}}\right) \tfrac{w_{m+n}}{w_{m+n+1}}. \end{aligned}$$

In the lower-range sum

$$\begin{aligned} \sum_{i=1}^{m} s_i\,A\,\tfrac{d_{m+n+1,i}}{w_{m+n+1}} + \left(\sum_{i=1}^{m} s_i\,d_{m+n+1,i}\right)/w_{m+n}\,B\,\tfrac{w_{m+n}}{w_{m+n+1}} \\ - \left(\sum_{i=1}^{m} \alpha_{n+1,i}\,s_i\right) F\,\tfrac{w_{m+n}}{w_{m+n+1}} \end{aligned}$$

we can apply the same determinant expansion as before to obtain

$$\begin{aligned} &\sum_{i=1}^{m} s_i\,A\,\tfrac{d_{m+n+1,i}}{w_{m+n+1}} + \left(-r_{n+1} - \sum_{j=1}^{n} r_j\,\tfrac{d_{m+n+1,m+j}}{w_{m+n}}\right) B\,\tfrac{w_{m+n}}{w_{m+n+1}} \\ &\qquad - a_{n+1}\,F\,\tfrac{w_{m+n}}{w_{m+n+1}} \\ &= \sum_{i=1}^{m} t_i\,C_i\,\tfrac{d_{m+n+1,i}}{w_{m+n+1}} + t_{m+n+1}\,C_{m+n+1}\,\tfrac{d_{m+n+1,m+n+1}}{w_{m+n+1}} \\ &\qquad - \sum_{j=1}^{n} r_j\,\tfrac{d_{m+n+1,m+j}}{w_{m+n}}\,B\,\tfrac{w_{m+n}}{w_{m+n+1}} - a_{n+1}\,F\,\tfrac{w_{m+n}}{w_{m+n+1}}; \end{aligned}$$

in the upper-range sum we get

$$\sum_{j=1}^{n} r_j\,\tfrac{d_{m+n+1,m+j}}{w_{m+n}}\,B\,\tfrac{w_{m+n}}{w_{m+n+1}} + \sum_{j=1}^{n} t_{m+j}\,C_{m+j}\,\tfrac{d_{m+n+1,m+j}}{w_{m+n+1}}.$$

Combining the lower-range with the upper-range sum, the first sum within the latter cancels with the second sum within the former, yielding

$$\sum_{i=1}^{m+n+1} t_i\,C_i\,\tfrac{d_{m+n+1,i}}{w_{m+n+1}} - a_{n+1}\,F\,\tfrac{w_{m+n}}{w_{m+n+1}}.$$

Now let us tackle the second sum in the above expression for G, namely

$$\begin{aligned}
&-\sum_{j=1}^{n} \tilde{a}_j \, F\big(-d_{m+j,m+n}\tfrac{w_{m+n+1}}{w_{m+n}^2}\big) \, B\,\tfrac{w_{m+n}}{w_{m+n+1}} \\
&= \sum_{j=1}^{n} \tilde{a}_j \, \big(\alpha_{n+1,m+j}\, F - F\,\tfrac{d_{m+n+1,m+j}}{w_{m+n}}\big)\,\tfrac{w_{m+n}}{w_{m+n+1}} \\
&= \left(\sum_{j=1}^{n} \alpha_{n+1,m+j}\,\tilde{a}_j\right) F\,\tfrac{w_{m+n}}{w_{m+n+1}} - \sum_{j=1}^{n} \tilde{a}_j\, F\,\tfrac{d_{m+n+1,m+j}}{w_{m+n+1}},
\end{aligned}$$

where the expression in parentheses is $a_{n+1} - \tilde{a}_{n+1}$ by the definition of the $\tilde{a}_j$. Altogether we obtain now

$$\begin{aligned}
G = &\sum_{i=1}^{m+n+1} t_i\, C_i\,\tfrac{d_{m+n+1,i}}{w_{m+n+1}} - \sum_{j=1}^{n} \tilde{a}_j\, F\,\tfrac{d_{m+n+1,m+j}}{w_{m+n+1}} \\
&\quad - \Big(a_{n+1}\, F\,\tfrac{w_{m+n}}{w_{m+n+1}} - (a_{n+1} - \tilde{a}_{n+1})\, F\,\tfrac{w_{m+n}}{w_{m+n+1}}\Big) \\
= &\sum_{i=1}^{m+n+1} t_i\, C_i\,\tfrac{d_{m+n+1,i}}{w_{m+n+1}} - \sum_{j=1}^{n+1} \tilde{a}_j\, F\,\tfrac{d_{m+n+1,m+j}}{w_{m+n+1}},
\end{aligned}$$

which is indeed (3.7) with $n+1$ in place of n. □

In concluding this appendix, let us also mention that Theorem 3.4 is also valid if B is taken to be the operator $\int_x^b$ with finite $b \in \mathbb{R}$ rather than $b = \infty$. The reason is that the operator identities from [18] are also valid in this case (and were actually set up for this case in the first place).

Acknowledgments. The authors would like to thank the Mathematical Institute of Wroclaw, Poland, the Radon Institute for Computational and Applied Mathematics, Austria, and the University of Lausanne, Switzerland, for accommodating several research visits.

REFERENCES

[1] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards Applied Mathematics Series 55, U.S. Government Printing Office, Washington, D.C., 1964.

[2] H. Albrecher, C. Constantinescu, G. Pirsic, G. Regensburger, and M. Rosenkranz, *An algebraic operator approach to the analysis of Gerber-Shiu functions*, Insurance Math. Econom., 46 (2010), pp. 42–51.

[3] H. Albrecher, J.L. Teugels, and R.F. Tichy, *On a gamma series expansion for the time-dependent probability of collective ruin*, Insurance Math. Econom., 29 (2001), pp. 345–355.

[4] S. Asmussen and H. Albrecher, *Ruin probabilities*, 2nd ed., Adv. Ser. Stat. Sci. Appl. Probab., 14, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2010.

[5] E.A. Coddington and N. Levinson, *Theory of Ordinary Differential Equations*, McGraw-Hill, New York, 1955.

[6] C. Constantinescu and E. Thomann, *Martingales for Renewal Jump-diffusion Processes*, Working paper, 2011.

[7] M. V. Fedoryuk, *Asymptotic Analysis*: *Linear Ordinary Differential Equations*, Springer-Verlag, Berlin, 1993.

[8] H. U. Gerber and E. S. W. Shiu, *On the time value of ruin*, N. Am. Actuar. J., 2 (1998), pp. 48–78.

[9] W. Gröbner, *Matrizenrechnung*, Bibliographisches Institut, Mannheim, Germany, 1966.

[10] E. L. Ince, *Ordinary Differential Equations*, Dover Publications, New York, 1944.
[11] C. Knessl and C. Peters, *Exact and asymptotic solutions for the time-dependent problem of collective ruin*. I, SIAM J. Appl. Math., 54 (1994), pp. 1745–1767.
[12] C. Knessl and C. Peters, *Exact and asymptotic solutions for the time-dependent problem of collective ruin*. II, SIAM J. Appl. Math., 56 (1996), pp. 1471–1521.
[13] D. Landriault and G. Willmot, *On the Gerber-Shiu discounted penalty function in the Sparre Andersen model with an arbitrary interclaim time distribution*, Insurance Math. Econom., 42 (2008), pp. 600–608.
[14] S. Li and J. Garrido, *On a general class of renewal risk process: Analysis of the Gerber-Shiu function*, Adv. Appl. Probab., 37 (2005), pp. 836–856.
[15] G. Pólya, *On the mean-value theorem corresponding to a given linear homogeneous differential equation*, Trans. Amer. Math. Soc., 24 (1922), pp. 312–324.
[16] G. Regensburger and M. Rosenkranz, *Symbolic Integral Operators and Boundary Problems*, Lecture Notes, 2010.
[17] R. Ristroph, *Pólya's property W and factorization—A short proof*, Proc. Amer. Math. Soc., 31 (1972), pp. 631–632.
[18] M. Rosenkranz, *A new symbolic method for solving linear two-point boundary value problems on the level of operators*, J. Symbolic Comput., 39 (2005), pp. 171–199.
[19] M. Rosenkranz and G. Regensburger, *Solving and factoring boundary problems for linear ordinary differential equations in differential algebras*, J. Symbolic Comput., 43 (2008), pp. 515–544.
[20] M. Rosenkranz, G. Regensburger, L. Tec, and B. Buchberger, *Symbolic analysis for boundary problems: From rewriting to parametrized Gröbner bases*, in Numerical and Symbolic Scientific Computing: Progress and Prospects, U. Langer and P. Paule, eds., Springer-Wien, New York, Vienna, 2012, pp. 273–331.
[21] A. Salem and K. Said, *A simple proof of Sylvester's (determinants) identity*, Appl. Math. Sci. (Ruse), 2 (2008), pp. 1571–1580.
[22] C.-O. Segerdahl, *Über einige risikotheoretische Fragestellungen*, Skand. Aktuartidsk., 25 (1942), pp. 43–83.
[23] R.F. Tichy, *Über eine zahlentheoretische Methode zur numerischen Integration und zur Behandlung von Integralgleichungen*, Österreich. Akad. Wiss. Math.-Natur. Kl. Sitzungsber. II, 193 (1984), pp. 329–358.
[24] A. Zettl, *General theory of the factorization of ordinary linear differential operators*, Trans. Amer. Math. Soc., 197 (1974), pp. 341–353.

5th IFAC Symposium on System Structure and Control
Part of 2013 IFAC Joint Conference SSSC, FDA, TDS
Grenoble, France, February 4-6, 2013

Polynomial Solutions and Annihilators of Ordinary Integro-Differential Operators ⋆

Alban Quadrat * **Georg Regensburger** **

* *INRIA Saclay – Île de France, Project DISCO, L2S, Supélec, 91192 Gif-sur-Yvette Cedex, France (e-mail: alban.quadrat@inria.fr).*
** *INRIA Saclay – Île de France, Project DISCO, L2S, Supélec, 91192 Gif-sur-Yvette Cedex, France (e-mail: georg.regensburger@ricam.oeaw.ac.at)*

Abstract: In this paper, we study algorithmic aspects of linear ordinary integro-differential operators with polynomial coefficients. Even though this algebra is not noetherian and has zero divisors, Bavula recently proved that it is coherent, which allows one to develop an algebraic systems theory. For an algorithmic approach to linear systems theory of integro-differential equations with boundary conditions, computing the kernel of matrices is a fundamental task. As a first step, we have to find annihilators, which is, in turn, related to polynomial solutions. We present an algorithmic approach for computing polynomial solutions and the index for a class of linear operators including integro-differential operators. A generating set for right annihilators can be constructed in terms of such polynomial solutions. For initial value problems, an involution of the algebra of integro-differential operators also allows us to compute left annihilators, which can be interpreted as compatibility conditions of integro-differential equations with boundary conditions. We illustrate our approach using an implementation in the computer algebra system Maple. Finally, system-theoretic interpretations of these results are given and illustrated on integro-differential equations.

1. INTRODUCTION

A standard RLC circuit is governed by the following linear *integro-differential (ID) equation*

$$L\,\frac{di(t)}{dt} + R\,i(t) + \frac{1}{C}\int_0^t i(s)\,ds = v(t), \qquad (1)$$

where L is the inductor, R the resistor, C the capacitor, i the current, and v the voltage source. ID equations is a class of equations that naturally appear while modeling natural phenomena and they appear in many applications.

Using operator notation, (1) can be written as:

$$\big(L\,\partial + R + C^{-1}\int\big)\; i(t) = v(t). \qquad (2)$$

The integral operator is generally eliminated by differentiating once (1) to get the following linear ordinary differential (OD) equation:

$$L\,\frac{d^2i(t)}{dt^2} + R\,\frac{di(t)}{dt} + \frac{1}{C}\,i(t) = \frac{dv(t)}{dt}. \qquad (3)$$

If the current source v is constant, we find again the classical second order OD equation defining a RLC circuit. Equation (3) was obtained by pre-multiplying (2) by the differential operator ∂ and using the *fundamental theorem of analysis* stating that $\partial\int = \mathrm{id}$, i.e., $\int$ is a right inverse of ∂. We note that $\int$ is in general not a two-sided inverse since applying the operator $\int\partial$ to a function y, we get

$$\int_0^t \dot{y}(s)\,ds = y(t) - y(0),$$

which shows that $\int\partial = \mathrm{id} - \mathtt{E}$, where $\mathtt{E}$ denotes the evaluation at 0. Initial value problems of linear OD systems can be algebraically investigated using the *evaluation* $\mathtt{E}$.

Rings of *functional operators* (e.g., rings of OD operators, partial differential (PD) operators, differential time-delay operators, differential difference operators) were recently introduced in mathematical systems theory. Since many control linear systems can be defined by means of a matrix with entries in a *skew polynomial ring* or in an *Ore algebra* of functional operators (i.e., classes of univariate or multivariate noncommutative polynomial rings) [6], the classical *polynomial approach* to linear systems theory can be generalized yielding a *module-theoretic approach* to linear functional systems [9, 16, 17, 24]. Symbolic computation techniques (e.g., Gröbner basis techniques) and computer algebra systems can then be used to develop dedicated packages for algebraic systems theory [7, 15].

Algebras of ordinary ID operators have recently been studied within an algebraic approach in [1, 2, 3, 4] and within an algorithmic approach in [20, 21, 22]. The goal of the latter works is to provide an algebraic and algorithmic framework for studying *boundary value problems and Green's operators*.

Even though linear systems of ID equations play an important role in different domains and applications (e.g., PID controllers), it does not seem that they have been extensively studied by the mathematical systems community. For *boundary value systems*, we refer to [10, 11] and the references therein. The first purpose of this paper is to introduce concepts, techniques, and results developed in

⋆ G.R. was supported by the Austrian Science Fund (FWF): J 3030-N18.

978-3-902823-25-0/13/$20.00 © 2013 IFAC 10.3182/20130204-3-FR-2033.00133

the above recents works. In particular, we emphasize that the algebraic structure of the ring of ID operators with polynomial coefficients

is much more involved (e.g., zero divisors, non noetherianity) than the one of the ring of OD operators with polynomial coefficients (the so-called *Weyl algebra*). The fundamental issue of computing left/right kernel of a matrix of ID operators has to be solved towards developing a system-theoretic approach to linear ID systems. The second goal of this paper is to study this problem for an ID operator $d \in \mathbb{I}$, by investigating the computation of zero divisors of d, which allows us to compute compatibility conditions of the inhomogeneous linear ID equation $d\,y = u$. Within a *representation approach*, we show that this problem is related to the computation of polynomial solutions of ID operators, a problem that is studied in detail in this paper.

2. ORDINARY INTEGRO-DIFFERENTIAL OPERATORS WITH POLYNOMIAL COEFFICIENTS

In what follows, let k be a field a field of characteristic zero (i.e., containing a subfield isomorphic to $\mathbb{Q}$). The k-algebra $A(k)$ of OD operators with coefficients in the polynomial ring $k[t]$ (*Weyl algebra*) can be defined in the following two ways (see, e.g., [8]):

(1) Let $k\langle X\rangle$ be the *free associative* k-algebra on $X = \{T, \Delta\}$ (i.e., the k-vector space with the basis formed by all words over X and the multiplication of basis elements defined by the concatenation). Then $A(k) = k\langle X\rangle/J$, where J is the two-sided ideal of $k\langle X\rangle$ generated by $\Psi := \Delta\,T - T\,\Delta - 1$, i.e., $J = k\langle X\rangle\,\Psi\,k\langle X\rangle$. If t (resp., ∂) is the residue class of T (resp., Δ) in $A(k) = k\langle t, \partial\rangle$, then using the relation $\partial\,t = t\,\partial + 1$, any element d of $A(k)$ can uniquely be written as a finite sum
$$d = \sum a_{ij}\,t^i\,\partial^j, \quad a_{ij} \in k,$$
which is called the *normal form* of $d \in A(k)$.

(2) Let $\mathrm{end}_k(k[t])$ be the k-algebra formed by all the k-endomorphisms of $k[t]$ (i.e., k-linear maps form $k[t]$ to $k[t]$). Then, $A(k)$ can also be be defined as the k-subalgebra of $\mathrm{end}_k(k[t])$ generated by the following three k-endomorphisms
$$\begin{cases} 1\colon t^n \longmapsto t^n, \\ t\colon t^n \longmapsto t^{n+1}, \\ \partial\colon t^n \longmapsto n\,t^{n-1}, \end{cases} \quad \forall\, n \in \mathbb{N}$$
defined on the basis $(t^n)_{n\in\mathbb{N}}$ of $k[t]$. In particular, they respectively correspond to the following operators
$$\begin{array}{lll} 1\colon k[t] \longrightarrow k[t] & t\colon k[t] \longrightarrow k[t] & \partial\colon k[t] \longrightarrow k[t] \\ \quad p \longmapsto p, & \quad p \longmapsto t\,p, & \quad p \longmapsto \dot{p}, \end{array} \tag{4}$$
namely, the identity, the multiplication operator, and the derivation operator on the polynomial ring $k[t]$.

The first definition of $A(k)$ is by *generators* (T and Δ) and relations (Ψ). The second one is in terms of *representation theory*. We recall that $\partial\,t = t\,\partial + 1$ translates the following Leibniz rule in the operator language:
$$\partial(t\,y(t)) = t\,\partial\,y(t) + y(t) = (t\,\partial + 1)\,y(t).$$

Let us now introduce an important ring of ID operators.

Definition 1. The *k-algebra of ordinary ID operators with polynomial coefficients* $\mathbb{I}(k)$ is defined as the k-subalgebra of $\mathrm{end}_k(k[t])$ generated by 1, t, ∂ as in (4), and by
$$\begin{aligned} \textstyle\int\colon k[t] &\longrightarrow k[t] \\ t^n &\longmapsto t^{n+1}/(n+1), \quad \forall\, n \in \mathbb{N}. \end{aligned}$$

The algebra $\mathbb{I}(k)$, simply be denoted by $\mathbb{I}$ in what follows, was studied in [1, 3] as a *generalized Weyl algebra*. See [20] for the construction of $\mathbb{I}$ as a factor algebra of a *skew polynomial ring*.

Note that the integral operator $\int$ corresponds to usual integral $p \in k[t] \longmapsto \int_0^t p(s)\,ds \in k[t]$. In the algebra $\mathbb{I}$, the *fundamental theorem* and a version of *integration by parts* can respectively be rewritten as:
$$\partial \textstyle\int = 1, \quad \int\int = t\int - \int t.$$

Moreover, the *evaluation* at 0 can be defined as follows:
$$\mathtt{E} = 1 - \textstyle\int\partial\colon p \in k[t] \longmapsto p(0) \in k. \tag{5}$$
The evaluation $\mathtt{E}$ can be used to study *initial value problems*.

Note that the operator $\mathtt{E}$ naturally induces the existence of *zero divisors* in $\mathbb{I}$ since, for instance, we have:
$$\partial\,\mathtt{E} = \mathtt{E}\textstyle\int = \mathtt{E}\,t = 0.$$
The *left annihilator* of $d \in \mathbb{I}$, namely,
$$\mathrm{ann}_{\mathbb{I}}(.d) := \{e \in \mathbb{I} \mid e\,d = 0\},$$
can be interpreted as *compatibility conditions* of the inhomogeneous ID equation $d\,y(t) = u(t)$. Indeed, we have:
$$\forall\, e \in \mathrm{ann}_{\mathbb{I}}(.d), \quad e\,u(t) = e\,d\,y(t) = 0.$$
If d is not a zero divisor, then $d\,y = u$ does not admit compatibility condition of the form $e\,u = 0$, where $e \in \mathbb{I}$.

Example 2. Let us consider the following trivial example:
$$\textstyle\int_0^t y(s)\,ds = u(t).$$
The compatibility condition $u(0) = 0$ corresponds to the left annihilator $\mathtt{E}$ of $\int$, i.e., $\mathtt{E}\int = 0$ in $\mathbb{I}$.

Let us consider the following inhomogeneous ID equation:
$$\begin{aligned} &t^2\,\ddot{y}(t) - 2\,t\,\dot{y}(t) + (t+2)\,y(t) \\ &\quad - (3\,t/5 + 2)\textstyle\int_0^t y(s)\,ds + 3/5\int_0^t s\,y(s)\,ds = u(t). \end{aligned} \tag{6}$$
The left annihilator of the following IO operator
$$d = t^2\,\partial^2 - 2\,t\,\partial + (t+2) - (3\,t/5 + 2)\textstyle\int + 3/5\int t \in \mathbb{I} \tag{7}$$
yields the compatibility conditions of (6). See Example 22.

For the general construction of the algebra of ID operators $\mathcal{F}_\Phi[\partial, \int]$ defined over an ordinary ID algebra $\mathcal{F}$ and endowed with a set of *characters* (i.e., multiplicative linear functionals) Φ, we refer to [21, 22]. We note that the algebra $\mathbb{I}$ can be seen as a special case of this construction with $\mathcal{F} = (k[t], \partial, \int)$ and $\Phi = \{\mathtt{E}\}$. Hence, $\mathbb{I}$ can be defined as $k\langle t, \partial, \int\rangle = k\langle T, \Delta, I\rangle/J$, where J is the two sided ideal of the free algebra $k\langle T, \Delta, I\rangle$ generated by:
$$\Delta\,T - T\,\Delta - 1, \quad \Delta\,I - 1, \quad I\,I - T\,I + I\,T, \quad T - I\,\Delta\,T.$$
In particular, we have the following relations in $\mathbb{I}$
$$\partial\,t = t\,\partial + 1, \quad \partial\textstyle\int = 1, \quad \int\int = t\int - \int t, \quad \mathtt{E}\,t = 0,$$
where $\mathtt{E} = 1 - \int\partial$.

More generally, we denote the evaluation at $\alpha \in k$ by
$$\mathtt{E}_\alpha\colon p \in k[t] \longmapsto p(\alpha) \in k.$$

The corresponding relations are

$$\forall\, \alpha,\, \beta \in k, \quad \mathtt{E}_\alpha\, t = \alpha \quad \text{and} \quad \mathtt{E}_\beta\, \mathtt{E}_\alpha = \mathtt{E}_\alpha.$$

In contrast to [1, 3], this last approach allows one to have more than one point evaluation, which is crucial for the study of *boundary value problems*.

Let $\Phi \subseteq k$. Identifying $\alpha \in \Phi$ with the evaluation $\mathtt{E}_\alpha$, we denote by $\mathbb{I}_\Phi$ the algebra of ID operators with polynomial coefficients endowed with the set of characters Φ. Then, every ID operator $d \in \mathbb{I}_\Phi$ can be uniquely written as a sum $d = d_1 + d_2 + d_3$, where $d_1 = \sum a_{ij}\, t^i\, \partial^j$ is an OD operator, $d_2 = \sum b_{ij}\, t^i \int t^j$ an *integral operator*, and

$$d_3 = \sum_{\alpha \in \Phi} \Big(\sum f_{ij}\, t^i\, \mathtt{E}_\alpha\, \partial^j + \sum g_{ij}\, t^i\, \mathtt{E}_\alpha \int t^j \Big) \qquad (8)$$

a *boundary operator*, where a_{ij}, b_{ij}, f_{ij}, and $g_{ij} \in k$, and d_1, d_2, and d_3 contain only finitely nonzero summands. See [21, 20] for details, in particular, for a Gröbner basis of the defining relations. For $\alpha = 0$, a boundary operator (8) is of the form $\sum c_{ij}\, t^i\, \mathtt{E}_0\, \partial^k$ since $\mathtt{E} \int = 0$.

In the following, we discuss some important algebraic properties of $\mathbb{I}$. First, since the integral operator $\int$ is a right but not a left inverse of the derivation ∂, it is known that the algebra $\mathbb{I}$ is necessarily non *noetherian* [12]. More explicitly, if $\int^i = \int \cdots \int$ denotes the product of i integral operators, one verifies that $e_{ij} = \int^i \mathtt{E}\, \partial^j$ satisfy

$$e_{ij}\, e_{lm} = \delta_{jl}\, e_{im}, \qquad (9)$$

where $\delta_{jl} = 1$ for $j = l$, and 0 otherwise; see [12] or [14, Ex. 21.26]. In particular, $\mathbb{I}$ contains infinitely many *orthogonal idempotents* e_{ii} for all $i \in \mathbb{N}$, i.e., $e_{ii}\, e_{jj} = \delta_{ij}$ for all i, $j \in \mathbb{N}$. If we introduce $e_k = e_{11} + \cdots + e_{kk} \in \mathbb{I}$ for $k \geq 1$, then using (9), we get $e_{ii} = e_{ii}\, e_k = e_k\, e_{ii}$ for $1 \leq i \leq k$, and the increasing sequence $\{I_k := \mathbb{I}\, e_k\}_{k \geq 1}$ (resp., $\{I_k := e_k\, \mathbb{I}\}_{k \geq 1}$) of principal left (resp., right) ideals of $\mathbb{I}$ is not stationary (see [12]), which proves that $\mathbb{I}$ is not a left (resp., a right) noetherian ring.

The following fundamental result was obtained by Bavula.

Theorem 3. ([1]). The ring $\mathbb{I}$ is *coherent*, i.e., for every $r \geq 1$, and for all $d_1, \ldots, d_r \in \mathbb{I}$, the left (resp., right) $\mathbb{I}$-module $S = \big\{(e_1, \ldots, e_r) \in \mathbb{I}^{1 \times r} \mid \sum_{i=1}^r e_i\, d_i = 0\big\}$ (resp., $S = \big\{(e_1, \ldots, e_r)^T \in \mathbb{I}^{r \times 1} \mid \sum_{i=1}^r d_i\, e_i = 0\big\}$) is finitely generated as a left (resp., right) $\mathbb{I}$-module.

Linear systems are usually described by means of finite matrices with entries in a certain ring D. As explained in [18], if D is a coherent ring, an algebraic systems theory can be developed as if D was a noetherian ring. Hence, Theorem 3 shows that an algebraic systems theory can be developed over $\mathbb{I}$. In particular, basic module-theoretic operations of *finitely presented* left/right $\mathbb{I}$-modules, namely, left/right $\mathbb{I}$-modules defined by matrices, are finitely presented, and thus, finitely generated. For more details, see, e.g., [14, 23]. It is shown in [4] that Theorem 3 cannot be generalized for more than one differential operator, i.e., for $\mathbb{I}_n$ and $n > 1$ (partial analogues).

Based on the normal forms for generalized Weyl algebras, it is shown in [3] that $\mathbb{I}$ admits the *involution* θ

$$\theta(\partial) = \int, \quad \theta(\textstyle\int) = \partial, \quad \theta(t) = t\, \partial^2 + \partial = (t\, \partial + 1)\, \partial, \quad (10)$$

i.e., an *anti-automorphism* of D of order two, namely, the k-linear map θ satisfies the following two properties:

$$\forall\, d,\, e \in D, \quad \theta(d\, e) = \theta(e)\, \theta(d), \quad \theta^2(d) = d.$$

An important consequence is that many algebraic properties of left $\mathbb{I}$-modules have a right analogue and conversely.

The computation of *syzygies*, namely, left/right kernel of a matrix with entries in $\mathbb{I}$ is a central task towards developing an algorithmic approach to linear systems of ID equations with boundary conditions based on module theory and homological algebra. See [6, 15, 19] and references therein. As a first step, we have to find left/right zero divisors of elements of $\mathbb{I}$. This problem leads, in turn, to computing polynomial solutions of ordinary ID equations with boundary conditions.

Finally, in [1, 2, 3], various algebraic properties of $\mathbb{I}$ and important results are proven amongst them a classification of *simple modules*, an analogue of *Stafford's theorem*, and of the *first conjecture of Dixmier*.

3. FREDHOLM AND FINITE-RANK OPERATORS

Several properties of *Fredholm operators* can be studied in the purely algebraic setting of linear maps on infinite-dimensional vector spaces. In [1], such properties are used to investigate $\mathbb{I}$. It turns out that Fredholm operators are also very useful for an algorithmic approach to operator algebras. We review some algebraic properties of Fredholm operators in the following.

Definition 4. A k-linear map $f : V \longrightarrow W$ between two k-vector spaces is called *Fredholm* if it has finite dimensional kernel and cokernel, where $\operatorname{coker} f = W / \operatorname{im} f$. The *index* of a Fredholm operator f is defined by:

$$\operatorname{ind}_k f = \dim_k(\ker f) - \dim_k(\operatorname{coker} f).$$

We have the *long exact sequence* of k-vector spaces ([23])

$$0 \longrightarrow \ker f \xrightarrow{\ i\ } V \xrightarrow{\ f\ } W \xrightarrow{\ p\ } \operatorname{coker} f \longrightarrow 0,$$

i.e., i is injective, $\ker f = \operatorname{im} i$, $\ker p = \operatorname{im} f$, and p is surjective. Then, $\dim_k(\operatorname{coker} f)$ gives the number of independent k-linear compatibility conditions $g(w) = 0$ on w for the solvability of the inhomogeneous linear system $f(v) = w$ (e.g., f is surjective iff $\operatorname{coker} f = 0$), while $\dim_k(\ker f)$ measures the degrees of freedom in a solution ($v + u$ is solution for all $u \in \ker f$).

Example 5. Viewing the basic operators 1, t, ∂, $\int \in \mathbb{I}$ as k-linear maps on $k[t]$, we get:

$$\ker 1 = \ker t = \ker \textstyle\int = 0, \quad \ker \partial = k,$$
$$\operatorname{im} 1 = \operatorname{im} \partial = k[t], \quad \operatorname{im} t = \operatorname{im} \textstyle\int = k[t]\, t.$$

Hence, they are also Fredholm with index:

$$\operatorname{ind}_k 1 = 0, \quad \operatorname{ind}_k t = \operatorname{ind}_k \textstyle\int = -1, \quad \operatorname{ind}_k \partial = 1.$$

If V and W are two finite-dimensional k-vector spaces, then $\dim_k(\operatorname{coker} f) = \dim_k(W) - \dim_k(\operatorname{im} f)$ and the rank-nullity theorem yields $\dim_k V = \dim_k(\operatorname{im} f) + \dim_k(\ker f)$,

$$\operatorname{ind}_k f = \dim_k V - \dim_k W, \qquad (11)$$

i.e., $\operatorname{ind}_k f$ depends only on the dimensions of V and W.

We also recall the index formula for Fredholm operators.

Proposition 6. Let $V' \xrightarrow{\ f\ } V \xrightarrow{\ g\ } V''$ be k-linear maps between k-vector spaces. If two of the maps f, g, and $g \circ f$ are Fredholm, then so is the third, and:

$$\operatorname{ind}_k(g \circ f) = \operatorname{ind}_k g + \operatorname{ind}_k f.$$

Definition 7. A k-linear map between two k-vector spaces is called *finite-rank* if its image is finite-dimensional.

Example 8. Let us consider $\mathtt{E} = 1 - \int \partial \in \mathbb{I} \subset \mathrm{end}_k(k[x])$. It has an infinite-dimensional kernel $\ker_k \mathtt{E} = k[t]\,t$, but its image $\mathrm{im}_k \mathtt{E} = k$ is one-dimensional. More generally, every boundary operator $d_3 \in \mathbb{I}_\Phi$ is obviously of finite rank since its image is contained in the k-vector space of polynomials with degree less than or equal n, where n is the maximal index i with a nonzero coefficient f_{ij} or g_{ij} in (8).

Clearly, composing a finite-rank map with a linear map from either side gives again finite-rank map and Proposition 6 shows that the composition of two Fredholm operators is a Fredholm operator.

Proposition 9. Let V be a k-vector space and A a k-subalgebra of $\mathrm{end}_k(V)$. Then, $\mathcal{F}_A = \{a \in A \mid a \text{ Fredholm}\}$ forms a monoid and $\mathcal{C}_A = \{c \in A \mid c \text{ finite-rank}\}$ is a two-sided ideal of A.

In particular, the boundary operators $(\Phi) \subset \mathbb{I}_\Phi$ form a two-sided subideal of $\mathcal{C}_{\mathrm{end}_k(k[t])}$ generated by the evaluations $\mathtt{E}_\alpha \in \Phi$, and all other ID operators $\mathbb{I}_\Phi \backslash (\Phi)$ are Fredholm as we shall see in Proposition 15. More generally, Bavula has introduced in [1] the notion of *(strong) compact-Fredholm alternative* for an arbitrary k-algebra A.

4. POLYNOMIAL SOLUTIONS OF RATIONAL INDICIAL MAPS AND POLYNOMIAL INDEX

Computing polynomial solutions of linear systems of OD is well-studied in symbolic computation since it appears as a subproblem of many important algorithms. See [5] and the references therein. In this section, we discuss an algebraic setting and an algorithmic approach for the computation of polynomial solutions (kernel), cokernel, and the "polynomial" index for a general class of linear operators including ID operators.

For computing the kernel and cokernel of a k-linear map $L\colon V \longrightarrow V'$ on infinite-dimensional k-vector spaces V and V', we can use the following simple consequence of the snake lemma.

Lemma 10. Let $L\colon V \longrightarrow V'$ be a k-linear map and $U \subseteq V$, $U' \subseteq V'$ k-subspaces such that $L(U) \subseteq U'$. Let $L' = L_{|U}\colon U \longrightarrow U'$ and $\overline{L}\colon V/U \longrightarrow V'/U'$ the induced k-linear map defined by $\overline{L}(\pi(v)) = \pi'(L(v))$ for all $v \in V$, where $\pi\colon V \longrightarrow V/U$ (resp., $\pi'\colon V' \longrightarrow V'/U'$) is the canonical projection onto V/U (resp., V'/U'). Then, we have the following commutative exact diagram:

$$\begin{array}{ccccccccc} 0 & \longrightarrow & U & \longrightarrow & V & \xrightarrow{\pi} & V/U & \longrightarrow & 0 \\ & & \downarrow{\scriptstyle L'} & & \downarrow{\scriptstyle L} & & \downarrow{\scriptstyle \overline{L}} & & \\ 0 & \longrightarrow & U' & \longrightarrow & V' & \xrightarrow{\pi'} & V'/U' & \longrightarrow & 0. \end{array} \qquad (12)$$

If $\overline{L}$ is an isomorphism, i.e., $V/U \cong V'/U'$, then:

$$\ker L' = \ker L, \quad \mathrm{coker}\, L' \cong \mathrm{coker}\, L.$$

Moreover, if U and U' are two finite-dimensional k-vector spaces, then L is Fredholm and $\mathrm{ind}_k\, L = \dim_k U - \dim_k U'$.

Proof. Since $\overline{L}$ is an isomorphism, applying the standard the snake lemma (see, e.g., [23]) to (12), we obtain the following long exact sequence of k-vector spaces

$$0 \longrightarrow \ker L' \longrightarrow \ker L \longrightarrow 0 \longrightarrow \mathrm{coker}\, L' \longrightarrow \mathrm{coker}\, L \longrightarrow 0,$$

and the statements about the kernel and cokernel follow. If U and U' are two finite-dimensional k-vector spaces, then so are $\ker L' = \ker L$ and $\mathrm{coker}\, L' \cong \mathrm{coker}\, L$ and $\mathrm{ind}_k\, L = \mathrm{ind}_k\, L' = \dim_k U - \dim_k U'$ by (11).

From an algorithmic point of view, we want to find finite-dimensional k-subspaces U and U', and an algorithmic criterion for $\overline{L}$ being an isomorphism on the remaining infinite-dimensional parts V/U and V'/U'.

The cokernel of a k-linear map $f\colon V \longrightarrow W$ between two finite-dimensional k-vector spaces V and W can be characterized as follows. Choosing bases of V and W, there exists a matrix $C \in k^{m\times n}$ such that $f(v) = C\,v$ for all $v \in V \cong k^n$. Computing a basis of the finite-dimensional k-vector space $\ker C^T$ and stacking the elements of this basis into a matrix $D \in k^{m\times p}$, we get $\ker C^T = \mathrm{im}\, D$. Then, $\mathrm{coker}\, f \cong \mathrm{im}\, D^T$ and, more precisely, if $\pi\colon W \longrightarrow \mathrm{coker}\, f$ is the canonical projection onto $\mathrm{coker}\, f$, then the k-linear map $\sigma\colon \mathrm{coker}\, f \longrightarrow \mathrm{im}\, D$ defined by $\sigma(\pi(w)) = D\,w$ for all $w \in W$, is an isomorphism of k-vector spaces.

Let us study when the k-linear map $\overline{L}\colon V/U \longrightarrow V'/U'$ is an isomorphism. In what follows, we shall focus on the polynomial case, namely, $V = V' = k[t]$. To do that, let us introduce the degree filtration of $k[t]$, namely,

$$k[t] = \bigcup_{i\in\mathbb{N}} k[t]_{\leq i}, \quad k[t]_{\leq i} = \bigoplus_{j=0}^{i} k\,t^j,$$

defined by the finite-dimensional k-vector spaces $k[t]_{\leq i}$ formed by the polynomials of $k[t]$ of degree less than or equal to i (we set $k[t]_{\leq -1} = 0$). Note that this filtration is induced by any basis $\{p_i\}_{i\in\mathbb{N}}$ of $k[t]$ with $\deg p_i = i$ for all $i \in \mathbb{N}$. We recall that the multiplication operator, derivation, and integral operator are defined by (4), and we can check that:

$$\begin{cases} (t^i\,\partial^j)(t^n) = \dfrac{n!}{(n-j)!}\,t^{n-j+i}, \\ (t^i \int t^j)(t^n) = \dfrac{1}{n+j+1}\,t^{i+j+n+1}. \end{cases}$$

Definition 11. A k-linear map $L\colon k[t] \longrightarrow k[t]$ is called *rational indicial* with *rational symbol* $\mathrm{rsym}(L) = (s,\, q)$ if there exist a nonzero rational function $q \in k(n)$, $c_n \in k^*$, and $M \in \mathbb{N}$ such that:
$\forall\, n \geq M \geq -s,\ L(t^n) = c_n\, q(n)\, t^{n+s} + \text{lower degree terms}$.

Example 12. The rational symbols of (4) are:

$$\mathrm{rsym}(1) = (0,\, 1), \qquad \mathrm{rsym}(t) = (1,\, 1),$$
$$\mathrm{rsym}(\partial) = (-1,\, n), \quad \mathrm{rsym}(\textstyle\int) = \left(1,\, \dfrac{1}{(n+1)}\right).$$

Operators such as shift, dilation, convolution operators on $k[t]$ are also rational indicial. The sum of a rational indicial map and a finite-rank map is also rational indicial with the same symbol, as one sees, by choosing the bound M large enough, e.g., for $1 + t^3\mathtt{E}_0$, one can take $M = 4$.

Let us now state a result for the computation of the kernel and cokernel of rational indicial maps (compare with Lemma 6.5. of [1]).

Proposition 13. Let $L\colon k[t] \longrightarrow k[t]$ be a k-linear map. Let $-1 \leq N$, $-(N+1) \leq s$, $\quad U = k[t]_{\leq N}$, $\quad U' = k[t]_{\leq N+s}$ be such that $L(U) \subseteq U'$. Let $L' = L_{|U}\colon U \longrightarrow U'$ be the induced map. If $\deg L(t^n) = n + s$ for all $n \geq N+1$, then:

$$\ker L' = \ker L, \quad \operatorname{coker} L' \cong \operatorname{coker} L.$$

Moreover, L is a Fredholm operator with $\operatorname{ind}_k L = -s$.

Proof. Let $V = V' = k[t]$ and $\pi : V \longrightarrow V/U$ (resp., $\pi' : V' \longrightarrow V'/U'$) be the canonical projection onto V/U (resp., V'/U'). Then, $\overline{L}(\pi(t^n)) = \pi'(L(t^n))$ for all $n \in N$.

Now, the condition on the degree of the image $L(t^n)$ for $n \geq N$ shows that $\overline{L}$ maps the basis $\{\pi(t^n)\}_{n \geq N}$ of V/U to a basis of V'/U', and thus, defines an isomorphism. The result then follows from Lemma 10 after noting that:

$$\dim_k U - \dim_k U' = N + 1 - (N + 1 + s) = -s.$$

Given a rational indicial operator with rational symbol (s, q) and bound M, we obtain a bound N for Proposition 13 by computing the largest nonnegative integer root l of q and taking $N = \max(l, M)$. Hence computing the kernel and cokernel of $L\colon k[t] \longrightarrow k[t]$ reduces to the same problem for the k-linear map $L' = L_{|U}\colon U \longrightarrow U'$ between two finite-dimensional k-vector spaces, which can be solved using basic linear algebra techniques. We have implemented in Maple the computation of kernel and cokernel of rational indicial maps.

Corollary 14. A rational indicial operator with rational symbol (s, q) is Fredholm with index $-s$ and its kernel and cokernel can be effectively computed.

We can explicitly compute the rational symbol (s, q) for $d \notin (\Phi)$ from its normal form. The following proposition is a purely algebraic version of an *index theorem* (compare with [1, Proposition 6.1]).

Proposition 15. Let $d = \sum a_{ij}\, t^i\, \partial^j + \sum b_{ij}\, t^i \int t^j + d_3 \in \mathbb{I}_\Phi$ be an ID operator, where $d_3 \in (\Phi)$, such that $d \notin (\Phi)$. Then, the k-linear map

$$\begin{aligned} L_d\colon k[t] &\longrightarrow k[t], \\ p &\longmapsto d(p), \end{aligned} \tag{13}$$

is rational indicial with rational symbol

$s = -\operatorname{ind}_k d = \max(\{i-j \mid a_{ij} \neq 0\} \cup \{i+j+1 \mid b_{ij} \neq 0\})$, and $q(n) = \sum_{i-j=s} a_{ij}\, \frac{n!}{(n-j)!} + \sum_{i+j+1=s} b_{ij}\, \frac{1}{n+j+1}$.

5. POLYNOMIAL SOLUTION AND ANNIHILATORS

In the proof of Theorem 3, the fact that the left and right annihilators are finitely generated $\mathbb{I}$-modules is used, for which a non-constructive argument is given in [1].

Theorem 16. ([1]). If $d \in \mathbb{I}$, then the left (resp., right) annihilator $\operatorname{ann}_{\mathbb{I}}(.d)$ (resp., $\operatorname{ann}_{\mathbb{I}}(d.) := \{e \in \mathbb{I} \mid d\, e = 0\}$) of d is a finitely generated left (resp., right) $\mathbb{I}$-module.

We generalize this result to the right annihilator of a Fredholm operator $d \in \mathbb{I}_\Phi$ with more than one evaluation using a constructive approach. It is based on the fact that we can identify (as for the Weyl algebra and $\mathbb{I}$) an integro-differential operator d with the corresponding linear map L_d on the polynomial ring $k[t]$.

Theorem 17. The k-algebra homomorphism

$$\begin{aligned} \chi\colon \mathbb{I}_\Phi &\longrightarrow \operatorname{end}_k(k[t]) \\ d &\longmapsto L_d, \end{aligned}$$

is a *faithful representation* of $\mathbb{I}_\Phi$, i.e., χ is injective.

For a proof that χ is injective, we first observe that for $d \notin (\Phi)$, the k-linear map L_d is obviously nonzero by Proposition 15. So, let $d \in (\Phi)$ be a boundary operator. By (8), d is a finite $k[t]$-linear combination of terms of the form $\mathtt{E}_\alpha\, \partial^i$ and $\mathtt{E}_\alpha \int t^i$, where $\alpha \in \Phi$, namely

$$d = \sum_{\alpha \in \Phi} \left(\sum_{i=0}^{l} p_{\alpha,i}\, \mathtt{E}_\alpha\, \partial^i + \sum_{i=0}^{m} q_{\alpha,i}\, \mathtt{E}_\alpha \int t^i \right), \tag{14}$$

where $p_{\alpha,i},\ q_{\alpha,i} \in k[t]$.

Lemma 18. The k-linear functionals $\mathtt{E}_\alpha\, \partial^i$ and $\mathtt{E}_\alpha \int t^i$ on $k[t]$ for $i \in \mathbb{N}$ and $\alpha \in k$ are k-linearly independent.

Proof. This can be seen by evaluating $\mathtt{E}_\alpha\, \partial^i$ and $\mathtt{E}_\alpha \int t^i$ on sufficiently many polynomials of the form $(t - c)^n$ for $c \in k$ and $n \in \mathbb{N}$ since

(1) Evaluating $\mathtt{E}_{\alpha_1}, \ldots, \mathtt{E}_{\alpha_m}$ for distinct $\alpha_1, \ldots, \alpha_m \in k$ on $1, t, \ldots, t^{m-1}$ gives a *Vandermonde* matrix.
(2) Evaluating the functionals $\mathtt{E}_\alpha\, \partial, \mathtt{E}_\alpha\, \partial^2, \ldots, \mathtt{E}_\alpha\, \partial^m$ at $(t-c), (t-c)^2, \ldots, (t-c)^m$, for arbitrary $c \in k$, gives a upper triangular matrix with diagonal $1, 2!, \ldots, m!$.
(3) Evaluating the functionals $\mathtt{E}_\alpha \int, \mathtt{E}_\alpha \int t, \ldots, \mathtt{E}_\alpha \int t^{m-1}$, for $\alpha \neq 0$, at $1, (t-c), (t-c)^2, \ldots, (t-c)^{m-1}$ gives matrices A_m with entries $\int_0^\alpha x^j (x-c)^n\, dx$. For $\alpha = 1$ and $c = 0$, this is a *Hilbert matrix* H_m of order m. It is well-known that Hilbert matrices and all its submatrices are invertible. One can verify that $\det A_m$ is independent of c and is a nonzero multiple of $\det H_m$.

We can therefore apply the following lemma for linear functionals on arbitrary vector spaces to describe the image of a finite-rank operator L_d for a $d \in (\Phi)$ in terms of its normal form (14).

Lemma 19. Let V be a k-vector space and $\lambda_1, \ldots, \lambda_n \in V^*$ k-linear functionals. Then, the λ_i are k-linearly independent iff there exist $v_1, \ldots, v_n \in V$ such that:

$$\forall\, i, j = 1, \ldots, n, \quad \lambda_i(v_j) = \delta_{ij}.$$

Proposition 20. Let $d \in (\Phi)$ as in (14). Then, we have:

$$\operatorname{im} L_d = \sum_{\alpha \in \Phi} \sum_{i=0}^{l} k\, p_{\alpha,i} + \sum_{\alpha \in \Phi} \sum_{i=0}^{m} k\, q_{\alpha,i}.$$

Proof. The inclusion $\subseteq$ is obvious since $\mathtt{E}_\alpha\, \partial^i$ and $\mathtt{E}_\alpha \int t^i$ are functionals. Let $\mathtt{E}_\alpha\, \partial^i$ or $\mathtt{E}_\alpha \int t^i$ be a linear functional corresponding to a nonzero summand in (14). By Lemma 19 with $V = k[t]$, there exists a polynomial $p \in k[t]$ such that $(\mathtt{E}_\alpha\, \partial^i)(p) = 1$ (resp., $(\mathtt{E}_\alpha \int t^i)(p) = 1$) and $(\mathtt{E}_\beta\, \partial^j)(p) = 0$ (resp., $(\mathtt{E}_\beta \int t^j)(p) = 0$) for all other functionals corresponding to nonzero summands of (14). Then, we get $L_d(p) = d(p) = p_{\alpha,i}$ or $L_d(p) = d(p) = q_{\alpha,i}$, which proves the reverse inclusion.

In particular, by Proposition 20, we know that $L_d = 0$ implies $d = 0$ also for $d \in (\Phi)$. Hence χ is injective and Theorem 17 is proved.

To characterize $\operatorname{ann}_{\mathbb{I}_\Phi}(d.)$, we use the equivalences

$$d\, e = 0 \quad \Leftrightarrow\ L_{d\,e} = L_d \circ L_e = 0 \ \Leftrightarrow\ \operatorname{im} L_e \subseteq \ker L_d. \tag{15}$$

If d is Fredholm, i.e., $d \in \mathbb{I} \setminus (\Phi)$, then $\ker L_d$ is a finite-dimensional k-vector space, and thus, e has to be finite-rank. Thus, we have to compute polynomial solutions of the Fredholm operator d, i.e., $\ker L_d$, and then find generators for all the e's satisfying $\operatorname{im} L_e \subseteq \ker L_d$.

Theorem 21. Let $\Phi \subset k$. Let $d \in \mathbb{I}_\Phi$ be Fredholm with $\ker L_d = \sum_{i=1}^n k\, r_i$, where $r_i \in k[t]$. Then, we have:

$$\operatorname{ann}_{\mathbb{I}}(d.) = \sum_{\alpha\in\Phi}\sum_{i=1}^{n} (r_i\, \mathtt{E}_\alpha)\, \mathbb{I}_\Phi.$$

If Φ is finite (i.e., only finitely many evaluations points), then $\operatorname{ann}_{\mathbb{I}}(d.)$ is a finitely generated right $\mathbb{I}$-module.

Proof. Since $\operatorname{im} L_{r_i\,\mathtt{E}_\alpha} = k\, r_i \subseteq \ker L_d$, the inclusion $\supseteq$ follows by (15). Conversely, let $e \in \mathbb{I}_\Phi$ as in (14) with $d\, e = 0$. Then, by (15) and Proposition 20, we have

$$\operatorname{im} L_e = \sum_{\alpha\in\Phi}\sum_{i=0}^{l} k\, p_{\alpha,i} + \sum_{\alpha\in\Phi}\sum_{i=0}^{m} k\, q_{\alpha,i} \subseteq \ker L_d = \sum_{i=1}^{n} k\, r_i.$$

Hence, every nonzero $p_{\alpha,i}$ and $q_{\alpha,i}$ can be written as a k-linear combination of the r_i's. The reverse inclusion then follows by post-multiplying the generators $r_i\, \mathtt{E}_\alpha$ with suitable ∂^i or $\int t^i$.

The computation of the left annihilator $\operatorname{ann}_{\mathbb{I}}(.d)$ (e.g., for initial value problems) can be solved by computing the right annihilator $\operatorname{ann}_{\mathbb{I}}(\theta(d).)$, where θ is defined by (10), and then apply θ to each generator of $\operatorname{ann}_{\mathbb{I}}(\theta(d).)$.

All necessary steps for computing right and left annihilators have been implemented based on the Maple package *IntDiffOp* [13] for ID operators and boundary problems.

Example 22. Let us compute the compatibility conditions of (6). Note $\operatorname{rsym}(\theta(d)) = (0, n^2 - 3\,n + 2)$, where $\theta(d) = (t^2+t-3/5)\,\partial^2 - (2\,t+1)\,\partial + 2$. The largest nonnegative integer root of q is 2. With this bound N for Proposition 13, we get for the kernel $\ker L_{\theta(d)} = k\,(t^2+3/5) + k\,(t+1/2)$. By Theorem 21, $\operatorname{ann}_{\mathbb{I}}(\theta(d).) = ((t^2+3/5)\,\mathtt{E})\,\mathbb{I} + ((t+1/2)\,\mathtt{E})\,\mathbb{I}$. Computing the involution of these generators yield the left annihilator $\operatorname{ann}_{\mathbb{I}}(.d) = \mathbb{I}\,(2\,\mathtt{E}\,\partial^2 + 3/5\,\mathtt{E}) + \mathbb{I}\,(\mathtt{E}\,\partial + 1/2\,\mathtt{E})$ for (7), which correspond to the compatibility conditions:

$$2\,\ddot{u}(0) + 3/5\,u(0) = 0, \quad \dot{u}(0) + 1/2\,u(0) = 0.$$

REFERENCES

[1] V. V. Bavula. The algebra of integro-differential operators on an affine line and its modules. *J. Pure Appl. Algebra* 217(3):495–529, 2013.

[2] V. V. Bavula. An analogue of the Conjecture of Dixmier is true for the algebra of polynomial integro-differential operators. *J. Algebra* 372:237–250, 2012.

[3] V. V. Bavula. The algebra of integro-differential operators on a polynomial algebra. *J. Lond. Math. Soc. (2)*, 83(2):517–543, 2011.

[4] V. V. Bavula. The algebra of polynomial integro-differential operators is a holonomic bimodule over the subalgebra of polynomial differential operators. 2011. `http://arxiv.org/abs/arXiv:1011.3009`.

[5] A. Bostan, T. Cluzeau, and B. Salvy. Fast algorithms for polynomial solutions of linear differential equations. In *Proceedings of ISSAC 2005*, pages 45–52. ACM, 2005.

[6] F. Chyzak, A. Quadrat, and D. Robertz. Effective algorithms for parametrizing linear control systems over Ore algebras. *Appl. Algebra Engrg. Comm. Comput.*, 16(5):319–376, 2005.

[7] F. Chyzak, A. Quadrat, and D. Robertz. OreModules: A symbolic package for the study of multidimensional linear systems. In *Applications of time delay systems*, volume 352 of *LNCIS*, pages 233–264. Springer, Berlin, 2007.

[8] S. C. Coutinho. *A primer of algebraic D-modules*, volume 33. Cambridge University Press, 1995.

[9] M. Fliess. Some basic structural properties of generalized linear systems. *Systems Control Lett.*, 15(5): 391–396, 1990.

[10] I. Gohberg and M. A. Kaashoek. Time varying linear systems with boundary conditions and integral operators. I. The transfer operator and its properties. *Integral Equations Operator Theory*, 7(3):325–391, 1984.

[11] I. Gohberg, M. A. Kaashoek, and L. Lerer. Minimality and irreducibility of time-invariant linear boundary value systems. *Internat. J. Control*, 44(2):363–379, 1986.

[12] N. Jacobson. Some remarks on one-sided inverses. *Proc. Amer. Math. Soc.*, 1:352–355, 1950.

[13] A. Korporal, G. Regensburger, and M. Rosenkranz. Regular and singular boundary problems in Maple. In *Proceedings of CASC 2011*, volume 6885 of *LNCS*, pages 280–293, 2011. Springer, `http://www.risc.jku.at/people/akorpora/index.html`.

[14] T. Y. Lam. *A First Course in Noncommutative Rings*. Springer-Verlag, New York, 1991.

[15] V. Levandovskyy and E. Zerz. Algebraic systems theory and computer algebraic methods for some classes of linear control systems. In *Proceedings of MTNS 2006*, Kyoto, Japan, 2006.

[16] U. Oberst. Multidimensional constant linear systems. *Acta Appl. Math.*, 20(1-2):1–175, 1990.

[17] J.-F. Pommaret and A. Quadrat. A functorial approach to the behaviour of multidimensional control systems. *Int. J. Appl. Math. Comput. Sci.*, 13(1):7–13, 2003.

[18] A. Quadrat. The fractional representation approach to synthesis problems: An algebraic analysis viewpoint. I. (Weakly) doubly coprime factorizations. *SIAM J. Control Optim.*, 42(1):266–299, 2003.

[19] A. Quadrat. An introduction to constructive algebraic analysis and its applications. Inria Research Report n. 7354, 2010, `http://hal.archives-ouvertes.fr/inria-00506104/fr/`.

[20] G. Regensburger, M. Rosenkranz, and J. Middeke. A skew polynomial approach to integro-differential operators. In *Proceedings of ISSAC 2009*, pages 287–294, New York, NY, USA, 2009. ACM.

[21] M. Rosenkranz and G. Regensburger. Solving and factoring boundary problems for linear ordinary differential equations in differential algebras. *J. Symbolic Comput.*, 43(8):515–544, 2008.

[22] M. Rosenkranz, G. Regensburger, L. Tec, and B. Buchberger. Symbolic analysis for boundary problems: From rewriting to parametrized Gröbner bases. In *Numerical and Symbolic Scientific Computing: Progress and Prospects*, pages 273–331. SpringerWienNew York, Vienna, 2012.

[23] J. Rotman. *An Introduction to Homological Algebra*. Springer, New York, second edition, 2009.

[24] J. Wood. Modules and behaviours in nD systems theory. *Multidim. Syst. Signal Process.*, 11(1/2):11–48, 2000.

Journal of Pure and Applied Algebra 218 (2014) 456–473

Contents lists available at ScienceDirect

Journal of Pure and Applied Algebra

journal homepage: www.elsevier.com/locate/jpaa

On integro-differential algebras

Li Guo [a,b,*], Georg Regensburger [c], Markus Rosenkranz [d]

[a] Department of Mathematics, Lanzhou University, Lanzhou, Ganzu 730000, China
[b] Department of Mathematics and Computer Science, Rutgers University, Newark, NJ 07102, United States
[c] Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, A-4040 Linz, Austria
[d] School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7NF, England, United Kingdom

ARTICLE INFO

Article history:
Received 25 January 2013
Received in revised form 7 March 2013
Available online 29 July 2013
Communicated by C.A. Weibel

MSC: 16W99; 12H05; 12H20; 47G20; 68W30; 34M15

ABSTRACT

The concept of integro-differential algebra has been introduced recently in the study of boundary problems of differential equations. We generalize this concept to that of integro-differential algebra with a weight, in analogy to the differential Rota–Baxter algebra. We construct free commutative integro-differential algebras with weight generated by a differential algebra. This gives in particular an explicit construction of the integro-differential algebra on one generator. Properties of the free objects are studied.

1. Introduction

1.1. Motivation and goal

Differential algebra [28,33] is the study of differentiation and nonlinear differential equations by purely algebraic means, without using an underlying topology. It has been largely successful in many important areas like uncoupling of nonlinear systems, classification of singular components, and detection of hidden equations. There are various implementations that offer the main algorithms needed for such tasks, for instance the `DifferentialAlgebra` package in the Maple™ system [10].

In view of applications, there is one crucial component that does not fit well in differential algebra—the treatment of initial or boundary conditions. The problem is that the elements of a differential algebra or field are abstractions that cannot be evaluated at a specific point. For bridging this gap (first in a specific context of two-point boundary problems), a new framework was set up in [34] with the following features:

- Differential algebras are enhanced by two evaluations (multiplicative functionals to the ground field) and two integral operators (Rota–Baxter operators), leading to the notion of analytic algebra.
- The usual ring of differential operators is generalized to a ring of integro-differential operators.
- Boundary problems are formulated in terms of the operator ring (differential equations as usual, boundary conditions in terms of the evaluations).
- The Green's operator of a boundary problem is computed as an element of the operator ring.

The algebraic framework of boundary problems was subsequently refined and extended by a multiplicative structure with results on the corresponding factorizations along a given factorization of the differential operator [35,38]. The factorization approach to boundary problems was applied in [2,3] to find closed-form and asymptotic expressions for ruin probabilities and associated quantities in risk theory.

* Corresponding author at: Department of Mathematics, Lanzhou University, Lanzhou, Ganzu 730000, China.
E-mail addresses: liguo@rutgers.edu (L. Guo), georg.regensburger@oeaw.ac.at (G. Regensburger), M.Rosenkranz@kent.ac.uk (M. Rosenkranz).

0022-4049/$ – see front matter
http://dx.doi.org/10.1016/j.jpaa.2013.06.015

Moreover, it was realized that the algebraic theory of boundary problems is intimately related to the theory of Rota–Baxter algebras, which can be regarded as an algebraic study of both the integral and summation operators, even though it originated from the probability study of G. Baxter [7] in 1960. Rota–Baxter algebras have found extensive applications in mathematics and physics, including quantum field theory and the classical Yang–Baxter equation [4,14,15,18,19,25]. In a nutshell, the relation with Rota–Baxter algebras is this: In the differential algebra $C^\infty(\mathbb{R})$, every point evaluation ϕ gives rise to a unique Rota–Baxter operator $(1-\phi)\circ\int$, where $\int$ is any fixed integral operator, say $f \mapsto \int_0^x f(\xi)\,d\xi$. See also Theorem 2.5 below for a more general relation between evaluations and integral operators. We refer to [5,6] for an extensive study on algebraic properties of integro-differential operators with polynomial coefficients and a single evaluation (corresponding to initial value problems).

The algebraic approach to boundary problems is currently developed for linear ordinary differential equations although some effort is under way to cover certain classes of linear partial differential equations [37]. Various parts of the theory have been implemented, first as external Mathematica®-Theorema reasoner [34], then as internal Theorema code [37,38], and recently in a Maple™ package with new features for singular boundary problems [29].

1.2. Main results and outline of the paper

Our main purpose in this paper is to explicitly construct free objects in the category of λ-integro-differential algebras, which is at the heart of the algebraic framework of boundary problems described above. The existence of such free objects is known from universal algebra via equivalence classes of terms modulo the identities they satisfy [9,12,30] and from category theory via adjoint functors and monads; see [31, Chapter VI] and the references therein. But to construct free objects explicitly in terms of normal forms is often a non-trivial task. In the case of λ-integro-differential algebras, we make use of the construction of free objects in a structure closely related to the λ-integro-differential algebra, namely the differential Rota–Baxter algebra. A Rota–Baxter algebra is an algebraic abstraction of a reformulation of the integral by parts formula where only the integral operator appears. Free commutative Rota–Baxter algebras were obtained in [21,22] in terms of shuffles and the more general mixable shuffles of tensor powers.

More recently the concept of a differential Rota–Baxter algebra was introduced [23] by putting a differential operator and a Rota–Baxter operator of the same weight together such that one is the one side inverse of the other as in the Fundamental Theorem of Calculus. One advantage of this relatively independent combination of the two operators in a differential Rota–Baxter algebra is that the free objects can be constructed quite easily by building the free Rota–Baxter algebra on top of the free differential algebra. Since the axiom of an integro-differential algebra requires more intertwined relationship between the differential and Rota–Baxter operators, a free integro-differential algebra is a quotient of a free differential Rota–Baxter algebra. With this as the starting point of our construction of free integro-differential algebras, our strategy is to find an explicitly defined linear basis for this quotient from the known basis of the free differential Rota–Baxter algebra by tensor powers. For this purpose we use regular differential algebras as our basic building block for the tensor powers.

In Section 2, we first introduce the concept of an integro-differential algebra of weight λ and study their various characterizations, especially those in connection with differential Rota–Baxter algebras. In Section 3, we start with recalling free commutative Rota–Baxter algebras of weight λ and then free commutative differential Rota–Baxter algebras of weight λ and derive the existence of free commutative integro-differential algebras. The explicit construction of free objects in the category of λ-integro-differential algebras is carried out in Section 4 (Theorem 4.6) with a preparation on regular differential algebras and a detailed discussion on the regularity of the differential algebras of differential polynomials and rational functions.

2. Integro-differential algebras of weight λ

We first introduce the concepts and basic properties related to λ-integro-differential algebras.

2.1. Definitions and preliminary examples

We recall the concepts of a derivation with weight, a Rota–Baxter operator with weight and a differential Rota–Baxter algebra with weight, before introducing our definition of an integro-differential algebra with weight.

Definition 2.1. Let $\mathbf{k}$ be a unitary commutative ring. Let $\lambda \in \mathbf{k}$ be fixed.

(a) A **differential k-algebra of weight** λ (also called a λ**-differential k-algebra**) is a unitary associative $\mathbf{k}$-algebra R together with a linear operator $d\colon R \to R$ such that

$$d(xy) = d(x)y + xd(y) + \lambda d(x)d(y) \quad \text{for all } x, y \in R, \tag{1}$$

and

$$d(1) = 0. \tag{2}$$

Such an operator is called a **derivation of weight** λ or a λ-**derivation**.

(b) A **Rota–Baxter k-algebra of weight** λ is an associative $\mathbf{k}$-algebra R together with a linear operator $P: R \to R$ such that

$$P(x)P(y) = P(xP(y)) + P(P(x)y) + \lambda P(xy) \quad \text{for all } x, y \in R. \tag{3}$$

Such an operator is called a **Rota–Baxter operator of weight** λ or a λ-**Rota–Baxter operator**.

(c) A **differential Rota–Baxter k-algebra of weight** λ (also called a λ-**differential Rota–Baxter k-algebra**) is a differential $\mathbf{k}$-algebra (R, d) of weight λ and a Rota–Baxter operator P of weight λ such that

$$d \circ P = \mathrm{id}_R.$$

(d) An **integro-differential k-algebra of weight** λ (also called a λ-**integro-differential k-algebra**) is a differential $\mathbf{k}$-algebra (R, D) of weight λ with a linear operator $\Pi: R \to R$ such that

$$D \circ \Pi = \mathrm{id}_R \tag{4}$$

and

$$\Pi(D(x))\Pi(D(y)) = \Pi(D(x))y + x\Pi(D(y)) - \Pi(D(xy)) \quad \text{for all } x, y \in R. \tag{5}$$

When there is no danger of confusion, we will suppress λ and $\mathbf{k}$ from the notations. We will also denote the set of non-negative integers by $\mathbb{N}$.

Note that we require that a derivation d satisfies $d(1) = 0$. This follows from Eq. (1) automatically when $\lambda = 0$, but is a non-trivial restriction when $\lambda \neq 0$. In the next section, we give equivalent characterizations of the **hybrid Rota–Baxter axiom** (5) and discuss its relation to the **Rota–Baxter axiom** (3) as well as consequences of the **section axiom** (4). Note that the hybrid Rota–Baxter axiom does not contain a term with the weight λ.

We next give some simple examples of differential Rota–Baxter algebras and integro-differential algebras. As we shall see below (Lemma 2.3), the latter are a special case of the former. Further examples will be given in later sections. In particular, the algebras of λ-Hurwitz series are integro-differential algebras (Proposition 3.2). By Theorem 4.6, every regular differential algebra naturally gives rise to the corresponding free integro-differential algebra.

Example 2.2. (a) By the First Fundamental Theorem of Calculus

$$\frac{d}{dx}\Big(\int_a^x f(t)dt\Big) = f(x)$$

and the conventional integration-by-parts formula

$$\int_a^x f(t)g'(t)dt = f(t)g(t) - f(a)g(a) - \int_a^x f'(t)g(t)dt, \tag{6}$$

$(C^\infty(\mathbb{R}), d/dx, \int_a^x)$ is an integro-differential algebra of weight 0. As we shall see later in Theorem 2.5, integration by parts is in fact equivalent to the hybrid Rota–Baxter axiom (5).

(b) The following example from [23] of a differential Rota–Baxter algebra is also an integro-differential algebra. Let $\lambda \in \mathbb{R}$, $\lambda \neq 0$. Let $R = C^\infty(\mathbb{R})$ denote the $\mathbb{R}$-algebra of smooth functions $f: \mathbb{R} \to \mathbb{R}$, and consider the usual "difference quotient" operator D_λ on R defined by

$$(D_\lambda(f))(x) = (f(x+\lambda) - f(x))/\lambda. \tag{7}$$

Then D_λ is a λ-derivation on R. When $\lambda = 1$, we obtain the usual difference operator on functions. Further, the usual derivation is $D_0 := \lim_{\lambda \to 0} D_\lambda$. Now let R be an $\mathbb{R}$-subalgebra of $C^\infty(\mathbb{R})$ that is closed under the operators

$$\Pi_0(f)(x) = -\int_x^\infty f(t)dt, \qquad \Pi_\lambda(f)(x) = -\lambda \sum_{n \geq 0} f(x + n\lambda).$$

For example, R can be taken to be the $\mathbb{R}$-subalgebra generated by e^{-x}: $R = \sum_{k \geq 1} \mathbb{R}e^{-kx}$. Then Π_λ is a Rota–Baxter operator of weight λ and, for the D_λ in Eq. (7),

$$D_\lambda \circ \Pi_\lambda = \mathrm{id}_R \quad \text{for all } x, y \in R, 0 \neq \lambda \in \mathbb{R},$$

reducing to the fundamental theorem $D_0 \circ \Pi_0 = \mathrm{id}_R$ when λ goes to 0. We note the close relations of $(R, D_\lambda, \Pi_\lambda)$ to the time scale calculus [1] and the quantum calculus [27].

The fact that $(R, D_\lambda, \Pi_\lambda)$ is actually an integro-differential algebra follows from Theorem 2.5(g) since the kernel of D_λ is just the constant functions (in the case $\lambda \neq 0$ one uses that $R = \sum_{k \geq 1} \mathbb{R}e^{-kx}$ does not contain periodic functions).

(c) Here is one example of a differential Rota–Baxter algebra that is not an integro-differential algebra [35, Ex. 3]. Let $\mathbf{k}$ be a field of characteristic zero, $A = \mathbf{k}[y]/(y^4)$, and $(A[x], d)$, where d is the usual derivation with $d(x^k) = k\,x^{k-1}$. We define a $\mathbf{k}$-linear map P on $A[x]$ by

$$P(f) = \Pi(f) + f(0,0)\,y^2,$$

where Π is the usual integral with $\Pi(x^k) = x^{k+1}/(k+1)$. Since the second term vanishes under d, we see immediately that $d \circ P = \mathrm{id}_{A[x]}$. For verifying the Rota–Baxter axiom (3) with weight zero, we compute

$$\begin{aligned} P(f)P(g) &= \Pi(f)\Pi(g) + g(0,0)\,y^2\,\Pi(f) + f(0,0)\,y^2\Pi(g) + f(0,0)g(0,0)\,y^4,\\ P(fP(g)) &= \Pi(f\,(\Pi(g) + g(0,0)\,y^2)) = \Pi(f\,\Pi(g)) + g(0,0)\,y^2\,\Pi(f),\\ P(P(f)g) &= \Pi((\Pi(f) + f(0,0)\,y^2)\,g) = \Pi(\Pi(f)g) + f(0,0)\,y^2\,\Pi(g). \end{aligned}$$

Since $y^4 \equiv 0$ and the usual integral Π fulfills the Rota–Baxter axiom (3), this implies immediately that P does also. However, it does not fulfill the hybrid Rota–Baxter (5) since for example

$$P(d(x))P(d(y)) = P(1)P(0) = 0$$

but we obtain

$$P(d(x))y + xP(d(y)) - P(d(xy)) = P(1)y + xP(0) - P(y) = (x + y^2)y - xy = y^3.$$

for the right-hand side.

2.2. Basic properties of integro-differential algebras with weight

We first show that an integro-differential algebra with weight is a differential Rota–Baxter algebra of the same weight. We then give several equivalent conditions for integro-differential algebras.

Lemma 2.3. *Let (R, D) be a differential algebra of weight λ with a linear operator $\Pi : R \to R$ such that $D \circ \Pi = \mathrm{id}_R$. Denote $J = \Pi \circ D$.*

(a) *The triple (R, D, Π) is a differential Rota–Baxter algebra of weight λ if and only if*

$$\Pi(x)\Pi(y) = J(\Pi(x)\Pi(y)) \quad \text{for all } x, y \in R, \tag{8}$$

and if and only if

$$J(x)J(y) = J(J(x)J(y)) \quad \text{for all } x, y \in R. \tag{9}$$

(b) *Every integro-differential algebra is a differential Rota–Baxter algebra.*

Note that Eq. (8) does not contain a term with λ. Also note Eq. (9) involves only the initialization J and shows in particular that $\mathrm{im}\,J$ is a subalgebra.

Proof. (a) Using Eq. (1), we see that

$$D(\Pi(x)\Pi(y)) = x\Pi(y) + \Pi(x)y + \lambda xy.$$

Hence the Rota–Baxter axiom

$$\Pi(x)\Pi(y) = \Pi(x\Pi(y)) + \Pi(\Pi(x)y) + \lambda\Pi(xy)$$

is equivalent to Eq. (8). Moreover, substituting $D(x)$ for x and $D(y)$ for y in Eq. (8), we get the identity (9). Since D is onto by $D \circ \Pi = \mathrm{id}_R$, we also obtain Eq. (8) from Eq. (9).
(b) Since $J \circ \Pi = \Pi \circ (D \circ \Pi) = \Pi \circ \mathrm{id}_R = \Pi$, we obtain Eq. (8) from the hybrid Rota–Baxter axiom (5) by substituting $\Pi(x)$ for x and $\Pi(y)$ for y. □

We now give several equivalent conditions for an integro-differential algebra by starting with a result on complementary projectors on algebras.

Lemma 2.4. *Let E and J be projectors on a unitary $\mathbf{k}$-algebra R such that $E + J = \mathrm{id}_R$. Then the following statements are equivalent:*

(a) *E is an algebra homomorphism,*
(b) *J is a derivation of weight -1,*
(c) *$\ker E = \mathrm{im}\,J$ is an ideal and $\mathrm{im}\,E = \ker J$ is a unitary subalgebra.*

Proof. ((a) ⇔ (b)) It can be checked directly that $E(xy) = E(x)E(y)$ if and only if $J(xy) = J(x)y + xJ(y) - J(x)J(y)$. Further it follows from $E + J = \mathrm{id}_R$ that $E(1) = 1$ if and only if $J(1) = 0$.
((a) ⇒ (c)) is clear once we see that the assumption of the lemma implies $\ker E = \mathrm{im}\,J$ and $\mathrm{im}\,E = \ker J$.
((c) ⇒ (a)) Let $x, y \in R$. Since $R = \mathrm{im}\,E \oplus \ker E$, we have $x = x_1 + x_2$ and $y = y_1 + y_2$ with $x_1 = E(x)$, $y_1 = E(y) \in \mathrm{im}\,E$ and $x_2, y_2 \in \ker E$. Then $E(x_1y_1) = x_1y_1$ since $\mathrm{im}\,E$ is by assumption a subalgebra. Thus

$$E(xy) = E(x_1y_1) + E(x_1y_2) + E(x_2y_1) + E(x_2y_2) = x_1y_1 = E(x)E(y),$$

where the last three summands vanish assuming that $\ker E$ is an ideal. Moreover, $1 \in \mathrm{im}\,E$ implies $E(1) = 1$. □

We have the following characterizations of integro-differential algebras.

Theorem 2.5. *Let (R, D) be a differential algebra of weight λ with a linear operator Π on R such that $D \circ \Pi = \mathrm{id}_R$. Denote $J = \Pi \circ D$, called the **initialization**, and $E = \mathrm{id}_R - J$, called the **evaluation**. Then the following statements are equivalent:*

(a) *(R, D, Π) is an integro-differential algebra;*
(b) *$E(xy) = E(x)E(y)$ for all $x, y \in R$;*
(c) *$\ker E = \mathrm{im}\, J$ is an ideal;*
(d) *$J(xJ(y)) = xJ(y)$ and $J(J(x)y) = J(x)y$ for all $x, y \in R$;*
(e) *$J(x\Pi(y)) = x\Pi(y)$ and $J(\Pi(x)y) = \Pi(x)y$ for all $x, y \in R$;*
(f) *$x\Pi(y) = \Pi(D(x)\Pi(y)) + \Pi(xy) + \lambda\Pi(D(x)y)$ and $\Pi(x)y = \Pi(\Pi(x)D(y)) + \Pi(xy) + \lambda\Pi(xD(y))$ for all $x, y \in R$;*
(g) *(R, D, Π) is a differential Rota–Baxter algebra and $\Pi(E(x)y) = E(x)\Pi(y)$ and $\Pi(xE(y)) = \Pi(x)E(y)$ for all $x, y \in R$;*
(h) *(R, D, Π) is a differential Rota–Baxter algebra and $J(E(x)J(y)) = E(x)J(y)$ and $J(J(x)E(y)) = J(x)E(y)$ for all $x, y \in R$.*

Remark 2.6. (I) Items (d) and (e) can be regarded as the invariance formulation of the hybrid Rota–Baxter axiom.
(II) Item (f) can be seen as a "weighted" noncommutative version of integration by parts: One obtains it in case of weight zero by substituting $\int g$ for g in the usual formula (6). This motivates also the name integro-differential algebra. Clearly, in the commutative case the respective left and right versions are equivalent.
(III) Since $\mathrm{im}\, E = \ker D$, the identities in Items (g) and (h) can be interpreted as left/right linearity of respectively Π and J over the constants of the derivation D, restricted to $\mathrm{im}\, J$ in the case of (h). Note again that (g) and (h) do not contain a term with λ.

Proof. We first note that under the assumption, we have $J^2 = \Pi \circ (D \circ \Pi) \circ D = \Pi \circ \mathrm{id}_R \circ D = J$ and so the initialization J and evaluation E are projectors. Therefore

$$\ker D = \ker J = \mathrm{im}\, E \quad \text{and} \quad \mathrm{im}\, \Pi = \mathrm{im}\, J = \ker E, \tag{10}$$

and

$$R = \ker D \oplus \mathrm{im}\, \Pi$$

is a direct sum decomposition.
((a) $\Leftrightarrow$ (b)). It follows from Lemma 2.4 since the hybrid Rota–Baxter axiom (5) can be rewritten as

$$J(x)J(y) = J(x)y + xJ(y) - J(xy) \quad \text{for all } x, y \in R. \tag{11}$$

((b) $\Leftrightarrow$ (c)). It follows from Lemma 2.4, since $\ker D = \ker J = \mathrm{im}\, E$ is a unitary subalgebra by Eqs. (1) and (2).
((a) $\Rightarrow$ (e)). We obtain (e) by substituting in Eq. (11) respectively $\Pi(y)$ for y and $\Pi(x)$ for x.
((e) $\Leftrightarrow$ (d)). Substituting respectively $D(y)$ for y and $D(x)$ for x in (e) gives (d). Conversely, substituting respectively $\Pi(y)$ for y and $\Pi(x)$ for x in (d) gives (e).
((e) $\Leftrightarrow$ (f)). It follows from Eq. (1).
((a) $\Rightarrow$ (g)). By Lemma 2.3, (R, D, Π) is a differential Rota–Baxter algebra. Furthermore, using Eq. (1) and $D \circ E = 0$, we see that

$$D(E(x)\Pi(y)) = E(x)y \quad \text{and} \quad D(\Pi(x)E(y)) = xE(y)$$

and so

$$J(E(x)\Pi(y)) = \Pi(E(x)y) \quad \text{and} \quad J(\Pi(x)E(y)) = \Pi(xE(y)).$$

Since we have proved (e) from (a), we can respectively substitute $E(x)$ for x and $E(y)$ for y in (e) to get (g).
((g) $\Leftrightarrow$ (h)). Further, from $\Pi(E(x)y) = E(x)\Pi(y)$ we obtain

$$J(E(x)J(y)) = \Pi(D(E(x)J(y))) = \Pi(E(x)D(y)) = E(x)J(y),$$

Conversely, from $J(E(x)J(y)) = E(x)J(y)$ we obtain

$$\Pi(E(x)y) = \Pi(D(E(x)\Pi(y))) = J(E(x)\Pi(y)) = J(E(x)J(\Pi(y))) = E(x)\Pi(y)$$

using $\Pi = J \circ \Pi$ and $D(E(x)\Pi(y)) = E(x)y$. This proves the equivalence of the first equations in (g) and (h); the same proof gives the equivalence of the second equations.
((d) $\Rightarrow$ (c)). This is clear since the identities imply that $\mathrm{im}\, J$ is an ideal.
((h) $\Rightarrow$ (e)). Note that $J(E(x)J(y)) = E(x)J(y)$ gives

$$J(xJ(y)) - J(J(x)J(y)) = xJ(y) - J(x)J(y)$$

and hence $J(xJ(y)) = xJ(y)$ with the Rota–Baxter axiom in the form of Eq. (9). The identity $J(J(x)y) = J(x)y$ follows analogously. □

3. Free commutative integro-differential algebras

We first review the constructions of free commutative differential algebra with weight, free commutative Rota–Baxter algebras and free commutative differential Rota–Baxter algebras. These constructions are then applied in Section 3.3 to obtain free commutative integro-differential algebras and will be applied in Section 4 to give an explicit construction of free commutative integro-differential algebras.

3.1. Free and cofree differential algebras of weight λ

We recall the construction [23] of free commutative differential algebras of weight λ.

Theorem 3.1. *Let X be a set. Let*

$$\Delta(X) = X \times \mathbb{N} = \{x^{(n)} \,\big|\, x \in X, n \geq 0\}.$$

Let $\mathbf{k}\{X\}$ be the free commutative algebra $\mathbf{k}[\Delta X]$ on the set ΔX. Define $d_X : \mathbf{k}\{X\} \to \mathbf{k}\{X\}$ as follows. Let $w = u_1 \cdots u_k$, $u_i \in \Delta X$, $1 \leq i \leq k$, be a commutative word from the alphabet set $\Delta(X)$. If $k = 1$, so that $w = x^{(n)} \in \Delta(X)$, define $d_X(w) = x^{(n+1)}$. If $k > 1$, recursively define

$$d_X(w) = d_X(u_1)u_2 \cdots u_k + u_1 d_X(u_2 \cdots u_k) + \lambda d_X(u_1) d_X(u_2 \cdots u_k).$$

Further define $d_X(1) = 0$ and then extend d_X to $\mathbf{k}\{X\}$ by linearity. Then $(\mathbf{k}\{X\}, d_X)$ is the free commutative differential algebra of weight λ on the set X.

The use of $\mathbf{k}\{X\}$ for free commutative differential algebras of weight λ is consistent with the notation of the usual free commutative differential algebra (when $\lambda = 0$).

We also review the following construction from [23]. For any commutative $\mathbf{k}$-algebra A, let $A^{\mathbb{N}}$ denote the $\mathbf{k}$-module of all functions $f : \mathbb{N} \to A$. We define the λ-**Hurwitz product** on $A^{\mathbb{N}}$ by defining, for any $f, g \in A^{\mathbb{N}}$, $fg \in A^{\mathbb{N}}$ by

$$(fg)(n) = \sum_{k=0}^{n} \sum_{j=0}^{n-k} \binom{n}{k} \binom{n-k}{j} \lambda^k f(n-j) g(k+j).$$

We denote the $\mathbf{k}$-algebra $A^{\mathbb{N}}$ with this product by DA, and call it the $\mathbf{k}$-algebra of λ-**Hurwitz series over** A. It was shown in [23] that DA is a differential Rota–Baxter algebra of weight λ with the operators

$$D : DA \to DA, \qquad (D(f))(n) = f(n+1),\ n \geq 0,\ f \in DA,$$
$$\Pi : DA \to DA, \qquad (\Pi(f))(n) = f(n-1),\ n \geq 1,\ (\Pi(f))(0) = 0,\ f \in DA.$$

In fact, DA is the cofree differential algebra of weight λ on A. We similarly have

Proposition 3.2. *The triple (DA, D, Π) is an integro-differential algebra of weight λ.*

Proof. Since (DA, D, Π) is a differential Rota–Baxter algebra, we only need to show that $\Pi(E(x)y) = E(x)\Pi(y)$ for $x, y \in DA$ by Theorem 2.5. But this is clear since $\operatorname{im} E = \ker D = A$ and Π is A-linear. □

3.2. Free commutative Rota–Baxter algebras

We briefly recall the construction of free commutative Rota–Baxter algebras. Let A be a commutative $\mathbf{k}$-algebra. Define

$$Ш(A) = \bigoplus_{k \in \mathbb{N}} A^{\otimes(k+1)} = A \oplus A^{\otimes 2} \oplus \cdots, \tag{12}$$

where and hereafter all the tensor products are taken over $\mathbf{k}$ unless otherwise stated. Let $\mathfrak{a} = a_0 \otimes \cdots \otimes a_m \in A^{\otimes(m+1)}$ and $\mathfrak{b} = b_0 \otimes \cdots \otimes b_n \in A^{\otimes(n+1)}$. If $m = 0$ or $n = 0$, define

$$\mathfrak{a} \diamond \mathfrak{b} = \begin{cases} (a_0 b_0) \otimes b_1 \otimes \cdots \otimes b_n, & m = 0, n > 0, \\ (a_0 b_0) \otimes a_1 \otimes \cdots \otimes a_m, & m > 0, n = 0, \\ a_0 b_0, & m = n = 0. \end{cases} \tag{13}$$

If $m > 0$ and $n > 0$, inductively (on $m + n$) define

$$\begin{aligned} \mathfrak{a} \diamond \mathfrak{b} \ = (a_0 b_0) \otimes \Big(& (a_1 \otimes a_2 \otimes \cdots \otimes a_m) \diamond (1_A \otimes b_1 \otimes \cdots \otimes b_n) + (1_A \otimes a_1 \otimes \cdots \otimes a_m) \diamond (b_1 \otimes \cdots \otimes b_n) \\ & + \lambda\, (a_1 \otimes \cdots \otimes a_m) \diamond (b_1 \otimes \cdots \otimes b_n) \Big). \end{aligned} \tag{14}$$

Extending by additivity, we obtain a $\mathbf{k}$-bilinear map

$$\diamond : Ш(A) \times Ш(A) \to Ш(A).$$

Alternatively,

$$\mathfrak{a} \diamond \mathfrak{b} = (a_0 b_0) \otimes (\bar{\mathfrak{a}} \sqcup\!\sqcup_\lambda \bar{\mathfrak{b}}),$$

where $\bar{\mathfrak{a}} = a_1 \otimes \cdots \otimes a_m$, $\bar{\mathfrak{b}} = b_1 \otimes \cdots \otimes b_n$ and $\sqcup\!\sqcup_\lambda$ is the mixable shuffle (quasi-shuffle) product of weight λ [19,21,26], which specializes to the shuffle product $\sqcup\!\sqcup$ when $\lambda = 0$.

Define a $\mathbf{k}$-linear endomorphism P_A on $\sqcup\!\sqcup(A)$ by assigning

$$P_A(a_0 \otimes a_1 \otimes \cdots \otimes a_n) = 1_A \otimes a_0 \otimes a_1 \otimes \cdots \otimes a_n,$$

for all $a_0 \otimes a_1 \otimes \cdots \otimes a_n \in A^{\otimes(n+1)}$ and extending by additivity. Let $j_A : A \to \sqcup\!\sqcup(A)$ be the canonical inclusion map.

Theorem 3.3 ([21,22]). *The pair $(\sqcup\!\sqcup(A), P_A)$, together with the natural embedding $j_A : A \to \sqcup\!\sqcup(A)$, is a free commutative Rota–Baxter $\mathbf{k}$-algebra on A of weight λ. In other words, for any Rota–Baxter $\mathbf{k}$-algebra (R, P) and any $\mathbf{k}$-algebra map $\varphi : A \to R$, there exists a unique Rota–Baxter $\mathbf{k}$-algebra homomorphism $\tilde{\varphi} : (\sqcup\!\sqcup(A), P_A) \to (R, P)$ such that $\varphi = \tilde{\varphi} \circ j_A$ as $\mathbf{k}$-algebra homomorphisms.*

Since $\diamond$ is compatible with the multiplication in A, we will suppress the symbol $\diamond$ and simply denote xy for $x \diamond y$ in $\sqcup\!\sqcup(A)$, unless there is a danger of confusion.

Let (A, d) be a commutative differential $\mathbf{k}$-algebra of weight λ. Define an operator d_A on $\sqcup\!\sqcup(A)$ by assigning

$$d_A(a_0 \otimes a_1 \otimes \cdots \otimes a_n) = d(a_0) \otimes a_1 \otimes \cdots \otimes a_n + a_0 a_1 \otimes a_2 \otimes \cdots \otimes a_n + \lambda d(a_0) a_1 \otimes a_2 \otimes \cdots \otimes a_n \tag{15}$$

for $a_0 \otimes \cdots \otimes a_n \in A^{\otimes(n+1)}$ and then extending by $\mathbf{k}$-linearity. Here we use the convention that $d_A(a_0) = d(a_0)$ when $n = 0$.

Theorem 3.4 ([23]). *Let (A, d) be a commutative differential $\mathbf{k}$-algebra of weight λ. Let $j_A : A \to \sqcup\!\sqcup(A)$ be the $\mathbf{k}$-algebra embedding (in fact a morphism of differential $\mathbf{k}$-algebras of weight λ). The quadruple $(\sqcup\!\sqcup(A), d_A, P_A, j_A)$ is a free commutative differential Rota–Baxter $\mathbf{k}$-algebra of weight λ on (A, d).*

3.3. The existence of free commutative integro-differential algebras

The free objects in the category of commutative integro-differential algebras of weight λ are defined in a similar fashion as for the category of commutative differential Rota–Baxter algebras.

Definition 3.5. Let (A, d) be a λ-differential algebra over $\mathbf{k}$. A **free integro-differential algebra of weight λ on A** is an integro-differential algebra $(\mathrm{ID}(A), D_A, \Pi_A)$ of weight λ together with a differential algebra homomorphism $i_A : (A, d) \to (\mathrm{ID}(A), d_A)$ such that, for any integro-differential algebra (R, D, Π) of weight λ and a differential algebra homomorphism $f : (A, d) \to (R, D)$, there is a unique integro-differential algebra homomorphism $\bar{f} : \mathrm{ID}(A) \to R$ such that $\bar{f} \circ i_A = f$.

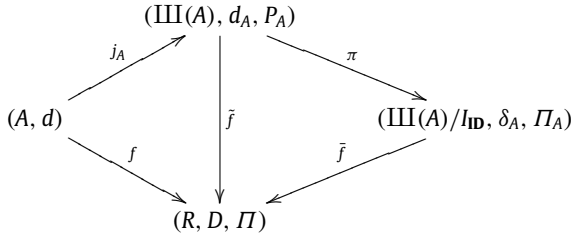
As in Theorem 3.4, let $(\sqcup\!\sqcup(A), d_A, P_A)$ be the free commutative differential Rota–Baxter algebra generated by the differential algebra (A, d). Then by Theorem 2.5, we have

Theorem 3.6. *Let (A, d) be a commutative differential $\mathbf{k}$-algebra of weight λ. Let $I_{\mathbf{ID}}$ be the differential Rota–Baxter ideal of $\sqcup\!\sqcup(A)$ generated by the set*

$$\{J\big(E(x)J(y)\big) - E(x)J(y) \big| x, y \in \sqcup\!\sqcup(A)\},$$

where J and E denote the projectors $P_A \circ d_A$ and $\mathrm{id}_A - P_A \circ d_A$, respectively. Let δ_A (resp. Π_A) denote d_A (resp. P_A) modulo $I_{\mathbf{ID}}$. Then the quotient differential Rota–Baxter algebra $(\sqcup\!\sqcup(A)/I_{\mathbf{ID}}, \delta_A, \Pi_A)$, together with the natural map $i_A : A \to \sqcup\!\sqcup(a) \to \sqcup\!\sqcup(A)/I_{\mathbf{ID}}$, is the free integro-differential algebra of weight λ on A.

Proof. Let a λ-integro-differential algebra (R, D, Π) be given. Then by Theorem 2.5, (R, D, Π) is also a λ-differential Rota–Baxter algebra. Thus by Theorem 3.4, there is a unique homomorphism $\tilde{f} : \sqcup\!\sqcup(A) \to R$ such that the left triangle of the following diagram commutes.



Since (R, D, Π) is a λ-integro-differential algebra, $\tilde{f}$ factors through $\sqcup\!\sqcup(A)/I_{\mathbf{ID}}$ and induces the λ-integro-differential algebra homomorphism $\bar{f}$ such that the right triangle commutes. Since $i_A = \pi \circ j_A$, we have $\bar{f} \circ i_A = f$ as needed.

Suppose $\bar{f}_1 : Ш(A)/I_{\mathbf{ID}} \to R$ is also a λ-integro-differential algebra homomorphism such that $\bar{f}_1 \circ i_A = f$. Define $\tilde{f}_1 = \bar{f}_1 \circ \pi$. Then $\tilde{f}_1 \circ j_A = f$. Thus by the universal property of $Ш(A)$, we have $\tilde{f}_1 = \tilde{f}$. Since π is surjective, we must have $\bar{f}_1 = \bar{f}$. This completes the proof. □

4. Construction of free commutative integro-differential algebras

As mentioned in Section 1, in integro-differential algebras the relation between d and Π is more intimate than in differential Rota–Baxter algebras. This makes the construction of their free objects more complex. Having ensured their existence in (Section 3.3), we introduce a vast class of differential algebras for which our construction applies (Section 4.1). Next we present the details of the construction and some basic properties (Section 4.2), leading on to the proof that it yields the desired free object (Section 4.3). The construction applies in particular to rings of differential polynomials $\mathbf{k}\{u\}$, yielding the free object over one generator, and to the ring of rational functions (Section 4.4).

4.1. Regular differential algebras

A free commutative integro-differential algebra can be regarded as a universal way of constructing an integro-differential algebra from a differential algebra. The easiest way of obtaining an integro-differential algebra from a differential algebra occurs when (A, d) already has an integral operator Π. This means in particular that $d \circ \Pi = \mathrm{id}_A$ so that the derivation d must be surjective. But often this will not be the case, for example when $A = \mathbf{k}\{u\}$ is the ring of differential polynomials (where u is clearly not in the image of d). But even if we cannot define an antiderivative (meaning a right inverse for d) on all of A, we may still be able to define one on $d(A)$ using an appropriate **quasi-antiderivative** Q. This means we require $d(Q(y)) = y$ for $y \in d(A)$ or equivalently $d(Q(d(x))) = d(x)$ for $x \in A$. For a general operator d, an operator Q with this property is called an inner inverse of d. It exists for many important differential algebras, in particular for differential polynomials (Proposition 4.10) and rational function (Proposition 4.12).

Before coming back to differential algebras, we recall some properties of generalized inverses for linear maps on $\mathbf{k}$-modules; for further details and references see [32, Section 8.1.].

Definition 4.1. Let $L : M \to N$ be a linear map between $\mathbf{k}$-modules.

(a) If a linear map $\bar{L} : N \to M$ satisfies $L \circ \bar{L} \circ L = L$, then $\bar{L}$ is called an **inner inverse** of L.
(b) If L has an inner inverse, then L is called **regular**.
(c) If a linear map $\bar{L} : N \to M$ satisfies $\bar{L} \circ L \circ \bar{L} = \bar{L}$, then $\bar{L}$ is called an **outer inverse** of L.
(d) If $\bar{L}$ is an inner inverse and outer inverse of L, then $\bar{L}$ is called a **quasi-inverse** or **generalized inverse** of L.

Proposition 4.2. *Let $L : M \to N$ be a linear map between $\mathbf{k}$-modules.*

(a) *If L has an inner inverse $\bar{L} : N \to M$, then $S = L \circ \bar{L} : N \to N$ is a projector onto $\operatorname{im} L$ and $E = \mathrm{id}_M - \bar{L} \circ L : M \to M$ is a projector onto $\ker L$.*
(b) *Given projectors $S : N \to N$ onto $\operatorname{im} L$ and $E : M \to M$ onto $\ker L$, there is a unique quasi-inverse $\bar{L}$ of L such that $\operatorname{im} \bar{L} = \ker E$ and $\ker \bar{L} = \ker S$. Thus a regular map has a quasi-inverse.*

Proof. (a) This statement is immediate.
(b) If L is regular, then by Item (a), there are submodules $\ker E \subseteq M$ and $\ker S \subseteq N$ such that

$$M = \ker L \oplus \ker E, \qquad N = \operatorname{im} L \oplus \ker S.$$

Thus L induces a bijection $L : \ker E \to \operatorname{im} L$. Define $\bar{L} : N \to M$ to be the inverse of this bijection on $\operatorname{im} L$ and to be zero on $\ker S$, then we check directly that $\bar{L}$ is a quasi-inverse of L and the unique one such that $\operatorname{im} \bar{L} = \ker E$ and $\ker \bar{L} = \ker S$. See also [32, Theorem 8.1.]. □

For a quasi-inverse $\bar{L}$ of L we note the direct sums

$$M = \operatorname{im} \bar{L} \oplus \ker L \quad \text{and} \quad N = \operatorname{im} L \oplus \ker \bar{L}.$$

Moreover, let

$$J = \mathrm{id}_M - E \quad \text{and} \quad T = \mathrm{id}_N - S,$$

then we have the relations

$$M_E := \operatorname{im} E = \ker L = \ker J, \qquad M_J := \operatorname{im} J = \operatorname{im} \bar{L} = \ker E$$
$$N_S := \operatorname{im} S = \operatorname{im} L = \ker T, \qquad N_T := \operatorname{im} T = \ker \bar{L} = \ker S$$

for the corresponding projectors.

The intuitive roles of the projectors E and J are similar as in Section 2.2, except that the "evaluation" E is not necessarily multiplicative and the image of the "initialization" J need not be an ideal. The projector S may be understood as extracting the solvable part of N, in the sense of solving $L(x) = y$ for x, as much as possible for a given $y \in N$.

Let us elaborate on this. Writing respectively $y_S = S(y)$ and $y_T = T(y)$ for the "solvable" and "transcendental" part of y, the equation $L(x) = y_S$ is clearly solved by $x^* = \bar{L}(y_S)$ while $L(x) = y_T$ is only solvable in the trivial case $y_T = 0$. So the identity $L(x^*) = y - T(y)$ may be understood in the sense that x^* solves $L(x) = y$ except for the transcendental part. We illustrate this in the following example.

Example 4.3. Consider the field $\mathbb{C}(x)$ of **complex rational functions** with its usual derivation d. We take d to be the linear map $L\colon M \to N$ where $M = N = \mathbb{C}(x)$. Any rational function can be represented by f/g with a monic denominator $g = (x-\alpha_1)^{n_1}\cdots(x-\alpha_k)^{n_k}$ having distinct roots $\alpha_i \in \mathbb{C}$. By partial fraction decomposition, it can be written uniquely as

$$r + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{\gamma_{ij}}{(x-\alpha_i)^j},$$

where $r \in \mathbb{C}[x]$ and $\gamma_{ij} \in \mathbb{C}$. Then for the domain $\mathbb{C}(x)$ of d, we have the decomposition

$$\mathbb{C}(x) = \ker d \oplus \mathbb{C}(x)_J$$

with $\ker d = \mathbb{C}$ and

$$\mathbb{C}(x)_J = \left\{ r + \sum_{i=1}^{k}\sum_{j=1}^{n_i} \frac{\gamma_{ij}}{(x-\alpha_i)^j} \;\middle|\; r \in x\,\mathbb{C}[x],\ \alpha_i \in \mathbb{C} \text{ distinct},\ \gamma_{ij} \in \mathbb{C} \right\}$$

as the initialized space. For the range $\mathbb{C}(x)$ of d, we have the decomposition

$$\mathbb{C}(x) = \operatorname{im} d \oplus \mathbb{C}(x)_T,$$

with

$$\operatorname{im} d = \left\{ r + \sum_{i=1}^{k}\sum_{j=2}^{n_i} \frac{\gamma_{ij}}{(x-\alpha_i)^j} \;\middle|\; r \in \mathbb{C}[x],\ \alpha_i \in \mathbb{C} \text{ distinct},\ \gamma_{ij} \in \mathbb{C} \right\}$$

and

$$\mathbb{C}(x)_T = \left\{ \sum_{i=1}^{k} \frac{\gamma_i}{x-\alpha_i} \;\middle|\; \alpha_i \in \mathbb{C} \text{ distinct},\ \gamma_i \in \mathbb{C} \right\}$$

as the transcendental space.

By Proposition 4.2 there exists a unique quasi-inverse $Q : \mathbb{C}(x) \to \mathbb{C}(x)$ of d corresponding to the above decompositions, which we can describe explicitly. On $\operatorname{im} d$ we define Q by setting $Q(x^k) = x^{k+1}/(k+1)$ for $k \geq 0$ and $Q(1/(x-\alpha)^j) = 1/(1-j)(x-\alpha)^{j-1}$ for $j > 1$, and we extend it by zero on $\mathbb{C}(x)_T$. Analytically speaking, the quasi-antiderivative Q acts as $\int_0^x$ on the polynomials and as $\int_{-\infty}^x$ on the solvable rational functions: Since $\mathbb{C}(x)$ is not an integro-differential algebra, it is not possible to use a single integral operator. The associated codomain projector $S = d \circ Q$ extracts the solvable part by filtering out the residues $1/(x-\alpha)$; their antiderivatives would need logarithms, which are not available in $\mathbb{C}(x)$. The domain projector $E = \mathrm{id}_{\mathbb{C}(x)} - Q \circ d$ is almost like evaluation at 0 but is not multiplicative according to Theorem 2.5 since $\mathbb{C}(x)_J$ cannot be an ideal of the field $\mathbb{C}(x)$. In fact, one checks immediately that $E(x \cdot 1/x) = E(1) = 1$ but $E(x) \cdot E(1/x) = 0 \cdot 0 = 0$.

See Proposition 4.12 for the case when d here is replaced by the difference operator or more generally the λ-difference quotient operator d_λ with $\lambda \neq 0$ (Example 2.2). We refer to [11] for details on effectively computing the above decomposition into solvable and transcendental part of rational functions in the context of symbolic integration algorithms. See also [13] for necessary and sufficient conditions for the existence of telescopers in the differential, difference, and q-difference case in terms of (generalizations of) residues.

We can now define what makes a differential algebra such as $A = \mathbf{k}\{u\}$ and $A = \mathbb{C}(x)$ adequate for the forthcoming construction of the free integro-differential algebra.

Definition 4.4. Let (A, d) be a differential algebra of weight λ with derivation $d\colon A \to A$.

(a) If $\lambda = 0$, then (A, d) is called **regular** if its derivation d is a regular map. Then a quasi-inverse of d is called a **quasi-antiderivative**.
(b) If $\lambda \neq 0$, then (A, d) is called **regular** if its derivation d is a regular map and the kernel of one of its quasi-inverses is a nonunitary $\mathbf{k}$-subalgebra of A. Such a quasi-inverse of d is called a **quasi-antiderivative**.

We observe that the class of regular differential algebras is fairly comprehensive in the zero weight case. It includes all differential algebras over a field $\mathbf{k}$ since in that case every subspace is complemented, so all $\mathbf{k}$-linear maps are regular. In

particular, all differential fields (viewed as differential algebras over their field of constants) are regular. The example $\mathbb{C}(x)$ is a case in point, but note that Example 4.3 provides an explicit quasi-antiderivative rather than plain existence.

The situation is more complex in the nonzero weight case due to the extra restriction on the derivation, which we need in our construction of free integro-differential algebras. If $\mathbf{k}$ is a field, the ring of differential polynomials $\mathbf{k}\{u\}$ is regular for any weight, and we will provide an explicit quasi-antiderivative that works also when $\mathbf{k}$ is a $\mathbb{Q}$-algebra but not a field (Proposition 4.10). Moreover, the field of complex rational functions $\mathbb{C}(x)$ with its usual difference operator is a regular differential ring of weight one, and this can be extended to arbitrary nonzero weight (Proposition 4.12).

4.2. Construction of $\mathrm{ID}(A)^*$

According to Theorem 3.6, the free integro-differential algebra $\mathrm{ID}(A)$ can be described by a suitable quotient. However, for studying this object effectively, a more explicit construction is preferable. We will achieve this, for a regular differential algebra A, by defining an integro-differential algebra $\mathrm{ID}(A)^*$, and by showing in the next subsection that it satisfies the relevant universal property. Hence we may take $\mathrm{ID}(A)^*$ to be $\mathrm{ID}(A)$.

4.2.1. Definition of $\mathrm{ID}(A)^*$ and the statement of Theorem 4.6

Let (A, d) be a regular differential algebra with a fixed quasi-antiderivative Q.

Denote

$$A_J = \operatorname{im} Q \quad \text{and} \quad A_T = \ker Q.$$

Then we have the direct sums

$$A = A_J \oplus \ker d \quad \text{and} \quad A = \operatorname{im} d \oplus A_T$$

with the corresponding projectors $E = \mathrm{id}_A - Q \circ d$ and $S = d \circ Q$, respectively. As before, we write $J = \mathrm{id}_A - E = Q \circ d$ and $T = \mathrm{id}_A - S$ for the complementary projectors. Furthermore, we use the notation $K := \ker d \supseteq \mathbf{k}$ in this subsection.

We give now an explicit construction of $\mathrm{ID}(A)^*$ via tensor products (all tensors are still over $\mathbf{k}$). First let

$$\text{Ш}_T(A) := \bigoplus_{k\geq 0} A \otimes A_T^{\otimes k} = A \oplus (A \otimes A_T) \oplus (A \otimes A_T^{\otimes 2}) + \cdots$$

be the $\mathbf{k}$-submodule of $\text{Ш}(A)$ in Eq. (12). Under our assumption that A_T is a subalgebra of A when $\lambda \neq 0$, $\text{Ш}_T(A)$ is clearly a $\mathbf{k}$-subalgebra of $\text{Ш}(A)$ under the multiplication in Eqs. (13) and (14). It is also closed under the derivation d_A defined in Eq. (15). Alternatively,

$$\text{Ш}_T(A) = A \otimes \text{Ш}^+(A_T)$$

is the tensor product algebra where $\text{Ш}^+(A_T) := \bigoplus_{n\geq 0} A_T^{\otimes n}$ is the mixable shuffle algebra [19,21,26] on the $\mathbf{k}$-algebra A_T. In the case $\lambda = 0$, this is the plain shuffle algebra, where it is sufficient for A_T to have the structure of a $\mathbf{k}$-module. So a pure tensor $\mathfrak{a}$ of $A \otimes \text{Ш}^+(A_T)$ is of the form

$$\mathfrak{a} = a \otimes \bar{\mathfrak{a}} \in A \otimes A_T^{\otimes n} \subseteq A^{\otimes(n+1)}. \tag{16}$$

We then define the **length** of $\mathfrak{a}$ to be $n + 1$.

Next let $\varepsilon : A \to A_\varepsilon$ be an isomorphism of K-algebras, where

$$A_\varepsilon := \{\varepsilon(a) \mid a \in A\}$$

denotes a replica of the K-algebra A, endowed with the zero derivation. We identify the image $\varepsilon(K) \subseteq A_\varepsilon$ with K so that $\varepsilon(c) = c$ for all $c \in K$. Finally let

$$\mathrm{ID}(A)^* := A_\varepsilon \otimes_K \text{Ш}_T(A) = A_\varepsilon \otimes_K A \otimes \text{Ш}^+(A_T) \tag{17}$$

denote the tensor product differential algebra of A_ε and $\text{Ш}_T(A)$, namely the tensor product algebra where the derivation (again denoted by d_A) is defined by the Leibniz rule.

4.2.2. Definition of Π_A

We will define a linear operator Π_A on $\mathrm{ID}(A)^*$. First require that Π_A is linear over A_ε. Thus we just need to define $\Pi_A(\mathfrak{a})$ for a pure tensor $\mathfrak{a}$ in $A \otimes \text{Ш}^+(A_T)$. We will accomplish this by induction on the length n of $\mathfrak{a}$. When $n = 1$, we have $\mathfrak{a} = a \in A$. Then we have

$$a = d(Q(a)) + T(a) \quad \text{with } T(a) \in A_T \tag{18}$$

and we define

$$\Pi_A(a) := Q(a) - \varepsilon(Q(a)) + 1 \otimes T(a). \tag{19}$$

Assume $\Pi_A(\mathfrak{a})$ has been defined for $\mathfrak{a}$ of length $n \geq 1$ and consider the case when $\mathfrak{a}$ has length $n+1$. Then $\mathfrak{a} = a \otimes \bar{\mathfrak{a}}$ where $a \in A$, $\bar{\mathfrak{a}} \in A_T^{\otimes n}$ and we define

$$\Pi_A(a \otimes \bar{\mathfrak{a}}) := Q(a) \otimes \bar{\mathfrak{a}} - \Pi_A(Q(a)\bar{\mathfrak{a}}) - \lambda\, \Pi_A(d(Q(a))\, \bar{\mathfrak{a}}) + 1 \otimes T(a) \otimes \bar{\mathfrak{a}}, \tag{20}$$

where the first and last terms are manifestly in $A \otimes \text{Ш}^+(A_T)$ while the middle terms are in $\mathrm{ID}(A)^*$ by the induction hypothesis. We write $E_A = \mathrm{id}_{\mathrm{ID}(A)^*} - \Pi_A \circ d_A$ for what will turn out to be the "evaluation" corresponding to Π_A (see the discussion before Example 4.3).

We display the following relationship between Π_A, P_A and ε for later application.

Lemma 4.5. (a) *For $a \in A$, we have $E_A(a) = \varepsilon(a)$.*
(b) *For $\bar{\mathfrak{a}} \in \text{Ш}^+(A_T)$, we have $\Pi_A(\bar{\mathfrak{a}}) = P_A(\bar{\mathfrak{a}}) = 1 \otimes \bar{\mathfrak{a}}$.*

Proof. (a) Using the direct sum $A = A_J \oplus \ker d$, we distinguish two cases. If $a \in \ker d = K$, then the left-hand side is $a - \Pi_A(d_A(a)) = a - \Pi_A(0) = a$; but the right-hand is a as well since $\varepsilon : A \to A_\varepsilon$ is a K-algebra homomorphism. Hence assume $a \in A_J = \mathrm{im} J$. In that case $a = J(a) = Q(d(a))$ and hence $T(d(a)) = d(a) - d(Q(d(a))) = 0$. So $\Pi_A(d_A(a)) = \Pi_A(d(a)) = a - \varepsilon(a)$ by Eq. (19).
(b) This is a special case of Eqs. (18) and (20) with $Q(a) = 0$ and $T(a) = a$ since $a \in A_T$. □

Theorem 4.6. *Let (A, d, Q) be a regular differential algebra of weight λ with quasi-antiderivative Q. Then the triple $(\mathrm{ID}(A)^*, d_A, \Pi_A)$, with the natural embedding*

$$i_A : A \to \mathrm{ID}(A)^* = A_\varepsilon \otimes_K A \otimes \text{Ш}^+(A_T)$$

to the second tensor factor, is the free commutative integro-differential algebra of weight λ generated by A.

The proof of Theorem 4.6 is given in Section 4.3.

Since $A_T \cong A/\mathrm{im}\, d$ as **k**-modules, for different choices of Q, the corresponding A_T are isomorphic as **k**-modules. Then for $\lambda = 0$ the mixable shuffle (i.e., shuffle) algebras $\text{Ш}^+(A_T)$ are isomorphic **k**-algebras since in that case the algebra structure of A_T is not used; see e.g. Section 2.1 of [24]. When $\lambda \neq 0$, for A_T from different choices of Q, they are still isomorphic as **k**-modules. But it is not clear that they are isomorphic as nonunitary **k**-algebras. Nevertheless, the free commutative integro-differential algebras derived by Theorem 4.6 are isomorphic due to the uniqueness of the free objects. See Remark 4.13 for further discussions.

The following is a preliminary discussion on subalgebras as direct sum factors.

Lemma 4.7. *Let T and S be projectors on a unitary **k**-algebra R such that $T + S = \mathrm{id}_R$. Then the following statements are equivalent:*

(a) $\mathrm{im}\, T = \ker S$ *is a subalgebra;*
(b) $T(T(x)T(y)) = T(x)T(y)$;
(c) $S(xy) = S(S(x)y + xS(y) - S(x)S(y))$.

Proof. ((a) ⇔ (b)) It is clear since T is a projector.
((a) ⇒ (c)) It follows from

$$S(T(x)T(y)) = S((x - S(x))(y - S(y)) = 0.$$

((c) ⇒ (a)) Clearly, the identity implies that $\ker S$ is a subalgebra. □

If $S = d \circ Q$ as above, we obtain from Lemma 4.7(c) an equivalent identity

$$Q(xy) = Q(d(Q(x))y + xd(Q(y)) - d(Q(x))d(Q(y)))$$

in terms of Q and d, since $Q \circ d \circ Q = Q$.

4.3. The proof of Theorem 4.6

We will verify that $(\mathrm{ID}(A)^*, d_A, \Pi_A)$ is an integro-differential algebra in Section 4.3.1 and verify its universal property in Section 4.3.2.

4.3.1. The integro-differential algebra structure on $\mathrm{ID}(A)^*$

Since d_A is clearly a derivation, by Theorem 2.5(b), we just need to check the two conditions

$$d_A \circ \Pi_A = \mathrm{id}_{\mathrm{ID}(A)^*}, \tag{21}$$
$$E_A(xy) = E_A(x)E_A(y), \quad x, y \in \mathrm{ID}(A)^*. \tag{22}$$

Since A_ε is in the kernel of d_A and in the ring of constants for Π_A, we just need to verify the equations for pure tensors $x = \mathfrak{a}, y = \mathfrak{b} \in A \otimes \text{Ш}^+(A_T)$.

We check Eq. (21) by showing $(d_A \circ \Pi_A)(\mathfrak{a}) = \mathfrak{a}$ for $\mathfrak{a} \in A \otimes \text{Ш}^+(A_T)$ by induction on the length $n \geq 1$ of $\mathfrak{a}$. When $n = 1$, we have $\mathfrak{a} = a \in A$ and obtain

$$d_A(\Pi_A(a)) = d_A(Q(a) - \varepsilon(Q(a)) + 1 \otimes T(a)) = d(Q(a)) + T(a) = a$$

by Eq. (18). Under the induction hypothesis, we consider $\mathfrak{a} = a \otimes \bar{\mathfrak{a}}$ with $\bar{\mathfrak{a}} \in A_T^{\otimes n}$, $n \geq 1$. Then we have

$$\begin{aligned} d_A(\Pi_A(a \otimes \bar{\mathfrak{a}})) &= d_A\big(Q(a) \otimes \bar{\mathfrak{a}} - \Pi_A(Q(a)\bar{\mathfrak{a}}) - \lambda\, \Pi_A(d(Q(a))\, \bar{\mathfrak{a}}) + 1 \otimes T(a) \otimes \bar{\mathfrak{a}}\big) \\ &= d(Q(a)) \otimes \bar{\mathfrak{a}} + Q(a)\bar{\mathfrak{a}} + \lambda\, d(Q(a))\bar{\mathfrak{a}} - Q(a)\bar{\mathfrak{a}} - \lambda\, d(Q(a))\bar{\mathfrak{a}} + T(a) \otimes \bar{\mathfrak{a}} \\ &= d(Q(a)) \otimes \bar{\mathfrak{a}} + T(a) \otimes \bar{\mathfrak{a}} \\ &= a \otimes \bar{\mathfrak{a}} \end{aligned}$$

by Eq. (18) again.

We next verify Eq. (22). If the length of both x and y are one, then x and y are in A. Then by Lemma 4.5(a), we have

$$E_A(xy) = \varepsilon(xy) = \varepsilon(x)\varepsilon(y) = E_A(x)E_A(y).$$

If at least one of x or y have length greater than one, then each pure tensor in the expansion of xy has length greater than one. Then the equation holds by the following lemma.

Lemma 4.8. *For any pure tensor $\mathfrak{a} = a \otimes \bar{\mathfrak{a}} \in A \otimes \text{Ш}^+(A_T)$ of length greater than one we have $E_A(\mathfrak{a}) = 0$.*

Remark 4.9. Combining Lemma 4.5(a) and Lemma 4.8 we have $\operatorname{im} E_A = A_\varepsilon$. Further, by Eq. (10), we have $\ker d_A = \operatorname{im} E_A = A_\varepsilon$.

Proof. For a given $\mathfrak{a} = a \otimes \bar{\mathfrak{a}}$ of length greater than one, we compute

$$\begin{aligned} E_A(a \otimes \bar{\mathfrak{a}}) &= a \otimes \bar{\mathfrak{a}} - \Pi_A(d_A(a \otimes \bar{\mathfrak{a}})) \quad \text{(by definition of } E_A\text{)} \\ &= a \otimes \bar{\mathfrak{a}} - \Pi_A(d(a) \otimes \bar{\mathfrak{a}}) - \Pi_A(a\bar{\mathfrak{a}}) - \Pi_A(\lambda d(a)\bar{\mathfrak{a}}) \quad \text{(by definition of } d_A\text{)} \\ &= a \otimes \bar{\mathfrak{a}} - Q(d(a)) \otimes \bar{\mathfrak{a}} + \Pi_A(Q(d(a))\bar{\mathfrak{a}}) + \lambda\, \Pi_A(d(Q(d(a)))\, \bar{\mathfrak{a}}) - 1 \otimes T(d(a)) \otimes \bar{\mathfrak{a}} \\ &\quad - \Pi_A(a\bar{\mathfrak{a}}) - \Pi_A(\lambda d(a)\bar{\mathfrak{a}}) \quad \text{(by definition of } \Pi_A\text{)} \\ &= a \otimes \bar{\mathfrak{a}} - Q(d(a)) \otimes \bar{\mathfrak{a}} + \Pi_A(Q(d(a))\bar{\mathfrak{a}}) - \Pi_A(a\bar{\mathfrak{a}}) \quad \text{(by } d \circ Q \circ d = d \text{ and } T(d(a)) = 0\text{)} \\ &= E(a) \otimes \bar{\mathfrak{a}} - \Pi_A(E(a)\bar{\mathfrak{a}}) \quad \text{(by definition of } E = \mathrm{id}_A - Q \circ d\text{)}. \end{aligned}$$

Since $E(A) = K \subseteq A_\varepsilon$ and Π_A is taken to be A_ε-linear, from Lemma 4.5(b), we obtain

$$E_A(a \otimes \bar{\mathfrak{a}}) = E(a)(1_A \otimes \bar{\mathfrak{a}} - \Pi_A(\bar{\mathfrak{a}})) = 0. \quad \square$$

4.3.2. The universal property

We now verify the universal property of $(\mathrm{ID}(A)^*, d_A, \Pi_A)$ as the free integro-differential algebra on (A, d): Let $i_A : A \to \mathrm{ID}(A)^*$ be the natural embedding of A into the second tensor factor of $\mathrm{ID}(A)^* = A_\varepsilon \otimes_K A \otimes \text{Ш}^+(A_T)$. Then for any integro-differential algebra (R, D, Π) and any differential algebra homomorphism $f : (A, d) \to (R, D)$, there is a unique integro-differential algebra homomorphism $\bar{f} : (\mathrm{ID}(A)^*, d_A, \Pi_A) \to (R, D, \Pi)$ such that $\bar{f} \circ i_A = f$.

The existence of $\bar{f}$: Let a differential algebra homomorphism $f : (A, d) \to (R, D)$ be given. Note that f is in fact a K-algebra homomorphism where the K-algebra structure on R is given by $f : K \to R$. Since (R, Π) is a commutative Rota–Baxter algebra, by the universal property of $\text{Ш}(A)$ as the free commutative Rota–Baxter algebra on the commutative algebra A, there is a homomorphism $\tilde{f} : (\text{Ш}(A), P_A) \to (R, \Pi)$ of commutative Rota–Baxter algebras such that $\tilde{f} \circ j_A = f$ where $j_A : A \to \text{Ш}(A)$ is the embedding into the first tensor factor. This means that $\tilde{f}$ is an A-algebra homomorphism and, in particular, a K-algebra homomorphism. Thus $\tilde{f}$ restricts to a K-algebra homomorphism

$$\tilde{f} : A \otimes \text{Ш}^+(A_T) \to R.$$

Further, f also gives a K-algebra homomorphism

$$f_\varepsilon : A_\varepsilon \to R,\ \varepsilon(a) \mapsto f(a) - \Pi(D(f(a))).$$

Thus we get an algebra homomorphism on the tensor product over K:

$$\bar{f} := f_\varepsilon \otimes_K \tilde{f} : A_\varepsilon \otimes_K (A \otimes \text{Ш}^+(A_T)) \to R$$

that extends $\tilde{f}$ and f_ε. Further, we have $\bar{f} \circ j_A = f$.

It remains to check the equations

$$\bar{f} \circ d_A = D \circ \bar{f}, \qquad \bar{f} \circ \Pi_A = \Pi \circ \bar{f}. \tag{23}$$

Since A_ε is in the kernel of d_A and in the ring of constants of Π_A, we only need to verify the equations when restricted to $A \otimes \text{Ш}^+(A_T)$.

Fix $a \otimes \bar{\mathfrak{a}} = a(1 \otimes \bar{\mathfrak{a}}) \in A \otimes \text{Ш}^+(A_T)$. By Lemma 4.5(b), we have

$$\Pi(\bar{f}(\bar{\mathfrak{a}})) = \Pi(\tilde{f}(\bar{\mathfrak{a}})) = \tilde{f}(\Pi_A(\bar{\mathfrak{a}})) = \bar{f}(1 \otimes \bar{\mathfrak{a}}).$$

Thus we obtain

$$\begin{aligned}
\bar{f}(d_A(a \otimes \bar{\mathfrak{a}})) &= \bar{f}(d(a) \otimes \bar{\mathfrak{a}}) + \bar{f}(a\bar{\mathfrak{a}}) + \bar{f}(\lambda d(a)\bar{\mathfrak{a}}) \\
&= f(d(a))\bar{f}(1 \otimes \bar{\mathfrak{a}}) + f(a)\bar{f}(\bar{\mathfrak{a}}) + \lambda f(d(a))\bar{f}(\bar{\mathfrak{a}}) \\
&= D(f(a))\bar{f}(1 \otimes \bar{\mathfrak{a}}) + f(a)D(\Pi(\bar{f}(\bar{\mathfrak{a}}))) + \lambda D(f(a))D(\Pi(\bar{f}(\bar{\mathfrak{a}}))) \\
&= D(f(a))\bar{f}(1 \otimes \bar{\mathfrak{a}}) + f(a)D(\bar{f}(1 \otimes \bar{\mathfrak{a}})) + \lambda D(f(a))D(\bar{f}(1 \otimes \bar{\mathfrak{a}})) \\
&= D(f(a)\bar{f}(1 \otimes \bar{\mathfrak{a}})) \\
&= D(\bar{f}(a \otimes \bar{\mathfrak{a}})).
\end{aligned}$$

This proves the first equation in Eq. (23). We next prove the second equation by induction on the length $k \geq 1$ of $\mathfrak{a} := a \otimes \bar{\mathfrak{a}} \in A \otimes \text{Ш}^+(A_T)$. When $k = 1$, we have $\mathfrak{a} = a \in A$ and

$$\begin{aligned}
\bar{f}(\Pi_A(a)) &= \bar{f}\,(Q(a) - \varepsilon(Q(a)) + 1 \otimes T(a)) \\
&= f(Q(a)) - f(Q(a)) + \Pi(D(f(Q(a)))) + \Pi(f(T(a))) \\
&= \Pi(f(d(Q(a)) + T(a))) \\
&= \Pi(f(a)),
\end{aligned}$$

using Lemma 4.5(a) and (b). Assume now that the claim has been proved for $k = n \geq 1$ and consider $\mathfrak{a} = a \otimes \bar{\mathfrak{a}}$ with length $n + 1$. Then we have

$$\begin{aligned}
\bar{f}(\Pi_A(a \otimes \bar{\mathfrak{a}})) &= \bar{f}\,(Q(a) \otimes \bar{\mathfrak{a}} - \Pi_A(Q(a)\bar{\mathfrak{a}}) - \lambda\,\Pi_A(d(Q(a))\bar{\mathfrak{a}}) + 1 \otimes T(a) \otimes \bar{\mathfrak{a}}) \\
&= \bar{f}(Q(a))\bar{f}(\Pi_A(\bar{\mathfrak{a}})) - \bar{f}(\Pi_A(Q(a)\bar{\mathfrak{a}})) - \lambda\bar{f}(\Pi_A(d(Q(a))\bar{\mathfrak{a}})) + \bar{f}(P_A(T(a) \otimes \bar{\mathfrak{a}})).
\end{aligned}$$

Here we have applied Lemma 4.5(b) in the last term. Applying the induction hypothesis to the first three terms and using the fact that the restriction $\tilde{f}$ of $\bar{f}$ to $A \otimes \text{Ш}^+(A_T)$ is compatible with the Rota–Baxter operators in the last term, we obtain

$$\begin{aligned}
\bar{f}(\Pi_A(a \otimes \bar{\mathfrak{a}})) &= f(Q(a))\Pi(\bar{f}(\bar{\mathfrak{a}})) - \Pi(\bar{f}(Q(a)\bar{\mathfrak{a}})) - \lambda\Pi(\bar{f}(d(Q(a))\bar{\mathfrak{a}})) + \Pi(\bar{f}(T(a) \otimes \bar{\mathfrak{a}})) \\
&= \Pi\left(D(f(Q(a)))\Pi(\bar{f}(\bar{\mathfrak{a}}))\right) + \Pi\left(f(T(a))\bar{f}(P_A(\bar{\mathfrak{a}}))\right),
\end{aligned}$$

where we have used integration by parts in Theorem 2.5(f) in the last step. On the other hand, we have

$$\begin{aligned}
\Pi(\bar{f}(a \otimes \bar{\mathfrak{a}})) &= \Pi(f(a)\bar{f}(P_A(\bar{\mathfrak{a}}))) \\
&= \Pi\left(f(d(Q(a)) + T(a))\bar{f}(P_A(\bar{\mathfrak{a}}))\right) \\
&= \Pi\left(D(f(Q(a)))\Pi(\bar{f}(\bar{\mathfrak{a}}))\right) + \Pi\left(f(T(a))\bar{f}(P_A(\bar{\mathfrak{a}}))\right).
\end{aligned}$$

Thus we have completed the proof of the existence of the integro-differential algebra homomorphism $\bar{f}$.

The uniqueness of $\bar{f}$: Suppose $\bar{f}_1 : \text{ID}(A)^* \to R$ is a homomorphism of integro-differential algebras such that $\bar{f}_1 \circ i_A = f$. For $1 \otimes a_1 \otimes \cdots \otimes a_n \in \text{Ш}^+(A_T)$, we have

$$\begin{aligned}
\bar{f}_1(1 \otimes a_1 \otimes \cdots \otimes a_n) &= \bar{f}_1\,(\Pi_A(a_1\Pi_A(\cdots\Pi_A(a_n)\cdots))) \\
&= \Pi(f(a_1)\Pi(\cdots\Pi(f(a_n))\cdots)) \\
&= \bar{f}\,(\Pi_A(a_1\Pi_A(\cdots\Pi_A(a_n)\cdots))) \\
&= \bar{f}(1 \otimes a_1 \otimes \cdots \otimes a_n).
\end{aligned}$$

Thus the restrictions of $\bar{f}$ and $\bar{f}_1$ to $A \otimes \text{Ш}^+(A_T)$ are the same. Further, by Lemma 4.5(a),

$$\bar{f}_1(\varepsilon(a)) = f(a) - \bar{f}_1(\Pi_A(d_A(a))) = f(a) - \Pi(D(f(a)) = \bar{f}(\varepsilon(a)).$$

Hence the restrictions of $\bar{f}$ and $\bar{f}_1$ to A_ε are also the same. As these restrictions to $A \otimes \text{Ш}^+(A_T)$ and A_ε are K-homomorphisms, by the universal property of the tensor product over K, $\bar{f}$ and $\bar{f}_1$ agree on $\text{ID}(A)^* = A_\varepsilon \otimes_K A \otimes \text{Ш}^+(A_T)$. This proves the uniqueness of $\bar{f}$ and thus completes the proof of Theorem 4.6.

4.4. Examples of regular differential algebras

In this section we show that some common examples of differential algebras, namely the algebra of differential polynomials and the algebra of rational functions, are regular where the weight can be taken arbitrary.

4.4.1. Rings of differential polynomials

Our main goal in this subsection is to prove that $(\mathbf{k}\{u\}, d)$ is a regular differential algebra for any weight, and to give an explicit quasi-antiderivative Q for d.

We start by introducing some definitions for classifying the elements of $A = \mathbf{k}\{u\}$. Let u_i, $i \geq 0$, be the i-th derivation of u. Then $\mathbf{k}\{u\}$ is the polynomial algebra on $\{u_i \mid i \geq 0\}$. For $\alpha = (\alpha_0, \ldots, \alpha_k) \in \mathbb{N}^{k+1}$, we write $u^\alpha = u_0^{\alpha_0} \cdots u_k^{\alpha_k}$. Furthermore, we use the convention that $u^\alpha = 1$ when $\alpha \in \mathbb{N}^0$ is the degenerate tuple of length zero. Then all monomials of $\mathbf{k}\{u\}$ are of the form u^α, where α contains no trailing zero. The **order** of such a monomial $u^{(\alpha_0,\ldots,\alpha_k)} \neq 1$ is defined to be k; the order of $u^{()} = 1$ is set to -1. The order of a nonzero differential polynomial is defined as the maximum of the orders of its monomials. The following classification of monomials is crucial [17,8]: A monomial u^α of order k is called **functional** if either $k \leq 0$ or $\alpha_k > 1$. We write

$$A_T = \mathbf{k}\{u^\alpha \mid u^\alpha \text{ is functional}\}$$

for the corresponding submodule. Since the product of two functional monomials is again functional, A_T is in fact a $\mathbf{k}$-subalgebra of A. Furthermore, we write A_J for the submodule generated by all monomials $u^\alpha \neq 1$.

Proposition 4.10. *For any $\lambda \in \mathbf{k}$, the canonical derivation $d\colon A \to A$ of weight λ defined in Theorem 3.1 admits a quasi-antiderivative Q with associated direct sums $A = A_T \oplus \operatorname{im} d$ and $A = A_J \oplus \ker d$.*

Proof. The main work goes into showing the direct sum $A = A_T \oplus \operatorname{im} d$. We first show $A_T \cap \operatorname{im} d = 0$. Let $x \in A$. If x has order -1, it is an element of $\mathbf{k}$ so that $d(x) = 0$. If x has order $k \geq 0$, we distinguish the two cases of $\lambda = 0$ and $\lambda \neq 0$. If $\lambda = 0$, then we have $d(x) = (\partial x/\partial u_k)\, u_{k+1} + \tilde{x}$, where all terms of $\tilde{x}$ have order at most k. Hence $d(x) \notin A_T$ and therefore we have $A_T \cap \operatorname{im} d = 0$.

We now turn to the case when $\lambda \neq 0$. By Eq. (1) and an inductive argument, we find that for a product $w = \prod_{i\in I} w_i$ in A, we have

$$d(w) = \sum_{\varnothing \neq J \subseteq I} \lambda^{|J|-1} \prod_{i\in J} d(w_i) \prod_{i\notin J} w_i.$$

Then for a given monomial $u^\alpha = u^{(\alpha_0,\ldots,\alpha_k)} = \prod_{i=0}^k u_i^{\alpha_i}$ of order k we have

$$\begin{aligned} d(u^\alpha) &= \sum_{0\leq\beta_i\leq\alpha_i,\, \sum_{i=0}^k \beta_i \geq 1} \lambda^{\beta_0+\cdots+\beta_k-1} \prod_{i=0}^k \binom{\alpha_i}{\beta_i} u_i^{\alpha_i-\beta_i} u_{i+1}^{\beta_i} \\ &= \sum_{0\leq\beta_i\leq\alpha_i,\, \sum_{i=0}^k \beta_i \geq 1} \lambda^{\beta_0+\cdots+\beta_k-1} \left(\prod_{i=0}^k \binom{\alpha_i}{\beta_i} u_i^{\alpha_i-\beta_i+\beta_{i-1}} \right) u_{k+1}^{\beta_k}, \end{aligned} \tag{24}$$

with the convention $\beta_{-1} = 0$. Consider the reverse lexicographic order on monomials of order $k+1$:

$$(\beta_0, \ldots, \beta_{k+1}) < (\gamma_0, \ldots, \gamma_{k+1}) \Leftrightarrow \exists\, 0 \leq n \leq k+1\ (\beta_i = \gamma_i \text{ for } n < i \leq k+1 \text{ and } \beta_n < \gamma_n).$$

The smallest monomial *of order* $k+1$ under this order in the sum in Eq. (24) is given by $u_0^{\alpha_0} \cdots u_{k-1}^{\alpha_{k-1}} u_k^{\alpha_k - 1} u_{k+1}$ when $\beta_k = 1$ and $\beta_0 = \cdots = \beta_{k-1} = 0$, coming from $u_0^{\alpha_0} \cdots u_{k-1}^{\alpha_{k-1}} d(u_k^{\alpha_k})$. Thus for two monomials of order k with $u^\alpha < u^\beta$ under this order, the least monomial of order $k+1$ in $d(u^\alpha)$ is smaller than the least monomial of order $k+1$ in $d(u^\beta)$. In particular, for the least monomial u^α of order k of our given element x of order $k \geq 0$, the least monomial of order $k+1$ in $d(u^\alpha)$ is the least monomial of order $k+1$ in $d(x)$ and is given by $u_0^{\alpha_0} \cdots u_{k-1}^{\alpha_{k-1}} u_k^{\alpha_k - 1} u_{k+1}$. Since this monomial is not in A_T, it follows that $d(x)$ is not in A_T, showing that $A_T \cap \operatorname{im} d = 0$.

Note that the previous argument shows in particular that $d(x) \neq 0$ for $x \notin \mathbf{k}$. Thus we have

$$A = A_J \oplus \mathbf{k}.$$

We next show that every monomial u^α in $\mathbf{k}\{u\}$ is in $A_T + \operatorname{im} d$. We prove this by induction on the order of u^α. If the order is -1 or 0, then $u^\alpha \in A_T$ by definition. Assuming the claim holds for differential monomials of order less than $k > 0$, consider now a monomial u^α of order k so that $\alpha = (\alpha_0, \ldots, \alpha_k)$. If $u^\alpha \in A_T$, we are done. If not, we must have $\alpha_k = 1$. Then we distinguish the cases when $\lambda = 0$ and $\lambda \neq 0$. If $\lambda = 0$, then

$$\begin{aligned} u^\alpha &= u_0^{\alpha_0} \cdots u_{k-1}^{\alpha_{k-1}} u_k \\ &= u_0^{\alpha_0} \cdots u_{k-2}^{\alpha_{k-2}} \tfrac{1}{\alpha_{k-1}+1}\, d(u_{k-1}^{\alpha_{k-1}+1}) \\ &= d(u_0^{\alpha_0} \cdots u_{k-2}^{\alpha_{k-2}} \tfrac{1}{\alpha_{k-1}+1}\, u_{k-1}^{\alpha_{k-1}+1}) - d\left(u_0^{\alpha_0} \cdots u_{k-2}^{\alpha_{k-2}}\right) \tfrac{1}{\alpha_{k-1}+1}\, u_{k-1}^{\alpha_{k-1}+1}. \end{aligned}$$

Now the first term in the result is in $\operatorname{im} d$ and the second term is in $A_T + \operatorname{im} d$ by the induction hypothesis, allowing us to complete the induction when $\lambda = 0$.

Now consider the case when $\lambda \neq 0$. Suppose the claim does not hold for some monomials $u^\alpha = u^{(\alpha_0,\dots,\alpha_{k-1},1)}$ of order k. Among these monomials, there is one such that the exponent vector $\alpha = (\alpha_0, \dots, \alpha_{k-1}, 1)$ is minimal with respect to the lexicographic order:

$$(\alpha_0, \dots, \alpha_{k-1}, 1) < (\beta_0, \dots, \beta_{k-1}, 1) \Leftrightarrow \exists\, 0 \leq n \leq k-1\ (\alpha_i = \beta_i \text{ for } 1 \leq i < n \text{ and } \alpha_n < \beta_n).$$

By Eq. (24), we have

$$\begin{aligned} d(u_{k-1}^{\alpha_{k-1}+1}) &= \sum_{\beta_{k-1}=1}^{\alpha_{k-1}+1} \binom{\alpha_{k-1}+1}{\beta_{k-1}} \lambda^{\beta_{k-1}-1} u_{k-1}^{\alpha_{k-1}+1-\beta_{k-1}} u_k^{\beta_{k-1}} \\ &= (\alpha_{k-1}+1) u_{k-1}^{\alpha_{k-1}} u_k + \sum_{\beta_{k-1}=2}^{\alpha_{k-1}+1} \binom{\alpha_{k-1}+1}{\beta_{k-1}} \lambda^{\beta_{k-1}-1} u_{k-1}^{\alpha_{k-1}+1-\beta_{k-1}} u_k^{\beta_{k-1}}. \end{aligned}$$

So

$$u_{k-1}^{\alpha_{k-1}} u_k = \tfrac{1}{\alpha_{k-1}+1} d(u_{k-1}^{\alpha_{k-1}+1}) - \sum_{\beta_{k-1}=2}^{\alpha_{k-1}+1} \tfrac{\lambda^{\beta_{k-1}-1}}{\alpha_{k-1}+1} \binom{\alpha_{k-1}+1}{\beta_{k-1}} u_{k-1}^{\alpha_{k-1}+1-\beta_{k-1}} u_k^{\beta_{k-1}}.$$

Thus

$$\begin{aligned} u^\alpha &= u_0^{\alpha_0} \cdots u_{k-1}^{\alpha_{k-1}} u_k \\ &= u_0^{\alpha_0} \cdots u_{k-2}^{\alpha_{k-2}} \tfrac{1}{\alpha_{k-1}+1} d(u_{k-1}^{\alpha_{k-1}+1}) - \sum_{\beta_{k-1}=2}^{\alpha_{k-1}+1} \tfrac{\lambda^{\beta_{k-1}-1}}{\alpha_{k-1}+1} \binom{\alpha_{k-1}+1}{\beta_{k-1}} u_0^{\alpha_0} \cdots u_{k-2}^{\alpha_{k-2}} u_{k-1}^{\alpha_{k-1}+1-\beta_{k-1}} u_k^{\beta_{k-1}}. \end{aligned}$$

The monomials in the sum are in A_T. For the first term, by Eq. (1), we have

$$\begin{aligned} &u_0^{\alpha_0} \cdots u_{k-2}^{\alpha_{k-2}} \tfrac{1}{\alpha_{k-1}+1} d(u_{k-1}^{\alpha_{k-1}+1}) \\ &= d\left(u_0^{\alpha_0} \cdots u_{k-2}^{\alpha_{k-2}} \tfrac{1}{\alpha_{k-1}+1} u_{k-1}^{\alpha_{k-1}+1}\right) - d(u_0^{\alpha_0} \cdots u_{k-2}^{\alpha_{k-2}}) \tfrac{1}{\alpha_{k-1}+1} u_{k-1}^{\alpha_{k-1}+1} \\ &\quad - \lambda\, d(u_0^{\alpha_0} \cdots u_{k-2}^{\alpha_{k-2}}) d\left(\tfrac{1}{\alpha_{k-1}+1} u_{k-1}^{\alpha_{k-1}+1}\right). \end{aligned}$$

As in the case of $\lambda = 0$, the first term in the result is in $\operatorname{im} d$ and the second term has the desired decomposition by the induction hypothesis. Applying Eq. (24) to both derivations in the third term, we see that the term is a linear combination of monomials of the form $u^\gamma = u^{(\gamma_0,\dots,\gamma_k)}$ where

$$\gamma = (\alpha_0 - \beta_0, \alpha_1 - \beta_1 + \beta_0, \dots, \alpha_{k-2} - \beta_{k-2} + \beta_{k-3}, \alpha_{k-1} + 1 - \beta_{k-1} + \beta_{k-2}, \beta_{k-1})$$

for some $0 \leq \beta_i \leq \alpha_i$, $0 \leq i \leq k-2$ with $\sum_{i=0}^{k-2} \beta_i \geq 1$ and $\beta_{k-1} \geq 1$. If such a monomials has $\beta_{k-1} \geq 2$, then the monomial is already in A_T. If such a monomial has $\beta_{k-1} = 1$, then it has order k and has lexicographic order less than u^α since $\sum_{i=0}^{k-2} \beta_i \geq 1$. By the minimality of u^α, this monomial is in $A_T + \operatorname{im} d$. Hence u^α is in $A_T + \operatorname{im} d$. This is a contradiction, allowing us to completes the induction when $\lambda \neq 0$.

With the two direct sum decompositions, the quasi-antiderivative Q is obtained by Proposition 4.2. □

We can thus conclude that $\mathbf{k}\{u\}$ is indeed a regular differential algebra, as claimed earlier. Hence the construction $\mathrm{ID}(\mathbf{k}\{u\})^*$ developed in Section 4.2 does yield the free integro-differential algebra over the single generator u.

Proposition 4.11. *Let* $\mathbf{k}$ *be a commutative* $\mathbb{Q}$*-algebra. Then the free integro-differential algebra* $\mathrm{ID}(\mathbf{k}\{u\})$ *is a polynomial algebra.*

Proof. We first take the coefficient ring to be $\mathbb{Q}$. Since $\mathrm{ID}(\mathbb{Q}\{u\})$ is isomorphic to $\mathrm{ID}(\mathbb{Q}\{u\})^*$, which is given by Eq. (17) with $A = \mathbb{Q}\{u\}$, it suffices to ensure that $\text{Ш}^+(A_T)$ is a polynomial algebra. Now observe that $A_T = \mathbb{Q}F$ is the monoid algebra generated over the set F of functional monomials. One checks immediately that the functional monomials F form a monoid under multiplication. Hence Theorem 2.3 of [24] is applicable, and we see that the mixable shuffle algebra $\text{Ш}^+(A_T) = \mathrm{MS}_{\mathbb{Q},\lambda}(F)$ is isomorphic to $\mathbb{Q}[\mathrm{Lyn}(F)]$, where $\mathrm{Lyn}(F)$ denotes the set of Lyndon words over F. This proves the proposition when $\mathbf{k} = \mathbb{Q}$. Then the conclusion follows for any commutative $\mathbb{Q}$-algebra $\mathbf{k}$ since $\mathrm{ID}(\mathbf{k}\{u\})^* \cong \mathbf{k} \otimes_{\mathbb{Q}} \mathrm{ID}(\mathbb{Q}\{u\})^*$. □

4.4.2. Rational functions

We show that the algebra of rational functions with derivation of any weight is regular.

Proposition 4.12. *Let* $A = \mathbb{C}(x)$*. For any* $\lambda \in \mathbb{C}$ *let*

$$d_\lambda : A \to A, f(x) \mapsto \begin{cases} \frac{f(x+\lambda)-f(x)}{\lambda}, & \lambda \neq 0, \\ f'(x), & \lambda = 0, \end{cases}$$

be the λ-derivation introduced in Example 2.2(b). Then d_λ is regular. In particular the difference operator on $\mathbb{C}(x)$ is a regular derivation of weight one.

Proof. We have considered the case of $\lambda = 0$ in Example 4.3. Modifying the notations there, any rational function can be uniquely expressed as

$$r + \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\gamma_{ij}}{(x-\alpha_{ij})^i}, \tag{25}$$

where $r \in \mathbb{C}[x]$, $\alpha_{ij} \in \mathbb{C}$ are distinct for any given i and $\gamma_{ij} \in \mathbb{C}$ are nonzero. Let $0 \neq \lambda \in \mathbb{C}$ be given. We have the direct sum of linear spaces

$$\mathbb{C}[x] \oplus \mathcal{R} = \mathbb{C}[x] \oplus \bigoplus_{i \geq 1} \mathcal{R}_i,$$

where $\mathcal{R}$ is the linear space from the fractions in Eq. (25), namely the linear space with basis $1/(x-\alpha)^i$, $\alpha \in \mathbb{C}$, $1 \leq i$, and $\mathcal{R}_i$, for fixed $i \geq 1$, is the linear subspace with basis $1/(x-\alpha)^i$, $\alpha \in \mathbb{C}$.

We note that the λ-**divided falling factorials**

$$\binom{x}{n}_\lambda := \frac{x(x-\lambda)(x-2\lambda)\cdots(x-(n+1)\lambda)}{n!}, \quad n \geq 0,$$

with the convention $\binom{x}{0}_\lambda = 1$, form a $\mathbb{C}$-basis of $\mathbb{C}[x]$. In fact,

$$\binom{x}{n}_\lambda = \frac{1}{n!} \sum_{k=0}^{n} s(n,k) \lambda^{n-k} x^k, \qquad x^n = n! \sum_{k=0}^{n} S(n,k) \lambda^{n-k} \binom{x}{n}_\lambda, \quad n \geq 0,$$

where $s(n,k)$ and $S(n,k)$ are Stirling numbers of the first and second kind, respectively; see [19,20] for example. By a direct computation, we have

$$d_\lambda \left(\binom{x}{n}_\lambda \right) = \frac{\binom{x+\lambda}{n}_\lambda - \binom{x}{n}_\lambda}{\lambda} = \binom{x}{n-1}_\lambda.$$

Thus $d_\lambda(\mathbb{C}[x]) = \mathbb{C}[x]$ and hence $\mathbb{C}[x] \subseteq \operatorname{im} d_\lambda$. We next note that $\mathcal{R}$, as well as $\mathcal{R}_k$, is also closed under the operator d_λ since

$$\lambda d_\lambda \left(\sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\gamma_{ij}}{(x-\alpha_{ij})^i} \right) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\gamma_{ij}}{(x-(\alpha_{ij}-\lambda))^i} - \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\gamma_{ij}}{(x-\alpha_{ij})^i}.$$

Further, for any $n \geq 0$ and $f(x) \in \mathbb{C}(x)$, we have

$$\lambda d_\lambda \left(\sum_{i=0}^{n} f(x+i\lambda) \right) = f(x+(n+1)\lambda) - f(x),$$

and similarly for $n < 0$,

$$\lambda d_\lambda \left(\sum_{i=n}^{-1} f(x+i\lambda) \right) = f(x) - f(x+n\lambda).$$

Thus for any $n \in \mathbb{Z}$, we have

$$f(x) \equiv f(x+n\lambda) \mod \operatorname{im} d_\lambda.$$

In particular,

$$1/(x-\alpha)^i \equiv 1/(x-(\alpha-n\lambda))^i \mod \operatorname{im} d_\lambda$$

and hence

$$1/(x-\alpha)^i \equiv 1/(x-\beta)^i \mod \operatorname{im} d_\lambda,$$

for some $\beta \in \mathbb{C}$ with the real part $\operatorname{Re}(\beta) \in [0, |\operatorname{Re}(\lambda)|)$. Consequently, any fraction in $\mathcal{R}$ is congruent modulo $\operatorname{im} d_\lambda$ to an element of

$$\mathbb{C}(x)_T := \left\{ \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\gamma_{ij}}{(x-\alpha_{ij})^i} \in \mathcal{R} \,\middle|\, \operatorname{Re}(\alpha_{ij}) \in [0, |\operatorname{Re}(\lambda)|) \right\}.$$

That is,

$$\mathbb{C}(x) = \operatorname{im} d_\lambda + \mathbb{C}(x)_T.$$

On the other hand, suppose there is a nonzero function

$$f(x) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\gamma_{ij}}{(x-\alpha_{ij})^i} \in \operatorname{im} d_\lambda \cap \mathbb{C}(x)_T.$$

Thus there is $g(x) = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \frac{\gamma_{ij}}{(x-\beta_{ij})^i}$ such that $d_\lambda(g(x)) = f(x)$. The range of i in $f(x)$ and $g(x)$ are the same since $d_\lambda(\mathcal{R}_i) \subseteq \mathcal{R}_i$. Let $f(x) = \sum_{i=1}^{k} f_i(x)$ and $g(x) = \sum_{i=1}^{k} g_i(x)$ be the homogeneous decompositions of f and g. Then $d_\lambda(g_i(x)) = f_i(x)$, $1 \le i \le k$. Fix $1 \le i \le k$ and take $\operatorname{Re}(\lambda) > 0$ for now. List $\beta_{i,1} < \cdots < \beta_{i,m_i}$ according to their lexicographic order from the pairs $(a, b) \leftrightarrow a + ib \in \mathbb{C}$. Then we have

$$\lambda d_\lambda(g_i(x)) = \sum_{j=1}^{m_i} \frac{\gamma_{ij}}{(x-(\beta_{ij}-\lambda))^i} - \sum_{j=1}^{m_i} \frac{\gamma_{ij}}{(x-\beta_{ij})^i}.$$

The first fraction in the first sum, $1/(x-(\beta_{i,1}-\lambda))^i$, is not the same as any other fraction in the first sum since they are translations by λ of distinct fractions in f_i, and is not the same as any fraction in the second sum since $\operatorname{Re}(\beta_{i,1}-\lambda) < \operatorname{Re}(\beta_{i,1}) \le \operatorname{Re}(\beta_{ij})$ for $1 \le j \le m_i$. Similarly the last fraction in the second sum, $1/(x-\beta_{i,m_i})^i$, is not the same as any other terms in the sums. Thus they both have nonzero coefficients in $d_\lambda(g_i(x))$. But

$$\operatorname{Re}(\beta_{i,m_i}) - \operatorname{Re}(\beta_{i,1}-\lambda) = \operatorname{Re}(\beta_{i,m_i} - (\beta_{i,1}-\lambda)) = \operatorname{Re}(\beta_{i,m_i} - \beta_{i,1}) + \operatorname{Re}(\lambda) \ge \operatorname{Re}(\lambda).$$

Hence $\operatorname{Re}(\beta_{i,m_i})$ and $\operatorname{Re}(\beta_{i,1}-\lambda)$ cannot both be in $[0, \operatorname{Re}(\lambda))$. Thus $d_\lambda(g_i)$ and hence $d_\lambda(g)$ cannot be in $\mathbb{C}(x)_T$. This is a contradiction, showing that $\operatorname{im} d_\lambda \cap \mathbb{C}(x)_T = 0$. When $\operatorname{Re}(\lambda) < 0$, we get analogously $\operatorname{im} d_\lambda \cap \mathbb{C}(x)_T = 0$. Thus we have proved

$$\mathbb{C}(x) = \operatorname{im} d_\lambda \oplus \mathbb{C}(x)_T. \tag{26}$$

Note that $\mathbb{C}(x)_T$ is closed under multiplication, hence is a nonunitary subalgebra of $\mathbb{C}(x)$.

The above argument shows that $d_\lambda(g)$ is in $\mathbb{C}(x)_T$ for $g \in \mathcal{R}$ only when $g = 0$. Thus $\ker d_\lambda \cap \mathcal{R} = 0$. Since d_λ preserves the decomposition $\mathbb{C}(x) = \mathbb{C}[x] \oplus \mathcal{R}$, we have $\ker d_\lambda = \ker(d_\lambda)\big|_{\mathbb{C}[x]} = \mathbb{C}$. Thus we have the direct sum decomposition

$$\mathbb{C}(x) = \ker d_\lambda \oplus (x\mathbb{C}[x] \oplus \mathcal{R}),$$

and hence d_λ is injective on $x\mathbb{C}[x] \oplus \mathcal{R}$ with image $\operatorname{im} d_\lambda$. Therefore d_λ is regular with quasi-antiderivative Q defined to be the inverse of

$$d_\lambda : x\mathbb{C}[x] \oplus \mathcal{R} \to \operatorname{im} d_\lambda$$

on $\operatorname{im} d_\lambda$ and to be zero on its complement $\mathbb{C}(x)_T$; see Proposition 4.2. □

Remark 4.13. We remark that the subalgebra of $\mathbb{C}(x)$ that is a complement of $\operatorname{im} d_\lambda$ is not unique, thus giving different quasi-antiderivatives. In fact, from the proof of Proposition 4.12 it is apparent that in the decomposition (26) one can replace $\mathbb{C}(x)_T$ by

$$\mathbb{C}(x)_{T,a} = \left\{ \sum_{i=1}^{k} \sum_{j=1}^{n_i} \frac{\gamma_{ij}}{(x-\alpha_{ij})^i} \in \mathcal{R} \;\middle|\; \operatorname{Re}(\alpha_i) \in [a, a + |\operatorname{Re}(\lambda)|) \right\},$$

for any given $a \in \mathbb{R}$. These two subalgebras are isomorphic since $\mathbb{C}(x)_{T,a}$ is isomorphic to the polynomial $\mathbb{C}$-algebra with generating set

$$\left\{ \frac{1}{x-\alpha} \;\middle|\; \alpha \in [a, a + |\operatorname{Re}(\lambda)|) \right\}.$$

Remark 4.14. In conclusion, we have given the first construction for the free integro-differential algebra $\mathrm{ID}(A)^*$ over a given regular differential algebra A. In several ways, this construction is similar to the integro-differential polynomials of [36,38]. This will be clear when one writes out the elements $a_0 \otimes a_1 \otimes a_2 \otimes \cdots$ of Eq. (16) in the form $a_0 \int a_1 \int a_2 \int \cdots$. But there are also some important differences:

(a) The integro-differential polynomials are the polynomial algebra in the variety of integro-differential algebras of weight zero, not the free algebra in this category. In fact, the polynomial algebra is always a free product of the coefficient algebra and the free algebra by Theorem 4.31 of [30].
(b) The construction of [36] uses the language of term algebras and rewrite systems whereas in this paper we use a more abstract approach through tensor products.

(c) In the integro-differential polynomials, the starting point is a given integro-differential algebra (A, D, Π) instead of a regular differential algebra as in the present paper. In the former case we can construct nested integrals over differential polynomials with coefficients in A, whereas in the latter case we can only treat differential polynomials with trivial coefficients (i.e., the derivation vanishes on them).

It would be interesting to apply the methods used in this paper to rederive and generalize the construction of the integro-differential polynomials of [36]. This would also shed some light on the constructive meaning of the free product mentioned in Item (a) above. An important step in this direction might be generalizing Section 4.4.1 to differential polynomials with nonzero derivation on the coefficient ring $\mathbf{k}$. See [16] for a construction of the free integro-differential algebra on an arbitrary set by the method of Gröbner–Shirshov bases.

Acknowledgments

L. Guo acknowledges the support from NSF grant DMS 1001855. G. Regensburger was supported by the Austrian Science Fund (FWF): J 3030-N18. M. Rosenkranz acknowledges the support from the EPSRC First Grant EP/I037474/1. The authors thank the editor and the anonymous referee for helpful suggestions.

References

[1] R. Agarwal, M. Bohner, D. O'Regan, A. Peterson, Dynamic equations on time scales: a survey, J. Comput. Appl. Math. 141 (2002) 1–26.
[2] H. Albrecher, C. Constantinescu, G. Pirsic, G. Regensburger, M. Rosenkranz, An algebraic operator approach to the analysis of Gerber-Shiu functions, Insurance Math. Econom. 46 (2010) 42–51.
[3] H. Albrecher, C. Constantinescu, Z. Palmowski, G. Regensburger, M. Rosenkranz, Exact and asymptotic results for insurance risk models with surplus-dependent premiums, SIAM J. Appl. Math. 73 (2013) 47–66.
[4] C. Bai, A unified algebraic approach to classical Yang–Baxter equation, Jour. Phys. A: Math. Theor. 40 (36) (2007) 11073–11082.
[5] V.V. Bavula, The algebra of integro-differential operators on a polynomial algebra, J. Lond. Math. Soc. (2) 83 (2011) 517–543.
[6] V.V. Bavula, The algebra of integro-differential operators on an affine line and its modules, J. Pure Appl. Algebra 217 (2013) 495–529.
[7] G. Baxter, An analytic problem whose solution follows from a simple algebraic identity, Pacific J. Math. 10 (1960) 731–742.
[8] A.H. Bilge, A REDUCE program for the integration of differential polynomials, Comput. Phys. Comm. 71 (1992) 263–268.
[9] G. Birkhoff, On the structure of abstract algebras, Proc. Cambridge Phil. Soc. 31 (1935) 433–454.
[10] F. Boulier, D. Lazard, F. Ollivier, M. Petitot, Computing representations for radicals of finitely generated differential ideals, Appl. Algebra Engrg. Comm. Comput. 20 (2009) 73–121.
[11] M. Bronstein, Symbolic Integration I: Transcendental Functions, second ed., Springer-Verlag, Berlin, 2005.
[12] S. Burris, H. P. Sankappanavar, A Course in Universal Algebra, Springer-Verlag, 1981, Available at http://www.math.uwaterloo.ca/~snburris/htdocs/ualg.html.
[13] S. Chen, M.F. Singer, Residues and telescopers for bivariate rational functions, Adv. Appl. Math. 49 (2012) 111–133.
[14] A. Connes, D. Kreimer, Renormalization in quantum field theory and the Riemann–Hilbert problem. I. The Hopf algebra structure of graphs and the main theorem, Comm. Math. Phys. 210 (2000) 249–273.
[15] K. Ebrahimi-Fard, L. Guo, D. Kreimer, Spitzer's identity and the algebraic Birkhoff decomposition in pQFT, J. Phys. A: Math. Gen. 37 (2004) 11037–11052.
[16] X. Gao, L. Guo, S. Zheng, Construction of free commutative integro-differential algebras by the method of Gröbner–Shirshov bases, arXiv:1302.0041 [math.AC].
[17] I.M. Gelfand, L.A. Dikiĭ, Fractional powers of operators and Hamiltonian systems, Funkcional. Anal. i Priložen. 10 (1976) 13–29. English translation: Functional Anal. Appl. 10 (1977), 259–273.
[18] L. Guo, WHAT IS a Rota–Baxter algebra, Notices Amer. Math. Soc. 56 (2009) 1436–1437.
[19] L. Guo, Introduction to Rota–Baxter Algebra, International Press, 2012.
[20] L. Guo, Baxter algebras, Stirling numbers and partitions, J. Algebra Appl. 4 (2005) 153–164.
[21] L. Guo, W. Keigher, Baxter algebras and shuffle products, Adv. Math. 150 (2000) 117–149.
[22] L. Guo, W. Keigher, On free Baxter algebras: completions and the internal construction, Adv. Math. 151 (2000) 101–127.
[23] L. Guo, W. Keigher, On differential Rota–Baxter algebras, J. Pure Appl. Algebra 212 (2008) 522–540.
[24] L. Guo, B. Xie, Structure theorems of mixable shuffle algebras and free commutative Rota–Baxter algebras, Comm. Algebra 41 (2013) 2629–2649.
[25] L. Guo, B. Zhang, Renormalization of multiple zeta values, J. Algebra 319 (2008) 3770–3809.
[26] M. Hoffman, Quasi-shuffle products, J. Algebraic Combin. 11 (2000) 49–68.
[27] V. Kac, P. Cheung, Quantum Calculus, Universitext, Springer, 2002.
[28] E.R. Kolchin, Differential Algebra and Algebraic Groups, Academic Press, New York, 1973.
[29] A. Korporal, G. Regensburger, M. Rosenkranz, Regular and singular boundary problems in Maple, in: V.P. Gerdt, W. Koepf, E.W. Mayr, E.H. Vorozhtsov (Eds.), Computer Algebra in Scientific Computing. Proceedings of the 13th International Workshop, CASC 2011, in: LNCS, vol. 6885, Springer, 2011, pp. 280–293.
[30] H. Lausch, W. Nöbauer, Algebra of Polynomials, North-Holland, Amsterdam, 1973.
[31] S. Mac Lane, Categories for the Working Mathematician, second ed., Springer-Verlag, New York, 1998.
[32] M.Z. Nashed, G.F. Votruba, A unified operator theory of generalized inverses, in: Generalized Inverses and Applications, Academic Press, New York, 1976, pp. 1–109.
[33] J.F. Ritt, Differential Algebra, American Mathematical Society, New York, 1950.
[34] M. Rosenkranz, A new symbolic method for solving linear two-point boundary value problems on the level of operator, J. Symbolic Comput. 39 (2005) 171–199.
[35] M. Rosenkranz, G. Regensburger, Solving and factoring boundary problems for linear ordinary differential equations in differential algebra, J. Symbolic Comput. 43 (2008) 515–544.
[36] M. Rosenkranz, G. Regensburger, Integro-differential polynomials and operators, in: D. Jeffrey (Ed.), Proceedings of the 2008 International Symposium on Symbolic and Algebraic Computation, ISSAC'08, ACM Press, 2008.
[37] M. Rosenkranz, G. Regensburger, L. Tec, B. Buchberger, A symbolic framework for operations on linear boundary problems, in: V.P. Gerdt, E.W. Mayr, E.H. Vorozhtsov (Eds.), Computer Algebra in Scientific Computing. Proceedings of the 11th International Workshop, CASC 2009, in: LNCS, vol. 5743, Springer, 2009, pp. 269–283.
[38] M. Rosenkranz, G. Regensburger, L. Tec, B. Buchberger, Symbolic analysis for boundary problems: from rewriting to parametrized Gröbner bases, in: U. Langer, P. Paule (Eds.), Numerical and Symbolic Scientific Computing: Progress and Prospects, Springer, Vienna, 2012, pp. 273–331.

Linear and Multilinear Algebra, 2014
Vol. 62, No. 12, 1567–1582, http://dx.doi.org/10.1080/03081087.2013.839672

On the product of projectors and generalized inverses

Anja Korporal and Georg Regensburger*

Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Linz, Austria

Communicated by R.B. Bapat

(*Received 25 February 2013; accepted 14 August 2013*)

We consider generalized inverses of linear operators on arbitrary vector spaces and study the question when their product in reverse order is again a generalized inverse. This problem is equivalent to the question when the product of two projectors is again a projector, and we discuss necessary and sufficient conditions in terms of their kernels and images alone. We give a new representation of the product of generalized inverses that does not require explicit knowledge of the factors. Our approach is based on implicit representations of subspaces via their orthogonals in the dual space. For Fredholm operators, the corresponding computations reduce to finite-dimensional problems. We illustrate our results with examples for matrices and linear ordinary boundary problems.

Keywords: generalized inverse; projector; reverse order law; Fredholm operator; linear boundary problem; duality

AMS Subject Classifications: 15A09; 47A05

1. Introduction

Analogues of the reverse order law $(AB)^{-1} = B^{-1}A^{-1}$ for bijective operators have been studied intensively for various kinds of generalized inverses. Most articles and books are concerned with the matrix case; see for example [1–11]. For infinite-dimensional vector spaces, usually additional topological structures like Banach or Hilbert spaces are assumed; see for example [12–15]. In our approach, we systematically exploit duality results that hold in arbitrary vector spaces and a corresponding duality principle for statements about generalized inverses and projectors; see Appendix A.

The validity of the reverse order law can be reduced to the question whether the product of two projectors is a projector (Section 2). This problem is studied in [16–18] for finite-dimensional vector spaces. We discuss necessary and sufficient conditions that carry over to arbitrary vector spaces and can be expressed in terms of the kernels and images of the respective operators alone (Section 4). Applying the duality principle leads to new conditions and a characterization of the commutativity of two projectors that generalizes a result from [19].

In Section 5, we translate the results for projectors to generalized inverses and obtain necessary and sufficient conditions for the reverse order law in arbitrary vector spaces.

*Corresponding author. Email: georg.regensburger@oeaw.ac.at

Based on these conditions, we give a short proof for the characterization in Theorem 5.3 of two operators such that the reverse holds for all inner inverses (also called g-inverses or {1}-inverses). Moreover, we show that there always exist algebraic generalized inverses (also called {1, 2}-inverses) of two operators A and B such that their product in reverse order is an algebraic generalized inverse of AB.

Assuming the reverse order law to hold, Theorem 6.2 gives a representation of the product of two outer inverses ({2}-inverses) that can be computed using only kernel and image of the outer inverses of the factors. In this representation, we rely on a description of the kernel of a composition using inner inverses (Section 3) and implicit representations of subspaces via their orthogonals in the dual space. Moreover, we avoid the computation of generalized inverses by using the associated transpose map. Examples for matrices illustrating the results are given in Section 7.

An important application for our results is given by linear boundary problems (Section 9). Their solution operators (Green's operators) are generalized inverses, and it is natural to express infinite dimensional solution spaces implicitly via the (homogeneous) boundary conditions they satisfy. Green's operators for ordinary boundary problems are Fredholm operators, for which we can check the conditions for the reverse order law algorithmically and compute the implicit representation of the product (Section 8). Hence we can test if the product of two (generalized) Green's operators is again a Green's operator, and we can determine which boundary problem it solves.

2. Generalized inverses

In this section, we first recall basic properties of generalized inverses. For further details and proofs, we refer to [15,20] and the references therein. Throughout this article, U, V and W always denote vector spaces over the same field F, and we use the notation $V_1 \leq V$ for a subspace V_1 of V.

Definition 2.1 Let $T\colon V \to W$ be linear. We call a linear map $G\colon W \to V$ an *inner inverse* of T if $TGT = T$ and an *outer inverse* of T if $GTG = G$. If G is an inner and an outer inverse of T, we call G an *algebraic generalized inverse* of T.

This terminology of generalized inverses is adopted from [20]; other sources refer to inner inverses as generalized inverses or g-inverses, whereas algebraic generalized inverses are also called reflexive generalized inverses. Also the notations {1}-inverse (resp. {2}- and {1, 2}-inverse) are used, which refer to the corresponding Moore–Penrose equations the generalized inverse satisfies.

PROPOSITION 2.2 *Let $T\colon V \to W$ and $G\colon W \to V$ be linear. The following statements are equivalent:*

(i) *G is an outer inverse of T.*
(ii) *GT is a projector and* $\mathrm{Im}GT = \mathrm{Im}G$.
(iii) *GT is a projector and* $V = \mathrm{Im}G \oplus \mathrm{Ker}GT$.
(iv) *GT is a projector and* $W = \mathrm{Im}T + \mathrm{Ker}G$.
(v) *TG is a projector and* $\mathrm{Ker}TG = \mathrm{Ker}G$.
(vi) *TG is a projector and* $W = \mathrm{Ker}G \oplus \mathrm{Im}TG$.
(vii) *TG is a projector and* $\mathrm{Im}G \cap \mathrm{Ker}T = \{0\}$.

Corresponding to (vii) and (vi), for subspaces $B \leq V$ and $E \leq W$ with

$$B \cap \mathrm{Ker}T = \{0\} \quad \text{and} \quad W = E \oplus T(B),$$

we can construct an outer inverse G of T with $\mathrm{Im}G = B$ and $\mathrm{Ker}G = E$ as follows; cf. [15, Cor. 8.2]. We consider the projector Q with

$$\mathrm{Im}Q = T(B), \quad \mathrm{Ker}Q = E. \tag{1}$$

The restriction $T|_B \colon B \to T(B)$ is bijective since $B \cap \mathrm{Ker}T = \{0\}$, and we can define $G = (T|_B)^{-1}Q$. One easily verifies that G is an outer inverse of T with $\mathrm{Im}G = B$ and $\mathrm{Ker}G = E$. Since by Proposition 2.2(iii) we have $V = B \oplus T^{-1}(E)$, we define the projector P in analogy to Q by

$$\mathrm{Im}P = T^{-1}(E), \quad \mathrm{Ker}P = B. \tag{2}$$

Then, by definition and by Proposition 2.2, we have

$$GTG = G, \quad TG = Q \quad \text{and} \quad GT = 1 - P,$$

and G is determined uniquely by these equations. Hence an outer inverse depends only on the choice of the defining spaces B and E. We use the notations $G = \mathrm{O}(T, B, E)$ and $G = \mathrm{O}(T, P, Q)$ for P and Q as in (2) and (1).

Obviously, G is an outer inverse of T if and only if T is an inner inverse of G. Therefore, we get a result analogous to Proposition 2.2 for inner inverses by interchanging the role of T and G. The construction of inner inverses is not completely analogous to outer inverses, see [20, Prop. 1.3]. For subspaces $B \leq V$ and $E \leq W$ such that

$$V = \mathrm{Ker}T \oplus B \quad \text{and} \quad W = \mathrm{Im}T \oplus E, \tag{3}$$

an inner inverse G of T is given on $\mathrm{Im}T$ by $(T|_B)^{-1}$ and can be chosen arbitrarily on E. For such an inner inverse with $B = \mathrm{Im}GT$ and $E = \mathrm{Ker}TG$, we write $G \in \mathrm{I}(T, B, E)$.

For constructing algebraic generalized inverses, we start with direct sums as in (3), but require $\mathrm{Ker}G = E$ and $\mathrm{Im}G = B$. We use the notation $G = \mathrm{G}(T, B, E)$.

The following result for inner inverses is well known in the matrix case [8,17,21] and its elementary proof remains valid for arbitrary vector spaces.

PROPOSITION 2.3 *Let $T_1 \colon V \to W$ and $T_2 \colon U \to V$ be linear with outer (resp. inner) inverses G_1 and G_2. Let $P = G_1T_1$ and $Q = T_2G_2$. Then G_2G_1 is an outer (resp. inner) inverse of T_1T_2 if and only if QP (resp. PQ) is a projector.*

Proof Let G_2G_1 be an outer inverse of T_1T_2, that is, $G_2G_1 = G_2G_1T_1T_2G_2G_1$. Multiplying with T_2 from the left and with T_1 from the right yields

$$T_2G_2G_1T_1 = T_2G_2G_1T_1T_2G_2G_1T_1,$$

thus $QP = T_2G_2G_1T_1$ is a projector. For the other direction, we multiply the previous equation with G_2 from the left and G_1 from the right and use that $G_1T_1G_1 = G_1$ and $G_2T_2G_2 = G_2$. The proof for inner inverses follows by interchanging the roles of T_i and G_i. □

3. Kernel of compositions

We now describe the inverse image of a subspace under the composition of two linear maps using inner inverses. For projectors, kernel and image of the composition can be expressed in terms of kernel and image of the corresponding factors alone. Note that a projector is an inner inverse of itself.

PROPOSITION 3.1 *Let $T_1 : V \to W$ and $T_2 : U \to V$ be linear and G_2 an inner inverse of T_2. For a subspace $W_1 \leq W$, we have*

$$(T_1T_2)^{-1}(W_1) = G_2(T_1^{-1}(W_1) \cap \mathrm{Im}T_2) \oplus \mathrm{Ker}T_2$$

for the inverse image of the composition. In particular,

$$\mathrm{Ker}T_1T_2 = G_2(\mathrm{Ker}T_1 \cap \mathrm{Im}T_2) \oplus \mathrm{Ker}T_2.$$

Proof Since T_2G_2 is a projector onto $\mathrm{Im}T_2$ by Proposition 2.2(ii) (interchanging the role of T and G), we have

$$\begin{aligned} T_1T_2(G_2(T_1^{-1}(W_1) \cap \mathrm{Im}T_2) + \mathrm{Ker}T_2) &= T_1Q_2(T_1^{-1}(W_1) \cap \mathrm{Im}T_2) + 0 \\ &= T_1(T_1^{-1}(W_1) \cap \mathrm{Im}T_2) \leq W_1 \cap \mathrm{Im}T_1T_2 \leq W_1. \end{aligned}$$

Conversely, let $u \in (T_1T_2)^{-1}(W_1)$. Then $T_2u = v$ with $v \in T_1^{-1}(W_1)$. Since also $v \in \mathrm{Im}T_2$, we have

$$T_2(u - G_2v) = T_2u - Q_2v = T_2u - v = v - v = 0,$$

that is, $u - G_2v \in \mathrm{Ker}T_2$. Writing $u = G_2v + u - G_2v$ yields $u \in G_2(T_1^{-1}(W_1) \cap \mathrm{Im}T_2) + \mathrm{Ker}T_2$. The sum is direct since by Proposition 2.2(vi) (interchanging the role of T and G), we have $U = \mathrm{Ker}T_2 \oplus \mathrm{Im}G_2T_2$. □

COROLLARY 3.2 *Let $T : V \to W$ be linear and let $P : V \to V$ and $Q : W \to W$ be projectors. Then*

$$\mathrm{Ker}TQ = (\mathrm{Ker}T \cap \mathrm{Im}Q) \oplus \mathrm{Ker}Q \quad \textit{and} \quad \mathrm{Im}PT = (\mathrm{Im}T + \mathrm{Ker}P) \cap \mathrm{Im}P.$$

Proof Applying Proposition 3.1 yields

$$\mathrm{Ker}TQ = Q(\mathrm{Ker}T \cap \mathrm{Im}Q) \oplus \mathrm{Ker}Q = (\mathrm{Ker}T \cap \mathrm{Im}Q) \oplus \mathrm{Ker}Q.$$

The statement for the image follows from the duality principle A.4. □

This result generalizes [17, Lemma 2.2], where the kernel and image of a product PQ of two projectors are computed as above, when PQ is again a projector.

4. Products of projectors

In view of Proposition 2.3, we study necessary and sufficient conditions for the product of two projectors to be a projector. Throughout this section let $P, Q : V \to V$ denote projectors.

The first of the following necessary and sufficient conditions for the product of P and Q to be a projector is mentioned as an exercise without proof in [22, p. 339]. In [16, Lemma 3]

the same result is formulated for matrices but the proof is valid for arbitrary vector spaces. The second necessary and sufficient condition for the matrix case is given in [17, Lemma 2.2]. The simpler proof from [18] carries over to arbitrary vector spaces.

LEMMA 4.1 *The composition PQ is a projector if and only if*

$$\operatorname{Im} PQ \leq \operatorname{Im} Q \oplus (\operatorname{Ker} P \cap \operatorname{Ker} Q)$$

if and only if

$$\operatorname{Im} Q \leq \operatorname{Im} P \oplus (\operatorname{Ker} P \cap \operatorname{Im} Q) \oplus (\operatorname{Ker} P \cap \operatorname{Ker} Q).$$

We obtain the following characterization of the idempotency of PQ in terms of the kernels and images of P and Q alone.

THEOREM 4.2 *The following statements are equivalent:*

(i) *The composition PQ is a projector.*
(ii) $\operatorname{Im} P \cap (\operatorname{Im} Q + \operatorname{Ker} P) \leq \operatorname{Im} Q \oplus (\operatorname{Ker} P \cap \operatorname{Ker} Q)$
(iii) $\operatorname{Im} Q \leq \operatorname{Im} P \oplus (\operatorname{Ker} P \cap \operatorname{Im} Q) \oplus (\operatorname{Ker} P \cap \operatorname{Ker} Q)$
(iv) $\operatorname{Ker} Q \oplus (\operatorname{Ker} P \cap \operatorname{Im} Q) \geq \operatorname{Ker} P \cap (\operatorname{Im} Q + \operatorname{Im} P)$
(v) $\operatorname{Ker} P \geq \operatorname{Ker} Q \cap (\operatorname{Im} Q + \operatorname{Ker} P) \cap (\operatorname{Im} Q + \operatorname{Im} P)$

Proof The equivalence of (i), (ii) and (iii) follows from the previous lemma and Corollary 3.2. By the duality principle A.4, the last two conditions are equivalent to (ii) and (iii), respectively. □

For algebraic generalized inverses, it is also interesting to have sufficient conditions for PQ as well as QP to be projectors; for example, if P and Q commute. This can again be characterized in terms of the images and kernels of P and Q alone. If $PQ = QP$, one sees with Corollary 3.2 that

$$\operatorname{Im} PQ = \operatorname{Im} P \cap \operatorname{Im} Q \quad \text{and} \quad \operatorname{Ker} PQ = \operatorname{Ker} P + \operatorname{Ker} Q. \tag{4}$$

In general, these conditions are necessary but not sufficient for commutativity of P and Q, see [16, Ex. 1].

Using Corollary 3.2, modularity (A1) and (A2), one obtains the following characterization of projectors with image or kernel as in (4); for further details see [23]. For the commutativity of projectors see also [22, p. 339].

PROPOSITION 4.3 *The composition PQ is a projector with*

(i) $\operatorname{Im} PQ = \operatorname{Im} P \cap \operatorname{Im} Q$ *if and only if*

$$\operatorname{Im} Q = (\operatorname{Im} P \cap \operatorname{Im} Q) \oplus (\operatorname{Ker} P \cap \operatorname{Im} Q).$$

(ii) $\operatorname{Ker} PQ = \operatorname{Ker} P + \operatorname{Ker} Q$ *if and only if*

$$\operatorname{Ker} P = (\operatorname{Ker} P \cap \operatorname{Ker} Q) \oplus (\operatorname{Ker} P \cap \operatorname{Im} Q).$$

COROLLARY 4.4 *We have $PQ = QP$ if and only if*

$$\operatorname{Im} Q = (\operatorname{Im} P \cap \operatorname{Im} Q) \oplus (\operatorname{Ker} P \cap \operatorname{Im} Q)$$

and

$$\operatorname{Ker} Q = (\operatorname{Im} P \cap \operatorname{Ker} Q) \oplus (\operatorname{Ker} P \cap \operatorname{Ker} Q).$$

In [16, Thm. 4] and [19, Thm. 3.2] different necessary and sufficient conditions for the commutativity of two projectors are given, but both require the computation of PQ as well as of QP.

5. Reverse order law for generalized inverses

Proposition 2.3 and Theorem 4.2 together give necessary and sufficient conditions for the reverse order law for outer inverses to hold, in terms of the defining spaces B_i and E_i alone.

THEOREM 5.1 *Let $T_1 \colon V \to W$ and $T_2 \colon U \to V$ be linear with outer inverses $G_1 = \mathrm{O}(T_1, B_1, E_1)$ and $G_2 = \mathrm{O}(T_2, B_2, E_2)$. The following conditions are equivalent:*

(i) G_2G_1 *is an outer inverse of* T_1T_2.
(ii) $T_2(B_2) \cap (B_1 + E_2) \leq B_1 \oplus (E_2 \cap T_1^{-1}(E_1))$
(iii) $B_1 \leq T_2(B_2) \oplus (E_2 \cap B_1) \oplus (E_2 \cap T_1^{-1}(E_1))$
(iv) $T_1^{-1}(E_1) \oplus (E_2 \cap B_1) \geq E_2 \cap (B_1 + T_2(B_2))$
(v) $E_2 \geq T_1^{-1}(E_1) \cap (B_1 + E_2) \cap (B_1 + T_2(B_2))$

Proof Recall that $\operatorname{Im} G_i = B_i$ and $\operatorname{Ker} G_i = E_i$, and $Q = T_2G_2$ and $P = G_1T_1$ are projectors with

$$\operatorname{Im} P = B_1, \quad \operatorname{Ker} P = T_1^{-1}(E_1), \quad \operatorname{Im} Q = T_2(B_2) \quad \text{and} \quad \operatorname{Ker} Q = E_2.$$

By Proposition 2.3, G_2G_1 is an outer inverse if and only if QP is a projector. Applying Theorem 4.2 proves the claim. □

In the following theorem, we give the analogous conditions for inner inverses, where $P = G_1T_1$ and $Q = T_2G_2$ are the projectors corresponding to the direct sums in (3). Note that the conditions for inner inverses only depend on the choice of B_1 and E_2, but not on B_2 and E_1.

The characterization of (iii) and the orthogonal of (v) in the following theorem generalize [17, Thm. 2.3] to arbitrary vector spaces.

THEOREM 5.2 *Let $T_1 \colon V \to W$ and $T_2 \colon U \to V$ be linear with inner inverses $G_1 \in \mathrm{I}(T_1, B_1, E_1)$ and $G_2 \in \mathrm{I}(T_2, B_2, E_2)$. The following conditions are equivalent:*

(i) G_2G_1 *is an inner inverse of* T_1T_2.
(ii) $B_1 \cap (\operatorname{Im} T_2 + \operatorname{Ker} T_1) \leq \operatorname{Im} T_2 \oplus (\operatorname{Ker} T_1 \cap E_2)$
(iii) $\operatorname{Im} T_2 \leq B_1 \oplus (\operatorname{Ker} T_1 \cap \operatorname{Im} T_2) \oplus (\operatorname{Ker} T_1 \cap E_2)$
(iv) $E_2 \oplus (\operatorname{Ker} T_1 \cap \operatorname{Im} T_2) \geq \operatorname{Ker} T_1 \cap (\operatorname{Im} T_2 + B_1)$
(v) $\operatorname{Ker} T_1 \geq E_2 \cap (\operatorname{Im} T_2 + \operatorname{Ker} T_1) \cap (\operatorname{Im} T_2 + B_1)$

The question when the reverse order law holds for all inner inverses of T_1 and T_2 was answered for matrices in [11, Thm. 2.3], and an alternative proof was given in [24]. Using the previous characterizations, we give a short proof that generalizes the result to arbitrary vector spaces.

THEOREM 5.3 *Let $T_1 : V \to W$ and $T_2 : U \to V$ be linear. Then G_2G_1 is an inner inverse of T_1T_2 for all inner inverses G_1 of T_1 and G_2 of T_2 if and only if $T_1T_2 = 0$ or* $\mathrm{Ker}T_1 \leq \mathrm{Im}T_2$.

Proof If $\mathrm{Ker}T_1 \leq \mathrm{Im}T_2$ then $\mathrm{Ker}T_1 \cap \mathrm{Im}T_2 = \mathrm{Ker}T_1$ and (iii) in Theorem 5.2 the previous theorem is satisfied since $\mathrm{Ker}T_1 + B_1 = V$. The case $T_1T_2 = 0$ is trivial.

For the reverse implication, assume that $\mathrm{Im}T_2$ is not contained in $\mathrm{Ker}T_1$ and $\mathrm{Ker}T_1$ is not contained in $\mathrm{Im}T_2$. Choose $V_1, V_2 \leq V$ such that we have two direct sums $\mathrm{Ker}T_1 = (\mathrm{Im}T_2 \cap \mathrm{Ker}T_1) \oplus V_1$ and $\mathrm{Im}T_2 = (\mathrm{Im}T_2 \cap \mathrm{Ker}T_1) \oplus V_2$. Then we have

$$\mathrm{Im}T_2 + \mathrm{Ker}T_1 = (\mathrm{Im}T_2 \cap \mathrm{Ker}T_1) \oplus V_1 \oplus V_2. \tag{5}$$

By assumption, we can choose non-zero $v_1 \in V_1$ and $v_2 \in V_2$. Let $v = v_1 + v_2$. Then $v \in \mathrm{Im}T_2 + \mathrm{Ker}T_1$ and $v \notin \mathrm{Ker}T_1$, $v \notin \mathrm{Im}T_2$. Hence we can choose B_1 and E_2 such that $v \in B_1$ and $v \in E_2$ and $V = \mathrm{Ker}T_1 \oplus B_1 = \mathrm{Im}T_2 \oplus E_2$. Then

$$v \in E_2 \cap (\mathrm{Im}T_2 + \mathrm{Ker}T_1) \cap (\mathrm{Im}T_2 + B_1)$$

but $v \in \mathrm{Ker}T_1$. Hence 5.2 in the previous theorem is not satisfied for inner inverses with $\mathrm{Im}G_1 = B_1$ and $\mathrm{Ker}G_2 = E_2$. □

Werner [17, Thm. 3.1] proves that for matrices, it is always possible to construct inner inverses such that the reverse order law holds. Using the necessary and sufficient condition for outer inverses above, we extend this result to algebraic generalized inverses in arbitrary vector spaces. The special case of Moore–Penrose inverses is treated in [8, Thm. 3.2], and explicit solutions are constructed in [25,26].

THEOREM 5.4 *Let $T_1 : V \to W$ and $T_2 : U \to V$ be linear. There always exist algebraic generalized inverses G_1 of T_1 and G_2 of T_2 such that G_2G_1 is an algebraic generalized inverse of T_1T_2.*

Proof Choose $V_1, V_2 \leq V$ as in the previous proof such that (5) holds. Moreover, choose $V_3 \leq V$ such that

$$V = (\mathrm{Im}T_2 + \mathrm{Ker}T_1) \oplus V_3 = (\mathrm{Im}T_2 \cap \mathrm{Ker}T_1) \oplus V_1 \oplus V_2 \oplus V_3.$$

Then $B_1 = V_2 \oplus V_3$ is a direct complement of $\mathrm{Ker}T_1$ and $E_2 = V_1 \oplus V_3$ is a direct complement of $\mathrm{Im}T_2$. Hence, there exist respectively an algebraic generalized inverse G_1 of T_1 with $\mathrm{Im}G_1 = B_1$ and G_2 of T_2 with $\mathrm{Ker}G_2 = E_2$. We verify that such G_1 and G_2 satisfy Theorem 5.1(iii), where $T_1^{-1}(E_1) = \mathrm{Ker}T_1$ and $T_2(B_2) = \mathrm{Im}T_2$ since G_1 and G_2 are algebraic generalized inverses:

$$\mathrm{Im}T_2 \oplus (E_2 \cap B_1) \geq \mathrm{Im}T_2 \oplus V_3 = (\mathrm{Im}T_2 \cap \mathrm{Ker}T_1) \oplus V_2 \oplus V_3 \geq B_1.$$

Similarly, we verify Theorem 5.2(iii)

$$B_1 \oplus (\mathrm{Ker}T_1 \cap \mathrm{Im}T_2) = V_2 \oplus V_3 \oplus (\mathrm{Ker}T_1 \cap \mathrm{Im}T_2) \geq V_2 \oplus (\mathrm{Ker}T_1 \cap \mathrm{Im}T_2) = \mathrm{Im}T_2.$$

Hence G_2G_1 is an algebraic generalized inverse of T_1T_2 for all $G_1 = \mathrm{G}(T_1, B_1, E_1)$ and $G_2 = \mathrm{G}(T_2, B_2, E_2)$, independent of the choice of E_1 and B_2. □

6. Representing the product of outer inverses

In this section, we assume that for two linear maps $T_1 \colon V \to W$ and $T_2 \colon U \to V$ with outer inverses G_1 and G_2, respectively, the reverse order law holds. Our goal is to find a description of the product G_2G_1 that does not require the explicit knowledge of G_1 and G_2. Using the representation via projectors, one immediately verifies that

$$\mathrm{O}(T_2, P_2, Q_2)\,\mathrm{O}(T_1, P_1, Q_1) = \mathrm{O}(T_1T_2, P_2 - G_2P_1T_2, T_1Q_2G_1)$$

but this expression involves both outer inverses G_1 and G_2. For the representation via defining spaces, we compute the kernel and the image of the product.

Lemma 6.1 *Let $T_1 \colon V \to W$ and $T_2 \colon U \to V$ be linear with outer inverses $G_1 = \mathrm{O}(T_1, B_1, E_1)$ and $G_2 = \mathrm{O}(T_2, B_2, E_2)$. Then*

$$\mathrm{Ker}G_2G_1 = E_1 \oplus T_1(B_1 \cap E_2) \quad \textit{and} \quad \mathrm{Im}G_2G_1 = G_2((B_1 + E_2) \cap \mathrm{Im}T_2).$$

Proof Recall that by definition $\mathrm{Ker}G_i = E_i$ and $\mathrm{Im}G_i = B_i$. The first identity follows directly from Proposition 3.1. For the second identity, we first note that for a linear map G and subspaces V_1, V_2, we have $G(V_1 \cap V_2) = G(V_1) \cap G(V_2)$ if $\mathrm{Ker}G \leq V_1$. Hence $G_2((B_1 + E_2) \cap \mathrm{Im}T_2)$ equals

$$G_2((\mathrm{Im}G_1 + \mathrm{Ker}G_2) \cap \mathrm{Im}T_2) = G_2(\mathrm{Im}G_1) \cap G_2(\mathrm{Im}T_2) = \mathrm{Im}G_2G_1,$$

since $G_2(\mathrm{Im}T_2) = \mathrm{Im}G_2$ by Proposition 2.2(ii). □

Note that the expression for the image of the composition requires the explicit knowledge of G_2. In particular, the reverse order law takes the form

$$\mathrm{O}(T_2, B_2, E_2)\,\mathrm{O}(T_1, B_1, E_1) = \mathrm{O}(T_1T_2, G_2((B_1 + E_2) \cap \mathrm{Im}T_2), E_1 + T_1(B_1 \cap E_2)).$$

Werner [17, Thm. 2.4] gives a result in a similar spirit for inner inverses of matrices.

Using an implicit description of $\mathrm{Im}G_i$, it is possible to state the reverse order law in a form that depends on the kernels and images of the respective outer inverses alone. This approach is motivated by our application to linear boundary problems (Section 9), where it is natural to define solution spaces via the boundary conditions they satisfy.

In more detail, the Galois connection from Appendix A allows to represent a subspace B implicitly via the orthogonally closed subspace $\mathscr{B} = B^\perp$ of the dual space. We will therefore use the notation $G = \mathrm{O}(T, \mathscr{B}, E)$ for the outer inverse with $\mathrm{Im}G = \mathscr{B}^\perp$ and $\mathrm{Ker}G = E$ as well as the analogue for inner inverses.

THEOREM 6.2 *Let $T_1 : V \to W$ and $T_2 : U \to V$ be linear with outer inverses $G_1 = \mathrm{O}(T_1, \mathscr{B}_1, E_1)$ and $G_2 = \mathrm{O}(T_2, \mathscr{B}_2, E_2)$. If G_2G_1 is an outer inverse of T_1T_2, then*

$$\mathrm{O}(T_2, \mathscr{B}_2, E_2)\,\mathrm{O}(T_1, \mathscr{B}_1, E_1) = \mathrm{O}(T_1T_2, \mathscr{B}_2 \oplus T_2^*(\mathscr{B}_1 \cap E_2^\perp), E_1 \oplus T_1(\mathscr{B}_1^\perp \cap E_2)), \quad (6)$$

where T_2^* denotes the transpose of T_2.

Proof From Lemma 6.1, we already know that $\mathrm{Ker}G_2G_1 = E_1 \oplus T_1(\mathscr{B}_1^\perp \cap E_2)$. From Proposition A.2 and 3.1, we get

$$\begin{aligned}(\mathrm{Im}G_2G_1)^\perp &= \mathrm{Ker}G_1^*G_2^* = T_2^*(\mathrm{Ker}G_1^* \cap \mathrm{Im}G_2^*) \oplus \mathrm{Ker}G_2^* \\ &= T_2^*((\mathrm{Im}G_1)^\perp \cap (\mathrm{Ker}G_2)^\perp) \oplus (\mathrm{Im}G_2)^\perp = T_2^*(\mathscr{B}_1 \cap E_2^\perp) \oplus \mathscr{B}_2,\end{aligned}$$

and thus (6) holds. □

A computational advantage of this representation is that one can determine G_2G_1 directly by computing only one outer inverse instead of computing both G_1 and G_2; see the next section for an example.

7. Examples for matrices

In this section, we illustrate our results for finite-dimensional vector spaces. In particular, we show how to compute directly the composition of two generalized inverses using the reverse order law in the form (6).

Consider the following linear maps $T_1 : \mathbb{Q}^4 \to \mathbb{Q}^3$ and $T_2 : \mathbb{Q}^3 \to \mathbb{Q}^4$ given by

$$T_1 = \begin{pmatrix} 1 & -1 & -1 & 1 \\ 0 & 2 & 2 & -2 \\ 3 & 1 & 1 & -1 \end{pmatrix} \quad \text{and} \quad T_2 = \begin{pmatrix} 1 & -2 & -1 \\ 1 & 1 & 2 \\ -1 & 5 & 4 \\ -1 & 5 & 4 \end{pmatrix}.$$

We first use Theorems 5.1 and 5.2 to check whether for algebraic generalized inverses $G_1 = \mathrm{G}(T_1, B_1, E_1)$ and $G_2 = \mathrm{G}(T_2, B_2, E_2)$, the composition G_2G_1 is an algebraic generalized inverse of T_1T_2.

For testing the conditions, we only need to fix $B_1 = \mathrm{Im}G_1$ and $E_2 = \mathrm{Ker}G_2$, such that $B_1 \oplus \mathrm{Ker}T_1 = \mathbb{Q}^4 = E_2 \oplus \mathrm{Im}T_2$. We have

$$\mathrm{Ker}T_1 = \mathrm{span}((0,1,0,1)^T, (0,0,1,1)^T), \quad \mathrm{Im}T_2 = \mathrm{span}((1,0,-2,-2)^T, (0,1,1,1)^T),$$

so we may choose for example

$$B_1 = \mathrm{span}((1,0,0,0)^T, (0,1,0,0)^T), \quad E_2 = \mathrm{span}((1,0,0,0)^T, (0,0,1,0)^T).$$

For algebraic generalized inverses, we obtain as a necessary and sufficient condition for being an outer inverse

$$B_1 \leq \mathrm{Im}T_2 \oplus (E_2 \cap B_1) \oplus (E_2 \cap \mathrm{Ker}T_1)$$

from Theorem 5.1(iii).

Since $E_2 \cap \mathrm{Ker}T_1 = \{0\}$ and $E_2 \cap B_1 = \mathrm{span}((1,0,0,0)^T)$, the right hand side yields that $\mathrm{span}((1,0,0,0)^T, (0,1,0,0)^T, (0,0,1,1)^T) \geq B_1$. Thus for all algebraic generalized

inverses G_1 and G_2 with $\mathrm{Im}G_1 = B_1$ and $\mathrm{Ker}G_2 = E_2$, the product G_2G_1 is an outer inverse of T_1T_2.

The corresponding condition for inner inverses by Theorem 5.2(iii) is

$$\mathrm{Im}T_2 \leq B_1 \oplus (\mathrm{Ker}T_1 \cap \mathrm{Im}T_2) \oplus (\mathrm{Ker}T_1 \cap E_2).$$

Since $\mathrm{Ker}T_1 \cap \mathrm{Im}T_2 = \{0\}$, the right hand side yields B_1, which does not contain $\mathrm{Im}T_2$. Hence for the above choices of G_1 and G_2, the product G_2G_1 is never an inner inverse of T_1T_2.

Since G_2G_1 is an outer inverse, Theorem 6.2 allows to determine G_2G_1 directly without knowing the factors. Identifying the dual space with row vectors, the orthogonals of B_1 and E_2 are given by

$$B_1^\perp = \mathscr{B}_1 = \mathrm{span}((0,0,1,0),(0,0,0,1)), \quad E_2^\perp = \mathrm{span}((0,1,0,0),(0,0,0,1)),$$

so we have $\mathscr{B}_1^\perp \cap E_2 = \mathrm{span}((1,0,0,0)^T)$ and $\mathscr{B}_1 \cap E_2^\perp = \mathrm{span}((0,0,0,1))$. For explicitly computing G_2G_1, we also have to choose $B_2 = \mathrm{Im}G_2$ and $E_1 = \mathrm{Ker}G_1$. Since we have

$$\mathrm{Im}T_1 = \mathrm{span}((1,0,3)^T),(0,1,2)^T), \quad \mathrm{Ker}T_2 = \mathrm{span}((1,1,-1)^T),$$

we may choose the complements $E_1 = \mathrm{Ker}G_1$ and $B_2 = \mathrm{Im}G_2$ as

$$E_1 = \mathrm{span}((0,0,1)^T) \quad \text{and} \quad B_2 = \mathrm{span}((1,0,0)^T,(0,1,0)^T).$$

Using (6), we can determine the kernel

$$E = \mathrm{Ker}G_2G_1 = E_1 \oplus T_1(\mathscr{B}_1^\perp \cap E_2) = \mathrm{span}((1,0,0)^T,(0,0,1)^T).$$

The image of G_2G_1 is by (6) given via the orthogonal

$$(\mathrm{Im}G_2G_1)^\perp = \mathscr{B}_2 \oplus T_2^*(\mathscr{B}_1 \cap E_2^\perp) = \mathrm{span}((0,0,1),(-1,5,4)),$$

which means that $B = \mathrm{Im}G_2G_1 = \mathrm{span}((5,1,0)^T)$. Therefore, we can directly determine G as the unique outer inverse

$$G = \mathrm{O}(T_1T_2, B, E) = \begin{pmatrix} 0 & \frac{5}{12} & 0 \\ 0 & \frac{1}{12} & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

One easily checks that G is an outer inverse of T.

8. Fredholm operators

We now turn to algorithmic aspects of the previous results. As already emphasized, for arbitrary vector spaces we can express conditions for the reverse order law in terms of the defining spaces alone. Nevertheless, in general it will not be possible to compute sums and intersections of infinite-dimensional subspaces. For algorithmically checking the conditions of Theorem 5.1 or 5.2 and for computing the reverse order law in the form (6), we consider finite (co)dimensional spaces and Fredholm operators.

Recall that a linear map T between vector spaces is called *Fredholm* operator if $\dim \mathrm{Ker}T < \infty$ and $\mathrm{codim}\,\mathrm{Im}T < \infty$. Moreover, for finite codimensional subspaces $V_1 \leq V$, we have $\mathrm{codim}\,V_1 = \dim V_1^\perp$. In this case, V_1 can be implicitly represented by

the finite-dimensional subspace $V_1^\perp \leq V^*$. For an application to linear ordinary boundary problems, see the next section.

We assume that for finite-dimensional subspaces, we can compute sums and intersections and check inclusions, both in vector spaces and in their duals. With the following lemma, the intersection of a finite-dimensional subspace with a finite codimensional subspace is reduced to computing kernels of matrices.

Definition 8.1 Let $u = (u_1, \ldots, u_m)^T \in V^m$ and $\beta = (\beta_1, \ldots, \beta_n)^T \in (V^*)^n$. We call

$$\beta(u) = \begin{pmatrix} \beta_1(u_1) & \ldots & \beta_1(u_m) \\ \vdots & \ddots & \vdots \\ \beta_n(u_1) & \ldots & \beta_n(u_m) \end{pmatrix} \in F^{n\times m}$$

the *evaluation matrix* of β and u.

LEMMA 8.2 *Let $U \leq V$ and $\mathscr{B} \leq V^*$ be generated respectively by $u = (u_1, \ldots, u_m)$ and $\beta = (\beta_1, \ldots, \beta_n)$. Let $k^1, \ldots, k^r \in F^m$ be a basis of* $\mathrm{Ker}\beta(u)$*, and $\kappa^1, \ldots, \kappa^s \in F^n$ a basis of* $\mathrm{Ker}(\beta(u))^T$*. Then*

(i) *$U \cap \mathscr{B}^\perp$ is generated by $\sum_{i=1}^m k_i^1 u_i, \ldots, \sum_{i=1}^m k_i^r u_i$ and*
(ii) *$U^\perp \cap \mathscr{B}$ is generated by $\sum_{i=1}^n \kappa_i^1 \beta_i, \ldots, \sum_{i=1}^n \kappa_i^s \beta_i$.*

Proof A linear combination $v = \sum_{\ell=1}^m c_\ell u_\ell$ is in $\mathscr{B}^\perp$ if and only if $\beta_i(v) = 0$ for $1 \leq i \leq n$, that is, $\sum_{\ell=1}^m c_\ell \beta_i(u_\ell) = 0$ for $1 \leq i \leq n$. Hence $\beta(u) \cdot (c_1, \ldots, c_m)^T = 0$. Analogously, one sees that the coefficients of linear combination in $U^\perp \cap \mathscr{B}$ are in the kernel of $(\beta(u))^T$. □

We reformulate the conditions of Theorem 5.1 such that for Fredholm operators they only involve operations on finite-dimensional subspaces and intersections like in the previous lemma. Similarly, one can rewrite the conditions of Theorem 5.2.

COROLLARY 8.3 *Let $T_1 \colon V \to W$ and $T_2 \colon U \to V$ be linear with outer inverses $G_1 =$* $\mathrm{O}(T_1, \mathscr{B}_1, E_1)$ *and $G_2 =$* $\mathrm{O}(T_2, \mathscr{B}_2, E_2)$*. Let $\mathscr{C}_2 = T_2(\mathscr{B}_2^\perp)^\perp$ and $K_1 = T_1^{-1}(E_1)$. The following conditions are equivalent:*

(i) *G_2G_1 is an outer inverse of T_1T_2.*
(ii) *$\mathscr{C}_2 + (\mathscr{B}_1 \cap E_2^\perp) \geq \mathscr{B}_1 \cap (E_2 \cap K_1)^\perp$*
(iii) *$\mathscr{B}_1 \geq \mathscr{C}_2 \cap (E_2 \cap \mathscr{B}_1^\perp)^\perp \cap (E_2 \cap K_1)^\perp$*
(iv) *$K_1 \oplus (E_2 \cap \mathscr{B}_1^\perp) \geq E_2 \cap (\mathscr{B}_1 \cap \mathscr{C}_2)^\perp$*
(v) *$E_2 \geq K_1 \cap (\mathscr{B}_1 \cap E_2^\perp)^\perp \cap (\mathscr{B}_1 \cap \mathscr{C}_2)^\perp$*

Proof Taking the orthogonal of both sides of 5.1 (ii) and (iii), respectively, and applying Proposition A.1 we get (ii) and (iii). For (iv) and (v), we can apply Proposition A.1 directly to the corresponding conditions of Theorem 5.1. □

We note that using Lemma 8.2, it also possible to determine constructively the implicit representation (6) of a product of generalized inverses; see the next section.

9. Examples for linear ordinary boundary problems

As an example involving infinite dimensional spaces and Fredholm operators, we consider solution (Green's) operators for linear ordinary boundary problems. Algebraically, linear boundary problems can be represented as a pair $(T, \mathscr{B})$, where $T\colon V \to W$ is a surjective linear map and $\mathscr{B} \leq V^*$ is an orthogonally closed subspace of (homogeneous) boundary conditions. We say that $v \in V$ is a solution of $(T, \mathscr{B})$ for a given $w \in W$ if $Tv = w$ and $v \in \mathscr{B}^\perp$.

For a regular boundary problem (having a unique solution for every right-hand side), the Green's operator is defined as the unique right inverse G of T with $\mathrm{Im}G = \mathscr{B}^\perp$; see [28] for further details. The product G_2G_1 of the Green's operators of two boundary problems $(T_1, \mathscr{B}_1)$ and $(T_2, \mathscr{B}_2)$ is then the Green's operator of the regular boundary problem $(T_1T_2, \mathscr{B}_2 \oplus T_2^*(\mathscr{B}_1))$, see also Theorem 6.2.

For boundary problems having at most one solution, that is $\mathscr{B}^\perp \cap \mathrm{Ker}T = \{0\}$, the linear algebraic setting has been extended in [23] by defining generalized Green's operators as outer inverses. More specifically, one first has to project an arbitrary right-hand side $w \in W$ onto $T(\mathscr{B}^\perp)$, the image of the 'functions' satisfying the boundary conditions, along a complement E of $T(\mathscr{B}^\perp)$. The corresponding generalized Green's operator is defined as the outer inverse $G = \mathrm{O}(T, \mathscr{B}, E)$, and we refer to $E \leq W$ as an *exceptional space* for the boundary problem $(T, \mathscr{B})$.

The question when the product of two outer inverses is again an outer inverse, is the basis for factoring boundary problems into lower order problems; see [28,29] for the case of regular boundary problems. This, in turn, provides a method to factor certain integral operators.

As an example, let us consider the boundary problem

$$\boxed{\begin{aligned} u'' &= f \\ u'(0) = u'(1) &= u(1) = 0. \end{aligned}} \tag{7}$$

In the above setting, this means we consider the pair $(T_1, \mathscr{B}_1)$ with $T_1 = \mathrm{D}^2$ and $\mathscr{B}_1 = \mathrm{span}(\mathrm{E}_0\,\mathrm{D}, \mathrm{E}_1\,\mathrm{D}, \mathrm{E}_1)$, where D denotes the usual derivation on smooth functions and E_c the evaluation at $c \in \mathbb{R}$. The boundary problem is only solvable for *forcing functions* f satisfying the *compatibility condition* $\int_0^1 f(\xi)\,\mathrm{d}\xi = 0$; more abstractly, we have $T_1(\mathscr{B}_1^\perp) = \mathscr{C}_1^\perp$ with $\mathscr{C}_1 = \mathrm{span}(\int_0^1)$, where $\int_0^1$ denotes the functional $f \mapsto \int_0^1 f(\xi)\,\mathrm{d}\xi$. For computing a generalized Green's operator of $(T_1, \mathscr{B}_1, E_1)$, we have to project f onto $\mathscr{C}_1^\perp$ along a fixed complement E_1. In [30], we computed the generalized Green's operator

$$G_1(f) = x\int_0^x f(\xi)\,\mathrm{d}\xi - \int_0^x \xi f(\xi)\,\mathrm{d}\xi - \frac{1}{2}(x^2+1)\int_0^1 f(\xi)\,\mathrm{d}\xi + \int_0^1 \xi f(\xi)\,\mathrm{d}\xi$$

of (7) for $E_1 = \mathbb{R}$ being the constant functions. It is easy to see that in this case we have $T_1^{-1}(E_1) = \mathrm{span}(1, x, x^2)$.

As a second boundary problem, we consider

$$\boxed{\begin{aligned} u'' - u &= f \\ u'(0) = u'(1) &= u(1) = 0, \end{aligned}}$$

or $(T_2, \mathscr{B}_2)$ with $T_2 = \mathrm{D}^2 - 1$ and $\mathscr{B}_2 = \mathrm{span}(\mathrm{E}_0\,\mathrm{D}, \mathrm{E}_1\,\mathrm{D}, \mathrm{E}_1)$. For the corresponding generalized Green's operator G_2 with exceptional space $E_2 = \mathrm{span}(x)$, we will now check

if the products G_1G_2 and G_2G_1 are again generalized Green's operators of $T_1T_2 = T_2T_1 = \mathrm{D}^4 - \mathrm{D}^2$, using condition (ii) of Corollary 8.3.

We use the algorithm from [30], implemented in the package `IntDiffOp` for the computer algebra system MAPLE, to compute the compatibility conditions. The algorithm is based on the identity

$$T(\mathscr{B}^\perp)^\perp = G^*(\mathscr{B} \cap (\mathrm{Ker}T)^\perp),$$

for any right inverse G of T, which follows from Propositions A.2 and 3.1. Moreover, a right inverse of the differential operator can be computed by the variation of constants and the intersection $\mathscr{B} \cap (\mathrm{Ker}T)^\perp$ using Lemma 8.2. Thus, we obtain $\mathscr{C}_2 = \mathrm{span}(\int_0^1(\exp(-x)+\exp(x)))$, where $\int_0^1(\exp(-x)+\exp(x))$ denotes the functional $f \mapsto \int_0^1(\exp(-\xi)+\exp(\xi))\, f(\xi)\,\mathrm{d}\xi$.

The space $T_2^{-1}(E_2) = \mathrm{span}(x, \exp(x), \exp(-x))$ can be computed using Proposition 3.1 and a right inverse of the differential operator; this is also implemented in the `IntDiffOp` package. Hence, we have $E_1 \cap T_2^{-1}(E_2) = \{0\}$ and therefore $\mathscr{B}_2 \cap (E_1 \cap T_2^{-1}(E_2))^\perp = \mathscr{B}_2$. Computing $\mathscr{B}_2 \cap E_1^\perp$ with Lemma 8.2 yields $\mathscr{B}_2 \cap E_1^\perp = \mathrm{span}(\mathrm{E}_0\,\mathrm{D}, \mathrm{E}_1\,\mathrm{D})$; thus G_1G_2 is not an outer inverse of the product $T_2T_1 = \mathrm{D}^4 - \mathrm{D}^2$ by Corollary 8.3(ii).

On the other hand, we have $E_2 \cap T_1^{-1}(E_1) = \mathrm{span}(x) = E_2$, hence we know by Corollary 8.3(ii) that G_2G_1 is an outer inverse of $T_1T_2 = \mathrm{D}^4 - \mathrm{D}^2$. Furthermore, by Theorem 6.2 we can determine which boundary problem it solves without computing G_1 and G_2. With Lemma 8.2, we obtain $\mathscr{B}_1^\perp \cap E_2 = \{0\}$ and $\mathscr{B}_1 \cap E_2^\perp = \mathrm{span}(\mathrm{E}_0\,\mathrm{D} - \mathrm{E}_1, \mathrm{E}_1\,\mathrm{D} - \mathrm{E}_1)$. Since applying the transpose T_2^* to $\mathrm{E}_0\,\mathrm{D} - \mathrm{E}_1$ and $\mathrm{E}_1\,\mathrm{D} - \mathrm{E}_1$ corresponds to multiplying $T_2 = \mathrm{D}^2 - 1$ from the right, G_2G_1 is the generalized Green's operator of

$$(\mathrm{D}^4 - \mathrm{D}^2, \mathrm{span}(\mathrm{E}_0\,\mathrm{D}, \mathrm{E}_1\,\mathrm{D}, \mathrm{E}_1, \mathrm{E}_0\,\mathrm{D}^3 - \mathrm{E}_1\,\mathrm{D}^2, \mathrm{E}_1\,\mathrm{D}^3 - \mathrm{E}_1\,\mathrm{D}^2), \mathbb{R})$$

by (6); or, in traditional notation, it solves the boundary problem

$$\boxed{\begin{aligned} &u'''' - u'' = f \\ &u'(0) = u'(1) = u(1) = u'''(0) - u''(1) = u'''(1) - u''(1) = 0, \end{aligned}}$$

with exceptional space $\mathbb{R}$.

Acknowledgements

We would like to thank the anonymous referee for his detailed comments.

Funding

The second author was supported by the Austrian Science Fund [grant number J 3030-N18].

References

[1] Ben-Israel A, Greville TNE. Generalized inverses. 2nd ed. New York: Springer-Verlag; 2003.
[2] De Pierro AR, Wei M. Reverse order law for reflexive generalized inverses of products of matrices. Linear Algebra Appl. 1998;277:299–311.
[3] Erdelyi I. On the 'reverse order law' related to the generalized inverse of matrix products. J. ACM. 1966;13:439–443.

[4] Greville TNE. Note on the generalized inverse of a matrix product. SIAM Rev. 1966;8:518–521.
[5] Liu Q, Wei M. Reverse order law for least squares g-inverses of multiple matrix products. Linear Multilinear Algebra. 2008;56:491–506.
[6] Mitra SK, Bhimasankaram P, Malik SB. Matrix partial orders, shorted operators and applications. Hackensack (NJ): World Scientific Publishing; 2010.
[7] Rao CR, Mitra SK. Generalized inverse of matrices and its applications. New York: Wiley; 1971.
[8] Shinozaki N, Sibuya M. The reverse order law $(AB)^- = B^-A^-$. Linear Algebra Appl. 1974;9:29–40.
[9] Takane Y, Tian Y, Yanai H. On reverse-order laws for least-squares g-inverses and minimum norm g-inverses of a matrix product. Aequationes Math. 2007;73:56–70.
[10] Tian Y, Cheng S. Some identities for Moore-Penrose inverses of matrix products. Linear Multilinear Algebra. 2004;52:405–420.
[11] Werner HJ. When is B^-A^- a generalized inverse of AB? Linear Algebra Appl. 1994;210:255–263.
[12] Dinčić NČ, Djordjević DS. Basic reverse order law and its equivalencies. Aequationes Math. 2013;85:505–517.
[13] Djordjević DS, Dinčić NČ. Reverse order law for the Moore–Penrose inverse. J. Math. Anal. Appl. 2010;361:252–261.
[14] Djordjević DS, Rakočević V. Lectures on generalized inverses. Niš: Faculty of Sciences and Mathematics, University of Niš;2008.
[15] Nashed M. Inner, outer, and generalized inverses in Banach and Hilbert spaces. Numer. Funct. Anal. Optimiz. 1987;9:261–325.
[16] Groß J, Trenkler G. On the product of oblique projectors. Linear Multilinear Algebra. 1998;44:247–259.
[17] Werner HJ. G-inverses of matrix products. In: Data analysis and statistical inference. Bergisch Gladbach: Eul-Verlag; 1992. p. 531–546.
[18] Takane Y, Yanai H. On oblique projectors. Linear Algebra Appl. 1999;289:297–310.
[19] Baksalary JK, Baksalary OM. Commutativity of projectors. Linear Algebra Appl. 2002;341:129–142.
[20] Nashed M, Vortuba G. A unified operator theory of generalized inverses. In: Generalized inverses and applications. New York: Academic Press; 1976. p. 1–109.
[21] Searle SR. Linear models. New York: Wiley; 1971.
[22] Brown AL, Page A. Elements of functional analysis. London: Van Nostrand Reinhold Company; 1970.
[23] Korporal A. Symbolic methods for generalized Green's operators and boundary problems. Ph.D. thesis. University of Linz, Research Institute for Symbolic Computation (RISC); 2012.
[24] Groß J. Some remarks concerning the reverse order law. Discuss. Math. Algebra Stochastic Methods. 1997;17:135–141.
[25] Shinozaki N, Sibuya M. Further results on the reverse-order law. Linear Algebra Appl. 1979;27:9–16.
[26] Wibker E, Howe R, Gilbert J. Explicit solutions to the reverse order law $(AB)^+ = B_{mr}^- A_{lr}^-$. Linear Algebra Appl. 1979;25:107–114.
[27] Regensburger G, Rosenkranz M. An algebraic foundation for factoring linear boundary problems. Ann. Mat. Pura Appl.(4) 2009;188:123–151.
[28] Rosenkranz M, Regensburger G. Solving and factoring boundary problems for linear ordinary differential equations in differential algebras. J. Symbolic Comput. 2008;43:515–544.
[29] Korporal A, Regensburger G, Rosenkranz M. Regular and singular boundary problems in Maple. In: Proceedings of CASC (Computer Algebra in Scientific Computing), Vol. 6885. Berlin: Springer; 2011. p. 280–293.
[30] Köthe G. Topological vector spaces. Vol. I. New York: Springer; 1969.

Appendix A. Duality

In the appendix, we summarize duality results for arbitrary vector spaces and their duals that generalize the standard duality for finite-dimensional vector spaces but do not require any topological assumptions; see [30, Sections 9.2 and 9.3] and [28] for further details. The notation should also remind of the analogous and well-known results for Hilbert spaces.

Let V and W be vector spaces over a field F and $\langle , \rangle \colon V \times W \to F$ be a bilinear map. For $V_1 \leq V$, we define the orthogonal

$$V_1^\perp = \{w \in W \mid \langle v, w\rangle = 0 \text{ for all } v \in V_1\} \leq W.$$

The orthogonal $W_1^\perp$ for $W_1 \leq W$ is defined analogously. A subspace U is called orthogonally closed if $U = U^{\perp\perp}$. It follows directly from the definition that for all subsets $X_1, X_2 \subseteq V$, we have $X_1 \subseteq X_2 \Rightarrow X_1^\perp \supseteq X_2^\perp$ and $X_1 \subseteq X_1^{\perp\perp}$; and the same holds for subsets of W. Let $\mathbb{P}(V)$ denote the projective geometry of V, that is, the partially ordered set (poset) of all subspaces ordered by inclusion. Then we have an order-reversing Galois connection between $\mathbb{P}(V)$ and $\mathbb{P}(W)$ defined by $U \mapsto U^\perp$.

We now consider the canonical bilinear form $V \times V^* \to F$ of a vector space V and its dual V^* defined by $\langle v, \beta\rangle \mapsto \beta(v)$. Then every subspace $V_1 \leq V$ is orthogonally closed with respect to the canonical bilinear form, and every finite-dimensional subspace $\mathscr{B} \leq V^*$ is orthogonally closed. The Galois connection gives an order-reversing bijection between $\mathbb{P}(V)$ and the poset of all orthogonally closed subspaces of V^*. So we can describe any subspace $V_1 \leq V$ implicitly by the corresponding orthogonally closed subspace $V_1^\perp$. We denote the poset of all orthogonally closed subspaces of V^* with $\overline{\mathbb{P}}(V^*)$.

The projective geometry $\mathbb{P}(V)$ is a modular lattice, where join and meet are defined as the sum and intersection of subspaces. Modularity means that for all $V_1, V_2, V_3 \in \mathbb{P}(V)$ with $V_1 \leq V_3$ we have

$$V_1 + (V_2 \cap V_3) = (V_1 + V_2) \cap V_3. \tag{A1}$$

Moreover, for spaces $V_1 \leq V_3$ and $V_2 \leq V_4$, we have

$$V = V_1 + V_2 = V_3 \oplus V_4 \;\Rightarrow\; V_1 = V_3 \text{ and } V_2 = V_4, \tag{A2}$$

since $V_3 \cap V_4 = \{0\}$ implies $V_3 = (V_1 \oplus V_2) \cap V_3 = V_1$ and $V_4 = (V_1 \oplus V_2) \cap V_4 = V_2$.

One can also show that $\overline{\mathbb{P}}(V^*)$ is a modular lattice, where the meet is the intersection and the join is the orthogonal closure of the sum of subspaces. Using this fact, one can prove in particular that the sum of two orthogonally closed subspaces is orthogonally closed. The following theorem summarizes Section 9.3 of [30].

Proposition A.1 *The map $V_1 \mapsto V_1^\perp$ gives an order-reversing lattice isomorphism with inverse $\mathscr{B}_1 \mapsto \mathscr{B}_1^\perp$ between the complemented modular lattices $\mathbb{P}(V)$ and $\overline{\mathbb{P}}(V^*)$. In particular, the intersection of orthogonally closed subspaces in V^* is orthogonally closed and*

$$(V_1 + V_2)^\perp = V_1^\perp \cap V_2^\perp \quad \textit{and} \quad (\mathscr{B}_1 \cap \mathscr{B}_2)^\perp = \mathscr{B}_1^\perp + \mathscr{B}_2^\perp.$$

The sum of two orthogonally closed subspaces in V^* is orthogonally closed and

$$(V_1 \cap V_2)^\perp = V_1^\perp + V_2^\perp \quad \textit{and} \quad (\mathscr{B}_1 + \mathscr{B}_2)^\perp = \mathscr{B}_1^\perp \cap \mathscr{B}_2^\perp.$$

Furthermore, orthogonality preserves direct sums, such that

$$V = V_1 \oplus V_2 \;\Rightarrow\; V^* = V_1^\perp \oplus V_2^\perp \quad \textit{and} \quad V^* = \mathscr{B}_1 \oplus \mathscr{B}_2 \;\Rightarrow\; V = \mathscr{B}_1^\perp \oplus \mathscr{B}_2^\perp.$$

For a linear map $A \colon V \to W$ between vector spaces, the *transpose* $A^* \colon W^* \to V^*$ is defined by $\gamma \mapsto \gamma \circ A$. The transposition map $A \mapsto A^*$ from $L(V, W)$ to $L(W^*, V^*)$ is linear, and it is injective since for all $w \neq 0$ there exists a linear map $h \in W^*$ with $h(w) \neq 0$. Moreover, the transpose of a composition is given by $(A_1 A_2)^* = A_2^* A_1^*$.

The image of an orthogonally closed space under the transpose map is orthogonally closed, and we have following identities, see, for example, [28, Prop. A.6].

PROPOSITION A.2 *Let V and W be vector spaces and $A\colon V \to W$ be linear. Then*

$$A(V_1)^\perp = (A^*)^{-1}(V_1^\perp), \qquad A(\mathscr{B}_1^\perp) = (A^*)^{-1}(\mathscr{B}_1)^\perp,$$
$$A^*(\mathscr{C}_1)^\perp = A^{-1}(\mathscr{C}_1^\perp), \qquad A^*(W_1^\perp) = A^{-1}(W_1)^\perp,$$

for subspaces $V_1 \leq V$, $W_1 \leq W$, $\mathscr{C}_1 \leq W^*$ and orthogonally closed subspaces $\mathscr{B}_1 \leq V^*$. In particular,

$$(\mathrm{Im}A)^\perp = \mathrm{Ker}A^*, \quad \mathrm{Im}A = (\mathrm{Ker}A^*)^\perp, \quad (\mathrm{Im}A^*)^\perp = \mathrm{Ker}A, \quad \mathrm{Im}A^* = (\mathrm{Ker}A)^\perp,$$

for the image and kernel of A and A^*.

The property of being a projector, outer/inner/algebraic generalized inverse carries over to the transpose.

PROPOSITION A.3 *A linear map $P\colon V \to V$ is a projector if and only if its transpose P^* is a projector. A linear map $G\colon W \to V$ is an outer/inner/algebraic generalized inverse of $T\colon V \to W$ if and only if G^* is an outer/inner/algebraic generalized inverse of T^*.*

Proof This follows from the defining equations for these properties. For example, if G is an outer inverse of T, we have $G^*T^*G^* = (GTG)^* = G^*$, and the reverse implication follows from the injectivity of the transposition map. □

With the results of this section, we obtain the following duality principle for generalized inverses.

Remark A.4 Given a valid statement for linear maps on arbitrary vector spaces V over a common field involving inclusions, $\{0\}$ and V, sums and intersections, direct sums, kernels and images, projectors, and outer/inner/algebraic generalized inverses, we obtain a valid dual statement by

- reversing the order of the linear maps and the corresponding domains and codomains,
- reversing inclusions and interchanging V and $\{0\}$,
- interchanging sums and intersections,
- interchanging kernels and images.

For example, one easily checks that in Proposition 2.2, the statements (v)–(vii) are the duals of (ii)–(iv) in this sense.

Journal of Symbolic Computation 85 (2018) 247–274

Contents lists available at ScienceDirect

Journal of Symbolic Computation

www.elsevier.com/locate/jsc

Algorithmic operator algebras via normal forms in tensor rings ☆

Jamal Hossein Poor [a], Clemens G. Raab [b], Georg Regensburger [b]

[a] *Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, 4040 Linz, Austria*
[b] *Institute for Algebra, Johannes Kepler University Linz (JKU), 4040 Linz, Austria*

ARTICLE INFO

Article history:
Received 30 November 2016
Accepted 1 April 2017
Available online 17 July 2017

Keywords:
Operator algebra
Tensor ring
Integro-differential operators
Linear substitutions
Noncommutative Gröbner basis
Reduction systems
Completion
Confluence

ABSTRACT

We propose a general algorithmic approach to noncommutative operator algebras generated by additive operators using quotients of tensor rings that are defined by tensor reduction systems. Skew polynomials are a well-established tool covering many cases arising in applications. However, integro-differential operators over an arbitrary integro-differential algebra do not fit this structure, for example. Instead of using parametrized Gröbner bases in free algebras, as has been used so far in the literature, we use Bergman's basis-free analog in tensor rings. Since reduction rules are given by module homomorphisms, the tensor setting often allows for a finite reduction system. A confluent tensor reduction system enables effective computations based on normal forms. Using tensor rings, we can also model integro-differential operators with matrix coefficients, where constants are not commutative.

To have smaller reduction systems, we develop a generalization of Bergman's setting. It allows overlapping domains of reduction homomorphisms, which also make the algorithmic verification of the confluence criterion more efficient. Moreover, we discuss a heuristic approach to complete a given reduction system to a confluent one in analogy to Buchberger's algorithm and Knuth–Bendix completion. Integro-differential operators are used to illustrate the tensor setting, verification of confluence, and completion of tensor reduction systems. We also introduce a confluent reduction system and normal forms for integro-differential operators with linear substitutions, which have applications in delay differential equations. Verification of the confluence criterion and

☆ All authors were supported by the Austrian Science Fund (FWF): P27229.
E-mail addresses: jamal.hossein.poor@ricam.oeaw.ac.at (J. Hossein Poor), clemens.raab@jku.at (C.G. Raab), georg.regensburger@jku.at (G. Regensburger).

http://dx.doi.org/10.1016/j.jsc.2017.07.011

completion based on S-polynomial computations is supported by the Mathematica package TenReS.

1. Introduction

Skew polynomial rings are used in the literature for an algebraic and algorithmic treatment of many common operators like differential and difference operators; see e.g. the works by Chyzak and Salvy (1998); Li (2002); Bueso et al. (2003); Chyzak et al. (2005); Levandovskyy (2005) or the recent overview by Gómez-Torrecillas (2014). Normal forms for skew polynomials are given by the standard polynomial basis. However, normal forms for univariate integral operators are sums of terms of the form $f \int g$. We show that quotients of tensor rings are useful for algebraic modeling of and algorithmic computations with additive operators. The framework provided uses a quotient of a tensor ring by a two-sided ideal for constructing a ring of operators, constructing quotients of such rings of operators by one-sided ideals would be a separate problem. Tensor rings naturally capture the multiadditivity of composition of additive operators. In addition, they allow basis-free treatment of multiplication operators resp. coefficients. In particular, the coefficient ring is not required to be finitely presented. Moreover, for integro-differential operators, they also cover arbitrary rings of constants which neither have to be fields nor commutative rings but need to contain a unit element.

We are not aware that tensor reduction systems in tensor rings have been used so far in the literature for an algorithmic treatment of operator algebras. For applications of noncommutative Gröbner bases in the free polynomial algebra to operator algebras, we refer to Helton et al. (1998), Helton and Stankus (1999), Rosenkranz et al. (2003) and the references on integro-differential operators in Section 4. An overview on Gröbner-Shirshov bases for various algebraic structures is given in Bokut and Chen (2014); see, in particular, Guo et al. (2013), Gao et al. (2014), Gao et al. (2015), Gao and Guo (2017) in connection with differential type, integro-differential, and Rota–Baxter type operators.

For computing in quotients of tensor rings by two-sided ideals, we use Bergman's analog (Bergman, 1978) of Gröbner bases in tensor rings, which we explain in Section 2 along with the underlying algebraic structures. Bergman's confluence criterion for tensor reduction systems involves computations in the tensor ring, but determining the structure of normal forms reduces to a combinatorial problem on words. We generalize Bergman's tensor setting in Section 3 by introducing the concept of specialization. As a first example for our setting with specialization, we present integro-differential operators (IDOs) over an arbitrary integro-differential ring in Section 4. There we give a confluent tensor reduction system together with the corresponding normal forms. In Section 5, we introduce IDOs with linear substitutions. For completing a tensor reduction system to a confluent one, we give a heuristic method along the lines of Buchberger's algorithm in Section 6 and we discuss various problems arising in this context. In each section, we comment about the computational aspects. The Mathematica package TenReS can be obtained at http://gregensburger.com/softw/tenres/ along with example files; see also Hossein Poor et al. (2016b) for further details on the package.

Throughout this paper rings are not necessarily commutative unless stated otherwise, but they are always assumed to have a unit element (of multiplication). Furthermore, we use operator notation, e.g. we write $\varphi 1$ instead of $\varphi(1)$ or $\partial f g = (\partial f)g + f\partial g$ for the Leibniz rule $\partial(fg) = \partial(f)g + f\partial(g)$. All our operators act from the left, in particular, a product AB acts on f as $(A \circ B)(f)$.

1.1. Comparison with conference paper

A two-level version of Bergman's setting in tensor algebras has been introduced already in Hossein Poor et al. (2016a). In contrast, in the present paper we deal with the more general structure of tensor rings instead of tensor algebras. We introduce a generalization and simplification of the two-level tensor setting in Section 3. New aspects treated are deletion criteria for excluding ambiguities from consideration (see Section 2.3.1) and the heuristic completion process discussed in Section 6.

The example presented in Section 4 is more general as it allows also noncommutative differential rings and Section 5 contains an entirely new example.

We also need to correct some minor mistakes in Hossein Poor et al. (2016a). The definition of Φ in Eq. (8) should include the requirement that $\varphi 1 = 1$. Lemma 4.2 should be replaced by the weaker statement of Lemma 15 of the present paper, the proof of Theorem 4.6 needs to be adapted accordingly, cf. the proof of Theorem 20. Also, the equation immediately before Lemma 4.4 has to be replaced by the equation immediately before Lemma 17 in the present paper.

1.2. Introductory example

We use the well-known example of differential operators to briefly discuss several approaches for modeling rings of operators. Recall that differential operators with polynomial coefficients (Weyl algebra) over a field $K \supseteq \mathbb{Q}$ can be defined as the quotient algebra

$$K\langle X, D\rangle/(DX - XD - 1)$$

of the free polynomial algebra $K\langle X, D\rangle$ by a two-sided ideal; see for example Coutinho (1995). Let now (R, ∂) be a commutative differential ring and let K denote its ring of constants. If R is a finitely presented K-algebra, then also the differential operators $R\langle\partial\rangle$ are a finitely presented K-algebra analogous to the Weyl algebra.

Skew polynomials are a well-established approach that only introduces finitely many rules for differential operators over arbitrary differential rings R (e.g. rational functions): they are represented by defining a multiplication on normal forms $\sum f_i \partial^i$ based on the commutation rule

$$\partial \cdot f = f\partial + \partial f.$$

Viewed as construction by generators and relations, this amounts to (potentially) infinitely many relations, one for each generator of R.

In the following, we motivate and illustrate informally tensor reduction systems. For a commutative differential ring, the construction leads to a quotient of the tensor algebra as in Hossein Poor et al. (2016a). The commutation rule for skew polynomials above corresponds to a reduction homomorphism for tensors below. The ring R is regarded as the coefficient ring of skew polynomials, whereas in the tensor construction below R is just considered as a K-module and we tensor over the ring K only. Hence for the multiplication in R, we need to introduce an additional reduction homomorphism for tensors.

Example 1. Consider a commutative differential ring (R, ∂) and let K denote its ring of constants. By the Leibniz rule, the derivation $\partial\colon R \to R$ is a K-module homomorphism. Since R is commutative, also the multiplication operators induced by $f \in R$ mapping $g \mapsto fg$ are K-module homomorphisms. Let $M_\mathsf{D} = K\partial$ denote the free left K-module generated by the symbol ∂. The identities in the K-tensor algebra $K\langle M\rangle$ on the K-module $M = R \oplus K\partial$ reflect the identities coming from the K-linearity of the operators and their compositions, where the tensor product is interpreted as composition of operators.

To incorporate the additional identities, we use reduction rules defined by K-module homomorphisms on certain submodules of the tensor algebra. Corresponding to the composition of multiplication operators and the Leibniz rule, we consider two homomorphisms defined by

$$f \otimes g \mapsto fg \quad \text{and} \quad \partial \otimes f \mapsto f \otimes \partial + \partial f.$$

These two reduction rules induce the two-sided ideal $J = (f \otimes g - fg, \partial \otimes f - f \otimes \partial - \partial f \mid f, g \in R)$ which we use to define the K-algebra of differential operators as the quotient algebra

$$R\langle\partial\rangle = K\langle M\rangle/J.$$

We want to obtain unique normal forms in the quotient by applying the reduction rules above. A tensor of the form

$$\partial \otimes f \otimes g$$

corresponds to an overlap ambiguity of these two rules, since it can be reduced by the homomorphisms in different ways to obtain either

$$(f \otimes \partial + \partial f) \otimes g \quad \text{or} \quad \partial \otimes (fg).$$

For checking resolvability of the ambiguity the S-polynomial formed by the difference of these alternatives should be reducible to zero. In the present case, it reduces to zero because of the Leibniz rule in R. More explicitly, for all $f, g \in R$ we have

$$\begin{aligned} \mathrm{SP}(\partial \otimes f, f \otimes g) &= (f \otimes \partial + \partial f) \otimes g - \partial \otimes (fg) \\ &\to f \otimes g \otimes \partial + f \otimes \partial g + (\partial f)g - fg \otimes \partial - \partial(fg) \\ &\to fg \otimes \partial + f\partial g + (\partial f)g - fg \otimes \partial - \partial(fg) \\ &= f\partial g + (\partial f)g - \partial(fg) = 0. \end{aligned}$$

Another ambiguity is expressed by tensors of the form $f \otimes g \otimes h$ and is resolvable as well. Since all ambiguities are resolvable, we obtain normal forms in terms of irreducible tensors

$$\partial^{\otimes j} \quad \text{and} \quad f \otimes \partial^{\otimes j}. \qquad \square$$

For differential operators with matrix coefficients, we let R be a ring of matrices over some (commutative) differential ring. Then not only R is a noncommutative differential ring, but also its ring of constants K is no longer commutative and elements of K do not commute with elements of R. Consequently, R is not a K-algebra anymore. More generally, we consider an arbitrary differential ring R. It is a bimodule over its ring of constants K and tensoring over K leads to a construction of the differential operators as a quotient of the tensor ring instead of the tensor algebra.

Example 2. For an arbitrary (not necessarily commutative) differential ring (R, ∂), ∂ is a K-bimodule homomorphism of R whereas multiplication operators $g \mapsto fg$ in general are only right K-module homomorphisms. We consider the K-tensor ring $K\langle M\rangle$ on the K-bimodule $M = R \oplus M_{\mathrm{D}}$, where M_{D} is a K-bimodule non-freely generated by ∂. The identities in the tensor ring $K\langle M\rangle$ reflect the identities coming from the additivity of the operators and their compositions. Reduction rules are K-bimodule homomorphisms defined by the same formulae as above. For details see Example 8 later. $\square$

2. Tensor reduction systems

In this section, we describe analogs of Gröbner bases in tensor rings following Bergman (1978) using standard notation for rewriting systems from Baader and Nipkow (1998). First we outline the construction and some properties of the K-tensor ring $K\langle M\rangle$ on a K-bimodule M over a arbitrary ring K with unit element. If K is commutative and the left and right scalar multiplication on M agree, then $K\langle M\rangle$ is the tensor algebra on M, which is a generalization of the noncommutative polynomial algebra on a set of indeterminates. In contrast to the noncommutative polynomials, in the tensor ring the "coefficients" in K do not commute with the "indeterminates". For further details on tensor rings and proofs see, for example, Cohn (2003), Rowen (1991). A Gröbner basis theory for free bimodules has been presented in Kobayashi (2005) and for bimodules over Poincaré-Birkhoff-Witt (PBW) algebras in Román García and Román García (2005), Levandovskyy (2005).

2.1. Basics of tensor rings

From now, K denotes a ring (not necessarily commutative) with unit element. A K-bimodule is a left K-module M which is also a right K-module satisfying the associativity condition $(km)l = k(ml)$ for all $m \in M$ and $k, l \in K$. By a K-ring we understand a ring R that is a K-bimodule such that $(xy)z = x(yz)$ for any x, y, z in R or K. Even when K is commutative, the notion of K-ring is more general than the notion of K-algebra, because the action of K need not centralize the ring, that is, we do not require $kr = rk$ for $k \in K$ and $r \in R$. In other words, the difference can be described by saying

that whereas a K-algebra (K commutative) is a ring R with a homomorphism from K to the center of R, a K-ring is a ring R with a ring homomorphism from K to R. In particular, if K is a subring of some ring R, then R is a K-ring.

We first recall basic properties of the tensor product on K-bimodules. Let $M_1, \dots, M_n$ be K-bimodules. Given an abelian group $(A, +)$, we say that $\beta\colon M_1 \times \cdots \times M_n \to A$ is a balanced map if it is multiadditive and it satisfies

$$\beta(m_1, \dots, m_i k, m_{i+1}, \dots, m_n) = \beta(m_1, \dots, m_i, k m_{i+1}, \dots, m_n)$$

for all $k \in K$, $m_j \in M_j$, where $i = 1, \dots, n-1$ and $j = 1, \dots, n$. By the definition of the tensor product, there exists an abelian group $M_1 \otimes \cdots \otimes M_n$ together with a balanced map

$$\otimes\colon M_1 \times \cdots \times M_n \to M_1 \otimes \cdots \otimes M_n.$$

We write $m_1 \otimes \cdots \otimes m_n$ for the image of $(m_1, \dots, m_n)$ under $\otimes$. The universal property of the tensor product states that if $\beta\colon M_1 \times \cdots \times M_n \to A$ is any balanced map, then there exists a unique homomorphism $\overline{\beta}\colon M_1 \otimes \cdots \otimes M_n \to A$ such that

$$\overline{\beta}(m_1 \otimes \cdots \otimes m_n) = \beta(m_1, \dots, m_n).$$

Note that, if $M_1, \dots, M_n$ are K-bimodules, then $M_1 \otimes \cdots \otimes M_n$ is again a K-bimodule with scalar multiplications

$$k(m_1 \otimes \cdots \otimes m_n) = k m_1 \otimes \cdots \otimes m_n \quad \text{and} \quad (m_1 \otimes \cdots \otimes m_n)k = m_1 \otimes \cdots \otimes m_n k.$$

We denote the tensor product of M with itself over K by $M^{\otimes n} = M \otimes \cdots \otimes M$ (n factors) and its elements are called *tensors*. In particular, $M^{\otimes 1} = M$ and we interpret $M^{\otimes 0}$ as the K-bimodule $K\epsilon$, where ϵ denotes the empty tensor. Elements of the form $m_1 \otimes \cdots \otimes m_n \in M^{\otimes n}$ with $m_1, \dots, m_n \in M$, are called *pure tensors* and they generate $M^{\otimes n}$ as a K-bimodule. As a K-bimodule, the tensor ring $K\langle M\rangle$ is defined as the direct sum $K\langle M\rangle = \bigoplus_{n=0}^{\infty} M^{\otimes n}$ with multiplication $M^{\otimes r} \times M^{\otimes s} \to M^{\otimes (r+s)}$ given by the balanced map

$$(m_1 \otimes \cdots \otimes m_r, \tilde{m}_1 \otimes \cdots \otimes \tilde{m}_s) \mapsto m_1 \otimes \cdots \otimes m_r \otimes \tilde{m}_1 \otimes \cdots \otimes \tilde{m}_s,$$

which can be extended to $K\langle M\rangle$ by biadditivity. In general, the K-bimodule $K\langle M\rangle$ with this multiplication is a ring with ϵ being its unit element. Note that by the homomorphism $K \to K\langle M\rangle$ mapping $k \mapsto k\epsilon$ the tensor ring $K\langle M\rangle$ is a K-ring.

The K-tensor algebra on a K-module M with K commutative is a special case of the K-tensor ring by viewing M as a K-bimodule with identical scalar multiplication from left and right. Note that for a free K-module M with basis X, the K-tensor algebra $K\langle M\rangle$ is isomorphic to the noncommutative polynomial algebra $K\langle X\rangle$. It has the set of all products $x_1 \otimes \cdots \otimes x_n$ for $x_1, \dots, x_n \in X$ as a K-module basis, i.e. elements in $K\langle X\rangle$ have a unique representation as K-linear combinations of such products.

The analogous situation for tensor rings is more involved. The free K-bimodule on a set X is given by $K \otimes_{\mathbb{Z}} \mathbb{Z}X \otimes_{\mathbb{Z}} K$, where $\mathbb{Z}X$ denotes the free left $\mathbb{Z}$-module on X. The K-tensor ring over the free K-bimodule on X is isomorphic to the free K-ring on X, which is generated as a K-bimodule by the set of all products $x_1 \otimes k_2 x_2 \otimes \cdots \otimes k_n x_n$ such that $x_1, \dots, x_n \in X$ and $k_2, \dots, k_n \in K$. Note that the representation of elements of the free K-ring on X in terms of such products is not unique, in contrast to the noncommutative polynomial algebra. Since bimodules have coefficients on both sides and coefficients do not commute with indeterminates, even the free K-bimodule generated by $\{x_1\}$ gives rise to non-uniqueness: $k_1 x_1 k_3 + k_2 x_1 k_1 = k_3 x_1 k_1 + k_1 x_1 k_2$ for $k_3 = k_1 + k_2 \in K$.

2.2. Diamond Lemma in tensor rings

Now we are ready to explain the setting for reduction systems in tensor rings following Bergman (1978, Sec. 6). Let $(M_x)_{x \in X}$ be a family of K-bimodules indexed by a set X. The modules M_x play the role of the indeterminates in the noncommutative polynomial algebra.

We denote the free monoid on X by $\langle X\rangle$ and its unit element by ϵ. The free monoid $\langle X\rangle$ can also be regarded as the word monoid over the alphabet X with ϵ as the empty word. For every word $W = x_1 \dots x_n \in \langle X\rangle$, we denote the tensor product of the corresponding bimodules by

$$M_W := M_{x_1} \otimes \cdots \otimes M_{x_n}.$$

In particular, we have $M_\epsilon = K\epsilon$ for the empty word/tensor ϵ. The pure tensors $m_1 \otimes \cdots \otimes m_n \in M_W$ with $m_i \in M_{x_i}$ play the role of the monomials in the tensor ring. We consider the direct sum

$$M := \bigoplus_{x\in X} M_x \tag{1}$$

and the K-tensor ring on M:

$$K\langle M\rangle = \bigoplus_{n=0}^{\infty} M^{\otimes n} = \bigoplus_{W\in\langle X\rangle} M_W. \tag{2}$$

Every tensor $t \in K\langle M\rangle$ can be written as a sum of pure tensors. However, in contrast to linear combinations of monomials in the noncommutative polynomial algebra, this representation is not unique. This happens because already $M^{\otimes n}$ is not freely generated as a K-bimodule by the pure tensors, e.g. $m_1 \otimes m_3 + m_2 \otimes m_1 = m_3 \otimes m_1 + m_1 \otimes m_2$ in $M^{\otimes 2}$ for $m_3 = m_1 + m_2 \in M$. Still, using bimodule homomorphisms, one can define reductions analogous to polynomial reduction for (non-)commutative Gröbner bases.

Definition 3. Let M be given by Eq. (1). A *reduction rule* for $K\langle M\rangle$ is given by a pair (W, h) of a word $W \in \langle X\rangle$ and a K-bimodule homomorphism $h\colon M_W \to K\langle M\rangle$. For a reduction rule $r = (W, h)$ and words $A, B \in \langle X\rangle$, we define a *reduction* as the K-bimodule homomorphism

$$h_{A,r,B}\colon K\langle M\rangle \to K\langle M\rangle$$

acting as $\mathrm{id}_A \otimes h \otimes \mathrm{id}_B$ on M_{AWB} and the identity on all other M_V with $V \in \langle X\rangle$ and $V \neq AWB$.

For a pure tensor $a \otimes w \otimes b \in M_{AWB}$ with $a \in M_A$, $w \in M_W$, and $b \in M_B$, the reduction $h_{A,r,B}$ is given by

$$a \otimes w \otimes b \mapsto a \otimes h(w) \otimes b.$$

So, as for polynomial reduction, we "replace" the "leading monomial" w by the "tail" $h(w)$ given by the homomorphism h.

Let $t \in K\langle M\rangle$. A reduction $h_{A,r,B}$ acts trivially on t, i.e. $h_{A,r,B}(t) = t$, if the summand of t in M_{AWB} is zero, see Eq. (2). A reduction rule $r = (W, h)$ *reduces* t to $s \in K\langle M\rangle$ if a reduction $h_{A,r,B}$ for some $A, B \in \langle X\rangle$ acts non-trivially on t and $h_{A,r,B}(t) = s$ and we write $t \to_r s$.

A *reduction system* for $K\langle M\rangle$ is a set Σ of reduction rules. Every reduction system Σ induces a *reduction relation* $\to_\Sigma$ on tensors by defining $t \to_\Sigma s$ for $t, s \in K\langle M\rangle$ if $t \to_r s$ for some reduction rule $r \in \Sigma$. Fixing a reduction system Σ, we say that $t \in K\langle M\rangle$ *can be reduced* to $s \in K\langle M\rangle$ by Σ if $t = s$ or there exists a finite sequence of reduction rules $r_1, \dots, r_n$ in Σ such that

$$t \to_{r_1} t_1 \to_{r_2} \cdots \to_{r_{n-1}} t_{n-1} \to_{r_n} s$$

and we write $t \overset{*}{\to}_\Sigma s$. In other words, $\overset{*}{\to}_\Sigma$ denotes the reflexive transitive closure of the reduction relation $\to_\Sigma$.

The set of *irreducible words* $\langle X\rangle_{\mathrm{irr}} \subseteq \langle X\rangle$ consists of those words having no subwords from the set $\{W \mid (W, h) \in \Sigma\}$. We define the K-subbimodule of *irreducible tensors* as

$$K\langle M\rangle_{\mathrm{irr}} = \bigoplus_{W\in\langle X\rangle_{\mathrm{irr}}} M_W. \tag{3}$$

We also need to consider partial orders on $\langle X\rangle$. A *semigroup partial order* on $\langle X\rangle$ is a partial order $\preceq$ on $\langle X\rangle$ such that $B \prec \tilde{B} \Rightarrow ABC \prec A\tilde{B}C$ for all $A, B, \tilde{B}, C \in \langle X\rangle$. If in addition $\epsilon \preceq A$, for all $A \in \langle X\rangle$, then it is called a *monoid partial order*. It is called *Noetherian* if there are no infinite descending chains.

Remark 4. Note that a lexicographic order on $\langle X\rangle$ is not a semigroup order. However, a (weighted) degree-lexicographic order of the words is a semigroup (total) order on $\langle X\rangle$ and it is Noetherian if X is finite. Given a semigroup S with a semigroup partial order $\preceq$ on it and a semigroup homomorphism $\varphi\colon \langle X\rangle \to S$, we can define the induced semigroup partial order on $\langle X\rangle$ by

$$V \leq W :\Leftrightarrow V = W \text{ or } \varphi(V) \prec \varphi(W).$$

For example, for $S = \mathbb{N}$ with the usual order and the homomorphism given by $\varphi(x_0) = 1$ for $x_0 \in X$ and $\varphi(x) = 0$ for $x \in X \setminus \{x_0\}$, the induced partial order just compares the degree in x_0. Given two semigroups S_1 and S_2 with corresponding semigroup partial orders $\leq_1$ and $\leq_2$ respectively, we can combine them lexicographically to obtain a semigroup partial order on $S = S_1 \times S_2$ by

$$(a_1, a_2) \leq (b_1, b_2) :\Leftrightarrow a_1 <_1 b_1 \text{ or } a_1 = b_1 \text{ and } a_2 \leq_2 b_2.$$

A semigroup partial order $\leq$ is *compatible* with a reduction system Σ if for all reduction rules $(W, h) \in \Sigma$,

$$h(M_W) \subseteq \bigoplus_{V<W} M_V.$$

If a compatible semigroup partial order is Noetherian, then there do not exist infinite sequences of reductions in Σ. In other words, the reduction relation $\to_\Sigma$ is *terminating* or *Noetherian*. So, in that case, every $t \in K\langle M\rangle$ can be reduced in finitely many steps to an irreducible tensor

$$t \xrightarrow{*}_\Sigma s \in K\langle M\rangle_{\mathrm{irr}}$$

and such an s is called a *normal form* of t. In general, a tensor can have different normal forms. If $t \in K\langle M\rangle$ has a *unique normal form*, we denote it by $t\downarrow_\Sigma$.

For ensuring unique normal forms for reduction systems on tensor rings, we state below Bergman's analog of Buchberger's criterion for Gröbner bases (Buchberger, 1965). In the context of Gröbner-Shirshov bases for various algebraic structures this is also referred to as the Composition-Diamond Lemma; see e.g. the survey by Bokut and Chen (2014).

Let Σ be a reduction system. We study the cases when two different reductions act non-trivially on tensors in M_W for $W \in \langle X\rangle$.

Definition 5. An *overlap ambiguity* is given by two (not necessarily distinct) reduction rules $(W, h), (\tilde{W}, \tilde{h}) \in \Sigma$ and nonempty words $A, B, C \in \langle X\rangle$ such that

$$W = AB \quad \text{and} \quad \tilde{W} = BC.$$

It is called *resolvable* if for all pure tensors $a \in M_A$, $b \in M_B$, and $c \in M_C$ the *S-polynomial* can be reduced to zero:

$$h(a \otimes b) \otimes c - a \otimes \tilde{h}(b \otimes c) \xrightarrow{*}_\Sigma 0.$$

An *inclusion ambiguity* is given by distinct reduction rules $(W, h), (\tilde{W}, \tilde{h}) \in \Sigma$ and words $A, B, C \in \langle X\rangle$ with $W = B$ and $\tilde{W} = ABC$. It is called *resolvable* if for all pure tensors $a \in M_A$, $b \in M_B$, and $c \in M_C$ the *S-polynomial* can be reduced to zero: $a \otimes h(b) \otimes c - \tilde{h}(a \otimes b \otimes c) \xrightarrow{*}_\Sigma 0$.

With slight abuse of notation, we refer to S-polynomials of an overlap or inclusion ambiguity, respectively, by

$$\mathrm{SP}(A\underline{B}, \underline{B}C) \quad \text{or} \quad \mathrm{SP}(\underline{B}, A\underline{B}C).$$

A reduction system Σ induces the two-sided *reduction ideal*

$$I_\Sigma := (t - h(t) \mid (W, h) \in \Sigma \text{ and } t \in M_W) \subseteq K\langle M\rangle. \tag{4}$$

For studying operator algebras, we want to compute in the factor ring $K\langle M\rangle/I_\Sigma$. If all ambiguities are resolvable, then we can do this effectively using reductions in $K\langle M\rangle$ and the corresponding normal

forms with respect to $\to_\Sigma$. This is the *confluence criterion* (condition 1. below) that we will check algorithmically, for a brief discussion see the following subsection.

Theorem 6. *(Bergman, 1978) Let K be a ring, let $(M_x)_{x\in X}$ be a family of K-bimodules indexed by a set X, and let $M = \bigoplus_{x\in X} M_x$. Let Σ be a reduction system on $K\langle M\rangle$ and let $\leq$ be a Noetherian semigroup partial order on $\langle X\rangle$ that is compatible with Σ. Then the following are equivalent:*

1. *All ambiguities of Σ are resolvable.*
2. *Every $t \in K\langle M\rangle$ has a unique normal form $t\downarrow_\Sigma$.*
3. *$K\langle M\rangle/I_\Sigma$ and $K\langle M\rangle_{irr}$ are isomorphic as K-bimodules.*

If these conditions hold, then we can define a multiplication on $K\langle M\rangle_{irr}$ by $s\cdot t := (s\otimes t)\downarrow_\Sigma$ so that $K\langle M\rangle/I_\Sigma$ and $K\langle M\rangle_{irr}$ are isomorphic as K-rings.

Note that our definition of resolvability above differs from the definition used by Bergman. Actually, he uses two different notions for resolvability of ambiguities, which we briefly describe below. Both of them are weaker than our Definition 5 in general. However, if every tensor has a unique normal form, then all three definitions of resolvability are equivalent. Hence Theorem 6 holds regardless which of these three notions of resolvability is used. One slightly weaker notion only requires the existence of a tensor $t \in K\langle M\rangle$ such that

$$h(a\otimes b)\otimes c \xrightarrow{*}_\Sigma t \xleftarrow{*}_\Sigma a\otimes \tilde{h}(b\otimes c) \quad \text{or} \quad a\otimes h(b)\otimes c \xrightarrow{*}_\Sigma t \xleftarrow{*}_\Sigma \tilde{h}(a\otimes b\otimes c),$$

respectively, in other words, the two different results of the reductions of $a\otimes b\otimes c$ are joinable. Another even weaker notion is the following, which depends on semigroup partial order $\leq$.

Definition 7. We call an overlap or inclusion ambiguity with words $A, B, C \in \langle X\rangle$ *$\leq$-resolvable* if and only if all its S-polynomials are contained in the bimodule I_{ABC} generated by

$$\bigcup_{\substack{V\in\langle X\rangle\\ V<ABC}} \{t - s \mid t \in M_V \text{ and } t \to_\Sigma s \in K\langle M\rangle\}.$$

If the semigroup partial order $\leq$ is compatible with Σ, then this bimodule is contained in a "truncation" $I_\Sigma \cap \bigoplus_{\substack{V\in\langle X\rangle\\ V<ABC}} M_V$ of the reduction ideal I_Σ.

Example 8. We revisit Example 2 to study it formally in the tensor ring setting. Let (R,∂) be a differential ring and let K denote its ring of constants. We consider the K-bimodule $M_\mathsf{R} = R$ (indexed by the letter R). In addition, we consider the free left K-module $M_\mathsf{D} = K\partial$ generated by ∂ (indexed by the letter D), which we view as a K-bimodule with right multiplication defined by

$$c\partial\cdot d = cd\partial,$$

for all $c, d \in K$. This definition is based on left K-linearity of the operation ∂ on R:

$$(c\partial d)f = c\partial(df) = (cd\partial)f.$$

Let $M = M_\mathsf{R}\oplus M_\mathsf{D}$ be the module of basic operators. Then words over the alphabet $X = \{\mathsf{R},\mathsf{D}\}$ index the direct summands of the K-tensor ring $K\langle M\rangle$.

We interpret elements $f \in R$ as multiplication operators, ∂ as the derivation on R, and the tensor product $\otimes$ as the composition of operators. So we consider the reduction system $\Sigma = \{r_{\mathsf{RR}}, r_{\mathsf{DR}}\}$ with the reduction rules

$$r_{\mathsf{RR}} = (\mathsf{RR}, f\otimes g \mapsto fg) \quad \text{and} \quad r_{\mathsf{DR}} = (\mathsf{DR}, \partial\otimes f \mapsto f\otimes\partial + \partial f)$$

corresponding to the composition of multiplication operators and the Leibniz rule. Then the ring of differential operators can be defined as the quotient

$$R\langle\partial\rangle = K\langle M\rangle / I_\Sigma$$

of the tensor ring by the two-sided reduction ideal. The informal definition of the reduction homomorphisms above can be made formal in the following way. First, since

$$M_\mathsf{R} \times M_\mathsf{R} \to M_\mathsf{R}$$
$$(f, g) \mapsto fg$$

is a balanced map, it induces a well-defined homomorphism $M_{\mathsf{RR}} \to M_\mathsf{R}$ of abelian groups. This homomorphism can be verified to be even a K-bimodule homomorphism, which we use to define r_{RR}. Extending the definition

$$\beta(\partial, f) := f \otimes \partial + \partial f$$

by

$$\beta(c\partial, f) := \beta(\partial, cf),$$

we obtain a balanced map $\beta\colon M_\mathsf{D} \times M_\mathsf{R} \to M_{\mathsf{RD}} \oplus M_\mathsf{R}$, since

$$\beta(c\partial \cdot d, f) = \beta(cd\partial, f) = \beta(\partial, cdf) = \beta(c\partial, df).$$

Like above, β induces a K-bimodule homomorphism $M_{\mathsf{DR}} \to M_{\mathsf{RD}} \oplus M_\mathsf{R}$ constituting r_{DR}.

So any semigroup partial order $\leq$ on $\langle X\rangle$ with $\mathsf{RR} > \mathsf{R}$, as well as $\mathsf{DR} > \mathsf{RD}$ and $\mathsf{DR} > \mathsf{R}$ is compatible with Σ, e.g. the degree-lexicographic order with $\mathsf{D} > \mathsf{R}$. There are two overlap ambiguities. The S-polynomials of the first ambiguity reduce to zero in two steps:

$$\mathrm{SP}(\underline{\mathsf{RR}}, \underline{\mathsf{RR}}) = (fg) \otimes h - f \otimes (gh) \to_{r_{\mathsf{RR}}} (fg)h - f(gh) = 0.$$

We already have seen in Example 2 that the S-polynomials $\mathrm{SP}(\underline{\mathsf{DR}}, \underline{\mathsf{RR}})$ reduce to the Leibniz rule in R. Hence by Theorem 6 every $t \in K\langle M\rangle$ has a unique normal form $t{\downarrow_\Sigma}$ in $K\langle M\rangle_{\mathrm{irr}}$, where

$$K\langle M\rangle_{\mathrm{irr}} = K\epsilon \oplus M_\mathsf{R} \oplus M_\mathsf{D} \oplus (M_\mathsf{R} \otimes M_\mathsf{D}) \oplus M_\mathsf{D}^{\otimes 2} \oplus (M_\mathsf{R} \otimes M_\mathsf{D}^{\otimes 2}) \oplus \cdots$$

since $\langle X\rangle_{\mathrm{irr}} = \{\epsilon, \mathsf{R}, \mathsf{D}, \mathsf{RD}, \mathsf{D}^2, \mathsf{RD}^2, \ldots\}$. In other words, $t{\downarrow_\Sigma}$ can be written as a sum of pure tensors of the form $\epsilon,\ f,\ \partial,\ f \otimes \partial,\ \partial \otimes \partial,\ f \otimes \partial \otimes \partial, \ldots$ and we recover the well-known normal forms of differential operators. □

Remark 9. If some $\alpha \in M_x$ corresponds to a left K-linear operator, like $\partial \in M_\mathsf{D}$ above, then for the right scalar multiplication of left multiples of α, we always have

$$c\alpha \cdot d = cd\alpha$$

with $c, d \in K$; see also Eq. (12). As soon as such an operator is present, the ring over which the tensors are formed has to contain K in order to incorporate the corresponding relations directly into the tensor ring.

2.3. Computational aspects

Considering the algorithmic aspects of Theorem 6, we assume that we have a finite reduction system Σ over a finite alphabet X. Moreover, a compatible Noetherian semigroup partial order has to be assumed.

For generating the set of ambiguities, we only need to work in the word monoid $\langle X\rangle$. Likewise, determining the set of irreducible words $\langle X\rangle_{\mathrm{irr}}$ is a purely combinatorial problem on words as well, cf. the proofs of Theorems 27 and 32. For checking resolvability of ambiguities, it suffices to work with S-polynomials constructed from general elements of the basic bimodules M_x. The result of a reduction step, i.e. the application of a homomorphism from the reduction system, needs to be simplified in the tensor ring. This involves application of properties of the tensor product and of identities in the bimodules, like the Leibniz rule in the example above. In practice, the reduction to zero often can be detected heuristically without having a canonical simplifier in the bimodules.

The package `TenReS` provides routines to generate all ambiguities and corresponding S-polynomials of a reduction system given by the user. It also includes routines for computing in the tensor ring. Identities needed for computing in the bimodules of Eq. (1) have to be implemented by the user in each concrete case.

In contrast to specifying new identities in the polynomial resp. term algebra, already the constructive specification of reduction homomorphisms in the tensor setting is not clear in general.

2.3.1. Deletion criteria

For polynomial rings there are two classical deletion criteria for excluding critical pairs from consideration: the *product criterion* and the *chain criterion*. We want to consider their analogs for excluding ambiguities from the confluence check for tensor reduction systems.

There is no need for an analog of the product criterion as it is already built into the definition of ambiguities of tensor reduction rules. If rules $(W, h), (\tilde{W}, \tilde{h}) \in \Sigma$ are such that no word of length less than $|W| + |\tilde{W}|$ contains both W and $\tilde{W}$ as subwords, then the rules do not have any ambiguities among them anyway. Hence we focus only on the chain criterion. The following lemma is an analog of Lemma 5.11 in Mora (1994).

Lemma 10. *Let $\leq$ be a semigroup partial order on $\langle X\rangle$ compatible with the reduction system Σ. Let $r_1, r_2 \in \Sigma$ have an overlap ambiguity with $A, B, C \in \langle X\rangle$, i.e. $r_1 = (AB, g)$ and $r_2 = (BC, h)$. Let $r_3 = (V, f) \in \Sigma$ where V is a subword of $W = ABC$ such that one of the following cases holds.*

1. *V is a subword of $A = LVR$ and the inclusion ambiguity of r_1 and r_3 with L, V, RB is $\leq$-resolvable.*
2. *V is a subword of $B = LVR$ and the two inclusion ambiguities of r_1 and r_3 with AL, V, R and of r_2 and r_3 with L, V, RC are $\leq$-resolvable.*
3. *V is a subword of $C = LVR$ and the inclusion ambiguity of r_2 and r_3 with BL, V, R is $\leq$-resolvable.*
4. *V is a subword of $AB = LVR$ (with nonempty V_1, V_2 such that $V = V_1V_2$ and $B = V_2R$) and the inclusion ambiguity of r_1 and r_3 with L, V, R as well as the overlap ambiguity of r_2 and r_3 with V_1, V_2, RC are $\leq$-resolvable.*
5. *V is a subword of $BC = LVR$ (with nonempty V_1, V_2 such that $V = V_1V_2$ and $B = LV_1$) and the overlap ambiguity of r_1 and r_3 with AL, V_1, V_2 as well as the inclusion ambiguity of r_2 and r_3 with L, V, R are $\leq$-resolvable.*
6. *There are nonempty L, R such that $V = LBR$ (with $A = A_1L$ and $C = RC_2$) and the overlap/inclusion ambiguity of r_1 and r_3 with A_1, LB, R as well as the overlap/inclusion ambiguity of r_2 and r_3 with L, BR, C_2 are $\leq$-resolvable.*

Then the overlap ambiguity of r_1 and r_2 with A, B, C is $\leq$-resolvable.

Proof. For all cases there are canonical choices for W_1, W_2 such that $W = W_1LVRW_2$ (resp. $W = W_1LBRW_2$ in the last case). For a pure tensor $t \in M_W$ we have that the corresponding S-polynomial is equal to $h_{\epsilon,r_1,C}(t) - h_{A,r_2,\epsilon}(t) = t_1 + t_2$ with $t_1 := h_{\epsilon,r_1,C}(t) - t_3$, $t_2 := t_3 - h_{A,r_2,\epsilon}(t)$, and $t_3 := h_{W_1L,r_3,RW_2}(t)$ (resp. $t_3 := h_{W_1,r_3,W_2}(t)$ in the last case). According to Definition 7, we show that $t_1, t_2 \in I_W$.

In Case 3, we directly verify $t_1 = g(a \otimes b) \otimes (c - h_{L,r_3,R}(c)) - (a \otimes b - g(a \otimes b)) \otimes h_{L,r_3,R}(c) \in I_W$ with $a \in M_A$, $b \in M_B$, and $c \in M_C$ such that $t = a \otimes b \otimes c$. Otherwise, by assumption, all S-polynomials of r_1 and r_3 are contained in I_{S_1}, where $S_1 = ABV_2$ in Case 5, $S_1 = ABR$ in Case 6, and $S_1 = AB$ in the remaining cases. Then, there is $m_1 \in M_{T_1}$, where $T_1 \in \langle X\rangle$ is such that $W = S_1T_1$, and an S-polynomial s_1 of r_1 and r_3 such that $t_1 = s_1 \otimes m_1$. Hence $t_1 \in I_{S_1} \otimes M_{T_1} \subseteq I_{S_1T_1} = I_W$.

Analogously, we directly verify $t_2 \in I_W$ in Case 1. In the remaining cases we let $S_2 := BC$ (resp. $S_2 := V_1BC$ in Case 4 and $S_2 := LBC$ in Case 6). Then, we have $t_2 = m_2 \otimes s_2$ for some S-polynomial s_2 of r_2 and r_3 and some $m_2 \in M_{T_2}$, where $T_2 \in \langle X\rangle$ is such that $W = T_2S_2$. We conclude $t_2 \in M_{T_2} \otimes I_{S_2} \subseteq I_W$, since $s_2 \in I_{S_2}$ by assumption.

Consequently, t_1 and t_2 are in I_W in all cases. Hence the same applies to the S-polynomial $h_{\epsilon,r_1,C}(t) - h_{A,r_2,\epsilon}(t)$ and the overlap ambiguity of r_1 and r_2 with A, B, C is $\leq$-resolvable. □

Note that V might be a subword of W in multiple ways, so we need to specify which ambiguities of r_1, r_3 resp. r_2, r_3 are $\leq$-resolvable in order to be able to conclude that the given ambiguity of r_1, r_2 is $\leq$-resolvable. A similar statement can be obtained for inclusion ambiguities of r_1 and r_2.

3. Tensor setting with specialization

Direct application of Bergman's tensor setting requires the sum in Eq. (1) to be direct. As a consequence, domains of reduction rules in a reduction system cannot overlap, even their tensor factors cannot overlap. In order to emulate overlapping domains (or factors), reduction rules have to be split into several smaller parts so that domains of those smaller rules do not overlap. Thus computations with such reduction systems can be inconvenient and inefficient in practice as the smaller rules technically are just individual rules that need to be applied separately. Moreover, this leads to some redundancy in the investigation of ambiguities and S-polynomials. Sticking to the above definition of reduction systems for tensor rings, this situation cannot be avoided.

Example 11. Note that in Example 8 irreducible tensors still have some relations among them when acting as operators. For instance, $k\epsilon \in M^{\otimes 0}$ and $k \in M$ both act by multiplying with $k \in K$. So we need an additional reduction rule reducing $k \in M$ to $k\epsilon \in M^{\otimes 0}$ for $k \in K$. Fixing a direct complement $R = K \oplus \tilde{R}$ in R for defining the reduction rule

$$r_{\mathsf{K}} = (\mathsf{K}, 1 \mapsto \epsilon),$$

would cause the splitting of the rule r_{RR} into four rules $r_{\mathsf{KK}}, r_{\mathsf{K}\tilde{\mathsf{R}}}, r_{\tilde{\mathsf{R}}\mathsf{K}}, r_{\tilde{\mathsf{R}}\tilde{\mathsf{R}}}$ and similarly r_{DR} would split into two rules. The aim of this section is to introduce a framework that allows the rule r_{K} to coexist with r_{RR} and r_{DR}. □

In order to remedy this situation, the aim of this section is to introduce a more flexible tensor setting where the definable reduction systems are much more general. While the induced reduction relations are also more general, the corresponding reduction ideals are not, however.

Definition 12. Let M be a K-bimodule. We call a family $(M_z)_{z\in Z}$ of K-subbimodules of M a *decomposition with specialization*, if $M = \sum_{z\in Z} M_z$ and there exists a subset $X \subseteq Z$ such that

1. we have the direct sum decomposition $M = \bigoplus_{x\in X} M_x$ and
2. for every $z \in Z$ the corresponding module M_z satisfies

$$M_z = \bigoplus_{x\in S(z)} M_x \tag{5}$$

where $S(z) := \{x \in X \mid M_x \subseteq M_z\}$ is the set of *specializations* of z.

Note that this definition implies $S(x) = \{x\}$ for $x \in X$. In the following, we define a framework for tensor reduction systems that are based on such a decomposition with specialization. To this end, we fix a K-bimodule M, alphabets $X \subseteq Z$, and a decomposition $(M_z)_{z\in Z}$ of M with specialization.

For words $W = w_1 \dots w_n \in \langle Z\rangle$ we define the corresponding subbimodule of $K\langle M\rangle$ as before by $M_W := M_{w_1} \otimes \cdots \otimes M_{w_n}$. Because of Eq. (5), any M_W is then a direct sum of certain M_V, $V \in \langle X\rangle$. For a precise statement we can extend the notion of specialization from the alphabet Z to the whole word monoid $\langle Z\rangle$ by the definition below such that we have the following generalization of Eq. (5):

$$M_W = \bigoplus_{V\in S(W)} M_V.$$

Definition 13. For $W = w_1 \dots w_n \in \langle Z\rangle$ we define the set of *specializations* of W by

$$S(W) := \{v_1 \dots v_n \in \langle X\rangle \mid \forall i : v_i \in S(w_i)\}$$

Remark 14. Note that for $V \in \langle X \rangle$ and $W \in \langle Z \rangle$ the bimodules M_V and M_W either intersect only in 0 or M_V is contained in M_W. Note further that the specializations of $W \in \langle Z \rangle$ are also given by

$$S(W) = \{V \in \langle X \rangle \mid M_V \subseteq M_W\}.$$

Definition 3 carries over by replacing X with Z. For such a reduction system Σ over Z we define the reduction ideal I_Σ by Eq. (4) and we define $\langle X \rangle_{\text{irr}}$ as the set of words from $\langle X \rangle$ containing no subwords from the set

$$\bigcup_{(W,h)\in\Sigma} S(W).$$

Based on $\langle X \rangle_{\text{irr}}$ we define $K\langle M \rangle_{\text{irr}}$ as in Eq. (3). Furthermore, for every reduction system Σ over Z we call its reformulation as a reduction system over X the *refined reduction system* Σ_X, which is given by

$$\Sigma_X := \bigcup_{(W,h)\in\Sigma} \{(V, h|_{M_V}) \mid V \in S(W)\}. \tag{6}$$

Lemma 15. *Let Σ be a reduction system over Z and let Σ_X be its refinement on X. Then the reduction ideals and the irreducible words are the same for Σ and for Σ_X. Moreover, also $K\langle M \rangle_{irr}$ stays the same.*

Proof. Follows immediately from the definitions. □

Note that, however, the refined reduction system does not define the same reduction relation. In general, we neither have $\to_{\Sigma_X} \subseteq \to_\Sigma$ nor $\to_\Sigma \subseteq \to_{\Sigma_X}$. We only have $\to_\Sigma \subseteq \overset{*}{\to}_{\Sigma_X}$ in general.

Definition 16. We call a partial order $\leq$ on $\langle Z \rangle$ *consistent with specialization* if for all words $V, W \in \langle Z \rangle$ with $V < W$ we also have $\tilde{V} < \tilde{W}$ for all specializations $\tilde{V} \in S(V)$ and $\tilde{W} \in S(W)$.

Note that the above definition implies that W is incomparable to all elements in $S(W)$, except possibly W itself, which can be seen by considering the two cases $V \in S(W)$ and $W \in S(V)$ in the definition.

A semigroup partial order $\leq$ on $\langle Z \rangle$ is *compatible* with a reduction system Σ over Z if for all $(W, h) \in \Sigma$ we have

$$h(M_W) \subseteq \sum_{\substack{V\in\langle Z\rangle \\ V<W}} M_V.$$

If $\leq$ is consistent with specialization, then for any $\tilde{W} \in S(W)$ we have

$$\sum_{\substack{V\in\langle Z\rangle \\ V<W}} M_V \subseteq \bigoplus_{\substack{V\in\langle X\rangle \\ V<\tilde{W}}} M_V.$$

Lemma 17. *Let Σ be a reduction system over Z and let $\leq$ be a semigroup partial order on $\langle Z \rangle$ consistent with specialization and compatible with Σ. Then the restricted order $\leq$ on $\langle X \rangle$ is compatible with Σ_X.*

Proof. By definition of Σ_X, For any reduction rule $(\tilde{W}, \tilde{h}) \in \Sigma_X$ there is $(W, h) \in \Sigma$ such that $\tilde{W} \in S(W)$ and $\tilde{h} = h|_{M_{\tilde{W}}}$. So, by our assumptions, we have

$$\tilde{h}(M_{\tilde{W}}) = h(M_{\tilde{W}}) \subseteq h(M_W) \subseteq \sum_{\substack{V\in\langle Z\rangle \\ V<W}} M_V \subseteq \bigoplus_{\substack{V\in\langle X\rangle \\ V<\tilde{W}}} M_V. \qquad \square$$

We need to generalize the notion of ambiguities to account for the fact that the sum $K\langle M\rangle = \sum_{W\in\langle Z\rangle} M_W$ is not necessarily direct anymore.

Definition 18. Let $(W,h),(\tilde{W},\tilde{h}) \in \Sigma$ be two (not necessarily distinct) reduction rules and let $A, B_1, B_2, C \in \langle Z\rangle$ be nonempty words with

$$W = AB_1, \quad \tilde{W} = B_2C, \quad \text{and} \quad S(B_1)\cap S(B_2) \neq \emptyset,$$

then we call this an *overlap ambiguity*. An overlap ambiguity is called *resolvable* if for all pure tensors $a \in M_A$, $b \in M_{B_1}\cap M_{B_2}$, and $c \in M_C$ the S-polynomial can be reduced to zero:

$$h(a\otimes b)\otimes c - a\otimes \tilde{h}(b\otimes c) \xrightarrow{*}_\Sigma 0.$$

Similarly, an *inclusion ambiguity* is given by two distinct reduction rules $(W,h),(\tilde{W},\tilde{h}) \in \Sigma$ and words $A, B_1, B_2, C \in \langle Z\rangle$ with $W = B_1$, $\tilde{W} = AB_2C$, and $S(B_1)\cap S(B_2)\neq\emptyset$. An inclusion ambiguity is called *resolvable* if for all pure tensors $a \in M_A$, $b \in M_{B_1}\cap M_{B_2}$, and $c \in M_C$ the S-polynomial can be reduced to zero: $a\otimes h(b)\otimes c - \tilde{h}(a\otimes b\otimes c) \xrightarrow{*}_\Sigma 0$.

If $B_1 \neq B_2$ for an overlap or inclusion ambiguity, then we say that the ambiguity is *with specialization*.

Again, we use $\mathrm{SP}(A\underline{B_1}, \underline{B_2}C)$ or $\mathrm{SP}(\underline{B_1}, A\underline{B_2}C)$, respectively, to refer to S-polynomials of an overlap or inclusion ambiguity.

Remark 19. Note that in total there now can be four types of ambiguities: in addition to the two types of ambiguities (without specialization) of Definition 5 there are also corresponding versions with specialization as defined above.

With these definitions we can prove the following generalization of Bergman's result. In order to prove properties of the reduction system Σ over Z we apply Bergman's result (Theorem 6) to the refined reduction system Σ_X over X.

Theorem 20. *Let M be a K-bimodule and let $(M_z)_{z\in Z}$ be a decomposition with specialization. Let Σ be a reduction system over Z on $K\langle M\rangle$ and let $\leq$ be a Noetherian semigroup partial order on $\langle Z\rangle$ consistent with specialization and compatible with Σ. Then the following are equivalent:*

1. *All ambiguities of Σ are resolvable.*
2. *Every $t \in K\langle M\rangle$ has a unique normal form $t\downarrow_\Sigma$.*
3. *$K\langle M\rangle/I_\Sigma$ and $K\langle M\rangle_{irr}$ are isomorphic as K-bimodules.*

Moreover, if these conditions are satisfied, then we can define a multiplication on $K\langle M\rangle_{irr}$ by $s\cdot t := (s\otimes t)\downarrow_\Sigma$ so that $K\langle M\rangle/I_\Sigma$ and $K\langle M\rangle_{irr}$ are isomorphic as K-rings.

Proof. First, we prove the implication 2. $\Rightarrow$ 1. Any S-polynomial of an ambiguity of Σ is of the form $h(t) - \tilde{h}(t)$ for some pure tensor $t \in K\langle M\rangle$ and reductions h and $\tilde{h}$ of Σ. Let H_1 be a composition of reductions of Σ such that $H_1(h(t)) \in K\langle M\rangle_{\mathrm{irr}}$ and let H_2 be a composition of reductions of Σ such that $H_2(H_1(\tilde{h}(t))) \in K\langle M\rangle_{\mathrm{irr}}$. Then $H_2\circ H_1$ reduces the S-polynomial to zero since t has a unique normal form w.r.t. Σ.

The rest of the proof is reduced to Theorem 6 via properties of the refined reduction system Σ_X. Lemma 15 shows that we can replace the reduction system Σ by its refinement Σ_X without changing the reduction ideal or $K\langle M\rangle_{\mathrm{irr}}$, hence statement 3. holds for Σ if and only if it holds for Σ_X. Furthermore, we note that every S-polynomial of Σ_X is also an S-polynomial of Σ and that $\xrightarrow{*}_\Sigma \subseteq \xrightarrow{*}_{\Sigma_X}$, hence statement 1. holds for Σ_X if it holds for Σ. If statement 2. holds for Σ_X, then by $\xrightarrow{*}_\Sigma \subseteq \xrightarrow{*}_{\Sigma_X}$ and the fact that $K\langle M\rangle_{\mathrm{irr}}$ does not change it also holds for Σ. Finally, Lemma 17 implies that Σ_X and the restriction of $\leq$ to $\langle X\rangle$ satisfy the assumptions of Theorem 6, which concludes the proof. □

Note that for $W, \tilde{W} \in \langle Z \rangle$ having a common specialization, i.e. $S(W) \cap S(\tilde{W}) \neq \emptyset$, there does not necessarily exist $V \in \langle Z \rangle$ such that $S(V) = S(W) \cap S(\tilde{W})$. In general, the intersection of two modules is given by

$$M_W \cap M_{\tilde{W}} = \bigoplus_{V \in S(W) \cap S(\tilde{W})} M_V = \bigotimes_{k=1}^{n} \bigoplus_{x \in S(w_k) \cap S(\tilde{w}_k)} M_x,$$

where $W = w_1 \dots w_n$ and $\tilde{W} = \tilde{w}_1 \dots \tilde{w}_n$.

Example 21. Consider alphabets $X = \{x_1, x_2, x_3\}$ and $Z = X \cup \{y_1, y_2\}$ with $M_{y_1} = M_{x_1} \oplus M_{x_3}$ and $M_{y_2} = M_{x_2} \oplus M_{x_3}$. The words $W = x_1 y_2 y_1$ and $\tilde{W} = y_1 y_2 y_2$ in $\langle Z \rangle$ satisfy $S(W) \cap S(\tilde{W}) = \{x_1 x_2 x_3, x_1 x_3 x_3\} \neq \emptyset$. We have $M_W \cap M_{\tilde{W}} = M_{x_1} \otimes M_{y_2} \otimes M_{x_3}$. So, in this case, there even exists a word $V = x_1 y_2 x_3$ that satisfies $S(V) = S(W) \cap S(\tilde{W})$ and $M_V = M_W \cap M_{\tilde{W}}$. □

Example 22. Consider alphabets $X = \{x_1, x_2, x_3, x_4\}$ and $Z = X \cup \{y_1, y_2\}$ with $S(y_i) = X \setminus \{x_{5-i}\}$. The words $W = y_1$ and $\tilde{W} = y_2$ satisfy $S(W) \cap S(\tilde{W}) = \{x_1, x_2\} \neq \emptyset$ and there is no word V with $S(V) = S(W) \cap S(\tilde{W})$. □

In order to describe the intersection of modules in terms of words again it will be convenient to also consider another partial order $\preceq$ on $\langle Z \rangle$, which is induced by the natural partial order, given by set inclusion, on all sets of the form $S(W) \subseteq \langle X \rangle$. In other words, we have $V \preceq W$ in $\langle Z \rangle$ if and only if $S(V) \subseteq S(W)$, which holds if and only if M_V is contained in M_W.

In addition, for a set $S \subseteq \langle Z \rangle$ we define the K-bimodule

$$M_S := \sum_{W \in S} M_W \subseteq K\langle M \rangle \tag{7}$$

with M_S being the trivial bimodule $\{0\}$ if S is empty. We also define

$$lb(S) := \{V \in \langle Z \rangle \mid V \preceq W \text{ for all } W \in S\}$$

as the set of all lower bounds of S with respect to the partial order $\preceq$. Note that this implies

$$\bigcap_{W \in S} M_W = M_{lb(S)} = M_{lb(S) \cap \langle X \rangle}$$

where we have $lb(S) \cap \langle X \rangle = \bigcap_{W \in S} S(W)$. If $\preceq$ satisfies the ascending chain condition, it is enough to consider only maximal elements of $lb(S)$ for $\bigcap_{W \in S} M_W = M_{lb(S)}$.

Example 23. Consider alphabets $X = \{x_1, x_2, x_3, x_4, x_5, x_6\}$ and $Z = X \cup \{y_1, y_2, y_3, z_1, z_2\}$ with $S(y_i) = \{x_i, x_{i+1}\}$ and $S(z_i) = X \setminus \{x_{7-i}\}$. The words $W = z_1$ and $\tilde{W} = z_2$ satisfy $S(W) \cap S(\tilde{W}) = \{x_1, x_2, x_3, x_4\} \neq \emptyset$ and there is no word V with $S(V) = S(W) \cap S(\tilde{W})$. We have $lb(W, \tilde{W}) = \{x_1, x_2, x_3, x_4, y_1, y_2, y_3\}$ and the maximal elements of $lb(W, \tilde{W})$ are y_1, y_2, y_3. As explained above, we have $M_W \cap M_{\tilde{W}} = M_{lb(W,\tilde{W})} = M_{lb(W,\tilde{W}) \cap \langle X \rangle} = M_{\{y_1, y_2, y_3\}}$. In this example, we can even find words such that the intersection is a direct sum of as few modules as possible: $M_W \cap M_{\tilde{W}} = M_{y_1} \oplus M_{y_3}$. □

3.1. Multi-level setting

Our two-level tensor setting presented at ISSAC 2016 (Hossein Poor et al., 2016a, Sec. 4) can be generalized to obtain a multi-level tensor setting, which in turn is a special case of the setting presented above. We briefly describe how the multi-level tensor setting looks like. To this end, we first recall when one direct sum decomposition of M is a refinement of another.

For two families of K-bimodules with $M = \bigoplus_{x \in X} M_x = \bigoplus_{y \in Y} M_y$, we say that $(M_x)_{x \in X}$ is a refinement of $(M_y)_{y \in Y}$ if there exists a partition $(X_y)_{y \in Y}$ of X such that

1. $X_y = \{x\}$ for all $y \in X \cap Y$ and
2. $M_y = \bigoplus_{x \in X_y} M_x$ for all $y \in Y$.

For the multi-level setting we consider a family of alphabets $(X_i)_{i\in I}$ each corresponding to a direct sum decomposition $M = \bigoplus_{x\in X_i} M_x$, the "levels". On the index set I we can define a partial order $\preceq$ such that $i \preceq j$ if and only if $(M_x)_{x\in X_i}$ is a refinement of $(M_x)_{x\in X_j}$. We require that the set I has a least element $0 \in I$ w.r.t. $\preceq$, i.e. there exists a finest level that is a refinement of all levels. Defining $X := X_0$ and $Z := \bigcup_{i\in I} X_i$ we easily recognize this as a special case of the above tensor setting with specialization.

Conversely, each instance of the tensor setting with specialization can be viewed as multi-level by letting $X_0 := X$ and completing each M_z, $z \in Z \setminus X$, into a level of its own: $M = M_z \oplus M_{\overline{z}}$ with $M_{\overline{z}} := \bigoplus_{x\in X\setminus S(z)} M_x$. The resulting order $\preceq$ on $I := \{0\} \cup (Z \setminus X)$ may be far from total, it may even be trivial apart from $0 \preceq i$.

The multi-level setting is worth mentioning mainly because of the following property. If $\preceq$ is a total order on I, i.e. if all levels are nested, then for any $W, \tilde{W} \in \langle Z\rangle$ with $S(W) \cap S(\tilde{W}) \neq \emptyset$, there exists (at least one) $V \in \langle Z\rangle$ such that $S(V) = S(W) \cap S(\tilde{W})$, i.e. $M_V = M_W \cap M_{\tilde{W}}$.

3.2. Computational aspects

Many properties that we discussed for Bergman's tensor setting also hold for the tensor setting with specialization we introduced above. For instance, determining ambiguities and irreducible words is done just on the level of words. In the following, we discuss the differences of the two settings.

The main computational benefit of Theorem 20 compared to Theorem 6 lies in the fact that for the confluence criterion we only need to check ambiguities of Σ over the alphabet Z and no computations with Σ_X are needed. Computing with the refined reduction system over X instead, generally would lead to a higher number of ambiguities, since one reduction rule in Σ can give rise to many reduction rules in Σ_X. Only for determination of irreducible words we restrict to $\langle X\rangle$.

If we formulate our reduction system Σ over the alphabet Z, instead of using some $\tilde{\Sigma}$ over the smaller alphabet X for the same reduction ideals $I_{\tilde{\Sigma}} = I_\Sigma$, we may be able to considerably reduce the size of the reduction system. This may happen in two different ways. First, assume a partition of X such that some homomorphisms in $\tilde{\Sigma}$ are defined by the same formula and the homomorphisms differ only by the choice of their domain and the corresponding words are obtained as specializations from some template. Then the corresponding reduction rules from $\tilde{\Sigma}$ could be merged into one reduction rule in Σ. This is exactly what happens for $\tilde{\Sigma} = \Sigma_X$. Second, also extending the domain of some homomorphism from $\tilde{\Sigma}$ may contribute to obtaining a smaller reduction system Σ. So usually we will have $\tilde{\Sigma} \subset \Sigma_X$.

The package `TenReS` also provides routines for generating all overlap and inclusion ambiguities with specialization together with their corresponding S-polynomials. For a detailed comparison of Bergman's setting and our generalization for the example of IDOs see (Hossein Poor et al., 2016a).

4. Integro-differential operators

Integro-differential operators over a field of constants were introduced in Rosenkranz (2005), Rosenkranz and Regensburger (2008) to study algebraic and algorithmic aspects of linear ordinary boundary problems. The construction made use of a parametrized Gröbner basis in infinitely many variables coming from a basis of the coefficient algebra; see also the survey (Rosenkranz et al., 2012) for an automated confluence proof and (Regensburger, 2016) for related references. For polynomial coefficients, also generalized Weyl algebras (Bavula, 2013), skew polynomials (Regensburger et al., 2009), and noncommutative Gröbner bases (Quadrat and Regensburger, 2017) have been used to study them. In this section, we apply the tensor setting with specialization introduced above to the construction of normal forms for integro-differential operators (IDOs) over an arbitrary integro-differential ring. First, we define an integro-differential ring analogous to the definition of an integro-differential algebra in Rosenkranz et al. (2012), Guo et al. (2014).

Definition 24. Let (R, ∂) be a differential ring such that $\partial R = R$. Moreover, let $\int : R \to R$ be a bimodule homomorphism over the ring of constants in R, such that

$$\partial \int f = f \tag{8}$$

for all $f \in R$. We call $(R, \partial, \int)$ an *integro-differential ring* if the *evaluation*

$$\mathrm{E}f := f - \int \partial f \tag{9}$$

is multiplicative, i.e. for all $f, g \in R$ we have

$$\mathrm{E}fg = (\mathrm{E}f)\mathrm{E}g.$$

The following lemma shows that in any integro-differential ring, the evaluation E maps to the constants such that it acts as the identity on them, in particular, it is also a homomorphism of rings with unit element. Moreover, the ring R can be decomposed as direct sum of constant and non-constant "functions".

Lemma 25. *Let $(R, \partial, \int)$ be an integro-differential ring with constants K. Then, we have $\mathrm{E}1 = 1$, $\mathrm{E}f \in K$ for all $f \in R$, and*

$$R = K \oplus \int R,$$

as direct sum of K-bimodules.

Proof. We first compute $\mathrm{E}1 = 1 - \int \partial 1 = 1$ and $\partial \mathrm{E}f = \partial(f - \int \partial f) = \partial f - \partial f = 0$. For any $f \in R$, we have $f = \mathrm{E}f + f - \mathrm{E}f = \mathrm{E}f + \int \partial f$ and hence $R = K + \int R$. Let $f \in K \cap \int R$ and $g \in R$ such that $f = \int g$. Then $0 = \partial f = \partial \int g = g$, which implies $f = 0$. □

For the rest of this section, we fix an arbitrary integro-differential ring $(R, \partial, \int)$ and we denote its ring of constants by K. By an operator, we understand in the following a K-bimodule homomorphism from R to R. For example, the operations $\partial, \int, \mathrm{E}$ can be viewed as operators.

Following Lemma 25, we consider the direct sum decomposition $R = K \oplus \int R$ and the corresponding K-bimodules

$$M_{\mathsf{K}} = K \quad \text{and} \quad M_{\tilde{\mathsf{R}}} = \int R \tag{10}$$

(indexed by the symbols K and $\tilde{\mathsf{R}}$). Note that the elements of M_{K} and $M_{\tilde{\mathsf{R}}}$ are not interpreted as functions but as left multiplication operators $g \mapsto fg$ induced by those functions. For studying boundary value problems algebraically, we also need to deal with other multiplicative "functionals" on R with the same properties as E, so we consider the set

$$\Phi := \{\varphi \colon R \to K \mid \varphi \text{ is a } K\text{-bimodule homomorphism with } \varphi f g = (\varphi f)\varphi g \text{ and } \varphi 1 = 1\}. \tag{11}$$

Instead, one can also consider Φ as a proper subset (containing E) of the full set defined above. This amounts to working with a smaller ring of operators later. For the operators ∂, $\int$, E, and $\varphi \in \tilde{\Phi}$ with $\tilde{\Phi} = \Phi \setminus \{\mathrm{E}\}$, we consider the free left K-modules

$$M_{\mathsf{D}} = K\partial, \quad M_{\mathsf{I}} = K\textstyle\int, \quad M_{\mathsf{E}} = K\mathrm{E}, \quad M_{\tilde{\Phi}} = K\tilde{\Phi} \tag{12}$$

generated by them (indexed by the symbols D, I, E, and $\tilde{\Phi}$). We view these modules as K-bimodules with right multiplication defined by

$$c\alpha \cdot d = cd\alpha$$

where $\alpha \in \{\partial, \int, \mathrm{E}\} \cup \tilde{\Phi}$ and $c, d \in K$, since the generators of these modules correspond to left K-linear operators. We define two alphabets

$$X = \{\mathsf{K}, \tilde{\mathsf{R}}, \mathsf{D}, \mathsf{I}, \mathsf{E}, \tilde{\Phi}\} \quad \text{and} \quad Z = X \cup \{\mathsf{R}, \Phi\}, \tag{13}$$

with the K-bimodules $(M_x)_{x \in X}$ defined in Eqs. (10) and (12) as well as

$$M_{\mathsf{R}} = M_{\mathsf{K}} \oplus M_{\tilde{\mathsf{R}}} \quad \text{and} \quad M_{\Phi} = M_{\mathsf{E}} \oplus M_{\tilde{\Phi}}. \tag{14}$$

Now, we define the module M by

$$M := M_{\mathsf{R}} \oplus M_{\mathsf{D}} \oplus M_{\mathsf{I}} \oplus M_{\Phi}, \tag{15}$$

which turns $(M_z)_{z\in Z}$ into a decomposition with specialization.

In order to compute with these operators, we need to collect identities they satisfy in form of a reduction system. To this end, we first list identities following immediately from their definitions (like multiplicativity of functionals, K-linearity, and the Leibniz rule) and some of their consequences that hold in R. For all $f, g \in R$ and $\varphi, \psi \in \Phi$:

$$\begin{array}{ll}
\varphi f g = (\varphi f)\varphi g & \partial\int g = g \\
\psi\varphi g = \varphi g & \int\partial g = g - \mathrm{E}g \\
\mathrm{E}\int g = 0 & \int f\varphi g = (\int f)\varphi g \\
\partial f g = f\partial g + (\partial f)g & \int f\partial g = fg - \int(\partial f)g - (\mathrm{E}f)\mathrm{E}g \\
\partial\varphi g = 0 & \int f\int g = (\int f)\int g - \int(\int f)g
\end{array}$$

The identities that do not follow immediately from the definitions are $\mathrm{E}\int g = 0$, integration by parts

$$\int f\partial g = fg - \int(\partial f)g - (\mathrm{E}f)\mathrm{E}g,$$

and the Rota–Baxter identity

$$\int f\int g = (\int f)\int g - \int(\int f)g$$

for the integral. They can either be verified directly or we obtain them in Section 6 as a consequence of S-polynomial computations. All identities listed above correspond to identities for operators acting on $g \in R$. The reduction system Σ over the alphabet $\langle Z\rangle$ is given by Table 1, defined in terms of all $f, g \in R$ and $\varphi, \psi \in \Phi$.

In analogy to the definition of reduction homomorphisms in Section 2, the informal definitions in Table 1 have to be made formal. For instance,

$$\beta_{\mathsf{ID}}(\textstyle\int, \partial) := \epsilon - \mathrm{E}$$

is extended to a balanced map on $M_{\mathsf{I}} \times M_{\mathsf{D}}$ via

$$\beta_{\mathsf{ID}}(c\textstyle\int, d\partial) := cd\beta_{\mathsf{ID}}(\int, \partial)$$

and similarly

$$\beta_{\mathsf{IR}\Phi}(\textstyle\int, f, \varphi) := \int f \otimes \varphi$$

Table 1
Reduction rules for IDOs.

K	$1 \mapsto \epsilon$
RR	$f \otimes g \mapsto fg$
ΦR	$\varphi \otimes f \mapsto (\varphi f)\varphi$
ΦΦ	$\psi \otimes \varphi \mapsto \varphi$
EI	$\mathrm{E} \otimes \int \mapsto 0$
DR	$\partial \otimes f \mapsto f \otimes \partial + \partial f$
DΦ	$\partial \otimes \varphi \mapsto 0$
DI	$\partial \otimes \int \mapsto \epsilon$
IΦ	$\int \otimes \varphi \mapsto \int 1 \otimes \varphi$
ID	$\int \otimes \partial \mapsto \epsilon - \mathrm{E}$
II	$\int \otimes \int \mapsto \int 1 \otimes \int - \int \otimes \int 1$
IRΦ	$\int \otimes f \otimes \varphi \mapsto \int f \otimes \varphi$
IRD	$\int \otimes f \otimes \partial \mapsto f - \int \otimes \partial f - (\mathrm{E}f)\mathrm{E}$
IRI	$\int \otimes f \otimes \int \mapsto \int f \otimes \int - \int \otimes \int f$

with $\varphi \in \Phi$ is extended to a balanced map on $M_{\mathsf{I}} \times M_{\mathsf{R}} \times M_{\Phi}$ by

$$\beta_{\mathsf{IR}\Phi}(c\textstyle\int, f, \sum_i c_i\varphi_i) := \sum_i \beta_{\mathsf{IR}\Phi}(\textstyle\int, cfc_i, \varphi_i).$$

Definition 26. Let $(R, \partial, \int)$ be an integro-differential ring with constants K. We call

$$R\langle \partial, \textstyle\int, \Phi\rangle := K\langle M\rangle / J$$

the *ring of integro-differential operators*, where J is the two-sided reduction ideal induced by the reduction system obtained from Table 1.

In order to compute in $R\langle \partial, \int, \Phi\rangle$ we want to analyze the reduction system defined by Table 1 according to Theorem 20 above and determine normal forms of tensors. Following the definition in Eq. (6), the refined reduction system Σ_X is obtained, according to Eq. (14), by splitting rules whose words contain R or Φ into "smaller" rules using $S(\mathsf{R}) = \{\mathsf{K}, \tilde{\mathsf{R}}\}$ and $S(\Phi) = \{\mathsf{E}, \tilde{\Phi}\}$. For example, the reduction rule $(\Phi\mathsf{R}, h) \in \Sigma$ is split into the rules $(W, h|_{M_W}) \in \Sigma_X$ where $W \in S(\Phi\mathsf{R}) = \{\mathsf{EK}, \mathsf{E}\tilde{\mathsf{R}}, \tilde{\Phi}\mathsf{K}, \tilde{\Phi}\tilde{\mathsf{R}}\}$.

Theorem 27. *Let $(R, \partial, \int)$ be an integro-differential ring with constants K and let Φ be the set of multiplicative K-bimodule homomorphisms given by Eq. (11). Let M be defined by Eqs. (14) and (15) and let the reduction system Σ be defined by Table 1.*

Then every $t \in K\langle M\rangle$ has a unique normal form $t{\downarrow_\Sigma}$, which is given by a sum of pure tensors of the form

$$f \otimes \varphi \otimes \partial^{\otimes j} \quad \text{or} \quad f \otimes \varphi \otimes \textstyle\int \otimes g$$

where $j \in \mathbb{N}_0$, each of $f, g \in M_{\tilde{\mathsf{R}}}$ and $\varphi \in \Phi$ may be absent, and $\varphi \otimes \int$ does not specialize to $\mathsf{E} \otimes \int$. Moreover,

$$R\langle \partial, \textstyle\int, \Phi\rangle \cong K\langle M\rangle_{irr}$$

as K-rings, where the multiplication on $K\langle M\rangle_{irr}$ is defined by $s \cdot t := (s \otimes t){\downarrow_\Sigma}$.

Proof. We consider the alphabets X and Z given by Eq. (13). This turns $(M_z)_{z\in Z}$ into a decomposition with specialization for the module M, see Definition 12. For defining a Noetherian monoid partial order $\preceq$ on $\langle Z\rangle$ that is compatible with Σ, it is sufficient to require the order to satisfy

$$\mathsf{DR} > \mathsf{RD},\ \mathsf{IRD} > \mathsf{E},\ \mathsf{ID} > \mathsf{E},\ \mathsf{I} > \tilde{\mathsf{R}}.$$

For instance, we could use a degree-lexicographic order with $\mathsf{I} > \mathsf{D} > \Phi > \mathsf{R}$ on $\langle\{\mathsf{R}, \mathsf{D}, \mathsf{I}, \Phi\}\rangle \subseteq \langle Z\rangle$ or other degree-lexicographic orders with $\mathsf{D} > \mathsf{R}$ and $\mathsf{I} > \mathsf{R}$. We extend it to a monoid partial order on $\langle Z\rangle$ based on Definition 16 in order to make it consistent with specialization. Then by the package `TenReS` we verify that all ambiguities of Σ are resolvable, see Section 4.1. Hence by Theorem 20 every element of $K\langle M\rangle$ has a unique normal form and $K\langle M\rangle/I_\Sigma \cong K\langle M\rangle_{\mathrm{irr}}$ as K-rings.

It remains to determine the explicit form of elements in $K\langle M\rangle_{\mathrm{irr}}$. To do so, we determine the set of irreducible words $\langle X\rangle_{\mathrm{irr}}$ in $\langle X\rangle$. Irreducible words containing only the letters K and $\tilde{\mathsf{R}}$ have to avoid the subwords K and $S(\mathsf{RR}) = \{\mathsf{KK}, \mathsf{K}\tilde{\mathsf{R}}, \tilde{\mathsf{R}}\mathsf{K}, \tilde{\mathsf{R}}\tilde{\mathsf{R}}\}$, hence only the words ϵ and $\tilde{\mathsf{R}}$ are left. The irreducible words containing only E and $\tilde{\Phi}$ are exactly ϵ, E, and $\tilde{\Phi}$, since they have to avoid the subwords $S(\Phi\Phi) = \{\mathsf{EE}, \mathsf{E}\tilde{\Phi}, \tilde{\Phi}\mathsf{E}, \tilde{\Phi}\tilde{\Phi}\}$. Altogether, we see that the irreducible words containing only the letters K, $\tilde{\mathsf{R}}$, E, and $\tilde{\Phi}$ are given by the set $\{\epsilon, \tilde{\mathsf{R}}, \mathsf{E}, \tilde{\Phi}, \tilde{\mathsf{R}}\mathsf{E}, \tilde{\mathsf{R}}\tilde{\Phi}\}$, since they also have to avoid the subwords $S(\Phi\mathsf{R}) = \{\mathsf{EK}, \mathsf{E}\tilde{\mathsf{R}}, \tilde{\Phi}\mathsf{K}, \tilde{\Phi}\tilde{\mathsf{R}}\}$. Allowing also the letter D, we have to avoid the subwords coming from $S(\mathsf{DR}) = \{\mathsf{DK}, \mathsf{D}\tilde{\mathsf{R}}\}$ and $S(\mathsf{D}\Phi) = \{\mathsf{DE}, \mathsf{D}\tilde{\Phi}\}$. Therefore, we can only append words D^j with $j \in \mathbb{N}_0$ to the irreducible words determined so far, in order to obtain all elements of $\langle X\rangle_{\mathrm{irr}}$ not containing the letter I. Finally, we also consider the letter I. Since subwords EI and DI have to be avoided, the first occurrence of I in an irreducible word can only be preceded by ϵ, $\tilde{\mathsf{R}}$, $\tilde{\Phi}$, or $\tilde{\mathsf{R}}\tilde{\Phi}$. We also have to avoid the subwords $S(\mathsf{I}\Phi) = \{\mathsf{IE}, \mathsf{I}\tilde{\Phi}\}$, ID, and II, so any letter immediately

following I has to be $\tilde{\mathsf{R}}$. In addition, we have to avoid the subwords $S(\mathsf{IR\Phi}) = \{\mathsf{IKE}, \mathsf{IK\tilde{\Phi}}, \mathsf{I\tilde{R}E}, \mathsf{I\tilde{R}\tilde{\Phi}}\}$, $S(\mathsf{IRD}) = \{\mathsf{IKD}, \mathsf{I\tilde{R}D}\}$, and $S(\mathsf{IRI}) = \{\mathsf{IKI}, \mathsf{I\tilde{R}I}\}$, so the letter I cannot be followed by a subword of length greater than one. Altogether, the elements of $\langle X\rangle_{\mathrm{irr}}$ are of the form

$$\tilde{\mathsf{R}} V \mathsf{D}^j \quad \text{or} \quad \tilde{\mathsf{R}}\tilde{\Phi}\mathsf{I}\tilde{\mathsf{R}},$$

where $j \in \mathbb{N}_0$ and each of $\tilde{\mathsf{R}}$, $\tilde{\Phi}$, and $V \in S(\Phi) = \{\mathsf{E}, \tilde{\Phi}\}$ may be absent. The normal forms follow from Eq. (3). □

Note that the formulae given in Table 1 above to define the reduction system for the tensor ring are the same as the formulae presented in Table 2 in Hossein Poor et al. (2016a) for the tensor algebra with commutative K. Here we use these formulae to define K-bimodule homomorphisms via balanced maps instead of defining K-module homomorphisms via multilinear maps. The same ambiguities need to be considered for checking confluence and we obtain the same structure of normal forms. Differences arise only from R now being a K-ring instead of a K-algebra.

4.1. Computational aspects

In the following, we briefly discuss computational details of the tensor setting with specialization for integro-differential operators. Applying `TenReS` to the reduction system Σ, in total 52 ambiguities and corresponding S-polynomials are generated. Among them, there are 4 ambiguities for which the corresponding S-polynomials are zero anyway, for instance

$$\mathrm{SP}(\mathsf{D}\underline{\Phi}, \underline{\mathsf{E}}\mathsf{I}) = 0 \otimes \textstyle\int - \partial \otimes 0 = 0.$$

The S-polynomials of 48 remaining ambiguities are reduced to zero by applying automatically the implementation of rules from Σ, identities in R and identities in M_{D}, M_{I} and M_{Φ}. The complete computation is included in the example files of the package. Here we consider a few concrete instances of ambiguities. For example, we use the definition of E in R in the reduction of the following S-Polynomial

$$\begin{aligned}
\mathrm{SP}(\mathsf{IR}\underline{\mathsf{D}}, \underline{\mathsf{D}}\Phi) &= (f - \textstyle\int \otimes \partial f - (\mathsf{E}f)\mathsf{E}) \otimes \varphi - \int \otimes f \otimes 0 \\
&\to_{r_{\mathsf{IR}\Phi}} f \otimes \varphi - (\textstyle\int \partial f) \otimes \varphi - (\mathsf{E}f)\mathsf{E} \otimes \varphi \\
&= f \otimes \varphi - (f - \mathsf{E}f) \otimes \varphi - (\mathsf{E}f)\mathsf{E} \otimes \varphi \\
&= \mathsf{E}f \otimes \varphi - (\mathsf{E}f)\mathsf{E} \otimes \varphi \to_{r_{\Phi\Phi}} \mathsf{E}f \otimes \varphi - (\mathsf{E}f)\varphi \to_{r_{\mathsf{K}}} 0.
\end{aligned}$$

As another example, we use the definition of the right multiplication in the K-bimodule M_{I} in the following reduction

$$\begin{aligned}
\mathrm{SP}(\mathsf{I}\underline{\Phi}, \underline{\Phi}\mathsf{R}) &= (\textstyle\int 1 \otimes \varphi) \otimes f - \int \otimes (\varphi f)\varphi \to_{r_{\mathsf{I}\Phi}} \int 1 \otimes \varphi \otimes f - \varphi f (\int 1 \otimes \varphi) \\
&\to_{r_{\Phi\mathsf{R}}} \textstyle\int 1 \otimes (\varphi f)\varphi - \varphi f (\int 1 \otimes \varphi) \\
&= (\textstyle\int 1 \varphi f) \otimes \varphi - \varphi f (\int 1 \otimes \varphi) = (\varphi f) \int 1 \otimes \varphi - \varphi f (\int 1 \otimes \varphi) \\
&= \varphi f (\textstyle\int 1 \otimes \varphi) - \varphi f (\int 1 \otimes \varphi) = 0.
\end{aligned}$$

There are 41 ambiguities without specialization. The remaining 11 ambiguities consist of 4 overlap ambiguities with specialization and 7 inclusion ambiguities with specialization. For example,

$$\mathrm{SP}(\mathsf{IR}\underline{\Phi}, \underline{\mathsf{E}}\mathsf{I}) = (\textstyle\int f \otimes \mathsf{E}) \otimes \int - \int \otimes f \otimes 0 \to_{r_{\mathsf{EI}}} 0,$$

and

$$\mathrm{SP}(\underline{\mathsf{K}}, \mathsf{D}\underline{\mathsf{R}}) = \partial \otimes \epsilon \otimes \epsilon - 1 \otimes \partial \ \to_{r_{\mathsf{K}}} \partial - \partial = 0.$$

We emphasize again that the confluence criterion of Theorem 20 directly works with the reduction system Σ, no computations with the refined reduction system Σ_X over X are needed.

5. Integro-differential operators with linear substitutions

In this section, we apply our tensor setting with specialization to extend the ring of integro-differential operators by adding linear substitution operators. An important motivation for studying this ring comes from the work by Quadrat (2015). In this paper, such operators and their commutation rules are used for an algorithmic approach to Artstein's integral transformation of linear differential systems with delayed inputs to linear differential system without delays. IDOs with linear substitutions also address the univariate case in Rosenkranz et al. (2015), where algebraic aspects of multivariate integration with linear substitutions are studied. Moreover, they provide an algebraic setting for dealing with delay differential equations and the corresponding initial and boundary problems in general.

A delay differential equation is an ordinary differential equation in which the derivative at a certain time depends on the solution at prior times; see, for example, Hale and Verduyn Lunel (1993), Smith (2011). A general first-order constant delay equation has the form

$$y'(x) = f(x, y(x), y(x-b_1), y(x-b_2), \ldots, y(x-b_n))$$

where the time delays b_j for $1 \leq j \leq n$ are positive constants. A homogeneous linear first-order time-delay equation with one constant delay has the form

$$y'(x) = A(x)y(x) + B(x)y(x-b).$$

The chain rule and integration by substitution from calculus describe the interaction of linear substitutions $f(ax-b)$ with differentiation and integration. More formally, let $\sigma_{a,b}$ denote the linear substitution operator mapping a smooth function $f(x)$ to $f(ax-b)$ for a nonzero constant a and an arbitrary constant b. Then

$$\partial_x \sigma_{a,b} f(x) = af'(ax-b) = a\sigma_{a,b}\partial_x f(x)$$

and

$$\int_0^x \sigma_{a,b} f(t)\,dt = \int_0^x f(at-b)\,dt = \frac{1}{a}\int_{-b}^{ax-b} f(t)\,dt = \frac{1}{a}\sigma_{a,b}\int_0^x f(t)\,dt - \frac{1}{a}\mathrm{E}\sigma_{a,b}\int_0^x f(t)\,dt.$$

Following these identities, we want to define an integro-differential ring with linear substitutions. In what follows, $C = K \cap \mathcal{Z}(R)$ denotes the ring of elements of K which commute with all elements of R and C^* denotes its group of units. In order to find a proper algebraic setting, we will add an axiomatization of linear substitution operations to an integro-differential ring.

Definition 28. Let $(R, \partial, \int)$ be an integro-differential ring with constants K and let

$$S := \{\sigma_{a,b} \mid a \in C^*,\ b \in C\}$$

where $\sigma_{a,b}\colon R \to R$ are multiplicative K-bimodule homomorphisms on R fixing the constants K such that

$$\sigma_{1,0} f = f, \quad \sigma_{a,b}\sigma_{c,d} f = \sigma_{ac,bc+d} f \tag{16}$$

and

$$\partial\sigma_{a,b} f = a\sigma_{a,b}\partial f \tag{17}$$

for all $a, c \in C^*$, $b, d \in C$ and $f \in R$. Then we call $(R, \partial, \int, S)$ an *integro-differential ring with linear substitutions*.

Remark 29. The set S along with composition can be considered as a group of K-bimodule homomorphisms on R. The neutral element is $\sigma_{1,0}$ and the inverse for $\sigma_{a,b} \in S$ is given by

$$\sigma_{a,b}^{-1} = \sigma_{a^{-1},-ba^{-1}}.$$

So the elements in S actually are automorphisms.

As in analysis, integration by substitution is a consequence of the chain rule and the fundamental theorem of calculus.

Lemma 30. *Let $(R, \partial, \int, S)$ be an integro-differential ring with linear substitutions. For all $\sigma_{a,b} \in S$ and $f \in R$,*

$$\int \sigma_{a,b} f = a^{-1}(\mathrm{id} - \mathrm{E})\sigma_{a,b} \int f. \tag{18}$$

Proof. We first apply $\int$ to Eq. (17). So

$$\int \partial \sigma_{a,b} f = \int a \sigma_{a,b} \partial f = a \int \sigma_{a,b} \partial f.$$

By Eq. (9), we substitute $\int \partial \sigma_{a,b} f$ with $(\mathrm{id} - \mathrm{E})\sigma_{a,b} f$ and multiply the resulting equation by a^{-1}. This gives the identity

$$\int \sigma_{a,b} \partial f = a^{-1}(\mathrm{id} - \mathrm{E})\sigma_{a,b} f,$$

which implies Eq. (18) by just replacing f with $\int f$. □

In the sequel, we fix an integro-differential ring with linear substitutions $(R, \partial, \int, S)$ with constants K and evaluation $\mathrm{E} = \mathrm{id} - \int \partial$. We consider the modules M_{K}, $M_{\tilde{\mathsf{R}}}$, M_{D}, M_{I}, M_{E}, $M_{\tilde{\Phi}}$, M_{R}, and M_{Φ} which are introduced in Eqs. (10), (12), and (14). In addition, we add the free left K-module

$$M_{\mathsf{G}} := KS.$$

We also view it as a K-bimodule with the right multiplication defined by $c\sigma_{a,b} \cdot d = cd\sigma_{a,b}$ with $c, d \in K$. It has the direct sum decomposition

$$M_{\mathsf{G}} = M_{\mathsf{N}} \oplus M_{\tilde{\mathsf{G}}}$$

such that $M_{\mathsf{N}} := K\sigma_{1,0}$ is the K-bimodule generated by the trivial substitution $\sigma_{1,0} = \mathrm{id}$ and $M_{\tilde{\mathsf{G}}} := K\tilde{S}$ is the K-bimodule generated by all linear substitutions in $\tilde{S} = S \setminus \{\sigma_{1,0}\}$. Therefore we take the alphabets

$$X := \{\mathsf{K}, \tilde{\mathsf{R}}, \mathsf{D}, \mathsf{I}, \mathsf{E}, \tilde{\Phi}, \mathsf{N}, \tilde{\mathsf{G}}\}, \quad Z := X \cup \{\mathsf{R}, \Phi, \mathsf{G}\}. \tag{19}$$

With the K-bimodules

$$M_{\mathsf{R}} = M_{\mathsf{K}} \oplus M_{\tilde{\mathsf{R}}}, \quad M_{\Phi} = M_{\mathsf{E}} \oplus M_{\tilde{\Phi}}, \quad M_{\mathsf{G}} = M_{\mathsf{N}} \oplus M_{\tilde{\mathsf{G}}}, \tag{20}$$

we define

$$M := M_{\mathsf{R}} \oplus M_{\mathsf{D}} \oplus M_{\mathsf{I}} \oplus M_{\Phi} \oplus M_{\mathsf{G}}. \tag{21}$$

Then $(M_z)_{z \in Z}$ is a decomposition with specialization.

In addition to the identities of IDOs that we collected in Section 4, the identities for IDOs with linear substitutions include additional identities involving the substitution operators. Again, we first collect some identities involving substitution operations that hold in R. For all $f, g \in R$, $\varphi \in \Phi$ and $\sigma_{a,b}, \sigma_{c,d} \in S$ we have:

$$\begin{aligned}
&\sigma_{1,0} g = g && \sigma_{a,b}\sigma_{c,d} g = \sigma_{ac,bc+d} g \\
&\sigma_{a,b} fg = (\sigma_{a,b} f)(\sigma_{a,b} g) && \partial \sigma_{a,b} g = a \sigma_{a,b} \partial g \\
&\sigma_{a,b} \varphi g = \varphi g && \int f \sigma_{a,b} g = a^{-1}(\mathrm{id} - \mathrm{E})\sigma_{a,b} \int (\sigma_{a,b}^{-1} f) g
\end{aligned}$$

The only identity above that does not follow immediately from Definition 28 is

$$\int f \sigma_{a,b} g = a^{-1}(\mathrm{id} - \mathrm{E})\sigma_{a,b} \int (\sigma_{a,b}^{-1} f) g.$$

Table 2
New reduction rules for IDOs with linear substitutions.

N	$\sigma_{1,0} \mapsto \epsilon$
GR	$\sigma_{a,b} \otimes f \mapsto \sigma_{a,b} f \otimes \sigma_{a,b}$
GΦ	$\sigma_{a,b} \otimes \varphi \mapsto \varphi$
GG	$\sigma_{a,b} \otimes \sigma_{c,d} \mapsto \sigma_{ac,bc+d}$
DG	$\partial \otimes \sigma_{a,b} \mapsto a\sigma_{a,b} \otimes \partial$
IG	$\int \otimes \sigma_{a,b} \mapsto a^{-1}(\epsilon - \mathrm{E}) \otimes \sigma_{a,b} \otimes \int$
IRG	$\int \otimes f \otimes \sigma_{a,b} \mapsto a^{-1}(\epsilon - \mathrm{E}) \otimes \sigma_{a,b} \otimes \int \otimes \sigma_{a,b}^{-1} f$

It can be verified by replacing f with $(\sigma_{a,b}^{-1} f)g$ in Lemma 30 and then using multiplicativity of $\sigma_{a,b}$. Corresponding reduction rules to these identities in R are listed in Table 2.

In order to obtain our reduction system Σ over the alphabet $\langle Z \rangle$, we consider reduction rules of the Table 1 along with the reduction rules of the Table 2 simultaneously.

Definition 31. Let $(R, \partial, \int, S)$ be an integro-differential ring with linear substitutions. We call

$$R\langle \partial, \int, \Phi, S \rangle := K\langle M \rangle / J$$

the *ring of integro-differential operators with linear substitutions*, where J is the two-sided reduction ideal induced by the reduction system obtained from adjoining Table 2 to Table 1.

Similar to the previous example, the refined reduction system Σ_X is obtained, according to Eq. (20), by splitting rules whose words contain R, Φ or G into "smaller" rules using $S(\mathsf{R}) = \{\mathsf{K}, \tilde{\mathsf{R}}\}$, $S(\Phi) = \{\mathsf{E}, \tilde{\Phi}\}$ and $S(\mathsf{G}) = \{\mathsf{N}, \tilde{\mathsf{G}}\}$. Following Theorem 20, we determine normal forms of tensors in $R\langle \partial, \int, \Phi, S \rangle$.

Theorem 32. *Let $(R, \partial, \int, S)$ be an integro-differential ring with linear substitutions and let M be as in Eqs. (21) and (20) and let the reduction system Σ be defined by Tables 1 and 2. Then every $t \in K\langle M \rangle$ has a unique normal form given by a sum of pure tensors*

$$f \otimes \varphi \otimes \sigma_{a,b} \otimes \partial^{\otimes j} \qquad \text{or} \qquad f \otimes \varphi \otimes \sigma_{a,b} \otimes \int \otimes g,$$

where $j \in \mathbb{N}_0$, each of $f, g \in M_{\tilde{\mathsf{R}}}$, $\varphi \in \Phi$ and $\sigma_{a,b} \in \tilde{S}$ may be absent, and $\varphi \otimes \sigma_{a,b} \otimes \int$ does not specialize to $\mathrm{E} \otimes \int$. Moreover, with defining the multiplication $s \cdot t := (s \otimes t)\downarrow_\Sigma$ on $K\langle M \rangle_{irr}$

$$R\langle \partial, \int, \Phi, S \rangle \cong K\langle M \rangle_{irr}.$$

Proof. We consider the alphabets X and Z as defined in Eq. (19). Then $(M_z)_{z \in Z}$ is a decomposition with specialization for the module M, see Definition 12. For defining a Noetherian monoid partial order $\leq$ on $\langle Z \rangle$ that is compatible with Σ, it is sufficient to require the order to satisfy

$$\mathsf{DR} > \mathsf{RD},\ \mathsf{IRD} > \mathsf{E},\ \mathsf{ID} > \mathsf{E},\ \mathsf{I} > \tilde{\mathsf{R}},\ \mathsf{GR} > \mathsf{RG},\ \mathsf{DG} > \mathsf{GD},\ \mathsf{IG} > \mathsf{EGI},\ \mathsf{IRG} > \mathsf{EGIR}.$$

For instance, on $\langle Y \rangle$ with $Y = \{\mathsf{R}, \mathsf{D}, \mathsf{I}, \Phi, \mathsf{G}\}$ we first define a monoid order by

$$V \preceq W :\Leftrightarrow \tilde{V} \prec \tilde{W} \text{ or } \tilde{V} = \tilde{W} \text{ and } V \preceq W,$$

where $\tilde{V}$ and $\tilde{W}$ are obtained by removing all occurrences of Φ, cf. Remark 4, and $\preceq$ is the degree-lexicographic order with $\mathsf{I} \succ \mathsf{D} \succ \mathsf{G} \succ \Phi \succ \mathsf{R}$ on $\langle Y \rangle$. Then, we extend $\preceq$ to a monoid partial order on $\langle Z \rangle$ based on Definition 16 in order to make it consistent with specialization.

Then by the package `TenReS` we verify that all ambiguities of Σ are resolvable, see Section 5.1. Hence by Theorem 20 every element of $K\langle M \rangle$ has a unique normal form and $K\langle M \rangle / I_\Sigma \cong K\langle M \rangle_{\mathrm{irr}}$ as K-rings.

It remains to determine the explicit form of elements in $K\langle M\rangle_{\mathrm{irr}}$. To do so, we determine the set of irreducible words $\langle X\rangle_{\mathrm{irr}}$ in $\langle X\rangle$. Note that $\Sigma_{\mathrm{IDO}} \subset \Sigma$, where Σ_{IDO} is given by Table 1. Hence the irreducible words w.r.t. Σ are among the irreducible words w.r.t. Σ_{IDO}. In Theorem 27, we already determined the irreducible words that do not contain the letters N and $\tilde{\mathsf{G}}$ to be of the form

$$\tilde{\mathsf{R}} V \mathsf{D}^j \quad \text{or} \quad \tilde{\mathsf{R}} \tilde{\Phi} \mathsf{I} \tilde{\mathsf{R}},$$

where $j \in \mathbb{N}_0$ and each of $\tilde{\mathsf{R}}$, $\tilde{\Phi}$, and $V \in S(\Phi)$ may be absent.

The irreducible words containing only N and $\tilde{\mathsf{G}}$ are exactly ϵ and $\tilde{\mathsf{G}}$, since they have to avoid the subwords N and $S(\mathsf{GG}) = \{\mathsf{NN}, \mathsf{N}\tilde{\mathsf{G}}, \tilde{\mathsf{G}}\mathsf{N}, \tilde{\mathsf{G}}\tilde{\mathsf{G}}\}$. The irreducible words in $\langle X\rangle_{\mathrm{irr}}$ also have to avoid subwords from $S(\mathsf{GR})$, $S(\mathsf{G}\Phi)$, $S(\mathsf{DG})$, $S(\mathsf{IG})$, and $S(\mathsf{IRG})$. Hence they are of the form

$$\tilde{\mathsf{R}} V \tilde{\mathsf{G}} \mathsf{D}^j \quad \text{or} \quad \tilde{\mathsf{R}} V \tilde{\mathsf{G}} \mathsf{I} \tilde{\mathsf{R}},$$

where $j \in \mathbb{N}_0$ and each of $\tilde{\mathsf{R}}$, $\tilde{\mathsf{G}}$, and $V \in S(\Phi)$ may be absent and $V\tilde{\mathsf{G}}\mathsf{I}$ does not specialize to EI. The normal forms follow from Eq. (3). □

5.1. Computational aspects

In the following, we shortly mention some computational details of the tensor setting with specialization for integro-differential operators with linear substitutions. Applying `TenReS` to the reduction system Σ given by Tables 1 and 2, in total 87 ambiguities and corresponding S-polynomials are generated. All ambiguities are resolvable and the automatic verification can be found in the example files of the package. There are 66 ambiguities without specialization. For instance,

$$\mathrm{SP}(\mathsf{IR}\underline{\Phi}, \underline{\mathsf{EI}}) = (\textstyle\int f \otimes \mathrm{E}) \otimes \int - \int \otimes f \otimes 0 \to_{r_{\mathsf{EI}}} \int f \otimes 0 = 0,$$

and

$$\begin{aligned}\mathrm{SP}(\mathsf{I}\underline{\mathsf{G}}, \underline{\mathsf{G}}\mathsf{R}) &= (a^{-1}\sigma_{a,b} \otimes \textstyle\int - a^{-1}\mathrm{E} \otimes \sigma_{a,b} \otimes \int) \otimes f - \int \otimes (\sigma_{a,b} f \otimes \sigma_{a,b}) \\ &= a^{-1}\sigma_{a,b} \otimes \textstyle\int \otimes f - a^{-1}\mathrm{E} \otimes \sigma_{a,b} \otimes \int \otimes f - \int \otimes \sigma_{a,b} f \otimes \sigma_{a,b} \to_{r_{\mathsf{IRG}}} 0.\end{aligned}$$

The remaining 21 ambiguities consist of 5 overlap ambiguities with specialization and 16 inclusion ambiguities with specialization. They all involve the following three reduction rules (over X)

$$(\mathsf{K}, 1 \mapsto \epsilon),\ (\mathsf{EI}, \mathrm{E} \otimes \textstyle\int \mapsto 0),\ (\mathsf{N}, \sigma_{1,0} \mapsto \epsilon)$$

and their S-polynomials can be reduced to zero. For example,

$$\mathrm{SP}(\underline{\mathsf{N}}, \mathsf{D}\underline{\mathsf{G}}) = \partial \otimes \epsilon - \sigma_{1,0} \otimes \partial \to_{r_{\mathsf{N}}} \partial - \partial = 0,$$

and

$$\mathrm{SP}(\underline{\mathsf{N}}, \mathsf{IR}\underline{\mathsf{G}}) = \textstyle\int \otimes f - (\epsilon - \mathrm{E}) \otimes \sigma_{1,0} \otimes \int \otimes f \to_{r_{\mathsf{N}}} \mathrm{E} \otimes \sigma_{1,0} \otimes \int \otimes f \to_{r_{\mathsf{N}}} \mathrm{E} \otimes \int \otimes f \to_{r_{\mathsf{EI}}} 0.$$

6. Completion of tensor reduction systems

For computing in the quotient ring $K\langle M\rangle/I_\Sigma$, we would like to compute with a system of representatives. By Theorem 6, the irreducible tensors $K\langle M\rangle_{\mathrm{irr}}$ are such a system if the tensor reduction system is confluent. If the reduction system is not confluent, we want to construct a confluent one that generates the same reduction ideal of Eq. (4).

Like Buchberger's algorithm (Buchberger, 1965) and Knuth–Bendix completion (Knuth and Bendix, 1970), the completion process involves adding new rules corresponding to non-resolvable ambiguities (S-polynomials resp. critical pairs); see also Buchberger (1987). Obstructions for general algorithms are inherited from the noncommutative polynomial algebra case (Mora, 1994), e.g., deciding existence of finite Gröbner bases and the undecidability of the word problem. Unlike noncommutative Gröbner basis computations and Knuth–Bendix completion, where we have semi-decision algorithms, the method we describe for completing tensor reduction systems involves also non-algorithmic steps. One

of the main difficulties is to define a new reduction homomorphism based on the S-polynomials of a non-resolvable ambiguity. Since for verification of confluence, a compatible semigroup partial order is sufficient, one can also start the completion process with a compatible semigroup partial order instead of a total one. Extending this order in a compatible way may not always be possible. We refer also to Caboara (1993), Gritzmann and Sturmfels (1993) for variants of Buchberger's algorithm in the commutative case that do not assume a total term ordering as input.

Before we discuss aspects of the completion process for tensor reduction systems more formally below, we have a look at a few concrete non-resolvable ambiguities. We start with the following reduction rules for integro-differential operators that follow immediately from the definition:

$$\Sigma_0 = \{(\mathsf{K}, 1 \mapsto \epsilon), (\mathsf{RR}, f \otimes g \mapsto fg), (\Phi\mathsf{R}, \varphi \otimes f \mapsto (\varphi f)\varphi), (\Phi\Phi, \psi \otimes \varphi \mapsto \varphi),$$
$$(\mathsf{DR}, \partial \otimes f \mapsto f \otimes \partial + \partial f), (\mathsf{D}\Phi, \partial \otimes \varphi \mapsto 0), (\mathsf{DI}, \partial \otimes \textstyle\int \mapsto \epsilon), (\mathsf{ID}, \int \otimes \partial \mapsto \epsilon - \mathrm{E})\}$$

On $\langle Z \rangle$ we define a partial order $\leq$ based on the length of words with the additional property that $\mathsf{DR} > \mathsf{RD}$. Generating from it the minimal partial order that is consistent with specialization means that we also have to define $\mathsf{DK} > \mathsf{KD}$, $\mathsf{DK} > \tilde{\mathsf{R}}\mathsf{D}$, $\mathsf{D}\tilde{\mathsf{R}} > \mathsf{KD}$, and $\mathsf{D}\tilde{\mathsf{R}} > \tilde{\mathsf{R}}\mathsf{D}$. In order to obtain the minimal semigroup partial order generated by that, we not only have to define $A\mathsf{DR}B > A\mathsf{RD}B$ for any $A, B \in \langle Z \rangle$, but also for all $k \geq 2$ the general condition $A_1\mathsf{DR}A_2\mathsf{DR}\dots\mathsf{DR}A_k > A_1\mathsf{RD}A_2\mathsf{RD}\dots\mathsf{RD}A_k$ for all $A_i \in \langle Z \rangle$ along with all 2^{2k-2} specializations $\mathsf{R} \in \{\mathsf{K}, \tilde{\mathsf{R}}\}$. The resulting semigroup partial order $\leq$ is compatible with Σ_0.

The rules r_{DI} and r_{ID} have two overlap ambiguities with each other, one is resolvable and one is not. The latter has S-polynomial

$$\mathrm{SP}(\underline{\mathsf{ID}}, \underline{\mathsf{DI}}) = (\epsilon - \mathrm{E}) \otimes \textstyle\int - \int \otimes \epsilon = -\mathrm{E} \otimes \int.$$

This trivially gives rise to the new rule

$$(\mathsf{EI}, \mathrm{E} \otimes \textstyle\int \mapsto 0).$$

The rules r_{ID} and r_{DR} have a non-resolvable overlap ambiguity with S-polynomials

$$\mathrm{SP}(\underline{\mathsf{ID}}, \underline{\mathsf{DR}}) = (\epsilon - \mathrm{E}) \otimes f - \textstyle\int \otimes (f \otimes \partial + \partial f) \to_{r_{\Phi\mathsf{R}}} f - (\mathrm{E}f)\mathrm{E} - \int \otimes f \otimes \partial - \int \otimes \partial f.$$

While we could reduce further, by using r_{K} for example, we will not be able to reduce to zero for all $f \in R$. Based on the expression above, however, we can introduce a new rule

$$(\mathsf{IRD}, \textstyle\int \otimes f \otimes \partial \mapsto f - (\mathrm{E}f)\mathrm{E} - \int \otimes \partial f)$$

that allows to reduce all the S-polynomials of the overlap ambiguity of r_{ID} and r_{DR} to zero. This rule gives rise to a non-resolvable overlap ambiguity with r_{DI} among others. The corresponding S-polynomials can be reduced to

$$\mathrm{SP}(\mathsf{IR}\underline{\mathsf{D}}, \underline{\mathsf{DI}}) = (f - (\mathrm{E}f)\mathrm{E} - \textstyle\int \otimes \partial f) \otimes \int - \int \otimes f \otimes \epsilon \to_{r_{\mathsf{EI}}} f \otimes \textstyle\int - \int \otimes \partial f \otimes \int - \int \otimes f.$$

We would like to have a new reduction homomorphism on M_{IRI} that reduces the tensor $\int \otimes \partial f \otimes \int$ to $f \otimes \int - \int \otimes f$. Replacing f by $\int f$, we arrive at the definition

$$(\mathsf{IRI}, \textstyle\int \otimes f \otimes \int \mapsto \int f \otimes \int - \int \otimes \int f).$$

Finally, we consider the inclusion ambiguity (with specialization) of this new rule with r_{K}, which has irreducible S-polynomials

$$\mathrm{SP}(\underline{\mathsf{K}}, \underline{\mathsf{IRI}}) = \textstyle\int \otimes \epsilon \otimes \int - (\int 1 \otimes \int - \int \otimes \int 1) = \int \otimes \int - \int 1 \otimes \int + \int \otimes \int 1.$$

At this point, the leading term is not determined by our partial order above. We decide to have the new rule

$$(\mathsf{II}, \textstyle\int \otimes \int \mapsto \int 1 \otimes \int - \int \otimes \int 1)$$

and extend $\leq$ accordingly to have it compatible with the new rule. Similarly, the overlap ambiguity of r_{IRD} and $r_{\mathsf{D}\Phi}$ gives rise to the rule $r_{\mathsf{IR}\Phi}$, which in turn has an inclusion ambiguity with r_{K} giving

rise to $r_{|\Phi}$. Thereby we obtain the reduction system given in Table 1. The whole completion process for both Table 1 and 2 can be found in the example files of the `TenReS` package.

In the following, we discuss these issues more formally. For a better overview we consider three different tensor settings starting with the special case of a total order for Bergman's original setting, which already covers most issues that may arise during the completion process. Incrementally we discuss the problems arising in more general situations below. After that we illustrate some of those problems by revisiting the computations done for Σ_0 above.

Bergman's tensor setting with a total order Based on the direct sum decomposition (2) into word modules M_W we define the support of a tensor $t \in K\langle M\rangle$ by

$$\operatorname{supp}(t) := \{W \in \langle X\rangle \mid \pi_W(t) \neq 0\}, \tag{22}$$

where π_W denotes the canonical projection onto the direct summand M_W of $K\langle M\rangle$. For each non-resolvable ambiguity, the following points have to be considered.

- We apply a sequence of reductions uniformly to the bimodule generated by S-polynomials to obtain a new bimodule S_{red} generated by reduced S-polynomials. It is not necessary to have $S_{\mathrm{red}} \subseteq K\langle M\rangle_{\mathrm{irr}}$.
- Among all possible supports $\operatorname{supp}(S_{\mathrm{red}}) = \{\operatorname{supp}(t) \mid t \in S_{\mathrm{red}}\}$ we pick some nonempty support $S \in \operatorname{supp}(S_{\mathrm{red}})$, e.g. a maximal element of $\operatorname{supp}(S_{\mathrm{red}})$ w.r.t. $\subseteq$. The total order $\leq$ determines a maximal element $W \in S$, determining the "leading term" of the corresponding tensors in S_{red}.
- A new homomorphism h should be defined on M_W that allows to reduce $t \in S_{\mathrm{red}} \cap M_S$ with $\pi_W(t) \neq 0$ to zero, where M_S is defined in Eq. (7) as the sum of all modules M_V with $V \in S$. In addition, h has to be defined such that $\mathrm{id} - h$ maps M_W into I_Σ, i.e. the reduction ideal stays the same $I_\Sigma = I_{\Sigma\cup\{(W,h)\}}$. To discuss this we consider the subbimodule N of S_{red} generated by all $t \in S_{\mathrm{red}} \cap M_S$ with $\pi_W(t) \neq 0$. This bimodule N is contained in $S_{\mathrm{red}} \cap M_S$, but they are not necessarily equal. If $\pi_W : N \to M_W$ is bijective, then it is natural to define h via $h(\pi_W(t)) = \pi_W(t) - t$. Such a homomorphism may not exist for two reasons.
  - If there are distinct $t_1, t_2 \in N$ with $\pi_W(t_1) = \pi_W(t_2)$, then we cannot have $h(\pi_W(t_1)) = \pi_W(t_1) - t_1$ and $h(\pi_W(t_2)) = \pi_W(t_2) - t_2$ at the same time. In that case, we need to be content with some homomorphism $g\colon M_W \to N$ such that $h(t) = t - g(t)$ and $\pi_W \circ g = \mathrm{id}$. As a consequence $t_1 - t_2 \in S_{\mathrm{red}}$ may still not be reducible to zero with $\Sigma \cup \{(W,h)\}$.
  - If there is a $t \in M_W$ that is not in $\pi_W(N)$, then it is not clear how to define h on all of M_W so that $\leq$ is still compatible with $\Sigma \cup \{(W,h)\}$, in particular $\pi_W(h(M_W)) = \{0\}$, without violating $I_\Sigma = I_{\Sigma\cup\{(W,h)\}}$. Instead of N, considering the larger bimodule $\tilde{N} := S_{\mathrm{red}} \cap \bigoplus_{V\leq W} M_V$ might satisfy $\pi_W(\tilde{N}) = M_W$. If not, it may be necessary to split some modules M_x, $x \in X$, further in order to turn $\pi_W(N)$ or $\pi_W(\tilde{N})$ into a word module M_V over some new alphabet X.
- Finally, we include the new reduction rule (W,h) into Σ. If $\operatorname{supp}(S_{\mathrm{red}}) \neq \{\emptyset, S\}$, then it may happen that the new rule is not sufficient to reduce all elements of S_{red} to zero. In that case, we need to check resolvability of the current ambiguity again.

Bergman's tensor setting with a partial order The only new issue that appears with a partial order $\leq$ on words that is not a total order, is that the "leading term" of tensors in S_{red} may not be determined by the order. If the selected support $S \in \operatorname{supp}(S_{\mathrm{red}})$ does not have a greatest element already, we need to choose a word $W \in S$ so that we can extend the semigroup partial order in a compatible way, i.e. W becomes the greatest element of S. Such a choice is not guaranteed to exist.

Tensor setting with specialization The first thing to note is that we cannot have a total order on $\langle Z\rangle$ that is consistent with specialization (as long as $Z \neq X$). All points of the above discussion apply also to decompositions of M with specialization except that $\operatorname{supp}(S_{\mathrm{red}})$ should now be defined as $\operatorname{supp}(S_{\mathrm{red}}) = \bigcup\{\operatorname{supp}(t) \mid t \in S_{\mathrm{red}}\}$ where for a particular tensor t we now define $\operatorname{supp}(t)$ as the set of "all possible supports"

$$\operatorname{supp}(t) := \{S \subseteq \langle Z\rangle \mid t \in M_S, \forall W, \tilde{W} \in S : \pi_W(t) \neq 0 \wedge S(W) \cap S(\tilde{W}) = \emptyset\}.$$

Other than that, no new fundamental obstacles arise in this setting. We just add a few remarks.

It can be advantageous to pick supports with words associated to bigger modules in order to construct reduction homomorphisms h with larger domains. Also, it can be useful to introduce additional letters to the alphabet Z in order to collect some of the bimodules appearing in the process. We illustrate some of the points discussed formally by revisiting the concrete ambiguities treated above.

The first and simplest case above was the overlap ambiguity of r_{ID} and r_{DI} with words I, D, and I. All S-polynomials are irreducible w.r.t. Σ_0 and the bimodule generated by them has $\operatorname{supp}(S_{\mathrm{red}}) = \{\emptyset, \{\mathsf{EI}\}, \{\Phi\mathsf{I}\}\}$. Picking $S = \{\Phi\mathsf{I}\}$ and $W = \Phi\mathsf{I}$ would lead to $\pi_W|_{S_{\mathrm{red}}}$ not being surjective onto M_W. So the choice $S = \{\mathsf{EI}\}$ and $W = \mathsf{EI}$ is preferable and we can define the homomorphism $h\colon M_W \to K\langle M\rangle$ of r_{EI} by $h(\pi_W(t)) = \pi_W(t) - t = 0$ in this case.

For the overlap ambiguity of r_{ID} and r_{DR} we applied the reduction $h_{\epsilon, r_{\Phi\mathsf{R}}, \epsilon}$ to all S-polynomials. The bimodule generated by them now has

$$\operatorname{supp}(S_{\mathrm{red}}) = \{\emptyset, \{\mathsf{K}, \mathsf{E}, \mathsf{IKD}\}, \{\tilde{\mathsf{R}}, \mathsf{I}\tilde{\mathsf{R}}\mathsf{D}, \mathsf{IK}\}, \{\tilde{\mathsf{R}}, \mathsf{I}\tilde{\mathsf{R}}\mathsf{D}, \mathsf{I}\tilde{\mathsf{R}}\}, \{\tilde{\mathsf{R}}, \mathsf{I}\tilde{\mathsf{R}}\mathsf{D}, \mathsf{IR}\},$$
$$\{\mathsf{R}, \mathsf{E}, \mathsf{IRD}, \mathsf{IK}\}, \{\mathsf{R}, \mathsf{E}, \mathsf{IRD}, \mathsf{I}\tilde{\mathsf{R}}\}, \{\mathsf{R}, \mathsf{E}, \mathsf{IRD}, \mathsf{IR}\}, \dots\}.$$

The chosen partial order $\leq$ determines a greatest element of most $S \in \operatorname{supp}(S_{\mathrm{red}})$. Picking $S \in \operatorname{supp}(S_{\mathrm{red}})$ with the largest M_S gives $S = \{\mathsf{R}, \Phi, \mathsf{IRD}, \mathsf{IR}\}$ and $W = \mathsf{IRD}$ so that $\pi_W\colon S_{\mathrm{red}} \to M_W$ is bijective. This allows for a straightforward definition of r_{IRD} again.

A more interesting case is the overlap ambiguity of r_{IRD} and r_{DR} with words IR, D, R. After applying the reduction $h_{\epsilon, r_{\mathsf{EI}}, \epsilon}$ to all S-polynomials the bimodule generated by them has

$$\operatorname{supp}(S_{\mathrm{red}}) = \{\emptyset, \{\mathsf{KI}, \mathsf{IK}\}, \{\tilde{\mathsf{R}}\mathsf{I}, \mathsf{IKI}, \mathsf{I}\tilde{\mathsf{R}}\}, \{\mathsf{RI}, \mathsf{IKI}, \mathsf{IR}\}, \{\tilde{\mathsf{R}}\mathsf{I}, \mathsf{I}\tilde{\mathsf{R}}\mathsf{I}, \mathsf{I}\tilde{\mathsf{R}}\}, \{\mathsf{RI}, \mathsf{I}\tilde{\mathsf{R}}\mathsf{I}, \mathsf{IR}\}, \{\tilde{\mathsf{R}}\mathsf{I}, \mathsf{IRI}, \mathsf{I}\tilde{\mathsf{R}}\},$$
$$\{\mathsf{RI}, \mathsf{IRI}, \mathsf{IR}\}, \{\mathsf{RI}, \mathsf{IR}\}, \{\mathsf{KI}, \tilde{\mathsf{R}}\mathsf{I}, \mathsf{IKI}, \mathsf{IK}, \mathsf{I}\tilde{\mathsf{R}}\}, \dots\}.$$

Picking again one $S \in \operatorname{supp}(S_{\mathrm{red}})$ with the largest M_S gives $S = \{\mathsf{RI}, \mathsf{IRI}, \mathsf{IR}\}$ and $W = \mathsf{IRI}$. Now $N = S_{\mathrm{red}}$ and $\pi_W\colon S_{\mathrm{red}} \to M_W$ is surjective but not injective. We choose the bimodule homomorphism $g\colon M_W \to S_{\mathrm{red}}$ to be defined by $g(\int \otimes f \otimes \int) = \int \otimes f \otimes \int + \int \otimes \int f - \int f \otimes \int$. It satisfies $\pi_W \circ g = \mathrm{id}$ and we define the homomorphism $h\colon M_W \to K\langle M\rangle$ of r_{IRI} by $h := \mathrm{id} - g$. While $h_{\epsilon, r_{\mathsf{IRI}}, \epsilon}$ does not map S_{red} to $\{0\}$, the image contains only elements of the form $c \otimes \int - \int \otimes c$ with $c \in K$, which are reducible to zero by Σ_0.

The last ambiguity dealt with explicitly above is the inclusion ambiguity (with specialization) of r_{IRI} and r_{K}. Its S-polynomials are irreducible and we have

$$\operatorname{supp}(S_{\mathrm{red}}) = \{\emptyset, \{\mathsf{II}, \tilde{\mathsf{R}}\mathsf{I}, \mathsf{I}\tilde{\mathsf{R}}\}, \{\mathsf{II}, \tilde{\mathsf{R}}\mathsf{I}, \mathsf{IR}\}, \{\mathsf{II}, \mathsf{RI}, \mathsf{I}\tilde{\mathsf{R}}\}, \{\mathsf{II}, \mathsf{RI}, \mathsf{IR}\}\}.$$

As pointed out already, the partial order does not determine a greatest element within any of the possible supports. Since $\pi_W\colon S_{\mathrm{red}} \to M_W$ is not surjective except for $W = \mathsf{II}$, we would have to split $M_{\tilde{\mathsf{R}}}$ further in order to define a new reduction rule on $\pi_W(S_{\mathrm{red}})$ in all other cases. So we choose $W = \mathsf{II}$ and extend the semigroup partial order such that $\mathsf{II} > \tilde{\mathsf{R}}\mathsf{I}$ and $\mathsf{II} > \mathsf{I}\tilde{\mathsf{R}}$.

7. Concluding remarks

A ring of operators may not be finitely presented by generators and relations, it may not even be finitely generated. The tensor setting nonetheless often allows to have a finite decomposition of the module M of basic operators together with a finite reduction system. Reduction rules need to be defined by homomorphisms due to non-uniqueness of the representation of tensors. In addition, homomorphisms collect families of relations into one reduction rule. If a reduction system is confluent, the normal forms are unique as tensors while tensors themselves do not have unique representations in terms of pure tensors. Both the theoretical concepts and the concrete formulae for the reduction systems in the examples presented essentially are the same when working in the tensor algebra or in the tensor ring.

In comparison to Bergman's tensor setting, our tensor setting with specialization allows more flexibility in defining a reduction system for a given ring of operators. This is achieved by relaxing the

restriction that the submodules of M that are used for defining the reduction homomorphisms have to form a direct sum. As a consequence, reduction systems can be smaller and reduction is more efficient by avoiding unnecessary splitting.

Already when we compare quotients of the tensor algebra with quotients of the free algebra we note some important differences. All computations in quotients of the free algebra happen on two levels: polynomial arithmetic in the free algebra and polynomial reduction modulo the ideal. Computations in the K-algebra $K\langle M\rangle$ actually take place on three levels. The additional level are computations in the module M and its submodules M_z. Analogous to the free algebra there are computations in $K\langle M\rangle$ coming from the properties of the tensor product and the reduction system that acts by applying the reduction homomorphisms.

Depending on the choice of the module M and its decomposition, certain identities of operators either are dealt with by the reduction system or only within the module M. One extreme case occurs when M already is the whole K-ring of operators. Then the reduction system only consists of the rules $1 \mapsto \epsilon$ and $m_1 \otimes m_2 \mapsto m_1 m_2$ which do not expose any structure of the ring of operators. Another extreme case occurs when M is some module that generates the ring of operators and all M_z are cyclic. Then the reduction system has to encode all identities among those generators, which makes it harder to have a finite reduction system. For instance, any confluent reduction system for IDOs with polynomial coefficients $K[x]\langle\partial, \int, \mathrm{E}\rangle$, $\mathbb{Q} \subseteq K$, is infinite if M is just generated by x, ∂, $\int$, and E. In between those two extreme cases there is the opportunity to encode only part of the identities by the reduction system and "hide" the remaining ones inside the modules M_z. For instance, following the construction of $K[x]\langle\partial, \int, \mathrm{E}\rangle$, $\mathbb{Q} \subseteq K$, given in Section 4 the module M consists of $K[x]$ and the modules generated by ∂, $\int$, and E and the confluent reduction system given in Table 1 with $R = K[x]$ is finite. Finiteness of this reduction system can be understood by recalling that reduction rules can collect many identities of the same form into one reduction homomorphism.

In principle, if M is a free module, one could reformulate each reduction rule in terms of reduction rules on individual basis elements and work in the free algebra without making use of tensors. Consequently, computations with the reduction system would then have to use basis expansion in each step. In the tensor setting, however, we do not need to fix a basis of the module M. It is enough to work with the decomposition into modules M_z, which also enables working with non-free modules. This even allows to consider arbitrary modules M that are not concrete but carry a certain algebraic structure. For example, the reduction systems and the computations for checking their confluence in Sections 4 and 5 do not rely on a concrete integro-differential ring R.

Based on the normal forms, a confluent reduction system for a ring of operators enables to automatize many computations and proofs involving these operators. The confluent reduction systems given for IDOs and IDOs with linear substitutions can be used e.g. to prove the Taylor formula, to compute Green's operators of linear ordinary boundary problems, or to support computations in Artstein's reduction of linear time-delay systems. Since K is neither required to be a field nor commutative, we can directly consider operators with matrix coefficients to model systems. Elements in R can even model matrices of generic size. The tensor setting can also be used to model other rings of operators. For example, we already have results for IDOs with more general types of functionals or a discrete analog of IDOs.

Acknowledgements

We would like to thank Thomas Cluzeau and Alban Quadrat for discussions. We would also like to thank the anonymous referees for their remarks that helped to improve the presentation of the material and for pointing out one small problem in an earlier version of the proof of Lemma 10.

References

Baader, F., Nipkow, T., 1998. Term Rewriting and All That. Cambridge University Press, Cambridge.
Bavula, V.V., 2013. The algebra of integro-differential operators on an affine line and its modules. J. Pure Appl. Algebra 217, 495–529.
Bergman, G.M., 1978. The diamond lemma for ring theory. Adv. Math. 29, 178–218.
Bokut, L.A., Chen, Y., 2014. Gröbner–Shirshov bases and their calculation. Bull. Math. Sci. 4, 325–395.

Buchberger, B., 1965. An Algorithm for Finding the Bases Elements of the Residue Class Ring Modulo a Zero Dimensional Polynomial Ideal (German). Ph.D. thesis. University of Innsbruck.
Buchberger, B., 1987. History and basic features of the critical-pair/completion procedure. J. Symb. Comput. 3, 3–38.
Bueso, J., Gómez-Torrecillas, J., Verschoren, A., 2003. Algorithmic Methods in Non-Commutative Algebra. Kluwer Academic Publishers, Dordrecht.
Caboara, M., 1993. A dynamic algorithm for Gröbner basis computation. In: Bronstein, M. (Ed.), Proceedings of ISSAC '93. ACM, New York, NY, USA, pp. 275–283.
Chyzak, F., Quadrat, A., Robertz, D., 2005. Effective algorithms for parametrizing linear control systems over Ore algebras. Appl. Algebra Eng. Commun. Comput. 16, 319–376.
Chyzak, F., Salvy, B., 1998. Non-commutative elimination in Ore algebras proves multivariate identities. J. Symb. Comput. 26, 187–227.
Cohn, P.M., 2003. Further Algebra and Applications. Springer-Verlag London, Ltd., London.
Coutinho, S.C., 1995. A Primer of Algebraic D-Modules. Cambridge University Press, Cambridge.
Gao, X., Guo, L., 2017. Rota's Classification Problem, rewriting systems and Gröbner–Shirshov bases. J. Algebra 470, 219–253.
Gao, X., Guo, L., Rosenkranz, M., 2015. Free integro-differential algebras and Gröbner–Shirshov bases. J. Algebra 442, 354–396.
Gao, X., Guo, L., Zheng, S., 2014. Construction of free commutative integro-differential algebras by the method of Gröbner–Shirshov bases. J. Algebra Appl. 13, 1350160, 38pp.
Gómez-Torrecillas, J., 2014. Basic module theory over non-commutative rings with computational aspects of operator algebras. In: Barkatou, M., Cluzeau, T., Regensburger, G., Rosenkranz, M. (Eds.), AADIOS 2012. In: LNCS, vol. 8372. Springer, Heidelberg, pp. 23–82. With an appendix by V. Levandovskyy.
Gritzmann, P., Sturmfels, B., 1993. Minkowski addition of polytopes: computational complexity and applications to Gröbner bases. SIAM J. Discrete Math. 6, 246–269.
Guo, L., Regensburger, G., Rosenkranz, M., 2014. On integro-differential algebras. J. Pure Appl. Algebra 218, 456–473.
Guo, L., Sit, W.Y., Zhang, R., 2013. Differential type operators and Gröbner–Shirshov bases. J. Symb. Comput. 52, 97–123.
Hale, J.K., Verduyn Lunel, S.M., 1993. Introduction to Functional–Differential Equations. Springer-Verlag, New York.
Helton, J.W., Stankus, M., 1999. Computer assistance for "discovering" formulas in system engineering and operator theory. J. Funct. Anal. 161, 289–363.
Helton, J.W., Stankus, M., Wavrik, J.J., 1998. Computer simplification of formulas in linear systems theory. IEEE Trans. Autom. Control 43, 302–314.
Hossein Poor, J., Raab, C.G., Regensburger, G., 2016a. Algorithmic operator algebras via normal forms for tensors. In: Rosenkranz, M. (Ed.), Proceedings of ISSAC '16. ACM, New York, NY, USA, pp. 397–404.
Hossein Poor, J., Raab, C.G., Regensburger, G., 2016b. Normal forms for operators via Gröbner bases in tensor algebras. In: Greuel, G.-M., Koch, T., Paule, P., Sommese, A. (Eds.), Proceedings of ICMS 2016. In: LNCS, vol. 9725. Springer, pp. 505–513.
Knuth, D.E., Bendix, P.B., 1970. Simple word problems in universal algebras. In: Computational Problems in Abstract Algebra. Pergamon, Oxford, pp. 263–297.
Kobayashi, Y., 2005. Gröbner bases of associative algebras and the Hochschild cohomology. Trans. Am. Math. Soc. 357, 1095–1124.
Levandovskyy, V., 2005. Non-Commutative Computer Algebra for Polynomial Algebras: Gröbner Bases, Applications and Implementation. Ph.D. thesis. Universität Kaiserslauten.
Li, H., 2002. Noncommutative Gröbner Bases and Filtered-Graded Transfer. Lecture Notes in Mathematics, vol. 1795. Springer-Verlag, Berlin.
Mora, T., 1994. An introduction to commutative and noncommutative Gröbner bases. Theor. Comput. Sci. 134, 131–173.
Quadrat, A., 2015. A constructive algebraic analysis approach to Artstein's reduction of linear time-delay systems. IFAC-PapersOnLine 48 (12), 209–214.
Quadrat, A., Regensburger, G., 2017. Computing polynomial solutions and annihilators of integro-differential operators with polynomial coefficients. hal-01413907, in press.
Regensburger, G., 2016. Symbolic computation with integro-differential operators. In: Rosenkranz, M. (Ed.), Proceedings of ISSAC '16. ACM, New York, NY, USA, pp. 17–18.
Regensburger, G., Rosenkranz, M., Middeke, J., 2009. A skew polynomial approach to integro-differential operators. In: Proceedings of ISSAC '09. ACM, New York, NY, USA, pp. 287–294.
Román García, M., Román García, S., 2005. Gröbner bases and syzygies on bimodules over PBW algebras. J. Symb. Comput. 40, 1039–1052.
Rosenkranz, M., 2005. A new symbolic method for solving linear two-point boundary value problems on the level of operators. J. Symb. Comput. 39, 171–199.
Rosenkranz, M., Buchberger, B., Engl, H.W., 2003. Solving linear boundary value problems via non-commutative Gröbner bases. Appl. Anal. 82, 655–675.
Rosenkranz, M., Gao, X., Guo, L., 2015. An algebraic study of multivariable integration and linear substitution. arXiv:1503.01694 [math.RA].
Rosenkranz, M., Regensburger, G., 2008. Solving and factoring boundary problems for linear ordinary differential equations in differential algebras. J. Symb. Comput. 43, 515–544.
Rosenkranz, M., Regensburger, G., Tec, L., Buchberger, B., 2012. Symbolic analysis for boundary problems: from rewriting to parametrized Gröbner bases. In: Langer, U., Paule, P. (Eds.), Numerical and Symbolic Scientific Computing: Progress and Prospects. Springer, Vienna, pp. 273–331.
Rowen, L.H., 1991. Ring Theory, student edition. Academic Press, Inc., Boston, MA.
Smith, H., 2011. An Introduction to Delay Differential Equations with Applications to the Life Sciences. Springer, New York.

SIAM J. APPL. MATH.
Vol. 72, No. 6, pp. 1926–1947

GENERALIZED MASS ACTION SYSTEMS: COMPLEX BALANCING EQUILIBRIA AND SIGN VECTORS OF THE STOICHIOMETRIC AND KINETIC-ORDER SUBSPACES*

STEFAN MÜLLER† AND GEORG REGENSBURGER‡

Abstract. Mass action systems capture chemical reaction networks in homogeneous and dilute solutions. We suggest a notion of generalized mass action systems that admits arbitrary power-law rate functions and serves as a more realistic model for reaction networks in intracellular environments. In addition to the complexes of a network and the related stoichiometric subspace, we introduce corresponding kinetic complexes, which represent the exponents in the rate functions and determine the kinetic-order subspace. We show that several results of chemical reaction network theory carry over to the case of generalized mass action kinetics. Our main result essentially states that if the sign vectors of the stoichiometric and kinetic-order subspace coincide, there exists a unique complex balancing equilibrium in every stoichiometric compatibility class. However, in contrast to classical mass action systems, multiple complex balancing equilibria in one stoichiometric compatibility class are possible in general.

Key words. chemical reaction network theory, generalized mass action kinetics, complex balancing, generalized Birch's theorem, oriented matroids

AMS subject classifications. 92C42, 37C25, 52C40

DOI. 10.1137/110847056

1. Introduction.

Dynamical systems arising from chemical reaction networks with mass action kinetics are the subject of chemical reaction network theory (CRNT), which was initiated by the work of Horn, Jackson, and Feinberg; cf. [25, 24, 13]. In particular, this theory provides results about existence, uniqueness, and stability of equilibria *independently* of rate constants (and initial conditions). However, the validity of the underlying mass action law is limited; it only holds for elementary reactions in homogeneous and dilute solutions. In intracellular environments, which are highly structured and characterized by macromolecular crowding, the rate law has to be modified; cf. [8, 23, 28].

Two types of modifications have been proposed: "fractal reaction kinetics" [26, 27, 38, 21] and the "power-law formalism" [34, 35, 36, 37]. The names of the two approaches are a bit misleading since both approaches address the problem of dimensional restriction (i.e., molecules confined to surfaces, channels, or fractal-like structures) and both use power-laws. More specifically, in fractal-like kinetics, rate constants are time-dependent (via a power-law), whereas the exponents of the species concentrations in the rate function are the corresponding stoichiometric coefficients (as in mass action kinetics). On the other hand, in the power-law formalism, rate constants are time-independent (as in mass action kinetics), whereas the exponents of the species concentrations may be (nonnegative) real numbers different from the respective stoichiometric coefficients. For model selection, data have to be collected

*Received by the editors September 6, 2011; accepted for publication (in revised form) August 2, 2012; published electronically December 19, 2012.
http://www.siam.org/journals/siap/72-6/84705.html
†Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, 4040 Linz, Austria (stefan.mueller@ricam.oeaw.ac.at).
‡INRIA Saclay – Île de France, Project DISCO, L2S, Supélec, 91192 Gif-sur-Yvette Cedex, France (georg.regensburger@ricam.oeaw.ac.at). This author's work was supported by the Austrian Science Fund (FWF): J3030-N18.

for many molecules and intracellular environments. Recent data of binding kinetics in crowded media [2, 30] suggest that the power-law formalism is the preferred model.

In this work, we study the consequences of the power-law formalism for chemical reaction networks. In particular, we demonstrate that several fundamental results of CRNT carry over to the case of generalized mass action kinetics (i.e., power-law rate functions). There has been an early approach to account for generalized mass action kinetics [25], which entails a redefinition of the complexes of a network. Here, we suggest a different approach, where we keep the original complexes, but introduce additional "kinetic complexes," which determine the exponents of the species concentrations in the rate functions. This has the advantage that the underlying chemical reaction network and thus properties like weak reversibility and deficiency remain the same.

From the kinetic complexes, we obtain (in addition to the stoichiometric subspace) a "kinetic-order subspace," and it turns out that the generalization of a central result of CRNT (concerned with the uniqueness and existence of a complex balancing equilibrium in every stoichiometric compatibility class) depends on the sign vectors of the two subspaces. Our main result, Theorem 3.10, essentially states that if these sign vectors are equal, there exists a unique complex balancing equilibrium in every stoichiometric compatibility class. In general, however, there may be more than one complex balancing equilibrium in a stoichiometric compatibility class; see Proposition 3.2 and Example 4.2.

Chemical reaction networks with nonmass action kinetics are also studied in [5, 4, 3]. In this approach, one is interested in conditions that guarantee the uniqueness of equilibria. If autocatalytic reactions are excluded and if the dependence of the rate functions on the species concentrations corresponds to the stoichiometric matrix, the structure of the stoichiometric matrix alone guarantees uniqueness. Moreover, the properties of the stoichiometric matrix can be translated into conditions for the species reaction graph. As a consequence, this theory is applicable to many types of kinetics; however, it does not address the existence of equilibria. Existence and uniqueness of equilibria for general kinetics are discussed in [12]. The methods are based on homotopy invariance of the Brouwer degree in a way related to the approach in section 3.3.

Organization of the work. In the next section, we recall the definition of mass action systems and several fundamental results of CRNT. Then we introduce generalized mass action systems and discuss the results that carry over easily to this framework. In section 3, we study uniqueness and existence of complex balancing equilibria; more specifically, we reformulate the problem and study injectivity and surjectivity of a certain map (a simplified version of), which appears, for example, in toric and computational geometry or statistics. In section 4, we discuss two examples of generalized mass action systems. Finally, we draw our conclusions and give an outlook to further lines of research. In the appendix, we recall the relevant results on sign vectors of vector spaces and face lattices of polyhedral cones and polytopes.

Notation. We denote the positive real numbers by $\mathbb{R}_>$ and the nonnegative real numbers by $\mathbb{R}_\geq$. For a finite index set I, we write $\mathbb{R}^I$ for the real vector space of formal sums $x = \sum_{i\in I} x_i\, i$ with $x_i \in \mathbb{R}$, and $\mathbb{R}^I_>$ and $\mathbb{R}^I_\geq$ for the corresponding subsets. Given $x \in \mathbb{R}^I$, we write $x > 0$ if $x \in \mathbb{R}^I_>$ and $x \geq 0$ if $x \in \mathbb{R}^I_\geq$. Further, we define $\mathrm{e}^x \in \mathbb{R}^I_>$ and $\ln(x) \in \mathbb{R}^I$ componentwise, i.e., $(\mathrm{e}^x)_i = \mathrm{e}^{x_i}$ and $(\ln(x))_i = \ln(x_i)$, the latter for $x \in \mathbb{R}^I_>$. Finally, we define $x \circ y \in \mathbb{R}^I$ for $x, y \in \mathbb{R}^I$ as $(x \circ y)_i = x_i y_i$ and $x^y \in \mathbb{R}_\geq$ for $x, y \in \mathbb{R}^I_\geq$ as $x^y = \prod_{i\in I} x_i^{y_i}$, where we set $0^0 = 1$.

2. Chemical reaction networks. In our presentation of CRNT, we follow the surveys by Feinberg [14, 15, 16] and Gunawardena [22].

DEFINITION 2.1. *A chemical reaction network* $(\mathscr{S},\mathscr{C},\mathscr{R})$ *consists of three finite sets:* (i) *a set* $\mathscr{S}$ *of species,* (ii) *a set* $\mathscr{C} \subset \mathbb{R}^{\mathscr{S}}_{\geq}$ *of complexes, and* (iii) *a set* $\mathscr{R} \subset \mathscr{C} \times \mathscr{C}$ *of reactions with the following properties:* (a) *for all* $y \in \mathscr{C}$: $\exists y' \in \mathscr{C}$ *such that* $(y,y') \in \mathscr{R}$ *or* $(y',y) \in \mathscr{R}$ *and* (b) *for all* $y \in \mathscr{C}$: $(y,y) \notin \mathscr{R}$.

Complexes are formal sums of species; they are the left-hand sides and right-hand sides of chemical reactions. For $y \in \mathscr{C}$, we may write $y = \sum_{s\in\mathscr{S}} y_s\, s$, where y_s is the stoichiometric coefficient of species s. As usual in chemistry, we write $y \to y'$ for a reaction $(y,y') \in \mathscr{R}$. In a chemical reaction network, each complex appears in at least one reaction; moreover, there are no reactions of the form $y \to y$.

A chemical reaction network $(\mathscr{S},\mathscr{C},\mathscr{R})$ gives rise to a directed graph with complexes as nodes and reactions as edges. Connected components $L_1,\dots,L_l \subseteq \mathscr{C}$ are called *linkage classes*, strongly connected components are called *strong linkage classes*, and strongly connected components without outgoing edges $T_1,\dots,T_t \subseteq \mathscr{C}$ are called *terminal strong linkage classes*. Each linkage class must contain at least one terminal strong linkage class, i.e., $t \geq l$. The network $(\mathscr{S},\mathscr{C},\mathscr{R})$ is called *weakly reversible* if the linkage classes coincide with the strong linkage classes and hence with the terminal strong linkage classes.

From a dynamic point of view, each reaction $y \to y' \in \mathscr{R}$ causes a change in species concentrations proportional to $y' - y \in \mathbb{R}^{\mathscr{S}}$. The change caused by all reactions lies in a subspace of $\mathbb{R}^{\mathscr{S}}$ such that any trajectory in $\mathbb{R}^{\mathscr{S}}_{\geq}$ lies in a coset of this subspace.

DEFINITION 2.2. *Let* $(\mathscr{S},\mathscr{C},\mathscr{R})$ *be a chemical reaction network. The stoichiometric subspace is defined as*

$$S = \operatorname{span}\{y' - y \in \mathbb{R}^{\mathscr{S}} \mid y \to y' \in \mathscr{R}\}\,.$$

Further, let $c' \in \mathbb{R}^{\mathscr{S}}_{\geq}$. *The corresponding stoichiometric compatibility class is defined as*

$$(c' + S)_{\geq} = (c' + S) \cap \mathbb{R}^{\mathscr{S}}_{\geq}\,.$$

2.1. Mass action systems. The rate of a reaction $y \to y' \in \mathscr{R}$ depends on the concentrations of the species involved. The explicit form of the rate function $\mathscr{K}_{y\to y'}\colon \mathbb{R}^{\mathscr{S}}_{\geq} \to \mathbb{R}_{\geq}$ is determined by the underlying kinetics. In the case of mass action kinetics, it is a monomial in the concentrations $c \in \mathbb{R}^{\mathscr{S}}_{\geq}$ of reactant species, i.e., $\mathscr{K}_{y\to y'}(c) = k_{y\to y'}\, c^y$ with rate constant $k_{y\to y'} \in \mathbb{R}_{>}$. In other words, the stoichiometric coefficient of a species on the left-hand side of the reaction equals the exponent of the corresponding concentration in the rate function. It remains to formally introduce the rate constants.

DEFINITION 2.3. *A mass action system* $(\mathscr{S},\mathscr{C},\mathscr{R},k)$ *is a chemical reaction network* $(\mathscr{S},\mathscr{C},\mathscr{R})$ *together with a vector* $k \in \mathbb{R}^{\mathscr{R}}_{>}$ *of rate constants.*

DEFINITION 2.4. *The ordinary differential equation (ODE) associated with a mass action system* $(\mathscr{S},\mathscr{C},\mathscr{R},k)$ *is defined as*

$$\frac{dc}{dt} = r(c)$$

with the species formation rate

$$r(c) = \sum_{y\to y'\in\mathscr{R}} k_{y\to y'}\, c^y\, (y' - y)\,.$$

In order to rewrite the species formation rate, we use the unit vectors $\omega_y \in \mathbb{R}^{\mathscr{C}}$ corresponding to complexes $y \in \mathscr{C}$ and define

- a linear map[1] $Y\colon \mathbb{R}^{\mathscr{C}} \to \mathbb{R}^{\mathscr{S}}$ with $Y\omega_y = y$,
- a nonlinear map $\Psi\colon \mathbb{R}^{\mathscr{S}}_{\geq} \to \mathbb{R}^{\mathscr{C}}$, $c \mapsto \sum_{y\in\mathscr{C}} c^y\,\omega_y$, and
- a linear map[2] $A\colon \mathbb{R}^{\mathscr{C}} \to \mathbb{R}^{\mathscr{C}}$, $x \mapsto \sum_{y\to y'\in\mathscr{R}} k_{y\to y'}\,x_y\,(\omega_{y'} - \omega_y)$.

Now, the species formation rate can be decomposed as

$$\begin{aligned} r(c) &= \sum_{y\to y'\in\mathscr{R}} k_{y\to y'}\,c^y\,(y' - y) \\ &= Y \sum_{y\to y'\in\mathscr{R}} k_{y\to y'}\,c^y\,(\omega_{y'} - \omega_y) \\ &= Y \sum_{y\to y'\in\mathscr{R}} k_{y\to y'}\,\Psi(c)_y\,(\omega_{y'} - \omega_y) \\ &= Y A\,\Psi(c)\,. \end{aligned} \tag{2.1}$$

Equilibria of the ODE associated with a mass action system satisfying $A\,\Psi(c) = 0$ and $c > 0$ are called *complex balancing equilibria*. The possibility of other (positive) equilibria suggests the definition of the *deficiency* of a mass action system.

DEFINITION 2.5. *Let $(\mathscr{S},\mathscr{C},\mathscr{R},k)$ be a mass action system. The set of complex balancing equilibria is defined as*

$$Z = \{c \in \mathbb{R}^{\mathscr{S}}_{>} \mid A\,\Psi(c) = 0\}\,.$$

The deficiency of the system is defined as

$$\delta = \dim(\ker(Y) \cap \operatorname{im}(A))\,.$$

Originally, the deficiency was defined differently. As we will see in Proposition 2.8, the two definitions coincide under certain conditions on the network structure. In Figure 2.1, we summarize the definitions associated with a mass action system and depict their dependencies.

Results. Now we are in position to present several results of CRNT related to the deficiency zero theorem. (The results are due to Horn, Jackson, and Feinberg [25, 24, 13]. For proofs, we refer the reader to the surveys [14, 16, 22].) As we will see later, corresponding statements also hold in the case of generalized mass action kinetics. We start with a foundational linear algebra result, which can be proved using the Perron–Frobenius theorem.

THEOREM 2.6. *Let $(\mathscr{S},\mathscr{C},\mathscr{R},k)$ be a mass action system with the associated map A, and let $T_1,\dots,T_t \subseteq \mathscr{C}$ be the terminal strong linkage classes. Then*

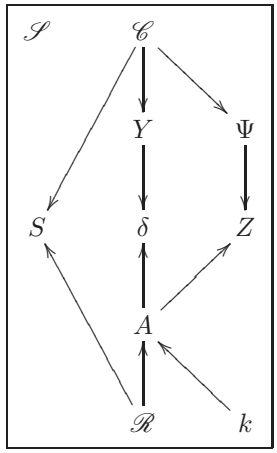
1. *for $i = 1,\dots,t$: $\exists \chi_i \in \mathbb{R}^{\mathscr{C}}_{\geq}$ with $\operatorname{supp}(\chi_i) = T_i$,*
2. *$\ker(A) = \operatorname{span}\{\chi_1,\dots,\chi_t\}$,*
3. *$\dim(\ker(A)) = t$.*

The next result is an immediate consequence of Theorem 2.6.

COROLLARY 2.7. *Let $(\mathscr{S},\mathscr{C},\mathscr{R})$ be a chemical reaction network. If there exist rate constants k such that the mass action system $(\mathscr{S},\mathscr{C},\mathscr{R},k)$ has a complex balancing equilibrium, then $(\mathscr{S},\mathscr{C},\mathscr{R})$ is weakly reversible.*

[1] The corresponding matrix amounts to $Y_{sy} = y_s$.

[2] The corresponding matrix amounts to $A_{yy'} = K_{y'y} - \delta_{yy'}\sum_{y''\in\mathscr{C}} K_{yy''}$, where $K \in \mathbb{R}^{\mathscr{C}\times\mathscr{C}}$ with $K_{yy'} = k_{y\to y'}$ if $y \to y' \in \mathscr{R}$ and $K_{yy'} = 0$ otherwise.



$$\frac{dc}{dt} = Y A\, \Psi(c)$$

FIG. 2.1. *The mass action system $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$: Associated definitions and their dependencies. (Definitions at arrowheads depend on tails.)*

If each linkage class contains exactly one terminal strong linkage class, the deficiency is independent of the rate constants and can be computed from basic parameters of the chemical reaction network. The resulting formula was the original definition of the deficiency.

PROPOSITION 2.8. *If a chemical reaction network $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ is weakly reversible (or more generally if $t = l$), then, for all rate constants k, the deficiency of the mass action system $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$ is given by $\delta = m - l - s$, where m is the number of complexes, l is the number of linkage classes, and s is the dimension of the stoichiometric subspace.*

In the case of deficiency zero, weak reversibility guarantees the existence of complex balancing equilibria.

PROPOSITION 2.9. *If a chemical reaction network $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ is weakly reversible and $\delta = 0$, then, for all rate constants k, the mass action system $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$ has a complex balancing equilibrium.*

Theorem 2.6 further implies that the set of complex balancing equilibria can be parametrized by the orthogonal of the stoichiometric subspace.

PROPOSITION 2.10. *Let $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$ be a mass action system with nonempty set Z of complex balancing equilibria. Then*

$$Z = \{c \in \mathbb{R}^{\mathscr{S}}_{>} \mid \ln(c) - \ln(c^*) \in S^{\perp}\} = \{c^* \circ \mathrm{e}^{v} \mid v \in S^{\perp}\}$$

for any $c^* \in Z$.

Finally, we recall a result concerned with the existence and uniqueness of a complex balancing equilibrium in every stoichiometric compatibility class. It can be proved using methods from convex analysis.

THEOREM 2.11. *Let $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$ be a mass action system with nonempty set Z of complex balancing equilibria. Then Z meets every stoichiometric compatibility class in exactly one point.*

In section 3, we study the conditions under which a result analogous to Theorem 2.11 holds in the case of generalized mass action kinetics.

2.2. Generalized mass action systems. Chemical reactions occur between entire molecules such that the stoichiometric coefficients are integers. Under the assumption of mass action kinetics, the rate functions are monomials in the concentrations of the reactant species. However, in Definition 2.1 we allowed nonnegative real stoichiometric coefficients and hence "generalized monomials" as rate functions, since all results presented above also hold in this generality. This observation can be used to account for generalized mass action kinetics. We outline two different approaches, the second of which is the focus of this paper.

In the first approach [25], chemical reactions are redefined as pseudoreactions with the same net balance, but real stoichiometric coefficients. For example, the reaction

$$n_A A + n_B B \to n_C C$$

with n_A, n_B, $n_C \in \mathbb{N}$ can be redefined as

$$\nu_A A + \nu_B B + \nu_C C \to (\nu_A - n_A)A + (\nu_B - n_B)B + (\nu_C + n_C)C$$

with $\nu_A, \nu_B, \nu_C \in \mathbb{R}_{\geq}$ and rate function $k\,[A]^{\nu_A}[B]^{\nu_B}[C]^{\nu_C}$. The redefinition of chemical reactions does not affect the stoichiometric subspace; however, it entails a new (and typically larger) set of complexes and hence a new mass action system (with different properties). For example, consider the (weakly) reversible chemical reaction network

$$A + B \rightleftharpoons C$$

with two complexes and one linkage class. Since the stoichiometric subspace $S = \operatorname{span}\{(-1, -1, 1)^T\}$ has dimension one, we obtain $\delta = 2-1-1 = 0$ by Proposition 2.8. In order to account for generalized mass action kinetics specified by the rate functions $k_{A+B\to C}[A]^a[B]^b$ and $k_{C\to A+B}[C]^c$ with $a, b, c \in \mathbb{R}_{>}$, the system can be redefined by the pseudoreactions

$$\begin{gathered} aA + bB \to (a-1)A + (b-1)B + C, \\ cC \to A + B + (c-1)C \end{gathered}$$

with four complexes and two linkage classes. This new system is not weakly reversible and has deficiency $\delta = 4 - 2 - 1 = 1$, again by Proposition 2.8.

In this paper, we present a different way to account for generalized mass action kinetics. Most importantly, we disentangle the definition of the rate functions from the stoichiometric coefficients. In particular, we keep the integer stoichiometric coefficients, but we allow "generalized monomials" as rate functions, in which the exponents of the concentrations can be arbitrary nonnegative real numbers. More formally, we do not change the chemical reaction network, but we associate with each complex a so-called *kinetic complex*, which determines the exponents of the concentrations in the rate function of the respective reaction. In the above example, we associate the kinetic complexes $aA + bB$ and cC with $A + B$ and C, thereby specifying the rate

functions $k_{A+B\to C}[A]^a[B]^b$ and $k_{C\to A+B}[C]^c$. We obtain the following network, where we indicate association of kinetic complexes by dots:

$$\begin{array}{ccc} A+B & \rightleftharpoons & C \\ \vdots & & \vdots \\ aA+bB & & cC \end{array}$$

For an arbitrary chemical reaction network with generalized mass action kinetics, the rate function $\mathscr{K}_{y\to y'}\colon \mathbb{R}^{\mathscr{S}}_{\geq} \to \mathbb{R}_{\geq}$ corresponding to reaction $y \to y' \in \mathscr{R}$ is given by $\mathscr{K}_{y\to y'}(c) = k_{y\to y'}\, c^{\tilde{y}}$, where $\tilde{y}$ is the kinetic complex associated with y.

DEFINITION 2.12. *A generalized chemical reaction network $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R})$ is a chemical reaction network $(\mathscr{S},\mathscr{C},\mathscr{R})$ together with a family $\tilde{\mathscr{C}} = (x_y)_{y\in\mathscr{C}}$ in $\mathbb{R}^{\mathscr{S}}_{\geq}$ of kinetic complexes, where $|\{x_y \,|\, y \in \mathscr{C}\}| = |\mathscr{C}|$. We write $\tilde{y} = x_y$ for the kinetic complex associated with the complex $y \in \mathscr{C}$.*

A generalized chemical reaction network $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R})$ contains the chemical reaction network $(\mathscr{S},\mathscr{C},\mathscr{R})$; moreover, it entails the fictitious chemical reaction network $(\mathscr{S},\underline{\tilde{\mathscr{C}}},\underline{\mathscr{R}})$, where the set $\underline{\tilde{\mathscr{C}}} = \{\tilde{y} \,|\, y \in \mathscr{C}\}$ has the same cardinality as $\mathscr{C}$ (by definition) and the relation $\underline{\mathscr{R}}$ is isomorphic to $\mathscr{R}$, i.e., $\tilde{y} \to \tilde{y}' \in \underline{\mathscr{R}}$ whenever $y \to y' \in \mathscr{R}$. Hence the networks $(\mathscr{S},\mathscr{C},\mathscr{R})$ and $(\mathscr{S},\underline{\tilde{\mathscr{C}}},\underline{\mathscr{R}})$ give rise to the same directed graph (up to the renaming of vertices). A generalized chemical reaction network $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R})$ is called weakly reversible if $(\mathscr{S},\mathscr{C},\mathscr{R})$ is weakly reversible. Also the definitions of the stoichiometric subspace and the stoichiometric compatibility classes carry over from $(\mathscr{S},\mathscr{C},\mathscr{R})$ to $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R})$; cf. Definition 2.2. Additionally, we introduce the kinetic-order subspace of a generalized chemical reaction network, which coincides with the stoichiometric subspace of the fictitious network.

DEFINITION 2.13. *Let $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R})$ be a generalized chemical reaction network. The kinetic-order subspace is defined as*

$$\tilde{S} = \operatorname{span}\{\tilde{y}' - \tilde{y} \,|\, y \to y' \in \mathscr{R}\}\,.$$

For consistency, the name *kinetic subspace* would be more appropriate for $\tilde{S}$ but this name has already been given to a certain subspace of the stoichiometric subspace [18], which coincides with the stoichiometric subspace if $t = l$.

For later use, we introduce the maps

- $\tilde{Y}\colon \mathbb{R}^{\mathscr{C}} \to \mathbb{R}^{\mathscr{S}}$ with $\tilde{Y}\omega_y = \tilde{y}$ and
- $\tilde{\Psi}\colon \mathbb{R}^{\mathscr{S}}_{\geq} \to \mathbb{R}^{\mathscr{C}}$, $c \mapsto \sum_{y\in\mathscr{C}} c^{\tilde{y}}\,\omega_y$,

where we identify $\mathbb{R}^{\mathscr{C}}$ and $\mathbb{R}^{\underline{\tilde{\mathscr{C}}}}$.

DEFINITION 2.14. *A generalized mass action system $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R},k)$ is a generalized chemical reaction network $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R})$ together with a vector $k \in \mathbb{R}^{\mathscr{R}}_{>}$ of rate constants.*

DEFINITION 2.15. *The ODE associated with a generalized mass action system $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R},k)$ is defined as*

$$\frac{dc}{dt} = \tilde{r}(c)$$

with the species formation rate

$$\tilde{r}(c) = \sum_{y\to y'\in\mathscr{R}} k_{y\to y'}\, c^{\tilde{y}}\,(y' - y)\,.$$

As in (2.1), we can decompose the species formation rate of a generalized mass action system as

$$\tilde{r}(c) = Y A \, \tilde{\Psi}(c) \, .$$

Analogous to Definition 2.5, equilibria satisfying $A \, \tilde{\Psi}(c) = 0$ and $c > 0$ are called complex balancing equilibria; they coincide with the complex balancing equilibria of the fictitious mass action system $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}}, k)$. The deficiency, which quantifies the possibility of other equilibria, coincides with the deficiency of the mass action system $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$.

DEFINITION 2.16. *Let $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ be a generalized mass action system. The set of complex balancing equilibria is defined as*

$$\tilde{Z} = \{c \in \mathbb{R}^{\mathscr{S}}_{>} \mid A \, \tilde{\Psi}(c) = 0\}$$

and the deficiency as

$$\delta = \dim(\ker(Y) \cap \operatorname{im}(A)) \, .$$

It remains to introduce the kinetic deficiency, which coincides with the deficiency of the fictitious system.

DEFINITION 2.17. *Let $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ be a generalized mass action system. The kinetic deficiency is defined as*

$$\tilde{\delta} = \dim(\ker(\tilde{Y}) \cap \operatorname{im}(A)) \, .$$

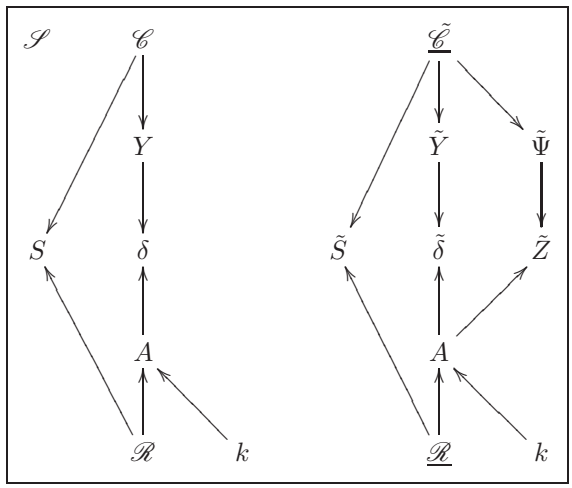
In Figure 2.2, we summarize the definitions associated with a generalized mass action system and depict their dependencies. From the mass action system $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$, we keep the stoichiometric subspace S and the deficiency δ, whereas we use all definitions associated with the fictitious mass action system $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}}, k)$; in particular, the kinetic-order subspace $\tilde{S}$, the kinetic deficiency $\tilde{\delta}$, and the set $\tilde{Z}$ of complex balancing equilibria.

Results. Now we return to the results of CRNT that have been derived for mass action systems. Since Theorem 2.6 is concerned with the kernel of the linear map A, the underlying kinetics is not relevant at all. But also Corollary 2.7 and Propositions 2.8–2.10 carry over easily to generalized mass action systems if we consider the fictitious chemical reaction network $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}})$ and the fictitious mass action system $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}}, k)$ defined above. For reference, we present the analogous results.

PROPOSITION 2.18. *Let $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ be a chemical reaction network. If there exists a generalized mass action system $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ with a complex balancing equilibrium, then $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ is weakly reversible.*

Proof. Assume that $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ and hence the mass action system $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}}, k)$ have a complex balancing equilibrium. By Corollary 2.7, the chemical reaction network $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}})$ and hence $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ are weakly reversible. □

PROPOSITION 2.19. *If a chemical reaction network $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ is weakly reversible (or more generally if $t = l$), then the deficiencies of any generalized mass action system $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ are given by $\delta = m - l - s$ and $\tilde{\delta} = m - l - \tilde{s}$, where m is the number of complexes, l is the number of linkage classes, s is the dimension of the stoichiometric subspace, and $\tilde{s}$ is the dimension of the kinetic-order subspace.*



$$\frac{dc}{dt} = Y A\, \tilde{\Psi}(c)$$

FIG. 2.2. *The generalized mass action system* $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$*: Associated definitions and their dependencies. (For better readability,* k *and* A *are plotted twice.)*

Proof. Assume that $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ and hence the chemical reaction network $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}})$ arising from $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ are weakly reversible (or more generally that $t = l$). The deficiency of the generalized mass action system equals the deficiency of $(\mathscr{S}, \mathscr{C}, \mathscr{R}, k)$, and the kinetic deficiency equals the deficiency of $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}}, k)$. By Proposition 2.8, the deficiencies of the two mass action systems are given by the formulas stated. ∎

PROPOSITION 2.20. *If a generalized chemical reaction network* $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R})$ *is weakly reversible and* $\tilde{\delta} = 0$*, then any generalized mass action system* $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ *has a complex balancing equilibrium.*

Proof. Assume that $(\mathscr{S}, \mathscr{C}, \mathscr{R})$ and hence the chemical reaction network $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}})$ arising from $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ are weakly reversible. Additionally, assume $\tilde{\delta} = 0$. By Proposition 2.9, the mass action system $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}}, k)$ and hence $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ have a complex balancing equilibrium. ∎

PROPOSITION 2.21. *Let* $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ *be a generalized mass action system with nonempty set* $\tilde{Z}$ *of complex balancing equilibria. Then*

$$\tilde{Z} = \{c \in \mathbb{R}^{\mathscr{S}}_{>} \mid \ln(c) - \ln(c^*) \in \tilde{S}^{\perp}\} = \{c^* \circ \mathrm{e}^{\tilde{v}} \mid \tilde{v} \in \tilde{S}^{\perp}\}$$

for any $c^* \in \tilde{Z}$*.*

Proof. The set of complex balancing equilibria of the mass action system $(\mathscr{S}, \underline{\tilde{\mathscr{C}}}, \underline{\mathscr{R}}, k)$ coincides with $\tilde{Z}$, and its stoichiometric subspace coincides with $\tilde{S}$, which is the kinetic-order subspace of $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$. By Proposition 2.10, the nonempty set $\tilde{Z}$ is given by the formula stated. ∎

One might conjecture that Theorem 2.11 also holds for generalized mass action systems. However, an analogous result depends on both the complexes $\mathscr{C}$ and the kinetic complexes $\tilde{\mathscr{C}}$, where $\mathscr{C}$ determines the stoichiometric subspace S (and hence the stoichiometric compatibility classes $(c' + S)_{\geq}$), whereas $\tilde{\mathscr{C}}$ determines the set $\tilde{Z}$

of complex balancing equilibria (and the related kinetic-order subspace $\tilde{S}$). It turns out that the result depends on additional assumptions concerning the sign vectors of the subspaces S and $\tilde{S}$; see Theorem 3.10.

3. Complex balancing equilibria. In the following, we consider a generalized mass action system $(\mathscr{S}, \mathscr{C}, \mathscr{C}, \mathscr{R}, k)$ with stoichiometric subspace S, kinetic-order subspace $\tilde{S}$, and nonempty set $\tilde{Z}$ of complex balancing equilibria.

From Proposition 2.21 we know that $\tilde{Z} = \{c^* \circ \mathrm{e}^{\tilde{v}} \mid \tilde{v} \in \tilde{S}^\perp\}$ for any $c^* \in \tilde{Z}$. We provide necessary and sufficient conditions such that in every stoichiometric compatibility class $(c' + S)_\geq$ there is at most one complex balancing equilibrium. Moreover, we provide sufficient conditions such that in every stoichiometric compatibility class there is at least one complex balancing equilibrium.

The question of uniqueness is answered by the following result for arbitrary subspaces S and $\tilde{S}$. It involves the corresponding sets of sign vectors denoted by $\sigma(S)$ and $\sigma(\tilde{S})$; for the definition of sign vectors and related notions we refer the reader to the appendix. We note that sign vectors also appear in the study of multiple equilibria that are not necessarily complex balancing [17, 32].

PROPOSITION 3.1. *Let $S, \tilde{S}$ be subspaces of $\mathbb{R}^n$. Then the following two statements are equivalent:*

1. *For all $c^* > 0$ and $c' > 0$, the intersection $(c' + S)_\geq \cap \{c^* \circ \mathrm{e}^{\tilde{v}} \mid \tilde{v} \in \tilde{S}^\perp\}$ contains at most one element.*
2. *$\sigma(S) \cap \sigma(\tilde{S}^\perp) = \{0\}$.*

Proof. ($\neg 1 \Rightarrow \neg 2$): Suppose there exist $u^1 \neq u^2 \in S$ and $\tilde{v}_1 \neq \tilde{v}_2 \in \tilde{S}^\perp$ such that $c' + u^1 = c^* \circ \mathrm{e}^{\tilde{v}_1}$ and $c' + u^2 = c^* \circ \mathrm{e}^{\tilde{v}_2}$ (for a certain c' and a certain c^*). Then $u^1 - u^2 = c^* \circ (\mathrm{e}^{\tilde{v}_1} - \mathrm{e}^{\tilde{v}_2})$ and by the monotonicity of the exponential function

$$\sigma(\underbrace{u^1 - u^2}_{\in S}) = \sigma(c^* \circ (\mathrm{e}^{\tilde{v}^1} - \mathrm{e}^{\tilde{v}^2})) = \sigma(\mathrm{e}^{\tilde{v}^1} - \mathrm{e}^{\tilde{v}^2}) = \sigma(\underbrace{\tilde{v}^1 - \tilde{v}^2}_{\in \tilde{S}^\perp}).$$

Hence $\sigma(S) \cap \sigma(\tilde{S}^\perp) \neq \{0\}$.

($\neg 2 \Rightarrow \neg 1$): Suppose that $0 \neq \tau \in \sigma(S) \cap \sigma(\tilde{S}^\perp)$. Then there exist $u \in S$ and $\tilde{v}^1 \in \tilde{S}^\perp$ such that $\sigma(u) = \sigma(\tilde{v}^1) = \tau$. Further, let $\tilde{v}^2 = \frac{1}{2}\tilde{v}^1$. Then $\sigma(\tilde{v}^1 - \tilde{v}^2) = \tau$ and

$$\sigma(u) = \sigma(\tilde{v}^1 - \tilde{v}^2) = \sigma(\mathrm{e}^{\tilde{v}^1} - \mathrm{e}^{\tilde{v}^2}) = \sigma(c^* \circ (\mathrm{e}^{\tilde{v}^1} - \mathrm{e}^{\tilde{v}^2}))$$

for all $c^* > 0$. In particular, there is c^* such that $u = c^* \circ (\mathrm{e}^{\tilde{v}^1} - \mathrm{e}^{\tilde{v}^2})$. With $c' = c^* \circ \mathrm{e}^{\tilde{v}^1}$, one has $c' - u = c^* \circ \mathrm{e}^{\tilde{v}^2}$ and hence both c' and $c' - u$ are elements of $(c' + S)_\geq \cap \{c^* \circ \mathrm{e}^{\tilde{v}} \mid \tilde{v} \in \tilde{S}^\perp\}$. ∎

It follows, in particular, that if the sign vectors are equal, $\sigma(S) = \sigma(\tilde{S})$, complex balancing equilibria are unique (in a stoichiometric compatibility class) since then

$$\sigma(S) \cap \sigma(\tilde{S}^\perp) = \sigma(S) \cap \sigma(\tilde{S})^\perp = \sigma(S) \cap \sigma(S)^\perp = \{0\}$$

using (A.1). Note that this is only a sufficient condition; for example, with $S = \mathrm{span}\{(-1, 1)\}$ and $\tilde{S} = \mathrm{span}\{(-1, 0)\}$, we have $\sigma(S) \cap \sigma(\tilde{S}^\perp) = \{0\}$ but $\sigma(S) \neq \sigma(\tilde{S})$. However, it includes classical mass action kinetics where $S = \tilde{S}$ and each stoichiometric compatibility class contains at most one complex balancing equilibrium. On the other hand, if $\sigma(S) \cap \sigma(\tilde{S}^\perp) \neq \{0\}$ and the underlying network is weakly reversible, then such a generalized chemical reaction network has the capacity for multiple complex balancing equilibria, as shown in the following result.

PROPOSITION 3.2. *If a generalized chemical reaction network $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R})$ is weakly reversible and $\sigma(S)\cap\sigma(\tilde{S}^\perp)\neq\{0\}$, there exist rate constants k such that the generalized mass action system $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R},k)$ has more than one complex balancing equilibrium in some stoichiometric compatibility class.*

Proof. Let $\sigma(S)\cap\sigma(\tilde{S}^\perp)\neq\{0\}$. By Proposition 3.1, there exist $c^*>0$ and $c'>0$ such that $(c'+S)_\geq\cap\{c^*\circ \mathrm{e}^{\tilde{v}}\mid \tilde{v}\in\tilde{S}^\perp\}$ contains more than one element. Using Proposition 2.21, it remains to show that there exist rate constants $k\in\mathbb{R}^{\mathscr{R}}_>$ such that c^* is a complex balancing equilibrium of $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R},k)$, i.e.,

$$A\,\tilde{\Psi}(c^*)=\sum_{y\to y'\in\mathscr{R}} k_{y\to y'}\,(c^*)^{\tilde{y}}\,(\omega_{y'}-\omega_y)=0\,.$$

Since $(\mathscr{S},\mathscr{C},\tilde{\mathscr{C}},\mathscr{R})$ and hence $(\mathscr{S},\mathscr{C},\mathscr{R})$ are weakly reversible, this is guaranteed by Lemma 3.3. ∎

In the proof of Proposition 3.2, we use the following result.

LEMMA 3.3. *Let $(\mathscr{S},\mathscr{C},\mathscr{R})$ be a chemical reaction network. Then, the following statements are equivalent:*

1. *$(\mathscr{S},\mathscr{C},\mathscr{R})$ is weakly reversible.*
2. *There exists $k\in\mathbb{R}^{\mathscr{R}}_>$ such that $\sum_{y\to y'\in\mathscr{R}} k_{y\to y'}\,(\omega_{y'}-\omega_y)=0$, where $\omega_y\in\mathbb{R}^{\mathscr{C}}$ denotes the unit vector corresponding to $y\in\mathscr{C}$.*

Proof. (1 ⇒ 2): By weak reversibility, there exists a cycle $y\to y'\to\cdots\to y$ for each reaction $y\to y'\in\mathscr{R}$ and we denote the set of reactions involved in this cycle by $C_{y\to y'}$. Clearly, $\sum_{z\to z'\in C_{y\to y'}}(\omega_{z'}-\omega_z)=0$ and hence

$$\sum_{y\to y'\in\mathscr{R}}\;\sum_{z\to z'\in C_{y\to y'}}(\omega_{z'}-\omega_z)=\sum_{y\to y'\in\mathscr{R}} k_{y\to y'}\,(\omega_{y'}-\omega_y)=0\,,$$

where $k_{y\to y'}>0$ records in how many cycles the reaction $y\to y'$ appears.

(2 ⇒ 1): We write $\sum_{y\to y'\in\mathscr{R}} k_{y\to y'}\,(\omega_{y'}-\omega_y)=A\,\Omega$ with $\Omega=(1,1,\dots,1)^T\in\mathbb{R}^{\mathscr{C}}_>$. By Theorem 2.6, if $A\,\Omega=0$, then $(\mathscr{S},\mathscr{C},\mathscr{R})$ is weakly reversible. ∎

The second implication is a basic fact from CRNT [24, 16].

3.1. The map F. In order to study uniqueness and existence in a common framework, we rephrase the problem. We suppose that $\mathscr{S}$ contains n species and fix an order among them. Then we can identify $\mathbb{R}^{\mathscr{S}}$ with $\mathbb{R}^n$ such that $S,\,\tilde{S}\subseteq\mathbb{R}^n$. Further, let $V=(v^1,\dots,v^d)$ and $\tilde{V}=(\tilde{v}^1,\dots,\tilde{v}^{\tilde{d}})$ be bases for $S^\perp$ and $\tilde{S}^\perp$, respectively. In other words, $S^\perp=\mathrm{im}(V)$ and $\dim(S^\perp)=d$ and analogously $\tilde{S}^\perp=\mathrm{im}(\tilde{V})$ and $\dim(\tilde{S}^\perp)=\tilde{d}$.

An element in $(c'+S)_\geq\cap\{c^*\circ\mathrm{e}^{\tilde{v}}\mid\tilde{v}\in\tilde{S}^\perp\}$ corresponds to $u\in S$ and $\tilde{v}\in\tilde{S}^\perp$ such that $c^*\circ\mathrm{e}^{\tilde{v}}=c'+u$ or equivalently to $\lambda\in\mathbb{R}^{\tilde{d}}$ such that

$$\langle c^*\circ\mathrm{e}^{\sum_{j=1}^{\tilde{d}}\lambda_j\tilde{v}^j},v^i\rangle=\langle c',v^i\rangle\quad\text{for}\quad i=1,\dots,d\,.$$

Hence, provided $c^*\in\tilde{Z}$, uniqueness and existence (of a complex balancing equilibrium in every stoichiometric compatibility class) correspond to injectivity and surjectivity of the following map:

$$\begin{aligned}(3.1)\qquad &F\colon\mathbb{R}^{\tilde{d}}\to C^\circ\subseteq\mathbb{R}^d\\ &\lambda\mapsto F(\lambda)\quad\text{with}\quad (F(\lambda))_i=\langle c^*\circ\mathrm{e}^{\sum_{j=1}^{\tilde{d}}\lambda_j\tilde{v}^j},v^i\rangle\,,\end{aligned}$$

where $c^* > 0$ and

$$C^\circ = \{\gamma \in \mathbb{R}^d \,|\, \gamma_i = \langle c', v^i \rangle,\, c' \in \mathbb{R}^n_> \}\,.$$

Note that F depends on c^*. It is instructive to reformulate the definition of F. To this end, we express the columns of V and $\tilde{V}$ by its rows,

$$V = (v^1, \dots, v^d) = (w^1, \dots, w^n)^T,$$
$$\tilde{V} = (\tilde{v}^1, \dots, \tilde{v}^{\tilde{d}}) = (\tilde{w}^1, \dots, \tilde{w}^n)^T\,,$$

or equivalently $v^j_i = w^i_j$ and $\tilde{v}^j_i = \tilde{w}^i_j$, and obtain

$$\begin{aligned}(F(\lambda))_i &= \langle c^* \circ \mathrm{e}^{\sum_{j=1}^{\tilde{d}} \lambda_j \tilde{v}^j}, v^i \rangle = \sum_{k=1}^n c^*_k\, \mathrm{e}^{\sum_{j=1}^{\tilde{d}} \lambda_j \tilde{v}^j_k}\, v^i_k \\ &= \sum_{k=1}^n c^*_k\, \mathrm{e}^{\sum_{j=1}^{\tilde{d}} \lambda_j \tilde{w}^k_j}\, w^k_i = \sum_{k=1}^n c^*_k\, \mathrm{e}^{\langle \lambda, \tilde{w}^k \rangle}\, w^k_i\end{aligned}$$

and

$$\gamma_i = \langle c', v^i \rangle = \sum_{k=1}^n c'_k\, v^i_k = \sum_{k=1}^n c'_k\, w^k_i\,.$$

Hence we can write $F(\lambda) = \sum_{k=1}^n c^*_k\, \mathrm{e}^{\langle \lambda, \tilde{w}^k \rangle}\, w^k$ and $\gamma = \sum_{k=1}^n c'_k\, w^k$.

Definition 3.4. *Let $V \in \mathbb{R}^{n\times d}$, $\tilde{V} \in \mathbb{R}^{n\times \tilde{d}}$ with $n \geq d, \tilde{d}$ have full rank. We write $V = (v^1, \dots, v^d) = (w^1, \dots, w^n)^T$ and $\tilde{V} = (\tilde{v}^1, \dots, \tilde{v}^{\tilde{d}}) = (\tilde{w}^1, \dots, \tilde{w}^n)^T$. Further, let $c^* > 0$. We define*

$$\begin{aligned} F\colon \mathbb{R}^{\tilde{d}} &\to C^\circ \subseteq \mathbb{R}^d \\ \lambda &\mapsto \sum_{k=1}^n c^*_k\, \mathrm{e}^{\langle \lambda, \tilde{w}^k \rangle}\, w^k\,,\end{aligned}$$

where

$$C^\circ = \left\{ \sum_{k=1}^n c'_k\, w^k \in \mathbb{R}^d \,|\, c' \in \mathbb{R}^n_> \right\}\,.$$

This definition is more transparent than the equivalent one given above. It becomes clear that the set C° is the interior of the polyhedral cone generated by the vectors $(w^1, \dots, w^n)$. The map F itself (in case $V = \tilde{V}$) appears in toric geometry [20], where it is related to moment maps, and in statistics [31], where it is related to exponential families. There is a useful result [20], which guarantees injectivity and surjectivity of F in case $V = \tilde{V}$.

Proposition 3.5. *Let V, $\tilde{V}$, and F be as in Definition* 3.4. *If $V = \tilde{V}$, then F is a real analytic isomorphism of $\mathbb{R}^d$ onto C° for all $c^* > 0$.*

This is a variant of Birch's theorem [31, 41, 9]; it implies Theorem 2.11. We will build on this result when we study the surjectivity of F, but first we deal with its injectivity in case $V \neq \tilde{V}$.

3.2. Injectivity of F. In the context of multiple equilibria in mass action systems [10] and geometric modeling [11], it was shown that the map F (in case $d = \tilde{d}$) is injective for all c^* if and only if F is a local isomorphism for all c^*. We give an alternative proof of this result and extend it to the case $d \neq \tilde{d}$, where we use the sign vectors of the spaces $\mathrm{im}(V)$ and $\mathrm{im}(\tilde{V})$.

THEOREM 3.6. *Let V, $\tilde{V}$, and F be as in Definition* 3.4. *Then, the following statements are equivalent:*

1. *F is injective for all $c^* > 0$.*
2. *F is an immersion for all $c^* > 0$. ($\frac{\partial F}{\partial \lambda}$ is injective for all λ and $c^* > 0$.)*
3. *$\sigma(\mathrm{im}(V)^\perp) \cap \sigma(\mathrm{im}(\tilde{V})) = \{0\}$.*

Proof. We use F in the form of (3.1).

($1 \Leftrightarrow 3$): By Proposition 3.1.
Using $S^\perp = \mathrm{im}(V)$ and $\tilde{S}^\perp = \mathrm{im}(\tilde{V})$, the injectivity of F for all c^* is equivalent to the existence of at most one element in $(c' + S)_{\geq} \cap \{c^* \circ \mathrm{e}^{\tilde{v}} \mid \tilde{v} \in \tilde{S}^\perp\}$ for all c' and c^*.

($\neg 2 \Rightarrow \neg 3$): Suppose that $\frac{\partial F}{\partial \lambda}$ is not injective (for a certain c^* and a certain λ), i.e., there exists a nonzero $\lambda' \in \mathbb{R}^{\tilde{d}}$ such that $\frac{\partial F}{\partial \lambda}\lambda' = 0$. Since

$$\sum_{j=1}^{\tilde{d}} \frac{\partial F_i}{\partial \lambda_j}\,\lambda'_j = \sum_{j=1}^{\tilde{d}} \langle c^* \circ \mathrm{e}^{\sum_{k=1}^{\tilde{d}} \lambda_k \tilde{v}^k} \circ \tilde{v}^j, v^i \rangle \lambda'_j = \left\langle \underbrace{c^* \circ \mathrm{e}^{\sum_{k=1}^{\tilde{d}} \lambda_k \tilde{v}^k}}_{c} \circ \underbrace{\sum_{j=1}^{\tilde{d}} \lambda'_j\, \tilde{v}^j}_{\tilde{v}'}, v^i \right\rangle ,$$

this is equivalent to the existence of $c > 0$ and $\tilde{v}' \in \mathrm{im}(\tilde{V})$ such that $\langle c \circ \tilde{v}', v^i \rangle = 0$ for $i = 1, \ldots, d$, which in turn is equivalent to $c \circ \tilde{v}' \in \mathrm{im}(V)^\perp$. Clearly $\sigma(c \circ \tilde{v}') = \sigma(\tilde{v}')$ and hence $\sigma(\mathrm{im}(V)^\perp) \cap \sigma(\mathrm{im}(\tilde{V})) \neq \{0\}$.

($\neg 3 \Rightarrow \neg 2$): Suppose that $0 \neq \tau \in \sigma(\mathrm{im}(V)^\perp) \cap \sigma(\mathrm{im}(\tilde{V}))$. Then, there exist $u \in \mathrm{im}(V)^\perp$ and $\tilde{v}' \in \mathrm{im}(\tilde{V})$ such that $\sigma(u) = \sigma(\tilde{v}') = \tau$. Clearly, one can choose $c > 0$ such that $u = c \circ \tilde{v}'$ and hence $c \circ \tilde{v}' \in \mathrm{im}(V)^\perp$. As demonstrated in the previous step, this is equivalent to the existence of $c^* > 0$ and $\lambda, \lambda' \neq 0 \in \mathbb{R}^{\tilde{d}}$ such that $\frac{\partial F}{\partial \lambda}\lambda' = 0$. ∎

Finally, we note that for $d = \tilde{d}$, statement 3 in Theorem 3.6 is symmetric with respect to V and $\tilde{V}$.

COROLLARY 3.7. *Let V, $\tilde{V}$ be as in Definition* 3.4 *with $d = \tilde{d}$. Then, $\sigma(\mathrm{im}(V)^\perp) \cap \sigma(\mathrm{im}(\tilde{V})) = \{0\}$ if and only if $\sigma(\mathrm{im}(\tilde{V})^\perp) \cap \sigma(\mathrm{im}(V)) = \{0\}$.*

Proof. Let F be in the form of (3.1) with $d = \tilde{d}$, and let $\tilde{F}$ be obtained from F by changing the roles of V and $\tilde{V}$,

$$\begin{aligned} \tilde{F} &\colon \mathbb{R}^d \to \mathbb{R}^d \\ &\lambda \mapsto \tilde{F}(\lambda) \quad \text{with} \quad (\tilde{F}(\lambda))_i = \langle c^* \circ \mathrm{e}^{\sum_{j=1}^{d} \lambda_j v^j}, \tilde{v}^i \rangle . \end{aligned}$$

We will show that $\frac{\partial F}{\partial \lambda}$ is injective for all c^* if and only if $\frac{\partial \tilde{F}}{\partial \lambda}$ is injective for all c^*. Then, by Theorem 3.6 we will obtain the desired result.

Suppose that $\frac{\partial \tilde{F}}{\partial \lambda}$ (or equivalently its transpose) is not injective (for a certain c^* and a certain λ), i.e., there exists $\lambda' \in \mathbb{R}^{\tilde{d}}$ such that $(\frac{\partial \tilde{F}}{\partial \lambda})^T \lambda' = 0$. Since

$$\sum_{j=1}^{d} \frac{\partial \tilde{F}_j}{\partial \lambda_i}\,\lambda'_j = \sum_{j=1}^{d} \langle c^* \circ \mathrm{e}^{\sum_{k=1}^{d} \lambda_k v^k} \circ v^i, \tilde{v}^j \rangle \lambda'_j = \left\langle \underbrace{c^* \circ \mathrm{e}^{\sum_{k=1}^{d} \lambda_k v^k}}_{c} \circ v^i, \underbrace{\sum_{j=1}^{d} \lambda'_j\, \tilde{v}^j}_{\tilde{v}'} \right\rangle$$

and $\langle c \circ v^i, \tilde{v}' \rangle = \langle c \circ \tilde{v}', v^i \rangle$, this is equivalent to the noninjectivity condition for $\frac{\partial F}{\partial \lambda}$ derived in the proof of Theorem 3.6. ∎

3.3. Surjectivity of F. It is more difficult to derive conditions for the surjectivity of F. Our main result is concerned with sufficient conditions; however, we start with a discussion of necessary conditions.

Let C and $\tilde{C}$ be the polyhedral cones generated by the vector configurations $V^T = (w^1, \dots, w^n)$ and $\tilde{V}^T = (\tilde{w}^1, \dots, \tilde{w}^n)$, respectively. Then $C^\circ = \mathrm{int}(C)$, and analogously we write $\tilde{C}^\circ = \mathrm{int}(\tilde{C})$. We note that C° and $\tilde{C}^\circ$ are nonempty since V and $\tilde{V}$ have full rank.

We write $\sigma(\mathrm{im}(V))_{\geq} = \sigma(\mathrm{im}(V)) \cap \{0, +\}^n$ for the face lattice of C (see the appendix), and analogously $\sigma(\mathrm{im}(\tilde{V}))_{\geq} = \sigma(\mathrm{im}(\tilde{V})) \cap \{0, +\}^n$ for the face lattice of $\tilde{C}$. A face f of C is characterized by a sign vector $\tau \in \sigma(\mathrm{im}(V))_{\geq}$ or equivalently by a supporting hyperplane with normal vector $\lambda \in \mathbb{R}^d$, where $\tau_k = 0$ whenever $\langle \lambda, w^k \rangle = 0$ (for w^k lying on f) and $\tau_k = +$ whenever $\langle \lambda, w^k \rangle > 0$.

Now, we can study a necessary condition for surjectivity: The image of F must contain points arbitrarily close to any point on a face of C. We assume that C is pointed, more specifically that $(+, \dots, +)^T \in \sigma(\mathrm{im}(V))$, and we consider the simplest nontrivial face, namely an extreme ray e. To begin with, we assume that e contains only one generator, say w^1; hence the characteristic sign vector amounts to $\tau = (0, +, \dots, +)^T$. If F is surjective, then the cone $\tilde{C}$ must have a corresponding extreme ray $\tilde{e}$ with the same sign vector τ. Only then is there $\mu \in \mathbb{R}^{\tilde{d}}$ with $\langle \mu, \tilde{w}^1 \rangle = 0$ and $\langle \mu, \tilde{w}^k \rangle > 0$ for $k = 2, \dots, n$ such that the limit

$$\lim_{a \to \infty} F(-a\mu + \nu) = \lim_{a \to \infty} \sum_{k=1}^{n} c_k^* \, \mathrm{e}^{-a\langle \mu, \tilde{w}^k \rangle + \langle \nu, \tilde{w}^k \rangle} \, w^k = c_1^* \, \mathrm{e}^{\langle \nu, \tilde{w}^1 \rangle} \, w^1$$

can be placed arbitrarily close to any point on e (by appropriate choice of $\nu \in \mathbb{R}^{\tilde{d}}$).

If the extreme ray e contains more than one generator, there may be several corresponding extreme rays $\tilde{e}$. For a particular $\tilde{e}$ with characteristic sign vector $\tilde{\tau}$, there is $\mu \in \mathbb{R}^{\tilde{d}}$ (with $\langle \mu, \tilde{w}^k \rangle = 0$ if $\tilde{\tau}_k = 0$ and $\langle \mu, \tilde{w}^k \rangle > 0$ if $\tilde{\tau}_k = +$) such that $\lim_{a \to \infty} F(-a\mu + \nu)$ lies on e. We note that if $\tilde{e}$ contains $\tilde{w}^k$, then e must contain w^k; otherwise the limit does not lie on e. This condition on the corresponding extreme rays $\tilde{e}$ and e can be expressed by their characteristic sign vectors $\tilde{\tau}$ and τ, namely as $\tilde{\tau} \geq \tau$. For higher-dimensional faces, similar (but more complicated) conditions can be formulated.

For the proof of the following surjectivity result, we will employ degree theory. In particular, we use two properties of the Brouwer degree $\mathrm{d}(f, D, y)$ of a continuous function $f \colon \bar{D} \to \mathbb{R}^d$ defined on the closure of an open and bounded subset $D \subset \mathbb{R}^d$ (with boundary ∂D) at a value $y \notin f(\partial D)$: (i) the degree is invariant under homotopy, and (ii) if the degree is nonzero, there exists x such that $y = f(x)$; see [29] or [19].

THEOREM 3.8. *Let V, $\tilde{V}$, and F be as in Definition 3.4. If there exists a lattice isomorphism $\Phi \colon \sigma(\mathrm{im}(\tilde{V}))_{\geq} \to \sigma(\mathrm{im}(V))_{\geq}$ with $\tilde{\tau} \geq \Phi(\tilde{\tau})$ and $(+, \dots, +)^T \in \sigma(\mathrm{im}(V))$, then F is surjective for all $c^* > 0$.*

Proof. In order to use the Brouwer degree, we require a map on a closed and bounded set. To this end, we define a map G equivalent to F from the interior of $\tilde{C}$ to the interior of C and extend G to the boundaries such that it maps faces to faces. Then, we cut the pointed cones such that we obtain polytopes $\tilde{P}$ and P. Finally, we define a homotopy between the map G and a homeomorphism between the polytopes guaranteed by the face lattice isomorphism. As a consequence, every point in the

interior of P has nonzero Brouwer degree and hence is in the image of G. Since the cut of the cone C can be placed at arbitrary distance from the origin, this holds for every point in the interior of C.

Since $(+,\dots,+)^T \in \sigma(\mathrm{im}(V))$, the face lattice isomorphism implies $(+,\dots,+)^T \in \sigma(\mathrm{im}(\tilde{V}))$, and hence the cones C and $\tilde{C}$ are pointed. We start by choosing a minimal set of generators for $\tilde{C}$, which (after reordering) we assume to be $(\tilde{w}^1,\dots,\tilde{w}^{n_E})$, where n_E is the number of extreme rays of $\tilde{C}$. We define an auxiliary map,

$$\begin{aligned}\tilde{F}\colon \mathbb{R}^{\tilde{d}} &\to \tilde{C}^\circ\\ \lambda &\mapsto \sum_{k=1}^{n_E} \tilde{c}_k^*\, \mathrm{e}^{\langle \lambda, \tilde{w}^k\rangle}\, \tilde{w}^k,\end{aligned}$$

which is a real analytic isomorphism by Proposition 3.5, and a composed map,

$$\begin{aligned}G^\circ\colon \tilde{C}^\circ &\to C^\circ\\ x &\mapsto F(\tilde{F}^{-1}(x))\,,\end{aligned}$$

which is surjective whenever F is surjective.

Since G° is defined only on $\tilde{C}^\circ$, we want to extend it continuously to the boundary $\partial\tilde{C}$, i.e., to the faces of the cone. Let $\tilde{f}$ be a face of $\tilde{C}$. It contains a subset of the minimal set of generators for $\tilde{C}$, which (after reordering) we assume to be $(\tilde{w}^1,\dots,\tilde{w}^{n_{\min}})$. There may be additional generators on $\tilde{f}$, which we assume to be $(\tilde{w}^{n_E+1},\dots,\tilde{w}^{n_E+n_{\mathrm{add}}})$, where $n_{\min}+n_{\mathrm{add}}$ is the total number of generators on $\tilde{f}$.

Now, let $(x^i)_{i\in\mathbb{N}}$ be a sequence with $x^i \in \tilde{C}^\circ$ and $\lim_{i\to\infty} x^i \in \tilde{f}$. Via the isomorphism $\tilde{F}$, there is a corresponding sequence $(\lambda^i)_{i\in\mathbb{N}}$ with $\lambda^i \in \mathbb{R}^{\tilde{d}}$. From

$$\lim_{i\to\infty} x^i = \sum_{k=1}^{n_{\min}} \tilde{c}_k^* \lim_{i\to\infty} \mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle}\, \tilde{w}^k + \sum_{k=n_{\min}+1}^{n_E} \tilde{c}_k^* \lim_{i\to\infty} \mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle}\, \tilde{w}^k,$$

we conclude that $\lim_{i\to\infty} \mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle} \geq 0$ for $k = 1,\dots,n_{\min}$ and $\lim_{i\to\infty} \mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle} = 0$ for $k = n_{\min}+1,\dots,n_E$. Additional generators $\tilde{w}^k$ on $\tilde{f}$ can be written as nonnegative linear combinations of the minimal generators $(\tilde{w}^1,\dots,\tilde{w}^{n_{\min}})$ and hence[3] we obtain $\lim_{i\to\infty} \mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle} \geq 0$. Generators $\tilde{w}^k$ not on $\tilde{f}$ can be written as nonnegative linear combinations containing at least one of the remaining minimal generators $(\tilde{w}^{n_{\min}+1},\dots,\tilde{w}^{n_E})$ and hence[4] we obtain $\lim_{i\to\infty} \mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle} = 0$. As a consequence, the image of the sequence converges and

$$\lim_{i\to\infty} G^\circ(x^i) = \sum_{k=1}^{n_{\min}} c_k^* \lim_{i\to\infty} \mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle}\, w^k + \sum_{k=n_E+1}^{n_E+n_{\mathrm{add}}} c_k^* \lim_{i\to\infty} \mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle}\, w^k\,.$$

The isomorphism Φ (between the face lattices of $\tilde{C}$ and C) with $\tilde{\tau} \geq \Phi(\tilde{\tau})$ implies that there is a face f of C with $w^k \in f$ if $\tilde{w}^k \in \tilde{f}$. That is, $w^1,\dots,w^{n_{\min}} \in f$ as well as $w^{n_E+1},\dots,w^{n_E+n_{\mathrm{add}}} \in f$ and hence $\lim_{i\to\infty} G^\circ(x^i) \in f$.

In other words, there is a continuous extension of G° to the face $\tilde{f}$, which maps $\tilde{f}$ to the corresponding face f. We set $G := G^\circ$ on $\tilde{C}^\circ$ and $G(x) := \lim_{i\to\infty} G^\circ(x^i)$ for

[3] By using $\mathrm{e}^{\langle \lambda^i, \sum_{k=1}^{n_{\min}} a_k \tilde{w}^k\rangle} = \prod_{k=1}^{n_{\min}} (\mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle})^{a_k}$.

[4] By using $\mathrm{e}^{\langle \lambda^i, \sum_{k=1}^{n_E} a_k \tilde{w}^k\rangle} = \prod_{k=1}^{n_{\min}} (\mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle})^{a_k} \prod_{k=n_{\min}+1}^{n_E} (\mathrm{e}^{\langle \lambda^i, \tilde{w}^k\rangle})^{a_k}$.

any sequence $(x^i)_{i\in\mathbb{N}}$ with $x^i \in \tilde{C}^\circ$ and $\lim_{i\to\infty} x^i = x \in \tilde{f}$. Since this can be done for all faces of $\tilde{C}$, there is a map $G\colon \tilde{C} \to C$ which extends G° continuously to $\partial\tilde{C}$ and maps faces to faces.

Due to the face lattice isomorphism, a minimal set of generators for C is given by $(w^1, \dots, w^{n_E})$. The isomorphism further implies $d = \tilde{d}$.

Since C is a pointed cone, we can choose a $(d-1)$-dimensional subspace of $\mathbb{R}^d$ such that C lies on one side of the subspace. We cut C with a hyperplane parallel to the subspace and obtain a polytope P (lying on one side of the hyperplane). In particular, we intersect the extreme rays of C with the hyperplane: the intersection of the extreme ray e^k (generated by w^k) is located at $\alpha_k w^k$ with $\alpha_k > 0$.

Analogously, we cut $\tilde{C}$ with a hyperplane and obtain a polytope $\tilde{P}$. The intersection of the extreme ray $\tilde{e}^k$ (generated by $\tilde{w}^k$) is located at $\tilde{\alpha}_k\tilde{w}^k$ with $\tilde{\alpha}_k > 0$.

From now on, we restrict the map G to $\tilde{P}$ and choose $\tilde{c}^*$ such that G maps corners of $\tilde{P}$ to corresponding corners of P. For example, the corner $\tilde{\alpha}_1\tilde{w}^1$ on $\tilde{e}^1$ corresponds (by $\tilde{F}^{-1}$) to the sequence $(\lambda^i)_{i\in\mathbb{N}}$ with $\lambda^i \in \mathbb{R}^n$, $\lim_{i\to\infty} \tilde{c}_1^*\, \mathrm{e}^{\langle\lambda^i,\tilde{w}^1\rangle} = \tilde{\alpha}_1$, and $\lim_{i\to\infty} \mathrm{e}^{\langle\lambda^i,\tilde{w}^k\rangle} = 0$ for $k = 2, \dots, n_E$. In turn, $(\lambda^i)_{i\in\mathbb{N}}$ corresponds (by F) to the corner $\alpha_1 w^1$ on e^1:

$$\lim_{i\to\infty}\left(c_1^*\, \mathrm{e}^{\langle\lambda^i,\tilde{w}^1\rangle}\; w^1 + \sum_{k=n_E+1}^{n_E+n_{\mathrm{add}}} c_k^*\, \mathrm{e}^{\langle\lambda^i,\tilde{w}^k\rangle}\; w^k\right) = \alpha_1 w^1\,.$$

Here, we have assumed that, in addition to $\tilde{w}^1$, there are additional generators $\tilde{w}^k$ (with $k = n_E+1, \dots, n_E+n_{\mathrm{add}}$) on $\tilde{e}^1$ with corresponding generators w^k on e^1. If we write $\tilde{w}^k = \tilde{\beta}_k\tilde{w}^1$, $w^k = \beta_k w^1$, and $x = \lim_{i\to\infty} \mathrm{e}^{\langle\lambda^i,\tilde{w}^1\rangle}$, we can determine $\tilde{c}_1^*$ from

$$\tilde{c}_1^*\, x = \tilde{\alpha}_1 \quad \text{with} \quad c_1^*\, x + \sum_{k=n_E+1}^{n_E+n_{\mathrm{add}}} c_k^*\, x^{\tilde{\beta}_k}\beta_k = \alpha_1\,.$$

If we choose $\tilde{c}_k^*$ accordingly for each extreme ray $\tilde{e}_k$, then G maps "side-edges" of $\tilde{P}$ to corresponding side-edges of P. The image of other faces of $\tilde{P}$ need not coincide with the corresponding faces of P. (However, due to the face lattice isomorphism, the image of a "side-face" of $\tilde{P}$ lying on a face of $\tilde{C}$, lies in the corresponding face of C.) In particular,[5] the image of the "cut-face" of $\tilde{P}$ (arising from the cut with the hyperplane) may lie outside the cut-face of P.

The isomorphism between the face lattices of $\tilde{C}$ and C has another important consequence. It guarantees the existence of a piecewise linear homeomorphism $G'\colon \tilde{P} \to P$, which restricts to homeomorphisms between corresponding faces of $\tilde{P}$ and P; see the appendix. We note that G' has nonzero Brouwer degree on $P^\circ = \mathrm{int}(P)$ and define a homotopy between G (restricted to $\tilde{P}$) and G',

$$\begin{aligned} H\colon \tilde{P}\times[0,1] &\to C \subset \mathbb{R}^d \\ (x,t) &\mapsto t\,G(x) + (1-t)\,G'(x)\,. \end{aligned}$$

(The homotopy H maps to C, since both G and G' map to C and C is convex.)

Now, let $y \in P^\circ$. Below we will show that $y \notin H(\partial\tilde{P}, t)$ for all $t \in [0,1]$. Writing $\tilde{P}^\circ = \mathrm{int}(\tilde{P})$, we conclude that $\mathrm{d}(G, \tilde{P}^\circ, y) = \mathrm{d}(G', \tilde{P}^\circ, y) \neq 0$ (by the homotopy

[5] A point on the cut-face of $\tilde{P}$ is a convex combination of the "corners" $\tilde{\alpha}_k\tilde{w}^k$, $k = 1, \dots, n_E$. By $\tilde{F}^{-1}$ it corresponds to some $\lambda \in \mathbb{R}^n$, which by F corresponds to a point on the cut-face of P, that is, a convex combination (with the same coefficients) of the corners $\alpha_k w^k$, $k = 1, \dots, n_E$, *plus* a positive linear combination of the additional generators w^k, $k = n_E+1, \dots, n$.

invariance of the Brouwer degree) and that there exists $x \in \tilde{P}^\circ$ with $G(x) = y$ (by the existence property of the Brouwer degree). In other words, the image of G restricted to $\tilde{P}^\circ$ contains P°. Since the cut of the cone C can be placed at an arbitrary distance from the origin, $G^\circ\colon \tilde{C}^\circ \to C^\circ$ and hence $F\colon \mathbb{R}^d \to C^\circ$ are surjective.

It remains to show that $y \notin H(\partial\tilde{P}, t)$ for all $t \in [0,1]$: For side-faces $\tilde{f} \subset \partial\tilde{P}$, one has $H(\tilde{f}, t) \subset \partial C$ for all $t \in [0,1]$ (since G and G' map side-faces to side-faces), whereas for the cut-face $\tilde{f} \subset \partial\tilde{P}$, one either has $H(\tilde{f}, t) \subset \partial P$ for all $t \in [0,1]$ (whenever G maps one cut-face to the other) or $H(\mathrm{int}(\tilde{f}), t) \cap P = \emptyset$ for all $t \in [0,1]$ (whenever G maps the cut-face out of P). In each case, one obtains $H(\partial\tilde{P}, t) \cap P^\circ = \emptyset$ for all $t \in [0,1]$. ∎

We think that the technical condition $(+,\dots,+)^T \in \sigma(\mathrm{im}(V))$ in Theorem 3.8, which requires the cone C to be pointed, is not necessary, and a similar result can be obtained for arbitrary cones. However, at the moment we do not have a complete proof for such a theorem.

3.4. Main results. The previous two theorems concerned with injectivity and surjectivity of F allow the following generalization of Proposition 3.5 (Birch's theorem).

PROPOSITION 3.9. *Let V, $\tilde{V}$, and F be as in Definition* 3.4. *If $\sigma(\mathrm{im}(V)) = \sigma(\mathrm{im}(\tilde{V}))$ and $(+,\dots,+)^T \in \sigma(\mathrm{im}(V))$, then F is a real analytic isomorphism of $\mathbb{R}^d$ onto C° for all $c^* > 0$.*

Proof. From $\sigma(\mathrm{im}(V)) = \sigma(\mathrm{im}(\tilde{V}))$ it follows that $d = \tilde{d}$ and with (A.1) that $\sigma(\mathrm{im}(V)^\perp) \cap \sigma(\mathrm{im}(\tilde{V})) = \{0\}$. Hence, F is injective and a local isomorphism by Theorem 3.6. Moreover, with Φ being the identity, F is surjective by Theorem 3.8. ∎

Note that the condition $\sigma(\mathrm{im}(V)) = \sigma(\mathrm{im}(\tilde{V}))$ in the previous proposition can be tested algorithmically using chirotopes; see the appendix. We can now formulate a result analogous to Theorem 2.11 in the case of generalized mass action kinetics.

THEOREM 3.10. *Let $(\mathscr{S}, \mathscr{C}, \tilde{\mathscr{C}}, \mathscr{R}, k)$ be a generalized mass action system with nonempty set $\tilde{Z}$ of complex balancing equilibria, stoichiometric subspace S, and kinetic-order subspace $\tilde{S}$. If $\sigma(S) = \sigma(\tilde{S})$ and $(+,\dots,+)^T \in \sigma(S^\perp)$, then $\tilde{Z}$ meets every stoichiometric compatibility class in exactly one point.*

Proof. Suppose $\tilde{Z} \neq \emptyset$. As discussed at the beginning of subsection 3.1, uniqueness and existence of a complex balancing equilibrium in every stoichiometric compatibility class correspond to injectivity and surjectivity of the map F as given in Definition 3.4, where V and $\tilde{V}$ are bases for $S^\perp$ and $\tilde{S}^\perp$, respectively. By (A.1), $\sigma(S) = \sigma(\tilde{S})$ is equivalent to $\sigma(\mathrm{im}(V)) = \sigma(\mathrm{im}(\tilde{V}))$, and obviously $(+,\dots,+)^T \in \sigma(S^\perp)$ is equivalent to $(+,\dots,+)^T \in \sigma(\mathrm{im}(V))$ such that F is injective and surjective by Proposition 3.9. ∎

In the terminology of CRNT, a chemical reaction network is *conservative* if $S^\perp \cap \mathbb{R}^{\mathscr{S}}_> \neq \emptyset$, i.e., if there is a "vector of molecular weights," relative to which all reactions are mass conserving. Note that the condition $(+,\dots,+)^T \in \sigma(S^\perp)$ in Theorem 3.10 means that the underlying chemical reaction network is conservative.

4. Examples. We discuss two examples of generalized mass action systems. First, we continue the example of the generalized chemical reaction network introduced in section 2.2,

$$\begin{array}{ccc} A+B & \rightleftharpoons & C \\ \vdots & & \vdots \\ aA+bB & & cC \end{array} \tag{4.1}$$

with $a, b, c \in \mathbb{R}_>$. The kinetic complexes $aA + bB$ and cC (associated with the complexes $A + B$ and C) determine the exponents in the rate functions $k_{A+B\to C}[A]^a[B]^b$ and $k_{C\to A+B}[C]^c$.

The network is (weakly) reversible and has two complexes and one linkage class. The stoichiometric and kinetic-order subspace amount to $S = \operatorname{span}\{(-1,-1,1)^T\}$ and $\tilde{S} = \operatorname{span}\{(-a,-b,c)^T\}$ with dimensions $d = \tilde{d} = 1$. By Proposition 2.19, $\delta = \tilde{\delta} = 2 - 1 - 1 = 0$, and by Proposition 2.20, $\tilde{Z} \neq \emptyset$. Further, the sign vectors of S and $\tilde{S}$ coincide, i.e., $\sigma(S) = \sigma(\tilde{S})$, and $(1,1,2)^T \in S^\perp$, which implies $(+,+,+)^T \in \sigma(S^\perp)$. Hence, by Theorem 3.10, every stoichiometric compatibility class contains exactly one complex balancing equilibrium.

In the rest of this section, we study an autocatalytic mechanism (for the overall reaction $A + B \rightleftharpoons C$) endowed with generalized mass action kinetics:

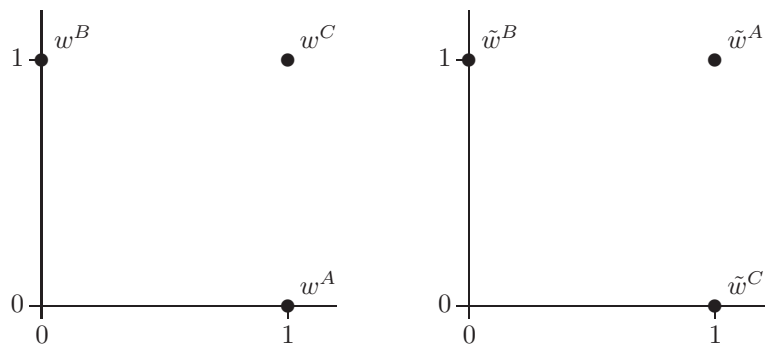
$$\begin{array}{ccc} A+2B & \rightleftharpoons & B+C \\ \vdots & & \vdots \\ A+B & & 2B+C \end{array} \tag{4.2}$$

The kinetic complexes $A+B$ and $2B+C$ (associated with the complexes $A+2B$ and $B+C$) determine the rate functions $k_{A+2B\to B+C}[A][B]$ and $k_{B+C\to A+2B}[B]^2[C]$. The particular kinetics may be unrealistic from a chemical point of view, however, it will serve to demonstrate how the conditions in Theorem 3.10 for existence and uniqueness of a complex balancing equilibrium (in every stoichiometric compatibility class) are violated.

The network is weakly reversible, $\delta = \tilde{\delta} = 0$, and hence $\tilde{Z} \neq \emptyset$. In particular, the stoichiometric and kinetic-order subspace amount to $S = \operatorname{span}\{(-1,-1,1)^T\}$ and $\tilde{S} = \operatorname{span}\{(-1,1,1)^T\}$. For the orthogonal complements $S^\perp$ and $\tilde{S}^\perp$ we choose the bases

$$V = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \tilde{V} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The cones C and $\tilde{C}$ generated by $V^T = (w^A, w^B, w^C)$ and $\tilde{V}^T = (\tilde{w}^A, \tilde{w}^B, \tilde{w}^C)$ both coincide with $\mathbb{R}^2_\geq$:



First, we address the question of existence. We observe that the cone C has an extreme ray generated by w^A, whereas the cone $\tilde{C}$ does not have a corresponding extreme ray generated by $\tilde{w}^A$ (since $\tilde{w}^A$ lies in the interior of $\tilde{C}$). As a consequence, the map F is not surjective for all c^*; cf. the argument at the beginning of subsection 3.3.

In other words, there may be a stoichiometric compatibility class that does not contain a complex balancing equilibrium.

Now, we turn to the question of uniqueness. In order to employ Propositions 3.1 or 3.2, we determine $\sigma(S) \cap \sigma(\tilde{S}^\perp)$. The sign vectors of S are $(-,-,+)^T$, its inverse, and 0, whereas the sign vectors of $\tilde{S}^\perp$ can be read off from the above figure: For every hyperplane of $\mathbb{R}^2$, i.e., for every line through $0 \in \mathbb{R}^2$, we check if $\tilde{w}^A$, $\tilde{w}^B$, and $\tilde{w}^C$ lie on the line or on its negative or positive side. We obtain

$$\sigma(S) = \begin{pmatrix} - & & 0 \\ - & \dots & 0 \\ + & & 0 \end{pmatrix} \quad \text{and} \quad \sigma(\tilde{S}^\perp) = \begin{pmatrix} + & + & + & 0 & - & - & & 0 \\ + & 0 & - & - & - & - & \dots & 0 \\ + & + & + & + & + & 0 & & 0 \end{pmatrix},$$

where we use matrix notation for sets of vectors and where we do not state vectors explicitly that are inverses of others. We find that $\sigma(S) \cap \sigma(\tilde{S}^\perp)$ contains $(-,-,+)^T$. Hence, by Proposition 3.2, there exist rate constants $k_{A+2B\to B+C}$ and $k_{B+C\to A+2B}$ such that some stoichiometric compatibility class contains more than one complex balancing equilibrium.

Due to the simplicity of the generalized mass action system, the equilibria of the associated ODE can be determined analytically. The equilibrium condition amounts to

$$k_{A+2B\to B+C}[A][B] = k_{B+C\to A+2B}[B]^2[C]\,,$$

and since $\delta = 0$ all equilibria are complex balancing. By using the conservation relations $[A] + [C] = [A]_0 + [C]_0 = \Sigma_{AC}$ and $[B] + [C] = [B]_0 + [C]_0 = \Sigma_{BC}$ and by writing $K = k_{A+2B\to B+C}/k_{B+C\to A+2B}$, we obtain a quadratic equation in $[C]$, which can be solved as

$$[C] = \frac{K + \Sigma_{BC}}{2} \pm \sqrt{\left(\frac{K + \Sigma_{BC}}{2}\right)^2 - K\,\Sigma_{AC}}\,.$$

Depending on the equilibrium constant K and the initial values Σ_{AC} and Σ_{BC} (which determine a stoichiometric compatibility class), the quadratic equation has 0, 1, or 2 solutions with $[C] > 0$. If, additionally, $[A] = \Sigma_{AC} - [C] > 0$ and $[B] = \Sigma_{BC} - [C] > 0$, then $([A], [B], [C])^T$ is a complex balancing equilibrium. Obviously, a stoichiometric compatibility class contains 0, 1, or 2 complex balancing equilibria; it turns out that each case is realized.

5. Conclusion. CRNT establishes intriguing results about the ODEs associated with mass action systems, in particular about the existence, uniqueness, and stability of equilibria. For application in molecular biology, however, one would like to have a framework that permits rate laws more general than mass action kinetics.

In this paper we show that the suggested notion of generalized mass action systems, which admits arbitrary nonnegative power-law rate functions, allows us to generalize several results of CRNT. In particular, Theorem 3.10 essentially states that if the sign vectors of the stoichiometric and the kinetic-order subspace coincide, there exists a unique complex balancing equilibrium in every stoichiometric compatibility class.

A natural next step is to study other results of CRNT in the case of generalized mass action kinetics, most importantly, to analyze the stability of complex balancing equilibria, which is guaranteed in the classical case. Further, genuinely biological notions such as the *robustness* [6, 39, 40] of chemical reaction networks can be addressed in a framework with more realistic kinetics.

Appendix. Sign vectors and face lattices. In this section, we outline some facts on the relation between sign vectors of vector spaces and face lattices of polyhedral cones and polytopes. For further details we refer to [1, Chap. 7] and [42, Chaps. 2 and 6] and to [33, 7] in the context of oriented matroids.

We obtain the *sign vector* $\sigma(x) \in \{-,0,+\}^n$ of a vector $x \in \mathbb{R}^n$ by applying the sign function componentwise, and we write

$$\sigma(S) = \{\sigma(x) \,|\, x \in S\}$$

for a subset $S \subseteq \mathbb{R}^n$.

Two sign vectors $\varsigma, \tau \in \{-,0,+\}^n$ are *orthogonal* if $\varsigma_k \tau_k = 0$ for all k or if there exist k, l with $\varsigma_k \tau_k = -$ and $\varsigma_l \tau_l = +$ (where the product on $\{-,0,+\}$ is defined in the obvious way); we write $\varsigma \perp \tau$. Note that $\varsigma \perp \tau$ if and only if there are orthogonal vectors $x, y \in \mathbb{R}^n$ such that $\sigma(x) = \varsigma$ and $\sigma(y) = \tau$.

The *orthogonal complement* $\Sigma^\perp$ of a set $\Sigma \subseteq \{-,0,+\}^n$ is defined by

$$\Sigma^\perp = \{\varsigma \in \{-,0,+\}^n \,|\, \varsigma \perp \tau \ \forall \tau \in \Sigma\}\,.$$

The sign vectors of the orthogonal complement of a subspace $S \subseteq \mathbb{R}^n$ are given by

$$\sigma(S^\perp) = \sigma(S)^\perp; \tag{A.1}$$

see, for example, [42, Prop. 6.8].

Let $V = (v^1, \dots, v^d) \in \mathbb{R}^{n \times d}$ with $n \geq d$ have full rank. Then $V^T = (w^1, \dots, w^n)$ is called a *vector configuration* (of n vectors in $\mathbb{R}^d$). With $\lambda \in \mathbb{R}^d$ and $v = \sum_{j=1}^d \lambda_j v^j \in \operatorname{im}(V)$, we obtain $v_k = \sum_{j=1}^d \lambda_j v_k^j = \sum_{j=1}^d \lambda_j w_j^k = \langle \lambda, w^k \rangle$. Hence, $\sigma(v)$ describes the positions of the vectors $w^1, \dots, w^n$ relative to the hyperplane with normal vector λ.

The *face lattice* of the cone C generated by $w^1, \dots, w^n$ can be recovered from the sign vectors of the subspace generated by $v^1, \dots, v^d$. It is the set $\sigma(\operatorname{im}(V)) \cap \{0,+\}^n$ with the partial order induced by the relation $0 < +$, which we denote by

$$\sigma(\operatorname{im}(V))_\geq = \sigma(\operatorname{im}(V)) \cap \{0,+\}^n.$$

A face f of C is characterized by a supporting hyperplane with normal vector $\lambda \in \mathbb{R}^d$ such that $\langle \lambda, w^k \rangle = 0$ for generators w^k lying on f and $\langle \lambda, w^k \rangle > 0$ for the remaining w^k (thus lying on the positive side of the hyperplane).

A cone C is called *pointed* if $C \cap (-C) = \{0\}$ or equivalently if it has vertex 0. A cone is pointed if and only if it has an extreme ray, and every pointed polyhedral cone is the conical hull of its finitely many extreme rays. Note that if $(+, \dots, +)^T \in \sigma(\operatorname{im}(V))$, the cone C generated by V^T is pointed.

As for polyhedral cones, the faces of a polytope form a lattice. Two polytopes are *combinatorially equivalent* if their face lattices are isomorphic. Combinatorial equivalence corresponds to the existence of a piecewise linear homeomorphism between the polytopes that restricts to homeomorphisms between faces.

The sign vectors $\sigma(\operatorname{im}(V))$ of the subspace $\operatorname{im}(V)$ can be equivalently characterized by the *chirotope* χ_{V^T} of the point configuration V^T, which is defined as the map

$$\begin{aligned} \chi_{V^T} \colon \{1, \dots, n\}^d &\to \{-,0,+\} \\ (i_1, \dots, i_d) &\mapsto \operatorname{sign}(\det(w^{i_1}, \dots, w^{i_d}))\,. \end{aligned}$$

The chirotope records for each d-tuple of vectors if it forms a positively (or negatively) oriented basis of $\mathbb{R}^d$ or it is not a basis. It can, for example, be used to test algorithmically if the sign vectors of two subspaces are equal, that is, to decide if $\sigma(\operatorname{im}(V)) = \sigma(\operatorname{im}(\tilde{V}))$ for two matrices $V, \tilde{V} \in \mathbb{R}^{n \times d}$.

Acknowledgments. We thank Josef Hofbauer for pointing our attention to degree theory and Günter M. Ziegler for answering our questions on oriented matroids. We also acknowledge fruitful discussions with François Boulier and François Lemaire on an earlier version of the paper and numerous helpful comments from an anonymous referee.

REFERENCES

[1] A. Bachem and W. Kern, *Linear Programming Duality*, Springer-Verlag, Berlin, 1992.
[2] Z. Bajzer, M. Huzak, K. L. Neff, and F. G. Prendergast, *Mathematical analysis of models for reaction kinetics in intracellular environments*, Math. Biosci., 215 (2008), pp. 35–47.
[3] M. Banaji and G. Craciun, *Graph-theoretic approaches to injectivity and multiple equilibria in systems of interacting elements*, Commun. Math. Sci., 7 (2009), pp. 867–900.
[4] M. Banaji and G. Craciun, *Graph-theoretic criteria for injectivity and unique equilibria in general chemical reaction systems*, Adv. in Appl. Math., 44 (2010), pp. 168–184.
[5] M. Banaji, P. Donnell, and S. Baigent, *P matrix properties, injectivity, and stability in chemical reaction systems*, SIAM J. Appl. Math., 67 (2007), pp. 1523–1547.
[6] E. Batchelor and M. Goulian, *Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system*, Proc. Natl. Acad. Sci. USA, 100 (2003), pp. 691–696.
[7] A. Björner, M. Las Vergnas, B. Sturmfels, N. White, and G. M. Ziegler, *Oriented Matroids*, 2nd ed., Encyclopedia Math. Appl. 46, Cambridge University Press, Cambridge, UK, 1999.
[8] J. S. Clegg, *Properties and metabolism of the aqueous cytoplasm and its boundaries*, Am. J. Physiol., 246 (1984), pp. R133–R151.
[9] G. Craciun, A. Dickenstein, A. Shiu, and B. Sturmfels, *Toric dynamical systems*, J. Symbolic Comput., 44 (2009), pp. 1551–1565.
[10] G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks.* I. *The injectivity property*, SIAM J. Appl. Math., 65 (2005), pp. 1526–1546.
[11] G. Craciun, L. Garcia-Puente, and F. Sottile, *Some geometrical aspects of control points for toric patches*, in Mathematical Methods for Curves and Surfaces, M. Daehlen, M. S. Floater, T. Lyche, J.-L. Merrien, K. Morken, and L. L. Schumaker, eds., Lecture Notes in Comput. Sci. 5862, Springer, Heidelberg, 2010, pp. 111–135.
[12] G. Craciun, J. W. Helton, and R. J. Williams, *Homotopy methods for counting reaction network equilibria*, Math. Biosci., 216 (2008), pp. 140–149.
[13] M. Feinberg, *Complex balancing in general kinetic systems*, Arch. Rational Mech. Anal., 49 (1972/73), pp. 187–194.
[14] M. Feinberg, *Lectures on Chemical Reaction Networks*, from notes of lectures given at the Mathematics Research Center of the University of Wisconsin in 1979, available online at http://www.che.eng.ohio-state.edu/FEINBERG/LecturesOnReactionNetworks (1979).
[15] M. Feinberg, *Chemical reaction network structure and the stability of complex isothermal reactors–*I. *The deficiency zero and deficiency one theorems*, Chem. Eng. Sci., 42 (1987), pp. 2229–2268.
[16] M. Feinberg, *The existence and uniqueness of steady states for a class of chemical reaction networks*, Arch. Rational Mech. Anal., 132 (1995), pp. 311–370.
[17] M. Feinberg, *Multiple steady states for chemical reaction networks of deficiency one*, Arch. Rational Mech. Anal., 132 (1995), pp. 371–406.
[18] M. Feinberg and F. J. M. Horn, *Chemical mechanism structure and the coincidence of the stoichiometric and kinetic subspaces*, Arch. Rational Mech. Anal., 66 (1977), pp. 83–97.
[19] I. Fonseca and W. Gangbo, *Degree Theory in Analysis and Applications*, Oxford Lecture Ser. Math. Appl. 2, The Clarendon Press, Oxford University Press, New York, 1995.
[20] W. Fulton, *Introduction to Toric Varieties*, Ann. of Math. Stud. 131, Princeton University Press, Princeton, NJ, 1993.
[21] R. Grima and S. Schnell, *A systematic investigation of the rate laws valid in intracellular environments*, Biophys. Chem., 124 (2006), pp. 1–10.
[22] J. Gunawardena, *Chemical Reaction Network Theory for in-silico Biologists*, available online at http://vcp.med.harvard.edu/papers/crnt.pdf (2003).
[23] P. J. Halling, *Do the laws of chemistry apply to living cells?*, Trends Biochem. Sci., 14 (1989), pp. 317–318.

[24] F. Horn, *Necessary and sufficient conditions for complex balancing in chemical kinetics*, Arch. Rational Mech. Anal., 49 (1972/73), pp. 172–186.
[25] F. Horn and R. Jackson, *General mass action kinetics*, Arch. Rational Mech. Anal., 47 (1972), pp. 81–116.
[26] R. Kopelman, *Rate processes on fractals: Theory, simulations, and experiments*, J. Stat. Phys., 42 (1986), pp. 185–200.
[27] R. Kopelman, *Fractal reaction kinetics*, Science, 241 (1988), pp. 1620–1626.
[28] H. Kuthan, *Self-organisation and orderly processes by individual protein complexes in the bacterial cell*, Prog. Biophys. Mol. Biol., 75 (2001), pp. 1–17.
[29] N. G. Lloyd, *Degree Theory*, Cambridge University Press, Cambridge, UK, 1978.
[30] K. L. Neff, C. P. Offord, A. J. Caride, E. E. Strehler, F. G. Prendergast, and Z. Bajzer, *Validation of fractal-like kinetic models by time-resolved binding kinetics of dansylamide and carbonic anhydrase in crowded media*, Biophys. J., 100 (2011), pp. 2495–2503.
[31] L. Pachter and B. Sturmfels, *Statistics*, in Algebraic Statistics for Computational Biology, Cambridge University Press, New York, 2005, pp. 3–42.
[32] M. Pérez Millán, A. Dickenstein, A. Shiu, and C. Conradi, *Chemical reaction systems with toric steady states*, Bull. Math. Biol., 74 (2012), pp. 1027–1065.
[33] J. Richter-Gebert and G. M. Ziegler, *Oriented Matroids*, in Handbook of Discrete and Computational Geometry, CRC, Boca Raton, FL, 1997, pp. 111–132.
[34] M. A. Savageau, *Biochemical systems analysis.* I. *Some mathematical properties of the rate law for the component enzymatic reactions*, J. Theoret. Biol., 25 (1969), pp. 365–369.
[35] M. A. Savageau, *Biochemical Systems Analysis: Study of Function and Design in Molecular Biology*, Addison-Wesley, Reading, MA, 1976.
[36] M. A. Savageau, *A critique of the enzymologist's test tube*, in Fundamentals of Medical Cell Biology, Vol. 3A, E. E. Bittar, ed., JAI Press Inc., Greenwich, CT, 1992.
[37] M. A. Savageau, *Michaelis-Menten mechanism reconsidered: Implications of fractal kinetics*, J. Theoret. Biol., 176 (1995), pp. 115–124.
[38] S. Schnell and T. E. Turner, *Reaction kinetics in intracellular environments with macromolecular crowding: Simulations and rate laws*, Prog. Biophys. Mol. Biol., 85 (2004), pp. 235–260.
[39] G. Shinar and M. Feinberg, *Structural sources of robustness in biochemical reaction networks*, Science, 327 (2010), pp. 1389–1391.
[40] R. Steuer, S. Waldherr, V. Sourjik, and M. Kollmann, *Robust signal processing in living cells*, PLoS Comput. Biol., 7 (2011), e1002218.
[41] B. Sturmfels, *Solving Systems of Polynomial Equations*, CBMS Reg. Conf. Ser. Math. 97, Conference Board of the Mathematical Sciences, Washington, DC, AMS, Providence, RI, 2002.
[42] G. M. Ziegler, *Lectures on Polytopes*, Springer-Verlag, New York, 1995.

Generalized Mass-Action Systems and Positive Solutions of Polynomial Equations with Real and Symbolic Exponents (*Invited Talk*)

Stefan Müller and Georg Regensburger

Johann Radon Institute for Computational and Applied Mathematics (RICAM),
Austrian Academy of Sciences, Linz, Austria
{stefan.mueller,georg.regensburger}@ricam.oeaw.ac.at

Abstract. Dynamical systems arising from chemical reaction networks with mass action kinetics are the subject of chemical reaction network theory (CRNT). In particular, this theory provides statements about uniqueness, existence, and stability of positive steady states for all rate constants and initial conditions. In terms of the corresponding polynomial equations, the results guarantee uniqueness and existence of positive solutions for all positive parameters.

We address a recent extension of CRNT, called generalized mass-action systems, where reaction rates are allowed to be power-laws in the concentrations. In particular, the (real) kinetic orders can differ from the (integer) stoichiometric coefficients. As with mass-action kinetics, complex balancing equilibria are determined by the graph Laplacian of the underlying network and can be characterized by binomial equations and parametrized by monomials. In algebraic terms, we focus on a constructive characterization of positive solutions of polynomial equations with real and symbolic exponents.

Uniqueness and existence for all rate constants and initial conditions additionally depend on sign vectors of the stoichiometric and kinetic-order subspaces. This leads to a generalization of Birch's theorem, which is robust with respect to certain perturbations in the exponents. In this context, we discuss the occurrence of multiple complex balancing equilibria.

We illustrate our results by a running example and provide a MAPLE worksheet with implementations of all algorithmic methods.

Keywords: Chemical reaction network theory, generalized mass-action systems, generalized polynomial equations, symbolic exponents, positive solutions, binomial equations, Birch's theorem, oriented matroids, multistationarity.

1 Introduction

In this work, we focus on dynamical systems arising from (bio-)chemical reaction networks with *generalized* mass-action kinetics and positive solutions of the corresponding systems of generalized polynomial equations.

V.P. Gerdt et al. (Eds.): CASC Workshop 2014, LNCS 8660, pp. 302–323, 2014.

In chemical reaction network theory, as initiated by Horn, Jackson, and Feinberg in the 1970s [15,33,32], several fundamental results are based on the assumption of mass action kinetics (MAK). Consider the reaction

$$1\,\mathrm{A} + 1\,\mathrm{B} \to \mathrm{C} \tag{1}$$

involving the reactant species A, B and the product C, where we explicitly state the stoichiometric coefficients of the reactants. The left- and right-hand sides of a reaction, in this case A+B and C, are called (stoichiometric) complexes. Let

$$[A] = [A](t)$$

denote the concentration of species A at time t, and analogously for B and C. Assuming MAK, the rate at which the reaction occurs is given by

$$v = k\,[\mathrm{A}]^1[\mathrm{B}]^1$$

with rate constant $k > 0$. In other words, the reaction rate is a monomial in the reactant concentrations [A] and [B] with the stoichiometric coefficients as exponents. Within a network involving additional species and reactions, the above reaction contributes to the dynamics of the species concentrations as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix}[\mathrm{A}]\\ [\mathrm{B}]\\ [\mathrm{C}]\\ [\mathrm{D}]\\ \vdots\end{pmatrix} = k\,[\mathrm{A}][\mathrm{B}]\begin{pmatrix}-1\\ -1\\ 1\\ 0\\ \vdots\end{pmatrix} + \cdots$$

In many applications, the reaction network is given, but the values of the rate constants are unknown. Surprisingly, there are results on existence, uniqueness, and stability of steady states that do not depend on the rate constants. See, for example, the lecture notes [16] and the surveys [17,19,30].

However, the validity of MAK is limited; it only holds for elementary reactions in homogeneous and dilute solutions. For biochemical reaction networks in intracellular environments, the rate law has to be modified. In previous work [40], we allowed generalized mass-action kinetics (GMAK) where reaction rates are power-laws in the concentrations. In particular, the exponents need not coincide with the stoichiometric coefficients and need not be integers. For example, the rate at which reaction (1) occurs may be given by

$$v = k\,[\mathrm{A}]^a[\mathrm{B}]^b$$

with kinetic orders $a, b > 0$. Formally, we specify the rate of a reaction by associating (here indicated by dots) with the reactant complex a kinetic complex, which determines the exponents in the generalized monomial:

$$\begin{array}{c}\mathrm{A} + \mathrm{B} \;\to \mathrm{C}\\ \vdots\\ a\mathrm{A} + b\mathrm{B}\end{array}$$

Before we give the definition of generalized mass action systems, we introduce a running example, which will be used to motivate and illustrate general statements. Throughout the paper, we focus on algorithmic aspects of the theoretical results. Additionally, we provide a MAPLE worksheet[1] with implementations of all algorithms applied to the running example. For other applications of computer algebra to chemical reaction networks, we refer to [7,14,36,45].

Notation. We denote the strictly positive real numbers by $\mathbb{R}_>$. We define $\mathrm{e}^x \in \mathbb{R}^n_>$ for $x \in \mathbb{R}^n$ component-wise, that is, $(\mathrm{e}^x)_i = \mathrm{e}^{x_i}$; analogously, $\ln(x) \in \mathbb{R}^n$ for $x \in \mathbb{R}^n_>$ and $x^{-1} \in \mathbb{R}^n$ for $x \in \mathbb{R}^n$ with $x_i \neq 0$. For $x, y \in \mathbb{R}^n$, we denote the component-wise (or Hadamard) product by $x \circ y \in \mathbb{R}^n$, that is, $(x \circ y)_i = x_i y_i$; for $x \in \mathbb{R}^n_>$ and $y \in \mathbb{R}^n$, we define $x^y \in \mathbb{R}_>$ as $\prod_{i=1}^n x_i^{y_i}$.

Given a matrix $B \in \mathbb{R}^{n \times m}$, we denote by $b^1, \ldots, b^m$ its column vectors and by $b_1, \ldots, b_n$ its row vectors. For $x \in \mathbb{R}^n_>$, we define $x^B \in \mathbb{R}^m_>$ as

$$(x^B)_j = x^{b^j} = \prod_{i=1}^n x_i^{b_{ij}}$$

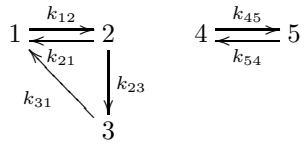
for $j = 1, \ldots, m$. As a consequence,

$$\ln(x^B) = B^T \ln x.$$

Finally, we identify a matrix $B \in \mathbb{R}^{n \times m}$ with the corresponding linear map $B \colon \mathbb{R}^m \to \mathbb{R}^n$ and write $\mathrm{im}(B)$ and $\ker(B)$ for the respective vector subspaces.

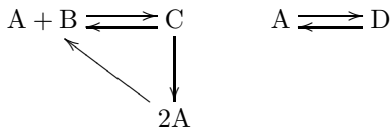
2 Running Example

We consider a reaction network based on the weighted directed graph



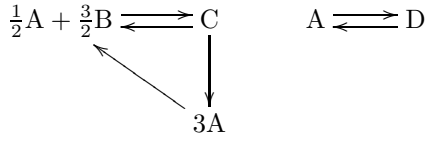
(2)

with 5 vertices, 6 edges and corresponding positive weights. Clearly, the edges represent reactions and the weights are rate constants. We assume that the network contains 4 species A, B, C, D and associate with each vertex a (stoichiometric) complex, that is, a formal sum of species:



[1] The worksheet is available at `http://gregensburger.com/software/GMAK.zip`.

In order to specify the reaction rates, e.g., $v_{12} = k_{12}[\mathrm{A}]^{\frac{1}{2}}[\mathrm{B}]^{\frac{3}{2}}$, we additionally associate a kinetic complex with each source vertex:



Writing

$$x = (x_1, x_2, x_3, x_4)^T$$

for the concentrations of species A, B, C, D, the dynamics of the generalized mass action system is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} = \begin{pmatrix} -1 & 1 & 2 & -1 & -1 & 1 \\ -1 & 1 & 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} k_{12}\,(x_1)^{\frac{1}{2}}(x_2)^{\frac{3}{2}} \\ k_{21}\,x_3 \\ k_{23}\,x_3 \\ k_{31}\,(x_1)^3 \\ k_{45}\,x_1 \\ k_{54}\,x_4 \end{pmatrix} = N\,v(x), \qquad (3)$$

where we fix an order on the edges, $E = \big((1,2),(2,1),(2,3),(3,1),(4,5),(5,4)\big)$, and introduce the stoichiometric matrix N and the vector of reaction rates $v(x)$.

We further decompose the system. Writing the stoichiometric and kinetic complexes as column vectors of the matrices

$$Y = \begin{pmatrix} 1 & 0 & 2 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad \tilde{Y} = \begin{pmatrix} \frac{1}{2} & 0 & 3 & 1 & 0 \\ \frac{3}{2} & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

and using the incidence matrix of the graph (2),

$$I_E = \begin{pmatrix} -1 & 1 & 0 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix},$$

we can write the stoichiometric matrix as

$$N = Y\,I_E.$$

The vector of reaction rates $v(x)$ can also be decomposed by introducing a diagonal matrix

$$\Delta_k = \mathrm{diag}(k_{12}, k_{21}, k_{23}, k_{31}, k_{45}, k_{54})$$

containing the rate constants, a matrix indicating the source vertex of each reaction,

$$I_s = \begin{pmatrix} 1&0&0&0&0&0\\ 0&1&1&0&0&0\\ 0&0&0&1&0&0\\ 0&0&0&0&1&0\\ 0&0&0&0&0&1 \end{pmatrix},$$

and the vector of monomials determined by the kinetic complexes,

$$x^{\tilde{Y}} = \begin{pmatrix} (x_1)^{\frac{1}{2}}(x_2)^{\frac{3}{2}}\\ x_3\\ (x_1)^3\\ x_1\\ x_4 \end{pmatrix}.$$

Then,

$$v(x) = \Delta_k\, I_s^T\, x^{\tilde{Y}},$$

and we can write

$$\frac{\mathrm{d}x}{\mathrm{d}t} = N\, v(x) = Y\, I_E\, \Delta_k\, I_s^T\, x^{\tilde{Y}}.$$

Note that the matrix

$$A_k = I_E\, \Delta_k\, I_s^T = \begin{pmatrix} -k_{12} & k_{21} & k_{31} & 0 & 0\\ k_{12} & -(k_{21}+k_{23}) & 0 & 0 & 0\\ 0 & k_{23} & -k_{31} & 0 & 0\\ 0 & 0 & 0 & -k_{45} & k_{54}\\ 0 & 0 & 0 & k_{45} & -k_{54} \end{pmatrix} \tag{4}$$

depends only on the weighted digraph, while Y and $x^{\tilde{Y}}$ are determined by the stoichiometric and kinetic complexes. The resulting decomposition

$$\frac{\mathrm{d}x}{\mathrm{d}t} = Y A_k\, x^{\tilde{Y}}$$

is due to [33], where A_k is called kinetic matrix and the stoichiometric and kinetic complexes are equal, that is, $Y = \tilde{Y}$. The interpretation of A_k as a weighted graph Laplacian was introduced in [24] and used in [12,47,31,37,34], in particular, in connection with the matrix-tree theorem.

3 Generalized Mass Action Systems

We consider *directed graphs* $G = (V, E)$ given by a finite set of *vertices*

$$V = \{1, \ldots, m\}$$

and a finite set of *edges* $E \subseteq V \times V$. We often denote an edge $e = (i, j) \in E$ by $i \to j$ to emphasize that it is directed from the *source* i to the *target* j. Further, we write

$$V_s = \{i \mid i \to j \in E\}$$

for the set of source vertices that appear as a source of some edge.

Definition 1. *A* generalized chemical reaction network $(G, y, \tilde{y})$ *is given by a digraph* $G = (V, E)$ *without self-loops, and two functions*

$$y\colon V \to \mathbb{R}^n \quad \textit{and} \quad \tilde{y}\colon V_s \to \mathbb{R}^n$$

assigning to each vertex a (stoichiometric) complex *and to each source a* kinetic complex.

We note that this definition differs from [40]. On the one hand, kinetic complexes were assigned also to non-source vertices, on the other hand, all (stoichiometric) complexes had to be different, and analogously the kinetic complexes.

Definition 2. *A* generalized mass action system $(G_k, y, \tilde{y})$ *is a generalized chemical reaction network* $(G, y, \tilde{y})$*, where edges* $(i, j) \in E$ *are labeled with* rate constants $k_{ij} \in \mathbb{R}_>$.

The contribution of reaction $i \to j \in E$ to the dynamics of the species concentrations $x \in \mathbb{R}^n$ is proportional to the *reaction vector* $y(j) - y(i) \in \mathbb{R}^n$. Assuming generalized mass action kinetics, the rate of the reaction is determined by the source kinetic complex $\tilde{y}(i)$ and the positive rate constant k_{ij}:

$$v_{i\to j}(x) = k_{ij}\, x^{\tilde{y}(i)}.$$

The ordinary differential equation associated with a generalized mass action system is defined as

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \sum_{i\to j\in E} k_{ij}\, x^{\tilde{y}(i)}\big(y(j) - y(i)\big).$$

The change over time lies in the *stoichiometric subspace*

$$S = \operatorname{span}\{y(j) - y(i) \in \mathbb{R}^n \mid i \to j \in E\},$$

which suggests the definition of a (positive) *stoichiometric compatibility class* $(c' + S) \cap \mathbb{R}^n_>$ with $c' \in \mathbb{R}^n_>$.

In case every vertex is a source, that is, $V_s = V$, we introduce also the *kinetic-order subspace*

$$\tilde{S} = \operatorname{span}\{\tilde{y}(j) - \tilde{y}(i) \in \mathbb{R}^n \mid i \to j \in E\}.$$

In order to decompose the right-hand side of the ODE system, we define the matrices $Y \in \mathbb{R}^{n\times m}$ as $y^j = y(j)$ and $\tilde{Y} \in \mathbb{R}^{n\times m}$ as $\tilde{y}^j = \tilde{y}(j)$ for $j \in V_s$ and $\tilde{y}^j = 0$ otherwise (see also the remark below). Further, we introduce the weighted

graph Laplacian $A_k \in \mathbb{R}^{m\times m}$: $(A_k)_{ij} = k_{ji}$ if $j \to i \in E$, $(A_k)_{ii} = -\sum_{i\to j\in E} k_{ij}$, and $(A_k)_{ij} = 0$ otherwise. We obtain:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = Y A_k\, x^{\tilde{Y}}.$$

Note that $\tilde{y}^j$ can be chosen arbitrarily for $j \notin V_s$, since in this case $(A_k)^j = 0$ and hence $(A_k)^j x^{\tilde{y}^j} = 0$.

Steady states of the ODE satisfying $x \in \mathbb{R}^n_>$ and $A_k\, x^{\tilde{Y}} = 0$ are called *complex balancing equilibria*. We denote the corresponding set by

$$Z_k = \{x \in \mathbb{R}^n_> \mid A_k\, x^{\tilde{Y}} = 0\}.$$

Finally, the *(stoichiometric) deficiency* is defined as

$$\delta = m - l - s,$$

where m is the number of vertices, l is the number of connected components, and $s = \dim S$ is the dimension of the stoichiometric subspace.

Using $S = \mathrm{im}(Y\, I_E)$, where I_E is the incidence matrix of the graph (for a fixed order on E), we obtain the equivalent definition

$$\delta = \dim(\ker(Y) \cap \mathrm{im}(I_E)),$$

see for example [34]. Further, note that $\mathrm{im}(A_k) \subseteq \mathrm{im}(I_E)$. Now, if $\delta = 0$, then $\ker(Y) \cap \mathrm{im}(A_k) \subseteq \ker(Y) \cap \mathrm{im}(I_E) = \{0\}$, and there are no $x \in \mathbb{R}^n_>$ such that $Y A_k\, x^{\tilde{Y}} = 0$, but $A_k\, x^{\tilde{Y}} \neq 0$. In other words, if $\delta = 0$, there are no steady states other than complex balancing equilibria.

4 Graph Laplacian

A basis for the kernel of A_k in (4) is given by

$$(k_{31}\, k_{21} + k_{31}\, k_{23}, k_{12}\, k_{31}, k_{23}\, k_{12}, 0, 0)^T \quad \text{and} \quad (0, 0, 0, k_{54}, k_{45})^T.$$

Obviously, the support of the vectors coincides with the connected components of the graph. In general, this holds for the strongly connected components without outgoing edges.

Let $G_k = (V, E, k)$ be a weighted digraph without self-loops and A_k its graph Laplacian. Further, let l be the number of connected components (aka linkage classes) and $T_1, \dots, T_t \subseteq V$ be the sets of vertices within the strongly connected components without outgoing edges (aka terminal strong linkage classes). Clearly, $t \geq l$. A fundamental result of CRNT [21] states that there exist linearly independent $\chi^1, \dots, \chi^t \in \mathbb{R}^n_\geq$, where $\chi^\lambda_\mu > 0$ if $\mu \in T_\lambda$ and $\chi^\lambda_\mu = 0$ otherwise, such that $\ker(A_k) = \mathrm{span}\{\chi^1, \dots, \chi^t\}$.

In fact, the non-zero entries in the basis vectors can be computed using the matrix-tree theorem:

$$\chi^\lambda_\mu = K_\mu, \quad \lambda \in \{1, \dots, t\}$$

with *tree constants*

$$K_\mu = \sum_{\mathcal{T}\in\mathcal{S}_\mu} \prod_{i\to j\in\mathcal{T}} k_{ij}, \quad \mu \in \{1,\dots,m\},$$

where $\mathcal{S}_\mu$ is the set of directed spanning trees (for the respective strongly connected component without outgoing edges) rooted at vertex μ; see [31,37,34]. We refer to [8] for further details and references on the graph Laplacian and a combinatorial proof of the matrix-tree theorem following [49].

If there exists $\psi \in \mathbb{R}^m_>$ with $A_k\,\psi = 0$, then every vertex resides in a strongly connected component without outgoing edges, that is, every connected component is strongly connected. In this case, the underlying unweighted digraph is called *weakly reversible*. Now, let $(G, y, \tilde{y})$ be a generalized chemical reaction network. If there exist rate constants k such that the generalized mass action system $(G_k, y, \tilde{y})$ admits a complex balancing equilibrium $x \in \mathbb{R}^n_>$, that is, $A_k\,x^{\tilde{Y}} = 0$, then G is weakly reversible.

5 Binomial Equations for Complex Balancing Equilibria

For a weakly reversible digraph, we know from the previous section that a basis for $\ker(A_k)$, parametrized by the weights, is given in terms of the l connected components and the m tree constants.

In our example, where $l = 2$ and $m = 5$, basis vectors of $\ker(A_k)$ are given by

$$(K_1, K_2, K_3, 0, 0)^T \quad \text{and} \quad (0, 0, 0, K_4, K_5)^T$$

with tree constants

$$(K_1, K_2, K_3, K_4, K_5) = (k_{31}\,k_{21} + k_{31}\,k_{23}, k_{12}\,k_{31}, k_{23}\,k_{12}, k_{54}, k_{45}).$$

Due to their special structure, we immediately find "binomial" basis vectors for the orthogonal complement $\ker(A_k)^\perp$,

$$(-K_2, K_1, 0, 0, 0)^T, \quad (0, -K_3, K_2, 0, 0)^T, \quad \text{and} \quad (0, 0, 0, -K_5, K_4)^T,$$

which are again determined by the connected components and tree constants. These vectors form a basis since they are linearly independent and

$$\dim\ker(A_k)^\perp = m - \dim\ker(A_k) = m - l = 5 - 2 = 3.$$

In our example, a complex balancing equilibrium $x \in \mathbb{R}^4_>$ with $\psi = x^{\tilde{Y}}$ and hence $A_k\,\psi = 0$, can equivalently be described as a positive solution of the binomial equations

$$\begin{pmatrix} -K_2 & K_1 & 0 & 0 & 0 \\ 0 & -K_3 & K_2 & 0 & 0 \\ 0 & 0 & 0 & -K_5 & K_4 \end{pmatrix} \psi = 0.$$

In other words, $\psi \in \ker(A_k)$ is equivalent to $\psi \perp \ker(A_k)^\perp$ or a basis thereof. Explicitly, we have $\psi = x^{\tilde{Y}} = ((x_1)^{\frac{1}{2}}(x_2)^{\frac{3}{2}},\, x_3,\, (x_1)^3,\, x_1,\, x_4)^T$ and

$$K_1\, x_3 - K_2\, (x_1)^{\frac{1}{2}}(x_2)^{\frac{3}{2}} = 0, \quad K_2\, (x_1)^3 - K_3\, x_3 = 0, \quad K_4\, x_4 - K_5\, x_1 = 0. \tag{5}$$

Clearly, these considerations generalize to arbitrary weakly reversible digraphs: Based on the (strongly) connected components, we can characterize complex balancing equilibria by $m - l$ binomial equations with tree constants as coefficients.

Proposition 1. *Let A_k be the graph Laplacian of a weakly reversible digraph with positive weights and m vertices ordered within l connected components,*

$$L_\lambda = (i_\mu^\lambda)_{\mu=1,\dots,m_\lambda} \quad \text{for } \lambda = 1, \dots, l, \quad \text{where } \textstyle\sum_{\lambda=1}^{l} m_\lambda = m.$$

Let $\tilde{Y} \in \mathbb{R}^{n\times m}$ and

$$Z_k = \{x \in \mathbb{R}^n_> \mid A_k\, x^{\tilde{Y}} = 0\}.$$

Then,

$$Z_k = \{x \in \mathbb{R}^n_> \mid K_i\, x^{\tilde{y}^j} - K_j\, x^{\tilde{y}^i} = 0,\ (i,j) \in \mathcal{E}\}$$

where

$$\mathcal{E} = \{(i_\mu^\lambda, i_{\mu+1}^\lambda) \mid \lambda = 1, \dots, l;\ \mu = 1, \dots, m_\lambda - 1\}.$$

Note that the actual binomial equations depend on the order of the vertices within the connected components, but the zero set does not.

6 Binomial Equations with Real and Symbolic Exponents

In this section, we collect basic facts about positive real solutions of binomial equations with real exponents. We present the results in full generality, in particular, not restricted to complex balancing equilibria, and emphasize algorithmic aspects. Moreover, by reducing computations to linear algebra, we outline the treatment of symbolic exponents.

In an algebraic perspective, one usually considers solutions of binomial equations with integer exponents. We refer to [13] for an introduction including algorithmic aspects and an extensive list of references. An algorithm with polynomial complexity for computing solutions with non-zero or positive coordinates of parametric binomial systems is presented in [29]. For recent algorithmic methods for binomial equations and monomial parametrizations, see [1]. Toric geometry and computer algebra was introduced to the study of mass action systems in [25,27,26] and further developed in [12]. So-called *toric steady states* are solutions of binomial equations arising from polynomial dynamical systems [42].

In chemical reaction networks, it is natural to consider real exponents: kinetic orders, measured by experiments, need not be integers. Also in S-systems [46,48], defined by binomial power-laws, the exponents are real numbers identified from data. We note that binomial equations are implicit in the original works on chemical reaction networks [33,32].

In the following, we consider binomial equations

$$\alpha_i\, x^{a^i} - \beta_i\, x^{b^i} = 0 \quad \text{for } i = 1, \ldots, r$$

for $x \in \mathbb{R}^n_>$, where $a^i, b^i \in \mathbb{R}^n$ and $\alpha_i, \beta_i \in \mathbb{R}_>$. Clearly, x is a solution iff

$$x^{a^i - b^i} = \frac{\beta_i}{\alpha_i} \quad \text{for } i = 1, \ldots, r.$$

By introducing the exponent matrix $M \in \mathbb{R}^{n \times r}$, whose ith column is the vector $a^i - b^i$, and the vectors $\alpha, \beta \in \mathbb{R}^r_>$ with entries α_i and β_i, respectively, we can rewrite the above equation system as

$$x^M = \frac{\beta}{\alpha}.$$

More generally, we are interested for which $\gamma \in \mathbb{R}^r_>$ the equations

$$x^M = \gamma$$

have a positive solution. Taking the logarithm, we obtain the equivalent linear equations

$$M^T \ln x = \ln \gamma, \tag{6}$$

which reduces the problem to linear algebra.

In the rest of this section, we fix a matrix $M \in \mathbb{R}^{n \times r}$ and write

$$Z_{M,\gamma} = \{x \in \mathbb{R}^n_> \mid x^M = \gamma\}$$

for the set of all positive solutions with right-hand side $\gamma \in \mathbb{R}^r_>$.

Proposition 2. *The following statements hold:*

$$Z_{M,\gamma} \neq \emptyset \quad \textit{for all } \gamma \in \mathbb{R}^r_> \quad \textit{iff} \quad \ker(M) = \{0\}.$$

If $\ker(M) \neq \{0\}$*, then*

$$Z_{M,\gamma} \neq \emptyset \quad \textit{for } \gamma \in \mathbb{R}^r_> \quad \textit{iff} \quad \gamma^C = 1,$$

where $C \in \mathbb{R}^{r \times p}$ *with* $\operatorname{im}(C) = \ker(M)$ *and* $\ker(C) = \{0\}$.

Proof. Using (6), $x^M = \gamma$ is equivalent to

$$\ln \gamma \in \operatorname{im}(M^T) = \ker(M)^\perp.$$

Hence, $Z_{M,\gamma} \neq \emptyset$ for all $\gamma \in \mathbb{R}^r_>$ iff $\ker(M) = \{0\}$. If $\ker(M) \neq \{0\}$, then

$$\ln \gamma \in \ker(M)^\perp = \operatorname{im}(C)^\perp \quad \Leftrightarrow \quad C^T \ln \gamma = 0 \quad \Leftrightarrow \quad \gamma^C = 1.$$

□

Computing an explicit positive solution $x^* \in Z_{M,\gamma}$ (if it exists) in terms of γ is equivalent to computing a particular solution for the linear equations (6). For this, we use an arbitrary generalized inverse H of M^T, that is, a matrix $H \in \mathbb{R}^{n\times r}$ such that

$$M^T H M^T = M^T.$$

We refer to [4] for details on generalized inverses.

Proposition 3. *Let $\gamma \in \mathbb{R}^r_>$ such that $\ln\gamma \in \operatorname{im}(M^T)$. Let $H \in \mathbb{R}^{n\times r}$ be a generalized inverse of M^T. Then,*

$$x^* = \gamma^{H^T} \in Z_{M,\gamma}.$$

Proof. By assumption, $\ln\gamma = M^T z$ for some $z \in \mathbb{R}^n$. Then,

$$M^T \ln x^* = M^T H \ln\gamma = M^T H M^T z = M^T z = \ln\gamma$$

and hence $x^* \in Z_{M,\gamma}$ as claimed. □

Given one positive solution $x^* \in Z_{M,\gamma}$, we have a generalized monomial parametrization for the set of all positive solutions.

Proposition 4. *Let $x^* \in Z_{M,\gamma}$. Then,*

$$Z_{M,\gamma} = \{x^* \circ \mathrm{e}^v \mid v \in \operatorname{im}(M)^\perp\}.$$

If $\operatorname{im}(M)^\perp \neq \{0\}$, then

$$Z_{M,\gamma} = \{x^* \circ \xi^{B^T} \mid \xi \in \mathbb{R}^q_>\},$$

where $B \in \mathbb{R}^{n\times q}$ with $\operatorname{im}(B) = \operatorname{im}(M)^\perp$ and $\ker(B) = \{0\}$.

Proof. The first equality follows from (6): $x \in Z_{M,\gamma}$ iff $v = \ln x - \ln x^* \in \ker(M^T) = \operatorname{im}(M)^\perp$, that is, $x = x^* \circ \mathrm{e}^v$ with $v \in \operatorname{im}(M)^\perp$.

Since the columns of B form a basis for $\operatorname{im}(M)^\perp$, we can write $v \in \operatorname{im}(M)^\perp$ uniquely as $v = B\,t$ for some $t \in \mathbb{R}^q$. By introducing $\xi = \mathrm{e}^t \in \mathbb{R}^q_>$, we obtain

$$(\mathrm{e}^v)_i = \mathrm{e}^{v_i} = \mathrm{e}^{\sum_j b_{ij} t_j} = \textstyle\prod_j \xi_j^{b_{ij}} = \xi^{b_i} = (\xi^{B^T})_i,$$

that is, $\mathrm{e}^v = \xi^{B^T}$. □

Note that the conditions for the existence of positive solutions and the parametrization of all positive solutions, respectively, depend only on the vector subspaces $\ker(M)$ and $\operatorname{im}(M)^\perp = \ker(M^T)$.

Summing up, we have seen that computing positive solutions for binomial equations reduces to linear algebra involving the exponent matrix M. The matrices C, H and B from Propositions 2, 3, and 4 can be computed effectively if $M \in \mathbb{Q}^{n\times r}$ and C, B can be chosen to have only integer entries.

Moreover, the linear algebra approach to binomial equations allows to deal algorithmically with indeterminate (symbolic) exponents. We can use computer algebra methods for matrices with symbolic entries like Turing factoring (generalized PLU decomposition) [10] and its implementation [11]. Based on these methods, we can compute explicit monomial parametrizations with symbolic exponents for generic entries and investigate conditions for special cases. See Section 8 for an example.

7 Kinetic Deficiency

Applying the results from the previous section, we rewrite the binomial equations (5) from our example,

$$K_1\, x_3 - K_2\, (x_1)^{\frac{1}{2}} (x_2)^{\frac{3}{2}} = 0, \quad K_2\, (x_1)^3 - K_3\, x_3 = 0, \quad K_4\, x_4 - K_5\, x_1 = 0,$$

as

$$x^M = \kappa_k,$$

where

$$M = \begin{pmatrix} -\frac{1}{2} & 3 & -1 \\ -\frac{3}{2} & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{7}$$

and

$$\kappa_k = (K_2/K_1, K_3/K_2, K_5/K_4)^T,$$

which depends on the weights k via the tree constants K.

Recall that the binomial equations depend on the basis vectors for $\ker(A_k)^\perp$ which are determined by the relation $\mathcal{E} = \{(1,2),(2,3),(4,5)\}$. To specify the resulting exponent matrix M and the right-hand side κ_k, we have fixed an order on the relation. By abuse of notation, we write

$$\mathcal{E} = ((1,2),(2,3),(4,5)).$$

Hence, $M = \tilde{Y}\, I_{\mathcal{E}}$ with

$$I_{\mathcal{E}} = \begin{pmatrix} -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & -1 \\ 0 & 0 & 1 \end{pmatrix}. \tag{8}$$

In general, for a weakly reversible digraph with m vertices and l connected components, let $\mathcal{E}$ be a relation as in Proposition 1 with fixed order. We denote by $I_{\mathcal{E}} \in \mathbb{R}^{m\times(m-l)}$ the matrix with columns

$$e^j - e^i \quad \text{for } (i,j) \in \mathcal{E},$$

where e^i denotes the ith standard basis vector in $\mathbb{R}^m$. Clearly, the columns of $I_{\mathcal{E}}$ are linearly independent and hence $\dim \mathrm{im}(I_{\mathcal{E}}) = m - l$. To rewrite the binomial equations in Proposition 1, we define the exponent matrix $M \in \mathbb{R}^{n\times(m-l)}$ as

$$M = \tilde{Y}\, I_{\mathcal{E}},$$

the right-hand side $\kappa_k \in \mathbb{R}^{m-l}_{>}$ as

$$(\kappa_k)_{(i,j)} = K_j/K_i \quad \text{for } (i,j) \in \mathcal{E}, \tag{9}$$

and obtain

$$Z_k = \{x \in \mathbb{R}^n_> \mid x^M = \kappa_k\}.$$

We note that the actual matrix M depends on $\mathcal{E}$, but $\operatorname{im}(M)$ does not. This can be seen using the following fact.

Proposition 5. *Let $G = (V, E)$ be a digraph with m vertices and l connected components. Let $I_E \in \mathbb{R}^{m\times|E|}$ denote its incidence matrix (for fixed order on E), and let $I_\mathcal{E} \in \mathbb{R}^{m\times(m-l)}$ be as defined above. Then,*

$$\operatorname{im}(I_\mathcal{E}) = \operatorname{im}(I_E).$$

Proof. From graph theory (see for example [35]) and the argument above, we know that $\dim\operatorname{im}(I_E) = \dim\operatorname{im}(I_\mathcal{E}) = m - l$. It remains to show that $\operatorname{im}(I_E) \subseteq \operatorname{im}(I_\mathcal{E})$. We consider the column $e^j - e^i$ of I_E corresponding to the edge $(i, j) \in E$. Clearly, i and j are in the same connected component L_λ, in particular, $i = i^\lambda_{\mu(i)}$ and $j = i^\lambda_{\mu(j)}$, where we assume $\mu(i) < \mu(j)$. Then,

$$e^j - e^i = \sum_{\mu=\mu(i),\dots,\mu(j)-1} e^{i^\lambda_{\mu+1}} - e^{i^\lambda_\mu},$$

where $e^{i^\lambda_{\mu+1}} - e^{i^\lambda_\mu}$ are columns of $I_\mathcal{E}$ corresponding to pairs $(i^\lambda_\mu, i^\lambda_{\mu+1})$ in $\mathcal{E}$. □

Now, we see that $\operatorname{im}(M)$ equals the kinetic-order subspace $\tilde{S}$:

$$\operatorname{im}(M) = \operatorname{im}(\tilde{Y} I_\mathcal{E}) = \operatorname{im}(\tilde{Y} I_E) = \tilde{S}.$$

Finally, we recall that the number of independent conditions on κ_k for the existence of a positive solution of $x^M = \kappa_k$ is given by $\dim\ker(M)$, cf. Proposition 2. Observing $M \in \mathbb{R}^{n\times(m-l)}$, we obtain

$$\dim\ker(M) = m - l - \dim\operatorname{im}(M) = m - l - \dim\tilde{S}. \tag{10}$$

Hence, for a digraph with m vertices and l connected components, we define the *kinetic deficiency* as

$$\tilde{\delta} = m - l - \tilde{s},$$

where $\tilde{s} = \dim\tilde{S}$ denotes the dimension of the kinetic-order subspace.

8 Computing Complex Balancing Equilibria

Combining the results from the previous sections, we obtain the following constructive characterization of complex balancing equilibria in terms of quotients of tree constants.

Theorem 1. *Let A_k be the graph Laplacian of a weakly reversible digraph with positive weights, m vertices, and l connected components. Let $\tilde{Y} \in \mathbb{R}^{n\times m}$ be the matrix of kinetic complexes, $\tilde{s} = \dim\tilde{S}$ the dimension of the kinetic-order*

subspace, and $\tilde{\delta} = m - l - \tilde{s}$ *the kinetic deficiency. Further, let* $M \in \mathbb{R}^{n\times(m-l)}$ *and* $\kappa_k \in \mathbb{R}^{m-l}_>$ *such that*

$$Z_k = \{x \in \mathbb{R}^n_> \mid A_k\, x^{\tilde{Y}} = 0\} = \{x \in \mathbb{R}^n_> \mid x^M = \kappa_k\}.$$

Then, the following statements hold:

(a) $Z_k \neq \emptyset$ *for all* k *iff* $\tilde{\delta} = 0$.

(b) *If* $\tilde{\delta} > 0$, *then*

$$Z_k \neq \emptyset \quad \textit{iff} \quad (\kappa_k)^C = 1,$$

where $C \in \mathbb{R}^{(m-l)\times\tilde{\delta}}$ *with* $\mathrm{im}(C) = \ker(M)$ *and* $\ker(C) = \{0\}$.

(c) *If* $Z_k \neq \emptyset$, *then*

$$x^* = (\kappa_k)^{H^T} \in Z_k,$$

where $H \in \mathbb{R}^{n\times(m-l)}$ *is a generalized inverse of* M^T.

(d) *If* $x^* \in Z_k$ *and* $\tilde{s} < n$, *then*

$$Z_k = \{x^* \circ \xi^{B^T} \mid \xi \in \mathbb{R}^{n-\tilde{s}}_>\},$$

where $B \in \mathbb{R}^{n\times(n-\tilde{s})}$ *with* $\mathrm{im}(B) = \tilde{S}^\perp$ *and* $\ker(B) = \{0\}$.

Proof. By Propositions 2, 3, and 4. In fact, it remains to prove one implication in (a). Assume $Z_k \neq \emptyset$ for all k, that is, there exists a solution to $x^M = \kappa_k$ for all k. By Lemma 1 below, for all $\gamma \in \mathbb{R}^{m-l}_>$, there exists k such that $\kappa_k = \gamma$. Hence, there exists a solution to $x^M = \gamma$ for all γ. Using (10) and Proposition 2, we obtain $\tilde{\delta} = \dim\ker(M) = 0$. □

Lemma 1. *Let* A_k *be the graph Laplacian of a weakly reversible digraph with positive weights,* m *vertices, and* l *connected components, and let* $\kappa_k \in \mathbb{R}^{m-l}_>$ *be the vector of quotients of tree constants defined in* (9). *For all* $\gamma \in \mathbb{R}^{m-l}_>$, *there exists* k *such that* $\kappa_k = \gamma$.

Proof. First, we show that every positive vector $\psi \in \mathbb{R}^m_>$ solves $A_k\,\psi = 0$ for some weights k. Indeed, for given k, the vector of tree constants $K \in \mathbb{R}^m_>$ solves $A_k\,K = 0$, and by choosing $k^*_{ij} = k_{ij}\,\frac{K_i}{\psi_i}$, one obtains

$$\begin{aligned}(A_{k^*}\,\psi)_i &= \sum_{j=1}^m (A_{k^*})_{ij}\,\psi_j = \sum_{j\to i\in E} k^*_{ji}\,\psi_j - \sum_{i\to j\in E} k^*_{ij}\,\psi_i \\ &= \sum_{j\to i\in E} k_{ji}\,K_j - \sum_{i\to j\in E} k_{ij}\,K_i = \sum_{j=1}^m (A_k)_{ij}\,K_j = (A_k\,K)_i = 0\end{aligned}$$

for all $i = 1,\dots,m$, that is, $A_{k^*}\,\psi = 0$.

Let $\mathcal{E}$ be a relation as in Proposition 1 with the obvious order. Using basis vectors of $\ker(A_k)$ having tree constants as entries, we find that

$$\frac{\psi_j}{\psi_i} = \frac{K_j}{K_i} = (\kappa_k)_{(i,j)} \quad \text{for all } (i,j) \in \mathcal{E}.$$

By choosing the entries of $\psi \in \mathbb{R}^m_>$ in the obvious order, every $\gamma \in \mathbb{R}^{m-l}_>$ can be attained by κ_k for some k. □

Remark 1. Theorem 1 is constructive in the following sense:

- To test if the digraph G is weakly reversible, we compute the connected and the strongly connected components and check whether they are equal.
- The tree constants are computed in terms of the weights k, using (fraction-free) Gaussian elimination on the sub-matrices of A_k determined by the (strongly) connected components.
- Given the kinetic complexes $\tilde{Y} \in \mathbb{Q}^{n\times m}$ and the (strongly) connected components of the digraph, we compute a matrix M and a vector κ_k as introduced in Section 7.
- All matrices involved are computed by linear algebra from the exponent matrix M. This can also be done algorithmically if the kinetic complexes $\tilde{Y}$ and hence M contain indeterminate (symbolic) entries; see the end of Section 6.

In our example, $\tilde{\delta} = 5 - 2 - 3 = 0$ and a monomial parametrization of all complex balancing equilibria is given by

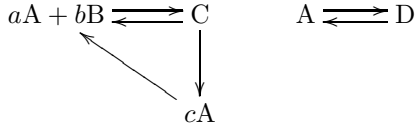
$$\left((\kappa_3)^{-1}, (\kappa_1)^{-\frac{2}{3}} (\kappa_2)^{-\frac{2}{3}} (\kappa_3)^{-\frac{5}{3}}, \kappa_2^{-1} (\kappa_3)^{-3}, 1\right)^T \circ (\xi^3, \xi^5, \xi^9, \xi^3)^T,$$

where

$$\kappa \equiv \kappa_k = \left(\frac{k_{12}}{k_{21}+k_{23}}, \frac{k_{23}}{k_{31}}, \frac{k_{45}}{k_{54}}\right)^T$$

and $\xi \in \mathbb{R}_>$.

To conclude, we associate with each vertex of the graph a kinetic complex possibly containing symbolic coefficients, thereby specifying monomials with symbolic exponents:



(11)

In this setting, a monomial parametrization with symbolic exponents of all complex balancing equilibria is given by

$$\left((\kappa_3)^{-1}, (\kappa_1)^{-\frac{1}{b}} (\kappa_2)^{-\frac{1}{b}} (\kappa_3)^{\frac{a-c}{b}}, (\kappa_2)^{-1} (\kappa_3)^{-c}, 1\right)^T \circ (\xi^b, \xi^{c-a}, \xi^{bc}, \xi^b)^T,$$

which is valid for non-zero $a, b, c \in \mathbb{R}$.

9 Generalized Birch's Theorem

Since the dynamics of generalized mass-action systems is confined to cosets of the stoichiometric subspace, we are interested in uniqueness and existence of complex balancing equilibria in every positive stoichiometric compatibility class.

Let G_k be a weakly reversible digraph with positive weights, m vertices and l connected components. For fixed rate constants k, a complex balancing equilibrium $x^* \in \mathbb{R}^n_>$ of the mass-action system $(G_k, y, \tilde{y})$ solves $A_k\, x^{\tilde{Y}} = 0$, where $A_k \in \mathbb{R}^{m\times m}$ is the graph Laplacian and $\tilde{Y} \in \mathbb{R}^{n\times m}$ is the matrix of kinetic complexes. Equivalently, it solves $x^M = \kappa_k$, where the columns of $M \in \mathbb{R}^{n\times(m-l)}$ are differences of kinetic complexes and the entries of $\kappa_k \in \mathbb{R}^{m-l}_>$ are quotients of the tree constants K, which depend on the weights k. In other words,

$$\begin{aligned} Z_k &= \{x \in \mathbb{R}^n_> \mid A_k\, x^{\tilde{Y}} = 0\} \\ &= \{x \in \mathbb{R}^n_> \mid x^M = \kappa_k\}. \end{aligned}$$

Given a complex balancing equilibrium $x^* \in \mathbb{R}^n_>$, we further know that

$$\begin{aligned} Z_k &= \{x^* \circ \mathrm{e}^v \mid v \in \mathrm{im}(M)^\perp\} \\ &= \{x^* \circ \xi^{B^T} \mid \xi \in \mathbb{R}^{\tilde{d}}_>\}, \end{aligned}$$

where the second equality holds if $\mathrm{im}(M)^\perp \neq \{0\}$ and $B \in \mathbb{R}^{n\times\tilde{d}}$ is defined as $\mathrm{im}(B) = \mathrm{im}(M)^\perp$ and $\ker(B) = \{0\}$.

For simplicity, we write $\tilde{W} = B^T \in \mathbb{R}^{\tilde{d}\times n}$ such that $\tilde{S} = \mathrm{im}(M) = \mathrm{im}(B)^\perp = \mathrm{im}(\tilde{W}^T)^\perp = \ker(\tilde{W})$. Analogously, we introduce a matrix $W \in \mathbb{R}^{d\times n}$ with full rank d such that $S = \ker(W)$.

If the intersection of the set of complex balancing equilibria with some compatibility class,

$$Z_k \cap (x' + S),$$

is non-empty, then there exist $\xi \in \mathbb{R}^{\tilde{d}}_>$ and $u \in S$ such that

$$x^* \circ \xi^{\tilde{W}} = x' + u.$$

Multiplication by W yields

$$W\,(x^* \circ \xi^{\tilde{W}}) = W\,x'$$

such that existence and uniqueness of complex balancing equilibria in every stoichiometric compatibility class are equivalent to surjectivity and injectivity of the generalized polynomial map

$$\begin{aligned} f_{x^*} \colon \mathbb{R}^{\tilde{d}}_> &\to C^\circ \subseteq \mathbb{R}^d \\ \xi &\mapsto W\,(x^* \circ \xi^{\tilde{W}}) = \sum_{i=1}^n x_i^*\, \xi^{\tilde{w}^i}\, w^i, \end{aligned} \tag{12}$$

where C° is the interior of the polyhedral cone

$$C = \left\{Wx' \in \mathbb{R}^d \mid x' \in \mathbb{R}^n_\geq\right\} = \left\{\sum_{i=1}^n x'_i\, w^i \in \mathbb{R}^d \mid x' \in \mathbb{R}^n_\geq\right\}.$$

In mass-action systems, where $S = \tilde{S}$ and hence $W = \tilde{W}$, one version [23] of Birch's theorem [5] states that f_{x^*} is a real analytic isomorphism of $\mathbb{R}^d_>$ onto C° for all $x^* \in \mathbb{R}^n_>$. We refer to [28, Sect. 5] for a recent overview on the use of Birch's theorem in CRNT and to [41] for the version used in algebraic statistics. Interestingly, Martin W. Birch's seminal paper on maximum likelihood methods for log-linear models was part of a PhD thesis at the University of Glasgow that was never submitted [22].

Recently, we have generalized Birch's theorem to $W \neq \tilde{W}$, cf. [40, Proposition 3.9]. To formulate the result, we define the sign vector $\sigma(x) \in \{-, 0, +\}^n$ of a vector $x \in \mathbb{R}^n$ by applying the sign function component-wise, and we write $\sigma(S) = \{\sigma(x) \mid x \in S\}$ for a subset $S \subseteq \mathbb{R}^n$.

Theorem 2. *Let $W \in \mathbb{R}^{d\times n}$, $\tilde{W} \in \mathbb{R}^{\tilde{d}\times n}$ and $S = \ker(W)$, $\tilde{S} = \ker(\tilde{W})$. If $\sigma(S) = \sigma(\tilde{S})$ and $(+, \dots, +)^T \in \sigma(S^\perp)$, then the generalized polynomial map f_{x^*} in (12) is a real analytic isomorphism of $\mathbb{R}^{\tilde{d}}_>$ onto C° for all $x^* \in \mathbb{R}^n_>$.*

If $\tilde{\delta} = 0$, there exists a complex balancing equilibrium for all rate constants k, by Theorem 1. If further the generalized polynomial map f_{x^*} is surjective and injective for all x^*, then, by Theorem 2, there exists a unique steady state in every positive stoichiometric compatibility class for all k.

To illustrate the result, we consider the minimal (weakly) reversible weighted digraph

$$1 \underset{k_{21}}{\overset{k_{12}}{\rightleftarrows}} 2,$$

and associate with each vertex a (stoichiometric) complex

$$\mathrm{A} + \mathrm{B} \rightleftarrows \mathrm{C}$$

as well as a kinetic complex

$$a\mathrm{A} + b\mathrm{B} \rightleftarrows \mathrm{C},$$

where $a, b > 0$. We find $S = \mathrm{im}(-1, -1, 1)^T$ and $\tilde{S} = \mathrm{im}(-a, -b, 1)^T$ and choose

$$W = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \quad \text{and} \quad \tilde{W} = \begin{pmatrix} 1 & 0 & a \\ 0 & 1 & b \end{pmatrix}$$

such that $S = \ker(W)$ and $\tilde{S} = \ker(\tilde{W})$. Clearly, our generalization of Birch's theorem applies since

$$\sigma(S) = \left\{ \begin{pmatrix} - \\ - \\ + \end{pmatrix}, \begin{pmatrix} + \\ + \\ - \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \right\} = \sigma(\tilde{S})$$

and $(1, 1, 2)^T \in S^\perp$. Hence, there exists a unique solution $\xi \in \mathbb{R}^2_>$ for the system of generalized polynomial equations

$$x_1^* \, \xi_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + x_2^* \, \xi_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} + x_3^* \, (\xi_1)^a \, (\xi_2)^b \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}$$

for all right-hand-sides $y \in C^\circ = \mathbb{R}^2_>$, all parameters $x^* \in \mathbb{R}^3_>$, and all exponents $a, b > 0$. Note that Birch's theorem guarantees the existence of a unique solution only for $a = b = 1$.

In terms of the generalized mass-action system above, we have the following result: Since $\tilde{\delta} = 2 - 1 - 1 = 0$, there exists a unique complex balancing equilibrium in every positive stoichiometric compatibility class for all $k_{12}, k_{21} > 0$ and all kinetic orders $a, b > 0$. Since $\delta = 2 - 1 - 1 = 0$, there are no other steady states.

10 Sign Vectors and Oriented Matroids

The characterization of surjectivity and injectivity of generalized polynomial maps involves sign vectors of real linear subspaces, which are basic examples of oriented matroids. (Whereas a matroid abstracts the notion of linear independence, an oriented matroid additionally captures orientation.)

The theory of oriented matroids provides a common framework to study combinatorial properties of various geometric objects, including point configurations, hyperplane arrangements, convex polyhedra, and directed graphs. See [2], [50, Chapters 6 and 7], and [44] for an introduction and overview, and [6] for a comprehensive study.

There are several sets of sign vectors associated with a linear subspace which satisfy the axiom systems for (co-)vectors, (co-)circuits, or chirotopes of oriented matroids. (In fact, there are non-realizable oriented matroids that do not arise from linear subspaces.)

For algorithmic purposes, the characterization of oriented matroids in terms of basis orientations is most useful. The chirotope of a matrix $W \in \mathbb{R}^{d\times n}$ (with rank d) is defined as the map

$$\begin{aligned}\chi_W : \{1,\dots,n\}^d &\to \{-,0,+\}\\ (i_1,\dots,i_d) &\mapsto \operatorname{sign}(\det(w^{i_1},\dots,w^{i_d})),\end{aligned}$$

which records for each d-tuple of vectors whether it forms a positively oriented basis of $\mathbb{R}^d$, a negatively oriented basis, or not a basis. Hence, chirotopes can be used to test algorithmically if the sign vectors of two subspaces are equal by comparing determinants of maximal minors.

More generally, the realization space of matrices defining the same oriented matroid as $W \in \mathbb{R}^{d\times n}$ (with rank d) is described by the semi-algebraic set

$$\begin{aligned}\mathcal{R}(W) = \{A \in \mathbb{R}^{d\times n} \mid &\operatorname{sign}(\det(a^{i_1},\dots,a^{i_d})) =\\ &\operatorname{sign}(\det(w^{i_1},\dots,w^{i_d})),\ 1 \le i_1 < \dots < i_d \le n\}.\end{aligned}$$

Mnëv's universality theorem [38] theorem states that already for oriented matroids with rank $d = 3$, the realization space can be "arbitrarily complicated"; see [6] for a precise statement and [3] for semi-algebraic sets and algorithms.

Concerning software, the `C++` package TOPCOM [43] allows to compute efficiently chirotopes with rational arithmetic and generate all cocircuits (covectors

with minimal support). There is also an interface to the open source computer algebra system SAGE.

In our running example, we have $\tilde{S} = \operatorname{im}(\tilde{Y}\, I_{\mathcal{E}}) = \operatorname{im}(M)$ with M as in (7). Analogously, $S = \operatorname{im}(Y\, I_{\mathcal{E}}) = \operatorname{im}(\mathcal{N})$ with

$$\mathcal{N} = \begin{pmatrix} -1 & 2 & -1 \\ -1 & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{13}$$

To check the sign vector condition $\sigma(S) = \sigma(\tilde{S})$, we compare the chirotopes of $\mathcal{N}^T$ and M^T. Computing the signs of the four maximal minors of $\mathcal{N}^T$, we see that its chirotope is given by

$$\chi_{\mathcal{N}^T}(1,2,3) = -, \quad \chi_{\mathcal{N}^T}(1,2,4) = +, \quad \chi_{\mathcal{N}^T}(1,3,4) = -, \quad \chi_{\mathcal{N}^T}(2,3,4) = +.$$

Analogously, we compute the chirotope of M^T and verify $\chi_{\mathcal{N}^T} = \chi_{M^T}$. Clearly, the other sign vector condition $(+,\dots,+)^T \in \sigma(S^\perp)$ also holds, for example, $(1,1,2,1)^T \in S^\perp$.

Since $\tilde{\delta} = 0$, we know from Theorems 1 and 2 that there exists a unique complex balancing equilibrium in every positive stoichiometric compatibility class for all rate constants k. Moreover, since $\delta = 5 - 2 - 3 = 0$, we know that there are no steady states other than complex balancing equilibria for the ODE (3).

In the setting of symbolic exponents (11), the exponent matrix amounts to

$$M = \begin{pmatrix} -a & c & -1 \\ -b & 0 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{14}$$

and the chirotope of M^T (in the same order as above) is given by

$$-\operatorname{sign}(b), \quad \operatorname{sign}(b\,c), \quad \operatorname{sign}(a-c), \quad \operatorname{sign}(b)$$

for $a, b, c \neq 0$. Hence, there exists a unique steady state in every positive stoichiometric compatibility class for all rate constants and all exponents with $a, b, c > 0$ and $a < c$.

11 Multistationarity

A (generalized) chemical reaction network $(G, y, \tilde{y})$ has *the capacity for multistationarity* if there exist rate constants k such that the generalized mass action system $(G_k, y, \tilde{y})$ admits more than one steady state in some stoichiometric compatibility class.

In mass-action systems, every stoichiometric compatibility class contains at most one complex balancing equilibrium. However, in generalized mass action systems, multiple steady states of this type are possible [40, Proposition 3.2].

Proposition 6. *Let $(G, y, \tilde{y})$ be a generalized chemical reaction network. If G is weakly reversible and $\sigma(S) \cap \sigma(\tilde{S}^{\perp}) \neq \{0\}$, then $(G, y, \tilde{y})$ has the capacity for multiple complex balancing equilibria.*

Analogously, multiple toric steady states are possible (for networks with mass-action kinetics) if the sign vectors of two subspaces intersect non-trivially [9,42]. For deficiency one networks (with mass-action kinetics), the capacity for multistationarity is also characterized by sign conditions [18,20].

For precluding multistationarity, injectivity of the right-hand side of the dynamical system on cosets of the stoichiometric subspace is sufficient. In [39], we characterize injectivity of generalized polynomial maps on cosets of the stoichiometric subspace in terms of sign vectors. There, we also give a survey on injectivity criteria and discuss algorithms to check sign vector conditions.

For the last time, we return to our example, in particular, to the setting of symbolic kinetic complexes. Considering the matrix M in (14), a matrix B with $\mathrm{im}(B) = \mathrm{im}(M)^{\perp} = \tilde{S}^{\perp}$ is given by

$$B = (b, c - a, b\,c, b)^T$$

for $a, b, c \neq 0$. Hence, for $a, b, c > 0$ and $a > c$, we have $(+, -, +, +)^T \in \sigma(\tilde{S}^{\perp})$.

On the other hand, considering the matrix $\mathcal{N}$ in (13) with $\mathrm{im}(\mathcal{N}) = S$, we also have $(+, -, +, +)^T \in \sigma(S)$, and hence $\sigma(S) \cap \sigma(\tilde{S}^{\perp}) \neq \{0\}$. By Proposition 6, if the inequalities $a, b, c > 0$ and $a > c$ hold, then there exist rate constants k that admit more than one complex balancing equilibrium in some stoichiometric compatibility class.

References

1. Adrovic, D., Verschelde, J.: A polyhedral method to compute all affine solution sets of sparse polynomial systems (2013), `http://arxiv.org/abs/1310.4128`, arXiv:1310.4128 [cs.SC]
2. Bachem, A., Kern, W.: Linear programming duality. Springer, Berlin (1992)
3. Basu, S., Pollack, R., Roy, M.F.: Algorithms in real algebraic geometry, 2nd edn. Springer, Berlin (2006)
4. Ben-Israel, A., Greville, T.N.E.: Generalized inverses, 2nd edn. Springer, New York (2003)
5. Birch, M.W.: Maximum likelihood in three-way contingency tables. J. Roy. Statist. Soc. Ser. B 25, 220–233 (1963)
6. Björner, A., Las Vergnas, M., Sturmfels, B., White, N., Ziegler, G.M.: Oriented matroids, 2nd edn. Cambridge University Press, Cambridge (1999)
7. Boulier, F., Lemaire, F., Petitot, M., Sedoglavic, A.: Chemical reaction systems, computer algebra and systems biology. In: Gerdt, V.P., Koepf, W., Mayr, E.W., Vorozhtsov, E.V. (eds.) CASC 2011. LNCS, vol. 6885, pp. 73–87. Springer, Heidelberg (2011)
8. Brualdi, R.A., Ryser, H.J.: Combinatorial matrix theory. Cambridge University Press, Cambridge (1991)
9. Conradi, C., Flockerzi, D., Raisch, J.: Multistationarity in the activation of a MAPK: parametrizing the relevant region in parameter space. Math. Biosci. 211, 105–131 (2008)

10. Corless, R.M., Jeffrey, D.J.: The turing factorization of a rectangular matrix. SIGSAM Bull. 31, 20–30 (1997)
11. Corless, R.M., Jeffrey, D.J.: Linear Algebra in Maple. In: CRC Handbook of Linear Algebra, 2nd edn. Chapman and Hall/CRC (2013)
12. Craciun, G., Dickenstein, A., Shiu, A., Sturmfels, B.: Toric dynamical systems. J. Symbolic Comput. 44, 1551–1565 (2009)
13. Dickenstein, A.: A world of binomials. In: Foundations of Computational Mathematics, Hong Kong, pp. 42–67. Cambridge Univ. Press, Cambridge (2009)
14. Errami, H., Seiler, W.M., Eiswirth, M., Weber, A.: Computing Hopf bifurcations in chemical reaction networks using reaction coordinates. In: Gerdt, V.P., Koepf, W., Mayr, E.W., Vorozhtsov, E.V. (eds.) CASC 2012. LNCS, vol. 7442, pp. 84–97. Springer, Heidelberg (2012)
15. Feinberg, M.: Complex balancing in general kinetic systems. Arch. Rational Mech. Anal. 49, 187–194 (1972)
16. Feinberg, M.: Lectures on chemical reaction networks (1979), `http://crnt.engineering.osu.edu/LecturesOnReactionNetworks`
17. Feinberg, M.: Chemical reaction network structure and the stability of complex isothermal reactors–I. The deficiency zero and deficiency one theorems. Chem. Eng. Sci. 42, 2229–2268 (1987)
18. Feinberg, M.: Chemical reaction network structure and the stability of complex isothermal reactors–II. Multiple steady states for networks of deficiency one. Chem. Eng. Sci. 43, 1–25 (1988)
19. Feinberg, M.: The existence and uniqueness of steady states for a class of chemical reaction networks. Arch. Rational Mech. Anal. 132, 311–370 (1995)
20. Feinberg, M.: Multiple steady states for chemical reaction networks of deficiency one. Arch. Rational Mech. Anal. 132, 371–406 (1995)
21. Feinberg, M., Horn, F.J.M.: Chemical mechanism structure and the coincidence of the stoichiometric and kinetic subspaces. Arch. Rational Mech. Anal. 66, 83–97 (1977)
22. Fienberg, S.E.: Introduction to Birch (1963) Maximum likelihood in three-way contingency tables. In: Kotz, S., Johnson, N.L. (eds.) Breakthroughs in statistics, vol. II, pp. 453–461. Springer, New York (1992)
23. Fulton, W.: Introduction to toric varieties. Princeton University Press, Princeton (1993)
24. Gatermann, K., Wolfrum, M.: Bernstein's second theorem and Viro's method for sparse polynomial systems in chemistry. Adv. in Appl. Math. 34, 252–294 (2005)
25. Gatermann, K.: Counting stable solutions of sparse polynomial systems in chemistry. In: Symbolic Computation: Solving Equations in Algebra, Geometry, and Engineering, pp. 53–69. Amer. Math. Soc., Providence (2001)
26. Gatermann, K., Eiswirth, M., Sensse, A.: Toric ideals and graph theory to analyze Hopf bifurcations in mass action systems. J. Symbolic Comput. 40, 1361–1382 (2005)
27. Gatermann, K., Huber, B.: A family of sparse polynomial systems arising in chemical reaction systems. J. Symbolic Comput. 33, 275–305 (2002)
28. Gopalkrishnan, M., Miller, E., Shiu, A.: A Geometric Approach to the Global Attractor Conjecture. SIAM J. Appl. Dyn. Syst. 13, 758–797 (2014)
29. Grigoriev, D., Weber, A.: Complexity of solving systems with few independent monomials and applications to mass-action kinetics. In: Gerdt, V.P., Koepf, W., Mayr, E.W., Vorozhtsov, E.V. (eds.) CASC 2012. LNCS, vol. 7442, pp. 143–154. Springer, Heidelberg (2012)

30. Gunawardena, J.: Chemical reaction network theory for in-silico biologists (2003), http://vcp.med.harvard.edu/papers/crnt.pdf
31. Gunawardena, J.: A linear framework for time-scale separation in nonlinear biochemical systems. PLoS ONE 7, e36321 (2012)
32. Horn, F.: Necessary and sufficient conditions for complex balancing in chemical kinetics. Arch. Rational Mech. Anal. 49, 172–186 (1972)
33. Horn, F., Jackson, R.: General mass action kinetics. Arch. Rational Mech. Anal. 47, 81–116 (1972)
34. Johnston, M.D.: Translated Chemical Reaction Networks. Bull. Math. Biol. 76, 1081–1116 (2014)
35. Jungnickel, D.: Graphs, networks and algorithms, 4th edn. Springer, Heidelberg (2013)
36. Lemaire, F., Ürgüplü, A.: MABSys: Modeling and analysis of biological systems. In: Horimoto, K., Nakatsui, M., Popov, N. (eds.) ANB 2010. LNCS, vol. 6479, pp. 57–75. Springer, Heidelberg (2012)
37. Mirzaev, I., Gunawardena, J.: Laplacian dynamics on general graphs. Bull. Math. Biol. 75, 2118–2149 (2013)
38. Mnëv, N.E.: The universality theorems on the classification problem of configuration varieties and convex polytopes varieties. In: Topology and geometry—Rohlin Seminar. Lecture Notes in Math., vol. 1346, pp. 527–543. Springer, Berlin (1988)
39. Müller, S., Feliu, E., Regensburger, G., Conradi, C., Shiu, A., Dickenstein, A.: Sign conditions for injectivity of generalized polynomial maps with applications to chemical reaction networks and real algebraic geometry (2013) (submitted), http://arxiv.org/abs/1311.5493, arXiv:1311.5493 [math.AG]
40. Müller, S., Regensburger, G.: Generalized mass action systems: Complex balancing equilibria and sign vectors of the stoichiometric and kinetic-order subspaces. SIAM J. Appl. Math. 72, 1926–1947 (2012)
41. Pachter, L., Sturmfels, B.: Statistics. In: Algebraic statistics for computational biology, pp. 3–42. Cambridge Univ. Press, New York (2005)
42. Pérez Millán, M., Dickenstein, A., Shiu, A., Conradi, C.: Chemical reaction systems with toric steady states. Bull. Math. Biol. 74, 1027–1065 (2012)
43. Rambau, J.: TOPCOM: triangulations of point configurations and oriented matroids. In: Mathematical Software (Beijing 2002), pp. 330–340. World Sci. Publ, River Edge (2002)
44. Richter-Gebert, J., Ziegler, G.M.: Oriented matroids. In: Handbook of Discrete and Computational Geometry, pp. 111–132. CRC, Boca Raton (1997)
45. Samal, S.S., Errami, H., Weber, A.: PoCaB: A software infrastructure to explore algebraic methods for bio-chemical reaction networks. In: Gerdt, V.P., Koepf, W., Mayr, E.W., Vorozhtsov, E.V. (eds.) CASC 2012. LNCS, vol. 7442, pp. 294–307. Springer, Heidelberg (2012)
46. Savageau, M.A.: Biochemical systems analysis: II. The steady state solutions for an n-pool system using a power-law approximation. J. Theor. Biol. 25, 370–379 (1969)
47. Thomson, M., Gunawardena, J.: The rational parameterisation theorem for multisite post-translational modification systems. J. Theoret. Biol. 261, 626–636 (2009)
48. Voit, E.O.: Biochemical systems theory: A review. In: ISRN Biomath. 2013, 897658 (2013)
49. Zeilberger, D.: A combinatorial approach to matrix algebra. Discrete Math. 56, 61–72 (1985)
50. Ziegler, G.M.: Lectures on polytopes. Springer, New York (1995)

Found Comput Math (2016) 16:69–97
DOI 10.1007/s10208-014-9239-3

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Sign Conditions for Injectivity of Generalized Polynomial Maps with Applications to Chemical Reaction Networks and Real Algebraic Geometry

Stefan Müller · Elisenda Feliu · Georg Regensburger · Carsten Conradi · Anne Shiu · Alicia Dickenstein

Received: 21 November 2013 / Revised: 30 September 2014 / Accepted: 26 October 2014 /
Published online: 6 January 2015

Abstract We give necessary and sufficient conditions in terms of sign vectors for the injectivity of families of polynomial maps with arbitrary real exponents defined on the positive orthant. Our work relates and extends existing injectivity conditions expressed in terms of Jacobian matrices and determinants. In the context of chemical

Communicated by Marie-Francoise Roy.

Stefan Müller, Elisenda Feliu, and Georg Regensburger have contributed equally to this work.

S. Müller · G. Regensburger
Johann Radon Institute for Computational and Applied Mathematics, Austrian Academy of Sciences, Altenbergerstraße 69, 4040 Linz, Austria
e-mail: stefan.mueller@ricam.oeaw.ac.at

G. Regensburger
e-mail: georg.regensburger@ricam.oeaw.ac.at

E. Feliu
Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark
e-mail: efeliu@math.ku.dk

C. Conradi
Max-Planck-Institut Dynamik komplexer technischer Systeme, Sandtorstr. 1, 39106 Magdeburg, Germany
e-mail: conradi@mpi-magdeburg.mpg.de

A. Shiu
Department of Mathematics, Texas A&M University, Mailstop 3368, College Station, TX 77843-3368, USA
e-mail: annejls@math.tamu.edu

A. Dickenstein (✉)
Dto. de Matemática, FCEN, Universidad de Buenos Aires, and IMAS (UBA-CONICET), Ciudad Universitaria, Pab. I, C1428EGA Buenos Aires, Argentina
e-mail: alidick@dm.uba.ar

Springer

reaction networks with power-law kinetics, our results can be used to preclude as well as to guarantee multiple positive steady states. In the context of real algebraic geometry, our work recognizes a prior result of Craciun, Garcia-Puente, and Sottile, together with work of two of the authors, as the first partial multivariate generalization of the classical Descartes' rule, which bounds the number of positive real roots of a univariate real polynomial in terms of the number of sign variations of its coefficients.

Keywords Sign vector · Restricted injectivity · Power-law kinetics · Descartes' rule of signs · Oriented matroid

Mathematics Subject Classification 13P15 · 12D10 · 70K42 · 37C10 · 80A30 · 52C40

1 Introduction

In many fields of science, the analysis of parametrized systems by way of sign vectors has a long history. In economics, market models depend on monotonic price and demand curves, leading to the theory of sign-solvable linear systems [15,49]. In electronics, devices such as diodes, transistors, and operational amplifiers are characterized by monotonic functions, and one studies whether the input-output relation of an electronic circuit is well posed, using the theory of oriented matroids [16,61]. In many settings, uniqueness of positive solutions is a desirable property, but deciding this is difficult in general [24,50]. If, however, the maps of interest are injective, then this precludes multiple solutions.

Motivated by applications to chemical reaction networks and real algebraic geometry, we characterize injectivity of parametrized families of polynomial maps with arbitrary real exponents, in terms of sign vectors. Our work builds on results from chemical engineering, by abstracting, relating, and extending existing injectivity conditions expressed in terms of Jacobian matrices and determinants.

The relevant literature from the theory of chemical reaction networks is discussed in Sect. 1.2. The main application to real algebraic geometry is addressed in Sect. 1.3.

1.1 Statement of the Main Theorem

Throughout this paper, we consider families of maps defined on the positive orthant, associated with two real matrices of coefficients and exponents, respectively, and a vector of positive parameters.

Definition 1.1 Let $A = (a_{ij}) \in \mathbb{R}^{m\times r}$, $B = (b_{ij}) \in \mathbb{R}^{r\times n}$, and $\kappa \in \mathbb{R}^r_+$. We define the associated *generalized polynomial map* $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^m$ as

$$f_{\kappa,i}(x) = \sum_{j=1}^{r} a_{ij}\,\kappa_j\, x_1^{b_{j1}} \dots x_n^{b_{jn}}, \quad i = 1, \dots, m.$$

The term *generalized* indicates that we allow polynomials with real exponents. In the literature, generalized polynomials occur under other names. For instance, they are called *signomials* in geometric programming [14].

We often use a more compact notation. By introducing $A_\kappa \in \mathbb{R}^{m\times r}$ as $A_\kappa = A\,\mathrm{diag}(\kappa)$ and $x^B \in \mathbb{R}^r_+$ via $(x^B)_j = x_1^{b_{j1}} \dots x_n^{b_{jn}}$ for $j = 1, \dots, r$, we can write

$$f_\kappa(x) = A_\kappa\, x^B. \tag{1}$$

A generalized polynomial map $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^n$ (1) with $A \in \mathbb{R}^{n\times r}$ and $B \in \mathbb{R}^{r\times n}$, induces a system of ordinary differential equations (ODEs) called a *power-law system*:

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f_\kappa(x). \tag{2}$$

For any initial value $x_0 \in \mathbb{R}^n_+$, the solution is confined to the coset $x_0 + S_\kappa$, where S_κ is the smallest vector subspace containing the image of f_κ. Hence, when studying positive steady states of (2), one is in general interested in the positive solutions to the equation $f_\kappa(x) = 0$ within cosets $x' + S_\kappa$ with $x' \in \mathbb{R}^n_+$. Due to the form of f_κ, one has $S_\kappa \subseteq S$ where $S = \mathrm{im}(A)$. In many applications, $S_\kappa = S$ for all $\kappa \in \mathbb{R}^r_+$, for example, if the rows of B are distinct. If f_κ is injective on $(x' + S) \cap \mathbb{R}^n_+$, then $f_\kappa(x) \neq f_\kappa(y)$ for all distinct $x, y \in (x' + S) \cap \mathbb{R}^n_+$, and hence the coset $x' + S$ contains at most one positive steady state. Clearly, for a vector subspace S of $\mathbb{R}^n$, two vectors $x, y \in \mathbb{R}^n$ lie in $x' + S$ for some $x' \in \mathbb{R}^n$, if and only if $x - y \in S$. This motivates the following definition of injectivity with respect to a subset.

Definition 1.2 Given two subsets $\Omega, S \subseteq \mathbb{R}^n$, a function g defined on Ω is called *injective with respect to* S if $x, y \in \Omega$, $x \neq y$, and $x - y \in S$ imply $g(x) \neq g(y)$.

We will in general consider functions defined on the positive orthant, that is, $\Omega = \mathbb{R}^n_+$. When S is a vector subspace, injectivity with respect to S is equivalent to injectivity on every coset $x' + S$.

When the matrix B has integer entries, determining the injectivity of the map f_κ for a fixed (computable) parameter value κ, with respect to a semialgebraic subset S, is a question of quantifier elimination and thus can be decided algorithmically, but is very hard in practice. This paper focuses on how to decide injectivity for the whole family, that is, for *all* possible values of $\kappa \in \mathbb{R}^r_+$, for a matrix B with real entries. Our results are given in terms of sign vectors characterizing the orthants that $\ker(A)$ and (a subset of) $\mathrm{im}(B)$ intersect nontrivially.

Definition 1.3 For a vector $x \in \mathbb{R}^n$, we obtain the *sign vector* $\sigma(x) \in \{-, 0, +\}^n$ by applying the sign function componentwise.

Note that a sign vector $v \in \{-, 0, +\}^n$ corresponds to the (possibly lower dimensional) orthant of $\mathbb{R}^n$ given by $\sigma^{-1}(v)$. For a subset $S \subseteq \mathbb{R}^n$, we write $\sigma(S) = \{\sigma(x) \mid x \in S\}$ for the set of all sign vectors of S and

$$\Sigma(S) = \sigma^{-1}(\sigma(S))$$

for the union of all (possibly lower dimensional) orthants that S intersects. For convenience, we introduce $S^* = S\backslash\{0\}$.

In order to state our main result, we require some more notation. Identifying $B \in \mathbb{R}^{r\times n}$ with the linear map $B\colon \mathbb{R}^n \to \mathbb{R}^r$, we write $B(S)$ for the image under B of the subset $S \subseteq \mathbb{R}^n$. In analogy to A_κ, we introduce $B_\lambda = B\,\mathrm{diag}(\lambda)$ for $\lambda \in \mathbb{R}^n_+$. Finally, we write J_{f_κ} for the Jacobian matrix associated with the map f_κ. Here is our main result, which brings together and extends various existing results (see Sect. 1.2).

Theorem 1.4 *Let $f_\kappa\colon \mathbb{R}^n_+ \to \mathbb{R}^m$ be the generalized polynomial map $f_\kappa(x) = A_\kappa\, x^B$, where $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{r\times n}$, and $\kappa \in \mathbb{R}^r_+$. Further, let $S \subseteq \mathbb{R}^n$. The following statements are equivalent:*

(inj) f_κ *is injective with respect to* S, *for all* $\kappa \in \mathbb{R}^r_+$.
(jac) $\ker\left(J_{f_\kappa}(x)\right) \cap S^* = \emptyset$, *for all* $\kappa \in \mathbb{R}^r_+$ *and* $x \in \mathbb{R}^n_+$.
(lin) $\ker(A_\kappa B_\lambda) \cap S^* = \emptyset$, *for all* $\kappa \in \mathbb{R}^r_+$ *and* $\lambda \in \mathbb{R}^n_+$.
(sig) $\sigma(\ker(A)) \cap \sigma(B(\Sigma(S^*))) = \emptyset$.

Note that, for a fixed exponent matrix B, condition (sig) depends only on the sign vectors of $\ker(A)$ and S. In particular, f_κ is injective with respect to S for all $\kappa \in \mathbb{R}^r_+$ if and only if it is injective with respect to $\Sigma(S) \subseteq \mathbb{R}^n$, which is the largest set having the same sign vectors as S.

To study unrestricted injectivity, we set $S = \mathbb{R}^n$ in Theorem 1.4, in which case condition (sig) is equivalent to

$$\ker(B) = \{0\} \quad \text{and} \quad \sigma(\ker(A)) \cap \sigma(\mathrm{im}(B)) = \{0\};$$

see Corollary 2.8. Assuming $\ker(B) = \{0\}$, condition (sig) depends only on the corresponding vector subspaces $\ker(A)$ and $\mathrm{im}(B)$; see also [53, Theorem 3.6].

Birch's theorem [12] in statistics corresponds to the unrestricted case $S = \mathbb{R}^n$ and $B = A^T$ with full rank n. Note that $\mathrm{im}(B) = \mathrm{im}(A^T) = \ker(A)^\perp$, and hence $\sigma(\ker(A)) \cap \sigma(\mathrm{im}(B)) = \{0\}$ is trivially fulfilled. Therefore, statement (inj) holds, so Theorem 1.4 guarantees that for any choice of vectors $y \in \mathbb{R}^n$ and $\kappa \in \mathbb{R}^r_+$, there is at most one solution $x \in \mathbb{R}^n_+$ to the equations

$$\sum_{j=1}^{r} a_{ij}\,\kappa_j\, x_1^{a_{1j}} \cdots x_n^{a_{nj}} = y_i, \quad i = 1, \ldots, n.$$

In fact, Birch's theorem also guarantees the existence of a solution, for all y in the interior of the polyhedral cone generated by the columns of A. A related result, due to Horn, Jackson, and Feinberg, asserts the existence and uniqueness of complex balancing equilibria [26,43,44], which is discussed in the next subsection and Sect. 3. Our generalization of Birch's theorem based on [53] is given in statement (ex) of Theorem 1.5.

1.2 Motivation from Chemical Reaction Networks

For chemical reaction networks with mass-action kinetics, the concentration dynamics are governed by dynamical systems (2) with polynomial maps $f_\kappa(x) = A_\kappa\, x^B$, as defined in (1). We introduce some terms that are standard in the chemical engineering literature. The components of $\kappa \in \mathbb{R}^r_+$ are called *rate constants* and are often unknown in practice. The vector subspace $S = \mathrm{im}(A)$ is called the *stoichiometric subspace*. One speaks of *multistationarity* if there exist a vector of rate constants $\kappa \in \mathbb{R}^r_+$ and two distinct positive vectors $x, y \in \mathbb{R}^n_+$ with $x - y \in S$ such that $f_\kappa(x) = f_\kappa(y) = 0$. Clearly, if f_κ is injective with respect to S for all values of κ, then multistationarity is ruled out. Therefore, Theorem 1.4 can be applied in this setting to preclude multistationarity.

Indeed, our work unifies and extends existing conditions for injectivity established in the context of chemical reaction networks. The first such result was given by Craciun and Feinberg for the special case of a fully open network, that is, when each chemical species has an associated outflow reaction and hence $S = \mathbb{R}^n$: injectivity of the corresponding family of polynomial maps was characterized by the nonsingularity of the associated Jacobian matrices, which could be assessed by determinantal conditions [20]. An elementary proof of this foundational result appeared in the context of geometric modeling [24], and extended Jacobian and determinantal criteria were subsequently achieved for arbitrary networks [32,37,46]. Also, for networks with uni- and bimolecular reactions and fixed rate constants, injectivity of the polynomial map has been characterized [56]. Injectivity results have been obtained also for families of kinetics different from mass-action, in particular, for nonautocatalytic kinetics [7,8], power-law kinetics and strictly monotonic kinetics [33,73], weakly monotonic kinetics [65], and other families [9]. Further, several injectivity criteria have been translated to conditions on the species-reaction graph or the interaction graph [7,22,41,52,66].

Sign conditions for the injectivity of monomial maps have been applied both to preclude and to assert multiple positive steady states for several special types of steady states, such as *detailed balancing* and *complex balancing equilibria* of mass-action systems [26,43,44], *toric steady states* of mass-action systems [57], and complex balancing equilibria of generalized mass-action systems [53]. Specifically, such special steady states are parametrized by a monomial map, and multistationarity occurs if and only if the sign vectors of two vector subspaces intersect nontrivially. Moreover, for given rate constants, existence of one complex balancing equilibrium in a mass-action system implies existence and uniqueness of such steady states within each coset of the stoichiometric subspace, and no other steady states are possible [44].

In this paper, we unify and extend the criteria for injectivity and multistationarity described above. Related results appear in the deficiency-oriented theory, as initiated by Horn, Jackson, and Feinberg [26,43,44] (see also [27–31]). This theory is named after the *deficiency* of a reaction network, a nonnegative integer that can be computed from basic network properties. Deficiency zero networks with mass-action kinetics admit positive steady states if and only if the network is strongly connected, and, in this case, there is a unique positive steady state, which is a complex balancing equilibrium. On the other hand, some networks with deficiency one admit multiple positive steady states, and the capacity for multistationarity is characterized by certain sign conditions

[29,31]. For other uses of sign conditions to determine multistationarity, see [17–19] and the related applications to particular biochemical networks [42].

1.3 Application to Real Algebraic Geometry

An interesting consequence in the realm of real algebraic geometry that emerges from the study of injectivity of generalized polynomial maps in applications is Theorem 1.5 below. Statement (bnd) in that result was first proved by Craciun, Garcia-Puente, and Sottile in their study of control points for toric patches [24, Corollary 8] based on a previous injectivity result by Craciun and Feinberg [20]. The surjectivity result underlying statement (ex) is due to Müller and Regensburger [53, Theorem 3.8], who use arguments of degree theory for differentiable maps. We recognize Theorem 1.5 as the first partial multivariate generalization of the following well-known rule proposed by René Descartes in 1637 in "La Géometrie," an appendix to his "Discours de la Méthode," see [71, pp. 96–99]. No multivariate generalization is known, and only a lower bound together with a disproven conjecture was proposed by Itenberg and Roy in 1996 [45].

Descartes' rule of signs *Given a univariate real polynomial* $f(x) = c_0 + c_1 x + \cdots + c_r x^r$*, the number of positive real roots of* f *(counted with multiplicity) is bounded above by the number of sign variations in the ordered sequence of the coefficients* $c_0, \ldots, c_r$*, more precisely, discard the zeros in this sequence and then count the number of times two consecutive entries have different signs. Additionally, the difference between these two numbers (the number of positive roots and the number of sign variations) is even.*

For instance, given the polynomial $f(x) = c_0 + x - x^2 + x^k$ with degree $k > 2$, the number of variations in the sequence $\operatorname{sign}(c_0), +, -, +$ equals 3 if $c_0 < 0$ and 2 if $c_0 \geq 0$. Hence, f admits at most 3 or 2 positive real roots, respectively, and this is independent of its degree.

An important consequence of Descartes' rule of signs is that the number of real roots of a real univariate polynomial f can be bounded in terms of the number of monomials in f (with nonzero coefficient), independently of the degree of f. In the multivariate case, Khovanskii [48, Corollary 7] proved the remarkable result that the number of nondegenerate solutions in $\mathbb{R}^n$ of a system of n real polynomial equations can also be bounded solely in terms of the number q of distinct monomials appearing in these equations. Explicitly, the number of nondegenerate positive roots is at most $2^{(q-1)(q-2)/2}\,(n+1)^{q-1}$. In contrast to Descartes' rule, this bound is far from sharp, and the only known refinements of this bound do not depend on the signs of the coefficients of f [69, Chapters 5–6]. Accordingly, we view Theorem 1.5 as the first partial multivariate generalization of Descartes' rule, as the conditions of the theorem for precluding more than one positive solution depend both on the coefficients and the monomials of f.

We require the following notation. We introduce $[r] = \{1, \ldots, r\}$ for any natural number r. For $A \in \mathbb{R}^{n\times r}$ and $B \in \mathbb{R}^{r\times n}$ with $n \leq r$, and some index set $J \subseteq [r]$ of

cardinality n, we write $A_{[n],J}$ for the submatrix of A indexed by the columns in J and $B_{J,[n]}$ for the submatrix of B indexed by the rows in J.

For any choice of $y \in \mathbb{R}^n$, we consider the system of n equations in n unknowns

$$\sum_{j=1}^{r} a_{ij}\, x_1^{b_{j1}} \cdots x_n^{b_{jn}} = y_i, \quad i = 1, \ldots, n. \tag{3}$$

We denote by $C^\circ(A)$ the interior of the polyhedral cone generated by the column vectors $a^1, \ldots, a^r$ of A:

$$C^\circ(A) = \left\{ \sum_{i=1}^{r} \mu_i\, a^i \in \mathbb{R}^n \mid \mu \in \mathbb{R}^r_+ \right\}.$$

Theorem 1.5 (Multivariate Descartes' rule for (at most) one positive real root) *Let $A \in \mathbb{R}^{n\times r}$ and $B \in \mathbb{R}^{r\times n}$ be matrices with full rank n. Then,*

(bnd) Assume that for all index sets $J \subseteq [r]$ of cardinality n, the product $\det\left(A_{[n],J}\right)\det\left(B_{J,[n]}\right)$ either is zero or has the same sign as all other nonzero products, and moreover, at least one such product is nonzero. Then, (3) has at most one positive solution $x \in \mathbb{R}^n_+$, for any $y \in \mathbb{R}^n$.

(ex) Assume that the row vectors of B lie in an open half-space and that the determinants $\det\left(A_{[n],J}\right)$ and $\det\left(B_{J,[n]}\right)$ have the same sign for all index sets $J \subseteq [r]$ of cardinality n, or the opposite sign in all cases. Then, (3) has exactly one positive solution $x \in \mathbb{R}^n_+$ if and only if $y \in C^\circ(A)$.

Note that the sign conditions in statement (ex) together with the full rank of the matrices imply the hypotheses of (bnd).

To analyze a univariate polynomial $f(x) = c_0 + c_1 x + \cdots + c_r x^r$ in the setting of Theorem 1.5, we have $A \in \mathbb{R}^{1\times r}$ with entries $c_1, \ldots, c_r$, $B \in \mathbb{R}^{r\times 1}$ with entries $1, \ldots, r$, and $y = -c_0$. In this univariate case, the hypotheses of (bnd) in Theorem 1.5 reduce to the conditions that $c_1, \ldots, c_r$ are all nonnegative (or nonpositive) and not all are zero. If these hold, Theorem 1.5 states that f has at most one positive real root and, furthermore, if c_0 has the opposite sign from the nonzero $c_1, \ldots, c_r$'s, (ex) guarantees the existence of this root. Indeed, there is at most one sign variation, depending on $\mathrm{sign}(c_0)$, and so the classical Descartes' rule yields the same conclusion. The result is also valid in the case of real, not necessarily natural, exponents.

In Proposition 3.12, we consider the more general system of m equations in n unknowns with r parameters: $f_\kappa(x) = y$, where f_κ is as in Definition 1.1. More precisely, we give a criterion via sign vectors for precluding multiple positive real solutions $x \in \mathbb{R}^n_+$ for all $y \in \mathbb{R}^m$ and $\kappa \in \mathbb{R}^r_+$.

We will give the proof of Theorem 1.5 in Sect. 3.3, where we restate the sign conditions on the minors of A and B in terms of oriented matroids. Based on this approach, a generalization for multivariate polynomials systems in n variables with $n + 2$ distinct monomials is given in [11]. This case shows the intricacy inherent in the pursuit of a full generalization of Descartes' rule to the multivariate case.

Outline of the paper In Sect. 2, we characterize, in terms of sign vectors, the injectivity of a family of generalized polynomial maps with respect to a subset. In particular, we prove Theorem 1.4, thereby isolating and generalizing key ideas in the literature. Further, we relate our results to determinantal conditions, in case the subset is a vector subspace. In Sect. 3, we apply our results to chemical reaction networks with power-law kinetics, thereby relating and extending previous results. We give conditions for precluding multistationarity in general, for precluding multiple "special" steady states, and for guaranteeing the existence of two or more such steady states. Further, we present applications to real algebraic geometry. We prove the partial multivariate generalization of Descartes' rule, Theorem 1.5, and we restate the hypotheses in the language of oriented matroids. Finally, in Sect. 4, we address algorithmic aspects of our results, in particular, the efficient computation of sign conditions to decide injectivity.

2 Sign Conditions for Injectivity

In this section, we characterize, in terms of sign vectors, generalized polynomial maps $f_\kappa(x) = A_\kappa\, x^B$ that are injective with respect to a subset for all choices of the positive parameters κ. We accomplish this through a series of results that lead to the proof of Theorem 1.4.

2.1 Notation

Here, we summarize the notation used throughout this work. Moreover, we elaborate on the concept of sign vectors defined in the introduction.

We denote the strictly positive real numbers by $\mathbb{R}_+$ and the nonnegative real numbers by $\overline{\mathbb{R}}_+$. We define $\mathrm{e}^x \in \mathbb{R}^n_+$ for $x \in \mathbb{R}^n$ componentwise, that is, $(\mathrm{e}^x)_i = \mathrm{e}^{x_i}$; analogously, $\ln(x) \in \mathbb{R}^n$ for $x \in \mathbb{R}^n_+$ and $x^{-1} \in \mathbb{R}^n$ for $x \in \mathbb{R}^n$ with $x_i \neq 0$. For $x, y \in \mathbb{R}^n$, we denote the componentwise (or Hadamard) product by $x \circ y \in \mathbb{R}^n$, that is, $(x \circ y)_i = x_i\, y_i$. Further, we define $x^b \in \mathbb{R}$ for $x \in \mathbb{R}^n_+$ and $b \in \mathbb{R}^n$ as $x^b = \prod_{i=1}^n x_i^{b_i}$.

Given a matrix $B \in \mathbb{R}^{r\times n}$, we denote by $b^1, \ldots, b^n$ its column vectors and by $b_1, \ldots, b_r$ its row vectors. Thus, the jth coordinate of the map $x^B \colon \mathbb{R}^n_+ \to \mathbb{R}^r_+$ is given by

$$(x^B)_j = x^{b_j} = x_1^{b_{j1}} \ldots x_n^{b_{jn}}.$$

Recall that we define B_λ for $B \in \mathbb{R}^{r\times n}$ and $\lambda \in \mathbb{R}^n_+$ as $B_\lambda = B\,\mathrm{diag}(\lambda)$.

We identify a matrix $B \in \mathbb{R}^{r\times n}$ with the corresponding linear map $B \colon \mathbb{R}^n \to \mathbb{R}^r$ and write $\mathrm{im}(B)$ and $\ker(B)$ for the respective vector subspaces. For a subset $S \subseteq \mathbb{R}^n$, we write $S^* = S\backslash\{0\}$ and denote the image of S under B by

$$B(S) = \{B\,x \mid x \in S\}.$$

For any natural number n, we define $[n] = \{1, \ldots, n\}$. Given sets $I \subseteq [n]$ and $J \subseteq [r]$, we denote the submatrix of B with row indices in J and column indices in I by $B_{J,I}$.

Now, we are ready to state some consequences of Definition 1.3. For $x, y \in \mathbb{R}^n$, we have the equivalence

$$\sigma(x) = \sigma(y) \Leftrightarrow x = \lambda \circ y \quad \text{for some} \quad \lambda \in \mathbb{R}^n_+,$$

and hence, for $S \subseteq \mathbb{R}^n$, we obtain

$$\Sigma(S) = \sigma^{-1}(\sigma(S)) = \{\lambda \circ x \mid \lambda \in \mathbb{R}^n_+ \text{ and } x \in S\}. \tag{4}$$

For subsets $X, Y \subseteq \mathbb{R}^n$, we have the equivalences

$$\Sigma(X) \cap Y = \emptyset \;\Leftrightarrow\; \sigma(X) \cap \sigma(Y) = \emptyset \;\Leftrightarrow\; X \cap \Sigma(Y) = \emptyset. \tag{5}$$

2.2 Families of Linear Maps

In this subsection, we consider the case of linear maps. We start with a useful lemma.

Lemma 2.1 *Let $B \in \mathbb{R}^{r\times n}$ and $S \subseteq \mathbb{R}^n$. The following statements are equivalent:*

(i) $\ker(B_\lambda) \cap S = \emptyset$, *for all* $\lambda \in \mathbb{R}^n_+$.
(ii) $\sigma(\ker(B)) \cap \sigma(S) = \emptyset$.

Proof Statement (i) holds if and only if $B_\lambda x = B(\lambda \circ x) \neq 0$ for all $\lambda \in \mathbb{R}^n_+$ and $x \in S$, that is, if and only if $\ker(B) \cap \Sigma(S) = \emptyset$. By (5), this is equivalent to statement (ii). □

We note that, if $0 \in S$, statements (i) and (ii) do not hold, so we instead apply Lemma 2.1 to S^*. In particular, if S is a vector subspace of $\mathbb{R}^n$, then $\ker(B_\lambda) \cap S^* = \emptyset$ reduces to $\ker(B_\lambda) \cap S = \{0\}$, that is, B_λ is injective on S.

Now, we are ready to prove the equivalence of statements (lin) and (sig) in Theorem 1.4.

Proposition 2.2 *Let $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{r\times n}$, and $S \subseteq \mathbb{R}^n$. The following statements are equivalent:*

(i) $\ker(A_\kappa B_\lambda) \cap S = \emptyset$, *for all* $\kappa \in \mathbb{R}^r_+$ *and* $\lambda \in \mathbb{R}^n_+$.
(ii) $\sigma(\ker(A)) \cap \sigma(B(\Sigma(S))) = \emptyset$.

Proof Clearly, statement (i) is equivalent to $\ker(A_\kappa) \cap B_\lambda(S) = \emptyset$, for all $\kappa \in \mathbb{R}^r_+$ and $\lambda \in \mathbb{R}^n_+$. Using (4), this is equivalent to $\ker(A_\kappa) \cap B(\Sigma(S)) = \emptyset$, for all $\kappa \in \mathbb{R}^r_+$. By Lemma 2.1 applied to the matrix A and the subset $B(\Sigma(S))$, this is in turn equivalent to statement (ii). □

Again, if S is a vector subspace, $\ker(A_\kappa B_\lambda) \cap S^* = \emptyset$ reduces to $\ker(A_\kappa B_\lambda) \cap S = \{0\}$, that is, $A_\kappa B_\lambda$ is injective on S. Clearly, the statements in Lemma 2.1 are necessary conditions for the statements in Proposition 2.2.

2.3 Families of Generalized Monomial/Polynomial Maps

In this subsection, we use the results on families of linear maps to give sign conditions for the injectivity of families of generalized polynomial maps with respect to a subset.

From Definition 1.2, we conclude that a function g defined on $\mathbb{R}^n_+$ is injective with respect to a subset $S \subseteq \mathbb{R}^n$ if and only if for every $x \in \mathbb{R}^n_+$ one has $g(x) \neq g(y)$ for all $y \in (x + S^*) \cap \mathbb{R}^n_+$, where $x + S^* := \{x + y \mid y \in S^*\}$. In case S is a vector subspace, then such a function g is injective on the intersection $(x + S) \cap \mathbb{R}^n_+$ of any coset $x + S$ with the domain $\mathbb{R}^n_+$.

We start with a key observation.

Lemma 2.3 *For $S \subseteq \mathbb{R}^n$, let*

$$\Lambda(S) := \{\ln x - \ln y \mid x, y \in \mathbb{R}^n_+ \text{ and } x - y \in S\}. \tag{6}$$

Then, $\Lambda(S) = \Sigma(S)$.

Proof Let $x, y \in \mathbb{R}^n_+$ such that $x - y \in S$. Then, using the strict monotonicity of the logarithm, we have $\sigma(\ln x - \ln y) = \sigma(x - y) \in \sigma(S)$ and hence $\ln x - \ln y \in \Sigma(S)$. This proves the inclusion $\Lambda(S) \subseteq \Sigma(S)$. Conversely, let $\lambda \in \mathbb{R}^n_+$ and $z \in S$. We construct $x, y \in \mathbb{R}^n_+$ such that $\ln x - \ln y = \lambda \circ z$ and $x - y = z$ as follows: if $z_i \neq 0$, then $e^{\lambda_i z_i} \neq 1$, so we may define $y_i := z_i/(\mathrm{e}^{\lambda_i z_i} - 1)$ and $x_i := y_i\, \mathrm{e}^{\lambda_i z_i}$; otherwise, set $x_i = y_i = 1$. This proves $\Sigma(S) \subseteq \Lambda(S)$. □

The construction of x, y such that $x - y = z \in S$ in the proof of Lemma 2.3 can be traced back at least to [29, Sect. 7]. See also [19, Lemma 1] and [57, Theorem 5.5].

Lemma 2.4 *For $B \in \mathbb{R}^{r\times n}$ and $S \subseteq \mathbb{R}^n$, let*

$$S_B := \{x^B - y^B \mid x, y \in \mathbb{R}^n_+ \text{ and } x - y \in S^*\}. \tag{7}$$

Then, $\sigma(S_B) = \sigma(B(\Sigma(S^*)))$.

Proof For $x, y \in \mathbb{R}^n_+$, we have $\sigma(x^B - y^B) = \sigma(B(\ln x - \ln y))$ by the strict monotonicity of the logarithm, and hence

$$\sigma(S_B) = \sigma\big(\{B(\ln x - \ln y) \mid x, y \in \mathbb{R}^n_+ \text{ and } x - y \in S^*\}\big) = \sigma(B(\Lambda(S^*))),$$

using (6). By Lemma 2.3, $\sigma(S_B) = \sigma(B(\Sigma(S^*)))$. □

Proposition 2.5 *Let $B \in \mathbb{R}^{r\times n}$ and $S \subseteq \mathbb{R}^n$. Further, let $\varphi_B \colon \mathbb{R}^n_+ \to \mathbb{R}^r_+$ be the generalized monomial map $\varphi_B(x) = x^B$. The following statements are equivalent:*

(i) φ_B is injective with respect to S.
(ii) $\sigma(\ker(B)) \cap \sigma(S^*) = \emptyset$.

Proof By (7), statement (i) is equivalent to $0 \notin S_B$. By Lemma 2.4, this is in turn equivalent to $0 \notin B(\Sigma(S^*))$, that is, $\ker(B) \cap \Sigma(S^*) = \emptyset$. By (5), this is equivalent to statement (ii). □

Comparing Proposition 2.5 with Lemma 2.1, we observe that φ_B being injective with respect to S is equivalent to $\ker(B_\lambda) \cap S^* = \emptyset$, for all $\lambda \in \mathbb{R}^n_+$. In case S is a vector subspace, then φ_B is injective on the intersection $(x + S) \cap \mathbb{R}^n_+$ of any coset of S with the domain $\mathbb{R}^n_+$ if and only if B_λ is injective on S for all $\lambda \in \mathbb{R}^n_+$.

Next, we prove the equivalence of statements (inj) and (sig) in Theorem 1.4.

Proposition 2.6 *Let $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^m$ be the generalized polynomial map $f_\kappa(x) = A_\kappa\, x^B$, where $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{r\times n}$, and $\kappa \in \mathbb{R}^r_+$. Further, let $S \subseteq \mathbb{R}^n$. The following statements are equivalent:*

(inj) *f_κ is injective with respect to S, for all $\kappa \in \mathbb{R}^r_+$.*
(sig) *$\sigma(\ker(A)) \cap \sigma(B(\Sigma(S^*))) = \emptyset$.*

Proof Statement (inj) asserts that for $x, y \in \mathbb{R}^n_+$ with $x - y \in S^*$, we have $A_\kappa\,(x^B - y^B) \neq 0$ for all $\kappa \in \mathbb{R}^r_+$. This is equivalent to asserting that $\ker(A_\kappa) \cap S_B = \emptyset$ for all $\kappa \in \mathbb{R}^r_+$, with S_B as in (7). By applying Lemma 2.1 to the matrix A and the subset S_B, this is in turn equivalent to $\sigma(\ker(A)) \cap \sigma(S_B) = \emptyset$. By Lemma 2.4, $\sigma(S_B) = \sigma(B(\Sigma(S^*)))$, and the equivalence to statement (sig) is proven. □

A necessary condition for (sig) to hold is $\ker(B) \cap \Sigma(S^*) = \emptyset$ or, equivalently, $\sigma(\ker(B)) \cap \sigma(S^*) = \emptyset$. By Proposition 2.5, this corresponds to the fact that for f_κ to be injective with respect to S for all $\kappa \in \mathbb{R}^r_+$, the monomial map φ_B must be injective with respect to S.

To prove the equivalence of statements (lin) and (jac) in Theorem 1.4, we will use the following observation.

Lemma 2.7 *Let $A = (a_{ij}) \in \mathbb{R}^{m\times r}$, $B = (b_{ij}) \in \mathbb{R}^{r\times n}$, $\kappa \in \mathbb{R}^r_+$, and $\lambda \in \mathbb{R}^n_+$. Further, let $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^m$ be the generalized polynomial map $f_\kappa(x) = A_\kappa\, x^B$. Then, the sets of all Jacobian matrices $J_{f_\kappa}(x)$ and all matrices $A_\kappa B_\lambda$ coincide:*

$$\left\{J_{f_\kappa}(x) \mid \kappa \in \mathbb{R}^r_+ \text{ and } x \in \mathbb{R}^n_+\right\} = \left\{A_\kappa B_\lambda \mid \kappa \in \mathbb{R}^r_+ \text{ and } \lambda \in \mathbb{R}^n_+\right\}.$$

Proof As $f_{\kappa,i}(x) = \sum_{j=1}^r a_{ij}\,\kappa_j\, x^{b_j}$, the (i,ℓ)th entry of the Jacobian matrix of f_κ amounts to

$$J_{f_\kappa}(x)_{i,\ell} = \frac{\partial f_{\kappa,i}(x)}{\partial x_\ell} = \sum_{j=1}^r a_{ij}\,\kappa_j\, x^{b_j}\, b_{j\ell}\, x_\ell^{-1}.$$

That is,

$$J_{f_\kappa}(x) = A\ \operatorname{diag}\left(\kappa \circ x^B\right) B\ \operatorname{diag}\left(x^{-1}\right) = A_{\kappa'} B_\lambda$$

with $\kappa' = \kappa \circ x^B$ and $\lambda = x^{-1}$. Clearly, quantifying over all $\kappa \in \mathbb{R}^r_+$ and $x \in \mathbb{R}^n_+$ is equivalent to quantifying over all $\kappa' \in \mathbb{R}^r_+$ and $\lambda \in \mathbb{R}^n_+$. □

We can now combine all the results in this section in the proof of our main theorem.

Proof of Theorem 1.4 The equivalences (lin) $\Leftrightarrow$ (sig) and (inj) $\Leftrightarrow$ (sig) are shown in Propositions 2.2 and 2.6, respectively. The equivalence (jac) $\Leftrightarrow$ (lin) follows from Lemma 2.7. □

In case S is a vector subspace, the injectivity of f_κ (on cosets $x + S$) can be directly related to the injectivity of the Jacobian of f_κ (on S). This line of thought underpins the original injectivity results on chemical reaction networks due to Craciun and Feinberg [20] and their extensions. In particular, the case $S = \mathrm{im}(A)$ and $m = n$ arises in applications to chemical reaction networks, which we address in Sect. 3.

As discussed in the introduction, a direct corollary of Theorem 1.4 characterizes unrestricted injectivity, that is, the case $S = \mathbb{R}^n$. See also [53, Theorem 3.6].

Corollary 2.8 *Let $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^m$ be the generalized polynomial map $f_\kappa(x) = A_\kappa\, x^B$, where $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{r\times n}$, and $\kappa \in \mathbb{R}^r_+$. The following statements are equivalent:*

(i) *f_κ is injective, for all $\kappa \in \mathbb{R}^r_+$.*
(ii) *$\ker(B) = \{0\}$ and $\sigma(\ker(A)) \cap \sigma(\mathrm{im}(B)) = \{0\}$.*

Proof Let $S = \mathbb{R}^n$ and hence $\Sigma(S^*) = S^*$. By Theorem 1.4, statement (i) is equivalent to

$$\sigma(\ker(A)) \cap \sigma(B(S^*)) = \emptyset.$$

Clearly, the above equality does not hold if $\ker(B) \neq \{0\}$. If $\ker(B) = \{0\}$, then $B(S^*) = B(S)^* = \mathrm{im}(B)^*$. Hence, statement (i) is equivalent to $\ker(B) = \{0\}$ and $\sigma(\ker(A)) \cap \sigma(\mathrm{im}(B)^*) = \emptyset$, which is in turn equivalent to statement (ii). □

The results presented so far concern the injectivity of maps defined on the positive orthant. In fact, the domain of $f_\kappa(x) = A_\kappa\, x^B$ can be extended to include certain points on the boundary of $\mathbb{R}^n_+$, and our next result concerns this setting. Given $B = (b_{ij}) \in \mathbb{R}^{r\times n}$, let $\Omega_B \subseteq \overline{\mathbb{R}}^n_+$ be the maximal subset on which the monomial map $\varphi_B(x) = x^B$ is well defined, that is,

$$\Omega_B := \left\{ x \in \overline{\mathbb{R}}^n_+ \mid x_j \neq 0 \text{ if } b_{ij} < 0 \text{ for some } i \in [r] \right\},$$

and let $\bar{f}_\kappa$ be the extension of f_κ to Ω_B. As it was shown in the context of chemical reaction networks [32,65,73], injectivity of f_κ with respect to S precludes the existence of distinct $x, y \in \Omega_B$ in the same coset of S that have the same image under $\bar{f}_\kappa$, i.e., with $x - y \in S$ and $\bar{f}_\kappa(x) = \bar{f}_\kappa(y)$.

The technical condition in Proposition 2.9 below is satisfied if at least one of the two vectors x and y is in the positive orthant, or if both contain some zero coordinates, but no coordinate of x^B and y^B vanishes simultaneously. In particular, if f_κ is injective with respect to S for all $\kappa \in \mathbb{R}^r_+$, then a coset of S cannot contain a vector in the interior of the positive orthant and a vector on the boundary that have the same image under $\bar{f}_\kappa$.

Proposition 2.9 *Let $f_\kappa\colon \mathbb{R}^n_+ \to \mathbb{R}^m$ be a generalized polynomial map $f_\kappa(x) = A_\kappa\, x^B$, where $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{r\times n}$, and $\kappa \in \mathbb{R}^r_+$. Assume that f_κ is injective with respect to $S \subseteq \mathbb{R}^n$, for all $\kappa \in \mathbb{R}^r_+$. As above, let $\bar{f}_\kappa$ denote the extension of f_κ to Ω_B. Consider $x, y \in \Omega_B$ with $x \neq y$ and $x - y \in S$, satisfying the following condition: for any $j \in [r]$, $x^{b_j} = y^{b_j} = 0$ implies that $x_i = y_i = 0$ for all $i \in [n]$ with $b_{ji} \neq 0$. Then, $\bar{f}_\kappa(x) \neq \bar{f}_\kappa(y)$ for all $\kappa \in \mathbb{R}^r_+$.*

Proof For $\varepsilon \in \mathbb{R}_+$, we define positive vectors $x_\varepsilon, y_\varepsilon \in \mathbb{R}^n_+$ coordinate-wise as follows: $(x_\varepsilon)_i = x_i + \varepsilon$ and $(y_\varepsilon)_i = y_i + \varepsilon$ whenever $x_i y_i = 0$, and $(x_\varepsilon)_i = x_i$ and $(y_\varepsilon)_i = y_i$ otherwise. Clearly, $x_\varepsilon - y_\varepsilon = x - y \in S$. We claim that we can choose ε small enough such that

$$\sigma(x_\varepsilon^B - y_\varepsilon^B) = \sigma(x^B - y^B).$$

If $x^{b_j} \neq y^{b_j}$, then clearly $\operatorname{sign}(x_\varepsilon^{b_j} - y_\varepsilon^{b_j}) = \operatorname{sign}(x^{b_j} - y^{b_j})$ for sufficiently small ε since the map $\varepsilon \mapsto x_\varepsilon^B - y_\varepsilon^B$ is continuous. Thus, it suffices to show that $x^{b_j} = y^{b_j}$ implies $x_\varepsilon^{b_j} = y_\varepsilon^{b_j}$. In fact, we only need to consider the case when $x_\ell y_\ell = 0$ for some $\ell \in [n]$ with $b_{j\ell} \neq 0$. Then, our hypothesis implies that $x_i = y_i = 0$ for all $i \in [n]$ with $b_{ji} \neq 0$. By construction, $(x_\varepsilon)_i = (y_\varepsilon)_i = \varepsilon$ for all such i and thus $x_\varepsilon^{b_j} = y_\varepsilon^{b_j}$, as claimed.

Suppose $\bar{f}_\kappa(x) - \bar{f}_\kappa(y) = A_\kappa(x^B - y^B) = 0$ for some $\kappa \in \mathbb{R}^r_+$. Since $x^B - y^B = \lambda \circ (x_\varepsilon^B - y_\varepsilon^B)$ for some $\lambda \in \mathbb{R}^r_+$, we obtain $0 = A_\kappa(x^B - y^B) = A_{\kappa'}(x_\varepsilon^B - y_\varepsilon^B) = f_{\kappa'}(x_\varepsilon) - f_{\kappa'}(y_\varepsilon)$, where $\kappa' = \kappa \circ \lambda$. Clearly, this contradicts the hypothesis that $f_{\kappa'}$ is injective with respect to S. □

A related result concerning injectivity up to the boundary in the two-dimensional case appears in [68].

2.4 Determinantal Conditions

In this subsection, we characterize the injectivity of a family of maps on the positive orthant $f_\kappa\colon \mathbb{R}^n_+ \to \mathbb{R}^n$, $x \mapsto A_\kappa\, x^B$, with respect to $S \subseteq \mathbb{R}^n$, in the case where S is a vector subspace with $\dim(S) = \operatorname{rank}(A)$. In particular, we provide injectivity conditions in terms of determinants and signs of maximal minors.

Given a proper vector subspace $S \subseteq \mathbb{R}^n$ of dimension s, it can be presented as the image of a full-rank matrix $C \in \mathbb{R}^{n\times s}$, or as the kernel of a full-rank matrix $Z \in \mathbb{R}^{(n-s)\times n}$, whose rows are a basis of $S^\perp$. To recall the relation between the maximal minors of C and Z, we need the following notation. For $n \in \mathbb{N}$ and a subset $I = \{i_1, \ldots, i_s\} \subseteq [n]$, let $I^c = \{j_1, \ldots, j_{n-s}\}$ be the complement of I in $[n]$. For $i_1 < \cdots < i_s$ and $j_1 < \cdots < j_{n-s}$, let $\tau(I) \in \{\pm 1\}$ denote the sign of the permutation that sends $1, \ldots, n$ to $j_1, \ldots, j_{n-s}, i_1, \ldots, i_s$, respectively.

Lemma 2.10 *Let s, n be natural numbers with $0 < s < n$ and $C \in \mathbb{R}^{n\times s}$, $Z \in \mathbb{R}^{(n-s)\times n}$ full-rank matrices with $\operatorname{im}(C) = \ker(Z)$. Then, there exists a nonzero real number δ such that*

$$\delta \det(C_{I,[s]}) = (-1)^{\tau(I)} \det(Z_{[n-s],I^c}),$$

for all subsets $I \subseteq [n]$ of cardinality s.

Lemma 2.10 is well known (see for instance, [35, p. 94, Equ. (1.6)] and [35, Appendix A] on the determinant of a complex, in particular, the proofs of Lemma 5 and Proposition 11 or Theorem 12.16 in [47]). The full-rank matrices Z and C are called *Gale dual*; see Definition 3.6 below.

Let $s \leq n$. For $A' \in \mathbb{R}^{s\times r}$, $B \in \mathbb{R}^{r\times n}$, and $Z \in \mathbb{R}^{(n-s)\times n}$, let $\Gamma_{\kappa,\lambda} \in \mathbb{R}^{n\times n}$ be the square matrix given in block form as

$$\Gamma_{\kappa,\lambda} = \begin{pmatrix} Z \\ A'_\kappa B_\lambda \end{pmatrix}, \quad \text{for } \kappa \in \mathbb{R}^r_+ \text{ and } \lambda \in \mathbb{R}^n_+. \tag{8}$$

For simplicity, we do not treat the case $s = n$ separately. Instead, we use $\Gamma_{\kappa,\lambda} = A'_\kappa B_\lambda$ and $\det(Z_{[n-s],I^c}) = 1$ in the statements below for this case.

We start with two useful lemmas.

Lemma 2.11 *Let $\Gamma_{\kappa,\lambda}$ be the matrix defined in (8), for $s \leq n$, $A' \in \mathbb{R}^{s\times r}$, $B \in \mathbb{R}^{r\times n}$, $Z \in \mathbb{R}^{(n-s)\times n}$, $\kappa \in \mathbb{R}^r_+$, and $\lambda \in \mathbb{R}^n_+$. Then,*

$$\det(\Gamma_{\kappa,\lambda}) = \sum\nolimits_{I,J} (-1)^{\tau(I)} \det(Z_{[n-s],I^c}) \det(A'_{[s],J}) \det(B_{J,I}) \kappa^J \lambda^I,$$

where we sum over all subsets $I \subseteq [n]$, $J \subseteq [r]$ of cardinality s, and $\kappa^J = \prod_{j\in J} \kappa_j$, $\lambda^I = \prod_{i\in I} \lambda_i$.

Proof By Laplace expansion on the bottom s rows of $\Gamma_{\kappa,\lambda}$, we have that

$$\det(\Gamma_{\kappa,\lambda}) = \sum\nolimits_{I} (-1)^{\tau(I)} \det(Z_{[n-s],I^c}) \det((A'_\kappa B_\lambda)_{[s],I}),$$

where we sum over all subsets $I \subseteq [n]$ of cardinality s. The Cauchy–Binet formula yields

$$\begin{aligned} \det((A'_\kappa B_\lambda)_{[s],I}) &= \sum\nolimits_{J} \det((A'_\kappa)_{[s],J}) \det((B_\lambda)_{J,I}) \\ &= \sum\nolimits_{J} \det(A'_{[s],J}) \det(B_{J,I}) \kappa^J \lambda^I, \end{aligned}$$

where we sum over all subsets $J \subseteq [r]$ of cardinality s. □

Lemma 2.12 *Let $q(c) \in \mathbb{R}[c_1, \ldots, c_\ell]$ be a nonzero homogeneous polynomial, with degree at most one in each variable. There exists $c^* \in \mathbb{R}^\ell_+$ such that $q(c^*) = 0$ if and only if $q(c)$ has both positive and negative coefficients.*

Proof If all coefficients of $q(c)$ have the same sign, it is clear that $q(c^*) \neq 0$, for all $c^* \in \mathbb{R}^\ell_+$. To prove the reverse implication, let $\alpha\, c^v$ be any monomial of q (so, $v \in \{0, 1\}^\ell$). For $\epsilon \in \mathbb{R}_+$, define $c(\epsilon) \in \mathbb{R}^\ell_+$ by $c_i(\epsilon) := \epsilon$ if $v_i = 1$ and $c_i(\epsilon) := 1$

if $v_i = 0$. Then, $q(c(\epsilon))$ is a univariate polynomial in ϵ of the same degree as q and with leading coefficient α. For sufficiently large ϵ, the sign of $q(c(\epsilon))$ is the sign of α. Therefore, if two nonzero coefficients have opposite signs, $q(c)$ takes both positive and negative values, and so by continuity, there exists $c^* \in \mathbb{R}^\ell_+$ such that $q(c^*) = 0$. □

The following result generalizes [73, Propositions 5.2–5.3].

Theorem 2.13 *Let $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^m$ be the generalized polynomial map $f_\kappa(x) = A_\kappa\, x^B$, where $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{r\times n}$, and $\kappa \in \mathbb{R}^r_+$.*

Assume that $\operatorname{rank}(A) = s$, *and consider a vector subspace $S \subseteq \mathbb{R}^n$ with* $\dim(S) = s$. *Let $Z \in \mathbb{R}^{(n-s)\times n}$ and $C \in \mathbb{R}^{n\times s}$ be matrices presenting S, that is, such that* $\operatorname{im}(C) = S = \ker(Z)$. *Given $A' \in \mathbb{R}^{s\times r}$ with* $\ker(A) = \ker(A')$, *call $\widetilde{A} = CA' \in \mathbb{R}^{n\times r}$, and let $\Gamma_{\kappa,\lambda} \in \mathbb{R}^{n\times n}$ be the square matrix associated to A', B, Z, $\kappa \in \mathbb{R}^r_+$, and $\lambda \in \mathbb{R}^n_+$ as in (8).*

The following statements are equivalent:

(inj) *f_κ is injective with respect to S, for all $\kappa \in \mathbb{R}^r_+$.*
(det) *Viewed as a polynomial in κ and λ,* $\det(\Gamma_{\kappa,\lambda})$ *is nonzero and all of its nonzero coefficients have the same sign.*
(min) *For all subsets $I \subseteq [n]$, $J \subseteq [r]$ of cardinality s, the product* $\det(\widetilde{A}_{I,J})\det(B_{J,I})$ *either is zero or has the same sign as all other nonzero products, and moreover, at least one such product is nonzero.*

Proof Using the equivalence (inj) ⇔ (lin) of Theorem 1.4 and that S is the solution set to the equation $Zx = 0$, statement (inj) is equivalent to $\Gamma_{\kappa,\lambda}(x) \neq 0$ for all $\kappa \in \mathbb{R}^r_+$, $\lambda \in \mathbb{R}^n_+$, and $x \in \mathbb{R}^n$ with $x \neq 0$. As $\Gamma_{\kappa,\lambda}$ is a square matrix, this is in turn equivalent to $\det(\Gamma_{\kappa,\lambda}) \neq 0$, for all $\kappa \in \mathbb{R}^r_+$ and $\lambda \in \mathbb{R}^n_+$. By Lemma 2.11, $\det(\Gamma_{\kappa,\lambda})$ is a homogeneous polynomial in κ, λ with degree at most one in each variable. Hence, the equivalence (inj) ⇔ (det) follows from Lemma 2.12.

By Cauchy–Binet, $\det(\widetilde{A}_{I,J}) = \det(C_{I,[s]})\det(A'_{[s],J})$, and hence the equivalence (det) ⇔ (min) follows from Lemmas 2.10 and 2.11. □

Therefore, injectivity of $A_\kappa x^B$ can be assessed by computing either the nonzero products of the $s \times s$ minors of $\widetilde{A}$ and B or the determinant of the symbolic matrix $\Gamma_{\kappa,\lambda}$. Further, it follows from Theorem 2.13 that $\det(\Gamma_{\kappa,\lambda})$ equals the sum of the principal minors of size s of $\widetilde{A}_\kappa B_\lambda$. This implies the interesting fact that if $\det(\Gamma_{\kappa,\lambda})$ is nonzero, it equals the product of the nonzero eigenvalues of $\widetilde{A}_\kappa B_\lambda$.

Clearly, the hypotheses of Theorem 2.13 are fulfilled for $S = \operatorname{im}(A)$. In this case, the matrix $C \in \mathbb{R}^{n\times s}$ can be chosen to satisfy $A = CA'$. Therefore, we obtain the following corollary, which was proven in [73].

Corollary 2.14 *Let $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^n$ be the generalized polynomial map $f_\kappa(x) = A_\kappa\, x^B$, where $A \in \mathbb{R}^{n\times r}$, $B \in \mathbb{R}^{r\times n}$, and $\kappa \in \mathbb{R}^r_+$. Further, let $s = \operatorname{rank}(A)$. The following statements are equivalent:*

(inj) *f_κ is injective with respect to* $\operatorname{im}(A)$, *for all $\kappa \in \mathbb{R}^r_+$.*
(min) *For all subsets $I \subseteq [n]$, $J \subseteq [r]$ of cardinality s, the product* $\det(A_{I,J})\det(B_{J,I})$ *either is zero or has the same sign as all other nonzero such products, and moreover, at least one such product is nonzero.*

Let $A \in \mathbb{R}^{m\times r}$ and $B \in \mathbb{R}^{r\times n}$ have full rank m and n, respectively. By Corollary 2.8, the (unrestricted) injectivity of $f_\kappa(x) = A_\kappa\, x^B$ is equivalent to $\sigma(\ker(A)) \cap \sigma(\mathrm{im}(B)) = \{0\}$. For $m < n$, the intersection $\ker(A) \cap \mathrm{im}(B)$ is always nontrivial since $(r - m) + n > r$. For $m = n$, determinantal conditions are given in Corollary 2.15 below; see also [16, Theorem 3.1] and [24, Corollary 8]. For $m > n$, the problem is NP-complete; see Sect. 4.

Corollary 2.15 *Let $A \in \mathbb{R}^{n\times r}$ and $B \in \mathbb{R}^{r\times n}$ be matrices of rank n. The following statements are equivalent:*

(i) $\sigma(\ker(A)) \cap \sigma(\mathrm{im}(B)) = \{0\}$.
(ii) *For all subsets $J \subseteq [r]$ of cardinality n, the product $\det(A_{[n],J})\det(B_{J,[n]})$ either is zero or has the same sign as all other nonzero products, and moreover, at least one such product is nonzero.*

Proof The (unrestricted) injectivity of $f_\kappa(x) = A_\kappa\, x^B$ for all $\kappa \in \mathbb{R}^r_+$ is equivalent to both (i) by Corollary 2.8 (since $\ker(B) = \{0\}$) and (ii) by Corollary 2.14 (since $\mathrm{im}(A) = \mathbb{R}^n$). □

3 Applications

The first application of our results is to study steady states (or equilibria) of dynamical systems induced by generalized polynomial maps. In Sect. 3.1, we introduce such *power-law systems* and state our results in this setting. In Sect. 3.2, we give sign conditions that preclude/guarantee the existence of multiple steady states of a particular form. In Sect. 3.3, we show how our results reveal the first partial multivariate generalization of Descartes' rule of signs in real algebraic geometry and interpret our results in the language of oriented matroids.

3.1 Power-Law Systems

Power-law systems arise naturally as models of systems of interacting species, such as chemical reaction networks. Other examples include the classical Lotka–Volterra model in ecology [55] and the SIR model in epidemiology [2].

For readers unfamiliar with chemical reaction networks, we elaborate on the construction of the corresponding dynamical systems. A chemical reaction network consists of a set of n molecular species and a set of r reactions, where the left- and right-hand sides of the reactions are formal sums of species, called reactant and product complexes, respectively. A *kinetic system* describes the dynamics of the species concentrations x, where each reaction contributes to the dynamics an additive term: namely, a corresponding reaction vector (the difference between the product and reactant complexes) multiplied by a particular reaction rate (a nonnegative function of the concentrations, called kinetics). Thus, a kinetic system has the form

$$\frac{\mathrm{d}x}{\mathrm{d}t} = N K(x), \tag{9}$$

where the columns of the *stoichiometric matrix* N are the reaction vectors and the ith coordinate of $K(x)$ is the rate function of the ith reaction. The right-hand side of (9) is called the *species-formation rate function*. In power-law systems, the kinetics are given by monomials with real exponents [62]. More precisely, the *power-law system* arising from the stoichiometric matrix $N \in \mathbb{R}^{n\times r}$, a *kinetic-order matrix* $V \in \mathbb{R}^{r\times n}$, and *rate constants* $\kappa \in \mathbb{R}^r_+$ is the kinetic system (9) with kinetics $K(x) = \kappa \circ x^V$. That is, the species-formation rate function $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^n$ is given by

$$f_\kappa(x) = N_\kappa\, x^V. \tag{10}$$

In fact, the domain of f_κ may be extended to $\Omega_V \subseteq \overline{\mathbb{R}}^n_+$, the maximal subset on which the monomial map $\varphi_V : x \mapsto x^V$ is well defined. We note that, without further restrictions on the matrix V, a power-law system may exhibit physically/chemically meaningless behavior. For example, a trajectory starting in the interior may reach the boundary of the positive orthant in finite time with nonzero velocity.

The vector subspace $S = \mathrm{im}(N)$ is the *stoichiometric subspace*, and the sets $(x' + S) \cap \mathbb{R}^n_+$ for $x' \in \mathbb{R}^n_+$ are the positive *compatibility classes*. As explained in the introduction, a trajectory starting at a point $x' \in \mathbb{R}^n_+$ is confined to the coset $x' + S$. As a consequence, we study power-law systems restricted to compatibility classes. In particular, we want to characterize whether there exist distinct $x, y \in \mathbb{R}^n_+$ such that $x - y \in S$ and $f_\kappa(x) = f_\kappa(y) = 0$ for some $\kappa \in \mathbb{R}^r_+$. In our terminology, if f_κ is injective with respect to S for all $\kappa \in \mathbb{R}^r_+$, then no such x, y can exist, that is, multiple *positive* steady states cannot occur within one compatibility class for any choice of the rate constants.

Example 3.1 (Mass-action systems) *Mass-action systems* form a family of power-law systems, and they are widely used to model the dynamics of chemical reaction networks. In mass-action systems, the rate of a chemical reaction is a monomial in the concentrations of the reactant species; more precisely, the exponents of the concentrations are the corresponding *stoichiometric coefficients*, i.e., the coefficients of the species in the reactant complex. As a consequence, the kinetic-order matrix V is a nonnegative integer matrix, which encodes for each reaction the stoichiometries of the reactant species, and the map $f_\kappa(x)$ is a polynomial map in the standard sense with domain $\Omega_V = \overline{\mathbb{R}}^n_+$. Mass-action systems are at the core of the so-called *chemical reaction network theory*, initiated by Horn, Jackson, and Feinberg in the 1970s [26,43,44]; see also the surveys [27,39].

Example 3.2 (Generalized mass-action systems) The *law of mass-action*, proposed by Guldberg and Waage in the 19th century [38], refers to both the formula for chemical equilibrium, which holds for all reactions, and the formula for the reaction rate (explained in Example 3.1), which holds only for elementary reactions in homogeneous and dilute solutions. To model the dynamics of chemical reaction networks in more general environments, power-law kinetics has been considered under different formalisms [44,62]. The notion of *generalized mass-action systems* as introduced in [53,54] is a direct extension of mass-action systems, in particular, it includes the inherent structure of chemical reaction networks.

Example 3.3 (S-systems) *S-systems* form another family of power-law systems. This research area was initiated by the work of Savageau in the late 1960s [62]. In S-systems, the formation rate of each species consists of one production term and one degradation term. In other words, the components $f_{\kappa,i}(x)$ are binomial, and each row of the stoichiometric matrix N contains the entries 1 and -1, and all other entries are zero. S-systems can be used to infer gene regulatory networks, for instance, if the regulation logic is not known or the precise mechanisms are inaccessible. Further, many common kinetic systems, including (generalized) mass-action systems, can be approximated by S-systems after a process called recasting [63].

An injectivity criterion for precluding multistationarity in fully open networks with mass-action kinetics was introduced by Craciun and Feinberg [20] and has been extended in various ways [8,21,23,32,37,73]. Our contribution to this topic builds on these results and is summarized in Theorem 3.4. It is a restatement of Theorem 1.4 in the setting of power-law systems; in particular, $m = n$ and $S = \mathrm{im}(N)$ is a vector subspace. In this case, Corollary 2.14 allows us to add the condition (min). Further, the condition (inj) concerns the injectivity of the generalized polynomial map *on compatibility classes*, and (jac) addresses the injectivity of the Jacobian matrix *on the stoichiometric subspace*.

Theorem 3.4 (Theorem 1.4 for power-law systems) *Let $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^n$ be the species-formation rate function $f_\kappa(x) = N_\kappa\, x^V$ of a power-law system, where $N \in \mathbb{R}^{n\times r}$, $V \in \mathbb{R}^{r\times n}$, and $\kappa \in \mathbb{R}^r_+$. Further, let $S = \mathrm{im}(N)$ and $s = \mathrm{rank}(N)$. The following statements are equivalent:*

- *(inj)* *f_κ is injective on every compatibility class, for all $\kappa \in \mathbb{R}^r_+$.*
- *(jac)* *The Jacobian matrix $J_{f_\kappa}(x)$ is injective on the stoichiometric subspace S, for all $\kappa \in \mathbb{R}^r_+$ and $x \in \mathbb{R}^n_+$.*
- *(min)* *For all subsets $I \subseteq [n]$, $J \subseteq [r]$ of cardinality s, the product $\det(N_{I,J})\det(V_{J,I})$ either is zero or has the same sign as all other nonzero such products, and moreover, at least one such product is nonzero.*
- *(sig)* *$\sigma(\ker(N)) \cap \sigma(V(\Sigma(S^*))) = \emptyset$.*

If the conditions of Theorem 3.4 hold, then multistationarity is precluded. In the context of chemical reaction networks, the equivalence of conditions (inj), (jac), and (min) was proven in [73]. Thus, our contribution is condition (sig).

Remark 3.5 Injectivity results for generalized polynomial maps also preclude multistationarity for *strictly monotonic* kinetics [33,73], which include power-law kinetics. In the study of *concordant* networks [65], sign conditions preclude multistationarity for *weakly monotonic* kinetics. Injectivity results for differentiable maps and various classes of kinetics using P-matrices appear in [6–9,34]. P-matrices are defined by the positivity of principal minors, which is related to condition (min) in this work. Analysis of the signs of minors of Jacobian matrices with applications to counting steady states appear in [25,40,41].

3.2 Precluding/Guaranteeing Special Steady States

In this subsection, we relate results on injectivity and sign vectors occurring in the chemical reaction literature for "special" steady states, under seemingly different hypotheses. On one side, we study *complex balancing equilibria* defined for mass-action systems [26,43,44] and extended to generalized mass-action systems [53]; on the other side, we consider *toric steady states* [57]. The common feature of all these cases is that the steady states under consideration lie in a generalized variety that has dual equivalent presentations: via generalized binomial equations and via a generalized monomial parametrization. Our results give conditions for precluding multiple special steady states (Proposition 3.9) and for guaranteeing multiple special steady states (Corollary 3.11).

Given $M \in \mathbb{R}^{d' \times n}$ and $x^* \in \mathbb{R}^n_+$, we denote the corresponding fiber of $x \mapsto x^M$ by

$$Z^M_{x^*} := \left\{ x \in \mathbb{R}^n_+ \mid x^M = (x^*)^M \right\}.$$

We note that in the literature on chemical reaction network theory, the alternate formulation $Z^M_{x^*} = \{x \in \mathbb{R}^n_+ \mid \ln(x) - \ln(x^*) \in \ker(M)\}$ is used. Also, if we denote by m_i the ith row vector of M and write it as $m_i = m_i^+ - m_i^-$ with $m_i^+, m_i^- \in \overline{\mathbb{R}}^n_+$, then for any positive γ_i, the generalized monomial equation $x^{m_i} = \gamma_i$ is equivalent to the generalized binomial equation $x^{m_i^+} - \gamma_i x^{m_i^-} = 0$, when we restrict our attention to $x \in \mathbb{R}^n_+$.

Definition 3.6 Two matrices $M \in \mathbb{R}^{d' \times n}$ and $B \in \mathbb{R}^{n \times d}$ with $\mathrm{im}(B) = \ker(M)$ and $\ker(B) = \{0\}$ are called *Gale dual*.

In the usual definition of Gale duality, the matrix M is required to have full rank $d' = n - d$.

The following lemma is classic.

Lemma 3.7 *Let $M \in \mathbb{R}^{d' \times n}$ and $B \in \mathbb{R}^{n \times d}$ be Gale dual. Then, for any $x^* \in \mathbb{R}^n_+$, the fiber $Z^M_{x^*}$ can be parametrized as follows:*

$$Z^M_{x^*} = \{x^* \circ \mathrm{e}^v \mid v \in \ker(M)\} = \{x^* \circ \xi^B \mid \xi \in \mathbb{R}^d_+\}.$$

Proof We start by proving the first equality. We have $x \in Z^M_{x^*}$ if and only if $x^M = (x^*)^M$, which is equivalent to $M(\ln x - \ln x^*) = 0$. Therefore, $x \in Z^M_{x^*}$ if and only if $v := \ln x - \ln x^* \in \ker(M)$, that is, $x = x^* \circ \mathrm{e}^v$ with $v \in \ker(M)$. Now, we turn to the second equality. Since the columns of B form a basis for $\ker(M)$, we can write $v \in \ker(M)$ uniquely as $v = Bt$ for some $t \in \mathbb{R}^d$. By introducing $\xi := \mathrm{e}^t \in \mathbb{R}^d_+$, we obtain

$$(\mathrm{e}^v)_i = \mathrm{e}^{v_i} = \mathrm{e}^{\sum_j b_{ij} t_j} = \prod_j \xi_j^{b_{ij}} = \xi^{b_i} = (\xi^B)_i,$$

that is, $\mathrm{e}^v = \xi^B$, so the inclusion $\subseteq$ holds. Similarly, $\supseteq$ holds via $v := B \log \xi$. □

We consider a power-law system (10) and assume that the set of steady states contains the positive part of a generalized variety defined by generalized binomials, according to the following definition. Recall the connection between certain monomial and binomial equations explained before Definition 3.6.

Definition 3.8 Let $f_\kappa\colon \mathbb{R}^n_+ \to \mathbb{R}^n$ be the species-formation rate function $f_\kappa(x) = N_\kappa\, x^V$ of a power-law system, where $N \in \mathbb{R}^{n\times r}$, $V \in \mathbb{R}^{r\times n}$, and $\kappa \in \mathbb{R}^r_+$. Further, let $M \in \mathbb{R}^{d'\times n}$ and $\gamma\colon \mathbb{R}^r_+ \to \mathbb{R}^{d'}_+$. Consider the family of generalized varieties

$$Y_\kappa^{M,\gamma} := \left\{x \in \mathbb{R}^n_+ \mid x^M = \gamma(\kappa)\right\} \quad \text{for } \kappa \in \mathbb{R}^r_+,$$

and assume that each such generalized variety consists of steady states of the corresponding power-law system:

$$Y_\kappa^{M,\gamma} \subseteq \{x \in \mathbb{R}^n_+ \mid f_\kappa(x) = 0\} \quad \text{for all} \quad \kappa \in \mathbb{R}^r_+.$$

An element $x^* \in Y_{\kappa^*}^{M,\gamma}$ is called a *special steady state for* κ^*.

According to the definition, x^* is a special steady state for κ^* if and only if $(x^*)^M = \gamma(\kappa^*)$, or, equivalently, $Y_{\kappa^*}^{M,\gamma} = Z_{x^*}^M$. Clearly, if $\gamma(\kappa^*)$ does not belong to the image of the monomial map $\varphi_M\colon x \mapsto x^M$, then $Y_{\kappa^*}^{M,\gamma} = \emptyset$. As already mentioned, special steady states include *complex balancing equilibria* of generalized mass-action systems [53] and *toric steady states* [57]. In both cases, the relevant map γ is a rational function.

Consider N, V, M, and γ as in Definition 3.8. Let $x^* \in \mathbb{R}^n_+$ be a special steady state for κ^*. By Lemma 3.7, the corresponding set of special steady states $Y_{\kappa^*}^{M,\gamma} = Z_{x^*}^M$ can be parametrized as $\{x^* \circ \xi^B \mid \xi \in \mathbb{R}^d_+\}$, where $B \in \mathbb{R}^{n\times d}$ with $\operatorname{im}(B) = \ker(M)$ and $\ker(B) = \{0\}$. In fact, we are interested in the intersection of the set of special steady states with some compatibility class,

$$Z_{x^*}^M \cap (x' + S).$$

If the intersection is nonempty, then there exist $\xi \in \mathbb{R}^d_+$ and $u \in S$ such that

$$x^* \circ \xi^B = x' + u,$$

and multiplication by a matrix $A \in \mathbb{R}^{m\times n}$ for which $\ker(A) = S$ yields

$$A\,(x^* \circ \xi^B) = A\,x'.$$

Thus, using $\ker(B) = \{0\}$, injectivity of the generalized polynomial map $f_{x^*}\colon \mathbb{R}^d_+ \to \mathbb{R}^n_+$,

$$f_{x^*}(\xi) = A\,(x^* \circ \xi^B) = A_{x^*}\,\xi^B,$$

is equivalent to the uniqueness of special steady states in every compatibility class. Therefore, if f_{x^*} is injective for *all* $x^* \in \mathbb{R}^n_+$, as characterized in Proposition 3.9 below, then multiple special steady states are precluded for *all* rate constants. Note that Theorem 3.4 precludes multiple "general" steady states.

Proposition 3.9 *Let $M \in \mathbb{R}^{d' \times n}$ and $B \in \mathbb{R}^{n \times d}$ be Gale dual, $S \subseteq \mathbb{R}^n$ be a vector subspace, and $A \in \mathbb{R}^{m \times n}$ such that $S = \ker(A)$. The following statements are equivalent:*

(i) *The monomial map $\varphi_M : \mathbb{R}^n \to \mathbb{R}^{d'}$, $x \mapsto x^M$ is injective on $(x' + S) \cap \mathbb{R}^n_+$, for all $x' \in \mathbb{R}^n_+$.*
(ii) $\sigma(\ker(M)) \cap \sigma(S) = \{0\}$.
(iii) *The polynomial map $f_{x^*} : \mathbb{R}^d_+ \to \mathbb{R}^m$, $\xi \mapsto A_{x^*}\, \xi^B$ is injective, for all $x^* \in \mathbb{R}^n_+$.*

Proof Statement (ii) is equivalent to $\sigma(\ker(A)) \cap \sigma(\mathrm{im}(B)) = \{0\}$, by the definitions of the matrices. (i) $\Leftrightarrow$ (ii) holds by Proposition 2.5 (for a vector subspace S). (iii) $\Leftrightarrow$ (ii) holds by Corollary 2.8.

In other words, injectivity of monomial maps on cosets of a vector subspace is equivalent to injectivity of a related family of polynomial maps on the positive orthant.

Remark 3.10 Related sign conditions for injectivity appear in [31, Lemma 4.1], [19, Lemma 1], [57, Theorem 5.5], and [53, Proposition 3.1 and Theorem 3.6]. In generalized mass-action systems [53], uniqueness of *complex balancing equilibria* is guaranteed by the sign condition $\sigma(S) \cap \sigma(\tilde{S}^\perp) = \{0\}$, where $\tilde{S}$ is the kinetic-order subspace with $\tilde{S}^\perp = \ker(M) = \mathrm{im}(B)$. In the specific case of mass-action systems, the stoichiometric and kinetic-order subspaces coincide, $S = \tilde{S}$, and hence $\sigma(S) \cap \sigma(S^\perp) = \{0\}$ holds trivially. Further, in this case, if complex balancing equilibria exist, all steady states are of this form [44, Theorem 6A] and multistationarity cannot occur. The sign condition for precluding multiple *toric steady states* [57] takes the form $\sigma(\mathrm{im}(\mathcal{A}^T)) \cap \sigma(\ker(\mathcal{Z}^T)) = \{0\}$, where we use calligraphic fonts to avoid confusion with symbols in this work. The matrix $\mathcal{A}$ specifies the parametrization of Z, whereas the matrix $\mathcal{Z}$ defines the stoichiometric subspace S: $\ker(M) = \mathrm{im}(\mathcal{A}^T)$ and $S = \ker(\mathcal{Z}^T)$.

We close this subsection by considering the case when statement (ii) in Proposition 3.9 does not hold. In this case, multiple special steady states in one compatibility class are possible, provided that every $x^* \in \mathbb{R}^n_+$ is a special steady state for some κ^*.

Corollary 3.11 *Let $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^n$ be the species-formation rate function $f_\kappa(x) = N_\kappa\, x^V$ of a power-law system with stoichiometric subspace $S = \mathrm{im}(N)$, where $N \in \mathbb{R}^{n \times r}$, $V \in \mathbb{R}^{r \times n}$, and $\kappa \in \mathbb{R}^r_+$. Further, let $M \in \mathbb{R}^{d' \times n}$, $\gamma : \mathbb{R}^r_+ \to \mathbb{R}^{d'}_+$, and $Y_\kappa^{M,\gamma}$ be a set of special state states as in Definition 3.8. Assume that:*

(i) $\sigma(\ker(M)) \cap \sigma(S) \neq \{0\}$.
(ii) *For all $x \in \mathbb{R}^n_+$, there exists $\kappa \in \mathbb{R}^r_+$ such that $x \in Y_\kappa^{M,\gamma}$.*

Then, there exist $\kappa^* \in \mathbb{R}^r_+$ and distinct $x^*, y^* \in \mathbb{R}^n_+$ such that

$$x^*, y^* \in Y_{\kappa^*}^{M,\gamma} \quad \text{and} \quad x^* - y^* \in S.$$

In other words, there exist multiple special steady states in some compatibility class.

Proof Assume $\sigma(\ker(M)) \cap \sigma(S) \neq \{0\}$. By (ii) $\Leftrightarrow$ (i) in Proposition 3.9, there exist $x^*, y^* \in \mathbb{R}^n_+$ with $x^* \neq y^*$, $x^* - y^* \in S$, and $(x^*)^M = (y^*)^M$, that is, $x^*, y^* \in Z^M_{x^*}$. By assumption (ii), there exists $\kappa^* \in \mathbb{R}^r_+$ such that $x^* \in Y^{M,\gamma}_{\kappa^*}$, that is, $Z^M_{x^*} = Y^{M,\gamma}_{\kappa^*}$. Hence, $x^*, y^* \in Y^{M,\gamma}_{\kappa^*}$. □

In the case of complex balancing equilibria, the crucial assumption (ii) in Corollary 3.11 follows from weak reversibility (cf. [53, Lemma 3.3]). In the case of toric steady states, it is guaranteed by the existence of a positive toric steady state for some κ or, equivalently, by the existence of a positive vector in the kernel of N (cf. [57, Theorem 5.5]).

3.3 Solving Systems of Generalized Polynomial Equations

In this subsection, we prove the partial multivariate generalization of Descartes' rule, Theorem 1.5. The bound on the number of positive solutions in statement (bnd) is a direct consequence of Corollaries 2.8 and 2.15, and it was proved in previous works, e.g., in [24, Corollary 8]. The existence of positive solutions in statement (ex) relies on the surjectivity result in [53, Theorem 3.8]. The framework of our results is the theory of oriented matroids, which is concerned with combinatorial properties of geometric configurations.

Proposition 3.12 *Let $A \in \mathbb{R}^{m\times r}$ and $B \in \mathbb{R}^{r\times n}$ with full rank n. The following statements are equivalent:*

(i) *For all $\kappa \in \mathbb{R}^r_+$ and $y \in \mathbb{R}^m$, the system of m generalized polynomial equations in n unknowns*

$$\sum_{j=1}^{r} a_{ij}\,\kappa_j\,x_1^{b_{j1}}\cdots x_n^{b_{jn}} = y_i, \quad i = 1,\ldots,m,$$

has at most one positive real solution $x \in \mathbb{R}^n_+$.

(ii) $\sigma(\ker(A)) \cap \sigma(\mathrm{im}(B)) = \{0\}$.

Proof The left-hand side of the equation system in (i) is the image of x under the generalized polynomial map $f_\kappa : \mathbb{R}^n_+ \to \mathbb{R}^m, x \mapsto A_\kappa\, x^B$. Thus, statement (i) is equivalent to the injectivity of f_κ for all $\kappa \in \mathbb{R}^r_+$. So, by Corollary 2.8, (i) $\Leftrightarrow$ (ii). □

We can now prove the bound in the partial multivariate generalization of Descartes' rule.

Proof of (bnd) in Theorem 1.5 By Corollary 2.15, the hypothesis of (bnd) in Theorem 1.5 is equivalent to statement (ii) in Proposition 3.12 for $m = n$. The equivalent condition (i) in Proposition 3.12 implies the conclusion of Theorem 1.5, by setting $\kappa = (1,\ldots,1)^T$. □

Next, we relate our results to the theory of oriented matroids. With a vector configuration $A = (a^1,\ldots,a^r) \in \mathbb{R}^{n\times r}$ of r vectors spanning $\mathbb{R}^n$, one can associate the

following data, each of which encodes the combinatorial structure of A. On one side, the *chirotope* of A, defined by the signs of maximal minors,

$$\begin{aligned}\chi_A \colon \{1,\dots,r\}^n &\to \{-,0,+\}\\ (i_1,\dots,i_n) &\mapsto \operatorname{sign}(\det(a^{i_1},\dots,a^{i_n})),\end{aligned}$$

records for each n-tuple of vectors whether it forms a positively oriented basis of $\mathbb{R}^n$, forms a negatively oriented basis, or is not a basis. On the other side, the set of *covectors* of A,

$$\mathcal{V}^*(A) = \left\{ \left(\operatorname{sign}(t^T a^1), \dots, \operatorname{sign}(t^T a^r) \right) \in \{-,0,+\}^r \mid t \in \mathbb{R}^n \right\},$$

encodes the set of all ordered partitions of A into three parts, induced by hyperplanes through the origin. Equivalently, the covectors of A are the sign vectors of A^T,

$$\mathcal{V}^*(A) = \sigma\left(\operatorname{im}\left(A^T\right)\right),$$

since for $x = A^T t \in \mathbb{R}^r$ with $t \in \mathbb{R}^n$, we have

$$\sigma(x)_i = \operatorname{sign}(x_i) = \operatorname{sign}\left(\textstyle\sum_j (A^T)_{ij}\, t_j\right) = \operatorname{sign}\left(\textstyle\sum_j t_j a_{ji}\right) = \operatorname{sign}(t^T a^i),$$

and hence $\sigma(x) = \left(\operatorname{sign}(t^T a^1), \dots, \operatorname{sign}(t^T a^r)\right)$.

Further, the set of *vectors* of A, denoted by $\mathcal{V}(A)$, is the orthogonal complement of $\mathcal{V}^*(A)$. We note that two sign vectors $\mu, \nu \in \{-,0,+\}^n$ are *orthogonal* if $\mu_i \nu_i = 0$ for all i or if there exist i, j with $\mu_i \nu_i = +$ and $\mu_j \nu_j = -$. We have

$$\mathcal{V}(A) = \mathcal{V}^*(A)^\perp = \sigma(\operatorname{im}(A^T))^\perp = \sigma(\operatorname{im}(A^T)^\perp) = \sigma(\ker(A)),$$

where we use $\sigma(S)^\perp = \sigma(S^\perp)$ for any vector subspace $S \subseteq \mathbb{R}^n$, cf. [74, Proposition 6.8].

The *oriented matroid* of A is a combinatorial structure that can be given by any of these data (chirotopes, covectors, or vectors) and defined/characterized in terms of any of the corresponding axiom systems [13,59,74]. The proofs for the equivalences among these data/axiom systems are nontrivial. We note that χ_A and $-\chi_A$ define the same oriented matroid.

We may now express the sign condition in Proposition 3.12 in terms of oriented matroids. Clearly, $\sigma(\ker(A)) \cap \sigma(\operatorname{im}(B)) = \{0\}$ if and only if $\mathcal{V}(A) \cap \mathcal{V}^*(B^T) = \{0\}$. In other words, no nonzero vector of A is orthogonal to all vectors of B^T, or, equivalently, no nonzero covector of B^T is orthogonal to all covectors of A.

Analogously, we translate the sign conditions in statement (ex) of Theorem 1.5. Indeed, the maximal minors of A and B have the same (opposite) sign(s) if and only if $\chi_A = \pm\chi_{B^T}$, that is, if and only if A and B^T define the same oriented matroid.

The proof of statement (ex) in Theorem 1.5 combines our injectivity result, Proposition 3.12, with a surjectivity result from previous work [53, Theorem 3.8] to guarantee

the existence and uniqueness of a positive solution. In fact, (ex) restates a generalization of Birch's theorem [53, Proposition 3.9] in terms of polynomial equations.

Proof of (ex) in Theorem 1.5 Clearly, if (3) has a positive solution $x \in \mathbb{R}^n_+$, then $y \in C^\circ(A)$. Conversely, the sign conditions in (ex), together with the full rank of the matrices, imply the hypotheses of (bnd), and hence there is at most one positive solution of (3). Further, they imply that A and B^T define the same oriented matroid, and hence $\sigma(\mathrm{im}(A^T)) = \sigma(\mathrm{im}(B))$. Finally, the assumption about the row vectors of B implies the sign condition $(+, \ldots, +)^T \in \sigma(\mathrm{im}(A^T))$.

By [53, Theorem 3.8], the generalized polynomial map $f_\kappa : \mathbb{R}^n_+ \to C^\circ \subseteq \mathbb{R}^n$, $x \mapsto A_\kappa\, x^B$ is surjective for all $\kappa \in \mathbb{R}^r_+$. Clearly, the left-hand side of the equation system (3) is the image of x under the generalized polynomial map f_κ for $\kappa = (1, \ldots, 1)^T$. Hence the equation system has at least one solution $x \in \mathbb{R}^n_+$ for all $y \in C^\circ(A) \subseteq \mathbb{R}^n$. (We note that the relevant objects in [53, Theorem 3.8] are $F(\lambda) = f_\kappa(x)$ with $\lambda = \ln x$, $V = A^T$, $\tilde{V} = B$, and $c^* = \kappa$.) □

Observe that statement (ex) in Theorem 1.5 can also be stated for a fixed exponent matrix $B \in \mathbb{R}^{r\times n}$ with full rank n and row vectors lying in an open half-space. Then, for any coefficient matrix $A \in \mathbb{R}^{n\times r}$ such that A and B^T define the same oriented matroid, the equation system (3) has exactly one positive solution $x \in \mathbb{R}^n_+$, for any $y \in C^\circ(A)$. Alternatively, the hypotheses of statement (ex) can be expressed in a more symmetric way: "consider matrices $A \in \mathbb{R}^{n\times r}$ and $B \in \mathbb{R}^{r\times n}$ such that A and B^T define the same oriented matroid and the column vectors of A (or, equivalently, the row vectors of B) lie in an open half-space."

Remark 3.13 In many applications, the existence of positive solutions is guaranteed, for instance, as in item (ex) above or by a fixed-point argument, in which case the sign condition in Proposition 3.12 suffices to ensure the existence and uniqueness of positive solutions. In this setting, homotopy continuation methods can be used to obtain the solution for a given system [67]. Namely, we identify one system in the family that has a unique solution—by choosing the coefficients a_{ij} and right-hand sides y_i appropriately, we can ensure that $x = (1, \ldots, 1)$ is the unique solution—and then, we follow the unique positive solution while performing the homotopy by deforming the parameters of the solved system to those of our given system. However, this need not always work, since the followed solution can fail to remain positive along the way.

4 Algorithmic Verification of Sign Conditions

In this section, we outline how the sign condition (sig) in Theorem 1.4 can be verified algorithmically. Recall that for matrices $A \in \mathbb{R}^{m\times r}$, $B \in \mathbb{R}^{r\times n}$, and a subset $S \subseteq \mathbb{R}^n$, the condition

$$\text{(sig)} \quad \sigma(\ker(A)) \cap \sigma(B(\Sigma(S^*))) = \emptyset$$

is equivalent to the injectivity of $f_\kappa(x) = A_\kappa\, x^B$ with respect to S, for all $\kappa \in \mathbb{R}^r_+$. A characterization of condition (sig) in terms of determinants and signs of maximal

minors is given in Theorem 2.13 for the special case where S is a vector subspace with $\dim(S) = \operatorname{rank}(A)$.

We assume that the two matrices have rational entries: $A \in \mathbb{Q}^{m\times r}$ and $B \in \mathbb{Q}^{r\times n}$. As discussed in the introduction, f_κ is injective with respect to S if and only if it is injective with respect to any subset S' for which $S \subseteq S' \subseteq \Sigma(S) = \sigma^{-1}(\sigma(S))$. The subset $\Sigma(S)$ depends only on the set of nonzero sign vectors $\sigma(S)$ of S. Therefore, we assume that a set of sign vectors $T \subseteq \{-, 0, +\}^n \backslash \{0\}$ is given and discuss how to check condition (sig) for the corresponding union of (possibly lower dimensional) orthants $\sigma^{-1}(T)$, that is, whether

$$\text{(sig)} \quad \sigma(\ker(A)) \cap \sigma(B(\sigma^{-1}(T))) = \emptyset$$

holds. Clearly, (sig) holds if and only if there do not exist sign vectors $\mu \in \{-, 0, +\}^r$ and $\tau \in T$ such that

$$\mu \in \sigma(\ker(A)) \cap \sigma(B(\sigma^{-1}(\tau))),$$

or, equivalently, if for all $\mu \in \{-, 0, +\}^r$ and all $\tau \in T$ the system of linear inequalities

$$Ax = 0, \quad \sigma(x) = \sigma(By) = \mu, \quad \sigma(y) = \tau \tag{11}$$

is infeasible, that is, the system (11) has no solution $z = (x, y) \in \mathbb{R}^{r+n}$.

Linear inequalities arise from the sign equalities in (11). Some of these are strict inequalities and hence techniques from linear programming do not directly apply. However, since the inequalities are homogeneous, the set of solutions to (11) forms a convex cone. In particular, if z is a solution, then so is λz, for all $\lambda \in \mathbb{R}_+$. Therefore, we can verify the infeasibility of (11) by checking the infeasibility of the system of linear inequalities obtained by replacing the inequalities >0 and <0 by $\geq \epsilon$ and $\leq -\epsilon$, respectively, for an arbitrary $\epsilon \in \mathbb{R}_+$. In this setting, one can apply methods for exact linear programming, which makes use of Farkas' lemma to guarantee the infeasibility of linear programs by way of rational certificates; see for example [1,3,36] and the exact linear programming solver QSopt_ex [4]. An alternative is to develop and adapt exact linear programming methods for strict inequalities using Theorems of the Alternative (Transposition theorems); see for example [51,64]. Using this approach, we might need to test the infeasibility of system (11) for 3^r times the cardinality of T choices of pairs $\mu \in \{-, 0, +\}^r$ and $\tau \in T$.

To apply this approach, we need to compute the set of sign vectors $\sigma(S)$ of S. In the applications in Sect. 3, the subset S is a vector subspace. In this case, the set of sign vectors $\sigma(S)$ are the covectors of the corresponding oriented matroid. Chirotopes can be used to compute covectors with minimal support, which are called cocircuits. Covectors can be computed from cocircuits. In general, the number of covectors can be exponentially large compared with the number of cocircuits. For example, $\mathbb{R}^n$ has 3^n covectors and n cocircuits corresponding to the vectors of the standard basis. Therefore, it is reasonable to measure the complexity of enumeration algorithms as a function of input and output sizes. By this measure, an efficient polynomial-time algorithm that generates all covectors from cocircuits is discussed in [5]. Note that one

also can use chirotopes to test directly whether the oriented matroids corresponding to two vector subspaces are equal, which is the condition for existence and uniqueness of positive real solutions in item (ex) of Theorem 1.5.

For the special case of unrestricted injectivity (cf. Corollary 2.8 and Proposition 3.9), condition (sig) reduces to the condition

$$\sigma(\ker(A)) \cap \sigma(\mathrm{im}(B)) = \{0\},$$

for matrices $A \in \mathbb{R}^{m\times r}$ and $B \in \mathbb{R}^{r\times n}$, such that A has full rank m and B has full rank n. In other words, we must check whether the two vector subspaces $\ker(A)$ and $\mathrm{im}(B)$ have a common nontrivial sign vector, or, equivalently, whether the corresponding oriented matroids have a common covector. For $m = n$, this condition is characterized in Corollary 2.15 in terms of signs of products of maximal minors. For $m > n$, it is shown in [16, Theorem 5.5] that for integer matrices the problem is strongly NP-complete.

We cannot hope for a polynomial-time algorithm to verify condition (sig) in general. A software to find nonzero sign vectors in $\sigma(\ker(A)) \cap \sigma(\mathrm{im}(B))$ is described in [72]. It uses mixed linear integer programming and branch-and-bound methods for enumerating all sign vectors and has been successfully applied to establish multistationarity for models arising in Systems Biology [17]. The C++ package TOPCOM [58] efficiently computes chirotopes with rational arithmetic and generates all cocircuits. It also has an interface to the open source computer algebra system SAGE [70]. For algorithmic methods to compute sign vectors of real algebraic varieties and semialgebraic sets, we refer to [10]. The software package RAGLIB [60] can test whether a system of polynomial equations and inequalities has a real solution.

Acknowledgments This project began during the Dagstuhl Seminar on "Symbolic methods for chemical reaction networks" held in November 2012 at Schloss Dagstuhl, Germany. The authors also benefited from discussions during the AIM workshop on "Mathematical problems arising from biochemical reaction networks" held in March 2013, in Palo Alto. EF was supported by a postdoctoral grant "Beatriu de Pinós" from the Generalitat de Catalunya and the Spanish research project MTM2012-38122-C03-01. CC was supported by BMBF grant Virtual Liver (FKZ 0315744) and the research focus dynamical systems of the state Saxony-Anhalt. AS was supported by the NSF (DMS-1004380 and DMS-1312473). AD was partially supported by UBACYT 20020130100207BA, CONICET PIP 11220110100580, and ANPCyT PICT-2013-1110, Argentina. The authors thank an anonymous referee for helpful comments.

References

1. E. Althaus and D. Dumitriu, *Certifying feasibility and objective value of linear programs*, Oper. Res. Lett. **40** (2012), 292–297.
2. R. M. Anderson and R. M. May, *Infectious diseases of humans: Dynamics and control*, Oxford University Press, Oxford, 1991.
3. D. L. Applegate, W. Cook, S. Dash, and D. G. Espinoza, *Exact solutions to linear programming problems*, Oper. Res. Lett. **35** (2007), 693–699.
4. D. L. Applegate, W. Cook, S. Dash, and D. G. Espinoza, *QSopt_ex* (2009). Available online at http://www.math.uwaterloo.ca/~bico//qsopt/ex/
5. E. Babson, L. Finschi, and K. Fukuda, *Cocircuit graphs and efficient orientation reconstruction in oriented matroids*, European J. Combin. **22** (2001), 587–600.
6. M. Banaji and G. Craciun, *Graph-theoretic approaches to injectivity and multiple equilibria in systems of interacting elements*, Commun. Math. Sci. **7** (2009), 867–900.

7. M. Banaji and G. Craciun, *Graph-theoretic criteria for injectivity and unique equilibria in general chemical reaction systems*, Adv. Appl. Math. **44** (2010), 168–184.
8. M. Banaji, P. Donnell, and S. Baigent, *P matrix properties, injectivity, and stability in chemical reaction systems*, SIAM J. Appl. Math. **67** (2007), 1523–1547.
9. M. Banaji and C. Pantea, *Some results on injectivity and multistationarity in chemical reaction networks*, Available online at arXiv:1309.6771, 2013.
10. S. Basu, R. Pollack, and M. F. Roy, *Algorithms in real algebraic geometry*, second ed., Algorithms and Computation in Mathematics, vol. 10, Springer-Verlag, Berlin, 2006, Updated online version available at http://perso.univ-rennes1.fr/marie-francoise.roy/bpr-ed2-posted2.html.
11. F. Bihan and A. Dickenstein, *Descartes' rule of signs for polynomial systems supported on circuits*, Preprint, 2014.
12. M. W. Birch, *Maximum likelihood in three-way contingency tables*, J. Roy. Stat. Soc. B Met. **25** (1963), 220–233.
13. A. Björner, M. Las Vergnas, B. Sturmfels, N. White, and G. M. Ziegler, *Oriented matroids*, second ed., Encyclopedia Math. Appl., vol. 46, Cambridge University Press, Cambridge, 1999.
14. S. Boyd, S. J. Kim, L. Vandenberghe, and A. Hassibi, *A tutorial on geometric programming*, Optim. Eng. **8** (2007), 67–127.
15. R. Brualdi and B. Shader, *Matrices of sign-solvable linear systems*, Cambridge University Press, 1995.
16. S. Chaiken, *Oriented matroid pairs, theory and an electric application*, Matroid theory (Seattle, WA, 1995), Contemp. Math., vol. 197, Amer. Math. Soc., Providence, RI, 1996, pp. 313–331.
17. C. Conradi and D. Flockerzi, *Multistationarity in mass action networks with applications to ERK activation*, J. Math. Biol. **65** (2012), 107–156.
18. C. Conradi and D. Flockerzi, *Switching in mass action networks based on linear inequalities*, SIAM J. Appl. Dyn. Syst. **11** (2012), 110–134.
19. C. Conradi, D. Flockerzi, and J. Raisch, *Multistationarity in the activation of a MAPK: parametrizing the relevant region in parameter space*, Math. Biosci. **211** (2008), 105–131.
20. G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks. I. The injectivity property*, SIAM J. Appl. Math. **65** (2005), 1526–1546.
21. G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks: extensions to entrapped species models*, Systems Biology, IEE Proceedings **153** (2006), 179–186.
22. G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks. II. The species-reaction graph*, SIAM J. Appl. Math. **66** (2006), 1321–1338.
23. G. Craciun and M. Feinberg, *Multiple equilibria in complex chemical reaction networks: semiopen mass action systems*, SIAM J. Appl. Math. **70** (2010), 1859–1877.
24. G. Craciun, L. Garcia-Puente, and F. Sottile, *Some geometrical aspects of control points for toric patches*, Mathematical Methods for Curves and Surfaces (Heidelberg) (M. Dæhlen, M. S. Floater, T. Lyche, J. L. Merrien, K. Morken, and L. L. Schumaker, eds.), Lecture Notes in Computer Science, vol. 5862, Springer, 2010, pp. 111–135.
25. G. Craciun, J. W. Helton, and R. J. Williams, *Homotopy methods for counting reaction network equilibria*, Math. Biosci. **216** (2008), 140–149.
26. M. Feinberg, *Complex balancing in general kinetic systems*, Arch. Rational Mech. Anal. **49** (1972/73), 187–194.
27. M. Feinberg, *Lectures on chemical reaction networks*, Available online at http://www.crnt.osu.edu/LecturesOnReactionNetworks, (1980).
28. M. Feinberg, *Chemical reaction network structure and the stability of complex isothermal reactors–I. The deficiency zero and deficiency one theorems*, Chem. Eng. Sci. **42** (1987), 2229–2268.
29. M. Feinberg, *Chemical reaction network structure and the stability of complex isothermal reactors–II. Multiple steady states for networks of deficiency one*, Chem. Eng. Sci. **43** (1988), 1–25.
30. M. Feinberg, *The existence and uniqueness of steady states for a class of chemical reaction networks*, Arch. Rational Mech. Anal. **132** (1995), 311–370.
31. M. Feinberg, *Multiple steady states for chemical reaction networks of deficiency one*, Arch. Rational Mech. Anal. **132** (1995), 371–406.
32. E. Feliu and C. Wiuf, *Preclusion of switch behavior in reaction networks with mass-action kinetics*, Appl. Math. Comput. **219** (2012), 1449–1467.
33. E. Feliu and C. Wiuf, *A computational method to preclude multistationarity in networks of interacting species*, Bioinformatics **29** (2013), 2327–2334.

34. D. Gale and H. Nikaidô, *The Jacobian matrix and global univalence of mappings*, Math. Ann. **159** (1965), 81–93.
35. I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky, *Discriminants, resultants, and multidimensional determinants*, reprint of the 1994 ed., Boston, MA: Birkhäuser, 2008.
36. A. M. Gleixner, D. E. Steffy, and K. Wolter, *Improving the accuracy of linear programming solvers with iterative refinement*, Proceedings of the 37th International Symposium on Symbolic and Algebraic Computation (ISSAC '12) (New York, NY, USA), ACM, 2012, pp. 187–194.
37. G. Gnacadja, *A Jacobian criterion for the simultaneous injectivity on positive variables of linearly parameterized polynomials maps*, Linear Algebra Appl. **437** (2012), 612–622.
38. C. M. Guldberg and P. Waage, *Studies Concerning Affinity*, C. M. Forhandlinger: Videnskabs-Selskabet i Christiana **35** (1864).
39. J. Gunawardena, *Chemical reaction network theory for* in-silico *biologists*, Available online at http://vcp.med.harvard.edu/papers/crnt.pdf, 2003.
40. J. W. Helton, V. Katsnelson, and I. Klep, *Sign patterns for chemical reaction networks*, J. Math. Chem. **47** (2010), 403–429.
41. J. W. Helton, I. Klep, and R. Gomez, *Determinant expansions of signed matrices and of certain Jacobians*, SIAM J. Matrix Anal. Appl. **31** (2009), 732–754.
42. K. Holstein, D. Flockerzi, and C. Conradi, *Multistationarity in sequential distributed multisite phosphorylation networks*, Bull. Math. Biol. **75** (2013), 2028–2058.
43. F. Horn, *Necessary and sufficient conditions for complex balancing in chemical kinetics*, Arch. Rational Mech. Anal. **49** (1972), 172–186.
44. F. Horn and R. Jackson, *General mass action kinetics*, Arch. Ration. Mech. Anal. **47** (1972), 81–116.
45. I. Itenberg and M. F. Roy, *Multivariate Descartes' rule*, Beiträge zur Algebra und Geometrie **37** (1996), 337–346.
46. B. Joshi and A. Shiu, *Simplifying the Jacobian criterion for precluding multistationarity in chemical reaction networks*, SIAM J. Appl. Math. **72** (2012), 857–876.
47. M. Joswig and T. Theobald, *Polyhedral and algebraic methods in computational geometry*, Universitext, Springer, London, 2013.
48. A. G. Khovanskiĭ, *Fewnomials*, Translations of Mathematical Monographs, vol. 88, American Mathematical Society, Providence, RI, 1991, Translated from the Russian by Smilka Zdravkovska.
49. V. Klee, R. Ladner, and R. Manber, *Signsolvability revisited*, Linear Algebra Appl. **59** (1984), 131–157.
50. F. Kubler and K. Schmedders, *Tackling multiplicity of equilibria with Gröbner bases*, Oper. Res. **58** (2010), 1037–1050.
51. O. L. Mangasarian, *Nonlinear programming*, Classics in Applied Mathematics, vol. 10, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1994, Corrected reprint of the 1969 original.
52. M. Mincheva and G. Craciun, *Multigraph conditions for multistability, oscillations and pattern formation in biochemical reaction networks*, Proceedings of the IEEE **96** (2008), 1281–1291.
53. S. Müller and G. Regensburger, *Generalized mass action systems: Complex balancing equilibria and sign vectors of the stoichiometric and kinetic-order subspaces*, SIAM J. Appl. Math. **72** (2012), 1926–1947.
54. S. Müller and G. Regensburger, *Generalized mass-action systems and positive solutions of polynomial equations with real and symbolic exponents*, Computer Algebra in Scientific Computing (V. P. Gerdt, W. Koepf, W. M. Seiler, and E. V. Vorozhtsov, eds.), Lecture Notes in Computer Science, vol. 8660, Springer International Publishing, 2014, pp. 302–323.
55. J. D. Murray, *Mathematical biology: I. An introduction*, third ed., Interdisciplinary Applied Mathematics, vol. 17, Springer-Verlag, New York, 2002.
56. C. Pantea, H. Koeppl, and G. Craciun, *Global injectivity and multiple equilibria in uni- and bi-molecular reaction networks*, Discrete Contin. Dyn. Syst. Ser. B **17** (2012), 2153–2170.
57. M. Pérez Millán, A. Dickenstein, A. Shiu, and C. Conradi, *Chemical reaction systems with toric steady states*, Bull. Math. Biol. **74** (2012), 1027–1065.
58. J. Rambau, *TOPCOM: Triangulations of point configurations and oriented matroids*, Mathematical software (Beijing, 2002), World Sci. Publ., River Edge, NJ, 2002, pp. 330–340.
59. J. Richter-Gebert and G. M. Ziegler, *Oriented matroids*, Handbook of discrete and computational geometry, CRC, Boca Raton, FL, 1997, pp. 111–132.
60. M. Safey El Din, *RAGLib*, Available online at http://www-polsys.lip6.fr/~safey/RAGLib/, 2013.

61. I. W. Sandberg and A. N. Willson, *Existence and uniqueness of solutions for the equations of nonlinear DC networks*, SIAM J. Appl. Math. **22** (1972), 173–186.
62. M. A. Savageau, *Biochemical systems analysis. I. Some mathematical properties of the rate law for the component enzymatic reactions*, J. Theor. Biol. **25** (1969), 365–369.
63. M. A. Savageau and E. O. Voit, *Recasting nonlinear differential equations as S-systems: a canonical nonlinear form*, Math. Biosci. **87** (1987), 83–115.
64. A. Schrijver, *Theory of linear and integer programming*, Wiley-Interscience Series in Discrete Mathematics, John Wiley & Sons Ltd., Chichester, 1986, A Wiley-Interscience Publication.
65. G. Shinar and M. Feinberg, *Concordant chemical reaction networks*, Math. Biosci. **240** (2012), 92–113.
66. G. Shinar and M. Feinberg, *Concordant chemical reaction networks and the species-reaction graph*, Math. Biosci. **241** (2013), 1–23.
67. A. J. Sommese and C. W. Wampler, II, *The numerical solution of systems of polynomials arising in engineering and science*, World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2005.
68. F. Sottile and C. Zhu, *Injectivity of 2D toric Bézier patches*, Proceedings of 12th International Conference on Computer-Aided Design and Computer Graphics (Jinan, China) (R. Martin, H. Suzuki, and C. Tu, eds.), IEEE CPS, 2011, pp. 235–238.
69. F. Sottile, *Real solutions to equations from geometry*, University Lecture Series, vol. 57, American Mathematical Society, Providence, RI, 2011.
70. W. A. Stein et al., *Sage Mathematics Software*, Available online at http://www.sagemath.org, 2013.
71. D. J. Struik (ed.), *A source book in mathematics, 1200-1800*, Source Books in the History of the Sciences. Cambridge, Mass.: Harvard University Press, XIV, 427 p., 1969.
72. M. Uhr, *Structural analysis of inference problems arising in systems biology*, Ph.D. thesis, ETH Zurich, 2012.
73. C. Wiuf and E. Feliu, *Power-law kinetics and determinant criteria for the preclusion of multistationarity in networks of interacting species*, SIAM J. Appl. Dyn. Syst. **12** (2013), 1685–1721.
74. G. M. Ziegler, *Lectures on polytopes*, Springer-Verlag, New York, 1995.

Annals of Combinatorics 9 (2005) 223-243
0218-0006/05/020223-21
DOI 10.1007/s00026-005-0253-7

Annals of Combinatorics

Symbolic Computation for Moments and Filter Coefficients of Scaling Functions

Georg Regensburger[1] and Otmar Scherzer[2]

[1]Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria
georg.regensburger@oeaw.ac.at

[2]Department of Computer Science, University of Innsbruck, Techniker Str. 25, A-6020 Innsbruck Austria
otmar.scherzer@uibk.ac.at

Received December 15, 2003

AMS Subject Classification: 42C40, 65T60, 13P10, 94A12, 05A10, 33C45

Abstract. Algebraic relations between discrete and continuous moments of scaling functions are investigated based on the construction of Bell polynomials. We introduce families of scaling functions which are parametrized by moments. Filter coefficients of scaling functions and wavelets are computed with computer algebra methods (in particular Gröbner bases) using relations between moments. Moreover, we propose a novel concept for data compression based on parametrized wavelets.

Keywords: scaling functions, moments, Bell polynomials, wavelets, Gröbner bases, data compression

1. Introduction

Discrete (real or complex) *filter coefficients* $\{h_k : k = 0, \dots, N\}$ in the *dilation equation* of a *scaling function*

$$\phi(x) = \sum_{k=0}^{N} h_k \phi(2x-k), \tag{1.1}$$

are used in many areas of applications, for instance *data compression*; scaling functions (and in turn filter coefficients) are the basis for constructing *wavelets* (see e.g. Daubechies [12, 13], Mallat [20], Strang & Nguyen [27]) and also play a fundamental role in subdivision schemes (see e.g. Cavaretta et al. [7] and Rioul [22]).

Filter coefficients are determined by the *continuous moments*

$$M_n = \int x^n \phi(x)\, dx,$$

and *discrete moments*

$$m_n = \sum_{k=0}^{N} h_k k^n,$$

respectively.

In Section 2 we study *algebraic* relations between discrete and continuous moments – in contrast to the literature where recursive relations have been established (see e.g. Strang & Nguyen [27]). In particular we express the nth continuous moment as a polynomial of the first n discrete moments and vice versa. The polynomials are related to *Bell polynomials*. The definition of Bell polynomials as well as some of their elementary properties are stated in the Appendix.

We recall and extend polynomial relations between moments of scaling functions associated with orthogonal wavelets in Section 3. In wavelet theory smoothness and the approximation order of scaling functions are related to vanishing moments conditions for wavelets (see Daubechies [12, 13], Strang & Nguyen [27] or Unser & Blu [29] for a recent survey). The study of *parametrized* scaling functions and wavelets is along the lines of Daubechies [14] who showed that more symmetry and regularity for scaling functions associated with wavelets can be achieved by using additional degrees of freedom obtained by giving up some higher order vanishing moment conditions in the constitutive equations. We compute analytical expressions of parametrized scaling function families using symbolic computation methods (in particular Gröbner bases) (cf. Section 3). In our work we use parametrization with respect to the discrete and continuous moments. In Subsection 3 we propose a novel concept of data compression using parametrized scaling functions and wavelets. For compression, the coefficients of the wavelet expansion of the data are computed for a series of parameters. The parameter yielding optimal compression rates is selected. The compressed data consists of the coefficients and the single parameter. These data are sufficient for decoding. A numerical example illustrating the compression idea is presented.

2. Continuous and Discrete Moments

We recall a well-known recursive relation between discrete and continuous moments (see for example Strang & Nguyen [27, p. 396]).

Lemma 2.1. *Let ϕ be a scaling function satisfying $M_0 = \int \phi = 1$. Then $m_0 = 2$ and*

$$\begin{aligned} M_n &= \frac{1}{2^{n+1}-2} \sum_{i=1}^{n} \binom{n}{i} m_i M_{n-i}, \\ m_n &= \left(2^{n+1}-2\right) M_n - \sum_{i=1}^{n-1} \binom{n}{i} m_i M_{n-i}, \quad \text{for } n = 1, 2 \ldots . \end{aligned} \tag{2.1}$$

In particular, for $n = 1, \ldots, 4$, subsequent application of (2.1) shows that

$$M_1 = \frac{1}{2} m_1,$$

$$M_2 = \frac{1}{6} m_1^2 + \frac{1}{6} m_2,$$

$$M_3 = \frac{1}{28} m_1^3 + \frac{1}{7} m_1 m_2 + \frac{1}{14} m_3,$$

$$M_4 = \frac{1}{210} m_1^4 + \frac{11}{210} m_1^2 m_2 + \frac{8}{105} m_1 m_3 + \frac{1}{30} m_2^2 + \frac{1}{30} m_4,$$

and

$$m_1 = 2M_1,$$

$$m_2 = -4M_1^2 + 6M_2,$$

$$m_3 = 12M_1^3 - 24M_1 M_2 + 14M_3,$$

$$m_4 = -48M_1^4 + 120M_1^2 M_2 - 64M_1 M_3 - 36M_2^2 + 30M_4.$$

These examples indicate that the continuous moments can be expressed as polynomials with rational coefficients in the discrete moment variables. The discrete moments are polynomials with integer coefficients in the continuous moments variables. In this section we derive the algebraic structure of these polynomials. To this end we set $p_0 = 1$ and define recursively the polynomials

$$p_n := p_n(x_1, \ldots, x_n) = \frac{1}{2^{n+1} - 2} \sum_{i=1}^{n} \binom{n}{i} x_i p_{n-i} \in \mathbb{Q}[x_1, \ldots, x_n], \tag{2.2}$$

and

$$q_n := q_n(x_1, \ldots, x_n) = \left(2^{n+1} - 2\right) x_n - \sum_{i=1}^{n-1} \binom{n}{i} x_{n-i} q_i \in \mathbb{Z}[x_1, \ldots, x_n]. \tag{2.3}$$

By induction it can easily be shown that p_n and q_n are weighted homogeneous of degree n with $\deg x_i = i$.* From Lemma 2.1 we see that

$$M_n = p_n(m_1, \ldots, m_n) \text{ and } m_n = q_n(M_1, \ldots, M_n).$$

In the following two subsections we analyze the polynomials p_n and q_n. We solve the recurrence Equations (2.2) and (2.3) by giving explicit formulas for the polynomials. Apart from the theoretical interest these formulas allow us to calculate nth polynomial without knowing the previous ones.

* Let d_i, $i = 1, \ldots, n$, be positive integers. The *weighted degree* of a monomial $x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ is $\sum_{i=1}^n \alpha_i d_i$. We refer to d_i as the *weight* (or *degree*) of x_i and write $\deg x_i = d_i$. A polynomial is called *weighted homogeneous* if all of its monomials have the same weighted degree.

2.1. Continuous ⇒ Discrete Moments

We derive formulas for the polynomials q_n, defined in (2.3). For this purpose we use linear combinations of partial Bell polynomials $B_{n,k}$ (see Definition 4.1). We define

$$Q_{n,k} = (-1)^k k! B_{n,k}, \quad \text{for } n,k \in \mathbb{N}.$$

From (4.2) in the Appendix we see that

$$\begin{aligned} Q_{n,k} :=& Q_{n,k}(x_1,\dots,x_{n-k+1}) \\ &= \sum_{\substack{i_1+\cdots+i_k=n \\ i_j>0}} (-1)^k \binom{n}{i_1,\dots,i_k} x_{i_1}\cdots x_{i_k} \in \mathbb{Z}[x_1,\dots,x_{n-k+1}], \end{aligned} \tag{2.4}$$

where

$$\binom{n}{i_1,\dots,i_k} = \frac{n!}{i_1!\cdots i_k!}.$$

Note that

$$\binom{n}{i_1,\dots,i_k} = \binom{n}{i_1}\binom{n-i_1}{i_2,\dots,i_k}. \tag{2.5}$$

We define

$$Q_{0,0} = 1, \quad Q_{n,0} = Q_{0,k} = 0, \quad \text{for } n,k \in \mathbb{N},$$

and

$$Q_n := Q_n(x_1,\dots,x_n) = \sum_{k=0}^{n} Q_{n,k} \in \mathbb{Z}[x_1,\dots,x_n], \quad \text{for } n \in \mathbb{N}_0. \tag{2.6}$$

The first polynomials Q_n are:

$$\begin{aligned} Q_0 &= 1, \\ Q_1 &= -x_1, \\ Q_2 &= 2x_1^2 - x_2, \\ Q_3 &= -6x_1^3 + 6x_1x_2 - x_3, \\ Q_4 &= 24x_1^4 - 36x_1^2x_2 + 6x_2^2 + 8x_1x_3 - x_4. \end{aligned}$$

The polynomials Q_n fit in the class of potential polynomials (see Comtet [10, p. 141]). The following lemma provides recurrence relations for the polynomials $Q_{n,k}$ and Q_n.

Lemma 2.2. *Let $n \in \mathbb{N}$. The polynomials $Q_{n,k}$ and Q_n satisfy*

(1) *For $1 \le k \le n$*

$$Q_{n,k} = -\sum_{i=1}^{n} \binom{n}{i} x_i Q_{n-i,k-1}. \tag{2.7}$$

(2)

$$Q_n = -\sum_{i=1}^{n} \binom{n}{i} x_i Q_{n-i} = -\sum_{i=0}^{n-1} \binom{n}{i} x_{n-i} Q_i. \tag{2.8}$$

Proof. Let $n \in \mathbb{N}$ and $2 \le k \le n$. From (2.4) and (2.5) it follows that

$$Q_{n,k} = -\sum_{i=1}^{n-1} \binom{n}{i} x_i Q_{n-i,k-1}.$$

This together with $Q_{0,k-1} = 0$ gives the first assertion for $2 \le k \le n$.

Let $k = 1$. Since $Q_{0,0} = 1$ and $Q_{n-i,0} = 0$ for $i = 1, \dots, n-1$, it follows that

$$Q_{n,1} = -x_n = -\sum_{i=1}^{n} \binom{n}{i} x_i Q_{n-i,0}.$$

Moreover, since $Q_{n,0} = 0$, we have

$$Q_n = \sum_{k=0}^{n} Q_{n,k} = \sum_{k=1}^{n} Q_{n,k}.$$

Using (2.7) this implies that

$$Q_n = -\sum_{k=1}^{n}\sum_{i=1}^{n} \binom{n}{i} x_i Q_{n-i,k-1} = -\sum_{i=1}^{n} \binom{n}{i} x_i \sum_{k=1}^{n} Q_{n-i,k-1}.$$

Since $Q_{n,k} = 0$ for $k > n$, we have $Q_{n-i,k-1} = 0$ for $k - 1 > n - i$. Therefore

$$\sum_{k=1}^{n} Q_{n-i,k-1} = \sum_{k=1}^{n-i+1} Q_{n-i,k-1} = Q_{n-i}.$$

The second assertion follows from the last two equations. ∎

The following theorem gives an explicit formula for the polynomials q_n in terms of the known polynomials Q_n.

Theorem 2.3. *For $n \in \mathbb{N}$*

$$q_n = \sum_{i=1}^{n} (2^{i+1} - 2) \binom{n}{i} x_i Q_{n-i}.$$

Proof. For $n = 1$ the assertion is true since $q_1 = 2x_1$ by (2.3) and

$$2\binom{1}{1} x_1 Q_0 = 2x_1.$$

Let

$$\tilde{q}_n = \sum_{i=1}^{n} (2^{i+1} - 2) \binom{n}{i} x_i Q_{n-i}, \quad n = 1, 2, \dots. \tag{2.9}$$

We prove that q_n and $\tilde{q}_n$ both satisfy the recurrence (2.3), which then implies that they are identical. To this end we show that

$$\begin{aligned} \tilde{q}_n &= \left(2^{n+1} - 2\right) x_n + \sum_{i=1}^{n-1} (2^{i+1} - 2) \binom{n}{i} x_i Q_{n-i} \\ &= \left(2^{n+1} - 2\right) x_n - \sum_{i=1}^{n-1} \binom{n}{i} \tilde{q}_i x_{n-i}, \quad \text{for } n \in \mathbb{N}. \end{aligned} \tag{2.10}$$

From (2.9) it follows that

$$-\sum_{i=1}^{n-1}\binom{n}{i}\tilde{q}_i x_{n-i} = -\sum_{i=1}^{n-1}\binom{n}{i}\left(\sum_{j=1}^{i}(2^{j+1}-2)\binom{i}{j}x_j Q_{i-j}\right)x_{n-i}.$$

Using the binomial identity

$$\binom{n}{i}\binom{i}{j} = \binom{n}{j}\binom{n-j}{i-j} \tag{2.11}$$

and interchanging the order of summation gives

$$-\sum_{i=1}^{n-1}\binom{n}{i}\tilde{q}_i x_{n-i} = -\sum_{j=1}^{n-1}\binom{n}{j}(2^{j+1}-2)x_j\left(\sum_{i=j}^{n-1}\binom{n-j}{i-j}Q_{i-j}x_{n-i}\right). \tag{2.12}$$

From (2.8) it follows that

$$\sum_{i=j}^{n-1}\binom{n-j}{i-j}Q_{i-j}x_{n-i} = \sum_{i=0}^{n-1-j}\binom{n-j}{i}Q_i x_{n-j-i} = -Q_{n-j}.$$

Using this identity in (2.12) yields

$$-\sum_{i=1}^{n-1}\binom{n}{i}\tilde{q}_i x_{n-i} = \sum_{i=1}^{n-1}\binom{n}{i}(2^{i+1}-2)x_i Q_{n-i}$$

and the assertion (2.10) is proved. ∎

The first polynomials q_n are:

$$\begin{aligned}
q_1 &= 2x_1,\\
q_2 &= -4x_1^2 + 6x_2,\\
q_3 &= 12{x_1}^3 - 24x_1x_2 + 14x_3,\\
q_4 &= -48{x_1}^4 + 120{x_1}^2x_2 - 64x_1x_3 - 36{x_2}^2 + 30x_4,\\
q_5 &= 240{x_1}^5 - 720{x_1}^3x_2 + 360{x_1}^2x_3 + 420x_1{x_2}^2 - 160x_1x_4 - 200x_2x_3 + 62x_5,\\
q_6 &= -1440x_1^6 + 5040x_1^4x_2 - 2400x_1^3x_3 - 4320x_1^2x_2^2 + 1020x_1^2x_4 + 2640x_1x_2x_3\\
&\quad + 540x_2^3 - 384x_1x_5 - 540x_2x_4 - 280x_3^2 + 126x_6.
\end{aligned}$$

2.2. Discrete ⇒ Continuous Moments

In this section we further analyze the polynomials p_n, defined in (2.2). We give explicit formulas for the polynomials as a sum over *compositions* (for a definition of compositions we refer to Definition 4.2).

Let $k, n \in \mathbb{N}$. We define

$$p_{n,k} := p_{n,k}(x_1, \ldots, x_n) = \sum_{\substack{i_1+\cdots+i_k=n \\ i_j>0}} c^n_{i_1\cdots i_k} \binom{n}{i_1, \ldots, i_k} x_{i_1} \cdots x_{i_k} \in \mathbb{Q}[x_1, \ldots, x_{n-k+1}], \tag{2.13}$$

with

$$c^n_{i_1\cdots i_k} = \frac{1}{(2^{n+1}-2)(2^{n+1-i_1}-2)\cdots\left(2^{n+1-i_1-\cdots-i_{k-1}}-2\right)}.$$

We define $p_{0,0} = 1$ and $p_{n,0} = p_{0,k} = 0$.

We note that

$$p_{n,k} = 0, \quad \text{for } k > n. \tag{2.14}$$

The sum (2.13) is over all compositions of n in k parts. We recall the analogy with the constitutive equations for the partial Bell polynomials (cf. (4.2)). However, here in contrast to Bell polynomials, the coefficients $c^n_{i_1\cdots i_k}$ depend on the particular order of the numbers $i_1, \ldots, i_k$.

In the following theorem we establish recurrence relations for the polynomials $p_{n,k}$ and give a formula for p_n.

Theorem 2.4. *The polynomials $p_{n,k}$ and p_n satisfy*

(1) *For $n \in \mathbb{N}$ and $1 \le k \le n$*

$$p_{n,k} = \frac{1}{(2^{n+1}-2)} \sum_{i=1}^{n} \binom{n}{i} x_i p_{n-i,k-1}. \tag{2.15}$$

(2) *For $n \in \mathbb{N}_0$*

$$p_n = \sum_{k=0}^{n} p_{n,k}.$$

Proof. Let $n \in \mathbb{N}$ and $2 \le k \le n$. The relation

$$c^n_{i_1\cdots i_k} = \frac{1}{(2^{n+1}-2)} c^{n-i_1}_{i_2\cdots i_k}$$

and the binomial identity (2.5) with (2.13) show that

$$p_{n,k} = \frac{1}{(2^{n+1}-2)} \sum_{i=1}^{n-1} \binom{n}{i} x_i p_{n-i,k-1}.$$

This together with $p_{0,k-1} = 0$ gives the first assertion for $2 \le k \le n$. For $k = 1$ it follows from $p_{0,0} = 1$ and $p_{n,0} = 0$ that

$$p_{n,1} = \frac{1}{(2^{n+1}-2)} x_n = \frac{1}{(2^{n+1}-2)} \sum_{i=1}^{n} \binom{n}{i} x_i p_{n-i,0}.$$

For the second claim we define

$$\tilde{p}_n = \sum_{k=0}^{n} p_{n,k}, \quad n = 0, 1, \ldots.$$

Since $p_0 = \tilde{p}_0 = 1$, it is sufficient to prove that $\tilde{p}_n$ and p_n both satisfy the recurrence Equation (2.2), that is, it suffices to show that

$$\tilde{p}_n = \frac{1}{2^{n+1}-2} \sum_{i=1}^{n} \binom{n}{i} x_i \tilde{p}_{n-i}, \quad \text{for } n \in \mathbb{N}.$$

Since $p_{n,0} = 0$ it follows from (2.15) that

$$\tilde{p}_n = \sum_{k=1}^{n} p_{n,k} = \frac{1}{(2^{n+1}-2)} \sum_{i=1}^{n} \binom{n}{i} x_i \sum_{k=1}^{n} p_{n-i,k-1}.$$

From (2.14) it follows that

$$\sum_{k=1}^{n} p_{n-i,k-1} = \sum_{k=1}^{n-i+1} p_{n-i,k-1} = \sum_{k=0}^{n-i} p_{n-i,k} = \tilde{p}_{n-i}.$$

This shows the assertion. ∎

The first polynomials p_n are:

$$p_0 = 1,$$

$$p_1 = \frac{1}{2} x_1,$$

$$p_2 = \frac{1}{6} x_1^2 + \frac{1}{6} x_2,$$

$$p_3 = \frac{1}{28} x_1^3 + \frac{1}{7} x_1 x_2 + \frac{1}{14} x_3,$$

$$p_4 = \frac{1}{210} x_1^4 + \frac{11}{210} x_1^2 x_2 + \frac{8}{105} x_1 x_3 + \frac{1}{30} x_2^2 + \frac{1}{30} x_4,$$

$$p_5 = \frac{1}{2604} x_1^5 + \frac{13}{1302} x_1^3 x_2 + \frac{43}{1302} x_1^2 x_3 + \frac{67}{2604} x_1 x_2^2 + \frac{4}{93} x_1 x_4 + \frac{25}{651} x_2 x_3 + \frac{1}{62} x_5.$$

Daubechies & Lagarias [15, 16] consider scaling functions satisfying a dilation equation of the form

$$\phi(x) = \sum_{k=0}^{N} h_k \phi(\alpha x - \beta_k), \tag{2.16}$$

with real numbers $\alpha > 1$ and $\beta_0 < \beta_1 < \cdots < \beta_N$. The results stated so far can easily be modified to this more general situation.

We define the nth discrete moment by

$$m_n = \sum_{k=0}^{N} h_k \beta_k^n.$$

Again we assume that $M_0 = \int \varphi = 1$. Then (2.16) implies $m_0 = \alpha$ and

$$M_n = \frac{1}{\alpha^{n+1} - \alpha} \sum_{i=1}^{n} \binom{n}{i} m_i M_{n-i},$$

$$m_n = \left(\alpha^{n+1} - \alpha\right) M_n - \sum_{i=1}^{n-1} \binom{n}{i} m_i M_{n-i}, \quad \text{for } n = 1, 2, \ldots.$$

In this case the polynomials p_n and q_n are defined recursively by $p_0^\alpha = 1$ and

$$p_n^\alpha = \frac{1}{\alpha^{n+1} - \alpha} \sum_{i=1}^{n} \binom{n}{i} x_i p_{n-i}^\alpha,$$

$$q_n^\alpha = \left(\alpha^{n+1} - \alpha\right) x_n - \sum_{i=1}^{n-1} \binom{n}{i} x_{n-i} q_i^\alpha.$$

Replacing the dilation factor 2 by α in the proof of Theorem 2.3 we see that

$$q_n^\alpha = \sum_{i=1}^{n} (\alpha^{i+1} - \alpha) \binom{n}{i} x_i Q_{n-i}, \quad \text{for } n \in \mathbb{N},$$

with Q_n defined as in (2.6). The generalization of Theorem 2.4 reads as follows

$$p_n^\alpha = \sum_{k=0}^{n} p_{n,k}^\alpha, \quad \text{for } n \in \mathbb{N}_0,$$

where

$$p_{n,k}^\alpha = \sum_{\substack{i_1 + \cdots + i_k = n \\ i_j > 0}} c_{i_1 \cdots i_k}^n \binom{n}{i_1, \ldots, i_k} x_{i_1} \cdots x_{i_k} \in \mathbb{Q}[x_1, \ldots, x_{n-k+1}]$$

and

$$c_{i_1 \cdots i_k}^n = \frac{1}{(\alpha^{n+1} - \alpha)(\alpha^{n+1-i_1} - \alpha) \cdots \left(\alpha^{n+1-i_1 - \cdots - i_{k-1}} - \alpha\right)}.$$

3. Moments and Filter Coefficients

The goal of this section is twofold. In the first subsection we recall and extend polynomial relations between moments of scaling functions associated with orthonormal wavelets. The second subsection is devoted to a study of *parametrized wavelets*. We give an application of parametrized wavelets to compression in the last subsection.

Daubechies [14] showed that more symmetry and regularity for scaling functions associated with wavelets can be obtained with the same number of filter coefficients, by neglecting some higher order vanishing moment conditions. She calculated a parametrized family of wavelets with four filter coefficients [12]. In the case of more than four filter coefficients, she replaced various vanishing moment conditions for the associated wavelet and computed solutions of the resulting system for the filter coefficients numerically. In our work we introduce moments of the scaling function as parameters and

give up one vanishing moment condition. The resulting algebraic system for the filter coefficients is solved with symbolic methods, in particular by using Gröbner bases. Our approach differs essentially from other work on symbolic computation of wavelets coefficients which aim to calculate a finite number of solutions. Here we calculate *infinitely* many solutions, which are due to the additional degree of freedom imposed by neglecting a vanishing moment condition. Applications of Gröbner bases to the design of wavelets and digital filters are for example described in Chyzak et al. [8], Lebrun & Selesnick [18], Lebrun & Vetterli [19] and Selesnick & Burrus [25]. Gröbner bases were introduced by Buchberger in [4] and [5]. For an introduction see for example Cox et al. [11]. In Subsection 3 we use the parametrized scaling functions and wavelets for data compression. The idea is to find the optimal parameter before storage or transmission.

3.1. Moments of Wavelets

In orthonormal wavelet theory scaling functions ϕ are considered with the additional property that their integer translates $\{\phi(x-k)\}_{k\in\mathbb{Z}}$ are orthonormal in $L^2(\mathbb{R})$. A *wavelet* function ψ is associated with ϕ via

$$\psi(x) = \sum_{k=0}^{N} (-1)^k h_{N-k} \phi(2x-k). \tag{3.1}$$

We denote by

$$N_n = \langle x^n, \psi(x) \rangle = \int x^n \psi(x)\, dx$$

the nth continuous moment of the wavelet.

In the sequel we discuss the relation between moments of orthonormal scaling functions ϕ and associated wavelets ψ. Gopinath & Burrus [17] and Sweldens & Piessens [28] established a relation between the first two moments of orthonormal scaling functions and wavelets. This result is generalized to an arbitrary number of moments in Theorem 3.4.

Theorem 3.1. [17, 28] *If $N_0 = N_1 = 0$, then*

$$M_2 = M_1^2.$$

To establish higher order moment identities we make use of the following two lemmas:

Lemma 3.2. *Let the first p moments of ψ vanish, that is*

$$N_j = 0, \quad \text{for } j = 0, \ldots, p-1.$$

Then

$$\sum_k (x-k)^n \phi(x-k) = M_n, \quad \text{for } 0 \le n \le p-1. \tag{3.2}$$

For a proof of this result we refer to Sweldens & Piessens [28].

Lemma 3.3. *Let the first p moments of* ψ *vanish. Then*

$$\sum_k k^n \phi(x-k) = \sum_{i=0}^{n} \binom{n}{i} (-1)^i x^{n-i} M_i, \quad \text{for } 0 \le n \le p-1. \tag{3.3}$$

Proof. The proof is done by induction. For $p = 1$ the assertion follows from the previous Lemma. Suppose that the assertion is true for p and we assume that $N_j = 0$ for $j = 0, \ldots, p$. We have to show that

$$\sum_k k^p \phi(x-k) = \sum_{i=0}^{p} \binom{p}{i} (-1)^i x^{p-i} M_i.$$

Again from the previous Lemma we know that

$$\sum_k (x-k)^p \phi(x-k) = M_p.$$

Expanding $(x-k)^p$ yields

$$\sum_{j=0}^{p-1} \binom{p}{j} (-1)^j x^{p-j} \sum_k k^j \phi(x-k) + (-1)^p \sum_k k^p \phi(x-k) = M_p.$$

From this equation and the induction hypothesis it follows that

$$(-1)^{p+1} \sum_k k^p \phi(x-k) + M_p = \sum_{j=0}^{p-1} \binom{p}{j} (-1)^j x^{p-j} \sum_{i=0}^{j} \binom{j}{i} (-1)^i x^{j-i} M_i. \tag{3.4}$$

Interchanging the order of summation and using the identities

$$\binom{p}{j}\binom{j}{i} = \binom{p}{i}\binom{p-i}{j-i}$$

and

$$\sum_{j=i}^{p-1} (-1)^j \binom{p-i}{j-i} = (-1)^{p+1}$$

in (3.4) shows that

$$\begin{aligned}
(-1)^{p+1} \sum_k k^p \phi(x-k) + M_p &= \sum_{i=0}^{p-1} (-1)^i x^{p-i} M_i \sum_{j=i}^{p-1} \binom{p}{j}\binom{j}{i} (-1)^j \\
&= \sum_{i=0}^{p-1} \binom{p}{i} (-1)^i x^{p-i} M_i \left(\sum_{j=i}^{p-1} (-1)^j \binom{p-i}{j-i} \right) \\
&= (-1)^{p+1} \sum_{i=0}^{p-1} \binom{p}{i} (-1)^i x^{p-i} M_i.
\end{aligned}$$

This shows the assertion. ∎

Using the last lemma we are able to prove a relation between higher order continuous moments of orthonormal scaling functions. This result generalizes Theorem 3.1.

Theorem 3.4. *Let $\phi \in L^2(\mathbb{R})$ be a scaling function with the additional property that its integer translates $\{\phi(x-k)\}_{k\in\mathbb{Z}}$ are orthonormal. Let $p \in \mathbb{N}$ be odd and let the first $p+1$ moments of the associated wavelet ψ vanish. Then*

$$M_{p+1} = \sum_{i=1}^{p} (-1)^{i+1} \binom{p}{i} M_i M_{p-i+1}. \tag{3.5}$$

Proof. Let

$$s_k = \langle x, \phi(x)\phi(x-k)\rangle, \quad \text{for } k \in \mathbb{Z}.$$

Since $\phi(x)$ and $\phi(x-k)$ are orthogonal it follows that

$$s_{-k} = \langle x, \phi(x)\phi(x+k)\rangle = \langle x-k, \phi(x-k)\phi(x)\rangle = s_k, \quad \text{for } k \in \mathbb{Z}.$$

Therefore we get using the assumption that p is odd

$$0 = \sum_k k^p s_k = \langle x, \phi(x) \sum_k k^p \phi(x-k)\rangle.$$

This identity together with (3.3) shows that

$$0 = \sum_{i=0}^{p} \binom{p}{i} (-1)^i M_i \langle x, x^{p-i}\phi(x)\rangle = \sum_{i=0}^{p} \binom{p}{i} (-1)^i M_i M_{p-i+1}$$

and the proposition follows. ∎

Remark 3.5. The above theorem reveals that the even moments of an orthonormal scaling function are completely determined by the odd up to the number of vanishing moments of the associated wavelet. We exemplarily give the equations for the even moments using (3.5) for $p = 1, 3, 5$:

$$M_2 = M_1^2,$$

$$M_4 = -3M_2^2 + 4M_1M_3 = -3M_1^4 + 4M_1M_3,$$

$$M_6 = 10M_3^2 + 6M_1M_5 - 15M_2M_4 = 45M_1^6 - 60M_1^3M_3 + 6M_1M_5 + 10M_3^2.$$

Using the relations between continuous and discrete moments from the previous sections, in particular the polynomials p_n, we obtain from the above equations the following equations for the even discrete moments. We use this observation for the construction of parametrized families of scaling functions in the following subsection.

$$m_2 = \frac{1}{2}m_1^2,$$

$$m_4 = -\frac{1}{2}m_1^4 + 2m_1^2m_2 + 2m_1m_3 - \frac{7}{2}m_2^2 = -\frac{3}{8}m_1^4 + 2m_1m_3,$$

$$m_6 = \frac{45}{32}m_1^6 - \frac{15}{2}m_1^3m_3 + 3m_1m_5 + 5m_3^2.$$

From the following lemma we obtain a different formulation of the previous theorem.

Lemma 3.6. *Let $n \in \mathbb{N}$ be odd. Let $x_1, \ldots, x_{n+1}$ be variables and $x_0 = 1$. Then*

$$\sum_{i=0}^{n} \binom{n}{i} (-1)^i x_i x_{n-i} = 0 \tag{3.6}$$

and

$$\sum_{i=0}^{n+1} \binom{n+1}{i} (-1)^i x_i x_{n+1-i} = 2x_{n+1} + 2\sum_{i=1}^{n} \binom{n}{i} (-1)^i x_i x_{n+1-i}. \tag{3.7}$$

Proof. Since n is odd the first assertion follows from

$$\begin{aligned}\sum_{i=0}^{n} \binom{n}{i} (-1)^i x_i x_{n-i} &= (-1)^n \sum_{i=0}^{n} \binom{n}{n-i} (-1)^{-i} x_{n-i} x_i \\ &= -\sum_{i=0}^{n} \binom{n}{i} (-1)^i x_i x_{n-i}.\end{aligned}$$

The second assertion follows from

$$\begin{aligned}&\sum_{i=0}^{n+1} \binom{n+1}{i} (-1)^i x_i x_{n+1-i} \\ &= x_0 x_{n+1} + (-1)^{n+1} x_0 x_{n+1} + \sum_{i=1}^{n} \binom{n+1}{i} (-1)^i x_i x_{n+1-i} \\ &= 2x_{n+1} + \sum_{i=1}^{n} \left(\binom{n}{i} + \binom{n}{i-1} \right) (-1)^i x_i x_{n+1-i}\end{aligned}$$

and

$$\begin{aligned}\sum_{i=1}^{n} \binom{n}{i-1} (-1)^i x_i x_{n+1-i} &= (-1)^{n+1} \sum_{i=1}^{n} \binom{n}{n-i} (-1)^{-i} x_{n+1-i} x_i \\ &= \sum_{i=1}^{n} \binom{n}{i} (-1)^i x_{n+1-i} x_i.\end{aligned}$$

∎

Corollary 3.7. *Let the first p moments of ψ vanish. Then*

$$\sum_{i=0}^{n} (-1)^i \binom{n}{i} M_i M_{n-i} = 0, \quad \textit{for } 1 \le n \le p. \tag{3.8}$$

Proof. Let n be odd. In this case (3.8) is trivially satisfied by (3.6). Let n be even and $1 \le n \le p$, then (3.8) follows from (3.5) from Theorem 3.4 and (3.7). ∎

Equation (3.8) is well known to hold if the continuous moments are replaced by discrete moments (see for example Bäni [2, p. 115], where the normalization $m_0 = 1$ is used).

3.2. Filter Coefficients Parametrized by Moments

To construct families of parametrized scaling functions and wavelets we use that the even moments are determined by the odd up to number of vanishing moments (cf. Remark 3.5). The parametrization is with respect to the first moment of the associated scaling function. We derive formulas for filter coefficients of scaling functions with four, six, and eight filter coefficients and, one, two respectively three vanishing moments using symbolic computation.

To this end we recall the basic polynomial equations for the filter coefficients of a scaling function implied by orthonormality and vanishing moments of the associated wavelet, see for example Daubechies [13] or Strang [26]. The orthonormality of the integer translates of the scaling function imply that the number of filter coefficients is even. In the following, it is convenient to number the filter coefficients by

$$h_k, \quad \text{for } 1-N \leq k \leq N.$$

Note that then the discrete moments m_n for the filter coefficients are equal to

$$m_n = \sum_{k=1-N}^{N} h_k (k+N-1)^n. \tag{3.9}$$

Orthonormality of the scaling function, $\int \phi(x)\phi(x-l) = \delta_{0,l}$, can be transformed using the dilation Equation (1.1) into

$$\sum_{k=1-N}^{N} h_k h_{k-2l} = 2\delta_{0,l}, \quad \text{for } l = 0, \ldots, N-1, \tag{3.10}$$

where $h_k = 0$, for $k < 1-N$ or $k > N$. The condition that the first p moments of the associated wavelet

$$\psi(x) = \sum_{k=1-N}^{N} (-1)^k h_{1-k} \phi(2x-k)$$

vanish, that is

$$N_j = \int x^j \psi(x)\, dx = 0, \quad \text{for } j = 0, \ldots, p-1$$

is equivalent to

$$\sum_{k=1-N}^{N} (-1)^k h_{1-k} k^l = 0, \quad \text{for } l = 0, \ldots, p-1. \tag{3.11}$$

Equation (3.10) for $l = 0$ is redundant and thus omitted.

In the following we present the conditional equations for four and six filter coefficients with one degree of freedom achieved by giving up a vanishing moment condition of the standard orthogonal wavelet setting. For four filter coefficients the conditional system consists of two linear equations, resulting from the normalization $m_0 = 2$ (cf. Lemma 2.1) and the vanishing moment condition (3.11). Using the first discrete moment $m := m_1$ as a parameter gives a third linear constraint on the filter coefficients.

Thus we have the following system of equations:

$$\left.\begin{aligned} h_{-1}+h_0+h_1+h_2 &= 2\\ -h_2+h_1-h_0+h_{-1} &= 0\\ h_0+2h_1+3h_2 &= m \end{aligned}\right\} \text{ linear equations,}$$

$$h_1h_{-1}+h_2h_0 = 0 \qquad \text{quadratic equation.}$$

Solving the system of linear equations for h_2 and substituting the solution into the quadratic equation gives

$$-2h_2^2+h_2m-h_2-1/4m^2+m-3/4=0.$$

This equation has two possible solutions – each of them gives feasible filter coefficients. Let

$$w=\sqrt{-5+6m-m^2} \text{ and } 1\le m\le 5,$$

then for $h_2=-1/4+1/4m-1/4w$ we obtain

$$h_{-1}=5/4-1/4m-1/4w,$$

$$h_0=5/4-1/4m+1/4w,$$

$$h_1=-1/4+1/4m+1/4w.$$

For $m=3-\sqrt{3}$ and $m=3+\sqrt{3}$ we obtain the classical Daubechies filters with two vanishing moments [12]. The Haar wavelet corresponds to $m=1$ (for $m=3,5$ we get translated versions). The smoothest scaling function with four filter coefficients with respect to the Hölder regularity is obtained for $m=1.4$, see Daubechies [13, p. 242] and Rioul [22].

For six filter coefficients with at least two vanishing moments the resulting system of equations for the filter coefficients is much more involved. Now, the system consists of linear equations resulting from the normalization $m_0=2$ (one equation) and the vanishing moment conditions (3.11) (two equations). Using again the first discrete moment $m:=m_1$ as a parameter gives a forth linear constraint on the filter coefficients. In Remark 3.5 it is shown that $m_2=m_1^2/2$ if the first two moments of the associated wavelet function vanish. This gives a further linear equation. Two quadratic equations follow from the orthonormality of the scaling function (3.10):

$$\left.\begin{aligned} h_{-2}+h_{-1}+h_0+h_1+h_2+h_3 &= 2\\ h_3-h_2+h_1-h_0+h_{-1}-h_{-2} &= 0\\ -2h_3+h_2-h_0+2h_{-1}-3h_{-2} &= 0\\ h_{-1}+2h_0+3h_1+4h_2+5h_3 &= m\\ h_{-1}+4h_0+9h_1+16h_2+25h_3 &= m^2/2 \end{aligned}\right\} \text{ linear equations,}$$

$$\left.\begin{aligned} h_0h_{-2}+h_1h_{-1}+h_2h_0+h_3h_1 &= 0\\ h_2h_{-2}+h_3h_{-1} &= 0 \end{aligned}\right\} \text{ quadratic equations.}$$

We have solved this system using *Gröbner bases* with the computer algebra software MAPLE and obtained the parametrized solutions:

$$
\begin{aligned}
h_{-2} &= \frac{21}{16} - \frac{7}{16}m + \frac{1}{32}m^2 - \frac{1}{32}w,\\
h_{-1} &= \frac{25}{16} - \frac{7}{16} + \frac{1}{32}m^2 + \frac{1}{32}w,\\
h_0 &= -\frac{5}{8} + \frac{5}{8}m - \frac{1}{16}m^2 + \frac{1}{16}w,\\
h_1 &= -\frac{5}{8} + \frac{5}{8}m - \frac{1}{16}m^2 - \frac{1}{16}w,\\
h_2 &= \frac{5}{16} - \frac{3}{16}m + \frac{1}{32}m^2 - \frac{1}{32}w,\\
h_3 &= \frac{1}{16} - \frac{3}{16}m + \frac{1}{32}m^2 + \frac{1}{32}w,
\end{aligned}
$$

with

$$w = \sqrt{-260 + 360m - 136m^2 + 20m^3 - m^4} \text{ and } 5 - \sqrt{15} \leq m \leq 5 + \sqrt{15}.$$

The Daubechies wavelet `db3` corresponds to

$$m = 5 - \sqrt{5 - 2\sqrt{10}} \text{ or } m = 5 - \sqrt{5 + 2\sqrt{10}},$$

coiflets belong to the case $m = 4$.

The parametrized solutions for eight filter coefficients with at least three vanishing moments:

$$
\begin{aligned}
h_{-3} &= -\tfrac{1}{512}\,\frac{m^5-42m^4+684m^3-5416m^2+20840m-31088+w}{m^2-14m+50},\\
h_{-2} &= -\tfrac{1}{512}\,\frac{m^6-52m^5+1124m^4-12880m^3+82344m^2-278080m-mw+6w+387072}{m^3-22m^2+162m-400},\\
h_{-1} &= \tfrac{1}{512}\,\frac{3m^5-110m^4+1508m^3-9432m^2+25016m-16464+3w}{m^2-14m+50},\\
h_0 &= \tfrac{1}{512}\,\frac{3m^6-140m^5+2636m^4-25360m^3+129144m^2-317760m-3mw+18w+265216}{m^3-22m^2+162m-400},\\
h_1 &= -\tfrac{1}{512}\,\frac{3m^5-94m^4+1092m^3-5944m^2+15416m-16464+3w}{m^2-14m+50},\\
h_2 &= -\tfrac{1}{512}\,\frac{3m^6-124m^5+2028m^4-16688m^3+71416m^2-142784m-3wm+18w+86016}{m^3-22m^2+162m-400},\\
h_3 &= \tfrac{1}{512}\,\frac{m^5-26m^4+268m^3-1416m^2+4072m-5488+w}{m^2-14m+50},\\
h_4 &= \tfrac{1}{512}\,\frac{m^6-36m^5+516m^4-3696m^3+13352m^2-20160m-wm+6w+3072}{m^3-22m^2+162m-400},
\end{aligned}
$$

with

$$w=\sqrt{-\left(m^8-56m^7+1336m^6-17696m^5+141792m^4-699328m^3+2049600m^2-3186176m+1891904\right)(m-8)^2}.$$

We recall that with Subsection 2.2 the filter coefficients can also be expressed via the continuous moment M_1.

3.3. Parametrized Wavelets and Compression

Here we discuss a novel concept of data compression using parametrized scaling functions and wavelets. For compression the coefficients of the expansion of the data with respect to scaled and dilated scaling functions and wavelets are computed. The coefficients of the expanded data are transmitted and decoded afterwards. For data compression with parametrized wavelets the expansion is computed for a series of parameters and the one that yields optimal compression rates is selected. The coefficients and the parameter are transmitted. These data are sufficient for decoding.

In the following we present a numerical example for data compression using parametrized wavelets. As data we use a one-dimensional signal from contact less ultrasound measurements for non destructive evaluation of an Aluminum sheet (see Figure 1 *left*). Figure 1 *right* shows the "optimal" scaling function and the according wavelet with 8 filter coefficients for approximating the data.

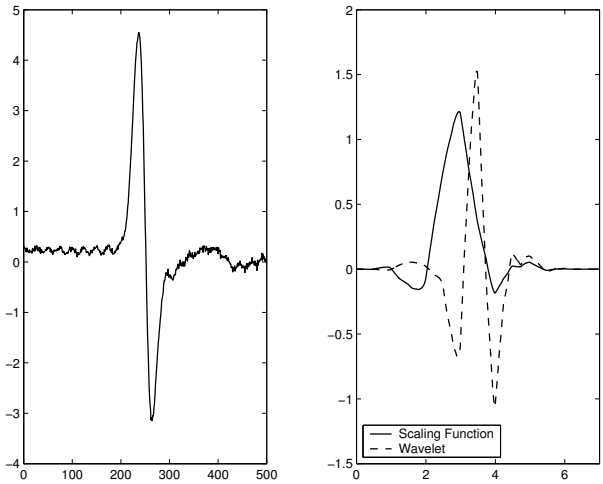


Figure 1: Signal and "optimal" scaling function and wavelet.

In the particular example we have only used the coefficients of the scaling function expansion and have set all wavelet coefficients to zero. This is of course not a realistic way of data compression, but more sophisticated approaches can be dealt with analogously.

The error for the decoded data is shown in Figure 2. We also show the result using the `db4` (Daubechies 4) scaling function. In comparison the squared error (SE) is about three times as high for `db4` and the maximal error (ME) is about double. The related software is available on request from the first named author.

4. Appendix: Bell Polynomials and Partitions

Bell polynomials have been introduced by E.T. Bell [3]. Since then Bell polynomials have become a fundamental tool in combinatorics. One classical application of

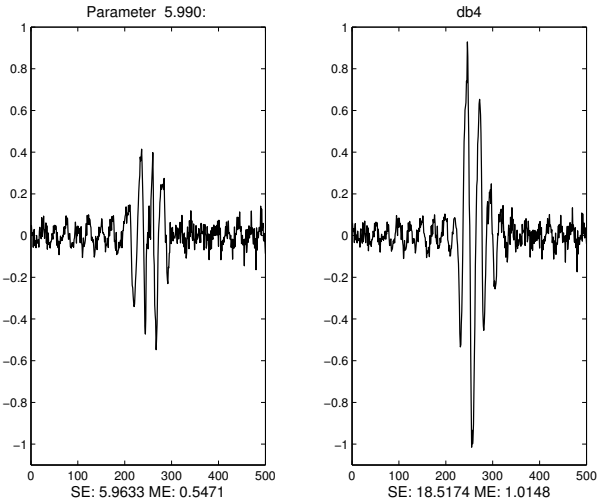


Figure 2: Error for the decoded data.

Bell polynomials is the compact representation of higher order derivatives of composite functions known as the Formula of Faa di Bruno (see Roman [23]). References on Bell polynomials are Comtet [10], Riordan [21], and Roman [24]. For some recent applications of Bell polynomials we refer to Cassisaricci [6] and Collins [9].

Definition 4.1. *Let $n, k \in \mathbb{N}$. The (exponential)* partial Bell polynomials *are defined by*

$$B_{n,k} := B_{n,k}(x_1, \dots, x_{n-k+1}) \tag{4.1}$$

$$= \frac{1}{k!} \sum_{\substack{i_1+\cdots+i_k=n \\ i_j>0}} \binom{n}{i_1, \dots, i_k} x_{i_1} \cdots x_{i_k} \tag{4.2}$$

$$= \sum_{\substack{k_1+\cdots+k_n=k \\ k_1+2k_2+\cdots+nk_n=n \\ k_i\geq 0}} \frac{n!}{k_1!k_2!\cdots k_n!} \left(\frac{x_1}{1!}\right)^{k_1} \left(\frac{x_2}{2!}\right)^{k_2} \cdots \left(\frac{x_n}{n!}\right)^{k_n}. \tag{4.3}$$

This identity in particular shows that $B_{n,k} \in \mathbb{Z}[x_1, \dots, x_{n-k+1}]$. We set $B_{n,0} = 0$ and $B_{0,0} = 1$. The (exponential) Bell polynomials *are defined by*

$$Y_n = \sum_{k=0}^{n} B_{n,k} \in \mathbb{Z}[x_1, \dots, x_n], \quad \text{for } n \in \mathbb{N}_0.$$

The first five Bell polynomials are:

$$\begin{aligned} Y_0 &= 1, \\ Y_1 &= x_1, \\ Y_2 &= x_1^2 + x_2, \\ Y_3 &= x_1^3 + 3x_1x_2 + x_3, \\ Y_4 &= x_1^4 + 6x_1^2x_2 + 3x_2^2 + 4x_1x_3 + x_4. \end{aligned}$$

They satisfy the recurrence relation

$$Y_{n+1}(x_1,\ldots,x_{n+1}) = \sum_{i=0}^{n} \binom{n}{i} Y_{n-i}(x_1,\ldots,x_{n-i})\, x_{i+1}, \tag{4.4}$$

see for example Riordan [21, p. 36].

Definition 4.2. *A* partition *of $n \in \mathbb{N}$ is a nonincreasing sequence of positive integers, denoted by $(\lambda_1,\ldots,\lambda_k)$, whose sum is n. Each λ_i is called a* part *of the partition. A* composition *of $n \in \mathbb{N}$ is a sequence of positive integers whose sum is n.*

Example 4.3. *There are five partitions of* 4:

$$(4), (31), (22), (211), (1111)$$

and eight compositions:

$$(4), (31), (13), (22), (211), (121), (112), (1111).$$

A partition of n in k parts is usually denoted by

$$1^{k_1}2^{k_2}\cdots n^{k_n}, \quad \text{with } k_1 + 2k_2 + \cdots + nk_n = n, \tag{4.5}$$

where $k_i \in \mathbb{N}_0$ is the number of parts equal to i and $k = k_1 + \cdots + k_n$. For instance

$$(211) = 1^2 2.$$

Using Definition 4.2 we see that the sums in (4.2), (4.3) are sums over all compositions, partitions, respectively, of n in k parts.

The number of partitions of n is denoted by $p(n)$. Equation (4.3) together with (4.5) show that the number of monomials in the Bell polynomials Y_n is the number of partitions of n, which increases rapidly with n. For example $p(7) = 15$ and $p(33) = 10143$. For further background on partitions we refer to Andrews [1].

Acknowledgments. This work has been supported by the FWF (Fonds zur Förderung der wissenschaftlichen Forschung), grant Y-123 INF-N12. The authors are grateful to Dr. Peter Burgholzer for providing the ultrasound measurement data Figure 1.

References

1. G.E. Andrews, The Theory of Partitions, Cambridge University Press, Cambridge - New York, 1998.
2. W. Bäni, Wavelets, Eine Einführung für Ingenieure, Oldenbourg Verlag, München, Wien, 2002.
3. E.T. Bell, Exponential polynomials, Ann. of Math. (2) **35** (1934) 258–277.
4. B. Buchberger, An algorithm for finding the bases elements of the residue class ring modulo a zero dimensional polynomial ideal, Ph.D. Thesis, University of Innsbruck, German, 1965.
5. B. Buchberger, Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems, Aequationes Math. **4** (1970) 374–383.
6. C. Cassisa and P.E. Ricci, Orthogonal invariants and the Bell polynomials, Dedicated to the memory of Gaetano Fichera (Italian), Rend. Mat. Appl. **20** (7) (2000) 293–303.
7. A.S. Cavaretta, W. Dahmen, and C.A. Micchelli, Stationary subdivision, Mem. Amer. Math. Soc. **93** (453) (1991) vi+ 186 pp.
8. F. Chyzak, P. Paule, O. Scherzer, A. Schoisswohl, and B. Zimmermann, The construction of orthonormal wavelets using symbolic methods and a matrix analytical approach for wavelets on the interval, Experiment. Math. **10** (1) (2001) 67–86.
9. C.B. Collins, The role of Bell polynomials in integration, J. Comput. Appl. Math. **131** (1-2) (2001) 195–222.
10. L. Comtet, Advanced Combinatorics, D. Reidel Publishing Co., Dordrecht, 1974.
11. D. Cox, J. Little, and D. O'Shea, Ideals, Varieties, and Algorithms, Second Edition, Undergraduate Texts in Mathematics, Springer-Verlag, New York, 1997.
12. I. Daubechies, Orthonormal bases of compactly supported wavelets, Comm. Pure Appl. Math. **41** (7) (1988) 909–996.
13. I. Daubechies, Ten Lectures on Wavelets, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.
14. I. Daubechies, Orthonormal bases of compactly supported wavelets. II. Variations on a theme, SIAM J. Math. Anal. **24** (2) (1993) 499–519.
15. I. Daubechies and J.C. Lagarias, Two-scale difference equations. I. Existence and global regularity of solutions, SIAM J. Math. Anal. **22** (5) (1991) 1388–1410.
16. I. Daubechies and J.C. Lagarias, Two-scale difference equations. II. Local regularity, infinite products of matrices and fractals, SIAM J. Math. Anal. **23** (4) (1992) 1031–1079.
17. R.A. Gopinath and C.S. Burrus, On the moments of the scaling function ψ_0, In: Proc. of the IEEE ISCAS, Vol 2, San Diego, CA, 1992, pp. 963–966.
18. J. Lebrun and I.W. Selesnick, Gröbner bases and wavelet design, J. Symbolic Comput. **37** (2) (2004) 227–259.
19. J. Lebrun and M. Vetterli, High-order balanced multiwavelets: theory, factorization, and design, IEEE Trans. Signal Process. **49** (9) (2001) 1918–1930.
20. S. Mallat, A Wavelet Tour of Signal Processing, Academic Press Inc., San Diego, CA, 1998.
21. J. Riordan, An Introduction to Combinatorial Analysis, Wiley Publications in Mathematical Statistics, John Wiley & Sons Inc., New York, 1958.
22. O. Rioul, Simple regularity criteria for subdivision schemes, SIAM J. Math. Anal. **23** (6) (1992) 1544–1576.
23. S. Roman, The formula of Faà di Bruno, Amer. Math. Monthly **87** (10) (1980) 805–809.
24. S. Roman, The Umbral Calculus, Academic Press Inc., New York, 1984.
25. I.W. Selesnick and C.S. Burrus, Maximally flat low-pass FIR filters with reduced delay, IEEE Trans. Circuits Systems II: Analog and Digital Signal Proc. **45** (1) (1998) 53–68.
26. G. Strang, Wavelets and dilation equations: a brief introduction, SIAM Rev. **31** (4) (1989) 614–627.
27. G. Strang and T. Nguyen, Wavelets and Filter Banks, Wellesley-Cambridge Press, Wellesley, MA, 1996.

28. W. Sweldens and R. Piessens, Quadrature formulae and asymptotic error expansions for wavelet approximations of smooth functions, SIAM J. Numer. Anal. **31** (4) (1994) 1240–1264.
29. M. Unser and T. Blu, Wavelet theory demystified, IEEE Trans. Signal Process. **51** (2) (2003) 470–483.

AAECC (2007) 18:583–601
DOI 10.1007/s00200-007-0054-9

Parametrizing compactly supported orthonormal wavelets by discrete moments

Georg Regensburger

Received: 30 December 2005 / Revised: 19 March 2007 / Published online: 16 August 2007

Abstract We discuss parametrizations of filter coefficients of scaling functions and compactly supported orthonormal wavelets with several vanishing moments. We introduce the first discrete moments of the filter coefficients as parameters. The discrete moments can be expressed in terms of the continuous moments of the related scaling function. To solve the resulting polynomial equations we use symbolic computation and in particular Gröbner bases. The cases of four to ten filter coefficients are discussed and explicit parametrizations are given.

Keywords Orthonormal wavelets · Parametrization · Filter coefficients · Moments · Gröbner bases

1 Introduction

Over the last two decades wavelets have become a fundamental tool in many areas of applied mathematics and engineering ranging from signal and image processing to numerical analysis, see for example Daubechies [13], Mallat [26], and Strang and Nguyen [35]. A function $\psi \in L^2(\mathbb{R})$ is an *orthonormal wavelet* if the family

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k), \quad \text{for } j, k \in \mathbb{Z},$$

This work was supported by the Austrian Science Fund (FWF) under the SFB grant F1322.

G. Regensburger ( )
Johann Radon Institute for Computational and Applied Mathematics (RICAM),
Austrian Academy of Sciences, Altenbergerstraße 69, 4040 Linz, Austria
e-mail: georg.regensburger@oeaw.ac.at

is an orthonormal basis of the Hilbert space $L^2(\mathbb{R})$. The first known example is the Haar wavelet [16]

$$\psi(x) = \begin{cases} 1, & \text{for } 0 \leq x < \frac{1}{2}, \\ -1, & \text{for } \frac{1}{2} \leq x < 1, \\ 0, & \text{otherwise.} \end{cases}$$

Daubechies [12] introduced a general method to construct compactly supported wavelets. It is based on *scaling functions* which satisfy a *dilation equation*

$$\phi(x) = \sum_{k=0}^{N} h_k \phi(2x - k) \tag{1}$$

given by a linear combination of real *filter coefficients* h_k and dilated and translated versions of the scaling function. We outline her construction in Sect. 2. The corresponding scaling function for the Haar wavelet is the box function

$$\phi(x) = \begin{cases} 1, & \text{for } 0 \leq x < 1, \\ 0, & \text{otherwise} \end{cases}$$

with the filter coefficients $h_0 = h_1 = 1$. In general, there is no closed analytic form for the scaling function, and for computations with wavelets only the filter coefficients are used.

Conditions on the scaling function imply, using the dilation equation (1), constraints on the filter coefficients. Orthonormality gives quadratic equations and vanishing moments of the associated wavelet and normalization linear constraints. For the existence of a wavelet at least one vanishing moment is necessary. Daubechies wavelets [12] have the maximal number of vanishing moments for a fixed number of filter coefficients and so there are only finitely many solutions. See Sect. 2 for details.

Parametrizing all possible filter coefficients that correspond to compactly supported orthonormal wavelets has been studied by several authors [20,25,29,31,34,37–39]. For a discussion and illustrations of scaling functions with six filter coefficients depending on two parameters see also [3] and [18]. Applications of parametrized wavelets to compression are for example discussed in [17] and [30]. In all parametrizations the filter coefficients are expressed in terms of trigonometric functions and there is no natural interpretation of the angular parameters for the resulting scaling function. Furthermore, one has to solve transcendental constraints for the parameters to find wavelets with more than one vanishing moment.

In the proposed parametrization we introduce the first discrete moments of the filter coefficients as parameters. The discrete moments can be expressed in terms of the continuous moments of the scaling function, see Sect. 3. Moreover, we do not want to parametrize all possible filter coefficients but only such with a high number of vanishing moments. More precisely, we omit one vanishing moment condition from the construction of Daubechies wavelets. We also use the fact that the even discrete

moments are determined by the odd up to the number of vanishing moments, see Sect. 3. We discussed a first parametrization using the same approach in [30]. In this paper, we present new simplified parametrizations, discuss all computational aspects and different cases in detail, and give a parametrization for ten filter coefficients and at least four vanishing moments.

We solve the resulting parametrized polynomial equations for the filter coefficients using symbolic computation and for the more involved equations in particular Gröbner bases. Gröbner bases were introduced by Buchberger in [4], see also [5]. For further details on Gröbner bases we refer to [1,6,11]. Applications of Gröbner bases to the design of wavelets and filter coefficients are for example discussed in [8,9,15,23,24, 27,28,32]. The idea of using the first discrete moment as a parameter to simplify the Gröbner basis computations was also used in Selesnick and Burrus [32] and Lebrun and Selesnick [23].

In Sects. 4–7 we describe in detail the cases of four to ten filter coefficients. We give explicit parametrizations and discuss several special parameter values, for example, for the Daubechies wavelets. The corresponding Maple worksheet with all computations, several MATLAB functions and a GUI to compute with and illustrate parametrized wavelets are available on request from the author.

2 Equations for the filter coefficients

We outline the construction of orthonormal wavelets based on scaling functions and recall the polynomial equations for the filter coefficients, see for example Daubechies [13] or Strang and Nguyen [35].

Orthonormality of the integer translates $\{\phi(x-l)\}_{l\in\mathbb{Z}}$ in $L^2(\mathbb{R})$, that is,

$$\int \phi(x)\phi(x-l)dx = \delta_{0,l}$$

implies, using the dilation equation (1), the quadratic equations

$$\sum_{k\in\mathbb{Z}} h_k h_{k-2l} = 2\delta_{0,l}, \quad \text{for } l \in \mathbb{Z}, \tag{2}$$

where we set $h_k = 0$ for $k < 0$ and $k > N$. We can assume that $h_0 h_N \neq 0$. Then with Eq. (2) we see that N must be odd and the number of filter coefficients even. We have one nonhomogeneous equation

$$\sum_{k=0}^{N} h_k^2 = 2 \tag{3}$$

and the homogeneous equations

$$\sum_{k=0}^{N} h_k h_{k-2l} = 0, \quad \text{for } l = 1, \ldots, (N-1)/2. \tag{4}$$

If the filter coefficients satisfy the necessary conditions for orthogonality (2) and the normalization

$$\sum_{k=0}^{N} h_k = 2, \tag{5}$$

there exists a unique solution of the dilation equation (1) in $L^2(\mathbb{R})$ with support $[0, N]$ and for which $\int \phi = 1$, see Lawton [21]. For almost all such scaling functions the integer translates $\{\phi(x-l)\}_{l\in\mathbb{Z}}$ are orthogonal, and then

$$\psi(x) = \sum_{k=0}^{N} (-1)^k h_{N-k} \phi(2x-k) \tag{6}$$

is an orthonormal wavelet. Necessary and sufficient conditions for orthonormality were given by Cohen [10] and Lawton [22], see also Daubechies [13, Chap. 6.3.]. The only example with four filter coefficients that satisfies the Eqs. (2) and (5) and where the integer translates of the corresponding scaling are not orthogonal is $h_0 = h_3 = 1$ and $h_1 = h_2 = 0$ with the scaling function

$$\phi(x) = \begin{cases} 1/3, & \text{for } 0 \le x < 3, \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

Vanishing moments of the associated wavelet are related to several properties of the scaling function and wavelet. For example, to the smoothness, the polynomial reproduction and the approximation order of the scaling function, and the decay of the wavelet coefficients for smooth functions, see Strang and Nguyen [35] and the survey by Unser and Blu [36] for details. The condition that the first p moments of the wavelet ψ vanish, that is,

$$\int x^l \psi(x)\, dx = 0, \quad \text{for } l = 0, \ldots, p-1$$

is using Eq. (6) equivalent to the *sum rules*

$$\sum_{k=0}^{N} (-1)^k k^l h_k = 0, \quad \text{for } l = 0, \ldots, p-1. \tag{8}$$

We say that ψ has p vanishing moments. Since the vector space of all polynomials with degree less then p is invariant under translation and dilation, we can equivalently require vanishing moments of $\psi(x+n-1)$ with $N = 2n-1$. This corresponds to Daubechies choice [12,13] where the wavelet has support $[1-n, n]$. For the computations we use the resulting linear equations

$$\sum_{k=0}^{2n-1} (-1)^{n-k} h_k (n-k)^l = 0, \quad \text{for } l = 0, \ldots, p-1$$

since they have smaller coefficients. Note that the normalization of the filter coefficients (5) and the first sum rule

$$\sum_{k=0}^{N}(-1)^k h_k = 0 \tag{9}$$

are equivalent to

$$\sum_{\substack{k=0\\ k\text{ even}}}^{N} h_k = \sum_{\substack{k=0\\ k\text{ odd}}}^{N} h_k = 1. \tag{10}$$

The following proposition is a consequence of the first Newton identities, which give a relation between power sums and elementary symmetric functions, see Bourbaki [2, A.IV. 70] and Knuth [19, p. 497].

Proposition 1 *Let $x_0, \ldots, x_n$ be variables of a polynomial ring over a commutative ring. Then*

$$\left(\sum_{k=0}^{n} x_k^2\right) = \left(\sum_{k=0}^{n} x_k\right)^2 - 2\left(\sum_{\substack{0\le i<j\le n\\ j-i\text{ even}}} x_i x_j\right) - 2\left(\sum_{\substack{k=0\\ k\text{ even}}}^{n} x_k\right)\left(\sum_{\substack{k=0\\ k\text{ odd}}}^{n} x_k\right). \tag{11}$$

Proof The Newton identities tell us in particular that

$$\left(\sum_{k=0}^{n} x_k^2\right) = \left(\sum_{k=0}^{n} x_k\right)^2 - 2\left(\sum_{0\le i<j\le n} x_i x_j\right).$$

The last sum in this equation is

$$\left(\sum_{0\le i<j\le n} x_i x_j\right) = \left(\sum_{\substack{0\le i<j\le n\\ j-i\text{ even}}} x_i x_j\right) + \left(\sum_{\substack{0\le i<j\le n\\ j-i\text{ odd}}} x_i x_j\right)$$

and the proposition follows by observing that

$$\left(\sum_{\substack{0\le i<j\le n\\ j-i\text{ odd}}} x_i x_j\right) = \left(\sum_{\substack{k=0\\ k\text{ even}}}^{n} x_k\right)\left(\sum_{\substack{k=0\\ k\text{ odd}}}^{n} x_k\right).$$

□

If the filter coefficients satisfy the homogeneous equations (4) from the orthonormality conditions then

$$\sum_{\substack{0\le i<j\le n\\ j-i \text{ even}}} h_i h_j = 0.$$

Therefore we see with the identity (11) that the normalization and the first sum rule, see Eqs. (5), (9) and (10) together with (4) imply the nonhomogeneous equation (3). So we can replace the quadratic equation (3) by the linear equation (9), which simplifies the computations.

3 Discrete and continuous moments

In this section, we discuss relations between the *discrete moments*

$$m_n = \sum_{k=0}^{N} h_k k^n$$

of the filter coefficients and the *continuous moments* of the scaling function

$$M_n = \int x^n \phi(x)\, dx.$$

We first recall a well-known recursive relation between discrete and continuous moments, see for example Strang and Nguyen [35, p. 396].

Let ϕ be a scaling function satisfying $M_0 = \int \phi = 1$. Then $m_0 = 2$ and

$$M_n = \frac{1}{2^{n+1}-2} \sum_{i=1}^{n} \binom{n}{i} m_i M_{n-i},$$

$$m_n = \left(2^{n+1}-2\right) M_n - \sum_{i=1}^{n-1} \binom{n}{i} m_i M_{n-i}, \quad \text{for } n > 0.$$

Using the recursion we obtain for the first moments

$$\begin{aligned} M_1 &= 1/2\, m_1 \\ M_2 &= 1/6\, m_1^2 + 1/6\, m_2 \\ M_3 &= 1/28\, m_1^3 + 1/7\, m_1 m_2 + 1/14\, m_3 \end{aligned}$$

and

$$\begin{aligned} m_1 &= 2\,M_1 \\ m_2 &= -4\,M_1^2 + 6\,M_2 \\ m_3 &= 12\,M_1^3 - 24\,M_1 M_2 + 14\,M_3. \end{aligned}$$

Explicit formulas expressing the discrete moments in terms of the continuous and vice versa are given in [30].

For the parametrization of the filter coefficients we use the fact that the even moments are determined by the odd moments up to the number of vanishing moments, see [30]. In more detail, if the first two moments of the associated wavelet vanish, then

$$m_2 = m_1^2/2, \tag{12}$$

and if the first four moments vanish, we additionally have

$$m_4 = -1/2\,m_1^4 + 2\,m_1^2 m_2 + 2\,m_1 m_3 - 7/2\,m_2^2 = -3/8\,m_1^4 + 2\,m_1 m_3. \tag{13}$$

4 Four filter coefficients

In the case of four filter coefficients, we have the following system equations (normalization, first sum rule, parameter $m = m_1$, and orthogonality):

$$\begin{aligned} h_0 + h_1 + h_2 + h_3 &= 2 \\ h_0 - h_1 + h_2 - h_3 &= 0 \\ h_1 + 2\,h_2 + 3\,h_3 &= m \\ h_0 h_2 + h_1 h_3 &= 0. \end{aligned}$$

We solve the three linear equations for h_0, substitute the solution into the quadratic equation, and obtain

$$-2\,{h_0}^2 + (5-m)h_0 - 1/4\,m^2 + 2\,m - 15/4 = 0. \tag{14}$$

We first consider the solution

$$h_0 = 5/4 - 1/4\,m - 1/4\,\sqrt{-m^2 + 6\,m - 5}.$$

Since

$$-m^2 + 6\,m - 5 = -(m-1)(m-5), \tag{15}$$

we can choose $m \in [1, 5]$ to get real filter coefficients. We set $m = a + 3$ to obtain parameter values symmetrically around zero. This correspond to a Tschirnhaus transformation for the polynomial (15) and simplifies the expression for the filter coefficients. Substituting the solution for h_0 into the solution for the linear equations we

get:

$$\begin{aligned} h_0 &= 1/2 - 1/4\,a - 1/4\,w \\ h_1 &= 1/2 - 1/4\,a + 1/4\,w \\ h_2 &= 1/2 + 1/4\,a + 1/4\,w \\ h_3 &= 1/2 + 1/4\,a - 1/4\,w \end{aligned} \tag{16}$$

with $w = \sqrt{4 - a^2}$ and $a = m - 3 \in [-2, 2]$.

Notice that for $a = -a$ we obtain the flipped filter coefficients.

4.1 Special parameter values

For $a = 0$ we get the filter coefficients $(0, 1, 1, 0)$, which correspond to a translated Haar scaling function and wavelet. The parameter values $a = -2, 2$ give also Haar scaling functions with the filter coefficients $(1, 1, 0, 0)$ and $(0, 0, 1, 1)$.

The Daubechies wavelet has two vanishing moments, so we have one more sum rule

$$2\,h_0 - h_1 + h_3 = 0.$$

Substituting the parametrized filter coefficients into this equations and solving for a, we get the two solutions $a = -\sqrt{3}, \sqrt{3}$ with the first discrete moments $m = 3 - \sqrt{3}, 3 + \sqrt{3}$. The first solution gives the famous Daubechies filters [12]

$$1/4\,(1 + \sqrt{3}, 3 + \sqrt{3}, 3 - \sqrt{3}, 1 - \sqrt{3}) \tag{17}$$

and the second the flipped version.

For $a = -8/5$ we get the rational filters $(3/5, 6/5, 2/5, -1/5)$. These rational filter coefficients give the smoothest scaling function with respect to the Hölder continuity, see Daubechies [13, p. 242].

4.2 Second root

If we choose the second root

$$h_0 = 5/4 - 1/4\,m + 1/4\,\sqrt{-m^2 + 6\,m - 5}$$

for the quadratic equation (14) and apply again the Tschirnhaus transformation $m = a + 3$, we obtain the parametrized filter coefficients:

$$
\begin{aligned}
h_0 &= 1/2 - 1/4\,a + 1/4\,w \\
h_1 &= 1/2 - 1/4\,a - 1/4\,w \\
h_2 &= 1/2 + 1/4\,a - 1/4\,w \\
h_3 &= 1/2 + 1/4\,a + 1/4\,w
\end{aligned}
$$

with $w = \sqrt{4 - a^2}$ and $a = m - 3 \in [-2, 2]$.

Comparing this solution with the parametrized filter coefficients (16), we see that w is replaced by $-w$ and so the two first and the two last filter coefficients are swapped. Notice that again for $a = -a$ we obtain the flipped filters.

For $a = 0$ we now get the filter coefficients $(1, 0, 0, 1)$, which give the scaling function (7) where the integer translates of the scaling function are not orthogonal. The parameter values $a = -2, 2$ also give Haar scaling functions with the filter coefficients $(1, 1, 0, 0)$ and $(0, 0, 1, 1)$. This parametrization does not contain filter coefficients with a second vanishing moment. The corresponding scaling functions are, compared to the parametrization (16), irregular.

5 Six filter coefficients

For six filter coefficients we have two vanishing moments, and we can use the relation $m_2 = m_1^2/2$, see Eq. (12). This gives an additional linear constraint, and we have the following linear equations with $m = m_1$:

$$
\begin{aligned}
h_0 + h_1 + h_2 + h_3 + h_5 + h_4 &= 2 \\
-h_0 + h_1 - h_2 + h_3 - h_4 + h_5 &= 0 \\
-3\,h_0 + 2\,h_1 - h_2 + h_4 - 2\,h_5 &= 0 \\
h_1 + 2\,h_2 + 3\,h_3 + 4\,h_4 + 5\,h_5 &= m \\
h_1 + 4\,h_2 + 9\,h_3 + 16\,h_4 + 25\,h_5 &= m^2/2
\end{aligned}
$$

and the quadratic equations

$$
\begin{aligned}
h_0h_2 + h_1h_3 + h_2h_4 + h_3h_5 &= 0 \\
h_0h_4 + h_1h_5 &= 0.
\end{aligned}
$$

We solve the linear equations for h_0, substitute the solution into the quadratic equations and obtain:

$$
\begin{aligned}
&-8\,{h_0}^2 + (1/2\,m^2 - 7\,m + 21)h_0 - \frac{1}{64}\,m^4 + \frac{3}{8}\,m^3 - \frac{13}{4}\,m^2 + 12\,m - \frac{253}{16} = 0 \\
&2\,{h_0}^2 + \left(-1/8\,m^2 + \frac{7}{4}\,m - \frac{21}{4}\right)h_0 + \frac{1}{256}\,m^4 - \frac{3}{32}\,m^3 + \frac{13}{16}\,m^2 - 3\,m + \frac{253}{64} = 0.
\end{aligned}
\tag{18}
$$

Since the first equation is minus four times the second equation, we have, as in the case of four filter coefficients, only one quadratic equation to solve. We first consider the solution

$$h_0 = \frac{21}{16} - \frac{7}{16}\,m + \frac{1}{32}\,m^2 - \frac{1}{32}\,\sqrt{-m^4 + 20\,m^3 - 136\,m^2 + 360\,m - 260}.$$

The Tschirnhaus transformation $m = a + 5$ for the polynomial

$$-m^4 + 20\,m^3 - 136\,m^2 + 360\,m - 260$$

yields

$$-a^4 + 14\,a^2 + 15 = -\left(a^2 - 15\right)\left(a^2 + 1\right).$$

So we get real filter coefficients for $a \in [-\sqrt{15}, \sqrt{15}]$ or the first discrete moment $m \in [5 - \sqrt{15}, 5 + \sqrt{15}]$. Substituting the solution for h_0 into the solution for the linear equations, we get the following parametrized filter coefficients with at least two vanishing moments:

$$\begin{aligned}
h_0 &= -3/32 - 1/8\,a + 1/32\,a^2 - 1/32\,w \\
h_1 &= 5/32 - 1/8\,a + 1/32\,a^2 + 1/32\,w \\
h_2 &= 15/16 - 1/16\,a^2 + 1/16\,w \\
h_3 &= 15/16 - 1/16\,a^2 - 1/16\,w \\
h_4 &= 5/32 + 1/8\,a + 1/32\,a^2 - 1/32\,w \\
h_5 &= -3/32 + 1/8\,a + 1/32\,a^2 + 1/32\,w
\end{aligned} \tag{19}$$

with $w = \sqrt{-a^4 + 14\,a^2 + 15}$ and $a = m - 5 \in [-\sqrt{15}, \sqrt{15}]$.

5.1 Special parameter values

The Daubechies wavelet has one more vanishing moment, that is, it satisfies the sum rule

$$-9\,h_0 + 4\,h_1 - h_2 - h_4 + 4\,h_5 = 0.$$

Substituting the parametrized filter coefficients into this equations and solving for a, we get one real solution $a = -\sqrt{5 + 2\,\sqrt{10}}$, which gives the filter coefficients

$$\begin{aligned}
1/16\,(&1 + \sqrt{10} + w, 5 + \sqrt{10} + 3\,w, 10 - 2\,\sqrt{10} + 2\,w, \\
&10 - 2\,\sqrt{10} - 2\,w, 5 + \sqrt{10} - 3\,w, 1 + \sqrt{10} - w)
\end{aligned} \tag{20}$$

with $w = \sqrt{5 + 2\,\sqrt{10}}$.

The Daubechies filters with four nonzero filter coefficients (17) satisfy two sum rules and are therefore contained in this parametrization. Their first discrete moment

is $m = 3 - \sqrt{3}$. So here the corresponding parameter is $a = -2 - \sqrt{3}$. We get a translated version for $a = -\sqrt{3}$.

For $a = -\sqrt{15}$ we obtain

$$1/8\,(3 + \sqrt{15}, 5 + \sqrt{15}, 0, 0, 5 - \sqrt{15}, 3 - \sqrt{15}).$$

The parameter $a = -1$ gives the first coiflet

$$1/16\,(1 - \sqrt{7}, 5 + \sqrt{7}, 14 + 2\sqrt{7}, 14 - 2\sqrt{7}, 1 - \sqrt{7}, -3 + \sqrt{7}),$$

see Daubechies [14] and [13, Chap. 8.2.]. For $a = 0$ we get

$$1/32\,(-3 - \sqrt{15}, 5 + \sqrt{15}, 30 + 2\sqrt{15}, 30 - 2\sqrt{15}, 5 - \sqrt{15}, -3 + \sqrt{15}).$$

The corresponding scaling functions and wavelets for $a > 0$ become increasingly irregular.

5.2 Second root

If we choose the second solution for the quadratic equation (18) and apply the Tschirnhaus transformation $m = a + 5$, we obtain:

$$\begin{aligned}
h_0 &= -3/32 - 1/8\,a + 1/32\,a^2 + 1/32\,w \\
h_1 &= 5/32 - 1/8\,a + 1/32\,a^2 - 1/32\,w \\
h_2 &= 15/16 - 1/16\,a^2 - 1/16\,w \\
h_3 &= 15/16 - 1/16\,a^2 + 1/16\,w \\
h_4 &= 5/32 + 1/8\,a + 1/32\,a^2 + 1/32\,w \\
h_5 &= -3/32 + 1/8\,a + 1/32\,a^2 - 1/32\,w
\end{aligned}$$

with $w = \sqrt{-a^4 + 14\,a^2 + 15}$ and $a = m - 5 \in [-\sqrt{15}, \sqrt{15}]$.

Notice that substituting $a = -a$ gives the flipped filter coefficients from the parametrization (19).

6 Eight filter coefficients

For eight filter coefficients we have three vanishing moments, and we can use as in the previous section the relation $m_2 = 1/2\,m_1^2$, see Eq. (12). We have the following

six linear equations with $m = m_1$:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 \\ 3 & -2 & 1 & 0 & -1 & 2 & -3 & 4 \\ -9 & 4 & -1 & 0 & -1 & 4 & -9 & 16 \\ 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 \\ 49 & 36 & 25 & 16 & 9 & 4 & 1 & 0 \end{pmatrix} \begin{pmatrix} h_7 \\ h_6 \\ h_5 \\ h_4 \\ h_3 \\ h_2 \\ h_1 \\ h_0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \\ m \\ 1/2\, m^2 \end{pmatrix} \tag{21}$$

and the quadratic equations

$$\begin{aligned} h_0h_2 + h_1h_3 + h_3h_5 + h_2h_4 + h_4h_6 + h_5h_7 &= 0 \\ h_0h_4 + h_1h_5 + h_3h_7 + h_2h_6 &= 0 \\ h_0h_6 + h_1h_7 &= 0. \end{aligned}$$

We solve the linear equations for h_0 and h_1 and substitute the solutions into the quadratic equations. Then we compute a Gröbner basis with respect to the lexicographic order with $h_1 >_{\text{lex}} h_0$ treating m as a parameter, that is, we compute a Gröbner basis in $\mathbb{Q}(m)[h_1, h_0]$.

The Gröbner basis has two elements. The first element is a quadratic polynomial in h_0 and the second linear in h_1 and h_0. We consider the following solution for the quadratic equation from the Gröbner basis

$$h_0 = -\frac{1}{512}\,\frac{m^5 - 42\,m^4 + 684\,m^3 - 5416\,m^2 + 20840\,m - 31088 + w}{m^2 - 14\,m + 50}$$

with $w =$

$$\sqrt{-(m^8 - 56\,m^7 + 1336\,m^6 - 17696\,m^5 + 141792\,m^4 - 699328\,m^3 + 2049600\,m^2 - 3186176\,m + 1891904)(m-8)^2}.$$

We set $m = a + 7$, which corresponds to a Tschirnhaus transformation for the first factor of the polynomial under the square root in w, and obtain

$$h_0 = -\frac{1}{512}\,\frac{a^5 - 7\,a^4 - 2\,a^3 + 30\,a^2 - 55\,a - 15 + w}{a^2 + 1}$$

with

$$w = \sqrt{-(a^8 - 36\,a^6 + 182\,a^4 - 1540\,a^2 + 945)(a-1)^2}. \tag{22}$$

To get real filter coefficients, we can choose a in

$$[-\sqrt{\beta}, -\sqrt{\alpha}] \quad \text{or} \quad [\sqrt{\alpha}, \sqrt{\beta}], \tag{23}$$

where α denotes the smaller and β the larger real root of

$$x^4 - 36\,x^3 + 182\,x^2 - 1540\,x + 945,$$

with numerical approximations

$$\sqrt{\alpha} = 0.8113601077\ldots \quad \text{and} \quad \sqrt{\beta} = 5.636256558\ldots$$

We substitute the solution for h_0 into the linear equation from the Gröbner basis, solve for h_1 and obtain with w as in (22)

$$h_1 = -\frac{1}{512}\,\frac{a^6 - 10\,a^5 + 39\,a^4 - 28\,a^3 - 25\,a^2 + 86\,a - 63 - (1+a)w}{a^3 - a^2 + a - 1}.$$

The denominator

$$a^3 - a^2 + a - 1 = (a-1)(a^2+1)$$

is zero for $a = 1$. We first assume $a < 1$. Then we can also simplify the root (22) and obtain with the solution for the linear equations (21) the following parametrized filter coefficients with at least three vanishing moments:

$$\begin{aligned}
h_0 &= -\frac{1}{512}\,\frac{a^5 - 7\,a^4 - 2\,a^3 + 30\,a^2 - 55\,a - 15 + (1-a)w}{a^2+1}\\
h_1 &= -\frac{1}{512}\,\frac{a^5 - 9\,a^4 + 30\,a^3 + 2\,a^2 - 23\,a + 63 + (1+a)w}{a^2+1}\\
h_2 &= \frac{1}{512}\,\frac{3\,a^5 - 5\,a^4 - 102\,a^3 + 186\,a^2 - 261\,a + 35 + 3(1-a)w}{a^2+1}\\
h_3 &= \frac{1}{512}\,\frac{3\,a^5 - 11\,a^4 - 70\,a^3 + 358\,a^2 - 229\,a + 525 + 3(1+a)w}{a^2+1}\\
h_4 &= -\frac{1}{512}\,\frac{3\,a^5 + 11\,a^4 - 70\,a^3 - 358\,a^2 - 229\,a - 525 + 3(1-a)w}{a^2+1}\\
h_5 &= -\frac{1}{512}\,\frac{3\,a^5 + 5\,a^4 - 102\,a^3 - 186\,a^2 - 261\,a - 35 + 3(1+a)w}{a^2+1}\\
h_6 &= \frac{1}{512}\,\frac{a^5 + 9\,a^4 + 30\,a^3 - 2\,a^2 - 23\,a - 63 + (1-a)w}{a^2+1}\\
h_7 &= \frac{1}{512}\,\frac{a^5 + 7\,a^4 - 2\,a^3 - 30\,a^2 - 55\,a + 15 + (1+a)w}{a^2+1}
\end{aligned} \tag{24}$$

with

$$w = \sqrt{-a^8 + 36\,a^6 - 182\,a^4 + 1540\,a^2 - 945},$$

$a = m - 7 < 1$ and a in the intervals (23).

If we choose the second root for the quadratic equation from the Gröbner basis and perform the same computations as before with the assumption $a < 1$, we obtain the filter coefficients (24) with w replaced by $-w$.

6.1 Different order on the variables

We now compute a Gröbner basis with respect to the lexicographic order with $h_0 >_{\text{lex}} h_1$. The Gröbner basis has again two elements. The first element is a quadratic polynomial in h_1 and the second linear in h_0 and h_1.

We consider the following solution for the quadratic equation from the Gröbner basis

$$h_1 = -\frac{1}{512}\,\frac{m^5 - 44\,m^4 + 772\,m^3 - 6704\,m^2 + 28712\,m - 48384 - w}{m^2 - 14\,m + 50}$$

with $w =$

$$\sqrt{-(m^8-56\,m^7+1336\,m^6-17696\,m^5+141792\,m^4-699328\,m^3+2049600\,m^2-3186176\,m+1891904)(m-6)^2}.$$

We set again $a = m + 7$ and obtain

$$h_1 = -\frac{1}{512}\,\frac{a^5 - 9\,a^4 + 30\,a^3 + 2\,a^2 - 23\,a + 63 - w}{a^2 + 1}$$

with

$$w = \sqrt{-(a^8 - 36\,a^6 + 182\,a^4 - 1540\,a^2 + 945)(a+1)^2}. \tag{25}$$

We get real filter coefficients for a in the same intervals (23) as in the previous section. We substitute the solution for h_1 into the linear equation from the second Gröbner basis, solve for h_0 and obtain with w as in (25)

$$h_0 = -\frac{1}{512}\,\frac{a^6 - 6\,a^5 - 9\,a^4 + 28\,a^3 - 25\,a^2 - 70\,a - 15 + (a-1)w}{a^3 + a^2 + a + 1}.$$

The denominator

$$a^3 + a^2 + a + 1 = (a+1)(a^2+1)$$

is zero for $a = -1$. We assume $a > -1$. Then we can also simplify the root (25) and obtain with the solution for the linear equations (21) the filter coefficients from Eq. (24) with w replaced by $-w$. From the previous section we know that this parametrization is also valid for $a < 1$ and hence for a in the intervals (23). Notice that substituting $a = -a$ in this parametrization gives the flipped filter coefficients from Eq. (24).

If we choose the second root for the quadratic equation from the Gröbner basis and perform the same computations as before with the assumption $a > -1$, we obtain the filter coefficients (24). Therefore the parametrization (24) is also valid for a in the intervals (23).

6.2 Special parameter values

The Daubechies wavelet satisfies one more sum rule

$$64\,h_0 - 27\,h_1 + 8\,h_2 - h_3 + h_5 - 8\,h_6 + 27\,h_7 = 0.$$

Substituting the parametrized filter coefficients (24) into this equations and solving for a, we get two real solution $a = -\sqrt{\beta}, -\sqrt{\alpha}$, where α denotes the smaller and β the larger real root of

$$x^4 - 28\,x^3 + 126\,x^2 - 1260\,x + 1225$$

or numerically

$$a = -4.989213573\ldots, -1.029063869\ldots$$

The first parameter gives the Daubechies wavelet with extremal phase [13, p. 195] and the second the "least asymmetric" [13, p. 198]. Notice that the symbolic expression for the parameter a with the parametrization (24) give us a closed form representation of the filter coefficients of the Daubechies wavelet. Compare this with the results obtained by Chyzak et al. [9], where also Gröbner bases are used, and the different approach by Shann and Yen [33].

The Daubechies wavelet with six nonzero filter coefficients (20) has the first discrete moment $m = 5 - \sqrt{5 + 2\sqrt{10}}$, so the corresponding parameter value for the parametrization (24) is $a = -2 - \sqrt{5 + 2\sqrt{10}}$.

7 Ten filter coefficients

For ten filter coefficients we require four vanishing moments. We can therefore use the two relations $m_2 = 1/2\,m_1^2$ and $m_4 = -3/8\,m_1^4 + 2\,m_1 m_3$, see Eqs. (12) and (13). This gives two additional linear constraints and we have the following linear equations with the two parameters $a = m_1$ and $c = m_3$:

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ -4 & 3 & -2 & 1 & 0 & -1 & 2 & -3 & 4 & -5 \\ 16 & -9 & 4 & -1 & 0 & -1 & 4 & -9 & 16 & -25 \\ -64 & 27 & -8 & 1 & 0 & -1 & 8 & -27 & 64 & -125 \\ 9 & 8 & 7 & 6 & 5 & 4 & 3 & 2 & 1 & 0 \\ 81 & 64 & 49 & 36 & 25 & 16 & 9 & 4 & 1 & 0 \\ 729 & 512 & 343 & 216 & 125 & 64 & 27 & 8 & 1 & 0 \\ 6561 & 4096 & 2401 & 1296 & 625 & 256 & 81 & 16 & 1 & 0 \end{pmatrix} \begin{pmatrix} h_9 \\ h_8 \\ h_7 \\ h_6 \\ h_5 \\ h_4 \\ h_3 \\ h_2 \\ h_1 \\ h_0 \end{pmatrix} = \begin{pmatrix} 2 \\ 0 \\ 0 \\ 0 \\ 0 \\ a \\ 1/2\,a^2 \\ c \\ -\frac{3}{8}\,a^4 + 2\,ac \end{pmatrix} \tag{26}$$

Table 1 Number of real solutions for f from (27)

Parameter a	# Real solutions for c
(1.6417, 7.6167]	Two
(7.6167, 9)	Four
9	Two, singular point
(9, 10.3832]	Four
(10.3832, 16.3583)	Two

and the quadratic equations

$$\begin{aligned} h_0h_2 + h_1h_3 + h_2h_4 + h_3h_5 + h_4h_6 + h_5h_7 + h_6h_8 + h_7h_9 &= 0\\ h_0h_4 + h_1h_5 + h_2h_6 + h_3h_7 + h_4h_8 + h_5h_9 &= 0\\ h_0h_6 + h_1h_7 + h_2h_8 + h_3h_9 &= 0\\ h_0h_8 + h_1h_9 &= 0. \end{aligned}$$

We solve the linear equations for h_0 and substitute the solutions into the quadratic equations. We compute a Gröbner basis with respect to the lexicographic order with $h_0 >_{\text{lex}} c$ treating a as a parameter, that is, a Gröbner basis in $\mathbb{Q}(a)[h_0, c]$. The Gröbner basis consists of two elements. The first is the polynomial

$$\begin{aligned} f = {} & 81\,a^{12} - 2916\,a^{11} + 40716\,a^{10} - 864\,a^9c - 155520\,a^9 + 31104\,a^8c - 2354328\,a^8\\ & - 496512\,a^7c + 2880\,a^6c^2 + 31658688\,a^7 + 3768768\,a^6c - 93312\,a^5c^2\\ & - 102669504\,a^6 - 4056192\,a^5c + 1540224\,a^4c^2 - 3072\,a^3c^3 - 590398848\,a^5\\ & - 176214528a^4c - 15303168a^3c^2 + 55296a^2c^3 + 6210049216a^4 + 1512364544a^3c\\ & + 97677312\,a^2c^2 - 489472\,ac^3 + 1024\,c^4 - 22429995264\,a^3 - 5357366784\,a^2c\\ & - 358511616\,ac^2 + 1419264\,c^3 + 41210318592\,a^2 + 8252955648\,ac\\ & + 548785152\,c^2 - 39607335936\,a - 4229148672\,c + 16394918400 \end{aligned} \tag{27}$$

in the two parameters a, c and has $\deg_a(f) = 12$ and $\deg_c(f) = 4$. All possible parameters must lie on the real algebraic curve defined by the polynomial f. This curve has genus eleven and two finite singular points with multiplicity two and coordinates

$$a = 9,\ \ c = 729/4 \pm 3/8\,\sqrt{210}. \tag{28}$$

We compute the discriminant f with respect to c. Approximating its zeros, we see that we have real solutions for c if the first discrete moment

$$a \in [1.641693500\ldots, 16.35830649\ldots].$$

The number of real solutions for c is given in Table 1.

The second element in the Gröbner basis is linear in h_0. We solve this polynomial for h_0 and obtain with the solution for the linear equations (26) the following parametrized filter coefficients with at least four vanishing moments:

$$
\begin{aligned}
h_0 &= \frac{1}{36864}\frac{9a^6-180a^5+948a^4-48a^3c+9840a^3+960a^2c-116824a^2-9568ac+32c^2+384480a+31680c-482976}{a-9}\\
h_1 &= -\frac{1}{36864}\frac{9a^6-144a^5+624a^4-48a^3c+1536a^3+768a^2c+12824a^2-5728ac+32c^2-237312a+12672c+665280}{a-9}\\
h_2 &= -\frac{1}{9216}\frac{9a^6-180a^5+948a^4-48a^3c+8976a^3+960a^2c-99064a^2-9472ac+32c^2+257760a+30816c-151200}{a-9}\\
h_3 &= \frac{1}{9216}\frac{9a^6-144a^5+624a^4-48a^3c+2544a^3+768a^2c-9976a^2-5824ac+32c^2-53280a+13536c+120960}{a-9}\\
h_4 &= \frac{1}{6144}\frac{9a^6-180a^5+948a^4-48a^3c+8304a^3+960a^2c-88408a^2-9376ac+32c^2+216288a+29952c-151200}{a-9}\\
h_5 &= -\frac{1}{6144}\frac{9a^6-144a^5+624a^4-48a^3c+3360a^3+768a^2c-24904a^2-5920ac+32c^2+27072a+14400c+12096}{a-9}\\
h_6 &= -\frac{1}{9216}\frac{9a^6-180a^5+948a^4-48a^3c+7824a^3+960a^2c-82552a^2-9280ac+32c^2+202464a+29088c-151200}{a-9}\\
h_7 &= \frac{1}{9216}\frac{9a^6-144a^5+624a^4-48a^3c+3984a^3+768a^2c-34264a^2-6016ac+32c^2+65952a+15264c-34560}{a-9}\\
h_8 &= \frac{1}{36864}\frac{9a^6-180a^5+948a^4-48a^3c+7536a^3+960a^2c-79192a^2-9184ac+32c^2+195552a+28224c-151200}{a-9}\\
h_9 &= -\frac{1}{36864}\frac{9a^6-144a^5+624a^4-48a^3c+4416a^3+768a^2c-40360a^2-6112ac+32c^2+88704a+16128c-60480}{a-9}
\end{aligned}
\tag{29}
$$

with $a \neq 9, c \in \mathbb{R}$ such that $f(a, c) = 0$ with f from (27).

7.1 Special parameter values

For the Daubechies wavelet we have an additional sum rule which we add to the linear equations (26). We solve the linear equations, substitute the solution into the quadratic equations and obtain four polynomials in the two parameters a and c. We compute a Gröbner basis with respect to the lexicographic order with $c >_{\text{lex}} a$. It consists of two polynomials. The first is a univariate polynomial of degree 16 in a. Solving for a, we obtain four real solutions $a = 9 \pm \sqrt{\alpha}, 9 \pm \sqrt{\beta}$, where α denotes the smaller and β the larger positive real root of

$$x^8 - 72x^7 + 1692x^6 - 20472x^5 - 3258\,x^4 + 1386504\,x^3 - 8218980\,x^2 - 1640520\,x + 16769025$$

or numerically

$$a = 2.387816036\ldots, 7.767314070\ldots, 10.23268592\ldots, 15.61218396\ldots$$

The second polynomial in the Gröbner basis has degree 15 in a but depends only linearly on c. So we can express the corresponding values for the parameter c in terms of the first discrete moment a and obtain the numerical approximations

$$c = 1.701845088\ldots, 109.6494477\ldots, 275.3639993\ldots, 953.0313413\ldots$$

The first choice for a and the corresponding c gives the Daubechies wavelet with extremal phase [13, p. 195] and the second the "least asymmetric" [13, p. 198]. The two other choices give the flipped versions. We have again a closed form of the filter coefficients of the Daubechies wavelet with the symbolic expression for the parameters a and c and the parametrization (29), compare with [9] and [33].

To compute the filter coefficients for $a = 9$, we solve the linear equations (26) with the parameter values (28) for h_0 and substitute the solution into the quadratic equations. Then we solve the four univariate polynomials and obtain two solutions for h_0 which give two different filter coefficients. The second choice for c from (28) gives the flipped filter coefficients.

Acknowledgments I would like to thank Otmar Scherzer for raising my interest in wavelets, Josef Schicho for interesting and helpful discussions, and the reviewers for useful remarks.

References

1. Becker, T., Weispfenning, V.: Gröbner Bases, Graduate Texts in Mathematics, vol. 141. Springer, New York (1993)
2. Bourbaki, N.: Algebra. II. Chap. 4–7. Elements of Mathematics (Berlin). Springer, Berlin (1990)
3. Bratteli, O., Jorgensen, P.: Wavelets through a Looking Glass. Applied and Numerical Harmonic Analysis. Birkhäuser Boston Inc., Boston (2002)
4. Buchberger, B.: An algorithm for finding the bases elements of the residue class ring modulo a zero dimensional polynomial ideal (German). Ph.D. Thesis, University of Innsbruck (1965) (English translation published in [7])
5. Buchberger, B.: Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems. Aequationes Math. **4**, 374–383 (1970)
6. Buchberger, B: Introduction to Gröbner bases. In: Buchberger, B., Winkler, F. (eds.) Gröbner Bases and Applications (Linz, 1998), Lond Math. Soc. Lect Note Ser., vol. 251, pp. 3–31. Cambridge University Press, Cambridge (1998)
7. Buchberger, B.: Bruno Buchberger's Ph.D. thesis 1965: An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal. J. Symb. Comput. **41**(3–4): 475–511 (2006) (Translated from the 1965 German original by Michael P. Abramson)
8. Charoenlarpnopparut, C., Bose, N.: Multidimensional FIR filter bank design using Gröbner bases. IEEE Trans. Circuits Syst., II, Analog Digit. Signal Process. **46**(12), 1475–1486 (1999)
9. Chyzak, F., Paule, P., Scherzer, O., Schoisswohl, A., Zimmermann, B.: The construction of orthonormal wavelets using symbolic methods and a matrix analytical approach for wavelets on the interval. Exp. Math. **10**(1), 67–86 (2001)
10. Cohen, A.: Ondelettes, analyses multirésolutions et filtres miroirs en quadrature. Ann. Inst. H. Poincaré Anal. Non Linéaire **7**(5), 439–459 (1990)
11. Cox, D., Little, J., O'Shea, D.: Ideals, Varieties, and Algorithms, 2nd edn. Undergraduate Texts in Mathematics. Springer, New York (1997). An introduction to computational algebraic geometry and commutative algebra
12. Daubechies, I.: Orthonormal bases of compactly supported wavelets. Comm. Pure Appl. Math. **41**(7), 909–996 (1988)
13. Daubechies, I.: Ten Lectures on Wavelets. Society for Industrial and Applied Mathematics (SIAM), Philadelphia (1992)
14. Daubechies, I.: Orthonormal bases of compactly supported wavelets. II. Variations on a theme. SIAM J. Math. Anal. **24**(2), 499–519 (1993)
15. Faugère, J.C., Moreau de Saint-Martin, F., Rouillier, F.: Design of regular nonseparable bidimensional wavelets using Gröbner basis techniques. IEEE Trans. Signal Process. **46**(4), 845–856 (1998)
16. Haar, A.: Zur Theorie der orthogonalen Funktionensysteme. (Erste Mitteilung.). Math. Ann. **69**, 331–371 (1910)
17. Hereford, J.M., Roach, D.W., Pigford, R.: Image compression using parameterized wavelets with feedback. pp. 267–277. SPIE (2003)
18. Jorgensen, P.E.T.: Matrix factorizations, algorithms, wavelets. Notices Am. Math. Soc. **50**(8), 880–894 (2003)
19. Knuth, D.E.: The art of computer programming. Vol. 1: Fundamental algorithms, 3rd edn. Addison-Wesley, Reading. xx, p. 650 (1997)

20. Lai, M.J., Roach, D.W.: Parameterizations of univariate orthogonal wavelets with short support. In: Approximation Theory, X, St Louis, MO, 2001, Innov. Appl. Math., pp. 369–384. Vanderbilt University Press, Nashville, TN (2002)
21. Lawton, W.M.: Tight frames of compactly supported affine wavelets. J. Math. Phys. **31**(8), 1898–1901 (1990)
22. Lawton, W.M.: Necessary and sufficient conditions for constructing orthonormal wavelet bases. J. Math. Phys. **32**(1), 57–61 (1991)
23. Lebrun, J., Selesnick, I.: Gröbner bases and wavelet design. J. Symb. Comput. **37**(2), 227–259 (2004)
24. Lebrun, J., Vetterli, M.: High-order balanced multiwavelets: Theory, factorization, and design. IEEE Trans. Signal Process. **49**(9), 1918–1930 (2001)
25. Lina, J.M., Mayrand, M.: Parametrizations for Daubechies wavelets. Phys. Rev. E (3) **48**(6), R4160–R4163 (1993)
26. Mallat, S.: A Wavelet Tour of Signal Processing. Academic, San Diego (1998)
27. Park, H.: Symbolic computation and signal processing. J. Symb. Comput. **37**(2), 209–226 (2004)
28. Park, H., Kalker, T., Vetterli, M.: Gröbner bases and multidimensional FIR multirate systems. Multidimensional Syst. Signal Process. **8**(1–2), 11–30 (1997)
29. Pollen, D.: $\mathrm{SU}_I(2, F[z, 1/z])$ for F a subfield of $\mathbf{C}$. J. Am. Math. Soc. **3**(3), 611–624 (1990)
30. Regensburger, G., Scherzer, O.: Symbolic computation for moments and filter coefficients of scaling functions. Ann. Comb. **9**(2), 223–243 (2005)
31. Schneid, J., Pittner, S.: On the parametrization of the coefficients of dilation equations for compactly supported wavelets. Computing **51**(2), 165–173 (1993)
32. Selesnick, I.W., Burrus, C.S.: Maximally flat low-pass FIR filters with reduced delay. IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process. **45**(1), 53–68 (1998)
33. Shann, W.C., Yen, C.C.: On the exact values of orthonormal scaling coefficients of lengths 8 and 10. Appl. Comput. Harmon. Anal. **6**(1), 109–112 (1999)
34. Sherlock, B.G., Monro, D.M.: On the space of orthonormal wavelets. IEEE Trans. Signal Process. **46**(6), 1716–1720 (1998)
35. Strang, G., Nguyen, T.: Wavelets and filter banks. Wellesley–Cambridge Press, Wellesley (1996)
36. Unser, M., Blu, T.: Wavelet theory demystified. IEEE Trans. Signal Process. **51**(2), 470–483 (2003)
37. Wang, S.H., Tewfik, A.H., Zou, H.: Correction to 'parametrization of compactly supported orthonormal wavelets'. IEEE Trans. Signal Process. **42**(1), 208–209 (1994)
38. Wells, R.O. Jr.: Parametrizing smooth compactly supported wavelets. Trans. Am. Math. Soc. **338**(2), 919–931 (1993)
39. Zou, H., Tewfik, A.H.: Parametrization of compactly supported orthonormal wavelets. IEEE Trans. Signal Process. **41**(3), 1428–1431 (1993)

Radon Series Comp. Appl. Math **3**, 191–214

Applications of filter coefficients and wavelets parametrized by moments

Georg Regensburger

Key words. Orthonormal wavelets, parametrized filter coefficients, moments, regularity, Hölder and Sobolev exponent, least asymmetric filters, rational filter coefficients.

AMS classification. 42C40, 65T60, 94A12, 68W30

1 Introduction

Wavelets and their generalizations are used in many areas of mathematics ranging from harmonic analysis over numerical analysis to signal and image processing, see for example Daubechies [11], Mallat [29], and Strang and Nguyen [42]. A function $\psi \in L^2(\mathbb{R})$ is an *orthonormal wavelet* if the family

$$\psi_{jk}(x) = 2^{j/2}\psi(2^j x - k) \quad \text{for } j,k \in \mathbb{Z}$$

of translated an dilated versions of ψ is an orthonormal basis of the Hilbert space $L^2(\mathbb{R})$. Alfred Haar gave in his dissertation from 1909, published in [17], the first example of an orthonormal wavelet

$$\psi(x) = \begin{cases} 1, & \text{for } 0 \le x < \frac{1}{2}, \\ -1, & \text{for } \frac{1}{2} \le x < 1, \\ 0, & \text{otherwise,} \end{cases}$$

which is now known as the *Haar wavelet*. Daubechies introduced in her seminal paper [10] a general method to construct compactly supported wavelets. Her construction is

This work was supported by the Austrian Science Fund (FWF) under the SFB grant F1322.

based on *scaling functions*, satisfying a *dilation equation*

$$\phi(x) = \sum_{k=0}^{N} h_k \phi(2x - k) \tag{1.1}$$

given by a linear combination of real *filter coefficients* h_k and dilated and translated versions of the scaling function; see the next section for an outline.

Imposing conditions on the scaling function gives, via the dilation equation (1.1), constraints on the filter coefficients. Orthonormality implies quadratic equations and vanishing moments of the associated wavelet and normalization linear constraints. Daubechies wavelets [10] have the maximal number of vanishing moments for a fixed number of filter coefficients, and so there are only finitely many solutions. Parametrizing all possible filter coefficients that correspond to compactly supported orthonormal wavelets has been studied by several authors [21, 28, 33, 38, 41, 46, 47, 49]. All parametrizations express the filter coefficients in terms of trigonometric functions, and there is no natural interpretation of the angular parameters for the resulting scaling function. Furthermore, one has to solve transcendental constraints for the parameters to find wavelets with more than one vanishing moment.

We gave parametrizations of filter coefficients such that the corresponding wavelets have several vanishing moments and that use the first discrete moments as parameters first in [36] and then simplified in [35]. See section 3 for the parametrizations of four to eight filter coefficients with one parameter and at least one, two, and three vanishing moments, respectively. To compute these parametrizations we used symbolic computation and for the more involved equations in particular Gröbner bases, which were introduced by Buchberger in [3], see also [4]. Other applications of Gröbner bases to the design of wavelets and filter coefficients are for example discussed in [6, 7, 16, 25, 26, 31, 32, 39].

As a first application of parametrized wavelets, we discussed in [36] how they can by used for compression by computing an optimal parameter for a given signal, see also [18]. In this paper, we describe several other applications. In section 4, we discuss the regularity of the scaling functions and wavelets corresponding to our parametrizations. We construct wavelets that have a higher Hölder exponent than the Daubechies wavelets. Filter design is another possible application of our parametrizations. We deal with the construction of least asymmetric orthonormal wavelets in section 5. Finally, we address the existence of rational filter orthogonal filter coefficients in section 6. For example, we show that there are no orthogonal filters with six nonzero filter coefficients and at least two sum rules. A Maple worksheet with all computations, several MATLAB functions to produce the figures and a GUI to compute with and illustrate parametrized wavelets are available on request from the author.

2 Wavelets and moments

We outline the construction of orthonormal wavelets based on scaling functions and recall the polynomial equations for the filter coefficients, see for example Daubechies [11] or Strang and Nguyen [42].

Orthonormality of the integer translates $\{\phi(x-l)\}_{l\in\mathbb{Z}}$ in $L^2(\mathbb{R})$, that is,

$$\int \phi(x)\phi(x-l)dx = \delta_{0,l}$$

implies, using the dilation equation (1.1), the quadratic equations

$$\sum_{k\in\mathbb{Z}} h_k h_{k-2l} = 2\delta_{0,l} \quad \text{for } l \in \mathbb{Z} \tag{2.1}$$

where we set $h_k = 0$ for $k < 0$ and $k > N$. We can assume that $h_0 h_N \neq 0$. Then with equation (2.1) we see that N must be odd and the number of filter coefficients even.

If the filter coefficients satisfy the necessary conditions for orthogonality (2.1) and the normalization

$$\sum_{k=0}^{N} h_k = 2, \tag{2.2}$$

there exists a unique solution of the dilation equation (1.1) in $L^2(\mathbb{R})$ with support $[0, N]$ and for which $\int \phi = 1$, see Lawton [23]. For almost all such scaling functions the integer translates $\{\phi(x-l)\}_{l\in\mathbb{Z}}$ are orthogonal, and then

$$\psi(x) = \sum_{k=0}^{N} (-1)^k h_{N-k} \phi(2x-k) \tag{2.3}$$

is an orthonormal wavelet.

Necessary and sufficient conditions for orthonormality were given by Cohen [8] and Lawton [24], see also Daubechies [11, section 6.3]. The only example with four filter coefficients that satisfies the equations (2.1) and (2.2) and where the integer translates of the corresponding scaling are not orthogonal is $h_0 = h_3 = 1$ and $h_1 = h_2 = 0$ with the scaling function

$$\phi(x) = \begin{cases} 1/3, & \text{for } 0 \le x < 3, \\ 0, & \text{otherwise.} \end{cases} \tag{2.4}$$

The corresponding scaling function for the Haar wavelet is the box function

$$\phi(x) = \begin{cases} 1, & \text{for } 0 \le x < 1, \\ 0, & \text{otherwise,} \end{cases}$$

with the filter coefficients $h_0 = h_1 = 1$. In general, there is no closed analytic form for the scaling function, and for computations with scaling functions and wavelets only the filter coefficients are used.

Vanishing moments of the associated wavelet are related to several properties of the scaling function and wavelet. For example, to regularity, the polynomial reproduction and the approximation order of the scaling function, and the decay of the wavelet coefficients for smooth functions, see Strang and Nguyen [42] and the survey [43] by Unser and Blu for details. The condition that the first p moments of the wavelet ψ vanish

$$\int x^l \psi(x)\, dx = 0 \quad \text{for } \; l = 0, \ldots, p-1$$

is using equation (2.3) equivalent to the *sum rules*

$$\sum_{k=0}^{N}(-1)^k k^l h_k = 0 \quad \text{for } l = 0, \ldots, p-1. \tag{2.5}$$

One then says that ψ has p *vanishing moments* or the filter coefficients satisfy p sum rules.

Since we use *discrete moments*

$$m_n = \sum_{k=0}^{N} h_k k^n$$

of the filter coefficients as a parameters, we recall a well-known recursive relation between discrete and *continuous moments*

$$M_n = \int x^n \phi(x)\, dx$$

of the scaling function. Let ϕ be a scaling function satisfying $M_0 = \int \phi = 1$. Then $m_0 = 2$ and

$$M_n = \frac{1}{2^{n+1}-2} \sum_{i=1}^{n} \binom{n}{i} m_i M_{n-i},$$

$$m_n = \left(2^{n+1}-2\right) M_n - \sum_{i=1}^{n-1} \binom{n}{i} m_i M_{n-i} \quad \text{for } n > 0,$$

see for example Strang and Nguyen [42, p. 396]. Using the recursion we obtain for the first moments

$$M_1 = 1/2\, m_1,$$
$$M_2 = 1/6\, m_1^2 + 1/6\, m_2,$$
$$M_3 = 1/28\, m_1^3 + 1/7\, m_1 m_2 + 1/14\, m_3$$

and

$$m_1 = 2\, M_1,$$
$$m_2 = -4\, M_1^2 + 6\, M_2,$$
$$m_3 = 12\, M_1^3 - 24\, M_1 M_2 + 14\, M_3.$$

Explicit formulas expressing the discrete moments in terms of the continuous and vice versa are given in [36].

3 Parametrizations

We discuss the parametrizations from [35] of four, six, and eight filter coefficients corresponding respectively to orthonormal wavelets with at least one, two, and three

vanishing moments. All families depend on the first discrete moment

$$m = m_1 = \sum_{k=0}^{N} h_k k$$

of the filter coefficients.

3.1 Four filter coefficients

We have the following parametrization of filter coefficients with at least one vanishing moments:

$$\begin{aligned} h_0 &= 1/2 - 1/4\,a - 1/4\,w, \\ h_1 &= 1/2 - 1/4\,a + 1/4\,w, \\ h_2 &= 1/2 + 1/4\,a + 1/4\,w, \\ h_3 &= 1/2 + 1/4\,a - 1/4\,w \end{aligned} \tag{3.1}$$

with $w = \sqrt{4 - a^2}$ and $a = m - 3 \in [-2, 2]$.

Note that for $a = -a$ we obtain the flipped filter coefficients. For $a = 0$ we get the filter coefficients $(0, 1, 1, 0)$, which correspond to a translated Haar scaling function and wavelet. The parameter values $a = -2, 2$ give also Haar scaling functions with the filter coefficients $(1, 1, 0, 0)$ and $(0, 0, 1, 1)$. The *Daubechies wavelet* has two vanishing moments, so we have one more sum rule

$$2\,h_0 - h_1 + h_3 = 0.$$

Substituting the parametrized filter coefficients into this equations and solving for a, we get the two solutions $a = -\sqrt{3}, \sqrt{3}$ with the first discrete moments $m = 3 - \sqrt{3}, 3 + \sqrt{3}$. The first solution gives the famous Daubechies filters [10]

$$1/4\,(1 + \sqrt{3}, 3 + \sqrt{3}, 3 - \sqrt{3}, 1 - \sqrt{3}) \tag{3.2}$$

and the second the flipped version. See Figure 3.1 for plots of scaling functions for various parameter values.

We have a second parametrization of filter coefficients with at least one vanishing moment:

$$\begin{aligned} h_0 &= 1/2 - 1/4\,a + 1/4\,w, \\ h_1 &= 1/2 - 1/4\,a - 1/4\,w, \\ h_2 &= 1/2 + 1/4\,a - 1/4\,w, \\ h_3 &= 1/2 + 1/4\,a + 1/4\,w \end{aligned} \tag{3.3}$$

with $w = \sqrt{4 - a^2}$ and $a = m - 3 \in [-2, 2]$.

Comparing this solution with the parametrized filter coefficients (3.1), we see that w is replaced by $-w$ and so the two first and the two last filter coefficients are swapped. Note that again for $a = -a$ we obtain the flipped filters.

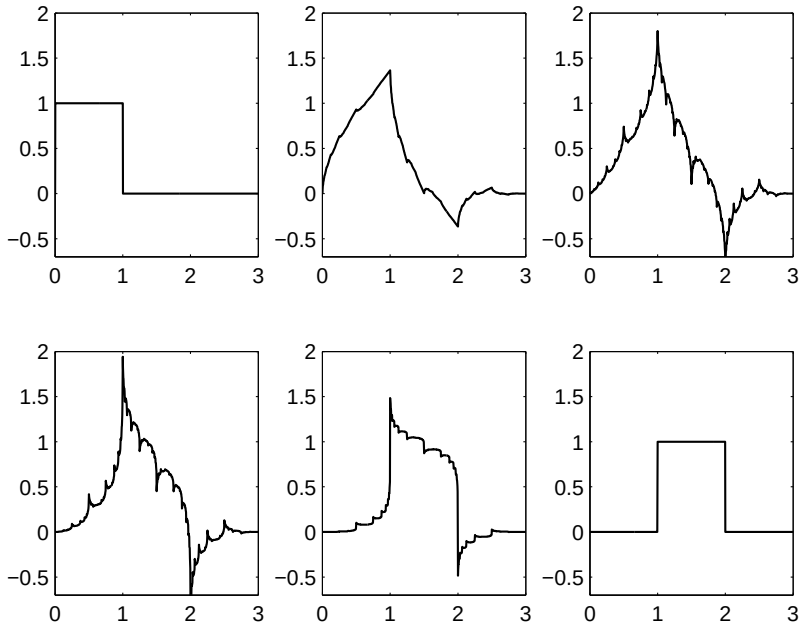


Figure 3.1: Scaling functions for $a = -2, -\sqrt{3}, -1/3\sqrt{3} - 2/3$ (first row) and $a = 1/3\sqrt{3} - 4/3, -2 + \sqrt{3}, 0$ (second row).

For $a = 0$ we now get the filter coefficients $(1, 0, 0, 1)$, which give the scaling function (2.4) where the integer translates of the scaling function are not orthogonal. The parameter values $a = -2, 2$ also give Haar scaling functions with the filter coefficients $(1, 1, 0, 0)$ and $(0, 0, 1, 1)$. This parametrization does not contain filter coefficients with a second vanishing moment. The corresponding scaling functions are, compared to the parametrization (3.1), irregular, see section 4 for details.

3.2 Six filter coefficients

We have the following parametrization of filter coefficients with with at least two vanishing moments:

$$\begin{aligned} h_0 &= -3/32 - 1/8\,a + 1/32\,a^2 - 1/32\,w, \\ h_1 &= 5/32 - 1/8\,a + 1/32\,a^2 + 1/32\,w, \\ h_2 &= 15/16 - 1/16\,a^2 + 1/16\,w \\ h_3 &= 15/16 - 1/16\,a^2 - 1/16\,w, \\ h_4 &= 5/32 + 1/8\,a + 1/32\,a^2 - 1/32\,w, \\ h_5 &= -3/32 + 1/8\,a + 1/32\,a^2 + 1/32\,w \end{aligned} \tag{3.4}$$

with $w = \sqrt{-a^4 + 14\,a^2 + 15}$ and $a = m - 5 \in [-\sqrt{15}, \sqrt{15}]$.

The Daubechies wavelet has one more vanishing moment, that is, it satisfies the sum rule

$$-9\,h_0 + 4\,h_1 - h_2 - h_4 + 4\,h_5 = 0.$$

Substituting the parametrized filter coefficients into this equations and solving for a, we get one real solution $a = -\sqrt{5 + 2\sqrt{10}}$, which gives the filter coefficients

$$\begin{aligned} 1/16\,(1 + \sqrt{10} + w, 5 + \sqrt{10} + 3\,w, 10 - 2\sqrt{10} + 2\,w, \\ 10 - 2\sqrt{10} - 2\,w, 5 + \sqrt{10} - 3\,w, 1 + \sqrt{10} - w) \end{aligned} \tag{3.5}$$

with $w = \sqrt{5 + 2\sqrt{10}}$. The Daubechies filters with four nonzero filter coefficients (3.2) satisfy two sum rules and are therefore contained in this parametrization. Their first discrete moment is $m = 3 - \sqrt{3}$. So here the corresponding parameter is $a = -2 - \sqrt{3}$. We get a translated version for $a = -\sqrt{3}$. For $a = -\sqrt{15}$ we obtain

$$1/8\,(3 + \sqrt{15}, 5 + \sqrt{15}, 0, 0, 5 - \sqrt{15}, 3 - \sqrt{15}).$$

The parameter $a = -1$ gives the first coiflet

$$1/16\,(1 - \sqrt{7}, 5 + \sqrt{7}, 14 + 2\sqrt{7}, 14 - 2\sqrt{7}, 1 - \sqrt{7}, -3 + \sqrt{7}),$$

see Daubechies [12] and [11, section 8.2]. For $a = 0$ we get

$$1/32\,(-3 - \sqrt{15}, 5 + \sqrt{15}, 30 + 2\sqrt{15}, 30 - 2\sqrt{15}, 5 - \sqrt{15}, -3 + \sqrt{15}).$$

See Figure 3.2 for plots of scaling functions for various parameter values. The corresponding scaling functions and wavelets for $a > 0$ become increasingly irregular, see section 4 for details.

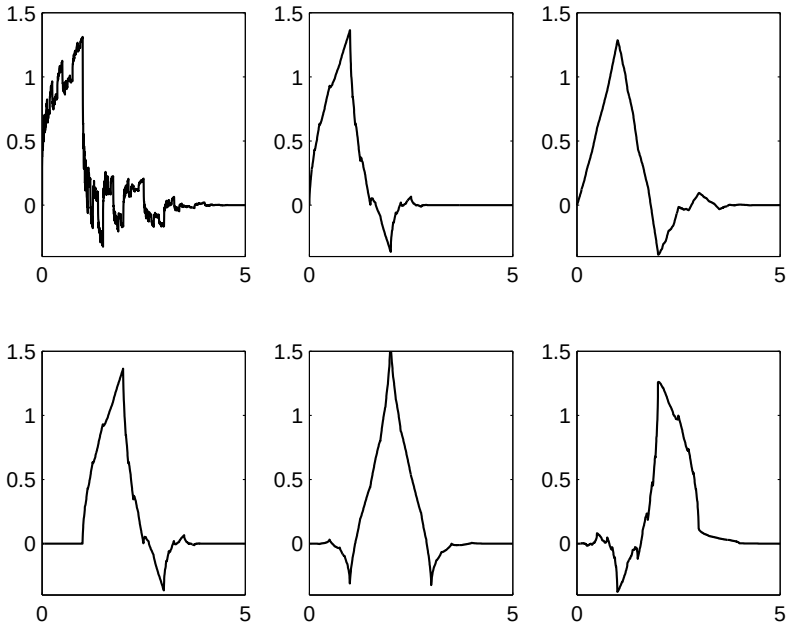


Figure 3.2: Scaling functions for $a = -\sqrt{15}, -2-\sqrt{3}, -\sqrt{5+2\sqrt{10}}$ (first row) and $a = -\sqrt{3}, -1, 0$ (second row).

3.3 Eight filter coefficients

We have the following parametrization of filter coefficients with at least three vanishing moments:

$$
\begin{aligned}
h_0 &= -\frac{1}{512}\,\frac{a^5 - 7\,a^4 - 2\,a^3 + 30\,a^2 - 55\,a - 15 + (1-a)w}{a^2+1},\\
h_1 &= -\frac{1}{512}\,\frac{a^5 - 9\,a^4 + 30\,a^3 + 2\,a^2 - 23\,a + 63 + (1+a)w}{a^2+1},\\
h_2 &= \frac{1}{512}\,\frac{3\,a^5 - 5\,a^4 - 102\,a^3 + 186\,a^2 - 261\,a + 35 + 3(1-a)w}{a^2+1},\\
h_3 &= \frac{1}{512}\,\frac{3\,a^5 - 11\,a^4 - 70\,a^3 + 358\,a^2 - 229\,a + 525 + 3(1+a)w}{a^2+1},\\
h_4 &= -\frac{1}{512}\,\frac{3\,a^5 + 11\,a^4 - 70\,a^3 - 358\,a^2 - 229\,a - 525 + 3(1-a)w}{a^2+1},\\
h_5 &= -\frac{1}{512}\,\frac{3\,a^5 + 5\,a^4 - 102\,a^3 - 186\,a^2 - 261\,a - 35 + 3(1+a)w}{a^2+1},\\
h_6 &= \frac{1}{512}\,\frac{a^5 + 9\,a^4 + 30\,a^3 - 2\,a^2 - 23\,a - 63 + (1-a)w}{a^2+1},\\
h_7 &= \frac{1}{512}\,\frac{a^5 + 7\,a^4 - 2\,a^3 - 30\,a^2 - 55\,a + 15 + (1+a)w}{a^2+1}
\end{aligned}
\tag{3.6}
$$

with

$$w = \sqrt{-a^8 + 36\,a^6 - 182\,a^4 + 1540\,a^2 - 945},$$

$a = m - 7$ and a in the intervals

$$[-\sqrt{\beta}, -\sqrt{\alpha}] \quad \text{or} \quad [\sqrt{\alpha}, \sqrt{\beta}],$$

where α denotes the smaller and β the larger real root of

$$x^4 - 36\,x^3 + 182\,x^2 - 1540\,x + 945,$$

with numerical approximations

$$\sqrt{\alpha} = 0.8113601077\ldots \quad \text{and} \quad \sqrt{\beta} = 5.636256558\ldots\,.$$

The Daubechies wavelet satisfies one more sum rule

$$64\,h_0 - 27\,h_1 + 8\,h_2 - h_3 + h_5 - 8\,h_6 + 27\,h_7 = 0.$$

Substituting the parametrized filter coefficients (3.6) into this equations and solving for a, we get two real solution $a = -\sqrt{\beta}, -\sqrt{\alpha}$, where α denotes the smaller and β the larger real root of

$$x^4 - 28\,x^3 + 126\,x^2 - 1260\,x + 1225$$

or numerically

$$a = -4.989213573\ldots, -1.029063869\ldots\ .$$

The first parameter gives the Daubechies wavelet with extremal phase [11, p. 195] and the second the "least asymmetric" [11, p. 198]. The Daubechies wavelet with six nonzero filter coefficients (3.5) has the first discrete moment

$$m = 5 - \sqrt{5 + 2\sqrt{10}},$$

so the corresponding parameter value for the parametrization (3.6) is

$$a = -2 - \sqrt{5 + 2\sqrt{10}} = -5.365197664\ldots\ .$$

See Figure 3.3 for plots of scaling functions for various parameter values.

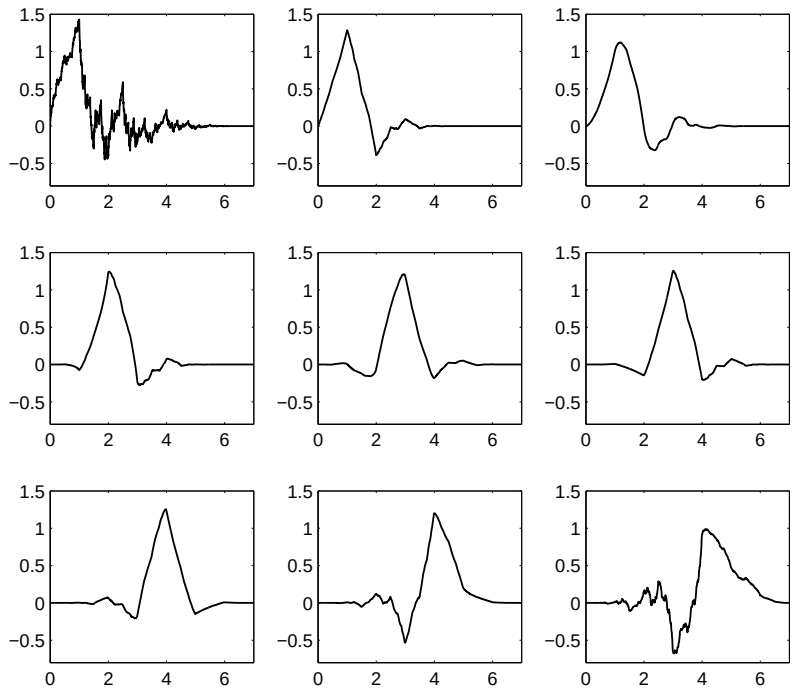


Figure 3.3: Scaling functions for $a = -5.636256559, -5.365197664, -4.989213573$ (first row), $a = -3.009138721, -1.029063869, -0.8113601077$ (second row), and $a = 0.8113601077, 2, 3$ (third row).

4 Regularity of scaling functions and wavelets

In this section, we discuss the regularity or smoothness of the scaling functions and wavelets corresponding to the parametrized filter coefficients from the previous section. The regularity of a function can be measured in different ways, we consider here the Hölder and Sobolev exponent.

We first recall the definitions. For $\alpha = n + \beta$, where $n \in \mathbb{N}$ and $0 \leq \beta < 1$, the set $C^\alpha = C^\alpha(\mathbb{R})$ is defined as the set of all functions f that are n times continuously differentiable and such that the nth derivative $f^{(n)}$ is uniformly *Hölder continuous* with exponent β, that is,

$$|f^{(n)}(x+h) - f^{(n)}(x)| \leq C|h|^\beta \quad \text{for all } x, h \in \mathbb{R}$$

where C is a constant. For $s \geq 0$ the *Sobolev space* $H^s = H^s(\mathbb{R})$ consists of all functions $f \in L^2(\mathbb{R})$ such that $(1 + |\xi|^2)^{s/2}\hat{f}(\xi) \in L^2(\mathbb{R})$, where $\hat{f}$ denotes the Fourier transform of f.

To measure the regularity or smoothness of a scaling function ϕ, one is interested respectively in the (optimal) *Sobolev*

$$s_{\max} = \sup\{s : \phi \in H^s\}$$

and *Hölder exponent*

$$\alpha_{\max} = \sup\{\alpha : \phi \in C^\alpha\}.$$

For a scaling function the Hölder exponent satisfies [44]

$$\alpha_{\max} \in [s_{\max} - 1/2, s_{\max}]. \tag{4.1}$$

The regularity of scaling functions is also related to vanishing moments of the corresponding wavelet. Villemoes [44] proved that if $\phi \in H^n$ with $n \in \mathbb{N}$, the filter coefficients satisfy $n+1$ sum rules or equivalently the corresponding wavelet has $n+1$ vanishing moments. So in particular if $\phi \in C^n$, then the filter coefficients satisfy $n+1$ sum rules, see also [11, pp. 153–156].

Eirola [14] and Villemoes [44] independently showed how the optimal Sobolev exponent can be computed from the spectral radius of a matrix depending on the filter coefficients, see also Strang and Nguyen [42] for further details. To find the optimal Hölder exponent is much more involved, see for example [9, 11, 13, 37], but Rioul [37] gave an algorithm to compute good lower bounds for the Hölder exponent. The algorithm produces monotonically increasing lower bounds with an increasing number of iterations, but the storage and computational costs approximately double for each additional iteration.

In Figures 4.1, 4.2 and 4.3 you can see plots of the Sobolev exponent of the corresponding scaling functions and wavelets depending on one parameter. For four filter coefficients the Sobolev exponents range from 0.5 to 1 (parametrization (3.1)) and from 0 to 0.5 (parametrization (3.3)). The maximum 1 is attained for the Daubechies wavelet since all other filter coefficients satisfy only one sum rule and hence their Sobolev exponent is necessarily less than one. We obtain numerically the maximal Sobolev exponent

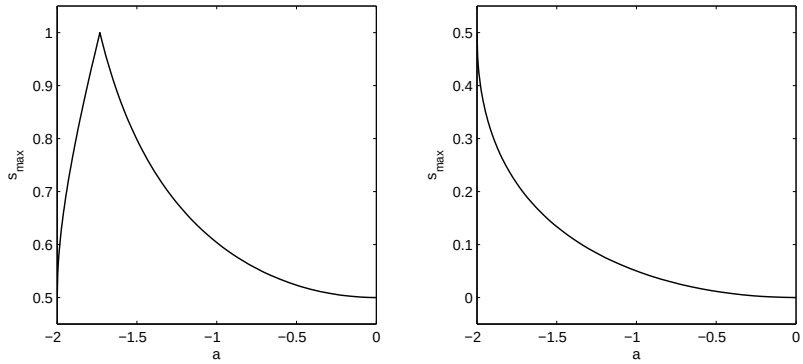


Figure 4.1: Sobolev exponent for scaling functions with four filter coefficients from equation (3.1) (left) and (3.3) (right).

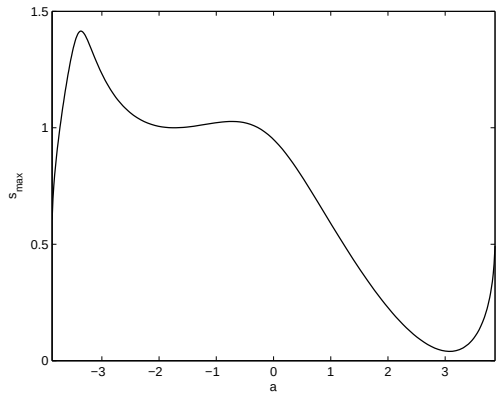


Figure 4.2: Sobolev exponent for scaling functions with six filter coefficients from equation (3.4).

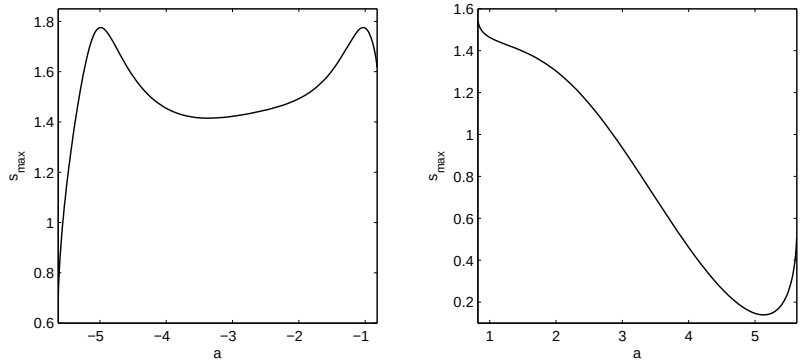


Figure 4.3: Sobolev exponent for scaling functions with eight filter coefficients from equation (3.6).

for respectively six and eight filter coefficients

$$s_{\max} = 1.4150, 1.7756,$$

at the parameter values for the Daubechies wavelets and the minimum is

$$s_{\max} = 0.0399, 0.1393$$

with parameter values

$$a = 3.077681946, 5.131603420.$$

For more than six filter coefficients it is possible to construct wavelets with a higher Sobolev exponents than the Daubechies wavelets by omitting more than one sum rule, see [27, 30, 45].

In Figures 4.4, 4.5 and 4.6 you can see plots of lower bounds for the Hölder exponent of the corresponding scaling functions and wavelets depending on one parameter, with the bounds from equation (4.1). We used 24 iteration in the algorithm from [37].

Note that for most, and for eight filter coefficients for all, parameters the computed lower bound is higher than the lower bound $s_{\max} - 1/2$. The negative lower bound in Figure 4.5 indicates that the corresponding scaling function is not continuous. We obtain numerically the maximal lower bound for the Hölder exponent for respectively four, six and eight filter coefficients

$$\alpha_{24} = 0.5776, 1.1386, 1.6344$$

with parameters

$$a = -1.66260325442517, -3.28211108661493, -4.93905744197576$$

and filter coefficients

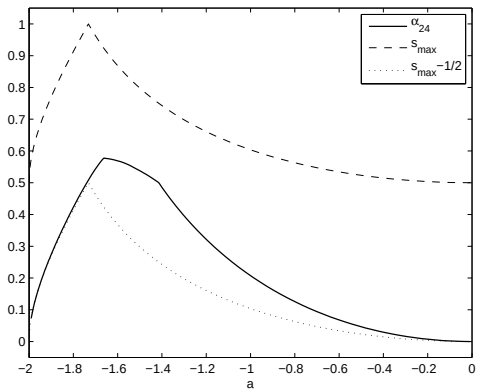


Figure 4.4: Lower bound for Hölder exponent for scaling functions with four filter coefficients from equation (3.1).

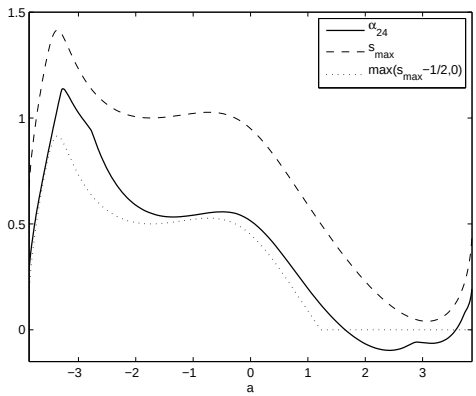


Figure 4.5: Lower bound for Hölder exponent for scaling functions with six filter coefficients from equation (3.4).

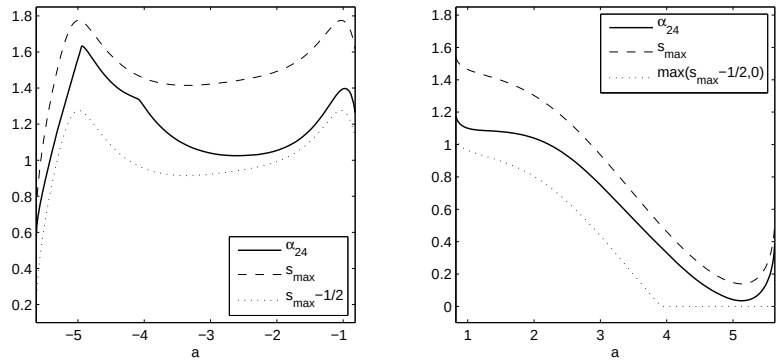


Figure 4.6: Lower bound for Hölder exponent for scaling functions with eight filter coefficients from equation (3.6).

```
 0.31887001724554, 0.59678079636075,
 0.18112998275446,-0.09678079636075

 0.21634225649014, 0.56180454136425, 0.35257937284659,
-0.08834519690163,-0.06892162933673, 0.02654065553738

 0.15488273436983, 0.49644876596501, 0.45767418856225,
-0.00833281609981,-0.13761439998701, 0.01970151455156,
 0.02505747705493,-0.00781746441676.
```

Daubechies and Lagarias [13] obtained the optimal Hölder exponents for the Daubechies wavelets with a different method (four, six, and eight filter coefficients)

$$\alpha_{\max} = 0.5500, 1.0878, 1.6179,$$

where the last one is for the Daubechies wavelet with extremal phase. So we obtained in all cases wavelets that have a higher Hölder exponent than the Daubechies wavelets.

Daubechies addressed in [12] and [11, p. 242] the question of finding wavelets with more regularity. For four filter coefficients she obtained the rational filter coefficients $(3/5, 6/5, 2/5, -1/5)$, which corresponds to $a = -8/5$ in (3.1), see also section 6. With the methods from [13] she found that the Hölder exponent of the corresponding scaling function is at least 0.5864.

Lang and Heller [22] discussed the general optimization problem of maximizing the Hölder exponent for a fixed number of filter coefficients. They found smoother wavelets than the Daubechies wavelets for more than eight filter coefficients, but the numerical method failed to find the more regular wavelets that we obtained using the explicit parametrizations of the filter coefficients. This might be due to the fact that Lang and Heller used a general purpose optimization routine while we could directly apply the golden section search for finding the maximum of a univariate function.

5 Least asymmetric filters

It is well known [11, p. 252] that if a compactly supported orthonormal wavelet is symmetric or antisymmetric around some axis, then it is the Haar wavelet. Symmetry of the scaling function is in turn equivalent to symmetry of the filter coefficients, see Belogay and Wang [2] and also Daubechies [12]. Here we say that the filter coefficients are *symmetric* around $n_0 \in \mathbb{Z}/2$ if

$$h_n = h_{2n_0-n},$$

where we set $h_k = 0$ for $k < 0$ and $k > N$. Symmetric filters are often called *linear phase filters* since the filter coefficients are symmetric around $n_0 \in \mathbb{Z}/2$ if and only if the phase of the *frequency response*

$$h(\xi) = \sum_n h_n e^{in\xi}$$

is a linear function of ξ, that is, if

$$h(\xi) = e^{in_0\xi}|h(\xi)|.$$

So we know that complete symmetry and orthogonality is not possible, and one can only try to find the least asymmetric filter coefficients out of a fixed family. For example, Daubechies discussed in [11] and [12] how to choose the least asymmetric out of the finitely many wavelets with a maximal number of vanishing moments. Another possibility is to omit some vanishing moments and use the additional degrees of freedom to find filters with partial symmetry. Several authors [1, 25, 40] discussed the use of Gröbner bases to find orthogonal filter coefficients with partial symmetry where several pairs of filters are equal. Zhao and Swamy [48] designed least asymmetric orthogonal wavelets with several vanishing moments via numerical optimization.

An immediate application of our parametrized filter coefficients is to find symbolically the least asymmetric filter coefficients using some criteria to measure symmetry. In the following, we discuss some examples, where we minimize the sum of squares error as in [48].

We want to find six filter coefficients satisfying two sum rules such that they are almost symmetric around 2, so that

$$h_0 \approx h_4, \quad h_1 \approx h_3, \quad h_6 \approx 0.$$

Using Maple, we find the minimum of the sum of squares error is attained at $a = \alpha$, where α denotes the largest negative real root of

$$25\,x^{10}-30\,x^9-702\,x^8+652\,x^7+5866\,x^6-3256\,x^5-13140\,x^4-1036\,x^3+5797\,x^2-2730\,x-5190$$

or numerically

$$a = -1.102986298\ldots\ .$$

The filter coefficients are:

```
-0.090589559870111, 0.504872307867382, 1.206925694336121,
 0.516001958861136,-0.116336134466010,-0.020874266728517.
```

See Figure 5.1 for the corresponding scaling function, which has a Sobolev exponent $s_{\max} = 1.0180$ and a lower bound for the Hölder exponent $\alpha_{24} = 0.5370$.

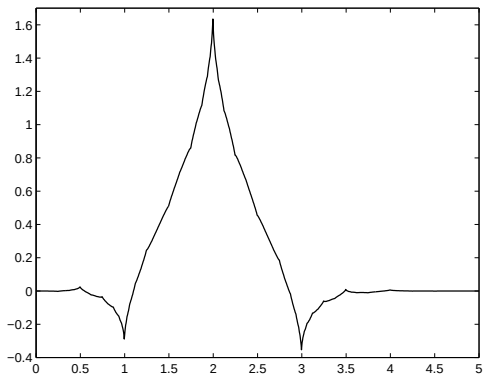


Figure 5.1: Least asymmetric (around 2) scaling function with six filter coefficients and two sum rules.

Now we consider eight filter coefficients. First we want to find filter coefficients that are almost symmetric around 3, so that

$$h_0 \approx h_6, \quad h_1 \approx h_5, \quad h_2 \approx h_4, \quad h_7 \approx 0.$$

The minimum of the sum of squares error is attained at $a = \alpha$, where α denotes the largest negative real root of

$$11025\,x^{24} - 21000\,x^{23} - 901900\,x^{22} + 1407480\,x^{21} + 25484946\,x^{20} - 23935800\,x^{19} - 280989500\,x^{18}$$
$$- 149785464\,x^{17} + 837190927\,x^{16} + 6460372400\,x^{15} + 4612440168\,x^{14} - 53422512976\,x^{13}$$
$$- 69302308420\,x^{12} + 344858640016\,x^{11} - 84085760856\,x^{10} - 294800719088\,x^{9} + 2435452393919\,x^{8}$$
$$- 1913025285928\,x^{7} - 18887356576348\,x^{6} + 10024351195096\,x^{5} + 51733811048402\,x^{4}$$
$$- 17259269191640\,x^{3} - 57876449779820\,x^{2} + 8466676099560\,x + 21625605062145$$

or numerically

$$a = -0.8395579286\ldots\ .$$

The filter coefficients are:

```
-0.073484394510424,-0.071424517120364, 0.556147092523951,
 1.154912201440016, 0.568048480655853,-0.135661369346454,
-0.050711178669381, 0.052173685026802.
```

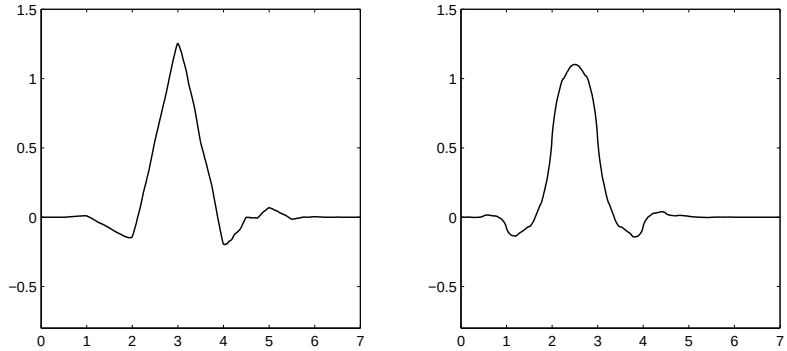


Figure 5.2: Least asymmetric (around 3 left and 2.5 right) scaling function with eight filter coefficients and three sum rules.

See Figure 5.2 (left) for the corresponding scaling function, which has a Sobolev exponent $s_{\max} = 1.6569$ and a lower bound for the Hölder exponent $\alpha_{24} = 1.3080$.

Finally, we want to design filters that are almost symmetric around 2.5, so that

$$h_0 \approx h_5, \quad h_1 \approx h_4, \quad h_2 \approx h_3, \quad h_6 \approx 0, \quad h_7 \approx 0.$$

This is related to the example considered in [1, 25], where the authors constructed using Gröbner bases eight orthogonal filters with two sum rules such that $h_0 = h_5$, $h_1 = h_4$ and $h_2 = h_3$. The minimum of the sum of squares error is attained at $a = \alpha$, where α denotes the second largest negative real root of

$$\begin{aligned}
&2025\,x^{24} - 9000\,x^{23} - 168020\,x^{22} + 823000\,x^{21} + 4733434\,x^{20} - 27869720\,x^{19} - 46538164\,x^{18} \\
&+ 437384872\,x^{17} - 40684609\,x^{16} - 3591330192\,x^{15} + 3105046936\,x^{14} + 20835868016\,x^{13} \\
&- 35438686580\,x^{12} - 64147246896\,x^{11} + 233849168056\,x^{10} - 48135550128\,x^{9} - 894126414729\,x^{8} \\
&+ 1033511750456\,x^{7} + 2682874758716\,x^{6} - 4634966862792\,x^{5} - 4762513155302\,x^{4} \\
&+ 10857513198280\,x^{3} + 182957235580\,x^{2} - 6268723929720\,x + 2258107786305
\end{aligned}$$

or numerically

$$a = -1.927469761\ldots\ .$$

The filter coefficients are:

```
-0.114678365799638, 0.127976021526492, 0.977783792709255,
 0.990754350911186, 0.120334952341046,-0.133569326041206,
 0.016559620749336, 0.014838953603528.
```

See Figure 5.2 (right) for the corresponding scaling function, which has a Sobolev exponent $s_{\max} = 1.5026$ and a lower bound for the Hölder exponent $\alpha_{24} = 1.0633$.

6 Rational filter coefficients

In this section, we address the existence of rational orthogonal filter coefficients. We know from section 2 that filter coefficients are determined by quadratic equations for orthonormality (2.1) and linear equations for normalization (2.2) and vanishing moments (2.5). Note that all these equations have integer coefficients, and we want to find a rational solution. This leads to "Hilbert's 10th Problem over $\mathbb{Q}$", which asks if there exists an algorithm for deciding the existence of rational points for a system of polynomial equations with integer coefficients. The answer is not known, and despite centuries of effort, even for curves it is an open problem although many results and computational methods are known, see for example Poonen [34] for an introduction and further references. Using our parametrizations, we can reduce the question of rational filter coefficients to finding rational points on curves and give some answers.

The case of four filter coefficients is not difficult. Daubechies [10] already gave a rational parametrization of all orthogonal filter coefficients

$$h_0 = \frac{t\,(t-1)}{t^2+1}, \quad h_1 = \frac{1-t}{t^2+1}, \quad h_2 = \frac{t+1}{t^2+1}, \quad h_3 = \frac{t\,(t+1)}{t^2+1}$$

with $t \in \mathbb{R}$. Note that for $t = -t$ we obtain the flipped filter coefficients. The interval $-1 \le t \le 1$ corresponds to the filter coefficients from (3.1) and $t \le -1, 1 \le t$ to (3.3), except for $(1, 0, 0, 1)$, which are approached for $t \to \infty$ and $t \to -\infty$.

The Daubechies wavelet corresponds to $t = -1/\sqrt{3}$. Computing the continued fraction expansion of $-1/\sqrt{3}$, we obtain the periodic expansion

$$-\frac{1}{\sqrt{3}} = [-1;\, 2,\, \overline{2,\, 1}\,] = -1 + \cfrac{1}{2 + \cfrac{1}{2 + \cfrac{1}{1 + \cfrac{1}{2 + \cfrac{1}{1 + \cdots}}}}}$$

with the first convergents

$$-1,\ -1/2,\ -3/5,\ -4/7,\ -\frac{11}{19},\ -\frac{15}{26},\ -\frac{41}{71},\ -\frac{56}{97},\ -\frac{153}{265},\ -\frac{209}{362}.$$

For further details on continued fractions see for example Khinchin [19] or Knuth [20].

Taking $t = -209/362$, we get a good rational approximation

$$1/174725\,(119339, 206702, 55386, -31977)$$

for the Daubechies filters. Surprisingly, we obtain the filter coefficients corresponding to the most regular scaling function found by Daubechies for the second convergent $t = -1/2$, see section 4.

In parametrization (3.4) for six filter coefficients there appears only the square root

$$w = \sqrt{-a^4 + 14\,a^2 + 15}.$$

So the question of the existence of rational filters reduces to finding a rational point $(a, b) \in \mathbb{Q}^2$ on the (hyperelliptic) algebraic curve defined by the equation

$$y^2 = -x^4 + 14\,x^2 + 15 = -(x^2 + 1)(x^2 - 15). \tag{6.1}$$

Proposition 6.1 *There are no rational points on the curve defined by equation* (6.1).

Proof. Substituting $x = X/Z$ and $y = Y/Z^2$ in (6.1) and multiplying by Z^4, we obtain

$$Y^2 = -(X^2 + Z^2)(X^2 - 15\,Z^2),$$

and we equivalently would have to find integers a, b, c with a and c coprime satisfying this equation. Suppose that we had integers a, b, c satisfying

$$b^2 = -(a^2 + c^2)(a^2 - 15\,c^2). \tag{6.2}$$

Then

$$b^2 \equiv (a^2 + c^2)^2 \pmod 2$$

and hence

$$b \equiv (a + c) \pmod 2.$$

This implies that either

$$a \equiv 1,\ c \equiv 0 \pmod 2 \quad \text{or} \quad a \equiv 0,\ c \equiv 1 \pmod 2$$

or, since a and c are coprime,

$$a \equiv c \equiv 1 \pmod 2.$$

In the first case, we get

$$(a^2 + c^2)^2 \equiv 1 \pmod 4.$$

But then by equation (6.2)

$$b^2 \equiv -1 \equiv 3 \pmod 4,$$

which is not possible since the only quadratic residues modulo 4, that is, the integers d for which

$$x^2 \equiv d \pmod 4$$

has a solution, are

$$d \equiv 0, 1 \pmod 4.$$

In the second case, we get

$$(a^2 + c^2)^2 \equiv 4 \pmod{16}.$$

But then by equation (6.2)

$$b^2 \equiv -4 \equiv 12 \pmod{16},$$

which is not possible since the only quadratic residues modulo 16 are

$$d \equiv 0, 1, 4, 9 \pmod{16},$$

and the proposition is proved. □

Corollary 6.2 *There are no rational orthogonal filters with six nonzero filter coefficients and at least two sum rules.*

In parametrization (3.6) for eight filter coefficients, we have the square root

$$w = \sqrt{-a^8 + 36\,a^6 - 182\,a^4 + 1540\,a^2 - 945}.$$

So we would have to find a rational point on the algebraic curve defined by the equation

$$y^2 = -x^8 + 36\,x^6 - 182\,x^4 + 1540\,x^2 - 945.$$

This is a nonsingular curve with genus 3. Hence by Falting's theorem [15] it has only finitely many rational points, and so there are at most finitely many rational orthogonal filters with eight nonzero filter coefficients and at least three sum rules. So far we could neither find rational points on this curve nor prove that there do not exist any.

Acknowledgements. I would like to thank Josef Schicho for his comments and help with the proof of Proposition 6.1 and the reviewers for useful remarks.

Bibliography

[1] A. F. Abdelnour and I. W. Selesnick, *Symmetric nearly orthogonal and orthogonal nearly symmetric wavelets*, Arab. J. Sci. Eng. Sect. C Theme Issues 29 (2004), pp. 3–16.

[2] E. Belogay and Y. Wang, *Compactly supported orthogonal symmetric scaling functions*, Appl. Comput. Harmon. Anal. 7 (1999), pp. 137–150.

[3] B. Buchberger, *An Algorithm for Finding the Bases Elements of the Residue Class Ring Modulo a Zero Dimensional Polynomial Ideal (German)*, Ph.D. thesis, Univ. of Innsbruck, 1965, English translation published in [5].

[4] ———, *Ein algorithmisches Kriterium für die Lösbarkeit eines algebraischen Gleichungssystems*, Aequationes Math. 4 (1970), pp. 374–383.

[5] ———, *Bruno Buchberger's PhD thesis 1965: An algorithm for finding the basis elements of the residue class ring of a zero dimensional polynomial ideal*, J. Symbolic Comput. 41 (2006), pp. 475–511, Translated from the 1965 German original by Michael P. Abramson.

[6] C. Charoenlarpnopparut and N. K. Bose, *Multidimensional FIR filter bank design using Gröbner bases.*, IEEE Trans. Circuits Syst., II, Analog Digit. Signal Process. 46 (1999), pp. 1475–1486.

[7] F. Chyzak, P. Paule, O. Scherzer, A. Schoisswohl, and B. Zimmermann, *The construction of orthonormal wavelets using symbolic methods and a matrix analytical approach for wavelets on the interval*, Experiment. Math. 10 (2001), pp. 67–86.

[8] A. Cohen, *Ondelettes, analyses multirésolutions et filtres miroirs en quadrature*, Ann. Inst. H. Poincaré Anal. Non Linéaire 7 (1990), pp. 439–459.

[9] D. Colella and C. Heil, *Characterizations of scaling functions: continuous solutions*, SIAM J. Matrix Anal. Appl. 15 (1994), pp. 496–518.

[10] I. Daubechies, *Orthonormal bases of compactly supported wavelets*, Comm. Pure Appl. Math. 41 (1988), pp. 909–996.

[11] ———, *Ten lectures on wavelets*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1992.

[12] ———, *Orthonormal bases of compactly supported wavelets. II. Variations on a theme*, SIAM J. Math. Anal. 24 (1993), pp. 499–519.

[13] I. Daubechies and J. C. Lagarias, *Two-scale difference equations. II. Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal. 23 (1992), pp. 1031–1079.

[14] T. Eirola, *Sobolev characterization of solutions of dilation equations*, SIAM J. Math. Anal. 23 (1992), pp. 1015–1030.

[15] G. Faltings, *Endlichkeitssätze für abelsche Varietäten über Zahlkörpern*, Invent. Math. 73 (1983), pp. 349–366.

[16] J.-C. Faugère, F. Moreau de Saint-Martin, and F. Rouillier, *Design of regular nonseparable bidimensional wavelets using Gröbner basis techniques*, IEEE Trans. Signal Process. 46 (1998), pp. 845–856.

[17] A. Haar, *Zur Theorie der orthogonalen Funktionensysteme. (Erste Mitteilung.)*, Math. Ann. 69 (1910), pp. 331–371.

[18] J. M. Hereford, D. W. Roach, and R. Pigford, *Image compression using parameterized wavelets with feedback*, vol. 5102, SPIE, 2003, pp. 267–277.

[19] A. Ya. Khinchin, *Continued fractions*, russian. ed., Dover Publications Inc., Mineola, NY, 1997, With a preface by B. V. Gnedenko, Reprint of the 1964 translation.

[20] D. E. Knuth, *The art of computer programming. Vol. 2: Seminumerical algorithms. 3rd ed.*, Bonn: Addison-Wesley. xiii, 1998.

[21] M.-J. Lai and D. W. Roach, *Parameterizations of univariate orthogonal wavelets with short support*, Approximation theory, X (St. Louis, MO, 2001), Innov. Appl. Math., Vanderbilt Univ. Press, Nashville, TN, 2002, pp. 369–384.

[22] M. Lang and P. N. Heller, *The design of maximally smooth wavelets*, IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing (Atlanta, GA), vol. 3, 1996, pp. 1463–1466.

[23] W. M. Lawton, *Tight frames of compactly supported affine wavelets*, J. Math. Phys. 31 (1990), pp. 1898–1901.

[24] ———, *Necessary and sufficient conditions for constructing orthonormal wavelet bases*, J. Math. Phys. 32 (1991), pp. 57–61.

[25] J. Lebrun and I. W. Selesnick, *Gröbner bases and wavelet design*, J. Symbolic Comput. 37 (2004), pp. 227–259.

[26] J. Lebrun and M. Vetterli, *High-order balanced multiwavelets: theory, factorization, and design*, IEEE Trans. Signal Process. 49 (2001), pp. 1918–1930.

[27] P. G. Lemarie-Rieusset and E. Zahrouni, *More regular wavelets*, Appl. Comput. Harmon. Anal. 5 (1998), pp. 92–105.

[28] J.-M. Lina and M. Mayrand, *Parametrizations for Daubechies wavelets*, Phys. Rev. E (3) 48 (1993), pp. R4160–R4163.

[29] S. Mallat, *A wavelet tour of signal processing*, Academic Press Inc., San Diego, CA, 1998.

[30] H. Ojanen, *Orthonormal compactly supported wavelets with optimal Sobolev regularity: Numerical results.*, Appl. Comput. Harmon. Anal. 10 (2001), pp. 93–98.

[31] H. Park, *Symbolic computation and signal processing*, J. Symbolic Comput. 37 (2004), pp. 209–226.

[32] H. Park, T. Kalker, and M. Vetterli, *Gröbner bases and multidimensional FIR multirate systems.*, Multidimensional Syst. Signal Process. 8 (1997), pp. 11–30.

[33] D. Pollen, $\mathrm{SU}_I(2, F[z, 1/z])$ *for* F *a subfield of* $\mathbf{C}$, J. Amer. Math. Soc. 3 (1990), pp. 611–624.

[34] B. Poonen, *Computing rational points on curves*, Number theory for the millennium, III (Urbana, IL, 2000), A K Peters, Natick, MA, 2002, pp. 149–172.

[35] G. Regensburger, *Parametrizing compactly supported orthonormal wavelets by discrete moments*, Appl. Algebra Engrg. Comm. Comput. (2007), to appear.

[36] G. Regensburger and O. Scherzer, *Symbolic computation for moments and filter coefficients of scaling functions*, Ann. Comb. 9 (2005), pp. 223–243.

[37] O. Rioul, *Simple regularity criteria for subdivision schemes*, SIAM J. Math. Anal. 23 (1992), pp. 1544–1576.

[38] J. Schneid and S. Pittner, *On the parametrization of the coefficients of dilation equations for compactly supported wavelets*, Computing 51 (1993), pp. 165–173.

[39] I. W. Selesnick and C. S. Burrus, *Maximally Flat Low-Pass FIR Filters with Reduced Delay*, IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process. 45 (1998), pp. 53–68.

[40] I. W. Selesnick, J. E. Odegard, and C. S. Burrus, *Nearly symmetric orthogonal wavelets with non-integer DC groupdelay*, Digital Signal Processing Workshop Proceedings, IEEE, September 1–4 1996, pp. 431–434.

[41] B. G. Sherlock and D. M. Monro, *On the space of orthonormal wavelets*, IEEE Trans. Signal Process. 46 (1998), pp. 1716–1720.

[42] G. Strang and T. Nguyen, *Wavelets and filter banks*, Wellesley-Cambridge Press, Wellesley, MA, 1996.

[43] M. Unser and T. Blu, *Wavelet theory demystified*, IEEE Trans. Signal Process. 51 (2003), pp. 470–483.

[44] L. F. Villemoes, *Energy moments in time and frequency for two-scale difference equation solutions and wavelets*, SIAM J. Math. Anal. 23 (1992), pp. 1519–1543.

[45] H. Volkmer, *Asymptotic regularity of compactly supported wavelets*, SIAM J. Math. Anal. 26 (1995), pp. 1075–1087.

[46] S. H. Wang, A. H. Tewfik, and H. Zou, *Correction to 'parametrization of compactly supported orthonormal wavelets'*, IEEE Trans. Signal Process. 42 (1994), pp. 208–209.

[47] R. O. Wells, Jr., *Parametrizing smooth compactly supported wavelets*, Trans. Amer. Math. Soc. 338 (1993), pp. 919–931.

[48] Y. Zhao and M. N. S. Swamy, *Design of least asymmetric compactly supported orthogonal wavelets via optimization*, Electrical and Computer Engineering, 1999 IEEE Canadian Conference on, vol. 2, May 1999, pp. 817–820.

[49] H. Zou and A. H. Tewfik, *Parametrization of compactly supported orthonormal wavelets*, IEEE Trans. Signal Process. 41 (1993), pp. 1428–1431.

Author information

Georg Regensburger, Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences,
Altenbergerstraße 69, A-4040 Linz, Austria.
Email: georg.regensburger@oeaw.ac.at