博士論文

Computing Valuations of Determinants
via Combinatorial Optimization:
Applications to Differential Equations

（組合せ最適化による行列式の付値計算：
微分方程式への応用）

Taihei Oki　（大城 泰平）

Computing Valuations of Determinants via Combinatorial Optimization: Applications to Differential Equations

Taihei Oki

Preface

Degrees of determinants of polynomial matrices often appear as algebraic formulations of weighted combinatorial optimization problems. For example, weighted Edmonds' problem (WEP), which is to compute the degree of the determinant of a polynomial matrix having symbols, reduces to the weighted bipartite matching problem and the weighted linear matroid intersection and parity problems depending on symbols' pattern. Conversely, the degree of the determinant of an arbitrary polynomial matrix serves as a lower bound on the maximum weight of a perfect matching in the associated edge-weighted bipartite graph. Based on this relation, the combinatorial relaxation algorithm of Murota (1995) computes the degree of the determinant of a polynomial matrix by iteratively solving the weighted bipartite matching problem.

The above property on degrees of determinants extends to valuations of determinants of matrices over valuation fields, or more generally, to valuations of the Dieudonné determinants of matrices over valuation skew fields. In combinatorial optimization, valuations of the Dieudonné determinants arise from a noncommutative version of WEP (nc-WEP). An algebraic abstraction of linear differential and difference equations gives rise to skew polynomials, which are a noncommutative generalization of polynomials. Valuations of Dieudonné determinants of skew polynomial matrices provide information on dimensions of solution spaces of linear differential and difference equations.

The combinatorial relaxation is of importance to preprocessing of differential-algebraic equations (DAEs). In numerical analysis of DAEs, consistent initialization and index reduction are necessary preprocessing prior to the numerical integration. Popular preprocessing methods of Pantelides (1988), Mattsson–Söderlind (1993), and Pryce (2001) are based on the assignment problem on a bipartite graph that represents variable occurrences in equations. The structural methods, however, fail for some DAEs due to inherent numerical or symbolic cancellations. The combinatorial relaxation provides a framework of modifying a DAE into another DAE to which the structural methods are applicable, whereas modification method used in the framework should be appropriately chosen according to the target DAEs.

In the first half of this thesis, we propose two algorithms for computing valuations

of the Dieudonné determinants of matrices over valuation skew fields. The algorithms are extensions of the combinatorial relaxation of Murota and the matrix expansion by Moriyama–Murota (2013), both of which are based on combinatorial optimization. We show that the skew polynomials arise as the most general algebraic structure to which these algorithms admit natural extensions. Applications are presented for the nc-WEP and analysis of linear differential and difference equations.

The last half of this thesis is devoted to DAEs' modification methods based on the combinatorial relaxation. This thesis presents three methods for modifying DAEs into other DAEs to which the preprocessing methods can be applied. One method is for linear DAEs whose coefficient matrices are mixed matrices, which are matrices having symbols representing physical quantities. We develop an efficient algorithm that relies on graph and matroid algorithms but not on symbolic computation. Other two deal with general nonlinear DAEs with the aid of symbolic computation engines to manipulate nonlinear formulas. In addition to theoretical guarantees, we conduct numerical experiments on real instances to present practical efficiency.

Acknowledgments

First of all, I would like to express my deepest gratitude to my supervisor, Professor Satoru Iwata. In these five years since he invited me to the research world, I have been learning from him a lot about research, explicitly or implicitly, such as tackling difficult problems, writing good papers, efficiently appealing research results, and winning research grants. Neither this dissertation nor myself in the present and future would be without his continuing and generous support.

I am also greatly indebted to my ex-supervisor in my undergraduate days, Dr. Hiroshi Hirai of the 2nd Laboratory. Even after my graduation, he has intermittently given me information on recent research topics, which significantly extended my research area. He also kindly took me to Vladimir Kolmogorov of IST Austria. I believe that this, my first experience to visit an overseas researcher, must be an epoch-making event in my research life.

I would like to thank my collaborators. Besides my supervisor, Chapter 7 benefited from discussions with Dr. Mizuyo Takamatsu of Chuo University. Her encouragement played an important role in my decision to go on to the Ph.D. course. I also thank other coauthors of a paper written during my Ph.D. course: Dr. Yutaro Yamaguchi of Kyushu University, Yuya Masumura of Fast Retailing Co., Ltd, and Kazuki Matoya of the 6th Laboratory. It was an invaluable experience to collaborate with them all.

Laboratory and other voluntary seminars have provided ample opportunities to obtain new knowledge and helpful comments. I am grateful to Dr. Shin-ichi Tanigawa, Dr. Tasuku Soma, Dr. Shinsaku Sakaue, and Nobutaka Shimizu of the 7th Laboratory, Dr. Ayumi Igarashi and Dr. Kaito Fujii of NII (National Institute of Informatics), Dr. Shinji Ito and Dr. Tatsuya Matsuoka of NEC Corporation and Dr. Yuni Iwamasa of Kyoto University for helpful comments in the seminars.

My proposals have been accepted as research projects of JST ACT-I “Information and Future” and its Acceleration Phase. As an area advisor, Professor Shin-ichi Minato of Kyoto University has kindly advised me about my research from a comprehensive perspective. The research supervisor Dr. Masataka Goto of AIST (National Institute of Advanced Industrial Science and Technology) and an area advisor Professor Ken-ichi Kawarabayashi

of NII have also given me helpful comments. It was invaluable for me to have stimulating discussions with other ACT-I members, including Dr. Yusuke Kobayashi of Kyoto University and Dr. Shuichi Hirahara of NII. During the Ph.D. course, my research was also financially supported by the JSPS Research Fellowship for Young Researchers (DC1), the JST CREST Iwata team, and the JST ERATO Maeda team. I offer my special thanks to Erika Hiruma for her secretarial support.

Last but not least, I would like to thank my family and friends for their warm support and love.

Contents

Chapter 1

Introduction

Edmonds [23] observed that the rank of a matrix does not exceed the maximum size of a matching in the associated bipartite graph. The weighted version of this relation holds in the following sense: the degree of the determinant of a polynomial matrix serves as a lower bound on the maximum weight of a perfect matching in the edge-weighted bipartite graph associated with the matrix. This relation leads us to an efficient deg-det computation algorithm, the combinatorial relaxation of Murota [67].

The first half of this thesis focuses on generalizing "degrees of determinants" in two directions: "degrees" to "valuations" and "determinants" to "the Dieudonné determinants", which are a noncommutative generalization of determinants by Dieudonné [19]. Valuations of the Dieudonné determinants arise from combinatorial optimization and analysis of linear differential and difference equations. We generalize two combinatorial algorithms for the deg-det computation, including the combinatorial relaxation, to the setting of valuations of the Dieudonné determinants.

The latter half of this thesis is attributed to modification methods for differential-algebraic equations (DAEs). The combinatorial relaxation has been used as a framework to modify DAEs into other DAEs that are more amenable to preprocessing methods prior to numerical integration. Based on the combinatorial relaxation, we develop modification methods for DAEs making use of combinatorial optimization algorithms, sometimes with symbolic computation support.

In what follows, we present the backgrounds and contributions of this thesis. In Section 1.1, we introduce basic notions of matrices and valuations over skew fields and how valuations of the Dieudonné determinants arise from applications. Next in Section 1.2, we describe the combinatorial relaxation algorithm for computing degrees of determinants. In Section 1.3, we introduce DAEs and structural preprocessing methods. In Section 1.4, we summarize our contributions presented in this thesis. The organization of this thesis is explained in Section 1.5.

1.1 Matrices and Valuations

1.1.1 Ranks and Determinants

We start with matrices over fields. Let A be an $n \times n'$ matrix over a field F. There are a large number of equivalent definitions of the *rank* of A. One definition is the dimension of the linear space spanned by row vectors of A. Another definition is the minimum nonnegative integer r such that A is decomposed as the product of two matrices of size $n \times r$ and $r \times n'$. The third definition is the maximum size of a nonsingular submatrix of A. Nonsingularity is equivalent to the nonzero-ness of the determinant.

This thesis deals with matrices over *skew fields*. A *skew field*, or a *division ring* is a ring F such that every nonzero element has a multiplicative inverse in F. The above three definitions of the rank are valid even for matrices over skew fields (by changing "dimension of the linear space" with "rank of the left F-module").

The determinant concept cannot be naturally extended to square matrices over noncommutative rings. Nevertheless, Dieudonné [19] defined the *Dieudonné determinant* for square matrices over skew fields as a noncommutative generalization of the usual determinant. While the Dieudonné determinant of $A \in F^{n \times n}$, denoted as $\operatorname{Det} A$, is no longer an element in F, it retains useful properties of the usual determinant such as $\operatorname{Det} AB = \operatorname{Det} A \operatorname{Det} B$.

Let A be a matrix over a skew field F with row set R and column set C. We associate A with the bipartite graph $G = (R \cup C, E)$ such that $\{i, j\}$ is in E if and only if the (i, j)th entry of A is nonzero for every $i \in R$ and $j \in C$. When F is a field and A is square, the determinant of A is the sum over all perfect matchings of G except that each matching has an associated sign. This means that G has a perfect matching only if A is nonsingular. More generally, the maximum size of a matching of G is an upper bound on the rank of A. This relation can also be shown for matrices over skew fields using the min-max theorem for bipartite matchings (see [27]).

1.1.2 Valuations of Determinants

Let F be a skew field. A (real, non-archimedean) *valuation* [101, Chapter IV] on F is a map $v : F \to \mathbb{R} \cup \{+\infty\}$ satisfying

(V1) $v(ab) = v(a) + v(b)$ for $a, b \in F$,

(V2) $v(a + b) \geq \min\{v(a), v(b)\}$ for $a, b \in F$,

(V3) $v(1) = 0$,

(V4) $v(0) = +\infty$.

The value $v(a)$ is called the *valuation* of $a \in F$. A valuation is called *discrete* if its image is a subset of $\mathbb{Z} \cup \{+\infty\}$. A (*discrete*) *valuation* (*skew*) *field* is a (skew) field equipped

with a (discrete) valuation. Discrete valuation skew fields and discrete valuation fields are abbreviated as *DVSFs* and *DVFs*, respectively.

The most basic example of a discrete valuation is the minus of the degree on the rational function field $K(s)$ over a skew field K. This is generalized to the degree on the *skew rational function field*, explained in Section 1.1.5. The *p-adic valuation* of the rational numbers is also a famous example of discrete valuations.

Let A be a square matrix over a DVF. The valuation of the determinant deserves to be the "weighted version" of ranks of matrices in the following sense. For every edge $\{i, j\}$ of the associate graph G with A, we set its weight as the valuation of the (i, j)th entry in A. By the definition of $\det A$ and the axioms of valuations, the minimum weight of a perfect matching of G serves as a lower bound on the valuation of the determinant of A.

For a square matrix over a DVSF, the valuation of the Dieudonné determinant of A is well-defined. The inequality between the valuation of $\mathrm{Det}\, A$ and the weighted bipartite matchings still hold.

In the rest of this section, we describe how valuations of the Dieudonné determinants arise in the study of combinatorial optimization and analysis of linear differential/difference equations.

1.1.3 Edmonds' Problem

In 1967, Edmonds [23] posed a question whether there exists a polynomial-time algorithm to compute the rank of a *linear* (*symbolic*) *matrix* B over a field K, which is in the form

$$B = B_0 + B_1 x_1 + \cdots + B_m x_m,$$

where $B_0, B_1 \ldots, B_m$ are matrices over K and $x_1, \ldots, x_m$ are commutative symbols. Here, B is regarded as a matrix over the polynomial ring $K[x_1, \ldots, x_m]$ or the rational function field $K(x_1, \ldots, x_m)$. If each B_i has only one nonzero entry, the rank computation for B corresponds to the bipartite matching problem. Similarly, Edmonds' problem coincides with the general matching problem if each B_i is skew-symmetric and of rank 2. More generally, Edmonds' problem is equivalent to the *linear matroid intersection problem* if all B_i are of rank 1, and to the *linear matroid parity problem* if all B_i are skew-symmetric matrices of rank 2; see Lovász [59]. Edmonds' problem is solvable in deterministic polynomial-time for these instances. It is known that these conditions on B_i can be eliminated for B_0 [31, 44, 72, 92]. As an other direction, Hirai–Iwamasa [41] gave a combinatorial algorithm when B is a 2×2 *partitioned matrix*. For general linear matrices, the celebrated Schwartz–Zippel lemma [88] provides a simple randomized algorithm if $|K|$ is large enough [59]. However, no deterministic polynomial-time algorithm still has been known; the existence of such an algorithm would imply nontrivial circuit complexity lower bounds [50, 95].

Hirai [39] introduced weighted versions of commutative and noncommutative Edmonds' problems. First, consider commutative symbols $x_1, \ldots, x_m$ and an extra commutative

symbol s. Define a matrix

$$A = A_0 + A_1 s + \cdots + A_\ell s^\ell, \tag{1.1}$$

where $A_d = A_{d,0} + A_{d,1}x_1 + \cdots + A_{d,m}x_m$ is a linear matrix over K for $d = 0, \ldots, \ell$. We call (1.1) a *linear polynomial matrix* over K. *Weighted Edmonds' problem* (WEP) is the problem of computing the degree (in s) of the determinant of A. Analogously to Edmonds' problem, the WEP includes a bunch of weighted combinatorial optimization problems as special cases, such as a maximum weighted perfect matching problem, a weighted linear matroid intersection problem, and a weighted linear matroid parity problem; see [39].

Mixed matrix and *mixed polynomial matrices* are important subclasses of linear matrices and linear polynomial matrices, respectively. A linear matrix B is called a *mixed matrix* if the constant term B_0 is arbitrary and each B_i with $i = 1, \ldots, m$ has only one nonzero entry; namely, each symbol x_i appears exactly once in entries of B. Mixed matrices were first introduced by Murota–Iri [72] as a faithful description of dynamical systems. In this use, entries in B_0 are "accurate constants" such as coefficients of conservation laws like Kirchhoff's law, and each symbol x_i is an "independent parameter" representing a physical quantity such as resistance values coming from Ohm's law. *Mixed polynomial matrices* are similarly defined. In systems analysis, mixed polynomial matrices appear as the transfer function matrices of linear systems. See [71] for details.

Efficient combinatorial algorithms have been given for dealing with mixed matrices and mixed polynomial matrices. Murota [65] showed that the rank computation of mixed matrices reduces to *independent matching problem* on linear matroids, which is equivalent to the (linear) matroid intersection problem. Murota [69] also gave a reduction of the deg-det computation for mixed polynomial matrices to the *valuated matroid intersection problem*, which is a generalization of the matroid intersection problem to *valuated matroids*. Subsequently, Iwata–Takamatsu [46] presented another deg-det computation algorithm for mixed polynomial matrices, which is based on the combinatorial relaxation.

1.1.4 Noncommutative Edmonds' Problem

Recent studies [28, 37, 43] address the noncommutative version of Edmonds' problem (nc-Edmonds' problem). This is a problem of computing the *noncommutative rank* (nc-rank) of B, which is the rank defined by regarding $x_1, \ldots, x_m$ as pairwise noncommutative, i.e., $x_i x_j \neq x_j x_i$ if $i \neq j$. In this way, B is viewed as a matrix over the free ring $K\langle x_1, \ldots, x_m \rangle$ generated by noncommutative symbols $x_1, \ldots, x_m$. The nc-rank of B is precisely the rank of B over a skew (noncommutative) field $K\langle\!\langle x_1, \ldots, x_m \rangle\!\rangle$, called a *free skew field*, which is the quotient of $K\langle x_1, \ldots, x_m \rangle$ defined by Amitsur [2]. We call a linear matrix over K having noncommutative symbols an *nc-linear matrix* over K. The recent studies [28, 37, 43] revealed that nc-Edmonds' problem is deterministically tractable. For the case where K is the set $\mathbb{Q}$ of rational numbers, Garg et al. [28] proved that Gurvits' *operator*

scaling algorithm [35] deterministically computes the nc-rank of B in $\mathrm{poly}(n, m)$ arithmetic operations on $\mathbb{Q}$. Algorithms over a general field K were later given by Ivanyos et al. [43] and Hamada–Hirai [37] exploiting the min-max theorem established for nc-rank.

We next define *noncommutative weighted Edmonds' problem* (nc-WEP). Let $x_1, \ldots, x_m$ be noncommutative symbols and s an extra symbol that commutes with any element in $K\langle x_1, \ldots, x_m\rangle$. An *nc-linear polynomial matrix* A over K is a matrix in the form of (1.1) with each A_d regarded as an nc-linear matrix. Then A can be viewed as a matrix over the rational function (skew) field $F(s)$ over $F \coloneqq K\langle\!\langle x_1, \ldots, x_m\rangle\!\rangle$. The nc-WEP is the problem of computing $\deg \mathrm{Det}$ of a given nc-linear polynomial matrix. Hirai [39] formulated the dual problem of the nc-WEP as the minimization of an *L-convex function* on a *uniform modular lattice*, and gave an algorithm based on the steepest gradient descent. Hirai's algorithm can also be regarded as a variant of combinatorial relaxation algorithms. While Hirai's algorithm uses only polynomially many arithmetic operations on K with respect to the matrix size, the number of symbols m, and the degree ℓ, no bit-length bound has been given for $K = \mathbb{Q}$. Very recently, Hirai–Ikeda [40] presented a strongly polynomial-time algorithm of the nc-WEP for a special class of "sparse" nc-linear polynomial matrices.

1.1.5 Linear Differential and Difference Equations

Consider a linear differential equation $A_0 y(t) + A_1 \dot{y}(t) + \cdots + A_\ell y^{(\ell)}(t) = 0$, where $A_0, \ldots, A_\ell$ are $n \times n$ matrices over $\mathbb{C}$. The number of initial values needed to uniquely determine a solution coincides with the dimension of the solution space (over $\mathbb{C}$). Chrystal's theorem [12] states that this dimension is equal to the degree of the determinant of $A \coloneqq \sum_{d=0}^{\ell} A_d s^d \in \mathbb{C}[s]^{n\times n}$. We see an algebraic extension of this relation for time-varying linear differential and difference equation systems.

Let K be a field and $\delta : K \to K$ a *derivation* on K; that is, it satisfies $\delta(a + b) = \delta(a) + \delta(b)$ and $\delta(ab) = \delta(a)b + a\delta(b)$ for any $a, b \in K$. A typical setting is $K = \mathbb{C}(t)$ and δ is the usual differentiation along t. An ℓth-order (ordinary, matrix) *linear differential equation* over F for an n-dimensional vector y is

$$A_0 y + A_1 \delta(y) + A_2 \delta^2(y) + \cdots + A_\ell \delta^\ell(y) = f,$$

where $A_0, \ldots, A_\ell \in K^{n\times n}$ and $f \in K^n$. This equation can be expressed as $A \bullet y = f$, where $A \coloneqq A_0 + A_1 s + \cdots + A_\ell s^\ell$ is a matrix over polynomials in the differential operator s that acts on K as $s \bullet a = \delta(a)$. Since $(sa) \bullet b = s \bullet (ab) = \delta(ab) = a\delta(b) + \delta(a)b = (as + \delta(a)) \bullet b$ for $a, b \in K$, the operator s satisfies $sa = as + \delta(a)$ for $a \in K$.

Similarly, let K be a field and $\sigma : K \to K$ a field automorphism on K, called a *difference operator*. A typical setting is $K = \mathbb{C}(t)$ and σ is the $\mathbb{C}$-automorphism that maps t to $t + 1$. An ℓth-order (matrix) *linear difference equation* over K for an n-dimensional

vector y is

$$A_0 y + A_1 \sigma(y) + A_2 \sigma^2(y) + \cdots + A_\ell \sigma^\ell(y) = f, \tag{1.2}$$

where $A_0, \ldots, A_\ell \in K^{n \times n}$ and $f \in K^n$. In the same way as the differential equation, (1.2) is expressed as $A \bullet y = f$, where $A \coloneqq A_0 + A_1 s + \cdots + A_\ell s^\ell$ is a matrix over polynomials in the difference operator s. By $(sa) \bullet b = s \bullet (ab) = \sigma(ab) = \sigma(a)\sigma(b) = (\sigma(a)s) \bullet b$ for $a, b \in K$, it holds $sa = \sigma(a)s$ for $a \in K$.

Polynomials in differential and difference operators can be treated in a unified way by *skew polynomials*, which were introduced by Ore [76]. Let K be a (skew) field, $\sigma : K \to K$ a ring automorphism, and $\delta : K \to K$ a *σ-derivation*; that is, it satisfies $\delta(a+b) = \delta(a)+\delta(b)$ and $\delta(ab) = \sigma(a)\delta(b) + \delta(a)b$ for all $a, b \in K$. The *skew polynomial* over (K, σ, δ) in indeterminate s is a polynomial over K with the usual addition and a twisted multiplication defined by the commutation rule

$$sa = \sigma(a)s + \delta(a) \tag{1.3}$$

for all $a \in K$. As we have seen above, s corresponds to the differential operator when σ is the identity map, and to the difference operator when $\delta = 0$.

Since both sides of (1.3) are of "degree one" with respect to s, the degree of a skew polynomial is well-defined. The degree can be extended to *skew rational functions*, the fractionals of skew polynomial rings. Then the skew field of skew rational functions forms a DVSF with valuation $-\deg$.

Taelman [93] showed that the dimension of the solution space of a homogeneous linear differential equation $Ay = 0$ is bounded by the degree of the Dieudonné determinant of A, where the equality is attained in the *Picard–Vessiot extension* of K. In this thesis, we extend Taelman's result to an inhomogeneous linear differential equation $Ay = f$ and to linear difference equations. This is how the computation of valuations of the Dieudonné determinants can be applied to linear differential/difference equations analysis.

1.2 Combinatorial Relaxation

The *combinatorial relaxation*, introduced by Murota [64, 67], is a framework of algorithms to compute the degrees of the determinants of polynomial matrices over fields. The combinatorial relaxation was first invented for computing the Newton polygon of the Puiseux series solutions of determinantal equations [64] and was later applied to $\deg \det$ computation [67].

We introduce some notions to describe the combinatorial relaxation. Let $A \in K[s]^{n \times n}$ be a matrix over a field K and G the bipartite graph associated with A. We set a weight of each edge in G as the degree of the corresponding entry in A. Put $d(A) \coloneqq \deg \det A$ and define $\hat{d}(A)$ as the maximum weight of a perfect matching in G. Since $-\deg$ is a

valuation, we have $d(A) \leq \hat{d}(A)$ as indicated in Section 1.1.2. We say that A is *upper-tight* if $d(A) = \hat{d}(A)$.

The upper-tightness of A is characterized by the *tight coefficient matrix*, which is a matrix $A^{\#}$ over K defined from A and a dual feasible solution of the linear programming relaxation of the weighted matching problem. By the complementary theorem, the bipartite graph associated with $A^{\#}$ has a perfect matching if and only if the dual solution corresponding to $A^{\#}$ is optimal. Moreover, A is upper-tight if and only if the tight coefficient matrix with respect to a dual optimal solution is nonsingular. These mean that numerical cancellations in $A^{\#}$ make A non-upper-tight.

The combinatorial relaxation consists of the following three phases:

Combinatorial Relaxation

Phase 1. Compute $\hat{d}(A)$ by solving the maximum-weight perfect bipartite matching problem. If $d(A) < 0$, output $-\infty$ and halt.

Phase 2. If A is upper-tight, output $\hat{d}(A)$ and halt.

Phase 3. Modify A into $\bar{A}$ such that $d(A) = d(\bar{A})$ and $\hat{d}(A) < \hat{d}(\bar{A})$. Go back to Phase 1.

Since $d(A)$ and $\hat{d}(A)$ are integral, the gap between $d(A)$ and $\hat{d}(A)$ decreases by at least 1 for each iteration (unless $d(A) = -\infty$). Thus, the combinatorial relaxation terminates in at most $\hat{d}(A)$ iterations. The upper-tightness testing in Phase 2 can be done without knowing $d(A)$ by checking the nonsingularity of the tight coefficient matrix.

Different modification methods in Phase 3 yield different combinatorial relaxation algorithms. The original algorithm by Murota [67] uses the *unimodular transformation* $A \mapsto UA$, where $U \in K[s]^{n \times n}$ is a *unimodular matrix*, i.e., a polynomial matrix whose determinant is in $K \backslash \{0\}$. Murota [66] presented another modification method by the *biproper transformation*, which is the multiplication by a *biproper matrix*: an invertible matrix over $K[s^{-1}]$. Combinatorial relaxation algorithms by biproper transformations have a merit in that they can be applied to computing the maximum degree of subdeterminants. For a *matrix pencil* $A = A_0 + A_1 s$ with $A_0, A_1 \in K^{n \times n}$, the algorithm of Iwata [47] modifies A by the *strict equivalence transformation* $A \mapsto UAV$, where U and V are nonsingular matrices over K. The combinatorial relaxation algorithm by Iwata–Takamatsu [46] for mixed polynomial matrices uses biproper transformations.

1.3 Differential-Algebraic Equations

Let $\mathbb{T} \subseteq \mathbb{R}$ be a nonempty open interval and $\Omega \subseteq \mathbb{R}^{(\ell+1)n}$ a nonempty open set. An ℓth-order *differential-algebraic equation* (DAE), which was introduced by Gear [29], is the

following equation

$$F(t, x(t), \dot{x}(t), \dots, x^{(\ell)}(t)) = 0 \tag{1.4}$$

for $x : T \to \mathbb{R}^n$, where $F : T \times \Omega \to \mathbb{R}^n$ is a sufficiently smooth function. DAEs have aspects of both *ordinary differential equations* (ODEs)

$$\dot{x}(t) = \varphi(t, x(t)) \tag{1.5}$$

and algebraic equations $G(t, x(t)) = 0$. DAEs are widely used for modeling dynamical systems such as mechanical systems, electrical circuits, and chemical reaction plants.

1.3.1 Consistent Initialization and Index Reduction

A fundamental and important problem in the study of DAEs is an *initial value problem*, which is to find a smooth trajectory $x : \mathbb{T} \to \mathbb{R}^n$ satisfying (1.4) with a specified initial value condition

$$x(t^*) = x^*_{(0)}, \quad \dot{x}(t^*) = x^*_{(1)}, \quad \dots, \quad x^{(\ell-1)}(t^*) = x^*_{(\ell-1)}, \tag{1.6}$$

where $t^* \in \mathbb{T}$ and $x^*_{(0)}, x^*_{(1)}, \dots, x^*_{(\ell-1)} \in \mathbb{R}^n$. Unlike ODEs, an initial value problem for a DAE may not have a solution because the DAE can involve algebraic constraints. The solution must satisfy not only the explicit constraints but also their differentiations, called *hidden constraints*. While giving a consistent initial value of a DAE is a crucial process prior to numerical integration, this is known to be a nontrivial task [4, 79, 89].

Another important preprocessing of the numerical simulation of DAEs is an *index reduction*, which is a process of reducing the *differentiation index* [9] of a DAE. The differentiation index of a first-order DAE

$$F(t, x(t), \dot{x}(t)) = 0 \tag{1.7}$$

is the minimum nonnegative integer ν such that the system of equations

$$F(t, x(t), \dot{x}(t)) = 0, \quad \frac{\mathrm{d}}{\mathrm{d}t} F(t, x(t), \dot{x}(t)) = 0, \quad \dots, \quad \frac{\mathrm{d}^\nu}{\mathrm{d}t^\nu} F(t, x(t), \dot{x}(t)) = 0$$

can determine $\dot{x}$ as a function of t and x. In other words, ν is the number of times one has to differentiate the DAE (1.7) to get an ODE. Intuitively, the differentiation index represents how far the DAE is from ODEs. The differentiation index of an ℓth-order DAE (1.4) is defined as that of the first-order DAE obtained by replacing higher-order derivatives of x with newly introduced variables. It is generally considered difficult to numerically solve high (≥ 2) index DAEs [4, 36, 89]. Therefore, it is important for accurate simulation of dynamical systems to convert a given DAE into a low (≤ 1) index DAE.

1.3.2 Structural Preprocessing Methods

Today, most simulation software libraries for dynamical systems, such as Dymola, OpenModelica, MapleSim, and Simulink, are equipped with graph-based preprocessing methods, which we call *structural preprocessing methods*. These methods were first presented by Pantelides [79] for consistent initialization of DAEs and was subsequently applied to an index reduction method by dummy derivative approach of Mattsson–Söderlind [60]. Pryce [82] proposed a consistent initialization method for DAEs, called the Σ-method, based on a variant of Pantelides' method. These structural preprocessing methods construct an edge-weighted bipartite graph from DAEs' structural information and solve the weighted bipartite matching problem.

These structural preprocessing methods, however, do not work for some DAEs. To explain this, consider a *linear DAE with constant coefficients*

$$\sum_{d=0}^{\ell} A_d x(t) = f(t), \tag{1.8}$$

where $A_0, \dots, A_\ell \in \mathbb{R}^{n\times n}$ and $f : \mathbb{R} \to \mathbb{R}^n$ is a smooth function. For the DAE (1.8), the structural methods construct the bipartite graph G, described in Section 1.2, associated with a polynomial matrix

$$A \coloneqq \sum_{d=0}^{\ell} A_d s^d \in \mathbb{R}[s]^{n\times n}. \tag{1.9}$$

Then the structural methods run assuming that the dimension N of the solution space of (1.8) is equal to the maximum weight of a perfect matching in G, whereas the correct statement is $N = \deg\det A$ as explained in Section 1.1.5. Hence the structural preprocessing methods might fail if A is not upper-tight. Since the upper-tightness of A is equivalent to the nonsingularity of the tight coefficient matrix $A^{\#}$ of A, the structural methods works for (1.8) if $A^{\#}$ is nonsingular.

Tight coefficient matrices are generalized for nonlinear DAEs as the *system Jacobian*, which is a kind of Jacobian matrices. The structural methods succeeds if the system Jacobian is singular and might fail if not.

1.3.3 DAE Modification via Combinatorial Relaxation

In order to overcome the above issue of structural preprocessing methods, we consider modifying a DAE into an equivalent DAE having nonsingular system Jacobian. Here, "equivalent" means that the modified DAE has the same solution set as the original DAE. For a linear DAE (1.8), the structural methods work if A in (1.9) is upper-tight. On the other hand, the combinatorial relaxation modifies A into an upper-tight matrix. Therefore, if a modification for A can be translated into an equivalent modification for the DAE (1.8),

one can make use of the combinatorial relaxation algorithm as a modification method for the DAE. Indeed, unimodular transformations and strict equivalence transformations are such modifications, whereas biproper transformations are not. Hence the combinatorial relaxation algorithm by Murota [67] and by Iwata [47] can be applied to the modification of linear DAEs with constant coefficients; see Wu et al. [103].

For nonlinear DAEs, Tan et al. [94] presented combinatorial-relaxation based modification methods, called the linear combination (LC) method and the expression substitution (ES) method. However, even the LC and ES-methods are not applicable to "highly nonlinear" DAEs, which are also included in the standard test set for DAE solvers [61].

1.4 Contributions

This thesis contains three results summarized below.

Computing Valuations of the Dieudonné Determinant. We develop two combinatorial algorithms to compute the valuations of the Dieudonné determinants of matrices over a certain type of DVSFs, called *split* [22]. The first algorithm is a generalization of the combinatorial relaxation. The second algorithm generalizes the *matrix expansion* by Van Dooren et al. [98] for $\deg \det$ of real rational function matrices and by Moriyama–Murota [62] for $\deg \det$ of polynomial matrices over fields. The matrix expansion algorithm essentially relies on the *Legendre conjugacy* between integer sequences of the valuations of minors and of ranks of matrices obtained by arranging coefficient matrices. The Legendre conjugacy is an important duality relation on discrete convex and concave functions treated in *discrete convex analysis* [70].

We carefully carry an argument so that algorithms can be applied as widely as possible. The splitness condition arises from natural requirements in dealing with DVSFs on computers. In commutative case, split DVSFs are nothing but subfields of the formal Laurent series fields. In general noncommutative case, split DVSFs are isomorphic to skew subfields of formal Laurent series (skew) fields having a commutation rule designated by a family of maps called *higher σ-derivations* [84]. Our algorithms additionally require an upper bound on the output. We show that matrices over a split DVSF have natural upper bounds if and only if the DVSF is obtained from skew polynomial rings. This gives a new characterization of skew polynomial rings as the most general ring structure that admits natural extensions of the combinatorial relaxation and matrix expansion algorithms.

Our algorithms can be applied to the computation of the dimension of the solution spaces of linear differential/difference equations and to the commutative and noncommutative WEP. In particular, we give the first deterministic polynomial-time algorithm for the nc-WEP over $\mathbb{Q}$ with bounded bit complexity.

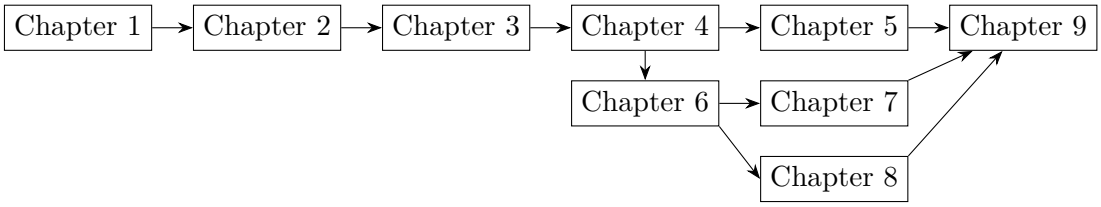


Figure 1.1: The dependence structure of chapters.

Structural Modification for Linear DAEs with Mixed Matrices. The second result is to develop a modification algorithm for linear DAEs whose coefficient matrices are mixed matrices; such DAEs naturally arise from dynamical systems. Based on the combinatorial relaxation framework, our algorithm transforms a DAE into an equivalent DAE whose tight coefficient matrix is nonsingular, i.e., the structural preprocessing methods are applicable. Technically, our contribution is to present a combinatorial relaxation algorithm for mixed polynomial matrices that uses unimodular transformations, whereas the algorithm of Iwata–Takamatsu [46] uses biproper transformations. Our algorithm does not rely on symbolic manipulations but fast combinatorial algorithms on graphs and matroids. We further provide an improved algorithm under an assumption based on dimensional analysis of dynamical systems. Through numerical experiments, it is confirmed that our algorithms run fast for large scale DAEs.

Structural Modification for Nonlinear DAEs. The third result is to present two combinatorial-relaxation based modification methods for nonlinear DAEs: the *substitution method* and the *augmentation method*. Both methods are aided by algebraic computation engines in manipulating mathematical formulations and are applicable to a large class of nonlinear DAEs. The substitution method symbolically solves equations for some derivatives based on the implicit function theorem and substitutes the solution back into the system. The augmentation method modifies DAEs by appending new variables and equations instead of solving equations. The augmentation method has advantages that the equation solving is not needed, and the sparsity of DAEs is retained.

Our methods are implemented as a MATLAB library using the MuPAD language, which is a core system of the Symbolic Math Toolbox in MATLAB. Through the application of it to practical DAEs, we show that our methods can be used as a promising preprocessing of DAEs that the index reduction procedure in MATLAB cannot handle.

1.5 Organization

Figure 1.1 illustrates the structure of this thesis. In Chapters 2 and 3, we introduce necessary preliminaries on algebra and combinatorial optimization, respectively. We present algorithms to computes valuations of the Dieudonné determinants in Chapter 4 and their

applications in Chapter 5.

Chapters 6–8 deal with topics of differential equations. In Chapter 6, we describe structural analysis for linear differential/difference equations and for DAEs. The first one is to provide an application of algorithms given in Chapter 4 and the latter one is to explain backgrounds and preliminaries of the subsequent chapters We then present modification methods for linear DAEs with mixed matrices in Chapter 7 and for nonlinear DAEs in Chapter 8.

Finally, we conclude this thesis in Chapter 9.

Chapter 2

Preliminaries on Valuated Skew Fields

We first mention basic notations and conventions used throughout this thesis. Let $\mathbb{N}$, $\mathbb{Z}$, $\mathbb{Q}$, $\mathbb{R}$, $\mathbb{C}$ denote the sets of nonnegative integers, integers, rationals, reals, and complex numbers, respectively. For $n \in \mathbb{N}$, define $[n] \coloneqq \{1, 2, \ldots, n\}$ and $[0, n] \coloneqq \{0, 1, 2, \ldots, n\}$. We sometimes make use of a special element $+\infty$ such that $a + \infty = +\infty + \infty = +\infty$ and $a < +\infty$ for all $a \in \mathbb{R}$.

All rings are assumed to have the multiplicative identity. We denote the multiplicative group of a ring R by $R^\times$. The *characteristic* $\mathrm{ch}(R)$ of a ring R is the minimum positive integer n such that $\underbrace{1 + \cdots + 1}_{n \text{ times}} = 0$. If such n does not exist, we define $\mathrm{ch}(R) \coloneqq 0$.

2.1 Valuations

2.1.1 Real Valuations

A *skew field*, or a *division ring* is a ring F such that every nonzero element has a multiplicative inverse in F. A (*real*) *valuation skew field* [101, Chapter IV] is a skew field F endowed with a (*real*) *valuation*, that is, a map $v : F \to \mathbb{R} \cup \{+\infty\}$ satisfying (V1)–(V4). A valuation skew field is called a *valuation field* if it is a field. The value $v(a)$ for $a \in F$ is called the *valuation* of a.

By (V1) and (V3), it holds $v(-a) = v(a)$ and $v(a^{-1}) = -v(a)$ for all $a \in F^\times$, where $F^\times = F \setminus \{0\}$ is the multiplicative group of F. In particular, we have $v(a) < +\infty$ for $a \in F^\times$. The equality in (V2) is attained whenever $v(a) \neq v(b)$; otherwise, if $v(a) < v(a+b)$ and $v(a) < v(b)$, it holds

$$v(a) = v((a + b) - b) \geq \min\{v(a + b), v(-b)\} = \min\{v(a + b), v(b)\} > v(a),$$

a contradiction.

The (*invariant*) *valuation ring* of a valuation skew field F with respect to a valuation v is a set

$$R \coloneqq \{a \in F \mid v(a) \geq 0\}.$$

Then R is a subring of F by (V1) and (V2), and is a *domain*, i.e., R has no zero-divisors. It also satisfies the following [54, Chapter 1]:

(VR1) either $a \in R$ or $a^{-1} \in R$ for $a \in F^\times$,

(VR2) $aR = Ra$ for $a \in F^\times$.

In addition, R is a *local ring*, i.e., it has a unique maximal right (and indeed a unique maximal left) ideal $J(R)$, which coincides with $R \setminus R^\times$ with $R^\times = \{a \in F \mid v(a) = 0\}$. Namely, it holds

$$J(R) = \{a \in F \mid v(a) > 0\}. \tag{2.1}$$

The quotient ring $R \,/\, J(R)$ forms a skew field, called the *residue skew field* of F (or a *residue field* if it is a field).

A *representative set* of F is a subset Q of R such that $0 \in Q$ and the restriction to Q of the canonical homomorphism from R to the residue skew field $K \coloneqq R \,/\, J(R)$ is a bijection from Q to K. Then for $a \in R$, there uniquely exists $a_0 \in Q$ such that $a \in a_0 + J(R)$. Hence $a - a_0 \in J(R)$, which means:

Proposition 2.1. *Let F be a valuation skew field with valuation v, valuation ring R, and representative set Q. Then any $a \in R$ is uniquely expressed as $a = a_0 + \tilde{a}$, where $a_0 \in Q$ and $\tilde{a} \in J(R)$.*

2.1.2 Discrete Valuations

Let F be a valuation skew field with valuation v. The *value group* of v is the additive subgroup $v(F^\times)$ of $\mathbb{R}$. A *discrete valuation* is a valuation F whose value group is $\mathbb{Z}$. A valuation skew field equipped with a discrete valuation is called a *discrete valuation skew field* (DVSF), which is of the main interest of this thesis. If F is a field, we call F a *discrete valuation field* (DVF).

Let F be a DVSF with discrete valuation v and the valuation ring R. Then (2.1) is

$$J(R) = \{a \in F \mid v(a) \geq 1\}. \tag{2.2}$$

Any element $\pi \in R$ with $v(\pi) = 1$ is called a *uniformizer* or a *prime element* of F. In addition to (VR1) and (VR2), R enjoys the following properties [54, Chapter 1]:

(DVR1) $J(R) = \pi R = R\pi$,

(DVR2) $\bigcap_{d=1}^{\infty} J(R)^d = \{0\}$.

Note that it holds

$$J(R)^d = \pi^d R = R\pi^d = \{a \in F \mid v(a) \geq d\} \tag{2.3}$$

by (2.2) and (DVR1) for $d \in \mathbb{N}$. In addition, any right ideal and left ideal of R are two-sided and are in the form of (2.3). This mean that R is a (right and left) *principal ideal domain* (PID), which is a domain whose every (right and left) ideal is generated by one element. More strongly, any DVR is a (right and left) *Euclidean domain* [7] as is well-known for commutative DVRs. Here, a domain R is said to be *Euclidean* if there exists a map $f : R \to \mathbb{N} \cup \{-\infty\}$, called an *Euclidean map*, such that for every $a, b \in R$ with $b \neq 0$, there exist $q, r, q', r' \in R$ such that $a = bq + r = q'b + r'$ and $f(r), f(r') < f(b)$. In case of a valuation ring of a DVSF, $-v$ serves as an Euclidean map. We remark that Euclidean domains are proper subclass of PIDs even for noncommutative rings [7].

Remark 2.2. In general, a local ring R satisfying (DVR1) and (DVR2) for some non-nilpotent element $\pi \in R$ is called a *discrete* (*invariant*) *valuation ring* (DVR). Here, an element $a \in R$ is said to be *nilpotent* if $a^k = 0$ for some $k \in \mathbb{N}$ and *non-nilpotent* if not. The valuation ring of any DVSF is a DVR as described above. Indeed, any DVR R is the valuation ring of some DVSF [54]; here we give a construction of the DVSF briefly. First, it follows from (DVR1) and (DVR2) that R is a PID. Then R is also a (right and left) *Ore domain*, which is a domain such that for each $s, t \in R \setminus \{0\}$, there exist $x, y, z, w \in R \setminus \{0\}$ satisfying $sx = ty$ and $zs = wt$ [33, Corollarly 6.7]. This property enables for R to have the *Ore quotient skew field* F, which is a skew field of fractions each of whose elements $a \in F$ is expressed as $a = sx^{-1} = y^{-1}t$ for some $s, t \in R$ and $x, y \in R \setminus \{0\}$. In particular, $a \in F^\times$ can be uniquely expressed as $a = \pi^k p = q\pi^k$ for some $p, q \in R^\times$ and $k \in \mathbb{Z}$. Denote this k by $v(a)$ for $a \in F^\times$ and let $v(0) \coloneqq +\infty$. Then $v : F \to \mathbb{Z} \cup \{+\infty\}$ is a discrete valuation on F, whose valuation ring coincides with R. We refer to the restriction of v onto R as the valuation of R and a representative set of R means that of F. See [54, Chapter 1] for details of DVRs and [33, Chapter 6] for Ore domains and quotient skew fields.

Let F be a DVSF with valuation v and uniformizer π. For an arbitrary real number $c > 1$, we define $d : F \times F \to \mathbb{R}$ as

$$d(a, b) \coloneqq c^{-v(a-b)}$$

for $a, b \in F$ (where $c^{-\infty} \coloneqq 0$). Then d forms a metric on F. The *π-adic topology* is the ring topology on F induced by d, which does not depend on the choice of c. On this topology, $\{a + J(R)^k \mid k \in \mathbb{N}\}$ is an open neighborhood system of $a \in F$ by (2.3). A DVSF is said to be *complete* if it is complete as a metric space. Then any DVSF can be

extended to a complete DVSF as follows.

Theorem 2.3 ([101, Theorem 17.2])**.** *Let F be a DVSF with discrete valuation v. Then there uniquely exists a complete DVSF $\hat{F}$ with discrete valuation $\hat{v}$ such that $\hat{F}$ contains F as a dense subring and $\hat{v}$ extends v. In addition, the residue skew field of $\hat{F}$ is isomorphic to that of F.*

The complete DVSF $\hat{F}$ in Theorem 2.3 is called the *completion* of F. By Theorem 2.3, it is convenient to consider complete DVSFs from the beginning. See [101] for details of topological rings and the π-adic topology.

Let F be a DVSF with uniformizer π, valuation ring R, and representative set Q. By Proposition 2.1 and (DVR1), we can express $a \in R$ as $a = a_0 + a'\pi$ by some $a_0 \in Q$ and $a' \in R$. By the same argument, there are unique $a_1 \in Q$ and $a'' \in R$ such that $a' = a_1 + a''\pi$. Therefore, we have $a = a_0 + a_1\pi + a''\pi^2$. Repeating this argument, we can represent a as a power series in π with coefficient Q, which is formally stated as follows.

Proposition 2.4 ([101, Theorem 18.5])**.** *Let F be a DVSF with discrete valuation v and let π and Q be a uniformizer and a representative set of F, respectively.*

(1) *For every $a \in F$, there uniquely exists a sequence $(a_d)_{d\in\mathbb{Z}}$ of elements in Q such that $a_d = 0$ for all but finitely many $d < 0$ and a power series*

$$\sum_{d\in\mathbb{Z}} a_d \pi^d \tag{2.4}$$

converges to a in the π-adic topology. If $\ell := v(a) \in \mathbb{Z}$, then $a_d = 0$ for $d < \ell$ and $a_\ell \neq 0$.

(2) *If F is complete and $(a_d)_{d\in\mathbb{Z}}$ is a sequence of elements in Q such that $a_d = 0$ for all but finitely many $d < 0$, the power series (2.4) converges to an element a of F. Its valuation $v(a)$ is equal to the minimum $\ell \in \mathbb{Z}$ such that $a_d = 0$ for $d < \ell$ and $a_\ell \neq 0$.*

We call (2.4) the *π-adic expansion* of $a \in F$.

2.1.3 Examples

We present several examples of valuation skew fields. All examples are DVSFs except for Example 2.6.

Example 2.5 (formal Laurent series)**.** Let K be a skew field. Denote by $K[s]$ the polynomial ring over K in indeterminate s that commutes with any element of K. Since $K[s]$ is an Ore domain, it has the quotient skew field $K(s)$, called the *rational function* (skew) *field*. The *order* $\operatorname{ord} p$ of $p \in K[s] \setminus \{0\}$ is the minimum $d \in \mathbb{N}$ such that the coefficient of s^d in p is nonzero. We also define $\operatorname{ord} f$ for $f \in K(s) \setminus \{0\}$ as $\operatorname{ord} f := \operatorname{ord} p - \operatorname{ord} q$, where $f = p/q$ with $p, q \in K[s] \setminus \{0\}$. Set $\operatorname{ord} 0 := +\infty$. Then it is well-known that the order is a

discrete valuation on $K(s)$ and the residue skew field is K. A canonical (but not unique) choice of a uniformizer is s. The completion of $K(s)$ is the *formal Laurent series* (skew) *field* $K((s))$ over K in s, whose each element is expressed as

$$f = \sum_{d=\ell}^{\infty} a_d s^d \tag{2.5}$$

with $\ell \in \mathbb{Z}$ and $a_\ell, a_{\ell+1}, \ldots \in K$. If $a_\ell \neq 0$, then $\ell = \operatorname{ord} f$. The valuation ring of $K((s))$ is called the *formal power series* (skew) *field* $K[[s]]$ over K in s, which is the subring of $K((s))$ consisting of formal power series

$$f = \sum_{d=0}^{\infty} a_d s^d \tag{2.6}$$

with $a_0, a_1, \ldots \in K$.

Similarly, the *degree* $\deg p$ of $p \in K[s] \setminus \{0\}$ is defined by replacing "minimum" with "maximum" in the definition of $\operatorname{ord} p$. Define $\deg f$ for $f = p/q \in K(s)^\times$ with $p, q \in K[s]\setminus\{0\}$ as $\deg f \coloneqq \deg p - \deg q$ and $\deg 0 \coloneqq -\infty$ as well. Since $\deg f(s) = -\operatorname{ord} f(s^{-1})$, the minus of the degree is a discrete valuation on $K(s)$ with uniformizer s^{-1} and residue skew field K. The completion of $K(s)$ with respect to the minus degree is $K((s^{-1}))$, which is a field isomorphic to $K((s))$. □

Example 2.6 (formal Laurent series with real exponents)**.** Let K be a skew field. A subset X of $\mathbb{R}$ is said to be *well-ordered* if any nonempty subset of X has the minimum element. We consider *formal Laurent series with real exponents*, each of which is in the following form

$$f = \sum_{x \in X} a_x s^x, \tag{2.7}$$

where $X \subsetneq \mathbb{R}$ is well-ordered, $a_x \in K^\times$ for $x \in X$, and s is a formal "indeterminate" that satisfies $s^{x+y} = s^x s^y$ and $as^x = s^x a$ for $x, y \in \mathbb{R}$ and $a \in K$. Addition on these series is naturally defined, and the multiplication of $f = \sum_{x\in X} a_x s^x$ and $g = \sum_{y\in Y} b_y s^y$ is given by

$$fg \coloneqq \sum_{z\in\mathbb{R}} \left(\sum_{\substack{x\in X, y\in Y \\ x+y=z}} a_x b_y \right) s^z.$$

For every $z \in \mathbb{R}$, the number of $(x, y) \in X \times Y$ satisfying $x + y = z$ is finite from the assumption that X and Y are well-ordered, and the set

$$\{z \in \mathbb{R} \mid \text{the coefficient of } s^z \text{ in } fg \text{ is nonzero}\}$$

is well-ordered as well. Hence fg is a formal Laurent series again in the sense defined above. By these operations, the set Σ of formal Laurent series with real exponents forms a skew field [74, Theorem 5.7].

Define the *order* $\operatorname{ord} f$ of (2.7) as the minimum $x \in X$. We also define $\operatorname{ord} 0 \coloneqq +\infty$. Then as Neumann [74] indicated, ord is a valuation on Σ that is not discrete. The residue skew field of Σ is K. The skew field Σ contains $K((s))$ as a subfield, and the restrictions of the order onto $K((s))$ coincides that on $K((s))$. Reversing the ordering of $\mathbb{R}$, we can also define $\deg f$ consistent with $K((s^{-1}))$ in the completely analogous way. □

Example 2.7 (p-adic numbers)**.** Let p be a prime number. The *p-adic valuation* $v_p(n)$ of $n \in \mathbb{Z} \setminus \{0\}$ is the maximum $k \in \mathbb{N}$ such that p^k divides n, and is extended to $\mathbb{Q}^\times$ as $v_p(x) \coloneqq v_p(n) - v_p(m)$ for $x = n/m \in \mathbb{Q}^\times$ with $n, m \in \mathbb{Z} \setminus \{0\}$. Also we define $v_p(0) \coloneqq +\infty$. Then v_p is a discrete valuation on $\mathbb{Q}$ with uniformizer p. The residue field is $\mathbb{F}_p$. The completion of $\mathbb{Q}$ with respect to v_p is the field $\mathbb{Q}_p$ of *p-adic numbers*. □

Example 2.8 (skew (inverse) Laurent series)**.** Let K be a skew field, $\sigma : K \to K$ a ring automorphism, and $\delta : K \to K$ a *left σ-derivation*; that is, it is additive, i.e., $\delta(a+b) = \delta(a) + \delta(b)$, and it satisfies $\delta(ab) = \sigma(a)\delta(b) + \delta(a)b$ for all $a, b \in K$. The (*left*) *skew polynomial ring*, or the *Ore polynomial ring* due to Ore [76] over (K, σ, δ) in indeterminate s, which is denoted by $K[s; \sigma, \delta]$, is a polynomial ring over K with the usual addition and a twisted multiplication defined by the commutation rule (1.3). Elements in $K[s; \sigma, \delta]$ are called *skew polynomials*. If $\delta = 0$, then $K[s; \sigma, 0]$ is denoted by $K[s; \sigma]$. When σ is the identity map id and $\delta = 0$, the skew polynomial ring is nothing but the polynomial ring $K[s]$, which means $K[s] = K[s; \mathrm{id}]$. A typical nontrivial example of skew polynomial rings is the ring $\mathbb{C}(t)[\partial; \mathrm{id}, ']$ of differential operators, where $' : \mathbb{C}(t) \to \mathbb{C}(t)$ is the usual differentiation. Another example of skew polynomial rings the ring $\mathbb{C}(t)[S; \tau]$ of shift operators, where $\tau : \mathbb{C}(t) \to \mathbb{C}(t)$ is defined by $f(t) \mapsto f(t+1)$ for $f \in \mathbb{C}(t)$.

Applying the commutation rule (1.3) iteratively, we can uniquely represent any skew polynomial $p \in K[s; \sigma, \delta] \setminus \{0\}$ as $p = a_0 + a_1 s + \cdots + a_\ell s^\ell$, where $\ell \in \mathbb{N}$ and $a_0, \ldots, a_\ell \in K$ with $a_\ell \neq 0$. This ℓ is called the *degree* of p and is denoted by $\deg p$. We set $\deg 0 \coloneqq -\infty$. Since a skew polynomial ring $K[s; \sigma, \delta]$ is an Ore domain (see, e.g., [33, Exercise 6F]), it has the quotient skew field $K(s; \sigma, \delta)$, called the *skew rational function field*. Its element $f \in K(s; \sigma, \delta)$, called a *skew rational function*, has the degree defined by $\deg f \coloneqq \deg p - \deg q$ with $f = pq^{-1}$ and $p, q \in K[s; \sigma, \delta]$. Then $-\deg$ is a discrete valuation on $K(s; \sigma, \delta)$ with residue skew field K. Its completion is the *skew inverse Laurent series field* $K((s^{-1}; \sigma, \delta))$, which is the skew field of formal power series over K in the form of

$$f = \sum_{d=\ell}^{\infty} a_d s^{-d}$$

for some $\ell \in \mathbb{Z}$ and $a_\ell, a_{\ell+1}, \ldots \in K$ [16, Section 2.3]. This skew field has the natural

addition and a multiplication defined by (1.3) and

$$s^{-1}a = \sum_{d=0}^{\infty} \delta_d(a) s^{-(d+1)}$$

for $a \in K$, where

$$\delta_d \coloneqq \sigma^{-1}\left(-\delta\sigma^{-1}\right)^d \tag{2.8}$$

for $d \in \mathbb{N}$ (the multiplication of maps means the composition) [80]. This is determined so that $ss^{-1}a = a$.

One can define the *order* $\operatorname{ord} p$ of a skew polynomial $p \in K[s;\sigma,\delta]$ similarly to the usual polynomials, i.e., $\operatorname{ord} p$ is the minimum $\ell \in \mathbb{N}$ such that p is represented as $p = a_\ell s^\ell + \cdots + a_L s^L$ for some $L \in \mathbb{N}$ and $a_\ell, \dots, a_L \in K$ with $a_\ell \neq 0$. Set $\operatorname{ord} 0 \coloneqq +\infty$ in the same way. However, if $a \in K^\times$ satisfies $\delta(a) \neq 0$, then $\operatorname{ord} s = 1$, $\operatorname{ord} a = 0$ and $\operatorname{ord} sa = \operatorname{ord}(\sigma(a)s + \delta(a)) = 0$, which violate (V1). Thus ord cannot be extended to a discrete valuation on $K(s;\sigma,\delta)$. Nevertheless, in case of $\delta = 0$, the order satisfies (V1)–(V3) and thus $K(s;\sigma) \coloneqq K(s;\sigma,0)$ becomes a DVSF equipped with a discrete valuation $\operatorname{ord} f \coloneqq \operatorname{ord} p - \operatorname{ord} q$ for $f = pq^{-1} \in K(s;\sigma)$ with $p, q \in K[s;\sigma]$. This is because the change of variable $\varphi : f(s) \mapsto f(s^{-1})$ provides an isomorphism between $K(s;\sigma)$ and $K(s;\sigma^{-1})$ and $\operatorname{ord} f = -\deg \varphi(f)$ for $f \in K(s;\sigma)$. The completion of $K(s;\sigma)$ with respect to ord is the *skew Laurent series field* $K((s;\sigma))$, whose elements are represented as formal Laurent series (2.5) [16, Section 2.3]. The residue skew field of $K((s;\sigma))$ is clearly K.

See [16, Chapter 2], [17, Section 7.3], and [33, Chapter 2] for details of skew polynomials, [80] for skew inverse Laurent series fields and Section 5.2 for the connection to differential and difference equations. □

2.2 Matrices

2.2.1 Basic Notions and Notations

For a ring R and $n, n' \in \mathbb{N}$, we denote the ring of $n \times n'$ matrices over R by $R^{n\times n'}$. We also denote by $Q^{n\times n'}$ the set of all $n \times n'$ matrices over a subset Q of R. A square matrix $A \in R^{n\times n}$ is said to be *invertible* if there (uniquely) exists an $n \times n$ matrix over R, denoted by A^{-1}, such that $AA^{-1} = A^{-1}A = I_n$, where I_n is the identity matrix of order n. When R can be extended to a skew field F, we call A *nonsingular* if A is invertible over F and *singular* if not; the nonsingularity does not depend on the choice of F. We denote by $\mathrm{GL}_n(R)$ the group of $n \times n$ invertible matrices over R, i.e., $\mathrm{GL}_n(R) \coloneqq (R^{n\times n})^\times$.

For $a \in R^\times$ and $\alpha = (\alpha_i)_{i\in[n]} \in \mathbb{Z}^n$, we define $D(a^\alpha) \coloneqq \operatorname{diag}\left(a^{\alpha_i}\right)_{i\in[n]}$, where diag denotes the diagonal matrix. For an additive map $\varphi : R \to R$ and $A \in R^{n\times n'}$, let $\varphi(A)$ denote the $n \times n'$ matrix over R obtained by applying φ to each entry in A.

For a matrix $A \in R^{n \times n'}$, let $\mathrm{Row}(A)$ and $\mathrm{Col}(A)$ denote the row and column sets of A, respectively. For $I \subseteq \mathrm{Row}(A)$ and $J \subseteq \mathrm{Col}(A)$, we denote by $A[I, J]$ the submatrix of A consisting of rows I and columns J. When $I = \mathrm{Row}(A)$, we simply write $A[J] \coloneqq A[\mathrm{Row}(A), J]$.

Let R be a commutative ring. The *determinant* of a square matrix $A = (A_{i,j}) \in R^{n \times n}$ is defined as

$$\det A \coloneqq \sum_{\sigma \in \mathfrak{S}_n} \operatorname{sgn} \sigma \prod_{i=1}^{n} A_{i,\sigma(i)}, \tag{2.9}$$

where $\mathfrak{S}_n$ is the group of all permutations on $[n]$ and $\operatorname{sgn} \sigma$ denotes the sign of a permutation $\sigma \in \mathfrak{S}_n$. It is well-known that $\det AB = \det A \det B$ and $\det A \neq 0$ if and only if A is invertible for $A, B \in R^{n \times n}$.

2.2.2 Matrices over Skew Fields

We consider matrices over a skew field F. A right (left) F-module is especially called a *right* (*left*) *F-vector space*. The *dimension* of a right (left) F-vector space V is defined as the rank of V as a module, that is, the cardinality of any basis of V. The usual facts from linear algebra on independent sets and generating sets in vector spaces are valid even on skew fields [57].

The *rank* $\operatorname{rank} A$ of a matrix $A \in F^{n \times n'}$ is the dimension of the right F-vector space spanned by the column vectors of A, and is equal to the dimension of the left F-vector space spanned by the row vectors of A. The rank is invariant under (right and left) multiplication of nonsingular matrices. It is observed that a square matrix $A \in F^{n \times n}$ is nonsingular if and only if $\operatorname{rank} A = n$. The rank of $A \in F^{n \times n'}$ is equal to the minimum $r \in \mathbb{N}$ such that there exists a decomposition $A = BC$ by some $B \in F^{n \times r}$ and $C \in F^{r \times n'}$ [15]. Here we give another characterization of the rank, which is well-known on the commutative case.

Proposition 2.9. *The rank of a matrix $A \in F^{n \times n'}$ over a skew field F is equal to the maximum $r \in \mathbb{N}$ such that A has a nonsingular $r \times r$ submatrix. In addition, A has a nonsingular $k \times k$ submatrix for all $k \in [0, r]$.*

Proof. We first show the latter part. For $k \in [0, \operatorname{rank} A]$, we can take a column subset $J \subseteq \mathrm{Col}(A)$ of cardinality k such that the column vectors of $A[J]$ are linearly independent. Since $\operatorname{rank} A[J] = k$, there must be $I \subseteq \mathrm{Row}(A)$ of cardinality k such that the row vectors of $A[I, J]$ is linearly independent. Then $A[I, J]$ is a $k \times k$ nonsingular submatrix of A due to $\operatorname{rank} A[I, J] = k$.

The former part is shown as follows. Let $r \in \mathbb{N}$ be the maximum size of a nonsingular submatrix of A. It holds $\operatorname{rank} A \leq r$ by the latter part of the claim. To show $\operatorname{rank} A \geq r$, take an $r \times r$ nonsingular submatrix $A[I, J]$ of A. Since $\operatorname{rank} A[I, J] = r$, the set of column vectors of A indexed by J is linearly independent. Thus we have $\operatorname{rank} A \geq r$. □

We next define the *Dieudonné determinant* for nonsingular matrices over a skew field. To describe this, we introduce the *Bruhat decomposition* as follows. A lower (upper) unitriangular matrix is a lower (resp. upper) triangular matrix whose diagonal entries are 1.

Proposition 2.10 (Bruhat decomposition [17, Theorem 9.2.2])**.** *A square matrix $A \in F^{n\times n}$ over a skew field F can be decomposed as $A = LDPU$, where L is lower unitriangular, D is diagonal, P is a permutation matrix, and U is upper unitriangular. If A is nonsingular, this decomposition is unique.*

Let $F^{\times}_{\mathrm{ab}} \coloneqq F^{\times} \,/\, [F^{\times}, F^{\times}]$ denote the abelianization of $F^{\times}$, where

$$[F^{\times}, F^{\times}] \coloneqq \langle \{ aba^{-1}b^{-1} \mid a, b \in F^{\times} \} \rangle$$

is the commutator subgroup of $F^{\times}$. The *Dieudonné determinant* $\operatorname{Det} A$ of $A \in \mathrm{GL}_n(F)$, which is decomposed as $A = LDPU$ by Proposition 2.10, is an element of $F^{\times}_{\mathrm{ab}}$ defined by

$$\operatorname{Det} A \coloneqq \operatorname{sgn}(P) \prod_{i=1}^{n} d_i \bmod [F^{\times}, F^{\times}],$$

where $\operatorname{sgn}(P) \in \{+1, -1\}$ is the sign of the permutation P and $d_i \in F^{\times}$ is the ith diagonal entry of D for $i \in [n]$ [19]. In case where F is commutative, the Dieudonné determinant coincides with the usual determinant.

An *elementary matrix* over F is a unitriangular matrix $E_n(i, j; a) \in \mathrm{GL}_n(F)$ whose (i, j)th entry $(i \neq j)$ is $a \in F$ and other off-diagonal entries are 0. An *elementary operation* on $A \in F^{n\times n'}$ is the (left or right) multiplication of A by an elementary matrix, which corresponds to adding a left (right) multiple of a row (resp. column) to another row (resp. column) of A. Denote by $\mathrm{E}_n(F)$ the subgroup of $\mathrm{GL}_n(F)$ generated by elementary matrices. If F is a field, $\mathrm{E}_n(F)$ is nothing but the special linear group $\mathrm{SL}_n(F) \coloneqq \{A \in \mathrm{GL}_n(F) \mid \det A = 1\}$ [17, Theorem 3.5.1]. This can be extended to the Dieudonné determinant as follows:

Theorem 2.11 ([17, Theorem 9.2.6])**.** *For a skew field F and $n \in \mathbb{N}$, the Dieudonné determinant gives rise to an exact sequence of groups*

$$1 \longrightarrow \mathrm{E}_n(F) \longrightarrow \mathrm{GL}_n(F) \xrightarrow{\operatorname{Det}} F^{\times}_{\mathrm{ab}} \longrightarrow 1.$$

Namely, $\operatorname{Det} : \mathrm{GL}_n(F) \to F^{\times}_{\mathrm{ab}}$ is a surjective map satisfying

(D1) $\operatorname{Det} AB = \operatorname{Det} A \operatorname{Det} B$ for $A, B \in \mathrm{GL}_n(F)$,

(D2) $\operatorname{Det} A = 1$ for $A \in \mathrm{E}_n(F)$,

where the inverse of (D2) also holds, i.e., $\mathrm{E}_n(F) = \{A \in \mathrm{GL}_n(F) \mid \operatorname{Det} A = 1\}$. It further follows immediately from the definition of Det that

(D3) $\operatorname{Det}\operatorname{diag}(d_1,\dots,d_n) = \prod_{i=1}^{n} d_i \bmod [F^\times, F^\times]$ for $d_1,\dots,d_n \in F^\times$.

Indeed, Det is the unique map satisfying (D1)–(D3) since unitriangular matrices are in $\mathrm{E}_n(F)$ and any permutation matrix P can be brought into $\operatorname{diag}(\operatorname{sgn}(P),1,\dots,1)$ by elementary operations.

2.2.3 Matrix Valuations

Let F be a valuation skew field with valuation v. For any $A \in \mathrm{GL}_n(F)$, we denote by $\zeta(A)$ the valuation of any representative of $\operatorname{Det} A$; this is well-defined because all commutators of $F^\times$ have valuation 0. We also define $\zeta(A) \coloneqq +\infty$ for singular $A \in F^{n\times n}$. By (V1), (V3) and (D1)–(D3), it holds

(VD1) $\zeta(AB) = \zeta(A) + \zeta(B)$ for $A, B \in F^{n\times n}$,

(VD2) $\zeta(A) = 0$ for $A \in \mathrm{E}_n(F)$,

(VD3) $\zeta(\operatorname{diag}(d_1,\dots,d_n)) = \sum_{i=1}^{n} v(d_i)$ for $d_1,\dots,d_n \in F$.

By the Bruhat decomposition, $\zeta : F^{n\times n} \to \mathbb{R}\cup\{+\infty\}$ is the unique map satisfying (VD1)–(VD3), as Taelman [93] observed for $\deg\operatorname{Det}$ of skew polynomials.

Let $\mathrm{M}(F)$ denote the set of all square matrices of finite order over F. If we see ζ as a function on $\mathrm{M}(F)$, it satisfies the (real) *matrix valuation* axioms. To describe this, we shall define the *determinantal sum* for two matrices $A, B \in F^{n\times n'}$ such that their columns are identical except for the first columns. The *determinantal sum* of A and B with respect to the first column is an $n \times n'$ matrix over F whose first column is the sum of those of A and B, and other columns are the same as A. The determinantal sums with respect to other columns and rows are also defined. We denote the determinantal sum of A and B (with respect to an appropriate column or row) by $A \nabla B$.

A (real) *matrix valuation* [38] on a skew field F is a map $V : \mathrm{M}(F) \to \mathbb{R}\cup\{+\infty\}$ that satisfies

(MV1) $V\begin{pmatrix} A & O \\ O & B \end{pmatrix} = V(A)+V(B)$ for $A, B \in \mathrm{M}(F)$, where O denotes the zero matrix of appropriate size,

(MV2) $V(A \nabla B) \geq \min\{V(A), V(B)\}$ for $A, B \in \mathrm{M}(F)$ such that $A \nabla B$ is defined,

(MV3) $V(1) = 0$,

(MV4) $V(A) = +\infty$ for singular $A \in \mathrm{M}(F)$,

(MV5) $V(A)$ is unchanged if a column or a row of A is multiplied by -1.

These axioms derive extra useful formulas as follows.

Proposition 2.12 ([38])**.** *For a matrix valuation V on a skew field F, the following hold:*

(1) $V(AB) = V(A) + V(B)$ *for* $A, B \in F^{n\times n}$.

(2) $V\begin{pmatrix} A & * \\ O & B \end{pmatrix} = V\begin{pmatrix} A & O \\ * & B \end{pmatrix} = V(A) + V(B)$ *for* $A, B \in \mathrm{M}(F)$, *where* $*$ *denotes any matrix of appropriate size.*

(3) *The equality in* (MV2) *holds whenever* $V(A) \neq V(B)$.

By Proposition 2.12 (1) and (MV2)–(MV4), a matrix valuation V restricted to F (1×1 matrices) is exactly a valuation v on F. This can be extended to $\mathrm{M}(F)$ as ζ, i.e., $V = \zeta$ holds. In general, for any valuation v of F, ζ is a matrix valuation on F [38]; the correspondence between v and V is clearly bijective. Therefore, a matrix valuation is nothing but a valuation of the Dieudonné determinant. See also [16, Section 9.3].

For a matrix $A \in F^{n\times n'}$ over a valuation skew field F with valuation v, we define

$$\zeta_k(A) \coloneqq \min\{\zeta(A[I, J]) \mid I \subseteq \mathrm{Row}(A), J \subseteq \mathrm{Col}(A), |I| = |J| = k\} \tag{2.10}$$

for $k \in [0, \min\{n, n'\}]$. Note that $\zeta_0(A) = 0$, $\zeta_1(A)$ is equal to the minimum of the valuation of an entry in A, and $\zeta_n(A) = \zeta(A)$ for $A \in F^{n\times n}$. In addition, $\zeta_k(A) \neq +\infty$ if and only if $k \leq \operatorname{rank} A$ by Proposition 2.9.

Propositions 2.1 and 2.4 are naturally extended to matrices over valuation skew fields and DVSFs as follows.

Proposition 2.13. *Let F be a valuation skew field with valuation v, valuation ring R, and representative set Q. Then any $A \in R^{n\times n'}$ is uniquely expressed as $A = A_0 + \tilde{A}$, where $A_0 \in Q^{n\times n'}$ and $\tilde{A} \in J(R)^{n\times n'}$.*

Proposition 2.14. *Let F be a DVSF with discrete valuation v and let π and Q be a uniformizer and a representative set of F, respectively.*

(1) *For every $A \in F^{n\times n'}$, there uniquely exists a sequence $(A_d)_{d\in\mathbb{Z}}$ of $n \times n'$ matrices over Q such that $A_d = O$ for all but finitely many $d < 0$ and*

$$A = \sum_{d\in\mathbb{Z}} A_d \pi^d \tag{2.11}$$

in the π-adic topology. If $\ell \coloneqq \zeta_1(A) \in \mathbb{Z}$, then $A_d = O$ for $d < \ell$ and $A_\ell \neq O$.

(2) *If F is complete and $(A_d)_{d\in\mathbb{Z}}$ is a sequence of elements in Q such that $A_d = O$ for all but finitely many $d < 0$, the power series* (2.4) *converges to an $n \times n'$ matrix A over F.*

For a matrix A over a DVR, the matrices A_0 in Propositions 2.13 and 2.14 are the same. As in the scalar case, we call (2.11) the *π-adic expansion* of A.

2.2.4 Smith–McMillan Form

Let F be a valuation skew field with valuation ring R. A matrix over F is called *proper* if its entries are in R. A proper matrix $A \in F^{n\times n}$ is particularly called *biproper* if it is nonsingular and its inverse is also proper, i.e., $A \in \mathrm{GL}_n(R)$. The (right or left) multiplication by biproper matrices are called *biproper transformations*. We establish the *Smith–McMillan form* of matrices over F, which is a canonical form under biproper transformations. This is well-known for matrices over $\mathbb{C}(s)$ as the *Smith–McMillan form at infinity* [71, 99] in the context of control theory.

Proposition 2.15 (Smith–McMillan form)**.** *Let F be a valuation skew field with valuation v and valuation ring R. For $A \in F^{n\times n'}$ of rank r, there exist $S \in \mathrm{GL}_n(R)$, $T \in \mathrm{GL}_{n'}(R)$ and $d_1, \dots, d_r \in F^\times$ such that $v(d_1) \le \cdots \le v(d_r)$ and*

$$SAT = \begin{pmatrix} \mathrm{diag}(d_1, \dots, d_r) & O \\ O & O \end{pmatrix}. \tag{2.12}$$

In addition, the element d_i for $i \in [r]$ is unique up to multiplication by a unit of R and its valuation satisfies

$$v(d_i) = \zeta_i(A) - \zeta_{i-1}(A). \tag{2.13}$$

Proof. We first construct the desired diagonalization. Suppose that $A \neq O$ and $d_1 \in F^\times$ is an entry in A such that $v(d_1) = \zeta_1(A)$. Multiplying permutation matrices to A from left and right, we move d_1 to the top-left entry. Note that permutation matrices are clearly biproper. Then we eliminate the first column of A other than the top entry using d_1. This can be achieved by multiplying an elementary matrix $E_n(1, i; ad_1{}^{-1})$ to A from left for $i = 2, \dots, n$, where a is the $(i,1)$st entry of A. Since $ad_1{}^{-1} \in R$ by $v(d_1) \le v(a)$, this elementary matrix is biproper. We similarly eliminate the first row of A other than the left entry. Now A is in the form $\left(\begin{smallmatrix} d_1 & 0 \\ 0 & B \end{smallmatrix}\right)$ with $B \in F^{(n-1)\times(n'-1)}$. Iteratively applying the same operation for B as long as $B \neq O$, we obtain the decomposition (2.12). Note that $\zeta_1(A) \le \zeta_1(B)$ by (V1) and (V2) and hence $v(d_1) \le \cdots \le v(d_r)$.

We next show the uniqueness part. Since units of R has valuation 0, the formula (2.13) implies the uniqueness of $v(d_1), \dots, v(d_r)$. Let D be the diagonal matrix constructed above. By the ordering of $d_1, \dots, d_r$, it holds $v(d_i) = \zeta_i(D) - \zeta_{i-1}(D)$. Therefore, it suffices to show that $\zeta_k(A)$ is invariant throughout the above procedure for $k \in [0, r]$. It is clear that $\zeta_k(A)$ does not change by row and column permutations. Consider multiplying an elementary matrix $E_n(i, j; a)$ to A from left, where $i, j \in \mathrm{Row}(A)$ with $i \neq j$ and $a \in R$. This corresponds to the operation of adding the ith row multiplied by a to the jth row. Put $A' \coloneqq E_n(i, j; e)A$ and consider a submatrix with rows $I \subseteq \mathrm{Row}(A)$ and columns $J \subseteq \mathrm{Col}(A)$ of cardinality k. If $j \notin I$, then $A'[I, J] = A[I, J]$. If $i, j \in I$, then $A[I, J] = EA[I, J]$ for some elementary matrix E of order k, which means $\zeta(A'[I, J]) = \zeta(A[I, J])$

by (VD1) and (VD2). In the remaining case, i.e., $i \notin I \ni j$, we have

$$A'[I, J] = A[I, J] \nabla (FA[I', J]),$$

where $I' \coloneqq (I \cup \{i\}) \setminus \{j\}$ and $C \in F^{n\times n}$ is the diagonal matrix having a for the ith diagonal entry and 1 for other diagonals. By (MV2), it holds

$$\begin{aligned} \zeta(A'[I, J]) &\geq \min\{\zeta(A[I, J]), \zeta(CA[I', J])\} \\ &= \min\{\zeta(A[I, J]), \zeta(A[I', J]) + v(a)\}. \end{aligned} \tag{2.14}$$

Since $a \in R$, we have $\zeta(A'[I, J]) \geq \zeta_k(A)$. Suppose $\zeta_k(A) = \zeta(A[I, J])$. If $\zeta_k(A) > \zeta(A[I', J]) + v(a)$, the equality of (2.14) is attained. If $\zeta_k(A) = \zeta(A[I', J]) + v(a)$, then $\zeta_k(A) = \zeta(A[I', J])$ by $v(a) \geq 0$ and $\zeta(A[I', J]) \geq \zeta_k(A)$. In addition, we have $\zeta(A'[I', J]) = \zeta(A[I', J])$ from $j \notin I'$, which means $\zeta(A'[I', J]) = \zeta_k(A)$. Hence we have $\zeta_k(A') = \zeta_k(A)$ in all cases. The proof of the right multiplication of elementary matrices is the same. □

Solving (2.13) for $\zeta_k(A)$, we have

$$\zeta_k(A) = \sum_{i=1}^{k} v(d_i) \tag{2.15}$$

for $k \in [0, \operatorname{rank} A]$. It is worth mentioning that $v(d_i) \geq 0$ for any $A \in R^{n\times n'}$ and $i \in [\operatorname{rank} A]$ since $v(d_1) = \zeta_1(A) \geq 0$.

If A is a matrix over a DVSF F, diagonal entries of the Smith–McMillan form of A can be taken as powers of a uniformizer of F as follows.

Proposition 2.16 (Smith–McMillan form for DVSFs)**.** *Let F be a DVSF with valuation ring R and uniformizer π. For $A \in F^{n\times n'}$ of rank r, there exist $S \in \mathrm{GL}_n(R)$, $T \in \mathrm{GL}_{n'}(R)$, and unique $\alpha = (\alpha_i)_{i\in[r]} \in \mathbb{Z}^r$ such that $\alpha_1 \leq \cdots \leq \alpha_r$ and*

$$SAT = \begin{pmatrix} D(\pi^\alpha) & O \\ O & O \end{pmatrix}. \tag{2.16}$$

For $i \in [r]$, the integer α_i is determined by

$$\alpha_i = \zeta_i(A) - \zeta_{i-1}(A). \tag{2.17}$$

Proof. Let $D = S'AT$ be the Smith–McMillan form of A given in Proposition 2.15. For $i \in [r]$, we define α_i as the valuation of the ith diagonal entry d_i of D. Then (2.17) follows

from (2.13). Define a biproper matrix

$$W \coloneqq \begin{pmatrix} \mathrm{diag}(\pi^{\alpha_1} {d_1}^{-1}, \dots, \pi^{\alpha_r} {d_r}^{-1}) & O \\ O & I_{n-r} \end{pmatrix} \in \mathrm{GL}_n(R).$$

Then $WD = WS'AT = UAV$ with $S \coloneqq WS'$ is equal to the right hand side of (2.16), as required. □

The equation (2.15) is rewritten as

$$\zeta_k(A) = \sum_{i=1}^{k} \alpha_i \tag{2.18}$$

for $k \in [0, \mathrm{rank}\, A]$. This equation plays an important role in Section 4.3.1.

We present two propositions for matrices over R which are obtained as corollaries of the Smith–McMillan form. The first one claims that $\zeta_k(A)$ is nonnegative for any proper matrix $A \in R^{n \times n'}$.

Proposition 2.17. *Let R be the valuation ring of a valuation skew field. For $A \in R^{n \times n'}$ and $k \in [0, \min\{n, n'\}]$, it holds $\zeta_k(A) \geq 0$.*

Proof. If $k > r$ with $r \coloneqq \mathrm{rank}\, A$, we have $\zeta_k(A) = +\infty > 0$. If $k \leq r$, the claim holds from (2.15) and $v(d_1), \dots, v(d_r) \geq 0$. □

The second proposition is a characterization of biproper matrices.

Proposition 2.18. *Let F be a valuation skew field with valuation ring R, residue skew field K, and representative set Q, and let $\varphi : R \to K$ be the natural homomorphism. Also, let $A \in R^{n \times n}$ be a square proper matrix and $A_0 \in Q^{n \times n}$ the matrix in Proposition 2.13 with respect to A. Then the following are equivalent:*

(1) *A is biproper.*

(2) *$\zeta(A) = 0$.*

(3) *$\varphi(A_0)$ is nonsingular.*

Proof. Let $SAT = D \coloneqq \mathrm{diag}(d_1, \dots, d_n)$ be the Smith–McMillan form of A. Since S and T are biproper, A is biproper if and only if so is D. This is equivalent to $v(d_i) = 0$ for all $i \in [n]$, where v is the valuation of F. Since $v(d_i)$ is nonnegative for $i \in [n]$, this condition is further equivalent to $\zeta(A) = \sum_{i=1}^{n} v(d_i) = 0$, where the first equality is from (2.15). Thus (1) and (2) are equivalent.

We next consider (3). Let $D_0 \in Q^{n \times n}$ be the matrix obtained from D by Proposition 2.13. By the above argument, A is biproper if and only if $v(d_i) = 0$ for every $i \in [n]$. This is equivalent to the nonsingularity of $\varphi(D)$ because for $i \in [n]$, the ith diagonal of $\varphi(D)$ is nonzero if and only if $v(d_i) = 0$. Applying φ to $D = SAT$ and $A = S^{-1}DT^{-1}$, we obtain $\varphi(D) = \varphi(S)\varphi(A)\varphi(T)$ and $\varphi(A) = \varphi(S^{-1})\varphi(D)\varphi(T^{-1})$. These

imply $\operatorname{rank}\varphi(D) = \operatorname{rank}\varphi(A)$. In addition, it holds $\varphi(A) = \varphi(A_0)$ and $\varphi(D) = \varphi(D_0)$ from $A - A_0, D - D_0 \in J(R)^{n\times n}$. Thus all the statements in Proposition 2.18 are equivalent. □

2.2.5 Jacobson Normal Form

Any DVR is a PID as stated in Section 2.1.2. For a commutative PID R, the *Smith normal form* is a celebrated canonical form of matrices over R under transformations by $\mathrm{GL}_n(R)$. The *Jacobson normal form* [49] is its generalization to general noncommutative PIDs. It can also be seen as a generalization of the Smith–McMillan form over DVRs. Recall from [17, 49] that a nonzero element c of a domain R is said to be *invariant* if $cR = Rc$ and $a \in R \setminus \{0\}$ is called a *total divisor* of $b \in R \setminus \{0\}$ if there exists invariant $c \in R$ such that $bR \subseteq cR \subseteq aR$.

Proposition 2.19 (Jacobson normal form [49, Theorem 16 in Chapter 3]; see [17, Theorem 7.2.1])**.** *Let $A \in R^{n\times m}$ be a matrix of rank r over a PID R*[1]*. There exist $U \in \mathrm{GL}_n(R)$, $V \in \mathrm{GL}_m(R)$ and $e_1, \dots, e_r \in R \setminus \{0\}$ such that e_i is a total divisor of e_{i+1} for $i \in [r-1]$ and*

$$UAV = \begin{pmatrix} \operatorname{diag}(e_1, \dots, e_r) & O \\ O & O \end{pmatrix}.$$

We can also prove Proposition 2.16 by using Proposition 2.19. Namely, the Smith–McMillan form over a DVR R can also be seen as a variant of the Jacobson normal form over R regarded as a PID.

[1] As explained in Section 2.1.1, any PID is an Ore domain, i.e., R can be extended to a skew field F. Thus the rank of A can be defined as that of a matrix over F.

Chapter 3

Preliminaries on Discrete Convex Analysis

In this chapter, we present preliminaries on bipartite matchings, matroids, and two kinds of discrete convex functions used in this thesis. All they are specific topics of discrete convex analysis, which is a field of combinatorial optimization.

3.1 Bipartite Matchings

Let $G = (V, E)$ be an undirected graph. In this thesis, all undirected and directed graphs are finite. A *matching* of G is an edge subset $M \subseteq E$ such that no two distinct edges in M share the same end. A matching M is said to be *perfect* if every vertex of G is covered by some edge in G. The *matching problem* on G is to find a maximum-cardinality matching of M. Given an edge weight $w : E \to \mathbb{R}$, the *minimum-weight perfect matching problem*, or simply the *weighted matching problem*, on G with respect to w is defined as the problem of finding a perfect matching M of G having the minimum weight $w(M)$ among all perfect matchings of G.

3.1.1 Unweighted Bipartite Matching

An undirected graph is called *bipartite* if there exists a bipartition of vertices such that every edge is between different parts in the bipartition. The bipartite matching problem is one of the fundamental and central problem in combinatorial optimization. It admits polynomial-time algorithms [42, 52, 55] and a min-max theorem [52], called the *Kőnig–Egerváry theorem*. To describe the formula, we shall define a *vertex cover* of a graph G as a vertex subset that includes at least one end of every edge of G.

Theorem 3.1 (Kőnig–Egerváry theorem [52]; see [87, Theorem 16.2])**.** *The maximum size of a matching in a bipartite graph G is equal to the minimum size of a vertex cover of G.*

Bipartite matching and ranks of matrices are closely related. Let $A = (A_{i,j}) \in F^{n \times n'}$

be a matrix over a skew field F and put $R \coloneqq \mathrm{Row}(A)$ and $C \coloneqq \mathrm{Col}(A)$. We associate to A a bipartite graph $G(A)$ with vertex set $R \cup C$ and edge set

$$E(A) \coloneqq \{\{i, j\} \mid i \in R,\, j \in C,\, A_{i,j} \neq 0\}$$

The *term-rank* of A, introduced by Ore [77], is the maximum size of a matching in $G(A)$. We denote the term-rank of A by $\text{t-rank}\, A$. By Theorem 3.1, $\text{t-rank}\, A$ is equal to the optimal value of the following problem:

$$\left|\begin{array}{ll} \text{minimize} & n + n' - s - t \\ \text{subject to} & A \text{ has a zero block of size } s \times t, \\ & s \in [0, n], t \in [0, n']. \end{array}\right.$$

Indeed, $\text{t-rank}\, A$ serves as a combinatorial upper bound on $\mathrm{rank}\, A$ as we well see below. When F is a field, this is well-known from the definition (2.9) of the determinant.

Proposition 3.2. *Let $A \in F^{n \times n'}$ be a matrix over a skew field F. Then it holds* $\mathrm{rank}\, A \leq \text{t-rank}\, A$.

Proof. Permuting rows and columns of A, we assume that A is in form of $A = \left(\begin{smallmatrix} X & Y \\ Z & O \end{smallmatrix}\right)$, where O is the zero matrix of size $s \times t$ and $\text{t-rank}\, A = n + n' - s - t$. Then we can decompose A as

$$A = \begin{pmatrix} X & Y \\ Z & O \end{pmatrix} = \begin{pmatrix} X & I_{n'-t} \\ Z & O \end{pmatrix} \begin{pmatrix} I_{n-s} & O \\ O & Y \end{pmatrix}. \tag{3.1}$$

The size of matrices in the right hand side of (3.1) is $n \times p$ and $p \times n'$ with $p \coloneqq \text{t-rank}\, A$. Hence $\mathrm{rank}\, A \leq \text{t-rank}\, A$ by the characterization of $\mathrm{rank}\, A$ (see Section 2.2.2). □

3.1.2 Weighted Bipartite Matching

We next consider the weighted bipartite matching problem, which is also called the *assignment problem*. This is solvable in strongly polynomial-time by the *Hungarian method* [55] for example. Let $G = (U \cup V, E)$ be a bipartite graph with $n \coloneqq |U| = |V|$ and $w : E \to \mathbb{R}$ an edge weight. The dual problem of the LP relaxation of the minimum-weight perfect bipartite matching problem on G is the following (see [87, Theorem 17.5]):

$$\left|\begin{array}{lll} \text{maximize} & \displaystyle\sum_{i \in U} p_i + \sum_{j \in V} q_j & \\ \text{subject to} & p_i + q_j \leq w(e) & (i \in U, j \in V, e = \{i, j\} \in E), \\ & p_i, q_j \in \mathbb{R} & (i \in U, j \in V). \end{array}\right.$$

By the strong duality of linear programming, the optimal value of the dual problem is equal to the minimum-weight of a perfect matching in G. In addition, if w is integer-valued,

then we can take optimal (p, q) as integer vectors.

The following *complementarity theorem* plays an important role in this thesis. Let $G = (U \cup V, E)$ be a bipartite graph equipped with an edge weight $w : E \to \mathbb{R}$. For a dual feasible solution (p, q), we define a bipartite graph $G^{\#} = (U \cup V, E^{\#})$ by

$$E^{\#} \coloneqq \{e \in E \mid p_i + q_j = w(e) \text{ with } e = \{i, j\}, i \in U, j \in V\}. \tag{3.2}$$

Namely, $G^{\#}$ is the subgraph of G obtained by collecting only the "tight" edges. Then the following holds from the complementarity theorem of linear programming.

Proposition 3.3 (complementarity theorem; see [67, Lemma 2.6])**.** *Under the above setting, (p, q) is optimal if and only if $G^{\#}$ has a perfect matching.*

Analogously to the relation between the bipartite matching problem and the rank computation, solving the weighted bipartite matching problem corresponds to computing the valuation of the Dieudonné determinant. Let $A = (A_{i,j}) \in F^{n \times n}$ be a square matrix over a valuation skew field F with valuation v. Recall from Section 2.2.3 that $\zeta(A)$ denotes the valuation of the Dieudonné determinant of A. For the bipartite graph $G(A)$ associated with A, we set an edge weight $w : E(A) \to \mathbb{R}$ as $w(e) \coloneqq v(A_{i,j})$ for $e = \{i, j\} \in E(A)$. We denote by $\hat{\zeta}(A)$ the minimum-weight of a perfect matching in $G(A)$ with respect to the edge weight w. If $G(A)$ has no perfect matching, put $\hat{\zeta}(A) \coloneqq +\infty$. If F is a field, then $\hat{\zeta}(A) \leq \zeta(A)$ by the definition (2.9) of the determinant and the axioms (V1), (V2) of valuations. This inequality is indeed valid even for noncommutative matrices:

Proposition 3.4. *Let $A \in F^{n \times n}$ be a square matrix over a valuation skew field F. Then it holds $\hat{\zeta}(A) \leq \zeta(A)$.*

Proof. By Proposition 3.2, $\hat{\zeta}(A) = +\infty$ implies $\zeta(A) = +\infty$. Suppose $\hat{\zeta}(A) < +\infty$, i.e., $G(A)$ has a perfect matching. Let (p, q) be a dual optimal solution of the maximum-weight perfect matching problem on A. We take diagonal matrices $P, Q \in \mathrm{GL}_n(F)$ such that the valuation of the ith and the jth diagonal entries of P and Q are p_i and q_j, respectively, for every $i \in \mathrm{Row}(A)$ and $j \in \mathrm{Col}(A)$[1]. Put $B \coloneqq P^{-1}AQ^{-1}$. Then the valuation of the (i, j)th entry of B is $w(\{i, j\}) - p_i - q_j \geq 0$ for all $\{i, j\} \in E(A)$. Thus B is a matrix over the valuation ring of F, and hence $\zeta(B) \geq 0$ by Proposition 2.17. By $\zeta(B) = \zeta(A) - \hat{\zeta}(A)$, the desired inequality is proved. □

3.2 Matroids

[1] By the existence of augmenting path algorithms for the weighted matching problem, we can assume that every component of p and q are integer combination of edge weights. Therefore, for every $i \in \mathrm{Row}(A)$ and $j \in \mathrm{Col}(A)$, there must exist $a, b \in F$ such that $v(a) = p_i$ and $v(b) = q_j$, where v is the valuation on F. The matrices P and Q are obtained by arranging these elements in diagonals.

3.2.1 Definitions and Properties

A *matroid* is a pair $\mathbf{M} = (E, \mathcal{I})$ of a finite set E and a family $\mathcal{I} \subseteq 2^E$ such that

(I1) $\varnothing \in \mathcal{I}$,

(I2) for each $I \subseteq J \subseteq E$, if $J \in \mathcal{I}$, then $I \in \mathcal{J}$,

(I3) for each $I, J \in \mathcal{I}$ with $|I| < |J|$, there exists $x \in J \setminus I$ such that $I \cup \{x\} \in \mathcal{I}$.

The set E is called a *ground set* and $I \in \mathcal{I}$ is an *independent set* of $\mathbf{M}$.

A *base* of $\mathbf{M}$ is an independent set that is maximal with respect to inclusion. Let $\mathcal{B}$ denote the family of bases. Then $\mathcal{B}$ is a nonempty set family which satisfies the following:

(BM$_-$) for each $B, B' \in \mathcal{B}$ and $x \in B \setminus B'$, there exists $y \in B' \setminus B$ such that $(B \setminus \{x\}) \cup \{y\} \in \mathcal{B}$.

This property shows that any base of $\mathbf{M}$ has the same cardinality, which is called the *rank* of $\mathbf{M}$. Conversely, a nonempty set family $\mathcal{B} \subseteq 2^E$ satisfying (BM$_-$) is the base family of the matroid $\mathcal{M} = (E, \mathcal{I})$ given by

$$\mathcal{I} \coloneqq \{I \subseteq E \mid \text{there exists } B \in \mathcal{B} \text{ containing } I\}.$$

We thus use both notations $\mathbf{M} = (E, \mathcal{I})$ and $\mathbf{M} = (E, \mathcal{B})$ to designate a matroid whenever convenient. See [71, 78] for proofs.

The *rank function* $\rho : 2^E \to \mathbb{N}$ of a matroid $\mathbf{M} = (E, \mathcal{I})$ is defined by

$$\rho(X) \coloneqq \max\{|I| \mid I \subseteq X, I \in \mathcal{I}\}$$

for $X \subseteq E$.

3.2.2 Examples

In this section, we enumerate several examples of matroids.

Example 3.5 (linear matroid)**.** Let $A \in F^{r \times n}$ be a matrix over a field F and put $E \coloneqq \mathrm{Col}(A)$. Define

$$\mathcal{B}(A) \coloneqq \{B \subseteq E \mid |B| = r,\ A[B] \text{ is nonsingular}\}.$$

If A is of row-full rank, $\mathbf{M}(A) \coloneqq (E, \mathcal{B}(A))$ forms a matroid, called a *linear matroid* represented by A. We refer to each element of $\mathcal{B}(A)$ as a *base* of A. The independent set family $\mathcal{I}$ and the rank function ρ on $\mathbf{M}(A)$ are given by

$$\begin{aligned} \mathcal{I} &= \{J \subseteq E \mid A[J] \text{ is of column-full rank}\}, \\ \rho(J) &= \operatorname{rank} A[J] \end{aligned}$$

for $J \subseteq E$. □

Example 3.6 (free matroid)**.** Let E be a finite set and put $\mathcal{I} \coloneqq 2^E$. Then $(E, \mathcal{I})$ is a matroid called the *free matroid* on E. We have $\rho(X) = |X|$ for $X \subseteq E$. The free matroid is the regular matroid represented by the identity matrix I_n with $n \coloneqq |E|$. □

Example 3.7 (transversal matroid)**.** Let $G = (U \cup V, E)$ be a bipartite graph. Define

$$\mathcal{I} \coloneqq \{I \subseteq U \mid \text{there exists a matching of } G \text{ covering } I\}.$$

Then $(U, \mathcal{I})$ forms a matroid, called a *transversal matroid*. □

3.2.3 Matroid Intersection Problem

The *matroid intersection problem* introduced by Edmonds [24, 25] is the following: given two matroids $\mathbf{M}_1 = (E, \mathcal{I}_1)$ and $\mathbf{M}_2 = (E, \mathcal{I}_2)$ over the same ground set E, we find a common independent set $I \in \mathcal{I}_1 \cap \mathcal{I}_2$ of maximum size. When both matroids are partition matroids, the matroid intersection problem coincides with the bipartite matching problem.

We can solve the matroid intersection problem in polynomial-time [24, 25], where we assumed that one can access given matroids via membership oracles of their independence sets. The matroid intersection problem admits the following min-max theorem.

Theorem 3.8 ([25])**.** *Let $\mathbf{M}_1 = (E, \mathcal{I}_1)$ and $\mathbf{M}_2 = (E, \mathcal{I}_2)$ be matroids over the same ground set E, and ρ_1 and ρ_2 the rank functions of $\mathbf{M}_1$ and $\mathbf{M}_2$, respectively. Then it holds*

$$\max\{|I| \mid I \in \mathcal{I}_1 \cap \mathcal{I}_2\} = \min\{\rho_1(X) + \rho_2(E \setminus X) \mid X \subseteq E\}.$$

When both the matroids $\mathbf{M}_1$ and $\mathbf{M}_2$ are linear and given as matrices A_1, A_2 with $\mathrm{Col}(A_1) = \mathrm{Col}(A_2)$ over the same field, the matroid intersection problem on $\mathbf{M}_1$ and $\mathbf{M}_2$ are called the *linear matroid intersection problem*.

3.3 Discrete Convex Functions

In this section, we introduce two types of discrete convex functions: valuated matroids and 1-dimensional discrete convex functions. The former one is defined on a set family, which is identified with $\{0,1\}^n$, and the latter one is on $\mathbb{Z}$. They are unified as *$M^\natural$-convex (concave) functions* in discrete convex analysis [70], though it is beyond the scope of this thesis.

3.3.1 Valuated Matroids

A *valuated matroid*, introduced by Dress–Wenzel [20, 21], on a finite set E is a function $\omega : 2^E \to \mathbb{R} \cup \{-\infty\}$ satisfying the following condition:

(VM) For any $j \in X \setminus Y$, there exists $j' \in Y \setminus X$ such that $\omega(X) + \omega(Y) \leq \omega(X \cup \{j'\} \setminus \{j\}) + \omega(Y \cup \{j\} \setminus \{j'\})$.

It is easily confirmed that the family $\{X \subseteq E \mid \omega(X) > -\infty\}$ forms a base family of a matroid over E (assuming the family is nonempty), which means that valuated matroids are a generalization of matroids. In addition, valuated matroids can be maximized by a greedy algorithm. Conversely, $\omega : 2^E \to \mathbb{R} \cup \{-\infty\}$ is a valuated matroid if and only if $\omega + p$ is maximized by the greedy algorithm for any linear function $p : 2^E \to \mathbb{R} \cup \{-\infty\}$ [20]. In this way, valuated matroids are recognized as a kind of "concave function" on $2^E \simeq \{0,1\}^n$.

A typical example of valuated matroids arises from the valuation of determinants of matrices over a valuation field [20, 21]. Since the proof essentially relies on the *Grassmann–Plücker identity*, which is an expansion formula of determinants, it cannot be directly applied to valuation skew fields. Nevertheless, Hirai [39, Proposition 2.12] presented another proof which is valid for the degree of rational functions over skew fields. This can be straightforwardly extended to general valuation skew fields as follows.

Proposition 3.9. *Let $A \in F^{n \times n'}$ be a matrix over a valuation skew field F and put $E \coloneqq \mathrm{Col}(A)$. The function $\omega : 2^E \to \mathbb{R} \cup \{-\infty\}$ given by*

$$\omega(J) \coloneqq \begin{cases} -\zeta(A[X]) & (|J| = n), \\ -\infty & (\textit{otherwise}) \end{cases}$$

for $X \subseteq E$ is a valuated matroid on E.

Proof. A local characterization [71, Theorem 5.2.25] of valuated matroids claims that ω is a valuated matroid if and only if (i) $\{X \subseteq E \mid \omega(X) \neq -\infty\}$ forms a base family of a matroid and (ii) ω satisfies (VM) for $X, Y \subseteq E$ with $|X \setminus Y| = |Y \setminus X| = 2$. The condition (i) holds since the linear independence of column vectors of A defines a matroid.

We show the condition (ii). Let $X, Y \subseteq E$ with $\omega(X), \omega(Y) \neq -\infty$ and $|X \setminus Y| = |Y \setminus X| = 2$. Put $A' \coloneqq A[X \cup Y]$. By a column permutation, we arrange columns of $X \cap Y$ in the left $n-2$ columns of A' without changing ω. In addition, by elementary row operations, we can assume without changing ω that A' is in the form of $\left(\begin{smallmatrix} S & T \\ O & U \end{smallmatrix}\right)$, where S is a nonsingular $(n-2) \times (n-2)$ matrix, T is an $(n-2) \times 4$ matrix, and U is a 2×4 matrix. Assume that $X \setminus Y = \{1, 2\}$ and $Y \setminus X = \{3, 4\}$. For distinct $j, j' \in \{1, 2, 3, 4\}$, define $u_{j,j'}$ as the valuation of the Dieudonné determinant of the 2×2 submatrix of U with column set $\{j, j'\}$. Then $\omega((X \cap Y) \cup \{j, j'\}) = -\zeta(S) - u_{j,j'}$ for any distinct $j, j' \in \{1, 2, 3, 4\}$. Hence (VM) is equivalent to the following:

(4PT) The minimum value of $u_{1,2} + u_{3,4}$, $u_{1,3} + u_{2,4}$, $u_{1,4} + u_{2,3}$ is attained at least twice.

Now $u_{1,2} \neq -\infty$ by $\omega(X) \neq -\infty$. By a column permutation, we assume that the $(1,1)$st entry of U is nonzero. In addition, we make the $(2,1)$st entry of U zero using an

elementary row operation. If the $(2,3)$rd entry is nonzero, make the $(1,3)$rd entry zero in the same way. Then U is in form of either

$$U = \begin{pmatrix} a & c & d & e \\ 0 & b & 0 & f \end{pmatrix} \quad \text{or} \quad \begin{pmatrix} a & c & 0 & e \\ 0 & b & d & f \end{pmatrix}.$$

In the left case, $u_{1,2}+u_{3,4} = u_{1,4}+u_{2,3} = v(a)+v(b)+v(d)+v(f)$ and $u_{1,3}+u_{2,4} = +\infty$, where v is the valuation of F. In the right case, $u_{1,2} + u_{3,4} = v(a) + v(b) + v(d) + v(e)$, $u_{1,4} + u_{2,3} = v(a) + v(f) + v(c) + v(d)$ and $u_{1,3} + u_{2,4} = v(a) + v(d) + \zeta(\begin{smallmatrix} c & e \\ b & f \end{smallmatrix}) \geq v(a) + v(d) + \max\{v(c) + v(f), v(b) + v(e)\}$ by Proposition 2.12 (3). The equality is attained if $v(c) + v(f) \neq v(b) + v(e)$. Hence (4PT) is satisfied for all cases. □

Let R and C be finite sets. Murota [68] introduced a *valuated bimatroid* over (R, C) as a function $w : 2^R \times 2^C \to \mathbb{R} \cup \{-\infty\}$ satisfying the following conditions:

(VBM1) For any $i' \in I' \setminus I$, at least one of the following holds:

(a1) $\exists j' \in J' \setminus J$: $w(I, J) + w(I', J') \leq w(I \cup \{i'\}, J \cup \{j'\}) + w(I' \setminus \{i'\}, J' \setminus \{j'\})$,

(b1) $\exists i \in I \setminus I'$: $w(I, J) + w(I', J') \leq w(I \cup \{i'\} \setminus \{i\}, J) + w(I \cup \{i\} \setminus \{i'\}, J')$.

(VBM2) For any $j' \in J' \setminus J$, at least one of the following holds:

(a2) $\exists i \in I \setminus I'$: $w(I, J) + w(I', J') \leq w(I \setminus \{i\}, J \setminus \{j\}) + w(I' \cup \{i\}, J' \cup \{j\})$,

(b2) $\exists j' \in J' \setminus J$: $w(I, J) + w(I', J') \leq w(I, J \cup \{j'\} \setminus \{j\}) + w(I', J \cup \{j\} \setminus \{j'\})$.

The following is a noncommutative generalization of [68, Remark 2].

Proposition 3.10. *Let $A \in F^{n \times n'}$ be a matrix over a valuation skew field F with rows $R := \mathrm{Row}(A)$ and columns $C := \mathrm{Col}(A)$. Define $w : 2^R \times 2^C \to \mathbb{R} \cup \{-\infty\}$ as*

$$w(I, J) := \begin{cases} -\zeta(A[I, J]) & (|I| = |J|), \\ -\infty & (\textit{otherwise}) \end{cases} \tag{3.3}$$

for $I \subseteq R$ and $J \subseteq C$. Then w is a valuated bimatroid.

Proof. Consider an $n \times (n + n')$ skew function matrix $B := \begin{pmatrix} I_n & A \end{pmatrix}$ with row set R and column set $E := R \cup C$. Then there is a one-to-one correspondence between a submatrix of A and a submatrix of B with row set R given by $2^R \times 2^C \ni (I, J) \mapsto (R, (R \setminus I) \cup J) \in 2^R \times 2^E$. In particular, if $|I| = |J| =: k$, then $|R| = |(R \setminus I) \cup J|$ and

$$\zeta(B[(R \setminus I) \cup J]) = \zeta \begin{pmatrix} I_k & A[R \setminus I, J] \\ O & A[I, J] \end{pmatrix} = \zeta(A[I, J]) = -w(I, J).$$

Define a map $\omega : E \to \mathbb{R} \cup \{-\infty\}$ by

$$\omega(X) \coloneqq \begin{cases} -\zeta(B[X]) \ (= w(R \setminus X, X \cap C)) & (|X| = n), \\ -\infty & (\text{otherwise}) \end{cases}$$

for $X \subseteq E$. Then w satisfies (VBM1) and (VBM2) if and only if ω is a valuated matroid, which was already shown in Proposition 3.9. □

By a kind of greedy algorithm, one can obtain sequences $\varnothing = I_0 \subseteq I_1 \subseteq \cdots I_{n^*} \subseteq R$ and $\varnothing = J_0 \subseteq J_1 \subseteq \cdots J_{n^*} \subseteq C$ with $n^* \coloneqq \min\{|R|, |C|\}$ such that (I_k, J_k) is a maximizer of the right-hand side in

$$d_k \coloneqq \{w(I, J) \mid |I| = |J| = k\} \tag{3.4}$$

for every $k \in [0, n^*]$ [68]. Therefore, from Proposition 3.10, any algorithm to compute valuations of the Dieudonné determinants can be applied to compute $\zeta_k(A)$ defined by (2.10).

3.3.2 Univariate Discrete Convex Functions

A *univariate discrete function*, or a *discrete function* for short, is a function $f : \mathbb{Z} \to \mathbb{R} \cup \{+\infty, -\infty\}$. A discrete function $f : \mathbb{Z} \to \mathbb{R} \cup \{+\infty\}$ is said to be *convex* if

$$f(x-1) + f(x+1) \geq 2f(x)$$

for all $x \in \mathbb{Z}$. We call a function $g : \mathbb{Z} \to \mathbb{R} \cup \{-\infty\}$ *concave* if $-g$ is convex.

Let $f : \mathbb{Z} \to \mathbb{R} \cup \{+\infty\}$ be a function such that $f(x) \in \mathbb{R}$ for some $x \in \mathbb{Z}$. The *concave conjugate* of f is a function $f^\circ : \mathbb{R} \to \mathbb{R} \cup \{-\infty\}$ defined by

$$f^\circ(y) \coloneqq \inf_{x \in \mathbb{R}} (f(x) - xy)$$

for $y \in \mathbb{R}$. Similarly, for a function $g : \mathbb{Z} \to \mathbb{R} \cup \{-\infty\}$ with $g(y) \in \mathbb{R}$ for some $y \in \mathbb{R}$, the *convex conjugate* of g is a function $g^\bullet : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ given by

$$g^\bullet(x) \coloneqq \sup_{y \in \mathbb{R}} (g(y) + xy)$$

for $x \in \mathbb{R}$. The maps $f \mapsto f^\circ$ and $g \mapsto g^\bullet$ are referred to as the *concave* and *convex discrete Legendre transform*, respectively.

Suppose that f and g are integer-valued. Then f° and $g^\bullet$ can be regarded as $f^\circ : \mathbb{Z} \to \mathbb{Z} \cup \{-\infty\}$ and $g^\bullet : \mathbb{Z} \to \mathbb{Z} \cup \{+\infty\}$. In this case, f° and $g^\bullet$ are discrete concave and convex functions, respectively. If f is convex and g is concave,

$$(f^\circ)^\bullet = f, \quad (g^\bullet)^\circ = g \tag{3.5}$$

hold. Hence the Legendre transformation establishes a one-to-one correspondence between integer-valued discrete convex and concave functions. See [70] for details of discrete convex/concave functions and their Legendre transform.

Example 3.11. Let w be a valuated bimatroid over (R, C) and d_k be defined by (3.4) for $k \in [0, n^*]$ with $n^* := \min\{|R|, |C|\}$. We identify a sequence $(d_0, \ldots, d_{n^*})$ with a function $\check{d} : \mathbb{Z} \to \mathbb{R} \cup \{-\infty\}$ defined by $\check{d}(k) := d_k$ if $k \in [0, n^*]$ and $-\infty$ otherwise. If w is obtained from a matrix A over a valuation skew field F by (3.3), then $d_k = -\zeta_k(A)$. From (2.13), the difference $\zeta_k(A) - \zeta_{k-1}(A)$ is non-decreasing with respect to k, which means $\zeta_{k-1}(A) + \zeta_{k+1}(A) \geq 2\zeta_k(A)$. Hence $(d_0, \ldots, d_{n^*})$ is concave in this case. Indeed, the sequence is concave for any valuated bimatroid [68, Theorem 1].

When w comes from a matrix over a DVSF, then $(d_0, \ldots, d_k)$ is integer-valued. We will encounter the Legendre conjugate of this sequence in Section 4.3. □

Chapter 4

Computing Valuations of the Dieudonné Determinants

In this chapter, we address the problem of computing valuations of the Dieudonné determinants of matrices over DVSFs. From natural requirements in dealing with a DVSF in computers, we need to assume that the DVSF satisfies a certain condition; such a DVSF is called *split*. Section 4.1 explains the problem of computational models and introduces properties of split DVSFs. We show in Sections 4.2 and 4.3 that two existing algorithms for computing the degree of determinants, the combinatorial relaxation and the matrix expansion, can be extended to matrices over split DVSFs.

These algorithms require an upper bound on the valuation of the Dieudonné determinant as a previous knowledge. Section 4.4 shows that the inverse skew Laurent series rings, which are the completion of the quotient of the skew polynomial rings, characterize DVSFs on which the upper bound can be easily estimated.

4.1 Computational Model of DVSFs

Let $A \in F^{n \times n}$ be a square matrix over a DVSF F with valuation v. Before discussing algorithms to compute $\zeta(A)\,(:= v \operatorname{Det} A)$, we need to clarify a computational model to deal with representation and operations on F.

The simplest model is the arithmetic model on F, i.e., an element in F is stored in a unit memory cell and we can perform arithmetic operations on F in constant time. In this model, one can compute $\zeta(A)$ in $\mathrm{O}(n^\omega)$-time by the Gaussian elimination, where ω is the exponent in the time complexity of multiplying two matrices. However, this model is too simplified and cannot catch the computational cost needed in the standard representation of some DVSF like the rational function field $K(s)$ over a field K.

In this chapter, we represent each element in F as the form of the π-adic expansion (2.11), where π is a uniformizer of F. By shifting valuations of matrix entries to

nonnegative numbers, we assume that A is given as

$$A = \sum_{d=0}^{\ell} A_d \pi^d, \tag{4.1}$$

where $\ell \in \mathbb{N}$ and $A_0, \ldots, A_\ell$ are matrices over a representative set Q of F. Note that A is a matrix over the valuation ring R of F.

We would like to adopt the "arithmetic model over Q". Now one difficulty arises: Q might not be a skew field, i.e., arithmetic operations on Q might not be closed. We thus require F to have a representative set that is a skew subfield of the valuation ring of F. Such a DVSF is called *split* [22].

4.1.1 Split DVSFs

A DVSF F is said to be *spilt* if it has a representative set Q such that it is a subring of the valuation ring R of F. Similarly, a DVR R is called *split* if its quotient skew field F (see Remark 2.2) is split. Such Q is called a *coefficient skew subfield* or a *Cohen skew subfield* of F and of R.

Let F be a split DVSF with coefficient skew subfield Q and residue skew field K. Since elements in Q and K correspond bijectively, Q and K must be isomorphic skew fields. We thus call Q "the" coefficient skew subfield of F. This observation also implies that F could be split only if F is *equicharacteristic*, i.e., F and K have the same characteristic. For example, the field of p-adic numbers is not split as the characteristics of $\mathbb{Q}$ and $\mathbb{F}_p$ are different. Indeed, if F is a field, then F is split if and only if F is equicharacteristic [13, Theorem 9]. Therefore, by Proposition 2.4, a complete split DVF F is isomorphic to the Laurent series field $K((s))$ over the residue field K of F. This is a special case of the Cohen structure theorem for complete commutative Noetherian local rings [13].

The situation is much more complicated in the general noncommutative case. No characterization of a DVSF to be split is yet known; Vidal [100] gave an equicharacteristic but non-split example of a DVSF. Nevertheless, as we have seen in Section 2.1.3, a skew inverse Laurent series field $K\big((s^{-1};\sigma,\delta)\big)$ and a skew Laurent series field $K((s;\sigma))$ over a skew field K are split, where their coefficient skew subfields are both K.

Let F be a complete split DVSF, K the coefficient skew subfield and π a uniformizer. Then Proposition 2.4 implies that the commutation rule between π and each $a \in K$ completely determines the ring structure of F. The element πa can be uniquely expressed as

$$\pi a = \sum_{d=0}^{\infty} \delta_d(a) \pi^{d+1}, \tag{4.2}$$

where $\delta_d : K \to K$ is some map for all $d \in \mathbb{N}$. The family of maps $(\delta_d)_{d\in\mathbb{N}}$ satisfies the following [84]:

(HD1) δ_d is additive for $d \in \mathbb{N}$.

(HD2) $\delta_d(ab) = \sum_{i=0}^{d} \delta_i(a)\Delta_i^d(b)$ for $d \in \mathbb{N}$ and $a, b \in K$, where $\Delta_i^d : K \to K$ is defined by

$$\Delta_i^d \coloneqq \sum_{\substack{j_0,\dots,j_i \in \mathbb{N} \\ j_0+\cdots+j_i=d-i}} \delta_{j_0} \cdots \delta_{j_i}$$

for $d \in \mathbb{N}$ and $i \in [0, d]$.

(HD3) δ_0 is an automorphism on K.

In fact, (HD1) and (HD2) are derived from the distributive law $\pi(a+b) = \pi a + \pi b$ and the associative law $\pi(ab) = (\pi a)b$, respectively [26, 91]. From (HD1), (HD2) for $d = 0$, and $\delta_0(1) = 1$ by $\pi 1 = 1\pi$, the leading map δ_0 must be a homomorphism on K. It further must be surjective by (DVR1), which implies (HD3).

Generally, a sequence $(\delta_d)_{d\in\mathbb{N}}$ of maps on a skew field K is called a *higher σ-derivation* [26, 91] of K (with $\sigma \coloneqq \delta_0$) if it satisfies (HD1)–(HD3). For a higher σ-derivation $(\delta_d)_{d\in\mathbb{N}}$, we denote by $K[[s; (\delta_d)]]$ the ring of formal power series over K in indeterminate s, whose every element f is uniquely expressed as (2.6). The addition on $K[[s; (\delta_d)]]$ is naturally defined and the multiplication is induced from

$$sa = \sum_{d=0}^{\infty} \delta_d(a) s^{d+1}$$

for $a \in K$. This ring is an Ore domain and thus has a quotient skew field $K((s; (\delta_d)))$. As the usual formal power series ring, each $f \in K((s; (\delta_d)))$ is represented as a formal Laurent series

$$f = \sum_{d=\ell}^{\infty} a_d s^d$$

with $a_d \in K$ for every $d \in \mathbb{Z}$. Defining the *order* of $f \in K((s; (\delta_d)))$ as the minimum $\ell \in \mathbb{N}$ with $a_\ell \neq 0$, the skew field $K((s; (\delta_d)))$ becomes a complete split DVSF with respect to the order [84]; its valuation ring is $K[[s; (\delta_d)]]$, its (one choice of a) uniformizer is s, and its coefficient skew subfield is K. Conversely, as seen above, we have:

Proposition 4.1 ([84, Proposition 1.6 in p. 292])**.** *Let F be a complete split DVSF with coefficient skew subfield K. Then F is isomorphic to $K((s; (\delta_d)))$, where $(\delta_d)_{d\in\mathbb{N}}$ is the higher δ_0-derivation of K determined by* (4.2)*.*

Corollary 4.2. *Let R be a complete split DVR with coefficient skew subfield K. Then R is isomorphic to $K[[s; (\delta_d)]]$, where $(\delta_d)_{d\in\mathbb{N}}$ is the higher δ_0-derivation of K determined by* (4.2)*.*

Note that since any split DVSF F and DVR R are a skew subfield and a subring of a complete split DVSF and DVR (see Theorem 2.3, F and R are isomorphic to a skew subfield of $K((s;(\delta_d)))$ and a subring of $K((s;(\delta_d)))$, respectively.

Example 4.3. We give some examples of higher σ-derivations and corresponding complete split DVSFs. Let K be a skew field and σ an automorphism on K. Then $(\sigma, 0, 0, \dots)$ is a higher σ-derivation and $K((s;(\sigma,0,0,\dots))) = K((s;\sigma))$. In particular, the case when K is a field and $\sigma = \mathrm{id}$ corresponds to the representation of complete equicharacteristic DVFs described above. More generally, let δ be a *right σ-derivation*, i.e., an additive map satisfying $\delta(ab) = \delta(a)\sigma(b) + a\delta(b)$ for $a, b \in K$. Then $(\sigma, \sigma\delta, \sigma\delta^2, \dots)$ is a higher σ-derivation [14, Section 2.1]. If δ is a left σ-derivation instead of the right one, $-\sigma^{-1}\delta$ is a right σ^{-1}-derivation, and hence $(\delta_d)_{d\in\mathbb{N}}$ defined by (2.8) is a higher σ^{-1}-derivation; this is consistent with the fact that $K\big((s^{-1};\sigma,\delta)\big)$ is isomorphic to $K((t;(\delta_d)))$. Another type of a higher σ-derivation is given in [8]. Dumas [22] provides a survey for higher σ-derivations. □

The following lemma provides a relation between coefficients in the π-adic expansions of $a \in R$ and πa.

Lemma 4.4. *Let R be a split DVR with coefficient skew subfield K and uniformizer π, and (δ_d) the higher δ_0-derivation such that R is isomorphic to $K[[s;(\delta_d)]]$ For $a = \sum_{d=0}^{\infty} a_d\pi^d \in R$ with $a_0, a_1, \dots \in K$, the coefficient b_d of π^d in the π-adic expansion of πa satisfies*

$$b_d = \begin{cases} \displaystyle\sum_{k=0}^{d-1} \delta_k(a_{d-k-1}) & (d \geq 1), \\ 0 & (d = 0). \end{cases} \tag{4.3}$$

Proof. Using (4.2), we can rewrite πa as

$$\pi a = \sum_{d=0}^{\infty} \pi a_d \pi^d = \sum_{d=0}^{\infty} \left(\sum_{k=0}^{\infty} \delta_k(a_d)\pi^{k+1} \right) \pi^d = \sum_{d=1}^{\infty} \left(\sum_{k=0}^{d-1} \delta_k(a_{d-k-1}) \right) \pi^d$$

as required. □

Let F be a split DVSF with coefficient skew subfield K and associated higher δ_0-derivative $(\delta_d)_{d\in\mathbb{N}}$. As a computational model, we adopt the arithmetic model on K and assume that one can compute $\delta_d(a)$ for every $d \in \mathbb{N}$ and $a \in K$ in constant time. In this model, if we know the leading $M+1$ coefficients $a_0, \dots, a_M$ in the π-adic expansion of $a \in K$, we can compute those of πa in $\mathrm{O}(M^2)$-time by (4.3).

4.1.2 Truncating Higher-Valuation Terms

Each entry in the input matrix A in (4.1) has only $\ell+1$ terms. However, once we multiply π from left to A, then the number of terms in each entry becomes $+\infty$ by (4.2). The

following proposition states that one can truncate the higher-valuation terms without changing $\zeta(A)$ drastically.

Proposition 4.5. *Let F be a DVSF with uniformizer π and let $A = \sum_{d=0}^{\ell} A_d \pi^d \in F^{n\times n}$ be a matrix in form of (4.1). For any $M \in \mathbb{N}$ and $\tilde{A} \coloneqq \sum_{d=0}^{M} A_d \pi^d$, the following hold:*

(1) *If $\zeta(A) \leq M$, then $\zeta(A) = \zeta(\tilde{A})$.*

(2) *If $\zeta(A) > M$, then $\zeta(\tilde{A}) > M$.*

Proof. Let v and R be the valuation and the valuation ring of F, respectively. Recall $J(R) = \pi R = R\pi$ from (DVR1) and let $\varphi : R \to R/J(R)^{M+1}$ be the natural homomorphism. It is easily checked that $\varphi(a) \neq 0$ if and only if $v(a) \leq M$ and $\varphi(a) = \varphi(b) \neq 0$ implies $v(a) = v(b) \leq M$ for $a, b \in R$.

Let $P = (P_{i,j}), Q = (Q_{i,j}) \in R^{n\times n}$ be any square matrices over R with $\varphi(P) = \varphi(Q)$. Let D and E be the Smith–McMillan forms of P and Q, respectively. We show $\varphi(D) = \varphi(E)$ by tracing the procedure to obtain the Smith–McMillan forms D, E given in the proof of Proposition 2.15. First, we find a matrix entry having the minimum valuation of each P and Q, and move it to the top-left. If the minimum valuation $\zeta_1(P)$ of an entry in P is larger than M, then $\varphi(P) = O$ and thus $\varphi(Q) = O$ by $\varphi(P) = \varphi(Q)$. Thus $\varphi(D) = \varphi(E) = O$ in this case. Suppose $v(P_{i,j}) = \zeta_1(P) \leq M$. By $\varphi(P_{i,j}) = \varphi(Q_{i,j}) \neq 0$, it holds $v(P_{i,j}) = v(Q_{i,j})$ and $\zeta_1(P) = \zeta_1(Q)$. Hence the top-left entries of $\varphi(D)$ and $\varphi(E)$ are the same. After moving the (i, j)th entries in P and Q to the top-left, we eliminate the first row and columns except for the top-left entries. Since φ is a homomorphism, $\varphi(P)$ remains to be the same as $\varphi(Q)$ after this elimination. Applying the above arguments to the bottom-right $(n-1)\times(n-1)$ submatrix recursively, we have $\varphi(D) = \varphi(E)$.

Let $\mathrm{diag}(d_1, \ldots, d_n)$ and $\mathrm{diag}(\tilde{d}_1, \ldots, \tilde{d}_n)$ be the Smith–McMillan forms of A and $\tilde{A}$, respectively. By $\varphi(A) = \varphi(\tilde{A})$ and the above arguments, the images of their Smith–McMillan forms by φ are the same, i.e., $\varphi(d_i) = \varphi(\tilde{d}_i)$ for $i \in [n]$.

Suppose that $\zeta(A) \leq M$. From $\sum_{i=1}^{n} v(d_i) = \zeta(A) \leq M$ and $v(d_i) \geq 0$ for $i \in [n]$, it holds $v(d_i) \leq M$ and thus $\varphi(\tilde{d}_i) = \varphi(d_i) \neq 0$. This means $v(d_i) = v(\tilde{d}_i)$ for $i \in [n]$. Hence $\zeta(A) = \sum_{i=1}^{n} v(d_i) = \sum_{i=1}^{n} v(\tilde{d}_i) = \zeta(\tilde{A})$

Next, suppose that $\zeta(A) > M$. If $v(d_i) \leq M$ for all $i \in [n]$, then $v(d_i) = v(\tilde{d}_i)$ and $\zeta(\tilde{A}) = \zeta(A) > M$ in the same way as above. If $v(d_n) > M$, then $\varphi(\tilde{d}_n) = \varphi(d_n) = 0$, which implies $\zeta(\tilde{A}) \geq v(\tilde{d}_n) > M$. □

By technical reasons, our algorithms assume that A is singular or an upper bound M on $\zeta(A)$ is given. From Proposition 4.5, we can compute $\zeta(A)$ by computing it for $\tilde{A} \coloneqq \sum_{d=0}^{M} A_d \pi^d$ instead of A. Hence we can assume $\ell = \mathrm{O}(M)$ by truncating higher-valuation terms in A if needed.

4.2 Combinatorial Relaxation Algorithm

This section presents the combinatorial relaxation algorithm for computing valuations of the Dieudonné determinants of matrices over DVSFs. First, Section 4.2.1 reviews the classical combinatorial relaxation algorithm of Murota [67] for polynomial matrices over fields. Then Section 4.2.2 describes an algorithm which is faithful to the original framework of the combinatorial relaxation described in Section 1.2. However, a naive implementation of the faithful algorithm requires an additional oracle that was not assumed in Section 4.1. We present in Section 4.2.3 an improved algorithm to avoid this problem and estimate time complexity.

4.2.1 Classical Algorithm for Polynomial Matrices

In this section, we review the classical combinatorial relaxation algorithm of Murota [67] to compute $\deg \det A$ for a polynomial matrix $A = (A_{i,j}) \in K[s]^{n \times n}$ over a field K. Algorithm's outline was described in Section 1.2 and here we give more concrete descriptions.

We begin with some preliminaries. Let $G(A) = (R \cup C, E(A))$ be the bipartite graph associated with A, where $R \coloneqq \mathrm{Row}(A)$ and $C \coloneqq \mathrm{Col}(A)$. We set a weight of an edge $e = \{i, j\} \in E(A)$ by $c_e = c_{i,j} \coloneqq \deg A_{i,j}$. Put

$$\begin{aligned} d(A) &\coloneqq \deg \det A, \\ \hat{d}(A) &\coloneqq \text{the maximum weight of a perfect matching in } G(A). \end{aligned}$$

Since $-\deg$ is a valuation on $\mathbb{R}(s)$, it holds $d(A) \leq \hat{d}(A)$ by Proposition 3.4. We say that A is *upper-tight* if $d(A) = \hat{d}(A)$.

Consider the following dual problem of the linear relaxation of the maximum-weight bipartite matching problem on $G(A)$:

$$\mathrm{D}(A) \left| \begin{array}{lll} \text{minimize} & \displaystyle\sum_{j \in C} q_j - \sum_{i \in R} p_i & \\ \text{subject to} & q_j - p_i \geq c_{i,j} & (i \in R, j \in C, \{i, j\} \in E(A)), \\ & p_i, q_j \in \mathbb{N} & (i \in R, j \in C). \end{array} \right.$$

Note that the problem $\mathrm{D}(A)$ slightly differs from the dual problem given in Section 3.1.2, though they are essentially equivalent. Let (p, q) be a feasible solution of $\mathrm{D}(A)$. The *tight coefficient matrix* of A with respect to (p, q) is a matrix $A^{\#} = (A^{\#}_{i,j}) \in K^{n \times n}$ defined by

$$A^{\#}_{i,j} \coloneqq \text{the coefficient of } s^{q_j - p_i} \text{ in } A_{i,j}$$

for $i \in R$ and $j \in C$. Since $G(A^{\#})$ coincides with the bipartite graph $G^{\#}$ obtained by collecting the tight edges, the following holds as a restatement of Proposition 3.3:

Proposition 4.6 ([67, Proposition 2.4])**.** *Let $A^{\#}$ be the tight coefficient matrix of A with respect to a feasible solution (p, q) of $\mathrm{D}(A)$. Then (p, q) is optimal if and only if* t-rank $A^{\#} = n$.

Let $A^{\#}$ be the tight coefficient matrix with respect to a dual optimal solution (p, q). By the definition of the determinant, $\det A^{\#}$ coincides with the coefficient of $s^{\hat{d}(A)}$ in $\det A$. This means:

Proposition 4.7 ([67, Proposition 2.6])**.** *A polynomial matrix A is upper-tight if and only if it has a nonsingular tight coefficient matrix.*

The combinatorial relaxation algorithm for A runs in accordance with the outline given in Section 1.2. Since $d(A)$ and $\hat{d}(A)$ are integral, the gap between $d(A)$ and $\hat{d}(A)$ decreases by at least 1 for each iteration (unless $d(A) = +\infty$). Thus, the algorithm terminates in at most $\hat{d}(A) \leq \ell n$ iterations, where ℓ is the maximum degree of an entry in A.

By Proposition 4.7, check of upper-tightness in Phase 2 can be done by testing the nonsingularity of the tight coefficient matrix $A^{\#}$ with respect to a dual optimal solution.

One example of the modification in Phase 3 is as follows. Suppose that A is not upper-tight and let $A^{\#}$ be the tight coefficient matrix with respect to a dual optimal solution (p, q). By Proposition 4.7, $A^{\#}$ is singular. Hence we can take $U \in \mathrm{GL}_n(K)$ such that t-rank $UA^{\#} = \operatorname{rank} UA^{\#} = \operatorname{rank} A < n$. Then $\bar{A} \coloneqq U'A$ with $U' \coloneqq D(s^p)UD(s^{-p})$ satisfies $d(\bar{A}) = d(A)$ and $\hat{d}(\bar{A}) \leq d(A) - 1$, as required. If we sort R in ascending order of p and take U upper-triangular, then U' is *unimodular*, i.e., $U' \in \mathrm{GL}_n(K[s])$. We remark that the modification in Phase 3 is not restricted to this algorithm; see Section 1.2.

4.2.2 Faithful Algorithm for Matrices over DVSFs

This section extends the combinatorial relaxation algorithm to matrices over split DVSFs. Let F be a split DVSF and $(\delta_d)_{d\in\mathbb{N}}$ the higher δ_0-derivations associated with F. Let $A = (A_{i,j}) \in F^{n\times n}$ be a square matrix in form (4.1) and M be an upper bound on $\zeta(A)$ which is valid when A is nonsingular.

Recall from Section 3.1.2 that A is associated with the bipartite graph $G(A)$ equipped with an integral edge weight and $\hat{\zeta}(A)$ denotes the minimum weight of a perfect matching in $G(A)$. By Proposition 3.4, $\hat{\zeta}(A)$ serves as a lower bound on $\zeta(A)$. We say that A is *upper-tight* if $\hat{\zeta}(A) = \zeta(A)$. The combinatorial relaxation framework to compute $\zeta(A)$ is the following:

Faithful Combinatorial Relaxation Algorithm over DVSFs

Phase 1a. Compute $\hat{\zeta}(A)$ by solving the minimum-weight perfect matching problem. If $\hat{\zeta}(A) > M$, output $+\infty$ and halt.

Phase 2a. If A is upper-tight, output $\hat{\zeta}(A)$ and halt.

Phase 3a. Modify A into $\bar{A}$ such that $\zeta(A) = \zeta(\bar{A})$ and $\hat{\zeta}(A) < \hat{\zeta}(\bar{A})$. Go back to Phase 1a.

Since the input matrix A is over the valuation ring R of F, each edge in $G(A)$ has a nonnegative weight, from which $\hat{\zeta}(A) \geq 0$ holds. Therefore, as in the classical combinatorial relaxation algorithm, the number of iterations is at most $\zeta(A) \leq M$. In the remaining of this section, we explain how the upper-testing testing in Phase 2a and the matrix modification in Phase 3a are generalized to matrices over DVSFs.

First, we consider Phase 2a. Let v be the valuation and π a uniformizer of F. Denote by $\mathrm{D}(A)$ the dual problem of the minimum-weight perfect matching problem on $G(A)$ given in Section 3.1.2. For $p, q \in \mathbb{Z}^n$, put

$$C = (C_{i,j}) \coloneqq D(\pi^{-p})AD(\pi^{-q}). \tag{4.4}$$

Then for every $i \in \mathrm{Row}(A)$ and $j \in \mathrm{Col}(A)$, we have

$$v(C_{i,j}) = v(\pi^{-p_i} A_{i,j} \pi^{-q_j}) = v(A_{i,j}) - p_i - q_j,$$

which is nonnegative if (p, q) is feasible to $\mathrm{D}(A)$. In particular, if (p, q) is feasible, then $C \in R^{n \times n}$.

The *tight coefficient matrix* of A with respect to a feasible solution (p, q) of $\mathrm{D}(A)$ is the coefficient matrix C_0 of π^0 in the π-adic expansion of C. In particular, when F is a field, the (i, j)th entry of the tight coefficient matrix is equal to the coefficient of $\pi^{p_i+q_j}$ in the π-adic expansion of $A_{i,j}$ for $i \in \mathrm{Row}(A)$ and $j \in \mathrm{Col}(A)$. Note that C_0 depends on (p, q). Then the following is a generalization of Proposition 4.6:

Proposition 4.8. *Let C_0 be the tight coefficient matrix of A with respect to an integral feasible solution (p, q) of $\mathrm{D}(A)$. Then (p, q) is optimal if and only if* $\text{t-rank}\, C_0 = n$.

Proof. For $i \in \mathrm{Row}(A)$ and $j \in \mathrm{Col}(A)$, the (i, j)th entry in C_0 is nonzero if and only if $v(C_{i,j}) = 0$, which is equivalent to $v(A_{i,j}) = p_i + q_j$. Thus $G(C_0)$ coincides with the subgraph $G^{\#}$ of $G(A)$ defined by (3.2) with respect to (p, q). By Proposition 3.3, having a perfect matching for $G(C_0)$ is equivalent to the optimality of (p, q). □

Proposition 4.7 is also generalized as follows:

Proposition 4.9. *Let C_0 be the tight coefficient matrix of A with respect to an optimal solution of $\mathrm{D}(A)$. Then A is upper-tight if and only if C_0 is nonsingular.*

Proof. Since $\zeta(C) = \zeta(A) - \hat{\zeta}(A)$, the matrix A is upper-tight if and only if $\zeta(C) = 0$. This is equivalent to the nonsingularity of C_0 by Proposition 2.18. □

By Proposition 4.9, we can check the upper-tightness of A just by checking the nonsingularity of C_0.

Modification in Phase 3a is almost the same as the classical algorithm described in Section 4.2.1. Suppose that A is not upper-tight. Since the tight coefficient matrix C_0

with respect to an integral dual optimal solution (p, q) is singular by Proposition 4.9, there exists $U \in \mathrm{GL}_n(K)$ such that

$$\text{t-rank}\, UC_0 = \operatorname{rank} UC_0 = \operatorname{rank} C_0 < n. \tag{4.5}$$

This U can be obtained by the Gaussian elimination applied to C_0. We modify A into $\bar{A} \coloneqq U'A$, where $U' \coloneqq D(\pi^p)UD(\pi^{-p})$.

Lemma 4.10. *It holds* $\zeta(A) = \zeta(\bar{A})$ *and* $\hat{\zeta}(A) < \hat{\zeta}(\bar{A})$.

Proof. We have

$$\zeta(U') = \zeta(D(\pi^p)) + \zeta(U) + \zeta(D(\pi^{-p})) = \zeta(U) = 0$$

and hence $\zeta(A) = \zeta(\bar{A})$.

To prove $\hat{\zeta}(A) < \hat{\zeta}(\bar{A})$, it suffices to show that (p, q) is feasible but not optimal to $\mathrm{D}(\bar{A})$. We first show the feasibility. Using C defined by (4.4), we can rewrite $\bar{A}$ as

$$\bar{A} = U'A = D(\pi^p)UD(\pi^{-p})D(\pi^p)CD(\pi^q) = D(\pi^p)UCD(\pi^q). \tag{4.6}$$

Since $U, C \in R^{n\times n}$, the matrix UC is also over R. Thus, the valuation of the (i, j)th entry of $\bar{A}$ is at least $p_i + q_j$. Hence (p, q) is feasible to $\mathrm{D}(\bar{A})$.

We next show the non-optimality of (p, q). By (4.6), the tight coefficient matrix T of $\bar{A}$ with respect to (p, q) is the coefficient matrix of π^0 in the π-adic expansion of UC. We thus have $T = UC_0$ and hence $\text{t-rank}\, T = \text{t-rank}\, UC_0 < n$. By Proposition 4.8, (p, q) is not optimal to $\mathrm{D}(\bar{A})$. □

4.2.3 Improved Algorithm

To compute $\bar{A}$ in Phase 3a, we need to multiply $D(\pi^{-p})$, U, and $D(\pi^p)$ in this order from left to A. This operation includes the computation of the coefficients in the π-adic expansion of $\pi^{-1}a$ for $a \in R$. This, however, is impossible for the computational model assumed in Section 4.1 because the oracle of computing the inverse of δ_0 is needed.

To avoid left-multiplying π^{-1}, we slightly improve the above faithful procedure of combinatorial relaxation. The improved algorithm does not modify the input matrix directly. Instead, the algorithm keeps track of $\gamma \coloneqq \hat{\zeta}(A)$ and the matrix $C \in R^{n\times n}$ defined by (4.4). The improved algorithm is outlined as follows.

Improved Combinatorial Relaxation Algorithm over DVSFs

Phase 0b. Set $\gamma \leftarrow 0$ and $C \leftarrow A$.

Phase 1b. Compute an integral optimal solution $(\Delta p, \Delta q)$ of $\mathrm{D}(C)$ such that Δp is nonpositive. Set $\gamma \leftarrow \gamma + \hat{\zeta}(C)$. If $\gamma > M$, report $\zeta(A) = +\infty$ and halt.

Set $C \leftarrow D(\pi^{-\Delta p})CD(\pi^{-\Delta q})$.

Phase 2b. If C_0 is nonsingular, report $\zeta(A) = \gamma$ and halt.

Phase 3b. Take $U \in \mathrm{GL}_n(K)$ satisfying (4.5) and modify C into UC. Go back to Phase 1b.

The validity of the improved algorithm is guaranteed by the following lemma. We denote by $\Pi(p, q)$ the objective function of the dual of the bipartite matching problem, i.e.,

$$\Pi(p, q) \coloneqq \sum_{i \in \mathrm{Row}(A)} p_i + \sum_{j \in \mathrm{Col}(A)} q_j.$$

Lemma 4.11. *Let A be the matrix at the beginning of the kth iteration of the faithful combinatorial relaxation algorithm. Let γ and C be the value and the matrix in the improved algorithm when* Phase 1b *at the kth iteration has just finished. Then $\gamma = \Pi(p, q)$ and $C = D(\pi^{-p})AD(\pi^{-q})$ hold for some optimal solution (p, q) of* $\mathrm{D}(A)$.

Proof. We show the claim by induction on k. The claim is clear when $k = 1$. Suppose that the claim holds the case when $k = m$. Let A be the matrices at Phase 3a in the mth iteration of the faithful algorithm. Similarly, let γ and C be the values in the improved algorithm when the mth Phase 1b has just finished. By the inductive assumption, $\gamma = \Pi(p, q)$ and $C = D(\pi^{-p})AD(\pi^{-q})$ hold for some optimal solution (p, q) of $\mathrm{D}(A)$.

Denote by $\bar{A}$, $\bar{\gamma}$, and $\bar{C}$ the values of A, γ, and C in the next iteration, i.e., $k = m + 1$. It holds $\bar{A} = D(\pi^{p})UD(\pi^{-p})A$, where $U \in \mathrm{GL}_n(K)$ is a matrix satisfying (4.5). Let $(\Delta p, \Delta q)$ be an optimal solution of $\mathrm{D}(UC)$ and put $\bar{p} \coloneqq p + \Delta p$ and $\bar{q} \coloneqq q + \Delta q$. Then

$$UC = UD(\pi^{-p})AD(\pi^{-q}) = D(\pi^{-p})\bar{A}D(\pi^{-q}),$$

which means that $G(UC) = G(\bar{A})$ and edge weights $w_{UC}(e)$ and $w_{\bar{A}}(e)$ for $e = \{i, j\} \in E(UC) = E(\bar{A})$ satisfies

$$w_{UC}(e) = w_{\bar{A}}(e) - p_i - q_j$$

for $i \in \mathrm{Row}(A)$ and $j \in \mathrm{Col}(A)$. Therefore, $(\bar{p}, \bar{q})$ is optimal to $\mathrm{D}(\bar{A})$ if and only if $(\Delta p, \Delta q)$ is optimal to $\mathrm{D}(UC)$. We have

$$\bar{\gamma} = \gamma + \hat{\zeta}(UC) = \gamma + \Pi(\Delta p, \Delta q) = \Pi(\bar{p}, \bar{q})$$

and

$$\begin{aligned}\bar{C} &= D(\pi^{-\Delta p})UCD(\pi^{-\Delta q}) \\ &= D(\pi^{-\Delta p})UD(\pi^{-p})AD(\pi^{-q})D(\pi^{-\Delta q}) \\ &= D(\pi^{-\Delta p})D(\pi^{-p})\bar{A}D(\pi^{-q})D(\pi^{-\Delta q}) \\ &= D(\pi^{-\bar{p}})\bar{A}D(\pi^{-\bar{q}}),\end{aligned}$$

as required. □

Corollary 4.12. *The improved combinatorial relaxation algorithm correctly outputs* $\zeta(A)$.

Proof. Follows from Propositions 3.4 and 4.9 and Lemmas 4.10 and 4.11, and the assumption on M. □

We require Δp in Phase 1b to be nonpositive so that we can avoid left-multiplying π^{-1} in the modification $C \leftarrow D(\pi^{-\Delta p})CD(\pi^{-\Delta q})$. Here we describe how we can obtain such an optimal solution $(\Delta p, \Delta q)$ of $\mathrm{D}(C)$. First, we initialize Δp and Δq as zero vectors, which is feasible to $\mathrm{D}(C)$ as the edge weight is nonnegative. We then iterate the following procedure. Construct the subgraph $G^{\#}$ of $G(C)$ defined by (3.2) with respect to $(\Delta p, \Delta q)$. If $G^{\#}$ has a perfect matching, then $(\Delta p, \Delta q)$ is optimal from Proposition 3.3 and we are done. Otherwise, by Theorem 3.1, there exists a vertex cover $W \subseteq \mathrm{Row}(A) \cup \mathrm{Col}(A)$ of $G^{\#}$ such that $|W| < n$. We change $(\Delta p, \Delta q)$ into $(\Delta p', \Delta q')$ by

$$\Delta p'_i := \begin{cases} \Delta p_i - 1 & (i \in \mathrm{Row}(A) \cap W), \\ \Delta p_i & (i \in \mathrm{Row}(A) \setminus W), \end{cases} \quad \Delta q'_j := \begin{cases} \Delta q_j & (j \in \mathrm{Col}(A) \cap W), \\ \Delta q_j + 1 & (j \in \mathrm{Col}(A) \setminus W). \end{cases} \tag{4.7}$$

Note that $\Delta p'_i \leq 0$ if $\Delta p_i \leq 0$ for $i \in \mathrm{Row}(A)$. The following lemma is well-known:

Lemma 4.13 ([55])**.** *Let* $(\Delta p, \Delta q)$ *be a feasible but not optimal dual solution. Then* $(\Delta p', \Delta q')$ *given by* (4.7) *is also feasible and* $\Pi(\Delta p, \Delta q) < \Pi(\Delta p', \Delta q')$.

By Lemma 4.13, the updated $(\Delta p, \Delta q)$ is an improved feasible solution of $\mathrm{D}(C)$. If $\gamma + \Pi(\Delta p, \Delta q) > M$, then report $\zeta(A) = +\infty$ and halt immediately. Otherwise, go back to the construction of $G^{\#}$ with respect to the updated $(\Delta p, \Delta q)$.

One more implementation issue is left: since the π-adic expansions of entries in $\bar{C}$ might have infinitely many terms, we cannot store all of them. We thus truncate higher-valuation terms relying on Proposition 4.5. Let

$$\tilde{C} := \sum_{d=0}^{M-\gamma} \bar{C}_d \pi^d, \tag{4.8}$$

where $\bar{C}_d \in K^{n \times n}$ is the coefficient matrix of π^d in the π-adic expansion of $\bar{C}$ for $d \in \mathbb{N}$. We update C into $\tilde{C}$ instead of $\bar{C}$ in Phase 3b.

Lemma 4.14. *The improved algorithm returns* $\zeta(A)$ *even if the above truncation procedure*

is executed.

Proof. We assume that C is truncated only once at the kth iteration; the general statement follows from this by induction. Let γ and C be the values in the improved algorithm when the kth Phase 1b has just finished. Let $\tilde{C}$ be the truncation (4.8) of C. From Corollary 4.12, if we replace C with $\tilde{C}$ at this point, the algorithm outputs $\zeta(\tilde{C})+\gamma$ if $\zeta(\tilde{C})+\gamma < M$ and $+\infty$ otherwise.

Suppose $\zeta(A) < M$. Since $\zeta(A) = \zeta(C) + \gamma$ by Lemma 4.11, it holds $\zeta(C) \leq M - \gamma$. This means $\zeta(\tilde{C}) = \zeta(C)$ by Proposition 4.5. Thus, the output of the improved algorithm with truncation coincides with $\zeta(A)$. Conversely, suppose $\zeta(A) = +\infty$. Then we have $\zeta(C) = +\infty > M - \gamma$, which implies $\zeta(\tilde{C}) > M - \gamma$ by Proposition 4.5 again. Thus, the improved algorithm with truncation outputs $+\infty$. □

The time complexity is analyzed as follows. Here, ω denotes the exponent in the time complexity to multiply two matrices over K.

Theorem 4.15. *Let $A \in F^{n\times n}$ be a square matrix over a split DVSF F in form of* (4.1) *and $M \in \mathbb{N}$ be an upper bound on $\zeta(A)$ valid when A is nonsingular. Then the improved combinatorial relaxation algorithm with truncation computes $\zeta(A)$ in $\mathrm{O}(M^3n^2 + M^2n^\omega + Mn^{2.5})$-time.*

Proof. The validity of the algorithm follows from Lemma 4.14. We analyze the running time. Let m be the number of times the algorithm applied (4.7) in total. It holds $m \leq M$ because one application increases γ at least by 1. In each application, we solve the bipartite matching problem, which can be solved in $\mathrm{O}(n^{2.5})$-time by the Hopcroft–Karp algorithm [42]. Thus the total time complexity of this part is $\mathrm{O}(mn^{2.5}) = \mathrm{O}(Mn^{2.5})$.

Each entry in C is multiplied by π from left at most m times because one application of (4.7) increases each Δp_i by at most 1. For $d \in [0, M-\gamma]$, we compute the coefficient of π^d in each entry of πC by calling the higher δ_0-derivative $(\delta_d)_{d\in\mathbb{N}}$. This can be done in $\mathrm{O}(M^2)$-time by (4.3). Since C has n^2 entries, the total running time of this process is $\mathrm{O}(mM^2n^2) = \mathrm{O}(M^3n^2)$.

Matrix computations in Phase 2b and Phase 3b can be done in $\mathrm{O}(Mn^\omega)$-time per each iteration as C contains $\mathrm{O}(M)$ terms. Summing it over $\mathrm{O}(M)$ iterations, we obtain $\mathrm{O}(M^2n^\omega)$-time in total. Thus the desired time complexity is attained. □

4.3 Matrix Expansion Algorithm

In this section, we describe another algorithm, the matrix expansion, for computing valuations of the Dieudonné determinants of matrices over DVSFs. First, Section 4.3.1 introduces *expanded matrices*, which is a key of the matrix expansion algorithm. Then Section 4.3.2 gives a formula connecting integer sequences of ranks of expanded matrices and valuations of subdeterminants, using the notion of Legendre conjugacy. Based on this formula, Section 4.3.3 describes an algorithm and time complexity analysis.

4.3.1 Expanded Matrices

Let F be a split DVSF with valuation ring R and coefficient skew subfield K such that the completion of F is isomorphic to $K((s;(\delta_d)))$. Let $A = (A_{i,j}) \in F^{n\times n}$ be a square matrix in form (4.1) and M be an upper bound on $\zeta(A)$ which is valid when A is nonsingular.

For $i, d \in \mathbb{N}$, let $A_d^{(i)} \in K^{n\times n}$ denote the coefficient matrix of π^d in the π-adic expansion of $\pi^i A$. Namely, for $i \in \mathbb{N}$, the matrix $\pi^i A$ is written as

$$\pi^i A = \sum_{d=0}^{\infty} A_d^{(i)} \pi^d.$$

Note that $A_d^{(i)} = O$ for $d < i$ as the valuations of entries in $\pi^i A$ are at least i. For $\mu \in \mathbb{N}$, we define the *μth-order expanded matrix* $\Omega_\mu(A)$ of A as the following $\mu n \times \mu n$ block matrix

$$\Omega_\mu(A) \coloneqq \begin{pmatrix} A_0^{(0)} & A_1^{(0)} & \cdots & \cdots & A_{\mu-1}^{(0)} \\ O & A_1^{(1)} & A_2^{(1)} & \cdots & A_{\mu-1}^{(1)} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & & A_{\mu-2}^{(\mu-2)} & A_{\mu-1}^{(\mu-2)} \\ O & \cdots & \cdots & O & A_{\mu-1}^{(\mu-1)} \end{pmatrix} \in K^{\mu n\times \mu n}. \tag{4.9}$$

Expanded matrices satisfy the multiplicativity as follows (see also [22, Section 1.2]). This is an extension of the result in [98] for rational function matrices over $\mathbb{C}$.

Lemma 4.16. *Let $A \in R^{n\times n}$ and $B \in R^{n\times n}$ be matrices over a split DVR R. Then it holds*

$$\Omega_\mu(AB) = \Omega_\mu(A)\Omega_\mu(B)$$

for $\mu \in \mathbb{N}$.

Proof. Fix $i \in [0, \mu-1]$ and let $\pi^i A = \sum_{d=0}^{\infty} A_d^{(i)} \pi^d$ be the π-adic expansion of $\pi^i A$, where π is a uniformizer of R. Similarly, for $d \in [0, \mu-1]$, let $\pi^d B = \sum_{j=0}^{\infty} B_j^{(d)} \pi^j$ be the π-adic expansion of $\pi^d B$. Then it holds

$$\pi^i AB = \left(\sum_{d=0}^{\infty} A_d^{(i)} \pi^d\right) B = \sum_{d=0}^{\infty} A_d^{(i)} \left(\sum_{j=0}^{\infty} B_j^{(d)} \pi^j\right) = \sum_{j=0}^{\infty} \left(\sum_{d=0}^{j} A_d^{(i)} B_j^{(d)}\right) \pi^j, \tag{4.10}$$

where the inner sum of the last term stops at $d = j$ by $B_j^{(d)} = O$ for $j < d$. The equality (4.10) implies that the coefficient matrix of π^j in the π-adic expansion of $\pi^i AB$

is

$$\sum_{d=0}^{j} A_d^{(i)} B_j^{(d)} = \sum_{d=0}^{\mu-1} A_d^{(i)} B_j^{(d)}$$

for $j < \mu$, which is equal to the $(i+1, j+1)$st entry of $\Omega_\mu(A)\Omega_\mu(B)$. □

Let $\omega_\mu(A)$ denote the rank of $\Omega_\mu(A)$. The following lemma claims that $\omega_\mu(A)$ coincides with that of the Smith–McMillan form (see Proposition 2.16) of A.

Lemma 4.17. *Let $A \in R^{n\times n}$ be a matrix over a split DVR R. Then it holds $\omega_\mu(A) = \omega_\mu(D)$ for $\mu \in \mathbb{N}$, where D is the Smith–McMillan form of A.*

Proof. Let $S \in R^{n\times n}$ and $T \in R^{n\times n}$ be biproper matrices such that $SAT = D$. From Lemma 4.16, we have

$$\omega_\mu(D) = \operatorname{rank} \Omega_\mu(SAT) = \operatorname{rank} \Omega_\mu(S)\Omega_\mu(A)\Omega_\mu(T).$$

For $i \in \mathbb{N}$, let $S_i^{(i)}$ be the coefficient matrix of π^i in the π-adic expansion of $\pi^i S$, where π is a uniformizer of R. Then $S_i^{(i)}$ is equal to the coefficient matrix of π^0 in the π-adic expansion of $\pi^{-i} S \pi^i$. Now $\pi^{-i} S \pi^i$ is biproper by $\left(\pi^{-i} S \pi^i\right)^{-1} = \pi^{-i} S^{-1} \pi^i$. Thus, $S_i^{(i)}$ is nonsingular from Proposition 2.18. Since $\Omega_\mu(S)$ is a block triangular matrix having $S_i^{(i)}$ for the $(i+1)$st diagonal block, it is nonsingular. Similarly, $\Omega_\mu(T)$ is nonsingular. Therefore, we have $\omega_\mu(D) = \omega_\mu(A)$. □

Let $0 \le \alpha_1 \le \cdots \le \alpha_r$ be the exponents of the Smith–McMillan form of $A \in R^{n\times n}$ with $r \coloneqq \operatorname{rank} A$. Put

$$N_d \coloneqq |\{i \in [r] \mid \alpha_i \le d\}| \tag{4.11}$$

for $d \in \mathbb{N}$. Lemma 4.17 leads us to the following lemma; a similar result based on the Kronecker canonical form is also known for matrix pencils over a field [45, Theorem 2.3].

Lemma 4.18. *Let $A \in R^{n\times n}$ be a matrix over a split DVR R. For $\mu \in \mathbb{N}$, it holds*

$$\omega_\mu(A) = \sum_{d=0}^{\mu-1} N_d, \tag{4.12}$$

where N_d is defined by (4.11).

Proof. Let D be the Smith–McMillan form of A and $D_d^{(i)} \in R^{n\times n}$ the coefficient matrix of π^d in the π-adic expansion of $\pi^i D$ for $i, d \in \mathbb{N}$. Since entries of D are powers of π, the matrix D commutes with π. This implies $D_d^{(i)} = D_{d-i}^{(0)} =: D_{d-i}$ for $d \ge i$. Now $\Omega_\mu(D)$ is

in the form

$$\Omega_\mu(D) = \begin{pmatrix} D_0 & D_1 & \cdots & D_{\mu-2} & D_{\mu-1} \\ O & \ddots & \ddots & \ddots & D_{\mu-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & D_1 \\ O & \cdots & \cdots & O & D_0 \end{pmatrix}. \tag{4.13}$$

Let $\alpha_1, \dots, \alpha_r$ be the exponents of the Smith–McMillan form D, where $r \coloneqq \operatorname{rank} A$. The ith diagonal entry of D_d is 1 if $i \le r$ and $\alpha_i = d$, and 0 otherwise. Thus from (4.13), each row and column in $\Omega_\mu(D)$ has at most one nonzero entry. Hence $\omega_\mu(D)$, which is equal to $\omega_\mu(A)$ by Lemma 4.17, is equal to the number of nonzero entries in $\Omega_\mu(D)$. It is easily checked that the $(\mu - d)$th block row of $\Omega_\mu(D)$ contains N_d nonzero entries for $d \in [0, \mu - 1]$. □

The equality (4.12) is a key identity that connects $\omega_\mu(A)$ and the Smith–McMillan form of A. We remark that (4.12) can be rewritten as

$$N_d = \omega_{d+1}(A) - \omega_d(A) \tag{4.14}$$

for $d \in \mathbb{N}$.

4.3.2 Legendre Conjugacy of $\zeta_k(A)$ and $\omega_\mu(A)$

Let $A \in R^{n\times n}$ be a matrix of rank r and put $\zeta_k \coloneqq \zeta_k(A)$ for $k = [0, r]$, where $\zeta_k(A)$ is defined by (2.10). As observed in Example 3.11, the integer sequence $(\zeta_0, \dots, \zeta_r)$ is convex in the sense described in Section 3.3.2. In addition, for $\mu \in \mathbb{N}$ put $\omega_\mu \coloneqq \omega_\mu(A)$ and define N_μ by (4.11). From $N_{\mu-1} \le N_\mu$ and (4.14), we have $\omega_{\mu-1} + \omega_{\mu+1} \ge 2\omega_\mu$ for all $\mu \ge 1$. This inequality also indicates the convexity of ω_μ

Indeed, the sequences of ζ_k and $-\omega_\mu$ are in the relation of Legendre conjugate. This can be shown from the key identities (2.18) and (4.12) that connect $\zeta_k(A)$ and $\omega_\mu(A)$ through the Smith–McMillan form of A.

Theorem 4.19. *Let $A \in R^{n\times n}$ be a matrix of rank r over a split DVR R. Then it holds*

$$\zeta_k(A) = \max_{\mu \ge 0}(k\mu - \omega_\mu(A)) \quad (0 \le k \le r), \tag{4.15}$$

$$\omega_\mu(A) = \max_{0 \le k \le r}(k\mu - \zeta_k(A)) \quad (\mu \ge 0). \tag{4.16}$$

Proof. Put $\zeta_k \coloneqq \zeta_k(A)$ for $k \in [0, r]$ and $\omega_\mu \coloneqq \omega_\mu(A)$ for $\mu \in \mathbb{N}$. Since $(\zeta_0, \zeta_1, \dots, \zeta_r)$ is convex and $(-\omega_0, -\omega_1, -\omega_2, \dots)$ is concave, (4.15) and (4.16) are equivalent by (3.5). We show (4.16) as follows.

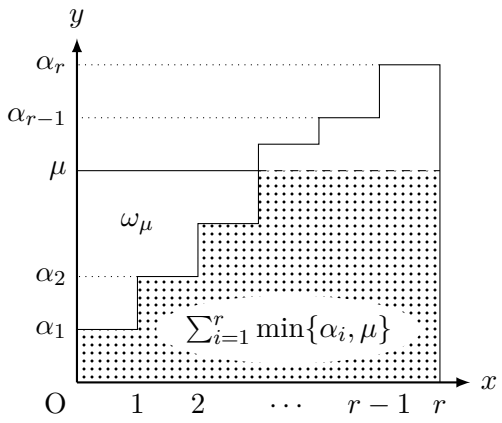


Figure 4.1: Graphic explanation of (4.17).

First we give an equality

$$\omega_\mu = r\mu - \sum_{i=1}^{r} \min\{\alpha_i, \mu\} \tag{4.17}$$

for $\mu \in \mathbb{N}$, where $\alpha_1 \leq \cdots \leq \alpha_r$ are the exponents of the Smith–McMillan form of A. Figure 4.1 graphically shows this equality. Let x and y be the coordinates along the horizontal and vertical axes in Figure 4.1, respectively. For $i \in [r]$, the height of the dotted rectangle with $i-1 \leq x < i$ is $\min\{\alpha_i, \mu\}$. Hence the area of the dotted region is equal to $\sum_{i=1}^r \min\{\alpha_i, \mu\}$. In addition, the width of the white rectangle with $d \leq y < d+1$ is equal to N_d for $d = 0, \ldots, \mu-1$, where N_d is defined by (4.11). Hence the area of the white stepped region is equal to $N_0 + \cdots + N_{\mu-1} = \omega_\mu$ by (4.12). Now we have (4.17) since the sum of the areas of these two regions is $r\mu$.

Substituting (2.18) into the right hand side of (4.16), we have

$$\max_{0\leq k\leq r} (k\mu - \zeta_k) = \max_{0\leq k\leq r} \sum_{i=1}^{k} (\mu - \alpha_i) = k^*\mu - \sum_{i=1}^{k^*} \alpha_i, \tag{4.18}$$

where k^* is the maximum $0 \leq k \leq r$ such that $\alpha_k \leq \mu$. Since $\min\{\alpha_i, \mu\}$ is α_i if $i \leq k^*$ and μ if $i > k^*$, it holds

$$\sum_{i=1}^{r} \min\{\alpha_i, \mu\} = (r - k^*)\mu + \sum_{i=1}^{k^*} \alpha_i. \tag{4.19}$$

From (4.18) and (4.19), we have

$$\max_{0\leq k\leq r} (k\mu - \zeta_k) = r\mu - \sum_{i=1}^{r} \min\{\alpha_i, \mu\},$$

in which the right hand side is equal to ω_μ by (4.17). □

4.3.3 Reductions and Algorithms

We finally apply Theorem 4.19 to the computation of $\zeta(A)$ via the following lemma.

Lemma 4.20. *Let $A \in F^{n\times n}$ be a matrix* (4.1) *of rank r over a split DVSF F such that $\zeta(A) \leq M$ or $\zeta(A) = +\infty$. Then A is nonsingular if and only if $\omega_{M+1}(A) - \omega_M(A) = n$. Furthermore, if A is nonsingular, then it holds*

$$\zeta(A) = Mn - \omega_M(A). \tag{4.20}$$

Proof. It holds $\omega_{M+1}(A) - \omega_M(A) = N_M \leq n$ by (4.14). If A is singular, then N_M must be less than n. If A is nonsingular, then α_i is at most M for all $i \in [r]$, which means $N_M = n$.

Suppose that A is nonsingular. From (4.12) and (4.15), it holds

$$\zeta(A) = \max_{\mu\geq 0} \sum_{d=0}^{\mu-1} (n - N_d). \tag{4.21}$$

Since $N_0 \leq N_1 \leq \cdots \leq N_M = N_{M+1} = \cdots = n$, the maximum value of the right hand side of (4.21) is attained by $\mu = M$. Thus we have (4.20). □

From Lemma 4.20, we can compute $\zeta(A)$ just by calculating $\omega_M(A)$ and $\omega_{M+1}(A)$; we call this the *matrix expansion algorithm*. These matrices can be constructed in $\mathrm{O}(M^3n^2)$-time by repeatedly applying (4.3) and the rank computation can be done in $\mathrm{O}(M^\omega n^\omega)$ arithmetic operations on K. Thus we have:

Theorem 4.21. *Let $A \in R^{n\times n}$ be a square matrix over a split DVSF F in form of* (4.1) *and $M \in \mathbb{N}$ be an upper bound on $\zeta(A)$ valid when A is nonsingular. Then the matrix expansion algorithm computes $\zeta(A)$ in $\mathrm{O}(M^3n^2 + M^\omega n^\omega)$-time.*

4.4 Estimating Upper Bounds

4.4.1 Bounds for Skew Polynomial Rings

Let R be a split DVR with coefficient skew subfield K. In the algorithms presented in Sections 4.2 and 4.3, we assume that an upper bound M of $\zeta(A)$ is known beforehand (or $\zeta(A) = +\infty$) for $A \in R^{n\times n}$. How can we know such M? Recall that entries in the input matrix $A \in R^{n\times n}$ in (4.1) contain terms having valuations at most ℓ. One optimistic estimation of the upper bound is ℓn. From the definition of the determinant (2.9), this is valid when R is commutative, or equivalently, R is isomorphic to a subring of $K[[s]]$. This can be extended to the case of skew polynomial rings (see Example 2.8) as follows.

Let K be a skew field equipped with an automorphism σ and a left σ-derivation

δ. As stated in Example 2.8, the skew inverse Laurent series field $K\big((s^{-1};\sigma,\delta)\big)$ forms a complete split DVR with valuation $-\deg$ and uniformizer s^{-1}. We denote by $K\big[\big[s^{-1};\sigma,\delta\big]\big]$ the valuation ring of $K\big((s^{-1};\sigma,\delta)\big)$. From Example 4.3, $K\big[\big[s^{-1};\sigma,\delta\big]\big]$ is isomorphic to $K\big[\big[t;(\delta_d)\big]\big]$ by an isomorphism $s^{-1}\mapsto t$, where δ_d is given by (2.8) for $d\in\mathbb{N}$.

Proposition 4.22. *Let $F \coloneqq K\big((s^{-1};\sigma,\delta)\big)$ be a skew inverse Laurent field over a skew field K. For a nonsingular matrix $A=\sum_{d=0}^{\ell} A_d s^{-d}\in F^{n\times n}$ with $A_0,\dots,A_\ell\in K^{n\times n}$, it holds $\zeta(A)=-\deg\operatorname{Det} A\le \ell n$.*

Proof. Consider

$$B \coloneqq As^{\ell}=\sum_{d=0}^{\ell} A_{\ell-d}s^{d}\in K[s;\sigma,\delta]^{n\times n}.$$

Since $\zeta(B)=\zeta(A)+\zeta(I_n s^{\ell})=\zeta(A)+n\ell$, it suffices to show $-\zeta(B)=\deg\operatorname{Det} B\ge 0$.

The skew polynomial ring $K[s;\sigma,\delta]$ is known to be a (left and right) PID [33, Theorem 2.8] as the usual polynomial ring $K[s]$. Let $D=UBV$ be the Jacobson normal form of B (see Proposition 2.19). Here, $U,V\in \mathrm{GL}_n(K[s;\sigma,\delta])\subseteq \mathrm{GL}_n(K\big[\big[s^{-1};\sigma,\delta\big]\big])$ are biproper matrices. By Proposition 2.18, we have $\zeta(D)=\zeta(U)+\zeta(B)+\zeta(V)=\zeta(B)$. Since diagonal entries in D are nonzero skew polynomials, they have nonnegative degrees. Thus we have $\zeta(B)=\zeta(D)\ge 0$. □

A *skew polynomial matrix* over K refers to a matrix over a skew polynomial ring over K. As we have shown in the proof of Proposition 4.22, for a skew polynomial matrix $A=\sum_{d=0}^{\ell} A_{\ell-d}s^{\ell}\in K[s;\sigma,\delta]^{n\times n}$, we can reduce the computation of $\deg\operatorname{Det} A$ into that of $-\det\operatorname{Det} As^{-\ell}$, where

$$As^{-\ell}=\sum_{d=0}^{\ell} A_d s^{-d}\in K\big((s^{-1};\sigma,\delta)\big)^{n\times n}.$$

From Proposition 4.22, we can set $M\coloneqq \ell n$ for $As^{-\ell}$. In addition, the coefficients of $s^{-1}a$ satisfy the following recursion formula.

Lemma 4.23. *Let $a=\sum_{d=0}^{\infty} a_d s^{-d}\in K\big[\big[s^{-1};\sigma,\delta\big]\big]$ with $a_d\in K$ for $d\in\mathbb{N}$. The coefficient b_d of s^{-d} in $s^{-1}a$ satisfies*

$$b_d=\begin{cases}\sigma^{-1}(a_{d-1}-\delta(b_{d-1})) & (d\ge 1),\\ 0 & (d=0).\end{cases} \tag{4.22}$$

Proof. By (1.3), we have

$$a = s(s^{-1}a) = s\sum_{d=0}^{\infty} b_d s^{-d} = \sum_{d=0}^{\infty}(\sigma(b_d)s + \delta(b_d))s^{-d} \tag{4.23}$$
$$= \sigma(b_0)s + \sum_{d=0}^{\infty}(\sigma(b_{d+1}) + \delta(b_d))s^{-d}.$$

By (4.23), it holds $\sigma(b_0) = 0$ and $\sigma(b_{d+1}) + \delta(b_d) = a_d$ for $d \in \mathbb{N}$, which imply (4.22). □

From (4.22), we can compute the leading M coefficients of $s^{-1}a$ by $\mathrm{O}(M)$ applications of σ^{-1} and δ. This is improved from $\mathrm{O}(M^2)$ based on (4.3). Applying this improvement and plugging ℓn into M in the time complexities in Theorem 4.15 and Theorem 4.21, we obtain the following.

Theorem 4.24. *Let $B = \sum_{d=0}^{\ell} A_{\ell-d}s^d \in K[s;\sigma,\delta]^{n\times n}$ be a square skew polynomial matrix over a skew field K. Suppose that arithmetic operations on K and applications of σ^{-1}, δ can be executed in constant time. Then we can compute* $\deg \mathrm{Det}\, B$ *in $\mathrm{O}(\ell^2 n^{\omega+2} + \ell n^{4.5})$-time by the combinatorial relaxation algorithm and in $\mathrm{O}(\ell^{\omega} n^{2\omega})$-time by the matrix expansion algorithm.*

Similarly, we can compute $\operatorname{ord}\mathrm{Det}$ of matrices over $K[s;\sigma]$ in the same time complexities as Theorem 4.24. See Section 5.2 for an application of these computations to differential equations.

Comparison to Existing Algorithms. In computer algebra, algorithms were proposed for computing various kinds of canonical forms of a skew polynomial matrix $A \in K[s;\sigma,\delta]^{n\times n}$ such as the Jacobson normal form [58] (see Section 2.2.5), the *Hermite normal form* [32], the *Popov normal form* [51] and their weaker form called a *row-reduced form* [1, 3]. One can use these algorithms to calculate $\deg \mathrm{Det}\, A$ since it is immediately obtained from the canonical forms of A. These algorithms iteratively solve systems of linear equations over K whose coefficient matrices are variants of expanded matrices $\Omega_{\mu}(A)$ under the name of "linearized matrices" [51] or "striped Krylov matrices" [3].

Our algorithms are faster than the existing algorithms. The fastest known algorithm given by Giesbrecht–Kim [32] runs in $\mathrm{O}(\ell^{\omega} n^{2\omega+2}\ell n)$ time, whereas our two algorithms require only $\mathrm{O}(\ell^2 n^{\omega+2} + \ell n^{4.5})$-time and $\mathrm{O}(\ell^{\omega} n^{2\omega})$-time as shown in Theorem 4.24.

4.4.2 Characterizing Split DVSFs with Bounds

In Section 4.4.1, we described that the valuation of the Dieudonné determinant of nonsingular $A = \sum_{d=0}^{\ell} A_d\pi^d \in F^{n\times n}$ is bounded by ℓn when F is a skew inverse Laurent series field. Indeed, the converse also holds in the following sense.

Theorem 4.25. *Let F be a complete split DVSF with coefficient skew subfield K and uniformizer π. Then every $A = \sum_{d=0}^{\ell} A_d\pi^{\ell} \in \mathrm{GL}_n(F)$ with $A_0, \ldots, A_d \in K^{n\times n}$ satisfies*

$\zeta(A) \leq \ell n$ if and only if F is isomorphic to $K((s^{-1};\sigma,\delta))$ with some automorphism σ and left σ-derivation δ.

Proof. The "if" part was shown in Proposition 4.22. We show the "only if" part. Let $(\delta_d)_{d\in\mathbb{N}}$ be the higher δ_0-derivatives corresponding to F. If F is isomorphic to $K((s^{-1};\sigma,\delta))$, then $\sigma = {\delta_0}^{-1}$ and $\delta = -{\delta_0}^{-1}\delta_1{\delta_0}^{-1}$ by (2.8). Hence we put $\sigma := {\delta_0}^{-1}$ and $\delta := -{\delta_0}^{-1}\delta_1{\delta_0}^{-1}$. This σ is an automorphism and δ is a left σ-derivation.

For $a \in K$, we put $\pi^{-1}a\pi =: a' = \sum_{d=0}^{\infty} a'_d\pi^d$ with $a'_0, a'_1, \ldots \in K$. We first show that if $a'_d = 0$ for any $a \in K$ and $d \geq 2$, then F is isomorphic to $K((s^{-1};\sigma,\delta))$. Suppose that F satisfies this assumption and put $s := \pi^{-1}$. Then it holds

$$sa = \pi^{-1}a = a'\pi^{-1} = a'_0\pi^{-1} + a'_1 = a'_0 s + a'_1 \tag{4.24}$$

for $a \in K$. From $a = \pi a'\pi^{-1}$ and (4.3) for $d = 0, 1$, we have $a = \delta_0(a'_0)$ and $0 = \delta_0(a'_1) + \delta_1(a'_0)$. Solving these qualities for a'_0 and a'_1, we obtain

$$a'_0 = {\delta_0}^{-1}(a) = \sigma(a), \tag{4.25}$$

$$a'_1 = {\delta_0}^{-1}(-\delta_1(a'_0)) = -({\delta_0}^{-1}\delta_1{\delta_0}^{-1})(a) = \delta(a). \tag{4.26}$$

Substituting (4.25) and (4.26) into (4.24), we have

$$sa = \sigma(a)s + \delta(a),$$

which is nothing but the commutation rule (1.3) of the skew polynomial ring $K[s;\sigma,\delta]$. Hence the ring generated by π^{-1} over K, its Ore quotient skew field, and its completion F with respect to the π-adic topology are isomorphic to $K[s;\sigma,\delta]$, $K(s;\sigma,\delta)$, and $K((s^{-1};\sigma,\delta))$, respectively.

Next, suppose that F is not isomorphic to $K((s^{-1};\sigma,\delta))$. From the contraposition of the above proof, there exists $a \in K$ such that $a'_d \neq 0$ for some $d \geq 2$; take such a and let $k \geq 2$ be the minimum number with $a'_k \neq 0$. Consider

$$A := \begin{pmatrix} 0 & 0 \\ 1 & a'_0 \end{pmatrix} + \begin{pmatrix} 1 & a \\ 0 & a'_1 \end{pmatrix}\pi = \begin{pmatrix} \pi & a\pi \\ 1 & a'_0 + a'_1\pi \end{pmatrix} \in F^{2\times 2}.$$

The values of ℓ and n for A are $\ell = 1$ and $n = 2$. Multiplying an elementary matrix, we can transform A into

$$B := \begin{pmatrix} 1 & 0 \\ -\pi^{-1} & 1 \end{pmatrix} A = \begin{pmatrix} \pi & a\pi \\ 0 & a'_0 + a'_1\pi - \pi^{-1}a\pi \end{pmatrix} = \begin{pmatrix} \pi & a\pi \\ 0 & -\sum_{d=k}^{\infty} a'_d\pi^d \end{pmatrix}.$$

Thus, A is nonsingular and it holds

$$\zeta(A) = \zeta(B) = v(\pi) + v\left(\sum_{d=k}^{\infty} a'_d \pi^d\right) = 1 + k > 2 = \ell n,$$

where v is the valuation on F. □

Theorem 4.25 means that the condition "$\zeta(A) \leq \ell n$ for any $A = \sum_{d=0}^{\ell} A_d \pi^d \in \mathrm{GL}_n(F)$" serves as a characterization of skew inverse Laurent series fields. In this way, skew polynomials arise not only from an algebraic abstraction of linear differential/difference equations but also from the most natural condition for which the combinatorial relaxation and the matrix expansion algorithms are applicable.

Chapter 5

Applications of Valuations of the Dieudonné Determinants

This chapter describes two applications of the computation of valuations of the Dieudonné determinants. First, Section 5.1 considers weighted Edmonds' problem (WEP). We show that both the combinatorial relaxation and matrix expansion algorithms can be applied for reducing the noncommutative WEP (nc-WEP) to the unweighted problem. In particular, the matrix expansion algorithm is also applicable to the commutative problem, and further yields a polynomial-time algorithm for the nc-WEP with bounded bit complexity. We also discuss the WEP for sparse matrices.

Second, Section 5.2 deals with linear differential and difference equations from algebraic viewpoint. These equations can be integrally handled as *σ-differential equations* by making use of the skew polynomials. We show that the dimension of the solution spaces of simultaneous σ-differential equations can be characterized by the degree (and the order) of the Dieudonné determinant.

5.1 Weighted Edmonds' Problem

5.1.1 Problem Definition

We briefly repeat definitions needed to explain Edmonds' problem. See Section 1.1.3 for more backgrounds.

Let K be a field. A *linear matrix* B over K is a matrix in the form

$$B = B_0 + \sum_{k=1}^{m} B_k x_k, \tag{5.1}$$

where $B_0, \ldots, B_m \in K^{n \times n'}$ and $x_1, \ldots, x_m$ are symbols which are commutative with any element in K. The linear matrix B is called *commutative* if $x_1, \ldots, x_m$ are pairwise commutative and *noncommutative* (nc) if they are pairwise noncommutative. Commutative

linear matrices are regarded as matrices over the rational function field $K(x_1, \dots, x_m)$ and nc-linear matrices are over the free skew field $K\langle\langle x_1, \dots, x_m\rangle\rangle$. Commutative and noncommutative *Edmonds' problem* [23] over K are to compute the rank of given commutative and noncommutative linear matrices over K, respectively.

A *linear polynomial matrix* A over K is a matrix

$$A = \sum_{d=0}^{\ell} A_{\ell-d} s^d, \tag{5.2}$$

where $A_d = A_{d,0} + \sum_{k=1}^{m} A_{d,k} x_k$ is a linear matrix over K for $d \in [0, \ell]$ and s is a symbol that commutes with any element in $K \cup \{x_0, \dots, x_m\}$. The matrix A is also called *commutative* or *noncommutative* according to the commutativity of $x_1, \dots, x_m$. Commutative linear polynomial matrices are over $K(x_1, \dots, x_m)(s)$ and nc-linear polynomial matrices are over $K\langle\langle x_1, \dots, x_m\rangle\rangle(s)$. The minus of the degree with respect to s serves as valuations on $K(x_1, \dots, x_m)(s)$ and $K\langle\langle x_1, \dots, x_m\rangle\rangle(s)$ (see Example 2.5). Commutative and noncommutative *weighted Edmonds' problem* (WEP) [39] over K are to compute the degree of the Dieudonné determinant of a given commutative and noncommutative linear polynomial matrices over K, respectively.

5.1.2 Solving Weighted Edmonds' Problem

Our algorithms can be applied for reducing the nc-WEP into nc-Edmonds' problem over a field K, where we assume the arithmetic model on K. Put $L \coloneqq K\langle\langle x_1, \dots, x_m\rangle\rangle$ and let R be the valuation ring of $L(s)$ with respect to the valuation $-\deg$. Instead of an $n \times n$ nc-linear polynomial matrix $A \in L(s)$ given in (5.1), we consider

$$As^{-\ell} = \sum_{d=0}^{\ell} A_d s^{-d}.$$

Then we can compute $\zeta(A) = -\deg \operatorname{Det} A$ from $\zeta(As^{-\ell})$ by $\zeta(A) = \zeta(As^{-\ell}) - \ell n$. Since $L(s)$ is a special case of skew rational function fields over L, i.e., $L(s) = L(s; \mathrm{id}, 0)$, it holds $\zeta(As^{-\ell}) \le \ell n$ when A is nonsingular by Proposition 4.22.

First, consider the combinatorial relaxation algorithm presented in Section 4.2. Since one cannot perform arithmetic operations on L efficiently, it is not immediate to apply the combinatorial relaxation algorithm to $As^{-\ell}$. In particular, the procedure of finding the matrix $U \in \mathrm{GL}_n(L)$ in Phase 3b based on the Gaussian elimination on L requires exponential number of arithmetic operations on K. Nevertheless, we can make use of the following property on nc-linear matrices given by Fortin–Reutenauer [27].

Theorem 5.1 ([27, Theorem 1])**.** *For an nc-linear matrix $B \in K\langle\langle x_1, \dots, x_m\rangle\rangle^{n \times n'}$ over a field K, there exist $U \in \mathrm{GL}_n(K)$ and $V \in \mathrm{GL}_{n'}(K)$ such that* t-rank $UBV = \operatorname{rank} B$.

The problem of finding U and V satisfying t-rank $UBV = \operatorname{rank} B$, which is a variant of

nc-Edmonds' problem by Theorem 5.1, is called the *maximum vanishing subspace problem* (MVSP) due to Hamada–Hirai [37]. The MVSP can be solved in deterministic polynomial-time [37, 43]. Therefore, by using the algorithms in [37, 43] as oracles, we obtain a deterministic polynomial-time algorithm for the nc-WEP. This algorithm indeed coincides with the steepest gradient descent algorithm given by Hirai [39].

Theorem 5.2 ([39, Theorem 4.4])**.** *The nc-WEP for* (5.2) *over a field K can be solved in deterministic $\mathrm{O}(\ell^2 mn^{\omega+2} + T_{\mathrm{MVSP}}(n,m)\ell n)$-time, where $T_{\mathrm{MVSP}}(n,m)$ denotes the time needed to solve the MVSP for an $n \times n$ nc-linear matrix with m symbols over K.*

Proof. In Phase 3b of each iteration, we solve the MVSP to obtain $U, V \in \mathrm{GL}_n(K)$ and update C into UCV. This matrix multiplication can be done in $\mathrm{O}(\ell mn^{\omega+1})$ arithmetic operations on K because C is expanded as $\sum_{d=0}^{\ell n} C_d s^{-d}$ and each C_d is an nc-linear matrix of m symbols. Since the number of iterations is $\mathrm{O}(\ell n)$, we obtain the desired time complexity. □

We remark that the time complexity in Theorem 5.2 is in terms of the arithmetic model on K. In case of $\mathbb{K} = \mathbb{Q}$, the bit-lengths of intermediate numbers are not bounded, even if an algorithm for MVSP guarantees the bounded bit-length. In addition, since Theorem 5.2 relies on Theorem 5.1, we cannot apply the combinatorial relaxation for the commutative problem.

We next apply the matrix expansion algorithm in Section 4.3 to the WEP. This application is rather immediate than that of the combinatorial relaxation algorithm. Namely, if A is a commutative (noncommutative) linear polynomial matrix (5.2) over a field K, then the expanded matrix $\Omega_\mu(As^{-\ell})$ given by (4.9) is a commutative (resp. noncommutative) linear matrix. Hence the rank computation of $\Omega_\mu(As^{-\ell})$ is nothing but solving the commutative (resp. noncommutative) Edmonds' problem. By Lemma 4.20 and Proposition 4.22, we obtain the following:

Theorem 5.3. *The commutative (noncommutative) WEP for* (5.2) *over a field K can be solved in deterministic $\mathrm{O}(T_{\mathrm{EP}}(\ell n^2, m))$-time, where $T_{\mathrm{EP}}(n,m)$ denotes the time needed to solve commutative (resp. noncommutative) Edmonds' problem for an $n \times n$ commutative (resp. noncommutative) linear matrix with m symbols over K.*

The algorithms of Gurvits [35] and Ivanyos et al. [43] deterministically solve nc-Edmonds' problem with polynomially bounded bit complexity when $K = \mathbb{Q}$. Using these algorithm as oracles, we obtain:

Theorem 5.4. *The nc-WEP over a field K can be deterministically solved using polynomially many arithmetic operations on K. When $K = \mathbb{Q}$, the algorithm runs in time polynomial in the binary encoding length of the input.*

5.1.3 Weighted Edmonds' Problem for Sparse Matrices

In view of combinatorial optimization, the algorithm given in Theorem 5.4 is regarded as pseudo-polynomial time algorithms since the running time depends on a polynomial of the maximum exponent ℓ of s instead of $\mathrm{poly}(\log \ell)$. Recently, Hirai–Ikeda [40] presented algorithms to solve the nc-WEP over K for an nc-linear polynomial matrix in form of

$$A = \sum_{k=0}^{m} A_k x_k s^{w_k}, \tag{5.3}$$

where $A_1, \ldots, A_m \in K^{n \times n}$ and $w_1, \ldots, w_m \in \mathbb{Z}$. The nc-WEP for (5.3) includes the weighted linear matroid intersection problem. An algorithm of Hirai–Ikeda runs in strongly polynomial time, i.e., it runs in time polynomial of n and m.

As an extension of a different direction, it is natural to try to solve the (commutative) WEP for

$$A = \sum_{k=0}^{m} A_k s^{w_k}, \tag{5.4}$$

where $A_1, \ldots, A_m \in K^{n \times n}$ and $w_1, \ldots, w_m \in \mathbb{Z}$. However, setting $w_k \coloneqq (n+1)^k$ for $k \in [m]$ would make the rank of (5.4) the same as that of a linear matrix $\sum_{k=0}^{m} A_k x_k \in K[x_1, \ldots, x_m]^{n \times n}$ (the *Kronecker substitution*). Since giving a deterministic polynomial-time algorithm for Edmonds' problem has been open for more than half a century, computing $\deg \det$ of (5.4) is also quite challenging.

5.2 Linear Differential and Difference Equations

5.2.1 σ-Differential Equations

Let R be a commutative ring endowed with a ring automorphism $\sigma : R \to R$ and a left σ-derivation $\delta : R \to R$ (see Example 2.8). A *σ-differential ring* is the triple (R, σ, δ), or R itself when σ and δ are clear. A *σ-differential field* is a σ-differential ring which is a field. If $\sigma = \mathrm{id}$, then σ-differential rings and fields are simply called *differential* rings and fields. Similarly, σ-differential rings and fields with $\delta = 0$ are called *difference* rings and fields.

A *constant* of a σ-differential ring (R, σ, δ) is an element $a \in R$ such that $\sigma(a) = a$ and $\delta(a) = 0$. The set of all constants of (R, σ, δ) is denoted by $\mathrm{Const}_{\sigma,\delta}(R)$ or by $\mathrm{Const}(R)$. It is easily checked that $\mathrm{Const}(R)$ is a subring of R, and if R is a field, so is $\mathrm{Const}(R)$.

An additive map $\theta : R \to R$ is said to be *pseudo-linear* if it satisfies

$$\theta(ab) = \sigma(a)\theta(b) + \delta(a)b \tag{5.5}$$

for all $a, b \in R$. Recall from Example 2.8 that $R[s;\sigma,\delta]$ denotes the skew polynomial ring over (R,σ,δ). Then θ induces a left $R[s;\sigma,\delta]$-module structure on R, where the action $\bullet : R[s;\sigma,\delta] \times R \to R$ is defined by

$$\left(\sum_{d=0}^{\ell} a_d s^d\right) \bullet b \coloneqq \sum_{d=0}^{\ell} a_d \theta^d(b) \tag{5.6}$$

for $a_0, \dots, a_\ell, b \in R$. It can be checked that $\bullet$ satisfies the axioms of actions; for example, by (1.3) and (5.5), it holds

$$(sa) \bullet b = (\sigma(a)s + \delta(a)) \bullet b = \sigma(a)\theta(b) + \delta(a)b = \theta(ab) = s \bullet (ab)$$

for $a, b \in R$. Abusing notations, we represent by θ in place of s the indeterminate of the skew polynomial ring that acts on R by (5.6). We also write $p \bullet b$ as $p(b)$ for $p \in R[\theta;\sigma,\delta]$.

An ℓth-order (scalar) *linear σ-differential equation* over R is an equation for $y \in R$ in the form of

$$a_0 y + a_1 \theta(y) + \cdots + a_{\ell-1}\theta^{\ell-1}(y) + a_\ell \theta^\ell(y) = f, \tag{5.7}$$

where $a_0, \dots, a_\ell, f \in R$. The equation (5.7) can be written as $p(y) = f$ by using a skew polynomial $p \coloneqq a_0 + a_1\theta + \cdots + a_\ell\theta^\ell \in R[\theta;\sigma,\delta]$. We call θ in (5.7) the *σ-differential operator*. If $\sigma = \mathrm{id}$ and $\theta = \delta$, then σ-differential equations are called *linear differential equations*. Similarly, if $\delta = 0$ and $\theta = \sigma$, then σ-differential equations are said to be *linear difference equations*. The equation (5.7) is said to be *homogeneous* when $f = 0$ and *inhomogeneous* when $f \neq 0$.

Recall from Section 2.2.1 that $\theta(y)$ denotes $(\theta(y_i))_{i\in[n]}$ for $y = (y_i)_{i\in[n]} \in R^n$. An ℓth-order n-dimensional (matrix) *linear σ-differential equation* over R is an equation for $y \in R^n$ in form of

$$A_0 y + A_1 \theta(y) + \cdots + A_{\ell-1}\theta^{\ell-1}(y) + A_\ell \theta^\ell(y) = f, \tag{5.8}$$

where $A_0, \dots, A_\ell \in R^{n\times n}$ and $f \in R^n$. Using a skew polynomial matrix $A \coloneqq A_0 + A_1\theta + \cdots + A_\ell\theta^\ell \in R[\theta;\sigma,\delta]^{n\times n}$, the equation (5.8) is simply expressed as

$$A(y) = f. \tag{5.9}$$

The *solution space* of (5.9) is defined as $V \coloneqq \{y \in R^n \mid A(y) = f\}$. It is easily checked that V forms an affine module[1] over $\mathrm{Const}(R)$ unless $V = \varnothing$.

Suppose that R is a field K. Indeed, any σ-differential equation over a σ-differential field is essentially either a (usual) differential or difference equation. This follows from the

[1] Affine modules are a generalization of affine spaces obtained by replacing tangent vector spaces with modules. They are nothing but affine spaces if $\mathrm{Const}(R)$ is a field.

following facts.

Proposition 5.5 ([5, Lemma 5], [6, Lemma 1])**.** *Let (K, σ, δ) be a σ-differential field. Then the following hold:*

(1) *An additive map $\theta : K \to K$ is pseudo-linear if and only if it is in the form of $\gamma\sigma + \delta$ for some $\gamma \in K$.*

(2) *If $\sigma \neq \mathrm{id}$, then there exists $\alpha \in K$ such that $\delta = \alpha(\sigma - \mathrm{id})$.*

By Proposition 5.5, a pseudo-linear map θ can be written as $\theta = \delta + \gamma$ if $\alpha = \mathrm{id}$ and as $\theta = (\alpha + \gamma)\sigma + \alpha$ if $\sigma \neq \mathrm{id}$. Expanding θ^d for $d = 1, \ldots, \ell$ using these equations, any σ-differential equation $p(y) = 0$ with $p \in K[\theta; \sigma, \delta]$ is represented as $q(y) = 0$ for some $q \in K[\delta; \mathrm{id}, \delta]$ if $\sigma = \mathrm{id}$ and as $q'(y) = 0$ for some $q' \in K[\sigma; \sigma, 0]$ if $\sigma \neq \mathrm{id}$. A typical example of this reduction is the replacement of the difference operator in a difference equation by the shift operator. Therefore, it essentially suffices to consider only differential equations $(\theta = \delta)$ over a differential field and difference equations $(\theta = \sigma)$ over a difference field. Nonetheless, we make use of the notion of σ-differential equations whenever possible since it provides a useful framework unifying differential and difference equations.

5.2.2 Dimensions of Solution Spaces

Let (K, σ, δ) be a differential $(\sigma = \mathrm{id})$ or difference $(\delta = 0)$ field. We put $\theta \coloneqq \delta$ in the differential case and $\theta \coloneqq \sigma$ in the difference case. Consider a differential or difference equation (5.9) over K and suppose that (5.9) has at least one solution. The solution space V of (5.9) forms an affine space over $C \coloneqq \mathrm{Const}(K)$ as stated above. Now our question is how large the dimension $\dim_C V$ of V over C is. This quantity is rephrased as the number of values we must designate to determine a solution of (5.9) uniquely. An upper bound on $\dim_C V$ is given in terms of $\deg \mathrm{Det}$ and $\mathrm{ord}\,\mathrm{Det}$ of A as follows. This is partially given in [97, Lemma 1.10], [90, Corollary 4.9], [1, Theorem 6], and [93, Corollary 2.2], whereas they assume $\mathrm{ch}(K) = 0$ which is not needed to show the following. Here, we describe complete a proof based on their proofs.

Proposition 5.6. *Let (K, σ, δ) be a differential or difference field with $C \coloneqq \mathrm{Const}(K)$. Let V be the solution space of $A(y) = f$ with $A \in K[\theta; \sigma, \delta]^{n \times n}$ and $f \in K^n$ and suppose $V \neq \varnothing$. Then the following hold:*

(1) *If the field extension $K \,/\, C$ is infinite, then $\dim_C V$ is finite if and only if A is nonsingular.*

(2) *If A is nonsingular, it holds $\dim_C V \leq \deg \mathrm{Det}\, A$ in the differential case and $\dim_C V \leq \deg \mathrm{Det}\, A - \mathrm{ord}\,\mathrm{Det}\, A$ in the difference case.*

Proof. For any $v \in V$, the C-vector space $V - v \coloneqq \{y - v \mid y \in V\}$ is the solution space of $A(y) = 0$. Hence it suffices to consider only homogeneous equations. Our proof consists of

three steps: we show the claims for first-order homogeneous equations in Step 1, for scalar homogeneous equations in Step 2, and for general homogeneous equations in Step 3.

(Step 1) Consider the case when $A = A_0 + I_n\theta$ and $f = 0$, i.e., the corresponding linear σ-differential equation is

$$\theta(y) = -A_0 y. \tag{5.10}$$

We further require A_0 to be nonsingular only in the difference case. Since A is nonsingular, it suffices to show only (2). Then $A\theta^{-1} = A_0\theta^{-1} + I_n$ is proper as a matrix over $K(\theta;\sigma,\delta)$ with valuation $-\deg$. Since I_n is nonsingular, it holds $\deg \operatorname{Det} A\theta^{-1} = 0$ by Proposition 2.18 and thus $\deg \operatorname{Det} A = n$. Similarly, in the difference case, it holds $\operatorname{ord} \operatorname{Det} A = 0$ by the nonsingularity of A_0. Therefore, our goal is to show $\dim_C V \leq n$ in both cases. Since $\dim_K V \leq n$ is clear, it suffices to prove $\dim_K V = \dim_C V$.

Let $v_1, \dots, v_m \in V$ be solutions of (5.10) that are linearly dependent over K. We show that they are also dependent over C, which implies $\dim_K V = \dim_C V$. Without loss of generality, we assume that $v_2, \dots, v_m$ are linearly independent over K. Then there uniquely exists $c_2, \dots, c_m \in K$ such that $v_1 = \sum_{i=2}^m c_i v_i$. Then it holds

$$\begin{aligned}
0 &= \theta\left(v_1 - \sum_{i=2}^m c_i v_i\right) = \theta(v_1) - \sum_{i=2}^m \theta(c_i v_i) \\
&= -A_0 v_1 - \sum_{i=2}^m (\sigma(c_i)\theta(v_i) + \delta(c_i)v_i) \\
&= -A_0 \sum_{i=2}^m c_i v_i - \sum_{i=2}^m (-\sigma(c_i)A_0 v_i + \delta(c_i)v_i) \\
&= A_0 \sum_{i=2}^m (\sigma(c_i) - c_i)v_i - \sum_{i=2}^m \delta(c_i)v_i.
\end{aligned}$$

In the differential case, we have $0 = -\sum_{i=2}^m \delta(c_i)v_i$ by $\sigma = \mathrm{id}$. From the independence of $v_2, \dots, v_m$, it must holds $\delta(c_i) = 0$, which means $c_i \in C$ for $i = 2, \dots, m$. In the difference case, we have $0 = \sum_{i=2}^m (\sigma(c_i) - c_i)v_i$ from $\delta = 0$ and the assumption that A_0 is nonsingular. Hence we obtain $\sigma(c_i) = c_i$ and thus $c_i \in C$ for $i = 2, \dots, m$. Thus $v_1, \dots, v_m$ are also linearly dependent over C in both cases.

(Step 2) Consider a scalar homogeneous linear differential or difference equation $p(y) = 0$ with $p = \sum_{d=0}^{\ell} a_d \theta^d \in K[\theta;\sigma,\delta]$. When $p = 0$, the solution space V coincides with K. Thus $\dim_C V = \dim_C K$ is infinite when $K \,/\, C$ is infinite. Suppose that $p \neq 0$ and $\deg p = \ell$, i.e., $a_\ell \neq 0$. In the difference case, as $\theta = \sigma$ is bijective, $p(y) = 0$ and $p'(y) = 0$ with $p' \coloneqq \theta^{-\operatorname{ord} p} p$ have the same solution spaces. Moreover, by $\deg p' = \deg p - \operatorname{ord} p$ and $\operatorname{ord} p' = 0$, it holds $\deg p' - \operatorname{ord} p' = \deg p - \operatorname{ord} p$. Therefore, in the difference case, we can assume $\operatorname{ord} p = 0$ (i.e., $a_0 \neq 0$) without loss of generality.

We construct the following ℓ-dimensional matrix linear differential or difference equa-

tion:

$$\theta \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{\ell-2} \\ y_{\ell-1} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & 0 \\ 0 & \cdots & \cdots & 0 & 1 \\ -\frac{a_0}{a_\ell} & -\frac{a_1}{a_\ell} & \cdots & -\frac{a_{\ell-2}}{a_\ell} & -\frac{a_{\ell-1}}{a_\ell} \end{pmatrix} \begin{pmatrix} y_0 \\ y_1 \\ \vdots \\ y_{\ell-2} \\ y_{\ell-1} \end{pmatrix}. \tag{5.11}$$

If $y \in K$ is a solution of $p(y) = 0$, then $\big(y, \theta(y), \dots, \theta^{\ell-1}(y)\big)^\top \in K^n$ is a solution of (5.11). Conversely, any solution of (5.11) is obtained in this way. Therefore, the solution space W of (5.11) is isomorphic to V as C-vector spaces. In the differential case, $\dim_C W = \ell = \deg p$ by the above proof of Step 1. In the difference case, the matrix in the right-hand side of (5.11) is nonsingular by $a_0 \neq 0$. Hence $\dim_C W = \ell = \deg p - \operatorname{ord} p$ again from Step 1.

(Step 3) Consider a matrix homogeneous differential or difference equation $A(y) = 0$ with $A \in K[\theta; \sigma, \delta]^{n \times n}$. Let $D = UAW = \operatorname{diag}(d_1, \dots, d_n)$ be the Jacobson normal form of A over $K[\theta; \sigma, \delta]$. Putting $z = (z_1, \dots, z_n) \coloneqq W(y)$, the solution space sof $A(y) = 0$ and $D(z) = 0$ are isomorphic as C-vector spaces. Since D is diagonal, the solution space of $D(z) = 0$ is the direct sum of the solution space V_i of $d_i(z_i) = 0$ for $i \in [n]$. Namely, it holds

$$\dim_C V = \sum_{i=1}^{n} \dim_C V_i. \tag{5.12}$$

If A (and thus D) is singular, there exists $i \in [n]$ such that $d_i = 0$. Thus $\dim_C V$ is infinite when $K \,/\, C$ is infinite by the above Step 2 and (5.12). Suppose that A is nonsingular. Since U and W are invertible over $K[\theta; \sigma, \delta]$, they are biproper over $K(\theta; \sigma, \delta)$ with valuation deg and over $K(\theta; \sigma, 0)$ with valuation ord in the difference case. Thus $\deg \operatorname{Det}$ of U and W are 0, which means $\deg \operatorname{Det} A = \deg \operatorname{Det} D = \sum_{i=1}^{n} \deg d_i$. Therefore, by Step 2 and (5.12), we have $\dim_C V \leq \deg \operatorname{Det} A$ in the differential case, as desired. The completely analog holds in the difference case by replacing $\deg \operatorname{Det}$ with $\deg \operatorname{Det} - \operatorname{ord} \operatorname{Det}$. □

The upper bound on $\dim_C V$ given in Proposition 5.6 may not be attained on some equations. For example, consider a first-order linear differential equation $y' + y = 0$ over $\mathbb{C}(t)$ with the usual differentiation $'$. The solution of this equation over $\mathbb{C}(t)$ is only $y = 0$ and thus the dimension of the solution space is 0. However, if the differential field $\mathbb{C}(t)$ is extended to $\mathbb{C}(t, \mathrm{e}^t)$, the solution space becomes $V \coloneqq \{c\mathrm{e}^{-t} \mid c \in \mathbb{C}\}$, which has dimension 1 over $\mathbb{C}$. This is analogous to the situation of extending a field to its algebraic closure in order for nth-order algebraic equations to have n solutions. We explain such an extension briefly.

Let (K, σ, δ) be a differential or difference field. A differential or difference ring $(R, \bar{\sigma}, \bar{\delta})$ is called a *differential* or *difference extension* of K if K is a subring of R and $\bar{\sigma}$ and $\bar{\delta}$ coincides with σ and δ on K, respectively. A differential or difference equation $A(y) = f$ over K is naturally extended to that over R. Following [1], we call an extension R of K *adequate* if it satisfies the following:

(AE1) $C \coloneqq \mathrm{Const}(R)$ is a field.

(AE2) Any scalar homogeneous differential or difference equation $p(y) = 0$ with $p \in K[\theta; \sigma, \delta] \setminus \{0\}$ has the solution space V over R such that $\dim_C V = \deg p$ in the differential case and $\dim_C V = \deg p - \operatorname{ord} p$ in the difference case.

Let K be a differential field. If $\mathrm{Const}(K)$ is algebraically closed, then there exists an adequate extension R of K such that $\mathrm{Const}(R) = \mathrm{Const}(K)$, called the *universal* (*differential*) *Picard–Vessiot ring* of K [97, Section 3.2]. In addition, any differential field K of characteristic 0 has a difference extension whose constant field is the algebraic closure of $\mathrm{Const}(K)$ [1]; see also [97, Exercise 1.5, 2:(c), (d), 3:(c)]. Therefore, there always exists an adequate extension of any differential field of characteristic 0.

Next, suppose that K is a difference field. If $\mathrm{Const}(K)$ is algebraically closed, there exists an adequate extension R of K such that $\mathrm{Const}(R) = \mathrm{Const}(K)$, called the *universal* (*difference*) *Picard–Vessiot ring* of K [96, Section 1.4]. Indeed, for any difference field K of characteristic 0, an adequate difference extension R can be easily constructed [1, Proposition 4], while $\mathrm{Const}(R) = \mathrm{Const}(K)$ is no longer guaranteed.

We then turn to matrix, inhomogeneous equations. As we will see below, (AE2) is indeed equivalent to the following:

(AE2') Any matrix differential or difference equation $A(y) = f$ with $A \in \mathrm{GL}_n(K[\theta; \sigma, \delta])$ and $f \in K^n$ has the solution space V over R such that $\dim_C V = \deg \operatorname{Det} A$ in the differential case and $\dim_C V = \deg \operatorname{Det} A - \operatorname{ord} \operatorname{Det} A$ in the difference case.

Lemma 5.7. (AE2) *and* (AE2') *are equivalent.*

Proof. It is clear that (AE2') implies (AE2); we show the converse holds. Let (K, σ, δ) be a differential or difference field and R its extension satisfying (AE1) and (AE2). As stated in the proof of Proposition 5.6, a matrix differential and difference equation is essentially reduced to n scalar equations by considering the Jacobson normal form. This means that it suffices to consider only a scalar inhomogeneous equation $p(y) = f$ with $p \in K[\theta; \sigma, \delta] \setminus \{0\}$ and $f \in K \setminus \{0\}$. In addition, the solution space of $p(y) = f$ over R is the translation of the solution space of $p(y) = 0$ over R by any solution of $p(y) = f$. Therefore, our goal is to show that $p(y) = f$ has at least one solution over R.

We first deal with the differential case. Let $q \coloneqq \theta f^{-1} p$. Then any solution $y \in R$ of $q(y) = 0$ is also a solution of $p(y) = cf$ for some $c \in C \coloneqq \mathrm{Const}(R)$ (see [97, Exercise 1.14,

1]). By (AE2), the dimension of the solution space W of $q(y) = 0$ is $\deg q = \deg p + 1$, whereas that of $p(y) = 0$ is $\deg p < \deg q$. Therefore, there exists $v \in W$ that is not a solution of $p(v) = 0$, i.e., $p(v) = cf$ for some nonzero $c \in C^\times$. Then $c^{-1}v$ is a solution of $p(y) = f$, as required. The difference case can be in the same way by considering $q \coloneqq (\theta - 1)(f^{-1}p) = \theta f^{-1}p - f^{-1}p$. □

Proposition 5.6 and Lemma 5.7 lead us to the following consequence.

Theorem 5.8. *Let (K, σ, δ) be a differential or difference field, R its adequate extension, and $C \coloneqq \mathrm{Const}(R)$. Let V be the solution space of $A(y) = f$ over R with $A \in \mathrm{GL}_n(K[\theta; \sigma, \delta])$ and $f \in K^n$. Then it holds $\dim_C V = \deg \mathrm{Det}\, A$ in the differential case and $\dim_C V = \deg \mathrm{Det}\, A - \mathrm{ord}\, \mathrm{Det}\, A$ in the difference case.*

Since deg and ord are discrete valuations, we can apply our algorithms given in Chapter 4 to compute the dimension of solution spaces of linear differential or difference equations over an adequate extension.

Chapter 6

Structural Methods for Differential-Algebraic Equations

This chapter reviews the literature of the structural preprocessing methods for *differential-algebraic equations* (DAEs). In Section 6.1, we introduce DAEs with some examples and explain two preprocessing processes, consistent initialization and index reduction, needed prior to numerical integration. In Section 6.2, we describe *structural preprocessing methods* for DAEs that are based on the assignment problem. While structural methods are efficient, they fail for some DAEs; we analyze the failures in Section 6.3. Finally, Section 6.4 describes that the combinatorial relaxation provides a framework of overcome the structural methods' failure.

6.1 Differential-Algebraic Equations

Let $\mathbb{T} \subseteq \mathbb{R}$ be a nonempty open interval and $\Omega \subseteq \mathbb{R}^{(\ell+1)n}$ a nonempty open set. An ℓth-order *differential-algebraic equation* (DAE) of size n for $x : \mathbb{T} \to \mathbb{R}^n$ is the equation (1.4), where $F : \mathbb{T} \times \Omega \to \mathbb{R}^n$ is a sufficiently smooth function.

6.1.1 DAE Examples from Dynamical Systems

DAEs are widely accepted as a mathematical model of dynamical systems since the pioneering work of Gear [29] and subsequent developments of theory and numerical methods. In this section, we demonstrate two examples of DAEs arising in dynamical systems: mechanical systems and electrical networks.

Example 6.1 (simple pendulumn)**.** Consider a simple planar pendulum illustrated in Figure 6.1, where a point of unit mass is suspended by a massless cord of length L from a pivot without friction. We take the Cartesian coordinate, in which the x_1-axis is in the horizontal direction and the x_2-axis is in the downward direction.

The (twice of) Lagrangian of this system is $\mathcal{L} = \dot{x}_1{}^2 + \dot{x}_2{}^2 + gy$, where g is (twice of)

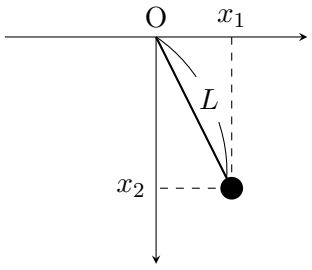


Figure 6.1: Simple pendulum.

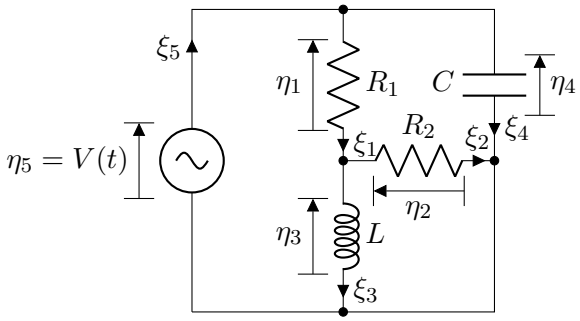


Figure 6.2: Simple RLC network.

the gravitational constant. According to the principle of least action in the Lagrangian mechanics, the motion from the time t_0 to t_1 minimizes the action $\int_{t_0}^{t_1} \mathcal{L}(t)\mathrm{d}t$. Thus, by solving the Euler–Lagrange equation with constraint ${x_1}^2 + {x_2}^2 - L^2 = 0$, we obtain an equation of motion

$$\begin{cases} \ddot{x}_1 + x_1 x_3 & = 0, \\ \ddot{x}_2 + x_2 x_3 - g & = 0, \\ {x_1}^2 + {x_2}^2 - L^2 = 0 \end{cases} \tag{6.1}$$

with Lagrange multiplier x_3. The equation system (6.1) is a DAE consisting of two differential equations and one purely algebraic equation with unknown functions $x_1(t)$, $x_2(t)$, and $x_3(t)$. □

Example 6.2 (RLC network)**.** Consider an electrical network illustrated in Figure 6.2, which is given in [71, Section 1.1]. The network consists of a voltage source of time-varying voltage $V(t)$, two resistances R_1 and R_2, an inductor L, and a capacitor C. State variables of this network is currents $\xi_1, \ldots, \xi_5$ and voltages $\eta_1, \ldots, \eta_5$ shown in Figure 6.2. One of

the equation systems representing this network is given by

$$\begin{cases} -\xi_1 - \xi_4 + \xi_5 = 0, \\ \xi_2 + \xi_3 + \xi_4 - \xi_5 = 0, \\ \eta_1 + \eta_3 - \eta_5 = 0, \\ -\eta_1 - \eta_2 + \eta_4 = 0, \\ \eta_2 - \eta_3 = 0, \\ R_1\xi_1 - \eta_1 = 0, \\ R_2\xi_2 - \eta_2 = 0, \\ L\dot{\xi}_3 - \eta_3 = 0, \\ -\xi_4 + C\dot{\eta}_4 = 0, \\ \eta_5 = V(t). \end{cases} \tag{6.2}$$

In this system (6.2), the first two equations come from Kirchhoff's current law (KCL), and the following three equations come from Kirchhoff's voltage law (KVL). These equations, which are called *structural equations*, are purely algebraic. On the other hand, the last five equations, called *constitutive equations*, represent the element characteristics coming from Ohm's law, the capacitor's differential equation and the inductor's one. □

6.1.2 Consistency of Initial Values

The *initial value problem* for a DAE (1.4) is to find a solution $x(t)$ of (1.4) satisfying the initial value condition (1.6) for given $t^* \in \mathbb{T}$ and initial values $x^*_{(0)}, x^*_{(1)}, \ldots, x^*_{(\ell-1)} \in \mathbb{R}^n$. Initial values are called *consistent* if there uniquely exists a smooth solution $x(t)$ of (1.4) in a neighborhood of the initial value on $\mathbb{T} \times \Omega$. Any initial values are consistent for an ordinary differential equation (ODE) in form of (1.5) under the smoothness assumption on φ. For DAEs, the initial value problem may not have a solution because DAEs can involve algebraic constraints.

Example 6.3. Consider the problem of giving a consistent initial value at $t^* = 0$ for the DAE (6.1) representing the simple pendulum. First, by the constraint ${x_1}^2 + {x_2}^2 - L^2 = 0$, one may choose

$$x_1(0) = x_2(0) = \frac{L}{\sqrt{2}}.$$

From the first and second equations in (6.1), we must set $\ddot{x}_1(0)$, $\ddot{x}_2(0)$ and $x_3(0)$ so that

they satisfy

$$\ddot{x}_1(0) = -x_1(0)x_3(0) = -\frac{L}{\sqrt{2}}x_3(0),$$
$$\ddot{x}_2(0) = -x_2(0)x_3(0) + g = -\frac{L}{\sqrt{2}}x_3(0) + g.$$

One possibility is

$$x_3(0) = 1, \quad \ddot{x}_1(0) = -\frac{L}{\sqrt{2}}, \quad \ddot{x}_2(0) = -\frac{L}{\sqrt{g}} + g.$$

We here try to set an initial velocity $\dot{x}_1(0)$ and $\dot{x}_2(0)$ as $\dot{x}_1(0) = \dot{x}_2(0) = 1$ since there are no equations in (6.1) constraining them. Now these initial values satisfy all the equations in (6.1).

Unfortunately, the DAE (6.1) has no solution with the above initial values because all solutions are subject to the differentiation $2x_1\dot{x}_1 + 2x_2\dot{x}_2 = 0$ of the constraint ${x_1}^2 + {x_2}^2 - L^2 = 0$, though the above initial value does not meet it. □

As seen in Example 6.3, consistent initial values must satisfy not only algebraic equations but also their differentiations, called *hidden constraints*. Hence the *consistent initialization*, the problem to give a consistent initial value, is not trivial for DAEs.

6.1.3 Differentiation Index

The index concept plays an important role in stability analysis for numerical analysis of DAEs. For a first-order DAE (1.7) and $k \in \mathbb{N}$, define

$$\boldsymbol{F}_k\big(t, x, \dot{x}, \ldots, x^{(k+1)}\big) \coloneqq \begin{pmatrix} F\big(t, x(t), \dot{x}(t)\big) \\ \frac{\mathrm{d}}{\mathrm{d}t}F\big(t, x(t), \dot{x}(t)\big) \\ \vdots \\ \frac{\mathrm{d}^k}{\mathrm{d}t^k}F\big(t, x(t), \dot{x}(t)\big) \end{pmatrix}.$$

The *differentiation index*, or the *index* for short, of a first-order DAE (1.7) is the minimum $\nu \in \mathbb{N}$ (if it exists) meeting the following: for each $(t, x_{(0)}) \in \mathbb{R}^{1+n}$, if there exists $\big(x_{(1)}, \ldots, x_{(\nu+1)}\big) \in \mathbb{R}^{n\mu}$ satisfying $\boldsymbol{F}_\nu\big(t, x_{(0)}, \ldots, x_{(\nu+1)}\big) = 0$, then such $x_{(1)}$ is unique [9]. Loosely speaking, the differentiation index is the minimum nonnegative integer ν such that the equation $\boldsymbol{F}_\nu = 0$ can determine $\dot{x}$ as a function of t and x. In this sense, the differentiation index measures how far the DAE is from ODEs.

Example 6.4. An ODE (1.5) has differentiation index 0 because $\dot{x}$ is always determined from t and x by φ. An algebraic equation $F(t, x(t)) = 0$ with nonsingular Jacobian matrix

$\frac{\partial F}{\partial x} = \left(\frac{\partial F_i}{\partial x_j}\right)_{i,j\in[n]}$ has differentiation index 1 because

$$\frac{\mathrm{d}}{\mathrm{d}t}F(t,x(t)) = \frac{\partial F}{\partial t}(t,x(t)) + \frac{\partial F}{\partial x}(t,x(t))\dot{x}(t) = 0$$

implies

$$\dot{x}(t) = -\left(\frac{\partial F}{\partial x}(t,x(t))\right)^{-1}\frac{\partial F}{\partial t}(t,x(t)),$$

from which $\dot{x}(t)$ is determined from t and $x(t)$ by $\boldsymbol{F}_1(t,x,\dot{x}) = 0$. □

For a higher-order DAE, we define its differentiation index as as that of the first-order DAE obtained by replacing higher-order derivatives with newly introduced variables.

Example 6.5. The DAE (6.1) for the simple pendulum is converted into the following first-order DAE

$$\begin{cases} \dot{x}_4 + x_1x_3 = 0, \\ \dot{x}_5 + x_2x_3 - g = 0, \\ {x_1}^2 + {x_2}^2 - L^2 = 0, \\ \dot{x}_1 - x_4 = 0, \\ \dot{x}_2 - x_5 = 0, \end{cases} \tag{6.3}$$

where x_4 and x_5 represent $\dot{x}_1$ and $\dot{x}_2$, respectively. We consider the system of 7 equations obtained by replacing the third equation in (6.3) with

$$2x_1\ddot{x}_4 + 6x_4\dot{x}_4 + 2x_2\ddot{x}_5 + 6x_5\dot{x}_5 = 0 \tag{6.4}$$

and appending two extra equations

$$\begin{cases} \ddot{x}_4 + x_4x_3 + x_1\dot{x}_3 = 0, \\ \ddot{x}_5 + x_5x_3 + x_2\dot{x}_3 = 0. \end{cases} \tag{6.5}$$

Here, (6.4) is the second-order differentiation of the original third equation and (6.5) is the first-order differentiations of the first and second equations (with $\dot{x}_1$ and $\dot{x}_2$ replaced with x_4 and x_5). The resulting system is a subsystem of $\boldsymbol{F}_3 = 0$ for (6.3).

The Jacobian matrix D of this system with respect to $(\dot{x}_1, \dot{x}_2, \dot{x}_3, \dot{x}_4, \dot{x}_5, \ddot{x}_4, \ddot{x}_5)$ is

$$D = \begin{pmatrix} & & 1 & & & & \\ & & & 1 & & & \\ & & 6x_4 & 6x_5 & 2x_1 & 2x_2 & \\ 1 & & & & & & \\ & 1 & & & & & \\ & x_1 & & & 1 & & \\ & x_2 & & & & 1 & \end{pmatrix}, \tag{6.6}$$

where empty entries indicate 0. Now we have

$$\det D = -2{x_1}^2 - 2{x_2}^2 = -2L^2 \neq 0 \tag{6.7}$$

for any consistent point. Thus, by the implicit function theorem, we can solve the system for $(\dot{x}_1, \dot{x}_2, \dot{x}_3, \dot{x}_4, \dot{x}_5, \ddot{x}_4, \ddot{x}_5)$ as a smooth function of $(x_1, x_2, x_3, x_4, x_5)$. This means that the differentiation indices of (6.3) and of (6.1) are at most 3, and are actually 3 [56]. □

The differentiation index often appears in validity conditions and convergence results of numerical methods for DAEs. For example, the code `DASSL` of Petzold [81] and the DAE solver `ode15i` in MATLAB, both of which are based on the *backward differentiation formula* (BDF), can handle only index-1 DAEs [4]. The code `RADAU5` of Hairer–Wanner [36], using the *implicit Runge–Kutta* (IRK) method, is applicable to DAEs of index less than 4. With regard to the convergence rate of numerical methods, for example, Gear et al. [30] showed that the global error in the numerical solution for k-step variable stepsize BDF applied to an index-ν linear DAE with constant coefficient is $\mathrm{O}({h_{\max}}^q)$, where $q \coloneqq \min\{k, k - \nu + 2\}$ and $h_{\max}$ is the maximum ratio of adjacent step size. It is generally considered that only index 0 or 1 DAEs can be numerically solved in practical accuracy. Therefore, the *index reduction*, which is a process to convert higher-index DAEs into lower-index ones, is important for accurate simulation of dynamical systems.

6.2 Structural Preprocessing Methods

Structural preprocessing methods for DAEs are popular methods of consistent initialization and index reduction prior to numerical integration. They make use of the assignment problem (weighted bipartite matching problem) on a bipartite graph constructed from a DAE. In this section, after introducing preliminaries and the assignment problem for DAEs, we explain two structural preprocessing methods: the Σ-method for consistent initialization and the Mattsson–Söderlind method for index reduction.

6.2.1 Griewank's Lemma

Structural methods utilize information on which variable each equation depends. We first introduce notations and a proposition to describe the structural preprocessing methods.

Let $\mathbb{T} \subseteq \mathbb{R}$ be a nonempty open interval and $\Omega \subseteq \mathbb{R}^{(\ell+1)n}$ a nonempty open set having coordinates $(x, \dot{x}, \dots, x^{(\ell)})$, where $x^{(k)} = (x_j^{(k)})_{j \in C} \in \mathbb{R}^n$ for $k \in [0, \ell]$. Here C is the set of indices with $|C| = n$. In the subsequent discussion, each $x_j^{(k)}$ is regarded not as the kth-order derivative of some trajectory but as an independent variable.

Let $f : \mathbb{T} \times \Omega \to \mathbb{R}$ be a smooth function. For $j \in C$ and $k \in [0, \ell]$, the function f is said to *depend on* $x_j^{(k)}$ if the partial derivative $\frac{\partial f}{\partial x_j^{(k)}}$ is not identically zero on the domain $\mathbb{T} \times \Omega$ of f. We denote the maximum nonnegative integer k such that f depends on $x_j^{(k)}$ by $\sigma(f, x_j)$. If f does not depend on $x_j^{(k)}$ for any k, we assign $\sigma(f, x_j) \coloneqq -\infty$ for convenience. The *derivative* $\dot{f}$ of f with respect to t is defined by

$$\dot{f}(t, x, \dot{x}, \dots, x^{(\ell+1)}) \coloneqq \frac{\partial f}{\partial t}(t, x, \dot{x}, \dots, x^{(\ell)}) + \sum_{k=0}^{\ell} \frac{\partial f}{\partial x^{(k)}}(t, x, \dot{x}, \dots, x^{(\ell)}) x^{(k+1)}.$$

For $d \in \mathbb{N}$, the dth-order derivative $f^{(d)}$ of f is recursively defined by $f^{(0)} \coloneqq f$ and $f^{(d)} \coloneqq \dot{f}^{(d-1)}$ for $d \geq 1$. It should be noted that the domain of $\dot{f}$ is not $\mathbb{T} \times \Omega$ but $\mathbb{T} \times \Omega \times \mathbb{R}^n$ because $\dot{f}$ (linearly) depends on $x^{(\ell+1)}$. Similarly, for a nonnegative integer d, we regard the domain of $f^{(d)}$ as $\mathbb{T} \times \Omega^{(d)}$, where $\Omega^{(d)} \coloneqq \Omega \times \mathbb{R}^{dn}$.

The following simple proposition plays an important role in structural preprocessing methods for DAEs.

Proposition 6.6 (Griewank's lemma [34, Section 2.2],[82, Lemma 3.7])**.** *Let $f : \mathbb{T} \times \Omega \to \mathbb{R}$ be a smooth function. For $j \in C$ and a nonnegative integer d, if $\sigma(f, x_j) \leq c$, then*

$$\frac{\partial f}{\partial x_j^{(c)}}(t, x, \dot{x}, \dots, x^{(\ell)}) = \frac{\partial f^{(d)}}{\partial x_j^{(c+d)}}(t, x, \dot{x}, \dots, x^{(\ell+d)}) \tag{6.8}$$

holds for all $(t, x, \dot{x}, \dots, x^{(\ell+d)}) \in \mathbb{T} \times \Omega^{(d)}$.

We sometimes regard the domain of $\frac{\partial f^{(d)}}{\partial x_j^{(c+d)}}$ not as $\mathbb{T} \times \Omega^{(d)}$ but as $\mathbb{T} \times \Omega$ to simply write the equality (6.8) as $\frac{\partial f}{\partial x_j^{(c)}} = \frac{\partial f^{(d)}}{\partial x_j^{(c+d)}}$. In addition, it follows from Proposition 6.6 that

$$\sigma(f^{(d)}, x_j) = \sigma(f, x_j) + d$$

holds for $j \in C$ and a nonnegative integer d.

6.2.2 Assignment Problem

Pryce [82] introduced an assignment problem for a reinterpretation of Pantelides' algorithm [79] as follows. Consider a DAE (1.4) of size n with equation index set R and variable index set C. Let $G(F)$ denote the bipartite graph with vertex set $R \cup C$ and edge set

$$E(F) \coloneqq \{\{i,j\} \mid i \in R, j \in C, \sigma(F_i, x_j) > -\infty\}.$$

We set the weight c_e of an edge $e = (i,j) \in E(F)$ by $c_e = c_{i,j} = \sigma(F_i, x_j)$. Denote the maximum-weight perfect matching problem on $G(F)$ by $\mathrm{P}(F)$ and consider the following formulation of the dual of $\mathrm{P}(F)$:

$$\mathrm{D}(F) \left| \begin{array}{lll} \text{minimize} & \displaystyle\sum_{j \in C} q_j - \sum_{i \in R} p_i & \\ \text{subject to} & q_j - p_i \geq c_{i,j} & (i \in R, j \in C, \{i,j\} \in E(F)), \\ & p_i, q_j \in \mathbb{N} & (i \in R, j \in C). \end{array} \right.$$

Define

$$\hat{d}(F) \coloneqq \text{the optimal value of the problem } \mathrm{D}(F).$$

For a dual feasible solution (p,q), the *system Jacobian* $D = (D_{i,j})_{i \in R, j \in C} : \mathbb{T} \times \Omega \to \mathbb{R}^{n \times n}$ of F with respect to (p,q) is a matrix defined by

$$D_{i,j} \coloneqq \frac{\partial F_i^{(p_i)}}{\partial x_j^{(q_j)}} = \frac{\partial F_i}{\partial x_j^{(q_j - p_i)}} \tag{6.9}$$

for each $i \in R$ and $j \in C$. The last equality in (6.9) for (i,j) with $q_j - p_i \geq 0$ is due to Proposition 6.6. The equality also holds for (i,j) with $q_j - p_i < 0$ by regarding $\frac{\partial F_i}{\partial x_j^{(q_j - p_i)}}$ as an identically zero function. By $q_j - p_i \geq c_{i,j}$, the entry $D_{i,j}$ is nonzero if and only if $q_j - p_i = c_{i,j}$. Therefore, the following holds as a restatement of Proposition 3.3.

Proposition 6.7. *For a DAE* (1.4) *of size* n*, let* D *be a system Jacobian of the DAE with respect to a feasible solution* (p,q) *of* $\mathrm{D}(F)$*. Then* (p,q) *is optimal if and only if* t-rank $D = n$*.*

Example 6.8. Let $F = 0$ be the DAE (6.1) representing the simple pendulum. The weight of edges in $G(F)$ is shown in the following matrix:

$$\begin{pmatrix} 2 & -\infty & 0 \\ -\infty & 2 & 0 \\ 0 & 0 & -\infty \end{pmatrix},$$

whose the (i,j)th entry indicates the weight of $\{i,j\} \in E(F)$ ($-\infty$ means $\{i,j\} \notin E(F)$) for $i \in R$ and $j \in C$. One choice of an optimal solution of $\mathrm{D}(F)$ is $p = (0,0,2)$ and $q = (2,2,0)$, which means $\hat{d}(F) = 2$. The corresponding system Jacobian is

$$D = \begin{pmatrix} 1 & & x_1 \\ & 1 & x_2 \\ 2x_1 & 2x_2 & \end{pmatrix},$$

which is nonsingular at any consistent point by $\det D = -2{x_1}^2 - 2{x_2}^2 = -L^2 \neq 0$. □

Example 6.9 (linear DAEs with constant coefficients)**.** Let $F = 0$ be a *linear DAE with constant coefficients* defined by (1.8), where $A_0, \ldots, A_\ell \in \mathbb{R}^{n \times n}$ and $f : \mathbb{R} \to \mathbb{R}^n$ is a smooth function. Applying the Laplace transformation, the DAE (1.8) is written as

$$A(s)\tilde{x}(s) = \tilde{f}(s), \tag{6.10}$$

where

$$A = A(s) \coloneqq \sum_{d=0}^{\ell} A s^d \in \mathbb{R}[s]^{n \times n}$$

and $\tilde{x}(s)$ and $\tilde{f}(s)$ are the Laplace transforms of $x(t)$ and $f(t)$, respectively (assuming $x^{(d)}(0) = 0$ for all $d \in \mathbb{N}$ for simplicity). From the algebraic viewpoint as in Section 5.2, we can also regard (6.10) as a DAE itself by regarding s as the differentiation operator.

Recall from Section 4.2.1 that the problem $\mathrm{D}(A)$ is defined for the polynomial matrix A. It is easily seen that $\mathrm{D}(F)$ and $\mathrm{D}(A)$ are the same problem. Moreover, the system Jacobian of (6.10) coincides with the tight coefficient matrix $A^{\#}$ of A defined in Section 4.2.1. In this sense, the system Jacobian can be seen as a nonlinear generalization of the tight coefficient matrix for polynomial matrices. □

6.2.3 Consistent Initialization by the Σ-Method

Pryce's Σ-method [82] is a structural preprocessing method for finding a consistent initial value of a DAE (1.4) at a given initial time $t^* \in \mathbb{T}$. The Σ-method is outlined as follows.

Σ-Method

Step 1. Compute an optimal solution (p,q) of $\mathrm{D}(F)$. If $\mathrm{D}(F)$ has no optimal solution, then the Σ-method terminates with failure.

Step 2. Collect $M \coloneqq \sum_{i \in R} p_i + n$ equations

$$\begin{cases} F_1 = 0, & \dot{F}_1 = 0, & \ldots, & F_1^{(p_1)} = 0, \\ F_2 = 0, & \dot{F}_2 = 0, & \ldots, & F_2^{(p_2)} = 0, \\ & & \vdots & \\ F_n = 0, & \dot{F}_n = 0, & \ldots, & F_n^{(p_n)} = 0. \end{cases} \tag{6.11}$$

Solve (6.11) as a system of algebraic equations for $N \coloneqq \sum_{j \in C} q_j + n$ unknown variables

$$X \coloneqq (x_1, \dot{x}_1, \ldots, x_1^{(q_1)}, x_2, \dot{x}_2, \ldots, x_2^{(q_2)}, \ldots, x_n, \dot{x}_n, \ldots, x_n^{(q_n)})$$

to obtain an initial value (t^*, X^*).

Step 3. If the Σ-Jacobian D with respect to (p, q) is singular at (t^*, X^*), the Σ-method terminates with failure. Otherwise, return (t^*, X^*).

Theorem 6.10 ([82, Theorem 4.2])**.** *Let $F(t, x, \dot{x}, \ldots) = F(t, X) = 0$ be a DAE with equation index set R and variable index set C. Suppose that $\mathrm{D}(F)$ has an optimal solution (p, q) and let D be the Σ-Jacobian with respect to (p, q). If the Σ-method finds an initial value (t^*, X^*) at which D is nonsingular, then (t^*, X^*) is consistent. Moreover, the differentiation index of the DAE is locally bounded by*

$$\max_{i \in R} p_i + \begin{cases} 0 & (q_j > 0 \text{ for all } j \in C), \\ 1 & (\text{otherwise}). \end{cases}$$

Example 6.11. Consider the DAE (6.1) representing the simple pendulum again. In Step 1, we obtain $p = (0, 0, 2)$ and $q = (2, 2, 0)$. In Step 2, we solve the following system

$$\begin{cases} \ddot{x}_1 + x_1 x_3 = 0, \\ \ddot{x}_2 + x_2 x_3 - g = 0, \\ {x_1}^2 + {x_2}^2 - L^2 = 0, \\ 2x_1 \dot{x}_1 + 2x_2 \dot{x}_2 = 0, \\ 2{\dot{x}_1}^2 + 2x_1 \ddot{x}_1 + 2{\dot{x}_2}^2 + 2x_2 \ddot{x}_2 = 0 \end{cases}$$

as an algebraic equation system for $x_1, \dot{x}_1, \ddot{x}_1, x_2, \dot{x}_2, \ddot{x}_2$, and x_3 to obtain an initial value. One solution (for $t^* = 0$) is

$$x_1^* = L, \quad x_2^* = \dot{x}_1^* = \dot{x}_2^* = \ddot{x}_1^* = 0, \quad \ddot{x}_2^* = g. \tag{6.12}$$

The Σ-Jacobian of (6.1) with respect to (p, q) is the matrix defined by (6.6). Since

$\det D = -L$ by (6.7), the Σ-method succeeds. From Theorem 6.10, the index of (6.1) is at most 3, which agrees with the consequence of Example 6.5. □

Theorem 6.10 also gives an information on the dimension of the solution manifold. Suppose that the Σ-method succeeds for a DAE $F = 0$ at (t^*, X^*). Then Theorem 6.10 indicates that one can determine a solution by solving $M \coloneqq \sum_{i \in R} p_i + n$ equations for $N \coloneqq \sum_{j \in C} q_j + n$ variables, where (p, q) is a dual optimal solution. This means that the dimension of the solution manifold, or solutions' degree of freedom, is $M - N = \sum_{i \in R} p_i - \sum_{j \in C} q_j = \hat{d}(F)$. Hence we have:

Theorem 6.12 ([82])**.** *For a DAE* (1.4)*, suppose that* $\mathrm{D}(F)$ *has an optimal solution* (p, q) *and let* D *be the system Jacobian of* (1.4) *with respect to* (p, q)*. If* D *is nonsingular at a consistent point* (t^*, X^*) *of* (1.4)*, then the dimension of the solution manifold of* (1.4) *in a neighborhood of* (t^*, X^*) *is* $\hat{d}(F)$*.*

Example 6.13. Consider the DAE (6.1) representing the simple pendulum. As we have seen in Example 6.8, the DAE has a nonsingular system Jacobian. Thus by Theorem 6.12, the dimension of its solution manifold is $\hat{d}(F) = 2$. This agrees with the fact that the state of the simple pendulum is determined from (i) the angle between the cord and the x_2-axis and (ii) the velocity of the mass along the normal direction of the cord. □

Example 6.14. Consider a linear DAE (6.10) with constant coefficients and let $A^{\#} \in \mathbb{R}^{n \times n}$ be the system Jacobian of (6.10) with respect to an optimal solution of $\mathrm{D}(F)$ for (6.10). By Proposition 4.7, if $A^{\#}$ is nonsingular, then A is upper-tight, i.e., $\deg \det A = \hat{d}(F)$. On the other hand, Chrystal's theorem [12] (and Theorem 5.8) state that the dimension of the solution space of (6.10) is equal to $\deg \det A$. Thus, $\hat{d}(F)$ coincides with the dimension if A is upper-tight. This is a special case of Theorem 6.12. □

6.2.4 Index Reduction by the Mattsson–Söderlind Method

We next review the Mattsson–Söderlind index reduction method [60] (MS-method). For an optimal solution (p, q) of $\mathrm{D}(F)$ and $h \in \mathbb{Z}$, define

$$R_h \coloneqq \{i \in R \mid p_i = h\}, \quad R_{\geq h} \coloneqq \{i \in R \mid p_i \geq h\},$$
$$C_h \coloneqq \{j \in C \mid q_j = h\}, \quad C_{\geq h} \coloneqq \{j \in C \mid q_j \geq h\}.$$

The input of the MS-method is a DAE and its consistent initial value (t^*, X^*). The MS-method is outlined as follows [60, Section 3.1].

Mattsson–Söderlind Index Reduction Method

Step 1. Compute an optimal solution (p, q) of $\mathrm{D}(F)$. If $\mathrm{D}(F)$ has no optimal solution, or the Σ-Jacobian D with respect to (p, q) is singular at (t^*, X^*), then the algorithm terminates in failure.

Step 2. For each $h \in [0, \eta+1]$ ($\eta \coloneqq \max_{i \in R} p_i$), obtain $J_h \subseteq C_{\geq h}$ such that $D[R_{\geq h}, J_h]$ is nonsingular at (t^*, X^*) and

$$C = J_0 \supseteq J_1 \supseteq J_2 \supseteq \cdots \supseteq J_\eta \supseteq J_{\eta+1} = \varnothing.$$

Step 3. For each $j \in C$, let k_j be an integer such that $j \in J_{k_j}$ and $j \notin J_{k_j+1}$. Introduce k_j dummy variables $z_j^{[q_j]}$, $z_j^{[q_j-1]}$, ..., $z_j^{[q_j-k_j+1]}$ corresponding to $x_j^{(q_j)}$, $x_j^{(q_j-1)}$, ..., $x_j^{(q_j-k_j+1)}$, respectively.

Step 4. For each $i \in R$, collect the 0th-, 1st-, ..., p_ith-order derivatives of the ith equation $F_i(t, x, \dot{x}, \ldots) = 0$. Replace variables in the collected system with the corresponding dummy variables.

Proposition 6.15 ([60, Section 3.2])**.** *For a DAE $F = 0$, suppose that $\mathrm{D}(F)$ has an optimal solution (p, q) and let D be the Σ-Jacobian with respect to (p, q). If D is nonsingular at a given consistent initial value, then the MS-method returns an equivalent DAE whose index is locally at most 1.*

Here, the term "equivalent" in Proposition 6.15 means that there is a trivial one-to-one correspondence between solutions of the original DAE and the returned DAE by the MS-method. Namely, for every solution x of the original DAE, there uniquely exists a function z corresponding to dummy variables such that (x, z) is a solution of the returned DAE, and conversely, for every solution (x, z) of the returned DAE, x is a solution of the original DAE.

Example 6.16. Consider the DAE (6.1) on the simple pendulum with consistent initial value given in (6.12). We find $p = (0, 0, 2)$ and $q = (2, 2, 0)$ in Step 1 of the MS-method. In Step 2, we choose $J_0, \ldots, J_3$ as

$$D[R_{\geq 0}, J_0] = \begin{pmatrix} 1 & & x_1 \\ & 1 & x_2 \\ 2x_1 & 2x_2 & \end{pmatrix}, \quad D[R_{\geq 1}, J_1] = D[R_{\geq 2}, J_2] = \begin{pmatrix} 2x_1 \end{pmatrix},$$

and $D[R_{\geq 3}, J_3]$ is the 0×0 matrix. In Step 3, we introduce two dummy variables $z_1^{[1]}$ and $z_1^{[2]}$ corresponding to $\dot{x}_1$ and $\ddot{x}_2$, respectively. Finally, we obtain a DAE

$$\begin{cases} z^{[2]} + x_1 x_3 = 0, \\ \ddot{x}_2 + x_2 x_3 - g = 0, \\ z_1^{[2]} + x_1^2 - L^2 = 0, \\ 2x_1 z_1^{[1]} + 2x_2 \dot{x}_2 = 0, \\ 2\left(z_1^{[1]}\right)^2 + 2x_1 z_1^{[2]} + 2\dot{x}_2^2 + 2x_2 \ddot{x}_2 = 0, \end{cases}$$

which is of index-1 as required. □

6.3 Failures of Structural Preprocessing Methods

As we have seen in Section 6.2, structural methods run only for DAEs having nonsingular system Jacobian. In practice, this is satisfied on many DAEs of real instances. For example, Pryce [82] showed that the Σ-method can be applied to any DAE which is of index 0, in standard canonical form, in Hessenberg form, a constrained mechanical system, or a triangular chain of systems for which the method works [82, Theorem 5.3]. The structural preprocessing methods succeed for seven instances out of nine DAE problems in the test set for IVP (initial value problem) solvers collected by Mazzia and Magherini [61].

However, it is also true that the structural preprocessing methods do not work for two DAEs in the test set, which model electrical circuits describing the behavior of a transistor amplifier and a ring modulator. Scholz [86] reports that the structural preprocessing methods fail even for a DAE modeling a simple RLC circuit.

Here we investigate how the structural preprocessing methods fail. The failures are classified into the following three scenarios:

(F1) The bipartite graph $G(F)$ has no perfect matching, or equivalently, the dual problem $\mathrm{D}(F)$ has no optimal solution.

(F2) The system Jacobian D with respect to an optimal solution of $\mathrm{D}(F)$ is not identically singular on $\mathbb{T} \times \Omega$ but singular at all consistent points.

(F3) D is identically singular.

Example DAEs of the failures are shown in the following.

Example 6.17. Consider the following DAE:

$$\begin{cases} {x_1}^2 + {x_2}^2 = 0, \\ \qquad\qquad 0 = 0. \end{cases} \tag{6.13}$$

The DAE (6.13) has a unique solution $x_1(t) = 0$ and $x_2(t) = 0$ for all $t \in \mathbb{R}$. However, since the bipartite graph $G(F)$ corresponding to (6.13) has no perfect matching, the structural preprocessing methods cannot be applied to (6.13) due to (F1). □

Remark 6.18. If we allow $x_1(t)$ and $x_2(t)$ to be complex-valued, solutions of the DAE (6.13) becomes $x_1(t) = c(t)$ and $x_2(t) = s(t)c(t)$ for an arbitrary $s : \mathbb{T} \to \{+\mathrm{i}, -\mathrm{i}\}$ and $c : \mathbb{T} \to \mathbb{C}$. This means that the solution set of the DAE (6.13) over $\mathbb{C}$ has the infinite degree of freedom. We conjecture that this happens for any DAEs over $\mathbb{C}$ with (F1). This is true for (possibly time-varying) linear DAEs due to Proposition 5.6.

Example 6.19. Consider the following DAE:

$$\begin{cases} {x_1}^2 = 0, \\ {x_2}^2 = 0. \end{cases} \tag{6.14}$$

The solution of (6.14) is the same as that of (6.13). The system Jacobian D with respect to a dual optimal solution $p = (0, 0)$ and $q = (0, 0)$ is

$$D = \begin{pmatrix} 2x_1 & 0 \\ 0 & 2x_2 \end{pmatrix},$$

which is not identically singular on $\Omega = \{(x_1, x_2) \mid x_1, x_2 \in \mathbb{R}\}$. However, D is singular at the unique consistent point $(0, 0)$ of (6.14). Hence (6.14) does not satisfy the validity condition of the Σ-method (and the MS-method) due to (F2). □

Example 6.20. Consider the following DAE

$$\begin{cases} \dot{x}_1 + \dot{x}_2 + x_3 = 0, \\ \dot{x}_1 + \dot{x}_2 \qquad = 0, \\ \qquad x_2 + \dot{x}_3 = 0. \end{cases} \tag{6.15}$$

The system Jacobian D corresponding to a dual optimal solution $p = (0, 0, 0)$ and $q = (1, 1, 1)$ is a singular constant matrix

$$D = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus the DAE (6.15) is in the case of (F3). □

Example 6.21. Consider the following index-1 nonlinear DAE

$$\begin{cases} F_1 : \dot{x}_1 \dot{x}_2 - 2\cos^2 t = 0, \\ F_2 : {\dot{x}_1}^2 {\dot{x}_2}^2 + x_1 + x_2 - 4\cos^4 t - 3\sin t - 2 = 0 \end{cases} \tag{6.16}$$

given in [94, Section 5.3]. On this DAE, it holds $\hat{d}(F) = 2$. The system Jacobian with respect to a dual optimal solution $p = (0, 0)$ and $q = (1, 1)$ is

$$D = \begin{pmatrix} \dot{x}_2 & \dot{x}_1 \\ 2\dot{x}_1 {\dot{x}_2}^2 & 2{\dot{x}_1}^2 \dot{x}_2 \end{pmatrix}, \tag{6.17}$$

which is identically singular. Hence (6.16) is also an example of (F3). □

The structural preprocessing methods indeed fail for the DAE (6.2) representing the

electrical network due to (F3). In this theses, we focus on (F3). The failure (F3) is attributed to the fact that the structural preprocessing methods use only combinatorial information and ignore numerical and symbolic information of DAEs assuming that nonzero entries in Jacobian matrices are generic. Then numerical or symbolic cancellations inherent in the DAEs make the system Jacobian identically singular.

6.4 DAE Modification via Combinatorial Relaxation

This section explains methods to modify a DAE into an equivalent DAE without (F3), i.e., the system Jacobian is not identically singular. All methods are based on the combinatorial relaxation framework.

6.4.1 Combinatorial Relaxation for Linear DAEs

Consider a linear DAE with constant coefficients $A(s)\tilde{x}(s) = \tilde{f}(s)$ given in (6.10). As seen in Example 6.9, the failure (F3) for this DAE is equivalent to the fact that A has a singular tight coefficient matrix. Combinatorial relaxation algorithms for A modify A into another polynomial matrix which is upper-tight (see Sections 1.2 and 4.2.1). Therefore, if the modification on A can be translated into a modification on the DAE which preserves the solution set, one can use the combinatorial relaxation algorithm for modifying DAEs to resolve (F3).

We consider the following three types of equivalent operations for linear DAEs: (i) multiplying a nonzero constant to an equation in the DAE, (ii) swapping two equations, and (iii) adding an equation in the DAE or its (possibly higher-order) derivative multiplied by a constant to another equation. The composition of these operations corresponds to a *unimodular transformation* in the form of

$$U(s)A(s)\tilde{x}(s) = U(s)\tilde{f}(s),$$

where $U(s)$ is a *unimodular matrix*, that is, an invertible matrix over $\mathbb{R}[s]$.

The original combinatorial relaxation algorithm by Murota [67] modifies $A(s)$ into $U(s)A(s)$ for some unimodular matrix $U(s)$. Therefore, by the above argument, Murota's algorithm can be used to modify linear DAEs to resolve (F3). Another combinatorial relaxation algorithm by Murota [66] uses a modification $A(s) \mapsto U(s)A(s)$ with *biproper matrix* $U(s)$, which is an invertible matrix over $\mathbb{R}[s^{-1}]$. Since multiplying by biproper matrices is an equivalent transformation of linear DAEs, the algorithm of [66] cannot be applied to modify DAEs. Wu et al. [103] indicated that first-order linear DAEs can be modified by the combinatorial relaxation algorithm of Iwata [47], which modifies $A(s) \coloneqq A_0 + A_1 s$ into $UA(s)V$ for some $U, V \in \mathrm{GL}_n(\mathbb{R})$. The right-multiplication of V is translated in DAEs as a change of variable coordinates.

6.4.2 Combinatorial Relaxation for Nonlinear DAEs

Tan et al. [94] observed that the modification method by the combinatorial relaxation can be generalized to nonlinear DAEs as follows:

Combinatorial Relaxation for DAEs

Phase 1d. Compute an optimal solution (p, q) of $\mathrm{D}(F)$. If $\mathrm{D}(F)$ has no optimal solution, the algorithm terminates with failure.

Phase 2d. If the system Jacobian D with respect to (p, q) is not identically singular, return the DAE $F = 0$ and halt.

Phase 3d. Modify the DAE $F = 0$ into an equivalent DAE $\bar{F} = 0$ such that $\hat{d}(\bar{F}) \leq \hat{d}(F) - 1$. Go back to Phase 1d.

Since $\mathrm{D}(F)$ has an optimal solution if and only if $\hat{d}(F) \geq 0$, the above process ends in at most $\hat{d}(F) \leq \ell n$ iterations. Therefore, given a DAE with (F3), the combinatorial relaxation method returns an equivalent DAE without (F3) (or with (F1) if the method has failed in Phase 1d; see Remark 6.18).

As an equivalent transformation for nonlinear DAEs in Phase 3d, the *linear combination method* (LC-method) of Tan et al. [94] replaces an equation in the DAE with a linear combination of (differentiations of) other equations. We review the LC-method as follows.

Suppose that we have a DAE (1.4) and its dual optimal solution (p, q) such that the system Jacobian D with respect to (p, q) is identically singular. First, we find a nonzero vector $u(t, x, \dot{x}, \dots) = (u_i)_{i \in R}$ in the cokernel of D, namely, u is a row vector such that uD is identically zero. Let $\operatorname{supp} u$ denote the support of u, i.e.,

$$\operatorname{supp} u \coloneqq \{i \in R \mid u_i \text{ is not identically zero}\}.$$

Take $r \in \operatorname{supp} u$ such that $p_r \leq p_i$ for all $i \in \operatorname{supp} u$ and put $I \coloneqq \operatorname{supp} u \setminus \{r\}$. Then we replace the rth equation $F_r = 0$ of the DAE by $\bar{F}_r^{\mathrm{LC}} = 0$, where

$$\bar{F}_r^{\mathrm{LC}} \coloneqq u_r F_r + \sum_{i \in I} u_i F_i^{(p_i - p_r)}.$$

It is shown that this modification decreases the value of $\hat{\delta}$ if

$$\sigma(u_i, x_j) < q_j - p_r \tag{6.18}$$

for all $i \in R$ and $j \in C$ [94, Theorem 4.1]. Intuitively, the condition (6.18) means that the highest-order derivatives appear linearly in DAEs. For (possibly time-varying) linear DAEs, (6.18) trivially holds since $\sigma(u_i, x_j) = -\infty$ for all i, j.

However, there still exist DAEs that violate the condition (6.18). Indeed, the DAE (6.16) is such an example as shown in [94, Section 5.3]. While [94] also presents another modification method called the *expression substitution method* (ES-method), it is also inapplicable to (6.16).

Chapter 7

Structural Modification for Linear DAEs with Mixed Matrices

Linear DAEs arising from dynamical systems are naturally modeled by means of *mixed matrices*, which distinguish between accurate constants and algebraically independent parameters. This chapter presents a combinatorial-relaxation based modification algorithm for a linear DAE (6.10) such that $A(s)$ is a mixed polynomial matrix. For such DAEs, we need to carefully design a modification algorithm avoiding arithmetic operations on the parameters.

In Section 7.1, we introduce an overview of mixed matrix theory. Then Section 7.2 presents our modification algorithm. Section 7.3 presents an improve algorithm for DAEs with *dimensionally consistent* mixed polynomial matrices. The dimensional consistency is a mathematical assumption on mixed matrices reflecting the principle of dimensional homogeneity in physical systems. Section 7.4 illustrates two examples. Section 7.5 shows results of numerical experiments. Finally, in Section 7.6, we discuss an application of the presented algorithms to nonlinear DAEs.

7.1 DAEs with Mixed Matrices

Mixed matrices and mixed polynomial matrices are mathematical tools introduced by Murota–Iri [72] for faithful model description in structural approach to systems analysis. Based on matroid theory, efficient algorithms are provided to compute the rank of mixed matrices and degree of minors of mixed polynomial matrices.

7.1.1 Mixed Matrices and Mixed Polynomial Matrices

Let L be a field and K a subfield of L. A typical setting in the context of DAEs is $K = \mathbb{Q}$ and L is the extension field of $\mathbb{Q}$ obtained by adjoining the set of independent physical parameters. A matrix T over L is said to be *generic* if the set of nonzero entries of T is

algebraically independent over K. A *mixed matrix* with respect to $L \,/\, K$ is a matrix in the form of $Q+T$, where Q is a matrix over K and T is a generic matrix. A mixed matrix $A = Q + T$ is called a *layered mixed matrix* (or *LM-matrix*) if there exists a bipartition $\{R_Q, R_T\}$ of $\mathrm{Row}(A)$ such that all nonzero entries of Q and T are in rows R_Q and R_T, respectively. An LM-matrix can is expressed as $A = \begin{pmatrix} Q \\ T \end{pmatrix}$.

A polynomial matrix $A(s) = \sum_{d=0}^{\ell} A_d s^d$ over L is called a *mixed polynomial matrix* if A_d is expressed as $A_d = Q_d + T_d$ with Q_d and T_d satisfying the following:

(MP-Q) Each Q_d is a matrix over K.

(MP-T) The set of nonzero entries of $T_0, \ldots, T_\ell$ is algebraically independent over K.

A *layered mixed polynomial matrix* (*LM-polynomial matrix*) is a mixed polynomial matrix such that nonzero rows of $Q(s) = \sum_{d=0}^{\ell} Q_d s^d$ and $T(s) = \sum_{d=0}^{\ell} T_d s^d$ are disjoint. An LM-polynomial matrix is expressed as $A(s) = \begin{pmatrix} Q(s) \\ T(s) \end{pmatrix}$.

Example 7.1. Consider the linear DAE (6.2) representing the electrical network illustrated in Figure 6.2. The Laplace transform of (6.2) is given by

$$\left(\begin{array}{ccccc|ccccc}
-1 & & & -1 & 1 & & & & & \\
 & 1 & 1 & 1 & -1 & & & & & \\
\hline
 & & & & & 1 & & 1 & & -1 \\
 & & & & & -1 & -1 & & 1 & \\
 & & & & & & 1 & -1 & & \\
\hline
R_1 & & & & & -1 & & & & \\
 & R_2 & & & & & -1 & & & \\
 & & sL & & & & & -1 & & \\
 & & & -1 & & & & & sC & \\
 & & & & & & & & & 1
\end{array}\right)
\begin{pmatrix} \tilde{\xi}_1(s) \\ \tilde{\xi}_2(s) \\ \tilde{\xi}_3(s) \\ \tilde{\xi}_4(s) \\ \tilde{\xi}_5(s) \\ \tilde{\eta}_1(s) \\ \tilde{\eta}_2(s) \\ \tilde{\eta}_3(s) \\ \tilde{\eta}_4(s) \\ \tilde{\eta}_5(s) \end{pmatrix}
=
\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ \tilde{V}(s) \end{pmatrix}. \tag{7.1}$$

Here, $\tilde{x} = (\tilde{\xi}_1, \ldots, \tilde{\xi}_5, \tilde{\eta}_1, \ldots, \tilde{\eta}_5)^\top$ is the Laplace transform of the vector $(\xi_1, \ldots, \xi_5, \eta_1, \ldots, \eta_5)^\top$ of variables and $\tilde{V}(s)$ is the Laplace transform of $V(t)$ (we assumed that all state variables and their derivatives were equal to 0 at $t = 0$ for simplicity). The coefficient matrix in (7.1) is naturally regarded as a mixed polynomial matrix with independent parameters R_1, R_2, L, and C since values of the parameters are supposed to be inaccurate. □

7.1.2 Rank of LM-matrices

Let $A = \begin{pmatrix} Q \\ T \end{pmatrix}$ be an LM-matrix. If A has no accurate constants, i.e., A is a generic matrix T, it holds that $\operatorname{rank} T = \text{t-rank}\, T$ from the independence of nonzero entries. From this equality, we can compute $\operatorname{rank} T$ by solving a maximum matching problem on the associated bipartite graph $G(T)$. For general LM-matrices, the following holds from the generalized Laplace expansion.

Proposition 7.2 ([73, Theorem 3.1])**.** *For an LM-matrix* $A = \begin{pmatrix} Q \\ T \end{pmatrix}$ *with* $R_Q = \mathrm{Row}(Q)$, $R_T = \mathrm{Row}(T)$ *and* $C = \mathrm{Col}(A)$*, the following rank identity holds:*

$$\operatorname{rank} A = \max\{\operatorname{rank} Q[R_Q, J] + \text{t-rank}\, T[R_T, C \setminus J] \mid J \subseteq C\}. \tag{7.2}$$

The problem of maximizing the right-hand side of (7.2) can be reduced to an *independent matching problem* on a linear matroid (Example 3.5) and a free matroid (Example 3.6), which is equivalent to the matroid intersection problem on a linear matroid and a transversal matroid (Example 3.7); see [71, Section 4.2] for detail. The following identity is obtained from Theorem 3.8.

Proposition 7.3 ([73, Theorem 3.1])**.** *For an LM-matrix* $A = \begin{pmatrix} Q \\ T \end{pmatrix}$ *with* $R_Q = \mathrm{Row}(Q)$, $R_T = \mathrm{Row}(T)$ *and* $C = \mathrm{Col}(A)$*, the following rank identity holds:*

$$\operatorname{rank} A = \min\{\operatorname{rank} Q[R_Q, J] + \text{t-rank}\, T[R_T, J] + |C \setminus J| \mid J \subseteq C\}. \tag{7.3}$$

Similarly, we give the following term-rank identity for LM-matrices, which will be used later in the proof of Lemma 7.12.

Proposition 7.4. *For an LM-matrix* $A = \begin{pmatrix} Q \\ T \end{pmatrix}$ *with* $R_Q = \mathrm{Row}(Q)$, $R_T = \mathrm{Row}(T)$ *and* $C = \mathrm{Col}(A)$*, the following term-rank identity holds:*

$$\text{t-rank}\, A = \min\{\text{t-rank}\, Q[R_Q, J] + \text{t-rank}\, T[R_T, J] + |C \setminus J| \mid J \subseteq C\}.$$

Proof. This immediately follows from the well-known rank formula of a union matroid [24] and the fact that the union of transversal matroids is also a transversal matroid [78, Corollary 11.3.8]. □

7.1.3 Dimensional Consistency

The principle of dimensional homogeneity claims that any equation describing a physical phenomenon must be consistent with respect to physical dimensions. In order to reflect the dimensional consistency in conservation laws of dynamical systems, Murota [63] introduced a class of mixed polynomial matrices $A(s) = Q(s) + T(s)$ that satisfy the following condition.

(MP-DC) $Q(s)$ is written as

$$Q(s) = \operatorname{diag}(s^{-\lambda_1}, \ldots, s^{-\lambda_m}) Q(1) \operatorname{diag}(s^{\mu_1}, \ldots, s^{\mu_n}) \tag{7.4}$$

for some integers $\lambda_1, \ldots, \lambda_m$ and $\mu_1, \ldots, \mu_n$.

A mixed polynomial matrix satisfying 7.1.3 is said to be *dimensionally consistent*. We abbreviate a dimensionally consistent mixed polynomial matrix and a dimensionally consis-

tent LM-polynomial matrix to a *DCM-polynomial matrix* and a *DCLM-polynomial matrix*, respectively.

Example 7.5. Let $A(s) = Q(s) + T(s)$ be the coefficient matrix of the DAE (7.1). Since $Q(s)$ is a constant matrix, $A(s)$ is a DCM-polynomial matrix with all λ_i and μ_j being 0. Note that (λ, μ) is not uniquely determined; the values

$$\lambda = (0, 0, -3, -3, -3, -3, -3, -3, 0, -3), \quad \mu = (0, 0, 0, 0, 0, -3, -3, -3, -3, -3) \tag{7.5}$$

also satisfy (7.4). □

The condition 7.1.3 can be "derived" from physical observations as follows. Suppose that a DAE $A(s)\tilde{x}(s) = \tilde{f}(s)$ arises from a dynamical system and the ith equation and the jth variable have physical dimensions X_i and Y_j, respectively. For example, in the DAE (7.1), the first, second, and ninth equations have the dimension of current and others have the dimension of voltage. Similarly, the first five variables $\tilde{\xi}_1, \ldots, \tilde{\xi}_5$ of (7.1) have the dimension of current and the last five variables $\tilde{\eta}_1, \ldots, \tilde{\eta}_5$ have the dimension of voltage. Then the dimension of each nonzero entry $A_{i,j}(s)$ of $A(s)$ must be $X_i Y_j^{-1}$ according to the principle of dimensional homogeneity. An important physical observation here is that all the nonzero coefficients of entries in $Q(s)$ are naturally regarded as dimensionless because they typically represent coefficients of conservation laws. In addition, since the indeterminate s corresponds to the time derivative, its dimension is the inverse T^{-1} of the dimension T of time. Thus if $Q_{i,j}(s) \neq 0$, then $Q_{i,j}(s)$ must be a monomial $Q_{i,j}(1)s^{d_{i,j}}$ of dimension $\mathrm{T}^{-d_{i,j}}$ with $d_{i,j} = \deg Q_{i,j}(s)$. Let $\lambda_i, \mu_j \in \mathbb{Q}$ such that X_i and Y_j are decomposed as $X_i = \mathrm{T}^{\lambda_i} X_i'$ and $Y_j = \mathrm{T}^{\mu_j} Y_j'$, where X_i' and Y_j' are physical dimensions that are not relevant to T in a using system of measurement. Now it holds $\mathrm{T}^{-d_{i,j}} = X_i Y_j^{-1} = \mathrm{T}^{\lambda_i - \mu_j} X_i' Y_j'^{-1}$ for $i \in R$ and $j \in C$ with $Q_{i,j}(s) \neq 0$. This implies $d_{i,j} = -\lambda_i + \mu_j$ and thus we have $Q_{i,j}(s) = Q_{i,j}(1)s^{-\lambda_i + \mu_j}$ for all $i \in R$ and $j \in C$. This is equivalent to 7.1.3 if every λ_i and μ_j are integral. Even if not, we can take integral (λ', μ') satisfying (7.4) [71, Theorem 2.2.35(2)]. See [71, Section 3] for more detail.

As described above, λ_i and μ_j can be taken as the exponents of T in the physical dimensions of the ith equation and the jth variable (if they are integral). In fact, the value (7.5) is taken from the DAE (7.1) in this way as the dimension of voltage is expressed as $\mathsf{L}^2\mathsf{T}^{-3}\mathsf{M}\mathsf{I}^{-1}$ by the SI base units, where L, M, and I are dimensions of length, mass, and current, respectively.

7.2 Algorithm Description

7.2.1 Overview

Consider a linear DAE (6.10) such that $A(s)$ is a nonsingular mixed polynomial matrix. As described in Section 6.4.1, our goal is to find a unimodular matrix $U(s)$ such that

$\bar{A}(s) \coloneqq U(s)A(s)$ is upper-tight. We emphasize again that the combinatorial relaxation algorithm by Iwata–Takamatsu [46] for mixed polynomial matrices cannot be used as a DAE modification because their algorithm modify matrices by biproper transformations.

Our first step is to convert a given DAE into another DAE whose coefficient matrix $A(s)$ is an LM-polynomial matrix expressed as $A(s) = \binom{Q(s)}{T(s)}$. Then we can transform $A(s)$ to

$$\bar{A}(s) \coloneqq \begin{pmatrix} U_Q(s) & O \\ O & I_{m_T} \end{pmatrix} \begin{pmatrix} Q(s) \\ T(s) \end{pmatrix}, \tag{7.6}$$

where $U_Q(s)$ is a unimodular matrix. Note that we are allowed to perform row operations only on $Q(s)$ even for an LM-polynomial matrix $A(s) = \binom{Q(s)}{T(s)}$, and thus we cannot always reduce the index to 1 only by row operations on $Q(s)$. We describe this conversion process from mixed polynomial matrices into LM-polynomial matrices in Section 7.2.2.

We then apply the combinatorial relaxation algorithm to $A(s)$ in accordance with the phases given in Sections 1.2 and 4.2.1. First, we obtain a nonnegative dual optimal solution (p, q) whose each entry is bounded by ℓn; we describe an algorithm to obtain such (p, q) in Section 7.2.3. In Phase 2, we check the upper-tightness of $A(s)$ by checking the nonsingularity of the tight coefficient matrix by Proposition 4.7. Since the tight coefficient matrix of $A(s)$ is an LM-matrix, we can compute its rank by solving an independent matching problem [73]. The matrix modification and an update procedure of (p, q) in Phase 3 are explained in Sections 7.2.4 and 7.2.5, respectively. Finally, Section 7.2.6 analyzes the time complexity of our algorithm.

7.2.2 Reduction to LM-polynomial Matrices

We first convert the DAE (6.10) of size n with a mixed polynomial coefficient matrix $A(s) = Q(s) + T(s)$ into the following augmented DAE

$$\begin{pmatrix} I_n & Q(s) \\ -D & DT(s) \end{pmatrix} \begin{pmatrix} \tilde{y}(s) \\ \tilde{z}(s) \end{pmatrix} = \begin{pmatrix} \tilde{f}(s) \\ 0 \end{pmatrix}, \tag{7.7}$$

where D is the diagonal matrix whose diagonal entries are independent parameters $\tau_1, \dots, \tau_n$. Note that the coefficient matrix of the augmented DAE (7.7) can be regarded as an LM-polynomial matrix as the set of nonzero coefficients of entries in $-D$ and $DT(s)$ is algebraically independent.

Proposition 7.6. *Let* $\binom{\tilde{y}(s)}{\tilde{z}(s)}$ *be a solution of the DAE* (7.7)*. Then* $\tilde{z}(s)$ *is a solution of the DAE* (6.10)*.*

Proof. By left-multiplying both sides of (7.7) by a nonsingular constant matrix $\begin{pmatrix} I_n & O \\ I_n & D^{-1} \end{pmatrix}$,

we obtain

$$\begin{pmatrix} I_n & Q(s) \\ O & A(s) \end{pmatrix} \begin{pmatrix} \tilde{y}(s) \\ \tilde{z}(s) \end{pmatrix} = \begin{pmatrix} \tilde{f}(s) \\ \tilde{f}(s) \end{pmatrix}.$$

Thus it holds $A(s)\tilde{z}(s) = \tilde{f}(s)$, which implies that $\tilde{z}(s)$ is a solution of the DAE (6.10). □

After preprocessing, we need to fill independent parameters by real numbers to start numerical methods. Indeed, we can substitute 1 for each diagonal entry τ_i of D, i.e., $D = I_n$. To explain this fact, let

$$B(s) \coloneqq \begin{pmatrix} Q_1(s) & Q_2(s) \\ -D & DT(s) \end{pmatrix} \tag{7.8}$$

be the coefficient matrix of a DAE that our algorithm returns for the augmented DAE (7.7), where $Q_1(s)$ and $Q_2(s)$ are some polynomial matrices. By substituting the identity matrix to D, we obtain

$$\bar{B}(s) \coloneqq \begin{pmatrix} Q_1(s) & Q_2(s) \\ -I_n & T(s) \end{pmatrix}. \tag{7.9}$$

Though $\bar{B}(s)$ is no longer an LM-polynomial matrix, the following lemma guarantees the upper-tightness of $\bar{B}(s)$.

Lemma 7.7. *Let $Q_1(s)$, $Q_2(s)$, and $T(s)$ are polynomial matrices and D a nonsingular diagonal matrix. Then $B(s)$ in (7.8) is upper-tight if and only if $\bar{B}(s)$ in (7.9) is upper-tight.*

Proof. Using $P \coloneqq \begin{pmatrix} I_n & O \\ O & D^{-1} \end{pmatrix}$, we have $\bar{B}(s) = PB(s)$. Since P is a nonsingular constant matrix, $d(B) = d(\bar{B})$ holds. In addition, since P is nonsingular, diagonal, and constant, the row transformation by P does not change the bipartite graph $G(B)$ and its edge weight c_e associated with $B(s)$. This fact implies that $\hat{d}(B) = \hat{d}(\bar{B})$. Thus the upper-tightness of $B(s)$ and $\bar{B}(s)$ are equivalent. □

From this lemma, we can "forget" the existence of D in the augmented DAE (7.7). That is, to modify the DAE (6.10) with $A(s) = Q(s) + T(s)$, it suffices to apply our algorithm to a DAE

$$\begin{pmatrix} I_n & Q(s) \\ -I_n & T(s) \end{pmatrix} \begin{pmatrix} \tilde{y}(s) \\ \tilde{z}(s) \end{pmatrix} = \begin{pmatrix} \tilde{f}(s) \\ 0 \end{pmatrix} \tag{7.10}$$

as if the set of nonzero coefficients of entries in $\begin{pmatrix} -I_n & T(s) \end{pmatrix}$ were independent.

Example 7.8. Consider the index-2 DAE

$$\begin{pmatrix} 1 & s+\alpha_1 \\ -1 & -s+\alpha_2 \end{pmatrix} \begin{pmatrix} \tilde{x}_1(s) \\ \tilde{x}_2(s) \end{pmatrix} = \begin{pmatrix} \tilde{f}_1(s) \\ \tilde{f}_2(s) \end{pmatrix}, \tag{7.11}$$

where α_1 and α_2 are independent parameters. Following (7.10), we convert this DAE into

$$\left(\begin{array}{cc|cc} 1 & & 1 & s \\ & 1 & -1 & -s \\ \hline -1 & & & \alpha_1 \\ & -1 & & \alpha_2 \end{array}\right) \begin{pmatrix} \tilde{y}_1(s) \\ \tilde{y}_2(s) \\ \tilde{z}_1(s) \\ \tilde{z}_2(s) \end{pmatrix} = \begin{pmatrix} \tilde{f}_1(s) \\ \tilde{f}_2(s) \\ 0 \\ 0 \end{pmatrix}. \tag{7.12}$$

Then we can obtain a solution $(\tilde{x}_1(s), \tilde{x}_2(s))$ of (7.11) by solving the augmented DAE (7.12). While the index of (7.12) is also 3, in general this conversion does not preserve the index of DAEs. □

7.2.3 Construction of Dual Optimal Solution

Let $A(s)$ be an $n \times n$ nonsingular LM-polynomial matrix with $R = \mathrm{Row}(A)$ and $C = \mathrm{Col}(A)$, and let ℓ be the maximum degree of an entry in $A(s)$. An optimal solution (p, q) of DA satisfying $0 \le p_i \le \ell n$ and $0 \le q_j \le \ell n$ for all $i \in R$ and $j \in C$ is constructed as follows.

Let $G(A) = (R \cup C, E(A))$ be the bipartite graph given in Section 4.2.1 and $c_e = c_{i,j}$ be the weight of an edge $e = \{i, j\} \in E(A)$. First, we obtain a maximum-weight perfect matching $M \subseteq E(A)$ in $G(A)$ by the Hungarian method [55]. Next, construct a residual graph $G_M = (W, E_M)$ with $W = R \cup C \cup \{r\}$ and $E_M = E^\circ \cup M \cup Z$, where r is a new vertex, $E^\circ = \{(j, i) \mid (i, j) \in E(A)\}$, and $Z = \{(r, i) \mid i \in R\}$. The arc length $\gamma : E_M \to \mathbb{Z}$ of G_M is defined by

$$\gamma(i,j) \coloneqq \begin{cases} -c_{j,i} & ((i,j) \in E^\circ), \\ +c_{i,j} & ((i,j) \in M), \\ 0 & ((i,j) \in Z) \end{cases}$$

for each $(i, j) \in E_M$.

Lemma 7.9. *For the residual graph G_M defined above, the following hold:*

(1) *All vertices are reachable from r.*

(2) *There is no negative-weight directed cycle with respect to γ.*

Proof. (1) Every vertex $i \in R$ is reachable from r through an edge $(r, i) \in Z$. In addition, since $G(A)$ has a perfect matching M, every vertex $j \in C$ is also reachable from r via $i \in R$ through edges $(r, i) \in Z$ and $(i, j) \in M \subseteq E(A)$.

(2) This immediately follows from an optimality criterion [53, Theorem 9.6] of the minimum cost flow problem. □

For $i, j \in W$ such that i is reachable to j, let $d(i, j)$ denote the length of a shortest path from i to j with respect to the arc length γ in G_M. Lemma 4.4 guarantees that $d(r, v)$ is defined for all $v \in W$. Using d, we define

$$p_i \coloneqq d(r, i) - \min_{i^* \in R} d(r, i^*), \tag{7.13}$$

$$q_j \coloneqq d(r, j) - \min_{i^* \in R} d(r, i^*) \tag{7.14}$$

for each $i \in R$ and $j \in C$.

The next lemma is easily shown in almost the same way as the case for $\ell = 1$ in [48, Lemma 2.2].

Lemma 7.10. *Let (p, q) be defined in* (7.13) *and* (7.14). *Then (p, q) is an optimal solution of* D(A) *satisfying $0 \leq p_i \leq \ell n$ for each $i \in R$ and $0 \leq q_j \leq \ell n$ for each $j \in C$.*

Proof. First, we prove that (p, q) is a feasible solution of DA. By the definition of (p, q), every $p_i\,(i \in R)$ and $q_j\,(j \in C)$ are clearly integer. For each $(i, j) \in E(A)$, it holds $d(r, i) \leq d(r, j) - c_{i,j}$. Thus

$$q_j - p_i = d(r, j) - d(r, i) \geq c_{i,j}$$

and this implies that (p, q) is a feasible solution of DA.

We second show the optimality of (p, q). For each $(i, j) \in M$, since $(i, j) \in E_M$ and $(j, i) \in E_M$, we obtain

$$q_j - p_i = d(r, j) - d(r, i) = c_{i,j}.$$

Thus it holds that

$$\sum_{j \in C} q_j - \sum_{i \in R} p_i = \sum_{j \in C} d(r, j) - \sum_{i \in R} d(r, i) = \sum_{(i,j) \in M} (d(r, j) - d(r, i)) = \sum_{(i,j) \in M} c_{i,j}$$

which implies that (p, q) is optimal to D(A).

Finally, we give the lower and upper bounds on p_i and q_j. The non-negativity of p_i clearly follows from the definition of p_i. In addition, since $G(A)$ has a perfect matching, each $j \in C$ is incident to at least one vertex $i \in R$ on $G(A)$. Thus we obtain $q_j \geq p_i + c_{i,j} \geq 0$ by $p_i, c_{i,j} \geq 0$. Let $i^* \in R$ denote a vertex such that $d(r, i^*) \leq d(r, i)$ for all $i \in R$. Fix $j \in C$. Let $P_j \subseteq E_M$ and $P_{i^*} \subseteq E_M$ be shortest paths from r to j and i^*, respectively. Let $v \in W$ be the last common vertex in P_j and P_{i^*}. Then it holds $q_j = d(r, j) - d(r, i^*) = d(v, j) - d(v, i^*)$. Let $Q_j \subseteq P_j$ and $Q_{i^*} \subseteq P_{i^*}$ denote subpaths from v to j and i^*, respectively. Note that $d(v, j)$ is at most ℓ times the number of edges

in $E(A)$ on Q_j, whereas $-d(v, i^*)$ is at most ℓ times the number of edges in M° on Q_{i^*}. The sum of these upper bounds is at most ℓn since Q_{i^*} and Q_j do not share the same vertex besides v. Thus $q_j \leq \ell n$ holds for each $j \in C$. In addition, for each $i \in R$, we have $p_i \leq q_j - c_{i,j} \leq q_j \leq \ell n$, where $j \in C$ is incident to i in M. □

Example 7.11. Consider the coefficient matrix

$$A(s) = \begin{pmatrix} 1 & & 1 & s \\ & 1 & -1 & -s \\ -1 & & & \alpha_1 \\ & -1 & & \alpha_2 \end{pmatrix} \tag{7.15}$$

in the DAE (7.12). An optimal solution of the assignment problem $\mathrm{P}(A)$ is given by

$$M = \{(1,3), (2,4), (3,1), (4,2)\}$$

with optimal value $\hat{d}(A) = 1$. According to (7.13) and (7.14), a dual optimal solution (p, q) is calculated as $p = (0, 0, 0, 0)$ and $q = (0, 0, 0, 1)$. □

7.2.4 Matrix Modification

Let $A(s) = \binom{Q(s)}{T(s)}$ be an $n \times n$ nonsingular LM-polynomial matrix that is not upper-tight. Let $A^\# = \binom{Q^\#}{T^\#}$ be the tight coefficient matrix with respect to an optimal solution (p, q) of $\mathrm{D}(A)$. Without loss of generality, we assume that $\mathrm{Row}(Q) = R_Q = [m_Q]$ and $p_1 \leq \cdots \leq p_{m_Q}$, where $m_Q \coloneqq |R_Q|$.

Recall the rank identity (7.3). Let $J^* \subseteq C$ be a column subset that minimizes the right-hand side of the identity for $A^\#$, i.e., it holds

$$\operatorname{rank} A^\# = \operatorname{rank} Q^\#[R_Q, J^*] + \text{t-rank}\, T^\#[R_T, J^*] + |C \setminus J^*|. \tag{7.16}$$

Such J^* is called a *minimizer* of (7.3). By a row transformation of $Q^\#$, we obtain a matrix $\bar{Q}^\# = UQ^\#$ such that

$$\operatorname{rank} \bar{Q}^\#[R_Q, J^*] = \text{t-rank}\, \bar{Q}^\#[R_Q, J^*]. \tag{7.17}$$

In particular, this transformation can be accomplished only by operations of adding a scalar multiple of a row $i \in R_Q$ to another row $j \in R_Q$ with $p_i > p_j$. Then the matrix U is upper-triangular due to the order of rows in R_Q. This is the *forward elimination* on $\bar{Q}^\#[R_Q, J^*]$ with the order of the rows reversed. Consider

$$U_Q(s) = \operatorname{diag}(s^{-p_1}, \ldots, s^{-p_{m_Q}}) U \operatorname{diag}(s^{p_1}, \ldots, s^{p_{m_Q}}), \tag{7.18}$$

where $\operatorname{diag}(a_1, \ldots, a_n)$ denotes a diagonal matrix with diagonal entries $a_1, \ldots, a_n$. Note

that each entry in $U_Q(s)$ is a polynomial because U is upper-triangular. In addition, since $\det U_Q(s) = \det U$ is a nonzero constant, $U_Q(s)$ is unimodular.

Recall that $D(s^p) = \mathrm{diag}(s^{p_1}, \ldots, s^{p_n})$ and $D(s^q) = \mathrm{diag}(s^{q_1}, \ldots, s^{q_n})$. Using $U_Q(s)$, we update $A(s)$ to $\bar{A}(s)$ as in (7.6):

$$\bar{A}(s) = \begin{pmatrix} U_Q(s) & O \\ O & I_{m_T} \end{pmatrix} A(s) = D(s^{-p}) \begin{pmatrix} U & O \\ O & I_{m_T} \end{pmatrix} D(s^p) A(s), \tag{7.19}$$

where $m_T \coloneqq |\mathrm{Row}(T)|$.

Lemma 7.12. *Let $A(s) = \begin{pmatrix} Q(s) \\ T(s) \end{pmatrix}$ be an $n \times n$ nonsingular LM-polynomial matrix that is not upper-tight, and $A^{\#} = \begin{pmatrix} Q^{\#} \\ T^{\#} \end{pmatrix}$ the tight coefficient matrix with respect to an optimal solution (p, q) of $\mathrm{D}(A)$. Then for the LM-polynomial matrix $\bar{A}(s)$ defined in (7.19), the value (p, q) is feasible on $\mathrm{D}(\bar{A})$ but not optimal.*

Proof. Consider a rational function matrix

$$H(s) \coloneqq D(s^p) \bar{A}(s) D(s^{-q}). \tag{7.20}$$

For each $i \in R$ and $j \in C$, it holds that $\deg H_{i,j}(s) = \bar{c}_{i,j} + p_i - q_j$, where $\bar{c}_{i,j} = \deg \bar{A}_{i,j}(s)$. By substituting (7.19) into (7.20), we obtain

$$H(s) = \begin{pmatrix} U & O \\ O & I_{m_T} \end{pmatrix} D(s^p) A(s) D(s^{-q}) = \begin{pmatrix} U & O \\ O & I_{m_T} \end{pmatrix} (A^{\#} + A^{\infty}(s)),$$

where $A^{\infty}(s)$ is a matrix whose entries are polynomials in s^{-1} without constant terms. Hence for each $i \in R$ and $j \in C$, it holds $\deg H_{i,j}(s) \leq 0$, which implies $\bar{c}_{i,j} \leq q_j - p_i$. Therefore (p, q) is feasible on $\mathrm{D}(\bar{A})$.

Next, we show that (p, q) is not optimal to $\mathrm{D}(\bar{A})$. From (7.19), the tight coefficient matrix $\bar{A}^{\#}$ of $\bar{A}(s)$ with respect to (p, q) is

$$\bar{A}^{\#} = \begin{pmatrix} U & O \\ O & I_{m_T} \end{pmatrix} A^{\#} = \begin{pmatrix} \bar{Q}^{\#} \\ T^{\#} \end{pmatrix}, \tag{7.21}$$

where $\bar{Q}^{\#} = UQ^{\#}$. From Proposition 7.4 and (7.17), it holds

$$\begin{aligned} \text{t-rank}\, \bar{A}^{\#} &= \min\{\text{t-rank}\, \bar{Q}^{\#}[R_Q, J] + \text{t-rank}\, T^{\#}[R_T, J] + |C \setminus J| \mid J \subseteq C\} \\ &\leq \text{t-rank}\, \bar{Q}^{\#}[R_Q, J^*] + \text{t-rank}\, T^{\#}[R_T, J^*] + |C \setminus J^*| \\ &= \mathrm{rank}\, \bar{Q}^{\#}[R_Q, J^*] + \text{t-rank}\, T^{\#}[R_T, J^*] + |C \setminus J^*|. \end{aligned}$$

Now since $Q^{\#}[R_Q, J^*]$ and $\bar{Q}^{\#}[R_Q, J^*] = UQ^{\#}[R_Q, J^*]$ have the same rank, we obtain

$$\text{t-rank}\, \bar{A}^{\#} \leq \mathrm{rank}\, Q^{\#}[R_Q, J^*] + \text{t-rank}\, T^{\#}[R_T, J^*] + |C \setminus J^*| = \mathrm{rank}\, A^{\#},$$

where the last equality comes from (7.16). In addition, since rank $\bar{A}^{\#} = \operatorname{rank} A^{\#}$ from (7.21), we have t-rank $\bar{A}^{\#} \leq \operatorname{rank} \bar{A}^{\#}$, which implies t-rank $\bar{A}^{\#} = \operatorname{rank} \bar{A}^{\#} = \operatorname{rank} A^{\#}$. Furthermore, since $A(s)$ is not upper-tight, we have $\operatorname{rank} A^{\#} < n$ by Proposition 6.7. Thus, t-rank $A^{\#} = \operatorname{rank} A^{\#} < n$ holds. It then follows from Proposition 6.7 again that (p, q) is not optimal to $\mathrm{D}(\bar{A})$. □

From Lemma 7.12 and the unimodularity of $U_Q(s)$, we obtain the following.

Corollary 7.13. *Let* $A(s) = \begin{pmatrix} Q(s) \\ T(s) \end{pmatrix}$ *be an* $n \times n$ *nonsingular LM-polynomial matrix that is not upper-tight, and* $\bar{A}(s)$ *the LM-polynomial matrix defined in* (7.19). *Then* $\hat{d}(\bar{A}) \leq \hat{d}(A) - 1$ *and* $d(A) = d(\bar{A})$ *hold.*

Example 7.14. Consider the LM-polynomial matrix (7.15) again. The tight coefficient matrix $A^{\#}$ with respect to $p = (0, 0, 0, 0)$ and $q = (0, 0, 0, 1)$ is

$$A^{\#} = \begin{pmatrix} Q^{\#} \\ T^{\#} \end{pmatrix} = \begin{pmatrix} 1 & & 1 & 1 \\ & 1 & -1 & -1 \\ -1 & & & \\ & -1 & & \end{pmatrix},$$

where the row sets R_Q of $Q^{\#}$ and R_T of $T^{\#}$ correspond to the first and last two rows in $A^{\#}$, respectively. A minimizer $J^* \subseteq C$ is the set of the right two columns as follows:

$$A^{\#} = \begin{array}{c} \overbrace{}^{C \setminus J^*} \quad \overbrace{}^{J^*} \\ \begin{pmatrix} 1 & & 1 & 1 \\ & 1 & -1 & -1 \\ -1 & & & \\ & -1 & & \end{pmatrix} \end{array} \begin{array}{l} \left.\vphantom{\begin{matrix} 1 \\ 1 \end{matrix}}\right\} R_Q \\ \left.\vphantom{\begin{matrix} 1 \\ 1 \end{matrix}}\right\} R_T \end{array}.$$

Then the rank of $A^{\#}$ is calculated by (7.16) as $Q^{\#}[R_Q, J^*] + T^{\#}[R_T, J^*] + |C \setminus J^*| = 1 + 0 + 2 = 3$. Since $A^{\#}$ is not upper-tight, we need to modify $A(s)$. By performing Gaussian elimination on $Q^{\#}[R_Q, J^*] = \begin{pmatrix} 1 & 1 \\ -1 & -1 \end{pmatrix}$, we obtain

$$\bar{Q}^{\#}[R_Q, J^*] = UQ^{\#}[R_Q, J^*] = \begin{pmatrix} & \\ -1 & -1 \end{pmatrix},$$

where $U = \begin{pmatrix} 1 & 1 \\ & 1 \end{pmatrix}$. The unimodular matrix $U_Q(s)$ defined by (7.18) coincides with U

since all p_i are 0. According to (7.19), we update $A(s)$ into

$$\bar{A}(s) = \begin{pmatrix} U_Q(s) & O \\ O & I_{m_T} \end{pmatrix} A(s) \tag{7.22}$$
$$= \begin{pmatrix} 1 & 1 & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} \begin{pmatrix} 1 & & 1 & s \\ & 1 & -1 & -s \\ -1 & & & \alpha_1 \\ & -1 & & \alpha_2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & & \\ & 1 & -1 & -s \\ -1 & & & \alpha_1 \\ & -1 & & \alpha_2 \end{pmatrix}. \qquad \square$$

7.2.5 Dual Updates

Let (p, q) be a feasible solution of $\mathrm{D}(\bar{A})$. We obtain an optimal solution of $\mathrm{D}(\bar{A})$ by iterating the following procedure.

Let $\bar{A}^{\#}$ be the tight coefficient matrix of $\bar{A}(s)$ with respect to (p, q). First we check t-rank $\bar{A}^{\#} = n$ or not. If it is, (p, q) is an optimal solution of $\mathrm{D}(\bar{A})$ from Proposition 4.6 and we are done. Otherwise, we construct a feasible solution (p', q') of $\mathrm{D}(\bar{A})$ such that the difference of the objective values is negative. Let $G^{\#} \coloneqq G(A^{\#})$ be the associated bipartite graph with $A^{\#}$. Since (p, q) is not optimal, $G^{\#}$ has no perfect matching by Proposition 3.3. This means that $G^{\#}$ has a vertex cover $S \subseteq R \cup C$ with $|S| < n$ by Theorem 3.1. Using this S, we define (p', q') as follows:

$$p'_i \coloneqq \begin{cases} p_i & (i \in R \cap S) \\ p_i + 1 & (i \in R \setminus S) \end{cases}, \quad q'_j \coloneqq \begin{cases} q_j + 1 & (j \in C \cap S) \\ q_j & (j \in C \setminus S) \end{cases} \tag{7.23}$$

for $i \in R$ and $j \in C$. The following lemma is a version of Lemma 4.13.

Lemma 7.15. *Let (p, q) be a feasible but not optimal solution of $\mathrm{D}(\bar{A})$ and (p', q') defined in (7.23). Then $(p'q')$ is feasible to $\mathrm{D}(\bar{A})$ and the objective value of (p', q') is less than that of (p, q).*

We update (p, q) to (p', q'), and go back to the optimality checking. From Lemma 7.15, it is guaranteed that (p, q) becomes an optimal solution of $\mathrm{D}(\bar{A})$ by iterating the update process above.

Example 7.16. Consider the modified LM-polynomial matrix (7.22). The tight coefficient matrix $\bar{A}^{\#}$ of $\bar{A}(s)$ with respect to $p = (0, 0, 0, 0)$ and $q = (0, 0, 0, 1)$ is

$$A^{\#} = \begin{pmatrix} 1 & 1 & & \\ & 1 & -1 & -1 \\ -1 & & & \\ & -1 & & \end{pmatrix}.$$

Let S be the set of the first and second columns and the second row of $A^{\#}$. Then S is

a vertex cover of $G^{\#}$ with $|S| < 4$. Following (7.23), we update (p, q) to $p' = (1, 0, 1, 1)$ and $q' = (1, 1, 0, 1)$. We then go back to Phase 2 for $\bar{A}(s)$. It is indeed confirmed in the next iteration that $\bar{A}(s)$ is upper-tight, and thus the iteration ends at this point. We can obtain a low-index DAE by applying the MS-algorithm. □

7.2.6 Complexity Analysis

This section is devoted to complexity analysis. The dominating part in our algorithm is the matrix multiplications in (7.19).

Let $A(s)$ be an $n \times n$ nonsingular LM-polynomial matrix and $A^{\#}$ the tight coefficient matrix with respect to an optimal solution (p, q) of $\mathrm{D}(A)$. From the definition of $A^{\#}$, we can express $A(s)$ as

$$A(s) = D(s^{-p}) \left(A^{\#} + \sum_{k=1}^{K} s^{-k} V_k \right) D(s^{q}) \tag{7.24}$$

for some K matrices $V_1, V_2, \ldots, V_K$ with $V_K \neq O$. By (7.19) and (7.24), we have

$$\bar{A}(s) = D(s^{-q}) \begin{pmatrix} U & O \\ O & I_{m_T} \end{pmatrix} \left(A^{\#} + \sum_{k=1}^{K} s^{-k} V_k \right) D(s^{q}).$$

Therefore, we can compute $\bar{A}(s)$ by performing $K + 1$ constant matrix multiplications.

By $V_K \neq O$, there exist $i \in R$ and $j \in C$ such that the (i, j) entry in V_K is nonzero. Then the degree of the corresponding term in $A_{i,j}(s)$ is equal to $q_j - p_i - K$. Since $A_{i,j}(s)$ is a polynomial, we have $q_j - p_i - K \geq 0$, which implies $K \leq q_j - p_i \leq q_j$. The following lemma bounds p_i and q_j at any iteration of our algorithm.

Lemma 7.17. *During the algorithm, the values p_i and q_j are at most $2\ell n$ for $i \in R$ and $j \in C$, where ℓ is the maximum degree of an entry in $A(s)$.*

Proof. From Lemma 7.10, the initial values of p_i and q_j are bounded by ℓn. In every update of (p, q), the values p_i and q_j increase by at most one from the update rule (7.23). In addition, (p, q) is updated at most $\hat{d}(A) - d(A) \leq \ell n$ times because the objective value $\sum_{j \in C} q_j - \sum_{i \in R} p_i$ of the dual problem decreases by at least one in every update. Therefore, at any iteration of the algorithm, it holds $p_i, q_j \leq \ell n + \hat{d}A \leq \ell n + \ell n = 2\ell n$. □

The time complexity of our algorithm is as follows. Recall that ω denotes the exponent of the time complexity of matrix multiplication.

Theorem 7.18. *Let $A(s)$ be an $n \times n$ nonsingular LM-polynomial matrix and ℓ the maximum degree of an entry in $A(s)$. Then our algorithm runs in $\mathrm{O}(\ell^2 n^{\omega+2})$-time.*

Proof. Phase 1 can be done in $\mathrm{O}(n^3)$-time by the Hungarian method [55] and shortest path algorithms such as the Bellman–Ford algorithm. Consider the time complexity in every iteration of Phases 2 and 3. In Phase 2, the nonsingularity of the tight coefficient

matrix $A^{\#}$ can be checked via the rank identity (7.3). Thus an efficient way is to obtain a minimizer J^* of (7.3) before Phase 2, and then check the nonsingularity of $A^{\#}$ by (7.3). The minimizer J^* can be found from a residual graph constructed by an augmenting path type algorithm [73], which runs in $\mathrm{O}(n^3 \log n)$-time [18]. The computation of $\bar{A}(s)$ in Phase 3 can be done in $\mathrm{O}(Nn^\omega) = \mathrm{O}(\max_{j \in C} q_j n^\omega) = \mathrm{O}(\ell n^{\omega+1})$-time from Lemma 7.17, where (p, q) is a dual optimal solution of $\mathrm{D}(A)$ and N is in (7.24). In addition, since the number of iterations of Phases 2 and 3 is at most $\hat{d}(A) - d(A) \leq \ell n$, the running time in Phases 2 and 3 is $\mathrm{O}(\ell^2 n^{\omega+2})$. Finally, the updates of (p, q) run in $\mathrm{O}(\ell n^4)$-time: (p, q) is updated at most $\hat{d}(A) \leq \ell n$ times, and in every update, we can find a vertex cover in $\mathrm{O}(n^3)$-time by the Ford–Fulkerson algorithm. Thus the total running time is $\mathrm{O}(\ell^2 n^{\omega+2})$. □

7.3 Exploiting Dimensional Consistency

This section improves the matrix modification procedure in Phase 3 for DCLM-polynomial matrices preserving their dimensional consistency.

Let $A(s) = \begin{pmatrix} Q(s) \\ T(s) \end{pmatrix}$ be a DCLM-polynomial matrix with $R_Q = \mathrm{Row}(Q)$, $R_T = \mathrm{Row}(T)$ and $C = \mathrm{Col}(A)$. Let (p, q) be an optimal solution of $\mathrm{D}(A)$. For $k \in \mathbb{Z}$, let

$$R_k = \{i \in R_Q \mid p_i - \lambda_i = k\}, \quad C_k = \{j \in C \mid q_j - \mu_j = k\}. \tag{7.25}$$

If $Q_{i,j}(s) \neq 0$, then we have $c_{i,j} \leq q_j - p_i$ from the feasibility of (p, q) and $c_{i,j} = \mu_j - \lambda_i$ by (7.4). Hence $p_i - \lambda_i \leq q_j - \mu_j$ follows, which implies $i \in R_h$ if and only if $j \in C_k$ with $h \leq k$. Thus, it holds $Q(s)[R_h, C_k] = O$ for integers $h, k \in \mathbb{Z}$ with $h > k$. Namely, $Q(s)$ forms a block triangular matrix.

Let $A^{\#} = \begin{pmatrix} Q^{\#} \\ T^{\#} \end{pmatrix}$ denote the tight coefficient matrix with respect to (p, q). From the definition of the tight coefficient matrix, $Q^{\#}$ forms a block diagonal matrix as

$$Q^{\#} = \begin{array}{c} \\ \vdots \\ R_{-1} \\ R_0 \\ R_1 \\ R_2 \\ \vdots \end{array} \begin{array}{c} \cdots\cdots C_{-1} \quad C_0 \quad C_1 \quad C_2 \cdots\cdots \\ \begin{pmatrix} \ddots & & & & & & \\ & Q^{\#}_{-1} & & & & & \\ & & Q^{\#}_0 & & & & \\ & & & Q^{\#}_1 & & & \\ & & & & Q^{\#}_2 & & \\ & & & & & \ddots & \end{pmatrix} \end{array},$$

where $Q^{\#}_k = Q^{\#}[R_k, C_k]$ for $k \in \mathbb{Z}$.

Let $J^* \subseteq C$ be a minimizer of the rank identity (7.3) for $A^{\#}$. Sorting rows in ascending order of p, the matrix modification process described in Section 7.2.4 finds a nonsingular

upper-triangular matrix U such that

$$\operatorname{rank} UQ^{\#}[R_Q, J^*] = \text{t-rank}\, UQ^{\#}[R_Q, J^*]. \tag{7.26}$$

For a DCLM-polynomial matrix, supposing that rows in R_k are sorted in ascending order of p, we find a nonsingular upper-triangular matrix U_k such that

$$\operatorname{rank} U_k Q_k^{\#}[R_k, C_k \cap J^*] = \text{t-rank}\, U_k Q_k^{\#}[R_k, C_k \cap J^*]$$

for $k \in \mathbb{Z}$. Then $U = \operatorname{diag}(\ldots, U_{-1}, U_0, U_1, U_2, \ldots)$ satisfies (7.26).

For $k \in \mathbb{Z}$, let $P_k(s)$ be a diagonal polynomial matrix with $\operatorname{Row}(P_k) = \operatorname{Col}(P_k) = R_k$ whose (i,i) entry is s^{p_i} for each $i \in R_k$. Then we have

$$D(s^p) = \operatorname{diag}(\ldots, P_{-1}(s), P_0(s), P_1(s), P_2(s), \ldots).$$

Now the unimodular matrix $U_Q(s)$ defined in (7.18) can be written as

$$\begin{aligned} U_Q(s) &= D(s^{-p}) \operatorname{diag}(\ldots, U_{-1}, U_0, U_1, U_2, \ldots) D(s^p) \\ &= D(s^{-p}) \operatorname{diag}(\ldots, U_{-1}P_{-1}(s), U_0P_0(s), U_1P_1(s), U_2P_2(s), \ldots). \end{aligned} \tag{7.27}$$

Then we update $A(s)$ into $\bar{A}(s) = \begin{pmatrix} U_Q(s)Q(s) \\ T(s) \end{pmatrix}$ as written in (7.19).

Lemma 7.19. *Let $A(s) = \begin{pmatrix} Q(s) \\ T(s) \end{pmatrix}$ be an $n \times n$ DCLM-polynomial matrix. Then $\bar{A}(s) = \begin{pmatrix} U_Q(s)Q(s) \\ T(s) \end{pmatrix}$ is also dimensionally consistent.*

Proof. Let $\lambda_1, \ldots, \lambda_{m_Q}$ and $\mu_1, \ldots, \mu_n$ defined in (7.4) for $A(s)$, where $m_Q = |\operatorname{Row}(Q|)$. For $k \in \mathbb{Z}$, let R_k and C_k defined in (7.25), and let $\Lambda_k(s)$ denote a diagonal polynomial matrix with $\operatorname{Row}(\Lambda_k) = \operatorname{Col}(\Lambda_k) = R_k$ whose (i,i) entry is s^{λ_i} for each $i \in R_k$, and $D(s^\mu) = \operatorname{diag}(s^{\mu_1}, \ldots, s^{\mu_n})$. Then the condition (7.4) for dimensional consistency is written as

$$Q(s) = \operatorname{diag}(\ldots, \Lambda_{-1}^{-1}(s), \Lambda_0^{-1}(s), \Lambda_1^{-1}(s), \Lambda_2^{-1}(s), \ldots) Q(1) D(s^\mu). \tag{7.28}$$

Combining (7.27) and (7.28), we obtain

$$\begin{aligned} &U_Q(s)Q(s) \\ &= P^{-1}(s) \operatorname{diag}(\ldots, U_{-1}P_{-1}(s)\Lambda_{-1}^{-1}(s), U_0P_0(s)\Lambda_0^{-1}(s), U_1P_1(s)\Lambda_1^{-1}(s), \ldots) Q(1) D(s^\mu) \\ &= P^{-1}(s) \operatorname{diag}(\ldots, s^{-1}U_{-1}, U_0, sU_1, s^2U_2, \ldots) Q(1) D(s^\mu) \\ &= \operatorname{diag}(\ldots, s^{-1}P_{-1}^{-1}(s), P_0^{-1}(s), sP_1^{-1}(s), s^2P_2^{-1}(s), \ldots) UQ(1) D(s^\mu), \end{aligned} \tag{7.29}$$

where we used $P_k(s)\Lambda_k^{-1}(s) = s^k I$ for $k \in \mathbb{Z}$. From (7.29), $\bar{A}(s)$ is also dimensionally consistent. □

For a DCLM-polynomial matrix $A(s)$, we can compute $\bar{A}(s) = U(s)A(s)$ only by

one constant matrix multiplication $UQ(1)$ from (7.29), whereas a general LM-polynomial matrix needs $\mathrm{O}(\ell n)$ multiplications. This improves the total running time as follows.

Theorem 7.20. *Let $A(s)$ be an $n \times n$ nonsingular DCLM-polynomial matrix and ℓ the maximum degree of an entry in $A(s)$. Then our algorithm for $A(s)$ runs in $\mathrm{O}(\ell n^4 \log n)$-time.*

Proof. For each iteration of Phases 2 and 3, the computation of $\bar{A}(s)$ in Phase 3 can be done in $\mathrm{O}(n^\omega)$-time. The most expensive part is the nonsingularity checking for a tight coefficient matrix in Phase 2, which requires $\mathrm{O}(n^3 \log n)$-time [18, 73]. Since the number of iterations of Phases 2 and 3 is at most $\hat{d}(A) - d(A) \leq \ell n$, the running time of Phases 2 and 3 is $\mathrm{O}(\ell n^4 \log n)$. We can check that other processes run in $\mathrm{O}(\ell n^4 \log n)$-time as in the proof of Theorem 7.18. □

7.4 Examples

We give two examples below. The first example is a simple index-4 DAE and the second example is the DAE (7.1) representing the electrical network shown in Figure 6.2. Throughout the execution of our algorithm, it is emphasized that: (i) we only use combinatorial operations and numerical calculations over rational numbers (over integers in the following examples), and (ii) we do not reference values of physical quantities.

7.4.1 Example of High-index DAE

The first example is the following index-4 DAE

$$\begin{cases} \ddot{x}_1 - \dot{x}_1 + \ddot{x}_2 - \dot{x}_2 + x_4 = f_1(t), \\ \ddot{x}_1 + \ddot{x}_2 + x_3 = f_2(t), \\ \alpha_1 x_2 + \alpha_2 \ddot{x}_3 + \alpha_3 \dot{x}_4 = f_3(t), \\ \alpha_4 x_3 + \alpha_5 \dot{x}_4 = f_4(t) \end{cases} \tag{7.30}$$

with independent parameters $\alpha_1, \ldots, \alpha_5$ and smooth functions $f_1, \ldots, f_4$. The coefficient matrix $A(s) = \binom{Q(s)}{T(s)}$ corresponding to (7.30) is an LM-polynomial matrix given by

$$A(s) = \begin{pmatrix} s^2 - s & s^2 - s & & 1 \\ s^2 & s^2 & 1 & \\ & \alpha_1 & \alpha_2 s^2 & \alpha_3 s \\ & & \alpha_4 & \alpha_5 s \end{pmatrix}. \tag{7.31}$$

The row sets R_Q of $Q(s)$ and R_T of $T(s)$ correspond to the first and last two rows in $A(s)$, respectively. Since $d(A) = \deg(-\alpha_1\alpha_5 s^3 - \alpha_1\alpha_4 s^2 + \alpha_1\alpha_5 s^2) = 3$ and $\hat{d}(A) = 7$, the structural preprocessing methods are not applicable to the DAE. This fact will be verified

in our algorithm.

Let us apply our algorithm to (7.31). First, we find a dual optimal solution $p = (0,0,0,0)$ and $q = (2,2,2,1)$. The corresponding tight coefficient matrix $A^{\#} = \left(\begin{smallmatrix} Q^{\#} \\ T^{\#} \end{smallmatrix}\right)$ is

$$A^{\#} = \begin{pmatrix} 1 & 1 & & \\ 1 & 1 & & \\ & & \alpha_2 & \alpha_3 \\ & & & \alpha_5 \end{pmatrix}.$$

A minimizer J^* of (7.3) for $A^{\#}$ is a set of the first two columns as follows:

$$A^{\#} = \overbrace{\begin{pmatrix} 1 & 1 & & \\ 1 & 1 & & \\ & & \alpha_2 & \alpha_3 \\ & & & \alpha_5 \end{pmatrix}}^{J^*\quad C\setminus J^*} \begin{matrix} \left.\vphantom{\begin{matrix}1\\1\end{matrix}}\right\}R_Q \\ \left.\vphantom{\begin{matrix}1\\1\end{matrix}}\right\}R_T \end{matrix}.$$

Then we can check that $\operatorname{rank} A^{\#} = Q^{\#}[R_Q, J^*] + T^{\#}[R_T, J^*] + |C \setminus J^*| = 1 + 0 + 2 = 3 < 4$, which implies that $A(s)$ is not upper-tight. We convert $Q^{\#}[R_Q, J^*] = \left(\begin{smallmatrix} 1 & 1 \\ 1 & 1 \end{smallmatrix}\right)$ by the backward elimination into

$$\bar{Q}^{\#}[R_Q, J^*] = UQ^{\#}[R_Q, J^*] = \begin{pmatrix} & \\ 1 & 1 \end{pmatrix},$$

where $U = \left(\begin{smallmatrix} 1 & -1 \\ & 1 \end{smallmatrix}\right)$. Using $U_Q(s) = U$, the LM-polynomial matrix $A(s)$ is modified to

$$A'(s) = \begin{pmatrix} 1 & -1 & & \\ & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} A(s) = \begin{pmatrix} -s & -s & -1 & 1 \\ s^2 & s^2 & 1 & \\ & \alpha_1 & \alpha_2 s^2 & \alpha_3 s \\ & & \alpha_4 & \alpha_5 s \end{pmatrix}.$$

The dual solution is updated to $p' = (1,0,0,1)$ and $q' = (2,2,2,2)$, and the corresponding tight coefficient matrix $A'^{\#} = \left(\begin{smallmatrix} Q'^{\#} \\ T'^{\#} \end{smallmatrix}\right)$ of $A'(s)$ is

$$A'^{\#} = \begin{pmatrix} -1 & -1 & & \\ 1 & 1 & & \\ & & \alpha_2 & \\ & & & \alpha_5 \end{pmatrix}.$$

The minimizer J^* that we used above also minimizes the right-hand side of the rank identity (7.3) for $A'^{\#}$. Since $A'^{\#}$ is still singular, we go on the modification. Noting the

order of rows, we transform $Q'^{\#}[R_Q, J^*] = \left(\begin{smallmatrix} -1 & -1 \\ 1 & 1 \end{smallmatrix}\right)$ by $U' = \left(\begin{smallmatrix} 1 & \\ 1 & 1 \end{smallmatrix}\right)$ into

$$\bar{Q}'^{\#}[R_Q, J^*] = U'Q'^{\#}[R_Q, J^*] = \begin{pmatrix} -1 & -1 \\ & \end{pmatrix}.$$

We have $U'_Q(s) = \mathrm{diag}(s^{-1}, 1)U' \,\mathrm{diag}(s, 1) = \left(\begin{smallmatrix} 1 & \\ s & 1 \end{smallmatrix}\right)$, and modify $A'(s)$ to

$$A''(s) = \begin{pmatrix} 1 & & & \\ s & 1 & & \\ & & 1 & \\ & & & 1 \end{pmatrix} A'(s) = \begin{pmatrix} -s & -s & -1 & 1 \\ & & -s+1 & s \\ & \alpha_1 & \alpha_2 s^2 & \alpha_3 s \\ & & \alpha_4 & \alpha_5 s \end{pmatrix}.$$

The dual solution is updated to $p'' = (1, 3, 2, 3)$ and $q'' = (2, 2, 4, 4)$. Our algorithm halts at this point since $A''(s)$ is upper-tight, which can be checked through the nonsingularity of the tight coefficient matrix $A''^{\#}$ again. Now $d(A)$ is computed as $d(A) = d(A'') = \hat{d}(A'') = 3$. The resulting DAE is

$$\begin{cases} -\dot{x}_1 - \dot{x}_2 - x_3 + x_4 = f_1(t) - f_2(t), \\ \qquad -\dot{x}_3 + x_3 + \dot{x}_4 = \dot{f}_1(t) - \dot{f}_2(t) + f_2(t), \\ \alpha_1 x_2 + \alpha_2 \ddot{x}_3 + \alpha_3 \dot{x}_4 = f_3(t), \\ \qquad \alpha_4 x_3 + \alpha_5 \dot{x}_4 = f_4(t), \end{cases} \tag{7.32}$$

which is index 2. An index-1 DAE is obtained by applying the MS-algorithm to the DAE (7.32).

7.4.2 Example from Electrical Network

The next example is the DAE (7.1) representing the electrical network in Figure 6.2. Since the coefficient matrix $A(s)$ is not LM-polynomial, it seems that we cannot directly apply our algorithm to $A(s)$. However, since each of the last five rows in $A(s)$ do not contain two or more accurate constants, we can convert $A(s)$ into an LM-polynomial matrix by multiplying an independent parameter to each of the rows. In addition, by the same logic to Lemma 7.7, our algorithm works without actually multiplying the independent parameters by regarding nonzero entries in the last five rows as independent parameters. Thus we see $A(s)$ as an LM-polynomial matrix $A(s) = \left(\begin{smallmatrix} Q(s) \\ T(s) \end{smallmatrix}\right)$, where $Q(s)$ and $T(s)$ correspond to the first and last five rows in $A(s)$, respectively. The matrix $A(s)$ meets the condition 7.1.3 for DCLM-polynomial matrices with $\lambda = (0, 0, 0, 0, 0)$ and $\mu = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$.

We are now ready for applying our algorithm to $A(s)$. In Phase 1, a dual optimal solution is obtained as $p = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$ and $q = (0, 0, 1, 0, 0, 0, 0, 0, 1, 0)$, which

implies $\hat{d}(A) = 2$. The corresponding tight coefficient matrix $A^{\#} = \begin{pmatrix} Q^{\#} \\ T^{\#} \end{pmatrix}$ is given by

$$A^{\#} = \left(\begin{array}{ccccc|ccccc} -1 & & & -1 & 1 & & & & & \\ & 1 & & 1 & -1 & & & & & \\ \hline & & & & & 1 & & 1 & -1 & \} R_Q \\ & & & & & -1 & -1 & & & \\ & & & & & & 1 & -1 & & \\ \hline R_1 & & & & & -1 & & & & \\ & R_2 & & & & & -1 & & & \} R_T \\ & & L & & & & & -1 & & \\ & & & -1 & & & & & C & \\ & & & & & & & & & 1 \end{array}\right).$$

A minimizer J^* of the rank identity (7.3) for $A^{\#}$ is the set of nine columns other than the rightmost column corresponding to the variable $\tilde{v}_5$. Thus we can check

$$\operatorname{rank} A^{\#} = Q^{\#}[R_Q, J^*] + T^{\#}[R_T, J^*] + |C \setminus J^*| = 4 + 4 + 1 = 9 < 10,$$

which implies that $A(s)$ is not upper-tight. We proceed to the matrix modification process for DCLM-polynomial matrices that we described in Section 7.3.

The row set R_k and the column set C_k for $k \in \mathbb{Z}$ defined in (7.25) are the following:

$$Q^{\#} = \left(\begin{array}{\overbrace{ccccc}^{C_0}\overbrace{c}^{C_1}\overbrace{ccccc}^{C_0}\overbrace{c}^{C_1}\overbrace{c}^{C_0}} \overbrace{}^{J^*} \\ -1 & & & -1 & 1 & & & & & \\ & 1 & & 1 & -1 & & & & & \\ & & & & & 1 & & 1 & -1 & \} R_0. \\ & & & & & -1 & -1 & & & \\ & & & & & & 1 & -1 & & \end{array}\right)$$

Now $Q^{\#}$ can be seen as a block diagonal matrix consisting of one diagonal block $Q_0^{\#} = Q^{\#}[R_0, C_0]$ by $Q^{\#}[R_0, C_1] = O$. We transform

$$Q_0^{\#}[R_0, C_0 \cap J^*] = \begin{pmatrix} -1 & & -1 & 1 & & & \\ & 1 & 1 & -1 & & & \\ & & & & 1 & & 1 \\ & & & & -1 & -1 & \\ & & & & & 1 & -1 \end{pmatrix}$$

into

$$UQ_0^{\#}[R_0, C_0 \cap J^*] = \begin{pmatrix} -1 & & -1 & 1 & & & \\ & 1 & 1 & -1 & & & \\ & & & & & & \\ & & & & -1 & -1 & \\ & & & & & 1 & -1 \end{pmatrix},$$

where $U = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ & & 1 & 1 & 1 \\ & & & 1 & \\ & & & & 1 \end{pmatrix}$. Using $U_Q(s) = U$, we modify $A(s)$ to

$$A'(s) = \left(\begin{array}{ccccc|ccccc} -1 & & & -1 & 1 & & & & & \\ & 1 & 1 & 1 & -1 & & & & & \\ \hline & & & & & & & & 1 & -1 \\ & & & & & -1 & -1 & & 1 & \\ & & & & & & 1 & -1 & & \\ \hline R_1 & & & & & -1 & & & & \\ & R_2 & & & & & -1 & & & \\ & & Ls & & & & & -1 & & \\ & & & -1 & & & & & Cs & \\ & & & & & & & & & 1 \end{array}\right),$$

where the third row is different between $A(s)$ and $A'(s)$. The dual solution is updated to $p' = (0, 0, 1, 0, 0, 0, 0, 0, 0, 1)$ and $q' = (0, 0, 1, 0, 0, 0, 0, 0, 1, 1)$. Since the corresponding tight coefficient matrix of $A'(s)$ is nonsingular, we stop the algorithm. The index of the modified DAE remains at 2. Applying the MS-algorithm, we can reduce the index of the modified DAE to 1.

7.5 Numerical Experiments

We conduct numerical experiments to reduce the index comparing our algorithm with the LC-method by Tan et al. [94]. Recall that the LC-method works for linear DAEs whose associated polynomial matrix $A(s)$ has only constants, whereas our algorithm can treat a DAE containing independent parameters.

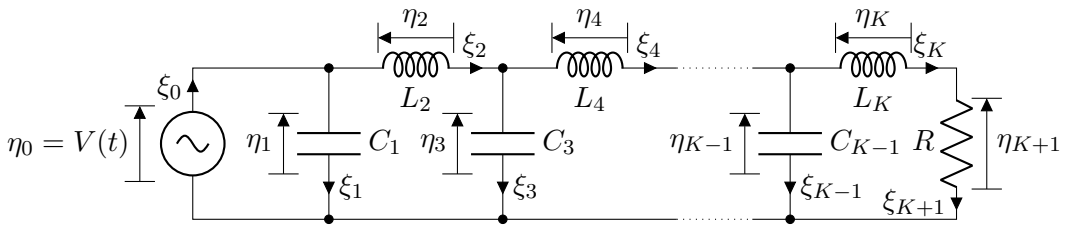


Figure 7.1: Butterworth filter via the K-th Cauer topology.

7.5.1 Experiment Description

For an even positive integer K, the *Butterworth filter* via the *K-th Cauer topology* is an electrical circuit shown in Figure 7.1. This circuit has $n \coloneqq 2K + 4$ state variables $\xi_0, \xi_1, \ldots, \xi_{K+1}, \eta_0, \eta_1, \ldots, \eta_{K+1}$, where ξ_j is a current shown in Figure 7.1 and η_j is a voltage across the branch carrying the current ξ_j for $j \in [0, K+1]$.

A DAE representing the circuit is given by

$$\begin{cases} -\xi_{k-1} + \xi_k + \xi_{k+1} = 0 \quad (k = 1, 3, 5, \ldots, K-1), \\ -\xi_0 + \xi_1 + \xi_3 + \cdots + \xi_{K+1} = 0, \\ \eta_0 + \eta_2 + \eta_4 + \cdots + \eta_K + \eta_{K+1} = 0, \\ -\eta_{k-1} + \eta_k + \eta_{k+1} = 0 \quad (k = 2, 4, 6, \ldots, K), \\ \eta_0 = V(t), \\ -\xi_k + C_k \dot{\eta}_k = 0 \quad (k = 1, 3, 5, \ldots, K-1), \\ L_k \dot{\xi}_k - \eta_k = 0 \quad (k = 2, 4, 6, \ldots, K), \\ R\xi_{K+1} - \eta_{K+1} = 0. \end{cases} \tag{7.33}$$

The index of the DAE (7.33) is 2 and the associated polynomial matrix $A(s)$ is a sparse matrix having $6K + 7$ nonzero coefficients. Though it suffices to use simpler equations $-\xi_K + \xi_{K+1} = 0$ and $\eta_0 + \eta_1 = 0$ instead of the second and the third equations in (7.33), respectively, we use them to make the tight coefficient matrix singular.

We apply our algorithm and the LC-method to the DAE (7.33) using the following two ways of implementations:

Dense Matrix Implementation, which stores a matrix in the memory as a two-dimensional array. While this implementation always requires $\mathrm{O}(nm)$ space for a matrix of size $m \times n$, it has less overhead than the sparse matrix implementation if the matrix is dense.

Sparse Matrix Implementation, which stores only nonzero entries of a matrix. A typical implementation of this type is in formats called the *compressed sparse column* (CSC) or the *compressed sparse row* (CSR). We adopt the CSR in our experiments.

Table 7.1: Running time (sec) of dense implementations for $K = 2^{11}$.

	LC-method		Proposed	
Phase 1	1.80×10^{-2}	(0.00%)	1.70×10^{-2}	(0.00%)
Phase 2	6.69×10^{2}	(29.61%)	9.69×10^{1}	(19.54%)
Phase 3	1.59×10^{3}	(70.26%)	3.97×10^{2}	(79.98%)
MS-algorithm	1.02×10^{0}	(0.04%)	7.28×10^{-1}	(0.15%)
Total	2.26×10^{3}	(100.00%)	4.96×10^{2}	(100.00%)

Table 7.2: Running time (sec) of sparse implementations for $K = 2^{11}$.

	LC-method		Proposed	
Phase 1	1.55×10^{-2}	(4.88%)	1.58×10^{-2}	(3.32%)
Phase 2	1.33×10^{-1}	(41.87%)	3.82×10^{-1}	(80.07%)
Phase 3	1.25×10^{-1}	(39.40%)	39.2×10^{-2}	(8.21%)
MS-algorithm	2.54×10^{-2}	(7.98%)	2.47×10^{-2}	(5.17%)
Total	3.18×10^{-1}	(100.00%)	4.78×10^{-1}	(100.00%)

The sparse matrix implementation has an advantage that it consumes only the space proportional to the number of nonzero entries, and thus algorithms using this implementation are expected to run efficiently for sparse matrices.

In our algorithm, we treat the coefficients R, C_k, L_k, and '± 1's in the last four equations in (7.33) as independent parameters similarly to the example in Section 7.4.2. Then the associated polynomial matrix $A(s) = \left(\begin{smallmatrix} Q(s) \\ T(s) \end{smallmatrix}\right)$ is dimensionally consistent, where $|\mathrm{Row}\, Q(s)| = |\mathrm{Row}\, T(s)| = K + 2$. In the LC-method, we substitute the following real numbers:

$$
\begin{aligned}
C_k &= 2 \sin \frac{2k-1}{2K} \pi \quad (k = 1, 3, 5, \ldots, K-1), \\
L_k &= 2 \sin \frac{2k-1}{2K} \pi \quad (k = 2, 4, 6, \ldots, K), \\
R &= \pi.
\end{aligned}
$$

Under this setting, we compare the running time for $K = 2, 4, 8, 16, \ldots, 2^{16}$. We implemented all algorithms in C++ using the library `Eigen3` for matrix computation. It is emphasized again that we do not rely on symbolic computation. The experiments are conducted on a laptop with Core i7 1.7 GHz CPU and 8 GB memory.

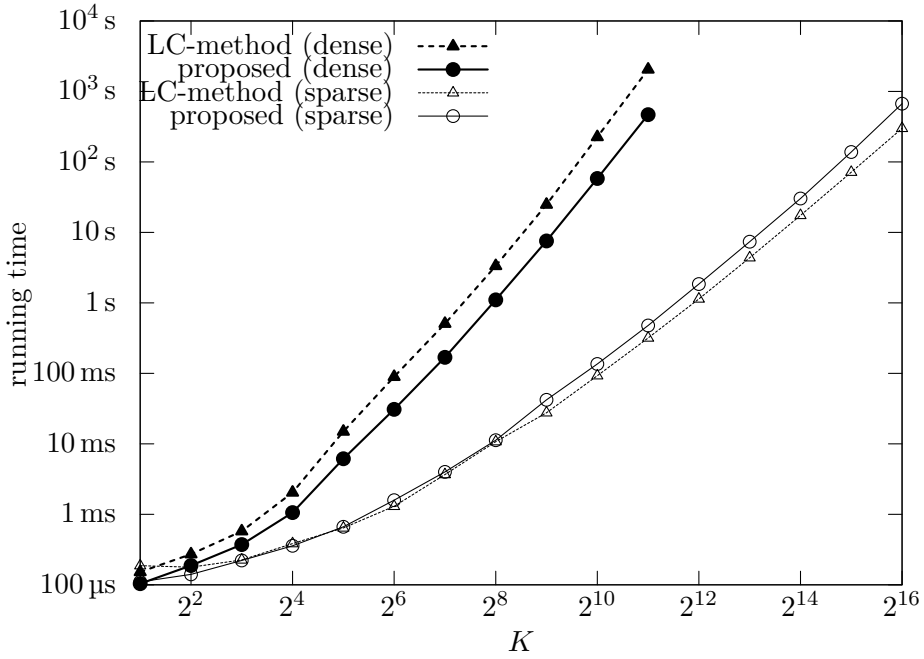
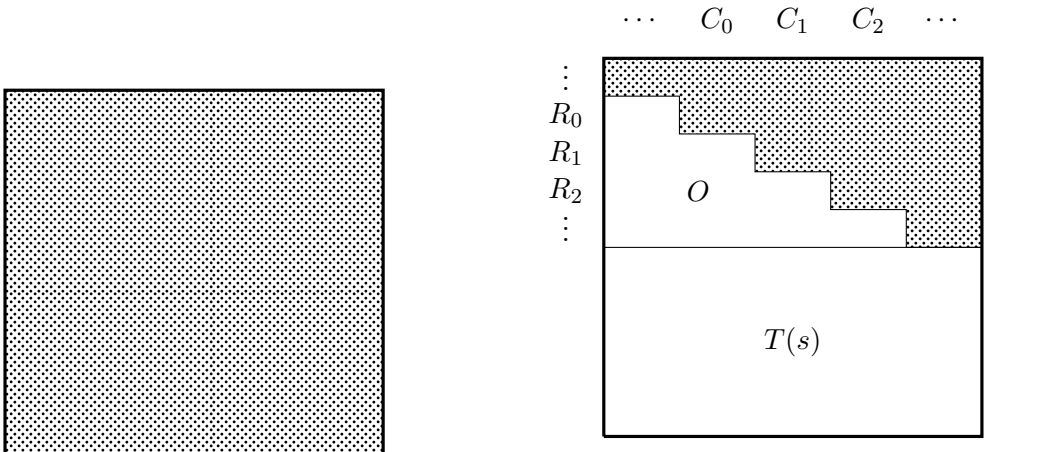


Figure 7.2: Log-log plot of the experimental result: K versus the running time.



(a) LC-method

(b) Our algorithm for DCLM-polynomial matrices

Figure 7.3: Hatched regions indicate submatrices in a polynomial matrix $A(s)$ to be modified by algorithms. In (b), we use notations given in Section 7.3.

7.5.2 Experimental Results

Tables 7.1 and 7.2 and Figure 7.2 show the running time of the algorithms. On the dense matrix implementations, both algorithms did not run for $K \geq 2^{12}$ due to the lack of memory capacity. The reasons are as follows. Our implementations express a polynomial as an array of coefficients using `std::vector<int>` or `std::vector<float>` in C++, and it consumes 32 bytes even for the zero polynomial. Since the number of entries in the input polynomial matrix $A(s)$ for $K = 2^{12}$ is $n^2 = (2K+4)^2 \geq 2^{26}$, we need at least $2^{26} \times 32\,\text{bytes} = 2\,\text{GB}$ to hold $A(s)$. Besides the input matrix, our implementations construct several constant and polynomial matrices of similar or larger size, such as a tight coefficient matrix $A^{\#}$, a unimodular matrix $U(s)$ for modification in Phase 3, and an output matrix. Thus, 2^{12} is near the borderline of the maximum K for which our implementations run on our laptop with 8 GB memory.

It can be seen from Figure 7.2 that our algorithm is faster than the LC-method on their dense matrix implementations, and it is converse for their sparse ones. This is attributed to the fact that in the process of multiplying polynomial matrices in (7.19) at Phase 3, the LC-method multiplies the entire of the given polynomial matrix $A(s)$ whereas our algorithm multiplies only submatrices of $A(s)$ as illustrated in Figure 7.3. Since this process is dominant on the dense matrix implementations as Table 7.1 indicates, the difference between the sizes of matrices to be multiplied directly affects the difference of the running times. This process, however, does not cost much in the sparse matrix implementations, and thus Phase 2 becomes relatively expensive. As a result, the difference between the running times on sparse matrix implementations reflect the difference between that of the independent matching algorithm and the Gaussian elimination used by our algorithm and the LC-method in Phase 2, respectively.

Recalling that the size of the DAE is $n = \mathrm{O}(K)$, Figure 7.2 shows that the running time of our algorithm grows proportionally to $\mathrm{O}(n^{2.84})$ in the dense matrix implementation and $\mathrm{O}(n^{1.97})$ in the sparse one for $K \geq 2^8$. Both are much faster than the theoretical guarantee $\mathrm{O}(n^4 \log n)$ given in Theorem 7.20.

7.6 Application to Nonlinear DAEs

In this section, we discuss the application of our algorithm to nonlinear DAEs. The $\sigma\nu$-method [11], which is implemented in Mathematica [102], adopts a strategy of treating nonlinear or time-varying terms as independent parameters in the Jacobian matrices of DAEs. We first describe the $\sigma\nu$-method briefly.

Consider an index-2 nonlinear DAE

$$\begin{cases} F_1 : \dot{x}_1 + g(x_2) = f_1(t), \\ F_2 : \dot{x}_1 + x_1 + x_3 = f_2(t), \\ F_3 : \dot{x}_1 + x_3 = f_3(t), \end{cases} \tag{7.34}$$

where $g : \mathbb{R} \to \mathbb{R}$ is a smooth nonlinear function. The $\sigma\nu$-method constructs two kinds of Jacobian matrices JD and JV as follows:

$$\mathrm{JD} = \left(\frac{\partial F_i}{\partial \dot{x}_j}\right)_{i,j} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}, \quad \mathrm{JV} = \left(\frac{\partial F_i}{\partial x_j}\right)_{i,j} = \begin{pmatrix} 0 & \mathrm{d}g/\mathrm{d}x_2 & 0 \\ 1 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

If JD is nonsingular, the DAE is index 0 from the implicit function theorem. Otherwise, the method performs Gaussian elimination on JD (and JV simultaneously) to make the bottom row of JD zero. Then the method differentiates the equation corresponding to the bottom row, and checks the nonsingularity of JD again. The main feature of the $\sigma\nu$-method is to treat nonlinear or time-varying terms as "independent parameters" to avoid complicated symbolic manipulations. The method works according to the rule that arithmetic operations and the differentiation of independent parameters generate new independent parameters.

The $\sigma\nu$-method may fail due to this rule. For example, let α_1 be an independent parameter representing $\mathrm{d}g/\mathrm{d}x_2$ in JV. By subtracting the first row from the second and third ones, we obtain

$$\mathrm{JD} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad \mathrm{JV} = \begin{pmatrix} 0 & \alpha_1 & 0 \\ 1 & \alpha_2 & 1 \\ 0 & \alpha_3 & 1 \end{pmatrix},$$

where $\alpha_2 = 0 - \alpha_1$ and $\alpha_3 = 0 - \alpha_1$ are newly generated parameters by the rule of arithmetic operations. We differentiate the second and third rows. Then JD and JV are

$$\mathrm{JD} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & \alpha_2 & 1 \\ 0 & \alpha_3 & 1 \end{pmatrix}, \quad \mathrm{JV} = \begin{pmatrix} 0 & \alpha_1 & 0 \\ 0 & \alpha_4 & 0 \\ 0 & \alpha_5 & 0 \end{pmatrix},$$

where α_4 and α_5 are parameters corresponding to the derivatives of α_2 and α_3, respectively. Although the Jacobian matrix JD is indeed singular due to $\alpha_2 = \alpha_3$, the $\sigma\nu$-method halts at this point as the method regards α_2 and α_3 as independent. This failure originates from the elimination of matrices involving the independent parameter α_1. We have confirmed that the implementation in Mathematica actually fails on this DAE.

Our algorithm is applied to the same DAE (7.34) as follows. Let

$$A(s) = \begin{pmatrix} s & \alpha & \\ s+1 & & 1 \\ s & & 1 \end{pmatrix},$$

where α is an independent parameter representing $\mathrm{d}g/\mathrm{d}x_2$. The tight coefficient matrix corresponding to a dual optimal solution $p = (0, 0, 0)$ and $q = (1, 0, 0)$ is

$$A^{\#} = \begin{pmatrix} 1 & \alpha & \\ 1 & & 1 \\ 1 & & 1 \end{pmatrix},$$

which is singular. Thus we need to modify the matrix. By the same logic as the discussion in Section 7.4.2, we can regard $A(s)$ as an LM-polynomial matrix $A(s) = \binom{T(s)}{Q(s)}$, where $T(s)$ corresponds to the first row and $Q(s)$ corresponds to the other two ones in $A(s)$. Then our algorithm modifies $A(s)$ to

$$A'(s) = \begin{pmatrix} s & \alpha & \\ 1 & & \\ s & & 1 \end{pmatrix},$$

which is upper-tight (we omit the detail of this modification).

This example shows that our algorithm works for a DAE to which the existing index reduction algorithm cannot be applied. Our algorithm is expected to rarely cause cancellations between nonlinear terms as it does not perform the row operations involving independent parameters. In particular, our algorithm can be applied to nonlinear DAEs in which cancellations occur only between linear terms like the transistor amplifier DAE in [61]; such DAEs often appear in practice. Therefore, although the application to nonlinear DAEs remains at the stage of a heuristic, it is anticipated that the proposed method can be useful for index reduction of nonlinear DAEs.

Chapter 8

Structural Modification for Nonlinear DAEs

In this chapter, we propose two modification methods for nonlinear DAEs: the *substitution method* in Section 8.1 and the *augmentation method* in Section 8.2. Both methods symbolically manipulate formulations of DAEs using a symbolic computation engine. We give an application example in Section 8.3 and conduct numerical experiments in Section 8.4.

8.1 Substitution Method

8.1.1 Outline of Method

In this section, we describe a new modification method for nonlinear DAEs, called the substitution method. This method is used in Phase 3d of the combinatorial relaxation framework.

Let $\mathbb{T} \subseteq \mathbb{R}$ be a nonempty open interval and $\Omega \subseteq \mathbb{R}^{(\ell+1)n}$ a nonempty open set. The input of the substitution method is a DAE (1.4) of size n with real analytic function $F : \mathbb{T} \times \Omega \to \mathbb{R}^n$ such that

(I1) $G(F)$ has a perfect matching,

(I2) for any square submatrix $D[I, J]$ of the system Jacobian D with respect to a dual optimal solution, if $D[I, J]$ is not identically singular on $\mathbb{T} \times \Omega$, then there exists a consistent point of (1.4) at which $D[I, J]$ is nonsingular,

(I3) D is identically singular.

The smoothness assumption on F is needed to avoid technical difficulties. We remark that (I2) is just a part of a sufficient condition for which the substitution method works, and it suffices in practice to check the condition only for a few submatrices of D that are used as pivots in the method.

The substitution method modifies the DAE (1.4) into another DAE

$$\bar{F}^{\mathrm{sub}}(t, x, \dot{x}, \ldots, x^{(\ell+\kappa)}) = 0 \tag{8.1}$$

of size n such that

(S1) $\bar{F}^{\mathrm{sub}}$ is a real analytic function defined on a nonempty open subset $\bar{\mathbb{T}}^{\mathrm{sub}} \times \bar{\Omega}^{\mathrm{sub}} \subseteq \mathbb{T} \times \Omega^{(\kappa)}$ with $\kappa \leq \ell n$,

(S2) the resulting DAE (8.1) is locally equivalent to the input DAE (1.4),

(S3) $\hat{d}(\bar{F}^{\mathrm{sub}}) \leq \hat{d}(F) - 1$.

See Lemma 8.8 for the precise meaning of "locally equivalent" in (S2).

We first introduce notations needed to describe the method. Let R and C be the equation index set and the variable index set of the DAE (1.4), respectively. For $I \subseteq R$, let F_I denote a "subvector" $(F_i)_{i\in I}$ of F indexed by I. Similarly, for $J \subseteq C$, let x_J denote a subvector $(x_j)_{j\in J}$ of x indexed by J. Let p and q be the vectors of variables in $\mathrm{D}(F)$. In addition, we use the following notations

$$F_I^{(p)} \coloneqq \left(F_i^{(p_i)}\right)_{i\in I}, \quad x_J^{(q)} \coloneqq \left(x_j^{(q_j)}\right)_{j\in J}, \quad \frac{\partial F_I^{(p)}}{\partial x_J^{(q)}} \coloneqq \left(\frac{\partial F_i^{(p_i)}}{\partial x_j^{(q_j)}}\right)_{i\in I, j\in J}$$

for $I \subseteq R$ and $J \subseteq C$.

Here we start to describe the method. Let D be the system Jacobian of (1.4) with respect to an optimal solution (p, q) of $\mathrm{D}(F)$ and suppose that D is identically singular. We regard D as a matrix over the quotient field K of the ring of real analytic functions on $\mathbb{T} \times \Omega$. The substitution method first finds $r \in R$, $I \subseteq R \setminus \{r\}$ and $J \subseteq C$ with $|I| = |J| =: m$ such that

(C1) $D[I, J]$ is nonsingular,

(C2) $\operatorname{rank} D[I \cup \{r\}, C] = m$,

(C3) $p_r \leq p_i$ for $i \in I$.

Here, both the nonsingularity in (C1) and the rank in (C2) are in the sense of those of matrices over K. Namely, these conditions can be rewritten as

(C1*) $D[I, J]$ is not identically singular, and

(C2*) the maximum size of a submatrix in $D[I\cup\{r\}, C]$ that is not identically singular is m.

The existence of (r, I, J) satisfying (C1)–(C3) is guaranteed through the algorithm explained in Section 8.1.2.

Example 8.1. Consider the DAE (6.16). We have seen in Example 6.21 that its system Jacobian D, which is (6.17), is identically singular. We can choose $r = 2$, $I = \{1\}$, and $J = \{1\}$. Then $D[I, J] = \dot{x}_2$ is nonsingular, $\operatorname{rank} D[I \cup \{r\}, \{1, 2\}] = \operatorname{rank} D = 1$, and $p_2 \leq p_1$ since $p = (0, 0)$. Hence this (r, I, J) satisfies (C1)–(C3). □

Let (r, I, J) be a triple satisfying the conditions (C1)–(C3). Define $S = R \setminus (I \cup \{r\})$ and $T = C \setminus J$. Then the DAE (1.4) is divided into three subsystems as follows:

$$\begin{cases} F_r\big(t, x, \dot{x}, \dots, x^{(\ell)}\big) = 0, \\ F_I\big(t, x, \dot{x}, \dots, x^{(\ell)}\big) = 0, \\ F_S\big(t, x, \dot{x}, \dots, x^{(\ell)}\big) = 0. \end{cases} \tag{8.2}$$

The system Jacobian D with respect to (p, q) forms a block matrix as follows:

$$D = \begin{array}{c} \\ \{r\} \\ I \\ S \end{array} \begin{array}{c} \begin{array}{cc} J & T \end{array} \\ \begin{pmatrix} \frac{\partial F_r^{(p_r)}}{\partial x_J^{(q)}} & \frac{\partial F_r^{(p_r)}}{\partial x_T^{(q)}} \\ \frac{\partial F_I^{(p)}}{\partial x_J^{(q)}} & \frac{\partial F_I^{(p)}}{\partial x_T^{(q)}} \\ \frac{\partial F_S^{(p)}}{\partial x_J^{(q)}} & \frac{\partial F_S^{(p)}}{\partial x_T^{(q)}} \end{pmatrix} \end{array}.$$

By the condition (C3) and Proposition 6.6, it holds that

$$\frac{\partial F_I^{(p)}}{\partial x_J^{(q)}} = \left(\frac{\partial F_i^{(p_i)}}{\partial x_j^{(q_j)}} \right)_{i \in I, j \in J} = \left(\frac{\partial F_i^{(p_i - p_r)}}{\partial x_j^{(q_j - p_r)}} \right)_{i \in I, j \in J} = \frac{\partial F_I^{(p - p_r \mathbb{1})}}{\partial x_J^{(q - p_r \mathbb{1})}},$$

where $\mathbb{1}$ is the vector of ones with appropriate dimension. In addition, from the condition (C1), the submatrix $D[I, J] = \frac{\partial F_I^{(p)}}{\partial x_J^{(q)}} = \frac{\partial F_I^{(p - p_r \mathbb{1})}}{\partial x_J^{(q - p_r \mathbb{1})}}$ is not identically singular on $\mathbb{T} \times \Omega$. Therefore, by (I2), there exists a point $(\hat{t}, \hat{X}) \in \mathbb{T} \times \Omega^{(\kappa)}$ such that $F_I^{(p - p_r \mathbb{1})}(\hat{t}, \hat{X}) = 0$ and $\frac{\partial F_I^{(p - p_r \mathbb{1})}}{\partial x^{(q - p_r \mathbb{1})}}(\hat{t}, \hat{X})$ is nonsingular, where

$$\kappa \coloneqq \max_{i \in I} p_i - p_r. \tag{8.3}$$

Then via the IFT, we can solve an equation

$$F_I^{(p - p_r \mathbb{1})}\big(t, x, \dot{x}, \dots, x^{(\ell + \kappa)}\big) = 0 \tag{8.4}$$

for $x_J^{(q - p_r \mathbb{1})}$ as

$$x_J^{(q - p_r \mathbb{1})} = \varphi\big(t, x, \dot{x}, \dots, x^{(\ell + \kappa)}\big), \tag{8.5}$$

where φ is a function that does not depend on $x_J^{(q-p_r\mathbb{1})}$. See Section 8.1.3 for a rigorous description of this part.

Example 8.2 (Continued from Example 8.1)**.** Since $p_1 - p_2 = 0$ and $q_1 - p_2 = 1$, the equation (8.4) on the DAE (6.16) is

$$F_1 : \dot{x}_1\dot{x}_2 - 2\cos^2 t = 0 \tag{8.6}$$

for $\dot{x}_1$. Solving (8.6) for $\dot{x}_1$, we obtain

$$\dot{x}_1 = -\frac{2\cos^2 t}{\dot{x}_2} \tag{8.7}$$

unless $\dot{x}_2 = 0$. The equation (8.7) corresponds to (8.5). □

Finally, we substitute the right-hand side of (8.5) into $x_J^{(q-p_r\mathbb{1})}$ in the first equation $F_r = 0$ of (8.2). The modified DAE (8.1) is

$$\begin{cases} \bar{F}_r^{\mathrm{sub}}(t, x, \dot{x}, \dots, x^{(\ell+\kappa)}) = 0, \\ \qquad F_I(t, x, \dot{x}, \dots, x^{(\ell)}) = 0, \\ \qquad F_S(t, x, \dot{x}, \dots, x^{(\ell)}) = 0, \end{cases} \tag{8.8}$$

where $\bar{F}_r^{\mathrm{sub}}$ is a function obtained from F_r by substituting (8.5).

Example 8.3 (Continued from Example 8.2)**.** We substitute (8.7) into F_2 in (6.16). The resulting DAE is

$$\begin{cases} F_1 \quad : & \dot{x}_1\dot{x}_2 - 2\cos^2 t = 0, \\ \bar{F}_2^{\mathrm{sub}} : & x_1 + x_2 - 3\sin t - 2 = 0. \end{cases} \tag{8.9}$$

and the substitution method is done.

According to the procedure of combinatorial relaxation, we go back to Phase 1d and check the nonsingularity of the system Jacobian again. It can be confirmed that $\hat{\delta}$ of (8.9) is 1, which is less than that of (6.16). The system Jacobian D' of (8.9) corresponding to a dual optimal solution $p' = (0, 1), q' = (1, 1)$ is

$$D' = \begin{pmatrix} \dot{x}_2 & \dot{x}_1 \\ 1 & 1 \end{pmatrix}.$$

Since D' is not identically singular, the combinatorial relaxation is done. □

8.1.2 Algorithm for Finding (r, I, J)

Let D be a singular $n \times n$ matrix over a field K with row index set R and column index set C, and $p = (p_i)_{i \in R}$ an integer vector indexed by R. On the setting in Section 8.1.1, K is the quotient field of the ring of analytic functions on $\mathbb{T} \times \Omega$. We give an algorithm, which uses arithmetic operations over K, to find $r \in R, I \subseteq R \setminus \{r\}$ and $J \subseteq C$ satisfying the conditions (C1)–(C3).

First, by column operations, we transform D into $D' = (D'_{i,j})_{i \in R, j \in C}$ in the form

$$D' = \begin{array}{c} \\ H \\ R \setminus H \end{array} \begin{array}{c} \begin{array}{cc} B & C \setminus B \end{array} \\ \begin{pmatrix} I_k & O \\ * & O \end{pmatrix} \end{array}, \tag{8.10}$$

where $H \subseteq R$ and $B \subseteq C$ with $k := |H| = |B| = \operatorname{rank} D$. Here, "$*$" indicates an arbitrary matrix. Let $h : B \to H$ denote the natural bijection represented by the top left block $D'[H, B]$ in (8.10). Namely, $h(j) = i$ if and only if $D'_{i,j} \neq 0$ for $j \in B$ and $i \in H$.

Next, we choose $l \in R \setminus H$ arbitrarily. Note that $R \setminus H$ is nonempty because D' is singular. Put

$$Z := \{l\} \cup \{h(j) \mid j \in B, D'_{l,j} \neq 0\} \subseteq R. \tag{8.11}$$

Finally, we take $r \in Z$ such that $p_r \leq p_i$ for all $i \in Z$. Put $I := Z \setminus \{r\}$ and choose $J \subseteq C$ such that $D[I, J]$ is nonsingular. The existence of J is guaranteed by the following lemma.

Lemma 8.4. *Let $D \in K^{n \times n}$ be a singular matrix and $Z \subseteq R$ defined in* (8.11)*. Then $D[Z, C]$ is not of full-row rank and $D[I, C]$ is of full-row rank for any proper subset $I \subsetneq Z$.*

Proof. Since D' in (8.10) is a matrix obtained from D by column operations, it suffices to show the statement for D'. By the definition of Z, it holds

$$D'[\{l\}, C] - \sum_{i \in Z \setminus \{l\}} D'[\{i\}, C] = 0.$$

This implies that $D'[Z, C]$ is not of full-row rank.

We next show that $D'[I, C]$ is of full-row rank for $I \subsetneq Z$. This is trivial if $l \notin I$ since $I \subsetneq Z \subseteq \{l\} \cup H$ and $D'[H, C]$ is of full-row rank. Suppose that $l \in I$. Then we can take $i \in Z \setminus I$. From the definition of Z, $D'[(Z \setminus \{i\}) \cup \{l\}, C]$ is of full-row rank. Since $I \subseteq (Z \setminus \{i\}) \cup \{l\}$, $D'[I, C]$ is also of full-row rank. □

The following theorem holds from the construction of (r, I, J) together with Lemma 8.4.

Theorem 8.5. *For a singular matrix $D \in K^{n \times n}$, the above algorithm returns (r, I, J) satisfying the conditions* (C1)–(C3)*.*

This algorithm uses $O(n^3)$ arithmetic operations on K.

8.1.3 Application of Implicit Function Theorem

This section gives a mathematically rigorous description of the application of the IFT to (8.4). The description in this section is used in proofs of the substitution method later.

We introduce additional notations. Let $\mathcal{C} \subseteq C \times \{0, 1, 2, \ldots, \ell\}$ be a finite set of index pairs such that $(j, k) \in \mathcal{C}$ indicates an argument $x_j^{(k)}$ of F in (1.4). Let $\mathbb{R}^{\mathcal{C}}$ denote a $|\mathcal{C}|$-dimensional real vector space with index set $\mathcal{C}$. For $X \in \mathbb{R}^{\mathcal{C}}$ and $\mathcal{J} \subseteq \mathcal{C}$, let $X_{\mathcal{J}}$ designate a subvector of X with index set $\mathcal{J}$.

The following is a version of the IFT which we use.

Theorem 8.6 (Implicit Function Theorem; IFT)**.** *Let $U \subseteq \mathbb{R}^{n+m}$ be an open set having coordinates (x, y) with $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Let $f : U \to \mathbb{R}^m$ be a real analytic function. Fix a point $(\xi, \eta) \in U$ such that $f(\xi, \eta) = 0$ and $\frac{\partial f}{\partial y}(\xi, \eta)$ is nonsingular. Then there exist open sets $V \subseteq \mathbb{R}^n$ and $W \subseteq \mathbb{R}^m$ with $(\xi, \eta) \in V \times W \subseteq U$ and a real analytic function $\varphi : V \to W$ such that*

(1) *$\varphi(\xi) = \eta$,*

(2) *$f(x, y) = 0$ if and only if $y = \varphi(x)$ for all $(x, y) \in V \times W$, and*

(3) *$\frac{\partial f}{\partial y}(x, \varphi(x))$ is nonsingular and*

$$\frac{\mathrm{d}\varphi}{\mathrm{d}x}(x) = -\left(\frac{\partial f}{\partial y}(x, \varphi(x))\right)^{-1} \frac{\partial f}{\partial x}(x, \varphi(x)) \tag{8.12}$$

for all $x \in V$.

The function φ in the IFT is called an *explicit function*. The formula (8.12) is called the *implicit differentiation formula*.

Let us start the description of the application of the implicit function theorem. Let (p, q) be an optimal solution of $\mathrm{D}(F)$ and (r, I, J) triple satisfying the conditions (C1)–(C3). Put

$$\mathcal{C} \coloneqq \{(j, k) \mid j \in C,\ 0 \leq k \leq q_j - p_r\}. \tag{8.13}$$

From Proposition 6.6 and the feasibility of (p, q), it holds that

$$\sigma(F_i^{(p_i - p_r)}, x_j) = \sigma(F_i, x_j) + p_i - p_r = c_{i,j} + p_i - p_r \leq q_j - p_r \tag{8.14}$$

for $i \in I \cup \{r\}$ and $j \in J$ with $\sigma(F_i, x_j) > -\infty$. Thus we regard both F_r and $F_I^{(p_I - p_r \mathbb{1})}$ as functions defined on $\mathbb{T} \times U$, where U is an open subset of $\mathbb{R}^{\mathcal{C}}$.

Take $(\hat{t}, \hat{X}) \in \mathbb{T} \times U$ such that $F_I^{(p - p_r \mathbb{1})}(\hat{t}, \hat{X}) = 0$ and $\frac{\partial F_I^{(p - p_r \mathbb{1})}}{\partial x_J^{(q - p_r \mathbb{1})}}(\hat{t}, \hat{X})$ is nonsingular.

Let

$$\mathcal{J} \coloneqq \{(j, q_j - p_r) \mid j \in J\} \subseteq \mathcal{C}. \tag{8.15}$$

Then the components of $\hat{X}$ is bipartitioned by $\mathcal{J}$ as $\hat{X} = (\hat{X}_{\mathcal{C}\setminus\mathcal{J}}, \hat{X}_{\mathcal{J}})$. Thus by the implicit function theorem, there exist open sets $\bar{\mathbb{T}}^{\mathrm{sub}} \subseteq \mathbb{T}$, $V \subseteq \mathbb{R}^{\mathcal{C}\setminus\mathcal{J}}$ and $W \subseteq \mathbb{R}^{\mathcal{J}}$ with $(\hat{t}, \hat{X}_{\mathcal{C}\setminus\mathcal{J}}, \hat{X}_{\mathcal{J}}) \in \bar{\mathbb{T}}^{\mathrm{sub}} \times V \times W \subseteq \mathbb{T} \times U$ and a real analytic function $\varphi : \bar{\mathbb{T}}^{\mathrm{sub}} \times V \to W$ such that $\hat{X}_{\mathcal{J}} = \varphi(\hat{t}, \hat{X}_{\mathcal{C}\setminus\mathcal{J}})$ and

$$F_I^{(p-p_r\mathbb{1})}(t, X_{\mathcal{C}\setminus\mathcal{J}}, \varphi(t, X_{\mathcal{C}\setminus\mathcal{J}})) = 0 \tag{8.16}$$

for every $(t, X_{\mathcal{C}\setminus\mathcal{J}}) \in V$. In addition, all zeros of $F_I^{(p-p_r\mathbb{1})}$ in $\bar{\mathbb{T}}^{\mathrm{sub}} \times V \times W$ are in the form of (8.16). Using φ, the modified function $\bar{F}_r^{\mathrm{sub}} : \bar{\mathbb{T}}^{\mathrm{sub}} \times V \to \mathbb{R}$ can be expressed as

$$\bar{F}_r^{\mathrm{sub}}(t, X_{\mathcal{C}\setminus\mathcal{J}}) = F_r(t, X_{\mathcal{C}\setminus\mathcal{J}}, \varphi(t, X_{\mathcal{C}\setminus\mathcal{J}})) \tag{8.17}$$

for $(t, X_{\mathcal{C}\setminus\mathcal{J}}) \in \bar{\mathbb{T}}^{\mathrm{sub}} \times V$. Since both F_r and φ are real analytic, so is $\bar{F}_r^{\mathrm{sub}}$.

We remark about the domain of the resulting system of functions $\bar{F}$ in (8.8). In the above argument, we treated the domain of $F_I^{(p-p_r\mathbb{1})}$ as $\mathbb{T} \times U$, which is an open subset of $\mathbb{T} \times \mathbb{R}^{\mathcal{C}}$. However, the domain of $F_I^{(p-p_r\mathbb{1})}$ can also be represented as $\mathbb{T} \times \Omega^{(\kappa)}$, where κ is defined by (8.3) (indeed, U is the projection of $\Omega^{(\kappa)}$ onto $\mathbb{R}^{\mathcal{C}}$). Since $\bar{F}_r$ is a function obtained from $F_I^{(p-p_r\mathbb{1})}$ and F_r by the above transformation, the domain of $\bar{F}_r$ (and thus of $\bar{F}$) can also be regarded as $\bar{\mathbb{T}}^{\mathrm{sub}} \times \bar{\Omega}^{\mathrm{sub}}$, where $\bar{\Omega}^{\mathrm{sub}}$ is a nonempty open subset of $\Omega^{(\kappa)}$.

8.1.4 Proofs

This section is devoted to the validity proofs of our method.

We first show (S1). In Section 8.1.3, we have already shown that F_r is a real analytic function defined on $\bar{\mathbb{T}}^{\mathrm{sub}} \times \bar{\Omega}^{\mathrm{sub}}$. Thus, what we should give is only the bound on κ. Applying the algorithm given in Section 7.2.3, we can obtain (p, q) such that $p_i \le \ell n$ for any $i \in R$. Then the following lemma immediately follows.

Lemma 8.7. *In the substitution method, κ defined in* (8.3) *is at most ℓn.*

Next, we focus on (S2), which claims about the equivalence of the original DAE and the modified DAE.

Lemma 8.8. *Consider a DAE* (1.4) *satisfying* (I1)–(I3). *Let $x : \bar{\mathbb{T}}^{\mathrm{sub}} \to \mathbb{R}^n$ be a sufficiently smooth trajectory satisfying the initial value condition* (1.6) *for $(t^*, X^*) \in \bar{\mathbb{T}}^{\mathrm{sub}} \times \bar{\Omega}^{\mathrm{sub}}$. Then there exists an open subinterval $\mathbb{I} \subseteq \bar{\mathbb{T}}^{\mathrm{sub}}$ containing t^* such that x is a solution of* (1.4) *on $\mathbb{I}$ if and only if x is a solution of* (8.8) *on $\mathbb{I}$.*

Proof. We show both the "if" and "only if" parts simultaneously. Suppose that there exists an open subinterval $\mathbb{I} \subseteq \bar{\mathbb{T}}^{\mathrm{sub}}$ with $t^* \in \mathbb{I}$ such that x is a solution of (1.4) or (8.8) on $\mathbb{I}$.

Then x satisfies $F_I(t, x(t), \dot{x}(t), \ldots, x^{(\ell)}(t)) = 0$ on $\mathbb{I}$, which is a subsystem of both (8.2) and (8.8). Thus x also satisfies (8.4) on $\mathbb{I}$.

We rewrite the equation (8.4) for $x(t)$ using $\mathcal{C}$ and $\mathcal{J}$ defined by (8.13) and (8.15), respectively. Let $\mathrm{D}^{\mathcal{C}}$ denote the differentiation operator that maps x to a trajectory $\mathrm{D}^{\mathcal{C}}x : \bar{\mathbb{T}}^{\mathrm{sub}} \to \mathbb{R}^{\mathcal{C}}$ such that

$$\big(\mathrm{D}^{\mathcal{C}}x(t)\big)_{(j,k)} = x_j^{(k)}(t)$$

for $t \in \bar{\mathbb{T}}^{\mathrm{sub}}$ and $(j,k) \in \mathcal{C}$. Then the initial value condition (1.6) can be represented as $\mathrm{D}^{\mathcal{C}}x(t^*) = X^*_{\mathcal{C}}$. Since the domain of $F_I^{(p-p_r\mathbb{1})}$ is an open subset of $\mathbb{T}\times\mathbb{R}^{\mathcal{C}}$, the equation (8.4) for $x(t)$ can also be represented as

$$F_I^{(p-p_r\mathbb{1})}\big(t, \mathrm{D}^{\mathcal{C}}x(t)\big) = 0$$

or

$$F_I^{(p-p_r\mathbb{1})}\big(t, \mathrm{D}^{\mathcal{C}\setminus\mathcal{J}}x(t), \mathrm{D}^{\mathcal{J}}x(t)\big) = 0 \tag{8.18}$$

for $t \in \mathbb{I}$, where $\mathrm{D}^{\mathcal{C}\setminus\mathcal{J}}$ and $\mathrm{D}^{\mathcal{J}}$ are differentiation operators defined in the same way as $\mathrm{D}^{\mathcal{C}}$.

Let $U \subseteq \mathbb{R}^{\mathcal{C}}, V \subseteq \mathbb{R}^{\mathcal{C}\setminus\mathcal{J}}$ and $W \subseteq \mathbb{R}^{\mathcal{J}}$ be open sets defined in Section 8.1.3. Here, since x is smooth, U is open and $\mathrm{D}^{\mathcal{C}}x(t^*) = X^*_{\mathcal{C}} \in U$, it holds $\mathrm{D}^{\mathcal{C}}x(t) \in U$ for all $t \in \mathbb{I}$ by taking $\mathbb{I}$ sufficiently small. This implies that $\mathrm{D}^{\mathcal{C}\setminus\mathcal{J}}x(t) \in V$ and $\mathrm{D}^{\mathcal{J}}x(t) \in W$ for $t \in \mathbb{I}$. Comparing (8.16) and (8.18), we obtain

$$\mathrm{D}^{\mathcal{J}}x(t) = \varphi\big(t, \mathrm{D}^{\mathcal{C}\setminus\mathcal{J}}x(t)\big)$$

for $t \in \mathbb{I}$. Therefore, we have

$$\begin{aligned} F_r(t, x(t), \dot{x}(t), \ldots, x^{(\ell)}(t)) &= F_r\big(t, \mathrm{D}^{\mathcal{C}\setminus\mathcal{J}}x(t), \mathrm{D}^{\mathcal{J}}x(t)\big) \\ &= F_r\big(t, \mathrm{D}^{\mathcal{C}\setminus\mathcal{J}}x(t), \varphi\big(t, \mathrm{D}^{\mathcal{C}\setminus\mathcal{J}}x(t)\big)\big) \\ &= \bar{F}_r^{\mathrm{sub}}\big(t, \mathrm{D}^{\mathcal{C}\setminus\mathcal{J}}x(t)\big) \\ &= \bar{F}_r^{\mathrm{sub}}\big(t, x(t), \dot{x}(t), \ldots, x^{(\ell+\kappa)}(t)\big), \end{aligned}$$

which means that x is a solution of (8.8) if x is a solution of (1.4), and vice versa. □

We finally show that the modified DAE satisfies (S3). In order to show (S3), it suffices to show that (p,q) is a feasible solution of $\mathrm{D}(\bar{F}^{\mathrm{sub}})$ but not an optimal solution. The feasibility is easily shown as follows.

Lemma 8.9. *Consider a DAE* (1.4) *satisfying* (I1)–(I3) *and let* (p,q) *be an optimal solution of* $\mathrm{D}(F)$. *Then* (p,q) *is feasible on* $\mathrm{D}(\bar{F}^{\mathrm{sub}})$.

Proof. Let (r, I, J) be a triple satisfying the conditions (C1)–(C3). Consider the explicit

function φ in (8.5). For $i \in I$ and $j \in C$, we have

$$\sigma(\varphi, x_j) \leq \sigma(F_i^{(p_i - p_r)}, x_j) \leq q_j - p_r,$$

where we used (8.14) in the last inequality. Because $\bar{F}_r^{\mathrm{sub}}$ is a function obtained from F_r by substituting (8.5), it holds

$$\sigma(\bar{F}_r^{\mathrm{sub}}, x_j) \leq \max\{\sigma(F_r, x_j), \sigma(\varphi, x_j)\} \leq q_j - p_r$$

for every $j \in C$. Thus (p, q) is feasible on $\mathrm{D}(\bar{F}^{\mathrm{sub}})$. □

We finally focus on the non-optimality of (p, q) on $\mathrm{D}(\bar{F}^{\mathrm{sub}})$. By Proposition 6.7, our goal is to show that t-rank $\bar{D} < n$ holds, where $\bar{D}$ be the system Jacobian of (8.1) with respect to (p, q). This is shown by the following lemma.

Lemma 8.10. *Consider a DAE* (1.4) *satisfying* (I1)–(I3)*. Let* (p, q) *be an optimal solution of* $\mathrm{D}(F)$ *and* (r, I, J) *a triple satisfying* (C1)–(C3)*. Then the modified function* $\bar{F}_r$ *in* (8.8) *does not depend on* $x_j^{(q_j - p_r)}$ *for all* $j \in C$*.*

Proof. The claim is easy to see for $j \in J$ because we have eliminated $x_j^{(q_j - p_r)}$ from F_r by substituting (8.5). Consider the variable $x_T^{(q - p_r \mathbb{1})}$ with $T = C \setminus J$. Let $\mathcal{C}$ and $\mathcal{J}$ be index sets defined in (8.13) and (8.15), respectively. For $(t, X) \in \bar{\mathbb{T}}^{\mathrm{sub}} \times \bar{\Omega}^{\mathrm{sub}}$, we denote $(t, X_{\mathcal{C}\setminus\mathcal{J}}, \varphi(t, X_{\mathcal{C}\setminus\mathcal{J}}))$ by $A_{t,X}$ for short, where φ is the explicit function given by (8.5). From the chain rule, the implicit differentiation formula (8.12) and Proposition 6.6, we obtain

$$\begin{aligned}
&\frac{\partial \bar{F}_r^{\mathrm{sub}}}{\partial x^{(q - p_r \mathbb{1})}}(t, X_{\mathcal{C}\setminus\mathcal{J}}) \\
&= \frac{\partial F_r}{\partial x_T^{(q - p_r \mathbb{1})}}(A_{t,X}) + \frac{\partial F_r}{\partial x_J^{(q - p_r \mathbb{1})}}(A_{t,X}) \frac{\partial \varphi}{\partial x_T^{(q - p_r \mathbb{1})}}(t, X_{\mathcal{C}\setminus\mathcal{J}}) \\
&= \frac{\partial F_r}{\partial x_T^{(q - p_r \mathbb{1})}}(A_{t,X}) - \frac{\partial F_r}{\partial x_J^{(q - p_r \mathbb{1})}}(A_{t,X}) \left(\frac{\partial F_I^{(p - p_r \mathbb{1})}}{\partial x_J^{(q - p_r \mathbb{1})}}(A_{t,X}) \right)^{-1} \frac{\partial F_I^{(p - p_r \mathbb{1})}}{\partial x_T^{(q - p_r \mathbb{1})}}(A_{t,X}) \\
&= \frac{\partial F_r^{(p_r)}}{\partial x_T^{(q)}}(A_{t,X}) - \frac{\partial F_r^{(p_r)}}{\partial x_J^{(q)}}(A_{t,X}) \left(\frac{\partial F_I^{(p)}}{\partial x_J^{(q)}}(A_{t,X}) \right)^{-1} \frac{\partial F_I^{(p)}}{\partial x^{(q)}}(A_{t,X})
\end{aligned} \tag{8.19}$$

for $(t, X) \in \bar{\mathbb{T}}^{\mathrm{sub}} \times \bar{\Omega}^{\mathrm{sub}}$. The right hand side of (8.19) coincides with the Schur complement of $\frac{\partial F_I^{(p)}}{\partial x_J^{(q)}}(A_{t,X})$ in the following matrix

$$\tilde{D}(t, X_{\mathcal{C}\setminus\mathcal{J}}) \coloneqq \begin{pmatrix} \frac{\partial F_r^{(p_r)}}{\partial x_J^{(q)}}(A_{t,X}) & \frac{\partial F_r^{(p_r)}}{\partial x_T^{(q)}}(A_{t,X}) \\ \frac{\partial F_I^{(p)}}{\partial x_J^{(q)}}(A_{t,X}) & \frac{\partial F_I^{(p)}}{\partial x_T^{(q)}}(A_{t,X}) \end{pmatrix}.$$

Thus, we have

$$\operatorname{rank} \tilde{D}(t, X_{\mathcal{C}\setminus\mathcal{J}}) = \operatorname{rank} \frac{\partial F_I^{(p)}}{\partial x_J^{(q)}}(A_{t,X}) + \operatorname{rank} \frac{\partial \bar{F}_r^{\mathrm{sub}}}{\partial x_T^{(q-p_r\mathbb{1})}}(t, X_{\mathcal{C}\setminus\mathcal{J}}) \tag{8.20}$$

for all $(t, X) \in \bar{\mathbb{T}}^{\mathrm{sub}} \times \bar{\Omega}^{\mathrm{sub}}$. Let D be a system Jacobian of F with respect to (p, q). Note that $\tilde{D}$ is a matrix obtained from $D[I \cup \{r\}, C]$ by substituting $\varphi(t, X_{\mathcal{C}\setminus\mathcal{J}})$ into $X_{\mathcal{J}}$. Hence it holds $\operatorname{rank} \tilde{D}(t, X_{\mathcal{C}\setminus\mathcal{J}}) \leq \operatorname{rank} D[I \cup \{r\}, C](t, X) \leq \operatorname{rank} D[I \cup \{r\}, C] = m$ with $m = |I|$, where the last equality comes from (C2). In addition, the rank of $\frac{\partial F_I^{(p)}}{\partial x_J^{(q)}}(A_{t,X})$ is m due to the invertibility. Therefore, the rank of $\frac{\partial \bar{F}_r^{\mathrm{sub}}}{\partial x_T^{(q-p_r\mathbb{1})}}(t, X_{\mathcal{C}\setminus\mathcal{J}})$ is zero from (8.20), which means that $\frac{\partial \bar{F}_r^{\mathrm{sub}}}{\partial x_T^{(q-p_r\mathbb{1})}}$ is identically zero on $\bar{\mathbb{T}}^{\mathrm{sub}} \times \bar{\Omega}^{\mathrm{sub}}$. Thus $\bar{F}_r^{\mathrm{sub}}$ does not depend on $x_j^{(q_j-p_r)}$ for $j \in T$. □

Corollary 8.11. *For a DAE* (1.4) *satisfying* (I1)–(I3)*, it holds* $\hat{d}(\bar{F}^{\mathrm{sub}}) \leq \hat{d}(F) - 1$.

Proof. Let (p, q) be an optimal solution of $\mathrm{D}(F)$ and (r, I, J) a triple satisfying the conditions (C1)–(C3). Let $\bar{D}$ be the system Jacobian of (8.8) with respect to (p, q). By Proposition 6.6, it holds

$$\bar{D}[\{r\}, C] = \frac{\partial \bar{F}_r^{(p_r)}}{\partial x^{(q)}} = \frac{\partial \bar{F}_r}{\partial x^{(q-p_r\mathbb{1})}},$$

whereas the right-hand side is identically zero from Lemma 8.10. Thus t-rank $\bar{D}$ is less than n, and from Proposition 6.7, (p, q) is not an optimal solution of $\mathrm{D}(\bar{F}^{\mathrm{sub}})$. This concludes the proof. □

We conclude this section with the following theorem.

Theorem 8.12. *For a DAE* (1.4) *satisfying* (I1)–(I3)*, the substitution method outputs a DAE* (8.1) *satisfying* (S1)–(S3)*.*

8.2 Augmentation Method

8.2.1 Method Description

This section describes another proposed modification method for nonlinear DAEs, which we call an augmentation method. The input of the augmentation method is a nonlinear DAE (1.4) of size n satisfying the conditions (I1)–(I3), where $F : \mathbb{T} \times \Omega \to \mathbb{R}^n$ is a real analytic function again. Instead of solving equations symbolically, the augmentation method augments the size of the DAE by introducing a new variable vector y and attaching

new equations. Formally, the augmentation method modifies (1.4) into a DAE

$$\bar{F}^{\mathrm{aug}}(t, x, \dot{x}, \ldots, x^{(\ell+\kappa)}, y) = 0 \tag{8.21}$$

of size $n + m$ such that

(A1) $\bar{F}^{\mathrm{aug}}$ is a real analytic function defined on a nonempty open subset $\bar{\mathbb{T}}^{\mathrm{aug}} \times \bar{\Omega}^{\mathrm{aug}} \times Y \subseteq \mathbb{T} \times \Omega^{(\kappa)} \times \mathbb{R}^m$ with $\kappa \leq \ell n$ and $m \leq n - 1$,

(A2) the resulting DAE (8.21) is locally equivalent to (1.4),

(A3) $\hat{d}(\bar{F}^{\mathrm{aug}}) \leq \hat{d}(F) - 1$.

See Lemma 8.15 for the precise meaning of "locally equivalent" in (A2).

The substitution method and the augmentation method are the same except for the last modification process. The overlapping part is described here briefly. Let R and C be the equation index set and the variable index set of the input DAE (1.4), respectively. Let (p, q) be an optimal solution of $\mathrm{D}(F)$ and D denote the system Jacobian with respect to (p, q). We first find $r \in R$, $I \subseteq R \setminus \{r\}$ and $J \subseteq C$ satisfying the conditions (C1)–(C3) described in Section 8.1.1. Define $\kappa \coloneqq \max_{i \in I} p_i - p_r, m \coloneqq |I|, S \coloneqq R \setminus (I \cup \{r\})$ and $T \coloneqq C \setminus J$.

The following modification step differs from the substitution method. Let $I' = \{i' \mid i \in I\}$ and $J' = \{j' \mid j \in J\}$ be copies of I and J, respectively. Take a point (τ, Ξ) arbitrary from the domain $\bar{\mathbb{T}}^{\mathrm{sub}} \times \bar{\Omega}^{\mathrm{sub}} \subseteq \mathbb{T} \times \Omega^{(\kappa)}$ of the resultant DAE $\bar{F}^{\mathrm{sub}}$ of the substitution method. We regard $\Omega^{(\kappa)}$ as a subset of $\mathbb{R}^{\mathcal{C}}$ hereafter, where $\mathcal{C} \coloneqq C \times \{0, 1, 2, \ldots, \ell + \kappa\}$. For $X \in \mathbb{R}^{\mathcal{C}}$ and a vector $y = (y_{j'})_{j' \in J'}$ with index set J', let $\psi_\Xi(X, y)$ be a vector of $\mathbb{R}^{\mathcal{C}}$ such that

$$(\psi_\Xi(X, y))_{(j,k)} \coloneqq \begin{cases} y_{j'} & (j \in J, k = q_j - p_r), \\ \Xi_{(j,k)} & (j \in T, k = q_j - p_r), \\ X_{(j,k)} & (\text{otherwise}) \end{cases}$$

for $(j, k) \in \mathcal{C}$. For each $i \in I$, we define a function

$$\bar{F}^{\mathrm{aug}}_{i'}(t, x, \dot{x}, \ldots, x^{(\ell+\kappa)}, y) \coloneqq F_i^{(p_i - p_r)}(t, \psi_\Xi(X, y)),$$

where $X = (x, \dot{x}, \ldots, x^{(\ell+\kappa)})$. Namely, $\bar{F}^{\mathrm{aug}}_{i'}$ is obtained by replacing $x_j^{(q_j - p_r)}$ in $F_i^{(p_i - p_r)}$ with a variable $y_{j'}$ for $j \in J$ and with a constant $\Xi_{(j, q_j - p_r)}$ for $j \in T$. Put $\bar{F}^{\mathrm{aug}}_{I'} \coloneqq \left(\bar{F}^{\mathrm{aug}}_{i'}\right)_{i' \in I'}$. We also define

$$\bar{F}^{\mathrm{aug}}_r(t, x, \dot{x}, \ldots, x^{(\ell+\kappa)}, y) \coloneqq F_r(t, \psi_\Xi(X, y))$$

in the same way.

The output (8.21) of the augmentation method is the following DAE

$$\begin{cases} \bar{F}_r^{\mathrm{aug}}(t, x, \dot{x}, \dots, x^{(\ell+\kappa)}, y) = 0, \\ F_I(t, x, \dot{x}, \dots, x^{(\ell)}) = 0, \\ F_S(t, x, \dot{x}, \dots, x^{(\ell)}) = 0, \\ \bar{F}_{I'}^{\mathrm{aug}}(t, x, \dot{x}, \dots, x^{(\ell+\kappa)}, y) = 0 \end{cases} \tag{8.22}$$

with unknown function $(x(t), y(t))$ of t. The domain $\bar{\mathbb{T}}^{\mathrm{aug}} \times \bar{\Omega}^{\mathrm{aug}}$ of (8.22) is given by $\bar{\mathbb{T}}^{\mathrm{aug}} := \bar{\mathbb{T}}^{\mathrm{sub}}$ and $\bar{\Omega}^{\mathrm{aug}} := \{(X, y) \in \bar{\Omega}^{\mathrm{sub}} \times \mathbb{R}^m \mid \psi_\Xi(X, y) \in \bar{\Omega}^{\mathrm{sub}}\}$.

Example 8.13 (Continued from Example 8.1)**.** The modified function $\bar{F}_{1'}^{\mathrm{aug}}$ is obtained from F_1 by replacing $x_1^{(q_1-p_2)} = \dot{x}_1$ with a new variable $y_{1'}$ and $x_2^{(q_2-p_2)} = \dot{x}_2$ with an arbitrary nonzero constant $\xi \in \mathbb{R}$. The function $\bar{F}_r^{\mathrm{aug}} = \bar{F}_2^{\mathrm{aug}}$ is obtained in the same manner. The output (8.22) of the augmentation method applied to (6.16) is

$$\begin{cases} F_1 \quad : & \dot{x}_1 \dot{x}_2 - 2\cos^2 t = 0, \\ \bar{F}_2^{\mathrm{aug}} : & {y_{1'}}^2 \xi^2 + x_1 + x_2 - 4\cos^4 t - 3\sin t - 2 = 0, \\ \bar{F}_{1'}^{\mathrm{aug}} : & y_{1'}\xi - 2\cos^2 t = 0 \end{cases} \tag{8.23}$$

with unknown function $(x_1, x_2, y_{1'})$. We can confirm that $\hat{\delta}$ of (8.23) is 1 and (8.23) has a nonsingular system Jacobian. □

The DAE (8.22) is obtained by copying some equations (or their derivatives), relabelling variables and substituting constants. Hence if the original DAE contains only a few variables in each equation, so does (8.22). Thus the augmentation method retains the sparsity of DAEs.

8.2.2 Proofs

Validity proofs of the augmentation method are given in this section. We first show (A1).

Lemma 8.14. *For a DAE* (1.4) *satisfying* (I1)–(I3)*, the resulting DAE* $\bar{F}^{\mathrm{aug}} = 0$ *satisfies* (A1)*.*

Proof. It is clear that $\bar{F}^{\mathrm{aug}}$ is real analytic from its construction, which is a combination of variable relabelling and partial substitution of constants on F. Let $\eta = (\eta_{j'})_{j' \in J'}$ be a vector defined by $\eta_{j'} = \Xi_{(j, q_j - p_r)}$ for $j' \in J'$. Then it holds $(\Xi, \eta) \in \bar{\Omega}^{\mathrm{aug}}$ from $\psi_\Xi(\Xi, \eta) = \Xi \in \bar{\Omega}^{\mathrm{sub}}$. Hence $\bar{\Omega}^{\mathrm{aug}}$ is nonempty. In addition, since ψ_Ξ is a continuous map and $\bar{\Omega}^{\mathrm{sub}}$ is an open set, $\bar{\Omega}^{\mathrm{aug}}$ is also open. Therefore the domain $\bar{\mathbb{T}}^{\mathrm{aug}} \times \bar{\Omega}^{\mathrm{aug}}$ of $\bar{F}^{\mathrm{aug}}$ is a nonempty open set.

The bounds on κ and m are given by Lemma 8.7 and $m = |I| \le n - 1$. □

We next show (A2) in the sense of the following lemma.

Lemma 8.15. *Consider a DAE* (1.4) *satisfying* (I1)–(I3)*. Let* $x : \bar{\mathbb{T}}^{\mathrm{aug}} \to \mathbb{R}^n$ *be a*

sufficiently smooth trajectory satisfying the initial value condition (1.6) *for* $(t^*, X^*) \in \bar{\mathbb{T}}^{\mathrm{aug}} \times \bar{\Omega}^{\mathrm{aug}}$. *Then there exists an open subinterval* $\mathbb{I} \subseteq \bar{\mathbb{T}}^{\mathrm{aug}}$ *containing* t^* *such that the following two statements are equivalent:*

(1) x *is a solution of* (1.4) *on* $\mathbb{I}$, *and*

(2) *there uniquely exists a trajectory* $y : \mathbb{I} \to \mathbb{R}^m$ *such that* (x, y) *is a solution of* (8.22) *on* $\mathbb{I}$.

Proof. From the argument on the substitution method, the last equation

$$\bar{F}^{\mathrm{aug}}_{I'}\big(t, x, \dot{x}, \dots, x^{(\ell+\kappa)}, y\big) = 0$$

in (8.22) can be solved for y on the domain of $\bar{F}^{\mathrm{aug}}$ as

$$y = \bar{\varphi}^{\mathrm{aug}}\big(t, x, \dot{x}, \dots, x^{(\ell+\kappa)}\big),$$

where $\bar{\varphi}^{\mathrm{aug}}$ is a function obtained by replacing $x_j^{(q_j - p_r)}$ of φ in (8.5) with the constant $\Xi_{(j, q_j - p_r)}$ for $j \in T$. Therefore, (8.22) is equivalent to

$$\begin{cases} \bar{F}^{\mathrm{aug}}_r\big(t, x, \dot{x}, \dots, , x^{(\ell+\kappa)}, \bar{\varphi}^{\mathrm{aug}}\big(t, x, \dot{x}, \dots, x^{(\ell+\kappa)}\big)\big) = 0, \\ \qquad\qquad F_I\big(t, x, \dot{x}, \dots, x^{(\ell)}\big) = 0, \\ \qquad\qquad F_S\big(t, x, \dot{x}, \dots, x^{(\ell)}\big) = 0, \\ \qquad\qquad\qquad y = \bar{\varphi}^{\mathrm{aug}}\big(t, x, \dot{x}, \dots, x^{(\ell+\kappa)}\big). \end{cases} \tag{8.24}$$

It can be seen from (8.17) that the left-hand side of the first equation in (8.24) is a function obtained by replacing $x_j^{(q_j - p_r)}$ of $\bar{F}^{\mathrm{sub}}_r$ with the constant $\Xi_{(j, q_j - p_r)}$ for $j \in T$. On the other hand, $\bar{F}^{\mathrm{sub}}_r$ does not depend on $x_j^{(q_j - p_r)}$ for all $j \in T$ from Lemma 8.10. Therefore, the first equation in (8.24) is equivalent to $\bar{F}^{\mathrm{sub}}_r\big(t, x, \dot{x}, \dots, x^{(\ell+\kappa)}\big) = 0$. Thus the system (8.24) is equivalent to

$$\begin{cases} \bar{F}^{\mathrm{sub}}\big(t, x, \dot{x}, \dots, x^{(\ell+\kappa)}\big) = 0, \\ \qquad\qquad y = \bar{\varphi}^{\mathrm{aug}}\big(t, x, \dot{x}, \dots, x^{(\ell+\kappa)}\big). \end{cases} \tag{8.25}$$

The statement of this lemma is shown by (8.25) together with Lemma 8.8. □

Let $\bar{R} \coloneqq R \cup I'$ and $\bar{C} \coloneqq C \cup J'$. We finally show (A3) as a corollary of the following lemma.

Lemma 8.16. *Consider a DAE* (1.4) *satisfying* (I1)–(I3) *and let* (p, q) *be a dual optimal*

solution. Define

$$\bar{p}_i \coloneqq \begin{cases} p_i & (i \in R), \\ p_r & (i \in I'), \end{cases} \quad \bar{q}_j \coloneqq \begin{cases} q_j & (j \in C), \\ p_r & (j \in J') \end{cases} \tag{8.26}$$

for $i \in \bar{R}$ and $j \in \bar{C}$. Then $(\bar{p}, \bar{q})$ is feasible but not optimal to $\mathrm{D}(\bar{F}^{\mathrm{aug}})$.

Proof. We first prove

$$\sigma(\bar{F}^{\mathrm{aug}}_{i'}, x_j) < q_j - p_r \tag{8.27}$$

for $i' \in I' \cup \{r\}$ and $j \in C$. Since $x_j^{(q_j - p_r)}$ in $\bar{F}^{\mathrm{aug}}_{i'}$ has been replaced with a dummy variable or a constant, it holds $\sigma(\bar{F}^{\mathrm{aug}}_{i'}, x_j) < \sigma(F_i^{(p_i - p_r)}, x_j)$, where $i = r$ if $i' = r$ here. In addition, $\sigma(F_i^{(p_i - p_r)}, x_j) = \sigma(F_i, x_j) + p_i - p_r \le q_j - p_r$ holds, where the first equality comes from Proposition 6.6 and the second inequality is due to the feasibility of (p, q) on $\mathrm{D}(F)$. Thus (8.27) is true.

We next show the feasibility of $(\bar{p}, \bar{q})$ on $\mathrm{D}(\bar{F}^{\mathrm{aug}})$. For $i \in R \setminus \{r\}$ and $j \in C$, it holds $\sigma(\bar{F}^{\mathrm{aug}}_i, x_j) = \sigma(F_i, x_j) \le q_j - p_i = \bar{q}_j - \bar{p}_i$ from the feasibility of (p, q) on $\mathrm{D}(F)$. For $i \in R \setminus \{r\}$ and $j' \in J'$, we have $\sigma(\bar{F}^{\mathrm{aug}}_i, y_{j'}) = \sigma(F_i, y_{j'}) = -\infty \le \bar{q}_j - \bar{p}_i$. For $i' \in I' \cup \{r\}$ and $j \in C$, it holds $\sigma(\bar{F}^{\mathrm{aug}}_{i'}, x_j) < q_j - p_r = \bar{q}_j - \bar{p}_{i'}$ from (8.27). In the last case with $i' \in I' \cup \{r\}$ and $j' \in J'$, we have $\sigma(\bar{F}^{\mathrm{aug}}_{i'}, y_{j'}) = 0 = p_r - p_r = \bar{p}_{i'} - \bar{q}_{j'}$. Thus $(\bar{p}, \bar{q})$ is feasible on $\mathrm{D}(\bar{F}^{\mathrm{aug}})$.

Finally, we show the non-optimality of $(\bar{p}, \bar{q})$ on $\mathrm{D}(\bar{F}^{\mathrm{aug}})$. From Proposition 6.7, it suffices to show t-rank $\bar{D} < n + m$, where $\bar{D}$ is the system Jacobian of (8.22) with respect to $(\bar{p}, \bar{q})$. Here, $\bar{D}_{i',j}$ is identically zero for $i' \in I' \cup \{r\}$ and $j \in C$ due to (8.27). Figure 8.1 shows the zero/nonzero pattern of $\bar{D}$, where $\bar{D}[I, J'] = O$ and $\bar{D}[S, J'] = O$ can also be checked from the definition of $\bar{F}^{\mathrm{aug}}$. Therefore, $I \cup S \cup J'$ is a vertex cover in the bipartite graph $G(D) = (\bar{R} \cup \bar{C}, E(D))$ associated with D. By the König–Egeváry theorem, we have

$$\text{t-rank}\, \bar{D} \le \left| I \cup S \cup J' \right| = m + (n - m - 1) + m = n + m - 1,$$

which completes the proof. □

Corollary 8.17. *For a DAE* (1.4) *satisfying* (I1)–(I3)*, the resulting DAE $\bar{F}^{\mathrm{aug}} = 0$ satisfies* (A3)*.*

Proof. Let (p, q) be a dual optimal solution on $\mathrm{D}(F)$ and $(\bar{p}, \bar{q})$ defined by (8.26). From Lemma 8.16, it holds that

$$\begin{aligned} \hat{d}(\bar{F}^{\mathrm{aug}}) &< \sum_{j \in \bar{C}} \bar{q}_j - \sum_{i \in \bar{R}} \bar{p}_i = \left(m p_r + \sum_{j \in C} q_j \right) - \left(m p_r + \sum_{i \in R} p_i \right) = \sum_{j \in C} q_j - \sum_{i \in R} p_i \\ &= \hat{d}(F) \end{aligned}$$

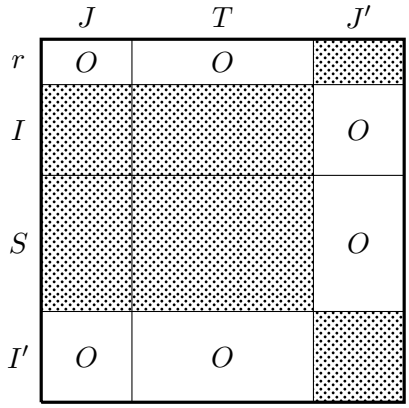


Figure 8.1: The zero/nonzero pattern of the system Jacobian $\bar{D}$ of $\bar{F}^{\text{aug}}$. The hatched region may contain nonzero elements.

as required. □

The above lemmas are summed up in the following theorem.

Theorem 8.18. *For a DAE* (1.4) *satisfying* (I1)–(I3)*, the augmentation method returns a DAE* (8.21) *satisfying* (A1)–(A3)*.*

8.3 More Example

We demonstrate our methods using an extra example arising from a real problem, a transistor amplifier problem on an electrical network [61]. The problem is described by an index-1 DAE in the following form

$$\begin{cases} F_1: & C_1(\dot{x}_1-\dot{x}_2)+(x_1-U_e(t))/R_0=0,\\ F_2: & -C_1(\dot{x}_1-\dot{x}_2)-U_b/R_2+x_2(1/R_1+1/R_2)-(\alpha-1)g(x_2-x_3)=0,\\ F_3: & C_2\dot{x}_3+x_3/R_3-g(x_2-x_3)=0,\\ F_4: & C_3(\dot{x}_4-\dot{x}_5)+(x_4-U_b)/R_4+\alpha g(x_2-x_3)=0,\\ F_5: & -C_3(\dot{x}_4-\dot{x}_5)-U_b/R_6+x_5(1/R_5+1/R_6)-(\alpha-1)g(x_5-x_6)=0,\\ F_6: & C_4\dot{x}_6+x_6/R_7-g(x_5-x_6)=0,\\ F_7: & C_5(\dot{x}_7-\dot{x}_8)+(x_7-U_b)/R_8+\alpha g(x_5-x_6)=0,\\ F_8: & -C_5(\dot{x}_7-\dot{x}_8)+x_8/R_9=0, \end{cases} \tag{8.28}$$

where $g(x)=\beta(\exp(x/U_F)-1)$ and $U_e(t)=0.1\sin(200\pi t)$ with nonzero parameters U_b, U_F, α, β, $R_0, R_1, \ldots, R_9$, and $C_1, \ldots, C_5$.

A dual optimal solution on (8.28) is given by $p=(0,\ldots,0)$ and $q=(1,\ldots,1)\in\mathbb{Z}^8$.

The system Jacobian corresponding (p, q) is a singular constant matrix

$$D = \begin{pmatrix} C_1 & -C_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -C_1 & C_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & C_2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & C_3 & -C_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & -C_3 & C_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & C_4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & C_5 & -C_5 \\ 0 & 0 & 0 & 0 & 0 & 0 & -C_5 & C_5 \end{pmatrix}.$$

One possible selection of (r, I, J) is $r = 1, I = \{2\}$ and $J = \{1\}$.

On the substitution method, we solve $F_2 = 0$ for $\dot{x}_1$ to get

$$\dot{x}_1 = \dot{x}_2 + (-U_b/R_2 + x_2(1/R_1 + 1/R_2) - (\alpha - 1)g(x_2 - x_3))/C_1 \tag{8.29}$$

and substitute (8.29) into $F_1 = 0$. Then the first equation is modified into

$$\bar{F}_1^{\text{sub}} : -U_b/R_2 + x_2(1/R_1 + 1/R_2) - (\alpha - 1)g(x_2 - x_3) + (x_1 - U_e(t))/R_0 = 0$$

and the dual optimal solution is updated to $p' = (1, 0, 0, 0, 0, 0, 0, 1)$ and $q' = q$. The substitution method modifies the DAE twice more in the same manner for $(r, I, J) = (4, \{5\}, \{4\})$ and $(7, \{8\}, \{7\})$, and outputs the following DAE

$$\begin{cases} \bar{F}_1^{\text{sub}} : & -U_b/R_2 + x_2(1/R_1 + 1/R_2) - (\alpha - 1)g(x_2 - x_3) + (x_1 - U_e(t))/R_0 = 0, \\ F_2 \quad : & -C_1(\dot{x}_1 - \dot{x}_2) - U_b/R_2 + x_2(1/R_1 + 1/R_2) - (\alpha - 1)g(x_2 - x_3) = 0, \\ F_3 \quad : & C_2\dot{x}_3 + x_3/R_3 - g(x_2 - x_3) = 0, \\ \bar{F}_4^{\text{sub}} : & -U_b/R_6 + x_5(1/R_5 + 1/R_6) - (\alpha - 1)g(x_5 - x_6) \\ & + (x_4 - U_b)/R_4 + \alpha g(x_2 - x_3) = 0, \\ F_5 \quad : & -C_3(\dot{x}_4 - \dot{x}_5) - U_b/R_6 + x_5(1/R_5 + 1/R_6) - (\alpha - 1)g(x_5 - x_6) = 0, \\ F_6 \quad : & C_4\dot{x}_6 + x_6/R_7 - g(x_5 - x_6) = 0, \\ \bar{F}_7^{\text{sub}} : & x_8/R_9 + (x_7 - U_b)/R_8 + \alpha g(x_5 - x_6) = 0, \\ F_8 \quad : & -C_5(\dot{x}_7 - \dot{x}_8) + x_8/R_9 = 0, \end{cases} \tag{8.30}$$

which has a nonsingular system Jacobian.

The augmentation method also modifies the DAE (8.28) three times for $(r, I, J) = (1, \{2\}, \{1\})$, $(4, \{5\}, \{4\})$, and $(7, \{8\}, \{7\})$. Due to limitations of space, we just describe

the resulting DAE in the following:

$$\begin{cases} \bar{F}_1^{\mathrm{aug}}: & C_1(y_1-\xi_2)+(x_1-U_e(t))/R_0=0, \\ F_2: & -C_1(\dot{x}_1-\dot{x}_2)-U_b/R_2+x_2(1/R_1+1/R_2)-(\alpha-1)g(x_2-x_3)=0, \\ \bar{F}_{2'}^{\mathrm{aug}}: & -C_1(y_1-\xi_2)-U_b/R_2+x_2(1/R_1+1/R_2)-(\alpha-1)g(x_2-x_3)=0, \\ F_3: & C_2\dot{x}_3+x_3/R_3-g(x_2-x_3)=0, \\ \bar{F}_4^{\mathrm{aug}}: & C_3(y_4-\xi_5)+(x_4-U_b)/R_4+\alpha g(x_2-x_3)=0, \\ F_5: & -C_3(\dot{x}_4-\dot{x}_5)-U_b/R_6+x_5(1/R_5+1/R_6)-(\alpha-1)g(x_5-x_6)=0, \\ \bar{F}_{5'}^{\mathrm{aug}}: & -C_3(y_4-\xi_5)-U_b/R_6+x_5(1/R_5+1/R_6)-(\alpha-1)g(x_5-x_6)=0, \\ F_6: & C_4\dot{x}_6+x_6/R_7-g(x_5-x_6)=0, \\ \bar{F}_7^{\mathrm{aug}}: & C_5(y_7-\xi_8)+(x_7-U_b)/R_8+\alpha g(x_5-x_6)=0, \\ F_8: & -C_5(\dot{x}_7-\dot{x}_8)+x_8/R_9=0, \\ \bar{F}_{8'}^{\mathrm{aug}}: & -C_5(y_7-\xi_8)+x_8/R_9=0, \end{cases}$$

where y_1, y_4, and y_7 are new variables corresponding to $\dot{x}_1, \dot{x}_4$, and $\dot{x}_7$, respectively, and ξ_2, ξ_5, and ξ_8 are arbitrary constants corresponding to $\dot{x}_2, \dot{x}_5$, and $\dot{x}_8$, respectively.

Indeed, the LC-method can also modify the DAE (8.28) into (8.30). In general, the substitution method and the LC-method return the same DAE under some reasonable restrictions; see the appendix for details.

8.4 Experiments

We have implemented combinatorial relaxation procedure equipped with our modification methods as a MATLAB library; our library is available at [75]. The most part of our method is implemented in the MuPAD language: a core system of the Symbolic Math Toolbox in MATLAB. For the rank computation of system Jacobian and the process of finding (r, I, J), we used the function `linalg::gaussJordan` in MuPAD (equivalent to `rref` in MATLAB applied to symbolic matrices). This function is based on the fast symbolic Gaussian elimination algorithm by Sasaki–Murao [85]. For solving symbolic equations in the substitution method, our library just applies `solve` in MuPAD (the same as that of MATLAB). On executing the augmentation method, our library introduces symbols that represent the constant (τ, Ξ) in modified DAEs. The experiments are conducted on a laptop with Core i7 2.8 GHz CPU and 16 GB memory.

8.4.1 Experiment Settings

We applied our library in practice to the following four DAEs. The DAEs have identically singular system Jacobian, and thus the MS-method, which is the index reduction method adopted by MATLAB, cannot be applied to them.

(a) Nonlinearly modified pendulum (index-3):

$$\begin{cases} \dot{x}_4 - x_1 x_2 \cos x_3 = 0, \\ \dot{x}_5 - {x_2}^2 \cos x_3 \sin x_3 + g = 0, \\ {x_1}^2 + {x_2}^2 \sin^2 x_3 - 1 = 0, \\ \tanh(\dot{x}_1 - x_4) = 0, \\ \dot{x}_2 \sin x_3 + x_2 \dot{x}_3 \cos x_3 - x_5 = 0 \end{cases}$$

with parameter g. This DAE is obtained by nonlinearly changing the variable $(y, z, \lambda, v_y, v_z)$ of a simple pendulum DAE

$$\begin{cases} \dot{v}_y - y\lambda = 0, \\ \dot{v}_z - z\lambda + g = 0, \\ y^2 + z^2 - 1 = 0, \\ \dot{y} - v_y = 0, \\ \dot{z} - v_z = 0 \end{cases}$$

by $(y, z, \lambda, v_y, v_z) = (x_1, x_2 \sin x_3, x_2 \cos x_3, x_4, x_5)$. In addition, we equivalently changed the fourth equation $\dot{x}_1 - x_4 = 0$ to $\tanh(\dot{x}_1 - x_4) = 0$.

(b) Robotic arm (index-5):

$$\begin{cases} \ddot{x}_1 - 2c(x_3)(\dot{x}_1 + \dot{x}_3)^2 - {\dot{x}_1}^2 d(x_3) + (x_2 - 2x_3)(a(x_3) + 2b(x_3)) \\ \qquad - a(x_3)(x_4 - x_5) = 0, \\ \ddot{x}_2 + 2c(x_3)(\dot{x}_1 + \dot{x}_3)^2 + {\dot{x}_1}^2 d(x_3) + (x_2 - 2x_3)(1 - 3a(x_3) - 2b(x_3)) \\ \qquad + a(x_3)(x_4 - x_5) + x_5 = 0, \\ \ddot{x}_3 + 2c(x_3)(\dot{x}_1 + \dot{x}_3)^2 + {\dot{x}_1}^2 d(x_3) + (x_2 - 2x_3)(a(x_3) - 9b(x_3)) \\ \qquad + 2\dot{x}_1^2 c(x_3) + d(x_3)(\dot{x}_1 + \dot{x}_3)^2 + (a(x_3) + b(x_3))(x_1 - x_2) = 0, \\ \cos x_1 + \cos(x_1 + x_3) - p_1(t) = 0, \\ \sin x_1 + \sin(x_1 + x_3) - p_2(t) = 0, \end{cases}$$

where

$$p_1(t) = \cos(1 - \mathrm{e}^t) + \cos(1 - t),\ p_2(t) = \sin(1 - \mathrm{e}^t) + \sin(1 - t),$$
$$a(s) = \frac{2}{2 - \cos^2 s},\ b(s) = \frac{\cos s}{2 - \cos^2 s},\ c(s) = \frac{\sin s}{2 - \cos^2 s},\ d(s) = \frac{\sin s \cos s}{2 - \cos^2 s}.$$

The robotic arm DAE arises from the path control of a two-link, flexible joint and planar robotic arm [10]. The above formulation is a slightly modified version given in the preliminary paper of [94] available on arXiv.

(c) Transistor amplifier (index-1): the DAE (8.28).

(d) Ring modulator (index-2):

$$\begin{cases} \dot{x}_1 + (x_1/R - x_8 + 0.5x_{10} - 0.5x_{11} - x_{14})/C = 0, \\ \dot{x}_2 + (x_2/R - x_9 + 0.5x_{12} - 0.5x_{13} - x_{15})/C = 0, \\ x_{10} - q(U_{D1}) + q(U_{D4}) = 0, \\ x_{11} - q(U_{D2}) + q(U_{D3}) = 0, \\ x_{12} + q(U_{D1}) - q(U_{D3}) = 0, \\ x_{13} + q(U_{D2}) - q(U_{D4}) = 0, \\ \dot{x}_7 + (x_7/R_p - q(U_{D1}) - q(U_{D2}) + q(U_{D3}) + q(U_{D4}))/C_p = 0, \\ \dot{x}_8 + x_1/L_h = 0, \\ \dot{x}_9 + x_2/L_h = 0, \\ \dot{x}_{10} + (-0.5x_1 + x_3 + R_{g2}x_{10})/L_{s2} = 0, \\ \dot{x}_{11} + (0.5x_1 - x_4 + R_{g3}x_{11})/L_{s3} = 0, \\ \dot{x}_{12} + (-0.5x_2 + x_5 + R_{g2}x_{12})/L_{s2} = 0, \\ \dot{x}_{13} + (0.5x_2 - x_6 + R_{g3}x_{13})/L_{s3} = 0, \\ \dot{x}_{14} + (x_1 + (R_{g1} + R_i)x_{14} - U_{\mathrm{in1}}(t))/L_{s1} = 0, \\ \dot{x}_{15} + (x_2 + (R_c + R_{g1})x_{15})/L_{s1} = 0, \end{cases}$$

where

$$\begin{aligned} &U_{D1} = x_3 - x_5 - x_7 - U_{\mathrm{in2}}(t), \quad U_{D2} = -x_4 + x_6 - x_7 - U_{\mathrm{in2}}(t), \\ &U_{D3} = x_4 + x_5 + x_7 + U_{\mathrm{in2}}(t), \quad U_{D4} = -x_3 - x_6 + x_7 + U_{\mathrm{in2}}(t), \\ &q(U) = \gamma(\mathrm{e}^{\delta U} - 1), \quad U_{\mathrm{in1}}(t) = 0.5 \sin 2000\pi t, \quad U_{\mathrm{in2}}(t) = 2 \sin 20000\pi t \end{aligned}$$

with parameters $C, C_p, L_h, L_{s1}, L_{s2}, L_{s3}, \gamma, \delta, R, R_p, R_{g1}, R_{g2}, R_{g3}, R_i$, and R_c. The DAE represents an electrical network describing the behavior of a ring modulator [61]. The above formulation is obtained by setting $C_s = 0$ in the original problem.

8.4.2 Experimental Results

Table 8.1 shows the running time of our implementation and the size of output DAEs. Except for the substitution method applied to the DAE (c), the substitution and the augmentation methods successfully modified the DAEs (a)–(d) within 1 second. We confirmed that the MS-method is applicable to all the resulting DAEs.

The reason of freezing of the substitution method for (d) is the following. Let $F_i = 0$ be the ith equation of (d) for $i = 1, \ldots, 15$. Our library first finds the following values of

Table 8.1: Experimental results.

DAE	DAE size	Modification method	Time (sec)	Modified DAE size
(a)	5	Substitution	0.1713	5
(a)	5	Augmentation	0.2290	13
(b)	5	Substitution	0.4334	5
(b)	5	Augmentation	0.1682	8
(c)	8	Substitution	0.2767	8
(c)	8	Augmentation	0.1819	11
(d)	15	Substitution	(more than 1 hour)	—
(d)	15	Augmentation	0.5114	18

(p, q, r, I, J):

$$\begin{aligned} &p = (0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),\\ &q = (1,1,0,0,0,0,1,1,1,1,1,1,1,1,1),\\ &r = 5,\ I = \{3,4,6\},\ J = \{3,4,5\}. \end{aligned} \tag{8.31}$$

Then the substitution method requires to solve the equation system

$$\begin{cases} F_3: & x_{10} - \gamma \mathrm{e}^{\delta(x_3 - x_5 - x_7 - U_{\mathrm{in}2}(t))} + \gamma \mathrm{e}^{\delta(-x_3 - x_6 + x_7 + U_{\mathrm{in}2}(t))} = 0,\\ F_4: & x_{11} - \gamma \mathrm{e}^{\delta(-x_4 + x_6 - x_7 - U_{\mathrm{in}2}(t))} + \gamma \mathrm{e}^{\delta(x_4 + x_5 + x_7 + U_{\mathrm{in}2}(t))} = 0,\\ F_6: & x_{13} + \gamma \mathrm{e}^{\delta(-x_4 + x_6 - x_7 - U_{\mathrm{in}2}(t))} - \gamma \mathrm{e}^{\delta(-x_3 - x_6 + x_7 + U_{\mathrm{in}2}(t))} = 0 \end{cases} \tag{8.32}$$

for (x_3, x_4, x_6) and substitute back it into $F_5 = 0$ to eliminate $x_j^{(q_j)}$ for $j = 1, \ldots, 15$. As we can see, however, solving the system (8.32) for (x_3, x_4, x_6) is not an easy task; the solution cannot be represented by a combination of the elementary functions. Hence the equation-solving engine in MuPAD could not accomplish the task to solve (8.32).

Indeed, while solving (8.32) for (x_3, x_4, x_6) is complicated, the modified 5th equation $\bar{F}_5^{\mathrm{sub}} = 0$ is quite simple; it coincides with the sum of the 3–6th equations, i.e.,

$$\bar{F}_5^{\mathrm{sub}} : x_{10} + x_{11} + x_{12} + x_{13} = 0.$$

Detecting and utilizing such a simple dependence structure is left for future investigation.

On the DAE (d), there exists another possible values of (p, q, r, I, J) as follows:

$$\begin{aligned} &p = (0,0,1,1,1,1,0,0,0,0,0,0,0,0,0),\\ &q = (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1),\\ &r = 11,\ I = \{3,4,5,6,10,12,13\},\ J = \{3,5,6,10,11,12,13\}. \end{aligned} \tag{8.33}$$

Since the equation system corresponding to (8.33) is linear, the substitution method would have gone on if our library had chosen not (8.31) but (8.33). This means that the success of the substitution method depends on the choice of (p, q, r, I, J). The experimental result shows that the augmentation method successfully serves as a remedy for this issue.

Chapter 9

Conclusion

In this thesis, we have considered computations of valuations of Dieudonné determinants and modifications of differential-algebraic equations (DAEs). We now conclude this thesis with a brief summary and discussions of prospective research directions.

In Chapter 4, we have presented two efficient algorithms to compute valuations of the Dieudonné determinants of matrices over split DVSFs. Both algorithms, the combinatorial relaxation and matrix expansion algorithms, are based on combinatorial optimization theory. We have shown that skew inverse Laurent fields are the most general DVSFs for which these algorithms are naturally applicable, in that the valuation of the Dieudonné determinant of a matrix admits a trivial upper bound if and only if the matrix is over a skew inverse Laurent field.

In Chapter 5, we have given two applications our algorithms for the weighted Edmonds' problem (WEP) and for linear differential/difference equations. In particular, the matrix expansion algorithm yields the first deterministic polynomial-time algorithm for the non-commutative WEP with polynomially bounded bit complexity when the base field is $\mathbb{Q}$, and is also applicable to the reduction of commutative problems. We have also shown that the dimension of the solution spaces of linear differential and difference equations can be calculated from degrees/orders of the Dieudonné determinants of skew polynomial matrices.

Chapters 6–8 have dealt with modification methods for DAEs. To give a consistent initial value and reduce the differentiation index of DAEs, most of the existing software libraries provide structural preprocessing methods based on the assignment problem. The structural methods, however, fail if the DAE has a singular system Jacobian. We thus consider modifying a DAE into an equivalent DAE whose system Jacobian is nonsingular. The combinatorial relaxation framework can be used for this modification, in which one needs to modify DAEs preserving their solution sets. For linear DAEs, we can use the combinatorial relaxation algorithm by Murota [67] that uses unimodular transformations. Tan et al. [94] generalized this to nonlinear DAEs, while their algorithms are rather limited to almost linear DAEs.

In Chapter 7, we have proposed a modification algorithm for linear DAEs whose coefficient matrices are mixed matrices. Technically, we have presented a combinatorial relaxation algorithm for LM-polynomial matrices that uses only unimodular transformations. Since mixed matrices represent physical quantities as independent parameters, one can avoid issues arising from measurement or numerical errors. Though our algorithm deals with matrices containing independent symbols, it does not depend on symbolic computation by making use of graph and matroid algorithms. We have also developed a faster algorithm for DAEs whose coefficient matrices are consistent with dimensional analysis. We also have confirmed through numerical experiments that our algorithm runs sufficiently fast for large scale DAEs.

A limitation of our algorithm is that it can handle only time-invariant systems. Generalizing mixed matrix theory to a time-varying setting is a promising future direction. Representing the differential operator as an indeterminate of skew polynomials, one can regard a time-varying linear DAE as an equation whose coefficient is a skew polynomial matrix. We can also consider adjoining independent parameters into skew polynomials, which result in a matrix something like "mixed skew polynomial matrix". There must be much work left for this type of matrices, such as giving efficient algorithms to compute characteristic quantities like the degrees of the Dieudonné determinants, properly extending the dimensional consistency reflecting the dimension of the differential operator, and of course, devising modification algorithms for a linear DAE having the matrix as the coefficient.

In Chapter 8, we have presented two modification methods for nonlinear DAEs, called the substitution method and the augmentation method. Using a symbolic computation engine as a black-box, both methods modify DAEs into other DAEs for which the structural preprocessing methods work. Both methods can be applied to "highly nonlinear DAEs" that the existing modification methods of Tan et al. [94] cannot handle. The substitution method modifies DAEs based on the implicit function theorem and has a merit that it retains the size of DAEs. The augmentation method modifies DAEs by appending new variables and equations, and is advantageous in that it does not require an equation-solving engine and keeps DAEs' sparsity.

Both (and all existing) methods cannot deal with DAEs which are nonsmooth or with (F1) or (F2); modifying such DAEs seems to require a new approach other than combinatorial relaxation. In addition, all methods require the symbolic Gaussian elimination for computing the rank of system Jacobian. In theory, this computation, or more specifically, testing if a mathematical formula is identically zero, is undecidable, i.e., there is no algorithm to solve the problem [83]. Symbolic computation engines implement heuristic algorithms for the zero testing problem, and they tend to get drastically slow or unreliable with respect to the growth of the size of mathematical formulas. Thus applications of both methods are limited to middle-sized DAEs. To keep parts in matrices that are eliminated smaller, dividing a DAE system into small subsystems according to the system Jacobian's

structure would be a promising future direction.

Bibliography

[1] S. A. Abramov and M. A. Barkatou. On solution spaces of products of linear differential or difference operators. *ACM Communications in Computer Algebra*, 48(4):155–165, 2014 (cited on pages 55, 63, 66).

[2] S. A. Amitsur. Rational identities and applications to algebra and geometry. *Journal of Algebra*, 3(3):304–359, 1966 (cited on page 4).

[3] B. Beckermann, H. Cheng, and G. Labahn. Fraction-free row reduction of matrices of Ore polynomials. *Journal of Symbolic Computation*, 41(5):513–543, 2006 (cited on page 55).

[4] K. E. Brenan, S. L. Campbell, and L. R. Petzold. *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*. SIAM, Philadelphia, 1996 (cited on pages 8, 73).

[5] M. Bronstein. On solutions of linear ordinary differential equations in their coefficient field. *Journal of Symbolic Computation*, 29(6):841–877, 2000 (cited on page 63).

[6] M. Bronstein and M. Petkovšek. An introduction to pseudo-linear algebra. *Theoretical Computer Science*, 157(1):3–33, 1996 (cited on page 63).

[7] H. H. Brungs. Left Euclidean rings. *Pacific Journal of Mathematics*, 45(1):27–33, 1973 (cited on page 15).

[8] H. H. Brungs and G. Törner. Skew power series rings and derivations. *Journal of Algebra*, 87(2):368–379, 1984 (cited on page 40).

[9] S. L. Campbell and C. W. Gear. The index of general nonlinear DAEs. *Numerische Mathematik*, 72(2):173–196, 1995 (cited on pages 8, 71).

[10] S. L. Campbell and E. Griepentrog. Solvability of general differential algebraic equations. *SIAM Journal on Scientific Computing*, 16(2):257–270, 1995 (cited on page 128).

[11] S. Chowdhry, H. Krendl, and A. A. Linninger. Symbolic numeric index analysis algorithm for differential algebraic equations. *Industrial & Engineering Chemistry Research*, 43:3886–3894, 2004 (cited on page 108).

[12] G. Chrystal. A fundamental theorem regarding the equivalence of systems of ordinary linear differential equations, and its application to the determination of the order and the systematic solution of a determinate system of such equations. *Transactions of the Royal Society of Edinburgh*, 38(1):163–178, 1897 (cited on pages 5, 78).

[13] I. S. Cohen. On the structure and ideal theory of complete local rings. *Transactions of the American Mathematical Society*, 59(1):54, 1946 (cited on page 38).

[14] P. M. Cohn. *Skew Field Constructions*. London Mathematical Society Lecture Note Series. Cambridge University Press, Cambridge, 1977 (cited on page 40).

[15] P. M. Cohn. *Free Rings and Their Relations*, volume 19 of *London Mathematical Society Monograph*. Academic Press, London, 2nd edition, 1985 (cited on page 20).

[16] P. M. Cohn. *Skew Fields: Theory of General Division Rings*, volume 57 of *Encyclopedia of Mathematics and Its Applications*. Cambridge University Press, Cambridge, 1995 (cited on pages 18, 19, 23).

[17] P. M. Cohn. *Further Algebra and Applications*. Springer, London, 2003 (cited on pages 19, 21, 27).

[18] W. H. Cunningham. Improved bounds for matroid partition and intersection algorithms. *SIAM Journal on Computing*, 15:948–957, 1986 (cited on pages 98, 100).

[19] J. Dieudonné. Les déterminants sur un corps non commutatif. *Bulletin de la Société Mathématique de France*, 71:27–45, 1943 (cited on pages 1, 2, 21).

[20] A. W. M. Dress and W. Wenzel. Valuated matroids: a new look at the greedy algorithm. *Applied Mathematics Letters*, 3(2):33–35, 1990 (cited on pages 32, 33).

[21] A. W. M. Dress and W. Wenzel. Valuated matroids. *Advances in Mathematics*, 93(2):214–250, 1992 (cited on pages 32, 33).

[22] F. Dumas. Skew power series rings with general commutation formula. *Theoretical Computer Science*, 98(1):99–114, 1992 (cited on pages 10, 38, 40, 49).

[23] J. Edmonds. Systems of distinct representatives and linear algebra. *Journal of Research of the National Bureau of Standards*, 71B(4):241–245, 1967 (cited on pages 1, 3, 59).

[24] J. Edmonds. Matroid partition. In G. B. Dantzig and A. F. Veinott, Jr., editors, *Mathematics of the Decision Sciences: Part I*. Volume 11, Lectures in Applied Mathematics, pages 335–345. AMS, Providence, RI, 1968 (cited on pages 32, 87).

[25] J. Edmonds. Submodular functions, matroids, and certain polyhedra. In R. Guy, H. Hanani, N. Sauer, and J. Schönheim, editors, *Combinatorial Structures and Their Applications*, pages 69–87. Gordon and Breach, New York, NY, 1970 (cited on page 32).

[26] S. Elliger. Potenzbasiserweiterungen. *Journal of Algebra*, 7(2):254–262, 1967 (cited on page 39).

[27] M. Fortin and C. Reutenauer. Commutative/noncommutative rank of linear matrices and subspaces of matrices of low rank. *Séminaire Lotharingien de Combinatoire*, 52, 2004 (cited on pages 2, 59).

[28] A. Garg, L. Gurvits, R. Oliveira, and A. Wigderson. A deterministic polynomial time algorithm for non-commutative rational identity testing. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science (FOCS '16)*, pages 109–117, 2016 (cited on page 4).

[29] C. W. Gear. Simultaneous numerical solution of differential-algebraic equations. *IEEE Transactions on Circuit Theory*, 18(1):89–95, 1971 (cited on pages 7, 68).

[30] C. W. Gear, H. H. Hsu, and L. R. Petzold. Differential-algebraic equations revisited. In *Proceedings of the Numerical Methods for Solving Stiff Initial Value Problems*, Oberwolfach, 1981 (cited on page 73).

[31] J. F. Geelen, S. Iwata, and K. Murota. The linear delta-matroid parity problem. *Journal of Combinatorial Theory. Series B*, 88(2):377–398, 2003 (cited on page 3).

[32] M. Giesbrecht and M. S. Kim. Computing the Hermite form of a matrix of Ore polynomials. *Journal of Algebra*, 376:341–362, 2013 (cited on page 55).

[33] K. R. Goodearl and R. B. Warfield, Jr. *An Introduction to Noncommutative Noetherian Rings*. Cambridge University Press, Cambridge, second edition, 2004 (cited on pages 15, 18, 19, 54).

[34] A. Griewank. On automatic differentiation. In M. Iri and K. Tanabe, editors, *Mathematical Programming: Recent Developments and Applications*, pages 83–108, Dordrecht. Kluwer Academic Publishers, 1989 (cited on page 74).

[35] L. Gurvits. Classical complexity and quantum entanglement. *Journal of Computer and System Sciences*, 69(3):448–484, 2004 (cited on pages 5, 60).

[36] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 1996 (cited on pages 8, 73).

[37] M. Hamada and H. Hirai. Computing the nc-rank via discrete convex optimization on CAT(0) spaces, 2020. arXiv: `2012.13651` (cited on pages 4, 5, 60).

[38] M. M. Hezavehi. Matrix valuations and their associated skew fields. *Results in Mathematics*, 5(1-2):149–156, 1982 (cited on pages 22, 23).

[39] H. Hirai. Computing the degree of determinants via discrete convex optimization on Euclidean buildings. *SIAM Journal on Applied Geometry and Algebra*, 3(3):523–557, 2019 (cited on pages 3–5, 33, 59, 60).

[40] H. Hirai and M. Ikeda. A cost-scaling algorithm for computing the degree of determinants, 2020. arXiv: 2008.11388 (cited on pages 5, 61).

[41] H. Hirai and Y. Iwamasa. A combinatorial algorithm for computing the rank of a generic partitioned matrix with 2×2 submatrices. In D. Bienstock and G. Zambelli, editors, *Proceedings of the 21st Conference on Integer Programming and Combinatorial Optimization (IPCO '20)*, volume 12125 of *Lecture Notes in Computer Science*, pages 196–208, Cham. Springer, 2020 (cited on page 3).

[42] J. E. Hopcroft and R. M. Karp. An $n^{5/2}$ algorithm for maximum matchings in bipartite graphs. *SIAM Journal on Computing*, 2:225–231, 1973 (cited on pages 28, 48).

[43] G. Ivanyos, Y. Qiao, and K. V. Subrahmanyam. Constructive non-commutative rank computation is in deterministic polynomial time. *Computational Complexity*, 27(4):561–593, 2018 (cited on pages 4, 5, 60).

[44] G. Ivanyos, M. Karpinski, and N. Saxena. Deterministic polynomial time algorithms for matrix completion problems. *SIAM Journal on Computing*, 39(8):3736–3751, 2010 (cited on page 3).

[45] S. Iwata and R. Shimizu. Combinatorial analysis of generic matrix pencils. *SIAM Journal on Matrix Analysis and Applications*, 29(1):245–259, 2007 (cited on page 50).

[46] S. Iwata and M. Takamatsu. Computing the maximum degree of minors in mixed polynomial matrices via combinatorial relaxation. *Algorithmica*, 66(2):346–368, 2013 (cited on pages 4, 7, 11, 89).

[47] S. Iwata. Computing the maximum degree of minors in matrix pencils via combinatorial relaxation. *Algorithmica*, 36(4):331–341, 2003 (cited on pages 7, 10, 82).

[48] S. Iwata and M. Takamatsu. Index reduction via unimodular transformations. *SIAM Journal on Matrix Analysis and Applications*, 39(3):1135–1151, 2018 (cited on page 92).

[49] N. Jacobson. *The Theory of Rings*, volume 2 of *Mathematical Surveys and Monographs*. AMS, Providence, RI, 1943 (cited on page 27).

[50] V. Kabanets and R. Impagliazzo. Derandomizing polynomial identity tests means proving circuit lower bounds. *Computational Complexity*, 13(1–2):1–46, 2004 (cited on page 3).

[51] M. Khochtali, J. Rosenkilde né Nielsen, and A. Storjohann. Popov form computation for matrices of Ore polynomials. In *Proceedings of the 42nd International Symposium on Symbolic and Algebraic Computation (ISSAC '17)*, pages 253–260, New York, NY. ACM Press, 2017 (cited on page 55).

[52] D. König. Gráfok és mátrixok. *Matematikai és Fizikai Lapok*, 38:116–119, 1931 (cited on page 28).

[53] B. Korte and J. Vygen. *Combinatorial Optimization*, volume 21 of *Algorithms and Combinatorics*. Springer, Berlin, 4th edition, 2008 (cited on page 92).

[54] P. A. Krylov and A. A. Tuganbaev. *Modules over Discrete Valuation Domains*, volume 145 of number 4 in *de Gruyter Expositions in Mathematics*. Walter de Gruyter, Berlin, 2008, pages 4997–5117 (cited on pages 14, 15).

[55] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955 (cited on pages 28, 29, 47, 91, 97).

[56] P. Kunkel and V. Mehrmann. *Differential-Algebraic Equations: Analysis and Numerical Solution*. EMS Textbooks in Mathematics. European Mathematical Society, Zürich, 2006, page 392 (cited on page 73).

[57] T. Y. Lam. *Lectures on Modules and Rings*, volume 189 of *Graduate Texts in Mathematics*. Springer, New York, NY, 1999 (cited on page 20).

[58] V. Levandovskyy and K. Schindelar. Computing diagonal form and Jacobson normal form of a matrix using Gröbner bases. *Journal of Symbolic Computation*, 46(5):595–608, 2011 (cited on page 55).

[59] L. Lovász. Singular spaces of matrices and their application in combinatorics. *Boletim da Sociedade Brasileira de Matemática*, 20(1):87–99, 1989 (cited on page 3).

[60] S. E. Mattsson and G. Söderlind. Index reduction in differential-algebraic equations using dummy derivatives. *SIAM Journal on Scientific Computing*, 14(3):677–692, 1993 (cited on pages 9, 78, 79).

[61] F. Mazzia and C. Magherini. Test set for initial value problem solvers. Technical report, Department of Mathematics, University of Bari, 2008. URL: `https://archimede.dm.uniba.it/~testset/`. (cited on pages 10, 80, 110, 125, 129).

[62] S. Moriyama and K. Murota. Discrete Legendre duality in polynomial matrices (in Japanese). *The Japan Society for Industrial and Applied Mathematics*, 23(2):183–202, 2013 (cited on page 10).

[63] K. Murota. Use of the concept of physical dimensions in the structural approach to systems analysis. *Japan Journal of Applied Mathematics*, 2(2):471–494, 1985 (cited on page 87).

[64] K. Murota. Computing Puiseux-series solutions to determinantal equations via combinatorial relaxation. *SIAM Journal on Computing*, 19(6):1132–1161, 1990 (cited on page 6).

[65] K. Murota. Mixed matrices: irreducibility and decomposition. In R. A. Brualdi, S. Friedland, and V. Klee, editors, *Combinatorial and Graph-Theoretical Problems in Linear Algebra*. Volume 50, The IMA Volumes in Mathematics and its Applications, pages 39–71. Springer, New York, NY, 1993 (cited on page 4).

[66] K. Murota. Combinatorial relaxation algorithm for the maximum degree of subdeterminants: Computing Smith-McMillan form at infinity and structural indices in Kronecker form. *Applicable Algebra in Engineering, Communication and Computing*, 6(4–5):251–273, 1995 (cited on pages 7, 82).

[67] K. Murota. Computing the degree of determinants via combinatorial relaxation. *SIAM Journal on Computing*, 24(4):765–796, 1995 (cited on pages 1, 6, 7, 10, 30, 42, 43, 82, 132).

[68] K. Murota. Finding optimal minors of valuated bimatroids. *Applied Mathematics Letters*, 8(4):37–41, 1995 (cited on pages 34–36).

[69] K. Murota. On the degree of mixed polynomial matrices. *SIAM Journal on Matrix Analysis and Applications*, 20(1):196–227, 1998 (cited on page 4).

[70] K. Murota. *Discrete Convex Analysis*. SIAM, Philadelphia, 2003 (cited on pages 10, 32, 36).

[71] K. Murota. *Matrices and Matroids for Systems Analysis*, volume 20 of *Algorithms and Combinatorics*. Springer, Berlin, 2010 (cited on pages 4, 24, 31, 33, 69, 87, 88).

[72] K. Murota and M. Iri. Structural solvability of systems of equations ——A mathematical formulation for distinguishing accurate and inaccurate numbers in structural analysis of systems—. *Japan Journal of Applied Mathematics*, 2:247–271, 1985 (cited on pages 3, 4, 85).

[73] K. Murota, M. Iri, and M. Nakamura. Combinatorial canonical form of layered mixed matrices and its application to block-triangularization of systems of linear/nonlinear equations. *SIAM Journal on Algebraic and Discrete Methods*, 8:123–149, 1987 (cited on pages 87, 89, 98, 100).

[74] B. H. Neumann. On ordered division rings. *Transactions of the American Mathematical Society*, 66(1):202, 1949 (cited on page 18).

[75] T. Oki. OptMist-Tokyo/DAEPreprocessingToolbox: MATLAB toolbox for preprocessing of differential-algebraic equations (DAEs). URL: `https://github.com/OptMist-Tokyo/DAEPreprocessingToolbox`. (accessed August 8, 2020) (cited on page 127).

[76] O. Ore. Theory of non-commutative polynomials. *Annals of Mathematics*, 34(3):480–508, 1933 (cited on pages 6, 18).

[77] O. Ore. Graphs and matching theorems. *Duke Mathematical Journal*, 22(4):625–639, 1955 (cited on page 29).

[78] J. G. Oxley. *Matroid Theory*. Oxford Graduate Texts in Mathematics. Oxford University Press, New York, NY, second edition, 2011. ISBN: 9780198566946 (cited on pages 31, 87).

[79] C. C. Pantelides. The consistent initialization of differential-algebraic systems. *SIAM Journal on Scientific and Statistical Computing*, 9(2):213–231, 1988 (cited on pages 8, 9, 75).

[80] K. Paykan and A. Moussavi. Study of skew inverse Laurent series rings. *Journal of Algebra and Its Applications*, 16(12):1750221, 2017 (cited on page 19).

[81] L. R. Petzold. A description of DASSL: a differential / algebraic system solver. In *Proceedings of the 10th IMACS World Congress on System Simulation and Scientific Computation*, pages 3–7, Montreal, 1982 (cited on page 73).

[82] J. D. Pryce. A simple structural analysis method for DAEs. *BIT Numerical Mathematics*, 41(2):364–394, 2001 (cited on pages 9, 74–78, 80).

[83] D. Richardson. Some undecidable problems involving elementary functions of a real variable. *The Journal of Symbolic Logic*, 33(4):514–520, 1968 (cited on page 133).

[84] B. Roux. Anneaux non commutatifs de valuation discrète ou finie. *Comptes Rendus de l'Académie des Sciences, Série I*, 302(9):259–262 and 291–293, 1986 (cited on pages 10, 38, 39).

[85] T. Sasaki and H. Murao. Efficient Gaussian elimination method for symbolic determinants and linear systems. *ACM Transactions on Mathematical Software*, 8(3):277–289, 1982 (cited on page 127).

[86] L. Scholz. The signature method for DAEs arising in the modeling of electrical circuits. *Journal of Computational and Applied Mathematics*, 332:107–139, 2018 (cited on page 80).

[87] A. Schrijver. *Combinatorial Optimization. Polyhedra and Efficiency*, volume 24 of *Algorithms and Combinatorics*. Springer, Berlin, 2003 (cited on pages 28, 29).

[88] J. T. Schwartz. Fast probabilistic algorithms for verification of polynomial identities. *Journal of the ACM*, 27(4):701–717, 1980 (cited on page 3).

[89] L. F. Shampine. Solving $0 = F(t, y(t), y'(t))$ in Matlab. *Journal of Numerical Mathematics*, 10(4):291–310, 2002 (cited on page 8).

[90] M. F. Singer. Algebraic and algorithmic aspects of linear difference equations. In *Galois Theories of Linear Difference Equations: An Introduction*. Volume 211, Mathematical Surveys and Monograph, pages 1–41. AMS, Providence, RI, 2016 (cited on page 63).

[91] T. H. M. Smits. Skew polynomial rings. *Indagationes Mathematicae*, 30(1):209–224, 1968 (cited on page 39).

[92] T. Soma. Fast deterministic algorithms for matrix completion problems. *SIAM Journal on Discrete Mathematics*, 28(1):490–502, 2014 (cited on page 3).

[93] L. Taelman. Dieudonné determinants for skew polynomial rings. *Journal of Algebra and Its Applications*, 5(1):89–93, 2006 (cited on pages 6, 22, 63).

[94] G. Tan, N. S. Nedialkov, and J. D. Pryce. Conversion methods for improving structural analysis of differential-algebraic equation systems. *BIT Numerical Mathematics*, 57(3):845–865, 2017. arXiv: `1505.03445` (cited on pages 10, 81, 83, 84, 104, 128, 132, 133).

[95] L. G. Valiant. Completeness classes in algebra. In *Proceedings of the 11th Annual ACM Symposium on Theory of Computing (STOC '79)*, pages 249–261, New York, NY. ACM Press, 1979 (cited on page 3).

[96] M. van der Put and M. F. Singer. *Galois Theory of Difference Equations*, volume 1666 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1997. ISBN: 978-3-540-63243-6 (cited on page 66).

[97] M. van der Put and M. F. Singer. *Galois Theory of Linear Differential Equations*, volume 328 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, 2003 (cited on pages 63, 66).

[98] P. M. Van Dooren, P. Dewilde, and J. Vandewalle. On the determination of the Smith-Macmillan form of a rational matrix from its Laurent expansion. *IEEE Transactions on Circuits and Systems*, 26(3):180–189, 1979 (cited on pages 10, 49).

[99] G. C. Verghese and T. Kailath. Rational matrix structure. *IEEE Transactions on Automatic Control*, 26(2):434–439, 1981 (cited on page 24).

[100] R. Vidal. Anneaux de valuation discrète complets non commutatifs. *Transactions of the American Mathematical Society*, 267(1):65–81, 1981 (cited on page 38).

[101] S. Warner. *Topological Rings*. L. Nachbin, editor, volume 178 of *North-Holand Mathematics Studies*. Elsevier, North Holland, 1993 (cited on pages 2, 13, 16).

[102] Wolfram Research, Inc. Numerical Solution of Differential-Algebraic Equations — Wolfram Language Documentation. 2017. URL: `http://reference.wolfram.com/language/tutorial/NDSolveDAE.html` (cited on page 108).

[103] X. Wu, Y. Zeng, and J. Cao. The application of the combinatorial relaxation theory on the structural index reduction of DAE. In *Proceedings of the 12th International Symposium on Distributed Computing and Applications to Business, Engineering & Science*, pages 162–166, London. IEEE, 2013 (cited on pages 10, 82).