Overcoming Degeneracy and Singularity: Techniques for Semidefinite Programs and Homotopy Continuation Endgames

by

Jeferson Leon Zapata Nieto

May, 2026

A thesis submitted to the
Graduate School
of the
Institute of Science and Technology Austria
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Committee in charge:

Matthew Kwan, Chair
Vladimir Kolmogorov
Uli Wagner
Simone Naldi
Gábor Pataki

Institute of
Science and
Technology
Austria

The thesis of Jeferson Leon Zapata Nieto, titled *Overcoming Degeneracy and Singularity: Techniques for Semidefinite Programs and Homotopy Continuation Endgames*, is approved by:

Supervisor:

Vladimir Kolmogorov, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member:

Uli Wagner, ISTA, Klosterneuburg, Austria

Signature: _____

Committee Member:

Simone Naldi, Université de Limoges, Limoges, France

Signature: _____

Committee Member:

Gábor Pataki, University of North Carolina at Chapel Hill, USA

Signature: _____

Defense Chair:

Matthew Kwan, ISTA, Klosterneuburg, Austria

Signature: _____

Signed page is on file

ISTA Thesis, ISSN: 2663-337X

ISBN: 978-3-99078-079-4

I hereby declare that this thesis is my own work and that it does not contain other people's work without this being so stated; this thesis does not contain my previous work without this being stated, and the bibliography contains all the literature that I used in writing the dissertation.

I accept full responsibility for the content and factual accuracy of this work, including the data and their analysis and presentation, and the text and citation of other work.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee, and that this thesis has not been submitted for a higher degree to any other university or institution.

I certify that any republication of materials presented in this thesis has been approved by the relevant publishers and co-authors.

Signature: _____

Jeferson Leon Zapata Nieto
May, 2026

Signed page is on file

Abstract

This thesis investigates algorithmic certification and approximation methods for degenerate semidefinite programs (SDPs) and the singular roots of polynomial systems. In the first part, we present a hybrid symbolic-numeric algorithm for certifying the feasibility of weakly feasible, degenerate SDPs. By reformulating linear matrix inequalities (LMIs) into a structured polynomial system via facial reduction and incidence varieties, we guarantee the existence of an isolated exact solution. This algebraic reduction enables the certification of maximum-rank numerical approximations using methods from algebraic geometry.

In the second part, we address the severe ill-conditioning and loss of quadratic convergence that plague standard path-tracking methods near isolated singular roots. To overcome this, we propose tracking algorithms that achieve superlinear convergence without the computational bloat characteristic of classical deflation techniques. By modeling the solution path as a generalized fractional Puiseux series, our approach combines an explicitly derived algebraic predictor with a localized hyperplane desingularization phase during the corrector step. Furthermore, we introduce a continuous path-limit method and an extension of the geometric sequence rule to directly extract exact fractional exponents. This bypasses traditional heuristic trial-and-error methods and explicitly accommodates sparse series expansions. Numerical experiments confirm that our method significantly reduces the cumulative number of matrix inversions while achieving high-accuracy root approximations, even for heavily degenerate systems exhibiting higher coranks.

About the Author

Jeferson Zapata completed his B.Sc. (2016) and M.Sc. (2019) in Mathematics at the Universidad Nacional de Colombia in Medellín. In 2019, he joined the Institute of Science and Technology Austria (ISTA), working within Vladimir Kolmogorov's research group. His research interests lie at the intersection of semidefinite programming and algebraic geometry, with a focus on algorithmic challenges and certification methods for degenerate and singular problems.

During his time at ISTA, Jeferson contributed to the development of symbolic–numeric frameworks. His work on the exact certification of degenerate semidefinite programs was published in the *SIAM Journal on Optimization* (2025). He also developed an adaptive power-series endgame for tracking singular polynomial roots by systematically navigating sparse Puiseux series. He has presented his research at international venues, including the International Symposium on Mathematical Programming (ISMP) and meetings of the Vienna Graduate School on Computational Optimization (VGSCO).

List of Collaborators and Publications

This thesis contains the following publications:

1. Vladimir Kolmogorov, Simone Naldi, and Jeferson Zapata (2025). "Certifying Solutions of Degenerate Semidefinite Programs." In: *SIAM Journal on Optimization* 35.3, pp. 1630–1654. DOI: 10.1137/24M1664691. (Incorporated as Chapter 2).
2. Mikhail Karapetyants, Vladimir Kolmogorov, and Jeferson Zapata (2026). *Computing singular solutions of polynomial systems*. In preparation. (Incorporated as Chapter 3).

The authors on the mentioned research articles are listed alphabetically. This reflects common practice in Mathematics, where joint research is a sharing of ideas and skills that cannot be attributed to the individuals separately.

Table of Contents

List of Figures

List of Tables

List of Algorithms

List of Abbreviations

AL ArcLength Endgame. 2, 66–68, 70, 81

CPM Critical Point Method. 19–24, 28

IPM Interior-Point Method. 1, 4, 22, 29

LAL Lifted ArcLength Endgame. 2, 53, 67, 70, 81, 82

LP Linear Programming. 3, 29

NAG Numerical Algebraic Geometry. 1, 4, 6, 22, 28, 35, 53, 69, 81

PSD Positive Semidefinite. 3–5, 7, 11, 13, 22, 33

SVD Singular Value Decomposition. 2, 43, 68

LMI Linear Matrix Inequality. 1, 3, 4, 9, 10, 19, 21, 22, 29, 81

SDP Semidefinite Programming. v, 1–8, 11, 22–24, 28, 29, 33, 81

Chapter 1

Introduction

Degeneracy represents a fundamental computational frontier in both continuous optimization and Numerical Algebraic Geometry (NAG). In well-posed problems, standard regularity conditions—such as strict feasibility in optimization or full-rank Jacobians in equation solving—guarantee the convergence of classical numerical algorithms. However, when the underlying geometry exhibits degeneracy, these guarantees cease to apply. In convex optimization, degenerate Semidefinite Programs (SDPs) and Linear Matrix Inequalities (LMIs) cause Interior-Point Method (IPM) solvers to stall or produce inaccurate outputs. Similarly, in NAG, isolated singular roots of polynomial systems force the Jacobian matrix to lose rank, causing Newton's method to lose its quadratic convergence and standard path trackers to fail due to severe ill-conditioning.

This thesis explores the mathematical and computational resolution of such degenerate geometries through two independent, yet theoretically complementary, research projects. The first project addresses the rigorous certification of degenerate SDPs via algebraic geometry, while the second develops efficient numerical endgames to compute the singular roots of degenerate polynomial systems that arise in such applications.

The first part of this thesis addresses the algorithmic challenge of solving weakly feasible semidefinite programming problems. Since feasible solutions in degenerate SDP instances often possess purely irrational entries, numerical solvers operating with floating-point arithmetic can only extract approximate solutions. Existing certification approaches typically rely on the assumption that an exact rational feasible solution exists, utilizing rounding and lattice reduction techniques.

In Chapter 2, we propose a hybrid (symbolic-numeric) alternative that bypasses the assumption of rational feasibility. We demonstrate how to construct a specialized system of polynomial equations whose set of real solutions is geometrically guaranteed to feature an exact, maximum-rank SDP solution as an isolated point. By translating the LMI feasibility problem into the domain of real algebraic geometry, this framework provides a pathway to certify numerical SDP solutions even in the presence of structural degeneracy in the feasible set.

It is important to clarify the terminology regarding degeneracy in this work. In the seminal work of Alizadeh et al. (1997), primal and dual degeneracy in SDPs are defined as local properties of a solution point, specifically relating to the transversality of the affine constraint space and the faces of the positive semidefinite cone. In contrast, this thesis adopts the terminology of Drusvyatskiy and Wolkowicz (2017), using "degeneracy" to describe the global structural property of weak feasibility—namely, the failure of Slater's condition and the absence of a strictly positive definite matrix in the feasible set. While these structural degeneracies often imply a loss of strict complementarity and Alizadeh-degeneracy at the optimum, our focus remains on resolving the geometric ill-conditioning of the feasible set itself.

The algebraic formulations derived in Chapter 2—and similar hybrid optimization techniques—inherently yield degenerate polynomial systems. To solve these systems via numerical homotopy continuation, solvers must rely on specialized techniques to track paths toward isolated singular roots. Currently, state-of-the-art methods are forced to choose between classical power-series endgames, which fail to capture the sparse geometry of the root, or deflation methods, which restore convergence at the cost of computational bloat and system inflation.

In Chapter 3, we propose an approach to achieve superlinear convergence without the heavy computational burden of deflation. We introduce ArcLength Endgame (AL) and Lifted ArcLength Endgame (LAL), which locally restore full rank by dynamically augmenting the Jacobian matrix with tangent and null-space hyperplanes derived via a Singular Value Decomposition (SVD). Utilizing explicitly derived algebraic predictors based on Puiseux series, our method respects sparse exponent gaps. Furthermore, we replace traditional heuristic trial-and-error techniques with quotient rules—namely, a continuous path-limit rule and an extended geometric sequence rule—to directly and accurately isolate the fractional exponents governing the singular path.

Chapter 2

Certifying solutions of degenerate semidefinite programs

This chapter contains the paper: Vladimir Kolmogorov, Simone Naldi, and Jeferson Zapata (2025). "Certifying Solutions of Degenerate Semidefinite Programs." In: *SIAM Journal on Optimization* 35.3, pp. 1630–1654. DOI: 10.1137/24M1664691

2.1 Introduction

Semidefinite Programming (SDP) is a nontrivial generalization of Linear Programming (LP) that involves minimizing a linear function over the space of real symmetric matrices, constrained to the Positive Semidefinite (PSD) cone. Despite this relatively simple formulation, the complexity analysis of SDP and its feasibility problem—defined by Linear Matrix Inequalities (LMIs)—remains a largely open question. Consequently, the development of efficient and reliable algorithms continues to be a central focus of research in convex optimization.

Beyond its interest as a conic optimization problem, the importance of SDP is underscored by its numerous applications, the most classical of which is perhaps the semidefinite relaxation of MAX-CUT by Goemans and Williamson (1995). More generally, SDP is used as a numerical tool for solving hard non-convex optimization problems through the so-called moment-SOS hierarchy (Henrion et al. 2020), which relaxes polynomial optimization to a sequence of SDP problems of increasing size with good convergence properties (Lasserre 2001). Specific applications include checking the stability of systems in control theory (Papachristodoulou and Prajna 2005) and analyzing the convergence rate of numerical algorithms for convex optimization (Drori and Teboulle 2014). In both cases, one needs

to find a Lyapunov function, and the search for such a function is cast as a small-scale SDP.

Unfortunately, naively relying on numerical SDP solvers may lead to incorrect conclusions; examples involving small dynamical systems can be found in Roux et al. (2018). This motivates the need for algorithms capable of certifying the feasibility of a given SDP. Below, we discuss two classes of such algorithms: *symbolic* methods, which compute an exact solution in a *Rational Univariate Representation* using techniques from real algebraic geometry; and *symbolic-numerical* (or *hybrid*) methods, which certify feasibility by employing a numerical SDP solver as a subroutine.

Symbolic methods. Since the constraints in an SDP are nonlinear in the components of the matrix yet algebraic (defined by polynomial inequalities), the solution is generally not rational but rather defined over an algebraic extension of the base field, which we usually assume to be $\mathbb{Q}$. The degree of this extension, known as the algebraic degree of the SDP, represents a measure of the intrinsic complexity of the program, for which exact formulas are known (Nie et al. 2010).

The underlying algebraic structure of SDP has motivated, in recent years, the development of computer algebra algorithms whose arithmetic complexity is essentially quadratic in a multilinear bound on the aforesaid algebraic degree (Henrion et al. 2016; Naldi 2015). The strategy employed by Henrion et al. (2015b), Henrion et al. (2016), Henrion et al. (2015a), and Naldi (2016) relies on reducing the problem of finding a matrix in a section of the PSD cone to that of finding low-rank elements in an affine space of matrices. This can be cast as a real-root finding problem, which requires computing at least one point in every real connected component of an algebraic set. This is modeled via systems of polynomial equations with finitely many solutions and solved using Gröbner-basis algorithms (Berthomieu et al. 2021). Due to their purely symbolic nature, such algorithms are not comparable with respect to scalability with numerical methods for SDP which are mainly based on variants of the interior-point method (IPM), and their correctness often depends on genericity assumptions on the input data.

Other algebraic approaches to SDP include the method of Naldi and Sinn (2020), which utilizes homogenization and projective geometry to identify the feasibility type of an LMI, and the certificate described by Klep and Schweighofer (2013), which uses the theory of sum of squares polynomials to certify SDP infeasibility.

Hybrid methods. These methods can potentially handle much larger SDP instances. Most existing work has focused on scenarios where the set of feasible solutions contains rational points (Harrison 2007; Kaltofen et al. 2012; Monniaux and Corbineau 2011; Peyrl and Parrilo 2007; Platzer 2009; Roux et al. 2018; Dostert et al. 2021). Some of these studies (Monniaux and Corbineau 2011; Dostert et al. 2021) employ lattice reduction via the Lenstra-Lenstra-Lovász (LLL) algorithm. Dostert et al. (2021) also

describes an extension to quadratic fields $\mathbb{Q}[\sqrt{\ell}]$, though it appears to assume ℓ is known.

Finally, we note the Numerical Algebraic Geometry (NAG) approach presented in Hauenstein et al. (2021), which does not strictly fall into the two aforementioned classes. This method is based on SDP duality and applies homotopy continuation to numerically track the central path.

Our contributions. In this chapter, we develop an alternative hybrid method that is not restricted to cases where the feasible region contains rational points. The method essentially reduces the problem of certifying the feasibility of an SDP to that of solving a system of polynomial equations whose set of real solutions has a zero-dimensional component containing a desired solution.

The SDP satisfiability problem consists of finding a PSD matrix subject to a linear constraint. More specifically, we aim to find a PSD matrix X such that $\mathcal{A}(X) = b$, where $b \in \mathbb{R}^m$ and $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is a linear map. Let P denote the set of matrices X that are solutions to this problem.

Our main result can be stated as follows (a more refined formulation is given later in Theorems 6 and 7).

Theorem 1. *There exists an algorithm that takes as input a feasible SDP system $\mathcal{A}(X) = b$ with $X \succeq 0$, an approximate numerical solution $\tilde{X}$, and a rank-revealing tolerance $\epsilon > 0$, with the following property:*

Assume that the feasible set P is non-empty and that X^* is a maximum-rank solution in P. Suppose the following conditions hold:

i) The tolerance ϵ is sufficiently small ($\epsilon \leq \epsilon_{max}$, where ϵ_{max} depends on $\mathcal{A}, b,$ and X^*).

ii) The approximate solution $\tilde{X}$ is sufficiently close to X^* ($\|\tilde{X} - X^*\| \leq \delta$, where δ depends on $\mathcal{A}, b, X^*,$ and ϵ).

Under these assumptions, the algorithm outputs a polynomial system in variables $X \in \mathbb{S}^n$ and auxiliary variables $Y \in \mathbb{R}^{n\times(n-r)}$ of the following form:

$$\mathcal{A}(X) = b, \quad X^\top = X \tag{2.1a}$$

$$\Pi^\top X \Pi \begin{pmatrix} Y \\ I_{n-r} \end{pmatrix} = 0^{n\times(n-r)} \tag{2.1b}$$

$$X_{ij} = \tilde{X}_{ij} \quad \forall (i,j) \in K \tag{2.1c}$$

where Π is a permutation matrix, $r \in [n]$ is the rank of X^*, and K is a computed index set (the quantities Π, r, and K are computed from $\mathcal{A}$, b, $\tilde{X}$, and ϵ). Furthermore, the set of real solutions of the resulting system is geometrically guaranteed to have a zero-dimensional component corresponding to an exact solution $X \in P$.

The correctness of this algorithm depends on the proximity of the numerical solution $\tilde{X}$ to a maximum-rank solution $X^* \in P$. Specifically, $\tilde{X}$ must lie within a certain neighborhood Ω of X^* to ensure that the non-zero eigenvalues of $\tilde{X}$ can be distinguished from numerical noise. By selecting numerically linearly independent columns in $\tilde{X}$, we determine a set of column indices ι that identifies a positive definite principal submatrix S^* within X^*. Similarly, to determine the set of indices K, the numerical approximations $\tilde{X}$ and $\tilde{S}$ are used to identify sets of row indices I and column indices J that can be fixed. These correspond to the complement of a maximal set of linearly independent columns for a new linear system (see Step 3 of Algorithm 2.1). Since there may be multiple valid choices for the sets ι, I, and J, the neighborhood Ω and the parameter δ are defined by the intersection of finitely many neighborhoods, each associated with a specific triplet (ι, I, J).

A precise definition of δ is provided in the proof of Theorem 6. The parameter $\epsilon_{\max}$ is a tolerance used to numerically compute maximal sets of linearly independent columns in Steps 2 and 4 of Algorithm 2.1.

Remark: The introduction of the auxiliary matrix Y in equation (2.1b) transforms the rank constraint into a bilinear system $XY = 0$. This formulation is crucial, as it allows us to bypass the computationally prohibitive task of enforcing rank via the vanishing of high-degree determinantal minors, shifting the problem instead to a much more tractable incidence variety. This is formally stated in the following section where we develop the necessary tools.

For example, consider an SDP whose exact solution is the rank-1 PSD matrix $X^* = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. To classically enforce the condition $\text{rank}(X) \leq 1$, one must compute the determinantal minor, yielding the quadratic constraint $x_{11}x_{22} - x_{12}^2 = 0$. Our formulation avoids this minor entirely by introducing a 1×1 auxiliary variable Y and enforcing the bilinear incidence relation:

$$X \begin{pmatrix} Y \\ 1 \end{pmatrix} = \begin{pmatrix} x_{11}Y + x_{12} \\ x_{12}Y + x_{22} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

At the target root X^*, this system evaluates to $\begin{pmatrix} 1(Y) + 0 \\ 0(Y) + 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$, whose unique solution is simply $Y = 0$. As the dimension n increases, replacing exponentially growing determinantal minors with this linear-in-Y incidence variety dramatically reduces the complexity of the resulting algebraic system.

There may be several ways to exploit this result. One possibility could be to refine the approximate solution $\tilde{X}$ and the associated matrix $\tilde{Y}$ (which is easily computable from $\tilde{X}$) by applying the Newton method for system (2.1).[1]

[1] The formula for computing $\tilde{Y}$ from $\tilde{X}$ will be specified later, in Algorithm 2.1. The mapping $\tilde{Y}(\tilde{X})$ will satisfy the following properties: (i) it is continuous in some open neighborhood of $\tilde{X}$; (ii) matrix $Y = \tilde{Y}(X^*)$ is the unique solution of (2.1b) for fixed $X = X^*$.

For some polynomial systems, this method converges very slowly near the optimum or diverges; in such cases, one could solve (2.1) using techniques from NAG, such as those implemented in the software Bertini(Bates et al. 2013a). The approximate solution $(\tilde{X}, \tilde{Y})$ enables the method to focus on the desired zero-dimensional component using the regime known as the *endgame*.

In this chapter, we explore a different direction: we solve (2.1) using exact methods from real algebraic geometry, similar to those described by Henrion et al. (2016). Accordingly, our experimental results focus on a comparison with Henrion et al. (2016). Note that the latter work also reduces the problem to a system of polynomial equations of the form (2.1a)–(2.1b), but with the following differences: (1) the algorithm of Henrion et al. (2016) has no a priori information about the feasible rank and thus needs to exhaustively search over all (exponentially many) kernel profiles until a solution is found; (2) it guarantees that, for one of these kernel profiles, the set of real solutions contains a (possibly positive-dimensional) connected component whose points (X, Y) satisfy $X \in P$; (3) it searches for a minimum-rank solution, whereas we search for a maximum-rank solution.

As in Henrion et al. (2016), we use a critical point method to find one real solution per component. This method works under certain assumptions (such as complete intersection of the corresponding variety or finiteness of the number of solutions of some critical equations); in the instances tested, these assumptions were often violated in the case of the method of Henrion et al. (2016), but not in our hybrid method. Consequently, the hybrid method was able to certify the feasibility of many instances on which Henrion et al. (2016) failed.

Outline. The remainder of this chapter is organized as follows. Section 2.2 introduces the primary theoretical foundations and notation. In Section 2.3, we establish the connection between determinantal varieties and facial reduction for SDPs. Section 2.4 describes our proposed hybrid algorithm, while Section 2.5 presents its application to a collection of benchmark instances from the literature (detailed in Section 2.8). Finally, a comprehensive example is developed in Section 2.6.

2.2 Preliminaries

Notation for matrices. All vector and matrix norms $\|\cdot\|$ used in this chapter are 2-norms. The Frobenius norm for matrices is denoted by $\|\cdot\|_F$.

For a matrix $A \in \mathbb{R}^{p\times q}$, we denote the i-th singular value of A as $\sigma_i(A) \geq 0$; if $i > \min\{p, q\}$, then $\sigma_i(A) = 0$ by definition. We denote by $\mathbb{S}^p_{\mathbb{F}}$ the set of $p \times p$ symmetric matrices with coefficients in a field $\mathbb{F}$, and for $\mathbb{F} = \mathbb{R}$, we simply write $\mathbb{S}^p$. If $A \in \mathbb{S}^p$, then $\lambda_i(A)$ denotes its i-th largest eigenvalue. A matrix $A \in \mathbb{S}^p$ is positive semidefinite ($A \succeq 0$) if $\lambda_p(A) \geq 0$, and positive

definite ($A \succ 0$) if $\lambda_p(A) > 0$. Recall that if $A \succeq 0$, then $\lambda_i(A) = \sigma_i(A)$ for all i.

The following inequalities are well known, see e.g., Golub and Van Loan (2013, Section 8):

$$|\sigma_i(A) - \sigma_i(B)| \leq \|A - B\| \qquad \forall A, B \in \mathbb{R}^{p\times q},\ \forall i \tag{2.2a}$$
$$|\lambda_i(A) - \lambda_i(B)| \leq \|A - B\| \qquad \forall A, B \in \mathbb{S}^p,\ \forall i \tag{2.2b}$$

We will also denote $\rho(A) = \sigma_r(A)$, where $r = \operatorname{rank}(A)$. Clearly, for every non-zero matrix A, we have $\rho(A) > 0$.

For a set of rows $I \subseteq [p]$, let $A_I \in \mathbb{R}^{|I|\times q}$ be the submatrix of A indexed by the rows in I. Similarly, for a set of columns $J \subseteq [q]$, let $A^J \in \mathbb{R}^{p\times|J|}$ be the submatrix of A indexed by the columns in J.

The transpose of a matrix A is denoted by $A^\top$. In this chapter, this operation is applied only to matrices with real entries. We do not employ the conjugate transpose for complex matrices; accordingly, we reserve the superscript * for a different purpose, namely to denote an exact solution X^* of the SDP system.

Semidefinite programming. We endow the vector space $\mathbb{S}^n$ of real symmetric matrices with the Frobenius inner product:

$$\langle M, N\rangle_F := \mathsf{Trace}(MN) = \sum_{i,j} M_{ij}N_{ij}$$

for $M = (M_{ij})$, $N = (N_{ij}) \in \mathbb{S}^n$. The cone of positive semidefinite matrices (PSD cone), denoted by $\mathbb{S}^n_+ = \{M \in \mathbb{S}^n : M \succeq 0\}$, is a convex closed basic semialgebraic set in $\mathbb{S}^n$; its interior is the open convex cone of positive definite matrices: $\mathbb{S}^n_{++} := \mathsf{int}(\mathbb{S}^n_+) \subseteq \mathbb{S}^n$. The Euclidean inner product on $\mathbb{R}^m$ is denoted by $\langle u, v\rangle_{\mathbb{R}} = \sum u_i v_i$.

An SDP in standard primal form is defined as

$$\begin{array}{rcll} p^* & := & \inf_{X\in\mathbb{S}^n} & \langle C, X\rangle_F \\ & & \text{s.t.} & \mathcal{A}(X) = b \\ & & & X \in \mathbb{S}^n_+ \end{array} \tag{2.3}$$

where $b \in \mathbb{R}^m$, $C \in \mathbb{S}^n$, and $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is a linear map defined by $\mathcal{A}(X) = (\langle A_1, X\rangle_F, \ldots, \langle A_m, X\rangle_F)$ for some real symmetric matrices $A_1, \ldots, A_m \in \mathbb{S}^n$.

The feasible set of program (2.3) is denoted by

$$P := \{X \in \mathbb{S}^n : \mathcal{A}(X) = b, X \succeq 0\}. \tag{2.4}$$

Problem (2.3) is assumed to be feasible ($P \neq \emptyset$). It is called *strongly feasible* if P contains a matrix X with $X \succ 0$, and *weakly feasible* otherwise. We will be mainly interested in weakly feasible SDPs. (Note that if (2.3) is strongly feasible, then P contains rational-valued matrices.)

Algebraic and semialgebraic geometry. We refer to Cox et al. (2007) for the foundations of computational algebraic geometry, and we recall here the main definitions needed in this work. A (complex) algebraic variety is a set of the form

$$V = \{x \in \mathbb{C}^n : f_1(x) = 0, \dots, f_s(x) = 0\} \tag{2.5}$$

for polynomials $f_1, \dots, f_s \in \mathbb{F}[x]$ with coefficients in some subfield $\mathbb{F} \subseteq \mathbb{C}$ (we usually fix $\mathbb{F} = \mathbb{Q}$ or $\mathbb{F} = \mathbb{R}$). Let $\mathcal{I} = \langle f_1, \dots, f_s \rangle \subseteq \mathbb{F}[x]$ be the ideal generated by $f_1, \dots, f_s$. Then, every $f \in \mathcal{I}$ vanishes over V; in other words, $V = V(\mathcal{I})$, with

$$V(\mathcal{I}) := \{x \in \mathbb{C}^n : f(x) = 0, \forall f \in \mathcal{I}\}.$$

The vanishing ideal of a set $W \subseteq \mathbb{C}^n$ is defined as

$$\mathcal{I}(W) := \{f \in \mathbb{C}[x] : f(x) = 0,\ \forall x \in W\}$$

and, by Hilbert's Nullstellensatz (Cox et al. 2007, §4.1), one has the correspondence $\mathcal{I}(V(\mathcal{J})) = \sqrt{\mathcal{J}}$, where $\sqrt{\mathcal{J}} := \{f \in \mathbb{C}[x] : f^m \in \mathcal{J}, \exists m \in \mathbb{N}\}$ is the radical ideal of $\mathcal{J}$. An ideal $\mathcal{I}$ is called radical if $\mathcal{I} = \sqrt{\mathcal{I}}$. The smallest algebraic variety containing a set W is denoted by $\overline{W}$ and called its Zariski closure.

A variety $V \subseteq \mathbb{C}^n$ is called irreducible if it cannot be expressed as a union of two proper subvarieties; every variety is a finite union of irreducible varieties, known as its irreducible components. The dimension of an algebraic variety $V \subseteq \mathbb{C}^n$ is the minimum integer $d = \dim(V)$ such that the intersection $V \cap H_1 \cap \cdots \cap H_d$ of V with d generic hyperplanes H_i is finite. The degree of an irreducible variety $V \subseteq \mathbb{C}^n$ is the cardinality of the intersection of V with $\dim(V)$ generic hyperplanes, and the degree of an arbitrary variety is the sum of the degrees of its irreducible components. A variety whose irreducible components have the same dimension is called equidimensional. An ideal $\mathcal{I}$ with $V(\mathcal{I}) \subseteq \mathbb{C}^n$ is a complete intersection if it can be generated by $\operatorname{codim}(V(\mathcal{I})) = n - \dim(V(\mathcal{I}))$ polynomials; with a slight abuse of notation, V is said to be a complete intersection if its vanishing ideal $\mathcal{I}(V)$ satisfies this condition.

Let $V \subseteq \mathbb{C}^n$ be a variety with $\mathcal{I}(V) = \langle f_1, \dots, f_s \rangle$. For a point $p \in V$, the tangent space T_pV is the kernel of the Jacobian matrix

$$J(f_1, \dots, f_s) := \left(\frac{\partial f_i}{\partial x_j}\right)_{\substack{1 \le i \le s \\ 1 \le j \le n}}$$

evaluated at p. A point $p \in V$ is said to be a *regular point* if the dimension of the tangent space T_pV coincides with the local dimension of V at p (the maximum of the dimensions of the irreducible components containing p). When V is equidimensional of dimension d, the Jacobian $J(f_1, \dots, f_s)$ has rank $\le n - d$, and equality holds exactly at regular points. We denote the set of regular points as $\operatorname{Reg}(V)$, and its complement, $\operatorname{Sing}(V) := V \setminus \operatorname{Reg}(V)$, is referred to as the set of *singular points* of V.

For $\mathbb{F} = \mathbb{R}$ and V as in (2.5), the set $V(\mathbb{R}) := V \cap \mathbb{R}^n$ is called a real algebraic variety, which is defined by the vanishing of finitely many real polynomials. Given $g_1, \dots, g_t \in \mathbb{R}[x]$, the set

$$S = \{x \in \mathbb{R}^n : g_1(x) \geq 0, \dots, g_t(x) \geq 0\}$$

is called a *basic closed semialgebraic set*.

Determinantal and incidence varieties. We refer to Eisenbud (1988) and Harris (1984) for the general theory of determinantal varieties, and recall the main definitions in the context of LMIs.

Let $A_1, \dots, A_m \in \mathbb{S}^n$ be the matrices defining the map $\mathcal{A}$ and $b \in \mathbb{R}^m$ the vector in (2.3). The set

$$\mathcal{D}_r := \{X \in \mathbb{S}^n_{\mathbb{C}} \ : \ \operatorname{rank}(X) \leq r, \mathcal{A}(X) = b\}$$

is called the determinantal variety associated with $\mathcal{A}$, b, and r. It is the algebraic variety defined by the vanishing of the $(r+1) \times (r+1)$ minors of the matrix X, subject to the linear constraints $\mathcal{A}(X) = b$. For some $r \in \{0, 1, \dots, n\}$, the real variety $\mathcal{D}_r(\mathbb{R}) := \mathcal{D}_r \cap \mathbb{S}^n$ intersects the feasible set P.

From a computational perspective, determinantal minors are difficult to handle. Instead, the rank constraint is equivalent to the existence of a matrix of full rank whose columns generate the kernel of the original matrix. The following set

$$\mathcal{V}_r := \{(X, Y) \in \mathbb{S}^n_{\mathbb{C}} \times \mathbb{C}^{n \times (n-r)} \ : \ \mathcal{A}(X) = b, XY = 0, \operatorname{rank}(Y) = n - r\}$$

is associated with the variety $\mathcal{D}_r$; indeed, $\mathcal{D}_r$ is the projection of $\mathcal{V}_r$ onto $\mathbb{S}^n_{\mathbb{C}}$. Note that $\mathcal{V}_r$ is not an algebraic variety, since the constraint on the rank of Y is not closed in the Zariski topology, but is a constructible set (an open subset of a Zariski closed set). We consider the algebraic set $\mathcal{V}_{r,\iota}$ whose projection on $\mathbb{S}^n_{\mathbb{C}}$ is contained in $\mathcal{D}_r$ (see Henrion et al. 2016):

$$\mathcal{V}_{r,\iota} := \{(X, Y) \in \mathbb{S}^n_{\mathbb{C}} \times \mathbb{C}^{n \times (n-r)} \ : \ \mathcal{A}(X) = b, XY = 0, Y_{[n]-\iota} = I_{n-r}\} \quad (2.6)$$

for $\iota \subseteq [n] = \{1, \dots, n\}$ of cardinality $|\iota| = r$, and $Y_{[n]-\iota}$ is the matrix obtained by selecting the rows of Y in $[n] - \iota$.

For a field $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$ and a subset $C \subseteq \mathbb{S}^n_{\mathbb{F}} \times \mathbb{F}^{n \times (n-r)}$, let $\pi_{\mathbb{S}^n_{\mathbb{F}}}(C) = \{X \in \mathbb{S}^n_{\mathbb{F}} : \exists Y \in \mathbb{F}^{n \times (n-r)} \text{ s.t. } (X, Y) \in C\}$ denote the projection of C onto $\mathbb{S}^n_{\mathbb{F}}$. Using this notation, we define the set

$$\mathcal{W}_{r,\iota} := \pi_{\mathbb{S}^n_{\mathbb{C}}}(\mathcal{V}_{r,\iota}).$$

We also define the real traces of $\mathcal{V}_{r,\iota}$ and $\mathcal{W}_{r,\iota}$ as follows:

$$\mathcal{V}_{r,\iota}(\mathbb{R}) = \mathcal{V}_{r,\iota} \cap (\mathbb{S}^n \times \mathbb{R}^{n \times (n-r)})$$
$$\mathcal{W}_{r,\iota}(\mathbb{R}) = \mathcal{W}_{r,\iota} \cap \mathbb{S}^n = \pi_{\mathbb{S}^n}(\mathcal{V}_{r,\iota}(\mathbb{R}))$$

where the last equality holds due to the following observation: if $(X, Y) \in \mathcal{V}_{r,\iota}$ with $X \in \mathbb{S}^n$, then $(X, \operatorname{Re} Y) \in \mathcal{V}_{r,\iota}(\mathbb{R})$. Note that $\mathcal{W}_{r,\iota}(\mathbb{R})$ might not be algebraic; it is a constructible set.

Incidence varieties are employed by Henrion et al. (2016) to solve feasibility problems of generic instances of LMIs, and the following theorem is a refinement of results in that work.

Theorem 2. *Assume that $P \neq \emptyset$. Let $X^* \in \mathbb{S}^n_+$ be a minimum-rank solution of (2.4), and let $\iota \subseteq [n]$ be a maximal set of linearly independent columns of X^*, with $r = |\iota|$.*

(a) The constructible set $\mathcal{W}_{r,\iota}(\mathbb{R})$ has a connected component C such that $X^* \in C \subseteq P$.

(b) The real variety $\mathcal{V}_{r,\iota}(\mathbb{R})$ has a connected component C' such that $X^* \in \pi_{\mathbb{S}^n}(C') \subseteq P$.

Proof. By definition of ι, there exists $Y^* \in \mathbb{R}^{n\times(n-r)}$ such that $X^*Y^* = 0$ and $Y^*_{[n]-\iota} = I_{n-r}$. Consequently, $(X^*, Y^*) \in \mathcal{V}_{r,\iota}(\mathbb{R})$, which implies $X^* \in \mathcal{W}_{r,\iota}(\mathbb{R})$. Let C be the connected component of $\mathcal{W}_{r,\iota}(\mathbb{R})$ containing X^*.

Part (a). We first demonstrate that $C \subseteq P$. If $r = 0$, then all points $(X, Y) \in \mathcal{V}_{r,\emptyset}(\mathbb{R})$ satisfy $X = 0$; hence, $C = \{X^*\} = \{0\}$, and the claim follows. Suppose that $r > 0$ and the claim does not hold. Then, there exists a continuous curve $\{X^{(t)}\}_{t\in[0,1]} \subseteq C \subseteq \mathcal{W}_{r,\iota}(\mathbb{R})$ with $X^{(0)} = X^*$ and $X^{(1)} = X \notin P$. For each $t \in [0, 1]$, there exists a matrix $Y^{(t)}$ such that $(X^{(t)}, Y^{(t)}) \in \mathcal{V}_{r,\iota}(\mathbb{R})$.

Let $\lambda_i^{(t)} = \lambda_i(X^{(t)})$ for $t \in [0, 1]$, so that $\lambda_1^{(t)} \geq \ldots \geq \lambda_n^{(t)}$. These values satisfy the following properties for each $t \in [0, 1]$:

(i) At least $n-r$ values in $\lambda_1^{(t)}, \ldots, \lambda_n^{(t)}$ are zero, or equivalently $\operatorname{rank}(X^{(t)}) \leq r$. This follows from the conditions $X^{(t)}Y^{(t)} = 0$ and $\operatorname{rank}(Y^{(t)}) = n - r$, which imply that $\operatorname{rank}(X^{(t)}) \leq n - \operatorname{rank}(Y^{(t)}) \leq n - (n - r) = r$.

(ii) If $\lambda_n^{(t)} \geq 0$, then $\lambda_{r+1}^{(t)} = \ldots = \lambda_n^{(t)} = 0$ and $\lambda_r^{(t)} > 0$. The first claim follows from (i); if $\lambda_r^{(t)} = 0$, then $\operatorname{rank}(X^{(t)}) \leq r - 1$, contradicting the minimality of r (note that $X^{(t)} \succeq 0$ and hence $X^{(t)} \in P$).

(iii) For each $i \in [n]$, $\lambda_i^{(t)}$ is a continuous function of t. This holds by the continuity of the curve and by (2.2b).

Conditions $X^{(0)} \in P$ and $X^{(1)} \notin P$ give $\lambda_n^{(0)} \geq 0$ and $\lambda_n^{(1)} < 0$. Denote $s = \sup\{t \in [0, 1] : \lambda_n^{(t)} \geq 0\}$, then $\lambda_n^{(s)} \geq 0$ by continuity and hence $s < 1$. By (ii) we have $\mu := \lambda_r^{(s)} > 0$. By continuity, there exists $u \in (s, 1]$ such that $\lambda_r^{(t)} > \mu/2$ for all $t \in [s, u]$. Condition (i) thus gives $\lambda_{r+1}^{(t)} = \ldots = \lambda_n^{(t)} = 0$ for all $t \in [s, u]$, and thus $\sup\{t \in [0, 1] : \lambda_n^{(t)} \geq 0\} \geq u > s$, which is a contradiction.

Part (b). Let C' be the connected component of $\mathcal{V}_{r,\iota}(\mathbb{R})$ containing (X^*, Y^*). Clearly, we have $\pi_{\mathbb{S}^n}(C') \subseteq C$, and so the claim holds by part (a). □

2.3 Incidence varieties and facial reduction

We consider the feasibility problem of the primal SDP defined in (2.3), which is given by the system:

$$\mathcal{A}(X) = b, \quad X \succeq 0. \tag{2.7}$$

Let $P = \{X \in \mathbb{S}^n \ : \ \mathcal{A}(X) = b, X \succeq 0\}$ be the set of feasible solutions as in (2.4), and let r be the maximum rank of a matrix in P. Throughout this section, we assume that $P \neq \emptyset$ and $r > 0$. A matrix $X \in P$ of rank r is referred to as a *maximum-rank solution*. We first recall some basic facts about facial reduction for SDPs(see, e.g., Drusvyatskiy and Wolkowicz 2017). Let $\mathcal{F}$ be the minimal face of the PSD cone $\mathbb{S}^n_+$ containing P. It can be represented by a full rank matrix $U \in \mathbb{R}^{n\times r}$ as follows:

$$\mathcal{F} = \{X \in \mathbb{S}^n_+ \ : \ \mathrm{range}(X) \subseteq \mathrm{range}(U)\} = \{UZU^\top \ : \ Z \in \mathbb{S}^r_+\}. \tag{2.8}$$

In other words, $\mathcal{F}$ is linearly isomorphic (via the mapping $UZU^\top \mapsto Z$) to a copy of the cone $\mathbb{S}^r_+$, and in particular it has dimension $\binom{r+1}{2}$. Any matrix in the relative interior of P is in the relative interior of $\mathcal{F}$ and hence has rank r; every other matrix in P has rank $< r$. Let $\iota \subseteq [n]$ be a maximal set of linearly independent rows of U, with $|\iota| = r$. We assume, for notational convenience, that $\iota = [r]$ (which can be achieved by permuting the rows and columns of X). Since U has full rank, we can apply Gaussian elimination to compute its column reduced echelon form without altering its range. Accordingly, we may assume without loss of generality that

$$U = \begin{pmatrix} I_r \\ U_0 \end{pmatrix}, \qquad \text{with } U_0 \in \mathbb{R}^{(n-r)\times r}.. \tag{2.9}$$

For the matrix $Y_U := -U_0^\top$, we have

$$\begin{aligned} \mathcal{F} &= \{X \in \mathbb{S}^n_+ \ : \ \mathrm{range}(X) \subseteq \mathrm{range}(U)\}, && \text{(2.10a)} \\ &= \{X \in \mathbb{S}^n_+ \ : \ \begin{pmatrix} Y_U^\top & I_{n-r} \end{pmatrix} X = 0^{(n-r)\times n}\}, && \text{(2.10b)} \\ &= \{X \in \mathbb{S}^n_+ \ : \ X \begin{pmatrix} Y_U \\ I_{n-r} \end{pmatrix} = 0^{n\times(n-r)}\}. && \text{(2.10c)} \end{aligned}$$

Now consider the following system of equations in the variables $X \in \mathbb{S}^n_{\mathbb{C}}$ and $Y \in \mathbb{C}^{n\times(n-r)}$:

$$\begin{aligned} \mathcal{A}(X) &= b, && \text{(2.11a)} \\ X \begin{pmatrix} Y \\ I_{n-r} \end{pmatrix} &= 0^{n\times(n-r)}. && \text{(2.11b)} \end{aligned}$$

Let $\mathcal{V}_{r,\iota} \subseteq \mathbb{S}^n_{\mathbb{C}} \times \mathbb{C}^{n\times(n-r)}$ be the set of complex solutions to (2.11). Note that $\mathcal{V}_{r,\iota}$ is essentially the same incidence variety defined in the previous section (for $\iota = [r]$), modulo notation: the matrix Y in (2.6) has a larger size, but all additional entries are fixed to constants. We define $\mathcal{W}_{r,\iota}$, $\mathcal{V}_{r,\iota}(\mathbb{R})$,

and $\mathcal{W}_{r,\iota}(\mathbb{R})$ in the same manner as before:

$$\begin{aligned}\mathcal{W}_{r,\iota} &:= \pi_{\mathbb{S}^n_{\mathbb{C}}}(\mathcal{V}_{r,\iota}) \\ &= \{X \in \mathbb{S}^n_{\mathbb{C}} \ : \ \exists Y \in \mathbb{C}^{r\times(n-r)} \text{ s.t. } (X,Y) \in \mathcal{V}_{r,\iota}\} \qquad (2.12a) \\ &= \{X \in \mathbb{S}^n_{\mathbb{C}} \ : \ \exists Y \in \mathbb{C}^{r\times(n-r)} \text{ s.t. } \mathcal{A}(X) = b,\ X\begin{pmatrix} Y \\ I_{n-r}\end{pmatrix} = 0^{n\times(n-r)}\}, \qquad (2.12b)\end{aligned}$$

and

$$\begin{aligned}\mathcal{V}_{r,\iota}(\mathbb{R}) &:= (\mathbb{S}^n \times \mathbb{R}^{n\times(n-r)}) \cap \mathcal{V}_{r,\iota}, \\ \mathcal{W}_{r,\iota}(\mathbb{R}) &:= \mathbb{S}^n \cap \mathcal{W}_{r,\iota} \qquad (2.13) \\ &= \pi_{\mathbb{S}^n}(\mathcal{V}_{r,\iota}(\mathbb{R})).\end{aligned}$$

By construction, we have $P \subseteq \mathcal{F} \cap \{X \in \mathbb{S}^n \ : \ \mathcal{A}(X) = b\} \subseteq \mathcal{W}_{r,\iota}(\mathbb{R})$.

Lemma 3. *Let $P \neq \emptyset$ and let $X^* = UZ^*U^\top \in P$ be a maximum-rank solution, with U as in (2.9) and $Z^* \in \mathbb{S}^r_{++}$. Define the open neighborhood Ω^* of X^* as $\Omega^* := \{X \in \mathbb{S}^n \ : \ \|X - X^*\| < \rho(X^*)\}$. The following properties hold:*

- *(a) $P \cap \Omega^* = \mathcal{W}_{r,\iota}(\mathbb{R}) \cap \Omega^*$;*
- *(b) for every $X \in \mathcal{W}_{r,\iota}(\mathbb{R}) \cap \Omega^*$, there exists a unique matrix Y such that $(X,Y) \in \mathcal{V}_{r,\iota}$, given by $Y = Y_U$.*

Proof. Part (a). If $X \in P$, then $\mathcal{A}(X) = b$ and $(X, Y_U) \in \mathcal{V}_{r,\iota}$ (since $P \subseteq \mathcal{F}$), which implies that $X \in \mathcal{W}_{r,\iota}(\mathbb{R})$. This shows that $P \subseteq \mathcal{W}_{r,\iota}(\mathbb{R})$ and hence $P \cap \Omega^* \subseteq \mathcal{W}_{r,\iota}(\mathbb{R}) \cap \Omega^*$.

Conversely, suppose that $X \in \mathcal{W}_{r,\iota}(\mathbb{R}) \cap \Omega^*$. Then there exists Y such that $(X,Y) \in \mathcal{V}_{r,\iota}$; that is, (X,Y) satisfies the system (2.11). In turn, Equation (2.11b) implies that X has at least $n-r$ zero eigenvalues. Since $\operatorname{rank}(X^*) = r$, we have $\lambda_r(X^*) = \rho(X^*) > 0$. By (2.2b), for all $X \in \Omega^*$ we have

$$|\lambda_r(X) - \lambda_r(X^*)| \leq \|X - X^*\| < \rho(X^*) = \lambda_r(X^*),$$

and thus $\lambda_r(X) > 0$; that is, X has at least r strictly positive eigenvalues. We conclude that X has exactly $n-r$ zero eigenvalues and exactly r strictly positive eigenvalues. Therefore, $X \succeq 0$ and $\operatorname{rank}(X) = r$. Combined with the fact that $\mathcal{A}(X) = b$, this implies that $X \in P$. This establishes the converse inclusion $\mathcal{W}_{r,\iota}(\mathbb{R}) \cap \Omega^* \subseteq P \cap \Omega^*$.

Part (b). As shown in part (a), we have $X \in P \subseteq \mathcal{F}$. Therefore, $(X, Y_U) \in \mathcal{V}_{r,\iota}$; that is, it satisfies the system (2.11). Suppose that there exists $Y \neq Y'$ such that (X,Y) and (X,Y') both satisfy (2.11). Let

$$C := \begin{pmatrix} Y & Y' \\ I_{n-r} & I_{n-r}\end{pmatrix};$$

then $XC = 0^{n\times 2(n-r)}$. It follows that $\operatorname{rank}(C) \geq n-r+1$: indeed, the first $n-r$ columns of C are linearly independent, and there exists a column among the remaining ones which is not in the range of the first $n-r$ columns, by the assumption $Y \neq Y'$. This implies that $\operatorname{rank}(X) \leq n - \operatorname{rank}(C) \leq n - (n-r+1) = r-1$, which contradicts the condition $\operatorname{rank}(X) = r$ shown in part (a). □

2.4 Description of the method

Motivated by Lemma 3, we propose the following hybrid method for certifying the feasibility of the system (2.7) (see Algorithm 2.1). (We assume that the system is feasible; that is, $P \neq \emptyset$). Several steps of Algorithm 2.1 require approximate computations. First, we compute an approximate solution $\tilde{X}$ close to some maximum-rank solution X^*; Theorems 6 and 7 will subsequently characterize how close $\tilde{X}$ must be to X^* for the method to work correctly. Second, Steps 2 and 4 invoke a procedure for *numerically* computing a maximal set of linearly independent columns of a given matrix. Such a procedure is described in Section 2.4.1.

Section 2.6 provides an illustration of Algorithm 2.1 on a small numerical example.

Algorithm 2.1: Hybrid algorithm for system $\mathcal{A}(X) = b,\ X \succeq 0$.

---

1 Solve system (2.7) numerically to obtain a solution $\tilde{X} \in \mathbb{S}^n$ close to some maximum-rank solution X^*.

2 Determine *numerically* a maximal set $\iota \subseteq [n]$ of linearly independent columns of $\tilde{X}$; let $r := |\iota|$. Assume, for notational convenience, that $\iota = [r]$ (which can be achieved by permuting variables). Partition $\tilde{X}$ as

$$\tilde{X} := \begin{pmatrix} \tilde{S} & \tilde{R}^\top \\ \tilde{R} & \tilde{W} \end{pmatrix},$$

where $\tilde{S} := \tilde{X}^\iota_\iota \in \mathbb{S}^r$. If $\tilde{S}$ is singular, terminate with failure; otherwise, define $\tilde{Y} := -\tilde{S}^{-1}\tilde{R}^\top \in \mathbb{R}^{r\times(n-r)}$. (By Lemma 4, $\tilde{X}\begin{pmatrix} \tilde{Y} \\ I_{n-r} \end{pmatrix} \approx 0^{n\times(n-r)}$).

3 Represent system (2.11) as

$$\mathcal{Q}(Y)X_{\mathsf{hvec}} = q, \tag{2.14}$$

where X_{hvec} is a vector of dimension $k := \frac{n(n+1)}{2}$ obtained by vectorizing the lower triangular part of X, $\mathcal{Q}(Y)$ is a matrix depending linearly on Y, and q is a fixed vector.

4 Determine *numerically* a maximal set $J \subseteq [k]$ of linearly independent columns of $\mathcal{Q}(\tilde{Y})$.

5 Augment system (2.11) by fixing the variables in $J' := [k] \setminus J$:

$$\mathcal{A}(X) = b, \tag{2.15a}$$

$$X\begin{pmatrix} Y \\ I_{n-r} \end{pmatrix} = 0^{n\times(n-r)}, \tag{2.15b}$$

$$(X_{\mathsf{hvec}})_j = (\tilde{X}_{\mathsf{hvec}})_j \quad \forall j \in J'. \tag{2.15c}$$

Following the methodology in Henrion et al. (2016) (cf. Section 2.4.2), we compute one point per connected component of the real algebraic variety defined by (2.15). Under this framework (specifically Henrion et al. (2016, Lemma 3.2)), the equations in (2.15b) at positions (i, j) with $i - j > r$ are algebraically dependent and thus omitted.

The following lemma justifies Step 2.

Lemma 4. *Consider a PSD matrix* $X = \begin{pmatrix} S & R^\top \\ R & W \end{pmatrix} \in \mathbb{S}^n_+$ *with* $S \in \mathbb{S}^r$*, and suppose that* $\iota = [r]$ *is a maximal set of linearly independent columns (implying the rank of* X *is* r*). Then* S *is nonsingular (and thus positive definite), and the matrix* $Y := -S^{-1}R^\top$ *satisfies* $X \begin{pmatrix} Y \\ I_{n-r} \end{pmatrix} = 0^{n\times(n-r)}$.

Proof. By hypothesis, the last $n-r$ columns of X are linearly dependent on the first r columns. Consequently, we can write

$$\begin{pmatrix} R^\top \\ W \end{pmatrix} = \begin{pmatrix} S \\ R \end{pmatrix} Z$$

for some matrix $Z \in \mathbb{R}^{r\times(n-r)}$, which implies that $R^\top = SZ$ and $W = RZ$. Hence, $R = Z^\top S$ and $W = Z^\top SZ$. We can therefore express X as

$$X = \begin{pmatrix} I_r & Z \end{pmatrix}^\top S \begin{pmatrix} I_r & Z \end{pmatrix}.$$

This implies that $r = \operatorname{rank}(X) \le \operatorname{rank}(S)$, and since $S \in \mathbb{S}^r$, it follows that S is nonsingular, establishing the first claim.

Since S is nonsingular, we have $Z = S^{-1}R^\top$ and $W = RS^{-1}R^\top$. Using these equations, it can be verified that $SY + R^\top = 0^{r\times(n-r)}$ and $RY + W = 0^{(n-r)\times(n-r)}$, which yields the second claim. □

Definition 5. *We say that the sets* $\iota \subseteq [n]$ *and* $J \subseteq [k]$ *are valid for a maximum-rank solution* X^* *if they are possible outputs of Steps 2 and 4, respectively, assuming that Step 1 computes the vector* X^**, and the computations in Steps 2 and 4 are performed exactly; that is, a maximal set of linearly independent columns is computed exactly rather than numerically.*

Theorem 6. *Each maximum-rank solution* X^* *has an open neighborhood* $\tilde{\Omega} \subset \mathbb{S}^n$ *with the following property. Suppose that Step 1 produces a point* $\tilde{X} \in \tilde{\Omega}$*, and Lines 2 and 4 output sets* ι *and* J*, respectively, which are valid for* X^**. Then the set of real solutions to the system (2.15) has a zero-dimensional component with a solution* (X, Y)*, where* $X \in \mathbb{S}^n_+$.

Note that the theorem's precondition essentially requires that the approximate computations output the same sets ι and J as the exact computations. In Section 2.4.1, we will show that this can be achieved even if X^* is known only approximately.

Proof. By assumption, the set ι is valid for X^*; that is, ι is a maximal set of linearly independent columns of X^* (and thus $|\iota| = r$). Assume, for notational convenience, that $\iota = [r]$. By Lemma 4, ι is also a maximal set of linearly independent rows of X^*. Choose a matrix $U \in \mathbb{R}^{n\times r}$ characterizing the face $\mathcal{F}$ as in (2.8). Since $\operatorname{range}(X^*) \subseteq \operatorname{range}(U)$ and $\operatorname{rank}(X^*) = r$, the set ι is a maximal set of linearly independent rows of U. Thus, we can choose U as in (2.9).

Let $\Omega^* := \{X \in \mathbb{S}^n \ : \ \|X - X^*\| < \rho(X^*)\}$. By Lemma 3(a,b), for any (X, Y) we have

$$\begin{cases} \mathcal{Q}(Y)X_{\mathsf{hvec}} = q \\ X \in \Omega^* \end{cases} \iff \begin{cases} \mathcal{Q}(Y_U)X_{\mathsf{hvec}} = q \\ X \in \Omega^* \\ Y = Y_U \end{cases} \implies X \in \mathbb{S}^n_+. \tag{2.16}$$

To simplify notation, let write

$$X_{\mathsf{hvec}} =: \begin{pmatrix} X^J \\ X^{J'} \end{pmatrix}$$

and $\mathcal{Q}(Y_U) =: \mathcal{Q} =: \begin{pmatrix} Q^J & Q^{J'} \end{pmatrix}$. By assumption, J is a maximal set of linearly independent columns of the matrix $\mathcal{Q}$. Let I be a maximal set of linearly independent rows of $\mathcal{Q}$; then $|I| = |J| = \operatorname{rank}(\mathcal{Q})$. The linear system $\mathcal{Q}X_{\mathsf{hvec}} = q$ has at least one solution, namely $X_{\mathsf{hvec}} = X^*_{\mathsf{hvec}}$. Consequently, removing equations corresponding to rows $i \notin I$ does not affect the set of feasible solutions. This implies that the system $\mathcal{Q}X_{\mathsf{hvec}} = q = \mathcal{Q}X^*_{\mathsf{hvec}}$ is equivalent to the system

$$\begin{pmatrix} Q^J_I & Q^{J'}_I \end{pmatrix} \begin{pmatrix} X^J - (X^*)^J \\ X^{J'} - (X^*)^{J'} \end{pmatrix} = 0,$$

which is equivalent to

$$X^J - (X^*)^J = -(Q^J_I)^{-1} Q^{J'}_I (X^{J'} - (X^*)^{J'})$$

since the matrix $\mathcal{Q}^J_I$ is nonsingular. Let denote $\mathcal{R}^J_I := -(\mathcal{Q}^J_I)^{-1}\mathcal{Q}^{J'}_I$. By adding the constraint $X^{J'} = \tilde{X}^{J'}$ to (2.16), we conclude the following for any (X, Y):

$$\begin{aligned} & \begin{cases} (X, Y) \text{ satisfies (2.15)} \\ X \in \Omega^* \end{cases} \\ \iff \quad & \begin{cases} X^J - (X^*)^J = \mathcal{R}^J_I(\tilde{X}^{J'} - (X^*)^{J'}) \\ X^{J'} = \tilde{X}^{J'} \\ X \in \Omega^* \\ Y = Y_U \end{cases} \\ \implies \quad & X \in \mathbb{S}^n_+. \end{aligned} \tag{2.17}$$

Let

$$\delta_{\iota,J,I} := \frac{\rho(X^*)}{4\max\{\|\mathcal{R}^J_I\|, 1\}}.$$

We claim that for any $\tilde{X} \in \mathbb{S}^n$ with $\|\tilde{X} - X^*\|_F < \delta_{\iota,J,I}$, the system in the middle of (2.17) has exactly one feasible solution (X, Y). Indeed, we only need to verify that X, defined by the equalities in this system, satisfies $X \in \Omega^*$. This holds since

$$\|X^J - (X^*)^J\| \le \|\mathcal{R}^J_I\| \cdot \|\tilde{X}^{J'} - (X^*)^{J'}\| < \frac{1}{4}\rho(X^*)$$

and $\|X^{J'} - (X^*)^{J'}\| < \frac{1}{4}\rho(X^*)$, implying that

$$\|X - X^*\| \leq \|X - X^*\|_F \leq 2\|X^J - (X^*)^J\| + 2\|X^{J'} - (X^*)^{J'}\| < \rho(X^*).$$

We can now define the set $\tilde{\Omega}$ in Theorem 6 as follows:

$$\tilde{\Omega} := \{\tilde{X} \in \mathbb{S}^n \ : \ \|\tilde{X} - X^*\|_F < \delta\},$$

where $\delta := \min_{\iota,J,I} \delta_{\iota,J,I}$ and the minimum is taken over the (finitely many) valid choices of ι, J, I. □

2.4.1 Numerical algorithms: implementing Steps 2 and 4

To implement Algorithm 2.1, we need a procedure for numerically computing a maximal set of linearly independent columns of a given matrix. This can be achieved, for example, by a rank-revealing Gaussian elimination algorithm (Pan 2000; Schork and Gondzio 2020). Such an algorithm takes a matrix $A \in \mathbb{R}^{p\times q}$ and a tolerance value ϵ as input, and produces an integer r and a subset $J \subseteq [q]$ of size r with the following guarantees: [2]

$$\sigma_r(A) \geq \epsilon, \tag{2.18a}$$
$$\sigma_{r+1}(A) \leq c_{pq}\epsilon, \tag{2.18b}$$
$$\sigma_r(A^J) \geq \sigma_r(A)/c_{pq}, \tag{2.18c}$$

where the constant $c_{pq} > 1$ depends polynomially on the dimensions p and q. We analyze Algorithm 2.1 assuming that Steps 2 and 4 employ the method above with tolerance values ϵ_1 and ϵ_2, respectively. The next result shows that if $\|\tilde{X} - X^*\|$ is sufficiently small, then one can compute sets ι and J from $\tilde{X}$ that are valid for X^*.

Theorem 7. *(a) Suppose that $\delta < \epsilon_1 < (\rho^* - \delta)/c_{nn}$, where $\delta := \|\tilde{X} - X^*\|$ and $\rho^* := \rho(X^*)$. Then the set ι computed in Step 2 is a maximal set of linearly independent columns of X^*. Furthermore,*

$$\lambda_r(\tilde{S}) \geq \phi(\delta) := \frac{((\rho^* - \delta)/c_{nn} - \delta)^2}{n\|X^*\|} - \delta. \tag{2.19}$$

(b) Suppose that the precondition in part (a) holds and $\phi(\delta) > \delta$. Partition X^* as

$$X^* := \begin{pmatrix} S^* & (R^*)^\top \\ R^* & W^* \end{pmatrix},$$

where $S^* := (X^*)^\iota_\iota$, and let $Y^* := -(S^*)^{-1}(R^*)^\top$. Let $\|\mathcal{Q}\|$ be the 2-norm of the linear operator $\mathcal{Q} \ : \ \mathbb{R}^{r\times(n-r)} \to \mathbb{R}^{p\times q}$, and let

$$\psi(\delta) := \|\mathcal{Q}\| \cdot \left[\frac{\|X^*\|}{\phi(\delta) - \delta} + 1\right] \cdot \frac{\delta}{\phi(\delta)}. \tag{2.20}$$

If $\psi(\delta) < \epsilon_2 < (\rho(\mathcal{Q}(Y^*)) - \psi(\delta))/c_{pq}$, then the set J computed in Step 4 is a maximal set of linearly independent columns of $\mathcal{Q}(Y^*)$.

[2] The algorithm actually produces subsets $I \subseteq [p]$ and $J \subseteq [q]$ of size r such that $\sigma_r(A^J_I) \geq \sigma_r(A)/c_{pq}$. (This follows by combining Lemma 3.1 and Theorem 4.1 in Schork and Gondzio (2020).) This implies (2.18c) since $\sigma_r(A^J) \geq \sigma_r(A^J_I)$ by a well-known property of singular values(see, e.g., Golub and Van Loan 2013, Corollary 8.6.3).

Note that $\lim_{\delta\to 0}\phi(\delta) = C > 0$ and $\lim_{\delta\to 0}\psi(\delta) = 0$. Consequently, the preconditions in parts (a) and (b) can be satisfied for a sufficiently small $\delta > 0$, provided that $\rho(\mathcal{Q}(Y^*)) > 0$. (The case where $\rho(\mathcal{Q}(Y^*)) = 0$ is trivial, as we then have $\mathcal{Q}(Y^*) = 0$ and $\mathcal{A}(X) = 0$ for all X). We remark that the linear operator $\mathcal{Q}$, the function $\psi(\cdot)$, and the value $\rho(\mathcal{Q}(Y^*)) - \psi(\delta)$ depend on ι; however, there are only finitely many subsets $\iota \subseteq [n]$ that correspond to maximal linearly independent columns of X^*. Therefore, one could take the minimum value over such ι when formulating the final condition.

The remainder of this section provides the proof of Theorem 7, beginning with part (a). Let $r^* := \mathrm{rank}(X^*)$ be the true rank and r be the rank computed by the numerical procedure in Step 2. Condition (2.18c) implies that

$$\sigma_r(\tilde{X}^\iota) \geq \frac{\sigma_r(\tilde{X})}{c_{nn}} \geq \frac{\sigma_r(X^*) - \|\tilde{X} - X^*\|}{c_{nn}} = \frac{\rho^* - \delta}{c_{nn}}.$$

Consequently, $\sigma_r((X^*)^\iota) \geq \sigma_r(\tilde{X}^\iota) - \|X^* - \tilde{X}\| \geq (\rho^* - \delta)/c_{nn} - \delta > 0$, which indicates that the columns in ι are linearly independent in X^* (and thus $r \leq r^*$). If $r < r^*$, then $\sigma_{r+1}(\tilde{X}) \leq c_{nn}\epsilon_1$ by condition (2.18b) and $\sigma_{r+1}(\tilde{X}) \geq \sigma_{r^*}(\tilde{X}) \geq \sigma_{r^*}(X^*) - \|\tilde{X} - X^*\| = \rho^* - \delta$ by (2.2a), which is a contradiction. This establishes that $r = r^*$.

The final claim (2.19) follows from Lemma 8 applied to the matrix $X := \tilde{X}$ and the value $\tau := \sigma_r(\tilde{X}^\iota) \geq (\rho^* - \delta)/c_{nn} > \delta$.

Lemma 8. *Consider symmetric matrices $X^* \in \mathbb{S}^n_+$ and $X \in \mathbb{S}^n$ such that $\|X^* - X\| = \delta$. Let $\iota \subseteq [n]$ be a subset of size r with $\sigma_r(X^\iota) = \tau \geq \delta$. Then*

$$\lambda_r(X^\iota_\iota) \geq \frac{(\tau - \delta)^2}{n\|X^*\|} - \delta.$$

Note that Lemma 8 also implies the first part of Lemma 4 (by setting $X = X^*$).

Proof. We utilize the well-known fact that for any matrix $A \in \mathbb{R}^{p\times q}$ with $p \geq q$, one has $\sigma_q(A) = \min\{\|Au\| \ : \ u \in \mathbb{R}^q, \|u\| = 1\}$.

Consider a vector $u \in \mathbb{R}^r$ with $\|u\| = 1$; then $\|X^\iota u\| \geq \tau$. Let $v \in \mathbb{R}^n$ be the vector with $v_i = u_i$ for $i \in \iota$ and $v_i = 0$ for $i \in [n] \setminus \iota$. We have $\|Xv\| = \|X^\iota u\| \geq \tau$, and hence

$$\|X^*v\| \geq \|Xv\| - \|(X^* - X)v\| \geq \tau - \|X^* - X\| \cdot 1 = \tau - \delta.$$

Letting $X^* = \sum_{i=1}^n \lambda_i v_i v_i^\top = \sum_{i=1}^n w_i w_i^\top$ where $||X^*|| = \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n \geq 0$, $v_1, \ldots, v_n$ are orthonormal vectors, and $w_i = \sqrt{\lambda_i} v_i$ (then $||w_i|| \leq \sqrt{\lambda_1}$). Denote $\alpha_i = |v^\top w_i|$, then $v^\top X^* v = \sum_{i=1}^n \alpha_i^2$. One can see that

$$\begin{aligned}\tau - \delta \leq ||X^*v|| = ||\sum_{i=1}^n w_i(v^\top w_i)|| \leq \sum_{i=1}^n \alpha_i ||w_i|| \\ \leq \sqrt{\lambda_1} \cdot \sum_{i=1}^n \alpha_i \leq \sqrt{\lambda_1} \cdot \sqrt{n\sum_{i=1}^n \alpha_i^2} = \sqrt{\lambda_1 n (v^\top X^* v)}\end{aligned}$$

where the last inequality is tight if and only if $\alpha_1 = \ldots = \alpha_n$. From this, we obtain

$$u^\top X_\iota' u = v^\top X v = v^\top X^* v + v^\top (X - X^*) v \geq \frac{(\tau - \delta)^2}{n\lambda_1} - \|v\|^2 \cdot \|X - X^*\| = \frac{(\tau - \delta)^2}{n\lambda_1} - \delta.$$

□

We now prove part (b) of Theorem 7. This proof requires the following result.

Theorem 9 (Atkinson (1991), Theorem 7.12)**.** *Let A and B be square matrices of the same size such that A is nonsingular and $\|A - B\| \leq 1/\|A^{-1}\|$. Then B is also nonsingular and*

$$\|B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|A - B\|},$$

and

$$\|A^{-1} - B^{-1}\| \leq \frac{\|A^{-1}\|^2 \cdot \|A - B\|}{1 - \|A^{-1}\| \cdot \|A - B\|}.$$

For brevity, let denote $S := S^*$ and $R := R^*$. One can express the difference as

$$\tilde{Y} - Y^* = S^{-1} R^\top - \tilde{S}^{-1} \tilde{R}^\top = (S^{-1} - \tilde{S}^{-1}) R^\top + \tilde{S}^{-1} (R^\top - \tilde{R}^\top).$$

From part (a), we have $\|\tilde{S}^{-1}\| \leq 1/\phi(\delta)$, and thus $\|\tilde{S} - S\| \leq \|\tilde{X} - X^*\| = \delta < \phi(\delta) \leq 1/\|\tilde{S}^{-1}\|$. Applying Theorem 9 with $A := \tilde{S}$ and $B := S$, we obtain

$$\begin{aligned}
\|\tilde{Y} - Y^*\| &\leq \|S^{-1} - \tilde{S}^{-1}\| \cdot \|R^\top\| + \|\tilde{S}^{-1}\| \cdot \|R^\top - \tilde{R}^\top\| \\
&\leq \frac{(\frac{1}{\phi(\delta)})^2 \cdot \delta}{1 - \frac{1}{\phi(\delta)} \cdot \delta} \cdot \|X^*\| + \frac{1}{\phi(\delta)} \cdot \delta \\
&= \left[\frac{\|X^*\|}{\phi(\delta) - \delta} + 1 \right] \cdot \frac{\delta}{\phi(\delta)}.
\end{aligned}$$

Let $\mathcal{Q}^* := \mathcal{Q}(Y^*)$ and $\tilde{\mathcal{Q}} := \mathcal{Q}(\tilde{Y})$; then $\|\tilde{\mathcal{Q}} - \mathcal{Q}^*\| \leq \|\mathcal{Q}\| \cdot \|\tilde{Y} - Y^*\| \leq \psi(\delta)$. We conclude that J is a maximal set of linearly independent columns of $\mathcal{Q}^*$ by the same argument as in part (a).

2.4.2 Symbolic algorithms: implementing Step 5

Step 5 of Algorithm 2.1 and Theorem 6 reduce the original problem of validating the feasibility of an LMI to a classical question in real algebraic geometry(as in Henrion et al. 2016): the computation of one point per connected component of the real trace $V(\mathbb{R}) := V \cap \mathbb{R}^n$ of a complex variety

$$V := \{x \in \mathbb{C}^n \ : \ f_1(x) = 0, \ldots, f_c(x) = 0\}.$$

A standard technique consists in computing the critical points of a well-chosen polynomial map $\varphi : \mathbb{C}^n \to \mathbb{C}$ (usually of low degree) restricted to

V, a procedure often referred to as the Critical Point Method (CPM). The expected output of a CPM is a finite set intersecting every connected component of $V(\mathbb{R})$. For an overview of this problem, we refer the reader to Basu et al. (2006, § 12.6).

Common choices for φ are linear functions. When φ is a generic linear function, and since $V(\mathbb{R})$ has finitely many connected components, there are finitely many critical points of its restriction to $V(\mathbb{R})$. In the case where $V(\mathbb{R})$ is noncompact, φ might not admit critical points on the (unbounded) connected components $C \subseteq V$ satisfying $\varphi(C) = \mathbb{R}$. In this case, the recursive method proposed by Safey El Din and Schost (2003) enables the computation of such components. However, our situation is special inasmuch as our target component is an isolated real point; consequently, the mentioned recursion is not required for Step 5 of Algorithm 2.1.

The CPM for a system of polynomials $f : \mathbb{C}^n \to \mathbb{C}^c$ and a polynomial map $\varphi : \mathbb{C}^n \to \mathbb{C}$ can be described as follows. Recall that $J(f) := (\partial f_i/\partial x_j)_{ij}$ is the Jacobian matrix of f. Define the extended Jacobian of (φ, f) as

$$J(\varphi, f) := \begin{pmatrix} \frac{\partial \varphi}{\partial x_1} & \cdots & \frac{\partial \varphi}{\partial x_n} \\ & J(f) & \end{pmatrix}.$$

Now consider the following Lagrange system associated with φ and V:

$$\mathcal{L}_u := \begin{cases} f_1 = \cdots = f_c = 0, \\ z^\top J(\varphi, f) = 0, \\ z^\top u - 1 = 0, \end{cases} \tag{2.21}$$

where $z := (z_0, z_1, \ldots, z_c) \in \mathbb{C}^{c+1}$ denotes a vector of new variables and $u \in \mathbb{C}^{c+1}$ is a fixed vector. Intuitively, the constraint $z^\top u = 1$ for a generic u is essentially equivalent to the constraint $z \neq 0$.

With abuse of notation, we denote by $\pi_x : \mathbb{C}^{n+c+1} \to \mathbb{C}^n$ the projection map sending (x, z) to the first n variables x. Let $V(\mathcal{L}_u) \subseteq \mathbb{C}^{n+c+1}$ be the set of solutions to the system (2.21). Under certain conditions (e.g., those described below), the set $\pi_x(V(\mathcal{L}_u))$ is finite. In this case, the set can be computed, for example, via the software MSOLVE based on Gröbner basis computations and represented through a so-called *rational univariate representation*. This representation is given by a sequence

$$(q, q_0, q_1, \ldots, q_n) \in \mathbb{Q}[t]^{n+2},$$

with q_0 and q coprime, such that

$$\pi_x(V(\mathcal{L}_u)) := \left\{ \left(\frac{q_1(t)}{q_0(t)}, \ldots, \frac{q_n(t)}{q_0(t)} \right) \in \mathbb{R}^n \ : \ q(t) = 0 \right\}.$$

In other words, the coordinates of the vectors in $\pi_x(V(\mathcal{L}_u))$ are represented by the evaluation of n univariate rational functions at the roots of a univariate polynomial q.

Properties of CPM. Abusing notation, define

$$\begin{aligned}\mathrm{Reg}(f) &:= \{x \in V \,:\, \mathrm{rank}(J(f)) = c\},\\ \mathrm{Sing}(f) &:= \{x \in V \,:\, \mathrm{rank}(J(f)) < c\}.\end{aligned}$$

Note that $\mathrm{Reg}(f) = \mathrm{Reg}(V)$ and $\mathrm{Sing}(f) = \mathrm{Sing}(V)$ (the sets of regular and singular points of $V = V(f)$, respectively), assuming that the following condition holds.

Assumption 1. *V is equidimensional of codimension c and $\mathcal{I}(V) = \langle f_1, \dots, f_c\rangle$. (Consequently, V is a complete intersection of dimension $d = n - c$, as defined in Section 2.2).*

A critical point of the restriction of φ to $\mathrm{Reg}(f)$ is a point $x \in \mathrm{Reg}(f)$ such that the differential of the restriction of φ to $\mathrm{Reg}(f)$ at x is not surjective; that is, an element of the constructible set

$$\begin{aligned}\mathrm{Crit}(\varphi, f) &:= \{x \in \mathrm{Reg}(f) \;:\; \mathrm{rank}(J(\varphi, f)) < c+1\}\\ &= \{x \in \mathrm{Reg}(f) \;:\; \mathrm{rank}(J(\varphi, f)) = c\}.\end{aligned}$$

Under Assumption 1, we denote $\mathrm{Crit}(\varphi, V) := \{x \in \mathrm{Reg}(V) \;:\; \mathrm{rank}(J(\varphi, f)) = c\}$.

Proposition 10. *Define*

$$\begin{aligned}\mathrm{Crit}_u(\varphi, f) &:= \{x \in \mathrm{Crit}(\varphi, f) \;:\; \exists z \in \mathbb{C}^{c+1} \text{ s.t. } z^\top J(\varphi, f) = 0,\ z^\top u = 1\},\\ \mathrm{Sing}_u(\varphi, f) &:= \{x \in \mathrm{Sing}(f) \;:\; \exists z \in \mathbb{C}^{c+1} \text{ s.t. } z^\top J(\varphi, f) = 0,\ z^\top u = 1\}.\end{aligned} \tag{2.22}$$

(a) the following property holds: $\pi_x(V(\mathcal{L}_u)) = \mathrm{Crit}_u(\varphi, f) \cup \mathrm{Sing}_u(\varphi, f)$;

(b) if Assumption 1 holds, then for a generic linear form $\varphi \in \mathbb{C}[x]_1$, the set $\mathrm{Crit}(\varphi, V)$ is finite;

(c) if $\mathrm{Crit}(\varphi, f)$ is finite, then $\mathrm{Crit}_u(\varphi, f) = \mathrm{Crit}(\varphi, f)$ for a generic u;

(d) if $\mathrm{Sing}(f)$ is finite, then $\mathrm{Sing}_u(\varphi, f) = \mathrm{Sing}(f)$ for a generic u.

Consequently, if Assumption 1 holds and $\mathrm{Sing}(V)$ is finite, then for generic φ and u, the set $\pi_x(V(\mathcal{L}_u))$ is finite and satisfies $\pi_x(V(\mathcal{L}_u)) = \mathrm{Crit}(\varphi, V) \cup \mathrm{Sing}(V)$.

Proof. **(a).** It is straightforward to verify using elementary linear algebra:

- if $x \in \mathrm{Reg}(f)$, then $x \in \mathrm{Crit}_u(\varphi, f)$ if and only if there exists $z \in \mathbb{C}^{c+1}$ such that $z^\top J(\varphi, f) = 0$ and $z^\top u = 1$, or equivalently if (x, z) satisfies (2.21);
- if $x \in \mathrm{Sing}(f)$, then $x \in \mathrm{Sing}_u(\varphi, f)$ if and only if there exists $z \in \mathbb{C}^{c+1}$ such that $z^\top J(\varphi, f) = 0$ and $z^\top u = 1$, or equivalently if (x, z) satisfies (2.21).

These facts imply that $\pi_x(V(\mathcal{L}_u)) \cap \mathrm{Reg}(f) = \mathrm{Crit}_u(\varphi, f)$ and $\pi_x(V(\mathcal{L}_u)) \cap \mathrm{Sing}(f) = \mathrm{Sing}_u(\varphi, f)$, which yields the desired claim.

(b). The claim follows from Bank et al. (2005, Lemma 7).

(c, d). For $\mathcal{X} \in \{\mathrm{Crit}(\varphi, f), \mathrm{Sing}(f)\}$, define $\mathcal{X}_u$ as the corresponding set in (2.22). We claim that if $\mathcal{X}$ is finite, then $\mathcal{X}_u = \mathcal{X}$ for a generic u. Indeed, suppose that $\mathcal{X} = \{x_1, \ldots, x_k\}$. By the definition of $\mathcal{X}$, for each $x_i \in \mathcal{X}$, there exists $z_i \in \mathbb{C}^{c+1} \setminus \{0\}$ such that $z_i^\top J(\varphi, f)|_{x_i} = 0$. Define $\mathcal{U}_i := \{u \in \mathbb{C}^{c+1} : z_i^\top u \neq 0\}$ and let $\mathcal{U} := \bigcap_{i=1}^k \mathcal{U}_i$. The set $\mathcal{U}$ is a nonempty Zariski-open set, and for each $u \in \mathcal{U}$, one has $\mathcal{X}_u = \mathcal{X}$. □

Remark 1. *The precondition of the final statement in Proposition 10 (namely, that $\mathrm{Sing}(f)$ is finite) can be shown to hold for the input system f in certain circumstances. For example, this occurs when the method of Henrion et al. (2016) (described in Section 2.2) is applied to an LMI problem $\mathcal{A}(X) = b$ with $X \in \mathbb{S}^n_+$ and generic input data $(\mathcal{A}, b)$. If $r_{\min}$ is the minimum rank, then for at least one subset $\iota \subseteq [n]$ of size $|\iota| = r_{\min}$, the system (2.11) satisfies Assumption 1, and the variety V is smooth (i.e., $\mathrm{Sing}(V) = \emptyset$) following Henrion et al. (2016, Theorem 2.1). Accordingly, in this case, the CPM solves the problem with generic φ and u; that is, it finds a rational univariate representation of at least one minimum-rank solution.*

However, for nongeneric instances $(\mathcal{A}, b)$, there are no such guarantees. It may happen, for example, that $\mathrm{Sing}(f)$ is positive-dimensional, in which case the Zariski closure of $\pi_x(V(\mathcal{L}_u))$ is expected to also be positive-dimensional, thus causing the CPM to fail.

Radical ideals. Instead of directly applying the CPM to f, one can first compute the radical ideal of $V(f)$ and attempt to derive a minimal set of polynomials g such that $\langle g \rangle = \mathcal{I}(V(f)) = \sqrt{\langle g \rangle}$. Subsequently, the CPM is applied to g. This approach is referred to as $\mathrm{CPM}_{\mathrm{RAD}}$.

In our experiments, we utilized the Macaulay2 software (Grayson and Stillman n.d.) to compute the radical and employed the command *mingens*, which attempts to find a smaller set of its generators. Note that this command is not guaranteed to produce a minimal generating set (unless the radical ideal is homogeneous). We observed that $\mathrm{CPM}_{\mathrm{RAD}}$ is capable of solving some systems that the CPM could not resolve. In the latter case, the ideal $\langle f \rangle$ is not radical, while the former is able to construct a polynomial system whose ideal is radical and a complete intersection.

2.5 Numerical Results

This section compares our method, denoted as Hybrid, with the method proposed by Henrion et al. (2016), which we refer to as HNS.[3] The details of these methods are described below.

[3] The method of Henrion et al. (2016) is implemented in the software Spectra (Henrion et al. 2019), but some details differ; for example, Spectra relies on the deprecated software FGb instead of the more recent msolve for solving zero-dimensional polynomial systems. For a fairer comparison, we used msolve both for the proposed

Hybrid. For its implementation, Algorithm 2.1 requires a numerical solution $\tilde{X}$ in the neighborhood $\tilde{\Omega} := \{\tilde{X} \in \mathbb{S}^n \ : \ \|\tilde{X} - X^*\|_F < \delta\}$ of an exact solution X^* as in Theorem 6. Since we are interested in weakly feasible SDPs, we cannot use standard IPMs, which are designed to work for instances in which both the primal and dual problems are strongly feasible. Fortunately, weakly feasible SDPs can be tackled via facial reduction approaches (Borwein and Wolkowicz 1981a; Borwein and Wolkowicz 1981b; Pataki 2013; Waki and Muramatsu 2013; Permenter et al. 2017; Permenter and Parrilo 2018; Zhu et al. 2019; Lourenço and Pataki 2022). We chose to employ the method presented in Hauenstein et al. (2021), which combines facial reduction with a NAG algorithm based on the Bertini software (Bates et al. 2013a). The resulting polynomial system (2.15) was solved with a critical point method (CPM or CPM_{RAD}), as described in Section 2.4.2.

HNS. This method constructs the system (2.11) for every subset $\iota \subseteq [n]$ of increasing cardinality $r := |\iota|$, ranging from 0 to $n-1$, and applies the critical point method to each such variety. The process stops if a PSD matrix is identified (with entries in a rational univariate representation). By Theorem 2, we can expect to obtain a minimal-rank solution in P for at least one ι of size $|\iota| = r_{\min}$. Note that in Henrion et al. (2016), only the CPM was used, which was shown to be sufficient for generic instances $(\mathcal{A}, b)$. Since we deal with nongeneric LMI instances, we tested both the CPM and the CPM_{RAD}.

The Lagrange system (2.21) was solved using the library msolve (Berthomieu et al. 2021). We say that this computation *succeeds* if the projection to the variables X is zero-dimensional and at least one of the solutions (represented in a rational univariate representation) is a PSD matrix. Otherwise, the computation *fails*.

Instances. We applied the methods to various instances of weakly feasible SDPs extracted from the literature; their descriptions are available at `https://git.ista.ac.at/jzapata/hybrid-method`. As these instances are relatively sparse, we also tested the methods on nonsparse instances. Specifically, for each SDP linear map

$$\mathcal{A}(X) := (\langle A_1, X\rangle_F, \dots, \langle A_n, X\rangle_F),$$

we created two input instances denoted as *clean* (I) and *rotated* (T). The latter uses the map

$$X \mapsto \mathcal{A}(TXT^\top) := \left(\langle T^\top A_1 T, X\rangle_F, \dots, \langle T^\top A_n T, X\rangle_F\right)$$

for some random matrix $T \in \mathrm{GL}(n, \mathbb{R})$. This transformation does not affect the feasibility of the system; however, it allows to generate additional weakly feasible SDPs that are less sparse and may present greater challenges for numerical methods (see Pataki 2020, Section 6).

method and for the algorithm in Henrion et al. (2016). Note that the method in Henrion et al. (2016) has been improved and adapted to the case when the feasible set P is nongeneric in Henrion et al. (2021), but this variant is not implemented and can be considered computationally more demanding, inasmuch as it would involve the computation of bivariate rational representations of algebraic curves instead of univariate representations.

Results. The outcomes of our experiments are shown in Table 2.1. The values $r_{\min}/r_{\max}$ given in the second column represent the ranks of the solutions found by HNS and Hybrid, respectively (recall that these methods search for minimum- and maximum-rank solutions, respectively). The results are presented in the format "t/t_{RAD}", where t denotes the runtime of the CPM and t_{RAD} is the runtime of the CPM_{RAD} (in seconds). For Hybrid, we report additionally the time required to compute the approximate solution; this value is given in square brackets. Note that this time is included in both t and t_{RAD}. If a method fails, the corresponding runtime is crossed out; for example, the entry "~~5.8~~/3.1" in the first row for HNS means that the CPM failed for all considered polynomial systems (corresponding to different subsets $\iota \subseteq [n]$), while the CPM_{RAD} succeeded for at least one system.

In cases where a method reaches the predefined time limit of 10 minutes, the computation is terminated, and the respective output is labeled with ∞.

Discussion of results. The Hybrid method resolved most of the instances both with the CPM and the CPM_{RAD}. In contrast, HNS failed, especially when the CPM was employed. In several cases, HNS with the CPM_{RAD} solved the clean version of the problem but failed on the rotated version. We conjecture that in these cases, the Macaulay2 software did not find *minimal* generators of the radical ideal.

In many cases, the polynomial systems constructed by Hybrid appear to be easier to solve compared to those in HNS; furthermore, Hybrid is usually faster than HNS (note that the latter requires solving many more systems). However, there are several exceptions: HNS with the CPM_{RAD} could solve the clean versions of the last four rows in Table 2.1, whereas Hybrid failed or did not terminate on these instances.

In Section 2.6, we detail the first example from Table 2.1 (the clean version of DruWo2017; cf. Section 2.8). Specifically, we demonstrate that in the failed cases, the varieties $\operatorname{Sing}(f)$ for the given polynomial systems f are positive-dimensional.

We confirm that the system (2.15), after fixing variables in Hybrid, indeed has a zero-dimensional real component, as predicted by Theorem 6. Furthermore, we observe that the system also has two other real components, which are positive-dimensional. After applying the CPM, the system becomes zero-dimensional, and thus Hybrid using CPM succeeds.

2.6 Numerical example

In this section, we illustrate how Hybrid and HNS work on the clean version of the DruWo2017 problem. (Recall that on this example, HNS with the CPM fails, while all other methods succeed.) Our conclusions are as follows:

- *Hybrid.* The resulting polynomial system F satisfies $\operatorname{Sing}(F) = \{(X^*, Y^*)\}$,

SDP	$n/r_{\min}/r_{\max}$	clean instances (I)		rotated instances (T)	
		HNS	HYBRID	HNS	HYBRID
DruWo2017-2.3.2P	3/1/2	~~5.8~~ / 3.1	1.1/1.6 [0.5]	~~4.6~~ / ~~12.7~~	1.2 / 1.6 [0.6]
Gupta2013-12.3P	3/1/2	~~5.9~~ / 3.1	1.7/1.9 [1.0]	~~4.8~~ / ~~12.4~~	1.7 / 2.0 [1.1]
Hauenstein2.6P	3/1/2	~~5.8~~ / 3.1	1.5/1.9 [0.8]	~~4.4~~ / ~~12.8~~	2.0 / 1.9 [1.3]
Helmberg2000-2.2.1P	3/1/2	2.6 / 2.8	1.7/1.9 [1.2]	2.1 / 3.4	1.8 / 1.8 [1.3]
LauVall2020-2.5.1P	2/1/1	1.6 / 1.8	1.1/1.3 [0.7]	1.2 / 2.0	1.1 / 1.5 [0.7]
LauVall2020-2.5.2P	3/1/2	~~6.2~~ / 3.1	1.6/1.9 [1.0]	~~4.4~~ / ~~12.5~~	1.7 / 2.4 [1.0]
Pataki2017-4P	3/1/2	~~5.9~~ / 3.0	1.5/1.9 [0.9]	~~4.5~~ / ~~12.6~~	1.6 / 2.0 [0.9]
deKlerk2002-2.1P	2/1/1	1.4 / 1.8	1.1/1.6 [0.7]	1.5 / 1.9	1.1 / 1.3 [0.7]
DruWo2017-2.3.2D	3/1/2	2.4 / 2.8	1.7/2.0 [1.2]	2.6 / 6.2	2.0 / 2.1 [1.3]
Gupta2013-12.3D	3/1/2	2.4 / 2.8	1.6/2.0 [1.1]	2.2 / 3.2	1.8 / 2.0 [1.3]
HNS2020-4.1D	4/2/2	8.7 / 9.7	2.4/2.9 [1.7]	7.0 / 17.6	2.9 / 3.1 [2.0]
Hauenstein2.6D	3/1/2	2.5 / 2.8	1.8/1.5 [1.2]	2.0 / 6.2	1.9 / 2.2 [1.3]
Helmberg2000-2.2.1D	3/1/2	~~5.9~~ / 3.0	1.6/1.8 [1.0]	~~4.1~~ / 12.3	1.8 / 2.1 [1.1]
Pataki2017-4D	3/1/2	2.4 / 2.9	1.7/1.9 [1.2]	2.0 / 3.4	2.5 / 1.8 [1.6]
Permenter2018-4.3.1D	5/0/1	11.5 / 1.0	5.5/8.1 [4.5]	1.0 / 1.5	6.6 / 6.7 [5.6]
Permenter2018-4.3.2D	4/2/2	1.0 / 1.5	2.3/2.7 [1.6]	9.1 / 129.6	2.6 / 3.1 [1.5]
PatakiCleanDim4P	4/1/-	~~49.5~~ / 4.4	~~2.3/3.3~~ [1.3]	~~77.3~~ / ~~133.4~~	~~27.2~~ / ∞ [2.6]
PatakiCleanDim5P	5/1/-	∞ / 6.9	~~154.6/519.2~~ [3.0]	∞ / ∞	∞ / ∞ [9.5]
PatakiCleanDim6P	6/1/-	∞ / 12.4	∞/∞ [11.5]	∞ / ∞	∞ / ∞ [68.4]
HeNaSa2016-6.2P	6/2/-	33.1 / 36.3	∞/∞ [9.3]	∞ / ∞	∞ / ∞ [10.5]

Table 2.1: Comparison between HYBRID and HNS for clean and rotated instances.

where (X^*, Y^*) is the desired solution, and the CPM succeeds. We provide the rational univariate representation of the solution obtained via MSOLVE.

- *HNS.* In this case, there are two polynomial systems to consider, F_1 and F_2. The sets $\mathrm{Sing}(F_1)$ and $\mathrm{Sing}(F_2)$ are both positive-dimensional, which causes the CPM to fail for both systems. Furthermore, the ideals $\langle F_1 \rangle$ and $\langle F_2 \rangle$ are not radical. We compute the generators of $\sqrt{\langle F_1 \rangle}$ and $\sqrt{\langle F_2 \rangle}$ using MACAULAY2. These generators are nonminimal in the first case and minimal in the second. Consequently, CPM_{RAD} fails for F_1 but succeeds for F_2.

This example comes from Drusvyatskiy and Wolkowicz (2017, Example 2.3.2) and can be expressed as follows for matrices $X \in \mathbb{S}^3$:

$$X_{33} = 0, \tag{2.23a}$$
$$2X_{13} + X_{22} = 1, \tag{2.23b}$$
$$X \succeq 0. \tag{2.23c}$$

Its solution set P is

$$P = \left\{ \begin{pmatrix} x & y & 0 \\ y & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} : \begin{pmatrix} x & y \\ y & 1 \end{pmatrix} \succeq 0 \right\}.$$

2.6.1 HYBRID method

The minimal face $\mathcal{F}$ containing P is given by

$$\mathcal{F} = \left\{ \begin{pmatrix} x & y & 0 \\ y & z & 0 \\ 0 & 0 & 0 \end{pmatrix} : \begin{pmatrix} x & y \\ y & z \end{pmatrix} \succeq 0 \right\}.$$

This face is described by the matrix

$$U = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{pmatrix}$$

as in (2.8) and (2.9) (with $\iota = \{1, 2\}$), which corresponds to the matrix

$$Y_U = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

Suppose that in Step 1, we find an approximate solution $\tilde{X}$ with entries $\tilde{X}_{11} = 1$ and $\tilde{X}_{12} = 0$ (and other entries are close to their unique true values). Also suppose that Step 2 correctly finds $\iota = \{1, 2\}$, which yields

$$\tilde{Y} \approx \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

If $\tilde{X}_{ij} \neq X^*_{ij}$ for all i, j, then the only way to obtain a system containing a zero-dimensional real component with a correct solution is to fix the variables $X_{11} = \tilde{X}_{11}$ and $X_{12} = \tilde{X}_{12}$. Suppose that Step 4 correctly identifies these variables; then, in Step 5, we obtain the system

$$\begin{cases} X_{33} = 0, \\ 2X_{13} + X_{22} = 1, \\ X_{11} = 1, \\ X_{12} = 0, \end{cases} \qquad \begin{pmatrix} X_{11} & X_{12} & X_{13} \\ X_{12} & X_{22} & X_{23} \\ X_{13} & X_{23} & X_{33} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ 1 \end{pmatrix} = 0^{3\times 1}. \tag{2.24}$$

Substituting $X_{11} = 1$, $X_{12} = 0$, $X_{33} = 0$, and $X_{13} = \frac{1}{2} - \frac{1}{2}X_{22}$, and eliminating these variables, yields the system

$$F := \begin{cases} Y_1 - \frac{X_{22}}{2} + \frac{1}{2} = 0, \\ X_{23} + X_{22}Y_2 = 0, \\ X_{23}Y_2 - Y_1\left(\frac{X_{22}}{2} - \frac{1}{2}\right) = 0. \end{cases} \tag{2.25}$$

Its set of solutions is

$$V(F) = \left\{ (X_{22}, X_{23}, Y_1, Y_2) = \left(2t + 1, \mp it\sqrt{2t+1}, t, \pm\tfrac{it}{\sqrt{2t+1}}\right) : t \in \mathbb{C} \setminus \left\{-\tfrac{1}{2}\right\} \right\}.$$

The solutions are real for $t \in \{0\} \cup (-\infty, -\frac{1}{2})$. Accordingly, the real variety $V(F)(\mathbb{R})$ contains a zero-dimensional component with the correct solution $(1, 0, 0, 0)$ (as predicted by Theorem 6), but also two one-dimensional components corresponding to $t \in (-\infty, -\frac{1}{2})$ with positive and negative square root branches.

To compute the set $\mathrm{Sing}(F)$, we evaluate the rank of the Jacobian matrix

$$J(F) = \begin{pmatrix} -\frac{1}{2} & 0 & 1 & 0 \\ Y_2 & 1 & 0 & X_{22} \\ -\frac{Y_1}{2} & Y_2 & \frac{1}{2} - \frac{X_{22}}{2} & X_{23} \end{pmatrix}$$

at points in $V(F)$. This rank equals two at the point corresponding to $t = 0$, and is three at all other points. Consequently,

$$\mathrm{Sing}(F) = \{(X_{22}, X_{23}, Y_1, Y_2) = (1, 0, 0, 0)\}.$$

Next, we form the Lagrange system $\mathcal{L}_u$ as in (2.21) for a randomly chosen linear map φ and a vector $u \in \mathbb{Z}^4$, and solve it using MSOLVE. The solver returns the following rational parametrization encoding the zero-dimensional solution set:

$$\pi_{X_{22},X_{23}}(V(\mathcal{L}_u)) = \left\{ \left(\frac{q_1(t)}{q'(t)}, \frac{q_2(t)}{q'(t)} \right) : q(t) = 0 \right\},$$

where $\pi_{X_{22},X_{23}}$ denotes the projection onto the variables X_{22} and X_{23}, and

$$\begin{aligned} q_1(t) &= -129t^4 - 240t^3 - 76t^2 - 16t - 4, \\ q_2(t) &= t\left(-135t^4 + 880t^3 + 60t^2 + 16t + 4\right), \\ q(t) &= 27t^5 - 220t^4 - 20t^3 - 8t^2 - 4t. \end{aligned}$$

The floating-point representation (up to five decimal places) is

$$\pi_{X_{22},X_{23}}(V(\mathcal{L}_u)) \approx \left\{ \begin{array}{l} (-10.18376, 17.84481), \\ (-0.08682, -0.16012), \\ (1, 0), \\ (0.04640 - 0.13399i, 0.08358 + 0.16089i), \\ (0.04640 + 0.13399i, 0.08358 - 0.16089i) \end{array} \right\}.$$

2.6.2 HNS method

Since HNS searches for a minimum-rank solution, we use $|\iota| = 1$. Consequently, three possibilities arise for the kernel matrix:

$$\begin{pmatrix} Y_{11} & Y_{12} \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ Y_{21} & Y_{22} \\ 0 & 1 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ Y_{31} & Y_{32} \end{pmatrix}.$$

It can be verified that for the third matrix Y, the system $XY = 0$ does not admit solutions with $X \in P$. We analyze the two remaining systems below.

(1) Substituting $X_{33} = 0$ and $X_{13} = \frac{1}{2} - \frac{1}{2}X_{22}$, and eliminating these variables, yields

$$\begin{pmatrix} X_{11} & X_{12} & \frac{1}{2} - \frac{X_{22}}{2} \\ X_{12} & X_{22} & X_{23} \\ \frac{1}{2} - \frac{X_{22}}{2} & X_{23} & 0 \end{pmatrix} \begin{pmatrix} Y_{11} & Y_{12} \\ 1 & 0 \\ 0 & 1 \end{pmatrix} = 0^{3\times 2}$$

$$\Longleftrightarrow$$

$$F_1 := \begin{cases} X_{12} + X_{11}Y_{11} = 0, \\ X_{22} + X_{12}Y_{11} = 0, \\ X_{11}Y_{12} - \frac{X_{22}}{2} + \frac{1}{2} = 0, \\ X_{23} + X_{12}Y_{12} = 0, \\ -Y_{12}\left(\frac{X_{22}}{2} - \frac{1}{2}\right) = 0. \end{cases}$$

Recall that the equation at position $(3,1)$ in the 3×2 matrix of polynomial equations is implied by the other equations (see Step 5 of Algorithm 2.1); accordingly, this equation is omitted in the description of F_1. This polynomial system has the following parametric description:

$$V(F_1) = \left\{ (X_{11}, X_{12}, X_{22}, X_{23}, Y_{11}, Y_{12}) = \left(\frac{1}{t^2}, -\frac{1}{t}, 1, 0, t, 0 \right) : t \in \mathbb{C} \setminus \{0\} \right\}.$$

By evaluating the Jacobian and substituting the above parameterization, it can be seen that the matrix has rank four:

$$J(F_1) := \begin{pmatrix} Y_{11} & 1 & 0 & 0 & X_{11} & 0 \\ 0 & Y_{11} & 1 & 0 & X_{12} & 0 \\ Y_{12} & 0 & -\frac{1}{2} & 0 & 0 & X_{11} \\ 0 & Y_{12} & 0 & 1 & 0 & X_{12} \\ 0 & 0 & -\frac{Y_{12}}{2} & 0 & 0 & \frac{1}{2} - \frac{X_{22}}{2} \end{pmatrix}$$

$$= \begin{pmatrix} t & 1 & 0 & 0 & \frac{1}{t^2} & 0 \\ 0 & t & 1 & 0 & -\frac{1}{t} & 0 \\ 0 & 0 & -\frac{1}{2} & 0 & 0 & \frac{1}{t^2} \\ 0 & 0 & 0 & 1 & 0 & -\frac{1}{t} \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

Moreover, one can see that the ideal $\mathcal{J} := \langle F_1 \rangle$ is not radical, since $(X_{22} - 1)^2 \in \mathcal{J}$ but $(X_{22} - 1) \notin \mathcal{J}$. Computing the radical ideal using Macaulay2 yields

$$\sqrt{\mathcal{J}} = \langle X_{23}, Y_{12}, X_{22} - 1, X_{12}^2 - X_{11}, X_{12}Y_{11} + 1, X_{12} + X_{11}Y_{11} \rangle.$$

However, in this case, the set of generators is not minimal, since

$$X_{12} + X_{11}Y_{11} = X_{12}(X_{12}Y_{11} + 1) - Y_{11}(X_{12}^2 - X_{11}).$$

(2) Similarly, for the second system we have

$$\begin{pmatrix} X_{11} & X_{12} & \frac{1}{2} - \frac{X_{22}}{2} \\ X_{12} & X_{22} & X_{23} \\ \frac{1}{2} - \frac{X_{22}}{2} & X_{23} & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ Y_{21} & Y_{22} \\ 0 & 1 \end{pmatrix} = 0^{3 \times 2}$$

$$\Updownarrow$$

$$F_2 : \begin{cases} X_{11} + X_{12}Y_{21} &= 0 \\ X_{12} + X_{22}Y_{21} &= 0 \\ X_{23}Y_{21} - \frac{X_{22}}{2} + \frac{1}{2} &= 0 \\ X_{23} + X_{22}Y_{22} &= 0 \\ X_{23}Y_{22} &= 0 \end{cases}$$

Again, the ideal $\mathcal{J} = \langle F_2 \rangle$ is not radical but when we obtain its radical in Macaulay2 we get a minimal set of generators:

$$\sqrt{\mathcal{J}} = \langle X_{23}, Y_{22}, X_{12} + Y_{21}, X_{22} - 1, X_{11} - Y_{21}^2 \rangle$$

To summarize, the CPM fails for F_1 and F_2, while the CPM_{RAD} fails for F_1 but succeeds for F_2.

Remark 2. *As Table 2.1 indicates, HNS with both the CPM and the CPM_{RAD} fails for the rotated version of the DruWo2017 problem. The resulting polynomial systems are too large to analyze manually; we conjecture that in these cases, the Macaulay2 software also fails to find minimal generators for the radical ideals, as was observed for system F_1.*

2.7 Conclusions and future work

This chapter has presented a hybrid method for certifying weakly feasible SDP problems that utilizes an approximate numerical SDP solver and an exact solver for polynomial systems over the real numbers. Our numerical results indicate that this hybrid method can outperform the pure exact algorithm proposed in Henrion et al. (2016) when provided with a good approximate solution.

In our current experiments, scalability was limited both by the numerical SDP solver (which was the facial reduction algorithm (Hauenstein et al. 2021) based on Bertini (Bates et al. 2013a)) and by the exact solver for polynomial systems (namely, msolve). We conjecture that our approach can handle larger instances if the method presented in Hauenstein et al. (2021) is replaced with an alternative facial reduction algorithm, and the exact solver for polynomial systems is replaced with a numerical solver; this is left for future work. Potential candidates for the latter could be algorithms from the field of NAG, such as Bertini. Employing such a solver could address the issue discussed in the previous sections: if a system of polynomial equations has a zero-dimensional component and a positive-dimensional component, a numerical solver would focus on the desired zero-dimensional component if the previous step found a good approximation.

2.8 Description of SDP Problems

The numerical efficacy of the proposed hybrid certification algorithm is evaluated against a collection of weakly feasible SDPs used for the numerical results shown in Table 2.1. These instances are drawn from the foundational literature and are widely recognized as benchmark problems due to their inherently degenerate geometries, which often cause standard IPMs to stall or fail. We incorporate instances from Laurent and Vallentin (2020), Klerk (2002), and Helmberg (2000), who construct pathological problems specifically to illustrate the stark differences in degeneracy behavior between SDPs and the classical LP case. We also present instances formulated by Drusvyatskiy and Wolkowicz (2017), which demonstrate the necessity of facial reduction techniques, alongside instances from Permenter and Parrilo (2018), which explore partial facial reduction methodologies. Additionally, we include instances from Henrion et al. (2016) and Henrion

et al. (2015b), which were utilized during the development of exact symbolic algorithms for LMIs and real root-finding for low-rank matrices. Finally, to rigorously test the scaling limits of our exact certification approach, we include the highly structured family of instances introduced by Pataki (2017) and Pataki (2020), which are known to induce severe ill-conditioning and illustrate the duality gap phenomena occurring in SDPs.

The following list details the weakly feasible SDPs evaluated in Table 2.1. Each instance includes its source reference and a formulation in the standard primal form:

$$\min_{X \in \mathbb{S}^n_+} \{ \langle C, X \rangle_F \ : \ \mathcal{A}(X) = b \},$$

with $X := (x_{ij}) \in \mathbb{S}^n$.

DruWo2017-2.3.2PStd (Drusvyatskiy and Wolkowicz 2017)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & x_{22} \\ \text{s.t.} \quad & 2x_{13} + x_{22} = 1, \\ & x_{33} = 0. \end{aligned}$$

Gupta2013-12.3PStd (Gupta 2013)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & x_{33} \\ \text{s.t.} \quad & 1 - x_{33} - 2x_{12} = 0, \\ & x_{22} = 0. \end{aligned}$$

Hauenstein2.6PStd (Hauenstein et al. 2021)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & x_{11} \\ \text{s.t.} \quad & x_{11} + 2x_{23} = 2, \\ & x_{22} = 0. \end{aligned}$$

Helmberg2000-2.2.1PStd (Helmberg 2000)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & x_{12} \\ \text{s.t.} \quad & x_{33} - x_{12} = 1, \quad x_{11} = 0, \\ & x_{13} = 0, \quad x_{23} = 0. \end{aligned}$$

LauVall2020-2.5.1PStd (Laurent and Vallentin 2020)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & -2x_{12} \\ \text{s.t.} \quad & x_{11} = 1, \\ & x_{22} = 0. \end{aligned}$$

LauVall2020-2.5.2PStd (Laurent and Vallentin 2020)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & -x_{11} - x_{22} \\ \text{s.t.} \quad & x_{11} = 0, \\ & 2x_{13} + x_{22} = 1. \end{aligned}$$

Pataki2017-4PStd (Pataki 2017)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & x_{11} + x_{22} \\ \text{s.t.} \quad & x_{11} = 0, \\ & 2x_{13} + x_{22} = 1. \end{aligned}$$

deKlerk2002-2.1PStd (Klerk 2002)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & 2x_{12} + x_{22} \\ \text{s.t.} \quad & x_{11} = 0, \\ & x_{22} - 1 = 0. \end{aligned}$$

DruWo2017-2.3.2DStd (Drusvyatskiy and Wolkowicz 2017)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & x_{22} \\ \text{s.t.} \quad & x_{11} = 0, \quad x_{12} = 0, \\ & x_{22} - x_{13} - 1 = 0, \quad x_{23} = 0. \end{aligned}$$

Gupta2013-12.3DStd (Gupta 2013)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & x_{33} \\ \text{s.t.} \quad & x_{11} = 0, \quad 2x_{13} = 0, \\ & 2x_{23} = 0, \quad x_{33} - x_{12} - 1 = 0. \end{aligned}$$

HNS2020-4.1DStd (Henrion et al. 2015a)

$$\begin{aligned} \min_{X \in \mathbb{S}^4_+} \quad & x_{11} + 2x_{22} + 4x_{34} \\ \text{s.t.} \quad & x_{11} = 1, \quad x_{13} = 0, \quad x_{14} = 0, \\ & x_{22} = 2, \quad x_{23} = 0, \quad x_{24} = 0, \\ & x_{33} - 2x_{12} = 0, \quad 2x_{34} = 4, \quad x_{44} - x_{12} = 0. \end{aligned}$$

Hauenstein2.6DStd (Hauenstein et al. 2021)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & x_{11} \\ \text{s.t.} \quad & 2x_{12} = 0, \quad 2x_{13} = 0, \\ & 2x_{23} - 2x_{11} + 2 = 0, \quad x_{33} = 0. \end{aligned}$$

Helmberg2000-2.2.1DStd (Helmberg 2000)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & x_{12} \\ \text{s.t.} \quad & x_{22} = 0, \\ & 2x_{12} + x_{33} - 1 = 0. \end{aligned}$$

Pataki2017-4DStd (Pataki 2017)

$$\begin{aligned} \min_{X \in \mathbb{S}^3_+} \quad & x_{11} + x_{22} \\ \text{s.t.} \quad & 2x_{12} = 0, \quad x_{22} - x_{13} - 1 = 0, \\ & 2x_{23} = 0, \quad x_{33} = 0. \end{aligned}$$

Permenter2018-4.3.2DStd (Permenter and Parrilo 2018)

$$\begin{aligned} \min_{X \in \mathbb{S}^4_+} \quad & x_{11} - x_{22} - x_{33} + x_{44} \\ \text{s.t.} \quad & x_{11} = 1, \quad x_{13} = 0, \quad x_{14} + x_{23} = 0, \\ & x_{24} = 0, \quad 2x_{12} + x_{33} + 1 = 0, \\ & x_{22} + 2x_{34} = -1, \quad x_{44} = 1. \end{aligned}$$

Permenter2018-4.3.1DStd (Permenter and Parrilo 2018)

$$\begin{aligned} \min_{X \in \mathbb{S}^5_+} \quad & 0 \\ \text{s.t.} \quad & x_{12} = 0, \quad x_{13} = 0, \quad x_{14} = 0, \quad x_{15} = 0, \\ & x_{11} + x_{22} = 0, \quad x_{24} = 0, \quad x_{25} = 0, \\ & x_{34} = 0, \quad x_{35} = 0, \quad x_{33} - x_{23} + x_{44} = 0, \\ & x_{45} = 0. \end{aligned}$$

PatakiCleanDim4PStd (Pataki 2020)

$$\begin{aligned} \min_{X \in \mathbb{S}^4_+} \quad & x_{11} + x_{22} + x_{33} \\ \text{s.t.} \quad & x_{11} = 0, \quad 2x_{14} + x_{22} = 0, \\ & 2x_{24} + x_{33} = 10. \end{aligned}$$

PatakiCleanDim5PStd (Pataki 2020)

$$\begin{aligned} \min_{X \in \mathbb{S}^5_+} \quad & x_{11} + x_{22} + x_{33} + x_{44} \\ \text{s.t.} \quad & x_{11} = 0, \quad 2x_{15} + x_{22} = 0, \\ & 2x_{25} + x_{33} = 0, \quad 2x_{35} + x_{44} = 10. \end{aligned}$$

PatakiCleanDim6PStd (Pataki 2020)

$$\begin{aligned} \min_{X \in \mathbb{S}^6_+} \quad & x_{11} + x_{22} + x_{33} + x_{44} + x_{55} \\ \text{s.t.} \quad & x_{11} = 0, \quad 2x_{16} + x_{22} = 0, \\ & 2x_{26} + x_{33} = 0, \quad 2x_{36} + x_{44} = 0, \\ & 2x_{46} + x_{55} = 10. \end{aligned}$$

HeNaSa2016-6.2PStd (Henrion et al. 2021)

$$
\begin{aligned}
\min_{X \in \mathbb{S}^6_+} \quad & x_{11} + 3x_{15} + 2x_{34} + x_{25} + x_{33} + 2x_{44} - x_{46} + x_{56} + x_{66} \\
\text{s.t.} \quad & x_{11} = 1, \quad x_{12} = 0, \quad x_{14} = 0, \\
& 2x_{13} + x_{22} = 0, \quad 2x_{23} = 1, \\
& 2x_{15} + 2x_{24} = -3, \quad x_{33} = 1, \\
& 2x_{25} + 2x_{34} = -4, \quad x_{35} = 0, \\
& 2x_{16} + x_{44} = 2, \quad x_{26} + x_{45} = 0, \\
& 2x_{46} = 1, \quad 2x_{36} + x_{55} = 0, \\
& 2x_{56} = 1, \quad x_{66} = 1.
\end{aligned}
$$

List of Notation for Chapter 2

$\langle M, N \rangle_F$	Frobenius inner product, defined as Trace(MN).
$\|\cdot\|_F$	Frobenius norm, $\|M\|_F = \sqrt{\langle M, M \rangle_F}$.
$\mathcal{A}(X)$	A affine mapping $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$.
$\mathcal{F}$	A face of the positive semidefinite cone.
I_n	The $n \times n$ identity matrix.
$\mathcal{V}_{r,\iota}$	The incidence variety associated with rank r and the index set ι.
$\mathcal{W}_{r,\iota}$	The projection of the incidence variety $\mathcal{V}_{r,\iota}$ onto the space of symmetric matrices.
$\lambda_i(A)$	The i-th largest eigenvalue of a symmetric matrix A.
$\mathbb{R}^n, \mathbb{C}^n$	n-dimensional real and complex Euclidean spaces.
$\sigma_i(A)$	The i-th singular value of matrix A.
$\mathbb{S}^n$	Space of $n \times n$ real symmetric matrices.
$\mathbb{S}^n_+$	Cone of positive semidefinite matrices (PSD cone).
$\mathbb{S}^n_{++}$	Interior of the PSD cone (positive definite matrices).
X^*	An exact solution of the SDP.
$\tilde{X}$	A numerical approximate solution of the SDP.
$X \succeq 0$	Matrix X is positive semidefinite.
$X \succ 0$	Matrix X is positive definite.

Chapter 3

Computing isolated singular solutions of polynomial systems

The content of this chapter is based on unpublished joint work (currently in preparation) with Vladimir Kolmogorov and Mikhail Karapetyants.

3.1 Introduction

We consider the problem of numerically solving a system of equations $f(z) = 0$ with a zero-dimensional set of solutions. Here f is an analytic mapping $\mathbb{C}^n \to \mathbb{C}^n$. A standard approach in NAG for tackling this problem is as follows. First, one constructs a homotopy function $h(z,t) = (1-t)f(z) + tg(z)$ where g is a polynomial system with easily computable roots. Consider one such root $z_{\texttt{root}}$. If g is chosen generically then there exists a unique smooth function $z : (0,1] \to \mathbb{C}^n$ with $z(1) = z_{\texttt{root}}$ and $h(z(t),t) = 0$ for all $t \in (0,1]$. Furthermore, if $z((0,1])$ is bounded then the limit $z^* = z(0) = \lim_{t\to 0} z(t)$ exists and is a root of f. The latter condition will always be satisfied if system f is homogeneous. If g is chosen to have sufficiently many roots then every root of f will be covered with probability 1 (Sommese and Wampler 2005, Theorem 8.4.1).

By differentiating equation $h(z(t),t) = 0$ with respect to t one obtains Davidenko ODE:

$$\dot{z}(t) = -h_z(z(t),t)^{-1} h_t(z(t),t) \tag{3.1}$$

We now need to track curve $z(t)$ by numerically solving this ODE.

This problem becomes especially challenging when t approaches zero and $z^* = z(0)$ is a singular solution z^*, i.e. the Jacobian $f_z(z^*)$ is singular. This regime, referred to as the "endgame", constitutes the main focus of the present chapter. The following assumptions are maintained throughout:

Assumption 2. (a) $h : \mathbb{C}^{n+1} \to \mathbb{C}^n$ *is a polynomial mapping, and* $f(z) = h(z,0)$.
(b) *Point* $z^* \in \mathbb{C}^n$ *is an isolated solution of* $f(z)$. *We denote* $x^* = (z^*, 0)$, $J^* = h_z(x^*) = f_z(z^*)$ *and* $h_t^* = h_t(x^*)$. *We also let* $\kappa = n - \texttt{rank}(J^*)$ *be the corank of* J^*.

(c) $\texttt{rank}(J^*) < n$, *i.e.* z^* *is a singular solution of* f.
(d) h_t^* *is linearly independent of columns in* J^*, *i.e.*

$$\texttt{rank}([J^*|h_t^*]) = \texttt{rank}(J^*) + 1.$$

(e) *There exists a finite set* Π *of formal Puiseux series of the form*

$$z(t) = z^* + \sum_{j=1}^{\infty} a_j t^{j/c} = z^* + a_{k_1} t^{k_1/c} + a_{k_2} t^{k_2/c} + \ldots \tag{3.2}$$

where $c \in \mathbb{N}$ *and* $1 \leq k_1 < k_2 < k_3 \ldots$ *is an increasing sequence of integers corresponding to non-zero coefficients* a_j *such that:*
(i) *each series*[1] $z(\cdot) \in \Pi$ *is convergent in some neighborhood* $\Omega_t \subseteq \mathbb{C}$ *of* 0 *and satisfies* $h(z(t), t) = 0$ *for* $t \in \Omega_t$*;*
(ii) *for every open set* $U \subseteq \Omega_t$ *and every* $z(\cdot) \in \Pi$*, function* $\phi(t) = \det[h_z(z(t), t)]$ *is not identically zero on* U*;*
(iii) *there exists neighborhood* Ω_z *of* z^* *such that for every* $(z, t) \in \Omega_z \times \Omega_t$ *with* $h(z, t) = 0$ *there exists* $z(\cdot) \in \Pi$ *with* $z(t) = z$.

Note, if (a)- (c) hold then conditions (d,e) will be satisfied with probability 1 if the start system $g(z)$ in the definition of homotopy h is sufficiently generic.

3.1.1 Corank-1 problems

Our first contribution is the following result.

Theorem 11. *Suppose that* $\kappa = 1$. *There exists a neighborhood* Ω *of* x^**, constants* $\alpha > 1, \beta_{\min} > 0$ *and an algorithm that takes point* $x_\circ = (z_\circ, t_\circ) \in \mathbb{C}^{n+1}$*, parameters* $\beta, k_1^{\max}$ *and does the following: if* $\beta > \beta_{\min}$*,* $\|h(x_\circ)\| \leq |t_\circ|^\beta$ *and index* k_1 *of each Puiseux series* $z(\cdot) \in \Pi$ *satisfies* $k_1 \leq k_1^{\max}$ *then it produces a (possibly infinite) sequence of points* $x_1, \ldots, x_K$ *such that (i)* $\|x_k - x^*\| \leq \|x_\circ - x^*\|^{\alpha^k}$ *for each* $k \in [K]$*, and (ii) if* $K < \infty$ *then the last point* $x = x_K = (z, t)$ *satisfies* $\|h(x)\| \leq |t|^\beta$*. It uses* $O(\log(1 + \beta) + K)$ *evaluations of function* $h(\cdot)$ *and its Jacobian.*

A recursive application of the algorithm in Theorem 11 immediately gives an algorithm with a superlinear convergence rate.

Corollary 12. *Suppose that* $\kappa = 1$. *There exists a neighborhood* Ω *of* x^**, constants* $\alpha > 1, \beta_{\min} > 0$ *and an algorithm that takes point* $x_\circ = (z_\circ, t_\circ) \in \mathbb{C}^{n+1}$*, parameters* $\beta, k_1^{\max}$ *and does the following: if* $\beta > \beta_{\min}$*,* $\|h(x_\circ)\| \leq |t_\circ|^\beta$ *and index* k_1 *of each Puiseux series* $z(\cdot) \in \Pi$ *satisfies* $k_1 \leq k_1^{\max}$ *then it produces a sequence of points* $x_1, x_2, \ldots$ *such that* $\|x_k - x^*\| \leq \|x_\circ - x^*\|^{\alpha^k}$ *for each* $k \geq 1$. *Computing each subsequent point uses* $O(\log(1 + \beta))$ *evaluations of function* $h(\cdot)$ *and its Jacobian.*

[1] When writing $z(\cdot) \in \Pi$, we will assume with some abuse of notation that we have chosen not only the formal series but also the specific branch $t \mapsto t^{1/c}$ used in (3.2) (unless noted otherwise).

The algorithm in Theorem 11 is achieved by combining two steps.

- Predictor phase: estimate coefficients $z^*, k_1/c, a_{k_1}$ of a Puiseux series $z(\cdot) \in \Pi$, then evaluate the obtained approximation of $z(\cdot)$ at $t = 0$ obtaining predictor point $\hat{x} = (\hat{z}, 0)$. We use a new rule for estimating k_1/c, and formally analyze the accuracy of the resulting approximation.
- Corrector phase: augment the system $h(x) = 0$ with a new linear equation given by a hyperplane passing through $\hat{x}$, whose normal is tangent to the solution curve. Solve the augmented system by applying several steps of the Newton's method.

The idea of adding an extra hyperplane is common in the *(pseudo)-arclength* continuation methods (Wempner 1971; Riks 1972; Keller 1977). Our scheme differs in the choice of the step size. To our knowledge, existing arclength methods control the step size via parameter Δs representing the Euclidean length along the solution curve. We are not aware of methods that explicitly combine an arclength method together with a Puiseux series-aware predictor.

Due to this connection, we call our method the *arclength endgame*, even though it does not use the Euclidean length of the curve in any way.

Related work. Below we discuss papers that give an explicit superlinear convergence rate when $t \to 0$. One approach is the *deflation* technique (Ojika et al. 1983; Leykin et al. 2006; Leykin et al. 2008), which introduces new auxiliary variables and new equations such that z^* becomes an isolated **nonsingular** root of the extended system. By classical results, applying the Newton's method for such system would give an algorithm with a quadratic convergence rate. In general, the size of the extended system can be $\Theta(n2^{\mu})$ where μ is the multiplicity of z^*. For corank-1 systems, more efficient techniques (without an exponential dependence on μ) with a guaranteed quadratic convergence rate have been proposed in Li and Zhi (2012) and Li and Zhi (2022).

Note that these techniques rely on computing polynomials in the Max Noether space; these are polynomials that are linear combinations of higher-order derivatives of the original equations, and evaluate to 0 at z^*. Thus, they require computing additional derivatives on the input system. In contrast, the algorithm in Theorem 11 uses only evaluations of function h and its Jacobian.

3.1.2 Corank $\kappa \geq 2$

Let us now consider systems with $\kappa \geq 2$. We investigate a heuristic algorithm that can be viewed as an extension of the arclength endgame. Given an initial point $x_\circ = (z_\circ, t_\circ)$, we first compute predictor $\hat{x} = (\hat{z}, \hat{t})$. We then introduce $\kappa - 1$ new variables ξ

and κ new linear hyperplanes passing through $\hat{x}$, obtaining an extended system with Jacobian J satisfying $\|J^{-1}\| \leq O(1)$. This system is solved using several steps of the Newton's method, producing new point (v, t, ξ). Experimentally, we observed that usually this step has a superlinear convergence rate (i.e. $\|v - z^*\| \leq \|z_\circ - z^*\|^{1+\Theta(1)}$), assuming that we in the endgame zone. However, new point (v, t) is no longer on the homotopy h. To continue, we change the homotopy to $h^v(z, t) = f(z) - tf(v)$, and continue the process starting with $(v, 1)$. Note that this becomes an algorithm for refining a solution close to z^* rather than for following a specified homotopy.

We call this procedure a *lifted arclength endgame*. We investigate its properties in section 3.4.2, and compare with the classical power-series endgame.

3.1.3 Estimating coefficients of the Puiseux series

Computing a predictor requires estimating ratios k_i/c for $i = 1, 2, ...$ of the current Puiseux series. One classical approach to tracking the solution path close to a singular root is the power-series endgame, which maintains a set of sample points and approximates the path using a truncated fractional power series (Sommese and Wampler 2005; Bates et al. 2013b). To project the path toward the target root $t = 0$, standard software typically utilizes a linear predictor or a cubic predictor. The classical cubic predictor is constructed via Hermite interpolation by matching the path positions and tangent derivatives at two consecutive sample points (Morgan et al. 1992a; Sommese and Wampler 2005). However, these classical formulations generally assume a dense, sequential set of fractional exponents (e.g., $1/c, 2/c, 3/c$). This assumption fails to capture the geometry of sparse Puiseux series, where the valid fractional powers are strictly governed by the value semigroup of the local ring at the singularity (Zariski and Samuel 1965; Wall 2004). Additionally, as demonstrated by polyhedral endgames (Huber and Sturmfels 1995), the vector of leading fractional exponents represents a fundamental geometric property of the variety. In this framework, the fractional exponents defining the path direction correspond directly to the inner normals of the facets characterizing the system's Newton polytope.

Accurately estimating the cycle number c and the fractional exponents k_i/c is another critical phase of the singular endgame. Traditional methods rely heavily on heuristic testing, such as the trial-and-error method, which evaluates prediction errors across a range of candidate integer values for c (Morgan et al. 1992a; Sommese and Wampler 2005). Another established technique is the geometric sequence sampling approach, which isolates the leading fractional exponent by analyzing the logarithmic differences of path samples taken at geometrically decreasing parameter values (Bates et al. 2011; Sommese and Wampler 2005). Other classical variants include Cauchy integral method (Bates et al. 2011; Sommese and Wampler 2005).

Contributions. The main contributions in this chapter regarding the prediction and estimation phases are detailed in Section 3.5. In the first

part (Sections 3.5.1 and 3.5.2), assuming we have a polynomial system $f(z)$ possessing an isolated singular root z^* and an initial tracked point $(z_\circ, t_\circ)$, we provide the explicit algebraic construction of the predictor function $\hat{z}(t)$. We rigorously derive the approximation error bounds, proving that the truncation error decays super-linearly as a fractional power of the current distance to the root. This guarantees that the newly predicted point $(\hat{z}, \hat{t})$ lies significantly closer to the target root z^*.

In the final part (Section 3.5.3), we provide an overview of practical methodologies for estimating the fractional exponents. Alongside the established trial-and-error and geometric sequence methods, we introduce a novel path-limit estimation approach that isolates the first fractional exponent directly from the continuous limit of a quotient of finite differences evaluated along the path. Furthermore, we extend the geometric sequence approach with a new recursive heuristic designed to systematically annihilate leading terms, allowing for the numerical estimation of higher-order fractional exponents.

3.2 Background and notation

For a function $F : \mathbb{C}^n \to \mathbb{C}^m$ and variables $x = (u, v)$ the Jacobian of F with respect to u is denoted either as $F_u(x)$ or as $D_u F(x)$.

Notations z and $z(\cdot)$ will denote different objects: $z(\cdot)$ is a function, while z is a specific value which is not necessarily related to $z(\cdot)$.

Throughout this chapter, for a point $(z, t) \in \mathbb{C}^{n+1}$ we denote

$$\dot{z} = -h_z(z(t), t)^{-1} h_t(z(t), t).$$

This definition depends also on t; the value of t should always be clear from the context. Note, if $z = \bar{z}(t)$ for a differentiable function $\bar{z}(\cdot)$ satisfying $h(\bar{z}(t), t) = 0$ in some neighborhood of t then $\dot{z} = \frac{d}{dt}\bar{z}(t)$.

We define variety $\mathcal{V} \subseteq \mathbb{C}^{n+1}$ as $\mathcal{V} = h^{-1}(0, 0)$.

3.2.1 Power-series endgame

One classical approach to tracking the path close to the root is the *power-series endgame*. It maintains a set of pairs of the form

$$\mathcal{X} = \{x_i = (z_i, t_i)\}_{i=0,1,\ldots}$$

where $t_i \in (0, 1]$ and z_i approximates $z(t_i)$. At each step it does the following.

- Using pairs in $\mathcal{X}$, estimate the first $\ell + 1$ coefficients of series (3.2) together with ratios k_i/c, obtaining approximation

$$\hat{z}(t) = \hat{z}^* + \hat{a}_{k_1} t^{k_1/c} + \ldots + \hat{a}_{k_\ell} t^{k_\ell/c}. \tag{3.3}$$

- Select "target" value $\hat{t} \in (0, 1]$. Usually one takes $\hat{t} = \rho \cdot t_\circ$ where $t_\circ$ is the smallest value present in $\mathcal{X}$, and parameter $\rho \in (0, 1)$ is either fixed or updated adaptively based on the success / failure status of previous steps.
- Predictor step: compute vector $\hat{z} = \hat{z}(\hat{t})$.
- Corrector step: compute z by applying several steps of the Newton's method to solve system $h(z, \hat{t}) = 0$ using $\hat{z}_p$ as the starting point. If the Newton's method converges according to a certain criterion then add pair $(z, \hat{t})$ to $\mathcal{X}$.

Popular choices for the predictor are a *linear predictor* (that estimates $z^*, a_{k_1}, k_1/c$) and a *cubic predictor* (that assumes that $(k_1, k_2, k_3) = (1, 2, 3)$ and estimates $z^*, a_{k_1}, a_{k_2}, a_{k_3}, c$). We refer to Section 3.5 for a further discussion of predictors.

3.2.2 Newton's method

In this section we state the classical Kantorovich theorem about convergence of the Newton's method which we will need later (Ferreira and Svaiter 2012).

Theorem 13. *Let X and Y be Banach spaces, Ω be a subset of X and F be a continuous non-linear operator, $F : \Omega \mapsto Y$, such that F is continuously Frechet-differentiable on $\operatorname{int}(\Omega)$. For an initial guess $x_0 \in \Omega$ and for positive reals $L, C \in \mathbb{R}_+$ assume that*

- $F'(x_0)$ *is non-singular;*
- $\left\| [F'(x_0)]^{-1} \left(F'(x) - F'(y) \right) \right\| \leq L\|x - y\| \qquad \forall x, y \in \Omega;$
- $\left\| [F'(x_0)]^{-1} F(x_0) \right\| \leq C;$
- $2CL < 1.$

Consider $r \in [r_-, r_+]$ where

$$r_- = \frac{1 - \sqrt{1 - 2CL}}{L}, \qquad r_+ = \frac{1 + \sqrt{1 - 2CL}}{L}.$$

If $B(x_0, r) = \{x \in X : \|x - x_0\| < r\} \subset \Omega$ then the sequence $\{x_k\}$ generated by Newton's method for solving non-linear equation $F(x) = 0$ with initial point x_0,

$$x_{k+1} = x_k - [F'(x_k)]^{-1} F(x_k) \qquad \forall k \geq 0,$$

is contained in $B(x_0, r)$, converges to the unique zero $x^* \in B(x_0, r)$ of F and the following error bound holds:

$$\|x_{k+1} - x^*\| \leq \frac{L}{2\sqrt{1 - 2CL}} \|x_k - x^*\|^2 \qquad \forall k \geq 0$$

3.3 Linear predictor

In this section we consider the following predictor at point $x = (z, t)$. It depends on parameters $\gamma, \beta, k_1^{\max}$; these are positive constants that will be specified later.

1. Set $t_1 = (1 - |t|^\gamma) \cdot t$.
2. Run the Newton's method to solve the system $h(z, t_1) = 0$ starting with a point
$$z_0 = z + \dot{z}(t_1 - t),$$
until getting a point z_1 with $\|h(z_1, t_1)\| \le |t_1|^\beta$.
3. Find positive integers c, k_1 with $k_1 \le k_1^{\max}$ that minimize
$$\left| \frac{\|t\dot{z}(t) - t_1\dot{z}(t_1)\|}{\|z(t) - z(t_1)\|} - \frac{k_1}{c} \right|.$$
4. Output predictor $\hat{z} = z - \frac{c}{k_1}\dot{z}t$.

We will prove the following result.

Theorem 14. *There exist constants $\gamma_{\min} > 0$, $\eta > 0$ with the following property. Suppose that $\gamma > \gamma_{\min}$, $\beta > \eta\gamma$, and index k_1 of each Puiseux series $z(\cdot) \in \Pi$ satisfies $k_1 \le k_1^{\max}$. Then there exists a neighborhood Ω of x^* such that any $x = (z, t) \in \Omega$ with $\|h(x)\| \le |t|^\beta$ satisfies the following.*
(i) $\|\hat{z} - z^*\| = O\left(\|z - z^*\|^{k_2/k_1}\right)$ *where k_1, k_2 are the indices in eq. (3.2) of the Puiseux series $z(\cdot) \in \Pi$ with the smallest ratio k_2/k_1.*
(ii) *The Newton's method in step 2 terminates after $O(\log(1 + \beta))$ iterations. If $\kappa = 1$ then $\gamma_{\min} < 1$.*

The remainder of this section is devoted to the proof of this theorem. In these proofs we will often omit the phrase "there exists a neighborhood Ω of x^* such that ...", making it implicit. For example, we will write $O(|t|^a) \le |t|^b$ when $a > b$; this would hold if $|t|$ is sufficiently small. Also, in each lemma we will implicitly assume that the current Ω is contained in the neighborhoods considered in all previous statements. One of them is the neighborhood $\Omega_z \times \Omega_t$ defined in Assumption 2(e), so all Puiseux series $z(\cdot) \in \Pi$ will be assumed to be convergent in the considered neighborhood.

First, we analyze what happens when the points lie exactly on the curve.

Lemma 15. *Consider Puiseux series $z(\cdot) \in \Pi$ associated with integers c, k_1, k_2. There exists a neighborhood Ω of x^* such points $x = (z, t) = (z(t), t) \in \Omega$ satisfy the following.*
(a) $\lim_{t \to 0} \sup_{t_1 \in [0,t)} \left| \frac{\|t\dot{z}(t) - t_1\dot{z}(t_1)\|}{\|z(t) - z(t_1)\|} - \frac{k_1}{c} \right| = 0$.
(b) $\|\hat{z} - z^*\| = O(\|z - z^*\|^{k_2/k_1})$ *assuming that $\hat{z}$ was computed with the correct value of k_1/c.*

Proof of Part (a). Let us analyze the behavior of the finite difference quotient along the solution path. By the local uniformization theorem, the solution path $z(t)$ and the auxiliary function $\zeta(t) := t\dot{z}(t)$ admit fractional Puiseux series expansions:

$$z(t) = z^* + a_{k_1} t^{k_1/c} + a_{k_2} t^{k_2/c} + O(t^{k_3/c})$$
$$\zeta(t) = \frac{k_1}{c} a_{k_1} t^{k_1/c} + \frac{k_2}{c} a_{k_2} t^{k_2/c} + O(t^{k_3/c})$$

We evaluate the spatial differences for both functions between parameter values t and t_1, where $t_1 \in [0, t)$. Factoring out the leading scalar term $(t^{k_1/c} - t_1^{k_1/c})$, we get:

$$z(t) - z(t_1) = (t^{k_1/c} - t_1^{k_1/c}) \left[a_{k_1} + a_{k_2} \frac{t^{k_2/c} - t_1^{k_2/c}}{t^{k_1/c} - t_1^{k_1/c}} + \dots \right]$$
$$\zeta(t) - \zeta(t_1) = (t^{k_1/c} - t_1^{k_1/c}) \left[\frac{k_1}{c} a_{k_1} + \frac{k_2}{c} a_{k_2} \frac{t^{k_2/c} - t_1^{k_2/c}}{t^{k_1/c} - t_1^{k_1/c}} + \dots \right]$$

By the Cauchy Mean Value Theorem, the fractional ratio $\frac{t^{k_2/c} - t_1^{k_2/c}}{t^{k_1/c} - t_1^{k_1/c}}$ is strictly bounded from above by $\frac{k_2}{k_1} t^{(k_2 - k_1)/c}$. Since this upper bound depends only on t and is independent of t_1, we can bundle all higher-order terms uniformly into a single error vector bounded by $O\left(t^{(k_2-k_1)/c}\right)$.

Applying the vector norm to these expressions, we can factor out the norm of the leading vector. By the reverse triangle inequality, the norms become:

$$\|z(t) - z(t_1)\| = |t^{k_1/c} - t_1^{k_1/c}| \cdot \|a_{k_1}\| \left[1 + O\left(t^{(k_2-k_1)/c}\right)\right]$$
$$\|\zeta(t) - \zeta(t_1)\| = |t^{k_1/c} - t_1^{k_1/c}| \cdot \frac{k_1}{c} \|a_{k_1}\| \left[1 + O\left(t^{(k_2-k_1)/c}\right)\right]$$

We now evaluate the ratio of their norms. The common scalar factor $|t^{k_1/c} - t_1^{k_1/c}| \cdot \|a_{k_1}\|$ cancels out perfectly from the numerator and the denominator:

$$\frac{\|\zeta(t) - \zeta(t_1)\|}{\|z(t) - z(t_1)\|} = \frac{|t^{k_1/c} - t_1^{k_1/c}| \cdot \frac{k_1}{c} \|a_{k_1}\| \left[1 + O\left(t^{(k_2-k_1)/c}\right)\right]}{|t^{k_1/c} - t_1^{k_1/c}| \cdot \|a_{k_1}\| \left[1 + O\left(t^{(k_2-k_1)/c}\right)\right]} = \frac{k_1}{c} \left[1 + O\left(t^{(k_2-k_1)/c}\right)\right]$$

Because this remaining error bound $O\left(t^{(k_2-k_1)/c}\right)$ is completely independent of t_1, taking the supremum over all $t_1 \in [0, t)$ preserves the bound:

$$\sup_{t_1 \in [0,t)} \left| \frac{\|\zeta(t) - \zeta(t_1)\|}{\|z(t) - z(t_1)\|} - \frac{k_1}{c} \right| \le O\left(t^{(k_2-k_1)/c}\right)$$

As $t \to 0$, the error term $t^{(k_2-k_1)/c} \to 0$. Consequently, the continuous limit of the supremum is squeezed to zero:

$$\lim_{t \to 0} \sup_{t_1 \in [0,t)} \left| \frac{\|\zeta(t) - \zeta(t_1)\|}{\|z(t) - z(t_1)\|} - \frac{k_1}{c} \right| = 0$$

which concludes the proof for part (a). □

Proof of Part (b). Assuming we have successfully extracted the exact leading exponent ratio k_1/c, the target prediction $\hat{z}$ (aiming for $t = 0$) is computed via the linear ideal predictor:

$$\hat{z} = z(t) - \frac{c}{k_1} t\dot{z}(t)$$

Substituting the series expansions into this predictor equation we have

$$\begin{aligned}\hat{z} &= z^* + \sum_{j=k_1}^{\infty} a_j t^{j/c} - \frac{c}{k_1} \sum_{j=k_1}^{\infty} \frac{j}{c} a_j t^{j/c} \\ &= z^* + \left(1 - \frac{k_2}{k_1}\right) a_{k_2} t^{k_2/c} + O(t^{k_3/c})\end{aligned}$$

The leading terms $a_{k_1} t^{k_1/c}$ cancel exactly. Therefore, isolating the error gives

$$\|\hat{z} - z^*\| = O(t^{k_2/c}) \quad \text{as } t \to 0.$$

To express this error strictly in terms of the distance to the root, we invert the leading term of the path expansion. Since $\|z(t) - z^*\| = \|a_{k_1}\| t^{k_1/c} + O(t^{k_2/c})$, we can asymptotically bound the parameter t as:

$$t^{1/c} = O\left(\|z(t) - z^*\|^{1/k_1}\right)$$

Substituting this relation back into our predictor error bound produces the final geometric bound

$$\|\hat{z} - z^*\| = O\left(\left(\|z(t) - z^*\|^{1/k_1}\right)^{k_2}\right) = O\left(\|z(t) - z^*\|^{k_2/k_1}\right)$$

□

Recall that by Assumption 2 matrix $h_z(z(t), t)$ is non-singular for each $z(\cdot) \in \Pi$ in some punctured neighborhood of 0. We will need to bound the norm of $h_z^{-1}(z(t), t)$.

Lemma 16. *Consider formal series $z(\cdot) \in \Pi$ associated with integers k_1, c. There exists a constant $\delta > 0$ and a punctured neighborhood Ω of x^* such that for any $z(\cdot) \in \Pi$ and $x = (z, t) = (z(t), t) \in \Omega$ there holds $\|h_z^{-1}(x)\| \le |t|^{-\delta}$, $\|\dot{z}(t)\| = O(|t|^{k_1/c-1})$ and $\|\ddot{z}(t)\| = O(|t|^{k_1/c-2})$. Furthermore, there exists component $i \in [n]$ such that $|\dot{z}_i(t)| = \Theta(|t|^{k_1/c-1})$.*

If $\kappa = 1$ then $\delta \in (0, 1)$ and $k_1/c < 1$.

Proof. Define function $\tilde{z}(s) = z(s^c)$. This is an analytic function at 0, as it is given by a convergent power series at some neighborhood of 0. Also, $z(t) = \tilde{z}(t^{1/c})$ for some branch $t \mapsto t^{1/c}$.

Define $J(t) = h_z(z(t), t)$ and $\tilde{J}(s) = J(s^c)$. Note that the entries of matrix $\tilde{J}$ are analytic functions of s since $\tilde{J}(s) = h_z(\tilde{z}(s), s^c)$, and thus $\det \tilde{J}(s)$ is also

an analytic function of s. By Assumption 2, $\det \tilde{J}(s)$ is not identically zero, therefore $\det \tilde{J}(s) = s^d \varphi(s)$ for some integer $d \geq 1$ and analytic function $\varphi(s)$ with $\varphi(0) \neq 0$. By Cramer's rule, $\tilde{J}(s)^{-1} = \frac{\mathtt{adj}\tilde{J}(s)}{\det \tilde{J}(s)} = \frac{\mathtt{adj}\tilde{J}(s)}{\varphi(s)} \cdot s^{-d}$. Function $\frac{\mathtt{adj}\tilde{J}(s)}{\varphi(s)}$ is analytic at 0, therefore $\|\tilde{J}(s)^{-1}\| = O(|s|^{-d})$ and hence $\|h_z(z(t), t)^{-1}\| = \|\tilde{J}(t^{1/c})\| = O(|t|^{-d/c}) < |t|^{-\delta}$ for any fixed $\delta > d/c$.

By differentiating the formal series (3.2) we obtain $\dot{z}(t) = \frac{k_1}{c} a_{k_1} t^{k_1/c-1}(1 + o(1))$ and $\ddot{z}(t) = \frac{k_1}{c}(\frac{k_1}{c} - 1) a_{k_1} t^{k_1/c-2}(1 + o(1))$. This implies that $\|\dot{z}(t)\| = O(|t|^{k_1/c-1})$ and $\|\ddot{z}(t)\| = O(|t|^{k_1/c-2})$, and also $|\dot{z}_i(t)| = \Theta(|t|^{k_1/c-1})$ for all components $i \in [n]$ with $(a_{k_1})_i \neq 0$.

Let us now assume that $\kappa = 1$. Let $J^* = U^* \Sigma^* (V^*)^\dagger$ and $h_z = U \Sigma V^\dagger$ be SVDs of $J^* = h_z(x^*)$ and $h_z(x)$ respectively, with $\Sigma^* = \mathtt{diag}(\sigma_1^*, \ldots, \sigma_n^*)$, $\sigma_1^* \geq \ldots \geq \sigma_{n-1}^* > \sigma_n^* = 0$, $\Sigma = \mathtt{diag}(\sigma_1, \ldots, \sigma_n)$, $\sigma_1 \geq \ldots \geq \sigma_n \geq 0$. Let $\{u_i^*\}, \{v_i^*\}, \{u_i\}, \{v_i\}$ be the columns of U^*, V^*, U, V respectively. Vector u_n^* is the left singular vector of J^* for value $\sigma_n^* = 0$ (i.e. $(u_n^*)^\dagger J^* = 0$); by Assumption 2(d), we have $(u_n^*)^\dagger h_t^* \neq 0$.

It follows from Wedin's theorem (Stewart and Sun 1990, Theorem 4.1) that σ_n, u_n, v_n depend continuously on matrix h_z (as long as singular value σ_n has multiplicity 1). Therefore, there exists a neighborhood of x^* in which points $x \in \mathcal{V}$ satisfy $\sigma_1 \geq \ldots \geq \sigma_{n-1} \geq \Theta(1)$, $\|h_t\| = \Theta(1)$ and $|u_n^\dagger h_t| = \Theta(1)$. For points $x = (z, t) \neq x^*$ in this neighborhood we have

$$\begin{aligned} \dot{z} &= -h_z^{-1} h_t = -(V \Sigma^{-1} U^\dagger) h_t \\ &= -\sum_{i=1}^{n} \sigma_i^{-1} v_i u_i^\dagger h_t \\ &= \left(-\sum_{i=1}^{n-1} \sigma_i^{-1} v_i u_i^\dagger h_t \right) - \left(\sigma_n^{-1} v_n u_n^\dagger h_t \right) \end{aligned} \tag{3.4}$$

The norm of the first term in (3.4) is bounded by a constant in a neighborhood of 0, while the norm of the second term goes to infinity as $t \to 0$ (since σ_n goes to zero, $\|v_n\| = 1$ and $|u_n^\dagger h_t| = \Theta(1)$). This implies that $\lim_{t \to 0} \|\dot{z}\| = +\infty$. Since $\|\dot{z}(t)\| = \Theta(|t|^{k_1/c-1})$, we must have $k_1/c < 1$.

From (3.4) we get

$$\sigma_n^{-1} v_n u_n^\dagger h_t = -\dot{z} - \sum_{i=1}^{n-1} \sigma_i^{-1} v_i u_i^\dagger h_t$$

Taking norms gives

$$\sigma_n^{-1} \cdot \|v_n\| \cdot |u_n^\dagger h_t| \leq \|\dot{z}\| + \sum_{i=1}^{n-1} \sigma_i^{-1} \|v_i\| \cdot \|u_i\| \cdot \|h_t\|$$

We have $\|v_i\| = \|u_i\| = 1$, and so $\|h_z^{-1}\| = \sigma_n^{-1} \leq \frac{1}{\Theta(1)}(O(|t|^{k_1/c-1}) + O(1)) = O(|t|^{k_1/c-1})$. Thus, any constant $\delta > 1 - k_1/c$ will satisfy the claim of the lemma, so we can indeed choose $\delta < 1$. □

Let us fix constants $\delta > 0, \Delta < 1, \Lambda > -1$ such that for each formal series $z(\cdot) \in \Pi$ we have $\delta > \delta^{z(\cdot)}$, $\Delta > 1 - \frac{k_1^{z(\cdot)}}{c^{z(\cdot)}}$, $\Lambda > \frac{k_1^{z(\cdot)}}{c^{z(\cdot)}} - 1$. (Here the superscript $z(\cdot)$ denotes the value associated with formal series $z(\cdot)$, and value $\delta^{z(\cdot)}$ comes from Lemma 16.) By the lemma, the following holds for all $x = (z,t) \in \mathcal{V}$ in some punctured neighborhood of x^*:

$$\|h_z^{-1}(x)\| \leq |t|^{-\delta} \tag{3.5a}$$
$$\|\dot{z}\| \leq |t|^{-\Delta} \tag{3.5b}$$
$$\|\ddot{z}\| \leq |t|^{-\Delta-1} \tag{3.5c}$$

Note, if $\kappa = 1$ then we can have $\delta < 1$ and $\Lambda < 0$.

We define $\gamma_{\min} = \delta + \frac{\Delta-1}{2}$. Note, if $\kappa = 1$ then $\gamma_{\min} < 1$. We thus assume from now on that

$$\gamma > \delta + \tfrac{\Delta-1}{2} \tag{3.6}$$

Lemma 17. *For any constant $\alpha > 0$ there exists another constant $\beta > 0$ and neighborhood Ω of x^* satisfying the following: if $x = (z,t) \in \Omega$ and $\|h(x)\| \leq |t|^\beta$ then there exists Puiseux series $\bar{z}(\cdot) \in \Pi$ such that $\|z - \bar{z}\| \leq |t|^\alpha$ and $\|\dot{z} - \dot{\bar{z}}\| \leq |t|^{\alpha-2\delta}$ where $\bar{z} = \bar{z}(t)$.*

Proof. We can assume w.l.o.g. that $\alpha \geq \delta$ and $\alpha + \delta > 1$ (by increasing α, if necessary; this will not affect the claim). By the classical Łojasiewicz inequality (Łojasiewicz 1959), there exist constants $C > 0$, $\theta > 0$ such that every x in some neighborhood of x^* satisfies

$$\texttt{dist}(x, \mathcal{V}) \leq C\|h(x)\|^\theta \tag{3.7}$$

We will show the lemma for any value β satisfying

$$\beta > \tfrac{\alpha+\delta}{\theta} \tag{3.8}$$

Consider $x = (z,t)$ with $\|h(x)\| \leq |t|^\beta$ in some neighborhood of x^*. By (3.7), there exists $x' = (z',t') \in \mathcal{V}$ with $\|x - x'\| \leq C\|h(x)\|^\theta \leq C(|t|^\beta)^\theta = C|t|^{\bar{\alpha}+\delta}$ where $\bar{\alpha} > \alpha$ is a constant. By shrinking the neighborhood if necessary, we can assume that $C|t|^{\bar{\alpha}+\delta} < \frac{1}{2}|t|$. We have $z' = \bar{z}(t')$ for some Puiseux series $\bar{z}(\cdot) \in \Pi$. By the mean value theorem, $\|\bar{z}(t) - \bar{z}(t')\| \leq \|\dot{\bar{z}}(\tau)\| \cdot |t - t'|$ for some $\tau \in [t, t']$. Since $|t - t'| \leq \|x - x'\| \leq C|t|^{\bar{\alpha}+1} < \frac{1}{2}|t|$, we must have $|\tau| = \Theta(|t|)$. Denoting $\bar{z} = \bar{z}(t)$, we get

$$\|z - \bar{z}\| \leq \|z - z'\| + \|z' - \bar{z}\| \leq \|x - x'\| + \|\dot{\bar{z}}(\tau)\| \cdot |t - t'| \leq \|x - x'\| \cdot (1 + \|\dot{\bar{z}}(\tau)\|)$$

since $\max\{\|z - z'\|, |t - t'|\} \leq \|x - x'\|$. Using eq. (3.5a), we get $\|\dot{\bar{z}}(\tau)\| = \|h_z(\bar{z}(\tau),\tau)^{-1} h_t(\bar{z}(\tau),\tau)\| \leq \|h_z(\bar{z}(\tau),\tau)^{-1}\| \cdot \|h_t(\bar{z}(\tau),\tau)\| \leq |\tau|^{-\delta} \cdot O(1) = O(|t|^{-\delta})$. This yields

$$\|z - \bar{z}\| \leq C|t|^{\bar{\alpha}+\delta} \cdot (1 + O(|t|^{-\delta})) = O(|t|^{\bar{\alpha}})$$

Since $\bar{\alpha} > \alpha$, taking a sufficiently small neighborhood will ensure that the last expression is at most $|t|^\alpha$.

To prove the bound on $\|\dot{z} - \dot{\bar{z}}\|$, we will use the following fact:

- *Suppose that $A, B \in \mathbb{C}^{n\times n}$, $a, b \in \mathbb{C}^{n\times 1}$, A is invertible and $\|A^{-1}\|\,\|A-B\| \le 1$. Then*

$$\|A^{-1}a - B^{-1}b\| \le \|A^{-1}\|\,\|a-b\| + \frac{\|A^{-1}\|^2\,\|A-B\|}{1-\|A^{-1}\|\,\|A-B\|}\|b\| \tag{3.9}$$

Indeed, the assumption implies that B is invertible and $\|B^{-1}\| \le \frac{\|A^{-1}\|}{1-\|A^{-1}\|\,\|A-B\|}$. We have $A^{-1}a - B^{-1}b = A^{-1}(a-b) + A^{-1}(B-A)B^{-1}b$ and hence $\|A^{-1}a - B^{-1}b\| \le \|A^{-1}\|\,\|a-b\| + \|A^{-1}\|\,\|B-A\|\,\|B^{-1}\|\,\|b\|$, which yields (3.9).

Let us plug $A = h_z(\bar z, t)$, $a = h_t(\bar z, t)$, $B = h_z(z,t)$, $b = h_t(z,t)$. Since h is analytic, we have $\|A-B\| \le O(\|z-\bar z\|) \le O(|t|^{\bar\alpha})$, $\|a-b\| \le O(\|z-\bar z\|) \le O(|t|^{\bar\alpha})$ and $\|b\| \le O(1)$. By eq. (3.5a), $\|A^{-1}\| \le |t|^{-\delta}$. Plugging this into (3.9) gives

$$\|\dot z - \dot{\bar z}\| = \|A^{-1}a - B^{-1}b\| \le |t|^{-\delta}\cdot O(|t|^{\bar\alpha}) + \frac{|t|^{-2\delta}\cdot O(|t|^{\bar\alpha})}{1-|t|^{-\delta}\cdot O(|t|^{\bar\alpha})}\cdot O(1) \le O(|t|^{\bar\alpha-2\delta})$$

since $\bar\alpha > \delta$. Since $\bar\alpha > \alpha$, taking a sufficiently small neighborhood will ensure that the last expression is at most $|t|^{\alpha-2\delta}$. □

We now proceed with the proof of Theorem 14. Fix α that satisfies

$$\alpha \;>\; \max\Big\{\; \Lambda+\gamma+\max\{2\delta,1\} \;,\; \tfrac{k_2}{c} \;,\; 2\delta - 1 - \tfrac{k_2}{c} \;\Big\} \tag{3.10}$$

for coefficients c, k_1, k_2 of all Puiseux series $z(\cdot) \in \Pi$. Let $\beta > 0$ be the constant specified in Lemma 17 for this value of α. Note that this value can be chosen so that $\beta = \Theta(\gamma)$ (see eq. (3.8)). Assume that the input point $x = (z,t)$ in the appropriate neighborhood of x^* satisfies $\|h(x)\| \le |t|^{\beta}$. By Lemma 17, there exists Puiseux series $\bar z(\cdot) \in \Pi$ such that $\bar z = \bar z(t)$ satisfies

$$\|z - \bar z\| \;\le\; |t|^{\alpha} \tag{3.11a}$$
$$\|\dot z - \dot{\bar z}\| \;\le\; |t|^{\alpha-2\delta} \tag{3.11b}$$

Let us denote $\bar z_1 = \bar z(t_1)$. Our next goal will be to show that point z_1 constructed by the algorithm is close to $\bar z_1$, and gives the correct value of the ratio k_1/c. Recall that $\bar z_1$ is obtained via the Newton's method for system $F(z) = 0$ where $F(z) = h(z, t_1)$.

Lemma 18. (a) *The starting point $z_0 = z + \dot z(t_1 - t)$ satisfies $\|z_0 - \bar z_1\| \le O(|t|^{2\gamma-\Delta+1})$.*
(b) *The preconditions of Theorem 13 hold with $L = \Theta(|t|^{-\delta})$, $C = \frac{4}{9L}$.*
(c) *Newton's method terminates after $O(\log(1+\beta))$ iterations.*
(d) *It produces point z_1 satisfying $\|z_1 - \bar z_1\| \le |t_1|^{\alpha}$ and $\|\dot z_1 - \dot{\bar z}_1\| \le |t_1|^{\alpha-2\delta}$.*
(e) *There exists a neighborhood Ω of x^* with the following property: if $x \in \Omega$ then the method produces the correct ratio k_1/c for the Puiseux series $\bar z(\cdot)$.*

Proof. **Part (a).** First, we will show the claim assuming that $n = 1$. We can write

$$\bar{z}_1 = \bar{z} + (t_1 - t)\dot{\bar{z}} + \tfrac{1}{2}(t_1 - t)^2 \ddot{\bar{z}}(\tau) \qquad \Rightarrow \qquad z_0 - \bar{z}_1 = (t_1 - t)(\dot{z} - \dot{\bar{z}}) - \tfrac{1}{2}(t_1 - t)^2 \ddot{\bar{z}}(\tau)$$

for some $\tau \in [t_1, t]$. We have $|t_1 - t| = |t|^{\gamma+1}$ and $|\ddot{\bar{z}}(\tau)| \le |\tau|^{-\Delta-1}$, and so

$$|z_0 - \bar{z}_1| \le |t|^{\gamma+1} \cdot |t|^{\alpha-2\delta} + \tfrac{1}{2}|t|^{2(\gamma+1)} \cdot |\tau|^{-\Delta-1} = O(|t|^{2\gamma-\Delta+1})$$

since $\gamma + 1 + \alpha - 2\delta \ge 2\gamma - \Delta + 1$. If $n > 1$ then $|(z_0 - \bar{z}_1)_i| \le O(|t|^{2\gamma-\Delta+1})$ for each coordinate $i \in [n]$ (by the argument above applied to the i-th coordinate of vectors), and hence $\|z_0 - \bar{z}_1\| \le O(|t|^{2\gamma-\Delta+1})$.

Part (b). Since $F_z(z) = h_z(z, t_1)$ is analytic in a neighborhood of z^*, we have $\|F_z(z') - F_z(z'')\| \le O(\|z' - z''\|)$ when z', z'' are in a certain neighborhood of z^*. In particular, we have $\|F_z(z_0) - F_z(\bar{z}_1)\| \le O(\|z_0 - \bar{z}_1\|) \le O(|t|^{2\gamma-\Delta+1})$. Also, $\|F_z^{-1}(\bar{z}_1)\| \le O(|t_1|^{-\delta}) = O(|t|^{-\delta})$ by eq. (3.5a). Therefore, $\|F_z(z_0) - F_z(\bar{z}_1)\| \, \|F_z^{-1}(\bar{z}_1)\| \le \frac{1}{2}$ if $|t|$ is sufficiently small (since $2\gamma - \Delta + 1 - \delta > 0$ by eq. (3.6)). This implies that

$$\|F_z^{-1}(z_0)\| \le \frac{\|F_z^{-1}(\bar{z}_1)\|}{1 - \|F_z(z_0) - F_z(\bar{z}_1)\| \, \|F_z^{-1}(\bar{z}_1)\|} \le O(|t|^{-\delta})$$

We conclude that $\|F_z^{-1}(z_0)(F_z(z') - F_z(z''))\| \le O(|t|^{-\delta}) \cdot O(\|z' - z''\|)$ when z', z'' are in a certain neighborhood of z^*, and hence the first two preconditions of Theorem 13 hold with $L = \Theta(|t|^{-\delta})$. We have $\|F(z_0)\| = \|F(z_0) - F(\bar{z}_1)\| \le O(\|z_0 - \bar{z}_1\|) \le O(|t|^{2\gamma-\Delta+1})$, and so the third precondition holds with $C \ge C_0 = O(|t|^{-\delta}) \cdot O(|t|^{2\gamma-\Delta+1})$. We obtain $C_0 L = O(|t|^{2\gamma-2\delta-\Delta+1}) = O(|t|^{\varepsilon})$ where $\varepsilon > 0$ by the choice of γ. Therefore, we can indeed set $C = \frac{4}{9L}$ if $|t|$ is sufficiently small.

Parts (c,d). Theorem 13 yields that equation $F(z) = 0$ has a unique solution in the ball $B(z_0, r)$ for any $r \in [r_-, r_+] = [\frac{2}{3L}, \frac{4}{3L}]$. We have $F(\bar{z}_1) = 0$ and $\|z_0 - \bar{z}_1\| \le O(|t|^{2\gamma-\Delta+1}) < r_-$ if $|t|$ is sufficiently small (since $2\gamma - \Delta + 1 > \delta$ by eq. (3.6)), so this unique solution must be $\bar{z}_1$.

Let us denote $w_k = z^{(k)} - \bar{z}_1$ where $z^{(k)}$ is the iterate at step k (with $z^{(0)} = z_0$). Theorem 13 gives that $\|w_{k+1}\| \le \bar{C} \cdot |t|^{-\delta} \|w_k\|^2$. We have $\|w_0\| \le O(|t|^{\lambda})$ where $\lambda = 2\gamma - \Delta + 1 > 2\delta$, therefore $\|w_k\| \le |t|^{O(\delta 2^k)}$. Since h is an analytic function, we have $\|h(z^{(k)}, t_1)\| = \|h(z^{(k)}, t_1) - h(\bar{z}_1, t_1)\| \le O(\|z^{(k)} - \bar{z}_1\|) \le |t|^{O(\delta 2^k)}$. We conclude that for any fixed β we will have $\|h(z^{(k)}, t_1)\| \le |t_1|^{\beta}$ after $k = O(\log(1 + \beta))$ iterations.

By Lemma 17, there exists $\bar{\bar{z}}(\cdot) \in \Pi$ with $\|z_1 - \bar{\bar{z}}_1\| \le |t_1|^{\alpha}$ and $\|\dot{z}_1 - \dot{\bar{\bar{z}}}_1\| \le |t_1|^{\alpha-2\delta}$ where $\bar{\bar{z}} = \bar{\bar{z}}(t_1)$. Since $\alpha > \delta$, we can assume that $|t_1|^{\alpha} < r_- = \Theta(|t|^{\delta})$. This implies that $\|z_0 - \bar{\bar{z}}_1\| \le \|z_0 - z_1\| + \|z_1 - \bar{\bar{z}}_1\| < r_- + r_- = r_+$. We have $F(\bar{z}_1) = F(\bar{\bar{z}}_1) = 0$ and $\bar{z}_1, \bar{\bar{z}}_1 \in B(z_0, r_+)$, therefore $\bar{z}_1 = \bar{\bar{z}}_1$ and hence $\dot{\bar{z}}_1 = \dot{\bar{\bar{z}}}_1$.

Part (e). Denote $A = \|t\dot{z} - t_1\dot{z}_1\|$, $B = \|z - z_1\|$ and $R = \frac{k_1}{c}$. Clearly, there exists constant $\varepsilon \in (0, \frac{1}{2}R)$ such that R is the only rational number p/q with

integers $p \in [1, k_1^{\max}]$, $q \geq 1$ satisfying $|R - p/q| < \varepsilon$. We will show that $|A/B - R| < \varepsilon$ when x is in some neighborhood of x^*; this will prove the claim. Denote $\bar{A} = \|t\dot{\bar{z}} - t_1\dot{\bar{z}}_1\|$ and $\bar{B} = \|\bar{z} - \bar{z}_1\|$. By Lemma 15 we can choose a neighborhood such that $|\bar{A}/\bar{B} - R| < \frac{1}{2}\varepsilon$. It now suffices to show that $|A/B - \bar{A}/\bar{B}| < \frac{1}{2}\varepsilon$ in some neighborhood of x^*.

By Lemma 16 there exists $i \in [n]$ such that $|\dot{\bar{z}}_i(\tau)| = \Theta(|\tau|^{k_1/c-1}) \geq \Theta(|\tau|^{\Lambda})$. By the mean value theorem, $\bar{B} \geq |\bar{z}_i(t) - \bar{z}_i(t_1)| = |\dot{\bar{z}}_i(\tau)| \cdot |t - t_1|$ for some $\tau \in [t, t_1]$. Therefore, $\bar{B} \geq \Theta(|t|^{\Lambda}) \cdot |t|^{\gamma+1}$. We have $|B - \bar{B}| \leq \|z - \bar{z}\| + \|z_1 - \bar{z}_1\| \leq |t|^{\alpha} + |t_1|^{\alpha} = O(|t|^{\alpha}) = o(\bar{B})$ since $\alpha > \Lambda + \gamma + 1$. Similarly, $\bar{A} = R\bar{B} = \Theta(|t|^{\Lambda+\gamma+1})$ and $|A - \bar{A}| \leq t\|\dot{z} - \dot{\bar{z}}\| + t_1\|\dot{z}_1 - \dot{\bar{z}}_1\| \leq |t|^{1+\alpha-2\delta} + |t_1|^{1+\alpha-2\delta} = O(|t|^{1+\alpha-2\delta}) = o(\bar{A})$ since $1 + \alpha - 2\delta > \Lambda + \gamma + 1$.

We showed that $A/B = (\bar{A}(1 + o(1))/(\bar{B}(1 + o(1)) = (\bar{A}/\bar{B})(1 + o(1))$. The claim follows.

□

We are now ready to prove Theorem 14. We have $\bar{z} - z^* = a_{k_1} t^{k_1/c}(1 + o(1))$, and hence $\|\bar{z} - z^*\| = \Theta(|t|^{k_1/c})$. Since $\|z - \bar{z}\| = O(|t|^{\alpha})$ and $\alpha > k_1/c$, we also have $\|z - z^*\| = \Theta(|t|^{k_1/c})$. By Lemma 15, the "ideal predictor" $\hat{\bar{z}} = \bar{z} - \frac{c}{k_1}\dot{\bar{z}}t$ satisfies $\|\hat{\bar{z}} - z^*\| = O(\|\bar{z} - z^*\|^{k_2/k_1}) = O(|t|^{k_2/c})$.

By the previous lemma, we can assume that the ratio k_1/c produced in step 3 is the correct ratio for the Puiseux series $\bar{z}(\cdot)$. Recall that our predictor is given by $\hat{z} = z - \frac{c}{k_1}\dot{z}t$. We then have $\hat{\bar{z}} = \bar{z} - \frac{c}{k_1}\dot{\bar{z}}t$, and so

$$\|\hat{z} - \hat{\bar{z}}\| \leq \|z - \bar{z}\| + \tfrac{c}{k_1}\|\dot{z} - \dot{\bar{z}}\|\, t \leq |t|^{\alpha} + \tfrac{c}{k_1}|t|^{\alpha-2\delta}\,|t|$$

$$\|\hat{z} - z^*\| \leq \|\hat{\bar{z}} - z^*\| + \|\hat{z} - \hat{\bar{z}}\| \leq O(|t|^{k_2/c}) + O(|t|^{\alpha}) + O(|t|^{\alpha-2\delta+1}) = O(|t|^{k_2/c})$$

since $\alpha > k_2/c$ and $\alpha - 2\delta + 1 > k_2/c$ by the choice of α in eq. (3.10). The RHS of the last expression is at most $O(\|z - z^*\|^{k_2/k_1})$.

3.4 Corrector

Let us now assume that we have initial point $x_\circ = (z_\circ, t_\circ)$ and predictor $\hat{x} = (\hat{z}, \hat{t})$ where $\hat{z}$ approximates $z(\hat{t})$ for some Puiseux series $z(\cdot) \in \Pi$. The predictor step moved us away from the homotopy $h(\cdot)$; the goal of the corrector is go back to this homotopy.

We will consider separately cases $\kappa = 1$ and $\kappa \geq 2$. We will use $\hat{t} = 0$ in the former case and $\hat{t} \neq 0$ in the latter.

3.4.1 Corank $\kappa = 1$: pseudo-arc length corrector

Recall that in the classical approach we are effectively solving the system

$$\begin{cases} h(x) & = 0 \\ t - \hat{t} & = 0 \end{cases} \tag{3.12}$$

over variables $x = (z, t)$. Its Jacobian is

$$\begin{pmatrix} h_z & h_t \\ 0 & 1 \end{pmatrix} \tag{3.13}$$

Note that if $\hat{t} = 0$ then x^* is a root of (3.12), and the Jacobian is singular at this root (since the columns of h_z are linearly dependent). This fact prevents us from setting $\hat{t} = 0$, since then the Newton's method may not converge.

We propose to do the following instead. Below β is the parameter used in Theorem 14.

1. Set $q = \frac{[\dot{z}_\circ^\dagger \ 1]}{\|[\dot{z}_\circ^\dagger \ 1]\|}$. Observe that the vector $q^\dagger$ is in the null space of $h_x(x_\circ)$, since $h_x(x_\circ) \cdot \begin{bmatrix} \dot{z}_\circ \\ 1 \end{bmatrix} = [h_z(x_\circ) \quad h_t(x_\circ)] \cdot \begin{bmatrix} -h_z^{-1}(x_\circ) h_t(x_\circ) \\ 1 \end{bmatrix} = -h_t(x_\circ) + h_t(x_\circ) = 0$.
2. Replace system (3.12) with

$$h[\hat{x}](x) \stackrel{\text{def}}{=} \begin{cases} h(x) \\ q \cdot (x - \hat{x}) \end{cases} = 0 \tag{3.14}$$

3. Apply Newton's method to solve equation $h[\hat{x}](x) = 0$ starting with $x_0 = \hat{x}$, generating a sequence of points $x_1, x_2, \ldots$. Stop once we get a point $x = x_K = (z, t)$ with $\|h(x)\| \le |t|^\beta$.

Lemma 19. *There exists a neighborhood Ω of x^* and constant $\lambda > 0$ such that $\|q\| = 1$ and $\|(D_x h[\hat{x}](x))^{-1}\| \le \lambda$ for any $x_\circ, x \in \Omega$.*

Proof. The Jacobian of $h[\hat{x}](\cdot)$ is given by

$$D_x h[\hat{x}] = \begin{pmatrix} h_z & h_t \\ q \end{pmatrix} = \begin{pmatrix} h_z(\hat{x}) & h_t(\hat{x}) \\ q(x_\circ) \end{pmatrix} \tag{3.15}$$

If $\hat{x} = x_\circ = x^*$ then matrix $D_x h[x^*] = \begin{pmatrix} h_z^* & h_t^* \\ q^* \end{pmatrix}$ is non-singular by Assumption 2, and matrix $[h_z^* \ \ h_t^*]$ has full rank. Singular vectors of a matrix corresponding to singular values of multiplicity 1 depend continuously on the matrix (by the Wedin's theorem which we used in the proof of Lemma 16). Therefore, matrix $\begin{pmatrix} h_z & h_t \\ q \end{pmatrix}$ depends continuously on $(\hat{x}, x_\circ)$ (since q is the singular vector of $h_x = [h_z \ \ h_t]$ corresponding to singular value 0 of multiplicity 1). The claim follows.

□

Note that the guarantee of Lemma 19 can be achieved by many other choices of q, e.g. if q is chosen randomly. For the result below we assume that vector q is chosen to satisfy the properties in Lemma 19 but is not necessarily in the null space of $h_x(x_\circ)$.

Theorem 20. *Let Ω be the neighborhoood of x^* from Lemma 19 (with constant $\lambda \geq 0$). There exists neighborhoods $\Omega^- \subseteq \Omega^+ \subseteq \Omega$ of x^* and constant $\beta_{\min} > 0$ with the following property: if $\hat{x} \in \Omega^-$ then equation $h[\hat{x}](x) = 0$ has a unique solution $\bar{x} = (\bar{z}, \bar{t}) \in \Omega^+$, and it satisfies $\|\bar{x} - x^*\| \leq \lambda \|\hat{x} - x^*\|$. Furthermore, the sequence of points $x_0 = \hat{x}, x_1, x_2, \ldots$ generated by the Newton's method satisfies the following:*
(a) *$\|x_k - x^*\| \leq O(\|\hat{x} - x^*\|)$ for each $k \geq 1$.*
(b) *If $\beta > \beta_{\min}$, $k \geq \ell \stackrel{\text{def}}{=} \lceil 2\log_2(1+\beta) \rceil$, $x_k = x = (z,t)$ and $\|h(x)\| > |t|^\beta$ then $\|x - x^*\| \leq \|\hat{x} - x^*\|^{O(2^{k-\ell})}$.*

Proof. We will apply Theorem 13 for function $F(x) = h[\hat{x}](x)$. By assumption, we have $\|F_x^{-1}(x)\| \leq \lambda$ for all $x \in \Omega$. For any $x, y \in \Omega$ we have

$$\|F'(x) - F'(y)\| = \left\| \begin{pmatrix} h_x(x) & h_x(y) \\ 0 & \end{pmatrix} \right\| = O(\|x - y\|)$$

since $h_x = [h_z \ h_t]$ is analytic in Ω. Thus, the first two preconditions of Theorem 13 hold if $L \geq L_0$ for some constant $L_0 > 0$. We have $\|F(\hat{x})\| = \left\| \begin{pmatrix} h(\hat{x}) \\ 0 \end{pmatrix} \right\| = \|h(\hat{x})\| = \|h(\hat{x}) - h(x^*)\| \leq C_0 \cdot \|\hat{x} - x^*\|$ for some constant C_0, when $\hat{x} \in \Omega$ (since h is analytic on Ω). Thus, the third precondition holds with any $C \geq \lambda C_0$. Let us choose value $L \geq \max\{L_0, \frac{9}{4}\lambda C_0\}$ so that $B(x^*, \frac{4}{3L}) \subseteq \Omega$, and set $C = 4L/9$. These values satisfy conditions of the theorem, and hence equation $F(x) = 0$ has a unique solution $\bar{x} \in B(\hat{x}, r)$ for any $r \in [r_-, r_+] = [\frac{2}{3L}, \frac{4}{3L}]$. Let us denote it as $\varphi(\hat{x})$. Define $\Omega^- = B(x^*, r^-)$ and $\Omega^+ = B(x^*, r^+)$, then for each $\hat{x} \in \Omega^-$ we have $\|\varphi(\hat{x}) - x^*\| \leq \|\varphi(\hat{x}) - \hat{x}\| + \|\hat{x} - x^*\| \leq r_- + r_- = r_+$ and hence $\varphi(x) \in \Omega^+$.

Function $\varphi(\cdot)$ must be continuous at each $\hat{x} \in \Omega^-$. Indeed, if x is an accumulation point of $\varphi(u)$ as $u \to \hat{x}$ then $h[\hat{x}](x) = 0$ by continuity, and thus is uniquely determined by $\hat{x}$ since $h[\hat{x}](x) = 0$ has a unique solution in $B(\hat{x}, r^+)$. The uniqueness of the accumulation point implies the claim.

Differentiating the equation $h[\hat{x}](\varphi(\hat{x})) = 0$ with respect to $\hat{x}$ gives

$$\begin{aligned} D_{\hat{x}}\varphi(\hat{x}) &= -(D_x h[\hat{x}](x))^{-1} D_{\hat{x}} h[\hat{x}](x) \\ &= -(D_x h[\hat{x}](x))^{-1} \begin{pmatrix} 0 \\ -q \end{pmatrix}, \end{aligned}$$

and hence

$$\|D_u \varphi(u)\| \leq \|(D_x h[\hat{x}](x))^{-1}\| \cdot \left\| \begin{pmatrix} 0 \\ -q \end{pmatrix} \right\| \leq \lambda \cdot 1.$$

This implies that $\|\bar{x} - x^*\| = \|\varphi(\hat{x}) - \varphi(x^*)\| \leq \lambda \cdot \|\hat{x} - x^*\|$.

Next, we show properties (a)- (b) of the sequence $\{x_k\}_{k=0,1,\ldots}$ generated by the Newton's method. We will denote $\Delta = \|\hat{x} - x^*\|$. Theorem 13 gives $\|x_{k+1} - \bar{x}\| \leq \frac{3L}{2}\|x_k - \bar{x}\|^2$, with $\|x_0 - \bar{x}\| \leq \lambda\Delta$. By shrinking Ω^-, if necessary, we can make sure for some constant $\alpha > 0$ we have

$$\|x_k - \bar{x}\| \leq \min\{\ \Delta^{\alpha 2^k}\ ,\ \|\hat{x} - \bar{x}\|\ \} \qquad \forall k \geq 1$$

Property (a). For any $k \geq 1$ we can write $\|x_k - x^*\| \leq \|x_k - \bar{x}\| + \|\hat{x} - \bar{x}\| \leq 2\lambda\|\hat{x} - x^*\|$.

Property (b). Suppose that $\|x_k - x^*\| > 2\Delta^{\alpha 2^{k-\ell}}$ for $k \geq \ell$. Then $\|\bar{x} - x^*\| \geq \|x_k - x^*\| - \|x_k - \bar{x}\| > 2\Delta^{\alpha 2^{k-\ell}} - \Delta^{\alpha 2^k} \geq \Delta^{\alpha 2^{k-\ell}}$. We have $h(\bar{x}) = 0$, and hence $\bar{z} = \bar{z}(\bar{t})$ for some Puiseux series $\bar{z}(\cdot) \in \Pi$. This implies that $\|\bar{z} - z^*\| \leq O(|\bar{t}|^{k_1/c})$ where k_1, c are the coefficients for $\bar{z}(\cdot)$. We can thus write

$$\Delta^{\alpha 2^{k-\ell}} \leq \|\bar{x} - x^*\| \leq \|\bar{z} - z^*\| + |\bar{t} - 0| < O(|\bar{t}|^{k_1/c}) + |\bar{t}| < |\bar{t}|^{\delta}$$

for some constant $\delta > 0$. This implies that $\|x_k - \bar{x}\| \leq \Delta^{\alpha 2^k} \leq |\bar{t}|^{\delta 2^\ell}$. Since $\ell = \lceil 2\log_2(1+\beta) \rceil \geq \log_2 \beta_{\min} + \log_2 \beta$, we have $\|x_k - \bar{x}\| \leq |t|^{\delta \beta_{\min} \beta}$. By choosing $\beta_{\min}$ sufficiently large, we can ensure that $\|x_k - \bar{x}\| \leq |\frac{1}{2}\bar{t}|^{\beta+\varepsilon}$ for some constant $\varepsilon > 0$ and $\|x_k - \bar{x}\| \leq |\frac{1}{2}\bar{t}|$. The latter condition implies that point $x_k = x = (z, t)$ satisfies $|t - \bar{t}| \leq |\frac{1}{2}\bar{t}|$ and hence $|t| \geq |\frac{1}{2}\bar{t}|$. This yields $\|x_k - \bar{x}\| \leq |\frac{1}{2}\bar{t}|^{\beta+\varepsilon} \leq |t|^{\beta+\varepsilon}$. It remains to observe that $\|h(x_k)\| = \|h(x_k) - h(\bar{x})\| \leq O(\|x_k - \bar{x}\|) \leq O(|t|^{\beta+\varepsilon}) < |t|^{\beta}$ if neighborhood Ω is sufficiently small.

□

We can finally prove Theorem 11. We use the following algorithm. Given point $x_\circ$, we compute predictor $\hat{x}$ as described in Section 3.3, then construct system (3.14) and run Newton's method, obtaining sequence $x_0 = \hat{x}, x_1, x_2, \ldots$. We stop when once we get a point $x_k = (x, t)$ with $\|h(x)\| \leq |t|^{\beta}$. If $k < \ell = \lceil 2\log_2(1+\beta) \rceil$ then we return x_k, otherwise we return the sequence $(x_\ell, x_{\ell+1}, \ldots, x_k)$. By combining Theorems 14 and 20 we conclude that this algorithm has the properties stated in Theorem 11.

Connection to the pseudo-arclength method. The corrector described above can be related to the pseudo-arclength method (Wempner 1971; Riks 1972; Keller 1977). The latter constructs the following system over variables $x = (z, t)$:

$$\begin{cases} h(x) \\ \frac{dz}{ds}(z - z_\circ) + \frac{dt}{ds}(t - t_\circ) - \Delta s \end{cases} = 0 \tag{3.16}$$

where parameter s represents the Euclidean length along the curve $(z(t), t)$, and quantities $\frac{dz}{ds}$, $\frac{dt}{ds}$, Δs are fixed. The first two quantities are computed from equations $\frac{dz}{ds} = \dot{z} \cdot \frac{dt}{ds}$ and $\left\|\frac{dz}{ds}\right\|^2 + \left|\frac{dt}{ds}\right|^2 = 1$. Thus, the last equation in. (3.16) can also be equivalently written as $q \cdot (x - \hat{x}) = 0$ for some $\hat{x} \in \mathbb{C}^{n+1}$, where $q = [\dot{z}_\circ \; 1]$, as in (3.14). The difference is that we set $\hat{x}$ explicitly via an endgame-aware linear predictor, while pseudo-arclength methods control parameter Δs instead.

3.4.2 Corank $\kappa \geq 2$: lifted pseudo-arc length corrector

In this case we will introduce $\kappa - 1$ new auxiliary variables $\xi \in \mathbb{C}^{\kappa-1}$. Let us denote $y = (x, \xi) = (z, t, \xi)$, $\hat{y} = (\hat{x}, 0) = (\hat{z}, \hat{t}, 0)$. We will solve the system

$$h[\hat{x}](y) \stackrel{\text{def}}{=} \begin{cases} h(x) + P \cdot \xi \\ Q \cdot (y - \hat{y}) \end{cases} = 0 \tag{3.17}$$

where $P \in \mathbb{C}^{n\times(\kappa-1)}$, $Q \in \mathbb{C}^{\kappa\times(n+\kappa)}$ are matrices computed from $(\hat{x}, x_\circ)$. The Jacobian of $h[\hat{x}]$ is

$$D_y h[\hat{x}] = \begin{pmatrix} h_z & h_t & P \\ & Q & \end{pmatrix} \tag{3.18}$$

By assumption, we have $\mathtt{rank}([h_z^* \;\; h_t^*]) = n-\kappa+1$. This means that we can find matrices P, Q such that $\|(D_y h[\hat{x}](y))^{-1}\| \le O(1)$ assuming that $(\hat{x}, y)$ lies in a certain neighborhood of (x^*, y^*). Using the same arguments as in the previous section, one can then show that system (3.17) has a unique solution $\bar{y}$ in a neighborhood of y^*, this solution satisfies $\|\bar{y} - y^*\| \le O(\|\hat{x} - x^*\|)$, and it can be efficiently computed with any desired accuracy using the Newton's method. Unfortunately, this does not lead to any guarantees on the convergence rate, so we leave this claim without proof.

We denote $y = (v, t, \xi)$ to be the output of this process. Experimentally, we observed that usually $\|f(v)\| \ll \|f(z_\circ)\|$ (and also $\|f(v)\| \ll \|f(\hat{z})\|$) where $f(z) = h(z, 0)$ is the system that we are trying to solve. In fact, very often we observed a superlinear convergence rate on $\|f(\cdot)\|$, i.e. $\|f(v)\| \le \|f(z_\circ)\|^\alpha$ for some constant $\alpha > 1$. The challenge here is that the obtained point does not lie on the homotopy h, so we would not be able to continue further with this homotopy.

We propose to replace $h(z, t)$ with the new homotopy $h^v(z, t)$ defined as follows:

$$h^v(z, t) = f(z) - t f(v)$$

This is known as the "Newton homotopy" (see, e.g., Morgan et al. 1992b). Note that point $(v, 1)$ lies on this homotopy. Furthermore, by tracking this homotopy starting from this point we can expect to arrive at z^*.

Lemma 21. *Call v* good *if there exists a continuous curve $\varphi^v : (0, 1] \to \mathbb{C}^n$ with $\varphi^v(1) = v$ and $h^v(\varphi^v(t), t) = 0$ for all $t \in (0, 1]$. There exists a neighborhood Ω of x^* such that for any good point $v \in \Omega$ there holds $\lim_{t\to 0} \varphi^v(t) = z^*$.*

Proof. Since z^* is an isolated root of f, there exists closed ball B around z^* such that z^* is the only solution of $f(z) = 0$ over $z \in B$. Since its boundary ∂B is compact and f is continuous, there exists $\alpha = \min_{z\in\partial B} \|f(z)\| > 0$. Define $\Omega = \{z \in B : \|f(z)\| < \alpha/2\}$; clearly, this is a neighborhood of z^*. We claim that $\varphi^v(t) \in B$ for any $v \in \Omega$ and $t \in (0, 1]$. Indeed, we have $\|f(v)\| < \alpha/2$ since $v \in \Omega$. Also, $f(\varphi^v(t)) - t f(v) = 0$ and hence $\|f(\varphi^v(t))\| = \|t f(v)\| \le \|f(v)\| < \alpha/2$ for any $t \in (0, 1]$. Suppose there exists $t \in (0, 1]$ with $\varphi^v(t) \notin B$, then by continuity there exists $t' \in (t, 1)$ with $\varphi^v(t') \in \partial B$. But then $\|f(\varphi^v(t'))\| \ge \alpha$ and $\|f(\varphi^v(t'))\| < \alpha/2$ – a contradiction.

We showed that $\varphi^v((0, 1]) \subseteq B$. Since B is compact, curve φ^v must have at least one accumulation point as $t \to 0$. Let $\bar{z} \in B$ be such point. Continuity of functions h^v and φ^v implies that $f(\bar{z}) = h^v(\bar{z}, 0) = 0$. Since z^* is the only root of f in B, we must have $\bar{z} = z^*$. This means that curve φ^v has exactly one accumulation point as $t \to 0$, and hence $\lim_{t\to 0} \varphi^v(t) = z^*$. □

Unfortunately, tracking homotopy h^v from $t = 1$ can be very difficult, since we may not be in the endgame zone yet. We illustrate this phenomenon on the following example.

Example 1. *Consider the system*

$$f(z_1, z_2) \stackrel{\text{def}}{=} \begin{cases} z_1 - z_2 - z_2^2 \\ z_1 - z_2 + z_2^2 \end{cases} = 0 \tag{3.19}$$

It has unique solution $z = (0, 0)$ of corank $\kappa = 1$. Now fix $v \in \mathbb{C}^2$, and define homotopy $h^v(z, t) = f(z) - tf(v)$. Solving equation $h^v(z, t) = 0$ gives

$$\begin{cases} z_1 = v_2 \cdot t^{1/2} + (v_1 - v_2) \cdot t \\ z_2 = v_2 \cdot t^{1/2} \end{cases}$$

We can define the "endgame zone" as those values of t for which term $v_2 \cdot t^{1/2}$ of the Puiseux series dominates term $(v_1 - v_2) \cdot t$; in that case the linear predictor that estimates only the first term would give a good approximation. Thus, the endgame regime is given by the condition $|t|^{1/2} \ll |\frac{v_2}{v_1 - v_2}|$.

Let us fix $r > 0$, and consider two processes for generating v.

- *sample $v \in \mathbb{C}^2$ uniformly at random subject to $\|v\| = r$. Then $\mathbb{E}[|\frac{v_2}{v_1 - v_2}|] = \Theta(1)$, and so value $t = 1$ is* ***not*** *in the endgame zone.*
- *sample $\alpha \in \mathbb{C}^2$ uniformly at random subject to $\|\alpha\| = r$, obtain v by solving $f(v) = \alpha$. Then $\frac{v_2}{v_1 - v_2} = \frac{\sqrt{2(\alpha_2 - \alpha_1)}}{\alpha_1 + \alpha_2}$ and $\mathbb{E}[|\frac{v_2}{v_1 - v_2}|] = \Theta(\frac{1}{\sqrt{r}})$. Thus, $t = 1$ will be in the endgame zone if r is sufficiently small.*

One might ask whether the "γ-trick" could help. This is a standard approach in NAG to ensure genericity of h (Sommese and Wampler 2005, Chapter 7). The idea is to choose a random value $\gamma \in \mathbb{C}$ with $|\gamma| = 1$ and then define homotopy

$$h(z, t) = (1 - t)f(z) - \gamma t(f(v) - f(z))$$

Note that we still have $h(v, 1) = 0$ and $h(z, 0) = f(z)$. Solving equation $h(z, t) = 0$ gives

$$z_2 = v \cdot \sqrt{\frac{\gamma t}{1 - t + \gamma t}} = v\sqrt{\gamma} \cdot \left(t^{1/2} + \frac{1 - \gamma}{2} \cdot t^{3/2} + \ldots \right)$$

If $\gamma \neq 1$ then the first two terms are of the same order when $t = 1$, and so $t = 1$ is ***not*** *in the endgame zone for any choice of v.*

For systems with larger corank ($\kappa \geq 2$), the predictor step within the LAL (ArcLength Endgame) method is constructed analogously to the $\kappa = 1$ case. However, we have empirically observed that aggressively projecting the path all the way to the target root (i.e., jumping directly to $t = 0$) often produces numerical instabilities in subsequent corrector iterations.

Such an aggressive jump can severely degrade both the quality of the newly predicted point and the reliability of the ongoing fractional exponent estimations.

To mitigate this instability, we restrict the parameter jump in the predictor phase by introducing an adaptive shrinking-factor exponent, η, which governs the step-size multiplier ρ. Specifically, the predicted parameter value is defined as $\hat{t} = \rho \cdot t_\circ$, where the step size is explicitly formulated as $\rho = |t_\circ|^\eta$. To systematically control the progression toward the singularity, the exponent is updated iteratively via the rule $\eta \leftarrow \eta^\alpha$ for some prescribed constant $\alpha > 0$. See Section 3.6.1.

3.5 Predictors

In this section, we establish error bounds for the truncated predictor polynomials $\hat{z}(t)$ defined in Section 3.2.1. Specifically, we extend the theoretical framework of Lemma 15 to a higher-order scenario where we sample two points (z_1, t_1) and (z_2, t_2) lying on the tracked path, assuming knowledge of the first ℓ-order derivatives at these points.

The structure of the section is as follows. In the first part, we formulate a method to obtain approximations for the non-zero coefficients a_j in the Puiseux series and provide an asymptotic bound for their accuracy. In the next subsection, we formally define the ℓ-term Puiseux predictor $\hat{z}_\ell(t)$ and derive rigorous bounds for its distance with respect to the true solution path $z(t)$. Finally, the last subsection deals with the numerical estimation of the fractional exponents k_i/c.

3.5.1 Estimation of coefficients.

We begin by reviewing relevant properties of the Vandermonde matrix, adapting them to our specific setting. With this matrix, we define a linear system whose solution provides specific bounds for every coefficient in $\hat{z}$, and this allows us to derive an asymptotic relationship with respect to the curve $z(t)$ for $t \in (0, 1)$.

Definition 22 (Scaled Vandermonde matrix)**.** *Let $k_1, \dots, k_\ell$ and c be scalars with $c \neq 0$. The $\ell \times \ell$ Vandermonde matrix*

$$V(k_1/c, \dots, k_\ell/c) := (V_{ij})_{1 \le i,j \le \ell}, \quad V_{ij} := \left(\frac{k_j}{c}\right)^{i-1},$$

is called the scaled Vandermonde matrix *associated with the nodes $k_1, \dots, k_\ell$.*

It is well known that if the nodes $k_1, \dots, k_\ell$ are pairwise distinct, then V is nonsingular (see, e.g., Higham (2002, Chapter 22)). In this case, the

entries of the inverse matrix V^{-1} admit the explicit representation

$$(V^{-1})_{ij} = \frac{(-1)^{\ell-j} c^{\ell-1}}{\prod_{\substack{m=1 \\ m\neq i}}^{\ell} (k_i - k_m)} \sum_{\substack{S\subseteq\{1,\dots,\ell\}\setminus\{i\} \\ |S|=\ell-j}} \prod_{s\in S} \frac{k_s}{c}, \quad 1 \le i,j \le \ell.$$

This formula follows from the classical expression for the inverse of a Vandermonde matrix in terms of elementary symmetric polynomials.

We consider a solution curve $z(t)$ of the homotopy equation $h(z(t),t) = 0$, which admits a Puiseux-type expansion of the form

$$z(t) = \sum_{j=0}^{\infty} a_j \, t^{j/c},$$

for some constant $c > 0$. Let $p_0(t) := z(t) - z^*$ and for each integer $r \ge 1$, we define the auxiliary series

$$p_r(t) := \sum_{j=1}^{\infty} \left(\frac{j}{c}\right)^r a_j t^{j/c}.$$

Lemma 23. *With $z(t)$ and $p_r(t)$ as above, the following identity holds:*

$$p_r(t) = \sum_{j=1}^{r} S_j^r \, t^j \, \frac{\mathrm{d}^j}{\mathrm{d}t^j} z(t), \qquad r > 0 \tag{3.20}$$

where S_k^n denotes the Stirling numbers of the second kind, defined by the recurrence

$$S_k^{n+1} = S_{k-1}^n + k\, S_k^n,$$

with initial conditions $S_1^n = 1$, $S_0^n = 0$ for $n > 0$, and $S_n^n = 1$ for $n \ge 0$.

Proof. The formula is proved by induction on r. The base case $r = 1$ follows directly from the definitions.

For the inductive step, we first observe the relation $p_{r+1}(t) = t\frac{\mathrm{d}}{\mathrm{d}t} p_r(t)$ for $r \ge 1$, which follows directly from the structure of the derivatives of the monomials $t^{j/c}$. Assuming Equation (3.20) holds for a given $r \ge 1$, we apply the product rule to obtain:

$$\begin{aligned} p_{r+1}(t) = t\frac{\mathrm{d}}{\mathrm{d}t} p_r(t) &= \sum_{j=1}^{r} S_j^r t^{j+1} \frac{\mathrm{d}^{j+1}}{\mathrm{d}t^{j+1}} z(t) + \sum_{j=1}^{r} j S_j^r t^j \frac{\mathrm{d}^j}{\mathrm{d}t^j} z(t) \\ &= \sum_{j=1}^{r+1} (S_{j-1}^r + j S_j^r) t^j \frac{\mathrm{d}^j}{\mathrm{d}t^j} z(t) \\ &= \sum_{j=1}^{r+1} S_j^{r+1} t^j \frac{\mathrm{d}^j}{\mathrm{d}t^j} z(t). \end{aligned}$$

The final equality follows immediately from the recurrence relation for the Stirling numbers of the second kind. □

Now we assume we are in the scenario described in Section 3.2.1, and we wish to predict a "target" value $\hat{t} \in (0,1]$ with $\hat{t} = \rho \cdot t_\circ$. Let $t_1 := t_\circ$ with $t_1 > 0$, and set $t_2 = \lambda t_1$ for some fixed $\lambda \in (0,1)$. For each $r = 1, \dots, \ell$, define

$$P_r := p_{r-1}(t_1) - p_{r-1}(t_2) = \sum_{j=1}^{\infty} \left(\frac{j}{c}\right)^{r-1} a_j (t_1^{j/c} - t_2^{j/c}),$$

and let P be the column vector $P = (P_1, \dots, P_\ell)^\top$ where $\top$ denotes the standard (non-conjugate) transpose of the vector. Let $V = V(k_1/c, \dots, k_\ell/c)$ be the Vandermonde matrix from Definition 22, whose entries are $V_{ij} = (k_j/c)^{i-1}$. Since the nodes $k_1, \dots, k_\ell$ are pairwise distinct, V is nonsingular and its inverse V^{-1} exists. To simplify the notation, we will assume $k_j = j$ for $j = 1, \dots, \ell$ in the following proposition; however, the exact same argument applies to the general case.

Proposition 24. *For each $i = 1, \dots, \ell$, the i-th coordinate of the product $V^{-1}P$ satisfies*

$$(V^{-1}P)_i = a_i t_1^{i/c}(1 - \lambda^{i/c}) + O\big(t_1^{(\ell+1)/c}\big)$$

as $t_1 \to 0$, where $z(t) = \sum_{j=0}^{\infty} a_j t^{j/c}$ is the Puiseux type solution curve of the homotopy equation $h(z(t), t) = 0$.

Define the approximated coefficients $\tilde{a}_i$ by

$$\tilde{a}_i := \left[\operatorname{diag}\left(\frac{1}{t_1^{i/c}(1 - \lambda^{i/c})} \right) V^{-1} P \right]_i, \quad i = 1, \dots, \ell.$$

Then

$$\|a_j - \tilde{a}_j\| = O\big(t_1^{(\ell+1-j)/c}\big), \quad j = 1, \dots, \ell,$$

as $t_1 \to 0$.

Proof. From the definition of P_r and by expanding the difference $t_1^{j/c} - t_2^{j/c} = t_1^{j/c}(1 - \lambda^{j/c})$, we can separate the first ℓ terms of the series from the higher-order tail. This allows us to write the column vector P as the following matrix equation:

$$P = \begin{bmatrix} P_1 \\ \vdots \\ P_\ell \end{bmatrix} = V \begin{bmatrix} t_1^{1/c}(1 - \lambda^{1/c}) a_1 \\ \vdots \\ t_1^{\ell/c}(1 - \lambda^{\ell/c}) a_\ell \end{bmatrix} + L(t_1) \tag{3.21}$$

where $L(t_1)$ represents the truncation error vector containing the tail of the Puiseux series for $j \geq \ell + 1$. Since the dominant term in the tail occurs at $j = \ell + 1$, we have $L(t_1) = O\big(t_1^{(\ell+1)/c}\big)\,\mathbf{1}$, where $\mathbf{1}$ is the $\ell \times 1$ vector of ones.

Multiplying both sides of the equation from the left by the inverse Vandermonde matrix V^{-1} yields

$$V^{-1}P = \begin{bmatrix} t_1^{1/c}(1 - \lambda^{1/c}) a_1 \\ \vdots \\ t_1^{\ell/c}(1 - \lambda^{\ell/c}) a_\ell \end{bmatrix} + V^{-1} L(t_1) \tag{3.22}$$

Because the entries of V^{-1} are constants determined entirely by the fixed exponents and the cycle number, the linear transformation $V^{-1}L(t_1)$ preserves the asymptotic bound of the error term. Therefore, the i-th vector entry of the product $V^{-1}P$ is exactly:

$$\left(V^{-1}P\right)_i = a_i t_1^{i/c}(1-\lambda^{i/c}) + O\left(t_1^{(\ell+1)/c}\right) \tag{3.23}$$

which proves the first claim.

To prove the second assertion regarding the truncation error of the approximated coefficients $\tilde{a}_i$, we substitute equation (3.23) into the definition of $\tilde{a}_i$

$$\tilde{a}_i = \frac{\left(V^{-1}P\right)_i}{t_1^{i/c}(1-\lambda^{i/c})} = a_i + \frac{O\left(t_1^{(\ell+1)/c}\right)}{t_1^{i/c}(1-\lambda^{i/c})}$$

Since $\lambda \in (0,1)$ is a fixed constant, the denominator scales strictly by $t_1^{i/c}$. Factoring this out gives

$$\tilde{a}_i = a_i + O\left(t_1^{(\ell+1-i)/c}\right) \quad \text{as } t_1 \to 0.$$

□

To bridge the gap between the theoretical Vandermonde matrix formulation and practical software implementation, we now provide the explicit algebraic expansions of the approximated coefficients $\tilde{a}_j$ for the 1-term (linear), 2-term (quadratic), and 3-term (cubic) Puiseux predictors. In the last two cases we use $\top$ to denote the non-conjugate transpose.

Case $\ell = 1$. In the simplest case of the 1-term predictor, the sequence of nodes contains only k_1/c. The Vandermonde matrix reduces to the scalar $V = 1$, making its inverse trivial. The sample differences vector consists of a single entry corresponding to the distance, $P = [z_1 - z_2]$. Applying the diagonal scaling factor directly gives the explicit coefficient:

$$\tilde{a}_1 = \frac{z_1 - z_2}{t_1^{k_1/c}(1-\lambda^{k_1/c})}. \tag{3.24}$$

It is worth remarking that the linear predictor introduced in Section 3.3 emerges as the special continuous case of this formulation when targeting $\hat{z}_1(0)$ (aiming for $t = 0$). See Section 3.5.2.

Case $\ell = 2$. In practice we can also expand the coefficients for the 2-term Puiseux predictor ($\ell = 2$), which utilizes the leading exponents k_1/c and k_2/c, and the information of the first derivatives at points z_1 and z_2.

Using the explicit definition of the inverse Vandermonde matrix for $\ell = 2$, we have:

$$V^{-1} = \frac{1}{k_2 - k_1}\begin{bmatrix} k_2 & -c \\ -k_1 & c \end{bmatrix}.$$

Applying this inverse to the vector $P = [z_1 - z_2,\ t_1\dot{z}_1 - t_2\dot{z}_2]^\top$, the linear system yields the scaled leading terms:

$$V^{-1}\begin{bmatrix} z_1 - z_2 \\ t_1\dot{z}_1 - t_2\dot{z}_2 \end{bmatrix} = \frac{1}{k_2 - k_1}\begin{bmatrix} k_2(z_1 - z_2) - c(t_1\dot{z}_1 - t_2\dot{z}_2) \\ -k_1(z_1 - z_2) + c(t_1\dot{z}_1 - t_2\dot{z}_2) \end{bmatrix}.$$

By applying the diagonal scaling factor specified in Proposition 24, we obtain the explicit formulas for the approximated coefficients $\tilde{a}_1$ and $\tilde{a}_2$:

$$\tilde{a}_1 = \frac{k_2(z_1 - z_2) - c(t_1\dot{z}_1 - t_2\dot{z}_2)}{(k_2 - k_1)\, t_1^{k_1/c}(1 - \lambda^{k_1/c})}, \tag{3.25a}$$

$$\tilde{a}_2 = \frac{-k_1(z_1 - z_2) + c(t_1\dot{z}_1 - t_2\dot{z}_2)}{(k_2 - k_1)\, t_1^{k_2/c}(1 - \lambda^{k_2/c})}. \tag{3.25b}$$

Case $\ell = 3$. For the 3-term Puiseux predictor, the algorithm incorporates second-order derivatives to reconstruct the first three leading coefficients. The sample differences vector expands to $P = [P_1, P_2, P_3]^\top$, defined as:

$$\begin{aligned} P_1 &= z_1 - z_2, \\ P_2 &= t_1\dot{z}_1 - t_2\dot{z}_2, \\ P_3 &= (t_1^2\ddot{z}_1 + t_1\dot{z}_1) - (t_2^2\ddot{z}_2 + t_2\dot{z}_2)\,. \end{aligned}$$

Using the explicit formula for the inverse Vandermonde matrix with $\ell = 3$, let us define the denominator constants $\Delta_i = \prod_{m\neq i}(k_i - k_m)$ for brevity:

$$\begin{aligned} \Delta_1 &= (k_1 - k_2)(k_1 - k_3), \\ \Delta_2 &= (k_2 - k_1)(k_2 - k_3), \\ \Delta_3 &= (k_3 - k_1)(k_3 - k_2). \end{aligned}$$

Evaluating the combinatorial sum for the entries of V^{-1} produces the 3×3 inverse matrix:

$$V^{-1} = \begin{bmatrix} \frac{k_2k_3}{\Delta_1} & \frac{-c(k_2+k_3)}{\Delta_1} & \frac{c^2}{\Delta_1} \\ \frac{k_1k_3}{\Delta_2} & \frac{-c(k_1+k_3)}{\Delta_2} & \frac{c^2}{\Delta_2} \\ \frac{k_1k_2}{\Delta_3} & \frac{-c(k_1+k_2)}{\Delta_3} & \frac{c^2}{\Delta_3} \end{bmatrix}.$$

Multiplying V^{-1} by the sample vector P and applying the diagonal scaling matrix yields

$$\tilde{a}_1 = \frac{k_2k_3P_1 - c(k_2 + k_3)P_2 + c^2P_3}{\Delta_1\, t_1^{k_1/c}(1 - \lambda^{k_1/c})}, \tag{3.26a}$$

$$\tilde{a}_2 = \frac{k_1k_3P_1 - c(k_1 + k_3)P_2 + c^2P_3}{\Delta_2\, t_1^{k_2/c}(1 - \lambda^{k_2/c})}, \tag{3.26b}$$

$$\tilde{a}_3 = \frac{k_1k_2P_1 - c(k_1 + k_2)P_2 + c^2P_3}{\Delta_3\, t_1^{k_3/c}(1 - \lambda^{k_3/c})}. \tag{3.26c}$$

Having established the explicit coefficients for the predictor and their corresponding error bounds, the following section derives the approximation error with respect to the true solution curve $z(t)$ along $t \in (0, 1)$. This derivation assumes knowledge of two sample points, the first ℓ fractional exponents, and the path derivatives at the sample points up to order ℓ.

3.5.2 Main result: error bound of the predictor

As before, let us assume that we have an exact point (z_1, t_1) lying on the solution path such that $h(z_1, t_1) = 0$, which admits the Puiseux series expansion around $t = 0$:

$$p_0(t_1) = z(t_1) - z^* = \sum_{j=1}^{\infty} a_j t_1^{j/c} = a_{k_1} t_1^{k_1/c} + a_{k_2} t_1^{k_2/c} + \dots \tag{3.27}$$

where $p_0(t)$ is the function defined in the previous section.

Definition 25 (ℓ-term Puiseux predictor)**.** *Let $z(t) = \sum_{j=0}^{\infty} a_j t^{j/c}$ be the Puiseux-type solution curve of the homotopy $h(z(t), t) = 0$, and let $t_1 > 0$ and $t_2 = \lambda t_1$ with $\lambda \in (0, 1)$ be fixed sampling times. Let $\tilde{a}_{k_m}$ for $m = 1, \dots, \ell$ be the solution of the linear system arising from the Vandermonde-based reconstruction of the first ℓ non-zero coefficients of $p_0(t)$, as defined in Proposition 15. The ℓ-term Puiseux predictor is defined as*

$$\hat{z}(t) := z_1 + \sum_{m=1}^{\ell} \tilde{a}_{k_m} \left(t^{k_m/c} - t_1^{k_m/c} \right)$$

where $z_1 = z(t_1)$.

We now state the main result of this section, which establishes the asymptotic error bound of the predictor $\hat{z}(t)$ defined via the reconstructed coefficients $\tilde{a}_{k_m}$ from Proposition 15.

Theorem 26. *Let $\hat{z}(t)$ be the ℓ-term Puiseux predictor approximation of the curve $z(t)$. For a prediction step $t = \rho t_1$ with $\rho \in (0, 1)$, the prediction error satisfies the following asymptotic bound:*

$$\|\hat{z}(t) - z(t)\| = O\left(t_1^{k_{\ell+1}/c} \right),$$

which implies that the absolute error, relative to the current distance to the solution z^*, obeys the following relationship:

$$\|\hat{z}(t) - z(t)\| = O\left(\|z_1 - z^*\|^{k_{\ell+1}/k_1} \right).$$

Proof. The first bound is obtained by expanding $\hat{z}(t)$ from its definition and substituting the exact distance $z_1 - z^* = \sum_{j=1}^{\infty} a_j t_1^{j/c}$. Using the asymptotic relation between the approximated coefficients $\tilde{a}_{k_m}$ and the exact

coefficients a_{k_m}, and expanding the step $t = \rho t_1$, one obtains:

$$\begin{aligned}\|\hat{z}(t) - z(t)\| &\\ &= \left\| \sum_{m=1}^{\ell} (\tilde{a}_{k_m} - a_{k_m}) t_1^{k_m/c} (\rho^{k_m/c} - 1) - \sum_{j=k_{\ell+1}}^{\infty} a_j t_1^{j/c} (\rho^{j/c} - 1) \right\| \\ &= O\left(t_1^{k_{\ell+1}/c}\right),\end{aligned}$$

where $k_{\ell+1} \geq k_\ell + 1 > k_1$ encodes the leading exponent gap of the uncomputed terms in the Puiseux expansion.

Because the current distance to the exact root is dominated by the leading term, $\|z_1 - z^*\| = O(t_1^{k_1/c})$, this implies that:

$$\|\hat{z}(t) - z(t)\| = O\left(\|z_1 - z^*\|^{k_{\ell+1}/k_1}\right),$$

showing that the prediction error decays super-linearly as a fractional power of the current distance to the target solution z^*. □

In particular, for the linear and quadratic predictors corresponding to $\ell = 1$ and $\ell = 2$, one has $p_0(t_1) - p_0(t_2) = z_1 - z_2$ and $p_1(t_1) - p_1(t_2) = t_1\dot{z}_1 - t_2\dot{z}_2$. When the fractional exponents are sequential integers $(k_1, k_2, k_3) = (1, 2, 3)$, we have:

$$\|\hat{z}_{lin}(t) - z(t)\| = O\left(\|z_1 - z^*\|^2\right) \quad \text{and} \quad \|\hat{z}_{quad}(t) - z(t)\| = O\left(\|z_1 - z^*\|^3\right)$$

where $\hat{z}_{lin}(t)$ and $\hat{z}_{quad}(t)$ approximate the first two and three coefficients of the Puiseux series of $z(t)$, respectively.

It is worth to check the linear predictor. If we expand the ideal linear predictor:

$$\begin{aligned}\hat{z}_1(t) &= z_1 + \tilde{a}_1(t^{k_1/c} - t_1^{k_1/c}) \\ &= (z_1 - \tilde{a}_1 t_1^{k_1/c}) + \tilde{a}_1 t^{k_1/c}\end{aligned}$$

By taking the limit of the coefficient $\tilde{a}_1$ as the second sample point approaches the first ($t_2 \to t_1$), we evaluate the instantaneous derivative:

$$\lim_{t_2 \to t_1} \frac{z_1 - z_2}{t_1^{k_1/c} - t_2^{k_1/c}} = \frac{c}{k_1} t_1^{1-k_1/c} \dot{z}_1.$$

Substituting this limit expression back into the predictor equation and evaluating at the root target $t = 0$, we obtain:

$$\hat{z}_1(0) = z_1 - \left(\frac{c}{k_1} t_1^{1-k_1/c} \dot{z}_1\right) t_1^{k_1/c} = z_1 - \frac{c}{k_1} t_1 \dot{z}_1,$$

which perfectly recovers the predictor formulation defined in Section 3.2.1.

Comparison with Classical Hermite Interpolation.

In the classical power-series endgame (Sommese and Wampler 2005; Bates et al. 2013b), the cubic predictor is constructed using Hermite interpolation by matching the path positions and tangent derivatives at two consecutive sample points t_1 and t_2 sharing the same cycle number. We can adapt this classical approach to a generalized Puiseux series with arbitrary fractional exponents k_j/c. By substituting $t_2 = \lambda t_1$ and factoring out $t_1^{k_j/c}$, we can absorb the t_1 dependency directly into the vector of unknowns. This cleanly places the constants and the λ terms into the matrix, resulting in the following 4×4 linear system to recover the target root $a_0 = z^*$ and the scaled coefficients:

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & \frac{k_1}{c} & \frac{k_2}{c} & \frac{k_3}{c} \\ 1 & \lambda^{k_1/c} & \lambda^{k_2/c} & \lambda^{k_3/c} \\ 0 & \frac{k_1}{c}\lambda^{k_1/c} & \frac{k_2}{c}\lambda^{k_2/c} & \frac{k_3}{c}\lambda^{k_3/c} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 t_1^{k_1/c} \\ a_2 t_1^{k_2/c} \\ a_3 t_1^{k_3/c} \end{bmatrix} = \begin{bmatrix} z_1 \\ t_1 \dot{z}_1 \\ z_2 \\ t_2 \dot{z}_2 \end{bmatrix}. \tag{3.28}$$

The explicit exact inversion of the 4×4 classical system yields the isolated root a_0 as a linear combination of the tracking data as follows:

$$\left[\frac{\lambda^{2/c}(\lambda^{1/c} - 3)}{(\lambda^{1/c} - 1)^3}\right] z_1 - \left[\frac{c\lambda^{2/c}}{(\lambda^{1/c} - 1)^2}\right] t_1 \dot{z}_1 + \left[\frac{3\lambda^{1/c} - 1}{(\lambda^{1/c} - 1)^3}\right] z_2 - \left[\frac{c}{(\lambda^{1/c} - 1)^2}\right] t_2 \dot{z}_2 \tag{3.29}$$

Furthermore, attempting to analytically invert the classical 4×4 Hermite system introduces coefficients heavily weighted by $(1 - \lambda^{1/c})^{-3}$. As adaptive tracking steps become arbitrarily small ($\lambda \to 1$), this formulation suffers from severe subtractive cancellation. Our algebraic decoupling naturally factors out these singular scaling terms prior to inversion, drastically improving the numerical conditioning of the final endgame steps.

Comparing this classical formulation with the predictor introduced in Section 3.5.2 reveals two major theoretical and computational advantages:

1. **General Fractional Exponents:** The classical formulation strictly assumes a dense, sequential set of exponents $\{1/c, 2/c, 3/c\}$ (Sommese and Wampler 2005). By contrast, our method intrinsically supports arbitrary exponent sequences $\{k_1/c, k_2/c, \dots\}$, explicitly accounting for the gaps dictated by the value semigroup of the singularity.
2. **Dimensionality Reduction and Precision Efficiency:** Rather than inverting the full, strongly coupled 4×4 matrix to isolate the root a_0, our method utilizes the finite differences $P_r = p_{r-1}(t_1) - p_{r-1}(t_2)$. In the matrix formulation above, this is algebraically equivalent to subtracting the third row from the first, and the fourth row from the second. This systematically annihilates the column of 1s associated with the constant target root a_0, decoupling it from the higher-order coefficients.

3.5.3 Estimations for k_i/c

The fractional exponents of a branch do not simply form a dense sequence $1/c, 2/c, 3/c, \ldots$; rather, they are restricted to a value semigroup $S \subseteq \mathbb{N}/c$. Gaps in this semigroup represent fractional powers that are topologically prohibited from appearing in the expansion due to the degeneracy of the tangent cone (Wall 2004, Chapter 4; Greuel et al. 2007; Zariski 1932).

Accurately estimating the cycle number (or winding number) c and the leading fractional exponent k_1/c is a critical phase of the singular endgame. In this section, we present three distinct methodologies for computing k_1/c and its higher-order counterparts.

First, we review the standard **trial-and-error** method traditionally used to estimate the cycle number (Morgan et al. 1992a; Sommese and Wampler 2005). Second, we introduce a novel **path-limit estimation** technique that isolates the cycle number directly from the continuous limit of a quotient of finite differences evaluated along the path, which is theoretically justified by Lemma 15. Finally, we examine the **geometric sequence sampling** approach (Huber and Verschelde 1998; Bates et al. 2011; Sommese and Wampler 2005), which we subsequently complement with a new recursive heuristic designed to systematically annihilate leading terms and estimate the larger fractional exponents k_i/c.

Method 1: Trial and Error. A standard, albeit heuristic, approach to estimate the winding number c is to track the solution curve $z(t)$ from some initial $t_\circ \in (0, 1]$ down to a smaller value $t_1 < t_\circ$ using small steps Δt. At this stage, a sequence of predicted points $\hat{z}_c$ is computed using fractional power series models for various integer candidate values of c (typically $c = 1, 2, \ldots, c_{max}$).

Because the true solution curve exhibits the behaviour of a Puiseux series as it approaches the singular root z^*, the prediction model utilizing the correct cycle number c will yield a significantly tighter approximation than the incorrect candidates. By evaluating the prediction error ϵ_c for each candidate, the algorithm selects the c that minimizes the error. While simple to implement, this method is highly dependent on entering the endgame operating zone and is generally limited to small winding numbers (e.g., $c \leq 6$) when operating in standard double-precision arithmetic (Sommese and Wampler 2005).

Method 2: Path-limit estimation. Once inside the endgame operating zone, we can directly estimate the cycle number c and the leading fractional exponents without relying on trial and error. Let us assume that each coordinate function $z_i(t)$ can be locally expressed as an analytic function of t in a neighborhood of z^*. Specifically, by applying the local uniformization theorem (Sommese and Wampler 2005, Lemma 10.2.1), we can reparameterize the curve using $t = s^c$ such that each coordinate around z_i^* is expressed as a convergent power series in s.

Let $h(z, t) = 0$ be a linear homotopy between functions g and f. Returning to the parameter t, we can expand the Puiseux series for the i-th coordinate

of the solution path as:

$$z_i(t) = z_i^* + \sum_{j=1}^{\infty} a_j t^{j/c} = z_i^* + a_{k_1} t^{k_1/c} + a_{k_2} t^{k_2/c} + \dots \tag{3.30}$$

where z_i^* is the i-th coordinate of the isolated singular root z^* of the target function f. Here, $c \in \mathbb{N}$ is the cycle number, and $1 \leq k_1 < k_2 < k_3 < \dots$ represents a strictly increasing sequence of integers corresponding to the non-zero coefficients (i.e., $a_{k_\ell} \neq 0$).

For the sake of notational simplicity in the subsequent derivations, we will drop the coordinate index i and simply refer to the scalar components as $z(t)$ and z^*, as the algebraic operations apply uniformly to all coordinates. Recalling the auxiliary series $p_r(t)$ defined in Section 3.5 for Lemma (23) yields:

$$p_1(t) := t\dot{z}(t) = \sum_{j=k_1}^{\infty} \frac{j}{c} a_j t^{j/c} \tag{3.31a}$$

$$p_2(t) := t^2\ddot{z}(t) + t\dot{z}(t) = \sum_{j=k_1}^{\infty} \frac{j^2}{c^2} a_j t^{j/c} \tag{3.31b}$$

By defining the function $\zeta(t) := t\dot{z}(t) \equiv p_1(t)$, we can analyze the behavior of the finite difference quotient along the solution path. Taking a parameter value t and a scaled step ρt for some $\rho \in (0, 1)$, we evaluate the continuous limit of the quotient as $\rho \to 1$:

$$\lim_{\rho \to 1} \frac{\zeta(t) - \zeta(\rho t)}{z(t) - z(\rho t)} = \lim_{\rho \to 1} \frac{t\dot{z}(t) - \rho t\dot{z}(\rho t)}{z(t) - z(\rho t)} = \frac{\zeta'(t)}{\dot{z}(t)} = \frac{\frac{d}{dt}\left(t\dot{z}(t)\right)}{\dot{z}(t)} = \frac{\dot{z}(t) + t\ddot{z}(t)}{\dot{z}(t)}$$

By multiplying the numerator and denominator of this result by t, we can express the limit exactly as the ratio of the series $p_2(t)$ and $p_1(t)$ defined above. Substituting their fractional power series expansions yields:

$$\frac{t\dot{z}(t) + t^2\ddot{z}(t)}{t\dot{z}(t)} = \frac{p_2(t)}{p_1(t)} = \frac{(k_1^2/c^2)a_{k_1} t^{k_1/c} + O(t^{k_2/c})}{(k_1/c)a_{k_1} t^{k_1/c} + O(t^{k_2/c})} \approx \frac{k_1}{c} \quad \text{as } t \to 0$$

where k_1 is the index of the leading non-zero coefficient in the Puiseux series.

In particular, this ratio becomes computationally efficient for the Newton homotopy, $h(z, t) = f(z) - tf(z_0) = 0$. For this homotopy, $f(z) = tf(z_0)$, and the Davidenko ODE dictates that $\dot{z} = -h_z^{-1} h_t = f_z(z)^{-1} f(z_0)$. Multiplying by t, we obtain the relation $t\dot{z} = f_z(z)^{-1} f(z)$.

Therefore, by taking two points $z_1 = z(t_1)$ and $z_2 = z(t_2)$ along the path with $t_2 \to t_1$, we can numerically approximate the inverse ratio $\frac{c}{k_1}$ directly from the Newton step residuals:

$$\frac{\|z_1 - z_2\|}{\|t_1\dot{z}_1 - t_2\dot{z}_2\|} = \frac{\|z_1 - z_2\|}{\|f_z^{-1}(z_1)f(z_1) - f_z^{-1}(z_2)f(z_2)\|} \approx \frac{c}{k_1}.$$

Method 3: Geometric Sequence Sampling. An alternative method to estimate k_1/c without relying directly on the continuous limits of the derivatives is to sample the path along a geometric sequence (Bates et al. 2011; Sommese and Wampler 2005). We can estimate this number using three tracked points $z_1, z_2, z_3 \in \mathbb{C}^N$ evaluated at geometrically decreasing parameter values, e.g., $z_\ell := z(\rho^\ell t)$ for some step ratio $0 < \rho < 1$.

To extract the exponents, we subtract the values of two consecutive points in this sequence and project them onto a randomly chosen generic vector $v \in \mathbb{C}^N$. Taking the dot product isolates a scalar sequence:

$$\langle v, z_\ell - z_{\ell+1}\rangle = \left\langle v, \sum_{j=k_1}^{\infty} a_j t^{j/c}(1-\rho^{j/c})(\rho^{j\ell/c})\right\rangle = \sum_{j=k_1}^{\infty} v_j \rho^{j\ell/c}$$

where $v_j := \langle v, a_j\rangle t^{j/c}(1-\rho^{j/c}) \in \mathbb{C}$. By taking the logarithmic difference of three consecutive projected distances, we can isolate the leading fractional exponent:

$$\begin{aligned}
&\log|\langle v, z_{\ell+1} - z_{\ell+2}\rangle| - \log|\langle v, z_\ell - z_{\ell+1}\rangle| \\
&= \log\left|\sum_{j=k_1}^{\infty} v_j \rho^{j(\ell+1)/c}\right| - \log\left|\sum_{j=k_1}^{\infty} v_j \rho^{j\ell/c}\right| \\
&= \log\left|\frac{\langle v, z_{\ell+1} - z_{\ell+2}\rangle}{\langle v, z_\ell - z_{\ell+1}\rangle}\right| \\
&= \log\left|\rho^{k_1/c}\frac{v_{k_1} + \sum_{j>k_1} v_j \rho^{(j-k_1)(\ell+1)/c}}{v_{k_1} + \sum_{j>k_1} v_j \rho^{(j-k_1)\ell/c}}\right| \\
&\approx \frac{k_1}{c}\log\rho
\end{aligned}$$

where the approximation becomes exact as $t \to 0$ (or as ℓ becomes large), forcing the higher-order terms to vanish.

Once we have computed k_1/c, we can use a similar quotient to systematically annihilate the leading term and obtain an approximation for k_2/c. Indeed, define the remainder vector function $R_\ell(t)$ as:

$$R_\ell(t) := z(\rho^\ell t) - \frac{c}{k_1}\left(\rho^\ell t \dot{z}(\rho^\ell t)\right)$$

Using the series expansions defined in (3.30) and (3.31a), the first term is cancelled:

$$R_\ell(t) = z^* + \sum_{j=k_2}^{\infty}\left(1 - \frac{j}{k_1}\right) a_j t^{j/c}\rho^{\ell j/c}$$

Taking the scalar projection of the differences, let $\Delta R_\ell := \langle v, R_\ell(t) - R_{\ell+1}(t) \rangle$. The leading term of this new sequence is now governed by k_2/c, giving:

$$\log \left| \frac{\Delta R_{\ell+1}}{\Delta R_\ell} \right| \approx \frac{k_2}{c} \log \rho$$

More generally, we can approximate the value for any higher-order exponent k_i/c recursively. Let $R^{(0)}(t) := z(t) - z^*$, and define the successive higher-order recursive differences as:

$$R^{(i)}(t) := R^{(i-1)}(t) - \frac{c}{k_i} \left(t \frac{d}{dt} R^{(i-1)}(t) \right)$$

Applying this operator repeatedly eliminates the first i fractional terms, yielding:

$$R^{(i)}(t) = \sum_{j=k_{i+1}}^{\infty} \left[\prod_{m=1}^{i} \left(1 - \frac{j}{k_m} \right) \right] a_j t^{j/c}$$

By sampling this function along the geometric sequence and defining the scalar projections $\Delta R_\ell^{(i)} := \langle v, R^{(i)}(\rho^\ell t) - R^{(i)}(\rho^{\ell+1} t) \rangle$, we can extract the $(i+1)$-th exponent:

$$\log \left| \frac{\Delta R_{\ell+1}^{(i)}}{\Delta R_\ell^{(i)}} \right| \approx \frac{k_{i+1}}{c} \log \rho.$$

With these values estimated, we are equipped with the precise fractional sequence required to approximate the homotopy path using tools developed in Sections 3.4 and 3.5.

3.6 Numerical results

In this section, we evaluate the computational performance of the proposed methods. To establish a rigorous baseline, we benchmark our approach against a classical *predictor–corrector path tracker* equipped with a standard *power-series endgame*, as detailed in the foundational literature (Sommese and Wampler 2005; Bates et al. 2013b).

3.6.1 Description of implementation.

On the estimation of the fractional exponents. The implemented trial-and-error method primarily relies on two points: an initial anchor $(z_\circ, t_\circ)$ and a subsequent point (z_1, t_1) obtained via the standard tracker through a sequence of small steps. The cycle number is estimated by identifying the integer $c \in \{1, \dots, c_{\max}\}$ that minimizes the prediction residual $\|z_1 - \hat{z}_c(t_1)\|$, where $\hat{z}_c(t)$ is the linear predictor parameterized by the candidate cycle number.

The path-limit rule (cLim) requires two consecutive tracked points, (z_1, t_1) and (z_2, t_2), separated by a relatively small parameter step $|t_1 - t_2|$ to accurately approximate the continuous limit.

For the geometric sequence approach (cLoG), three tracked points lying strictly on a geometric parameter sequence are required to estimate the leading fractional exponent k_1/c. To systematically compute the subsequent exponent k_2/c, the method requires these same three points, their respective first-order derivatives, and the previously isolated value of k_1/c.

To extract the third exponent k_3/c, the estimation further relies on k_1/c, k_2/c, and the second-order parameter derivatives at each of the three geometric nodes. Because direct access to exact second derivatives is generally unavailable in standard continuation implementations, we utilize an unequally spaced finite difference approximation (Fornberg 1988). Explicitly, we evaluate the remainder sequence values R_i for each point $i \in \{0, 1, 2\}$ as:

$$R_i = z_i - \left(\frac{c}{k_1} + \frac{c}{k_2}\right) t_i \dot{z}_i + \left(\frac{c}{k_1}\frac{c}{k_2}\right) t_i \cdot \mathcal{D}^2(t_i),$$

where $\mathcal{D}^2(t_i)$ is a 3-point finite difference operator that approximates the exact derivative of the auxiliary function $t\dot{z}$ at t_i with $O(\Delta t^2)$ accuracy.

Let t_0, t_1, t_2 denote the previous, current, and next parameter values in the tracking sequence, and let $\zeta_j = t_j \dot{z}_j$ be the corresponding scaled first derivatives. By differentiating the Lagrange interpolating polynomial over these three nodes, the approximated derivative at any target node t_i in the sequence is given by:

$$\mathcal{D}^2(t_i) = w_{i,0}\zeta_0 + w_{i,1}\zeta_1 + w_{i,2}\zeta_2,$$

with the respective weights for $j \in \{0, 1, 2\}$ defined as:

$$w_{i,j} = \sum_{\substack{m=0 \\ m \neq j}}^{2} \frac{1}{t_j - t_m} \prod_{\substack{n=0 \\ n \neq j,m}}^{2} \frac{t_i - t_n}{t_j - t_n}.$$

Using these evaluated values, the third fractional exponent k_3/c is directly isolated from the logarithmic ratio of the successive differences of the remainder sequence R_i.

Classic Power Series Endgame (Classic). For our numerical experiments, the classical power-series endgame serves as the baseline method for comparison (Sommese and Wampler 2005; Bates et al. 2013b). First, the tracking procedure is initialized using a fixed-point Newton homotopy of the form $h(z, t) = (1 - t)f(z) + t\gamma(z - z_0) = 0$, where z_0 is an initial approximation close to the isolated singular root z^* and γ is a random complex constant (the "gamma trick") used to ensure the path avoids singularities prior to $t = 0$.

Next, we track the solution path as t moves from 1 to 0 using the standard predictor-corrector method equipped with a linear predictor, as described in Section 3.2.1. The Newton corrector is restricted to a maximum number of allowed steps (typically 5 to 7) to achieve convergence within a prescribed tolerance. If convergence is not detected, the step is rejected, the

step size Δt is decreased, and the predictor–corrector step is performed again. Conversely, the method employs an adaptive step size; after a certain number of successive successful iterations, the step size is multiplied by a constant, which in our case is 2.

As the path progresses and enters the *endgame operating zone*, the cycle number c is estimated using the Trial and Error method described in Section 3.5.3. This cycle number is dynamically updated in the linear predictor to correctly anticipate the fractional power series behaviour of the path. Additionally, the method utilizes a cubic predictor to accelerate progress; this is achieved by performing Hermite interpolation using the positions and tangent derivatives (z and dz/dt) of two consecutive points on the tracker path that share the same detected cycle number.

ArcLength Endgame (AL). The AL method introduces a specialized endgame strategy designed to actively circumvent the ill-conditioning of the Jacobian matrix as the path approaches a singular root. Initially, the path is tracked using the classical predictor–corrector method, which continuously monitors the evolution of the cycle number estimations via the macroscopic limit (cLim) and logarithmic (cLog) rules. Once both estimators converge to a shared integer value within a strict tolerance (typically 10^{-2}), the AL method effectively hijacks the tracking process, inheriting the state data from the classical tracker.

Instead of relying on the standard fractional power-series endgame, the AL method dynamically restructures the Newton corrector into an augmented, well-conditioned linear system. At each tracking step, the method evaluates the augmented matrix $h_x = [h_z, h_t]$ (i.e. the Jacobian matrix evaluated at the last point in the tracking process) to isolate its null-space vector. Using this approximate kernel, it constructs an augmented square Jacobian matrix, denoted as $\bar{h}_x$. By appending an orthogonal hyperplane constraint—derived from the tangent vector of the path—the system mathematically regularizes the singularity. This localized augmentation ensures that the modified Jacobian $\bar{h}_x$ retains full rank, allowing the Newton iterations to maintain robust, quadratic convergence deep into the singular regime.

For path progression, the AL method replaces the standard power-series endgame with a highly adaptive, data-driven schedule based on the locked cycle number estimator (c). Crucially, we assume that the fractional powers governing the path geometry follow the regular sequence $1/c, 2/c, 3/c$, which bypasses the need to computationally estimate arbitrary high-order exponents. The algorithm continuously monitors the variance of the sequence of cycle number estimates over successive steps. While the variance remains above a prescribed threshold, the method advances cautiously, calculating the next step size using an adaptively scaling fractional power η. However, once the variance drops below the threshold and the cycle estimates achieve stationarity (indicating that the asymptotic geometry of the path is fully resolved), the algorithm triggers a termination protocol. At this stage, it aggressively sets the target parameter to $t = 0$, effectively

"jumping" directly to the singular root using the final, well-conditioned augmented corrector. Furthermore, we observe that the performance of both methods can exhibit very different behavior when certain tracking parameters are modified. For example, changing the adaptive step-size schedule for the Classic method to a more aggressive scheme induces a boost in performance for some instances, allowing it to take larger leaps along the path. However, this aggressive scaling can also significantly increase the total number of matrix inversions, as the tracker accumulates many failed steps—forcing the corrector to reject the point, shrink the step size, and recompute the matrix inverse repeatedly.

Lifted ArcLength Endgame (LAL). Extending the approach from the corank-1 case, the LAL method relies on a similar predictor-corrector scheme. Much like the AL method, it begins by tracking the path with classical techniques until the fractional exponent estimators achieve stationarity. However, critical modifications are introduced to sustain the tracking process toward the singular root. The primary difference is the implementation of a dynamic homotopy reset, introducing the modified Newton homotopy $h^v(z,t)$ discussed in Section 3.4.2. This is defined as:

$$h^v(z,t) = f(z) - t\frac{f(v)}{\|f(v)\|}, \tag{3.32}$$

with the starting point initialized at $(z_\circ, t_\circ) = (v, \|f(v)\|)$.

During the predictor phase, we first estimate the fractional exponents as detailed in Section 3.5.3. Specifically, we employ the continuous path-limit (cLim) rule to estimate the leading fractional exponent, and the extended geometric sequence (cLog) rule for the subsequent fractional exponents. Additionally, when computing the third fractional exponent, we utilize the $O(\Delta t^2)$ finite difference approximation for the derivative of the auxiliary function $t\dot{z}$, as described at the beginning of this section.

As discussed in Section 3.4.2, aggressively setting the step size to jump directly to the target root at $t = 0$ (i.e., setting $\rho = 0$) is prohibitive, as the subsequent corrector steps exhibit highly unstable progress. Thus, we introduce a dynamic shrinking factor updated after every iteration. Specifically, the predicted parameter value is defined as $\hat{t} = \rho \cdot t_\circ$, where the step ratio is explicitly formulated as $\rho = |t_\circ|^\eta$. To systematically govern the progression toward the singularity, the scaling exponent is initialized at $\eta = 1.1$ and updated iteratively via the geometric rule $\eta \leftarrow \eta^\alpha$. For failed iterations, the step size is penalized by setting $\alpha = 0.5$. For successful iterations, we attempt to accelerate tracking by setting $\alpha = 1.2$; however, this acceleration is strictly conditional.

The algorithm continuously monitors the variance across the coordinate-wise estimates for each exponent order (e.g., tracking distinct variances for the k_1/c ensemble and the k_2/c ensemble). The scaling exponent η is permitted to increase only when these variances drop below a predefined tolerance threshold, ensuring the local asymptotic geometry has stabilized. Furthermore, to prevent the step size ρ from shrinking dangerously close

to 0 prematurely, we impose an upper bound on η. This bound is set to $(k_2/k_1)^\beta$ for the quadratic predictor (Q) and $(k_3/k_1)^\beta$ for the cubic predictor (C). In our implementation, we evaluate thresholds of $\beta = 0.5$ and $\beta = 0.9$, yielding the variants labeled Q5, Q9, C5, and C9 in the convergence plots.

In the corrector phase, we augment the Jacobian matrix $h_x = [h_z, h_t]$ similarly to the standard AL method, which effectively reduces the corank by exactly one. However, a second extension is required to completely recover the numerical rank of the system. The algorithm performs a SVD on the standard augmented matrix $h_x^v =: [H_z, H_t]$ to isolate the near-zero singular values (falling below a strict tolerance threshold, set to 10^{-5} in our implementation). Let U be the matrix whose columns u_i correspond to these near-null left-singular vectors.

Using this basis, we construct the full row-rank block matrix $[H_z, H_t, U]$. To formulate a well-posed, square augmented Jacobian $\bar{H}_x$, we append an orthogonal constraint block Q derived from the tangent space:

$$Q = \Big[\left(H_z^{-1} [H_t, U] \right)^\dagger, -I \Big],$$

where $\dagger$ denotes the Hermitian transpose. This yields the fully augmented polynomial system:

$$\bar{H}(z, t, \xi) = \begin{bmatrix} h^v(z,t) + U\xi \\ Q \begin{bmatrix} z - \hat{z} \\ t - \hat{t} \\ \xi \end{bmatrix} \end{bmatrix} = 0,$$

where ξ represents the vector of auxiliary artificial variables, and $(\hat{z}, \hat{t})$ is the predicted state. Finally, this well-conditioned system is solved using the Newton-Raphson method, terminating once the Newton step-ratio contracts below a prescribed tolerance. Because of the geometric relaxation introduced by the orthogonal hyperplanes, the newly corrected point (z_n, t_n) no longer lies strictly on the exact solution curve $h^v(z,t) = 0$. Thus, to continue tracking toward the root, the algorithm dynamically resets the Newton homotopy by setting $v \leftarrow z_n$, and the predictor-corrector sequence repeats for the subsequent step.

Instances. To rigorously evaluate the performance of our endgame strategies, we benchmark against several well-known zero-dimensional polynomial systems from the NAG literature. These include the Griewank-Osborne system (Griewank and Osborne 1981), characterized by its narrow convergence trumpet; Lecerf's deflation benchmark (Lecerf 2002); and the Caprasse system (Hauenstein et al. 2015). To systematically isolate the impact of the Jacobian corank (κ) on predictor stability, we evaluate our methods on a family of crafted polynomial systems. These instances are constructed with a prescribed singular root at the origin, $z^* = \mathbf{0} \in \mathbb{C}^n$, using the form:

$$F(z) = Az + D(z) = 0 \tag{3.33}$$

Here, $A \in \mathbb{Z}^{n \times n}$ is a dense random integer matrix explicitly constructed to have rank $n - \kappa$. Because the Jacobian at the origin reduces strictly to this linear part ($JF(\mathbf{0}) = A$), the system guarantees a defect of exactly κ. The higher-order term $D(z)$ is a diagonal mapping of monomials $z_i^{\alpha_i}$ with randomly sampled integers $\alpha_i > 1$. Additionally, we obfuscate the system using dense random linear transformations on both the coordinates and equations, ultimately tracking $N \cdot F(Mz) = 0$.

Additionally, we use a second family of instances where the higher-order components are crossed monomials with prescribed total degree. These systems take the form:

$$F(z) = Az + T(z) = 0 \tag{3.34}$$

As before, $A \in \mathbb{Z}^{n \times n}$ is a random matrix explicitly constructed to have rank $n - \kappa$. However, instead of a simple diagonal mapping, the higher-order component $T(z)$ consists of multivariate polynomials. Each component $T_i(z)$ is formed by summing randomly generated monomials of the form $c \prod_{k=1}^{n} z_k^{\alpha_k}$, where the coefficients c and non-negative integer exponents α_k are randomly sampled.

To guarantee that the linear part strictly dominates the Jacobian at the origin—thereby perfectly preserving the prescribed corank κ—we enforce a total degree constraint of $\sum \alpha_k \geq 2$ on every monomial. This ensures that $\lim_{z \to \mathbf{0}} JT(z) = \mathbf{0}$. Finally, the system is subjected to the same dense obfuscation tracking, $N \cdot F(Mz) = 0$.

3.6.2 Plots

Convergence comparisons. In the following experiments, we present a convergence comparison between the classical power-series endgame (denoted as Classic) and the proposed ArcLength Endgame (AL, LAL) methods. To provide a clear, hardware-independent measure of algorithmic efficiency, the horizontal axis in all convergence plots denotes the cumulative number of matrix inversions. The vertical axis displays the tracking residual on a logarithmic scale, $\log_{10} \|f(z)\|$, illustrating the depth of convergence as the path approaches the singular root.

Throughout this section, test instances are identified in the figure captions by the source of the polynomial system and the parameter tuple (n, κ, c), denoting the number of variables, the prescribed corank of the Jacobian matrix at the singular root, and the winding cycle number, respectively. For certain instances, we also indicate the number of high-order monomials, denoted as $\#mon$, added to each equation, alongside the interval from which the monomial degrees α_i are sampled.

The evaluated LAL variants are distinguished in the plots by their predictor degree and adaptive step-size thresholds. Specifically, LAL-Q designates methods utilizing a quadratic predictor, while LAL-C designates those utilizing a cubic predictor. The appended numerical suffixes (e.g., Q5,

C9) indicate the specific bound thresholds used to govern the predictor step-size logic.

Finally, to effectively track paths deep within the endgame operating zone and mitigate severe ill-conditioning, the numerical trackers are initialized with a baseline of 500 digits of multiprecision arithmetic. To prevent numerical underflow, this precision is expanded dynamically in direct proportion to the order of magnitude of the homotopy parameter, $O(-\log_{10}|t|)$. Additionally, in instances where a tracking method fails to converge—whether due to stalling or divergence—the trajectory is truncated. The final valid step recorded before the tracker failed is explicitly marked on the plot with a hollow black circle (○). In our implementation, the floating-point outputs are rounded to the closest rational number p/q within a prescribed tolerance of tol $= 10^{-2}$. Consequently, the exact integer values reported in the experiment descriptions and figure captions (e.g., $c = 4$ and $k_{1,2,3} = \{1, 2, 3\}$) reflect these recovered rational representations.

Comparing c/k_1 Estimation Methods. In the final section, we present plots comparing the performance of the three approaches discussed in Section 3.5.3 to estimate the cycle number. To evaluate each method, we take a current state pair (z_1, t_1) and utilize the history of points generated along the tracking path using the classical power series endgame previously described.

For the trial-and-error rule (cSort), we compare the predicted values across all possible cycle numbers ranging from 1 to a predefined maximum, $C_{\max} = 16$. For the path-limit estimation (cLim), the estimation relies on two consecutive points from the history, supplemented by an additional point (cLim+) computed using a stepsize λ that is small relative to $|t|$ (e.g., $\lambda = |t| \cdot 10^{-10}$).

Finally, the geometric sequence sampling method (cLog) requires three specific values (t_1, t_2, t_3) that form a geometric progression such that $t_2 = \rho t_1$ and $t_3 = \rho^2 t_1$. Consequently, we only perform this computation when three consecutive history points satisfy this geometric property. If this condition is not met at a given step, the computation is skipped. In the accompanying plots, these skipped computations are bridged using linear interpolation to maintain the continuity of the curves.

Comparison performance AL vs CLASSIC:

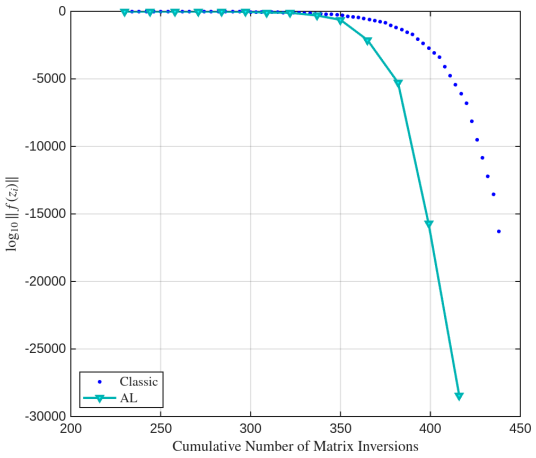

Figure 3.1: **Griewank-Osborne system.** (Griewank and Osborne 1981) System parameters: $(n, \kappa, c) = (2, 1, 3)$.

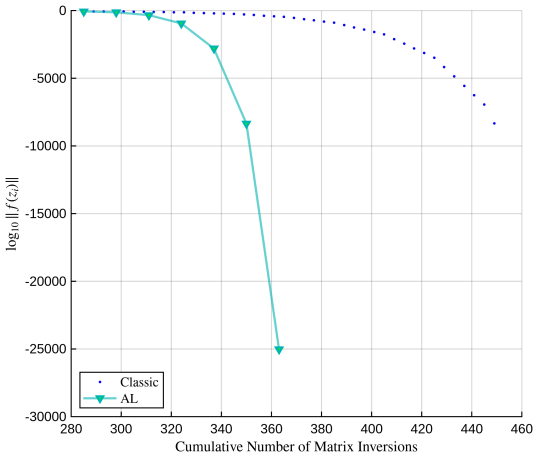

Figure 3.2: **Diagonal generator** $D(z)$**.** System parameters: $(n, \kappa, c) = (5, 1, 5)$, with sampled monomial degrees $|\alpha_i| \in [3, 7]$.

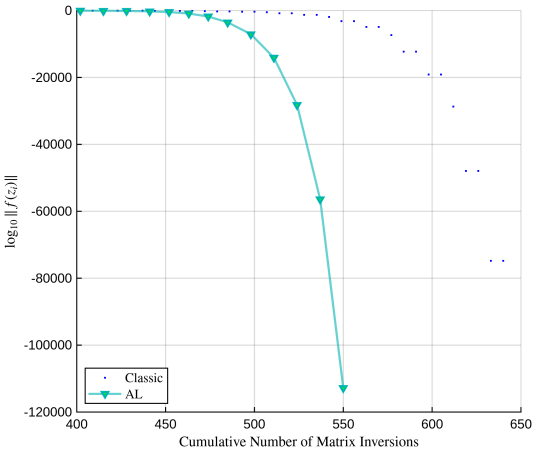

Figure 3.3: **Diagonal generator** $D(z)$**.** System parameters: $(n, \kappa, c) = (5, 1, 5)$, with sampled monomial degrees $|\alpha_i| \in [3, 7]$.

Instances from literature:

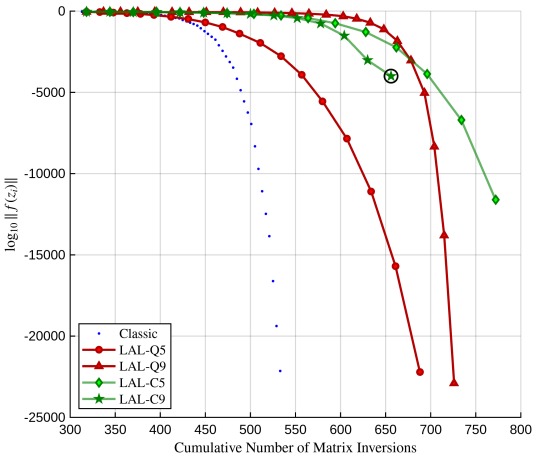

Figure 3.4: **Lecerf system** (Lecerf 2002). System parameters: $(n, \kappa, c) = (3, 2, 6)$, with $k_{1,2,3} = \{1, 2, 3\}$.

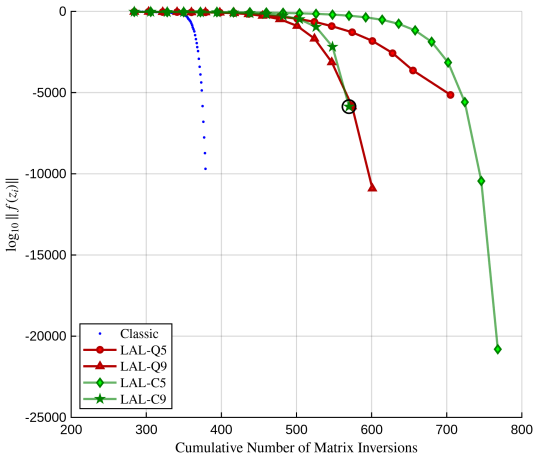

Figure 3.5: **Caprasse system** (Hauenstein et al. 2015). System parameters: $(n, \kappa, c) = (3, 2, 2)$, with $k_{1,2,3} = \{1, 2, 3\}$.

Problems first generator:

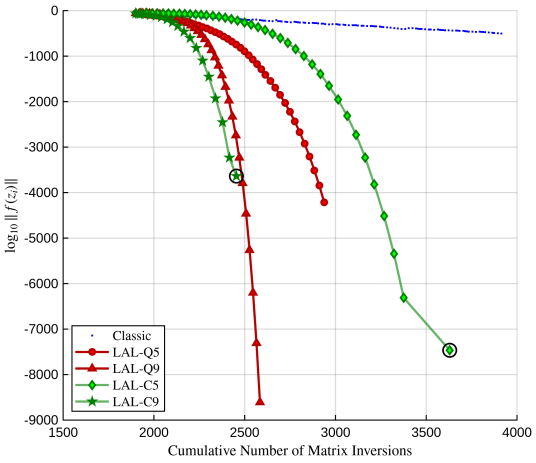

Figure 3.6: **Diagonal generator** $D(z)$**.**
System parameters: $(n, \kappa, c) = (7, 5, 30)$, with $k_{1,2,3} = \{5, 6, 7\}$ and $\alpha \in [3, 7]$.

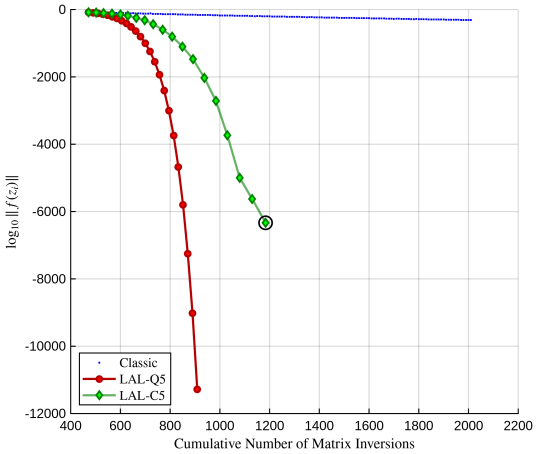

Figure 3.7: **Diagonal generator** $D(z)$**.**
System parameters: $(n, \kappa, c) = (9, 2, 20)$, with $k_{1,2,3} = \{4, 5, 6\}$ and $\alpha \in [3, 10]$.

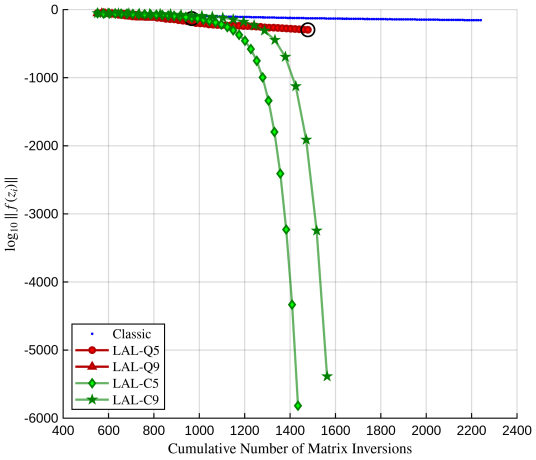

Figure 3.8: **Diagonal generator** $D(z)$**.**
System parameters: $(n, \kappa, c) = (6, 4, 90)$, with $k_{1,2,3} = \{10, 15, 18\}$ and $\alpha \in [3, 10]$.

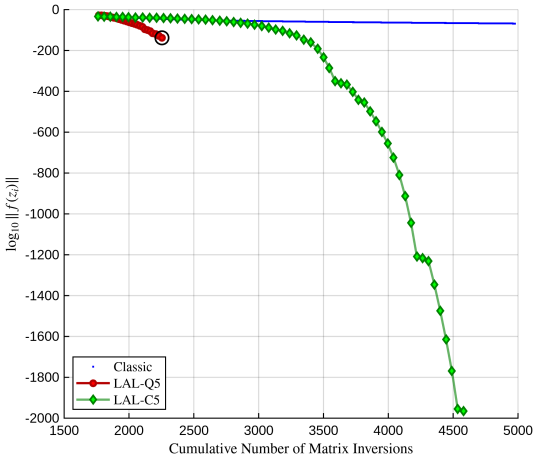

Figure 3.9: **Diagonal generator** $D(z)$**.**
System parameters: $(n, \kappa, c) = (8, 7, 210)$, with $k_{1,2,3} = \{30, 35, 42\}$ and $\alpha \in [3, 5]$.

Problems second generator:

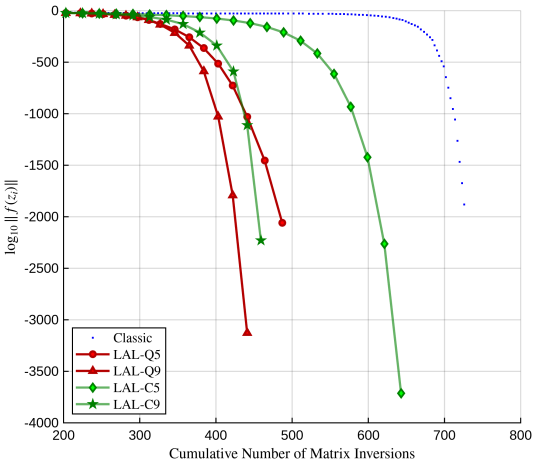

Figure 3.10: **Multivariate generator** $T(z)$**.** System parameters: $(n, \kappa, c, \#mon) = (4, 2, 4, 2)$, with $k_{1,2,3} = \{1, 2, 3\}$ and degree constraint $\alpha \in [1, 4]$.

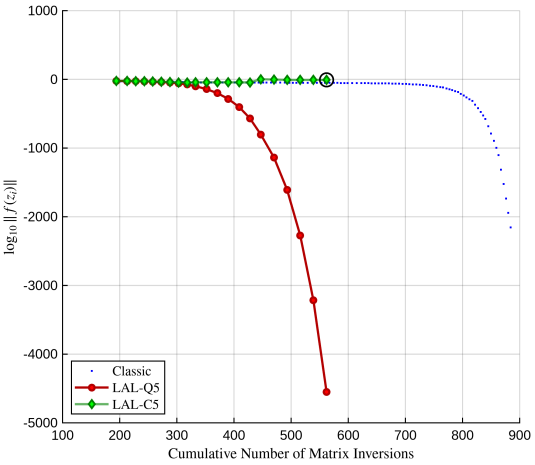

Figure 3.11: **Multivariate generator** $T(z)$**.** System parameters: $(n, \kappa, c, \#mon) = (5, 2, 4, 2)$, with $k_{1,2,3} = \{1, 2, 3\}$ and degree constraint $\alpha \in [1]$.

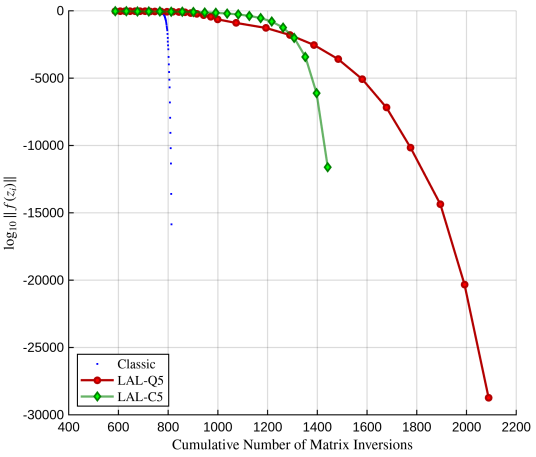

Figure 3.12: **Multivariate generator** $T(z)$**.** System parameters: $(n, \kappa, c, \#mon) = (4, 3, 2, 1)$, with $k_{1,2,3} = \{1, 2, 3\}$ and degree constraint $\alpha \in [1]$.

Comparison estimations c/k_1:

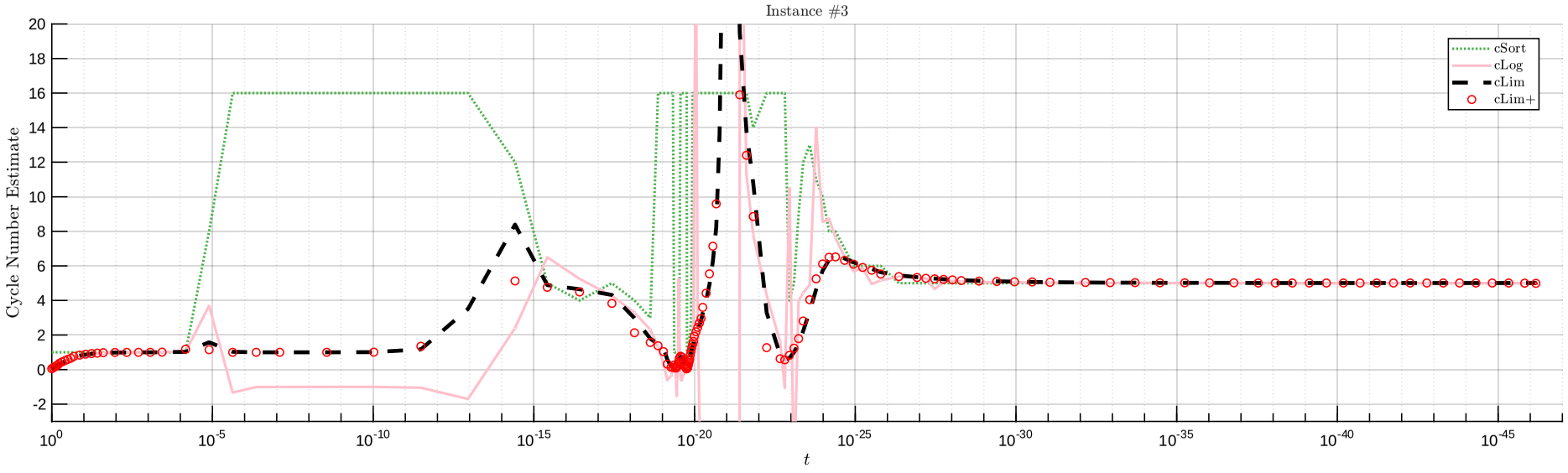

Figure 3.13: [n, κ]=[5,1]

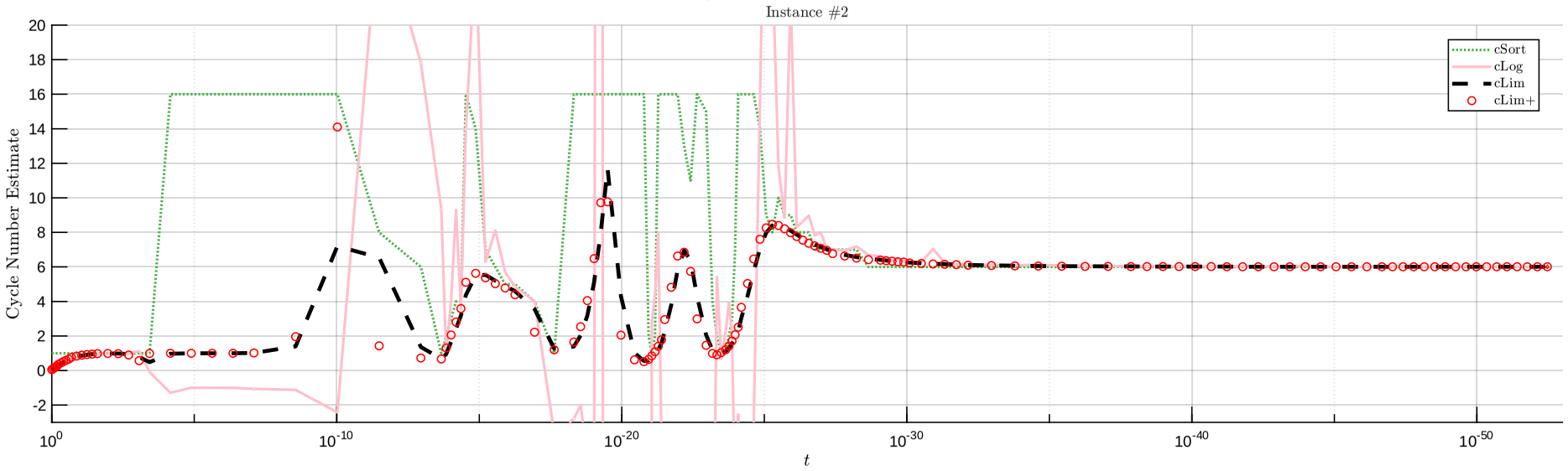

Figure 3.14: [n, κ]=[8,7]

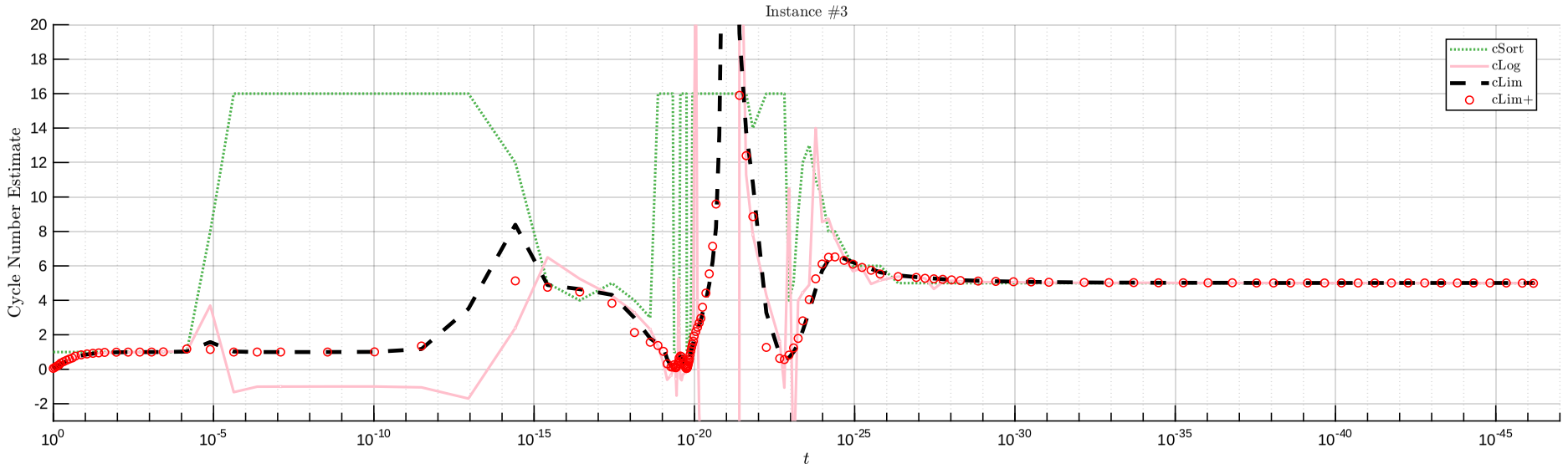

Figure 3.15: [n, κ]=[5,4]

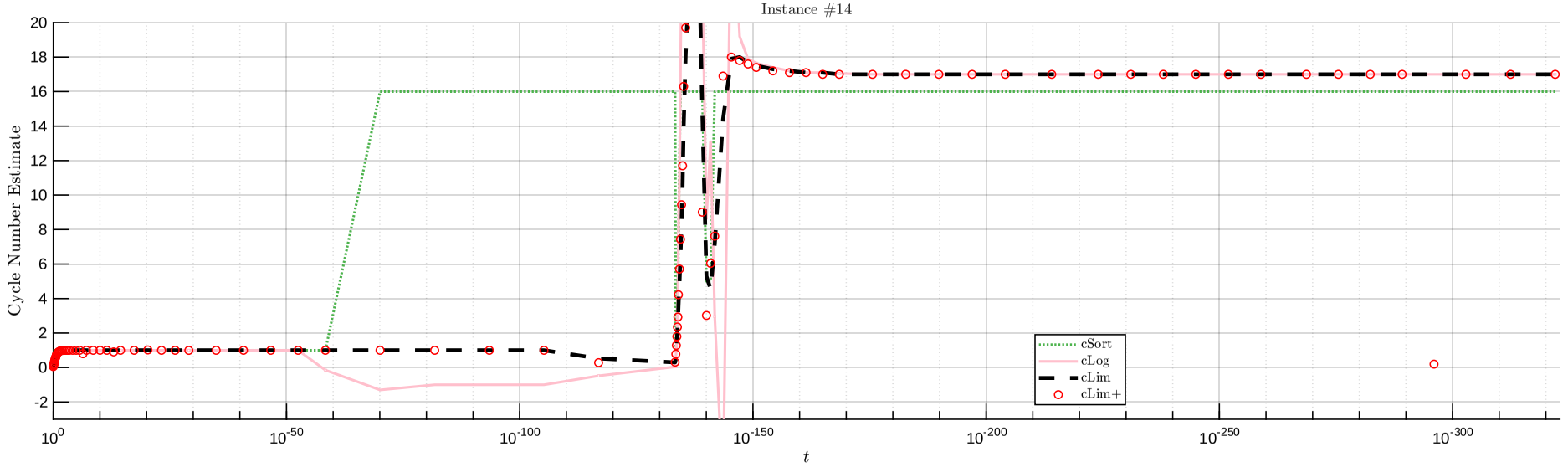

Figure 3.16: [n, κ]=[3,2]

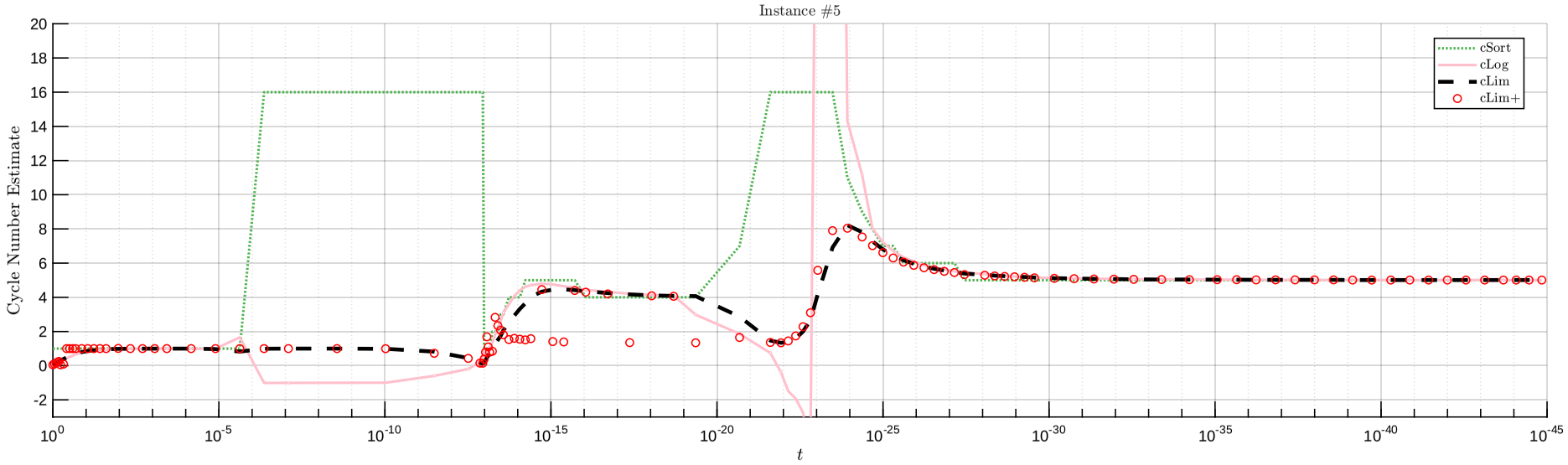

Figure 3.17: [n, κ]=[5,4]

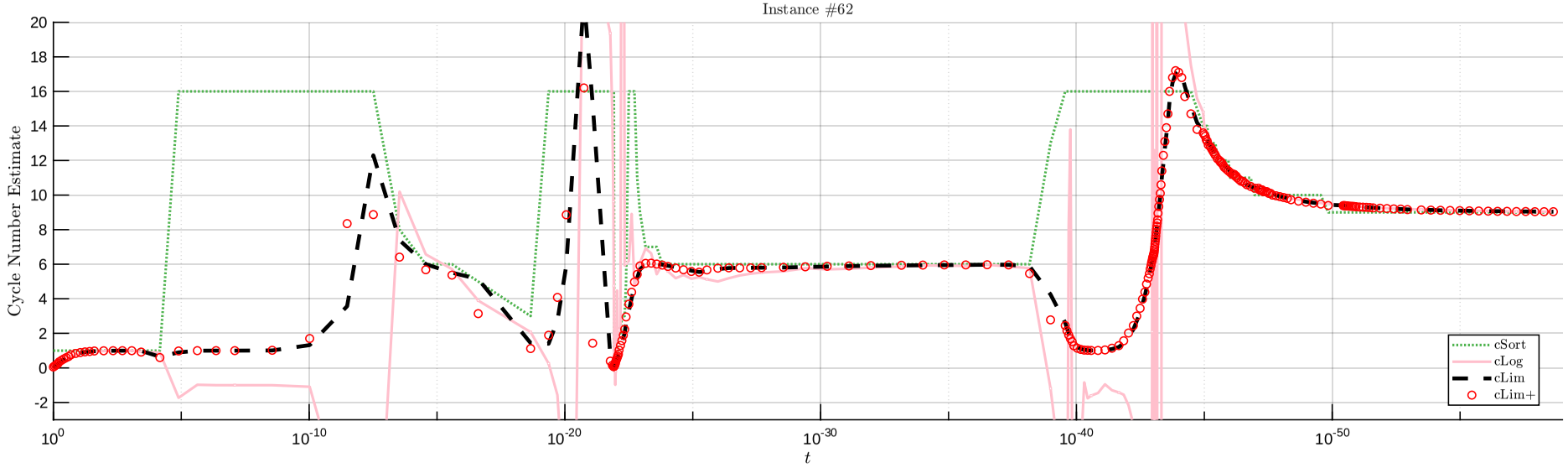

Figure 3.18: [n, κ]=[6,4]

List of Notation for Chapter 3

c	The cycle number (the denominator of the fractional exponents).
$f(z) = 0$	The target system of polynomial equations.
$h(z, t)$	The homotopy mapping.
$[h_z, h_t]$	The partial derivatives (Jacobians) of the homotopy function $h(z, t)$ with respect to z and t.
$[\bar{H}_z, \bar{H}_t, \bar{H}_\xi]$	The augmented square Jacobian matrix used to restore full rank.
J^*	The Jacobian matrix $f_z(z^*)$ at the singular point.
κ	The corank of the Jacobian matrix J^*, i.e., $n - \mathrm{rank}(J^*)$.
k_i/c	The i-th fractional exponent in the Puiseux series expansion.
σ_i, u_i, v_i	The singular values, left singular vectors, and right singular vectors of the Jacobian matrix h_z.
$x_\circ$	A pair $(z_\circ, t_\circ)$ such that $h(z_\circ, t_\circ) = 0$ and $z_\circ = z(t_\circ)$ belongs to the convergence zone of the Puiseux expansion.
$\widehat{x}$	The predictor point $(\widehat{z}, \widehat{t})$ evaluated at $\widehat{t} = \rho \cdot t_\circ$.
$\widehat{z}(t)$	The ℓ-term Puiseux series predictor of the solution path.
$\dot{z}(t)$	The derivative of the solution path $z(t)$ with respect to t.
z^*	An isolated singular solution of the system $f(z) = 0$.
AL	ArcLength endgame implementation.
LAL	Lifted ArcLength endgame implementation.
LAL-Q	LAL variant utilizing a quadratic predictor.
LAL-C	LAL variant utilizing a cubic predictor.
cLim	Implementation of the continuous path-limit quotient rule.
cLog	Implementation of the extended geometric sequence sampling rule.
cSort	Implementation of the trial-and-error sorting rule.

Chapter 4

Conclusions

This thesis advances the computational treatment of degeneracy in both semidefinite programming (SDP) and numerical algebraic geometry (NAG). While the two chapters present independent theoretical frameworks, together they establish a connection between heuristic numerical approximation and rigorous, certifiable algebraic solutions.

In the first part of this work, we established a hybrid pipeline for certifying the feasibility of weakly feasible, degenerate SDPs without relying on the existence of rational solutions. By proving that the exact maximum-rank solution to an LMI can be isolated as a solution to a carefully constructed system of polynomial equations, we eliminated a theoretical bottleneck in optimization certification. Our numerical evaluations demonstrate that, when initialized with a high-quality approximate solution, this hybrid framework can outperform purely exact algorithms presented in the literature.

However, our current experiments reveal that the scalability of this approach remains constrained by the intermediate solvers—specifically, the facial reduction algorithm based on Bertini and the exact solver for polynomial systems (msolve). To address these bottlenecks, a promising direction for future work is the replacement of the exact polynomial solver with an advanced NAG solver. Such an approach would be particularly advantageous in geometrically complex cases where the formulated polynomial system exhibits both zero-dimensional and positive-dimensional components. Provided a strong initial approximation from the prior step, a purely numerical solver could efficiently isolate the desired zero-dimensional target, paving the way for the certification of much larger, real-world SDP instances.

In the second part of this work, we addressed the computational bottleneck encountered when numerically solving degenerate polynomial systems. The proposed AL and LAL endgames bridge the gap between classical heuristic path tracking and singular root isolation, circumventing the computational bloat of traditional deflation techniques.

By explicitly reconstructing the coefficients of sparse Puiseux series, our algebraic predictors leverage the local geometric information obtained when only small, conservative steps are viable. This reconstruction enables

the tracker to execute accurate parameter leaps deep into the endgame zone. Our numerical experiments underscore the efficacy of the adaptive step-size parameter η. By continuously monitoring the statistical variance of the coordinate-wise fractional exponent estimates via the cLim and cLog rules, the LAL method maintains a systematic path progression. Evaluated against the cumulative number of matrix inversions, the localized hyperplane augmentation in LAL demonstrates effective superlinear convergence, even for heavily degenerate systems exhibiting higher coranks ($\kappa \geq 2$).

Future work in this domain will focus on enhancing the stability and robustness of the LAL approach to reliably solve systems with larger coranks. Ultimately, the integration of these stable singular endgames provides the numerical machinery required to address the bottlenecks identified in the development of the present work.

References

Alizadeh, Farid, Jean-Pierre A. Haeberly, and Michael L. Overton (1997). "Complementarity and nondegeneracy in semidefinite programming." In: *Mathematical Programming* 77.1, pp. 111–128. DOI: 10.1007/BF02611508.

Atkinson, Kendall (1991). *An Introduction to Numerical Analysis*. 2nd. New York, NY: Wiley.

Bank, Bernd, Marc Giusti, Joos Heintz, and Luis Miguel Pardo (2005). "Generalized polar varieties: geometry and algorithms." In: *Journal of Complexity* 21.4, pp. 377–412. DOI: 10.1016/j.jco.2004.09.001.

Basu, Saugata, Richard Pollack, and Marie-Françoise Roy (2006). *Algorithms in Real Algebraic Geometry*. 2nd. Vol. 10. Algorithms and Computation in Mathematics. Berlin: Springer-Verlag. DOI: 10.1007/3-540-33099-2.

Bates, Daniel J., Jonathan D. Hauenstein, and Andrew J. Sommese (2011). "A Parallel Endgame." In: *Contemporary Mathematics* 556, pp. 25–35. DOI: 10.1090/conm/556/11021.

Bates, Daniel J., Jonathan D. Hauenstein, Andrew J. Sommese, and Charles W. Wampler (2013a). *Numerically Solving Polynomial Systems with Bertini*. Philadelphia, PA, USA: SIAM. DOI: 10.1137/1.9781611973167.

— (2013b). *Numerically Solving Polynomial Systems with Bertini*. Philadelphia, PA, USA: SIAM. DOI: 10.1137/1.9781611972702.

Berthomieu, Jérémy, Christian Eder, and Mohab Safey El Din (2021). "MSOLVE: A library for solving polynomial systems." In: *Proceedings of the 2021 International Symposium on Symbolic and Algebraic Computation*, pp. 51–58. DOI: 10.1145/3452143.3465554.

Borwein, Jon M. and Henry Wolkowicz (1981a). "Facial reduction for a cone-convex programming problem." In: *Journal of the Australian Mathematical Society. Series A. Pure Mathematics and Statistics* 30.3, pp. 369–380. DOI: 10.1017/S144678870002164X.

— (1981b). "Regularizing the abstract convex program." In: *Journal of Mathematical Analysis and Applications* 83.2, pp. 495–530. DOI: 10.1016/0022-247X(81)90137-9.

Cox, David, John Little, and Donal O'Shea (2007). *Ideals, Varieties, and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. New York, NY: Springer.

Dostert, Maria, David de Laat, and Philippe Moustrou (2021). "Exact semidefinite programming bounds for packing problems." In: *SIAM Journal on Optimization* 31.2, pp. 1433–1458. DOI: 10.1137/20M1346067.

Drori, Yoel and Marc Teboulle (2014). “Performance of first-order methods for smooth convex minimization: A novel approach.” In: *Mathematical Programming* 145.1, pp. 451–482. DOI: 10.1007/s10107-013-0653-0.
Drusvyatskiy, Dmitriy and Henry Wolkowicz (2017). “The many faces of degeneracy in conic optimization.” In: *Foundations and Trends in Optimization* 3.2, pp. 77–170. DOI: 10.1561/2400000017.
Eisenbud, David (1988). “Linear sections of determinantal varieties.” In: *American Journal of Mathematics* 110.3, pp. 541–575. DOI: 10.2307/2374622.
Ferreira, O. P. and B. F. Svaiter (2012). “Kantorovich’s Theorem on Newton’s Method.” In: *arXiv preprint*. arXiv: 1209.5704.
Fornberg, Bengt (1988). “Generation of finite difference formulas on arbitrarily spaced grids.” In: *Mathematics of Computation* 51.184, pp. 699–706. DOI: 10.1090/S0025-5718-1988-0935077-0.
Goemans, Michel X. and David P. Williamson (1995). “Improved approximation algorithms for maximum cuts and satisfiability problems using semidefinite programming.” In: *Journal of the ACM* 42.6, pp. 1115–1145. DOI: 10.1145/227683.227684.
Golub, Gene H. and Charles F. Van Loan (2013). *Matrix Computations*. Johns Hopkins Studies in the Mathematical Sciences. Baltimore, MD, USA: Johns Hopkins University Press. ISBN: 9781421407944.
Grayson, Daniel R. and Michael E. Stillman (n.d.). *Macaulay2, a software system for research in algebraic geometry*. Available at http://www.math.uiuc.edu/Macaulay2/.
Greuel, Gert-Martin, Christoph Lossen, and Eugenii Shustin (2007). *Introduction to Singularities and Deformations*. Berlin, Germany: Springer Science & Business Media. DOI: 10.1007/978-3-540-28381-2.
Griewank, Andreas and Michael R. Osborne (1981). “Newton’s method for singular problems when the dimension of the null space is > 1.” In: *SIAM Journal on Numerical Analysis* 18.1, pp. 145–149. DOI: 10.1137/0718011.
Gupta, Garv (2013). “The facial structure of the cone of positive semidefinite matrices.” PhD thesis. Iowa City, IA: University of Iowa.
Harris, Lawrence A. (1984). “Symmetric semidefinite matrices with rational entries.” In: *Linear Algebra and its Applications* 58, pp. 103–107. DOI: 10.1016/0024-3795(84)90205-4.
Harrison, John (2007). “Verifying nonlinear real formulas via sums of squares.” In: *Theorem Proving in Higher Order Logics*. Springer, pp. 102–118. DOI: 10.1007/978-3-540-74591-4_9.
Hauenstein, Jonathan D., Alan C. Liddell Jr., Sanesha McPherson, and Yi Zhang (2021). “Numerical algebraic geometry and semidefinite programming.” In: *Results in Applied Mathematics* 11, p. 100166. DOI: 10.1016/j.rinam.2021.100166.
Hauenstein, Jonathan D., Bernard Mourrain, and Agnes Szanto (2015). “Certifying isolated singular points and their multiplicity structure.” In: *Proceedings of the 2015 International Symposium on Symbolic and Algebraic Computation*. New York, NY, USA: ACM, pp. 213–220. DOI: 10.1145/2755996.2756651.

Helmberg, Christoph (2000). "Semidefinite Programming for Combinatorial Optimization." In: *ZIB-Report* 00-34. Habilitationsschrift, TU Berlin.
Henrion, Didier, Milan Korda, and Jean Bernard Lasserre (2020). *The Moment-SOS Hierarchy: Lectures In Probability, Statistics, Computational Geometry, Control And Nonlinear PDEs*. Vol. 4. Singapore: World Scientific.
Henrion, Didier, Simone Naldi, and Mohab Safey El Din (2015a). "Real Root Finding for Determinantal Varieties and SDP Relaxation." In: *Proceedings of the 2015 ACM International Symposium on Symbolic and Algebraic Computation*. ISSAC '15. New York, NY, USA: ACM, pp. 205–212. DOI: 10.1145/2755996.2756645.
— (2015b). "Real root finding for determinants of linear matrices." In: *Journal of Symbolic Computation* 74, pp. 205–238. DOI: 10.1016/j.jsc.2015.06.010.
— (2016). "Exact algorithms for linear matrix inequalities." In: *SIAM Journal on Optimization* 26.4, pp. 2512–2539. DOI: 10.1137/15M1031168.
— (2019). "SPECTRA: a Maple library for solving linear matrix inequalities in exact arithmetic." In: *Optimization Methods and Software* 34.1, pp. 62–78. DOI: 10.1080/10556788.2017.1352011.
— (2021). "Exact algorithms for semidefinite programs with degenerate feasible set." In: *Journal of Symbolic Computation* 104, pp. 942–959. DOI: 10.1016/j.jsc.2020.09.011.
Higham, Nicholas J. (2002). *Accuracy and Stability of Numerical Algorithms*. 2nd. Philadelphia, PA, USA: SIAM. DOI: 10.1137/1.9780898718027.
Huber, Birkett and Bernd Sturmfels (1995). "A Polyhedral Method for Solving Sparse Polynomial Systems." In: *Mathematics of Computation* 64.212, pp. 1541–1555. DOI: 10.2307/2153370.
Huber, Birkett and Jan Verschelde (1998). "Polyhedral endgames for polynomial continuation." In: *Numerical Algorithms* 18.1, pp. 91–108. DOI: 10.1023/A:1019134102607.
Kaltofen, Erich L., Bin Li, Zhengfeng Yang, and Lihong Zhi (2012). "Exact certification in global polynomial optimization via sums-of-squares of rational functions with rational coefficients." In: *Journal of Symbolic Computation* 47.1, pp. 1–15. DOI: 10.1016/j.jsc.2011.08.002.
Karapetyants, Mikhail, Vladimir Kolmogorov, and Jeferson Zapata (2026). *Computing singular solutions of polynomial systems*. In preparation.
Keller, Herbert B. (1977). "Numerical solution of bifurcation and nonlinear eigenvalue problems." In: *Applications of Bifurcation Theory*. Ed. by Paul H. Rabinowitz. New York, NY, USA: Academic Press, pp. 359–384. DOI: 10.1016/B978-0-12-574250-4.50015-3.
Klep, Igor and Markus Schweighofer (2013). "An exact duality theory for semidefinite programming based on sums of squares." In: *Mathematics of Operations Research* 38.3, pp. 569–590. DOI: 10.1287/moor.1120.0573.
Klerk, Etienne de (2002). *Aspects of Semidefinite Programming: Interior Point Algorithms and Selected Applications*. Vol. 65. Applied Optimization. Dordrecht: Kluwer Academic Publishers. DOI: 10.1007/b116127.

Kolmogorov, Vladimir, Simone Naldi, and Jeferson Zapata (2025). "Certifying Solutions of Degenerate Semidefinite Programs." In: *SIAM Journal on Optimization* 35.3, pp. 1630–1654. DOI: 10.1137/24M1664691.
Lasserre, Jean Bernard (2001). "Global optimization with polynomials and the problem of moments." In: *SIAM Journal on Optimization* 11.3, pp. 796–817. DOI: 10.1137/S1052623499362230X.
Laurent, Monique and Frank Vallentin (2020). *A Course on Semidefinite Optimization: Draft Lecture Notes*. Lecture notes, Spring 2020, Centrum Wiskunde & Informatica and University of Cologne. Amsterdam, The Netherlands; Cologne, Germany.
Lecerf, Grégoire (2002). "Quadratic Newton iteration for systems with multiplicity." In: *Journal of Symbolic Computation* 33.5, pp. 747–794. DOI: 10.1006/jsco.2002.0528.
Leykin, Anton, Jan Verschelde, and Ailing Zhao (2006). "Newton's method with deflation for isolated singularities of polynomial systems." In: *Theoretical Computer Science* 359.1, pp. 111–122. DOI: 10.1016/j.tcs.2006.02.016.
— (2008). "Higher-Order Deflation for Polynomial Systems with Isolated Singular Solutions." In: *Algorithms in Algebraic Geometry*. Vol. 146. The IMA Volumes in Mathematics and its Applications. New York, NY, USA: Springer, pp. 79–97. DOI: 10.1007/978-0-387-75155-9_5.
Li, Nan and Lihong Zhi (2012). "Computing the multiplicity structure of an isolated singular solution: Case of breadth one." In: *Journal of Symbolic Computation* 47.6, pp. 700–710. DOI: 10.1016/j.jsc.2011.12.027.
— (2022). "Improved two-step Newton's method for computing simple multiple zeros of polynomial systems." In: *Numerical Algorithms* 91.1, pp. 19–50. DOI: 10.1007/s11075-021-01254-4.
Łojasiewicz, Stanisław (1959). "Sur le problème de la division." In: *Studia Mathematica* 18.1, pp. 87–136. DOI: 10.4064/sm-18-1-87-136.
Lourenço, Bruno F. and Gábor Pataki (2022). "A simplified treatment of Ramana's exact dual for semidefinite programming." In: *Optimization Letters* 17.2, pp. 219–243. DOI: 10.1007/s11590-022-01861-1.
Monniaux, David and Pierre Corbineau (2011). "On the generation of Positivstellensatz witnesses in degenerate cases." In: *International Conference on Interactive Theorem Proving*. Springer, pp. 249–264. DOI: 10.1007/978-3-642-22863-6_19.
Morgan, Alexander P., Andrew J. Sommese, and Charles W. Wampler (1992a). "A power series method for computing singular solutions to nonlinear analytic systems." In: *Numerische Mathematik* 63.1, pp. 391–409. DOI: 10.1007/BF01385817.
— (1992b). "Computing singular solutions to polynomial systems." In: *Advances in Applied Mathematics* 13.3, pp. 305–327. DOI: 10.1016/0196-8858(92)90014-N.
Naldi, Simone (2015). "Exact algorithms for real algebraic systems and semi-definite programming." PhD thesis. Université de Toulouse. URL: https://theses.hal.science/tel-01212502.
— (2016). "A Semi-Algebraic Proof of the First Integral Method." In: *Proceedings of the 2016 ACM International Symposium on Symbolic and*

Algebraic Computation. ISSAC '16. ACM, pp. 365–372. DOI: 10.1145/2930889.2930910.
Naldi, Simone and Rainer Sinn (2020). "Conic programming: infeasibility certificates and projective geometry." In: *Journal of Pure and Applied Algebra* 224.11, p. 106605. DOI: 10.1016/j.jpaa.2020.106605.
Nie, Jiawang, Kristian Ranestad, and Bernd Sturmfels (2010). "The algebraic degree of semidefinite programming." In: *Mathematical Programming* 122.2, pp. 379–405. DOI: 10.1007/s10107-008-0253-6.
Ojika, Takeo, Satoshi Watanabe, and Taketomo Mitsui (1983). "Deflation algorithm for the multiple roots of a system of nonlinear equations." In: *Journal of Mathematical Analysis and Applications* 96.2, pp. 463–479. DOI: 10.1016/0022-247X(83)90204-0.
Pan, C.-T. (2000). "On the existence and computation of rank-revealing LU factorizations." In: *Linear Algebra and its Applications* 316.1-3, pp. 199–222. DOI: 10.1016/S0024-3795(00)00150-1.
Papachristodoulou, Antonis and Stephen Prajna (2005). "A Tutorial on Sum of Squares Techniques for Systems Analysis." In: *American Control Conference*, pp. 2686–2700. DOI: 10.1109/ACC.2005.1470373.
Pataki, Gábor (2013). "Strong Duality in Conic Linear Programming: Facial Reduction and Extended Duals." In: *Computational and Analytical Mathematics*. Ed. by D. Bailey, H. H. Bauschke, F. Garvan, M. Théra, J. D. Vanderwerff, and H. Wolkowicz. Springer, pp. 613–634. DOI: 10.1007/978-1-4614-7621-4_33.
— (2017). "Bad Semidefinite Programs: They All Look the Same." In: *SIAM Journal on Optimization* 27.1, pp. 146–172. DOI: 10.1137/140995168.
— (2020). "On Positive Duality Gaps in Semidefinite Programming." In: *arXiv preprint*, pp. 1–30. arXiv: 1812.11796.
Permenter, Frank, Henrik A. Friberg, and Erling D. Andersen (2017). "Solving Conic Optimization Problems via Self-Dual Embedding and Facial Reduction: A Unified Approach." In: *SIAM Journal on Optimization* 27.3, pp. 1257–1282. DOI: 10.1137/16M1074712.
Permenter, Frank and Pablo A. Parrilo (2018). "Partial facial reduction: simplified, equivalent SDPs via approximations of the PSD cone." In: *Mathematical Programming* 171.1, pp. 1–54. DOI: 10.1007/s10107-017-1174-x.
Peyrl, Hannes and Pablo A. Parrilo (2007). "Computing Sum of Squares Decompositions with Rational Coefficients." In: *Symbolic-Numeric Computation*. Ed. by Dongming Wang and Lihong Zhi. Birkhäuser Basel, pp. 117–132. DOI: 10.1007/978-3-7643-7984-1_7.
Platzer, André (2009). "Differential-Algebraic Proofs for Polynomial Hybrid Systems." In: *Automated Deduction – CADE-22*. Ed. by Renate A. Schmidt. Springer Berlin Heidelberg, pp. 446–462. DOI: 10.1007/978-3-642-02959-2_30.
Riks, E. (1972). "The application of Newton's method to the problem of elastic stability." In: *Journal of Applied Mechanics* 39.4, pp. 1060–1065. DOI: 10.1115/1.3422829.
Roux, Pierre, Yuen-Lam Voronin, and Sriram Sankaranarayanan (2018). "Validating numerical semidefinite programming solvers for polynomial

invariants." In: *Formal Methods in System Design* 53.2, pp. 286–312. DOI: 10.1007/s10703-018-0322-8.
Safey El Din, Mohab and Éric Schost (2003). "Polar Varieties and Computation of one Point in each Connected Component of a Smooth Real Algebraic Set." In: *Proceedings of the 2003 International Symposium on Symbolic and Algebraic Computation (ISSAC'03)*. ACM, pp. 224–231. DOI: 10.1145/860854.860901.
Schork, Lukas and Jacek Gondzio (2020). "Rank revealing Gaussian elimination by the maximum volume concept." In: *Linear Algebra and its Applications* 592, pp. 1–19. DOI: 10.1016/j.laa.2019.12.012.
Sommese, Andrew J. and Charles W. Wampler (2005). *The Numerical Solution of Systems of Polynomials Arising in Engineering and Science*. Singapore: World Scientific. DOI: 10.1142/5728.
Stewart, Gilbert W. and Ji-guang Sun (1990). *Matrix Perturbation Theory*. New York, NY, USA: Academic Press.
Waki, Hayato and Masakazu Muramatsu (2013). "A facial reduction algorithm for semidefinite programming." In: *Operations Research Letters* 41.6, pp. 584–589. DOI: 10.1016/j.orl.2013.08.006.
Wall, Charles Terence Clegg (2004). *Singular Points of Plane Curves*. Vol. 63. London Mathematical Society Student Texts. Cambridge, UK: Cambridge University Press. DOI: 10.1017/CBO9780511617560.
Wempner, George A. (1971). "Discrete approximations related to nonlinear theories of solids." In: *International Journal of Solids and Structures* 7.11, pp. 1581–1599. DOI: 10.1016/0020-7683(71)90038-2.
Zariski, Oscar (1932). "On the topology of algebroid singularities." In: *American Journal of Mathematics* 54.2, pp. 453–465. DOI: 10.2307/2370887.
Zariski, Oscar and Pierre Samuel (1965). *Commutative Algebra, Volume II*. New York, NY, USA: Springer-Verlag. DOI: 10.1007/978-3-662-22830-2.
Zhu, Yuzixuan, Gábor Pataki, and Quoc Tran-Dinh (2019). "Sieve-SDP: a simple facial reduction algorithm to preprocess semidefinite programs." In: *Mathematical Programming Computation* 11.3, pp. 503–586. DOI: 10.1007/s12532-019-00156-5.

Appendix A

Declaration of the use of Generative AI and AI tools

In this thesis, generative AI tools, specifically Google Gemini, were utilized to support the research process by assisting with the refinement of MATLAB code for numerical path-tracking experiments, aiding in data visualization, and formatting text in LaTeX. Gemini also played a role in refining the academic writing style by paraphrasing complex mathematical concepts for improved clarity. The author is fully aware of the potential limitations of AI systems, including the risk of hallucinated citations or imprecise mathematical statements. To mitigate these risks, all AI-assisted text, theoretical explanations, and code were rigorously reviewed, manually edited, and independently verified against established academic literature. The author takes full responsibility and accountability for all content submitted in this thesis, regardless of the involvement of Generative AI tools, ensuring that it strictly meets established scientific and academic standards.