Computing Nearby Non-trivial Smith Forms

Mark Giesbrecht
Cheriton School of Computer Science
University of Waterloo
mwg@uwaterloo.ca

Joseph Haraldson
Cheriton School of Computer Science
University of Waterloo
jharalds@uwaterloo.ca

George Labahn
Cheriton School of Computer Science
University of Waterloo
glabahn@uwaterloo.ca

ABSTRACT

We consider the problem of computing the nearest matrix polynomial with a non-trivial Smith Normal Form. We show that computing the Smith form of a matrix polynomial is amenable to numeric computation as an optimization problem. Furthermore, we describe an effective optimization technique to find a nearby matrix polynomial with a non-trivial Smith form. The results are later generalized to include the computation of a matrix polynomial having a maximum specified number of ones in the Smith Form (i.e., with a maximum specified McCoy rank).

We discuss the geometry and existence of solutions and how our results can used for a backwards error analysis. We develop an optimization-based approach and demonstrate an iterative numerical method for computing a nearby matrix polynomial with the desired spectral properties. We also describe the implementation of our algorithms and demonstrate the robustness with examples in `Maple`.

ACM Reference Format:
Mark Giesbrecht, Joseph Haraldson, and George Labahn. 2018. Computing Nearby Non-trivial Smith Forms. In *ISSAC '18: 2018 ACM International Symposium on Symbolic and Algebraic Computation, July 16–19, 2018, New York, NY, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3208976.3209024

1 INTRODUCTION

Matrix polynomials appear in many areas of computational algebra, control systems theory, differential equations and mechanics. The algebra of matrix polynomials is typically described assuming that the coefficients are from the field of real or complex numbers. However, in some applications, coefficients can come from measured data or contain some amount of uncertainty. As such, arithmetic may contain numerical errors and algorithms are prone to numerical instability.

One problem of computational importance is finding the Smith Normal Form (SNF, or simply Smith form) of a matrix polynomial. Given $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$, the Smith form $\mathcal{S}$ of $\mathcal{A}$ is a matrix polynomial

$$\mathcal{S} = \begin{pmatrix} s_1 & & & \\ & s_2 & & \\ & & \ddots & \\ & & & s_n \end{pmatrix} \in \mathbb{R}[t]^{n\times n},$$

where $s_1, \ldots, s_n$ are monic and $s_i \mid s_{i+1}$ for $1 \le i < n$, such that there exist unimodular $\mathcal{U}, \mathcal{V} \in \mathbb{R}[t]^{n\times n}$ (i.e., with determinants in $\mathbb{R}^*$) with $\mathcal{S} = \mathcal{U}\mathcal{A}\mathcal{V}$. The Smith form always exists and is unique though the matrices $\mathcal{U}$, $\mathcal{V}$ are not unique [14, 18]. The diagonal entries $s_1, \ldots, s_n$ are referred to as the *invariant factors* of $\mathcal{A}$.

The Smith form is important as it reveals the structure of the polynomial lattice of rows and columns, as well as the effects of localizing at individual eigenvalues. That is, it characterizes how the rank decreases as the variable t is set to various eigenvalues. The form is closely related to the more general *Smith-McMillan form* for matrices of rational functions, a form that reveals the structure of eigenvalues at infinity.

In an exact setting, computing the Smith form has been well studied and very efficient procedures are available (see [19] and the references therein). However, in the case that coefficients contain uncertainties, the problem is much less understood. Numerical methods to compute the Smith form of a matrix polynomial typically rely on linearization and orthogonal transformations [3, 6, 24] to infer the Smith form of a nearby matrix polynomial via the Jordan blocks in the Kronecker canonical form (see [18]). These linearization techniques are backwards stable, and for many problems this is sufficient to ensure that the computed solutions are computationally useful when a problem is continuous. However, the eigenvalues of a matrix polynomial are not necessarily continuous functions of the coefficients of the matrix polynomial, and backwards stability is not always sufficient to ensure computed solutions are useful in the presence of discontinuities. These methods are also unstructured in the sense that the computed non-trivial Smith form may not be the Smith form of a matrix polynomial with a prescribed coefficient structure. In extreme instances, the unstructured backwards error can be arbitrarily small, while the structured distance to an interesting Smith form is relatively large. This is often seen in problems with prescribed sparsity patterns or zero-coefficients. Numerical methods can also fail to compute meaningful results on some problems due to uncertainties. Examples of such problems include nearly rank deficient matrix polynomials, repeated eigenvalues or eigenvalues that are close together and other ill-posed instances. The above issues are largely resolved by our optimization-based approach, though at a somewhat higher computational cost.

ISSAC '18, July 16–19, 2018, New York, NY, USA

ACM ISBN 978-1-4503-5550-6/18/07…$15.00
https://doi.org/10.1145/3208976.3209024

The invariant factors $s_1, \ldots, s_n$ of a matrix $A \in \mathbb{R}[t]^{n\times n}$ can also be defined via the *determinantal divisors* $\delta_1, \ldots, \delta_n \in \mathbb{R}[t]$, where

$$\delta_i = \text{GCD}\left\{\text{all } i \times i \text{ minors of } \mathcal{A}\right\} \in \mathbb{R}[t].$$

Then $s_1 = \delta_1$ and $s_i = \delta_i/\delta_{i-1}$ for $2 \le i \le n$ (and $\delta_n = \det(\mathcal{A})$). In the case of 2×2 matrix polynomials, computing the nearest non-trivial Smith form is thus equivalent to finding the nearest matrix polynomial whose polynomial entries have a non-trivial GCD. This points to a significant difficulty: approximate GCD problems can have infima that are *unattainable*. That is, there are co-prime polynomials with nearby polynomials with a non-trivial GCD at distances arbitrarily approaching an infimum, while at the infimum itself the GCD is trivial (see, e.g., [11]). This issue extends to Smith forms as is seen in the following example.

Example 1.1. Let $f = t^2 - 2t + 1$ and $g = t^2 + 2t + 2$. We first seek $\widetilde{f}, \widetilde{g} \in \mathbb{R}[t]$ of degree at most 2 such that $\gcd(\widetilde{f}, \widetilde{g}) = \gamma t + 1$ at minimal distance $\|f - \widetilde{f}\|_2^2 + \|g - \widetilde{g}\|_2^2$ for some $\gamma \in \mathbb{R}$. Using the approach of Karmarkar & Lakshman [20] it is shown [16, Example 3.3.6] that this distance is $(5\gamma^4 - 4\gamma^3 + 14\gamma^2 + 2)/(\gamma^4 + \gamma^2 + 1)$. This distance has an infimum of 2 at $\gamma = 0$. However, at $\gamma = 0$ we have $\gcd(\widetilde{f}, \widetilde{g}) = 1$ even though $\deg \gcd(\widetilde{f}, \widetilde{g}) > 0$ for all $\gamma \neq 0$.

Now consider the matrix $\mathcal{A} = \text{diag}(f, g) \in \mathbb{R}[t]^{2\times 2}$. For $\mathcal{A}$ to have a non-trivial Smith form we must perturb f, g such that they have a non-trivial GCD, and thus any such perturbation must be at a distance of at least 2. However, the perturbation of distance precisely 2 has a trivial Smith form. There is clearly no merit to perturbing the off-diagonal entries of $\mathcal{A}$.

Our work indirectly involves measuring the sensitivity to the eigenvalues of $\mathcal{A}$ and the determinant of $\mathcal{A}$. Thus we differ from most sensitivity and perturbation analysis [1, 23] since we also study how perturbations affect the invariant factors, instead of the roots of the determinant. Additionally our theory is able to support the instance of $\mathcal{A}$ being rank deficient and having degree exceeding one. One may also approach the problem geometrically in the context of manifolds [7, 8]. We do not consider the manifold approach directly since it does not yield numerical algorithms.

We address two fundamental questions in this paper: (1) what does it mean for a matrix polynomial $\mathcal{A}$ to have a non-trivial Smith form numerically and (2) how far is $\mathcal{A}$ from another matrix polynomial with an interesting or non-trivial Smith form?

We formulate the answers to these questions as solutions to continuous optimization problems. The main contributions of this paper are deciding when $\mathcal{A}$ has an interesting Smith form, providing bounds on a "radius of triviality" around $\mathcal{A}$ and a structured stability analysis on iterative methods to compute a structured matrix polynomial with desired spectral properties.

The remainder of the paper is organized as follows. In Section 2 we give the notation and terminology along with some needed background used in our work. Section 3 discusses the approximate Smith form computation as an optimization problem and provide some new bounds on the distance to non-triviality. We present an optimization algorithm in Section 4 with local stability properties and rapid local convergence to compute a nearby matrix polynomial with a non-trivial Smith form and discuss implementation details. A method to compute a matrix polynomial with a prescribed lower bound on the number of ones is discussed in Section 5. The paper ends with a discussion of our implementation and examples.

2 PRELIMINARIES

In this section we explore the topology of the approximate Smith normal form and discuss basic results concerning the notion of both *attainable* and *unattainable* solutions.

We make extensive use of the following terminology and definitions. A matrix polynomial $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$ is an $n \times n$ matrix whose entries consist of polynomials of degree at most d. Alternatively, we may express matrix polynomials as $\mathcal{A} = \sum_{1\le j\le d} A_j t^j$ where $A_j \in \mathbb{R}^{n\times n}$. The *degree* of a matrix polynomial d is defined to be the degree of the highest-order non-zero entry of $\mathcal{A}$, or the largest index j such that $A_j \neq 0$. We say that $\mathcal{A}$ has *full rank* or is *regular* if $\det(\mathcal{A}) \not\equiv 0$ and that $\mathcal{A}$ is *unimodular* if $\det(\mathcal{A}) \in \mathbb{R}\backslash\{0\}$. The *(finite) eigenvalues* are the roots of $\det(A) \in \mathbb{R}[t]$.

We define the norm of a polynomial $a \in \mathbb{R}[t]$ as $\|a\| = \|a\|_2 = \|(a_0, a_1, \ldots, a_d, 0, \ldots, 0)\|_2$ and for matrix polynomials we define $\|\mathcal{A}\| = \|\mathcal{A}\|_F = \sqrt{\sum_{i,j} \|\mathcal{A}_{i,j}\|_2^2}$. Our choice of norm is a distributed coefficient norm, sometimes known as the Frobenius norm.

Definition 2.1 (SVD [15]). The Singular Value Decomposition (SVD) of $A \in \mathbb{R}^{n\times n}$ is given by $U^T \Sigma V$, where $U, V \in \mathbb{R}^{n\times n}$ satisfy $U^T U = I$, $V^T V = I$ and $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_n)$ is a diagonal matrix with non-negative entries of the singular values of A in descending order. The distance to the nearest (unstructured) matrix of rank $m < n$ is $\sigma_{m+1}(A)$.

For scalar matrices we frequently write $\|\cdot\|_2$ for the largest singular value, and $\sigma_{\min}(\cdot)$ for the smallest singular value.

Definition 2.2 (Affine/Linear Structure). A matrix polynomial $\mathcal{A} \in \mathbb{R}[t]^{n\times n}\backslash\{0\}$ of degree at most d has a *linear structure* from the set $\mathcal{K}$ if $\mathcal{A} \in \text{span}(\mathcal{K})$ as a vector space over $\mathbb{R}$, where

$$\mathcal{K} = \{C_{0,0}, \ldots, C_{0,k}, tC_{1,0}, \ldots, tC_{1,l}, \ldots, t^d C_{d,0}, \ldots, t^d C_{d,k}\},$$

with $C_{l,j} \in \mathbb{R}^{n\times n}$ and the vectors in $\mathcal{K}$ are linearly independent. If $\mathcal{A} = \mathcal{C}_0 + \mathcal{C}_1$ where $\mathcal{C}_0 \in \mathbb{R}[t]^{n\times n}\backslash\{0\}$ and $\mathcal{C}_1 \in \text{span}(\mathcal{K})$, then the structure is said to be *affine*.

Examples of matrix polynomials with a linear structure include symmetric matrices, matrices with prescribed zero coefficients, prescribed zero entries, tri-diagonal matrices and several other classes. Affinely structured matrix polynomials include monic matrix polynomials, matrix polynomials with prescribed constant coefficients and banded matrix polynomials are a few of many possible. In this paper we are mainly concerned with preserving the zero structure of a matrix polynomial, that is we do not change zero-coefficients or increase the degree of entries.

The *rank* of a matrix polynomial is the maximum number of linearly independent rows or columns as a vector space over $\mathbb{R}(t)$. This is the rank of the matrix $\mathcal{A}(\omega)$ for any $\omega \in \mathbb{C}$ except when ω is an eigenvalue of $\mathcal{A}(t)$. The McCoy rank of $\mathcal{A}$ is $\min_{\omega\in\mathbb{C}}\{\text{rank}\,\mathcal{A}(\omega)\}$, which is the lowest rank when $\mathcal{A}$ is evaluated at an eigenvalues. The McCoy rank is also the number of ones in the Smith form, or equivalently, if $\mathcal{A}$ has m non-trivial invariant factors, then the McCoy rank of $\mathcal{A}$ is $n - m$. The matrix polynomial $\mathcal{A}$ is said to have a *non-trivial Smith form* if the McCoy rank is at most $n - 2$, or equivalently, has two or more invariant factors of positive degree.

Problem 2.3 (Approximate SNF and Low McCoy Rank). *Given a matrix polynomial $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$, find the distance to a non-trivial SNF and, if possible, a matrix polynomial $\widehat{\mathcal{A}} \in \mathbb{R}[t]^{n\times n}$ of prescribed coefficient structure that has a prescribed McCoy rank of $n-m$ for $m \geq 2$ such that $\|\mathcal{A} - \widehat{\mathcal{A}}\|$ is minimized under $\|\cdot\|$. If such $\widehat{\mathcal{A}}$ exists, then the Smith form of $\widehat{\mathcal{A}}$ is the approximate Smith form of A if $m = 2$ and a lower McCoy rank approximation if $m > 2$.*

As described in Section 1, it is possible that the distance to a non-trivial SNF is not attainable. That is, there is a solution that is approached asymptotically, but where the Smith form is trivial at the infimum. Fortunately, in most instances of interest, solutions will generally be attainable, and we will later discuss how to identify and compute unattainable solutions. This problem admits the nearest rank deficient matrix polynomial as a special case (see [12, 13]). However the computational challenges are fundamentally different for non-trivial instances.

2.1 Basic Results

In this subsection we review some basic results needed to analyze the topology of the approximate Smith form problem. We introduce the notion of a generalized Sylvester matrix, drawing on the theory of resultants.

Definition 2.4. The *adjoint* of a matrix polynomial is the $n \times n$ matrix $\text{Adj}(\mathcal{A})$ that satisfies $\text{Adj}(\mathcal{A})\mathcal{A} = \mathcal{A}\,\text{Adj}(\mathcal{A}) = \det(\mathcal{A})I$. The entries of $\text{Adj}(\mathcal{A})$ are the $(n-1)\times(n-1)$ minors of $\mathcal{A}$ up to a multiple of ± 1.

As $s_n = \delta_n/\delta_{n-1}$ it follows that $\mathcal{A}$ has a non-trivial Smith form if and only if the GCD of all entries of the adjoint is non-trivial, that is, $\deg\gcd(\text{Adj}(\mathcal{A})) \geq 1$. In order to obtain bounds on the distance to a matrix having a non-trivial Smith form, we consider an approximate GCD problem of the form

$$\min\left\{\|\Delta\mathcal{A}\| \text{ such that } \deg\gcd\left\{\text{Adj}\,(\mathcal{A}+\Delta\mathcal{A})_{ij}\right\} \geq 1\right\}.$$

If this was a classical approximate GCD problem, then the use of Sylvester-like matrices would be sufficient. However, in our problem the degrees of the entries of the adjoint may change under perturbations. In order to perform an analysis, we need to study a family of generalized Sylvester matrices that allow higher-degree zero coefficients to be perturbed.

For $a = \sum_{i=0}^{d} a_i t^i \in \mathbb{R}[t]$ of degree at-most d, we define the r^{th} *convolution matrix* of a as

$$\phi_r(a) = \begin{pmatrix} a_0 & \cdots & a_d & & \\ & \ddots & & \ddots & \\ & & a_0 & \cdots & a_d \end{pmatrix} \in \mathbb{R}^{r\times(r+d)}.$$

Let $\mathbf{f} = (f_1, \ldots, f_k) \in \mathbb{R}[t]$ be a vector of polynomials with degrees $\mathbf{d} = (d_1, \ldots, d_k)$ ordered as $d_j \leq d_{j+1}$ for $1 \leq j \neq k-1$ (with $\deg 0 = -\infty$). Set $d = d_1$ and $r = \max(d_2, \ldots, d_k)$ and suppose that for each $2 \leq i \leq k$ f_i is viewed as a polynomial of degree at most r. Then we define the *generalized Sylvester matrix* of f as

$$\text{Syl}(\mathbf{f}) = \text{Syl}_{\mathbf{d}}(\mathbf{f}) = \begin{pmatrix} \phi_r(f_1) \\ \phi_d(f_2) \\ \vdots \\ \phi_d(f_k) \end{pmatrix} \in \mathbb{R}^{(r+(k-1)d)\times(r+d)}.$$

Some authors [10, 25] refer to such a matrix as an expanded Sylvester matrix or generalized resultant matrix. The generalized Sylvester matrix has many useful properties pertaining to the Bézout coefficients. However we are only concerned with the well known result that $\gcd(\mathbf{f}) = 1$ if and only if $\text{Syl}_{\mathbf{d}}(\mathbf{f})$ has full rank.

It will be useful to treat a polynomial of degree d as one of larger degree. This can be accomplished by constructing a similar matrix and padding rows and columns with zero entries. The generalized Sylvester matrix of degree at most $\mathbf{d}' \geq \mathbf{d}$ of $\mathbf{f}$ is defined analogously as $\text{Syl}_{\mathbf{d}'}(\mathbf{f})$, taking d to be the largest degree entry and r to be the largest degree of the remaining entries of $\mathbf{d}'$. Note that $r = d$ is possible and typical. If $\mathbf{f}$ has a non-trivial GCD (possibly unattainable) under a perturbation structure $\Delta\mathbf{f}$, then it is necessary that $\text{Syl}_{\mathbf{d}'}(\mathbf{f})$ is rank deficient, and often this will be sufficient.

If we view the entries of $\mathbf{f}$ of polynomials of degree $\mathbf{d}'$ and $\mathbf{d}'_i > \mathbf{d}_i$ for all i, then the entries of $\mathbf{f}$ has an unattainable GCD of distance zero, typically of the form $1 + \varepsilon t \sim t + \varepsilon^{-1}$. In other words, the underlying approximate GCD problem is ill-posed.

Lemma 2.5. *If* $\max(\mathbf{d}) = \max(\mathbf{d}')$ *then the kernels of* $\text{Syl}_{\mathbf{d}}(\mathbf{f})$ *and* $\text{Syl}_{\mathbf{d}'}(\mathbf{f})$ *have the same dimension.*

Proof. Let d and r be the largest and second largest entries of $\mathbf{d}$ and r' be the second largest entry of $\mathbf{d}'$. The result follows from [25] by considering the case of $r' = d$. □

This lemma characterizes the (generic) case when elements of maximal degree of $\mathbf{f}$ do not change under perturbations, then the generalized Sylvester matrix still meaningfully encodes GCD information. However, it is possible that $\text{Syl}_{\mathbf{d}}(\mathbf{f})$ has full rank and $\text{Syl}_{\mathbf{d}'}(\mathbf{f})$ is rank deficient but the distance to a non-trivial gcd is not zero. This can occur when $\mathbf{d}_j = \mathbf{d}'_j$ for some j and $\mathbf{d}' \geq \mathbf{d}$.

Definition 2.6. The *degree d reversal* of an $f \in \mathbb{R}[t]$ of degree at most d is defined as $\text{rev}_d(f) = t^d f(t^{-1})$. For a vector of polynomials $\mathbf{f} \in \mathbb{R}[t]^\ell$ of degrees at most $\mathbf{d} = (d_1, \ldots, d_\ell)$ the *degree* $\mathbf{d}$ *reversal* of $\mathbf{f}$ is the vector $\text{rev}_{\mathbf{d}}(\mathbf{f}) = (\text{rev}_{d_1}(f_1), \ldots, \text{rev}_{d_\ell}(f_\ell))$.

The following lemma enables us to determine if unattainable solutions are occurring in an approximate GCD problem with an arbitrary (possibly non-linear) structure on the coefficients.

Lemma 2.7. *Let $\mathbf{f}$ be a vector of non-zero polynomials of degree at most d. Suppose that $\text{Syl}_{\mathbf{d}}(\mathbf{f})$ has full rank and $\text{Syl}_{\mathbf{d}'}(\mathbf{f})$ is rank deficient, where the perturbations $\Delta\mathbf{f}$ have degrees at most $\mathbf{d}'$ and the entries of $\mathbf{f}$ have degrees $\mathbf{d}$. Then $\mathbf{f}$ has an unattainable non-trivial GCD of distance zero under the perturbation structure $\Delta\mathbf{f}$ if and only if $\text{Syl}(\text{rev}_{\mathbf{d}'}(\mathbf{f}))$ is rank deficient.*

Proof. First suppose that $\text{Syl}(\text{rev}_{\mathbf{d}'}(\mathbf{f}))$ has full rank. Then $\gcd(\text{rev}_{\mathbf{d}'}(\mathbf{f})) = 1$, but $\text{rev}_{\mathbf{d}}(\gcd(\mathbf{f})) = \gcd(\text{rev}_{\mathbf{d}}(\mathbf{f})) = \gcd(\text{rev}_{\mathbf{d}'}(\mathbf{f}))$. Hence $\mathbf{f}$ does not have an unattainable non-trivial GCD. Conversely, suppose that $\text{Syl}(\text{rev}_{\mathbf{d}'}(\mathbf{f}))$ is rank deficient. Then, t is a factor of $\gcd(\text{rev}_{\mathbf{d}'}(\mathbf{f}))$ but t is not a factor of $\gcd(\text{rev}_{\mathbf{d}}(\mathbf{f}))$. Accordingly, all non-zero entries of $\mathbf{f}+\Delta\mathbf{f}$ may increase in degree and so the distance of $\mathbf{f}$ having a non-trivial GCD is zero, and so is unattainable. □

If the generalized Sylvester matrix of $\mathbf{f}$ has full rank, but the generalized Sylvester matrix that encodes the perturbations $\mathbf{f} + \Delta\mathbf{f}$ is rank deficient, then either there is an unattainable solution, or the

generalized Sylvester matrix is rank deficient due to over-padding with zeros. Lemma 2.7 provides a reliable way to detect this over padding.

Definition 2.8. We say that $\mathcal{A}$ has an *unattainable non-trivial Smith form* if $\gcd(\mathrm{Adj}(\mathcal{A})) = 1$ and $\gcd(\mathrm{Adj}(\mathcal{A} + \Delta\mathcal{A})) \neq 1$ for an infinitesimal perturbation $\Delta\mathcal{A}$ of prescribed affine structure.

Example 2.9. Let $\mathcal{A} = \begin{pmatrix} t & t-1 \\ t+1 & t \end{pmatrix}$. Then the 4×4 matrix polynomial $\mathcal{C} = \begin{pmatrix} \mathcal{A} & \\ & \mathcal{A} \end{pmatrix}$ has an unattainable non-trivial Smith form if all perturbations to $\mathcal{A}$ are support or degree preserving (preserve zero entries or do not increase the degree of each entry), both linear structures. Note that $\mathcal{C}$ and $\mathcal{A}$ are both unimodular. However small perturbations to the non-zero coefficients of $\mathcal{A}$ make $\mathcal{A} + \Delta\mathcal{A}$ non-unimodular.

These examples are non-generic. Generically the degree of the adjoint will be $(n-1)d$ and will remain unchanged locally under perturbations to the coefficients. We can formulate computing the distance to the nearest matrix polynomial with a non-trivial Smith form under a prescribed perturbation structure as finding the nearest rank deficient generalized Sylvester matrix of the adjoint or the $\mathbf{d}'$ reversal of the adjoint.

3 WHEN DOES A NUMERICAL MATRIX POLYNOMIAL HAVE A TRIVIAL SNF?

In this section we consider the question of determining if a matrix polynomial has a non-trivial SNF, or rather how much do the coefficients need to be perturbed to have a non-trivial SNF. We provide a lower bound on the quantity by analyzing the distance to a reduced rank generalized Sylvester matrix.

3.1 Nearest Rank Deficient Structured Generalized Sylvester Matrix

Suppose that $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$ of degree at most d has a trivial Smith form and does not have an unattainable non-trivial Smith form. Then one method to compute a lower bound on the distance the entries of $\mathcal{A}$ need to be perturbed to have an attainable or unattainable non-trivial Smith form is to solve

$$\inf \|\Delta\mathcal{A}\| \text{ such that } \begin{cases} \mathrm{rank}(\mathrm{Syl}_{\mathbf{d}'}(\mathrm{Adj}(\mathcal{A} + \Delta\mathcal{A}))) < e, \\ e = \mathrm{rank}(\mathrm{Syl}_{\mathbf{d}'}(\mathrm{Adj}(\mathcal{A}))). \end{cases} \tag{3.1}$$

Here $\mathbf{d}'$ is the vector of the largest possible degrees each entry of $\mathrm{Adj}(\mathcal{A} + \Delta\mathcal{A})$ and $\Delta\mathcal{A}$ has in a prescribed linear or affine perturbation structure.

It is sufficient to compute $\max(\mathbf{d}')$, and this quantity will generically be $(n-1)d$. For non-generic instances we require the computation of $\mathbf{d}'$. This optimization problem is non-convex, but multi-linear in each coefficient of $\Delta\mathcal{A}$.

We do not attempt to solve this problem directly via numerical techniques, since it enforces a necessary condition that is often sufficient. Instead we use it to develop a theory of solutions which can be exploited by faster and more robust numerical methods.

Lemma 3.1. *Let* $\mathbf{f}$ *be a vector of polynomials with degrees* $\mathbf{d}$ *and admissible perturbations* $\Delta\mathbf{f}$ *of degrees* $\mathbf{d}'$ *where* $\max(\mathbf{d}) \leq \max(\mathbf{d}')$. *Then the family of generalized Sylvester matrices* $\mathrm{Syl}_{\mathbf{d}'}(\mathbf{f})$ *of rank at least* e *form an open set under the perturbations* $\Delta\mathbf{f}$.

Proof. By the degree assumption on $\Delta\mathbf{f}$ we have that for an infinitesimal $\Delta\mathbf{f}$ that $\mathrm{Syl}_{\mathbf{d}'}(\mathbf{f})$ and $\mathrm{Syl}_{\mathbf{d}'}(\Delta\mathbf{f})$ have the same dimension. Accordingly, let us suppose that $\mathrm{Syl}_{\mathbf{d}'}(\mathbf{f})$ has rank at least e. Then it must have rank at least e in an open-neighborhood around it. In particular, when $\|\mathrm{Syl}_{\mathbf{d}'}(\Delta\mathbf{f})\|_2 < \sigma_e(\mathrm{Syl}_{\mathbf{d}'}(\mathbf{f}))$ then $\mathrm{rank}\,\mathrm{Syl}_{\mathbf{d}'}(\mathbf{f} + \Delta\mathbf{f}) \geq \mathrm{rank}\,\mathrm{Syl}_{\mathbf{d}'}(\mathbf{f})$ and the result follows. □

Theorem 3.2. *The optimization problem* (3.1) *has an attainable global minimum under linear perturbation structures.*

Proof. Let $\mathcal{S}$ be the set of all rank at most $e-1$ generalized Sylvester matrices of prescribed shape by $\mathbf{d}'$ and $\mathrm{Adj}(\mathcal{A})$. Lemma 3.1 implies that $\mathcal{S}$ is topologically closed.

Let $\mathcal{R} = \{\mathrm{Syl}_{\mathbf{d}'}(\mathrm{Adj}(C)) \text{ such that } \|C\| \leq \|\mathcal{A}\|\}$, where the generalized Sylvester matrices are padded with zeros to have the appropriate dimension if required. Since $\Delta\mathcal{A}$ has a linear perturbation structure, a feasible point is always $C = -\mathcal{A}$. By inspection $\mathcal{R}$ is seen to be a non-empty set that is bounded and closed.

The functional $\|\cdot\|$ is continuous over the non-empty closed and bounded set $\mathcal{S} \cap \mathcal{R}$. Let $\mathcal{B} \in \mathcal{S} \cap \mathcal{R}$. By Weierstrass's theorem $\|\mathcal{A} - B\|$ has an attainable global minimum over $\mathcal{S} \cap \mathcal{R}$. □

Note that if a feasible point exists under an affine perturbation structure, then a solution to the optimization problem exists as well. What this result says is that computing the distance to non-triviality is generally a well-posed problem, even though computing a matrix polynomial of minimum distance may be ill-posed. The same results also hold when working over the $\mathbf{d}'$ reversed coefficients.

3.2 Bounds on the Distance to non-triviality

Suppose that $\mathcal{A} \in \mathbb{R}[t]^{n\times n}$, of degree at most d, has a trivial Smith form and does not have an unattainable non-trivial Smith form. This section provides some basic bounds on the distance coefficients of $\mathcal{A}$ need to be perturbed to have a non-trivial Smith form.

If we consider the mapping $\mathrm{Adj}(\cdot)$ as a vector-valued function from $\mathbb{R}^{n^2(d+1)} \to \mathbb{R}^{n^2((n-1)d+1)}$ (with some coordinates possibly fixed to zero), then we note that the mapping is locally Lipschitz. More precisely, there exists $c > 0$ such that

$$\|\mathrm{Adj}(\mathcal{A}) - \mathrm{Adj}(\mathcal{A} + \Delta\mathcal{A})\| \leq c\|\Delta\mathcal{A}\|.$$

The quantity c can be bounded above by the (scalar) Jacobian matrix $\nabla\,\mathrm{Adj}(\cdot)$ evaluated at $\mathcal{A}$. A local upper bound for c is approximately $\|\nabla\,\mathrm{Adj}(\mathcal{A})\|_2$.

The entries of $\nabla\,\mathrm{Adj}(\mathcal{A})$ consist of the coefficients of the $(n-2)\times(n-2)$ minors of $\mathcal{A}$. This follows because $\mathrm{Adj}(\cdot)$ is a multi-linear vector mapping and the derivative of each entry is a coefficient of the leading coefficient with respect to the variable of differentiation. The size of each minor can be bounded above by Hadamard's inequality. As such, we have the sequence of bounds

$$\|\nabla\,\mathrm{Adj}(\mathcal{A})\|_2 \leq n\sqrt{d+1}\|\nabla\,\mathrm{Adj}(\mathcal{A})\|_\infty \leq n^3(d+1)^{3/2}\|\mathcal{A}\|_\infty^n n^{n/2},$$

where $\|\mathcal{A}\|_\infty$ is understood to be a vector norm and $\|\nabla\,\mathrm{Adj}(\mathcal{A})\|_\infty$ is understood to be a matrix norm. The bound in question can be used in conjunction with the SVD to obtain a lower bound on the distance to a matrix polynomial with a non-trivial Smith form.

Theorem 3.3. *Suppose that* $\mathbf{d}' = (\gamma, \gamma \ldots, \gamma)$ *and* $\mathrm{Syl}_{\mathbf{d}'}(\mathrm{Adj}(\mathcal{A}))$ *has rank* e. *Then a lower bound on the distance to non-triviality is*

$$\frac{1}{\gamma \|\nabla \operatorname{Adj}(\mathcal{A})\|_F} \sigma_e(\mathrm{Syl}_{\mathbf{d}'}(\operatorname{Adj}(\mathcal{A}))).$$

Proof. We note that for polynomials $\mathbf{f}$ with degrees $\mathbf{d}'$ that $\|\mathrm{Syl}_{\mathbf{d}'}(\mathbf{f})\| = \gamma\|\mathbf{f}\|$. Accordingly, if $\Delta\mathcal{A}$ is a minimal perturbation to non-triviality, then

$$\begin{aligned}\frac{1}{\gamma}\sigma_e(\mathrm{Syl}_{\mathbf{d}'}(\operatorname{Adj}(\mathcal{A}))) &\le \|\operatorname{Adj}(\mathcal{A}) - \operatorname{Adj}(\mathcal{A} + \Delta\mathcal{A})\|_F \\ &\le \|\nabla \operatorname{Adj}(\mathcal{A})\|_F \|\Delta\mathcal{A}\|_F,\end{aligned}$$

and the theorem follows by a simple rearrangement. □

If $\mathbf{d}'$ has different entries, then $r\|\mathbf{f}\| \le \|\mathrm{Syl}_{\mathbf{d}'}(\mathbf{f})\| \le \gamma\|\mathbf{f}\|$, where γ and r are the largest and second-largest entries of $\mathbf{d}'$. The lower bound provided can also be improved using the Karmakar-Lakshman distance [20] in lieu of the smallest singular value of the generalized Sylvester matrix, the $\mathbf{d}'$ reversal of the adjoint or other approximate GCD lower bounds [2].

4 APPROXIMATE SNF VIA OPTIMIZATION

In this section we formulate the approximate Smith form problem as the solution to a continuous constrained optimization problem. We assume that the solutions in question are attainable and develop a method with rapid local convergence. As the problem is non-convex, our convergence analysis will be local.

4.1 Constrained Optimization Formulation

An equivalent statement to $\mathcal{A}$ having a non-trivial attainable Smith form is that $\operatorname{Adj}(\mathcal{A}) = f^* h$ where f^* is a vector (matrix) of scalar polynomials and h is a divisor of $\gcd(\operatorname{Adj}(\mathcal{A}))$. This directly leads to the following optimization problem.

$$\min \|\Delta\mathcal{A}\|_F^2 \text{ where } \begin{cases} \operatorname{Adj}(\mathcal{A} + \Delta\mathcal{A}) = f^* h & f^* \in \mathbb{R}[t]^{n\times n}, h \in \mathbb{R}[t] \\ \mathcal{N}_h \operatorname{vec}(h) = 1 & \mathcal{N}_h \in \mathbb{R}^{1\times(\deg h+1)}. \end{cases} \tag{4.1}$$

This is a multi-linearly structured approximate GCD problem, which is a non-convex optimization problem. Instead of finding a rank deficient Sylvester matrix, we directly enforce that the entries of $\operatorname{Adj}(\mathcal{A})$ have a non-trivial GCD. The normalization requirement that $\mathcal{N}_h \operatorname{vec}(h) = 1$ is chosen to force h to have a non-zero degree, so that h is not a scalar. One useful normalization is to define $\mathcal{N}_h$ such that $\operatorname{lcoeff}(h) = 1$, that is, assume the degree of the approximate GCD is known and make it monic. Of course, other valid normalizations also exist.

Since we are working over $\mathbb{R}[t]$, there will always be a quadratic, linear or zero factor of attainable solutions. If we ignore the zero solution for now, then we can assume generically that $\deg h = 1$ or $\deg h = 2$. We note that if $h = 0$ then the approximate SNF of $\mathcal{A}$ is rank deficient and computing approximate SNF reduces to the nearest rank at-most $n-1$ or $n-2$ matrix polynomial problems, both of which are well-understood [12, 13]. Accordingly, for the remainder of the paper we will suppose that $h \neq 0$ is monic and that the degree is prescribed. The case of $h = 0$ is mentioned here for completeness but is not considered further.

4.2 Lagrange Multipliers and Optimality Conditions

In order to solve our problem we will employ the method of Lagrange multipliers. The Lagrangian is defined as

$$L = \|\Delta\mathcal{A}\|_F^2 + \lambda^T \begin{pmatrix} \operatorname{vec}(\operatorname{Adj}(\mathcal{A} + \Delta\mathcal{A}) - f^* h) \\ \mathcal{N}_h \operatorname{vec}(h) - 1 \end{pmatrix},$$

where $\operatorname{vec}(\cdot)$ stacks a matrix polynomial by columns into a column vector, and λ is a vector of Lagrange multipliers.

A necessary first-order condition (KKT condition, see [4]) for a tuple $z^* = z^*(\Delta\mathcal{A}, f^*, h, \lambda)$ to be a regular (attainable) minimizer is that the gradient of L vanishes, that is,

$$\nabla L(z^*) = 0. \tag{4.2}$$

Let J be the Jacobian matrix of the constraints defined as

$$J = \nabla_{\Delta\mathcal{A}, f^*, h} \left(\operatorname{vec}(\operatorname{Adj}(\mathcal{A} + \Delta\mathcal{A}) - f^* h) \right).$$

The second-order sufficiency condition for optimality at a local minimizer z^* is that

$$\ker(J(z^*))^T \nabla^2_{xx} L(z^*) \ker(J(z^*)) \succ 0, \tag{4.3}$$

or that the Hessian with respect to $x = x(\Delta\mathcal{A}, f^*, h)$ is positive definite over the kernel of the Jacobian of the constraints. The vector x corresponds to the variables in the affine structure of $\Delta\mathcal{A}$,f^*, and h. If (4.2) and (4.3) both hold, then z^* is necessarily a local minimizer of (4.1). Of course, it is also necessary that $\ker(J(z^*))^T \nabla^2_{xx} L(z^*) \ker(J(z^*)) \succeq 0$ at a minimizer, which is the second-order necessary condition. Our strategy for computing a local solution is to solve $\nabla L = 0$ by a Newton-like method.

4.3 An Implementation with Local Quadratic Convergence

A problem with Newton type methods is that when the Hessian is rank deficient or ill-conditioned, the Newton step becomes ill-defined or the rate of convergence degrades. The proposed formulation of our problem can encounter a rank deficient Hessian. Despite this we are still able to obtain a method with rapid local convergence under a very weak normalization assumption.

In order to obtain rapid convergence we make use of the Levenberg-Marquart (LM) algorithm. If $H = \nabla^2 L$, then the LM iteration is defined as repeatedly solving for $z^{(k+1)} = z^{(k)} + \Delta z^{(k)}$ by

$$(H^T H + \mu_k I)\Delta z^{(k)} = -H^T \nabla L(z^{(k)}) \text{ where } z = \begin{pmatrix} x \\ \lambda \end{pmatrix} \in \mathbb{R}^\ell,$$

for some $\ell > 0$ and using $\|\nabla L\|_2$ as a merit function. The speed of convergence depends on the choice of $\mu_k > 0$.

Fukushima and Yamashita [26] show that, under a local-error bound condition, a system of non-linear equations $g(z) = 0$ approximated by LM will converge quadratically to a solution with a suitable initial guess.

Essentially, what this says is that to obtain rapid convergence it is sufficient for regularity (J having full rank) to hold or second-order sufficiency, but it is not necessary to satisfy both. The advantage of LM over other quasi-Newton methods is that this method is globalized* in exchange for an extra matrix multiplication, as $H^T H +$

*Here "globalized" means that the method will converge to a stationary point of the merit function, not a local extrema of the problem.

$\mu_k I$ is always positive definite, and hence always a descent direction for the merit function. We make the choice of $\mu_k \approx \|g(z)\|_2$ based on the results of Fan and Yuan [9].

Definition 4.1 (Local Error Bound). Let Z^* be the set of all solutions to $g(z) = 0$ and X be a subset of $\mathbb{R}^\ell$ such that $X \cap Z^* \neq \emptyset$. We say that $\|g(z)\|$ provides a local error bound on $g(z) = 0$ if there exists a positive constant c such that $c \cdot \text{dist}(z, Z^*) \leq \|g(z)\|$ for all $z \in X$, where $\text{dist}(\cdot)$ is the distance between a point and a set.

Theorem 4.2. *If the second-order sufficiency condition* (4.3) *holds at an attainable solution to* (4.1)*, then the local error-bound property holds.*

Note that this result applies to all equality constrained optimization problems, and not just our specific problem.

Proof. Let $z = z(x, \lambda)$ and define $g(z) = \nabla L(z)$. First suppose that both the second-order sufficiency condition (4.3) and first-order necessary condition (4.2) hold at the point z^*. We can write the first-order expansion

$$g(z^* + \Delta z) = H(z^*)(\Delta z) + O(\|\Delta z\|_2^2) \approx H(z^*)(\Delta z),$$

noting that $g(z^*) = 0$. Next, we note that $g(x + \Delta x, \lambda + \Delta\lambda) = g(x + \Delta x, \lambda) + g(x, \lambda + \Delta\lambda)$, since g is linear in λ. It is useful to observe that

$$H(z^*) = \begin{pmatrix} H_{xx}(z^*) & J^T(z^*) \\ J(z^*) & \end{pmatrix}.$$

If $\Delta x = 0$ then the error-bound from Hoffman [17] applies and we have that there exists $c_{hof} > 0$ such that $c_{hof}\|\Delta\lambda\| \leq \|g(x, \lambda + \Delta\lambda)\|$. If $\Delta x \neq 0$ then $\left\| \begin{pmatrix} H_{xx}(z^*) \\ J(z^*) \end{pmatrix} \Delta x \right\| \approx \|g(x + \Delta x, \lambda)\|$ and (4.3) implies that $H(z^*)(\Delta z) = 0 \implies \Delta x = 0$, so

$$\sigma_{\min} \begin{pmatrix} H_{xx}(z^*) \\ J(z^*) \end{pmatrix} \|\Delta x\| \leq \|g(x + \Delta x, \lambda)\|.$$

Thus,

$$\min \left\{ \sigma_{\min} \begin{pmatrix} H_{xx}(z^*) & J^T(z^*) \end{pmatrix}, c_{hof} \right\} \|\Delta z\| \leq \|g(z^* + \Delta z)\|. \quad \square$$

Theorem 4.3. *The second-order sufficiency condition holds at minimal solutions with Lagrange multipliers of minimal norm if h is of maximal degree, monic and the minimal structured perturbation $\|\Delta\mathcal{A}^*\|$ is sufficiently small.*

Proof. The Hessian of L with respect to $x = x(\Delta A, f^*, h)$ is

$$\nabla^2_{xx} L = H_{xx} = \begin{pmatrix} F + 2I & & E \\ & E^T & \end{pmatrix},$$

where F is a square matrix with zero diagonal whose entries are a multi-linear polynomial in λ and $\Delta\mathcal{A}$ and E^T is a symmetric matrix whose entries are homogeneous linear functions in λ.

If $\Delta\mathcal{A}^* = 0$ then $\lambda^* = 0$ and hence both $E = 0$ and $F = 0$. Accordingly, if $y \in \ker(H_{xx}) \cap \ker(J)$ then $y = \begin{pmatrix} 0 & y_2 & y_3 \end{pmatrix}^T$. Note that

$$J = \begin{pmatrix} * & \mathcal{C}_{f^*} & \mathcal{C}_h \\ & \mathcal{H}_h & \end{pmatrix},$$

where $*$ are blocks corresponding to differentiating with respect to variables in $\Delta\mathcal{A}$ and the blocks $\mathcal{C}_{f^*}$ and $\mathcal{C}_h$ are block convolution and convolution matrices that respectively correspond to multiplication by f^* and h. The block $\mathcal{H}_h$ contains a normalization vector.

$Jy = 0$ implies that there exists a vector of polynomials v and a polynomial u with the same degrees as f^* and h such that $f^*u + vh = 0$ and $\mathcal{H}_h \text{vec}(u) = 0$.

We have that h is a factor of both f^*u and vh. Since $\gcd(f^*, h) = 1$ it must be that h is a factor of u. It follows that $\deg u = \deg h$, so there exists some $\alpha \neq 0$ such that $\alpha u = h$. Since h is monic, we have that $\mathcal{H}_h \text{vec}(h) = 1$ but $\mathcal{H}_h u = 0$, which implies that $\alpha = 0$, and so $u = 0$. We have that $vh = 0$ and so $v = 0$. Hence $\ker(J) \cap \ker(H_{xx}) = 0$ and second-order sufficiency holds when $\|\Delta\mathcal{A}^*\| = 0$.

If $\|\Delta\mathcal{A}^*\|$ is sufficiently small, then $\|F\|$ will be sufficiently small so that $F + 2I$ has full rank. Accordingly, we have that

$$\ker \begin{pmatrix} F + 2I & & \\ & 0 & E \\ & E^T & 0 \end{pmatrix} \subseteq \ker \begin{pmatrix} 2I & & \\ & 0 & \\ & & 0 \end{pmatrix}. \quad \square$$

We remark that the techniques in the proof are very similar to those of [27] and [11] to show that a Jacobian matrix appearing in approximate GCD computations of two polynomials has full rank.

The implication of the local-error bound property holding is that one can reasonably approximate when quadratic convergence occurs by estimating $\sigma_{\min}\left(\begin{pmatrix} H_{xx} & J^T \end{pmatrix}\right)$ and c_{hof}. In particular, these quantities act as a structured condition number on the system. A structured backwards-error analysis of existing techniques can be performed using these quantities. Additionally, it is somewhat generic that $F + 2I$ has full rank, hence the local error-bound will hold for most instances of the approximate SNF problem with an attainable solution.

It is also important to note that we did not explicitly use the adjoint matrix. Indeed the result remains valid if we replace the adjoint with minors of prescribed dimension. Likewise, if $\mathcal{A}$ is an ill-posed instance of lower McCoy rank or approximate SNF without an attainable global minimum, then optimizing over a reversal of each entry of $(\text{Adj}(\mathcal{A} + \Delta\mathcal{A}))$ would yield a non-trivial answer and the same stability properties would hold. Thus, poorly posed-problems also remain poorly posed if slightly perturbed.

Corollary 4.4. *The LM algorithm for solving $\nabla L = 0$ has quadratic convergence under the assumptions of Theorem 4.2 and using $\mu_k = \|\nabla(L(z^k))\|_2$.*

Proof. The quantity ∇L is a multivariate polynomial, hence it is locally Lipschitz. Second-order sufficiency holds, thus we have the local error bound property is satisfied. The method converges rapidly with a suitable initial guess. $\square$

These results can also be generalized to form a low McCoy rank approximation. In the next section we discuss a technique that possibly forgoes rapid local convergence, but has a polynomial per iteration cost to compute a low McCoy rank approximation.

4.4 Computational Challenges & Initial Guesses

The most glaring problem in deriving a fast iterative algorithm is that the matrix $\text{Adj}(\mathcal{A} + \Delta\mathcal{A})$ has exponentially many coefficients as a multivariate polynomial in $\Delta\mathcal{A}$. This means computing the adjoint matrix symbolically as an ansatz is not feasible. In order to solve (4.2) we instead approximate the derivatives of the coefficients of the

adjoint numerically, which our implementation does asymptotically faster than inverting the Hessian. A detailed discussion of this is left as a future paper.

To compute an initial guess, we can use $\Delta\mathcal{A}_{init} = 0$ and take f^* and h to be a reasonable approximation to an approximate GCD of $\mathrm{Adj}(\mathcal{A})$, which will often be valid as per Theorem 4.2. In the next section we will discuss more sophisticated techniques.

5 LOWER MCCOY RANK APPROXIMATION

In this section we describe how to perform a lower McCoy rank approximation of a matrix polynomial via linearization theory.

Another way to formulate $\mathcal{A}$ having a non-trivial SNF is to solve

$$\min \|\Delta\mathcal{A}\| \text{ such that } (\mathcal{A}(\omega) + \Delta\mathcal{A}(\omega))B = 0 \text{ and } B^*B = I_2,$$

for $B \in \mathbb{C}^{n\times n}$ and $\omega \in \mathbb{C}$. The method is unstable if ω is reasonably large, since the largest terms appearing are of the size $\|\mathcal{A}\|_\infty|\omega|^d$. To remedy this, if we assume that the solution is a full-rank matrix polynomial, we can use linearization theory. If there is no full-rank solution one can simply take a lower-rank approximation [13] and extracts a square pencil of full rank. Alternatively, one may forgo the linearization and work directly with a problem that is more poorly conditioned. We assume for the rest of this section that the solutions to the low McCoy rank problem have full rank.

We can encode the eigenstructure and SNF of $\mathcal{A}$ as a linear pencil of the form $\mathcal{P} \in \mathbb{R}[t]^{nd\times nd}$, defined as

$$\mathcal{P} = \begin{pmatrix} I & & \\ & \ddots & \\ & & A_d \end{pmatrix} t - \begin{pmatrix} & I & & \\ & & \ddots & \\ & & & \\ -A_0 & -A_1 & \cdots & -A_{d-1} \end{pmatrix}.$$

This particular linearization encodes the SNF of $\mathcal{A}$, as $\mathrm{SNF}(\mathcal{P}) = \mathrm{diag}(I, I, \ldots, I, \mathrm{SNF}(\mathcal{A}))$. It follows that $\mathcal{A}$ has a non-trivial SNF if and only if $\mathcal{P}$ has a non-trivial SNF. If we preserve the affine structure of $\mathcal{P}$ and only perturb blocks corresponding to $\mathcal{A}$, then the reduction to a linear pencil will be sufficient. Other linearizations are possible as well. The pencil is generally better behaved numerically since the largest entry is proportional to $\|\mathcal{A}\|_\infty|\omega|$ (it is no longer exponential in ω).

5.1 Low McCoy Rank via Optimization

The theory of low McCoy rank approximations is an immediate generalization of the previous sections if exponentially many more minors were used in the underlying computation.

We formulate the following real optimization problem for $\omega \in \mathbb{C}$.

$$\min \|\Delta\mathcal{A}\|_F^2 \text{ such that } \begin{cases} \Re((\mathcal{P} + \Delta\mathcal{P})(\omega)B) = 0 \\ \Im((\mathcal{P} + \Delta\mathcal{P})(\omega)B) = 0 \\ \Re(B^*B) = I_m \\ \Im(B^*B) = 0, \end{cases} \tag{5.1}$$

where $n-m$ is the desired McCoy rank. The instance of $\Im(\omega_{opt}) = 0$ corresponds to $t - \omega_{opt}$ as an invariant factor of order m, while $\Im(\omega_{opt}) \neq 0$ corresponds to the real irreducible quadratic $(t - \omega_{opt})(t - \overline{\omega_{opt}})$.

In order to approach the problem using the method of Lagrange multipliers we define the Lagrangian as

$$L = \|\Delta\mathcal{A}\|_F^2 + \lambda^T \begin{pmatrix} \Re((\mathcal{P} + \Delta\mathcal{P})(\omega)B) \\ \Im((\mathcal{P} + \Delta\mathcal{P})(\omega)B) \\ \Re(B^*B) - I \\ \Im(B^*B) \end{pmatrix},$$

and proceed to solve $\nabla L = 0$. In our implementation we again make use of the LM method, although given the relatively cheap gradient cost, a first-order method will often be sufficient and faster. The problem is tri-linear, and structurally similar to affinely structured low rank approximation.

5.2 Computing an Initial Guess

In order to compute an initial guess we exploit the tri-linearity of the problem. First we approximate the determinant of $\mathcal{A}$ and consider initial guesses where $\sigma_{n-m}(\mathcal{A}(\omega_{init}))$ is reasonably small. The zeros and local extrema of $\det(\mathcal{A})$ are suitable candidates. B_{init} can be approximated from the smallest m singular vectors of $\mathcal{A}(\omega_{init})$. One can take $\Delta\mathcal{A}_{init} = 0$ or solve a linear least squares problem using B_{init} and ω_{init} as initial guesses.

5.3 Convergence & Prescribed Spectral Structure

The linearization may converge to a solution where the invariant factors are reducible quadratics or degree larger than two. Accordingly, the rate of convergence will be linear with a first-order method and super-linear (but not always quadratic) with reasonable quasi-Newton methods. To obtain a prescribed spectral structure one simply adds constraints of the form (5.1) in conjunction with a "staircase form" constraint [24] to force invariant factors to be repeated or have higher degree.

5.4 About Global Optimization Methods

The discussed problems are NP hard to solve exactly and to approximate with coefficients from $\mathbb{Q}$. This follows because affinely structured low rank approximation [5, 22] is a special case. If we consider a matrix polynomial of degree zero, then this is a scalar matrix with an affine structure. The approximate SNF will be a matrix of rank at most $n-2$, and finding the nearest affinely structured singular matrix is NP hard.

Despite the problem being intractable in the worst case, not all instances are necessarily hard. The formulation (5.1) is multi-linear and polynomial, hence amenable to the sum of squares hierarchy. Lasserre's sum of squares hierarchy [21] is a global framework for polynomial optimization that asymptotically approximates a lower bound. Accordingly, if $\|\omega_{opt}\|$ is bounded, then sum of squares techniques should yield insight into the problem.

6 IMPLEMENTATION AND EXAMPLES

We implemented our methods in `Maple 2016` We use the variant of Levenberg-Marquardt discussed in Section 4 in all instances to solve the first-order necessary condition. All computations are done using hardware precision and measured in floating point operations, or flops. The input size of our problem is measured in the dimension and degree of $\mathcal{A}$, which are n and d respectively. The cost of most quasi-Newton methods is roughly proportional to inverting the Hessian matrix, which is $O(\ell^3)$, where ℓ is the number of variables in the problem.

The method of Section 4 requires approximately $O((n^3d)^3) = O(n^9d^3)$ flops per iteration in an asymptotically optimal implementation with cubic matrix inversion, which is the cost of inverting the Hessian. Computing the Hessian costs roughly $O(n^3d^2 \times (n^2)^2) = O(n^7d^2)$ flops using a blocking procedure, assuming the adjoint computation runs in time $O(n^3d^2)$. There are $O(n^3d)$ Lagrange multipliers since the adjoint has degree at most $(n-1)d$.

The method of Section 5 has a Hessian matrix of size $O(n^2d^2) \times O(n^2d^2)$ in the case of a rank zero McCoy rank approximation. Accordingly, the per iteration cost is roughly $O(n^6d^6)$ flops. Given the lack of expensive adjoint computation, a first-order method will typically require $O(n^2d^2)$ flops per iteration (ignoring the initial setup cost), with local linear convergence.

Example 6.1 (Nearest Interesting SNF). Consider the matrix polynomial $\mathcal{A}$ with a trivial SNF

$$\begin{pmatrix} t^2+.1t+1 & 0 & .3t-.1 & 0 \\ 0 & .9t^2+.2t+1.3 & 0 & .1 \\ .2t & 0 & t^2+1.32+.03t^3 & 0 \\ 0 & .1t^2+1.2 & 0 & .89t^2+.89 \end{pmatrix}$$

of the form $\mathrm{diag}(1,\ldots,1,\det(\mathcal{A}))$.

If we prescribe the perturbations to leave zero coefficients unchanged, then using the methods of Section 4 and Section 5 we compute a local minimizer $\mathcal{A}+\Delta\mathcal{A}_{opt}$ of

$$\begin{pmatrix} 1.0619t^2+.018349t+.94098 & 0 & .27477t-.077901 & 0 \\ 0 & .90268t^2+.22581t+1.2955 & 0 & .058333 \\ .13670t & 0 & .027758t^3+.97840t^2+1.3422 & 0 \\ 0 & .10285t^2+1.1977 & 0 & .84057t^2+.93694 \end{pmatrix},$$

with $\|\Delta\mathcal{A}_{opt}\| \approx .164813183138322$. The SNF of $\mathcal{A}+\Delta\mathcal{A}_{opt}$ is approximately

$$\mathrm{diag}(1,1,s1,s1(t^5+35.388t^4+6.4540t^3+99.542t^2+5.6777t+70.015)),$$

where $s_1 \approx t^2 + 0.0632934647739423t + 0.960572576466186$. s_1 corresponds to $\omega_{opt} \approx -0.0316467323869714 - 0.979576980535687i$.

The method discussed in Section 4 converges to approximately 14 decimal points of accuracy[†] after 69 iterations and the method of Section 5 converges to the same precision after approximately 34 iterations. The initial guess in both instances was $\Delta\mathcal{A}_{init} = 0$. The initial guesses of f^* and h were computed by an approximate GCD routine. For the initial guess of ω we chose a root or local extrema of $\det(\mathcal{A})$ that minimized the second-smallest singular value of $\mathcal{A}(\omega)$, one of which is $\omega_{init} \approx -.12793 - 1.0223i$.

Example 6.2 (Lowest McCoy Rank Approximation). With $\mathcal{A}$ as in the previous example, consider the 0-McCoy rank approximation problem with the same prescribed perturbation structure.

We compute a local minimizer $\mathcal{A}+\Delta\mathcal{A}_{opt}$ to be approximately

$$\begin{pmatrix} .80863t^2+1.1362 & 0 & 0 & 0 \\ 0 & .91673t^2+1.2881 & 0 & 0 \\ 0 & 0 & .95980t^2+1.3486 & 0 \\ 0 & .60052t^2+.84378 & 0 & .71968t^2+1.0112 \end{pmatrix},$$

with $\|\Delta\mathcal{A}_{opt}\| \approx .824645447014665$ after 34 iterations to 14 decimal points of accuracy. We compute $\omega_{opt} \approx -1.18536618732372i$ which corresponds to the single invariant factor $s_1 \approx t^2 + 1.4051$. The SNF of $\mathcal{A}+\Delta\mathcal{A}_{opt}$ is of the form (s_1, s_1, s_1, s_1).

[†] $\nabla L = 0$ is solved to 14 digits of accuracy; the extracted quantities are accurate to approximately the same amount.

7 FUTURE DIRECTIONS

We will continue our research towards a more complete theoretical understanding of computing the nearest matrix polynomial with prescribed finite and infinite spectral structure (or determining non-existence thereof). We also plan to investigate formulating the Smith-McMillan form as an optimization problem and determine if similar existence and stability results can be derived. A detailed exploration of computing the adjoint matrix in a numerically robust manner and corresponding error analysis will also be made.

REFERENCES

[1] S. Ahmad and R. Alam. 2009. Pseudospectra, critical points and multiple eigenvalues of matrix polynomials. *Linear Algebra Appl.* 430, 4 (2009), 1171–1195.
[2] B. Beckermann and G. Labahn. 1998. When are two numerical polynomials relatively prime? *J. Symb. Comp.* 26 (1998), 677–689.
[3] Th. Beelen and P. Van Dooren. 1988. An improved algorithm for the computation of Kronecker's canonical form of a singular pencil. *Linear Algebra Appl.* 105 (1988), 9–65.
[4] D. Bertsekas. 1999. *Nonlinear programming.* Athena Scientific, USA.
[5] R. P. Braatz, P. M. Young, J. C. Doyle, and Manfred M. 1994. Computational complexity of /spl mu/ calculation. *IEEE Trans. Automat. Control* 39, 5 (1994), 1000–1002.
[6] J. W. Demmel and A. Edelman. 1995. The dimension of matrices (matrix pencils) with given Jordan (Kronecker) canonical forms. *Linear Algebra Appl.* 230 (1995), 61–87.
[7] A. Edelman, E. Elmroth, and B. Kågström. 1997. A geometric approach to perturbation theory of matrices and matrix pencils. Part I: Versal deformations. *SIAM J. Matrix Anal. Appl.* 18, 3 (1997), 653–692.
[8] A. Edelman, E. Elmroth, and B. Kågström. 1999. A geometric approach to perturbation theory of matrices and matrix pencils. Part II: A stratification-enhanced staircase algorithm. *SIAM J. Matrix Anal. Appl.* 20, 3 (1999), 667–699.
[9] J-Y. Fan and Y-X. Yuan. 2005. On the quadratic convergence of the Levenberg-Marquardt method without nonsingularity assumption. *Computing* 74, 1 (2005), 23–39.
[10] S. Fatouros and N. Karcanias. 2003. Resultant properties of gcd of many polynomials and a factorization representation of gcd. *Internat. J. Control* 76, 16 (2003), 1666–1683.
[11] M. Giesbrecht, J. Haraldson, and E. Kaltofen. 2016. Computing Approximate Greatest Common Right Divisors of Differential Polynomials. (2016). Submitted.
[12] M. Giesbrecht, J. Haraldson, and G. Labahn. 2017. Computing the Nearest Rank-Deficient Matrix Polynomial. In *Proc. ACM on International Symposium on Symbolic and Algebraic Computation (ISSAC'17)*. 181–188.
[13] M. Giesbrecht, J. Haraldson, and G. Labahn. 2017. Lower Rank Approximations of Matrix Polynomials. *J. of Symbolic Computation* (2017). Submitted.
[14] I. Gohberg, P. Lancaster, and L. Rodman. 2009. *Matrix polynomials.* SIAM, USA.
[15] G. Golub and C. Van Loan. 2013. *Matrix Computations* (4 ed.). Johns Hopkins University Press, USA.
[16] J. Haraldson. 2015. *Computing Approximate GCRDs of Differential Polynomials.* Master's thesis. University of Waterloo.
[17] A. J. Hoffman. 1952. On Approximate Solutions of Systems of Linear Inequalities. *J. Res. Nat. Bur. Standards* 49, 4 (1952).
[18] T. Kailath. 1980. *Linear systems.* Vol. 156. Prentice-Hall, USA.
[19] E. Kaltofen and A. Storjohann. 2015. The complexity of computational problems in exact linear algebra. In *Encyclopedia of Applied and Computational Mathematics.* Springer, Germany, 227–233.
[20] N. Karmarkar and Y. N. Lakshman. 1996. Approximate Polynomial Greatest Common Divisors and Nearest Singular Polynomials. In *Proc. International Symposium on Symbolic and Algebraic Computation (ISSAC'96).* ACM Press, Zurich, Switzerland, 35–39.
[21] J-B. Lasserre. 2001. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization* 11, 3 (2001), 796–817.
[22] S. Poljak and J. Rohn. 1993. Checking robust nonsingularity is NP-hard. *Mathematics of Control, Signals, and Systems (MCSS)* 6, 1 (1993), 1–9.
[23] G. Stewart. 1994. Perturbation theory for rectangular matrix pencils. *Linear algebra and its applications* 208 (1994), 297–301.
[24] P. Van Dooren and P. Dewilde. 1983. The eigenstructure of an arbitrary polynomial matrix: computational aspects. *Linear Algebra Appl.* 50 (1983), 545–579.
[25] A. Vardulakis and P. Stoyle. 1978. Generalized resultant theorem. *IMA Journal of Applied Mathematics* 22, 3 (1978), 331–335.
[26] N. Yamashita and M. Fukushima. 2001. On the rate of convergence of the Levenberg-Marquardt method. In *Topics Num. Analysis.* Springer, 239–249.
[27] Z. Zeng and B. H. Dayton. 2004. The Approximate GCD of Inexact Polynomials. In *Proc. International Symposium on Symbolic and Algebraic Computation (ISSAC'04).* Santander, Spain, 320–327.