University of Alberta

Numerical Computation of Padé-Hermite Systems

by

Anthony Robert Jones ©

A thesis
submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree
of Masters of Science

Department of Computing Science

Edmonton, Alberta
Fall 1992

ISBN 0-315-77271-9

UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Anthony Robert Jones
TITLE OF THESIS: Numerical Computation of Padé-Hermite Systems

DEGREE: Masters of Science
YEAR THIS DEGREE GRANTED: 1992

(Signed) Anthony R. Jones

Permanent Address:
15207 77 Avenue,
Edmonton, Alberta,
Canada

Date: June 2, 1992

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled **Numerical Computation of Padé-Hermite Systems** submitted by **Anthony Robert Jones** in partial fulfillment of the requirements for the degree of Masters of Science.

supervisor: Stan Cabay

Gerald Cliff
external: Gerald Cliff (Mathematics)

Hong Zhang
examiner: Hong Zhang

Date: May 22, 1992

To Donna

and my parents

Abstract

For a vector $(n_0, \ldots, n_k)$ of nonnegative integers, and a vector of formal power series $A(z) = (A_0(z), \ldots, A_k(z))$, a *Padé-Hermite System (PHS)* $S(z)$ is a $k+1 \times k+1$ matrix of polynomials satisfying $A(z) \cdot S(z) = z^{\|n\|+1} R(z)$ where $\|n\| = n_0 + \cdots + n_k$. Computing a PHS involves solving two linear systems with a block Sylvester coefficient matrix. We present an iterative algorithm based on that of Cabay, Labahn and Beckermann for numerically computing Padé-Hermite Systems along a diagonal of the Padé-Hermite table. An easily computed stability parameter γ, which estimates the condition number of the block Sylvester matrix at a given point, is defined in order to determine if a given Padé-Hermite table point is stable. Stable points have well conditioned block Sylvester matrices. The iterative algorithm requires approximately $\mathcal{O}(\|n\|^2 + s^2\|n\|)$ operations, where s is the largest step-size taken along the Padé-Hermite diagonal.

To test the algorithm, a method is developed which enables the construction of power series with unstable blocks in predetermined Padé-Hermite table locations. Experiments using a Fortran implementation of the algorithm are reported for a variety of values of k. The relative error in the iteratively computed PHS is found to be comparable to that of a PHS determined by direct solution of the linear systems using Gaussian elimination requiring $\mathcal{O}(\|n\|^3)$ operations. Based on the findings of these experiments, a new stability parameter $\hat{\gamma}$ is proposed.

Acknowledgements

I would like to thank my supervisor, Dr. Stan Cabay for the time, patience, and valuable insight he provided during the development of this thesis. His vast knowledge of numerical analysis and Padé theory led to many of the results contained herein and his wisdom and guidance were much appreciated.

I would also like to thank the members of my examining committee, Dr. H. Zhang, Dr. G. Cliff for their time and effort in reviewing this research, and Dr. J. You for chairing the committee.

Special thanks to Dr. G. Labahn for his assistance with Maple.

Without the resources and technical support provided by The Department of Computing Science at the University of Alberta, this thesis would not have been possible. The financial support provided by an NSERC postgraduate scholarship was also much appreciated.

Finally I would like to thank Donna Anderson for her patience and support throughout this work and my parents who gave me the encouragement and opportunity to pursue my dreams.

Contents

List of Tables

Chapter 1

Introduction

The concept of a Padé-Hermite approximant was first introduced over 100 years ago in the thesis of Padé [23]. His work was based on previous results of Hermite [16] [17]. The **Latin** [18] or **type I polynomial problem** as it was known, emerged from Hermite's earlier study of a similar type of approximant for a vector of power series, the simultaneous Padé approximant.

Simultaneous Padé approximants (also known as the **Roman** [18] or **type II polynomial problem**) were used by Hermite when he proved the transcendence of e. A general definition for Padé-Hermite and simultaneous Padé approximants, as well as an extensive study of their properties is originally due to Mahler [21]. Additional properties have been presented by Jager [18] and Coates [12].

Padé-Hermite approximants can be defined as follows. Let

$$A_i(z) = \sum_{j=0}^{\infty} a_{j,i}\, z^j, \qquad i = 0, \ldots, k, \tag{1.1}$$

be power series with coefficients $a_{j,i}$ from a field $\mathcal{F}$. For nonnegative integers n_i, a Padé-Hermite approximant of type $(n_0, \ldots, n_k)$ is a set of $k+1$ polynomials $P_i(z)$

satisfying

$$A_0(z) \cdot P_0(z) + \cdots + A_k(z) \cdot P_k(z) = z^{\|n\|+k} R(z),$$

where $\|n\| = \sum_{i=0}^{k} n_i$. The power series $R(z)$ is called the residual and the vector of integers $(n_0, \ldots, n_k)$ is often referred to as a multi-index. The degree of the polynomial $P_i(z)$ is bounded by the integer n_i.

Given k power series $A_i(z), \quad i = 1, \ldots, k$, a simultaneous Padé approximant of type $(n_0, \ldots, n_k)$ is a set of $k+1$ polynomials $P_i(z)$ satisfying

$$A_i(z) \cdot P_0(z) + P_i(z) = z^{\|n\|+1} R(z), \qquad i = 1, \ldots, k.$$

The degree of each polynomial $P_i(z)$ in a simultaneous Padé approximant is bounded by $\|n\| - n_i$.

If $k = 1$ and $A_1(z) = -1$, the Padé-Hermite approximant problem reduces to the classical notion of a Padé approximant of a single power series. Other examples of Padé-Hermite approximation include the D-Log approximants of Baker [3] and the quadratic approximants of Shafer [26]. Basic properties of Padé-Hermite approximants accompanied by additional examples can be found in Baker and Graves-Morris [4].

The applications of Padé-Hermite approximants and simultaneous Padé approximants are diverse. As mentioned earlier, Hermite used Padé-Hermite approximants to prove the transcendence of the number e. These approximants can also be used to form the inverse of block Hankel, block Toeplitz and block Sylvester matrices. Simultaneous Padé approximants have proven to be a valuable tool in computing partial realizations [15]. To illustrate the use of simultaneous Padé approximants, we will briefly discuss the partial realization problem for single input, multi-output systems.

The minimal partial realization problem is described by Van Barel and Bultheel [27] as follows. Let $h_1, h_2, h_3, \ldots$ be a sequence of complex numbers called the **Markov parameters.** The rational function $u(z)/v(z)$ is called a **realization** of this sequence if and only if

$$\frac{u(z)}{v(z)} = h_1 z^{-1} + h_2 z^{-2} + \ldots, \quad z \to \infty. \tag{1.2}$$

If N is finite and

$$\frac{u(z)}{v(z)} = h_1 z^{-1} + h_2 z^{-2} + \ldots + h_N z^{-N} + \mathcal{O}(z^{-N-1}), \quad z \to \infty, \tag{1.3}$$

then $u(z)/v(z)$ is called a **partial realization** of order N. When $u(z)$ and $v(z)$ are coprime, the degree m of the denominator $v(z)$ is called the degree of the partial realization. A partial realization is called **minimal** if there exists no other partial realization of lower degree. If

$$\begin{aligned} v(z) &= v_0 + v_1 z + \ldots + v_n z^m \quad (v_m \neq 0) \quad \text{and} \\ u(z) &= u_0 + u_1 z + \ldots + u_{m-1} z^{m-1}, \end{aligned}$$

then we can express (1.3) in matrix notation as

$$\begin{pmatrix} 0 & 0 & \ldots & h_1 \\ \vdots & \vdots & & \vdots \\ 0 & h_1 & \ldots & h_m \\ h_1 & h_2 & \ldots & h_{m+1} \\ \vdots & \vdots & & \vdots \\ h_N & h_{N+1} & \ldots & h_{N+m} \end{pmatrix} \begin{pmatrix} v_0 \\ \vdots \\ v_m \end{pmatrix} = \left(\begin{array}{c} u_{m-1} \\ \vdots \\ u_0 \\ \hline 0 \\ \vdots \\ 0 \end{array}\right) \tag{1.4}$$

By making a substitution of variables, we can formulate (1.4) as a Padé approximation problem. Let

$$h(z) = h_1 + h_2 z^{-1} + \cdots. \tag{1.5}$$

Then (1.3) becomes

$$\frac{u(z)}{v(z)} = z^{-1}\, h(z) + \mathcal{O}\left(z^{-N-1}\right). \tag{1.6}$$

Moving $v(z)$ to the right gives

$$u(z) = z^{-1}\, v(z)\, h(z) + \mathcal{O}\left(z^{-N-1}\right). \tag{1.7}$$

Replacing z by $1/z$ results in

$$\begin{aligned} u\left(\frac{1}{z}\right) &= z\, v\left(\frac{1}{z}\right)\, h\left(\frac{1}{z}\right) + \mathcal{O}\left(z^{N+1}\right) && (1.8) \\ z^{m-1}\, u\left(\frac{1}{z}\right) &= \left[z^{m}\, v\left(\frac{1}{z}\right)\right]\, h\left(\frac{1}{z}\right) + \mathcal{O}\left(z^{m+N}\right). && (1.9) \end{aligned}$$

Thus,

$$u^{*}(z) = v^{*}(z)\, h^{*}(z) + \mathcal{O}\left(z^{m+N}\right) \tag{1.10}$$

where

$$u^{*}(z) = \sum_{i=0}^{m-1} u_{m-i-1}\, z^{i}, \qquad v^{*}(z) = \sum_{i=0}^{m} v_{m-i}\, z^{i}, \qquad h^{*}(z) = \sum_{i=1}^{N} h_{i}\, z^{i-1}. \tag{1.11}$$

These polynomials $u^{*}(z)$ and $v^{*}(z)$ give the Padé approximant of type $(m-1, m)$ for $h^{*}(z)$. Hence the minimal partial realization of $h^{*}(z)$ can be found by computing Padé approximants for increasing m until the required order condition (1.3) is obtained. Note that a solution of (1.4) always exists if $N \leq m$. The polynomials $u(z)$, $v(z)$ can be used to construct a controller canonical realization of the system [19].

Graves-Morris and Wilkins [15] generalize the partial realization problem by considering vectors of Markov parameters. Let

$$h(z) = \begin{pmatrix} h_{1}^{(1)} \\ \vdots \\ h_{1}^{(k)} \end{pmatrix} + \begin{pmatrix} h_{2}^{(1)} \\ \vdots \\ h_{2}^{(k)} \end{pmatrix} z^{-1} + \ldots. \tag{1.12}$$

Let $n = (m-1, m, \ldots, m)$ be a vector of $k+1$ nonnegative integers. Then define $\|n\| = (k+1)m - 1$. We wish to find a set of vectors $v(z), u^{(1)}(z), \ldots, u^{(k)}(z)$ such that for some finite integer N,

$$\frac{u^{(i)}(z)}{v(z)} = h_1^{(i)} z^{-1} + h_2^{(i)} z^{-2} + \ldots + h_N^{(i)} z^{-N} + \mathcal{O}(z^{-N-1}) \qquad i = 1, \ldots, k. \quad (1.13)$$

Let

$$v(z) = \sum_{j=0}^{km} v_i z^i, \qquad (v_{km} \neq 0), \quad (1.14)$$

$$u^{(i)}(z) = \sum_{j=0}^{km-1} u_j^{(i)}, \qquad i = 1, \ldots, k. \quad (1.15)$$

To compute the coefficients of $v(z)$ we can solve the system

$$(v_0, \ldots, v_{km}) \left(\begin{array}{ccc|c|ccc} h_1^{(1)} & \cdots & h_N^{(1)} & & h_1^{(k)} & \cdots & h_N^{(k)} \\ \vdots & & \vdots & \cdots & \vdots & & \vdots \\ h_{km+1}^{(1)} & \cdots & h_{km+N}^{(1)} & & h_{km+1}^{(k)} & \cdots & h_{km+N}^{(k)} \end{array} \right) = 0. \quad (1.16)$$

Then the components $u^{(i)}(z)$ are given by

$$(v_0, \ldots, v_{km}) \begin{pmatrix} & & 0 \\ & \cdot^{\cdot^{\cdot}} & h_1^{(i)} \\ 0 & \cdot^{\cdot^{\cdot}} & \vdots \\ h_1^{(i)} & \cdots & h_{km}^{(i)} \end{pmatrix} = \left(u_{km-1}^{(i)}, \ldots, u_0^{(i)}\right), \quad i = 1, \ldots, k. \quad (1.17)$$

A solution of (1.16) always exists for $N \leq m$. Thus

$$\frac{u^{(i)}(z)}{v(z)} = z^{-1} h(z) + \mathcal{O}\left(z^{-N-1}\right). \quad (1.18)$$

Moving $v(z)$ to the right, letting $z = \frac{1}{z}$, and multiplying by z^{km-1} gives

$$z^{km-1} u^{(i)}\left(\frac{1}{z}\right) = \left[z^{km} v\left(\frac{1}{z}\right)\right] \left[h\left(\frac{1}{z}\right)\right] + \mathcal{O}\left(z^{km+N}\right) \quad (1.19)$$

Hence

$$
\begin{aligned}
u^{(i)^*}(z) &= v^*(z)\,h^*(z) + \mathcal{O}\left(z^{km+N}\right) && && (1.20)\\
&= v^*(z)\,h^*(z) + \mathcal{O}\left(z^{(k+1)m}\right) && m \geq N && (1.21)\\
&= v^*(z)\,h^*(z) + \mathcal{O}\left(z^{\|m\|+1}\right) && m \geq N, && (1.22)
\end{aligned}
$$

where

$$u^{(i)^*}(z) = \sum_{j=0}^{km-1} u_{km-j-1}\, z^j, \qquad v^*(z) = \sum_{j=0}^{m} v_{km-j}\, z^j, \qquad h^*(z) = \sum_{j=1}^{N} h_j\, z^{j-1}. \quad (1.23)$$

The polynomials $v(z), u^{(1)^*}(z), \ldots, u^{(k)^*}(z)$ give the simultaneous Padé approximant of type n. If we substitute $z = 1/z$ for $v(z), u^{(1)}(z), \ldots, u^{(k)}(z)$, the resulting simultaneous Padé approximant is

$$\left(\frac{u_1^{(1)}z^{-1} + \ldots + u_{km-1}^{(1)}z^{-km+1}}{1 + v_1 z^{-1} + \ldots + v_{km}z^{-km}}, \ldots, \frac{u_1^{(k)}z^{-1} + \ldots + u_{km-1}^{(k)}z^{-km+1}}{1 + v_1 z^{-1} + \ldots + v_{km}z^{-km}}\right). \quad (1.24)$$

The polynomial coefficients of (1.24) can be used to construct the controller canonical form of a single input multi-output system [15].

Fundamental to the study of Padé approximants is the two-dimensional Padé table. The m^{th} row and n^{th} column of the Padé table contain the (m, n) Padé approximant to a power series. For Padé-Hermite approximants, we can generalize this to the $k+1$ dimensional Padé-Hermite table. Della Dora and Discrescenzo [13] present a number of relationships between neighboring entries in the table resulting in an algorithm to compute such approximants. Recurrence relations involving Padé-Hermite table elements led to the algorithm of Paszkowski [24] with cost complexity of $\mathcal{O}(\|n\|^2)$ operations.

The algorithms of Paszkowski and of Della Dora and Discrescenzo are only valid for **perfect**[1] power series. To compute a Padé-Hermite approximant we can solve an

[1] Paszkowski refers to this as being *normal*.

associated linear system with a block Hankel coefficient matrix of dimension $\|n\| \times \|n\|$ containing coefficients of the input power series $A_i(z)$. For a vector of power series to be perfect, this Hankel matrix along with a specific set of submatrices must be nonsingular. This restriction requires for example, that all constant terms in the power series' be nonzero for the system to be perfect. Non-perfect Padé-Hermite approximants correspond to *singular* blocks in the Padé-Hermite table.

Algorithms for computing Padé-Hermite approximants are often characterized as being **fast** or **superfast**. Gaussian elimination with pivoting requires $\mathcal{O}\left(\|n\|^3\right)$ operations to compute a Padé-Hermite approximant. Fast algorithms are considered to have a cost complexity of $\mathcal{O}\left(\|n\|^2\right)$ while superfast algorithms require $\mathcal{O}\left(\|n\| \log^2 \|n\|\right)$ operations. The algorithm we will develop has is fast with a cost of $\mathcal{O}(\|n\|^2 + s_i^2\|n\|)$ operations, where s_i is the largest step-size taken along the Padé-Hermite diagonal.

Several authors including Antoulas [2], Beckerman [5], and Cabay et al. [10] and Van Barel and Bultheel [28] give fast algorithms for computing Padé-Hermite approximants for non-perfect systems. These methods are based on computing a set of polynomial vectors which describe all possible solutions of the Padé-Hermite problem. Recently, Beckerman and Labahn [6] introduced a uniform approach to computing both Padé-Hermite and simultaneous Padé approximants using a *power Hermite-Padé* approximant. Cabay and Labahn [9] have also proposed a superfast algorithm for computing Padé-Hermite, and simultaneous Padé approximants which also works for non-perfect systems.

A characteristic common to all the algorithms mentioned is their algebraic approach to computing Padé-Hermite approximants. Exact arithmetic is implicitly assumed in all the aforementioned algorithms. Algebraic programming systems such as Maple, Mathematica, and Macsyma are available for coding these algorithms. These systems are computationally expensive and are thus limited to small problems. To our knowledge no attempt has been made to analyze the numerical properties of algorithms for computing the Padé-Hermite approximant.

The numerical algorithm we will present for computing Padé-Hermite approximants will be patterned after that of Cabay et al. [10]. This algorithm was chosen over all others because of the natural way it can be extended from algebraic to numerical. In the algebraic version, singular Padé-Hermite table blocks points are skipped when iteratively computing a Padé-Hermite approximant. If we view the Padé-Hermite table numerically as being composed of stable and unstable blocks, we need only define a stability measure to determine if a point should be accepted.

Some work has been done investigating numerical algorithms for computing Padé approximants (i.e. the case $k = 1$) in a stable manner (see Cabay and Meleshko [11] for discussion and references). In their work, Cabay and Meleshko propose an algorithm (based on that of Cabay and Choi [8]) which they show to be weakly stable (cf. Bunch [7]). To obtain the approximant, a number of linear systems involving Hankel matrices are solved. Successive approximants are computed along a diagonal path in the Padé table. The stability of each relevant Hankel system is estimated by a single stability parameter γ. This parameter is determined directly from the current and previous Padé approximant and estimates the condition number of the Hankel matrix. Points in the Padé table whose approximant is computed using a poorly conditioned Hankel matrix are deemed unstable and are jumped over. Error bounds on the polynomial coefficients are derived in terms of this stability parameter. Experimental evidence is provided which supports these error bounds.

The goal of this thesis is to extend to arbitrary k, the method of Cabay and Meleshko [11]. The algorithm so developed, can be applied to any vector of power series regardless of whether the perfect condition is met. Padé-Hermite approximants are computed iteratively by solving a set of linear systems involving a Sylvester matrix. To obtain a Padé-Hermite approximant along some diagonal in the Padé-Hermite table, a recurrence relation is defined which uses the solution of linear systems at the last computed point and the solution of two linear systems at the current point.

This algorithm will compute Padé-Hermite approximants at *stable* points along a diagonal path in the Padé-Hermite table. We develop a stability parameter γ which theoretically provides an *a posteriori* estimate for the inverse of the condition number of the Sylvester matrix. Solving a poorly conditioned linear system generally results in a solution with large error. The parameter γ enables us to decide if the Padé-Hermite approximant computed has sufficient accuracy. If the value of γ is greater than some user specified tolerance, the approximant is accepted (we consider this point in the Padé-Hermite table to be stable). Otherwise, we jump over this point as in the algorithm of Cabay and Meleshko. The parameter γ also provides an estimate on the number of digits of accuracy in the given approximant. By choosing a tolerance in an appropriate way, a user can specify the accuracy of the requested approximant.

It is of particular importance to understand that the approach taken in this thesis in developing such an algorithm is largely based on intuition and experimentation. We proceed without proof in many cases and do not substantiate all decisions leading to the algorithm. A formal error analysis will not be given for the algorithm.

Analysis of the algorithm will primarily consist of examining numerical results obtained through experimentation. The parameters of interest include the relative error in the Padé-Hermite approximant as well as the error in the residual. We compare results obtained from a Fortran implementation with those obtained using exact arithmetic in Maple to obtain the relative error in Padé-Hermite approximant polynomial coefficients. In addition, we obtain the relative error for computing a Padé-Hermite approximant directly using Gaussian elimination with pivoting. We illustrate that the error introduced in Padé-Hermite approximant coefficients by our iterative algorithm is comparable to that of the direct method. We will show that the growth in residual error and relative error in the coefficients is linear as opposed to exponential. Several specific conjectures will be made regarding the behavior of the algorithm. These conjectures will be supported by the numerical experiments.

Random power series coefficients generally result in well conditioned Hankel matrices (stable Padé-Hermite table points). To thoroughly test the algorithm we required a method to generate power series whose corresponding Padé-Hermite table contained unstable blocks in predictable locations. To accomplish this, a method was established which, in exact arithmetic under Maple, could generate power series with singular blocks of arbitrary size. This method we consider to be a secondary but important contribution of this thesis. It provides a platform for comparing methods and their effectiveness in dealing with singular blocks. By perturbing such power series by small amounts, we can construct a variety of interesting problems. These singular blocks will no longer be singular when floating point arithmetic is used. Instead, the singular blocks in the Padé-Hermite table will be extremely unstable as the condition number of the Sylvester matrix corresponding to the point is large. As we perturb the power series coefficients by larger amounts, these "singular blocks" become more stable.

We will compute Padé-Hermite approximant for four classes of problems. The first class will consist of random power series with coefficients between -1 and 1. The second class of problem utilizes power series which are artificially generated to contain singular blocks in the Padé-Hermite table. The third and fourth problem classes will be generated by perturbing the singular block power series by varying amounts.

The purpose of this experimental approach is to study the characteristics of Padé-Hermite approximants for a variety of k and power series class. By better understanding the behavior of Padé-Hermite approximants, it is hoped that a formal proof of the stability of the algorithm can be found in the future. This work is also intended to lay groundwork for a numerically stable superfast algorithm related to that of Cabay and Labahn. This we consider to be the main contribution of this thesis.

The thesis will be organized as follows. In chapter 2 the concept of a Padé-Hermite System will be introduced. A recurrence relation for computing successive Padé-Hermite Systems will be given along with the pseudo-code algorithm of Cabay et al. [10]. In chapter 3 we define several power series and matrix norms and prove their compatibility. A stability measure γ for a Padé-Hermite System is also given. Chapter 4 presents the pseudo-code for the numerical Padé-Hermite algorithm that will be tested throughout the balance of the thesis. Several normalizations and scalings adopted to promote stability are discussed at this time. In chapter 5 we present some error bounds due to Cabay and Meleshko [11] and postulate the error behavior of the Padé-Hermite algorithm. Details of the experimental procedure are highlighted in chapter 6. A method for generating power series that result in singular blocks in the Padé-Hermite table is given. The four problem classes used in numerical experiments, along with a description of the resulting tables are provided. Experimental results and analysis are the main focus of chapter 7. Results for all four problem classes are reported for a number of values of k. Based on the experimental findings, a new stability parameter $\hat{\gamma}$ is introduced. Finally chapter 8 summarizes the results of this thesis and suggests some future research topics to investigate.

Chapter 2

The Padé-Hermite System

In this chapter we introduce the fundamental structure used by Cabay et al. [10] to compute Padé-Hermite approximants. The Padé-Hermite System is a matrix of polynomials formed from the solutions of two linear systems with a block Sylvester coefficient matrix. Padé-Hermite Systems exist for nonsingular points in the Padé-Hermite table and contain the Padé-Hermite approximant. Cabay et al. [10] provide a recurrence relation allowing iterative computation of Padé-Hermite Systems at all nonsingular points along some diagonal in the Padé-Hermite table. This recurrence relation leads to a fast iterative algorithm for computing Padé-Hermite Systems.

In this chapter we will show how to compute a Padé-Hermite System given a vector of power series and an index vector. We will state the important theorem of Cabay et al. which motivates the iterative algorithm for computing Padé-Hermite Systems. Finally, a pseudo-code description of algorithm is given. Many of the results which follow are adapted from [10] and [9]. We begin by formally defining a Padé-Hermite System.

2.1 Definitions

Let

$$A(z) = (A_0(z) \,|\, A_1(z), \ldots, A_k(z)) = (B(z) \,|\, C(z)) \tag{2.1}$$

be a $1 \times k+1$ vector of power series, where

$$A_i(z) = \sum_{j=0}^{\infty} a_{j,i}\, z^j, \qquad 0 \le i \le k, \tag{2.2}$$

and the power series coefficients $a_{j,i}$ come from a field $\mathcal{F}$.

Let $n = (n_0, \ldots, n_k)$ be a vector of nonnegative integers with $n_i > 0$ for at least one i. Define

$$\|n\| = \sum_{i=0}^{k} n_i, \tag{2.3}$$

and let

$$S(z) = \left(\begin{array}{c|ccc} S_{0,0}(z) & S_{0,1}(z) & \cdots & S_{0,k}(z) \\ \hline S_{1,0}(z) & S_{1,1}(z) & \cdots & S_{1,k}(z) \\ \vdots & \vdots & & \vdots \\ S_{k,0}(z) & S_{k,1}(z) & \cdots & S_{k,k}(z) \end{array}\right) = \left(\begin{array}{c|c} z^2 P(z) & U(z) \\ \hline z^2 Q(z) & V(z) \end{array}\right) \tag{2.4}$$

be a $(k+1) \times (k+1)$ matrix of polynomials such that the degrees of $S(z)$ componentwise satisfy

$$\partial S(z) \le \left(\begin{array}{c|ccc} n_0+1 & n_0 & \ldots & n_0 \\ \hline n_1+1 & n_1 & \ldots & n_1 \\ \vdots & \vdots & & \vdots \\ n_k+1 & n_k & \ldots & n_k \end{array}\right). \tag{2.5}$$

Definition 2.1 (Cabay et al. [9]) *The matrix $S(z)$ given by (2.4) is a* **Padé-Hermite System (PHS)** *of type n if:*

I) *$S(z)$ satisfies the degree bounds (2.5);*

II) $$A(z) \cdot S(z) = z^{\|n\|+1} \left[R_0(z), \ldots, R_k(z)\right] = z^{\|n\|+1} R(z),$$

where $R_i(z)$, $i = 0, \ldots, k$, are power series;

III) $\det(V(0)) \neq 0$ *and* $R_0(0) \neq 0$.

□

A PHS is said to be normalized if $R_0(0) = 1$ and $V(0) = I_k$ (the $k \times k$ identity matrix). An arbitrary PHS can be normalized by multiplying it on the right by the matrix

$$\left(\begin{array}{c|c} R_0^{-1}(0) & 0 \\ \hline 0 & V^{-1}(0) \end{array}\right). \tag{2.6}$$

To obtain the polynomial components of the PHS matrix $S(z)$, two sets of linear equations are solved. We first consider the polynomials $S_{0,0}(z), S_{1,0}(z), \ldots, S_{k,0}(z)$ which give $P(z)$ and $Q(z)$. From (2.4) and Definition 2.1,

$$B(z) \cdot P(z) + C(z) \cdot Q(z) = z^{\|n\|-1} R_0(z). \tag{2.7}$$

We can express (2.7) as the matrix problem

$$\left(\begin{array}{ccc|c|ccc} a_{0,0} & & & & a_{0,k} & & \\ & \ddots & & & & \ddots & \\ \vdots & & a_{0,0} & \cdots & \vdots & & a_{0,k} \\ & & \vdots & & & & \vdots \\ a_{\|n\|-2,0} & \cdots & a_{\|n\|-n_0-1,0} & & a_{\|n\|-2,k} & \cdots & a_{\|n\|-n_k-1,k} \end{array}\right) X = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix} \tag{2.8}$$

where

The component $P(z)$ is then given by

$$P(z) = \sum_{i=0}^{n_0-1} p_i \cdot z^i \tag{2.10}$$

and the $j-th$ component $Q_j(z)$, $j = 1, \ldots, k$, of $Q(z)$ is given by

$$Q_j(z) = \sum_{i=0}^{n_j-1} q_{i,j} \cdot z^i. \tag{2.11}$$

The system (2.8) consists of $\|n\| - 1$ equations and $\|n\|$ unknowns and is therefore under-determined. Clearly a solution X and consequently a solution $(P(z), Q(z))$ always exists for this homogeneous system. By setting $S_{0,0}(z) = z^2 P(z)$ and $S_{i,0}(z) = z^2 Q_i(z)$, we satisfy conditions I) and II) of Definition 2.1 for the first column of $S(z)$.

In order to ensure that $R_0(0) = 1$, we add a row to the matrix in (2.8) and obtain a new system. For notational convenience we make the following definition. Let

$$T_n = \left(\begin{array}{ccc|c|ccc} a_{0,0} & & & & a_{0,k} & & \\ & \ddots & & & & \ddots & \\ \vdots & & a_{0,0} & \cdots & \vdots & & a_{0,k} \\ & & \vdots & & & & \vdots \\ a_{\|n\|-1,0} & \cdots & a_{\|n\|-n_0,0} & & a_{\|n\|-1,k} & \cdots & a_{\|n\|-n_k,k} \end{array} \right) \tag{2.12}$$

be a generalized $\|n\| \times \|n\|$ block Sylvester matrix. Then the system (2.8) with $R_0(0) = 1$ in (2.7) corresponds to finding the solution X of

$$T_n \cdot X = (0, \ldots, 0, r_0)^t, \tag{2.13}$$

where $r_0 = R_0(0) = 1$ for a normalized PHS. Because the matrix T_n is square, its nonsingularity provides a sufficient condition for the existence of a solution for (2.13).

The pair $(P(z), Q(z))$ satisfying (2.7) and the degree requirements in (2.5), but not necessarily the condition that $R_0(0) = 1$, is commonly called a $(n_0-1, \ldots, n_k-1)$

Example 2.1 Let $A(z) = (A_0(z), A_1(z), A_2(z))$, where

$A_0(z) = 1 - z + 2z^2 - 2z^3 + 3z^4 - 3z^5 + 4z^6 - 4z^7 + 5z^8 - 5z^9 + \ldots$

$A_1(z) = 2z + 3z^3 + 4z^5 + 5z^7 + 6z^9 + \ldots$

$A_2(z) = -1 + z + 5z^2 + 3z^3 + 2z^4 - 2z^5 - 6z^6 + z^7 - 8z^8 + 5z^9 + \ldots$

and let $n = [2, 3, 1]$. Then $\|n\| = 6$ and the resulting Sylvester matrix is

$$T_n = \left(\begin{array}{rr|rrr|r} 1 & 0 & 0 & 0 & 0 & -1 \\ -1 & 1 & 2 & 0 & 0 & 1 \\ 2 & -1 & 0 & 2 & 0 & 5 \\ -2 & 2 & 3 & 0 & 2 & 3 \\ 3 & -2 & 0 & 3 & 0 & 2 \\ -3 & 3 & 4 & 0 & 3 & -2 \end{array} \right), \quad \det(T_n) = -37.$$

Using T_n, we compute the vector X as

$$X = \left(\frac{-4}{37}, \frac{44}{37}, \frac{-22}{37}, \frac{36}{37}, \frac{-9}{37}, \frac{-4}{37} \right)^t$$

resulting in

$$\begin{aligned} S_{0,0}(z) &= z^2 P(z) = \tfrac{-4}{37} z^2 + \tfrac{44}{37} z^3 \\ S_{1,0}(z) &= z^2 Q_1(z) = -\tfrac{22}{37} z^2 + \tfrac{36}{37} z^3 - \tfrac{9}{37} z^4 \\ S_{2,0}(z) &= z^2 Q_2(z) = -\tfrac{4}{37} z^2. \end{aligned} \tag{2.14}$$

□

Next we focus on computing the components $(U(z), V(z))$. From Definition 2.1 we have the requirement that

$$B(z) \cdot U(z) + C(z) \cdot V(z) = z^{\|n\|+1} R(z). \tag{2.15}$$

To determine $(U(z), V(z))$ in (2.15) we solve the homogeneous system

$$\left(\begin{array}{ccc|ccc|c|ccc}
a_{0,0} & & & a_{0,1} & & & & a_{0,k} & & \\
& \ddots & & & \ddots & & & & \ddots & \\
& & a_{0,0} & & & a_{0,1} & \cdots & & & a_{0,k} \\
& & \vdots & & & \vdots & & & & \vdots \\
a_{\|n\|,0} & \cdots & a_{\|n\|-n_0,0} & a_{\|n\|,1} & \cdots & a_{\|n\|-n_1,1} & & a_{\|n\|,k} & \cdots & a_{\|n\|-n_k,k}
\end{array}\right) \cdot \tilde{Y} = 0, \tag{2.16}$$

where

$$\tilde{Y} = \left(u_0^{(j)}, \ldots, u_{n_0}^{(j)} | v_{1,0}^{(j)}, \ldots, v_{1,n_1}^{(j)} | \cdots | v_{k,0}^{(j)}, \ldots, v_{k,n_k}^{(j)}\right)^t. \tag{2.17}$$

This system consists of $\|n\|+1$ equations and $(n_0+1)+\cdots+(n_k+1) = \|n\|+k+1$ unknowns. Therefore at least k linearly independent solutions must exist and (2.16) is solved for $j = 1, \ldots, k$. A solution to (2.15) obtained in this manner will satisfy conditions I and II of Definition 2.1 for columns $1, \ldots, k$ of $S(z)$, but not necessarily condition III. We will obtain $(U(z), V(z))$ with $V(0) = I_k$ in the following way.

For $j = 1, \ldots, k$, set

$$u_0^{(j)} = -\frac{a_{j,0}}{a_{0,0}}, \tag{2.18}$$

$$v_{i,0}^{(j)} = \begin{cases} 1, & i = j, \qquad i = 1, \ldots, k, \\ 0, & i \neq j. \end{cases} \tag{2.19}$$

Then the remaining components

$$Y = \left(\begin{array}{ccc|ccc|c|ccc}
u_1^{(1)} & \cdots & u_{n_0}^{(1)} & v_{1,1}^{(1)} & \cdots & v_{n_1,1}^{(1)} & & v_{1,k}^{(1)} & \cdots & v_{n_k,k}^{(1)} \\
\vdots & & \vdots & \vdots & & \vdots & \cdots & \vdots & & \vdots \\
u_1^{(k)} & \cdots & u_{n_0}^{(k)} & v_{1,1}^{(k)} & \cdots & v_{n_1,1}^{(k)} & & v_{1,k}^{(k)} & \cdots & v_{n_k,k}^{(k)}
\end{array}\right)^t \tag{2.20}$$

of Y in (2.16) satisfy

$$T_n \cdot Y = \begin{pmatrix} -a_{1,1} - a_{1,0}\left(\frac{-a_{0,1}}{a_{0,0}}\right) & \cdots & -a_{1,k} - a_{1,0}\left(\frac{-a_{0,k}}{a_{0,0}}\right) \\ \vdots & & \vdots \\ -a_{||n||,1} - a_{||n||,0}\left(\frac{-a_{0,1}}{a_{0,0}}\right) & \cdots & -a_{||n||,k} - a_{||n||,0}\left(\frac{-a_{0,k}}{a_{0,0}}\right) \end{pmatrix}. \qquad (2.21)$$

The $j-th$ component, $j = 1 \ldots k$, of the row vector $U(z)$ then corresponds to

$$U_j(z) = \frac{-a_{j,0}}{a_{0,0}} + \sum_{l=1}^{n_0} u_l^{(j)} \cdot z^l. \qquad (2.22)$$

Similarly, the $i,j-th$ component $i,j = 1,\ldots,k$ of the matrix $V(z)$ is given by

$$V_{i,j}(z) = \begin{cases} \sum_{l=1}^{n_i} v_{l,i}^{(j)} \cdot z^l, & i \neq j, \\ 1 + \sum_{l=1}^{n_i} v_{l,i}^{(j)} \cdot z^l, & i = j. \end{cases} \qquad (2.23)$$

Example 2.2 Continuing with Example 2.1, the system (2.21) becomes

$$T_n \cdot Y = \begin{pmatrix} -2 & 0 \\ 0 & -7 \\ -3 & -1 \\ 0 & -5 \\ -4 & 5 \\ 0 & 2 \end{pmatrix}$$

resulting in the solution

$$Y = \begin{pmatrix} -\frac{73}{37} & -\frac{44}{37} \\ -\frac{48}{37} & \frac{3}{37} \\ -\frac{13}{37} & -\frac{131}{37} \\ -\frac{9}{37} & \frac{137}{37} \\ -\frac{7}{37} & \frac{123}{37} \\ \frac{1}{37} & -\frac{44}{37} \end{pmatrix}. \qquad (2.24)$$

This yields the following terms

$$\begin{pmatrix} S_{0,1}(z) & S_{0,2}(z) \\ S_{1,1}(z) & S_{1,2}(z) \\ S_{2,1}(z) & S_{2,2}(z) \end{pmatrix} = \tag{2.25}$$

$$\begin{pmatrix} -\frac{73}{37}z - \frac{48}{37}z^2 & 1 - \frac{44}{37}z + \frac{3}{37}z^2 \\ 1 - \frac{13}{37}z - \frac{9}{37}z^2 - \frac{7}{37}z^3 & -\frac{131}{37}z + \frac{137}{37}z^2 + \frac{123}{37}z^3 \\ \frac{1}{37}z & 1 - \frac{44}{37}z \end{pmatrix}.$$

Hence from (2.14) and (2.25) the Padé-Hermite System $S(z)$ for $n = (2, 3, 1)$ is given by

$$\begin{pmatrix} \frac{-4}{37}z^2 + \frac{44}{37}z^3 & -\frac{73}{37}z - \frac{48}{37}z^2 & 1 - \frac{44}{37}z + \frac{3}{37}z^2 \\ -\frac{22}{37}z^2 + \frac{36}{37}z^3 - \frac{9}{37}z^4 & 1 - \frac{13}{37}z - \frac{9}{37}z^2 - \frac{7}{37}z^3 & -\frac{131}{37}z + \frac{137}{37}z^2 + \frac{123}{37}z^3 \\ -\frac{4}{37}z^2 & \frac{1}{37}z & 1 - \frac{44}{37}z \end{pmatrix}. \tag{2.26}$$

Note that $S(z)$ satisfies the degree bounds in Definition 2.1. In addition

$$A(z)\,S(z) = z^7 R(z),$$

where

$$R(z) = \left(1 + \frac{20}{37}z + \frac{42}{37}z^2 + \ldots, -\frac{5}{37} + \frac{8}{37}z - \frac{4}{37}z^2 + \ldots, \frac{516}{37} - \frac{130}{37}z + \frac{805}{37}z^2 + \ldots\right)$$

so that conditions II and III of Definition 2.1 are also satisfied.

□

When T_n is nonsingular, the systems (2.13) and (2.21) have unique solutions up to multiplication by some nonzero value of $\mathcal{F}$. Cabay et al. [10] showed that $\det(T_n) \neq 0$ provides both a necessary and sufficient condition for the existence of a Padé-Hermite System of type n.

We have seen that even if T_n is singular, the components $(P(z), Q(z))$ and $(U(z), V(z))$ can still be computed by solving a set of homogeneous equations (2.13) and (2.21). These solutions may not be unique and may no longer satisfy condition III of Definition 2.1. In this case, it may be of interest to construct a basis for the solution space of (2.8) which in turn provides a basis for all Padé-Hermite Forms for the given vector n and power series $A(z)$. We will further discuss the relevance of this in Section 2.3.

2.2 A Recurrence Relation

In the previous section we have seen that, for a given vector of integers n and vector of power series $A(z)$, the computation of a Padé-Hermite System equivalent to solving sets of linear equations. These systems (2.13) and (2.21) can be computed using a method such as Gaussian elimination. Computing a PHS in this manner, while not restricting the input vector of power series, does not take advantage of the inherent structure of the coefficient matrix T_n.

Cabay et al. [10] describe a recurrence relation for computing Padé-Hermite Systems. This recurrence relation usually leads to a more efficient algorithm for computing the PHS of any type. The purpose of this section is to briefly describe this recurrence relation (using a more concise notation).

Given a vector of power series (2.1) and a vector $n = (n_0, \ldots, n_k)$ of nonnegative integers, permute the components of $A(z)$ and n so that

$$A_0(0) \neq 0, \qquad n_1 \geq \cdots \geq n_k. \tag{2.27}$$

Notice that, if $A_i(z) = 0$ for all $0 \leq i \leq k$, then by removing the largest factor z^β from each power series and reordering, (2.27) will be satisfied. A PHS of type n for

$(z^{-\beta} \cdot A_0(z), \ldots, z^{-\beta} \cdot A_k(z))$ is also a PHS of the same type for $(A_0(z), \cdots, A_k(z))$

Let $e_0 = (1,0,\ldots,0)$ be a $1 \times k+1$ vector. Set $M = \min\{n_0, n_1\} + 1$ and define integer vectors $n^{(i)} = (n_0^{(i)}, \ldots, n_k^{(i)})$ for $0 \le i \le M$ by $n^{(0)} = -e_0$ and, for $i > 0$,

$$n_j^{(i)} = \max\{0, n_j - M + i\}, \qquad j = 0, \ldots, k. \tag{2.28}$$

Then the sequence $\{n^{(i)}\}_{i=0,1,\ldots}$ lies on a piecewise linear path with $n_j^{(i+1)} \ge n_j^{(i)}$ for each i, j and

$$\begin{aligned} n^{(0)} &= -e_0, & (2.29)\\ n^{(1)} &= \begin{cases} (0, n_1 - n_0, \ldots), & n_1 \ge n_0, \\ (n_0 - n_1, 0, \ldots, 0), & n_1 < n_0, \end{cases} & (2.30)\\ n^{(M)} &= (n_0, \ldots, n_k) = n. & (2.31)\end{aligned}$$

The sequence $\{n^{(\sigma)}\}$ contains a subsequence $\{m^{(i)}\}$ called the **sequence of nonsingular points.** This sequence is defined by $m^{(i)} = n^{(\sigma_i)}$ where

$$\sigma_i = \begin{cases} 0, & i = 0, \\ \min\{\sigma > \sigma_{i-1} : \det(T_{n^{(\sigma)}}) \ne 0\}, & i \ge 1. \end{cases} \tag{2.32}$$

Observe that the ordering (2.27) implies $m_0^{(i)} \ge 0$ for all $i \ge 1$. Therefore, for all $0 < \sigma_i < \sigma$ it is true that

$$\sigma - \sigma_i = n_0^{(\sigma)} - m_0^{(i)}. \tag{2.33}$$

Corresponding to the sequence of nonsingular points $\{m^{(i)}\}$, is a sequence $\left\{S^{(i)}(z)\right\}$ of PHS's with residuals $\left\{R^{(i)}(z)\right\}$.

Two special cases of n which are excluded from Definition 2.1 arise in the process of developing the algorithm. As a new but small contribution, and for algorithmic purposes, we provide definitions to deal with these special cases. Note that the systems as defined below are, strictly speaking, not Padé-Hermite Systems. They act as initial conditions for a recurrence relation given later.

The first case occurs when $n^{(0)} = m^{(0)} = -e_0$. We define the corresponding matrix

$$S^{(0)}(z) = I_{k+1}, \tag{2.34}$$

where I_{k+1} denotes the $(k+1) \times (k+1)$ identity matrix. Also note

$$A(z)\, S^{(0)}(z) = z^{\|n^{(0)}\|+1}\, R^{(0)}(z) = A(z).$$

Thus $S^{(0)}(z)$ satisfies conditions I and II, and III of Definition 2.1. However, $S^{(0)}(z)$ is, strictly speaking, not a PHS since the factor z^2 cannot be removed from the first column giving the polynomials $P(z)$ and $Q_j(z)$ according to (2.4)

The second case occurs when $n = (s, s, \ldots, s)$ for some integer s. In this case, $n^{(1)} = m^{(1)} = (0, 0, \ldots, 0)$. In this case let

$$S^{(1)}(z) = \begin{pmatrix} z & \frac{-a_{0,1}}{a_{0,0}} & & \frac{-a_{0,k}}{a_{0,0}} \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \\ 0 & & & 1 \end{pmatrix}. \tag{2.35}$$

Clearly from (2.35),

$$A(z) \cdot S^{(1)}(z) = z\, (R_0(z), \ldots R_k(z)),$$

where $R_i(z)$, $i = 0, \ldots, k$, are power series so condition II of Definition 2.1 is satisfied. By inspection we can see that the $R_0(0) \neq 0$ and that $V(0) = I_k$ so Definition 2.1 is satisfied. Finally, it is easy to verify that the degree condition I is also satisfied. Once again, $S^{(1)}(z)$ is not a PHS due to the absence of a common factor z^2 cannot be removed from the first column.

Adjusting Definition 2.1 to include the exceptions (2.34) and (2.35), we have that for $i \geq 0$, the polynomial matrix $S^{(i)}(z)$ is the uniquely defined PHS of type $m^{(i)}$ with residual vector $R^{(i)}(z)$. Then

$$A(z) \cdot S^{(i)}(z) = z^{\|m^{(i)}\|+1} R^{(i)}(z) \tag{2.36}$$

and $R_0^{(0)}(0) = 1$, $V^{(0)} = I_k$. In addition, with the exceptions (2.34) and (2.35) we can partition $S^{(i)}(z)$ as

$$S^{(i)}(z) = \begin{pmatrix} z^2 P^{(i)}(z) & U^{(i)}(z) \\ z^2 Q^{(i)}(z) & V^{(i)}(z) \end{pmatrix}, \tag{2.37}$$

and $R_0^{(0)}(0) = 1$, $V^{(0)} = I_k$.

The following theorem provides a relationship of the $(i+1)$-st PHS of the sequence in terms of the i-th PHS.

Theorem 2.1 (Cabay et al. [10]) *For $i \geq 1$ and $\sigma > \sigma_i$, let $\nu = n^{(\sigma)} - m^{(i)} - e_0$. Then $n^{(\sigma)}$ is a nonsingular point for $A = (A_0(z), \ldots, A_k(z))$ if and only if ν is a nonsingular point for $R^{(i)}(z) = (R_0^{(i)}(z), \ldots, R_k^{(i)}(z))$. Furthermore we have the recurrence relations*

$$S^{(i+1)}(z) = S^{(i)}(z) \cdot \hat{S}(z) \quad \text{and} \quad R^{(i+1)}(z) = \hat{R}(z) \tag{2.38}$$

where $\hat{S}(z)$ is the PHS of type $(m^{(i+1)} - m^{(i)} - e_0)$ for the system $R^{(i)}(z)$ and $\hat{R}(z)$ is its residual.

Proof: The proof is given in [10].

□

Theorem 2.1 reduces the problem of determining a PHS of type n to two smaller problems: determine a PHS of type $m^{(i)}$ and then determine a PHS of type $\nu =$

$m^{(i+1)} - m^{(i)} - e_0$. Let $R^{(i)}(z)$ be a $1 \times (k+1)$ vector of power series with

$$R_j^{(i)}(z) = \sum_{l=0}^{\infty} r_{l,j} z^l \qquad j = 0, \ldots, k. \tag{2.39}$$

Then define

$$\hat{T}_\nu^{(i)} = \left(\begin{array}{ccc|c|ccc} r_{0,0} & & & & r_{0,k} & & \\ & \ddots & & & & \ddots & \\ \vdots & & r_{0,0} & \cdots & \vdots & & r_{0,k} \\ & & \vdots & & & & \vdots \\ r_{\|n\|-1,0} & \cdots & r_{\|n\|-n_0,0} & & r_{\|n\|-1,k} & \cdots & r_{\|n\|-n_k,k} \end{array}\right). \tag{2.40}$$

To compute the PHS $\hat{S}(z)$ we solve, assuming $\hat{T}_\nu^{(i)}$ is nonsingular, the equations

$$\hat{T}_\nu^{(i)} \cdot X = (0, \ldots, 0, r_0)^t, \tag{2.41}$$

$$\hat{T}_\nu^{(i)} \cdot Y = \begin{pmatrix} -r_{1,1} - r_{1,0}\left(\frac{-r_{0,1}}{r_{0,0}}\right) & \cdots & -r_{1,k} - r_{1,0}\left(\frac{-r_{0,k}}{r_{0,0}}\right) \\ \vdots & & \vdots \\ -r_{\|n\|,1} - r_{\|n\|,0}\left(\frac{-r_{0,1}}{r_{0,0}}\right) & \cdots & -r_{\|n\|,k} - r_{\|n\|,0}\left(\frac{-r_{0,k}}{r_{0,0}}\right) \end{pmatrix}. \tag{2.42}$$

The solutions are combined as in (2.10), (2.11), (2.22), and (2.23) to form $\hat{S}(z)$.

The overhead cost of each step of this iterative scheme is the cost of determining the residual power series and the cost of combining the solutions, i.e. the cost of computing $S^{(i+1)}(z)$ in equation (2.38). This overhead cost, in general, is an order of magnitude less than the cost of simply solving the linear systems (2.41) and (2.42).

Theorem 2.1 does not give consideration to the cases (2.34) and (2.35). The following discussion is intended to show that these cases support the theorem.

Consider first the case (2.34). Let $A(z)$ be a $1 \times k$ vector of power series and n a vector of integers both arranged according to (2.27). Let $m^{(0)} = -e_0$. From (2.34), $S^{(0)}(z) = I_{k+1}$ and $R^{(0)}(z) = A(z)$ is its residual. Let $\hat{S}(z)$ be the PHS of type

$m^{(1)} - m^{(0)} - e_0 = m^{(1)}$ for $R^{(0)}(z)$ and let $\hat{R}(z)$ be its residual. The degree conditions (2.5) for $S^{(1)}(z)$ correspond to those of $\hat{S}(z)$ so condition I of Definition 2.1 is satisfied. In addition,

$$\begin{aligned} A(z)S^{(1)}(z) &= A(z)S^{(0)}(z)\hat{S}(z) \\ &= R^{(0)}(z)\hat{S}(z) \\ &= z^{\|m^{(1)} - m(0) - e_0\|+1}\, \hat{R}(z) \\ &= z^{\|m^{(1)}\|+1}\, \hat{R}(z). \end{aligned}$$

Thus condition II holds. Finally, $R^{(1)}(z) = \hat{R}(z) = 1$ and $V^{(1)}(z) = \hat{V}(z) = I_k$ so condition III of Definition 2.1. The recurrence relation (2.38) holds for this special case.

For the case (2.35), let $n = (x, x, \ldots, x)$ for some nonnegative integer x. From (2.28) we can see that $m^{(1)} = (0, \cdots, 0)$ is the second vector in the sequence. Let $S^{(1)}(z)$ be given by (2.35) with $R^{(1)}(z)$ its residual and let $\hat{S}(z)$ be the PHS of type $\nu = (s-1, s, \ldots, s)$, $s \geq 1$, for $R^{(1)}(z)$ with $\hat{R}(z)$ its residual. We will show that conditions I, II and III of Definition 2.1 are satisfied for $S^{(2)}(z) = S^{(1)}(z) \cdot \hat{S}(z)$ of type $m^{(2)} = (s, s, \ldots, s)$ where $s \leq x$. From this we can then conclude that the recurrence relation (2.38) holds. To verify condition I observe that

$$\partial S^{(0)}(z) \leq \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

and

$$\partial\hat{S}(z) \leq \begin{pmatrix} s & s-1 & \cdots & s-1 \\ s+1 & s & \cdots & s \\ \vdots & \vdots & & \vdots \\ s+1 & s & \cdots & s \end{pmatrix}$$

which implies

$$\partial S^{(2)}(z) = \begin{pmatrix} s+1 & s & \cdots & s \\ s+1 & s & \cdots & s \\ \vdots & \vdots & & \vdots \\ s+1 & s & \cdots & s \end{pmatrix}.$$

The degree conditions above corresponds to a PHS of type $m^{(2)} = (s, \ldots, s)$. Next, condition II is satisfied because

$$\begin{aligned} A(z)\, S^{(2)}(z) &= A(z)\, S^{(1)}(z)\, \hat{S}(z) \\ &= z \cdot R^{(1)}(z) \cdot \hat{S}(z) \\ &= z \cdot z^{\|\nu\|+1} \hat{R}(z) \\ &= z^{\|m^{(2)}\|+1} \hat{R}(z). \end{aligned}$$

Finally,

$$R^{(2)}(0) = \hat{R}(0) = 1$$

and

$$V^{(2)}(0) = V^{(1)}(0) \cdot \hat{V}(0) = I_k$$

so that condition III is also satisfied.

Example 2.3 Continuing with Example 2.1 we can compute the PHS of type $(3,4,2)$ by utilizing the recurrence relation (2.38) and (2.26). In order to do this we must compute the PHS of type $\nu = (3,4,2)-(2,3,1)-(1,0,0) = (0,1,1)$. The coefficients of the matrix $\hat{T}_\nu$ will come from the residual $R(z)$ given by (2.27). The matrix $\hat{T}_\nu$ will be given by

$$\hat{T}_\nu = \begin{pmatrix} -\frac{5}{37} & \frac{516}{37} \\ \frac{8}{37} & -\frac{130}{37} \end{pmatrix}. \tag{2.43}$$

Using (2.43), equations (2.41) and (2.42) are solved to obtain the PHS

$$\hat{S}(z) = \begin{pmatrix} 0 & \frac{105}{148} & -\frac{3096}{4255} \\ \frac{2064}{1739}z^2 & 1-\frac{24}{37}z & -\frac{10432}{60865}z \\ \frac{4025}{3478}z^2 & -\frac{805}{296}z & 1+\frac{2175}{3478}z \end{pmatrix}. \tag{2.44}$$

By multiplying the $S(z)$ in (2.26) on the right by $\hat{S}(z)$ we obtain the new PHS of type $(3,4,2)$,

$$S(z) = \begin{pmatrix} \frac{5}{94}z^2 - \frac{512}{47}z^3 - \frac{669}{94}z^4 & -2z+z^3 & 1-\frac{53}{94}z+\frac{1639}{47}z^2+\frac{549}{94}z^3 \\ \frac{258}{47}z^2 - \frac{199}{94}z^3 - \frac{107}{94}z^4 - \frac{81}{94}z^5 & 1-z & -\frac{977}{47}z+\frac{1489}{94}z^2-\frac{351}{94}z^3+\frac{821}{94}z^4 \\ \frac{5}{94}z^2+\frac{4}{47}z^3 & 0 & 1-\frac{53}{94}z+\frac{14}{47}z^2 \end{pmatrix}.$$

□

2.3 Interpretations of the Padé-Hermite System

Padé-Hermite approximants have many applications, some of which are given in the introduction. It is the primary focus of this thesis to compute them in a stable fashion.

The first column $(S_{0,0}(z), \ldots, S_{k,0}(z))^t$ of $S(z)$ is commonly referred to as a **Padé-Hermite Form (PHFo)** of type $(n_0-1,\ldots,n_k-1)$ for $A(z)$ [10] [28] [25]. The terms Padé-Hermite Form and Padé-Hermite approximant are used interchangeably.

The remaining k columns of the PHS are referred to as a **Weak Padé-Hermite Form (WPHFo)** [10]. The main purpose for introducing these remaining components $(U(z), V(z))$ of the Padé-Hermite System $S(z)$ is to facilitate the computation of $(P(z), Q(z))$. There are however some uses for all the components of $S(z)$ such as that of obtaining the closed form inverse of a block Hankel matrix [20].

In addition, WPHFo's yield a set of simultaneous Padé approximants for the quotient power series $A_i(z) / A_0(z)$. We can restate condition II of (2.1) as

$$A_0(z)\,U(z) + (A_1(z), \ldots, A_k(z))\,V(z) = z^{\|n\|+1}\,(R_0(z), \ldots, R_k(z)); \qquad (2.45)$$

that is,

$$A_0(z)\,U(z) + (A_1(z), \ldots, A_k(z))\,V(z) = 0 \mod z^{\|n\|+1}. \qquad (2.46)$$

Since $V(0)$ is nonsingular, the determinant $D(z) = \det(V(z))$ evaluated at $z = 0$ is nonzero. Thus, since $A_0(0)$ is also nonzero, it follows from (2.46) that

$$\left(\frac{A_1(z)}{A_0(z)}, \ldots, \frac{A_k(z)}{A_0(z)}\right) = \frac{N(z)}{D(z)} \mod z^{\|n\|+1}, \qquad (2.47)$$

where

$$N(z) = -U(z) \cdot adjoint(V(z)). \qquad (2.48)$$

Equations (2.47) and (2.48) give a simultaneous rational approximation for each power series

$$\frac{A_i(z)}{A_0(z)} \approx \frac{N_i(z)}{D(z)}, \quad i = 1, \ldots, k. \qquad (2.49)$$

From the degree conditions of (2.5) we see that $N_i(z)$ has at most degree $\|n\| - n_i$ and that D(z) has at most degree $\|n\| - n_0$. Therefore the polynomials

$$(D(z), N_1(z), \ldots, N_k(z)) \qquad (2.50)$$

form a set of simultaneous Padé approximants to the power series'

$$\frac{A_1(z)}{A_0(z)}, \ldots, \frac{A_k(z)}{A_0(z)}$$

of type n.

Example 2.4 Continuing with Example 2.1 and using the PHS (2.26) we have

$$U(z) = \left(-\tfrac{73}{37}z - \tfrac{48}{37}z^2, 1 - \tfrac{44}{37}z + \tfrac{3}{37}z^2\right),$$

$$V(z) = \begin{pmatrix} 1 - \frac{13}{37}z - \frac{9}{37}z^2 - \frac{7}{37}z^3 & -\frac{131}{37}z + \frac{137}{37}z^2 + \frac{123}{37}z^3 \\ \frac{1}{37}z & 1 - \frac{44}{37}z \end{pmatrix}.$$

We can compute the determinant of $V(z)$ as

$$\det(V(z)) = 1 - \frac{57}{37}z + \frac{10}{37}z^2 + \frac{5}{37}z^4.$$

Likewise, the adjoint of $V(z)$ is given by

$$adjoint(V(z)) = \begin{pmatrix} 1 - \frac{44}{37}z & \frac{131}{37}z - \frac{137}{37}z^2 - \frac{123}{37}z^3 \\ -\frac{1}{37}z & 1 - \frac{13}{37}z - \frac{9}{37}z^2 - \frac{7}{37}z^3 \end{pmatrix}.$$

The approximants are therefore given by

$$\frac{A_1(z)}{A_0(z)} = \frac{-2z + \frac{40}{37}z^2 + \frac{57}{37}z^3}{1 - \frac{57}{37}z + \frac{10}{37}z^2 + \frac{5}{37}z^4} \mod z^{\|n\|+1},$$

$$\frac{A_2(z)}{A_0(z)} = \frac{1 - \frac{57}{37}z - \frac{249}{37}z^2 + \frac{103}{37}z^3 + \frac{428}{37}z^4 + \frac{159}{37}z^5}{1 - \frac{57}{37}z + \frac{10}{37}z^2 + \frac{5}{37}z^4} \mod z^{\|n\|+1}$$

when here $\|n\| = 6$.

□

Given that $S^{(i)}(z)$ is a PHS of type $m^{(i)}$, elements of the residual $R^{(i)}$ are used to form the matrix $\hat{T}_\nu^{(i)}$. Because $\hat{T}_\nu^{(i)}$ is generally small in size, (2.13) and (2.42) can be solved quickly.

In the worst case, all points in the Padé-Hermite Table will be singular with the exception of the final point $n = n^{(M)}$. The problem reduces to that of solving (2.13) and (2.21) with Gaussian elimination.

The optimal case occurs when there are no singular points. In such a situation, $\nu = (0, 1, \ldots, 1)$ and $\hat{T}_{\nu}^{(i)}$, is a $k \times k$ square matrix for all $i \leq M$. The systems (2.41) and (2.42) are solved using Gaussian elimination with pivoting and the result multiplied on the left by $S^{(i)}(z)$ to form $S^{(i+1)}(z)$.

Recall that in Section 2.1, the homogeneous equations (2.8) and (2.16) could be solved to obtain solutions for $(P(z), Q(z))$ and $(U(z), V(z))$. We modified these homogeneous equations in order to satisfy some additional requirements of the PHS, namely $R_0(0) = 1$ and $V(0) = I_k$. Regardless of whether T_n is singular or not, a solution to the homogeneous systems will exist because they are under-determined. These systems may have an infinite number of solutions due to the presence of one or more arbitrary parameters. Note that the systems (2.41) and (2.42) have corresponding homogeneous formulations.

If arbitrary parameters exist in the solution of the homogeneous systems, they do not yield a Padé-Hermite System. Instead, a basis for the solution space of (2.41) could be used to form a basis of all the PHFo's of type n. Theorem 2.1 does not permit subsequent PHS's to be computed using a basis. Therefore, Padé-Hermite table points for which $\hat{T}_{\nu}^{(i)}$ is singular, are jumped over by the iterative algorithm. Jumping over singular blocks in the Padé table allows the PHS at a nonsingular point $m^{(i+1)}$ in the Padé-Hermite table to be uniquely determined using the previously computed system $S^{(i)}(z)$ and the solutions of (2.41) and (2.42). If the last point $n^{(M)} = n$ in the sequence is singular, the basis for the solution space of (2.8) may be of practical use to compute.

2.4 An Iterative Algorithm For Computing A PHS

Next we introduce the algorithm of Cabay et al. [10] for iteratively computing Padé-Hermite Systems. The notation has been modified to conform to previous definitions. The special case where $n = [s, \ldots, s]$ has also been added. If $\hat{T}^{(i)}_{\nu}$ is singular for all remaining points in the Padé table, the associated homogeneous equations are not solved.

The algorithm consists of two parts. The subroutine INITIAL_PH takes as input a vector of power series $R(z)$, with $R_0(0) \neq 0$ and an integer vector ν ordered according to (2.27). If $\nu^{(1)} \neq [0, \ldots, 0]$, the PHS at the first nonsingular point is returned (if one exists). If $\nu^{(1)} = [0, \ldots, 0]$ the matrix returned is that given in (2.35). The identity matrix is returned if no nonsingular point is encountered.

The main routine, PADE_HERMITE invokes INITIAL_PH to iteratively construct PHS for residuals $R^{(i)}(z)$. The PHS $S^{(i)}(z)$ are computed according to Theorem 2.1. In the case where INITIAL_PH does not return a PHS, PADE_HERMITE returns the last successfully computed PHS.

```
INITIAL_PH($R(z)$, $\nu$)
  I-1) $d \leftarrow 0, \quad M \leftarrow \max\{\nu_0, \nu_1\} + 1, \quad \sigma \leftarrow 0$
  I-2) Do while $\sigma < M$ and $d = 0$
  I-3)      $\sigma \leftarrow \sigma + 1$
  I-4)      $\nu_j^{(\sigma)} \leftarrow \max\{0, \nu_j - M + \sigma\}, j = 0, \ldots, k$
  I-5)      If $\|\nu^{(\sigma)}\| \neq 0$ then compute $d \leftarrow \det(\hat{T}_{\nu^{(\sigma)}})$, using Gaussian elimination;
            else set $d \leftarrow 1$
        End While
  I-6) If $d \neq 0$ then set $\hat{S}(z)$ according to (2.35) or solve equations (2.41)
```

```
        and (2.42) for $R(z)$ and arrange the solutions into a $(k+1) \times (k+1)$
        matrix $\hat{S}(z)$, the PHS of type $\nu^{(\sigma)}$ for $R(z)$;
        else set $\hat{S}(z) = I_{k+1}$ and $\sigma \leftarrow M + 1$.
  I-7) Return($\sigma, \hat{S}(z)$).

PADE_HERMITE($A(z)$, $n$)
  PH-1) Find the largest $\beta$ such that $A_i(z) = z^{\beta} \cdot \hat{A}_i(z)$ are still power series.
        Set $A_i(z) = z^{-\beta} \cdot A_i(z)$. Reorder the power series according to (2.27).
  PH-2) $M \leftarrow \min\{n_0, n_1\} + 1$
  PH-3) $S^{(0)}(z)) \leftarrow I_{k+1}$,   $\sigma \leftarrow 0$,   $m^{(0)} \leftarrow -e_0$,   $i \leftarrow 0$
  PH-4) While $\sigma \leq M$ do
  PH-5)     Determine $R^{(i)}(z)$ using equation 2.36, $\nu \leftarrow n - m^{(i)} - e_0$
  PH-6)     $(s_i, \hat{S}(z)) \leftarrow$ INITIAL_PH($R^{(i)}, \nu$)
  PH-7)     $S^{(i+1)}(z) \leftarrow S^{(i)}(z) \cdot \hat{S}(z)$
  PH-8)     $\sigma \leftarrow \sigma + s_i$,  $m_j^{(i+1)} \leftarrow \max\{0, n_j - M + \sigma\}$,  $j = 0, \ldots, k$,
            $i \leftarrow i + 1$
        End While
  PH-9) Return $\left(\sigma, S^{(i)}(z)\right)$.
```

In order to evaluate the performance of the above algorithm, it was implemented in the MAPLE symbolic computation environment. MAPLE provides a built-in programming language and comprehensive library of common mathematical functions to facilitate rapid prototyping of algorithms. In addition, calculations can be carried out in exact arithmetic. Since the algorithm is algebraic in nature, the MAPLE setting is well suited to simulate performance.

As expected, the algorithm accurately computed Padé-Hermite Systems of varying size and degree. Real time performance of the system in the MAPLE environment was unfortunately, very slow.

A number of factors contributed to the poor performance of the algorithm in the MAPLE environment. Intermediate expression swell in solving the systems (2.41) and (2.42) made the memory requirements of MAPLE significant. Continuous paging by the computer often resulted. Numerous expensive gcd computations were required by MAPLE to minimize the number of digits stored for each polynomial coefficient. This overhead began to dominate as the parameters $\|n\|$ and k were increased.

Despite its shortcomings, the MAPLE implementation served to illustrate some important points about the nature of the coefficients of a PHS and the growth of the residual vector $R(z)$. In addition, MAPLE played an invaluable role in developing a method for fabricating power series which contained singular points.

However, for any practical application to utilize Padé-Hermite Systems, a numerical algorithm is necessary. The algorithm must be numerically stable because of the finite precision associated with floating point arithmetic. The primary goal of this thesis is to develop, without proof, a numerical algorithm and test it experimentally.

Chapter 3

Towards A Stable Numerical Algorithm

In the last chapter we saw how Padé-Hermite Systems could be computed algebraically by skipping points at which the Sylvester matrix $T_{n(\sigma)}$ was singular. The goal of a similar numerical algorithm is to compute Padé-Hermite Systems at all points except those for which $T_{n(\sigma)}$ is ill-conditioned. If the condition number $k(T_{n(\sigma)}) = \|T_{n(\sigma)}\| \cdot \|T_{n(\sigma)}^{-1}\|$ is large, the resulting components $P(z)$, $Q(z)$, $U(z)$, and $V(z)$ may contain large errors. Although we can compute $k(T_{n(\sigma)})$ directly, this is impractical as it involves inverting the matrix $T_{n(\sigma)}$.

The purpose of this chapter is to provide a simpler measure by which we can estimate the condition number of $T_{n(\sigma)}$ and determine if the point $n^{(\sigma)}$ in the Padé-Hermite table is stable. A stability parameter which can be easily computed from the Padé-Hermite System $S^{(i)}(z)$ will be given.

Before proceeding, we must define the notion of stability as it applies to Padé-Hermite Systems. Computing a PHS involves the solving of linear systems. A common definition of algorithmic stability is given by the following.

Definition 3.1 (Bunch [7]) *An algorithm for solving linear equations is* **stable** *for a class of matrices $\mathcal{M}$ if for each M in $\mathcal{M}$ and for each b, the computed solution x_c of $Mx = b$ satisfies $\hat{M}x_c = \hat{b}$, where $\hat{M}$ is close to M and $\hat{b}$ is close to b.*

□

We can think of $\mathcal{M}$ as the class of all nonsingular block Sylvester matrices. Definition 3.1 implies that the computed solution x_c as the exact solution of a *nearby* problem. Note that the condition of the problem is not constrained in the definition.

For an application that uses Padé-Hermite Systems, it is more important to obtain solutions that are close to the actual solution. It is common knowledge that the accuracy of a solution of a linear system depends not only on the method used to solve it, but also on the condition of the problem. Merely formulating a poorly conditioned problem numerically can introduce perturbations in the solution. If the problem is well-conditioned, we would like to compute an accurate solution.

By only solving well conditioned systems, we can construct an algorithm which is *weakly stable.* Weak stability is defined as follows.

Definition 3.2 (Bunch [7]) *An algorithm for solving linear equations is* **weakly stable** *for a class of matrices $\mathcal{M}$ if for each well-conditioned M in $\mathcal{M}$ and for each b, the computed solution x_c to $M\,x = b$ is such that $\|x - x_c\|/\|x\|$ is small.*

□

Although stability implies weak stability, the converse is not true. Weak stability says nothing about those problems which are ill-conditioned.

Thus by choosing to solve equations (2.13) and (2.21) for all well conditioned matrices $T_{n(\sigma)}$, we expect the relative error in the components of $S(z)$ to be small.

3.1 Power Series and Polynomial Matrix Norms

We begin by introducing power series and matrix polynomial norms used throughout the thesis and proving their compatibility. We assume that $\mathcal{F}$ is the field of complex numbers so that the norm of an element of $\mathcal{F}$ is the absolute value. Let

$$a(z) = \sum_{j=0}^{\infty} a_j z^j \in \mathcal{F}[[z]], \tag{3.1}$$

where $\mathcal{F}[[z]]$ is the set of power series with coefficients from the field $\mathcal{F}$.

Lemma 3.1 *A norm for $\mathcal{F}[[z]]$ is given by*

$$\|a(z)\| = \sup_{0 \le j < \infty} \{| a_j |\}. \tag{3.2}$$

Proof: If $a(z) \in \mathcal{F}[[z]]$, then

$$\begin{aligned} \|a(z)\| = 0 \quad &\Leftrightarrow \quad \sup_{0 \le j < \infty} \{| a_j |\} = 0 \\ &\Leftrightarrow \quad a_j = 0, \quad 0 \le j < \infty \\ &\Leftrightarrow \quad a(z) = 0. \end{aligned} \tag{3.3}$$

Also, for $\alpha \in \mathcal{F}$

$$\begin{aligned} \|\alpha\, a(z)\| &= \sup_{0 \le j < \infty} \{| \alpha\, a_j |\} \\ &= \sup_{0 \le j < \infty} \{|\alpha| \, | a_j |\} \\ &= |\alpha| \cdot \sup_{0 \le j < \infty} \{| a_j |\} \\ &= |\alpha| \cdot \|a(z)\|. \end{aligned} \tag{3.4}$$

Finally, for $a(z),\ b(z) \in \mathcal{F}[[z]]$,

$$\|a(z) + b(z)\| = \sup_{0 \le j < \infty} \{| a_j + b_j |\}$$

$$
\begin{aligned}
&\leq \sup_{0\leq j<\infty}\{|\, a_j \,| + |\, b_j \,|\} \\
&\leq \sup_{0\leq j<\infty}\{|\, a_j \,|\} + \sup_{0\leq l<\infty}\{|\, b_l \,|\} \\
&= \|a(z)\| + \|b(z)\| \qquad (3.5)
\end{aligned}
$$

□

For some integer ∂, let

$$s(z) = \sum_{j=0}^{\partial} s_j\, z^j \in \mathcal{F}[z], \qquad (3.6)$$

where $\mathcal{F}[z]$ is the set of polynomials with coefficients over $\mathcal{F}$. Then, as in the proof of Lemma 3.1, it is easy to show that a norm of $s(z)$ is

$$\|s(z)\| = \sum_{j=0}^{\partial} |\, s_j \,| \,. \qquad (3.7)$$

Lemma 3.2 *Let $a(z)$ be given by (3.1) and $s(z)$ by (3.6). Then*

$$\|a(z)\cdot s(z)\| \leq \|a(z)\| \cdot \|s(z)\|.$$

Proof: Conventionally, let $a_i = 0$ for $i < 0$. Then

$$
\begin{aligned}
\|a(z)\cdot s(z)\| &= \left\| \sum_{i=0}^{\infty} \left(\sum_{j=0}^{\partial} a_{i-j}\, s_j \right) z^i \right\| \\
&= \sup_{0\leq i<\infty} \left\{ \left| \sum_{j=0}^{\partial} a_{i-j}\, s_j \right| \right\} \\
&\leq \sup_{0\leq i<\infty} \left\{ \sum_{j=0}^{\partial} |\, a_{i-j} \,| \cdot |\, s_j \,| \right\} \\
&\leq \sup_{0\leq i<\infty} \left\{ \sum_{j=0}^{\partial} \left(\sup_{0\leq l<\infty} |\, a_l \,| \right) |\, s_j \,| \right\} \\
&= \sup_{0\leq l<\infty} \{|\, a_l \,|\}. \sum_{j=0}^{\partial} |\, s_j \,| \\
&= \|a(z)\| \cdot \|s(z)\|
\end{aligned}
$$

□

Thus, the norm (3.7) for $\mathcal{F}[z]$ is compatible with the norm (3.2) for $\mathcal{F}[[z]]$. In addition, for fixed $s(z)$, the bound is reached for $a(z) = 1$. Therefore,

$$\|s(z)\| = \sup_{a(z)\neq 0} \frac{\|a(z)\, s(z)\|}{\|a(z)\|}. \tag{3.8}$$

Thus, (3.7) is the operator norm for $\mathcal{F}[z]$ induced by the norm (3.2) for $\mathcal{F}[[z]]$.

Now let $A(z) \in \mathcal{F}_{k+1}[[z]]$ be a $1 \times k+1$ vector of power series. That is, let $A(z) = (A_0(z), \ldots, A_k(z))$ where

$$A_i(z) = \sum_{j=0}^{\infty} a_{j,i} z^j, \quad i = 0, \ldots, k.$$

Lemma 3.3 *A norm for $A(z)$ is given by*

$$\|A(z)\| = \max_{0\leq i\leq k} \{\|A_i(z)\|\}. \tag{3.9}$$

Proof: If $A(z) \in \mathcal{F}_{k+1}[[z]]$, then using (3.3), we have that

$$\begin{aligned} \|A(z)\| = 0 \iff & \max_{0\leq i\leq k} \{\|A_i(z)\|\} = 0 \\ \iff & \|A_i(z)\| = 0 \qquad i = 0, \ldots, k \\ \iff & A_i(z) = 0 \qquad i = 0, \ldots, k. \end{aligned}$$

Now let $\alpha \in \mathcal{F}$. Then, using (3.4), it follows that

$$\begin{aligned} \|\alpha\, A(z)\| &= \max_{0\leq i\leq k} \{\|\alpha\, A_i(z)\|\} \\ &= \max_{0\leq i\leq k} \{|\, \alpha\, | \cdot \|A_i(z)\|\} \\ &= |\, \alpha\, | \cdot \max_{0\leq i\leq k} \{\|A_i(z)\|\} \\ &= |\, \alpha\, | \cdot \|A(z)\|. \end{aligned}$$

Finally, if, in addition, $B(z) \in \mathcal{F}_{k+1}[[z]]$, then using (3.5), we have that

$$\|A(z) + B(z)\| = \max_{0\leq i\leq k} \{\|A_i(z) + B_i(z)\|\}$$

$$
\begin{aligned}
&\leq \max_{0\leq i\leq k}\{\|A_i(z)\| + \|B_i(z)\|\}\\
&\leq \max_{0\leq i\leq k}\{\|A_i(z)\|\} + \max_{0\leq j\leq k}\{\|B_j(z)\|\}\\
&= \|A(z)\| + \|B(z)\|.
\end{aligned}
$$

□

Let $S(z)^1 \in \mathcal{F}_{(k+1)\times(k+1)}[z]$. Then $S(z)$ defines a mapping

$$
\begin{aligned}
\mathcal{F}_{(k+1)}[[z]] &\Rightarrow \mathcal{F}_{(k+1)}[[z]] \\
A(z) &\mapsto A(z)S(z)
\end{aligned}
\tag{3.10}
$$

That is, the polynomial matrix $S(z)$ maps the vector of power series $A(z)$ to the power series $A(z)\,S(z)$. We will use the norm

$$
\|S(z)\| = \max_{0\leq j\leq k}\left\{\sum_{i=0}^{k}\|S_{i,j}(z)\|\right\}. \tag{3.11}
$$

for $\mathcal{F}_{(k+1)\times(k+1)}[z]$. As in the proof of Lemma 3.3 it is easy to verify that (3.11) satisfies the requirements for a norm.

We now prove the norms (3.9) and (3.11) for $A(z)$ and $S(z)$ respectively, are compatible and tight. Thus, the norm (3.11) is a operator norm induced by the norm (3.9).

Lemma 3.4 *With $\|A(z)\|$ defined by (3.9) and $\|S(z)\|$ by (3.11),*

$$
\|A(z)\cdot S(z)\| \leq \|A(z)\|\cdot\|S(z)\|. \tag{3.12}
$$

Proof: Using (3.5), (3.8), (3.9), and (3.11), it follows that

$$
\|A(z)\cdot S(z)\| \;=\; \max_{0\leq j\leq k}\left\{\left\|\sum_{i=0}^{k} A_i(z)\cdot S_{i,j}(z)\right\|\right\}
$$

$$
\begin{aligned}
&\leq \max_{0\leq j\leq k}\left\{\sum_{i=0}^{k}\|A_i(z)\cdot S_{i,j}(z)\|\right\} \\
&\leq \max_{0\leq j\leq k}\left\{\sum_{i=0}^{k}\|A_i(z)\|\cdot\|S_{i,j}(z)\|\right\} \\
&\leq \max_{0\leq j\leq k}\left\{\sum_{i=0}^{k}\max_{0\leq l\leq k}\{\|A_l(z)\|\}\cdot\|S_{i,j}(z)\|\right\} \\
&= \|A(z)\|\max_{0\leq j\leq k}\left\{\sum_{i=0}^{k}\|S_{i,j}(z)\|\right\} \\
&= \|A(z)\|\cdot\|S(z)\|.
\end{aligned}
$$

□

Theorem 3.5 below shows that the bound (3.12) is tight.

Theorem 3.5 *For $\|A(z)\|$ defined by (3.9) and $\|S(z)\|$ by (3.11),*

$$\|S(z)\| = \sup_{A(z)\neq 0}\left\{\frac{\|A(z)\cdot S(z)\|}{\|A(z)\|}\right\}. \tag{3.13}$$

Proof: From Lemma 3.4 we have that $\|A(z)\cdot S(z)\| \leq \|A(z)\|\cdot\|S(z)\|$. Therefore, in order to prove (3.13), we must show

$$\exists\, A(z) \ni \|A(z)\cdot S(z)\| \geq \|A(z)\|\cdot\|S(z)\| \tag{3.14}$$

Let m be such that

$$\max_{0\leq j\leq k}\left\{\sum_{i=0}^{k}\|S_{i,j}(z)\|\right\} = \sum_{i=0}^{k}\|S_{i,m}(z)\|$$

and for $i = 0,\ldots,k$ define[2]

$$S_{i,m}(z) = \sum_{l=0}^{\partial} S_{i,m}^{(l)}\, z^l.$$

The vector $A(z)$ that we seek is given by

$$\begin{aligned} A_i(z) &= \sum_{l=0}^{\partial} \mathrm{sign}(S_{i,m}^{(l)}) z^{\partial - l} \\ &\overset{def}{\equiv} \sum_{l=0}^{n} A_i^{(\partial - l)} z^{\partial - l}, \quad i = 0, \ldots, k. \end{aligned}$$

Note that $\|A(z)\| = 1$. Let $C(z) = A(z) \cdot S(z)$. In particular, the $m - th$ component $C_m(z)$ of $C(z)$ is given by

$$\begin{aligned} C_m(z) &\overset{def}{\equiv} \sum_{l=0}^{\infty} c_m^{(l)} z^l \\ &= \sum_{i=0}^{k} A_i(z) S_{i,m}(z) \end{aligned}$$

Then the coefficient z^{∂} in $C_m(z)$ is given by

$$\begin{aligned} c_m^{(\partial)} &= \sum_{i=0}^{k} \left(\sum_{l=0}^{\partial} A_i^{(\partial - l)} S_{i,m}^{(l)} \right) \\ &= \sum_{i=0}^{k} \left(\sum_{l=0}^{\partial} \mathrm{sign} \left(S_{i,m}^{(l)} \right) \cdot S_{i,m}^{(l)} \right) \\ &= \sum_{i=0}^{k} \sum_{l=0}^{\partial} \left| S_{i,m}^{(l)} \right| \\ &= \sum_{i=0}^{k} \|S_{i,m}(z)\| \\ &= \|S(z)\|. \end{aligned}$$

Thus,

$$\begin{aligned} \|A(z) \cdot S(z)\| &= \|C(z)\| \\ &= \max_{0 \leq i \leq k} \{\|C_i(z)\|\} \\ &\geq \|C_m(z)\| \end{aligned}$$

$$
\begin{aligned}
&\geq \left|c_m^{(n)}\right| \\
&= \|S(z)\| \\
&= \|A(z)\| \cdot \|S(z)\|.
\end{aligned}
$$

The theorem follows.

□

Thus (3.11) is the operator norm induced by the norm (3.9).

3.2 Scaled Padé-Hermite Systems

Next we define the notion of a **scaled Padé-Hermite System** in which the sum of the norm of elements in each column of $S(z)$ is 1. We will work with scaled Padé-Hermite Systems in our numerical algorithm due to their desirable roundoff properties. Also, scaling the PHS after it is computed will reduce the probability of encountering overflow or underflow during the successive iteration.

Definition 3.3 *Let $S(z)$ be a Padé-Hermite System of type n. $S(z)$ is said to be a* **scaled PHS** *of type n if*

$$R_0(0) = \gamma_0, \quad V(0) = \begin{pmatrix} \gamma_1 & & 0 \\ & \ddots & \\ 0 & & \gamma_k \end{pmatrix} \tag{3.15}$$

is nonsingular and

$$\sum_{i=0}^{k} \|S_{i,j}(z)\| = 1, \quad 0 \leq j \leq k. \tag{3.16}$$

If $\hat{S}(z)$ is a normalized PHS with $V(0) = I_k$ and $R_0(0) = 1$, we can transform $\hat{S}(z)$ into a scaled PHS in the following way. Determine

$$\gamma_j = \frac{1}{\sum_{i=0}^{k} \|\hat{S}_{i,j}(z)\|}, \quad 0 \leq j \leq k. \tag{3.17}$$

Then

$$S(z) = \hat{S}(z) \cdot \Gamma \quad \text{where} \quad \Gamma = \begin{pmatrix} \gamma_0 & & 0 \\ & \ddots & \\ 0 & & \gamma_k \end{pmatrix} \tag{3.18}$$

is a scaled PHS.

Let

$$\gamma = \prod_{i=0}^{k} \gamma_i. \tag{3.19}$$

As a consequence of Definitions 2.1 and 3.3 we have that for a scaled Padé-Hermite System $S(z)$,

$$\gamma = R_0(0) \cdot \det(V(0)). \tag{3.20}$$

The parameter γ will be used as our estimate of the inverse of the condition number of the matrix T_n. If T_n is well conditioned, γ will be close to 1. For a poorly conditioned T_n, γ will be small.

In the next section we will justify our choice of stability parameter based on the inverse of a block Hankel matrix.

3.3 Hankel Systems and Their Inverses

system.

To view the problem in this way, we can, without loss of generality, assume that $A_0(z) \equiv 1$. This can be achieved by multiplying $A(z)$ by $A_0^{-1}(z)$. By doing this, the leftmost block of the Sylvester matrix will have 1 along the diagonal with all other elements of this block being zero. We can algebraically decouple the systems (2.13) and (2.21) eliminating the uppermost n_0 rows and leftmost n_0 columns. The resulting matrix H_n allows us to compute the components $Q(z)$ and $V(z)$ by solving two linear systems. The components $P(z)$ and $U(z)$ are obtained as functions of $Q(z)$ and $V(z)$ respectively.

Consider the $1 \times k$ vector of power series

$$D(z) = B^{-1}(z) \cdot C(z) \tag{3.21}$$

The block Hankel matrix associated with $D(z)$ is defined to be the $(\|n\| - n_0) \times (\|n\| - n_0)$ matrix

$$H_n = \left(\begin{array}{ccc|c|ccc} d_{n_0,1} & \cdots & d_{n_0-n_1+1,1} & & d_{n_0,k} & \cdots & d_{n_0-n_k+1,k} \\ d_{n_0+1,1} & \cdots & d_{n_0-n_1+2,1} & & d_{n_0+1,k} & \cdots & d_{n_0-n_k+2,k} \\ \vdots & & \vdots & \cdots & \vdots & & \vdots \\ d_{\|n\|-1,1} & \cdots & d_{\|n\|-n_1,1} & & d_{\|n\|-1,k} & \cdots & d_{\|n\|-n_k,k} \end{array} \right). \tag{3.22}$$

We obtain the components $P(z)$, $Q(z)$, $U(z)$, and $V(z)$ of a normalized Padé-Hermite System by solving linear systems with H_n as the coefficient matrix. The component $Q(z)$ is determined by solving the system

$$H_n \cdot X = (0, \ldots, 0, 1)^t \tag{3.23}$$

where X is the $(\|n\| - n_0) \times 1$ vector partitioned as

$$X = (q_{0,1}, \ldots, q_{n_1-1,1} | \cdots | q_{0,k}, \ldots, q_{n_k-1,k})^t . \tag{3.24}$$

The component $Q_j(z),\ \ j = 1,\ldots,k$ is given by (2.11). The remaining components of the first column of $S(z)$ are given by

$$P(z) = -D(z) \cdot Q(z) \mod z^{||n||-1}. \tag{3.25}$$

Similarly, the component $V(z)$ (with $V(0) = I_k$) is determined from the solution of

$$H_n \cdot Y = -\begin{pmatrix} d_{n_0+1,1} & d_{n_0+1,2} & \cdots & d_{n_0+1,k} \\ d_{n_0+2,1} & d_{n_0+2,2} & \cdots & d_{n_0+2,k} \\ \vdots & \vdots & & \vdots \\ d_{||n||,1} & d_{||n||,2} & \cdots & d_{||n||,k} \end{pmatrix} \tag{3.26}$$

where Y is a $(||n|| - n_0) \times k$ matrix partitioned as

$$Y = \left(\begin{array}{ccc|c|ccc} v_{1,1}^{(1)} & \cdots & v_{n_1,1}^{(1)} & \cdots & v_{1,k}^{(1)} & \cdots & v_{n_k,k}^{(1)} \\ \vdots & & \vdots & & \vdots & & \vdots \\ v_{1,1}^{(k)} & \cdots & v_{n_1,1}^{(k)} & \cdots & v_{1,k}^{(k)} & \cdots & v_{n_k,k}^{(k)} \end{array}\right)^t . \tag{3.27}$$

The $i, j-th$ component $i = 1,\ldots,k \quad j = 1,\ldots,k$ of $V(z)$ is given as in (2.23). The $1 \times k$ vector $U(z)$ of $S(z)$ is obtained from

$$U(z) = -D(z) \cdot V(z) \mod z^{||n||+1} \tag{3.28}$$

Theorem 3.6 (Labahn [20]) *Let $A(z)$ be a vector of $k+1$ formal power series (with $A_0(z) = 1$) and n be a vector of nonnegative integers. Let H_n be the block Hankel matrix (3.22) (where $D(z) = (A_1(z),\ldots,A_k(z))$) and $S(z)$ be a PHS of type n for $A(z)$. Recall from the definition of a PHS that*

$$Q(z) = \begin{pmatrix} Q_1(z) \\ \vdots \\ Q_k(z) \end{pmatrix}, \quad \text{where} \quad Q_j(z) = \sum_{i=0}^{n_j-1} q_{i,j} \cdot z^i. \tag{3.29}$$

Also recall that

$$V(z) = \begin{pmatrix} V_{1,1}(z) & \cdots & V_{1,k}(z) \\ \vdots & & \vdots \\ V_{k,1}(z) & \cdots & V_{k,k}(z) \end{pmatrix}, \tag{3.30}$$

where

$$V_{i,j}(z) = \sum_{\alpha=0}^{n_i} v_{i,j}^{(\alpha)} \cdot z^{\alpha}. \tag{3.31}$$

Let $m = \|n\| - n_0$ *and define*

$$T(z) = \det(V(z)) = \sum_{i=0}^{m} t_i \cdot z^i. \tag{3.32}$$

Furthermore let

$$G(z) = V^{adj}(z)\, Q(z) = \begin{pmatrix} G_1(z) \\ \vdots \\ G_k(z) \end{pmatrix}, \quad \text{where} \quad G_j(z) = \sum_{i=0}^{m-1} g_{i,j} \cdot z^i. \tag{3.33}$$

Given these definitions, we have that

$$H_n \cdot \left(\frac{1}{\gamma} \cdot \hat{H}_n \right) = I, \tag{3.34}$$

where $\hat{H}_n$ *is given by*

$$\hat{H}_n = \sum_{j=1}^{k} \left[\left(\begin{array}{ccc} v_{1,j}^{(n_1-1)} & \cdots & v_{1,j}^{(0)} \\ \vdots & \iddots & 0 \\ v_{1,j}^{(0)} & & \\ \hline & \vdots & \\ \hline v_{k,j}^{(n_k-1)} & \cdots & v_{k,j}^{(0)} \\ \vdots & \iddots & 0 \\ v_{k,j}^{(0)} & & \end{array} \right) \begin{pmatrix} g_{m-1,j} & \cdots & g_{0,j} \\ & \ddots & \vdots \\ 0 & & g_{m-1,j} \end{pmatrix} \right] \tag{3.35}$$

$$- \left(\begin{array}{cccc} q_{n_1-2,1} & \cdots & q_{0,1} & 0 \\ \vdots & \iddots & & \\ q_{0,1} & \iddots & & \\ 0 & & & 0 \\ \hline & \vdots & & \\ \hline q_{n_k-2,k} & \cdots & q_{0,k} & 0 \\ \vdots & \iddots & & \\ q_{0,k} & \iddots & & \\ 0 & & & 0 \end{array}\right) \left(\begin{array}{ccc} t_m & \cdots & t_1 \\ & \ddots & \vdots \\ 0 & & t_m \end{array}\right). \tag{3.36}$$

Proof:

The proof is given in [20].

□

As a direct consequence of Theorem 3.6, if we are working in exact arithmetic,

$$H_n^{-1} = \gamma \cdot \hat{H}_n \tag{3.37}$$

Suppose that we scale H_n so that $\|H_n\|_1 = 1$. We can accomplish this by scaling the power series $D(z)$ so that

$$\sum_{i=n_0}^{\|n\|-1} d_{i,j} = 1 \qquad j = 1, \ldots, k. \tag{3.38}$$

Then $\|\hat{H}_n\|_1$ will also be approximately 1. Thus γ indicates the inverse of the condition number of the matrix H_n. If γ is small, the condition number of H_n must be large and conversely if γ is large, then the condition number of H_n will be small.

Solving (3.23) and (3.26) when H_n is poorly conditioned will result in a loss of significance in the residual $R(z)$ and in the coefficients of $S(z)$. Therefore we use γ to help us determine whether or not a point in the Padé-Hermite table is stable.

Because of the close relationship between the Hankel and Sylvester systems, we can also use γ as a stability indicator for Sylvester systems.

Chapter 4

A Numerical Algorithm for Computing A Padé-Hermite System

Until now we have been assuming that all calculations were performed in exact arithmetic. By computing a PHS numerically, we invite error in the polynomial coefficients of $S(z)$ and the power series coefficients of the residual $R(z)$. A goal of a stable numerical algorithm is to minimize these errors.

To improve the numerical stability of our algorithm, we introduce a collection of component scalings. Although intuition plays a role in motivating our choice of scalings, their purpose is to reduce the likelihood of overflow and underflow occurring, and reduce the roundoff error arising from solving linear systems using Gaussian elimination. The components to be scaled include the input power series $A(z)$, the residuals $R^{(i)}(z)$, and the Padé-Hermite Systems $S^{(i)}(z)$. In each case, scaling can be accomplished by multiplying by a diagonal matrix. The scalings can be combined to form a new recurrence relation for computing the sequence of successive Padé-Hermite Systems at stable points in the Padé-Hermite table. After defining each transformation, we give a pseudo-code description of our numerical algorithm.

4.1 Scaling The Residual

Recall that in the iterative method for computing a Padé-Hermite System, the matrix $\hat{T}_{\nu}^{(i)}$ is constructed from coefficients of the previous PHS's residual $R^{(i)}(z)$. If the coefficients in $R^{(i)}(z)$ are very large, very small, or contain large variances between series, then a scaling may be appropriate. By scaling, we reduce the likelihood of overflow and underflow occurring when solving the systems (2.41) and (2.42) and reduce the accumulation of roundoff errors [14].

We scale each residual power series independently. Given $R^{(i)}(z)$ is the residual for the PHS $S^{(i)}(z)$ of type $m^{(i)}$, the matrix $\hat{T}_{\nu}^{(i)}$ is of order $\|\nu\|$, where $\nu = n^{(\sigma)} - m^{(i)} - e_0$. Also, define the modulo operation on the power series $R_j^{(i)}(z)$ to be another power series

$$R_j^{(i)}(z) \bmod z^{\|\nu\|+1} = \sum_{l=0}^{\|\nu\|} r_{l,j}^{(i)} \cdot z^l \; + \sum_{l=\|\nu\|+1}^{\infty} 0 \cdot z^l, \qquad j = 0, \ldots, k \tag{4.1}$$

In an effort to reduce the effect of roundoff errors in the solution of (2.41) and (2.42) obtained by Gaussian elimination, we scale the columns of $\hat{T}_{\nu}^{(i)}$ by determining, for $j = 0, \ldots, k$,

$$\lambda_j = \begin{cases} 1, & R_j^{(i)}(z) \bmod z^{\|\nu\|+1} = 0, \\ \|R_j^{(i)}(z) \bmod z^{\|\nu\|+1}\|, & \text{otherwise,} \end{cases} \tag{4.2}$$

where the norm used is given by (3.2) for power series. Then scaling the columns of

$\hat{T}_\nu^{(i)}$ is achieved by multiplying $R^{(i)}(z)$ on the right by Λ^{-1}, where

$$\Lambda = \begin{pmatrix} \lambda_0 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix}. \tag{4.3}$$

Next we show that we can recover a normalized PHS from one computed using a scaled residual vector.

Lemma 4.1 *If $\hat{S}(z)$ is a normalized PHS of type ν for $R^{(i)}(z)\,\Lambda^{-1}$, then $S'(z) = \Lambda^{-1}\,\hat{S}(z)\,\Omega$, where*

$$\Omega = \begin{pmatrix} 1 & & & 0 \\ & \lambda_1 & & \\ & & \ddots & \\ 0 & & & \lambda_k \end{pmatrix} \tag{4.4}$$

is a normalized PHS of type ν for $R^{(i)}(z)$.

Proof: Let $\hat{R}(z)$ be such that

$$R^{(i)}(z)\,\Lambda^{-1}\,\hat{S}(z) = z^{\|\nu\|+1}\,\hat{R}(z),$$

where $\hat{R}_0(0) = 1$ and $\hat{V}(0) = I_k$. Then

$$\begin{aligned} R^{(i)}(z)\,S'(z) &= R^{(i)}(z)\,\Lambda^{-1}\,\hat{S}(z)\,\Omega \\ &= z^{\|\nu\|+1}\,\hat{R}(z)\,\Omega \\ &\stackrel{def}{\equiv} z^{\|\nu\|+1}\,R'(z). \end{aligned}$$

Since $S'(z)$ also satisfies degree requirements, then $S'(z)$ is a PHS of type ν for $R^{(i)}(z)$.

Also note that $R_0'(0) = \hat{R}_0(0) = 1$ and

$$V'(0) = \begin{pmatrix} \lambda_1^{-1} & & 0 \\ & \ddots & \\ 0 & & \lambda_k^{-1} \end{pmatrix} \hat{V}(0) \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_k \end{pmatrix} = I_k$$

so that $S'(z)$ is also normalized.

□

Theorem 4.2 *Let $S^{(i)}(z)$ be a normalized PHS of type $m^{(i)}$ for $A(z)$ with residual $R^{(i)}(z)$ and let $\hat{S}(z)$ be the normalized PHS of type ν for $R^{(i)}(z)\,\Lambda^{-1}$. Then*

$$S'(z) = S^{(i)}(z)\,\Lambda^{-1}\,\hat{S}(z)\,\Omega \tag{4.5}$$

is a normalized PHS of type $m^{(i)} + \nu + e_0$ for $A(z)$.

Proof: The result follows immediately from Theorem 2.1 and Lemma 4.1.

□

From Lemma 4.1, computing a PHS of type n for $A(z)\,\Lambda^{-1}$ enables us to then immediately obtain a PHS of type n for $A(z)$. Although we can recover the original normalized PHS, we will not do so in the algorithm given in section 4.5. The next section describes why.

4.2 Scaling The Padé-Hermite System

In Section 3.2 we defined the concept of a scaled PHS. For the numerical algorithm we will compute scaled Padé-Hermite Systems instead of normalized ones. The reason for using scaled PHS's is twofold. First, by dividing each polynomial in a given column by the sum of the norms of polynomials we effectively reduce the variation between columns. That is, if a normalized PHS contains one column with polynomials having small coefficients and a different column contains polynomials with very large coefficients, then the corresponding scaled PHS will not have this variation. The second reason we use scaled PHS's is that the stability parameter $\gamma^{(i)}$ can be extracted easily from them.

In light of this, we can see that there is no need to compute normalized PHS's at each iteration. The work expended in obtaining a normalized PHS is negated when it is scaled. Specifically, the recurrence relation (4.5) need not include multiplication by the diagonal matrix Ω. Before stating what the actual recurrence relation is, we mention one additional scaling.

4.3 Scaling The Input Power Series

If we were to compute a PHS directly (rather than iteratively) by assembling coefficients of $A(z)$ to form T_n and solving the appropriate linear equations, we would once again be faced with the possibility of overflow, underflow and unnecessarily large roundoff error. To reduce the probability of these events occurring, we could scale the input power series $A(z)$ much as we did the residual $R(z)$ for the iterative method. It is of primary interest in this thesis to compare the errors that arise from computing a PHS directly and iteratively. If we do not scale the input power series before computing a PHS directly by the Gaussian elimination method, we are penalizing this method and will not achieve as good results as we could. As with scaling $R(z)$, we can scale $A(z)$ by multiplying by a diagonal matrix. However, if we do scale $A(z)$, we are no longer solving the original problem since we have modified the input. We can however recover from this scaling by multiplying the computed Padé-Hermite System by the inverse of the diagonal scaling matrix.

The scaling that we use is determined by only those terms of $A(z)$ used in obtaining the Padé-Hermite System. For the system of type $m^{(i)}$, the normalization is given by

$$||A_j^{(i)}(z) \bmod z^{||m^{(i)}||+1}|| = 1, \qquad j = 0, \ldots, k, \tag{4.6}$$

where Υ is a diagonal matrix such that

$$A^{(i)}(z) = A(z)\,\Upsilon^{-1}. \tag{4.7}$$

The norm in (4.6) corresponds to the power series norm (3.2) with $A_j^{(i)}(z) \bmod z^{||m^{(i)}||+1}$ defined as in (4.1). Thus, if $S^{(i)}(z)$ is a normalized PHS of type $m^{(i)}$ for $A^{(i)}(z)$ satisfying

$$A^{(i)}(z) \cdot S^{(i)}(z) = z^{||m^{(i)}||+1} R^{(i)}(z), \tag{4.8}$$

then $\Upsilon^{-1}\, S^{(i)}(z)$ is an (unnormalized) PHS of type $m^{(i)}$ for $A(z)$ satisfying

$$A(z)\,\Upsilon^{-1}\, S^{(i)}(z) = z^{||m^{(i)}||+1} R^{(i)}(z). \tag{4.9}$$

4.4 Combining the Scalings

In this section we will describe how to obtain the next PHS in the sequence of stable PHS and test for stability using the parameter γ. To do this we will assume a number of things.

Let $A(z)$ be a $1 \times (k+1)$ vector of power series. Suppose $S^{(i)}(z)$ is the scaled PHS of type $m^{(i)}$ for the scaled input power series $A^{(i)}(z) = A(z)\Upsilon^{(i)^{-1}}$ and let $R^{(i)}(z)$ be the residual. We want to compute a PHS $\hat{S}(z)$ of type ν for $R^{(i)}(z)$ such that

$$S^{(i+1)}(z) = S^{(i)}(z) \cdot \hat{S}(z)$$

is the scaled PHS of type $m^{(i+1)} = m^{(i)} + \nu + e_0$ for the scaled input power series $A^{(i+1)}(z) = A(z)\Upsilon^{(i+1)^{-1}}$, which we accomplish by the following steps.

Step 1: Determine the scaling of $R^{(i)}(z)$ by the matrix Λ^{-1}, where Λ is given by (4.2) and (4.3). Note that $R^{(i)}(z)\,\Lambda^{-1} \bmod z^{\|\nu\|+1}$ is a vector of power series whose norm is a single scalar.

Step 2: Solve the systems (2.41) and (2.42) to obtain the normalized PHS $\hat{S}(z)$ such that

$$R^{(i)}(z)\,\Lambda^{-1}\,\hat{S}(z) = z^{\|\nu\|+1}\hat{R}(z) \tag{4.10}$$

with $\hat{V}(0) = I_k$ and $\hat{R}_0(0) = 1$. Then

$$\begin{aligned} A^{(i)}(z)\,S^{(i)}(z)\,\Lambda^{-1}\,\hat{S}(z) &= z^{\|m^{(i)}\|+1}R^{(i)}(z)\,\Lambda^{-1}\,\hat{S}(z) \qquad (4.11)\\ &= z^{\|m^{(i+1)}\|+1}\hat{R}(z) \qquad (4.12)\end{aligned}$$

Step 3: Scale $A(z)$ by $\Upsilon^{(i+1)}$ to obtain

$$A^{(i+1)}(z) = A(z)\,\left(\Upsilon^{(i+1)}\right)^{-1}, \tag{4.13}$$

where $\Upsilon^{(i+1)}$ is determined so that

$$\|A_j^{(i+1)}(z) \bmod z^{\|m^{(i+1)}+1\|}\| = 1, \qquad j = 0, \ldots, k. \tag{4.14}$$

Step 4: Let

$$S^{(i+1)}(z) = \Upsilon^{(i+1)}\,\Upsilon^{(i)^{-1}}\,S^{(i)}(z)\,\Lambda^{-1}\,\hat{S}(z). \tag{4.15}$$

Then

$$\begin{aligned} A^{(i+1)}(z)S^{(i+1)}(z) &= A^{(i+1)}(z)\,\Upsilon^{(i+1)}\,\left(\Upsilon^{(i)}\right)^{-1}\,S^{(i)}(z)\,\Lambda^{-1}\,\hat{S}(z) \qquad (4.16)\\ &= A^{(i)}(z)\,S^{(i)}(z)\,\Lambda^{-1}\,\hat{S}(z) \qquad (4.17)\\ &= z^{\|m^{(i+1)}\|+1}\,\hat{R}(z). \qquad (4.18)\end{aligned}$$

Also note that the degree bounds (2.5) are satisfied for the PHS $S^{(i+1)}(z)$ of type $m^{(i+1)}$ since all the diagonal scaling matrices consist of constant entries.

Step 5: Convert $S^{(i+1)}(z)$ to a scaled PHS and compute $\gamma^{(i+1)}$.

Step 6: Accept $S^{(i+1)}(z)$ as a stable point if $\gamma^{(i+1)}$ is less than some specified tolerance ϵ.

4.5 The Algorithm

For a given vector of nonnegative integers n, the algorithm NPADE_HERMITE makes use of Theorem 2.1 to compute the sequence $\{S^{(i)}(z)\}$ of PHS for a given vector of power series $A(z)$. The points $m^{(i)}$ now correspond to stable points rather than nonsingular points and we step over unstable blocks. A quantitative measure of the stability of a point $m^{(i)}$ is provided by the stability parameter $\gamma^{(i)}$. The user supplies the tolerance value ϵ.

Let

$$A(z) = (A_0(z), A_1(z), \ldots, A_k(z)),$$

where

$$A_i(z) = \sum_{j=0}^{\infty} a_{j,i}\, z^j$$

```
NPADE_HERMITE($A(z)$, $n$, $k$, $\epsilon$)
$i \leftarrow 0$,    $m^{(0)} \leftarrow -e_0$,    $S^{(0)} \leftarrow I_{k+1}$,    $\Upsilon^{(0)} \leftarrow I_{k+1}$
$M \leftarrow \min(n_0, n_1) + 1$,
$\sigma \leftarrow 0$,    $stable \leftarrow$ true
While (($\sigma < M$) and $stable$) do
    $\nu \leftarrow n - m^{(i)} - e_0$
    $s \leftarrow 0$, $\textit{stable} \leftarrow$ false
    While $(s < M - \sigma)$ and (not $\textit{stable}$) do
        $s \leftarrow s + 1$
        $\nu_j^{(s)} \leftarrow \max(0, \nu_j + \sigma - M + s), \qquad j = 0, \ldots, k$
        /* compute the residual */
        $R^{(i)}(z) \leftarrow A(z) \cdot \left(\Upsilon^{(i)}\right)^{-1} \cdot S^{(i)}(z) \,/\, z^{\|m^{(i)}\|+1} \mod z^{\|\nu^{(s)}\|+1}$
        /* scale the residual */
        $\hat{R}^{(i)}(z) \leftarrow R^{(i)}(z)\, \Lambda^{-1}$, where $\Lambda$ is given in (4.3)
        Construct $\hat{T}^{(i)}_{\nu^{(s)}}$
        If $\hat{T}^{(i)}_{\nu^{(s)}}$ is numerically nonsingular then
            $m^{(i+1)} \leftarrow m^{(i)} + \nu^{(s)} + e_0$
            Obtain $\hat{S}(z)$ using (2.41) and (2.42)
            Obtain $\Upsilon^{(i+1)}$ satisfying (4.6) and (4.7)
            $S^{(i+1)}(z) \leftarrow \Upsilon^{(i+1)} (\Upsilon^{(i)})^{-1} S^{(i)}(z)\, \Lambda^{-1} \hat{S}(z)$
            Obtain $\Gamma^{(i+1)}$ satisfying (3.17) and (3.18)
            /* scale the PHS */
            $S^{(i+1)}(z) \leftarrow S^{(i+1)}(z) \left(\Gamma^{(i+1)}\right)^{-1}$
            $\textit{stable} \leftarrow \gamma^{(i+1)} > \epsilon$, where $\gamma^{(i+1)}$ is given by (3.15) and (3.19)
        end if
    end While
    If $\textit{stable}$ then $\sigma \leftarrow \sigma + s, \quad i \leftarrow i + 1$
end While
```

Chapter 5

Previous Numerical Results

5.1 Padé-Hermite Systems In a Numerical Setting

Previously we have considered the elements $S(z)$ and $R(z)$ to be exact. Rather than introducing new notation, we will now use these components to represent values computed by our algorithm using finite precision arithmetic. When the exact values are of interest, we will indicate them with the subscript E as in $S_E(z)$ and $R_E(z)$. Hence condition II of Definition 2.1, for the exact and the computed scaled PHS of type n for $A(z)$, become respectively

$$A(z) \cdot S_E(z) = z^{\|n\|+1} R_E(z) \tag{5.1}$$

and

$$A(z) \cdot S(z) = \delta R(z) + z^{\|n\|+1} R(z), \tag{5.2}$$

where $\delta S(z)$ is the consequence of the numerical error in $S(z)$. Note that $\delta R(z) = 0$ if and only if $S(z) = S_E(z)$. If we define

$$\delta S(z) \;=\; S(z) - S_E(z); \tag{5.3}$$

then from (5.1) and (5.2)

$$\delta R(z) \;=\; A(z) \cdot \delta S(z) \bmod z^{||n||+1}, \tag{5.4}$$

and

$$R(z) \;=\; R_E(z) \;+\; \left(A(z) \cdot \delta S(z) - \delta R(z)\right) / z^{||n||+1}. \tag{5.5}$$

In (5.2), for the purpose of norm compatibility used later, we consider $\delta R(z)$ to be a vector of power series whose j^{th} component we represent by

$$\delta R_j(z) = \sum_{l=0}^{||n||} r_{l,j} \cdot z^l \;+\; \sum_{l=||n||+1}^{\infty} 0 \cdot z^l \qquad j = 0, \ldots, k. \tag{5.6}$$

5.2 Estimating the Error

As with the case of Padé-Hermite approximants, numerous algebraic algorithms have been proposed for computing Padé Approximants. However, few of these algorithms has been analyzed from a numerical standpoint. Still fewer have been proven stable. For a restricted class of power series (those which correspond to positive definite Hankel matrices) Ammar and Gragg [1] give an algorithm which is stable. Recently, Cabay and Meleshko [11] generalized the algebraic algorithm of Cabay and Choi [8] to establish a fast weakly stable numerical algorithm for power series without restrictions. While they dealt exclusively with Padé approximants, we can apply many of their results to Padé-Hermite approximants. In fact the Padé-Hermite approximant for $k = 1$ corresponds directly to the classical notion of a Padé Approximant. Specifically, if the point (n_0, n_1) is a nonsingular point in the Padé table, then the 2×2 matrix composed of the (n_0, n_1) and $(n_0 - 1, n_1 - 1)$ Padé Approximants would be identical to that of a (n_0, n_1) Padé-Hermite System.

In this section, we highlight some of the important results reported by Cabay and Meleshko. Error bounds for the residual, $\delta R(z)$, and the relative error, $\delta S(z)$, establish that the key parameters governing the performance of their algorithm is the stability parameter $\gamma^{(i)}$. In the following chapters, we will confirm experimentally that these results hold for our algorithm with $k = 1$, and draw conclusions along similar lines for $k > 1$.

Cabay and Meleshko only considered Padé approximants where $n = (x - 1, x)$ for some nonnegative integer x. This corresponds to computing Padé approximants along the superdiagonal of the Padé table. This restriction is made without loss of generality since a Padé approximant at any point in the Padé table can be computed given that those on the superdiagonal are known.

Let $n = (x - 1, x, \ldots, x)$, then

$$m^{(i)} = \left(m_0^{(i)}, m_1^{(i)}, \ldots, m_k^{(i)}\right) \tag{5.7}$$

where $m_j^{(i)} = m_0^{(i)} + 1, \quad j = 1, \ldots, k$. Define the step size s_i (from $m^{(i)}$ to $m^{(i+1)}$) to be

$$s_i = m_0^{(i+1)} - m_0^{(i)}. \tag{5.8}$$

Note that $m^{(i+1)} = m^{(i)} + \nu + e_0$, where

$$\nu = (s_i - 1, s_i, \ldots, s_i)\,. \tag{5.9}$$

Cabay and Meleshko provide an error bound for the residual error $\delta R^{(i)}(z)$.

Theorem 5.1 (Cabay and Meleshko [11]) *Let $k = 1$. Assuming that $\gamma^{(j)}$ is large and $\|\delta R^{(j)}(z)\|$ is small so that*

$$22\,(m_0^{(i)}+1)^2\,(\|\delta R^{(i)}(z)\| + 5\,(m_0^{(i)}+1)\,\mu) \le \gamma^{(i)} \tag{5.10}$$

for each $i \le j$, then

$$\|\delta R^{(j+1)}(z)\| \le \left(\frac{\Delta_0^{(j)}}{\gamma^{(j)}} + \sum_{l=0}^{j-1} \frac{\Delta_1^{(l)}}{\gamma^{(l)}\,\gamma^{(l+1)}} \right) \mu + \mathcal{O}\left(\mu^2\right), \tag{5.11}$$

where

$$\begin{aligned}
\Delta_0^{(i)} &= 4.04\,(s_i+1)\,(m_0^{(i)}+1)^3 + 512\,s_i^4\,\rho_l\,(m_0^{(i)}+1)^2 && (5.12)\\
&+ 8.08\,(m_0^{(i+1)}+1)\,(m_0^{(i)}+1)\,(s_i+1)^2 &&\\
\Delta_1^{(i)} &= 32.32\,(s_i+1)\,(m_0^{(i)}+1)^3\,(m_0^{(i+1)}+1) && (5.13)\\
&+ 4096\,s_i^5\,\rho_l\,(m_0^{(i)}+1)^2\,(m_0^{(i+1)}+1) && (5.14)\\
&+ 32.32\,(m_0^{(i+1)}+1)^3\,(s_i+1)^2\,(m_0^{(i)}+1) &&
\end{aligned}$$

with $\gamma^{(i)} = \left|R_0^{(i)}(0)\;V_0^{(i)}(0)\right|$, ρ_i is the growth factor associated with the Gaussian elimination performed at the i^{th} step, and μ is the machine unit roundoff.

□

Notice that Theorem 5.1 assures us that if $\|\delta R^{(j)}(z)\|$ is small and $\gamma^{(i)}$ is large, then $\|\delta R^{(j+1)}(z)\|$ will also be small. If the step size s_j is chosen so that $\gamma^{(j+1)}$ is large as well, then Theorem 5.1 can be applied at the next computed point. Hence $\|\delta R^{(j+1)}(z)\|$ will remain small for all j, as long as s_j is chosen appropriately. Equivalently stated, the error in the residual will be small as long as we step from stable point to stable point.

Meleshko and Cabay observed experimentally that for power series with randomly generated coefficients, s_j could be chosen so that $\gamma^{(j+1)} > 10^{-2}$. Also $\|\delta R^{(j+1)}(z)\|$

depends on $\gamma^{(i)}$ in the denominator of the summation and not $\gamma^{(i)}\,\gamma^{(i+1)}$. The overall error in $\|\delta R^{(j+1)}(z)\|$ was observed to be inversely proportional to the smallest $\gamma^{(i)}$ encountered.

Theorem 5.2 (Cabay and Meleshko [11]) *Let $k = 1$. If (5.10) holds for $0 \leq i \leq j$, then the approximants computed by NPADE_HERMITE satisfy*

$$\|\delta S^{(j)}(z)\| \;\leq\; \frac{2.2\,(m_0^{(j)}+1)^3}{\gamma^{(j)}}\left(\frac{\Delta_0^{(j-1)}}{\gamma^{(j-1)}} + \sum_{i=0}^{j-2}\frac{\Delta_1^{(i)}}{\gamma^{(i)}\,\gamma^{(i+1)}}\right)\mu \;+\; \mathcal{O}\left(\mu^2\right), \quad (5.15)$$

where $\Delta_0^{(i)}$ and $\Delta_1^{(i)}$ are defined in 5.12.

□

Experimentally, Cabay and Meleshko [11] observed that the large constants and powers of $m_0^{(i)}$ and s_i that occur in Theorem 5.2 were not manifested in the experiments.

The question that arises from these is "How do these error bounds relate to those for Padé-Hermite approximants?" For $k = 1$ we would expect similar results due to the fact that our algorithm and Cabay and Meleshko's behave comparably. However for larger k it is unclear as to whether these bounds are significant.

The actual error bounds will contain different $\Delta_0^{(i)}$, $\Delta_1^{(i)}$ composed of low degree polynomials in $m_0^{(i)}$ and s_i. In order for the error bounds for $k > 1$ to be similar, it must be the case that $\gamma^{(i)}$ be a good estimate of the condition number $H_{m^{(i)}}$. If this is the case, a forward error analysis similar to Cabay and Meleshko [11] could be done to obtain the actual error bounds for the Padé-Hermite approximant algorithm.

Chapter 6

Experimental Method

Before presenting the experimental results, we will discuss some of the details involved in gathering the data. Several programs were written in order to collect the results. We will describe their implementation and role in the experimental process. Several problem classes were used in the execution of the experiments. These, along with a method for generating power series with singular points will be highlighted. The next chapter contains a collection of tables detailing the numerical values obtained. The format of these tables will be explained and an overview of the different types of experiments will be provided.

6.1 Software Implementations

One of the main goals of this thesis is to study the numerical calculation of Padé-Hermite approximants. Naturally a numerical implementation of the algorithm in section 4.5 was required. This algorithm was implemented using Sun Fortran 1.3.1. Fortran was chosen due to its widespread acceptance as the numerical language of choice and the availability of numerical libraries such as LINPACK. All calculations were performed in double precision. The linear systems (2.41) and (2.42) were solved using the LINPACK routines SGEFA and SGESL[1]. Numerical singularity of the matrix $\hat{T}_\nu^{(i)}$ was tested using the determinant returned by LINPACK's SGEDI. Numerical PHS's were written to a file in a format (nearly) readable by Maple. A C program was used to convert the floating point format of the output text file to one compatible with Maple.

Previously, it was mentioned that the results that are of most interest are the relative error in the Padé-Hermite System and the error in the residual. In order to compute the relative error in the PHS, we require a means to compute the exact solution $S_E(z)$ of a PHS for a given n. We also wish to examine the relative error in the intermediate systems $S^{(i)}(z)$. To accomplish this, an implementation of the iterative numerical algorithm was done using the programming language offered by Maple V. The Maple symbolic algebra environment enabled all calculations to be done in exact arithmetic. All the scalings involved in the numerical algorithm were incorporated into this implementation. Each PHS was saved in a separate file for later comparison with its numerical counterpart.

Although this program accomplished the desired goal, it proved very costly in terms of execution time. As a compromise, a Maple program was written which directly computed all Padé-Hermite Systems along the desired diagonal in the Padé-Hermite table using 50 digits of accuracy per polynomial coefficient. This solution utilized singular value decomposition [14]. Because the numerical implementation of the algorithm offered only double precision, and due to the fact that only scaled

[1] Although these routines are written for single precision computation, a compiler flag was used to force all single precision values to be converted to double precision

PHS's were computed, it was felt that these results could be considered to be exact for our purposes. Note that this Maple implementation contained a test to identify $T_{n(\sigma)}$ matrices with condition number > 10^{30}. Such systems were considered numerically singular. Solving a linear system with a condition number $\approx 10^{p}$ results in a loss of accuracy of p digits in the solution [7]. Therefore, by doing all computations with 50 digits of accuracy and rejecting those with condition number > 10^{30}, a minimum of 20 digits of accuracy could always be expected in the resulting PHS.

With the approximate and exact results computed by the two implementations, a Maple program was written which read in the results and computed the relative error

$$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|} = \frac{\|S_E^{(i)}(z) - S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|} \tag{6.1}$$

in exact arithmetic for the various stable Padé-Hermite Systems.

The error in the residual was determined by the Fortran implementation. For a given PHS $S^{(i)}(z)$, we compute

$$\|\delta R^{(i)}(z)\| = \| \left(A(z) \left(\Upsilon^{(i)}\right)^{-1} S^{(i)}(z)\right) \bmod z^{\|m^{(i)}\|+1} \|. \tag{6.2}$$

In order to judge how good the numerical solution is, a program was developed which would compute the Padé-Hermite System directly by solving equations (2.13) and (2.21) directly using Gaussian elimination with partial pivoting. This was also implemented in Fortran using the LINPACK routines SGEFA and SGESL. If we let the PHS's obtained in this way be denoted by $S_G^{(i)}(z)$, then the relative error for these systems is given by

$$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|} = \frac{\|S_E^{(i)}(z) - S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}. \tag{6.3}$$

In developing our algorithm, we have relied on $\gamma^{(i)}$ providing a gcod estimate of the condition number of the matrix $T_{m^{(i)}}$. Therefore a Fortran program was written which computed the condition number of $T_{m(i)}$ using $\|\cdot\|_\infty$. Using the LINPACK routine SGEDI, the inverse of $T_{m(i)}$ was established and the condition number computed as

$$\kappa(T_{m(i)}) = \|T_{m(i)}\|_\infty \cdot \|T^{-1}_{m(i)}\|_\infty \tag{6.4}$$

Because these values were computed using double precision floating point arithmetic, they are only estimates of the actual condition number. Estimates greater than 10^{17} may be subject to considerable error.

All floating point computations were carried out on a Sun Sparcstation 1. The unit round μ for this computer was 2.22×10^{-16}. All Maple results were obtained on a Silicon Graphics 4D/340 and Sun 4/40FC-24.

6.2 Problem Classes

To thoroughly test the numerical algorithm, it must be tested on a variety of different problems. We can group problems with power series sharing certain common features into a *problem class*. Within each problem class, the parameters k and n can be varied. Note that for all problem classes, it must be the case that $A_0(0) \neq 0$ as required by Definition 2.1.

For the numerical experiments performed, we can consider four classes. The first class of problem (class I) contains power series with random integer coefficients. Problems of this type were fabricated by randomly generating numbers between -128 and 128 and using them as power series coefficients. This class includes problems where we set $A_0(z) = 1$ and generate random coefficients for the remaining k power series. Assuming that $A_0(z) = 1$ is a common assumption used in many applications (and is not a further restriction since this can be achieved by multiplying $A(z)$ by $A_0^{-1}(z)$).

We will see in Section 7.2 that $A_0(z)$ plays a key role in the growth of the relative error in Padé-Hermite Systems.

The class II problem consists of power series whose corresponding Padé-Hermite table contain singular points at predetermined positions. The problem of numerically computing PHS's at these singular points will be extremely unstable. These problems are explicitly constructed using relations involving the adjoint of the PHS matrix. We introduce the notion of a **template** PHS which is used to generate a power series with certain properties. To prove this construction, we require the following lemmas.

Lemma 6.1 *Let $S(z)$ be a normalized Padé-Hermite System of type n for a vector of power series $A(z)$ with $A_0(0) = 1$. Then*

$$\det(S(z)) = z^{\|n\|+1}. \tag{6.5}$$

Proof: Let $S^{adj}(z) = \text{adjoint}\,(S(z))$. We have that

$$A(z) \cdot S(z) = z^{\|n\|+1} R(z), \qquad R_0(0) = 1, \; V(0) \approx I_k. \tag{6.6}$$

So

$$\begin{aligned} A(z) \cdot \det(S(z)) &= A(z) \cdot S(z) \cdot S^{adj}(z) \\ &= z^{\|n\|+1} R(z) \cdot S^{adj}(z). \end{aligned} \tag{6.7}$$

The first component of (6.7) satisfies

$$A_0(z) \cdot \det(S(z)) = z^{\|n\|+1} \sum_{i=0}^{k} R_i(z) \cdot S^{adj}_{i,0}(z) \tag{6.8}$$

But from the degree bounds (2.5), $\partial\,[\det(S(z))] \leq \|n\| + 1$, so that (6.8) becomes

$$\det(S(z)) = z^{\|n\|+1} A_0^{-1}(z) \cdot \sum_{i=0}^{k} R_i(0) \cdot S^{adj}_{i,0}(0) \tag{6.9}$$

The result (6.5) now follows from (6.6) and (6.9) after observing that $S^{adj}_{0,0}(0) = \det(V(0)) = 1$ and that for $i = 1,\ldots,k$, $S^{adj}_{i,0}(0) = 0$ (since $S^{ac}_{i,0}$), $i = 1,\ldots,k$ each has z^2 as a factor).

□

Lemma 6.2 *Let $A(z)$ is a vector of $k+1$ power series and $\hat{S}(z)$ be a* **template** *PHS for $A(z)$ of type $\nu = (s-1, s, \ldots, s)$ where ν is the first nonsingular point along the superdiagonal of the Padé-Hermite table. Then*

$$A(z) \cdot \hat{S}(z) = z^{\|\nu\|+1} R(z), \quad \hat{V}(0) = I, \quad R_0(0) = 1. \tag{6.10}$$

Now let $B(z)$ be a vector of $k+1$ power series such that

$$B(z) = W(z) \left(\hat{S}(z)\right)^{adj} \tag{6.11}$$

for any $W(z) \in \mathcal{F}_{k+1}[[z]]$ such that $W_0(0) \neq 0$. Then $\hat{S}(z)$ is a PHS of type ν for $B(z)$. Furthermore, ν is the first nonsingular point along the superdiagonal for $B(z)$.

Proof:

$$\begin{aligned} B(z) \cdot \hat{S}(z) &= W(z) \left(\hat{S}(z)\right)^{adj} \hat{S}(z) & (6.12) \\ &= z^{\|\nu\|+1} W(z). & (6.13) \end{aligned}$$

Hence the PHS order condition is satisfied. Also note that the remaining two conditions of Definition 2.1 are satisfied, so $\hat{S}(z)$ is a PHS of type ν for $B(z)$.

Next, suppose that $\tilde{S}(z)$ is a PHS of type $\tilde{\nu}$ such that $\tilde{\nu}$ lies on the same diagonal path as ν, $\|\tilde{\nu}\| < \|\nu\|$, and

$$B(z) \cdot \tilde{S}(z) = z^{\|\tilde{\nu}\|+1} \tilde{W}(z). \tag{6.14}$$

Then by Theorem 2.1, there must exist a PHS $\bar{S}(z)$ of type $\bar{\nu} = \nu - \tilde{\nu} - e_0$ where

$$\hat{S}(z) = \tilde{S}(z) \cdot \bar{S}(z). \tag{6.15}$$

Then

$$A(z)\tilde{S}(z)\bar{S}(z) \;=\; z^{\|\nu\|+1}R(z) \tag{6.16}$$

$$A(z)\tilde{S}(z) \;=\; z^{\|\nu\|+1}R(z)\cdot\left(\bar{S}(z)\right)^{adj}. \tag{6.17}$$

Also note that the remaining PHS requirements are met for $\tilde{S}(z)$. But (6.16) implies that $\tilde{S}(z)$ is a PHS of type $\tilde{\nu}$ for $A(z)$. This contradicts the fact that ν is the first nonsingular point along the superdiagonal. This contradiction establishes the lemma.

□

Theorem 6.3 *Let $\hat{S}^{(i)}(z)$ be a normalized template PHS of type $\nu^{(i)} = (s_i-1, s_i, \ldots, s_i)$ for $A^{(i)}(z)$, $i = 1, \ldots, j$. Consider*

$$A(z) = W(z)\cdot\left(\hat{S}^{(j)}(z)\right)^{adj}\cdot\cdots\cdot\left(\hat{S}^{(1)}(z)\right)^{adj}, \tag{6.18}$$

where $W_0(0) \neq 0$. Then the nonsingular subsequence for $A(z)$ along the superdiagonal is given by

$$m^{(i+1)} = m^{(i)} + \nu^{(i)} + e_0, \qquad i = 0, 1, \ldots, j \tag{6.19}$$

and the normalized PHS at $m^{(i)}$ is given by

$$S^{(i)}(z) \;=\; \hat{S}^{(1)}(z)\cdot\cdots\cdot\hat{S}^{(i)}(z) \tag{6.20}$$

Proof: Let

$$A^{(i)}(z) = W(z)\left(\hat{S}^{(j)}(z)\right)^{adj}\cdots\left(\hat{S}^{(i+1)}(z)\right)^{adj}, \;\; i = 0, \ldots, j-1. \tag{6.21}$$

Then $A^{(0)}(z) = A(z)$ and

$$A^{(i)}(z) = A^{(i+1)}(z)\left(\hat{S}^{(i+1)}(z)\right)^{adj}, \;\; i = 0, \ldots, j-1. \tag{6.22}$$

If we assume inductively that $A_0^{(j)}(0) \neq 0, \ldots, A_0^{(i+1)}(0) \neq 0$, it then follows from (6.21) using arguments similar to those in the proof of Lemma 6.1 that

$$A_0^{(i)}(0) \neq 0, \quad i = j-1, \ldots, 0. \tag{6.23}$$

Let

$$S^{(i)}(z) = \hat{S}^{(1)}(z) \cdot \cdots \cdot \hat{S}^{(i)}(z), \quad i = 1, \ldots, j. \tag{6.24}$$

It is easy to show that $S^{(i)}(z)$ is a PHS of type $m^{(i)}$ for $A(z)$ satisfying

$$A(z) \cdot S^{(i)}(z) = z^{\|m^{(i)}\|+1} A^{(i)}(z). \tag{6.25}$$

From (6.21), (6.23) and Lemma 6.2, the first nonsingular point along the superdiagonal for $A^{(i)}(z)$ is $\hat{S}^{(i+1)}(z)$. Thus, from Theorem 2.1 the next nonsingular point for $A(z)$ is $m^{(i+1)}$ and the PHS at $m^{(i+1)}$ is $S^{(i+1)}(z) = S^{(i)}(z) \cdot \hat{S}^{(i+1)}(z)$. The result now follows by induction on i.

□

This relationship allows us to construct a vector of power series $A(z)$ with singular and nonsingular points at chosen locations along a specific path in the Padé-Hermite table. Because this method requires exact arithmetic, it was performed in Maple. We illustrate the method with an example.

Example 6.1 Suppose that we want to generate a vector of power series $A(z) =$

$(A_0(z), A_1(z))$ (i.e. $k = 1$) such that along the superdiagonal the points are as follows

					n_1			
		0	1	2	3	4	5	6
	0		nonsing					
	1			sing				
n_0	2				sing			
	3					nonsing		
	4						sing	
	5							nonsing

(6.26)

where nonsing represents a nonsingular point and sing denotes a singular one. To do this we require three **template** PHS's, $\hat{S}^{(1)}(z)$ of type $\nu_1 = (0,1)$, $\hat{S}^{(2)}(z)$ of type $\nu_2 = (1,2)$, and $\hat{S}^{(3)}(z)$ of type $\nu_3 = (2,3)$.

To obtain such systems we can randomly generate power series and modify them so that the associated Sylvester matrix T_{ν_i} is nonsingular and all submatrices of lower order are singular. For example, let

$$\begin{aligned} A_0(z) &= 1 + 3z - 2z^2 + \ldots & (6.27) \\ A_1(z) &= z - z^3 + \ldots . & (6.28) \end{aligned}$$

Now

$$T_{(0,1)} = (0) \quad \text{and} \quad T_{(1,2)} = \begin{pmatrix} 1 & 0 & 0 \\ 3 & 1 & 0 \\ -2 & -1 & 1 \end{pmatrix} \tag{6.29}$$

so the solutions of (2.13) and (2.21) for $T_{(1,2)}$ would generate the PHS $\hat{S}^{(1)}(z)$ since $\det(T_{(1,2)}) \neq 0$.

Once the template Padé-Hermite Systems have been computed, compute their adjoints $\left(\hat{S}^{(1)}(z)\right)^{adj}, \left(\hat{S}^{(2)}(z)\right)^{adj}, \left(\hat{S}^{(3)}(z)\right)^{adj}$. Obtain a vector of power series $R(z) = (R_0(z), R_1(z))$ with $R_0(0) \neq 0$ and compute

$$A(z) = R(z) \cdot \left(\hat{S}^{(2)}(z)\right)^{adj} \cdot \left(\hat{S}^{(3)}(z)\right)^{adj} \cdot \left(\hat{S}^{(1)}(z)\right)^{adj}. \tag{6.30}$$

The power series $A(z)$ will have the Padé-Hermite table up to the point $(5, 6)$ given in (6.26).

□

Note that this method cannot generate problems with $A_0(z) = 1$. If we desire, we can multiply $A(z)$ by $A_0^{-1}(z)$ to obtain this type of problem. The disadvantage of constructing power series in this manner is that it produces very divergent power series with rapidly growing coefficients. For this reason, only relatively small problems (small k and $||n||$) can be computed.

The final two classes of power series used in the experiments are perturbed versions of the class II problem just illustrated. In Maple, the coefficients of class II power series are perturbed by either 10^{-12} or 10^{-6} yielding class III and IV problems respectively. The larger the perturbation, the more stable the system will become at the points which were originally constructed to be singular. These four classes will give us opportunity to examine problems of varying instability.

6.3 Explanation of Table Headings

The headings of the experimental results look like

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$

Although the table headings indicate that we are only reporting values for stable points $m^{(i)}$, the results for unstable rows are also listed. This allows us to illustrate why such points were rejected by the algorithm. The first column labeled by i indicates if the point was accepted as stable. If an entry is blank then the point was not accepted as stable, otherwise an integer indicating the points position in the sequence of stable points appears. The next column provides the condition number of the Sylvester matrix $T_{m^{(i)}}$. Following that, the value $\|m^{(i)}\|$ as given by (2.3) appears. Next, the value of the stability parameter $\gamma^{(i)}$ is given. The norm of the error in the residual occupies the fifth column. Finally the relative error for the iterative and Gaussian elimination methods is presented. Points decided unstable by the iterative algorithm are labeled as such. All numerical entries are given in scientific notation with two digits of accuracy and the exponent appearing in brackets. For example, 10000 would be entered as 1.0(4).

6.4 Overview of the Experiments

As the main purpose of this thesis is to evaluate the proposed numerical algorithm for computing Padé-Hermite Systems, numerous experiments were performed. Due to the amount of CPU time required to compute the exact solution of some systems, we were forced to restrict the parameter k to be relatively small. A short description of the parameters used for each experiment appears with the tables in the next chapter. Experiments with values of k equaling 1, 2, and 7 were performed. For $k = 1,2$, power series with singular points were generated. These were perturbed and rerun to study the effect. An experiment was run with $k = 2$ and very large $\|n\|$ to illustrate the growth of error when descending deep into the Padé-Hermite table. Several other specific experiments were run in order to validate some of the conjectures given in the next chapter.

Chapter 7

Numerical Results

In this chapter we present some of the experimental results obtained using the Fortran implementation of our numerical Padé-Hermite algorithm. Several important results will be given including the effectiveness of $\gamma^{(i)}$ in estimating the condition number of the Sylvester matrix $T_{m^{(i)}}$, the relationship between the tolerance ϵ and the relative and residual error, and a comparison between the relative error for our iterative numerical algorithm and that of the direct method of solving $T_{m^{(i)}}$ using Gaussian elimination. We also compare how the relative and residual errors compare with bounds of Cabay and Meleshko highlighted in Section 5.2.

The results will be separated into three sections. In Section 7.1, we consider results for various values of k when $A_0(z) = 1$. Because our method for generating power series with singular points does not produce such problems, we rely on randomly generated power series coefficients. Section 7.2 deals with problems in which $A_0(z)$ is a formal power series not equal to 1. In this section we consider power series from classes II, III and IV for various k. In Section 7.3 we study the effect of modifying the power series $A(z)$ by multiplying it by $A_0^{-1}(z)$ before testing. A new definition for the stability parameter γ will be provided in Section 7.4 motivated by the experimental results of the previous sections. Within each section we adopt the format of presenting a set of observations in point form, a discussion of the implication of each observation, and a set of tables corroborating the observations.

7.1 Results for $A_0(z) = 1$

As is often the case in the literature, we consider the case in which the power series $A_0(z)$ is equal to the constant value 1. For our tests, all the remaining input power series coefficients are randomly generated (class I problem).

Observation 1: The stability parameter $\gamma^{(i)}$ provides a good estimate for the condition number of $T_{m^{(i)}}$ when $k = 1$.

One of the premises upon which our algorithm is based is that $\gamma^{(i)}$ is a good estimate of the inverse of the condition number of $T_{m^{(i)}}$. By specifying the tolerance $\epsilon = 10^{-3}$ we expect to reject points with a corresponding condition number $\kappa(T_{m^{(i)}}) \geq 1.0(3)$. In Table 7.1, there is a strong correspondence between the value of $\kappa(T_{m^{(i)}})$ and the value of $\gamma^{(i)}$. For example the point corresponding to $i = 4$, the condition number estimate $1/\gamma^{(i)}$ is 58.8 whereas the computed condition number is 28. Similarly when $i = 24$, the inverse of $\gamma^{(i)}$ is 526 and the computed condition number is approximately 890. Observe that for unstable points such as $\|m^{(i)}\| = 43$, $1/\gamma^{(i)} = 1.8(5)$ is still a reasonable estimate of the computed condition number 3.5(4). From this table it appears that $1/\gamma^{(i)}$ is within an order of magnitude of the actual condition number of $T_{m^{(i)}}$. The numerical results of Table 7.1 confirm the numerical results of Meleshko and Cabay [22].

Observation 2: For $k > 1$, $\gamma^{(i)}$ is a gross over-estimate of $T_{m^{(i)}}$. The discrepancy

Table 7.1: $k = 1, \quad \epsilon = 10^{-3}, \quad$ Class I

i	$\kappa(T_{m^{(i)}})$	$\Vert m^{(i)}\Vert$	$\gamma^{(i)}$	$\Vert\delta R^{(i)}(z)\Vert$	$\frac{\Vert\delta S^{(i)}(z)\Vert}{\Vert S_E^{(i)}(z)\Vert}$	$\frac{\Vert\delta S_G^{(i)}(z)\Vert}{\Vert S_E^{(i)}(z)\Vert}$
1	1.0	1	.37	0.0	5.6(-17)	5.6(-17)
2	7.0	3	5.5(-2)	2.8(-17)	6.1(-17)	8.8(-17)
3	25	5	1.0(-2)	4.2(-17)	1.6(-16)	6.4(-17)
4	28	7	1.7(-2)	1.7(-16)	4.7(-16)	9.7(-17)
5	67	9	9.0(-3)	1.7(-16)	7.8(-16)	3.1(-16)
6	1.5(2)	11	6.7(-3)	1.7(-16)	8.2(-16)	2.9(-16)
7	50	13	2.2(-2)	1.8(-16)	1.2(-15)	3.6(-16)
8	1.7(2)	15	2.2(-3)	2.7(-16)	2.3(-15)	2.7(-16)
9	78	17	1.4(-2)	5.4(-16)	4.0(-15)	4.6(-16)
10	1.4(2)	19	9.9(-3)	5.4(-16)	4.0(-15)	5.1(-16)
11	3.9(2)	21	1.9(-3)	3.7(-16)	9.5(-15)	8.1(-16)
12	5.5(2)	23	1.3(-3)	3.9(-16)	1.0(-14)	1.1(-15)
13	3.8(2)	25	4.9(-3)	3.9(-16)	1.2(-14)	3.5(-15)
14	5.1(2)	27	5.2(-3)	4.4(-16)	1.5(-14)	4.4(-15)
15	3.9(2)	29	1.1(-2)	5.4(-16)	1.4(-14)	2.1(-15)
16	3.7(2)	31	6.0(-3)	5.3(-16)	1.1(-14)	6.9(-15)
17	5.8(2)	33	4.6(-3)	4.7(-16)	2.7(-14)	8.1(-15)
18	6.8(2)	35	5.3(-3)	5.9(-16)	2.7(-14)	1.0(-14)
19	1.0(3)	37	1.8(-3)	6.0(-16)	1.8(-14)	1.1(-14)
20	9.5(2)	39	3.6(-3)	5.8(-16)	1.7(-14)	8.5(-15)
21	7.9(2)	41	5.7(-3)	5.9(-16)	1.4(-14)	2.4(-14)
-	3.5(4)	43	5.6(-6)	6.0(-16)	unstable	-
-	6.0(3)	45	3.6(-4)	5.7(-16)	unstable	-
22	1.6(3)	47	2.4(-3)	4.1(-16)	1.1(-14)	4.1(-14)
23	1.1(3)	49	2.1(-3)	5.6(-16)	7.9(-15)	1.4(-13)
-	7.8(3)	51	2.3(-5)	5.8(-16)	unstable	-
24	8.9(2)	53	1.9(-3)	5.6(-16)	6.8(-15)	1.6(-13)
-	5.1(3)	55	1.6(-4)	5.9(-16)	unstable	-
-	5.4(3)	57	1.5(-4)	5.6(-16)	unstable	-
-	3.7(3)	59	1.3(-4)	6.1(-16)	unstable	-

Table 7.2: $k = 2$, $\epsilon = 10^{-8}$, Class I

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	7.6	2	9.1(-3)	5.6(-17)	1.4(-16)	1.4(-16)
2	16	5	6.5(-3)	1.7(-16)	5.3(-16)	1.9(-16)
3	27	8	6.8(-3)	1.9(-16)	1.0(-15)	3.0(-16)
4	89	11	4.8(-4)	2.3(-16)	1.4(-15)	4.1(-16)
5	94	14	7.8(-4)	1.9(-16)	1.2(-15)	4.1(-16)
-	2.2(4)	17	5.7(-11)	1.7(-16)	unstable	-
6	1.7(2)	20	3.1(-4)	2.4(-16)	1.8(-15)	4.1(-16)
7	2.1(2)	23	1.2(-4)	2.3(-16)	1.3(-15)	5.0(-16)
8	5.0(2)	26	1.4(-5)	3.2(-16)	1.8(-15)	5.3(-16)
9	5.0(2)	29	4.3(-5)	4.3(-16)	4.9(-15)	2.1(-15)
10	3.1(2)	32	4.4(-5)	5.3(-16)	4.4(-15)	1.1(-15)
-	1.3(4)	35	3.7(-9)	5.8(-16)	unstable	-
11	2.7(2)	38	1.6(-4)	7.9(-16)	9.2(-15)	6.5(-16)
12	1.2(3)	41	6.6(-6)	6.9(-16)	1.1(-14)	7.5(-16)
13	4.2(2)	44	6.6(-5)	6.8(-16)	1.0(-14)	8.2(-16)
14	9.7(2)	47	1.4(-6)	8.0(-16)	1.3(-14)	1.9(-15)
15	1.9(3)	50	1.2(-6)	1.0(-15)	1.7(-14)	2.3(-15)
16	1.3(3)	53	8.4(-6)	1.3(-15)	3.0(-14)	1.1(-15)
17	5.4(3)	56	4.1(-7)	1.6(-15)	1.8(-13)	2.0(-15)
18	8.6(2)	59	7.5(-6)	2.0(-15)	4.1(-14)	1.9(-15)
19	8.5(3)	62	1.8(-8)	2.3(-15)	3.1(-14)	2.4(-15)
20	5.7(2)	65	2.3(-5)	3.5(-15)	4.2(-14)	1.4(-15)
⋮	⋮	⋮	⋮	⋮	⋮	⋮
85	3.8(4)	299	2.5(-8)	3.7(-14)	3.7(-12)	8.7(-15)
86	2.2(4)	302	2.7(-8)	3.5(-14)	3.6(-12)	8.6(-15)
87	2.2(4)	305	7.6(-8)	3.2(-14)	4.4(-12)	1.4(-14)
88	3.6(4)	308	2.5(-8)	2.8(-14)	6.4(-12)	1.4(-14)
89	3.0(4)	311	4.0(-8)	2.7(-14)	8.5(-12)	1.4(-13)
-	1.2(5)	314	1.0(-9)	2.4(-13)	unstable	-
-	1.5(5)	317	5.2(-10)	3.4(-14)	unstable	-
90	3.0(4)	320	6.3(-8)	3.5(-14)	4.7(-12)	1.5(-14)
91	4.8(4)	323	3.6(-8)	3.5(-14)	7.1(-12)	2.6(-14)

Table 7.3: $k = 7, \quad \epsilon = 10^{-19}, \quad$ Class I

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	85	7	5.2(-8)	8.3(-17)	2.7(-16)	2.7(-16)
2	4.5(2)	15	3.0(-14)	2.6(-16)	1.4(-15)	1.7(-15)
-	1.5(5)	23	5.3(-32)	9.7(-16)	unstable	-
3	2.4(3)	31	8.0(-16)	1.3(-14)	1.1(-12)	2.8(-15)
4	5.2(2)	39	2.7(-12)	1.3(-13)	1.1(-12)	2.9(-15)
5	7.4(2)	47	1.2(-12)	1.4(-13)	1.4(-12)	2.1(-15)
6	3.7(3)	55	7.8(-16)	1.2(-13)	2.7(-12)	8.9(-15)
7	4.0(3)	63	7.6(-15)	2.5(-13)	6.6(-12)	9.8(-15)
8	3.8(3)	71	1.1(-14)	3.3(-13)	2.4(-11)	6.1(-15)
9	6.5(3)	79	1.3(-16)	3.5(-13)	4.3(-11)	6.3(-15)
10	2.8(3)	87	1.9(-13)	4.6(-13)	1.8(-11)	5.8(-15)
11	5.7(3)	95	3.4(-17)	4.5(-13)	1.2(-11)	1.7(-14)
12	1.5(4)	103	5.0(-18)	5.8(-13)	1.9(-11)	1.2(-14)
-	1.1(6)	111	5.8(-36)	3.2(-13)	unstable	-
-	1.8(4)	119	6.4(-21)	2.6(-13)	unstable	-
13	1.8(4)	127	1.9(-19)	5.3(-13)	1.7(-11)	2.1(-14)
14	2.5(4)	135	5.3(-19)	7.2(-13)	1.4(-11)	1.9(-14)
15	1.2(4)	143	9.0(-17)	7.7(-13)	4.1(-11)	1.3(-14)
16	1.4(4)	151	2.7(-18)	4.5(-13)	2.6(-11)	8.0(-15)
17	1.4(4)	159	1.7(-17)	6.9(-13)	2.0(-11)	1.5(-14)
-	7.7(4)	167	8.7(-23)	6.4(-13)	unstable	-
18	9.2(3)	175	1.5(-16)	5.7(-13)	1.4(-11)	1.0(-14)
-	5.4(5)	183	2.2(-29)	2.9(-13)	unstable	-
19	2.3(4)	191	5.1(-19)	5.4(-13)	1.5(-11)	1.8(-14)
-	7.1(4)	199	8.6(-22)	5.0(-13)	unstable	-
-	5.0(4)	207	9.0(-21)	4.1(-13)	unstable	-
-	4.1(4)	215	1.7(-20)	4.5(-13)	unstable	-
-	7.7(4)	223	4.0(-22)	3.5(-13)	unstable	-
20	3.4(4)	231	1.2(-19)	4.7(-13)	2.4(-11)	2.0(-14)
-	4.5(4)	239	7.5(-21)	6.2(-13)	unstable	-

Table 7.4: $k = 7, \quad \epsilon = 10^{-16}, \quad$ Class I

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	85	7	5.2(-8)	8.3(-17)	2.7(-16)	2.7(-16)
2	4.5(2)	15	3.0(-14)	2.6(-16)	1.4(-15)	1.7(-15)
-	1.5(5)	23	5.3(-32)	9.7(-16)	unstable	-
3	2.4(3)	31	8.0(-16)	1.3(-14)	1.1(-12)	2.8(-15)
4	5.2(2)	39	2.7(-12)	1.3(-13)	1.1(-12)	2.9(-15)
5	7.4(2)	47	1.2(-12)	1.4(-13)	1.4(-12)	2.1(-15)
6	3.7(3)	55	7.8(-16)	1.2(-13)	2.7(-12)	8.9(-15)
7	4.0(3)	63	7.6(-15)	2.5(-13)	6.6(-12)	9.8(-15)
8	3.8(3)	71	1.1(-14)	3.3(-13)	2.4(-11)	6.1(-15)
9	6.5(3)	79	1.3(-16)	3.5(-13)	4.3(-11)	6.3(-15)
10	2.8(3)	87	1.9(-13)	4.6(-13)	1.8(-11)	5.8(-15)
-	5.7(3)	95	3.4(-17)	4.5(-13)	unstable	-
-	1.5(4)	103	5.0(-18)	5.8(-13)	unstable	-
-	1.1(6)	111	5.8(-36)	3.2(-13)	unstable	-
-	1.8(4)	119	6.4(-21)	2.6(-13)	unstable	-
-	1.8(4)	127	1.9(-19)	5.3(-13)	unstable	-
-	2.5(4)	135	5.3(-19)	7.2(-13)	unstable	-
-	1.2(4)	143	9.0(-17)	7.8(-13)	unstable	-
-	1.4(4)	151	2.7(-18)	4.5(-13)	unstable	-
-	1.4(4)	159	1.7(-17)	7.0(-13)	unstable	-
-	7.7(4)	167	8.7(-23)	6.4(-13)	unstable	-
11	9.2(3)	175	1.5(-16)	5.7(-13)	1.5(-11)	1.0(-14)
-	5.4(5)	183	2.2(-29)	3.0(-13)	unstable	-
-	2.3(4)	191	5.1(-19)	5.3(-13)	unstable	-
-	7.1(4)	199	8.6(-22)	5.2(-13)	unstable	-
-	5.0(4)	207	9.0(-21)	4.2(-13)	unstable	-
-	4.1(4)	215	1.7(-20)	4.6(-13)	unstable	-
-	7.7(4)	223	4.0(-22)	3.4(-13)	unstable	-
-	3.4(4)	231	1.2(-19)	4.5(-13)	unstable	-
-	4.5(4)	239	7.5(-21)	6.0(-13)	unstable	-

is magnified with increased k. A closer approximation is given by[1]

$$\kappa\left(T_{m^{(i)}}\right) = \left(\frac{1}{\gamma^{(i)}}\right)^{\frac{2}{k+1}} \tag{7.1}$$

Consider the point $i = 14$ in Table 7.2 with $\kappa(T_{m^{(i)}}) = 970$. The stability parameter $\gamma^{(i)}$ estimates the condition number to be 7.1(5). Similarly, at the point $i = 12$ in Table 7.3 with $k = 7$, $1/\gamma^{(i)} = 2.0(17)$ whereas the actual condition number is 1.5(4), a difference of approximately 10^{13}. We should not be completely surprised by the fact that $\gamma^{(i)}$ decreases as k increases. The parameter $\gamma^{(i)}$ is the product of the values of $V_{j,j}^{(i)}(0)$, $j = 1, k$, and the residual element $R_0^{(i)}(0)$. If these values are each $\approx 10^{-1}$, then $\gamma^{(i)}$ will be about $10^{-(k+1)}$.

Applying (7.1) to the points listed above we have that for the point $i = 14$ in Table 7.2 $(1/\gamma^{(i)})^{2/3} = 7.9(3)$ and for the point $i = 12$ in Table 7.3, $(1/\gamma^{(i)})^{2/3} = 2.1(4)$. Each of these values lies within an order of magnitude of the actual condition number of $T_{m^{(i)}}$.

Observation 3: Although $\gamma^{(i)}$ does not accurately estimate the condition number $\kappa(T_{m^{(i)}})$, there is a relationship between the two values. Observe that for two points in a given table, if $\kappa(T_{m^{(i)}}) > \kappa(T_{m^{(j)}})$, then the relationship $\gamma^{(i)} < \gamma^{(j)}$ will hold. This inverse ordering applies as long as there is a reasonable difference in the values of $\kappa(T_{m^{(i)}})$ and $\kappa(T_{m^{(j)}})$.

As an example, this ordering holds for the points $i = 12$ and $i = 87$ in Table 7.2. We see that $\kappa(T_{m^{(87)}}) = 2.2(4) > 1.2(3) = \kappa(T_{m^{(12)}})$ and that $\gamma^{(87)} = 7.6(-8) < 6.6(-6) = \gamma^{(12)}$.

Observation 4: An approximation of the relative error of the iterative algorithm

[1] A better approximation is given in Section 7.4.

for the case $A_0(z) = 1$ is given by

$$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|} \approx \max_{1 \le j \le i} \kappa(T_{m^{(j)}}) \cdot \|m^{(i)}\| \cdot \mu. \tag{7.2}$$

In Section 5.2 we saw the error bound of Cabay and Meleshko for the relative error of a PHS with $k = 1$. This error bound involves the sum of the ratio of some low degree polynomials to the stability parameter $\gamma^{(i)}$. However, we have shown that $1/\gamma^{(i)}$ is not a very good estimate of $\kappa(T_{m^{(i)}})$. Hence a more accurate formulation of (5.15) might be

$$\begin{aligned} \|\delta S^{(j)}(z)\| \;\le\; & 2.2\,(m_0^{(j)} + 1)^3\, \kappa(T_{m^{(j)}})\, [\Delta_0^{(j-1)} \cdot \kappa(T_{m^{(j-1)}}) + \\ & \sum_{i=0}^{j-2} \Delta_1^{(i)} \cdot \kappa(T_{m^{(i)}}) \cdot \kappa(T_{m^{(i+1)}})]\, \mu \,+\, \mathcal{O}\left(\mu^2\right), \end{aligned} \tag{7.3}$$

where $\Delta_0^{(j)}$ and $\Delta_1^{(j)}$ are low degree polynomials in $m_0^{(j)}$ and s_j. Although these error bounds are supported by Tables 7.1, 7.2, and 7.3, this formulation shows how weak the bounds actually are. The operational bounds (7.2) are considerably closer to the actual error. For example in Table 7.3, the relative error at $i = 5$ would be predicted by (7.2) to be $2.4(3) \cdot 47 \cdot 2.2(-16) = 2.5(-11)$ and the actual relative error was $\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|} = 1.4(-12)$. Using the error bound (7.3) with $\Delta_0^{(j)}$ and $\Delta_1^{(j)}$ given by (5.12), the predicted relative error is .14. Note that this operational bound implies that the growth of the relative error in the PHS is linear as opposed to exponential. Furthermore, $\|\delta S(z)\|$ appears to be proportional to the maximum $\kappa(T_{m^{(i)}})$ as opposed to the sum of condition numbers indicated by (7.3).

Observation 5: The approximate error in the residual $\delta R^{(i)}(z)$ obeys the relationship

$$\|\delta R^{(i)}(z)\| \approx \max_{1 \le j \le i} \kappa(T_{m^{(j)}}) \cdot \mu. \tag{7.4}$$

As in Observation 4, we can convert the statement of Theorem 5.1 to use $\kappa(T_{m^{(i)}})$ instead of $\gamma^{(i)}$. This becomes

$$\|\delta R^{(j+1)}(z)\| \leq \left(\Delta_0^{(j)} \cdot \kappa(T_{m^{(j)}}) + \sum_{l=0}^{j-1} \Delta_1^{(l)} \cdot \kappa(T_{m^{(l)}}) \cdot \kappa(T_{m^{(l+1)}})\right) \mu + O\left(\mu^2\right). \quad (7.5)$$

The error bound (7.5) is also supported by Tables 7.1 - 7.4. Using the point $i = 5$ of Table 7.3 the error)) in $\delta R^{(5)}(z)$ is predicted by (7.4) to be $2.4(3) \times \mu = 5.3(-13)$. The actual error was found to be $\delta R^{(5)}(z) = 1.4(-13)$ and the value predicted by (7.5) is 6.0(-3). Once again, the bound (7.4) implies that the growth of error in the residual is linear. Notice also that the growth in the residual error depends on the maximum $\kappa(T_{m^{(i)}})$ rather than the sum of condition numbers as suggested by (7.5).

Observation 6: With the tolerance ϵ properly chosen, the relative error occurring in Padé-Hermite System polynomial coefficients is comparable to the relative error of the direct method using Gaussian elimination. Unfortunately, because of the lack of correspondence between $\gamma^{(i)}$ and $\kappa(T_{m^{(i)}})$, the tolerance ϵ was chosen *a posteriori* for many of the experiments.

In Table 7.1 we see that the our iterative algorithm is marginally better than the direct method whereas in Tables 7.2 and 7.3, a slight advantage is held by the direct method. Note that in these experiments, no points with large $\kappa(T_{m^{(i)}})$ were accepted. In Section 7.2 we will see that accepting such points has a dismal effect on the accuracy of PHS polynomial entries. In such cases, Gaussian elimination on the matrix $T_{m^{(i)}}$ produces much more accurate solutions.

Observation 7: For well conditioned problems, the choice of the tolerance ϵ is not critical to the performance of the algorithm. In these Tables 7.3 and 7.4 we can see that increasing the tolerance causes fewer points to be accepted but little improvement in the PHS relative error and residual error. In fact if we were to accept all points as stable by using a tolerance of zero, then by (7.2) we could expect to have a relative error of approximately 5.8(-8) and error in the residual equivalent to 2.4(-10). In Section 7.2 we shall see that the choice of tolerance is an integral part of achieving accurate results for problem classes II, III and IV.

7.2 Results for $A_0(z) \neq 1$

In this section we present a number of experiments for class II, III and IV problems where $A_0(z)$ is a formal power series not identically 1. Using the method of Section 6.2 for generating a vector of power series containing singularities, we can assemble problems with a wide range of instabilities. Because of the growth of power series coefficients associated with this method, only experiments where $k = 1, 2$ were performed. For the most part, the observations made in Section 7.1 will again be supported by these experiments. We will begin by presenting a brief description of the experiments.

Tables 7.5 - 7.10 document class II, III, and IV problems for $k = 1$. In Maple, a vector of power series was generated so that the points corresponding to $\|m^{(i)}\| = 1, 5, 7, 11, 19, 23, 29, 35, 37, 41, 47, 49, 53,$ and 57 would be singular. The results of experiments using these power series (without perturbations in the coefficients except for those arising from roundoff) are given in Tables 7.5 and 7.6. Perturbing the coefficients each by 10^{-12} led to the class III results of Tables 7.7 and 7.8. Finally, perturbing the original power series coefficients by 10^{-6} yielded class IV input for experiments documented in Tables 7.9 - 7.11.

Similar experiments were run for $k = 2$. A class II problem with singularities at $\|m^{(i)}\| = 8, 17, 20, 32, 41, 44, 47,$ and 53 was generated. Table 7.12 report some results for this unperturbed problem. Type III problems with 10^{-12} perturbations of this the class II data are given in Tables 7.13 and 7.14. Perturbing the class II problem by 10^{-6} led to the class III problem reported in Tables 7.15 and 7.16.

Before listing the observations that can be made from these data sets, we note that the error bounds given in Observations 4 and 5 also hold for the case where $A(z) \neq 1$. Also note that for these experiments, Observation 6, concerning the correlation between the relative error of the iterative and direct methods, is valid.

Observation 8: When $A_0(z) \neq 1$, $\gamma^{(i)}$ is an unreliable estimate of $\kappa(T_{m^{(i)}})$ for all k. Consider Table 7.5 at the point $i = 10$. The value of $\kappa(T_{m^{(i)}})$ is 1.3(9) and the estimate given by $1/\gamma^{(i)}$ is 9.0(5). In Table 7.9 at the unstable point $\|m^{(i)}\| = 8$, the condition number of $T_{m^{(i)}}$ is 2.2(10) whereas the value of $1/\gamma^{(i)}$ is 5.0(17). The estimate $1/\gamma^{(i)}$ is no longer a consistent overestimate of $\kappa(T_{m^{(i)}})$ as it was in Section 7.1. As these examples show, $1/\gamma^{(i)}$ may be an overestimate or an underestimate of $\kappa(T_{m^{(i)}})$. Hence we cannot predict the condition number with the alternative approximation (7.1) given in Observation 2.

Observation 9: The choice of ϵ is extremely important for problems in which $A(z) \neq 1$. In choosing a tolerance ϵ, several criteria must be considered. The prime motivating factor in choosing the tolerance is the amount of accuracy required in the PHS coefficients. The larger the tolerance, the more accurate the solution will be due to selective nature of the algorithm in accepting Padé-Hermite table points as stable. However, if the tolerance is too large, large unstable blocks will appear in the table. In the worst case, all points are rejected and the solution is computed by solving T_n using Gaussian elimination.

Table 7.6 gives an example of choosing a tolerance too large. All points after $\|m^{(i)}\| = 27$ are rejected due to the tolerance $\epsilon = 10^{-5}$. We see that in trying to find

Table 7.5: $k = 1$, $\epsilon = 10^{-8}$, Class II

i	$\kappa(T_{m^{(i)}})$	$\lVert m^{(i)}\rVert$	$\gamma^{(i)}$	$\lVert\delta R^{(i)}(z)\rVert$	$\frac{\lVert\delta S^{(i)}(z)\rVert}{\lVert S_E^{(i)}(z)\rVert}$	$\frac{\lVert\delta S_G^{(i)}(z)\rVert}{\lVert S_E^{(i)}(z)\rVert}$
-	singular	1	singular	singular	singular	singular
1	1.5	3	0.12	1.4(-17)	2.4(-17)	4.8(-17)
-	8.8(17)	5	2.2(-17)	5.6(-17)	unstable	-
-	1.6(17)	7	7.6(-18)	3.1(-17)	unstable	-
2	5.7(3)	9	4.6(-3)	3.6(-17)	6.4(-15)	1.8(-14)
-	4.0(18)	11	1.2(-30)	8.9(-17)	unstable	-
3	2.3(5)	13	8.4(-4)	8.3(-17)	7.1(-14)	1.1(-13)
4	1.4(6)	15	2.1(-4)	4.2(-17)	7.1(-14)	5.2(-13)
5	4.0(6)	17	9.2(-5)	4.2(-17)	1.8(-13)	3.9(-13)
-	4.2(19)	19	3.6(-31)	5.9(-17)	unstable	-
6	1.5(7)	21	3.4(-5)	5.9(-17)	3.9(-13)	2.5(-12)
-	3.7(18)	23	4.6(-30)	5.9(-17)	unstable	-
7	1.1(8)	25	1.6(-5)	8.7(-17)	1.0(-12)	1.1(-11)
8	1.7(8)	27	7.0(-6)	6.2(-17)	1.0(-12)	1.7(-11)
-	1.3(19)	29	4.0(-29)	1.2(-16)	unstable	-
9	7.4(8)	31	2.7(-6)	1.1(-16)	1.6(-12)	1.8(-10)
10	1.3(9)	33	1.1(-6)	6.8(-17)	1.6(-12)	4.4(-10)
-	1.9(19)	35	4.3(-19)	1.2(-16)	unstable	-
-	1.2(19)	37	4.3(-19)	1.5(-16)	unstable	-
11	1.9(9)	39	6.0(-7)	7.9(-17)	4.0(-12)	4.2(-10)
-	1.4(19)	41	8.5(-30)	1.4(-16)	unstable	-
12	5.9(9)	43	2.8(-7)	1.2(-16)	4.7(-12)	1.3(-9)
13	9.2(9)	45	1.0(-7)	1.4(-16)	1.6(-12)	1.7(-9)
-	1.8(19)	47	1.1(-19)	1.4(-16)	unstable	-
-	1.2(19)	49	1.1(-19)	1.2(-16)	unstable	-
14	9.4(9)	51	5.8(-8)	1.5(-16)	7.1(-12)	1.8(-9)
-	3.3(19)	53	4.2(-29)	1.8(-16)	unstable	-
15	1.3(10)	55	2.8(-8)	1.1(-16)	4.4(-11)	1.5(-9)
-	5.0(18)	57	6.6(-30)	1.0(-16)	unstable	-
16	3.7(10)	59	1.4(-8)	1.4(-16)	4.4(-11)	5.5(-9)

Table 7.6: $k = 1$, $\epsilon = 10^{-5}$, Class II

i	$\kappa\,(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
-	singular	1	singular	singular	singular	singular
1	1.5	3	0.12	1.4(-17)	2.4(-17)	4.8(-17)
-	8.8(17)	5	2.2(-17)	5.6(-17)	unstable	-
-	1.6(17)	7	7.6(-18)	3.1(-17)	unstable	-
2	5.7(3)	9	4.6(-3)	3.6(-17)	6.4(-15)	1.8(-14)
-	4.0(18)	11	1.2(-30)	8.9(-17)	unstable	-
3	2.3(5)	13	8.4(-4)	8.3(-17)	7.1(-14)	1.1(-13)
4	1.4(6)	15	2.1(-4)	4.2(-17)	7.1(-14)	6.0(-13)
5	4.0(6)	17	9.2(-5)	4.2(-17)	1.8(-13)	3.9(-13)
-	4.2(19)	19	3.6(-31)	5.9(-17)	unstable	-
6	1.5(7)	21	3.4(-4)	5.9(-17)	3.9(-13)	2.5(-12)
-	3.7(18)	23	4.6(-30)	5.9(-17)	unstable	-
7	1.1(8)	25	1.6(-5)	8.7(-17)	1.0(-12)	1.1(-11)
8	1.7(8)	27	7.0(-6)	6.2(-17)	1.0(-12)	1.7(-11)
-	1.3(19)	29	4.1(-29)	9.5(-17)	unstable	-
-	7.4(8)	31	2.7(-6)	1.6(-16)	unstable	-
-	1.3(9)	33	1.1(-6)	1.1(-16)	unstable	-
-	1.9(19)	35	6.2(-19)	1.1(-16)	unstable	-
-	1.2(19)	37	6.1(-19)	7.5(-17)	unstable	-
-	1.9(9)	39	6.0(-7)	1.0(-16)	unstable	-
-	1.4(19)	41	6.0(-30)	1.4(-16)	unstable	-
-	5.9(9)	43	2.8(-7)	1.1(-16)	unstable	-
-	9.2(9)	45	1.0(-7)	1.3(-16)	unstable	-
-	1.8(19)	47	3.4(-20)	7.9(-17)	unstable	-
-	1.2(19)	49	3.1(-20)	1.3(-16)	unstable	-
-	9.4(9)	51	5.8(-8)	9.0(-17)	unstable	-
-	3.3(19)	53	3.3(-29)	9.2(-17)	unstable	-
-	1.3(10)	55	2.8(-8)	1.0(-16)	unstable	-
-	5.0(18)	57	6.1(-30)	1.4(-16)	unstable	-
-	3.7(10)	59	1.4(-8)	1.6(-16)	unstable	-

Table 7.7: $k = 1$, $\epsilon = 10^{-8}$, Class III

i	$\kappa\left(T_{m^{(i)}}\right)$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
-	1.0	1	1.0(-24)	2.0(-28)	unstable	-
1	1.5	3	.12	2.0(-28)	1.0(-16)	1.0(-16)
-	7.2(13)	5	6.1(-14)	5.6(-17)	unstable	-
-	1.8(14)	7	2.1(-14)	4.1(-17)	unstable	-
2	5.7(3)	9	4.6(-3)	5.6(-17)	1.1(-14)	1.2(-14)
-	4.4(14)	11	1.9(-23)	8.3(-17)	unstable	-
3	2.3(5)	13	8.5(-4)	8.3(-17)	2.1(-13)	1.1(-13)
4	1.4(6)	15	2.1(-4)	4.2(-17)	2.1(-13)	3.3(-13)
5	4.0(6)	17	9.2(-5)	7.6(-17)	3.5(-13)	9.8(-13)
-	5.5(15)	19	9.3(-23)	7.6(-17)	unstable	-
6	1.5(7)	21	3.4(-5)	1.2(-16)	1.1(-12)	2.7(-12)
-	2.9(16)	23	5.5(-23)	1.0(-16)	unstable	-
7	1.1(8)	25	1.6(-5)	9.0(-17)	3.3(-12)	1.9(-11)
8	1.7(8)	27	7.0(-6)	9.4(-17)	3.3(-12)	2.0(-11)
-	9.1(16)	29	1.0(-22)	1.2(-16)	unstable	-
9	7.4(8)	31	2.7(-6)	1.2(-16)	1.0(-11)	1.8(-10)
10	1.3(9)	33	1.1(-6)	1.2(-16)	6.6(-12)	1.5(-10)
-	6.6(17)	35	5.9(-16)	1.6(-16)	unstable	-
-	8.7(17)	37	5.9(-16)	2.0(-16)	unstable	-
11	1.9(9)	39	6.0(-7)	1.4(-16)	5.6(-12)	1.8(-10)
-	6.5(18)	41	1.7(-25)	1.4(-16)	unstable	-
12	5.9(9)	43	2.8(-7)	1.3(-16)	9.1(-12)	8.4(-10)
13	9.2(9)	45	1.0(-7)	1.1(-16)	2.1(-12)	1.2(-9)
-	3.2(18)	47	4.3(-18)	1.4(-16)	unstable	-
-	2.0(19)	49	4.3(-18)	1.6(-16)	unstable	-
14	9.4(9)	51	5.8(-8)	1.5(-16)	1.1(-11)	1.3(-9)
-	7.9(18)	53	2.9(-25)	1.6(-16)	unstable	-
15	1.3(10)	55	2.8(-8)	1.1(-16)	4.1(-11)	1.5(-9)
-	4.4(18)	57	1.0(-24)	1.3(-16)	unstable	-
16	3.7(10)	59	1.4(-8)	1.1(-16)	4.[illegible](-11)	4.0(-9)

Table 7.8: $k = 1$, $\epsilon = 10^{-14}$, Class III

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
-	1.0	1	1.0(-24)	2.0(-28)	unstable	-
1	1.5	3	.12	2.0(-28)	1.0(-16)	1.0(-16)
2	7.2(13)	5	6.1(-14)	5.6(-17)	1.8(-4)	1.3(-4)
3	1.8(14)	7	2.1(-14)	3.4(-17)	1.8(-4)	2.8(-4)
4	5.7(3)	9	4.6(-3)	5.8(-6)	1.8(-3)	1.2(-14)
5	4.4(14)	11	5.5(-12)	3.3(-6)	2.0	7.5(-14)
6	2.3(5)	13	8.5(-4)	2.7(-6)	4.3(-3)	1.1(-13)
7	1.4(6)	15	2.1(-4)	1.0(-6)	1.4(-2)	3.3(-13)
8	4.0(6)	17	9.3(-5)	1.0(-6)	1.4(-2)	9.8(-13)
9	5.5(15)	19	7.1(-8)	5.0(-7)	2.0	8.9(-13)
10	1.5(7)	21	3.5(-5)	6.6(-7)	5.9(-2)	2.7(-12)
11	2.9(16)	23	1.0(-6)	6.8(-7)	.13	5.7(-12)
12	1.1(8)	25	1.4(-5)	6.8(-7)	.17	1.9(-11)
13	1.7(8)	27	4.9(-6)	3.6(-7)	.17	2.0(-11)
14	9.1(16)	29	4.7(-7)	3.8(-7)	2.0	1.6(-11)
15	7.4(8)	31	2.0(-6)	3.8(-7)	.54	1.8(-10)
16	1.3(9)	33	8.9(-7)	4.2(-7)	.14	1.5(-10)
17	6.6(17)	35	2.0(-9)	4.2(-7)	2.0	7.3(-2)
18	8.7(17)	37	8.0(-9)	4.6(-7)	2.0	9.0(-2)
19	1.9(9)	39	6.1(-7)	4.6(-7)	7.6(-2)	1.8(-10)
20	6.5(18)	41	4.4(-10)	4.6(-7)	2.0	8.8(-10)
21	5.9(9)	43	2.9(-7)	4.6(-7)	4.2(-2)	8.4(-10)
22	9.2(9)	45	1.0(-7)	3.3(-7)	1.5(-2)	1.2(-9)
23	3.2(18)	47	1.3(-11)	3.3(-7)	2.0	2.0
24	2.0(19)	49	4.9(-11)	3.4(-7)	2.0	2.0
25	9.4(9)	51	5.7(-8)	3.4(-7)	.14	1.3(-9)
26	7.9(18)	53	2.1(-10)	3.5(-7)	.20	2.8(-9)
27	1.3(10)	55	2.8(-8)	3.5(-7)	.20	1.5(-9)
28	4.4(18)	57	1.2(-10)	1.9(-7)	.23	2.3(-9)
29	3.7(10)	59	1.2(-8)	2.3(-7)	.23	4.0(-9)

Table 7.9: $k = 1$, $\epsilon = 10^{-8}$, Class IV

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_F^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_F^{(i)}(z)\|}$
-	1.0	1	1.0(-12)	0.0	unstable	-
1	1.5	3	.12	2.8(-17)	9.4(-17)	9.4(-17)
2	7.2(7)	5	6.1(-8)	5.6(-17)	3.7(-10)	1.1(-10)
3	1.8(8)	7	2.1(-8)	3.6(-17)	3.7(-10)	3.6(-10)
4	5.7(3)	9	4.6(-3)	5.5(-12)	4.5(-9)	1.8(-14)
-	4.4(8)	11	1.9(-11)	3.1(-12)	unstable	-
5	2.3(5)	13	8.6(-4)	2.5(-12)	1.7(-8)	4.2(-13)
6	1.4(6)	15	2.1(-4)	9.6(-13)	4.5(-8)	1.1(-13)
7	4.0(6)	17	9.2(-5)	6.4(-13)	4.9(-8)	8.7(-13)
-	5.5(9)	19	9.3(-11)	5.3(-13)	unstable	-
8	1.5(7)	21	3.4(-5)	5.3(-13)	1.2(-7)	2.1(-12)
-	2.9(10)	23	5.5(-11)	4.5(-13)	unstable	-
9	1.1(8)	25	1.6(-5)	4.5(-13)	6.4(-7)	2.0(-11)
10	1.7(8)	27	6.9(-6)	4.2(-13)	6.4(-7)	1.7(-11)
-	9.0(10)	29	1.1(-10)	4.2(-13)	unstable	-
11	7.6(8)	31	2.7(-6)	4.2(-13)	4.6(-7)	1.2(-10)
12	1.3(9)	33	1.1(-6)	4.0(-13)	4.6(-7)	2.5(-10)
-	7.3(11)	35	5.8(-10)	3.7(-13)	unstable	-
-	9.1(11)	37	5.7(-10)	3.7(-13)	unstable	-
13	1.9(9)	39	6.0(-7)	3.7(-13)	1.2(-7)	4.5(-10)
-	5.1(12)	41	2.2(-13)	2.7(-13)	unstable	-
14	6.1(9)	43	2.8(-7)	2.7(-13)	7.6(-8)	7.2(-10)
15	9.2(9)	45	1.0(-7)	2.5(-13)	7.1(-8)	1.1(-9)
-	4.2(13)	47	4.4(-12)	2.2(-13)	unstable	-
-	4.6(13)	49	4.4(-12)	2.2(-13)	unstable	-
16	9.5(9)	51	5.8(-8)	2.1(-13)	4.9(-7)	1.5(-9)
-	3.7(12)	53	2.8(-13)	1.9(-13)	unstable	-
17	1.3(10)	55	2.8(-8)	1.9(-13)	8.6(-7)	1.6(-9)
-	3.3(12)	57	9.9(-13)	1.6(-13)	unstable	-
18	3.8(10)	59	1.5(-8)	2.0(-13)	1.1(-6)	1.1(-8)

Table 7.10: $k = 1$, $\epsilon = 10^{-12}$, Class IV

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
-	1.0	1	1.0(-12)	0.0	unstable	-
1	1.5	3	.12	2.8(-17)	9.4(-17)	9.4(-17)
2	7.2(7)	5	6.1(-8)	5.6(-17)	3.7(-10)	1.1(-10)
3	1.8(8)	7	2.1(-8)	3.6(-17)	3.7(-10)	3.6(-10)
4	5.7(3)	9	4.6(-3)	5.5(-12)	4.5(-9)	1.8(-14)
5	4.4(8)	11	1.9(-11)	3.1(-12)	5.2(-9)	1.8(-13)
6	2.3(5)	13	8.6(-4)	2.3(-9)	4.2(-7)	4.2(-13)
7	1.4(6)	15	2.1(-4)	1.5(-9)	1.1(-6)	1.1(-13)
8	4.0(6)	17	9.2(-5)	1.0(-9)	2.7(-6)	8.7(-13)
9	5.5(9)	19	9.3(-11)	6.3(-10)	2.7(-6)	1.0(-12)
10	1.5(7)	21	3.4(-5)	6.4(-10)	4.1(-6)	2.1(-12)
11	2.9(10)	23	5.5(-11)	6.4(-10)	5.1(-6)	7.6(-12)
12	1.1(8)	25	1.6(-5)	6.4(-10)	5.1(-6)	2.0(-11)
13	1.7(8)	27	6.9(-6)	6.6(-10)	5.6(-6)	1.7(-11)
14	9.0(10)	29	1.1(-10)	6.6(-10)	5.6(-6)	1.2(-11)
15	7.6(8)	31	2.7(-6)	6.6(-10)	4.6(-6)	1.2(-10)
16	1.3(9)	33	1.1(-6)	6.3(-10)	1.5(-5)	2.5(-10)
17	7.3(11)	35	5.8(-10)	5.9(-10)	6.1(-4)	5.0(-8)
18	9.1(11)	37	5.7(-10)	5.9(-10)	6.1(-4)	4.8(-8)
19	1.9(9)	39	6.0(-7)	5.9(-10)	4.0(-5)	4.5(-10)
-	5.1(12)	41	2.1(-13)	4.3(-10)	unstable	-
20	6.1(9)	43	2.8(-7)	4.3(-10)	5.6(-5)	7.2(-10)
21	9.2(9)	45	1.0(-7)	3.4(-10)	2.7(-5)	1.1(-9)
22	4.2(13)	47	4.9(-12)	3.7(-10)	1.7(-2)	3.6(-7)
23	4.6(13)	49	4.9(-12)	3.7(-10)	1.7(-2)	2.1(-6)
24	9.5(9)	51	5.8(-8)	3.2(-10)	9.7(-5)	1.5(-9)
-	3.7(12)	53	2.7(-13)	2.8(-10)	unstable	-
25	1.3(10)	55	2.8(-8)	2.8(-10)	1.8(-4)	1.6(-9)
-	3.3(12)	57	9.8(-13)	2.7(-10)	unstable	-
26	3.8(10)	59	1.5(-8)	3.1(-10)	1.4(-4)	1.1(-8)

Table 7.11: $k = 1$, $\epsilon = 10^{-7}$, Class IV

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
-	1.0	1	1.0(-12)	0.0	unstable	-
1	1.5	3	.12	2.8(-17)	9.4(-17)	9.4(-17)
-	7.2(7)	5	6.1(-8)	5.6(-17)	unstable	-
-	1.8(8)	7	2.1(-8)	2.8(-17)	unstable	-
2	5.7(3)	9	4.6(-3)	8.3(-17)	1.1(-14)	1.8(-14)
-	4.4(8)	11	1.9(-11)	9.3(-17)	unstable	-
3	2.3(5)	13	8.6(-4)	5.6(-17)	1.5(-13)	4.2(-13)
4	1.4(6)	15	2.1(-4)	4.2(-17)	2.0(-13)	1.1(-13)
5	4.0(6)	17	9.2(-5)	9.7(-17)	2.4(-13)	8.7(-13)
-	5.5(9)	19	9.3(-11)	4.7(-17)	unstable	-
6	1.5(7)	21	3.4(-5)	1.1(-16)	6.5(-13)	2.1(-12)
-	2.9(10)	23	5.5(-11)	9.7(-17)	unstable	-
7	1.1(8)	25	1.6(-5)	7.6(-17)	2.2(-12)	2.0(-11)
8	1.7(8)	27	6.9(-6)	1.0(-16)	2.7(-12)	1.7(-11)
-	9.0(10)	29	1.1(-10)	9.7(-17)	unstable	-
9	7.6(8)	31	2.7(-6)	8.9(-17)	6.7(-12)	1.2(-10)
10	1.3(9)	33	1.1(-6)	2.4(-16)	2.3(-12)	2.5(-10)
-	7.3(11)	35	5.8(-10)	1.8(-16)	unstable	-
-	9.1(11)	37	5.7(-10)	1.8(-16)	unstable	-
11	1.9(9)	39	6.0(-7)	1.1(-16)	7.3(-12)	4.5(-10)
-	5.1(12)	41	2.2(-13)	1.2(-16)	unstable	-
12	6.1(9)	43	2.8(-7)	9.4(-17)	7.3(-12)	7.2(-10)
13	9.2(9)	45	1.0(-7)	1.3(-16)	6.9(-12)	1.1(-9)
-	4.2(13)	47	4.4(-12)	1.1(-16)	unstable	-
-	4.6(13)	49	4.4(-12)	1.6(-16)	unstable	-
-	9.5(9)	51	5.8(-8)	1.7(-16)	unstable	-
-	3.7(12)	53	2.8(-13)	1.3(-16)	unstable	-
14	1.3(10)	55	2.8(-8)	1.7(-16)	unstable	-
-	3.3(12)	57	9.9(-13)	1.3(-16)	unstable	-
15	3.8(10)	59	1.5(-8)	1.6(-16)	unstable	-

Table 7.12: $k = 2$, $\epsilon = 10^{-15}$, Class II

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	83	2	4.1(-3)	7.8(-17)	1.9(-15)	1.9(-15)
2	5.7(3)	5	1.0(-4)	5.6(-17)	1.3(-14)	5.6(-14)
-	2.8(18)	8	7.5(-44)	4.2(-17)	unstable	-
3	1.5(7)	11	5.1(-7)	1.4(-17)	9.0(-13)	7.5(-13)
4	6.4(8)	14	3.4(-8)	4.9(-17)	3.0(-11)	1.4(-11)
-	1.3(20)	17	3.8(-38)	4.3(-17)	unstable	-
-	6.6(20)	20	2.7(-32)	4.2(-17)	unstable	-
5	8.4(11)	23	8.8(-10)	7.3(-17)	5.4(-9)	4.1(-9)
6	2.8(13)	26	8.1(-11)	7.0(-17)	3.6(-8)	1.6(-8)
7	2.0(14)	29	5.5(-12)	5.9(-17)	3.5(-8)	1.2(-7)
-	1.3(21)	32	2.2(-33)	1.0(-16)	unstable	-
8	2.0(15)	35	3.3(-13)	7.6(-17)	4.0(-7)	3.7(-7)
9	1.4(16)	38	9.2(-14)	1.1(-16)	3.0(-6)	2.7(-5)
-	1.2(23)	41	6.8(-27)	2.5(-16)	unstable	-
-	2.4(21)	44	7.0(-31)	3.1(-16)	unstable	-
-	1.0(21)	47	2.9(-22)	3.3(-16)	unstable	-
10	1.0(20)	50	6.8(-14)	4.3(-16)	1.3(-2)	.10
-	7.4(20)	53	5.8(-17)	2.7(-16)	unstable	-
11	1.6(21)	56	6.3(-15)	3.2(-16)	.49	1.8
12	4.6(20)	59	5.2(-15)	4.8(-16)	1.1	1.5

Table 7.13: $k = 2, \quad \epsilon = 10^{-13}, \quad$ Class III

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	83	2	4.1(-3)	5.6(-17)	5.7(-15)	5.8(-15)
2	5.7(3)	5	1.0(-4)	2.4(-17)	3.9(-14)	1.3(-14)
-	2.2(16)	8	2.0(-35)	8.3(-17)	unstable	-
3	1.5(7)	11	5.1(-7)	3.5(-17)	4.9(-13)	2.3(-12)
4	6.4(8)	14	3.4(-8)	7.1(-17)	8.3(-12)	9.2(-12)
-	9.6(17)	17	1.1(-30)	3.6(-17)	unstable	-
-	8.4(18)	20	1.2(-25)	3.4(-17)	unstable	-
5	8.4(11)	23	8.8(-10)	5.7(-17)	1.7(-9)	4.9(-9)
6	2.8(13)	26	8.1(-11)	5.9(-17)	6.7(-8)	6.5(-8)
7	2.0(14)	29	5.5(-12)	6.5(-17)	6.7(-8)	1.9(-7)
-	2.1(19)	32	2.6(-25)	1.3(-16)	unstable	-
8	2.0(15)	35	3.3(-13)	1.1(-16)	5.3(-7)	1.0(-6)
-	1.4(16)	38	9.2(-14)	1.2(-16)	unstable	-
-	2.9(19)	41	1.3(-19)	2.2(-16)	unstable	-
-	2.5(20)	44	4.7(-24)	2.3(-16)	unstable	-
-	1.2(20)	47	1.1(-17)	1.8(-16)	unstable	-
-	5.7(19)	50	2.5(-14)	2.4(-16)	unstable	-
-	3.5(19)	53	5.3(-14)	3.9(-16)	unstable	-
-	6.0(19)	56	3.2(-15)	2.2(-16)	unstable	-
-	3.5(19)	59	5.5(-15)	4.1(-16)	unstable	-

Table 7.14: $k = 2$, $\epsilon = 10^{-15}$, Class III

i	$\kappa\left(T_{m^{(i)}}\right)$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	83	2	4.1(-3)	5.6(-17)	5.7(-15)	5.8(-15)
2	5.7(3)	5	1.0(-4)	2.4(-17)	3.9(-14)	1.3(-14)
-	2.2(16)	8	2.0(-35)	8.3(-17)	unstable	-
3	1.5(7)	11	5.1(-7)	3.5(-17)	4.9(-13)	2.3(-12)
4	6.4(8)	14	3.4(-8)	7.1(-17)	8.3(-12)	9.2(-12)
-	9.6(17)	17	1.1(-30)	3.6(-17)	unstable	-
-	8.4(18)	20	1.2(-25)	3.4(-17)	unstable	-
5	8.4(11)	23	8.8(-10)	5.7(-17)	1.7(-9)	4.9(-9)
6	2.8(13)	26	8.1(-11)	5.9(-17)	6.7(-8)	6.5(-8)
7	2.0(14)	29	5.5(-12)	6.5(-17)	6.7(-8)	1.9(-7)
-	2.1(19)	32	2.6(-25)	1.3(-16)	unstable	-
8	2.0(15)	35	3.3(-13)	1.1(-16)	5.3(-7)	1.0(-6)
9	1.4(16)	38	9.2(-14)	1.2(-16)	4.1(-6)	1.3(-7)
-	2.9(19)	41	1.3(-19)	2.3(-16)	unstable	-
-	2.5(20)	44	4.7(-24)	2.7(-16)	unstable	-
-	1.2(20)	47	1.1(-17)	1.7(-16)	unstable	-
10	5.7(19)	50	2.5(-14)	2.5(-16)	1.5(-3)	7.0(-2)
11	3.5(19)	53	5.3(-14)	3.6(-16)	7.7(-4)	8.2(-2)
12	6.0(19)	56	3.2(-15)	3.6(-16)	1.9(-4)	9.7(-2)
13	3.5(19)	59	5.5(-15)	4.6(-16)	3.2(-4)	6.8(-2)

Table 7.15: $k = 2$, $\epsilon = 10^{-13}$, Class IV

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	83	2	4.1(-3)	2.8(-17)	6.3(-15)	6.3(-15)
2	5.7(3)	5	1.0(-4)	4.2(-17)	3.2(-14)	3.3(-14)
-	2.2(10)	8	2.0(-18)	4.8(-17)	unstable	-
3	1.5(7)	11	5.1(-7)	1.5(-17)	8.7(-13)	3.3(-12)
4	6.4(8)	14	3.4(-8)	3.3(-17)	1.6(-11)	8.2(-12)
5	9.5(11)	17	9.9(-13)	4.2(-17)	5.1(-10)	8.0(-10)
-	8.9(12)	20	1.4(-14)	6.5(-17)	unstable	-
6	7.1(11)	23	1.0(-9)	1.6(-14)	3.8(-8)	2.8(-9)
7	1.7(13)	26	4.3(-11)	5.7(-15)	7.5(-8)	1.8(-8)
8	1.3(14)	29	5.3(-13)	5.4(-15)	1.9(-7)	3.1(-8)
9	2.3(14)	32	5.2(-13)	4.4(-15)	3.2(-6)	4.2(-7)
-	3.3(14)	35	5.9(-14)	2.4(-15)	unstable	-
10	1.4(14)	38	1.4(-12)	3.6(-15)	5.1(-7)	8.3(-8)
11	4.3(14)	41	1.1(-13)	4.1(-15)	1.3(-5)	2.3(-7)
-	8.2(14)	44	1.2(-15)	4.0(-15)	unstable	-
-	2.7(15)	47	1.6(-15)	3.9(-15)	unstable	-
-	1.3(15)	50	1.1(-15)	3.5(-15)	unstable	-
-	8.5(14)	53	3.8(-14)	4.4(-15)	unstable	-
-	1.9(15)	56	6.7(-15)	3.8(-15)	unstable	-
-	1.7(15)	59	6.4(-15)	3.5(-15)	unstable	-

Table 7.16: $k = 2$, $\epsilon = 10^{-15}$, Class IV

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	83	2	4.1(-3)	2.8(-17)	6.3(-15)	6.3(-15)
2	5.7(3)	5	1.0(-4)	4.2(-17)	3.3(-14)	3.3(-14)
-	2.2(10)	8	2.0(-18)	4.8(-17)	unstable	-
3	1.5(7)	11	5.1(-7)	1.5(-17)	8.7(-13)	3.3(-12)
4	6.4(8)	14	3.4(-8)	3.3(-17)	1.6(-11)	8.2(-12)
5	9.5(11)	17	9.9(-13)	4.2(-17)	5.1(-10)	8.0(-10)
6	8.9(12)	20	1.4(-14)	6.5(-17)	1.5(-9)	2.8(-9)
7	7.1(11)	23	1.0(-9)	2.4(-14)	2.4(-8)	2.8(-9)
8	1.7(13)	26	4.3(-11)	8.9(-15)	5.1(-8)	1.8(-8)
9	1.3(14)	29	5.3(-13)	7.8(-15)	1.3(-7)	3.1(-8)
10	2.3(14)	32	5.2(-13)	6.7(-15)	1.4(-6)	4.2(-7)
11	3.3(14)	35	5.9(-14)	3.6(-15)	1.7(-7)	8.3(-8)
12	1.4(14)	38	1.4(-12)	3.6(-15)	3.1(-7)	1.3(-7)
13	4.3(14)	41	1.1(-13)	3.7(-15)	3.3(-6)	2.3(-7)
14	8.2(14)	44	1.2(-15)	4.0(-15)	9.2(-6)	1.3(-6)
15	2.7(15)	47	1.6(-15)	3.1(-15)	1.9(-5)	1.5(-6)
16	1.3(15)	50	1.1(-15)	3.2(-15)	4.7(-6)	3.9(-6)
17	8.5(14)	53	3.8(-14)	3.4(-15)	5.0(-6)	9.0(-7)
18	1.9(15)	56	6.7(-15)	3.7(-15)	1.8(-5)	1.8(-6)
19	1.7(15)	59	6.4(-15)	3.5(-15)	2.6(-5)	3.5(-6)

the next stable point, at $\|m^{(i)}\| = 59$. $\hat{T}_{\nu}^{(i)}$ is a 31×31 matrix. Letting $\hat{T}_{\nu}^{(i)}$ grow large is undesirable if it can be avoided because of the $\|\nu\|^3$ complexity of solving (2.41) and (2.42) using Gaussian elimination.

Just as setting the tolerance too large is undesirable, setting it too low can also have dire consequences. From our error bounds (7.2) and (7.4), we see that if we accept a point where the condition number of $T_{m^{(i)}}$ is large, the relative error and residual error will grow proportionally. For example, Table 7.8 shows that for $k = 1$ and a type III problem, 12 digits of accuracy in the PHS are lost at the point $i = 2$. It is interesting to note that the error in the residual (the value of $\delta R^{(i)}(z)$) does not jump until a point with small $\kappa(T_{m^{(i)}})$ is accepted. This occurs at $i = 4$ in Table 7.8.

Accepting a poorly conditioned point has another negative effect. Upon accepting such a point, all future values of $\gamma^{(i)}$ tend to be larger than they would have been if the point had been skipped. This leads to acceptance of more poorly conditioned points and larger errors. Comparing the values of $\gamma^{(i)}$ in Tables 7.7 and 7.8 illustrates this observation.

To summarize then, the tolerance should be chosen such that as many points as possible are accepted and the desired accuracy in the solution is maintained. Because of the lack of correspondence between $\kappa(T_{m^{(i)}})$ and $\gamma^{(i)}$, the choice of tolerance was made a posteriori.

7.3 Normalizing by $A_0^{-1}(z)$

In this section we examine the effects of normalizing class III and IV problems by multiplying each component of $A(z)$ by $A_0^{-1}(z)$. Performing such a transformation will cause the coefficients of $A_1(z), \ldots, A_k(z)$ to increase by an amount depending largely upon the value of $A_0(0)$. If $A_0(0)$ is small relative to the other coefficients of $A_0(z)$, the growth can be dramatic.

The Tables 7.17 - 7.20 illustrate the effect of this normalization on two of the problems considered in Section 7.2. Tables 7.17 and 7.18 show the result of transforming the input power series' used to obtain the data in Tables 7.7 and 7.8. Similarly, we document the effect of normalizing $A(z)$ for the data of Tables 7.15 and 7.16 in Tables 7.19 and 7.20.

Observation 10: By multiplying a vector of power series $A(z)$ (where $A_0(z) \neq 1$) by $A_0^{-1}(z)$ and then obtaining a PHS for a given n, poorer error results are obtained than if the problem had been solved without modifying $A(z)$.

In Table 7.17, the tolerance was chosen such so that the point $m^{(i)} = (29, 30)$ would be accepted (as it was in Table 7.7). If we compare the relative error in the PHS at that point, we see that by not normalizing we retain approximately 8 more digits of accuracy. Similarly, in Table 7.19 the relative error of our iterative algorithm at the point $m^{(20)} = (19, 20, 20)$ is 2.6(-5) compared with that in Table 7.16 of 2.6(-2). Notice that, in order to accept the point (29, 30) as stable in Table 7.17, the tolerance chosen caused several poorly conditioned points (not accepted in table 7.7) to be accepted as stable.

Observation 11: The condition numbers $\kappa(T_{m^{(i)}})$ of the modified series' at a point $m^{(i)}$ tend to be larger than those for the original series at the same point.

For example, $\kappa(T_{m^{(i)}})$ for $||m^{(i)}|| = 44$ on Table 7.15 is 8.2(14) compared with the value 3.7(16) in Table 7.19. Note that this observation is not always true as in the case of the first 14 points of Table 7.19.

The reason for this growth in condition number relates back to the growth in the power series coefficients as a result of multipiying by $A_0^{-1}(z)$. Although not listed in

Table 7.17: $k = 1$, $\epsilon = 10^{-13}$, Class III

i	$\kappa\left(T_{m^{(i)}}\right)$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
-	1.0	1	1.0(-23)	6.0(-40)	unstable	-
1	1.9(7)	3	.48	1.1(-19)	1.2(-16)	1.0(-16)
2	8.5(18)	5	4.6(-12)	5.6(-17)	4.2(-6)	3.9(-6)
3	8.5(18)	7	4.6(-13)	5.6(-17)	4.2(-6)	3.9(-6)
4	2.5(7)	9	.15	4.2(-5)	8.4(-5)	1.5(-15)
5	3.9(17)	11	7.6(-11)	2.5(-5)	2.0	1.4(-14)
6	6.0(7)	13	4.0(-2)	2.5(-5)	7.1(-5)	6.8(-14)
7	1.5(8)	15	1.5(-2)	1.5(-5)	5.4(-5)	1.0(-13)
8	1.8(8)	17	4.7(-3)	6.0(-6)	7.0(-5)	1.1(-13)
9	2.2(17)	19	5.4(-10)	4.8(-6)	2.0	1.2(-13)
10	5.4(8)	21	8.4(-4)	2.0(-6)	8.4(-5)	2.4(-13)
11	4.4(17)	23	2.5(-10)	2.0(-6)	1.8(-4)	7.9(-13)
12	9.7(8)	25	2.7(-4)	1.4(-6)	1.8(-4)	1.7(-12)
13	1.5(9)	27	5.8(-5)	5.6(-7)	2.2(-4)	1.7(-12)
14	6.3(17)	29	2.6(-11)	3.4(-7)	2.0	1.5(-12)
15	4.3(9)	31	4.4(-6)	1.0(-7)	4.9(-4)	1.8(-12)
16	6.6(9)	33	1.1(-6)	5.8(-8)	6.7(-4)	1.8(-12)
17	8.2(18)	35	3.8(-10)	2.1(-8)	2.0	5.2(-5)
18	8.2(18)	37	1.7(-10)	9.4(-9)	2.0	5.0(-5)
19	2.8(10)	39	4.3(-8)	4.1(-8)	5.6(-4)	1.4(-12)
-	3.4(19)	41	6.8(-14)	1.3(-9)	unstable	-
21	3.2(10)	43	3.8(-9)	6.0(-10)	6.0(-4)	3.0(-12)
22	6.8(10)	45	5.9(-10)	1.7(-10)	6.0(-4)	2.5(-11)
-	8.2(20)	47	5.6(-14)	8.2(-11)	unstable	-
-	1.0(21)	49	2.4(-14)	7.9(-11)	unstable	-
23	2.0(11)	51	2.6(-11)	1.1(-11)	1.1(-3)	3.7(-10)
-	5.7(19)	53	1.3(-17)	4.8(-12)	unstable	-
24	2.9(11)	55	2.3(-11)	2.1(-12)	1.5(-3)	1.3(-9)
-	4.2(19)	57	8.5(-20)	5.5(-13)	unstable	-
25	1.0(12)	59	2.2(-13)	1.3(-9)	1.3(-3)	6.1(-9)

Table 7.18: $k = 1$, $\epsilon = 10^{-12}$, Class III

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
-	1.0	1	1.0(-23)	1.8(-40)	unstable	-
1	1.9(7)	3	.48	1.1(-19)	1.2(-16)	1.0(-16)
-	8.5(18)	5	4.6(-12)	5.6(-17)	unstable	-
-	8.5(18)	7	4.6(-13)	5.6(-17)	unstable	-
2	2.5(7)	9	.15	7.1(-17)	1.6(-15)	1.5(-15)
-	3.9(17)	11	8.0(-22)	5.6(-17)	unstable	-
3	6.0(7)	13	4.0(-2)	2.8(-17)	6.8(-14)	6.8(-14)
4	1.5(8)	15	1.5(-2)	6.9(-17)	1.0(-13)	1.0(-13)
5	1.8(8)	17	4.7(-3)	2.8(-17)	1.1(-13)	1.1(-13)
-	2.2(17)	19	5.6(-21)	6.5(-17)	unstable	-
6	5.4(8)	21	8.4(-4)	1.4(-17)	2.4(-13)	2.4(-13)
-	4.4(17)	23	1.3(-21)	2.3(-17)	unstable	-
7	9.7(8)	25	2.7(-4)	2.1(-17)	1.7(-12)	1.7(-12)
8	1.5(9)	27	5.8(-5)	1.0(-17)	1.7(-12)	1.7(-12)
-	6.3(17)	29	5.4(-22)	8.3(-18)	1.5(-12)	1.5(-12)
9	4.3(9)	31	4.4(-6)	6.9(-18)	1.8(-12)	1.8(-12)
10	6.6(9)	33	1.1(-6)	3.5(-18)	1.8(-12)	1.8(-12)
-	8.2(18)	35	2.1(-16)	1.7(-18)	unstable	-
11	8.2(18)	37	9.5(-17)	1.5(-18)	unstable	-
12	2.8(10)	39	4.3(-8)	1.7(-18)	1.3(-12)	1.4(-12)
-	3.4(19)	41	5.4(-27)	2.6(-18)	unstable	-
13	3.2(10)	43	3.8(-9)	1.7(-18)	4.5(-12)	3.0(-12)
14	6.8(10)	45	5.9(-10)	1.3(-18)	1.1(-11)	2.5(-11)
-	8.2(20)	47	1.4(-20)	1.1(-18)	unstable	-
-	1.0(21)	49	5.8(-21)	7.1(-19)	unstable	-
15	2.0(11)	51	2.6(-11)	1.3(-18)	3.1(-10)	3.7(-10)
-	5.7(19)	53	2.2(-29)	5.4(-19)	unstable	-
16	2.9(11)	55	2.3(-12)	2.4(-19)	7.0(-10)	1.3(-9)
-	4.2(19)	57	1.1(-28)	3.6(-19)	unstable	-
-	1.0(12)	59	2.2(-13)	4.3(-19)	unstable	-

Table 7.19: $k = 2$, $\epsilon = 10^{-24}$, Class IV, $A(z) = A_0^{-1}(z) \cdot A(z)$

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	3.7	2	6.7(-2)	2.8(-17)	8.8(-17)	8.8(-17)
2	33	5	2.5(-3)	2.8(-17)	2.1(-16)	2.2(-16)
3	1.7(7)	8	2.8(-17)	3.4(-17)	1.7(-15)	4.8(-15)
4	3.3(3)	11	2.6(-5)	1.0(-8)	1.5(-6)	3.9(-15)
5	3.1(4)	14	5.6(-7)	1.9(-9)	1.2(-5)	3.3(-14)
6	2.0(8)	17	1.7(-12)	1.3(-10)	5.1(-10)	1.5(-12)
7	1.0(9)	20	1.1(-14)	2.2(-11)	1.4(-4)	1.4(-11)
8	2.8(7)	23	6.9(-10)	5.6(-12)	1.1(-4)	2.5(-11)
9	4.9(8)	26	9.9(-12)	1.4(-12)	1.5(-4)	4.8(-11)
10	3.4(10)	29	1.9(-14)	2.3(-13)	6.1(-4)	5.5(-11)
11	6.1(11)	32	2.9(-15)	3.3(-14)	3.3(-3)	1.1(-8)
12	6.1(12)	35	4.8(-17)	4.0(-15)	5.8(-4)	6.5(-9)
13	7.5(12)	38	6.4(-16)	1.4(-15)	9.2(-4)	2.1(-8)
14	3.9(14)	41	7.7(-18)	2.8(-16)	1.6(-2)	3.6(-7)
15	3.7(16)	44	3.8(-21)	2.9(-17)	8.3(-3)	2.1(-6)
16	2.6(17)	47	1.7(-21)	4.6(-18)	1.6(-2)	3.0(-4)
17	1.4(18)	50	3.9(-22)	1.9(-18)	1.8(-2)	6.7(-5)
18	2.0(18)	53	3.6(-21)	2.7(-18)	5.6(-2)	3.7(-4)
19	2.7(19)	56	5.0(-23)	4.5(-19)	4.0(-2)	1.9(-3)
20	2.2(20)	59	5.7(-24)	9.2(-19)	2.6(-2)	8.5(-3)

Table 7.20: $k = 2$, $\epsilon = 10^{-16}$, Class IV, $A(z) = A_0^{-1}(z) \cdot A(z)$

i	$\kappa(T_{m^{(i)}})$	$\|m^{(i)}\|$	$\gamma^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	3.7	2	6.7(-2)	2.8(-17)	8.8(-17)	8.8(-17)
2	33	5	2.5(-3)	2.8(-17)	2.1(-16)	2.2(-16)
-	1.7(7)	8	2.8(-17)	3.4(-17)	unstable	-
3	3.3(3)	11	2.6(-5)	4.9(-17)	8.4(-15)	3.9(-15)
4	3.1(4)	14	5.6(-7)	1.7(-16)	9.0(-14)	3.3(-14)
5	2.0(8)	17	1.7(-12)	5.2(-18)	2.1(-12)	1.5(-12)
6	1.0(9)	20	1.1(-14)	7.4(-18)	8.0(-12)	1.4(-11)
7	2.8(7)	23	6.9(-10)	1.2(-15)	2.1(-9)	2.5(-11)
8	4.9(8)	26	9.9(-12)	1.4(-16)	2.8(-9)	4.8(-11)
9	3.4(10)	29	1.8(-14)	2.2(-17)	1.8(-8)	5.5(-11)
10	6.1(11)	32	3.0(-15)	6.4(-18)	1.3(-7)	1.1(-8)
-	6.1(12)	35	4.7(-17)	1.1(-17)	unstable	-
11	7.5(12)	38	6.4(-16)	4.4(-17)	4.1(-7)	2.1(-8)
-	3.9(14)	41	8.0(-18)	7.8(-18)	unstable	-
-	3.7(16)	44	3.8(-21)	5.4(-19)	unstable	-
-	2.6(17)	47	1.2(-21)	2.3(-19)	unstable	-
-	1.4(18)	50	3.9(-22)	4.9(-19)	unstable	-
-	2.0(18)	53	3.3(-21)	3.8(-19)	unstable	-
-	2.7(19)	56	5.2(-23)	3.3(-19)	unstable	-
-	2.2(20)	59	5.9(-24)	2.1(-19)	unstable	-

the tables, for the data of Table 7.7, $\|A(z) \bmod z^{\|m^{(16)}\|} \approx 10^2$. By multiplying this vector of power series by $A_0^{-1}(z)$ the norm of the relevant terms of $A(z)$ became $\approx 10^7$. These large coefficients cause a corresponding increase in the condition numbers of $T_{m^{(i)}}$.

7.4 An Alternate Choice of γ

The primary deficiency of our iterative algorithm is the lack of a stability parameter that accurately predicts the condition number of the matrix $T_{m^{(i)}}$. In an attempt to find such a parameter, numerous experiments were conducted using different norms for $S^{(i)}(z)$ and using different choices of $\gamma^{(i)}$. In this section we give an alternate definition of $\gamma^{(i)}$ and provide some results as to its suitability in approximating the condition number of $T_{m^{(i)}}$.

Let

$$\begin{aligned} \hat{\gamma}^{(i)} &= \gamma_0 \cdot \min_{1 \le j \le k} \{\gamma_j\} && (7.6) \\ &= R_0^{(i)}(0) \cdot \min_{1 \le j \le k} \left\{ V_{j,j}^{(i)}(0) \right\}. && (7.7) \end{aligned}$$

where γ_j, $j = 0, \ldots, k$, are consistent with (3.15). Using $\hat{\gamma}^{(i)}$ as the stability parameter, we obtained the results of Table 7.21 for $k = 7$, $A_0(z) = 1$ and Table 7.22 for $k = 2$, $A(z) = A_0^{-1}(z) \cdot A(z)$.

Observation 12: The parameter $\hat{\gamma}^{(i)}$ is a much better approximate of $\kappa(T_{m^{(i)}})$ than was $\gamma^{(i)}$.

Let us examine the effectiveness of $1/\hat{\gamma}^{(i)}$ in estimating $\kappa(T_{m^{(i)}})$ by comparing the Tables 7.3 and 7.21. Recall that in Section 7.1, we saw how the point $i = 12$ in Table 7.3, $1/\gamma^{(i)}$ provided the condition number estimate 2.0(17) to the actual con-

Table 7.21: $k = 7$, $\epsilon = 5.2 * 10^{-5}$, Class I

i	$\kappa\left(T_{m^{(i)}}\right)$	$\|m^{(i)}\|$	$\hat{\gamma}^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	85	7	1.2(-2)	8.3(-17)	2.7(-16)	2.7(-16)
2	4.5(2)	15	3.6(-4)	2.6(-16)	1.4(-15)	1.7(-15)
-	1.5(5)	23	9.4(-8)	9.7(-16)	unstable	-
3	2.4(3)	31	2.7(-4)	1.3(-14)	1.1(-12)	2.8(-15)
4	5.2(2)	39	2.7(-3)	1.3(-13)	1.1(-12)	2.9(-15)
5	7.4(2)	47	1.9(-3)	1.4(-13)	1.4(-12)	2.1(-15)
6	3.7(3)	55	9.9(-5)	1.2(-13)	2.7(-12)	8.9(-15)
7	4.0(3)	63	5.6(-4)	2.5(-13)	6.6(-12)	9.8(-15)
8	3.8(3)	71	9.3(-4)	3.3(-13)	2.4(-11)	6.1(-15)
9	6.5(3)	79	3.1(-4)	3.5(-13)	4.3(-11)	6.3(-15)
10	2.8(3)	87	9.1(-4)	4.6(-13)	1.8(-11)	5.8(-15)
11	5.7(3)	95	2.4(-4)	4.5(-13)	1.2(-11)	1.7(-14)
12	1.5(4)	103	1.1(-4)	5.8(-13)	1.9(-11)	1.2(-14)
-	1.1(6)	111	2.8(-9)	3.2(-13)	unstable	-
-	1.8(4)	119	3.9(-5)	2.6(-13)	unstable	-
13	1.8(4)	127	2.9(-5)	5.3(-13)	unstable	-
14	2.5(4)	135	7.6(-5)	7.2(-13)	1.4(-11)	1.9(-14)
15	1.2(4)	143	7.2(-5)	7.7(-13)	4.5(-11)	1.3(-14)
-	1.4(4)	151	2.6(-5)	4.4(-13)	unstable	-
16	1.4(4)	159	2.0(-4)	7.1(-13)	2.1(-11)	1.5(-14)
-	7.7(4)	167	5.1(-6)	6.4(-13)	unstable	-
17	9.2(3)	175	2.0(-4)	5.5(-13)	1.6(-11)	1.0(-14)
-	5.4(5)	183	6.2(-8)	3.2(-13)	unstable	-
18	2.3(4)	191	5.3(-4)	5.3(-13)	1.1(-11)	1.8(-14)
-	7.1(4)	199	8.9(-6)	5.2(-13)	unstable	-
-	5.0(4)	207	1.3(-5)	4.3(-13)	unstable	-
-	4.1(4)	215	1.1(-5)	4.5(-13)	unstable	-
-	7.7(4)	223	3.7(-6)	3.6(-13)	unstable	-
19	3.4(4)	231	5.2(-5)	4.6(-13)	1.6(-11)	2.0(-14)
-	4.5(4)	239	2.6(-5)	6.2(-13)	unstable	-

Table 7.22: $k = 2$, $\epsilon = 10^{-13}$, Class IV, $A(z) = A_0^{-1}(z) \cdot A(z)$

i	$\kappa\left(T_{m^{(i)}}\right)$	$\|m^{(i)}\|$	$\hat{\gamma}^{(i)}$	$\|\delta R^{(i)}(z)\|$	$\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$	$\frac{\|\delta S_G^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$
1	3.7	2	.15	2.8(-17)	8.8(-17)	8.8(-17)
2	33	5	1.5(-2)	2.8(-17)	2.1(-16)	2.2(-16)
3	1.7(7)	8	1.2(-12)	3.4(-17)	1.7(-15)	4.8(-15)
4	3.3(3)	11	2.0(-4)	1.0(-8)	1.5(-6)	3.9(-15)
5	3.1(4)	14	9.4(-6)	1.9(-9)	1.2(-5)	3.3(-14)
6	2.0(8)	17	3.2(-10)	1.3(-10)	1.1(-4)	1.5(-12)
7	1.0(9)	20	2.4(-11)	2.2(-11)	1.4(-4)	1.4(-11)
8	2.8(7)	23	2.0(-8)	5.6(-12)	1.1(-4)	2.5(-11)
9	4.9(8)	26	8.8(-10)	1.4(-12)	1.5(-4)	4.8(-11)
10	3.4(10)	29	7.3(-12)	2.3(-13)	6.1(-4)	5.5(-11)
11	6.1(11)	32	3.9(-13)	3.3(-14)	3.3(-3)	1.1(-8)
-	6.1(12)	35	4.6(-14)	4.0(-15)	unstable	-
12	7.5(12)	38	1.8(-13)	1.4(-15)	9.3(-4)	2.1(-8)
-	3.9(14)	41	2.7(-15)	2.8(-16)	unstable	-
-	3.7(16)	44	1.1(-17)	2.9(-17)	unstable	-
-	2.6(17)	47	2.1(-18)	4.6(-18)	unstable	-
-	1.4(18)	50	1.7(-18)	9.5(-19)	unstable	-
-	2.0(18)	53	7.0(-18)	4.6(-19)	unstable	-
-	2.7(19)	56	2.0(-18)	3.8(-19)	unstable	-
-	2.2(20)	59	2.0(-20)	3.5(-19)	unstable	-

dition number $\kappa(T_{m^{(12)}}) = 1.5(4)$. Our new parameter $1/\hat{\gamma}^{(i)}$ estimates the condition number to be 9.0(3).

Further evidence of the suitability of $\hat{\gamma}^{(i)}$ over $\gamma^{(i)}$ can be seen by comparing Tables 7.19 and 7.22. From Table 7.19, $\gamma^{(i)}$ predicts the condition number of $T_{m^{(i)}}$ at the point $\|m^{(i)}\| = 38$ to be 1.6(15) whereas the estimate given by $\hat{\gamma}^{(i)}$ is 5.6(12). The actual value is $\kappa(T_{m^{(i)}}) = 7.5(12)$.

Note that the parameter $\hat{\gamma}^{(i)}$ only provides a good estimate of $\kappa(T_{m^{(i)}})$ if $A_0(z) = 1$. If necessary, we must normalize $A(z)$ by multiplying by $A_0^{-1}(z)$. As we have seen, this transformation often results in the power series' $A_1(z), \ldots, A_k(z)$ becoming extremely divergent.

Chapter 8

Conclusions

As more applications are developed which utilize Padé-Hermite approximants, the need for a fast (or superfast) and numerically stable method of computing them will become necessary. For large problems, exact methods using systems such as Maple or Mathematica are not practical due to their slow performance. The only alternative is to consider numerical methods; and the question now becomes one of stability. An obvious choice is the Gaussian elimination method which is known to be numerically stable. But the cost complexity of the Gaussian elimination method for the Padé-Hermite problem is $\mathcal{O}(\|n\|^3)$, and we seek a faster numerically stable algorithm. The work in this thesis has brought the reality of such an algorithm one step closer.

We have provided experimental evidence that the algebraic algorithm of Cabay et al. [10] for computing Padé-Hermite Systems can be adapted to function in a numerical setting with the introduction of a stability parameter γ. Cabay et al.'s algorithm iteratively computes PHS's at points along a diagonal path in the Padé-Hermite table where the block Sylvester matrix $T_{n(\sigma)}$ is nonsingular. The stability parameter γ predicts the condition number of the block Sylvester matrices allowing iterative computation of Padé-Hermite Systems at points along the diagonal path where $\kappa(T_{n(\sigma)})$ is within a specified tolerance. The choice of γ was suggested by the inverse formula for a block Hankel matrix.

To test the efficiency of our algorithm, a method was devised for generating a vector of power series containing singular points at predetermined locations in the Padé-Hermite table. This method provided a means of generating problems with varying instabilities. It also provides a framework upon which future numerical methods can be compared.

Experimental evidence supported the choice of γ when $A_0(z) = 1,\ k = 1$ and the matrices $T_{n(\sigma)}$ were reasonably well conditioned. As k was increased, the effectiveness of γ in estimating the condition number faltered. When tested with power series for which $A_0(z) \neq 1$, γ again failed to provide a good estimate of $\kappa(T_{n(\sigma)})$, even when $k = 1$.

We found that the relative error $\frac{\|\delta S^{(i)}(z)\|}{\|S_E^{(i)}(z)\|}$ and error in the residual $\delta R^{(i)}(z)$ agreed for $k = 1$ with the error bounds of Cabay and Meleshko [10]. Based on the experimental findings, operational bounds for the two error terms were supplied.

We compared the relative error of the PHS obtained iteratively, with that arising from solving the $\|n^{(\sigma)}\| \times \|n^{(\sigma)}\|$ block Sylvester matrix $T_{n(\sigma)}$ directly by the Gaussian elimination method. The error of the two methods was found to be comparable (provided the tolerance ϵ was chosen carefully).

We also examined the effect of normalizing a vector of power series $A(z)$ by multiplying each element by $A_0^{-1}(z)$ and using this as the input to our problem. Modifying the input power series in this way resulted in a growth in the power series coefficients of $A_i(z),\ i = 1, \ldots, k$. Because the power series coefficients grew, so did the condition numbers of $T_{n(\sigma)}$ which led to increased relative error in computing the PHS's.

Having studied the performance of the algorithm, we proposed a new stability parameter $\hat{\gamma}$. This parameter was found to provide a much more accurate estimate of $\kappa(T_{n(\sigma)})$ when $A_0(z) = 1$. This perhaps is the single most important contribution of this thesis.

8.1 Future Research

Since the stability parameter is an integral aspect of our algorithm, we would like to prove that the choice of $\hat{\gamma}$ is correct. Experimental evidence would suggest that a closed form inverse for block Sylvester matrices, similar to that of Theorem (3.6), may be found which involves $\hat{\gamma}$. The existence of such a theorem would enable the proof of weak stability of our numerical algorithm. As a by-product, a new set of numerical error bounds for $\delta R(z)$ and $\frac{\|\delta S(z)\|}{\|S_E(z)\|}$ could be derived by taking the approach of Cabay and Meleshko [11].

By scaling the residual $R^{(i)}(z)$ we are imposing a column scaling on the systems involving $\hat{T}_{\nu}^{(i)}$. If the linear systems (2.41) and (2.42) were solved using a row-column equilibration as described by Golub and Van Loan [14], the error in the solutions may be reduced further. This could be particularly effective for power series vectors in which $A_0(z) = 1$. However, as Golub and Van Loan emphasize, this method may render a worse solution than using no scaling at all. Scaling is very problem dependent, and identifying the most appropriate scaling to be used in computing Padé-Hermite Systems is left as an open problem.

Padé-Hermite approximants are themselves a special case of a more general rational interpolation problem. M-Padé approximants are a generalization of the Padé-Hermite problem by requiring that the residual $R(z)$ have specific zeros. The sequence

$\{z_i\}$ $i = 0, \ldots, \|n\|$ of (not necessarily distinct) complex numbers are called knots and represent roots of the residual. That is, using the notation of Chapter 2,

$$B(z) \cdot P(z) + C(z)Q(z) = (z - z_0) \cdots (z - z_{\|n\|})R(z). \tag{8.1}$$

When these knots are all identically zero, the Padé-Hermite approximant problem results. Beckerman [5] gives a good explanation of the M-Padé approximation problem as well as providing a reliable method for computing them. We believe that much of the results we have presented can be generalized to the rational interpolation problem.

We have evidence to suggest that Padé-Hermite Systems can be used to compute the greatest common divisor (gcd) of a set of polynomials over the field of integers. The motivation for this intuition is the following. We have

$$A(z) \cdot S(z) = z^{\|n\|+1}\, R(z)$$

and from Lemma 6.1

$$A(z) = R(z) \cdot S^{adj}(z).$$

Thus,

$$G \mid \{A_0(z), \ldots, A_k(z)\} \Rightarrow G \mid \{R_0(z), \ldots, R_k(z)\}, \tag{8.2}$$

and

$$G \mid \{R_0(z), \ldots, R_k(z)\} \Rightarrow G \mid \{A_0(z), \ldots, A_k(z)\}. \tag{8.3}$$

If we consider $A(z)$ and $R(z)$ to be vectors of polynomials, then $\gcd\{A_0(z), \ldots, A_k(z)\} = \gcd\{R_0(z), \ldots, R_k(z)\}$. The guiding premise for this proposed algorithm is that for nontrivial n, the problem of finding $\gcd\{R_0(z), \ldots, R_k(z)\}$ is simpler than finding $\gcd\{A_0(z), \ldots, A_k(z)\}$ because the polynomials are of lower degree.

The results of our fast numerical algorithm may prove fruitful in developing a superfast $\mathcal{O}(\|n\| \cdot \log^2 \|n\|)$ numerical algorithm for computing Padé-Hermite and simultaneous Padé approximants. The incorporation of a stability parameter γ into the algorithm of Cabay and Labahn [9] is needed to establish if Padé-Hermite or simultaneous Padé table points along a diagonal path are stable or unstable. A crucial result in formulating a superfast algorithm involves the recurrence relationship relating consecutive points along a table diagonal. Ignoring the various scaling matrices we have in the Padé-Hermite case, the relationship $S^{(i+1)}(z) = S^{(i)}(z) \cdot \hat{S}(z)$, where now for the superfast algorithm the degrees of $\hat{S}(z)$ are about as large as the degrees of $S^{(i)}(z)$. We would like to be able to predict the value of $\gamma^{(i+1)}$, based on the known values $\gamma^{(i)}$ and the value of γ for $\hat{S}(z)$. Because of the recursive nature of this algorithm, detection of an unstable point after multiplication may require expensive calculations to be repeated. Such occurrences may drastically affect the cost complexity of the algorithm.

An important issue in the algebraic computation of Padé-Hermite approximants is the manner in which singular points are handled. Van Barel and Bultheel [28] present a fast algebraic algorithm that takes a different approach than we have taken. Rather than computing Padé-Hermite Systems at nonsingular points along the diagonal path, their algorithm iteratively computes a set of auxiliary vectors at all points (singular and nonsingular) along the path. They show that a basis for all Padé-Hermite forms (see Section 2.3) at a given point on the diagonal path can be obtained from the auxiliary vectors computed at that point. If the point is nonsingular, the basis is unique (except for multiplication by some element of the field $\mathcal{F}$). To show that the method of Van Barel and Bultheel is stable, what must be established is that the construction of the auxiliary set at a point, using the auxiliary set at the previous point, is a stable process. Although we have no evidence to support such a claim, we suspect that it is not stable. The affirmation of this and our previous claims, is left to future research.

Bibliography

[1] G. S. Ammar and W. B. Gragg. Superfast solution of real positive definite Toeplitz systems. *SIAM Journal on Matrix Analysis and Applications*, 9:61–76, 1988.

[2] A. C. Antoulas. On Recursiveness and Related Topics in Linear Systems. *IEEE Transactions on Automatic Control*, 31:1121–1135, 1986.

[3] G. A. Baker. *Essentials of Padé Approximants.* Academic Press, New York, 1975.

[4] G. A. Baker and P. R. Graves-Morris. *Padé Approximants Part II: Extensions and Applications.* Addison Wesley, Reading, 1981.

[5] B. Beckerman. A reliable method for computing M-Padé approximants on arbitrary staircases. *Journal of Computational and Applied Mathematics*, to appear.

[6] B. Beckerman and G. Labahn. A uniform approach for the fast, reliable computation of Matrix-type Padé approximants. preprint, January 1992.

[7] J. R. Bunch. The weak and strong stability of algorithms in numerical linear algebra. *Linear Algebra and its Applications*, 88/89:49–66, 1987.

[8] S. Cabay and D. K. Choi. Algebraic Computation of Scaled Padé Fractions. *SIAM Journal on Computing*, 15:243–270, 1986.

[9] S. Cabay and G. Labahn. A Superfast Algorithm For Multi-Dimensional Padé Systems. *Numerical Algorithms*, to appear.

[10] S. Cabay, G. Labahn, and B. Beckerman. On the Theory and Computation of Non-perfect Padé-Hermite Approximants. *Journal of Computional and Applied Mathematics*, 39:295–313, 1992.

[11] S. Cabay and R. Meleshko. A Weakly Stable Algorithm for Padé Approximants and the Inversion of Hankel Matrices. *SIAM Journal on Matrix Analysis*, to appear.

[12] J. Coates. On the Algebraic Approximation of Functions I-III. *Indagationes Mathematicae*, 28:421–461, 1966.

[13] J. Della Dora and C. Dicrescenzo. Approximants de Padé-Hermite. *Numerische Mathematik*, 43:23–57, 1984.

[14] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Johns Hopkins University Press, Baltimore, Maryland, second edition, 1990.

[15] P. R. Graves-Morris and J. M. Wilkins. A Fast Algorithm to Solve Kalman's Partial Realization Problem. In A. Dold and B. Eckmann, editors, *Approximation Theory*, pages 1–20. Springer-Verlag, 1987.

[16] M. Ch. Hermite. Sur l'Expression U sinx + V cosx + W, Extrait d'une lettre a M. Paul Gordan. *Oeuvres Tome III*, pages 135–145, 1873.

[17] M. Ch. Hermite. Sur Quelques Approximations Algebriques, Extrait d'une lettre a M. Borchardt. *Oeuvres Tome III*, pages 146–149, 1873.

[18] H. Jager. A multidimensional generalization of the Padé table I-VI. *Indagationes Mathematicae*, 26:192–249, 1964.

[19] T. Kailath. *Linear Systems.* Prentice-Hall, 1980.

[20] G. Labahn. Inverse of block Hankel matrices using Padé-Hermite Approximants. unpublished material.

[21] K. Mahler. Zur Approximation der Exponontialfunktion und des Logarithmus. *Teil I, Journ. f.d.r.u.a*, 166:118–136, 1932.

[22] R. Meleshko and S. Cabay. On Computing Padé Approximants Quickly and Accurately. *Congressus Numerantium*, 80:245–255, 1991.

[23] H. Padé. *Sur la Representation Approchee d'une Fonction par des Fractions Rationelles.* PhD thesis, Ann. Ecole Nor., Paris, 1892.

[24] S. Paszkowski. Recurrence Relations in Padé-Hermite Approximation. *Journal of Computational and Applied Mathematics*, 19:99–107, 1987.

[25] S. Paszkowski. Hermite Padé approximation: basic notions and theorems. *Journal of Computational and Applied Mathematics*, 32:229–236, 1990.

[26] R. E. Shafer. On Quadratic Approximation. *SIAM Journal on Numerical Analysis*, 11:447–460, 1974.

[27] M. Van Barel and A. Bultheel. An Algebraic Method to Solve the Minimal Partial Realization Problem for Scalar Sequences. *Linear Algebra and its Applications*, 104:117–129, 1988.

[28] M. Van Barel and A. Bultheel. The computation of non-perfect Padé-Hermite approximants. preprint, 1991.