Computation of Numerical Padé-Hermite and Simultaneous Padé Systems II: A Weakly-Stable Algorithm

Stan Cabay*, Anthony R. Jones†, and George Labahn‡

Abstract. For $k+1$ power series $a_0(z), \ldots, a_k(z)$, we present a new iterative, look-ahead algorithm for numerically computing Padé-Hermite systems and simultaneous Padé systems along a diagonal of the associated Padé tables. The algorithm computes the systems at all those points along the diagonal at which the associated striped Sylvester and mosaic Sylvester matrices are well-conditioned. The operation and the stability of the algorithm is controlled by a single parameter τ which serves as a threshold in deciding if the Sylvester matrices at a point are sufficiently well-conditioned. We show that the algorithm is weakly stable, and provide bounds for the error in the computed solutions as a function of τ. Experimental results are given which show that the bounds reflect the actual behavior of the error.

The algorithm requires $\mathcal{O}(\|n\|^2 + s^3\|n\|)$ operations, to compute Padé-Hermite and simultaneous Padé systems of type $n = [n_0, \ldots, n_k]$, where $\|n\| = n_0 + \cdots + n_k$ and s is the largest step-size taken along the diagonal. An additional application of the algorithm is the stable inversion of striped and mosaic Sylvester matrices.

Key Words. Padé-Hermite approximants, simultaneous Padé approximants, striped Sylvester inverses, mosaic Sylvester inverses, numerical algorithm, numerical stability

AMS(MOS) Subject Classification. 41A21, 65F05, 65G05

1. Introduction.

Let $A^t(z) = [a_0(z), \ldots, a_k(z)]$, $k \geq 1$, be a vector of formal power series over the real numbers[1] with $a_0(0) \neq 0$ and let $n = [n_0, \ldots, n_k]$ be a vector of integers with $n_\beta \geq -1, 0 \leq \beta \leq k$, and with at least one $n_\beta \geq 0$. A *Padé-Hermite approximant* of type n for $A(z)$ is a nontrivial vector $[q_0(z), \ldots, q_k(z)]$ of polynomials $q_\beta(z)$ over the real numbers having degrees[2] at most $n_\beta, 0 \leq \beta \leq k$, such that

$$a_0(z)q_0(z) + \cdots + a_k(z)q_k(z) = c_{\|n\|+k}z^{\|n\|+k} + c_{\|n\|+k+1}z^{\|n\|+k+1} + \cdots, \tag{1}$$

with $\|n\| = n_0 + \ldots + n_k$.

The *Padé-Hermite approximation problem* was introduced in 1873 by Hermite and has been widely studied by several authors (for a bibliography, see, for example [27, 2, 4, 5, 23]). Note that for $A^t(z) = [-1, a(z)]$, (1) becomes

$$a(z)q_1(z) - q_0(z) = O(z^{n_0+n_1+1}).$$

Thus, as a special case we have the classical Padé approximation problem for the power series $a(z)$. The Padé-Hermite approximation problem also includes other

* Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada, T6G 2H1. The research of this author as partially supported by Natural Sciences and Engineering Research Council of Canada grant A8035.

† Bell Northern Research, P.O. Box 3511, Station C, Ottawa, Ontario, Canada, K1Y 4H7

‡ Department of Computer Science, University of Waterloo, Waterloo, Ontario, Canada, N2L3G1. The research of this author was partially supported by Natural Sciences and Engineering Research Council of Canada grant FS1525C.

[1] The restriction to real numbers is made in order to simplify floating-point analysis. All of the results given in this paper also hold with minor modifications for the field of complex numbers.

[2] By convention, a polynomial of degree -1 is the zero polynomial.

classical approximation problems such as the algebraic approximants where $A^t(z) = [1, a(z), a(z)^2, \ldots, a(z)^k]$ (see [25] for the special case $k = 2$) and G^3J approximants where $A^t(z) = [1, a(z), a'(z)]$. Additional examples can be found in [1].

Closely related to Padé-Hermite approximants are *simultaneous Padé approximants*. A simultaneous Padé approximant of type n for $A(z)$ is a nontrivial vector $[q_0^*(z), \ldots, q_k^*(z)]$ of polynomials $q_\beta^*(z)$ over the real numbers having degrees of at most $\|n\| - n_\beta, 0 \leq \beta \leq k$, such that

$$q_0^*(z) \cdot a_\beta(z) + q_\beta^*(z) \cdot a_0(z) = c_{\|n\|+1}^{(\beta)} z^{\|n\|+1} + c_{\|n\|+2}^{(\beta)} z^{\|n\|+2} + \cdots, \tag{2}$$

for $\beta = 1, \cdots, k$. Simultaneous Padé approximants were also defined by Hermite and were used in his famous proof of the transcendence of e. Again, for $A^t(z) = [-1, a(z)]$, the simultaneous Padé approximation problem becomes the classical Padé approximation problem for $a(z)$.

By equating coefficients in (1), the Padé-Hermite approximation problem can be viewed as solving a system of linear equations of size $\|n\| \times \|n\|$. Thus, one can use Gaussian elimination to solve this problem with a complexity of $\mathcal{O}(\|n\|^3)$ operations. However, the coefficient matrix of the corresponding linear system has a structured form so it is not surprising that there are a number of fast [27, 14] $\mathcal{O}(\|n\|^2)$ and superfast [5, 12] $\mathcal{O}(\|n\| \log^2 \|n\|)$ algorithms for determining Padé-Hermite approximants. All these algorithms have the property that they work for any input vector of power series. In addition, these algorithms all make important use of exact arithmetic; in particular, they all depend on knowing that certain quantities are known to be 0 or not. A similar statement also applies for the fast and superfast computation of simultaneous Padé approximants.

In the special case of Padé approximants it has long been known that most fast and superfast algorithms for their computation have problems with numerical stability. The first known numerically stable algorithm for fast Padé approximation was presented by Cabay and Meleshko [15]. Alternate algorithms for fast Padé computation that also consider the issue of numerical stability include [6], [13] and [18], and for superfast computation [19]. Algorithms dealing with the closely associated problem of stably computing fast rational interpolation include [8].

In this paper, we present a new algorithm for the computation of Padé-Hermite and simultaneous Padé *systems*. These systems are matrix polynomials which contain the desired multi-dimensional Padé approximant along with quantities that can be used to recursively or iteratively compute the next approximant along a diagonal path in the associated Padé tables. The algorithm works for all vectors of power series and is fast in the sense that it computes a system in $\mathcal{O}(\|n\|^2)$ operations in the generic case. In addition, we show that this algorithm is *weakly stable* in the sense that it provides good answers to well-conditioned problems. The algorithm is a look-ahead procedure that computes the systems of type n by computing all the Padé systems at the well-conditioned locations along the diagonal path passing through the point n. In the case of Padé approximation ($k = 1$), the algorithm reduces to the Cabay and Meleshko algorithm.

It is known (cf. [12] or [23]) that in exact arithmetic a Padé-Hermite system exists uniquely if and only if the striped Sylvester coefficient matrix of the corresponding associated linear system is nonsingular. This is also true for simultaneous Padé systems where the coefficient matrix of the associated linear system is now a mosaic Sylvester rather than a striped Sylvester matrix. However, in the case of floating-point arithmetic determining that such coefficient matrices are nonsingular is not good enough.

Instead one must know, at least in a reasonably computable way that the linear systems are also well-conditioned. Central to the stable operation of our algorithm is the ability to estimate the condition numbers of the associated striped Sylvester and mosaic Sylvester matrices. The estimates follow from some "near" inverse formulae for these matrices that are derived in the companion paper [11] and which are expressed in terms of both Padé-Hermite and simultaneous Padé systems. This is the reason why our algorithm computes Padé-Hermite and simultaneous Padé systems in tandem; the inverse formulae, and consequently the estimates for the condition numbers, require that both the Padé-Hermite and the simultaneous Padé systems be available. The striped Sylvester and mosaic Sylvester matrices are deemed to be well-conditioned if the computed estimates of the condition numbers are bounded by some specified "stability" tolerance τ.

As a corollary to the results [11], there is a formula which gives the inverse of a striped Sylvester matrix expressed in terms of the associated Padé-Hermite system only. One attempt to use this formula to develop a stable algorithm for computing Padé-Hermite systems (independent of simultaneous Padé systems) was only partly successful [22]; bounds for the inverse of the associated striped Sylvester matrix (and consequently bounds for its condition number) using the formula were often too pessimistic and impractical.

This paper is organized as follows. Preliminary definitions and basic facts about Padé-Hermite and simultaneous Padé systems are given in the next two sections, and the algorithm for computing these systems is given §4. The remainder of the paper is devoted to showing that the algorithm is weakly stable for the computation of either system. To this end, §5 discusses the errors that result from the iterative steps of the algorithm, while §6 gives the proof of stability. §7 provides results of some numerical experiments that reflect the theoretic results of the previous sections. The final section gives some conclusions and a discussion of further areas of research.

We conclude this section by defining some norms which are used in the analysis of the errors made by the algorithm. Let

$$a(z) = \sum_{\ell=0}^{\infty} a^{(\ell)}\, z^{\ell} \in \mathcal{R}\,[[z]]\,,$$

where $\mathcal{R}[[z]]$ is the domain of power series with coefficients from $\mathcal{R}$, and define the bounded power series

$$\mathcal{R}^{B}[[z]] = \left\{ a(z) \;\middle|\; a(z) \in \mathcal{R}\,[[z]]\,,\ \sum_{\ell=0}^{\infty} |a^{(\ell)}| < \infty \right\}.$$

A norm for $a(z) \in \mathcal{R}^{B}\,[[z]]$ is

$$\|a(z)\| = \sum_{\ell=0}^{\infty} |a^{(\ell)}|.$$

$\mathcal{R}^{B}[[z]]$ includes the domain of polynomials $\mathcal{R}[z]$. So, for

$$s(z) = \sum_{\ell=0}^{\partial} s^{(\ell)}\, z^{\ell} \ \in \mathcal{R}\,[z]\,,$$

we use the norm

$$\|s(z)\| = \sum_{\ell=0}^{\partial} \mid s^{(\ell)} \mid .$$

For vectors and matrices over $\mathcal{R}^B[[z]]$, we use the 1-norm unless otherwise specified. So, for example, the norm for $A^t(z)$ is

$$\|A^t(z)\| = \max_{0\leq\beta\leq k} \{\|a_\beta(z)\|\}$$

and the norm for $S(z) \in \mathcal{R}_{(k+1)\times(k+1)}[z]$ is

$$\|S(z)\| = \max_{0\leq\beta\leq k} \left\{ \sum_{\alpha=0}^{k} \|S_{\alpha,\beta}(z)\| \right\}.$$

It is easy to verify that various compatibility conditions are satisfied. For example,

$$\|A^t(z) \cdot S(z)\| \leq \|A^t(z)\| \cdot \|S(z)\|$$

and

$$\|a(z) \cdot b(z)\| \ \leq \ \|a(z)\| \cdot \|b(z)\|,$$

where $b(z)$ is also a bounded power series. In addition, for $S^*(z) \in \mathcal{R}_{(k+1)\times(k+1)}[z]$ and $A^*(z) \in \mathcal{R}^B_{(k+1)\times k}[[z]]$,

$$\|S^*(z) \cdot A^*(z)\| \leq \|S^*(z)\| \cdot \|A^*(z)\|,$$

$$\|S(z) \cdot S^*(z)\| \ \leq \ \|S(z)\| \cdot \|S^*(z)\|.$$

In the subsequent development, we also make use of the inequality

$$\|a(z) \ (mod \ z^{\|n\|+1})\| \ \leq \ \|a(z)\|,$$

where

$$a(z) \ (mod \ z^{\|n\|+1}) = \sum_{\ell=0}^{\|n\|} a^{(\ell)} \, z^\ell + \sum_{\ell=\|n\|+1}^{\infty} 0 \cdot z^\ell \in \mathcal{R}^B[[z]]$$

2. Padé-Hermite Systems. In this section, we give the definition of a Padé-Hermite system for a vector of formal power series. Let $n = [n_0, \ldots, n_k]$ and define $\|n\| = n_0 + \cdots + n_k$. Let

$$A^t(z) = [a_0(z), \ldots, a_k(z)] \,, \tag{3}$$

where

$$a_\beta(z) = \sum_{\ell=0}^{\infty} a_\beta^{(\ell)} \, z^\ell, \qquad \beta = 0, \ldots, k,$$

with $a_\beta^{(\ell)} \in \mathcal{R}$, the field of real numbers. Assume that $a_0^{(0)} \neq 0$, which means that $a_0^{-1}(z)$ exists. Assume also that $A^t(z)$ is scaled so that $\|a_\beta(z)\,(mod\ z^{\|n\|+1})\| = 1$, $0 \leq \beta \leq k$.

The $(k+1) \times (k+1)$ matrix of polynomials

$$S(z) = \left[\begin{array}{c|c} z^2p(z) & U^t(z) \\ \hline z^2Q(z) & V(z) \end{array}\right] = \left[\begin{array}{c|ccc} z^2p(z) & u_1(z) & \cdots & u_k(z) \\ \hline z^2q_1(z) & v_{1,1}(z) & \cdots & v_{1,k}(z) \\ \vdots & \vdots & & \vdots \\ z^2q_k(z) & v_{k,1}(z) & \cdots & v_{k,k}(z) \end{array}\right] \tag{4}$$

is a Padé-Hermite system (PHS) [14] of type n for $A(z)$ if the following conditions are satisfied.

I. **(Degree conditions)**: For $1 \leq \alpha, \beta \leq k$,

$$\begin{aligned} p(z) &= \sum_{\ell=0}^{n_0-1} p^{(\ell)} z^\ell, & u_\beta(z) &= \sum_{\ell=0}^{n_0} u_\beta^{(\ell)} z^\ell, \\ q_\alpha(z) &= \sum_{\ell=0}^{n_\alpha-1} q_\alpha^{(\ell)} z^\ell, & v_{\alpha,\beta}(z) &= \sum_{\ell=0}^{n_\alpha} v_{\alpha,\beta}^{(\ell)} z^\ell. \end{aligned} \tag{5}$$

II. **(Order condition)**:

$$A^t(z)S(z) = z^{\|n\|+1}T^t(z), \tag{6}$$

where $T^t(z) = [r(z), W^t(z)]$ with $W^t(z) = [w_1(z), \ldots, w_k(z)]$ is the residual.

III. **(Nonsingularity condition)**: The constant term of $V(z)$ is a diagonal matrix,

$$V(0) = diag\ [\gamma_1, \ldots, \gamma_k]\,, \tag{7}$$

and

$$\gamma \equiv (a_0^{(0)})^{-1} \prod_{\alpha=0}^{k} \gamma_\alpha\ \neq 0, \tag{8}$$

where $\gamma_0 = r(0)$.

Remark 1: Only the first column of $S(z)$ is a Padé-Hermite approximant as defined in §1; this being of type $[n_0 - 1, \ldots, n_k - 1]$. The remaining columns $S(z)$ do not quite satisfy the order condition (1) and are therefore not Padé-Hermite approximants; these columns serve primarily to facilitate the computation of the first column using the algorithm given later in §4. But there are other uses for these columns of $S(z)$, such as that of expressing the inverse of a striped Sylvester matrix [9, 11].

Remark 2: The nonsingularity condition III is equivalent to the condition that $r(0) \neq 0$ and that $V(0)$ be a nonsingular diagonal matrix.

Remark 3: The PHS is said to be **normalized** [14] if the nonsingularity condition III is replaced by $r(0) = 1$ and $V(0) = I_k$. This can be achieved by multiplying $S(z)$ on the right by Γ^{-1}, where

$$\Gamma = diag\,[\gamma_0, \ldots, \gamma_k]\,. \tag{9}$$

The PHS is said to be **scaled** [22] if each column of $S(z)$ has norm equal to 1 for some norm and if, in addition, $\gamma_\beta > 0$, $0 \le \beta \le k$. Here, also, scaling a PHS is accomplished by multiplying it on the right by an appropriate diagonal matrix.

Remark 4: The nonsingularity condition III, namely $\gamma \neq 0$, refers to the nonsingularity of $S(z)$; that is, $S(z)$ is nonsingular iff $\gamma \neq 0$ (in [9], it is shown that $\det S(z) = \gamma \cdot z^{\|n\|+1}$). Equivalently, the nonsingularity condition refers to the nonsingularity of the associated striped Sylvester matrix $\mathcal{M}_n$ defined in (11) below; in [14] it is shown that a PHS (with $\gamma \neq 0$) exists iff $\mathcal{M}_n$ is nonsingular.

If the order condition (6) is not satisfied exactly, but rather

$$A^t(z)S(z) = z^{\|n\|+1}T^t(z) + \delta T^t(z), \tag{10}$$

where $\delta T^t(z) = \left[z^2\ \delta r(z), \delta W^t(z)\right]$ with $\delta W^t(z) = [\delta w_1(z), \ldots, \delta w_k(z)]$ is a relatively "small" residual error, then $S(z)$ is called a numerical Padé-Hermite system (NPHS). In (10), for $1 \le \beta \le k$,

$$\begin{aligned} \delta r(z) &= \sum_{\ell=0}^{\|n\|-2} \delta r^{(\ell)}\ z^\ell, \\ \delta w_\beta(z) &= \sum_{\ell=0}^{\|n\|} \delta w_\beta^{(\ell)}\ z^\ell. \end{aligned}$$

If $\delta T^t(z) = 0$, then $S(z)$ is an exact (rather than a numerical) Padé-Hermite system. To distinguish it from a NPHS $S(z)$, an exact system is denoted by $S_E(z)$.

Associated with $A(z)$, let $\mathcal{M}_n$ be the striped Sylvester matrix of order $\|n\|$,

$$\mathcal{M}_n = \left[\begin{array}{ccc|c|ccc} a_0^{(0)} & & & & a_k^{(0)} & & \\ & \ddots & & & & \ddots & \\ \vdots & & a_0^{(0)} & \cdots & \vdots & & a_k^{(0)} \\ & & \vdots & & & & \vdots \\ a_0^{(\|n\|-1)} & \cdots & a_0^{(\|n\|-n_0)} & & a_k^{(\|n\|-1)} & \cdots & a_k^{(\|n\|-n_k)} \end{array}\right]. \tag{11}$$

Then $S(z)$ can be obtained by solving two sets of linear equations with $\mathcal{M}_n$ as the coefficient matrix [14]. ¿From (10),

$$a_0(z)\ p(z) + \sum_{\alpha=1}^{k} a_\alpha(z)\ q_\alpha(z) = z^{\|n\|-1}r(z) + \delta r(z), \tag{12}$$

which gives rise to

$$\mathcal{M}_n \cdot \mathcal{X} = [0, \ldots, 0, \gamma_0]^t, \tag{13}$$

where

$$\mathcal{X} = \left[p^{(0)}, \ldots, p^{(n_0-1)} | q_1^{(0)}, \ldots, q_1^{(n_1-1)} | \cdots | q_k^{(0)}, \ldots, q_k^{(n_k-1)}\right]^t.$$

The solution $\mathcal{X}$ yields the first column $S_{0,0}(z), S_{1,0}(z), \ldots, S_{k,0}(z)$ of $S(z)$. In (13), we require that $\gamma_0 = r(0) \neq 0$; $\gamma_0 = 1$ for a normalized NPHS. The existence of a solution to (13) is assured if $\mathcal{M}_n$ is nonsingular. The term $\delta r(z)$ in (12) represents the residual error made in solving (13)

Next, to compute $U^t(z)$ and $V(z)$ (i.e., the remaining columns of $S(z)$), again we use (6), namely,

$$a_0(z)\, u_\beta(z) + \sum_{\alpha=1}^{k} a_\alpha(z)\, v_{\alpha,\beta}(z) = z^{\|n\|+1} w_\beta(z) + \delta w_\beta(z),\ 1 \leq \beta \leq k. \tag{14}$$

For $\alpha, \beta = 1, \ldots, k$, set

$$\begin{aligned} u_\beta^{(0)} &= -\frac{a_\beta^{(0)}}{a_0^{(0)}}\gamma_\beta, \\ v_{\alpha,\beta}^{(0)} &= \begin{cases} \gamma_\beta, & \alpha = \beta, \\ 0, & \alpha \neq \beta. \end{cases} \end{aligned} \tag{15}$$

This yields the constant terms $U^t(0)$ and $V(0)$ of $U^t(z)$ and $V(z)$, respectively. The remaining components

$$\mathcal{Y} = \left[\begin{array}{ccc|ccc|c|ccc} u_1^{(1)} & \cdots & u_1^{(n_0)} & v_{1,1}^{(1)} & \cdots & v_{1,1}^{(n_1)} & & v_{k,1}^{(1)} & \cdots & v_{k,1}^{(n_k)} \\ \vdots & & \vdots & \vdots & & \vdots & \cdots & \vdots & & \vdots \\ u_k^{(1)} & \cdots & u_k^{(n_0)} & v_{1,k}^{(1)} & \cdots & v_{1,k}^{(n_1)} & & v_{k,k}^{(1)} & \cdots & v_{k,k}^{(n_k)} \end{array}\right]^t \tag{16}$$

can be obtained by solving

$$\mathcal{M}_n \cdot \mathcal{Y} = -\left[\begin{array}{ccc} a_0^{(1)} & \cdots & a_k^{(1)} \\ \vdots & & \vdots \\ a_0^{(\|n\|)} & \cdots & a_k^{(\|n\|)} \end{array}\right] \left[\frac{U^t(0)}{V(0)}\right]. \tag{17}$$

In (17), we require that $\gamma_\beta \neq 0, 1 \leq \beta \leq k$; $\gamma_\beta = 1$ for a normalized NPHS. Again, the existence of a solution to (17) is assured if $\mathcal{M}_n$ is nonsingular. The terms $\delta w_\beta(z)$, $1 \leq \beta \leq k$, in (14) represent the residual errors made when solving (15) and (17).

For the special case when $n = [n_0, 0, \ldots, 0]$ the NPHS becomes

$$S(z) = \left[\begin{array}{c|c} [a_0^{(0)}]^{-1} z^{n_0+1} & U^t(z) \\ \hline 0 & I_k \end{array}\right] \cdot diag[\gamma_0, \cdots, \gamma_k], \tag{18}$$

where $U^t(z) = -[a_0(z)]^{-1} \cdot [a_1(z), \cdots, a_k(z)]$ $(mod\ z^{n_0+1})$. For initialization purposes in the algorithm given later in §4, we adopt (18) even in the cases $n_0 = 0$ and $n_0 = -1$, despite the fact that it no longer strictly meets all the requirements of an NPHS.

3. Simultaneous Padé Systems. A Padé-Hermite system gives an approximation to a vector of formal power series using matrix multiplication on the right. In this section we give the definition of a simultaneous Padé system which corresponds to a similar approximation but with matrix multiplication on the left and with degree constraints that can be thought of as being "dual" to the degree constraints of a

Padé-Hermite system. As in the previous section, a simultaneous Padé system exists if and only if a particular matrix of Sylvester type is nonsingular, in this case it is a mosaic Sylvester matrix.

Let

$$A^*(z) = \left[\begin{array}{ccc} a^*_{0,1}(z) & \cdots & a^*_{0,k}(z) \\ \hline a^*_{1,1}(z) & \cdots & a^*_{1,k}(z) \\ \vdots & & \vdots \\ a^*_{k,1}(z) & \cdots & a^*_{k,k}(z) \end{array}\right] = \left[\begin{array}{c} B^{*t}(z) \\ \hline C^*(z) \end{array}\right] \tag{19}$$

be a $(k+1)\times k$ matrix of power series with $\det(C^*(0)) \neq 0$. The $(k+1)\times(k+1)$ matrix of polynomials

$$S^*(z) = \left[\begin{array}{c|c} v^*(z) & U^{*t}(z) \\ \hline z^2Q^*(z) & z^2P^*(z) \end{array}\right] = \left[\begin{array}{c|ccc} v^*(z) & u^*_1(z) & \cdots & u^*_k(z) \\ \hline z^2q^*_1(z) & z^2p^*_{1,1}(z) & \cdots & z^2p^*_{1,k}(z) \\ \vdots & \vdots & & \vdots \\ z^2q^*_k(z) & z^2p^*_{k,1}(z) & \cdots & z^2p^*_{k,k}(z) \end{array}\right] \tag{20}$$

is a simultaneous Padé system (SPS) [12, 14] of type n for $A^*(z)$ if the following conditions are satisfied.

I. **(Degree conditions)**: For $1 \leq \alpha, \beta \leq k$,

$$\begin{aligned} v^*(z) &= \sum_{\ell=0}^{\|n\|-n_0} v^{*(\ell)} z^\ell, & u^*_\beta(z) &= \sum_{\ell=0}^{\|n\|-n_\beta} u^{*(\ell)}_\beta z^\ell, \\ q^*_\alpha(z) &= \sum_{\ell=0}^{\|n\|-n_0-1} q^{*(\ell)}_\alpha z^\ell, & p^*_{\alpha,\beta}(z) &= \sum_{\ell=0}^{\|n\|-n_\beta-1} p^{*(\ell)}_{\alpha,\beta} z^\ell. \end{aligned} \tag{21}$$

II. **(Order condition)**:

$$S^*(z)A^*(z) = z^{\|n\|+1}T^*(z), \tag{22}$$

where $T^{*t}(z) = [W^*(z)|R^{*t}(z)]$ with $R^*(z)$ a $k\times k$ matrix.

III. **(Nonsingularity condition)**: The constant term of $R^*(z)$ is a diagonal matrix

$$R^*(0) = diag\ [\gamma^*_1, \ldots, \gamma^*_k], \tag{23}$$

and

$$\gamma^* \equiv (a^{(0)}_0)^{-1} \prod_{\alpha=0}^{k} \gamma^*_\alpha\ \neq 0, \tag{24}$$

where $\gamma^*_0 = v^*(0)$.

Remark 5: The SPS is said to be **normalized** [12] if the nonsingularity condition III is replaced by $v^*(0) = 1$ and $R^*(0) = I_k$. This can be achieved by multiplying $S^*(z)$ on the left by Γ^{*-1}, where

$$\Gamma^* = diag\,[\gamma^*_0, \ldots, \gamma^*_k]. \tag{25}$$

The SPS is said to be **scaled** when each row of $S^*(z)$ has norm equal to 1 for some norm and if, in addition, $\gamma_\alpha^* > 0$, $0 \le \alpha \le k$. Here, also, scaling a SPS is accomplished by multiplying it on the left by an appropriate diagonal matrix.

Remark 6: The nonsingularity condition III, namely $\gamma^* \neq 0$, refers to the nonsingularity of $S^*(z)$; that is, $S^*(z)$ is nonsingular iff $\gamma^* \neq 0$ (this follows from Theorem 1 given below and from an observation about $\det(S(z))$ made in Remark 4). Equivalently, the nonsingularity condition refers to the nonsingularity of the associated mosaic Sylvester matrix $\mathcal{M}_n^*$ defined in (27); in [12] it is shown that a SPS exists iff $\mathcal{M}_n^*$ is nonsingular.

As for the Padé-Hermite system, if the order condition (22) is not satisfied exactly, but rather

$$S^*(z)A^*(z) = z^{\|n\|+1}T^*(z) + \delta T^*(z), \tag{26}$$

where $\delta T^{*t}(z) = \left[\delta W^*(z)|z^2\ \delta R^{*t}(z)\right]$ (with $\delta R^*(z)$ a $k \times k$ matrix) is a relatively "small" residual error, then $S^*(z)$ is called a numerical simultaneous Padé system (NSPS). In (26), for $1 \le \alpha, \beta \le k$,

$$\begin{aligned} \delta w_\beta^*(z) &= \sum_{\ell=0}^{\|n\|} \delta w_\beta^{*(\ell)}\ z^\ell, \\ \delta r_{\alpha,\beta}^*(z) &= \sum_{\ell=0}^{\|n\|-2} \delta r_{\alpha,\beta}^{*(\ell)}\ z^\ell. \end{aligned}$$

As with the NPHS $S(z)$, a NSPS for which $\delta T^*(z) = 0$ is denoted by $S_E^*(z)$.

Associated with $A^*(z)$, let $\mathcal{M}_n^*$ be the mosaic Sylvester matrix of order $k\|n\|$,

$$\mathcal{M}_n^* = \begin{bmatrix} \mathcal{S}_{0,1}^* & \cdots & \mathcal{S}_{0,k}^* \\ \vdots & & \vdots \\ \mathcal{S}_{k,1}^* & \cdots & \mathcal{S}_{k,k}^* \end{bmatrix}, \tag{27}$$

where, for $0 \le \alpha \le k$ and $1 \le \beta \le k$,

$$\mathcal{S}_{\alpha,\beta}^* = \begin{bmatrix} a_{\alpha,\beta}^{*(0)} & & \cdots & & a_{\alpha,\beta}^{*(\|n\|-1)} \\ & \ddots & & & \vdots \\ & & a_{\alpha,\beta}^{*(0)} & \cdots & a_{\alpha,\beta}^{*(n_\alpha)} \end{bmatrix}.$$

Also define the order $k(\|n\|+1)$ matrix

$$\mathcal{N}_n^* = \left[\begin{array}{c|ccc|c|ccc} & a_{1,1}^{*(1)} & \cdots & a_{1,1}^{*(\|n\|)} & & a_{1,k}^{*(1)} & \cdots & a_{1,k}^{*(\|n\|)} \\ C^*(0) & \vdots & & \vdots & \cdots & \vdots & & \vdots \\ & a_{k,1}^{*(1)} & \cdots & a_{k,1}^{*(\|n\|)} & & a_{k,k}^{*(1)} & \cdots & a_{k,k}^{*(\|n\|)} \\ \hline \mathbf{0} & \multicolumn{7}{c}{\mathcal{M}_n^*} \end{array}\right]. \tag{28}$$

Then, as for the NPHS, $S^*(z)$ can be obtained by solving two sets of linear equations with $\mathcal{M}_n^*$ and $\mathcal{N}_n^*$ as the coefficient matrices (also see [12]).

To obtain $S_{0,1}^*(z), \ldots, S_{0,k}^*(z)$ of $S^*(z)$, we use

$$(29)\ v^*(z)\ a_{0,\beta}^*(z) + \sum_{\alpha=1}^{k} u_\alpha^*(z) a_{\alpha,\beta}^*(z)\ = z^{\|n\|+1} w_\beta^*(z) + \delta w_\beta^*(z),\ \ 1 \le \beta \le k,$$

which is the first row of (26). Matching coefficients of $1, z, \ldots, z^{\|n\|}$ in (29) gives

$$(30)\quad \mathcal{X}^{*t} \cdot \mathcal{N}_n^* = \\ -v^{*(0)} \left[\ B^{*t}(0)\ \middle|\ a_{0,1}^{*(1)},\quad \ldots,\quad a_{0,1}^{*(\|n\|)}\ \middle|\ \cdots\ \middle|\ a_{0,k}^{*(1)},\quad \ldots,\quad a_{0,k}^{*(\|n\|)}\ \right],$$

where

$$\mathcal{X}^{*t} = [u_1^{*(0)}, \ldots, u_k^{*(0)} | v^{*(1)}, \ldots, v^{*(\|n\|-n_0)} | u_1^{*(1)}, \ldots, u_1^{*(\|n\|-n_1)} | \cdots \\ \cdots | u_k^{*(1)}, \ldots, u_k^{*(\|n\|-n_k)}].$$

With $v^{*(0)} = \gamma_0^* \neq 0$ specified ($\gamma_0^* = 1$ for a normalized NSPS), a unique solution to (30) is assured if $\mathcal{M}_n^*$ is nonsingular, since by assumption $\det[C^*(0)] \neq 0$. The terms $\delta w_\beta^*(z)$ in (29) represent the residual errors made in solving (30).

Next, to compute $P^*(z)$ and $Q^*(z)$ (i.e., the remaining rows of $S^*(z)$), again we use (26), namely,

$$(31)\ q_\alpha^*(z)\ a_{0,\beta}^*(z) + \sum_{\rho=1}^{k} p_{\alpha,\rho}^*(z)\ a_{\rho,\beta}^*(z) = z^{\|n\|-1} r_{\alpha,\beta}^*(z) + \delta r_{\alpha,\beta}^*(z),\ \ 1 \le \alpha, \beta \le k.$$

Let

$$\mathcal{Y}_\alpha^{*t} = \left[q_\alpha^{*(0)}, \ldots, q_\alpha^{*(\|n\|-n_0-1)} | p_{\alpha,1}^{*(0)}, \ldots, p_{\alpha,1}^{*(\|n\|-n_1-1)} | \cdots | p_{\alpha,k}^{*(0)}, \ldots, p_{\alpha,k}^{*(\|n\|-n_k-1)} \right].$$

Then, (31) and the requirement that $R^*(0) = diag[\gamma_1^*, \ldots, \gamma_k^*]$ yields

$$(32)\qquad \mathcal{Y}_\alpha^{*t} \cdot \mathcal{M}_n^* = \gamma_\alpha^* E_{\alpha\|n\|}^t,\ \ 1 \le \alpha \le k,$$

where $E_{\alpha\|n\|}^t$ is the unit row vector of length $k\|n\|$ with a single 1 in position $\alpha\|n\|$. With $diag[\gamma_1^*, \ldots, \gamma_k^*]$ specified ($\gamma_\alpha^* = 1$ for a normalized NSPS), a solution of (32) exists uniquely if $\mathcal{M}_n^*$ is nonsingular. The solution $\mathcal{Y}_\alpha^*$ provides the αth row of $S^*(z)$; namely, $S_{\alpha,0}^*(z) = z^2 \cdot q_\alpha^*(z)$ and $S_{\alpha,\beta}^*(z) = z^2 \cdot p_{\alpha,\beta}^*(z),\quad 1 \le \beta \le k$. The terms $\delta r_{\alpha,\beta}^*(z)$ in (31) represent the residual errors made in solving (32).

In the remainder of the paper, without loss of generality, we make the simplifying assumption that

$$(33)\qquad A^*(z) = \left[\begin{array}{ccc|c} -a_1(z) & \cdots & -a_k(z) & \\ \hline a_0(z) & & & \mathbf{0} \\ & \ddots & & \\ \mathbf{0} & & a_0(z) & \end{array}\right].$$

With $A^*(z)$ defined by (33), for the special case when $n = [n_0, 0, \ldots, 0]$, the NSPS becomes

$$(34)\qquad S^*(z)\ =\ diag[\gamma_0^*, \cdots, \gamma_k^*] \left[\begin{array}{c|c} 1 & U^{*t}(z) \\ \hline 0 & [a_0^{(0)}]^{-1} z^{n_0+1} I_k \end{array}\right],$$

where $U^{*t}(z) = [a_0(z)]^{-1} \cdot [a_1(z), \cdots, a_k(z)]$ $(mod\ z^{n_0+1})$. For initialization purposes in the algorithm given in §4, we adopt (34) even in the case when $n_0 = 0$ and $n_0 = -1$, despite the fact that it no longer strictly meet all the requirements of a NSPS.

In addition, with $A^*(z)$ defined by (33), there is an important commutativity relationship between Padé-Hermite systems and simultaneous Padé systems, given in Theorem 1 below. This relationship is used later in §5. But, in our presentation, the residual $T^*(z)$ continues to take the more general form (19) rather than (33); because, for the computation of the NSPS for $T^*(z)$, which is required by the algorithm given in §4, the conversion of $T^*(z)$ from the form (19) to the form (33) by means of multiplication on the right by $R^{*-1}(z)$ introduces undesirable instabilities.

THEOREM 1. *If $S(z)$ is a NPHS of type n for $A(z)$ and $S^*(z)$ is a NSPS of type n for $A^*(z)$, then*

$$S^*(z) \cdot S(z) = z^{\|n\|+1}(a_0^{(0)})^{-1}\Gamma^*\Gamma + \theta_I(z), \tag{35}$$

where

$$\theta_I(z) \;=\; a_0^{-1}(z)\left\{\begin{bmatrix} v^*(z) \\ z^2Q^*(z) \end{bmatrix}\delta T^t(z) + \delta T^*(z)\left[\begin{array}{c|c} z^2Q(z) & V(z) \end{array}\right]\right\} \;\; (mod\ z^{D+1})$$

with

$$D = \left[\begin{array}{c|ccc} \|n\|+1 & \|n\| & \cdots & \|n\| \\ \hline \|n\|+2 & \|n\|+1 & \cdots & \|n\|+1 \\ \vdots & \vdots & & \vdots \\ \|n\|+2 & \|n\|+1 & \cdots & \|n\|+1 \end{array}\right]$$

and with the modulo operation applied component-wise.

Proof. See [9]. ∎

Thus, given a NPHS, a NSPS can be computed using (35). However, the stability of such a computation is not known, and we choose instead to compute NPHS and NSPS systems in tandem by the algorithm described in the next section.

4. The Algorithm. To compute a NPHS of type n for $A(z)$ and a NSPS of type n for $A^*(z)$, the systems (13), (17), (30) and (32) can be solved using a method such as Gaussian elimination. This method, while not restricting the input power series, does not take advantage of the inherent structure of the coefficient matrices $\mathcal{M}_n$ and $\mathcal{M}_n^*$. Alternatively, a variety of recurrence relations which do take advantage of this structure have been described in the literature ([27],[4],[12],[14]). These recurrence relations usually lead to much more efficient algorithms for algebraically computing Padé-Hermite systems and simultaneous Padé systems. The recurrence relations given in [12] and [14] appear to be the most easily adaptable to numerical computation and it is the detailed study of the numerical behavior of these recurrences to which we devote the remainder of this paper. We begin by briefly describing these recurrences in the algebraic case.

Let $e_0 = [1, 0, \ldots, 0]$ be a $1 \times k+1$ vector, set

$$M \;=\; \min\left\{n_0,\; \max_{1\le\beta\le k}\{n_\beta\}\right\} \;+\; 1,$$

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 $i > 0$,

$$n_\beta^{(i)} = \max\{0, n_\beta - M + i\}, \qquad \beta = 0, \ldots, k.$$

Then the sequence $\{n^{(i)}\}_{i=0,1,\ldots}$ lies on a piecewise linear path with $n_\beta^{(i+1)} \ge n_\beta^{(i)}$ for each i, β and[3] $n^{(M)} = n$. The sequence $\{n^{(i)}\}$ contains a subsequence $\{m^{(\sigma)}\}$ called the **sequence of nonsingular points** for $A(z)$ and $A^*(z)$. This sequence is defined by $m^{(\sigma)} = n^{(i_\sigma)}$, where

$$i_\sigma = \begin{cases} 0, & \sigma = 0, \\ \min\{i > i_{\sigma-1} : \det(\mathcal{M}_{n^{(i)}}) \neq 0\}, & \sigma \ge 1, \end{cases}$$

where $\det(\mathcal{M}_{n^{(i)}})$ is the determinant[4] of $\mathcal{M}_{n^{(i)}}$. Corresponding to the sequence of nonsingular points $\{m^{(\sigma)}\}$ is the sequence $\left\{S_E^{(\sigma)}(z)\right\}$ of Padé-Hermite systems with residuals $\left\{T_E^{(\sigma)^t}(z)\right\}$ and the sequence $\left\{S_E^{*(\sigma)}(z)\right\}$ of simultaneous Padé systems with residuals $\left\{T_E^{*(\sigma)}(z)\right\}$. For $\sigma = 0$, set $\left\{S_E^{(0)}(z)\right\} = \left\{S_E^{*(0)}(z)\right\} = I_{k+1}$. We have that

$$A^t(z) \cdot S_E^{(\sigma)}(z) = z^{\|m^{(\sigma)}\|+1}\, T_E^{(\sigma)t}(z)$$

and

$$S_E^{*(\sigma)}(z) \cdot A^*(z) = z^{\|m^{(\sigma)}\|+1}\, T_E^{*(\sigma)}(z).$$

The following theorem provides a relation of the $(\sigma+1)$th exact systems in terms of the σth exact systems.

THEOREM 2. *For $\sigma \ge 0$ and $i > i_\sigma$, let $\nu = n^{(i)} - m^{(\sigma)} - e_0$. Then, the following statements are equivalent.*

1. *$n^{(i)}$ is a nonsingular point for $A(z)$ and $A^*(z)$.*
2. *ν is a nonsingular point for $T_E^{(\sigma)}(z)$.*
3. *ν is a nonsingular point for $T_E^{*(\sigma)}(z)$.*

Furthermore, we have the recurrence relations

$$S_E^{(\sigma+1)}(z) = S_E^{(\sigma)}(z) \cdot \widehat{S}_E(z), \quad T_E^{(\sigma+1)}(z) = \widehat{T}_E(z), \tag{36}$$

and

$$S_E^{*(\sigma+1)}(z) = \widehat{S}_E^*(z) \cdot S_E^{*(\sigma)}(z), \quad T_E^{*(\sigma+1)}(z) = \widehat{T}_E^*(z), \tag{37}$$

where $\widehat{S}_E(z)$ is the Padé-Hermite system of type $(m^{(\sigma+1)} - m^{(\sigma)} - e_0)$ for $T_E^{(\sigma)}(z)$ with residual $\widehat{T}_E(z)$ and $\widehat{S}_E^*(z)$ is the simultaneous Padé system of type $(m^{(\sigma+1)} - m^{(\sigma)} - e_0)$ for $T_E^{*(\sigma)}(z)$ with residual $\widehat{T}_E^*(z)$.

Proof. The proof for the NPHS is given in [14] and for the NSPS in [12]. ∎

[3] We assume here with loss of generality that $n_\beta \ge 0, 0 \le \beta \le k$, because if $n_\beta = -1$ for some β, we can simply remove n_β from n and $a_\beta(z)$ from $A^t(z)$ and decrease k by 1.

[4] By convention, the determinant of a null matrix is defined to be equal to 1.

Theorem 2 reduces the problem of determining a Padé-Hermite system and a simultaneous Padé system of types $m^{(\sigma+1)}$ to two smaller problems: determine systems of type $m^{(\sigma)}$ for the original power series and then determine systems of type $\nu = m^{(\sigma+1)} - m^{(\sigma)} - e_0$ for the residual power series. For the residual power series, the system $\widehat{S}_E(z)$ is obtained by solving the linear equations (13) and (17), where in the following the associated matrix is now denoted by $\widehat{\mathcal{M}}_\nu$ rather than by $\mathcal{M}_\nu$; and, the system $\widehat{S}^*_E(z)$ is obtained by solving the linear equations (30) and (32), where in the following the associated matrix is now denoted by $\widehat{\mathcal{M}}^*_\nu$ rather than by $\mathcal{M}^*_\nu$. 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 $S_E^{(\sigma+1)}(z)$ and $S_E^{*(\sigma+1)}(z)$ in (36) and (37). This overhead cost summed over all the steps, in general, is an order of magnitude less than the cost of solving the linear systems (13), (17), (30) and (32) directly.

Numerically, the recurrences (36) and (37) perform badly if $\mathcal{M}_{m^{(\sigma)}}$ and $\mathcal{M}^*_{m^{(\sigma)}}$ are ill-conditioned at any point $m^{(\sigma)}$. Rather than moving from nonsingular point to nonsingular point along the diagonal, what we would like to do is move from a well-conditioned point to the next well-conditioned point. This is the motivation for the algorithm PHS_SPS given below, where the points $m^{(\sigma)},\ \sigma = 0, 1, \ldots$, correspond to stable points rather than to nonsingular points and we step over unstable blocks.

A quantitative measure of the stability of a point $m^{(\sigma)}$ is provided by the stability parameter

$$\kappa^{(\sigma)} = \sum_{\beta=0}^{k} (\gamma_\beta^{(\sigma)} \gamma_\beta^{*(\sigma)})^{-1}. \tag{38}$$

It is shown in [9, 11] that $2\kappa^{(\sigma)} |a_0^{(0)}| \cdot \|a_0^{-1}(z)\ (mod\ z^{\|n\|+1})\|$ is an upper bound for the condition numbers $\|\mathcal{M}_{m^{(\sigma)}}\| \cdot \|\mathcal{M}^{-1}_{m^{(\sigma)}}\|$ of $\mathcal{M}_{m^{(\sigma)}}$ and $\|\mathcal{M}^*_{m^{(\sigma)}}\|_\infty \cdot \|\mathcal{M}^{*-1}_{m^{(\sigma)}}\|_\infty$ of $\mathcal{M}^*_{m^{(\sigma)}}$. For the parameter (38), as well as considerations of §5 and §6 it is assumed that $S^{(\sigma)}(z)$ and $S^{*(\sigma)}(z)$ are both scaled and that $\|a_\beta(z)\| \le 1,\ \ 0 \le \beta \le k$. The norms used for the various scaling are defined in §1. In (38), it is also assumed that the residual errors $\delta T^{(\sigma)}(z)$ and $\delta T^{*(\sigma)}(z)$ in the order equations

$$A^t(z) \cdot S^{(\sigma)}(z) \ = \ z^{\|m^{(\sigma)}\|+1}\, T^{(\sigma)t}(z) \ + \ \delta T^{(\sigma)t}(z) \tag{39}$$

and

$$S^{*(\sigma)}(z) \cdot A^*(z) \ = \ z^{\|m^{(\sigma)}\|+1}\, T^{*(\sigma)}(z) \ + \ \delta T^{*(\sigma)}(z), \tag{40}$$

at the point $m^{(\sigma)}$ are relatively insignificant. We say that $m^{(\sigma)}$ is a **stable point** (or, a well-conditioned point) if for some preassigned tolerance τ, $\kappa^{(\sigma)} \le \tau$. In the algorithm below, the user supplies the tolerance value τ.

PHS_SPS($A(z),\ n,\ k, \tau$)
$\sigma \leftarrow 0$; $m^{(0)} \leftarrow -e_0$; $S^{(0)} \leftarrow I_{k+1}$; $S^{*(0)} \leftarrow I_{k+1}$;
$M \leftarrow \min\{n_0, \max_{1\le\beta\le k}\{n_\beta\}\} + 1$
$i \leftarrow 0$; *stable* $\leftarrow$ true
While (($i < M$) and *stable*) do
    $\nu \leftarrow n - m^{(\sigma)} - e_0$
    $s \leftarrow 0$; *stable* $\leftarrow$ false

While ($s < M - i$) and (not *stable*) do
$s \leftarrow s + 1$
$\nu_\beta^{(s)} \leftarrow \max\{0, \nu_\beta + i - M + s\}, \qquad \beta = 0, \ldots, k$
Compute the residuals $T^{(\sigma)}(z)$ and $T^{*(\sigma)}(z)$ in (39) and (40)
Construct the matrices $\mathcal{M}_{\nu^{(s)}}$ for $T^{(\sigma)}(z)$ and $\mathcal{M}^*_{\nu^{(s)}}$ for $T^{*(\sigma)}(z)$
If $\mathcal{M}_{\nu^{(s)}}$ is numerically nonsingular then
$m^{(\sigma+1)} \leftarrow m^{(\sigma)} + \nu^{(s)} + e_0$
Obtain $\widehat{S}(z)$ by solving (13) and (17) by Gaussian elimination
$S^{(\sigma+1)}(z) \leftarrow S^{(\sigma)}(z)\ \widehat{S}(z)$
Scale $S^{(\sigma+1)}(z)$ and compute $\Gamma^{(\sigma+1)}$
Obtain $\widehat{S}^*(z)$ by solving (30) and (32) by Gaussian elimination
$S^{*(\sigma+1)}(z) \leftarrow \widehat{S}^*(z)\ S^{*(\sigma)}(z)$
Scale $S^{*(\sigma+1)}(z)$ and compute $\Gamma^{*(\sigma+1)}$
Using (38), compute $\kappa^{(\sigma+1)}$
stable $\leftarrow \kappa^{(\sigma+1)} \leq \tau$
end If
end While
If *stable* then $\sigma \leftarrow \sigma + 1$; $\quad i \leftarrow i + s$
end While
If *stable* then return $(S^{(\sigma)}(z), S^{*(\sigma)}(z), \kappa^{(\sigma)})$ else return $(S^{(\sigma+1)}(z), S^{*(\sigma+1)}(z), \kappa^{(\sigma+1)})$
∎

In the algorithm above, by the numerical nonsingularity of a matrix, we mean that no zero pivot elements are encountered during the triangularization of the matrix by the Gaussian elimination method with partial pivoting..

5. Bounds on Errors in the Order Conditions. In this section, we give bounds for the errors in the order conditions for the NPHS and the NSPS computed by the algorithm PHS_SPS. Some of the details of the derivations are omitted and can be found in [9]; in particular, for the NSPS the final result only (without proof) is given.

We begin by giving some standard results from the field of floating-point error analysis. Let μ denote the unit floating-point error and assume that the degrees of all polynomials and the orders of all matrices are bounded by some N, where $N\mu \leq 0.01$ (this restriction comes from Forsythe and Moler [16]). Indeed, as an assumption for all the lemmas and theorems below, we require that $(\|n\| + k + 1)\mu \leq 0.01$. After Wilkinson [28], we denote a floating-point operation by $fl[\cdot]$. In the following results, it is assumed that the operands consist of floating-point numbers.

LEMMA 3. *If* $\partial\mu \leq 0.01$, *then*

$$fl[\sum_{k=1}^{\partial} u_k v_k] = \sum_{k=1}^{\partial} u_k v_k (1 + \delta_k),$$

where $|\delta_k| \leq 1.01\partial\mu$.

LEMMA 4. *If* $S(z)$ *is a NPHS of type* n *for* $A(z)$, *then*

$$fl[A^t(z) \cdot S(z)] = A^t(z) \cdot S(z) + \Psi^t(z),$$

where

$$\|\Psi^t(z)\| \leq 1.01\mu(\|n\| + k + 1)\|A^t(z)\| \cdot \|S(z)\|.$$

Proof. Using Lemma 3, for $0 \le \beta \le k$,

$$\begin{aligned} fl[\sum_{\alpha=0}^{k} a_\alpha(z) S_{\alpha,\beta}(z)] &= \sum_{\ell=0}^{\infty} z^\ell fl[\sum_{\alpha=0}^{k}\sum_{j=0}^{n_\alpha} a_\alpha^{(\ell-j)} S_{\alpha,\beta}^{(j)}] \\ &= \sum_{\ell=0}^{\infty} z^\ell \sum_{\alpha=0}^{k}\sum_{j=0}^{n_\alpha} a_\alpha^{(\ell-j)} S_{\alpha,\beta}^{(j)}(1+\delta_{\alpha,\beta,j,\ell}), \end{aligned}$$

where $|\delta_{\alpha,\beta,j,\ell}| \le 1.01(n_\alpha + k + 1)\mu$. So,

$$\Psi_\beta(z) = \sum_{\ell=0}^{\infty} z^\ell \sum_{\alpha=0}^{k}\sum_{j=0}^{n_\alpha} a_\alpha^{(\ell-j)} S_{\alpha,\beta}^{(j)} \delta_{\alpha,\beta,j,\ell},$$

and

$$\begin{aligned} \|\Psi^t(z)\| &= \max_{0\le\beta\le k}\{\|\Psi_\beta(z)\|\} \\ &\le \max_{0\le\beta\le k}\left\{\sum_{\ell=0}^{\infty}\sum_{\alpha=0}^{k}\sum_{j=0}^{n_\alpha} |a_\alpha^{(\ell-j)}|\cdot|S_{\alpha,\beta}^{(j)}|\cdot|\delta_{\alpha,\beta,j,\ell}|\right\} \\ &\le 1.01\mu \max_{0\le\beta\le k}\left\{\sum_{\alpha=0}^{k}(n_\alpha+k+1)\sum_{j=0}^{n_\alpha}|S_{\alpha,\beta}^{(j)}|\sum_{\ell=0}^{\infty}|a_\alpha^{(\ell-j)}|\right\} \\ &\le 1.01\mu \max_{0\le\alpha\le k}\{n_\alpha+k+1\}\|A^t(z)\| \max_{0\le\beta\le k}\left\{\sum_{\alpha=0}^{k}\|S_{\alpha,\beta}(z)\|\right\} \\ &\le 1.01\mu(\|n\|+k+1)\|A^t(z)\|\cdot\|S(z)\|. \end{aligned}$$

∎

We begin the analysis of the error in the order condition in the NSPS by first examining the floating-point errors introduced by one iteration of the algorithm. At the σth iteration, the NPHS $S^{(\sigma)}(z)$ of type $m^{(\sigma)}$ for $A^t(z)$ is available and satisfies

$$A^t(z)\cdot S^{(\sigma)}(z) = \delta {T^{(\sigma)}}^t(z) + \mathcal{O}(z^{\|m^{(\sigma)}\|+1}).$$

The algorithm proceeds to compute $S^{(\sigma+1)}(z)$ of type $m^{(\sigma+1)}$.

An iterative step consists of three parts. In the first part, the first $\|\nu^{(\sigma)}\|+1$ terms of $T^{(\sigma)}(z)$ are computed; a bound for the floating-point errors introduced in this part is given in Lemma 5 below. In the second part, the NPHS $\widehat{S}^{(\sigma)}(z)$ of type $\nu^{(\sigma)}$ for $T^{(\sigma)}(z)$ is computed; an error analysis is given Lemma 6. In the third part, Lemma 7 provides bounds for the floating-point errors introduced in computing $S^{(\sigma+1)}(z) = S^{(\sigma)}(z)\cdot\widehat{S}^{(\sigma)}(z)$. At this point in the algorithm, $S^{(\sigma+1)}(z)$ is scaled so that the norm of each column is 1. We assume for the sake of simplicity that this scaling introduces no additional errors. This is reasonable assumption because errors due to scaling are comparatively insignificant[5].

[5] Note also that $\widehat{S}^{(\sigma)}(z)$ can be determined a posteriori with appropriate values of $\hat{\gamma}^{(\sigma)}$ so that $S^{(\sigma+1)}(z)$ is already scaled. None of the subsequent error bounds would change, and so in reality this assumption is made without loss of generality.

LEMMA 5. *The computed residual* $T^{(\sigma)}(z)$ *satisfies*

$$z^{\|m^{(\sigma)}\|+1}{T^{(\sigma)}}^t(z) = A^t(z)\cdot S^{(\sigma)}(z) - \delta{T^{(\sigma)}}^t(z) + z^{\|m^{(\sigma)}\|+1}{\theta_{II}^{(\sigma)}}^t(z),$$

where

$$\|{\theta_{II}^{(\sigma)}}^t(z)\| \le 1.01(\|m^{(\sigma)}\| + k + 1)\cdot\mu.$$

Proof. The result is an immediate consequence of Lemma 4 since $A^t(z)$ and $S^{(\sigma)}(z)$ are both scaled. For details, see [9]. ∎

LEMMA 6. *If* $\widehat{\mathcal{M}}_{\nu^{(\sigma)}}$ *is nonsingular and* $\widehat{S}^{(\sigma)}(z)$ *is obtained by solving (13) and (17), then*

$$T^{(\sigma)^t}(z)\cdot\widehat{S}^{(\sigma)}(z) = {\theta_{III}^{(\sigma)}}^t(z) + O(z^{\|\nu^{(\sigma)}\|+1}),$$

where

$$\|{\theta_{III}^{(\sigma)}}^t(z)\| \le (8\|\nu^{(\sigma)}\|^3\cdot\rho_\sigma\cdot\mu + O(\mu^2))\cdot\|\widehat{S}^{(\sigma)}(z)\|.$$

Proof. First we obtain bounds for the first component of ${\theta_{III}^{(\sigma)}}^t(z)$. The first column of $\widehat{S}^{(\sigma)}(z)$ corresponds to the solution $\widehat{\mathcal{X}}$ of (13) obtained by Gaussian elimination. The vector $\widehat{\mathcal{X}}$ is the exact solution of

$$(\widehat{\mathcal{M}}_{\nu^{(\sigma)}} + \mathcal{E})\cdot\widehat{\mathcal{X}} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix},$$

where[6]

$$\|\mathcal{E}\|_1 \le 8\|\nu^{(\sigma)}\|^3\cdot\rho_\sigma\cdot\|\widehat{\mathcal{M}}_{\nu^{(\sigma)}}\|_1\cdot\mu + O(\mu^2)$$

and ρ_σ is the growth factor associated with the LU-decomposition of $\widehat{\mathcal{M}}_{\nu^{(\sigma)}}$ ([17](page 67)). But, from Lemma 4,

$$\|{T^{(\sigma)}}^t(z)\| \le 1 + 1.01\cdot(\|m^{(\sigma)}\| + k + 1)\cdot\mu,$$

since $A(z)$ and $S^{(\sigma)}(z)$ are both scaled. So,

$$\|\widehat{\mathcal{M}}_{\nu^{(\sigma)}}\|_1 \le \|{T^{(\sigma)}}^t(z)\| \le 1 + O(\mu)$$

Thus,

$$\widehat{\mathcal{M}}_{\nu^{(\sigma)}}\cdot\widehat{\mathcal{X}} - \begin{pmatrix} 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix} = -\mathcal{E}\cdot\widehat{\mathcal{X}},$$

[6] The results in [17] use the ∞-norm, but it is easy to show that they are also valid using the 1-norm. With partial pivoting, ρ_σ is of order unity in practice. Examples can be constructed, however, where the growth factor ρ grows exponentially if partial pivoting is used, but in practice ρ_σ is usually comparable to the modest growth that results when complete pivoting is used (which is approximately 10 in practice) [17, page 69]. Further discussion and new results regarding the growth factor ρ_σ can be found in [21] and [26].

where

$$\|\mathcal{E}\cdot\widehat{\mathcal{X}}\|_1 \le \left\{8\|\nu^{(\sigma)}\|^3\cdot\rho_\sigma\cdot\mu \;+\; O(\mu^2)\right\}\cdot\|\widehat{\mathcal{X}}\|_1.$$

A similar analysis can be done for solving (17) to obtain $\widehat{\mathcal{Y}}$. But $\widehat{\mathcal{X}}$ yields the first column of $\widehat{S}^{(\sigma)}(z)$ with residual error $\mathcal{E}\cdot\widehat{\mathcal{X}}$ and $\widehat{\mathcal{Y}}$ yields the remaining columns of $\widehat{S}^{(\sigma)}(z)$ with a corresponding residual error. Thus,

$$T^{(\sigma)^t}(z)\cdot\widehat{S}^{(\sigma)}(z) = \theta_{III}^{(\sigma)^t}(z) + O(z^{\|\nu^{(\sigma)}\|+1}),$$

where

$$\|\theta_{III}^{(\sigma)^t}(z)\| \le \left\{8\|\nu^{(\sigma)}\|^3\cdot\rho_\sigma\cdot\mu + O(\mu^2)\right\}\cdot\|\widehat{S}^{(\sigma)}(z)\|.$$

∎

LEMMA 7. *If* $S^{(\sigma+1)}(z) = fl(S^{(\sigma)}(z)\cdot\widehat{S}^{(\sigma)}(z))$, *then*

$$S^{(\sigma+1)}(z) \;=\; S^{(\sigma)}(z)\cdot\widehat{S}^{(\sigma)}(z) \;+\; \theta_{IV}^{(\sigma)}(z),$$

where

$$\|\theta_{IV}^{(\sigma)}(z)\| \;\le\; 1.01(\|\nu^{(\sigma)}\|+k+1)\cdot\|S^{(\sigma)}(z)\|\cdot\|\widehat{S}^{(\sigma)}(z)\|\mu.$$

Proof. For $1\le\alpha,\beta\le k$, the (α,β)-component of $S^{(\sigma+1)}(z)$ is

$$\begin{aligned}
&fl\left[z^2 q_\alpha(z)\cdot\widehat{u}_\beta(z) + \sum_{\rho=1}^{k} v_{\alpha,\rho}(z)\cdot\widehat{v}_{\rho,\beta}(z)\right]\\
&= \;fl\left[z^2\sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_0^{(\sigma)}-1} z^\ell\sum_{j=0}^{\nu_0^{(\sigma)}} q_\alpha^{(\ell-j)}\widehat{u}_\beta^{(j)} \;+\; \sum_{\rho=1}^{k}\sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_\rho^{(\sigma)}} z^\ell\sum_{j=0}^{\nu_\rho^{(\sigma)}} v_{\alpha,\rho}^{(\ell-j)}\widehat{v}_{\rho,\beta}^{(j)}\right]\\
&= \sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_0^{(\sigma)}-1} z^{\ell+2}\sum_{j=0}^{\nu_0^{(\sigma)}} q_\alpha^{(\ell-j)}\widehat{u}_\beta^{(j)}\cdot(1+\delta_{\alpha,\beta,j,\ell,0}).\\
&\qquad\qquad + \sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_\rho^{(\sigma)}} z^\ell\sum_{\rho=1}^{k}\sum_{j=0}^{\nu_\rho^{(\sigma)}} v_{\alpha,\rho}^{(\ell-j)}\widehat{v}_{\rho,\beta}^{(j)}\cdot(1+\delta_{\alpha,\beta,j,\ell,\rho}),
\end{aligned}$$

where $|\delta_{\alpha,\beta,j,\ell,\rho}| \;\le\; 1.01\cdot(\nu_\rho^{(\sigma)}+k+1)\cdot\mu$. Here, we have used Lemma 3 with the assumption that $(\|\nu^{(\sigma)}\|+k+1)\mu\le 0.01$. So,

$$\begin{aligned}
\left(\theta_{IV}^{(\sigma)}(z)\right)_{\alpha,\beta} \;&=\; z^2\sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_0^{(\sigma)}-1} z^\ell\sum_{j=0}^{\nu_0^{(\sigma)}} q_\alpha^{(\ell-j)}\cdot\widehat{u}_\beta^{(j)}\cdot\delta_{\alpha,\beta,j,\ell,0}\\
&\quad + \sum_{\rho=1}^{k}\sum_{\ell=0}^{m_\alpha^{(\sigma)}+\nu_\rho^{(\sigma)}} z^\ell\sum_{j=0}^{\nu_\rho^{(\sigma)}} v_{\alpha,\rho}^{(\ell-j)}\cdot\widehat{v}_{\rho,\beta}^{(j)}\cdot\delta_{\alpha,\beta,j,\ell,\rho}.
\end{aligned}$$

Thus,

$$\|\left(\theta_{IV}^{(\sigma)}(z)\right)_{\alpha,\beta}\| \;\le\; 1.01\cdot(\|\nu^{(\sigma)}\|+k+1)\cdot\{\|q_\alpha(z)\|\cdot\|\widehat{u}_\beta(z)\| + \sum_{\rho=1}^{k}\|v_{\alpha,\rho}(z)\|\cdot\|\widehat{v}_{\rho,\beta}(z)\|\}\mu.$$

An equivalent result holds for $\alpha = \beta = 0$. The lemma now follows. ∎

The use of the results of the three lemmas above enables us to express the residual error $\delta T^{(\sigma+1)^t}(z)$ in the order condition at the $(\sigma+1)$th iteration in terms of the residual error $\delta T^{(\sigma)^t}(z)$ at the σth iteration plus the floating-point errors introduced "locally" by the σth iteration.

LEMMA 8.

$$\delta T^{(\sigma+1)^t}(z) \;=\; \delta T^{(\sigma)^t}(z) \cdot \widehat{S}^{(\sigma)}(z) + \mathcal{L}^{(\sigma)^t}(z), \tag{41}$$

where

$$\begin{aligned}\mathcal{L}^{(\sigma)^t}(z) \;=\; &\Big\{ A^t(z) \cdot \theta_{IV}^{(\sigma)}(z) \\ &+\; z^{\|m^{(\sigma)}\|+1} \left[\theta_{III}^{(\sigma)^t}(z) \;-\; \theta_{II}^{(\sigma)^t}(z) \cdot \widehat{S}^{(\sigma)}(z) \right] \Big\} \quad (mod \;\; z^{\|m^{(\sigma+1)}\|+1}).\end{aligned}$$

Proof. The result is an immediate consequence of Lemmas 5, 6 and 7. ∎

Thus, the residual error $\delta T^{(\sigma+1)^t}(z)$ is composed of the local error $\mathcal{L}^{(\sigma)^t}(z)$ introduced by the σth iteration plus the residual error $\delta T^{(\sigma)^t}(z)$ from the previous iteration propagated by $\widehat{S}^{(\sigma)}(z)$. Applying (41) recursively, we obtain the following.

THEOREM 9. *The residual error satisfies*

$$\delta T^{(\sigma+1)^t}(z) = \sum_{j=0}^{\sigma} \mathcal{L}^{(j)^t}(z) \cdot \mathcal{G}_j^{(\sigma)}(z), \tag{42}$$

where

$$\mathcal{G}_j^{(\sigma)}(z) = \begin{cases} \widehat{S}^{(j+1)}(z) \cdot \widehat{S}^{(j+2)}(z) \cdots \widehat{S}^{(\sigma)}(z), & 0 \le j < \sigma, \\ I_{k+1}, & j = \sigma. \end{cases} \tag{43}$$

Proof. The result follows by induction from Lemma 8. ∎

¿From (42), we see that the residual error $\delta T^{(\sigma+1)^t}(z)$ is composed of the local errors $\mathcal{L}^{(j)^t}(z)$ propagated by $\mathcal{G}_j^{(\sigma)}$. Lemmas 5, 6 and 7 provide bounds for $\mathcal{L}^{(j)^t}(z)$. To obtain a bound for $\delta T^{(\sigma+1)^t}(z)$, it remains to determine bounds for the propagation matrices $\mathcal{G}_j^{(\sigma)}$. The concern is that the $\widehat{S}^{(j)}(z)$ making up $\mathcal{G}_j^{(\sigma)}$ will cause $\mathcal{G}_j^{(\sigma)}$ to grow exponentially with σ. The next Lemma and Theorem show that this is not the case; a bound is obtained for $\mathcal{G}_j^{(\sigma)}$ which is independent of σ. Hence, the local error $\mathcal{L}^{(j)^t}(z)$ introduced at iteration j and propagated to iteration $\sigma+1$ by $\mathcal{G}_j^{(\sigma)}$ does not grow with σ. Thus, in this sense, the error grows additively; that is, $\delta T^{(\sigma+1)^t}(z)$ is bounded by the sum of the bounds of the local errors at each iteration j.

LEMMA 10. *If μ is so small and $\delta T^{(\sigma)^t}(z)$ and $\delta T^{*(\sigma)}(z)$ are not too large so that*

$$\begin{aligned}\kappa^{(\sigma)} \cdot |a_0^{(0)}| \cdot \Big\{ &\|a_0^{-1}(z) \,(mod\, z^{\|n\|+1})\| \left[(k+1)\|\delta T^{(\sigma)^t}(z)\| + \|\delta T^{*(\sigma)}(z)\| \right] \\ &+\; 1.01(k+1)(\|\nu^{(\sigma)}\| + k + 1) \cdot \mu \Big\} \le \frac{1}{2},\end{aligned}$$

then

$$\|\widehat{S}^{(\sigma)}(z)\| \le 2\kappa^{(\sigma)} \cdot (k+1) \cdot |a_0^{(0)}|.$$

Proof. ¿From (38),

$$\begin{aligned}\|(\Gamma^{*(\sigma)}\cdot\Gamma^{(\sigma)})^{-1}\cdot S^{*(\sigma)}(z)\cdot S^{(\sigma+1)}(z)\| &\leq \|(\Gamma^{*(\sigma)}\cdot\Gamma^{(\sigma)})^{-1}\|\cdot\|S^{*(\sigma)}(z)\|\cdot\|S^{(\sigma+1)}(z)\|\\ &\leq \kappa^{(\sigma)}\cdot(k+1).\end{aligned}$$

But, using Lemma 7 and Theorem 1 (adjusted to apply at the point $m^{(\sigma)}$ rather than at n)

$$\begin{aligned}&\|(\Gamma^{*(\sigma)}\cdot\Gamma^{(\sigma)})^{-1}\cdot S^{*(\sigma)}(z)\cdot S^{(\sigma+1)}(z)\|\\ &= \|(\Gamma^{*(\sigma)}\cdot\Gamma^{(\sigma)})^{-1}\cdot S^{*(\sigma)}(z)\cdot\Big\{S^{(\sigma)}(z)\cdot\widehat{S}^{(\sigma)}(z)+\theta_{IV}^{(\sigma)}(z)\Big\}\|\\ &= \|(\Gamma^{*(\sigma)}\cdot\Gamma^{(\sigma)})^{-1}\cdot\Big\{(z^{\|m^{(\sigma)}\|+1}\cdot(a_0^{(0)})^{-1}\cdot\Gamma^{*(\sigma)}\cdot\Gamma^{(\sigma)}+\theta_I^{(\sigma)}(z)\Big\}\cdot\widehat{S}^{(\sigma)}(z)\\ &\quad + (\Gamma^{*(\sigma)}\cdot\Gamma^{(\sigma)})^{-1}\cdot S^{*(\sigma)}(z)\cdot\theta_{IV}^{(\sigma)}(z)\|\\ &\geq |a_0^{(0)}|^{-1}\cdot\|\widehat{S}^{(\sigma)}(z)\|\\ &\quad - \|(\Gamma^{*(\sigma)}\cdot\Gamma^{(\sigma)})^{-1}\|\cdot\|a_0^{-1}(z)\,(mod\,z^{\|n\|+1})\|\cdot\Big[(k+1)\|\delta T^{(\sigma)^t}(z)\| + \|\delta T^{*(\sigma)}(z)\|\Big]\cdot\|\widehat{S}^{(\sigma)}(z)\|\\ &\quad - \|(\Gamma^{*(\sigma)}\cdot\Gamma^{(\sigma)})^{-1}\|\cdot\Big\{1.01\cdot(\|\nu^{(\sigma)}\|+k+1)\Big\}\cdot\|S^{(\sigma)}(z)\|\cdot\|\widehat{S}^{(\sigma)}(z)\|\cdot\|S^{*(\sigma)}(z)\|\cdot\mu\\ &\geq \|\widehat{S}^{(\sigma)}(z)\|\cdot\Big\{|a_0^{(0)}|^{-1}-\kappa^{(\sigma)}\|a_0^{-1}(z)\,(mod\,z^{\|n\|+1})\|\\ &\quad\cdot\Big[(k+1)\|\delta T^{(\sigma)^t}(z)\| + \|\delta T^{*(\sigma)}(z)\|\Big] - 1.01\,\kappa^{(\sigma)}\cdot(\|\nu^{(\sigma)}\|+k+1)\cdot(k+1)\cdot\mu\Big\}\\ &\geq |a_0^{(0)}|^{-1}\cdot\|\widehat{S}^{(\sigma)}(z)\|/2.\end{aligned}$$

The result now follows. ∎

THEOREM 11. *If μ is so small and $\delta T^{(j)^t}(z)$ and $\delta T^{*(j)}(z)$ are not too large so that*

$$\begin{aligned}\kappa^{(\sigma)}\cdot|a_0^{(0)}|\cdot\Big\{\|a_0^{-1}(z)\,(mod\,z^{\|n\|+1})\|\Big[(k+1)\|\delta T^{(j)^t}(z)\|+\|\delta T^{*(j)}(z)\|\Big]\\ + 1.01(k+1)(\|\nu^{(\sigma)}\|+k+1)\cdot\mu\Big\}\leq\frac{1}{2},\end{aligned}$$

then

$$\|\mathcal{G}_{j-1}^{(\sigma)}(z)\|\leq 2\kappa^{(j)}\cdot(k+1)\cdot|a_0^{(0)}|+O(\mu),\quad j\leq\sigma.$$

Proof. ¿From (43) and from Lemma 7

$$S^{(\sigma+1)}(z)=S^{(j)}(z)\cdot\mathcal{G}_{j-1}^{(\sigma)}(z)+\sum_{\ell=j}^{\sigma}\theta_{IV}^{(\ell)}(z)\cdot\mathcal{G}_{\ell}^{(\sigma)}(z).$$

We proceed by induction. Assume the theorem is true for $\mathcal{G}_{\sigma-1}^{(\sigma)}(z)$, $\mathcal{G}_{\sigma-2}^{(\sigma)}(z)$, $\cdots$, $\mathcal{G}_{j}^{(\sigma)}(z)$ (the initial case, $j=\sigma-1$, is proved in Lemma 10 since $\mathcal{G}_{\sigma-1}^{(\sigma)}(z)=\widehat{S}^{(\sigma)}(z)$). From (38),

$$\|(\Gamma^{*(j)}\cdot\Gamma^{(j)})^{-1}S^{*(j)}(z)\cdot S^{(\sigma+1)}(z)\|\leq\kappa^{(j)}(k+1).$$

But, using Lemma 7, Theorem 1 and the inductive hypothesis,

$$\begin{aligned}
&\|(\Gamma^{*(j)}\cdot\Gamma^{(j)})^{-1}\cdot S^{*(j)}(z)\cdot S^{(\sigma+1)}(z)\| \\
&\quad\geq \|(\Gamma^{*(j)}\cdot\Gamma^{(j)})^{-1}\cdot S^{*(j)}(z)\cdot S^{(j)}(z)\cdot \mathcal{G}_{j-1}^{(\sigma)}(z) \\
&\qquad +\sum_{\ell=j}^{\sigma}(\Gamma^{*(j)}\cdot\Gamma^{(j)})^{-1}\cdot S^{*(j)}(z)\cdot\theta_{IV}^{(\ell)}(z)\cdot\mathcal{G}_{\ell}^{(\sigma)}(z)\| \\
&\quad\geq \|(\Gamma^{*(j)}\cdot\Gamma^{(j)})^{-1}\cdot\left\{z^{\|m^{(j)}\|+1}(a_0^{(0)})^{-1}\Gamma^{*(j)}\cdot\Gamma^{(j)}+\theta_I^{(j)}(z)\right\}\cdot\mathcal{G}_{j-1}^{(\sigma)}(z)\| \\
&\qquad -\kappa^{(j)}\sum_{\ell=j}^{\sigma}\{k+1\}\cdot\left\{2.02\kappa^{(\ell)}\cdot(k+1)\cdot(\|\nu^{(\ell)}\|+k+1)\cdot|a_0^{(0)}|\cdot\mu\right\}\cdot \\
&\qquad\qquad\left\{2\kappa^{(\ell+1)}\cdot(k+1)\cdot|a_0^{(0)}|+O(\mu)\right\} \\
&\quad\geq \|\mathcal{G}_{j-1}^{(\sigma)}(z)\|\left\{|a_0^{(0)}|^{-1}-\kappa^{(j)}\|a_0^{-1}(z)\,(mod\, z^{\|n\|+1})\|\right. \\
&\qquad\qquad\left.\left[(k+1)\|\delta T^{(j)^t}(z)\|+\|\delta T^{*(j)}(z)\|\right]\right\}-O(\mu) \\
&\quad\geq |a_0^{(0)}|^{-1}\|\mathcal{G}_{j-1}^{(\sigma)}(z)\|/2-O(\mu).
\end{aligned}$$

∎

In the above theorem, we have taken the liberty of replacing a summation involving terms linear in μ with an $O(\mu)$ expression. We could have left the summation in explicitly, but, as we shall see, this summation becomes quadratic in μ when it is used to obtain a bound on $\delta T^{(\sigma)^t}(z)$.

Finally, we can give the bound on the residual error.

THEOREM 12. *If μ is so small and $\delta T^{(j)^t}(z)$ and $\delta T^{*(j)}(z)$ are not too large so that*

$$(\|n\|+k+1)\mu\leq 0.01$$

and

$$\begin{aligned}
\kappa^{(j)}\cdot|a_0^{(0)}|\cdot\Big\{&\|a_0^{-1}(z)\,(mod\, z^{\|n\|+1})\|\left[(k+1)\|\delta T^{(j)^t}(z)\|+\|\delta T^{*(j)}(z)\|\right] \\
&+\ 1.01(k+1)(\|\nu^{(j)}\|+k+1)\cdot\mu\Big\}\leq\frac{1}{2},\ j\leq\sigma,
\end{aligned}$$

then

$$\|\delta T^{(\sigma+1)^t}(z)\|\leq F_\sigma+2(k+1)\cdot|a_0^{(0)}|\sum_{j=0}^{\sigma-1}\kappa^{(j+1)}F_j, \tag{44}$$

where

$$\begin{aligned}
F_j &= 4\kappa^{(j)}(k+1)\cdot|a_0^{(0)}|\cdot\mu \\
&\quad\cdot\left\{(\|m^{(j)}\|+k+1)+4\rho_j\|\nu^{(j)}\|^3+(\|\nu^{(j)}\|+k+1)\right\}
\end{aligned} \tag{45}$$

and ρ_j is the growth factor associated with the LU-decomposition of $\widehat{\mathcal{M}}_{\nu^{(j)}}$ by Gaussian elimination.

Proof. To simplify the analysis, we now split the local error $\mathcal{L}^{(\sigma)^t}(z)$ into three parts and analyze the propagation of each part separately. Let

$$\mathcal{L}_1^{(\sigma)^t}(z) = \begin{cases} 0, & \sigma = 0, \\ -z^{\|m^{(\sigma)}\|+1}\ \theta_{II}^{(\sigma)^t}(z)\ \widehat{S}^{(\sigma)}(z)\ (\bmod z^{\|m^{(\sigma+1)}\|+1}), & \sigma \geq 1, \end{cases} \tag{46}$$

$$\mathcal{L}_2^{(\sigma)^t}(z) = z^{\|m^{(\sigma)}\|+1}\theta_{III}^{(\sigma)^t}(z)\ (\bmod z^{\|m^{(\sigma+1)}\|+1}), \qquad \sigma \geq 0, \tag{47}$$

$$\mathcal{L}_3^{(\sigma)^t}(z) = \begin{cases} 0, & \sigma = 0, \\ A^t(z)\theta_{IV}^{(\sigma)}(z)\ (\bmod z^{\|m^{(\sigma+1)}\|+1}), & \sigma \geq 1, \end{cases} \tag{48}$$

and define

$$\mathcal{E}_i^{(\sigma+1)}(z) = \sum_{j=0}^{\sigma} \mathcal{L}_i^{(j)^t}(z) \cdot \mathcal{G}_j^{(\sigma)}(z), \qquad i = 1, 2, 3. \tag{49}$$

Then, according to Lemma 8 and Theorem 9,

$$\delta T^{(\sigma+1)^t}(z) = \sum_{i=1}^{3} \mathcal{E}_i^{(\sigma+1)^t}(z)\ .$$

We now bound $\mathcal{E}_i^{(\sigma+1)^t}(z),\ 1 \leq i \leq 3$.

¿From (46) and (49), from Lemmas 5 and 10 and from Theorem 11,

$$\begin{aligned} \|\mathcal{E}_1^{(\sigma+1)^t}(z)\| &= \|\sum_{j=0}^{\sigma} \mathcal{L}_1^{(j)^t}(z) \cdot \mathcal{G}_j^{(\sigma)}(z)\| \\ &\leq \|\theta_{II}^{(\sigma)^t}(z)\| \cdot \|\widehat{S}^{(\sigma)}(z)\| + \sum_{j=0}^{\sigma-1} \|\theta_{II}^{(j)^t}(z)\| \cdot \|\widehat{S}^{(j)}(z)\| \cdot \|\mathcal{G}_j^{(\sigma)}(z)\| \qquad (50) \\ &\leq \Big\{1.01(\|m^{(\sigma)}\| + k + 1) \cdot \mu\Big\} \Big\{2\kappa^{(\sigma)}(k+1)|a_0^{(0)}|\Big\} \\ &\quad + \sum_{j=0}^{\sigma-1} \Big\{1.01(\|m^{(j)}\| + k + 1)\mu\Big\} \cdot \Big\{2\kappa^{(j)}(k+1)|a_0^{(0)}|\Big\} \\ &\qquad \cdot \Big\{2\kappa^{(j+1)}(k+1)|a_0^{(0)}| + O(\mu)\Big\} \\ &\leq 4\kappa^{(\sigma)} \cdot (k+1) \cdot (\|m^{(\sigma)}\| + k + 1) \cdot |a_0^{(0)}| \cdot \mu \\ &\quad + \ 8(k+1)^2 \cdot |a_0^{(0)}|^2 \cdot \mu \sum_{j=0}^{\sigma-1} \kappa^{(j)} \cdot \kappa^{(j+1)} \cdot (\|m^{(j)}\| + k + 1) \\ &\quad + \ O(\mu^2). \end{aligned}$$

¿From (47) and (49), from Lemmas 6 and 10 and from Theorem 11,

$$\begin{aligned} \|\mathcal{E}_2^{(\sigma+1)^t}(z)\| &= \|\sum_{j=0}^{\sigma} \mathcal{L}_2^{(j)^t}(z)\mathcal{G}_j^{(\sigma)}(z)\| \\ &\leq \|\theta_{III}^{(\sigma)^t}(z)\| + \sum_{j=0}^{\sigma-1} \|\theta_{III}^{(j)^t}(z)\| \cdot \|\mathcal{G}_j^{(\sigma)}(z)\| \qquad (51) \end{aligned}$$

$$
\begin{aligned}
&\leq \Big\{8\|\nu^{(\sigma)}\|^3\cdot\rho_\sigma\cdot\mu+O(\mu^2)\Big\}\cdot\|\widehat{S}^{(\sigma)}(z)\| \\
&\quad+\sum_{j=0}^{\sigma-1}\Big\{8\|\nu^{(j)}\|^3\cdot\rho_j\cdot\mu+O(\mu^2)\Big\}\cdot\|\widehat{S}^{(j)}(z)\|\cdot\|\mathcal{G}_j^{(\sigma)}(z)\| \\
&\leq \Big\{8\|\nu^{(\sigma)}\|^3\cdot\rho_\sigma\cdot\mu+O(\mu^2)\Big\}\cdot\Big\{2\kappa^{(\sigma)}(k+1)|a_0^{(0)}|\Big\} \\
&\quad+\sum_{j=0}^{\sigma-1}\Big\{8\|\nu^{(j)}\|^3\cdot\rho_j\cdot\mu+O(\mu^2)\Big\}\cdot\Big\{2\kappa^{(j)}(k+1)|a_0^{(0)}|\Big\} \\
&\qquad\qquad\cdot\Big\{2\kappa^{(j+1)}(k+1)|a_0^{(0)}|+O(\mu)\Big\}. \\
&\leq 16\cdot\kappa^{(\sigma)}\cdot(k+1)\|\nu^{(\sigma)}\|^3\cdot\rho_\sigma\cdot|a_0^{(0)}|\cdot\mu \\
&\quad+32(k+1)^2\cdot|a_0^{(0)}|^2\sum_{j=0}^{\sigma-1}\kappa^{(j)}\cdot\kappa^{(j+1)}\cdot\rho_j\cdot\|\nu^{(j)}\|^3\cdot\mu \\
&\quad+O(\mu^2).
\end{aligned}
$$

¿From (48) and (49), from Lemmas 7 and 10 and from Theorem 11,

$$
\begin{aligned}
\|\mathcal{E}_3^{(\sigma+1)^t}(z)\| &= \|\sum_{j=0}^{\sigma}\mathcal{L}_3^{(j)^t}(z)\cdot\mathcal{G}_j^{(\sigma)}(z)\| \\
(52)\qquad &\leq \|A^t(z)\cdot\theta_{IV}^{(\sigma)}(z)\|+\sum_{j=0}^{\sigma-1}\|A^t(z)\cdot\theta_{IV}^{(j)}(z)\|\cdot\|\mathcal{G}_j^{(\sigma)}(z)\| \\
&\leq 1.01(\|\nu^{(\sigma)}\|+k+1)\cdot\|\widehat{S}^{(\sigma)}(z)\|\cdot\mu \\
&\quad+\sum_{j=0}^{\sigma-1}\Big\{1.01(\|\nu^{(j)}\|+k+1)\cdot\mu\Big\}\cdot\|\widehat{S}^{(j)}(z)\|\cdot\|\mathcal{G}_j^{(\sigma)}(z)\| \\
&\leq \Big\{1.01\cdot(\|\nu^{(\sigma)}\|+k+1)\cdot\mu\Big\}\cdot\Big\{2\kappa^{(\sigma)}(k+1)|a_0^{(0)}|\Big\} \\
&\quad+\sum_{j=0}^{\sigma-1}\Big\{1.01(\|\nu^{(j)}\|+k+1)\cdot\mu\Big\}\cdot\Big\{2\kappa^{(j)}(k+1)|a_0^{(0)}|\Big\} \\
&\qquad\qquad\cdot\Big\{2\kappa^{(j+1)}(k+1)|a_0^{(0)}|+O(\mu)\Big\} \\
&\leq 4\kappa^{(\sigma)}\cdot(k+1)\cdot(\|\nu^{(\sigma)}\|+k+1)\cdot|a_0^{(0)}|\cdot\mu \\
&\quad+8(k+1)^2\cdot|a_0^{(0)}|^2\cdot\mu\sum_{j=0}^{\sigma-1}\kappa^{(j)}\kappa^{(j+1)}(\|\nu^{(j)}\|+k+1) \\
&\quad+O(\mu^2).
\end{aligned}
$$

The result follows by summing (50), (51) and (52). ∎

In Theorem 12 above, the bound for $\delta T^{(\sigma+1)^t}(z)$ involves the products $\kappa^{(j)}\kappa^{(j+1)}$. These result from inequalities involving the expression $\|\widehat{S}^{(j)}(z)\|\cdot\|\mathcal{G}_j^{(\sigma)}(z)\|$. However, it is seen that $\widehat{S}^{(j)}(z)\cdot\mathcal{G}_j^{(\sigma)}(z)=\mathcal{G}_{j-1}^{(\sigma)}(z)$, so it is felt that the inequalities are crude and the bounds should just involve a single $\kappa^{(j)}$. Experimental results [10] support this conjecture.

This completes the analysis of the error in the order condition for computing a NPHS. Proceeding in an analogous manner we can obtain the following theorem which gives bounds for the error in the order condition for the NSPS computed by PHS_SPS.

THEOREM 13. *If the conditions of Theorem 12 are satisfied, then*

$$\|\delta T^{*(\sigma+1)}(z)\| \leq F_\sigma^* + 2(k+1)\cdot|a_0^{(0)}|\sum_{j=0}^{\sigma-1}\kappa^{(j+1)}F_j^*, \tag{53}$$

where

$$F_j^* = 8\kappa^{(j)}(k+1)^2\cdot|a_0^{(0)}|\cdot\mu \left\{(\|m^{(j)}\|+1)+4(k+1)^5\rho_j^*\|\nu^{(j)}\|^3+(\|\nu^{(j)}\|+k+1)\right\} \tag{54}$$

and ρ_j^* is the growth factor associated with the LU-decomposition of $\widehat{\mathcal{M}}^*_{\nu^{(j)}}$ by Gaussian elimination.

Proof. See [9]. ∎

Theorems 12 and 13 assure us that if $\|\delta T^{(\sigma)^t}(z)\|$ and $\delta T^{*(\sigma)}(z)$ are small and $\kappa^{(\sigma)}$ is not too large, then $\|\delta T^{(\sigma+1)^t}(z)\|$ and $\delta T^{*(\sigma+1)}(z)$ will also be small. Thus, $\|\delta T^{(\sigma)^t}(z)\|$ and $\delta T^{*(\sigma)}(z)$ will remain small for all σ as long as, at every iteration j, a step $\nu^{(j)}$ is chosen (stepping over unstable blocks) so that $\kappa^{(j)}$ is not too large. Consequently, the assumptions of Theorems 12 and 13 are satisfied in practice.

6. Stability. In this section, bounds for the errors $\delta S(z) = S(z) - S_E(z)$ and $\delta S^*(z) = S^*(z) - S_E^*(z)$ are obtained. Since $S(z)$ and $S^*(z)$ are scaled, these same bounds serve also as bounds for the relative errors in $S(z)$ and $S^*(z)$. To make the comparisons meaningful in the above, we insist that $S_E(z)$ and $S_E^*(z)$ are such that

$$\begin{aligned} V_E(0) &= V(0) = diag[\gamma_1,\ldots,\gamma_k], \\ r_E(0) &= r(0) = \gamma_0, \end{aligned}$$

and

$$\begin{aligned} v_E^*(0) &= v^*(0) = \gamma_0^*, \\ R_E^*(0) &= R^*(0) = diag[\gamma_1^*,\ldots,\gamma_k^*]. \end{aligned}$$

We begin by first finding bounds for $\delta S(z)$. From (6) and (10)

$$A^t(z)\cdot\delta S(z) = \delta T^t(z) + \mathcal{O}(z^{\|n\|+1}).$$

So, the constant terms[7] $\delta u_\beta^{(0)}$ and $\delta v_{\alpha,\beta}^{(0)}$ for $0 \leq \alpha,\beta \leq k$ of $S(z)$ are zero. It then follows that the remaining components of $\delta S(z)$ satisfy

$$\mathcal{M}_n\cdot\delta\mathcal{X} = [\delta r^{(0)},\ldots,\delta r^{(\|n\|-1)}]^t, \tag{55}$$

where

$$\delta\mathcal{X} = \left[\delta p^{(0)},\ldots,\delta p^{(n_0-1)}|\delta q_1^{(0)},\ldots,\delta q_1^{(n_1-1)}|\cdots|\delta q_k^{(0)},\ldots,\delta q_k^{(n_k-1)}\right]^t,$$

[7] In actual fact, the computations in (15) may yield errors resulting in nonzero values of $\delta u_\beta^{(0)}$ for $1 \leq \beta \leq k$. But, these errors, each resulting from two floating-point operations, are comparatively small and are ignored in order to simplify the analysis.

and

$$\mathcal{M}_n \cdot \delta\mathcal{Y} = \begin{bmatrix} \delta w_1^{(1)} & \cdots & \delta w_k^{(1)} \\ \vdots & & \vdots \\ \delta w_1^{(\|n\|)} & \cdots & \delta w_k^{(\|n\|)} \end{bmatrix}, \tag{56}$$

where

$$\delta\mathcal{Y} = \left[\begin{array}{ccc|ccc|c|ccc} \delta u_1^{(1)} & \cdots & \delta u_1^{(n_0)} & \delta v_{1,1}^{(1)} & \cdots & \delta v_{1,1}^{(n_1)} & \cdots & \delta v_{k,1}^{(1)} & \cdots & \delta v_{k,1}^{(n_k)} \\ \vdots & & \vdots & \vdots & & \vdots & & \vdots & & \vdots \\ \delta u_k^{(1)} & \cdots & \delta u_k^{(n_0)} & \delta v_{1,k}^{(1)} & \cdots & \delta v_{1,k}^{(n_1)} & & \delta v_{k,k}^{(1)} & \cdots & \delta v_{k,k}^{(n_k)} \end{array}\right]^t.$$

¿From (55) and (56), it follows that

$$\begin{aligned} \|\delta S(z)\| &\leq \max\{\|\delta\mathcal{X}\|_1, \|\delta\mathcal{Y}\|_1\} \\ &\leq \|\mathcal{M}_n^{-1}\|_1 \cdot \max\left\{\|\delta r(z)\|, \|\delta W^t(z)\|\right\} \\ &\leq \|\mathcal{M}_n^{-1}\|_1 \cdot \|\delta T^t(z)\|. \end{aligned} \tag{57}$$

Thus, to obtain a bound for $\delta S(z)$, we need only to obtain bounds for $\mathcal{M}_n^{-1}$ and $\delta T^t(z)$. This is done formally in Theorem 15 below. But first, in a similar fashion, we show that bounds for $\delta S^*(z)$ can be expressed in terms of bounds for $\mathcal{M}_n^{*-1}$ and $\delta T^*(z)$.

¿From (22) and (26)

$$S^*(z)A^*(z) = \delta T^*(z) + \mathcal{O}(z^{\|n\|+1}).$$

As for the NSPS, for the sake of simplicity, here again we ignore that the constant term errors, $\delta w_\beta^{*(0)}$, for $1 \leq \beta \leq k$. This is done with no great loss of generality, since these are the comparatively small errors made in computing $\delta u_\beta^{*(0)}(z)$ from

$$u_\beta^{*(0)} a_0^{(0)} + v^{*(0)} a_\beta^{(0)} = 0$$

with $v^{*(0)} = \gamma_0^*$. It then follows, in a fashion similar to solving (30) and (32), that the remaining components of $\delta S^*(z)$ satisfy

$$\delta\mathcal{X}^{*t} \cdot \mathcal{M}_n^* = [\delta w_1^{*(1)}, \ldots, \delta w_1^{*(\|n\|)} | \cdots | \delta w_k^{*(1)}, \ldots, \delta w_k^{*(\|n\|)}], \tag{58}$$

where

$$\delta\mathcal{X}^{*t} = \left[\delta v^{*(1)}, \ldots, \delta v^{*(\|n\|-n_0)} | \delta u_1^{*(1)}, \ldots, \delta u_1^{*(\|n\|-n_1)} | \cdots | \delta u_k^{*(1)}, \ldots, \delta u_k^{*(\|n\|-n_k)}\right],$$

and, for $1 \leq \alpha \leq k$,

$$\delta\mathcal{Y}_\alpha^{*t} \cdot \mathcal{M}_n^* = [\delta r_{\alpha,1}^{*(0)}, \ldots, \delta r_{\alpha,1}^{*(\|n\|-1)} | \cdots | \delta r_{\alpha,k}^{*(0)}, \ldots, \delta r_{\alpha,k}^{*(\|n\|-1)}], \tag{59}$$

where

$$\delta\mathcal{Y}_\alpha^{*t} = \left[\delta q_\alpha^{*(0)}, \ldots, \delta q_\alpha^{*(\|n\|-n_0-1)} | \delta p_{\alpha,1}^{*(0)}, \ldots, \delta p_{\alpha,1}^{*(\|n\|-n_1-1)} | \cdots | \delta p_{\alpha,k}^{*(0)}, \ldots, \delta p_{\alpha,k}^{*(\|n\|-n_k-1)}\right].$$

¿From (58) and (59), we get

$$\begin{aligned} \|\delta S^*(z)\| &\leq (k+1) \max_{1\leq\alpha\leq k} \{\|\delta\mathcal{X}^*\|_1, \|\delta\mathcal{Y}^*_\alpha\|_1\} \\ &\leq (k+1)^2 \|\mathcal{M}_n^{*-1}\|_\infty \cdot \|\delta T^*(z)\|. \end{aligned} \tag{60}$$

We are now ready to give the main results of this paper in the two theorems below; the first theorem shows that the algorithm PHS_SPS is weakly stable, whereas the second provides bounds for the errors $\delta S(z)$ and $\delta S^*(z)$. But, first note some notational details. Let $\delta T^t(z)$ and $\delta T^*(z)$ denote the residual errors corresponding, respectively, to the NPHS and NSPS computed by the algorithm PHS_SPS in $\sigma+1$ steps. So, $n = m^{(\sigma+1)}$ and a bound for $\|\delta T^t(z)\|$ is given by Theorem 12 in which $\delta T^{(\sigma+1)^t}(z) = \delta T^t(z)$ and a bound for $\|\delta T^*(z)\|$ is given by Theorem 13 in which $\delta T^{*(\sigma+1)}(z) = \delta T^*(z)$. At the point $m^{(\sigma+1)}$, we drop the superscript $\sigma+1$ so that $\kappa = \kappa^{(\sigma+1)}$, $S(z) = S^{(\sigma+1)}(z)$, $S^*(z) = S^{*(\sigma+1)}(z)$ and so on. The point $m^{(\sigma)}$ is the last stable point (i.e., $\kappa^{(\sigma)} \leq \tau$) prior to the point n along the diagonal passing through n. The point n itself need not be stable.

THEOREM 14. *The algorithm PHS_SPS for computing $S(z)$ and $S^*(z)$ is weakly stable.*

Proof. ¿From (44), (53), (57) and (60), it follows that, if the problem is well-conditioned (i.e., if the condition number κ associated with the matrices $\mathcal{M}_n$ and $\mathcal{M}_n^*$ is not too large), then the computed solution $S(z)$ is close to the exact solution $S_E(z)$ and $S^*(z)$ is close to the exact solution $S_E^*(z)$. The algorithm is therefore weakly stable [7]. ∎

Note that the bounds (44) and (53) for the residual errors $\delta T^t(z)$ and $\delta T^*(z)$ (and therefore also the weak stability of PHS_SPS) do not depend on $a_0^{-1}(z)$. So $\kappa^{(j)}$ defined by (38) (i.e., excluding the term $\|a_0^{-1}(z) \ (mod\ z^{\|n\|+1})\|$ that appears in the bounds for $\mathcal{M}_n^{-1}$ and $\mathcal{M}_n^{*-1}$ [11]) is an appropriate choice for a stability parameter. Bounds for the errors $\delta S(z)$ and $\delta S^*(z)$ in the solutions, given in Theorem (15) do, however, depend on $a_0^{-1}(z)$.

THEOREM 15. *If κ is not too large and $\delta T^t(z)$ and $\delta T^*(z)$ are sufficiently small,*[8] *then*

$$\|\delta S(z)\| \leq 2\kappa \cdot |a_0^{(0)}| \cdot \|a_0^{-1}(z) \ (mod\, z^{\|n\|+1})\| \left\{ \bar{F}_\sigma + 2\tau(k+1) \cdot |a_0^{(0)}| \cdot \sum_{j=0}^{\sigma-1} \bar{F}_j \right\},$$

where

$$\bar{F}_j = 4\tau(k+1) \cdot |a_0^{(0)}| \cdot \mu \left\{ (\|m^{(j)}\| + k + 1) + 4\rho_j \|\nu^{(j)}\|^3 + (\|\nu^{(j)}\| + k + 1) \right\}$$

and

$$\|\delta S^*(z)\| \leq 2\kappa(k+1)^2 \cdot |a_0^{(0)}| \cdot \|a_0^{-1}(z) \ (mod\, z^{\|n\|+1})\| \left\{ \bar{F}^*_\sigma + 2\tau(k+1) \cdot |a_0^{(0)}| \sum_{j=0}^{\sigma-1} \bar{F}^*_j \right\},$$

[8] In addition to satisfying the assumptions of Theorems 12 at all the stable points $m^{(j)}$, $1 \leq j \leq \sigma$, at the final point $n = m^{(\sigma+1)}$, we require $\delta T^t(z)$ and $\delta T^*(z)$ to be sufficiently small so that

$$\left[(\kappa+1)(k+2)|a_0^{(0)}|(\|a_0^{-1}(z) \ (mod\, z^{\|n\|+1})\| + 1) \right]^2 \left[(k+2)\|\delta T^t(z)\| + \|\delta T^*(z)\| \right] \leq 1/8.$$

This assumption at the last point n is used in [11] in obtaining bounds for $\mathcal{M}_n^{-1}$ and $\mathcal{M}_n^{*-1}$. All these assumptions are easily satisfied if all the points, including the last one, are reasonably stable.

where

$$\bar{F}_j^* = 8\tau(k+1)^2 \cdot |a_0^{(0)}| \cdot \mu \left\{ (\|m^{(j)}\| + 1) + 4(k+1)^5 \rho_j^* \|\nu^{(j)}\|^3 + (\|\nu^{(j)}\| + k + 1) \right\}.$$

Proof. For κ not too large and $\delta T^t(z)$ and $\delta T^*(z)$ sufficiently small, bounds for $\mathcal{M}_n^{-1}$ and $\mathcal{M}_n^{*-1}$ are derived in [11] to be

$$\|\mathcal{M}_n^{-1}\|_1,\ \|\mathcal{M}_n^{*-1}\|_\infty \leq 2\kappa \cdot |a_0^{(0)}| \cdot \|a_0^{-1}(z)\,(mod\, z^{\|n\|+1})\|.$$

The results of the theorem now follows from (57) and (60) using (44) and (53). ∎

7. Experimental Results. Numerical experiments have been performed to compare the analysis of the algorithm with its practice. A summary of the conclusions is presented here; details appear in [10].

The algorithm PHS_SPS was implemented using Sun Fortran 1.3.1. All calculations were performed in double precision. The linear systems (13), (17), (30) and (32) arising at intermediate steps of the algorithm were solved using the LINPACK routines SGEFA and SGESL. The results were then compared to the exact answers, obtained via the Maple computer algebra system.

Tables A1 and A2 give the results of a small but typical experiment for which $n = (18, 19, 19)$ and $A^t(z) = [a_0(z), a_1(z), a_2(z)]$ with $a_0(z) = 1$ and with coefficients of $a_1(z)$, $a_2(z)$ randomly and uniformly distributed between -1 and 1 and then scaled. The tables give results at all intermediate points along the diagonal through n. In these tables, the errors (represented in scientific notation with two digits of accuracy and the exponent enclosed in parenthesis) in the computed $S^{(j)}(z)$ and $S^{*(j)}$ and in the order conditions are given for two values of the stability parameter τ. The value $\tau = 10^4$ in Table A1 indicates a willingness to accept only those striped Sylvester matrices $\mathcal{M}_{m^{(j)}}$ and mosaic Sylvester matrices $\mathcal{M}^*_{m^{(j)}}$ with condition numbers less than 10^4, approximately (i.e., those for which $\kappa^{(j)} \leq 10^4$). Striped and mosaic Sylvester matrices not satisfying this criterion are assumed to lie in an unstable block and are skipped over. An unstable point is identified by the value "-" in the column labeled "j". In Table A2, the value $\tau = 10^9$ permits a much greater tolerance for ill-conditioning and results in an expected deterioration in the accuracy.

Tables B1 and B2 give the results of a similar experiment but for which $a_0(z)$, $a_1(z)$ and $a_2(z)$ were all first randomly generated (except that $a_0^{(0)}$ is initially set to 1) and then modified so as to introduce some pronounced instabilities. To introduce an instability at $m^{(j+1)}$, the coefficients of $a_1(z)$ and $a_2(z)$ were changed to make almost dependent the columns of coefficient matrix $\widehat{\mathcal{M}}_\nu$ corresponding to the residual $T^{(j)^t}(z)$ at the point $m^{(j)}$. The power series were then scaled. For this particular experiment, $\|a_0^{-1}(z)\ (mod\ z^{\|n\|+1})\| = 2.3 \times 10^2$, approximately.

It was observed that the large powers of k that occur in the bounds derived above are not manifested in the experiments. Also, $\|\delta T^t(z)\|$ and $\|\delta T^*(z)\|$ appear to depend on $\kappa^{(j)}$ and not $\kappa^{(j)}\kappa^{(j+1)}$ and the overall error is proportional to the largest $\kappa^{(j)}$ encountered. Thus, the bounds are crude, but they do appear to reflect the behavior of the error. As Wilkinson points out [29, page 567], "The main object of such an analysis is to expose the potential instabilities, if any, of an algorithm so that hopefully from the insight thus obtained one might be led to improved algorithms. Usually the bound itself is weaker than it might have been because of the necessity

Table A1: $a_0(z) = 1$
Errors at intermediate steps: $\tau = 10^4$

j	$\kappa^{(j)}$	$\|\delta {T^{(j)}}^t(z)\|$	$\frac{\|\delta S^{(j)}(z)\|}{\|S_E^{(j)}(z)\|}$	$\|\delta T^{*(j)}(z)\|$	$\frac{\|\delta S^{*(j)}(z)\|}{\|S_E^{*(j)}(z)\|}$
1	1.2(2)	1.7(-18)	1.4(-16)	2.2(-18)	7.0(-17)
2	1.8(2)	1.0(-17)	6.5(-16)	2.0(-17)	6.5(-16)
3	1.6(2)	1.7(-17)	9.8(-16)	3.3(-17)	1.8(-15)
4	9.5(2)	1.6(-17)	8.3(-16)	6.6(-17)	2.0(-15)
5	6.6(2)	2.0(-17)	1.7(-15)	9.0(-17)	2.9(-15)
-	4.1(7)	2.3(-17)	1.2(-15)	8.7(-17)	2.1(-15)
6	1.1(3)	3.3(-17)	2.3(-15)	1.3(-16)	4.8(-15)
7	1.5(3)	3.6(-17)	1.2(-15)	1.2(-16)	6.6(-15)
8	9.1(3)	5.6(-17)	1.8(-15)	1.9(-16)	4.4(-15)
9	3.7(3)	8.2(-17)	4.9(-15)	2.2(-16)	1.3(-14)
10	2.9(3)	1.2(-16)	3.3(-15)	3.9(-16)	1.2(-14)
-	3.2(6)	7.7(-17)	5.6(-15)	5.7(-16)	4.6(-14)
11	2.0(3)	2.8(-16)	8.1(-15)	5.6(-16)	1.4(-14)
-	1.6(4)	2.8(-16)	7.4(-15)	4.5(-16)	1.8(-14)
12	2.9(3)	2.9(-16)	9.5(-15)	6.8(-16)	2.2(-14)
-	4.1(4)	2.5(-16)	1.0(-14)	7.5(-16)	2.3(-14)
-	6.3(4)	2.7(-15)	2.4(-14)	8.2(-16)	2.9(-14)
-	1.1(4)	2.3(-16)	1.7(-14)	8.9(-16)	3.3(-14)
-	1.1(5)	2.5(-16)	1.3(-13)	8.0(-16)	1.4(-13)

Table A2: $a_0(z) = 1$
Errors at intermediate steps: $\tau = 10^9$

j	$\kappa^{(j)}$	$\|\delta {T^{(j)}}^t(z)\|$	$\frac{\|\delta S^{(j)}(z)\|}{\|S_E^{(j)}(z)\|}$	$\|\delta T^{*(j)}(z)\|$	$\frac{\|\delta S^{*(j)}(z)\|}{\|S_E^{*(j)}(z)\|}$
1	1.2(2)	1.7(-18)	1.4(-16)	2.2(-18)	7.0(-17)
2	1.8(2)	1.0(-17)	6.5(-16)	2.0(-17)	6.5(-16)
3	1.6(2)	1.7(-17)	9.8(-16)	3.3(-17)	1.8(-15)
4	9.5(2)	1.6(-17)	8.3(-16)	6.6(-17)	2.0(-15)
5	6.6(2)	2.0(-17)	1.7(-15)	9.0(-17)	2.9(-15)
6	4.1(7)	2.3(-17)	1.2(-15)	8.7(-17)	2.1(-15)
7	1.1(3)	6.9(-13)	3.4(-11)	3.2(-12)	1.2(-10)
8	1.5(3)	6.8(-13)	1.9(-11)	3.7(-12)	1.5(-10)
9	9.1(3)	1.1(-12)	3.7(-11)	6.6(-12)	5.6(-10)
10	3.7(3)	1.6(-12)	9.5(-11)	6.4(-12)	3.7(-10)
11	2.9(3)	1.1(-12)	7.3(-11)	9.5(-12)	3.5(-10)
12	3.2(6)	1.2 -12)	1.7(-10)	1.2(-11)	1.9(-9)
13	2.0(3)	5.0(-12)	1.3(-10)	8.6(-12)	2.2(-10)
14	1.6(4)	4.9(-12)	1.4(-10)	8.3(-12)	1.9(-10)
15	2.9(3)	3.3(-12)	1.1(-10)	1.5(-11)	3.5(-10)
16	4.1(4)	3.6(-12)	1.1(-10)	9.8(-12)	3.5(-10)
17	6.3(4)	2.4(-12)	1.5(-10)	1.3(-11)	6.5(-10)
18	1.1(4)	2.8(-12)	1.8(-10)	1.1(-11)	4.4(-10)
19	1.1(5)	3.7(-12)	2.2(-10)	1.3(-11)	8.1(-10)

Table B1: Random $a_0(z)$
Errors at intermediate steps: $\tau = 10^5$

j	$\kappa^{(j)}$	$\|\delta T^{(j)^t}(z)\|$	$\frac{\|\delta S^{(j)}(z)\|}{\|S_E^{(j)}(z)\|}$	$\|\delta T^{*(j)}(z)\|$	$\frac{\|\delta S^{*(j)}(z)\|}{\|S_E^{*(j)}(z)\|}$
1	3.2(0)	0.0	9.8(-17)	6.9(-18)	7.6(-17)
2	3.9(3)	1.5(-17)	7.1(-17)	1.7(-17)	4.7(-16)
3	3.7(3)	3.6(-17)	6.6(-16)	2.5(-17)	2.9(-15)
4	7.7(3)	1.0(-16)	5.7(-15)	3.6(-17)	2.7(-15)
-	6.4(14)	1.1(-16)	1.0(-14)	4.5(-17)	3.6(-10)
5	1.1(4)	9.3(-17)	1.5(-14)	5.7(-17)	8.4(-15)
-	3.8(5)	9.2(-17)	1.3(-14)	4.1(-16)	2.0(-14)
6	1.1(4)	1.1(-16)	8.5(-15)	4.2(-16)	2.2(-14)
-	1.3(14)	1.1(-16)	2.1(-14)	2.1(-16)	8.2(-10)
7	3.9(4)	1.2(-16)	7.7(-15)	4.2(-16)	3.5(-14)
-	3.8(8)	9.4(-17)	3.2(-11)	4.3(-16)	5.1(-10)
-	1.9(9)	8.9(-17)	1.7(-10)	4.1(-16)	7.1(-10)
-	1.1(15)	9.0(-17)	2.7(-10)	4.0(-16)	2.8(-9)
-	1.3(9)	9.2(-17)	1.9(-10)	4.5(-16)	1.0(-9)
-	2.1(5)	3.3(-16)	6.2(-14)	4.4(-16)	3.5(-14)
8	3.0(4)	3.2(-16)	6.9(-14)	4.2(-16)	4.9(-14)
-	1.4(13)	3.2(-16)	2.3(-13)	5.1(-16)	6.9(-10)
9	6.4(4)	5.4(-16)	7.6(-13)	6.0(-16)	2.1(-13)
-	2.3(5)	5.5(-16)	5.3(-13)	2.3(-15)	4.6(-13)

Table B2: Random $a_0(z)$
Errors at intermediate steps: $\tau = 10^9$

j	$\kappa^{(j)}$	$\|\delta T^{(j)^t}(z)\|$	$\frac{\|\delta S^{(j)}(z)\|}{\|S_E^{(j)}(z)\|}$	$\|\delta T^{*(j)}(z)\|$	$\frac{\|\delta S^{*(j)}(z)\|}{\|S_E^{*(j)}(z)\|}$
1	3.2(0)	0.0	9.8(-17)	6.9(-18)	7.6(-17)
2	3.9(3)	1.5(-17)	7.1(-17)	1.7(-17)	4.7(-16)
3	3.7(3)	3.6(-17)	6.6(-15)	2.5(-17)	2.9(-15)
4	7.7(3)	1.0(-16)	5.7(-15)	3.6(-17)	2.7(-15)
-	6.4(14)	1.1(-16)	1.0(-14)	4.5(-17)	3.6(-10)
5	1.1(4)	9.3(-17)	1.5(-14)	5.7(-17)	8.4(-15)
6	3.8(5)	9.2(-17)	1.3(-14)	4.1(-16)	2.0(-14)
7	1.1(4)	2.2(-16)	1.1(-14)	1.6(-15)	1.1(-13)
-	1.3(14)	1.1(-16)	1.1(-14)	6.7(-15)	7.7(-9)
8	3.9(4)	2.5(-16)	1.1(-14)	4.8(-15)	2.3(-13)
9	3.8(8)	1.7(-16)	1.6(-10)	6.0(-15)	4.1(-9)
-	1.9(9)	1.6(-16)	2.9(-10)	8.9(-15)	1.6(-8)
-	1.1(15)	1.1(-16)	1.0(-9)	8.2(-15)	4.1(-8)
-	1.3(9)	1.3(-16)	1.6(-10)	6.9(-15)	1.3(-8)
10	2.1(5)	1.3(-12)	1.9(-10)	2.2(-13)	2.1(-10)
11	3.0(4)	1.9(-11)	2.3(-9)	8.3(-13)	2.8(-10)
-	1.4(13)	7.2(-12)	1.1(-9)	1.6(-12)	1.4(-6)
12	6.4(4)	1.7(-11)	1.3(-9)	3.8(-12)	1.0(-9)
13	2.3(5)	3.4(-11)	1.1(-9)	2.1(-11)	3.7(-9)

of restricting the mass of detail to a reasonable level and because of the limitations imposed by expressing the errors in terms of matrix norms."

¿From these and other experiments [10], operational bounds on the errors in the order conditions (as for the case k=1 reported in [15]) appear to be

$$\|\delta T^t(z)\| \leq C(k+1)\mu \left(\sum_{j=0}^{\sigma} \kappa^{(j)} \rho_j \|m^{(j)}\| \right) + O(\mu^2)$$

and

$$\|\delta T^*(z)\| \leq C(k+1)^2\mu \left(\sum_{j=0}^{\sigma} \kappa^{(j)} \rho_j \|m^{(j)}\|^2 \right) + O(\mu^2),$$

where C is a moderate constant. In addition, for the errors in the solutions, operational bounds appear to be

$$\|\delta S(z)\| \leq C\kappa(k+1)\mu \left(\sum_{j=0}^{\sigma} \kappa^{(j)} \rho_j \|m^{(j)}\| \right) + O(\mu^2)$$

and

$$\|\delta S^*(z)\| \leq C\kappa(k+1)^3\mu \left(\sum_{j=0}^{\sigma} \kappa^{(j)} \rho_j \|m^{(j)}\|^2 \right) + O(\mu^2).$$

8. Conclusions. In this paper we have presented a new fast, weakly stable algorithm for the computation of Padé-Hermite and simultaneous Padé systems. The algorithm requires $\mathcal{O}(\|n\|^2+s^3\|n\|)$ operations to compute a Padé-Hermite system and a simultaneous Padé system of type $n = [n_0, \ldots, n_k]$, where $\|n\| = n_0+\cdots+n_k$ and s is the largest distance from one well-conditioned subproblem to the next. The algorithm can also be used for fast stable inversion of striped or mosaic Sylvester matrices (see [20] for the case $k = 1$ and $a_0(z) = 1$). The algorithm relies on the ability to specify when a given subproblem is well-conditioned. The stability estimates come as a result of "near" inversion formulae for striped and mosaic Sylvester matrices given in [11]. In addition to a complete stability analysis, we have also provided some numerical experiments that verify that the algorithm performs as the theoretic results imply.

There is a number of open research problems that result from this work. The algorithm that has been presented is fast rather than superfast as is possible in the case of exact arithmetic [12]. It is possible to modify the algorithm so that it takes steps in a quadratic fashion as done in [12]. However, while this approach will work in the generic case, it is possible to find examples where not all the required subproblems are well-conditioned. In these cases the algorithm might not be numerically stable. It would be of interest to find a superfast algorithm that works in all cases and in addition is numerically stable.

In cases where the largest step-size is small the algorithm has complexity $\mathcal{O}(\|n\|^2)$. However, there are cases where the algorithm may require a very large step-size and and then have a higher cost than Gaussian elimination. This will happen if there is a very large unstable block, or if the stability parameter τ is chosen to be too low. It would be of interest to find a fast, stable algorithm that has complexity $\mathcal{O}(\|n\|^2)$ in all cases.

Our algorithm proceeds along a diagonal path in the corresponding Padé tables of our approximants. It would be of interest to find fast, stable algorithms that proceed along alternate paths in the Padé tables. An example of this in the Padé case is found in [18] where the computation proceeds along straight-line paths. In the context of matrix solvers this is the difference between giving a Toeplitz solver instead of a Hankel solver as is done in [15].

The *M-Padé approximation problem* is a generalization of the Padé-Hermite approximation problem which requires that the residual in (1) vanishes at a given set of knots $z_0, z_1, \ldots, z_{N-1}$, counting multiplicities [2, 3, 4, 24]. The case where all the z_i are equal to 0 is just the Padé-Hermite problem. In this case the coefficient matrix for the associated linear system is the matrix of divided differences. It would be of interest to determine stability parameters for such matrices, with a view to developing fast, stable algorithms for computing this approximation problem. Along these lines, some experiments for the case k=1 are reported in [8].

Acknowledgement. We are very greatful to a referee who contributed much in terms of the correctness of results and the clarity of presentation.

REFERENCES

[1] G. Baker and P. Graves-Morris, *Padé Approximants, Part II*, Addison-Wesley, Reading, MA, 1981.

[2] B. Beckermann, *Zur Interpolation mit polynomialen Linearkombinationen beliebiger Funktionen*, PhD thesis, Institut für Angewandte Mathematik, Universität Hannover, 1988.

[3] B. Beckermann, *The structure of the singular solution table of the M-Padé approximation problem*, Journal of Computational and Applied Mathematics, 32 (1990), pp. 3–15.

[4] ———, *A reliable method for computing M-Padé approximants on arbitrary staircases*, Journal of Computational and Applied Mathematics, 40 (1992), pp. 19–42.

[5] B. Beckermann and G. Labahn, *A uniform approach for the fast computation of matrix-type Padé approximants*, SIAM Journal on Matrix Analysis and Applications, (1994), pp. 804–823.

[6] A. W. Bojanczyk, R. P. Brent, F. D. De Hoog, and D. R. Sweet, *On the stability of the Bariss and related Toeplitz factorization algorithms*, SIAM Journal on Matrix Analysis and Applications, 16 (1995), pp. 40–57.

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

[8] S. Cabay, M. Gutknecht, and R. Meleshko, *Stable rational interpolation?*, Systems and Networks: Mathematical Theory and Applications, Proceedings of MTNS 93, (1994), pp. 631–633.

[9] S. Cabay, A. Jones, and G. Labahn, *A stable algorithm for multi-dimensional Padé systems and the inversion of generalized Sylvester matrices*, Tech. Report TR 94-07, Dept. Comp. Sci., Univ. Alberta, 1994.

[10] ———, *Experiments with a stable algorithm for computing Padé-Hermite and simultaneous Padé approximants*, (in preparation).

[11] ———, *Computation of numerical Padé-Hermite and simultaneous Padé systems I: Near inversion of generalized Sylvester matrices*, SIAM Journal on Matrix Analysis and Applications, (to appear).

[12] S. Cabay and G. Labahn, *A superfast algorithm for multi-dimensional Padé systems*, Numerical Algorithms, 2 (1992), pp. 201–224.

[13] ———, *Fast, stable inversion of mosaic Hankel matrices*, Systems and Networks: Mathematical Theory and Applications, Proceedings of MTNS 93, (1994), pp. 625–630.

[14] S. Cabay, G. Labahn, and B. Beckermann, *On the theory and computation of non-perfect Padé-Hermite approximants*, Journal of Computational and Applied Mathematics, 39 (1992), pp. 295–313.

[15] S. Cabay and R. Meleshko, *A weakly stable algorithm for Padé approximants and inversion of Hankel matrices*, SIAM Journal on Matrix Analysis and Applications, 14 (1993), pp. 735–765.

[16] G. E. Forsythe and C. B. Moler, *Computer Solution of Linear Algebraic Systems*, Prentice-

Hall, 1967.
[17] G. H. Golub and C. F. van Loan, *Matrix Computations*, Johns Hopkins University Press, 1983.
[18] M. Gutknecht, *Stable row recurrences in the Padé table and generically superfast lookahead solvers for non-hermitian Toeplitz solvers*, Linear Algebra and Its Applications, 188/189 (1993), pp. 351–421.
[19] M. H. Gutknecht and M. Hochbruck, *Look-ahead Levinson and Schur algorithms for non-Hermitian Toeplitz systems*, Tech. Report IPS 93–11, IPS-Zürich, 1993.
[20] ———, *The stability of inversion formulas for Toeplitz matrices*, Tech. Report IPS 93–13, IPS-Zürich, 1993.
[21] N. J. Higham and D. J. Higham, *Large growth factors in Gaussian elimination with pivoting*, SIAM Journal on Matrix Analysis and Applications, 10 (1989), pp. 155–164.
[22] T. Jones, *The numerical computation of Padé-Hermite systems*, master's thesis, Dept. Comp. Sci., Univ. Alberta, 1992.
[23] G. Labahn, *Inversion components for block Hankel-like matrices*, Linear Algebra and Its Applications, 177 (1992), pp. 7–48.
[24] K. Mahler, *Perfect systems*, Compositio Math., 19 (1968), pp. 95–166.
[25] R. Shafer, *On quadratic approximation*, SIAM J. Numerical Analysis, 11 (1974), pp. 447–460.
[26] L. N. Trefethen and R. S. Schreiber, *Average-case stability of Gaussian elimination*, SIAM Journal on Matrix Analysis and Applications, 11 (1990), pp. 335–360.
[27] M. Van Barel and A. Bultheel, *The computation of non-perfect Padé-Hermite approximants*, Numerical Algorithms, 1 (1991), pp. 285–304.
[28] J. Wilkinson, *Rounding Errors in Algebraic Processes*, Prentice-Hall, 1963.
[29] J. H. Wilkinson, *Modern error analysis*, SIAM Review, 13 (1971), pp. 548–568.