Application of Interval Analysis Techniques to Linear Systems: Part II—The Interval Matrix Exponential Function

EDWARD P. OPPENHEIMER, MEMBER, IEEE, AND ANTHONY N. MICHEL, FELLOW, IEEE

***Abstract* —In Part I of the present three-part paper [3] we established new results for continuous and rational interval functions which are of interest in their own right. In the present paper we use these results to study interval matrix exponential functions and to devise a method of constructing augmented partial sums which approximate interval matrix exponential functions as closely as desired. *In the third part* of this three-part paper we will use the above results to generate an algorithm which enables us to obtain estimates of bounds for the set of all solutions of initial-value problems of linear systems of autonomous first-order ordinary differential equations that linearly depend on a parameter belonging to an interval.**

The motivation for the present work includes many interesting applications. We cite here as examples the tolerance problem in electric circuits, optimal control problems with large tolerances on a parameter (where the usual sensitivity analysis methods fail), and the like.

I. Introduction

In a companion paper [3] we established new results for continuous and rational interval functions. In the present paper we use these results to study interval exponential functions and to develop an algorithm which will enable us to obtain reasonably sharp estimates for the range of interval matrix exponential functions of an interval (parameter) variable. Since this algorithm will be implemented on a finite wordlength digital computer, it is necessary to incorporate machine bounding arithmetic.

The results of the present paper as well as those given in [3] will be used in Part III [16] to study linear initial-value problems which are endowed with a parameter belonging to an interval (e.g., a parameter with a specified tolerance).

The results of this paper are presented in the following manner.

In Sections III and IV we consider "scalar" and matrix interval exponential functions, respectively. These functions are represented by infinite power series and their properties are studied in terms of rational functions (using the results of [3]) obtained from truncations of these infinite series.

To determine optimal estimates of error bounds for the truncated series representation of the interval matrix exponential function, we establish appropriate results dealing with Householder norms. This is accomplished in Section V.

The conservativeness of interval arithmetic operations can be reduced by considering the nested form for interval polynomials and the centered form for interval arithmetic representations. These notions are introduced in Section VI. In this section we also discuss briefly machine bounding arithmetic in digital computers.

Finally, Section VII presents an algorithm for the computation of the interval matrix exponential function which yields prespecified error bounds. This algorithm incorporates: machine bounding arithmetic; the perturbation parameter interval partitioning philosophy of theorem 14M and proposition 13M established in [3]; the nested and centered form techniques; and an optimal Householder norm.

II. Notation

In the subsequent sections, we will make use of the notation established in the first part of the present three-part paper [3]. This notation will not be restated here.

III. An Interval Exponential Function Computation Technique

In this section we present a technique by means of which the interval exponential function may be approximated by an "augmented" truncated series representation that set theoretically includes the infinite series result, with prescribed relative error bounds for the interval result endpoints.

Let $\{f_n\}$ denote the sequence of rational interval functions defined by

$$f_n(J) \triangleq \sum_{j=0}^{n} \frac{J^j}{j!}, \qquad J \triangleq [c,d] \in \mathscr{T}_I \tag{1}$$

where, for simplicity, $j!$ denotes the degenerate interval

Manuscript received January 21, 1987; revised December 2, 1987. The work of A. N. Michel was supported in part by the National Science Foundation under Grant ECS88-02924. This paper was recommended by Associate Editor F. M. A. Salam.

E. P. Oppenheimer is with the Applied Physics Laboratory, Johns Hopkins University, Laurel, MD 20707.

A. N. Michel is with the Department of Electrical and Computer Engineering, University of Notre Dame, Notre Dame, IN 46556.

IEEE Log Number 8822597.

0098-4094/88/1000-1230$01.00 ©1988 IEEE

$[j!, j!]$ and where the zeroth power of any interval is assumed to be the degenerate interval $[1,1]$. Now for any n and k,

$$\rho(f_n(J),f_{n+k}(J)) \leqslant \rho(f_n(J),f_{n+1}(J))+\cdots + \rho(f_{n+k-1}(J),f_{n+k}(J)).$$

Let $\phi=[0,0]$. From the definition of the metric ρ (see [3, eq. (4)] and the interval arithmetic operation of addition (see [3, eq. (1)] it follows that

$$\rho(f_n(J),f_{n+k}(J))\leqslant\rho\left(\phi,\frac{J^{n+1}}{(n+1)!}\right)+\cdots+\rho\left(\phi,\frac{J^{n+k}}{(n+k)!}\right).$$

Similarly, it is obvious that

$$\rho\left(\phi,\frac{J^{n+l}}{(n+l)!}\right)=\frac{1}{(n+l)!}\rho(\phi,J^{n+l})\leqslant\frac{[\rho(\phi,J)]^{n+l}}{(n+l)!}.$$

Following Moore [1], we define the *magnitude* of $J=[c,d]$ by $|J|=\rho(\phi,J)\triangleq\max(|c|,|d|)$. Then

$$\begin{aligned}
&\rho(f_n(J),f_{n+k}(J))\\
&\quad\leqslant\frac{|J|^{n+1}}{(n+1)!}+\cdots+\frac{|J|^{n+k}}{(n+k)!}\\
&\quad=\frac{|J|^{n+1}}{(n+1)!}\left[1+\frac{|J|}{n+2}+\cdots+\frac{|J|^{k-1}}{(n+k)\cdots(n+2)}\right]\\
&\quad\leqslant\frac{|J|^{n+1}}{(n+1)!}\left[1+\frac{|J|}{n+2}+\cdots+\left(\frac{|J|}{n+2}\right)^{k-1}\right]\\
&\quad=\frac{|J|^{n+1}}{(n+1)!}\cdot\frac{1-\left(\dfrac{|J|}{n+2}\right)^{k}}{1-\left(\dfrac{|J|}{n+2}\right)}\leqslant\frac{|J|^{n+1}}{(n+1)!}\cdot\frac{1}{1-\left(\dfrac{|J|}{n+2}\right)}
\end{aligned}$$

where it is assumed that n is sufficiently large so that

$$\frac{|J|}{n+2}<1.$$

Thus

$$\rho(f_n(J),f_{n+k}(J))\leqslant\frac{|J|^{n+1}}{(n+1)!}\cdot\frac{1}{1-\left(\dfrac{|J|}{n+2}\right)}.$$

But this relation can be made independent of J in the sense that $J\in\mathscr{T}_I$, $\mathscr{T}_I$ is compact and for every $J\in\mathscr{T}_I$, $|J|\leqslant|I|$. Thus for any $\epsilon>0$ it is possible to select N sufficiently large so that for every $J\in\mathscr{T}_I$ and for all $n,m\geqslant N$,

$$\rho(f_n(J),f_m(J))\leqslant\frac{|I|^{N+1}}{(N+1)!}\cdot\frac{1}{1-\left(\dfrac{|I|}{N+2}\right)}<\epsilon.$$

To obtain the above inequality, we use Stirling's formula (see [2, p. 384]) to note that

$$\left\{\frac{n!}{\left(\dfrac{n}{e}\right)^{n}\sqrt{2\pi n}}\right\}$$

is a monotonically nonincreasing positive sequence with limit equal to one. Thus

$$\left(\frac{n}{e}\right)^{n}\sqrt{2\pi n}\leqslant n!$$

and therefore

$$\begin{aligned}
\lim_{N\to\infty}\frac{|I|^{N+1}}{(N+1)!}=\lim_{N\to\infty}\frac{|I|^{N}}{N!}&\leqslant\lim_{N\to\infty}\frac{|I|^{N}}{\left(\dfrac{N}{e}\right)^{N}\sqrt{2\pi N}}\\
&=\lim_{N\to\infty}\frac{1}{\left(\dfrac{N}{e|I|}\right)^{N}\sqrt{2\pi N}}=0.
\end{aligned}$$

Therefore, by selecting N sufficiently large, we obtain

$$\frac{|I|^{N+1}}{(n+1)!}\,\frac{1}{1-\left(\dfrac{|I|}{N+2}\right)}<\epsilon.$$

It now follows that $\{f_n\}$ is a Cauchy sequence in the complete metric space $\{\mathscr{F},\mu\}$ (see [3, proposition 5]) and as such it will converge to an element of $\mathscr{F}$. Let us denote this element by

$$f(J)\triangleq\sum_{j=0}^{\infty}\frac{J^j}{j!}\triangleq e^J.$$

Note that $\{f_n\}$ is a Cauchy sequence of rational interval functions and by proposition 11 [3] it follows that

$$f(J)\triangleq e^J\supset\bar{f}(J)\triangleq\bigcup_{x\in J}f([x,x])=\bigcup_{x\in J}e^x$$

since $f([x,x])\triangleq e^x$.

For the present, assume that the interval arithmetic operations can be exactly computed in the reals (i.e., on an "infinite-decimal" machine) and that it is desired to determine an interval result which will contain the value of the interval function

$$f(J)\triangleq\sum_{j=0}^{\infty}\frac{J^j}{j!}\triangleq e^J$$

within some predetermined relative error. Let the interval function be defined by

$$f(J)\triangleq e^J=\left[f^L(J),f^R(J)\right]$$

and let the computable truncated interval series be defined by

$$f_n(J)\triangleq\sum_{i=0}^{n}\frac{J^i}{i!}=\left[f_n^L(J),f_n^R(J)\right].$$

Let the remainder of the interval series be defined by

$$r_n(J)\triangleq\sum_{i=n+1}^{\infty}\frac{J^i}{i!}=\left[r_n^L(J),r_n^R(J)\right].$$

It has been shown previously that there exists an upper bound on the metric measure of how closely the truncated series approximates the actual function,

$$\begin{aligned}\rho(f_n(J), f(J)) &= \rho(\phi, r_n(J))\\ &\triangleq |r_n(J)| \leqslant \frac{|J|^{n+1}}{(n+1)!}\cdot\frac{1}{1-\dfrac{|J|}{n+2}}\\ &\triangleq \tau, \quad \text{provided that } \frac{|J|}{n+2} < 1. \qquad (2)\end{aligned}$$

Since the truncated series is computable, assume by appropriate programming techniques that it is possible to satisfy the relation

$$\begin{aligned}|r_n(J)| &\triangleq \max\left(|r_n^L(J)|, |r_n^R(J)|\right)\\ &\leqslant \tau \leqslant 10^{-P}\cdot\min\left\{|f_n^L(J)|, |f_n^R(J)|\right\} \qquad (3)\end{aligned}$$

where P is a positive integer, so that

$$0 < |r_n^L(J)| \leqslant \tau \leqslant 10^{-P}|f_n^L(J)|. \qquad (4)$$

Since

$$\begin{aligned}f(J) &\triangleq f_n(J) + r_n(J)\\ f^L(J) &\triangleq f_n^L(J) + r_n^L(J) \geqslant f_n^L(J) - |r_n^L(J)|\end{aligned}$$

combining the above two inequalities, we obtain

$$f^L(J) \geqslant f_n^L(J) - \tau \geqslant f_n^L(J) - 10^{-P}|f_n^L(J)|.$$

Taking the negative of this inequality and adding $f^L(J)$, we obtain

$$\begin{aligned}0 &\leqslant f^L(J) - [f_n^L(J) - \tau]\\ &\leqslant f^L(J) - \left[f_n^L(J) - 10^{-P}|f_n^L(J)|\right].\end{aligned}$$

Now $f^L(J) - [f_n^L(J) - \tau]$ is the error in the approximation endpoint $[f_n^L(J) - \tau]$ and it is bounded above, since

$$\begin{aligned}0 &\leqslant |f^L(J) - [f_n^L(J) - \tau]|\\ &\leqslant |f^L(J) - f_n^L(J)| + 10^{-P}|f_n^L(J)|.\end{aligned}$$

Assuming

$$|f^L(J)| \geqslant \left||f^L(J) - f_n^L(J)| - |f_n^L(J)|\right| > 0$$

then

$$\begin{aligned}&\frac{|f^L(J) - [f_n^L(J) - \tau]|}{|f^L(J)|}\\ &\quad\leqslant \frac{|f^L(J) - f_n^L(J)| + 10^{-P}|f_n^L(J)|}{\left||f^L(J) - f_n^L(J)| - |f_n^L(J)|\right|}\\ &\quad= \frac{|f^L(J) - f_n^L(J)|/|f_n^L(J)|}{|1 - |f^L(J) - f_n^L(J)|/|f_n^L(J)||}\\ &\qquad + \frac{10^{-P}}{|1 - |f^L(J) - f_n^L(J)|/|f_n^L(J)||}.\end{aligned}$$

But from above, $|f^L(J) - f_n^L(J)| \triangleq |r_n^L(J)| \leqslant 10^{-P}|f_n^L(J)|$. Therefore,

$$\frac{|f^L(J) - [f_n^L(J) - \tau]|}{|f^L(J)|} \leqslant 2\left(\frac{10^{-P}}{1 - 10^{-P}}\right). \qquad (5)$$

In a similar manner, we can show that

$$\frac{|[f_n^R(J) + \tau] - f^R(J)|}{|f^R(J)|} \leqslant 2\frac{10^{-P}}{1 - 10^{-P}}. \qquad (6)$$

Thus by selecting n sufficiently large, so that the conditions $|J|/(n+2) < 1$ and

$$\frac{|J|^{n+1}}{(n+1)!}\cdot\frac{1}{1-\dfrac{|J|}{n+2}} \triangleq \tau \leqslant 10^{-P}\cdot\min\left\{|f_n^L(J)|, |f_n^R(J)|\right\}$$

are satisfied, and denoting

$$\epsilon \triangleq 2\frac{10^{-P}}{1 - 10^{-P}}$$

we obtain the estimate

$$[1-\epsilon, 1+\epsilon]\cdot e^J \supset \sum_{i=0}^{n}\frac{J^i}{i!} + [-\tau, \tau] \supset e^J \supset \bigcup_{x\in J} e^x.$$

Recall that $0 \leqslant f^L(J) - [f_n^L(J) - \tau]$ and $0 \leqslant [f_n^R(J) + \tau] - f^R(J)$. Then (5) and (6) above imply that

$$\begin{aligned}f^L(J) - \epsilon|f^L(J)| &\leqslant f_n^L(J) - \tau\\ f_n^R(J) + \tau &\leqslant f^R(J) + \epsilon|f^R(J)|\end{aligned}$$

which provides that

$$[1-\epsilon, 1+\epsilon]\cdot e^J \supset (f_n(J) + [-\tau, \tau]).$$

In other words, by including a sufficient number of terms in the truncated computable series (1) so that the algorithmic inequality (3) is satisfied, it is possible to augment this truncated series result so that the augmented interval result bounds the actual interval exponential function and does so within the specified relative endpoint error bounds (5) and (6).

It should be remarked at this point that nothing has been said to indicate how well e^J approximates the corresponding united extension, $\bigcup_{x\in J} e^x$. Proposition 13 in the first part of the present three part paper [3] and Theorem 14 in [3] indicate the direction which will be followed in improving this approximation and the corresponding augmented result approximating e^J.

IV. An Interval Matrix Exponential Function Computation Technique

In the present section we extend the interval exponential function approximation technique of Section III to the matrix counterpart. To this end, we let $\{F_k\}$ denote the sequence of rational interval matrix functions defined by

$$F_k(J) \triangleq ((f_{ij_k}(J))) \triangleq (([f_{ij_k}^L(J), f_{ij_k}^R(J)])) \triangleq \sum_{i=0}^{k}\frac{A^i(J)}{i!}$$

where

$$A(J) \triangleq ((a_{ij}(J))) \triangleq (([a_{ij}^L(J), a_{ij}^R(J)])). \qquad (7)$$

$A(J)$ is the rational interval matrix function defined by

$$A(J) \triangleq A_1 + JA_2, \qquad J \in \mathscr{T}_I$$

where $A_1, A_2 \in \mathscr{T}^{n^2}$ are constant interval matrices. (Refer to (10) in [3] for the definition of $\mathscr{T}^{n^2}$.) For any k and l, we have

$$\sigma(F_k(J), F_{k+l}(J)) \leqslant \sigma(F_k(J), F_{k+1}(J)) + \cdots + \sigma(F_{k+l-1}(J), F_{k+l}(J)).$$

(Refer to [3, eq. (11)] for the definition of the metric σ.) Let 0 denote the interval matrix where each element is a degenerate zero interval. From the definitions of the metric σ and the interval arithmetic operation of addition, it follows then that

$$\sigma(F_k(J), F_{k+l}(J)) \leqslant \sigma\left(0, \frac{A^{k+1}(J)}{(k+1)!}\right) + \cdots + \sigma\left(0, \frac{A^{k+l}(J)}{(k+l)!}\right). \quad (8)$$

It is apparent that

$$\sigma\left(0, \frac{A^m(J)}{m!}\right) = \frac{1}{m!}\sigma(0, A^m(J)).$$

(Recall that $m!$ represents the degenerate interval $[m!, m!]$ and assume that the zeroth power of any interval matrix is the degenerate interval identity matrix in the above expression.) Also, we claim that

$$\sigma(0, A^m(J)) \leqslant [\sigma(0, A(J))]^m.$$

To see this, we first define the real matrix $|B|$ from the interval matrix $B \in \mathscr{T}^{n^2}$ by

$$|B| \triangleq ((|b_{ij}|)) \triangleq ((|[b_{ij}^L, b_{ij}^R]|)) \triangleq ((\max\{|b_{ij}^L|, |b_{ij}^R|\})). \quad (9)$$

Let $G \triangleq \operatorname{diag}(u_i)$, where $\boldsymbol{u}$ denotes the fixed positive real vector used in the definition of the metric σ (see [3, eq. (11)]). From the properties of interval arithmetic operations and from (9) we have

$$|A^m(J)| \leqslant |A(J)|^m. \quad (10)$$

Note that $G^{-1}|A^m(J)|G$ and$G^{-1}|A(J)|^m G$ are real $n \times n$ matrices. For the real $n \times n$ matrix $C = ((c_{ij}))$, define the matrix norm $\|C\|_\infty$ by

$$\|C\|_\infty \triangleq \max_i \sum |c_{ij}|. \quad (11)$$

From (9) and (10) and the definition of the metric σ, we have

$$\sigma(0, A^m(J)) \triangleq \sigma(0, |A^m(J)|) \leqslant \sigma(0, |A(J)|^m). \quad (12)$$

But

$$\begin{aligned}\sigma(0, |A(J)|^m) &\triangleq \|G^{-1}|A(J)|^m G\|_\infty \\ &= \|G^{-1}|A(J)|_{(1)} G G^{-1}|A(J)|_{(2)} G \cdots \\ &\quad G^{-1}|A(J)|_{(m)} G\|_\infty \\ &\leqslant \left(\|G^{-1}|A(J)|G\|_\infty\right)^m \qquad (13)\end{aligned}$$

by the properties of the matrix norm. Also,

$$\|G^{-1}|A(J)|G\|_\infty \triangleq \sigma(0, A(J)). \quad (14)$$

Combining (12)–(14), we obtain

$$\sigma(0, A^m(J)) \leqslant [\sigma(0, A(J))]^m \quad (15)$$

which was the claim to be proved.

For convenience, we define

$$\lambda \triangleq \sigma(0, A(J)).$$

Combining (8) and (15), we obtain

$$\begin{aligned}\sigma(F_k(J), F_{k+l}(J)) &\leqslant \frac{\lambda^{k+1}}{(k+1)!} + \cdots + \frac{\lambda^{k+l}}{(k+l)!} \\ &= \frac{\lambda^{k+1}}{(k+1)!}\left(1 + \frac{\lambda}{k+2} + \cdots + \frac{\lambda^{l-1}}{(k+2)\cdots(k+l)}\right) \\ &\leqslant \frac{\lambda^{k+1}}{(k+1)!}\left(1 + \frac{\lambda}{k+2} + \cdots + \left(\frac{\lambda}{k+2}\right)^{l-1}\right) \\ &= \frac{\lambda^{k+1}}{(k+1)!} \cdot \frac{1 - \left(\frac{\lambda}{k+2}\right)^l}{1 - \left(\frac{\lambda}{k+2}\right)} \\ &\leqslant \frac{\lambda^{k+1}}{(k+1)!} \cdot \frac{1}{1 - \left(\frac{\lambda}{k+2}\right)} \triangleq \tau \qquad (16)\end{aligned}$$

provided that $\lambda/(k+2) < 1$. But $J \in \mathscr{T}_I$ and $|J| \leqslant |I|$ and hence $|A(J)| \leqslant |A(I)|$ in the sense that for each $i, j, |a_{ij}(J)| \leqslant |a_{ij}(I)|$. Then

$$\lambda \triangleq \sigma(0, A(J)) \leqslant \bar{\lambda} \triangleq \sigma(0, A(I))$$

and $\bar{\lambda}$ is independent of J. For arbitrary $\epsilon > 0$ it is obvious then that there exists N such that

$$\sigma(F_k(J), F_{k+l}(J)) \leqslant \frac{\bar{\lambda}^{N+1}}{(N+1)!} \cdot \frac{1}{1 - \left(\frac{\bar{\lambda}}{N+2}\right)} \triangleq \bar{\tau} < \epsilon$$

for all $J \in \mathscr{T}_I$, for all $k \geqslant N$ and for all $l = 1, 2, \cdots$. Recalling the definition of the metric ζ defined on the metric space $\mathscr{F}^{n^2}$ (see (12) in [3]), we obtain

$$\zeta(F_k, F_{k+l}) \leqslant \bar{\tau} < \epsilon, \qquad k \geqslant N, \; l = 1, 2, \cdots.$$

It follows that $\{F_k\}$ is a Cauchy sequence in the complete metric space $\{\mathscr{F}^{n^2}, \zeta\}$, that $\{F_k\} \to F \in \{\mathscr{F}^{n^2}, \zeta\}$ and that

$$F(J) \triangleq \sum_{i=0}^{\infty} \frac{A^i(J)}{i!} \triangleq e^{A(J)} \in \mathscr{T}^{n^2}, \qquad \text{for all } J \in \mathscr{T}_I.$$

Moreover, from proposition 11M in [3], it follows that for any $J \in \mathscr{T}_I$,

$$F(J) \triangleq e^{A(J)} \supset \bar{F}(J) \triangleq \bigcup_{x \in J} F([x, x]) \triangleq \bigcup_{x \in J} e^{A([x, x])}.$$

Proceeding, as in the case of the scalar interval exponential function, we assume for the time being that the interval arithmetic operations are calculated on an infinite-decimal machine. Let the actual interval matrix exponential function be represented as

$$F(J) \triangleq ((f_{ij}(J))) \triangleq (([f_{ij}^L(J), f_{ij}^R(J)])) \triangleq e^{A(J)} \quad (17a)$$

and let the computable truncated interval matrix series $F_k(J)$ be defined by (7). Define the remainder of the interval matrix infinite series by

$$R_k(J) \triangleq ((r_{ij_k}(J))) \triangleq (([r_{ij_k}^L(J), r_{ij_k}^R(J)])) \triangleq \sum_{i=k+1}^{\infty} \frac{A^i(J)}{i!}. \quad (17b)$$

From the previous arguments then, there exists an upper bound on the metric measure σ of how closely the truncated series approximates the actual function,

$$\sigma(F_k(J), F(J)) \triangleq \sigma(0, R_k(J)) \leqslant \frac{\lambda^{k+1}}{(k+1)!} \cdot \frac{1}{1 - \dfrac{\lambda}{k+2}} \triangleq \tau \quad (18)$$

provided that $\lambda/(k+2) < 1$.

From the definition of the metric σ defined on the set $\mathscr{T}^{n^2}$ (see (11) in [3]), we see that

$$\left(\frac{u_j}{u_i}\right)|r_{ij_k}(J)| \triangleq \left(\frac{u_j}{u_i}\right) \max\left\{|r_{ij_k}^L(J)|, |r_{ij_k}^R(J)|\right\} \leqslant \sigma(0, R_k(J)) \leqslant \tau. \quad (19)$$

Assume then that for the computable truncated series, it is possible by appropriate programming techniques to satisfy the relation

$$|r_{ij_k}(J)| \leqslant \left(\frac{u_i}{u_j}\right) \cdot \sigma(0, R_k(J)) \leqslant \left(\frac{u_i}{u_j}\right)\tau \leqslant 10^{-P} \min\left\{|f_{ij_k}^L(J)|, |f_{ij_k}^R(J)|\right\} \quad (20)$$

for every i, j. Then

$$0 < |r_{ij_k}^L(J)| \leqslant \left(\frac{u_i}{u_j}\right)\tau \leqslant 10^{-P}|f_{ij_k}^L(J)|. \quad (21)$$

Also, for each i, j,

$$f_{ij}(J) \triangleq f_{ij_k}(J) + r_{ij_k}(J)$$

and, therefore,

$$f_{ij}^L(J) \triangleq f_{ij_k}^L(J) + r_{ij_k}^L(J) \geqslant f_{ij}^L(J) - |r_{ij_k}^L(J)|. \quad (22)$$

Combining (21) and (22), we obtain

$$f_{ij}^L(J) \geqslant f_{ij_k}^L(J) - \left(\frac{u_i}{u_j}\right)\tau \geqslant f_{ij_k}^L(J) - 10^{-P}|f_{ij_k}^L(J)|$$

and taking the negative of this inequality and adding $f_{ij}^L(J)$, we obtain

$$0 \leqslant f_{ij}^L(J) - \left(f_{ij_k}^L(J) - \frac{u_i}{u_j}\tau\right) \leqslant f_{ij}^L(J) - \left(f_{ij_k}^L(J) - 10^{-P}|f_{ij_k}^L(J)|\right). \quad (23a)$$

Hence, the error in the approximation endpoint

$$f_{ij_k}^L - \frac{u_i}{u_j}\tau$$

is bounded above by the inequality

$$0 \leqslant \left| f_{ij}^L(J) - \left(f_{ij_k}^L(J) - \frac{u_i}{u_j}\tau\right)\right| \leqslant |f_{ij}^L(J) - f_{ij_k}^L(J)| + 10^{-P}|f_{ij_k}^L(J)|. \quad (23b)$$

Assuming

$$|f_{ij}^L(J)| \geqslant \left| |f_{ij_k}^L(J) - f_{ij_k}^L(J)| - |f_{ij_k}^L(J)| \right| > 0$$

and using (23b), then

$$\frac{\left| f_{ij}^L(J) - \left(f_{ij_k}^L(J) - \dfrac{u_i}{u_j}\tau\right)\right|}{|f_{ij}^L(J)|} \leqslant \frac{|f_{ij}^L(J) - f_{ij_k}^L(J)| + 10^{-P}|f_{ij_k}^L(J)|}{\left| |f_{ij}^L(J) - f_{ij_k}^L(J)| - |f_{ij_k}^L(J)| \right|} = \frac{\left[\dfrac{|f_{ij}^L(J) - f_{ij_k}^L(J)|}{|f_{ij_k}^L(J)|} + 10^{-P}\right]}{\left[\left|1 - \dfrac{|f_{ij}^L(J) - f_{ij_k}^L(J)|}{|f_{ij_k}^L(J)|}\right|\right]}. \quad (24)$$

By (20) and (21),

$$|f_{ij}^L(J) - f_{ij_k}^L(J)| \triangleq |r_{ij_k}^L(J)| \leqslant 10^{-P}|f_{ij_k}^L(J)|. \quad (25)$$

Combining (24) and (25) we obtain

$$\left| f_{ij}^L(J) - \left(f_{ij_k}^L(J) - \frac{u_i}{u_j}\tau\right)\right| \Big/ |f_{ij}^L(J)| \leqslant 2\left(\frac{10^{-P}}{1 - 10^{-P}}\right) \quad (26)$$

and this holds for each i, j.

In a similar manner it can be shown that for each i, j,

$$\left| \left(f_{ij_k}^R(J) + \frac{u_i}{u_j}\tau\right) - f_{ij}^R(J)\right| \Big/ |f_{ij}^R(J)| \leqslant 2\left(\frac{10^{-P}}{1 - 10^{-P}}\right). \quad (27)$$

As in the case of scalar exponential interval functions, by selecting k sufficiently large so that for all i, j the conditions $\lambda/(k+2)<1$ and

$$\left(\frac{u_i}{u_j}\right)\cdot\frac{\lambda^{k+1}}{(k+1)!}\cdot\frac{1}{1-\dfrac{\lambda}{k+2}}$$

$$\triangleq\frac{u_i}{u_j}\tau\leqslant 10^{-P}\min\left\{\left|f^{L}_{ij_k}(J)\right|,\left|f^{R}_{ij_k}(J)\right|\right\}$$

are satisfied, then for

$$\epsilon\triangleq 2\left(\frac{10^{-P}}{1-10^{-P}}\right)$$

we obtain

$$[1-\epsilon,1+\epsilon]e^{A(J)}\supset\sum_{i=0}^{k}\frac{A^i(J)}{i!}+Z\supset e^{A(J)}\supset\bigcup_{x\in J}e^{A([x,x])}$$

where

$$Z\triangleq\left(\left(-\frac{u_i}{u_j}\tau,\frac{u_i}{u_j}\tau\right)\right).$$

In other words, by including a sufficient number of terms in the truncated computable series (7) so that the algorithmic inequality (20) is satisfied, it is possible to augment this truncated series so that the interval result bounds the actual interval matrix exponential function $e^{A(J)}$ and does so within the specified relative endpoint error bounds (26) and (27).

As in the scalar case, nothing has been said regarding how well $e^{A(J)}$ approximates the corresponding united extension $\bigcup_{x\in J}e^{A[x,x]}$. Both proposition 13M and theorem 14M in [3] indicate the direction which will be followed in improving this approximation and the corresponding augmented result approximating $e^{A(J)}$.

V. Estimates of Error Bounds for the Truncated Series Representation of the Exponential Matrix Function

In the present section we obtain sharp bounds for the errors created by truncating the series representation for e^{At} (where A is a real $n\times n$ matrix and $t\geqslant 0$) by making use of Householder matrix norms. In Section VII we will apply the results of the present section to interval matrix exponential functions.

Now recall that the following norms of vectors in R^n induce the corresponding norms on matrices in $R^{n\times n}$ (see e.g., [4]): for $x\in R^n$, $x^T=(x_1,\cdots,x_n)$, $A=((a_{ij}))$,

(a) if $\|x\|_1\triangleq\sum_{i=1}^{n}|x_i|$, then $\|A\|_1\triangleq\max_j\left\{\sum_{i=1}^{n}|a_{ij}|\right\}$;

(b) if $\|x\|_2\triangleq\left(\sum_{i=1}^{n}x_i^2\right)^{1/2}$, then $\|A\|_2=\lambda^{1/2}_{\max}$ where $\lambda_{\max}$ denotes the largest eigenvalue of A^TA; and

(c) if $\|x\|_\infty\triangleq\max_j\{|x_j|\}$, then $\|A\|_\infty\triangleq\max_i\left\{\sum_{j=1}^{n}|a_{ij}|\right\}$.

Recall also that a matrix norm is said to be *consistent* with a given vector norm if for every A and x it is true that

$$\|Ax\|\leqslant\|A\|\,\|x\|$$

and *subordinate* to the vector norm in the case where the matrix norms are consistent and for every A there exists an $x\neq 0$ such that

$$\|Ax\|=\|A\|\,\|x\|.$$

In particular, it can be shown that the matrix norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ are subordinate to their corresponding vector norms.

The subsequent two vector norms were studied by Householder ([5, pp. 9–16] [6, pp. 9, 10]). Let

$$\|x\|_u=\|G^{-1}x\|_\infty\qquad\text{where } G\triangleq\operatorname{diag}(g_i),\quad g_i>0.\tag{28}$$

Since the $\|\cdot\|_\infty$ matrix norm is subordinate to the vector norm, it follows readily that

$$\|A\|_u=\max_i\left\{\frac{1}{g_i}\sum_{j=1}^{n}g_j|a_{ij}|\right\}.\tag{29}$$

Also, let

$$\|x\|_{u'}\triangleq\|Hx\|_1\qquad\text{where}\quad H=\operatorname{diag}(h_i),\quad h_i>0.\tag{30}$$

Again, since the $\|\cdot\|_1$ matrix norm is subordinate to the vector norm, it follows readily that

$$\|A\|_{u'}=\max_j\left\{\frac{1}{h_j}\sum_{i=1}^{n}h_i|a_{ij}|\right\}.\tag{31}$$

Next, recall that a real matrix $B=((b_{ij}))$, $b_{ij}\geqslant 0$, is called *irreducible* in the sense of Frobenius ([7, p. 50]) if it is not possible to split the index set $\{i,\cdots,n\}$ into two nonvoid disjoint sets α and β such that $b_{ij}=0$ for all $i\in\alpha$, $j\in\beta$. Recall also that for any real matrix $A=((a_{ij}))$, the matrix $|A|=((|a_{ij}|))$ is called the *abmatrix of A* ([5, p. 9]). Furthermore, recall that the *Frobenius theorem* on the spectral properties of irreducible nonnegative matrices ([7, p. 53]) states that for such a matrix there is a simple real positive eigenvalue r which is greater than or equal to the modulus of every other characteristic value and that the corresponding eigenvector w has strictly positive components. It is interesting to note that if this irreducible matrix is $|A|$, the Frobenius theorem proof develops the characterization ([7, p. 65]) $r=\|A\|_u$ in (29) with $G=\operatorname{diag}(w_i)$. Recall also that the modulus of the largest characteristic value of a matrix is called the *spectral radius*. For any real matrix A, the infimum of the set of values of the matrix norms on A induced by the family of all vector norms on R^n is the spectral radius and in this sense then, for an irreducible matrix $|A|$, the Householder matrix norm with $G=\operatorname{diag}(w_i)$ produces the optimally infimum value, the spectral radius of $|A|$ (see [8, p. 249]). When $|A|$ is irreducible, [8] additionally characterizes the subfamily of vector norms and induced matrix norms on A for which this Householder matrix norm yields an optimally infimum value. This family includes the $\|\cdot\|_\infty$ and $\|\cdot\|_1$ vector norms.

It must be remarked that the class of nonnegative real matrices $|A|$ for which either $\|A\|_u$ or $\|A\|_{u'}$ yields an infimum value is larger than the set of irreducible matrices and necessary and sufficient conditions characterizing this additional set of matrices are stated precisely in [7, p. 77].

The consequence of the above remark is that $\|A\|_{u'}$ may be defined while $\|A\|_u$ is not (or the converse). For example,

$$A=\begin{bmatrix} 0 & 1 & 0 \\ -1 & -0.4 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

results in $g_3=0$ while A^T yields $h_i>0$, $i=1,2,3$ and thus $\|A\|_{u'}=0.2+\sqrt{1.04}$, the spectral radius of $|A|$ and $|A^T|$.

It is apparent from the definition of $\|\cdot\|_1$ and $\|\cdot\|_\infty$ that for any i, j, $|a_{ij}|\leqslant\|A\|_y$ when y denotes 1 or ∞. It turns out that this relationship is also true when $y=2$. This claim is verified as follows. The real symmetric matrix $B\triangleq A^TA$ is normal and, therefore, has at least one eigenvalue λ_t such that

$$|\lambda_t|\geqslant\max_{i,j}|b_{ij}|$$

(see e.g., [9, p. 161]). But A^TA is nonnegative definite and thus all of its eigenvalues are real and nonnegative. Now for any i, j,

$$|b_{ii}|\triangleq\left|\sum_{k=1}^{n}a_{ki}^2\right|\geqslant a_{ji}^2$$

and therefore, using the definition of $\|\cdot\|_2$, it follows that

$$\|A\|_2\geqslant|\lambda_t|^{1/2}\geqslant|a_{ij}|$$

which was the claim to be proved. In the case of the two Householder matrix norms, we have

$$\frac{g_j}{g_i}|a_{ij}|\leqslant\|A\|_u \quad\text{and}\quad \frac{h_i}{h_j}|a_{ij}|\leqslant\|A\|_{u'}. \tag{32}$$

Let us now turn to the numerical computation of the real matrix exponential

$$\Phi(t)\triangleq e^{At}\triangleq\sum_{i=0}^{\infty}\frac{A^it^i}{i!},\qquad t>0.$$

Let

$$M\triangleq((m_{ij}))\triangleq\sum_{i=0}^{k}\frac{A^it^i}{i!}\quad\text{and}$$

$$R\triangleq((r_{ij}))\triangleq\sum_{i=k+1}^{\infty}\frac{A^it^i}{i!}$$

where it is assumed that $\|A\|_u$ and $\|A\|_{u'}$ exist. In a manner similar to the numerical scheme described previously for computation of the interval matrix exponential, letting y denote 1, 2 or ∞ for any i, j, we obtain

$$|r_{ij}|\leqslant\|R\|_y\leqslant\tau_y\triangleq\frac{(\|A\|_yt)^{k+1}}{(k+1)!}\cdot\frac{1}{1-\dfrac{\|A\|_yt}{k+2}} \tag{33}$$

provided that

$$0<\frac{\|A\|_yt}{k+2}<1.$$

In the case of the Householder matrix norms (τ denotes the same calculation with the norms replaced appropriately in (33)), for any i, j we have

$$\frac{g_j}{g_i}|r_{ij}|\leqslant\tau_u \quad\text{and}\quad \frac{h_i}{h_j}|r_{ij}|\leqslant\tau_{u'}. \tag{34}$$

Then, if for each i, j

$$|r_{ij}|\leqslant\tau_{(1,2\text{ or }\infty)}\leqslant10^{-P}|m_{ij}|$$

$$|r_{ij}|\leqslant\frac{g_i}{g_j}\tau_u\leqslant10^{-P}|m_{ij}|\quad\text{and}$$

$$|r_{ij}|\leqslant\frac{h_j}{h_i}\tau_{u'}\leqslant10^{-P}|m_{ij}| \tag{35}$$

the error in any term of the truncated matrix series M approximating $\Phi(t)\triangleq e^{At}$ is less than $10^{-P}|m_{ij}|$. (This relationship can always be satisfied by increasing k sufficiently, as previously shown.)

Hopefully, the effect of using the "optimally infimum" Householder matrix norm in the computation of τ is to obtain a sharper result for this bound, thereby minimizing the number of terms in the series that is required to yield an approximation which is accurate to within a given decimal.

The "optimal" Householder matrix norms require of course the computation of the maximal eigenvalue and a corresponding eigenvector of $|A|$. However, these computations are readily available (e.g., in the fast double-precision EISPAC routine.)

Table I gives results for the two examples,

$$A=\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ -0.75 & -2.75 & -3 \end{bmatrix} \tag{36}$$

$$B=\begin{bmatrix} -\ 0.05 & -6.0 & 0 & 0 \\ -10^{-3} & -0.15 & 0 & 0 \\ 1.0 & 0 & 0 & 13.0 \\ 0 & 1.0 & 0 & 0 \end{bmatrix}. \tag{37}$$

The abmatrix $|A|$ is irreducible while the abmatrix $|B|$ is reducible. (The spectral radii and the norms $\|\cdot\|_2$ and $\|\cdot\|_u$ were computed using the double-precision EISPAC routine while the remaining computations were evaluated on a 12-digit calculator.)

In Table I it should be noted that for matrix A with $k=8$ and $t=0.1$, the

$$\max_{i,j}\left\{\frac{g_i}{g_j}\right\}\tau_u=\frac{1.051973}{0.07362408}\cdot\tau_u\triangleq W\cdot\tau_u \tag{38}$$

bound is an order of magnitude sharper than the τ_∞ bound obtained in [10, pp. 104–107] but more conservative than the τ_2 and τ_1 bounds. However, given

$$W=\max_{i,j}\left\{\frac{g_i}{g_j}\right\} \tag{39}$$

TABLE I
NUMERICAL RESULTS FOR THE COMPUTATION OF τ IN THE TWO EXAMPLES (36) AND (37), WITH VARIOUS MATRIX NORMS

Spectral Radius of A	= 1.500000	$\frac{\|A\|_y t}{k+2}$ $k=8$ $t=0.1$	τ_y	$\|r_{ij}\| \leqslant$
$\|A\|_\infty$	6.500000	6.5×10^{-2}	6.104436×10^{-8}	τ_∞
$\|A\|_1$	4.000000	4.0×10^{-2}	7.524985×10^{-10}	τ_1
$\|A\|_2$	4.253694	4.253694×10^{-2}	1.312234×10^{-9}	τ_2
$\|A\|_u$	3.780003	3.780003×10^{-2}	4.512331×10^{-10}	$W\cdot\tau_u = 6.447407\times10^{-9}$
Spectral Radius of B	= 0.1921954	$\frac{\|B\|_y t}{k+2}$ $k=20$ $t=0.1$*	τ_y	$\|r_{ij}\| \leqslant$
$\|B\|_\infty$	14.0	6.363636×10^{-2}	2.448501×10^{-17}	τ_∞
$\|B\|_1$	13.0	5.909091×10^{-2}	5.139476×10^{-18}	τ_1
$\|B\|_2$	13.03841	5.926550×10^{-2}	5.468980×10^{-18}	τ_2
$\|B\|_u$	0.1921954	3.494462×10^{-2}	8.108984×10^{-23}	$\frac{907.58886}{1.588150}\cdot\tau_u = 4.634085\times10^{-20}$

*$t = 4.0$ vice 0.1 in the case of $\|B\|_u$.

since $\|A\|_u \leqslant \|A\|_y$, where y denotes 1, 2 or ∞, by selecting k sufficiently large, for each i, j it is always possible to satisfy

$$|r_{ij}| \leqslant W\tau_u < \tau_y.$$

This can be seen by observing that for k sufficiently large, it is always possible to satisfy

$$\frac{\tau_y}{\tau_u} = \left(\frac{\|A\|_y}{\|A\|_u}\right)^{k+1}\left[\frac{1}{1-\frac{\|A\|_y t}{k+2}}\right]\Bigg/\left[\frac{1}{1-\frac{\|A\|_u t}{k+2}}\right] > W.$$

In the case of $\|A\|_1/\|A\|_u > 1$, $k \geqslant 47$ yields $W\cdot\tau_u < \tau_1$.

For matrix B given by equation (37), Table I points up the importance in selecting an optimal matrix norm in computing the bound τ. With a fixed maximal number of terms in the truncated series approximating the fundamental matrix $\Phi(t = 4.0)$, the optimal matrix norm bound τ_u is several orders of magnitude sharper than the other results for the same series approximating $\Phi(t = 1.0)$. It should be noted, however, that as t increases, some or all of the $|m_{ij}|$ elements may be decreasing and, therefore,

$$|r_{ij}| \leqslant W\tau_u \leqslant 10^{-P}|m_{ij}|$$

may be more difficult to satisfy for all i, j, even though τ_u is decreasing with increasing k. (This would certainly be true in the case where $\Phi(t)$ represents the state transition matrix for a system where all the eigenvalues of A have negative real parts.)

VI. Reduction of Conservativeness for Interval Arithmetic Operations

In the present section we introduce the *nested form* for interval polynomials and the *centered form* for interval arithmetic representations. These concepts will enable us in the next section (along with the results of Section V) to reduce the conservativeness of the interval arithmetic operations used in the computation of interval matrix exponential functions. In this section we also briefly discuss *machine bounding arithmetic* in digital computers. This type of arithmetic (which gave rise to the introduction of interval analysis in the first place) will be used to implement our algorithms on a digital computer (to compute interval matrix exponential functions and to solve linear (interval) initial-value problems) in order to obtain *true* estimates of interval bounds.

1) The subdistributivity property of interval arithmetic (see [3]) points to the utilization of the *nested form* for interval polynomials. To be more specific, consider the interval polynomial (a rational interval function) with interval coefficients $A_i \in \mathscr{T}$, $i = 0,1,\cdots,n$ and interval variable $J \in \mathscr{T}_I$. The nested form of the polynomial yields an interval result contained in and frequently "narrower" than that produced using the sum of powers,

$$(\cdots(A_nJ + A_{n-1})J + \cdots + A_1)J + A_0 \subset A_nJ^n + \cdots + A_1J + A_0. \quad (40)$$

(This same property obviously holds for interval matrix polynomials. The term "narrower" is used in the sense that Moore [1, p. 7] defines the width of an interval as $w(J) \triangleq w([c,d]) \triangleq d - c \geqslant 0$.) The nested form of computations also is reasonable from the computer programmer's point of view of improving the speed of an algorithm by minimizing the number of calculations required to obtain a result or from the numerical analyst's attempt to minimize the accumulation of local rounding errors [11, pp. 51, 302] by the same technique. This is in effect a reduction in the number of occurrences of the interval variable.

As a specific example, consider the rational function (of a real variable)

$$\frac{x}{x-2} = \frac{x-2}{x-2} + \frac{2}{x-2} = 1 + \frac{2}{x-2}.$$

TABLE II
SUMMARY OF RESULTS FOR VARIOUS INTERVAL ARITHMETIC REPRESENTATIONS OF $f(x) = x - x^2,\ x \in J = [(1/2) - r, (1/2) + r]$

Computational technique	Representation		Interval result
United extension (exact range of values)	$\bar{f}(J) = \{f(x) \mid x \in J\}$		$\left[\frac{1}{4} - r^2, \frac{1}{4}\right]$
Centered form	$\frac{1}{4} - \left(J - \frac{1}{2}\right)^2 \supset \bar{f}(J)$		$\left[\frac{1}{4} - r^2, \frac{1}{4} + r^2\right]$
Nested form	$J(1 - J) \supset \bar{f}(J)$	$\left(r \leqslant \frac{1}{2}\right)$,	$\left[\left(\frac{1}{2} - r\right)^2, \left(\frac{1}{2} + r\right)^2\right]$
		$\left(r > \frac{1}{2}\right)$,	$\left[\frac{1}{4} - r^2, \left(\frac{1}{2} + r\right)^2\right]$
Sum of powers form	$J - J^2 \supset \bar{f}(J)$	$\left(r \leqslant \frac{1}{2}\right)$,	$\left[\frac{1}{4} - 2r - r^2, \frac{1}{4} + 2r - r^2\right]$
		$\left(r > \frac{1}{2}\right)$,	$\left[\frac{1}{4} - 2r - r^2, \left(\frac{1}{2} + r\right)^2\right]$
"Mean-value" form*	$\frac{1}{4} + \left(1 - 2\left(\frac{1}{2} + \left(J - \frac{1}{2}\right)[0,1]\right)\right) \cdot \left(J - \frac{1}{2}\right) \supset \bar{f}(J)$		$\left[\frac{1}{4} - 2r^2, \frac{1}{4} + 2r^2\right]$

*This result is simply included for completeness of the example by Moore and since the mean-value form will not be used further, its derivation will not be included here [1, p. 47].

The corresponding interval function results are

$$\frac{[10,12]}{[10,12]-2} = \left[1, 1\frac{1}{2}\right], 1 + \frac{2}{[10,12]-2} = \left[1\frac{1}{5}, 1\frac{1}{4}\right]$$

and

$$\left\{\frac{x}{x-2} : x \in [10,12]\right\} = \left[1\frac{1}{5}, 1\frac{1}{4}\right]$$

which also points out that in the special case where the interval variable occurs only once, the interval function and the united extension yield the same interval result.

2) The *centered form* is another method of selecting a rational interval expression which may produce a less conservative interval result which contains the corresponding united extension interval result [1, p. 42].

Suppose it is desired to calculate the range of the rational function of a real variable $f(x) = x - x^2$ for $x \in J \triangleq [(1/2) - r, (1/2) + r]$, $r \geqslant 0$. If c denotes the center or midpoint of the interval J, the real function representation desired for the centered form is $f(x) = f(c) + g(x - c)$. Obviously, $g(x - c) = -(x - (1/2))^2$ and the centered form of the interval arithmetic representation is $(1/4) - (J - (1/2))^2$, where for simplicity of notation it will be the practice to denote the degenerate intervals by the reals.

It should be clear that the centered form is an interval arithmetic representation in which the result is computed in terms of the value of the original function at the interval center plus an interval arithmetic computation which is a function of an interval variable that is symmetric and centered at zero.

Table II lists the results of the above example for the various interval representations. The table shows that for r small the mean-value form gives a result which is less conservative than the nested form but more conservative than the centered form.

Moore [1, p. 45] conjectures from the various centered form examples studied (letting f_c denote the centered form representation of f) that

$$w(f_c(J)) \leqslant w(\bar{f}(J)) + \mathcal{O}(w^2(J)) \tag{41}$$

where $\mathcal{O}(h)$ denotes the usual "order of h" notation (i.e., $|\mathcal{O}(h)/h|$ has a finite upper bound for all h when $|h|$ is less than some positive constant).

3) The interval arithmetic operations (see [3]) are accurately defined insofar as computation in the reals are concerned. Suppose, however, for purposes of discussion, that a hypothetical decimal computer can retain only one digit after each computation and that an interval result for the square of 0.899 is required. The machine representable interval [0.8, 0.9] contains this number and the square of this interval is [0.64, 0.81]. But this would be represented on the one-digit machine as [0.6, 0.8], assuming that truncation of the excess interval endpoint result digits occurs. Obviously the exact result, 0.808201, is not contained in the resulting machine interval.

This simplified analogy begins with part of the original philosophy leading to the topic of interval analysis (the inability of a finite wordlength machine to exactly represent the real numbers) and it additionally points to the requirement for a machine bounding arithmetic to successfully accomplish the numerical interval arithmetic operations. (This certainly is expected since the best that any finite wordlength computer can do is to represent a bounded subset of the rational real numbers.)

The implementations of *numerical bounding interval arithmetic* are largely machine dependent and will not be discussed here. Examples of specific computer programs which were used by the present investigators may be found in [12]–[20].

4) The computational techniques which we will subsequently employ (in the next section and in [16]) combine the centered form representations, subdivision of the interval (see [3, propositions 13 and 13M and theorems 14 and 14M in [3]) and the nested form computations discussed above. Furthermore, all of our computer programs incorporate numerical bounding interval arithmetic.

VII. Nested Centered Form Computations for the Interval Fundamental Matrix

The purpose of the present section is to reformulate the interval matrix exponential computation technique by implementing: (1) the perturbation parameter interval partitioning philosophy of theorem 14M and proposition 13M (see [3]); (2) the nested and centered form techniques; and (3) the optimal Householder matrix norm. Items (1) and (2) enable us to reduce the conservativeness of the interval arithmetic evaluations while item (3) gives us sharper estimates of error bounds for the truncated series representation of the interval matrix exponential function.

In this sense then, we return to the computational scheme previously developed in Section IV which begins with (7).

Suppose that $A \triangleq (([a_{ij}^L, a_{ij}^R]))$ is the computer representation of the input "interval" matrix and T is the degenerate interval $T = [t, t]$. Using the bounding interval arithmetic with

$$\hat{a}_{ij}^L \triangleq \left[a_{ij}^L, a_{ij}^L\right] \quad \text{and} \quad \hat{a}_{ij}^R \triangleq \left[a_{ij}^R, a_{ij}^R\right]$$

compute the interval matrices

$$G_1 \triangleq \left(\left(\frac{\hat{a}_{ij}^L + \hat{a}_{ij}^R}{2}\right)\right) \quad \text{and} \quad G_2 \triangleq \left(\left(\frac{\hat{a}_{ij}^R - \hat{a}_{ij}^L}{2}\right)\right). \tag{42}$$

(As will subsequently be seen in the interval computation (42) for G_2, in order to accommodate a "signed" interval matrix element dependence on the perturbation parameter, $a_{ij}^R < a_{ij}^L$ is allowed for the input "interval" matrix A. Here, a_{ij}^L, a_{ij}^R, and t are single-precision floating-point machine numbers.) Then

$$A_{\text{True}} \subset G_1 + \theta G_2, \qquad \text{where } \theta = [-1, 1] \tag{43}$$

where

$$A_{\text{True}} = \left(\left(\left[\min\left\{a_{ij}^L, a_{ij}^R\right\}, \max\left\{a_{ij}^L, a_{ij}^R\right\}\right]\right)\right).$$

Now let the Cauchy sequence of rational interval matrix functions $\{F_k(\theta)\}$ be defined by

$$F_k(\theta) \triangleq \sum_{i=0}^{K} \frac{(G_1 + \theta G_2)^i T^i}{i!}. \tag{44}$$

For convenience in notation and programming, using the bounding interval arithmetic recursively, compute the interval matrices for $i = 2, \cdots, K$ and $p = 0, \cdots, i$

$$G_{\left[\frac{i(i+1)}{2} + p\right]} \triangleq \left(\frac{1}{i}\right) \times \sum_{l = \left\{\begin{matrix} 1,\ p = 0, \cdots, i-1 \\ 2,\ p = i \end{matrix}\right\}}^{\left\{\begin{matrix} 1,\ p = 0 \\ 2,\ p = 1, \cdots, i \end{matrix}\right\}} G_{\left[\frac{(i-1)i}{2} + p - (l-1)\right]} G_l. \tag{45}$$

Algebraically expanding (44), substituting (45) and rearranging, obtain

$$I + (G_1 + \theta G_2)T + \left(G_3 + \theta G_4 + \theta^2 G_5\right)T^2 + \cdots + \left(G_{\left[\frac{K(K+1)}{2}\right]} + \cdots + \theta^K G_{\left[\frac{K(K+3)}{2}\right]}\right)T^K \tag{46a}$$

and then

$$I + \sum_{l=1}^{K} G_{\left[\frac{l(l+1)}{2}\right]} T^l + \sum_{p=1}^{K} \left(\sum_{l=p}^{K} G_{\left[\frac{l(l+1)}{2} + p\right]} T^l \right) \theta^p. \tag{46b}$$

Now subdivide the perturbation parameter interval $\theta = [-1, 1]$ into M "equal" width subintervals (using single-precision arithmetic),

$$\theta_i \triangleq \left[\frac{2(i-1) - M}{M}, \frac{2i - M}{M}\right] \triangleq \left[\theta_i^L, \theta_i^R\right], \qquad i = 1, \cdots, M. \tag{47}$$

Represent each subinterval in the centered interval arithmetic form (computing with single-precision arithmetic),

$$\theta_{i_c} \triangleq \left[\frac{2i - M - 1}{M}, \frac{2i - M - 1}{M}\right] \triangleq [c_i, c_i]$$

and

$$\eta_i \triangleq [-w_i, w_i], \text{ where } w_i = \max\left\{c_i - \theta_i^L, \theta_i^R - c_i\right\}. \tag{48}$$

Note that

$$\theta_i \subset \theta_{i_c} + \eta_i. \tag{49}$$

Then (46b) assumes the form

$$I + \sum_{l=1}^{K} G_{\frac{l(l+1)}{2}} T^l + \sum_{p=1}^{K} \left(\sum_{l=p}^{K} G_{\frac{l(l+1)}{2} + p} T^l \right) \cdot \sum_{j=0}^{p} C_j^p \theta_{i_c}^{p-j} \eta_i^j \tag{50}$$

where the binomial coefficients are defined by

$$C_j^p \triangleq \frac{p!}{j!(p-j)!}. \tag{51}$$

(When the computations for the binomial coefficients are actually programmed, it is simpler and more accurate to obtain these interval representations recursively completing Pascal's triangle [11, p. 53].)

Using bounding interval arithmetic computations, define the interval matrices

$$B_{K+1} \triangleq \left(\cdots \left(G_{\left[\frac{K(K+1)}{2}\right]} T + G_{\left[\frac{(K-1)K}{2}\right]} \right) T + \cdots + G_1 \right) T + I$$

$$\subset I + \sum_{l=1}^{K} G_{\left[\frac{l(l+1)}{2}\right]} T^l \tag{52a}$$

and

$$B_{K+1-p} \triangleq \left(\cdots \left(G_{\left[\frac{K(K+1)}{2}+p\right]} T + G_{\left[\frac{(K-1)K}{2}+p\right]} \right) T + \cdots + G_{\left[\frac{p(p+1)}{2}+p\right]} \right) T^p \subset \sum_{l=p}^{K} G_{\left[\frac{l(l+1)}{2}+p\right]} T^l, \quad p = 1, \cdots, K. \tag{52b}$$

Using (52a) and (52b), (50) may be rewritten as

$$B_{K+1} + \sum_{p=1}^{K} B_{K+1-p} \theta_{i_c}^p + \sum_{p=1}^{K} B_{K+1-p} \sum_{j=1}^{p} C_j^p \theta_{i_c}^{p-j} \eta_i^j. \tag{53}$$

Rearranging the last term of (53) in powers of η_i, obtain

$$\sum_{j=1}^{K} \sum_{p=j}^{K} B_{K+1-p} C_j^p \theta_{i_c}^{p-j} \eta_i^j \tag{54}$$

Using the nested computations, define the interval matrices

$$E_{K+1} \triangleq \left(\cdots \left(B_1 \theta_{i_c} + B_2 \right) \theta_{i_c} + \cdots + B_K \right) \theta_{i_c} + B_{K+1}$$

$$\subset B_{K+1} + \sum_{p=1}^{K} B_{K+1-p} \theta_{i_c}^p \tag{55a}$$

and

$$E_{K+1-j} \triangleq \left(\cdots \left(B_1 C_j^K \theta_{i_c} + B_2 C_j^{K-1} \right) \theta_{i_c} + \cdots + B_{K-j} C_j^{j+1} \right) \theta_{i_c} + B_{K-j+1} C_j^j \subset \sum_{p=j}^{K} B_{K+1-p} C_j^p \theta_{i_c}^{p-j}, \quad j = 1, \cdots, K. \tag{55b}$$

Then using (55a) and (55b), equation (53) may be rewritten in nested form as

$$\left(\cdots \left(E_1 \eta_i + E_2 \right) \eta_i + \cdots + E_K \right) \eta_i + E_{K+1} \subset E_{K+1} + \sum_{j=1}^{K} E_{K+1-j} \eta_i^j. \tag{56}$$

This is the centered form interval expression for $F_K(\theta_i)$ in (44). E_{K+1} is the nested expression for the interval center result $F_K(\theta_{i_c})$ and $(\cdots(E_1\eta_i + E_2)\eta_i + \cdots + E_K)\eta_i$ is the nested expression for the balance of the centered form, explicitly in terms of the "zero-symmetric" interval variable $\eta_i = \theta_i - \theta_{i_c}$ and indirectly in terms of the interval center variable θ_{i_c} (see (55b)). Denote the centered interval form of $F_K(\theta)$ given in (44) by

$$\tilde{F}_K\left(\theta_{i_c} + \eta_i\right) \triangleq \left(\cdots \left(E_1 \eta_i + E_2 \right) \eta_i + \cdots + E_K \right) \eta_i + E_{K+1}. \tag{57}$$

From the definition of the united extension in Propositions 6M and 9M and the relation (49), obviously

$$\bar{F}_K(\theta_i) \subset \tilde{F}_K\left(\theta_{i_c} + \eta_i\right). \tag{58}$$

For all of the numerical initial-value examples which we tried (some of which are given in [16]), bounding interval arithmetic computations have demonstrated that

$$\tilde{F}_K\left(\theta_{i_c} + \eta_i\right) \subset F_K(\theta_i). \tag{59}$$

Letting the remainder associated with $F_K(\theta_i)$ in (44) be denoted by

$$R_K(\theta_i) \triangleq \left(\left(r_{lm_K}(\theta_i) \right)\right) \triangleq \sum_{j=K+1}^{\infty} \frac{(G_1 + \theta_i G_2)^j T^j}{j!} \tag{60}$$

and letting the remainder term associated with $\tilde{F}_K(\theta_{i_c} + \eta_i)$ in (57) be denoted by

$$\tilde{R}_K\left(\theta_{i_c} + \eta_i\right) \triangleq \left(\left(\tilde{r}_{lm_K}\left(\theta_{i_c} + \eta_i\right) \right)\right) \tag{61}$$

assume also that

$$\tilde{R}_K\left(\theta_{i_c} + \eta_i\right) \subset R_K(\theta_i). \tag{62}$$

(In view of the nature of the remainder terms involved for increasing K and the less conservative interval results produced by the centered form (see (59)), this is a reasonable assumption.) Then for each l and m ($|\cdot|$ denotes the "magnitude" of an interval),

$$\left| \tilde{r}_{lm_K}\left(\theta_{i_c} + \eta_i\right) \right| \leqslant \left| r_{lm_K}(\theta_i) \right|. \tag{63}$$

Assume that

$$|G_1 + \theta_i G_2| \triangleq \left(\left(|g_{lm_1} + \theta_i g_{lm_2}| \right)\right)$$

allows the optimal Householder matrix norm $\|\cdot\|_u$ given in (29). Then if

$$\lambda_i \triangleq \|(|G_1 + \theta_i G_2|)\|_u \triangleq \|G^{-1}|G_1 + \theta_i G_2|G\|_\infty \ll \|(|G_1 + \theta_i G_2|)\|_y \tag{64}$$

where y indicates 1, 2 or ∞ and $G \triangleq \operatorname{diag}(g_p)$, $g_p > 0$, $p = 1, \cdots, n$, applying the definition of the metric σ on the set $\mathscr{T}^{n^2}$ (see [3, eqs. (10) and (11)], and applying (18), (34), and (63) for each l and m,

$$\left(\frac{g_m}{g_l}\right) \left| \tilde{r}_{lm_K}\left(\theta_{i_c} + \eta_i\right) \right| \leqslant \left(\frac{g_m}{g_l}\right) \left| r_{lm_K}(\theta_i) \right| \leqslant \tau_{u_i} \triangleq \frac{\lambda_i^{K+1}}{(K+1)!} \cdot \frac{1}{1 - \frac{\lambda_i}{K+2}} \tag{65}$$

assuming $\lambda_i/(K+2) < 1$. (In (64), the symbol $\ll$ is intended to indicate that for any l and m, $(g_l/g_m)\tau_{u_i} < \tau_{y_i}$, where y indicates the 1, 2, or ∞ matrix norm use in (65).)

Turn now to the computational relationship which is equivalent to (20) for a hexadecimal machine,

$$\begin{aligned}\left|\tilde{r}_{lm_K}\left(\theta_{i_c}+\eta_i\right)\right| &\leqslant \left(\frac{g_l}{g_m}\right)\tau_{u_i}\\ &\leqslant 16^{-P}\min\left\{\left|\tilde{f}^L_{lm_K}\left(\theta_{i_c}+\eta_i\right)\right|,\right.\\ &\qquad\left.\left|\tilde{f}^R_{lm_K}\left(\theta_{i_c}+\eta_i\right)\right|\right\} \qquad (66)\end{aligned}$$

where $(P+1)$ is one of the integers $\{1,\cdots,6\}$ which indicates the number of the single-precision hexadecimal fraction digit to which the final interval result endpoints are required to be accurate when approximating the infinite series. (Relationships (59), (62), and (65) are implicitly used in obtaining (66).)

In the actual implementation of the numerical scheme for the calculation of the matrix exponential, if (66) is satisfied for all l and m for the preselected integer P and the initial selection of K, then the final interval matrix result

$$\tilde{F}_K\left(\theta_{i_c}+\eta_i\right)+Z_i, \text{ where } Z_i \triangleq \left(\left(\left[-\left(\frac{g_l}{g_m}\right)\tau_{u_i}, \left(\frac{g_l}{g_m}\right)\tau_{u_i}\right]\right)\right) \qquad (67)$$

will contain the centered form infinite series result for

$$e^{(G_1+(\theta_{i_c}+\eta_i)G_2)T}. \qquad (68)$$

(This assumes the neglecting of possible additional accumulation of bounding errors which would result in the continuing calculation.) Furthermore, the relative interval endpoint error bounds are given by (26) and (27), except that the bounds here are computed with respect to the hexadecimal base 16.

In general, (66) will not always be satisfied for each l and m. The following discussion describes such occurrences and the programming techniques employed.

The most obvious case where (66) will never be satisfied occurs when the pth row of the original A matrix (and consequently the same rows of the G_1 and G_2 matrices) consists of degenerate zero interval elements. Then

$$\tilde{f}_{pm_K}\left(\theta_{i_c}+\eta_i\right)=\begin{cases}[0,0], & m\neq p\\ [1,1], & m=p\end{cases}. \qquad (69)$$

(A similar case occurs with respect to the pth column.) In this case, a logical test will indicate interval result degeneracy and the corresponding setting of

$$z_{pm_i}=[0,0]$$

will yield the correct result for

$$\tilde{f}_{pm_K}\left(\theta_{i_c}+\eta_i\right)+z_{pm_i}.$$

Initially, for the preselected algorithm "accuracy" (input integer P), it may happen that the estimated starting value of K is not sufficiently large to satisfy (66) for all l and m. For this reason, an automatic increase in the value of K must be programmed into the routine. However, it must be pointed out that computer storage limits place a final constraint on this technique. (In the algorithm which we used a limiting value of $K=20$ for 5×5 interval matrices results in a 46 kbyte array for $G_1,\cdots,G_1,\cdots,G_{(20)(23)/2}=G_{230}$.)

If the minimum of the two values on the right-hand side of (66) does not satisfy the relation but the maximum does, this is termed a "single-fault" and computational experience has demonstrated that $K+1$ usually results in clearing the single-fault. If additionally the maximum does not satisfy the relation, this is termed a "double-fault" and it may be necessary to "run-up" K beyond $K+1$.

Since situations may obviously occur where there is a great disparity between elements of

$$\tilde{F}_K\left(\theta_{i_c}+\eta_i\right)$$

some predetermined input judgment should be programmed into the routine so that there will be a preset limit number for the two types of faults which are allowed. This "fault acceptance" will not contradict the set containment of (68) by (67) but the relative error bounds given by relations (26) and (27) will no longer be valid. This has been incorporated into our algorithm.

This completes the development and discussion of the reformulated interval matrix exponential computation technique. To summarize, this technique uses the concepts of subdivision of the parameter interval and the centered form representations for the resulting parameter subintervals, the centered and nested form matrix interval arithmetic computations, bounding of the interval matrix function metric employing the optimally infimum Householder matrix norm (for irreducible nonnegative real matrices) and interval augmenting of the computable truncated interval matrix series for set containment of the interval matrix infinite series form of the interval matrix exponential with prescribed relative error bounds.

The linear interval integration technique described briefly following theorem 14M in [3], which will subsequently be implemented in [16] in the solution of initial-value problems, requires the computation of interval fundamental matrices for each partition subinterval. In this sense then, the above technique provides the necessary computation method.

VIII. Concluding Remarks

In the previous sections we introduced and studied "scalar" and matrix interval exponential functions. These functions are represented as infinite power series and their properties were studied in terms of rational functions obtained from truncations. To determine optimal estimates of error bounds for the truncated series representation of the exponential matrix function, we established appropriate results dealing with Householder norms. In order to reduce the conservativeness for interval arithmetic operations, we considered the nested form for interval polynomials and the centered form for interval arithmetic representations. We also discussed briefly machine bounding arithmetic in digital computers. Finally, we presented an algorithm for the computation of the interval matrix exponential function which yields prespecified error bounds. This algorithm incorporates: machine bounding arithmetic; the perturbation parameter interval partitioning philosophy of theorems 14M and proposition 13M in [3]; the nested and centered form techniques; and the optimal Householder norm.

In [16] we will use the results of [3] and of the present paper to study initial-value problems.

References

[1] R. E. Moore, *Interval Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1966.

[2] D. V. Widder, *Advanced Calculus*, Englewood Cliffs, NJ: Prentice-Hall, 1961, second edition.

[3] E. P. Oppenheimer and A. N. Michel, "Application of Interval Analysis Techniques to Linear Systems—Part I: Fundamental Results," *IEEE Circuits Syst.*, vol. CAS-35, pp. 1129–1138, Sept. 1988.

[4] A. N. Michel and C. J. Herget, *Mathematical Foundations in Engineering and Science: Algebra and Analysis*. Englewood Cliffs, NJ: Prentice-Hall, 1981.

[5] A. S. Householder, "On Norms of Vectors and Matrices," U.S. Atomic Energy Commission Rep. ORNL-1756, Physics (Oak Ridge National Lab., TN, Tennessee), 1954.

[6] ____, "On the Convergence of Matrix Iterations," U.S. Atomic Energy Commission Report ORNL-1883, Physics (Oak Ridge National Lab. TN), 1955.

[7] F. R. Gantmacher, *The Theory of Matrices*. vol. II, New York: Chelsea, 1960.

[8] R. S. Varga, "On a connection between infima of norms and eigenvalues of associated operations," *J. Linear Algebra Appl.*, vol. 6, pp. 249–256, 1973.

[9] M. Marcus and H. Minc, *A Survey of Matrix Theory and Matrix Inequalities*. Boston, MA: Allyn and Bacon, 1964.

[10] *System Analysis by Digital Computer*. (F. F. Kuo and J. F. Kaiser, Eds.), New York: Wiley, 1966.

[11] P. Henrici, *Elements of Numerical Analysis*. New York, Wiley, 1964.

[12] C. A. Clark, "Implementation of the Fortran precompiler CLUDGE, for the IBM 360/67," unpublished report, Computing Center Library, Washington State Univ., 1971.

[13] ____, "Interval arithmetic package for the IBM 360/67," unpublished report, Computing Center Library, Washington State Univ., 1971.

[14] ____, "Implementation of best possible floating point arithmetic for the IBM 360/67," unpublished report, Computing Center Library, Washington State Univ., 1971.

[15] E. P. Oppenheimer, "Application of interval analysis to problems of linear control systems," Ph.D. dissertation, Iowa State Univ. Library, Ames, IA.

[16] E. P. Oppenheimer and A. N. Michel, "Application of interval analysis techniques to linear systems—Part III: Initial-value problems" pp. 1243–1256, this issue.

[17] "High accuracy arithmetic subroutine library," IBM Program Number 5664-185, Sept. 1983.

[18] "*IBM high-accuracy arithmetic subroutine library*," Version 1, Release 3, Program Numbers 5664-185, 5665-337, and 5666-320, Apr. 1986.

[19] "*IBM high-accuracy arithmetic subroutine library, General, Information Manual*," Order No. GC33-6163-02, File No. 5370/4300-82, third ed., Apr. 1986.

[20] L. B. Rall, "An introduction to the scientific computing language PASCAL-SC," *Trans. Second Army Conf. on Applied Mathematics and Computing*, Washington, DC, May 1984.

Edward P. Oppenheimer (M'57), for a photograph and biography please see page 1138 of the September 1988 issue of this Transactions.

Anthony N. Michel (S'55–M'59–SM'79–F'82), for a photograph and biography please see page 1138 of the September 1988 issue of this Transactions.