A Study of the Matrix Exponential

Van Loan, Charles F.

2006

MIMS EPrint: **2006.397**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

ISSN 1749-9097

This EPrint is a reissue of the 1975 technical report

> Charles F. Van Loan. A study of the matrix exponential. Numerical Analysis Report No. 10, University of Manchester, Manchester, UK, August 1975.

That report—one of the first in the Numerical Analysis Report series that ran from 1974 to 2005—has been out of print for a long time, yet it is still cited, for example in [2] and [4]. Since the report contains some material that is not readily available elsewhere, it seems appropriate to re-issue it in the MIMS EPrint series that succeeded the earlier series (see [1] for a brief history of the latter). The following pages are scanned from a surviving version of the original report.

This EPrint should be cited as

> Charles F. Van Loan. A study of the matrix exponential. Numerical Analysis Report No. 10, University of Manchester, Manchester, UK, August 1975. Reissued as MIMS EPrint 2006.397, Manchester Institute for Mathematical Sciences, The University of Manchester, UK, November 2006.

For an overview of the modern e^A literature, see [4] and [3].

References

[1] Manchester Centre for Computational Mathematics. Annual report: January–December 2004. Numerical Analysis Report No. 472, Manchester Centre for Computational Mathematics, Manchester, England, September 2005.

[2] Gene H. Golub and Charles F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, Baltimore, MD, USA, third edition, 1996.

[3] Nicholas J. Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM J. Matrix Anal. Appl.*, 26(4):1179–1193, 2005.

[4] Cleve B. Moler and Charles F. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.*, 45(1):3–49, 2003.

Nicholas J. Higham
Director of MIMS
November 2006

A Study of the Matrix Exponential

Charles Van Loan[†]

Numerical Analysis Report No. 10

August 1975

[†] Department of Computer Science
Cornell University
Ithaca, New York 14850
USA

Formerly: S.R.C. Research Fellow
Department of Mathematics
Manchester University
Manchester, England

Preface

I became interested in the matrix exponential during the preparation of a talk I gave on the subject in 1974 here at Manchester. Since then I have been motivated by the work of B.N. Parlett[20]and by C.B. Moler with his "n bad ways to compute the matrix exponential $(n \geq 9)$".

Although this numerical analysis report is more analysis than numerical, I hope that it will provide a framework within which more practical research on the subject can take place. To this end I have included in the references some papers of a computational nature which, though not actually cited in the text, may be worth scrutinizing in the future.

Readers will find that the Schur decomposition figures heavily in this report. It was Parlett who, with his algorithm for computing functions of triangular matrices, convinced me that this decomposition had an important role to play in the analysis and computation of the matrix exponential. This view is consistent with one of the most basic tenets of numerical algebra, namely, *anything that the Jordan decomposition can do, the Schur decomposition can do better!*

Abstract

This report brings together a wide variety of facts concerning the matrix exponential. Against a background of familiar results, we present an analysis of matrix functions (the exponential in particular) which exploits the Schur decomposition theorem. This helps us explore the behavior of a function of a matrix whose eigensystem is poorly conditioned. Finally, we investigate Padé approximation of the matrix exponential and feature in the discussion, a potentially useful inverse error analysis.

Outline:

Notation:

$\mathbb{C}^n$ and $\mathbb{C}^{n\times n}$ stand for complex n-vectors and nxn matrices respectively. If $x \in \mathbb{C}^n$ then $\|x\| = (x^*x)^{\frac{1}{2}}$ denotes the Euclidean or 2-norm of x. Similarly, $A \in \mathbb{C}^{n\times n} \Rightarrow \|A\| = \max\limits_{0\neq x\in \mathbb{C}^n} \|Ax\|/\|x\|$. If $A = (a_{ij}) \in \mathbb{C}^{n\times n}$ then $A^* = (\bar{a}_{ji})$, $|A| = (|a_{ij}|)$, and $\|A\|_F = (\sum\limits_{i,j} |a_{ij}|^2)^{\frac{1}{2}}$. Notice that $\|A\|_F = \||A|\|_F$. For nxn matrices A and B, $|A| \leqslant |B|$ means that $|a_{ij}| \leqslant |b_{ij}|$ for all i and j and $\frac{A}{B}$ denotes AB^{-1} (provided B is invertible.).

If f(t) is differentiable, $\frac{d}{dt} f(t) = \dot{f}(t) = f^{(1)}(t)$. If higher order derivatives exist, $\frac{d^k}{dt^k} f(t) = f^{(k)}(t)$. If $A(t) = (a_{ij}(t)) \in \mathbb{C}^{n\times n}$ where t is a real variable, then $\frac{d}{dt} A(t) = A'(t) = (a'_{ij}(t))$ and $\int_a^b A(t)dt = (\int_a^b a_{ij}(t)dt)$. Similar definitions hold for vectors.

1. Functions of Matrices.

Let A be an nxn complex matrix with eigenvalues $\lambda_1, \ldots, \lambda_n$. We shall define the matrix f(A) by

$$f(A) = \frac{1}{2\pi i} \oint_{\Gamma} f(z)(zI-A)^{-1} dz \tag{1.1}$$

Here Γ consists of a finite number of simple, closed curves Γ_k with interiors Ω_k such that (a) f(z) is analytic on Γ_k and Ω_k and (b) each λ_i is contained in some Ω_k. Equation (1.1) is just the matrix version of Cauchy's integral formula and we refer the reader to Dunford and Schwartz [9 ,pp.566-577] for a discussion of it.

There are other ways of defining f(A). For example, if $f(z) = \sum_{k=o}^{\infty} a_k z^k$ and each eigenvalue of A lies inside the circle of convergence for this series, then

$$f(A) = \sum_{k=0}^{\infty} a_k A^k \quad . \tag{1.2}$$

A discussion of this power series representation can be found in MacDuffee [17] .

Some leading examples of matrix functions are

$$e^A = \sum_{k=0}^{\infty} \frac{A^k}{k!}$$

which converges for all A and

$$(I - A)^{-1} = \sum_{k=0}^{\infty} A^k$$

which converges for all A with spectral radius less than one.

Suppose $A = X B X^{-1}$ represents either the Jordan or Schur decompositions of A. As will be shown in the next two sections, it is possible to express f(B) in "closed" form. This implies that an explicit representation of f(A) can be given because

$$A = X B X^{-1} \quad \Rightarrow \quad f(A) = X f(B) X^{-1} \quad . \tag{1.3}$$

2. The Jordan Canonical Form and f(A) .

The Jordan Canonical Form (JCF) Theorem guarantees the existence of an invertible matrix X such that

$$X^{-1}A\,X \;=\; J_{m_1}(\lambda_1) \;\oplus \cdots \oplus\; J_{m_p}(\lambda_p) \;=\; J \tag{2.1}$$

where on the right we have the direct sum of Jordan "blocks":

$$J_k = J_{m_k}(\lambda_k) \;=\; \begin{bmatrix} \lambda_k & 1 & 0 & & 0 \\ 0 & \lambda_k & 1 & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ \vdots & \vdots & & \ddots & 1 \\ 0 & 0 & \cdots & 0 & \lambda_k \end{bmatrix} \qquad (m_k \times m_k) \tag{2.2}$$

The matrix X in (2.1) is not unique, *but we shall always assume that it is chosen to minimize* $\kappa(X) = \|X\|\,\|X^{-1}\|$. We say that the eigenvalue λ_i occurs with algebraic multiplicity m_i $(m_1 + \ldots + m_p = n)$. Because J has such simple structure it is possible to directly specify f(J) and hence, $f(A) = X\,f(J)\,X^{-1}$.

Theorem 1. (See MacDuffee [17])

If the JCF of A is given by (2.1) and (2.2) and if f(A) is defined by (1.1), then

$$f(A) = X\left[\, f(J_1) \oplus \cdots \oplus f(J_p)\right] X^{-1} \tag{2.3}$$

where

$$f(J_k) = \begin{bmatrix} f(\lambda_k) & f^{(1)}(\lambda_k) & \cdots & \dfrac{f^{(m_k-1)}(\lambda_k)}{(m_k-1)!} \\ 0 & f(\lambda_k) & \ddots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \ddots & f^{(1)}(\lambda_k) \\ 0 & 0 & \cdots & f(\lambda_k) \end{bmatrix} \tag{2.4}$$

As an application of Theorem 1, we prove the following result:

Theorem 2.

If the Jordan decomposition of A is given by (2.1) and (2.2) and if f(A) is defined by (1.1), then

$$\|f(A)\| \leq \kappa(X)\, m \max_{\substack{z \in \lambda(A) \\ 0 \leqslant r \leqslant m-1}} \frac{|f^{(r)}(z)|}{r!} \tag{2.5}$$

where $m = \max\{m_1, \ldots, m_p\}$.

Proof.

If $C = (c_{ij})$ is a qxq matrix, it can be shown that

$$\|C\| \leq q \max|c_{ij}| \quad .$$

Thus, from (2.4)

$$\|f(J_k)\| \leq m_k \max_{0 \leqslant r \leqslant m_k - 1} \frac{|f^{(r)}(\lambda_k)|}{r!}$$

$$\leq m \max_{\substack{0 \leqslant r \leqslant m-1 \\ z \in \lambda(A)}} \frac{|f^{(r)}(z)|}{r!}$$

The theorem now follows because from (2.3)

$$\|f(A)\| \leq \|X\| \, \|X^{-1}\| \max_{1 \leq k \leq p} \|f(J_k)\|$$

$$\leq \kappa(X)\, m \max_{\substack{0 \leqslant r \leqslant m-1 \\ z \in \lambda(A)}} \frac{|f^{(r)}(z)|}{r!}$$

Q..E. D.

3. The Schur Decomposition and f(A).

The Schur decomposition provides an interesting alternative when it comes to the specification and analysis of f(A). This decomposition states that there exists a unitary matrix Q such that

$$(3.1) \qquad Q^* A\, Q = T = \operatorname{diag}(\lambda_i) + N$$

where $T = (t_{ij})$ is upper triangular ($t_{ij} = 0$, $i > j$) and $\lambda_i = t_{ii}$ ($i=1,\ldots,n$). Since $Q^* = Q^{-1}$, (3.1) represents a similarity transformation and thus, $\lambda(A) = \{\lambda_1,\ldots,\lambda_n\}$. We also observe that the matrix N in (3.1) is <u>strictly</u> upper triangular and hence, $N^n = 0$. The matrix Q can be chosen such that the eigenvalues λ_i appear in any order. A discussion of the Schur decomposition can be found in Stewart [26] .

In order to obtain an explicit representation of f(A) using (3.1) we must investigate f(T). To motivate the general result we consider the example e^T where

$$T = \begin{bmatrix} \lambda_1 & t_{12} & t_{13} \\ 0 & \lambda_2 & t_{23} \\ 0 & 0 & \lambda_3 \end{bmatrix}$$

It can be shown that

$$e^T = \begin{bmatrix} e^{\lambda_1} & t_{12}E_{12} & t_{13}E_{13} + t_{12}t_{23}E_{123} \\ 0 & e^{\lambda_2} & t_{23}E_{23} \\ 0 & 0 & e^{\lambda_3} \end{bmatrix}$$

where $E_{ij} = \dfrac{e^{\lambda_i} - e^{\lambda_j}}{\lambda_i - \lambda_j}$ $(i < j)$ and $E_{123} = \dfrac{E_{12} - E_{23}}{\lambda_1 - \lambda_3}$

From this example we observe that e^T is an upper triangular matrix whose entries involve divided differences of the e^{λ_i} . As we shall find, similar remarks apply for general $f(T)$. To make this precise we shall need some definitions and a lemma.

Definition 1.

Suppose $f(z)$ is analytic on some open set Ω containing points $\mu_o, \ldots, \mu_k$. We denote the k-th order divided difference of f at these points by $[\mu_o, \ldots, \mu_k]$. I.e.

$$(k \geq 1) \quad [\mu_o, \ldots, \mu_k] = \frac{[\mu_o, \ldots, \mu_{k-1}] - [\mu_1, \ldots, \mu_k]}{\mu_o - \mu_k}$$

$$(k = 0) \qquad [\mu_o] = f(\mu_o)$$

If any of the μ_i's are repeated, then a limit argument can give meaning to $[\mu_o, \ldots, \mu_k]$. (See Ostrowski [19].)

Definition 2.

For $i < j$, S_{ij} denotes the set of all strictly increasing sequences of integers which begin at i and end at j:

$$S_{ij} = \{ \sigma \mid \sigma = (\sigma_o, \ldots, \sigma_k), \ i = \sigma_o < \sigma_1 < \cdots < \sigma_k = j \}$$

For example, $S_{25} = \{ (2,5), (2,3,5), (2,4,5), (2,3,4,5) \}$

Definition 3.

If $\sigma = (\sigma_o, \ldots, \sigma_k) \in S_{ij}$, then $\ell(\sigma) = k$. That is, $\ell(\sigma)$ is the "length" of σ.

Definition 4.

$$S_{ij}^{(k)} = \{ \sigma \in S_{ij} \mid \ell(\sigma) = k \} .$$

Thus, $S_{ij} = \bigcup_{k=1}^{j-i} S_{ij}^{(k)}$.

Lemma 1. (Parlett [20])

Suppose $T = (t_{ij})$ is an nxn upper triangular matrix and that $F = f(T) = (f_{ij})$ is defined by (1.1). If the diagonal entries $t_{ii} = \lambda_i$ are distinct, then

$$(3.2)\quad f_{ij} = \begin{cases} 0 & (i > j) \\ f(\lambda_i) & (i = j) \\ t_{ij}\dfrac{f_{ii} - f_{jj}}{\lambda_i - \lambda_j} + \displaystyle\sum_{k=i+1}^{j-1} \dfrac{f_{ik}t_{kj} - t_{ik}f_{kj}}{\lambda_i - \lambda_j} & (i < j) \end{cases}$$

Proof.

From (1.1) we have that $f_{ij} = \dfrac{1}{2\pi i}\oint_{\Gamma} f(z)\left[(zI - T)^{-1}\right]_{ij} dz$. Since $(zI - T)^{-1}$ is upper triangular for all $z \in \Gamma$, we have that $f_{ij} = 0$ for $i > j$. Since $\left[(zI - T)^{-1}\right]_{ii} = (z - \lambda_i)^{-1}$, we obtain $f_{ii} = f(\lambda_i)$. Thus, f_{ij} is correctly specified for all $i \geq j$.

Now assume $i < j$ and equate the (i,j) entries of the identity FT = TF which follows from (1.1). We obtain

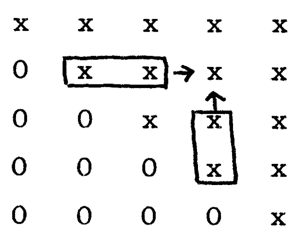
$$\sum_{k=i}^{j} f_{ik}t_{kj} = \sum_{k=i}^{j} t_{ik}f_{kj}$$

whereupon

$$f_{ij} = t_{ij}\frac{f_{ii} - f_{jj}}{\lambda_i - \lambda_j} + \sum_{k=i+1}^{j-1} \frac{f_{ik}t_{kj} - t_{ik}f_{kj}}{\lambda_i - \lambda_j}$$

Q.E.D.

Two remarks are in order. First, sense can be made of (3.2) if any of the λ_i are repeated. However, for our purposes we do not have to worry about this possibility. Second, (3.2) indicates a systematic way in which f(T) can be computed "superdiagonal at a time". We refer the reader to Parlett [20] for more details . For now we remark that an element on the p-th superdiagonal of F ($f_{i,i+p}$, $i=1,\ldots,n-p$) is a linear combination of elements from the superdiagonals 0 ,..., p-1 . Schematically:



(f_{24} is a linear combination of f_{22}, f_{23}, f_{44}, and f_{34})

This observation follows from the ($i < j$) case of (3.2)

We are now ready to give an explicit expression for f(T).

<u>Theorem 3.</u>

Let $T = (t_{ij})$ be an upper triangular matrix with $\lambda_i = t_{ii}$. Suppose f(T) is defined and given by (1.1). If $F = (f_{ij}) = f(T)$, then for $i > j$, $f_{ij} = 0$; for $i = j$, $f_{ij} = f(\lambda_i)$; and for $i < j$,

$$(3.3) \qquad f_{ij} = \sum_{(\sigma_0,\ldots,\sigma_k)\ \varepsilon\ S_{ij}} t_{\sigma_0,\sigma_1} t_{\sigma_1,\sigma_2} \cdots t_{\sigma_{k-1},\sigma_k} [\lambda_{\sigma_0},\ldots,\lambda_{\sigma_k}]$$

<u>Proof.</u>

By observations made in the proof of Lemma 1, we see that the theorem correctly specifies f_{ij} for $i \not> j$. To verify (3.3), we first assume that the λ_i are distinct. Hence, (3.2) is applicable and when we set $j = i+1$ in that formula we obtain

$$f_{ij} = t_{i,i+1} \frac{f_{ii} - f_{i+1,i+1}}{\lambda_i - \lambda_{i+1}} = t_{i,i+1} [\lambda_i , \lambda_{i+1}]$$

This shows that (3.3) is true whenever $1 = j-i$. Now assume (3.3) is true whenever $1 \leq j-i \leq p$ for some $p \geq 1$. To establish (3.3) by induction we must show that it holds whenever $1 \leq j-i \leq p+1$. Without loss of generality, it suffices to set $i = 1$, $j = n$, and $p = n - 2$ and show

$$(3.4) \qquad f_{1n} = \sum_{\sigma\ \varepsilon\ S_{ij}} t_{\sigma_0,\sigma_1} \cdots t_{\sigma_{k-1},\sigma_k} [\lambda_{\sigma_0},\ldots,\lambda_{\sigma_k}]$$

From (3.2) we have

$$(3.5) \qquad f_{1n} = t_{1n} \; \lambda_1, \lambda_n + \sum_{q=2}^{n-1} \frac{f_{1q} t_{qn} - t_{1q} f_{qn}}{\lambda_1 - \lambda_n}$$

By the inductive hypothesis,

$$(3.6) \qquad f_{1q} = \sum_{\sigma \, \varepsilon \, S_{1q}} t_{\sigma_0,\sigma_1} \cdots\cdot \; t_{\sigma_{k-1},\sigma_k} \left[\lambda_{\sigma_0}, \ldots, \lambda_{\sigma_k}\right]$$

and

$$(3.7) \qquad f_{qn} = \sum_{\tau \, \varepsilon \, S_{qn}} t_{\tau_0,\tau_1} \cdots\cdot \; t_{\tau_{k-1},\tau_k} \left[\lambda_{\tau_0}, \ldots, \lambda_{\tau_k}\right]$$

for $q = 2,\ldots,n-1$. Now in each term of (3.6) $\sigma_0 = 1$ and $\sigma_k = q$ and thus

$$(3.8) \qquad \sum_{q=2}^{n-1} f_{1q} t_{qn} = \sum_{q=2}^{n-1} \sum_{\sigma \, \varepsilon \, S_{1q}} t_{\sigma_0,\sigma_1} \cdots \; t_{\sigma_{k-1},\sigma_k} \; t_{qn} \left[\lambda_{\sigma_0}, \ldots, \lambda_{\sigma_k}\right]$$

$$= \sum_{\substack{\alpha \, \varepsilon \, S_{1n} \\ \ell(\alpha) > 1}} t_{\alpha_0,\alpha_1} \cdots \; t_{\alpha_{k-1},\alpha_k} \left[\lambda_{\alpha_0}, \ldots, \lambda_{\alpha_{k-1}}\right]$$

Similar manipulation of (3.7) gives

$$(3.9) \qquad \sum_{q=2}^{n-1} t_{1q} f_{qn} = \sum_{\substack{\alpha \, \varepsilon \, S_{1n} \\ \ell(\alpha) > 1}} t_{\alpha_0,\alpha_1} \cdots \; t_{\alpha_{k-1},\alpha_k} \left[\lambda_{\alpha_1}, \ldots, \lambda_{\alpha_k}\right]$$

Substitution of (3.8) and (3.9) into (3.5) gives

$$f_{1n} = \left[\lambda_1, \lambda_n\right] t_{1n} + \sum_{\substack{\alpha \, \varepsilon \, S_{1n} \\ \ell(\alpha) > 1}} t_{\alpha_0,\alpha_1} \cdots \; t_{\alpha_{k-1},\alpha_k} \frac{\left[\lambda_{\alpha_0}, \ldots, \lambda_{\alpha_{k-1}}\right] - \left[\lambda_{\alpha_1}, \ldots, \lambda_{\alpha_k}\right]}{\lambda_1 - \lambda_n}$$

$$= \sum_{\alpha \, \varepsilon \, S_{1n}} t_{\alpha_0,\alpha_1} \cdots \; t_{\alpha_{k-1},\alpha_k} \left[\lambda_{\alpha_0}, \ldots, \lambda_{\alpha_k}\right]$$

which completes the proof of the inductive step.

There remains the detail of repeated eigenvalues. Suppose we write $T = \mathrm{diag}(\lambda_i) + N$ where N is the strictly upper triangular portion of T ($n_{ij} = (1 - \delta_{ij})t_{ij}$). Define a sequence of upper triangular matrices T_q by

$$T_q = \mathrm{diag}(\lambda_i^{(q)}) + N$$

such that (a) $\lim T_q = T$ and (b) each T_q has distinct eigenvalues $\lambda_1^{(q)}, \ldots, \lambda_n^{(q)}$. Clearly we can choose the $\lambda_i^{(q)}$ to be interior to the contour Γ in (1.1). Thus,

$$f(T) = \frac{1}{2\pi i}\oint_\Gamma f(z)(zI - T)^{-1}dz = \lim_{q\to\infty} \frac{1}{2\pi i}\oint_\Gamma f(z)(zI - T_q)^{-1}dz$$

$$= \lim_{q\to\infty} f(T_q)$$

Another continuity argument shows that

$$\lim_{q\to\infty} \left[\lambda_{\sigma_0}^{(q)}, \ldots, \lambda_{\sigma_k}^{(q)}\right] = \left[\lambda_{\sigma_0}, \ldots, \lambda_{\sigma_k}\right]$$

for any $(\sigma_0, \ldots, \sigma_k) \in S_{ij}$ $(i < j)$. If $F_q = (f_{ij}^{(q)}) = f(T_q)$, then we clearly obtain

$$f_{ij} = \lim_{q\to\infty} f_{ij}^{(q)}$$

$$= \lim_{q\to\infty} \sum_{\sigma \in S_{ij}} t_{\sigma_0,\sigma_1} \cdots t_{\sigma_{k-1},\sigma_k} \left[\lambda_{\sigma_0}^{(q)}, \ldots, \lambda_{\sigma_k}^{(q)}\right]$$

$$= \sum_{\sigma \in S_{ij}} t_{\sigma_0,\sigma_1} \ldots, t_{\sigma_{k-1},\sigma_k} \left[\lambda_{\sigma_0}, \ldots, \lambda_{\sigma_k}\right]$$

This shows that (3.3) holds even though T might have repeated eigenvalues.

Q .E. D.

(Professor Parlett has mentioned in private correspondence that Theorem 3 was known to him although the result does not appear in [20].)

As a preliminary application of Theorem 3, we prove the following result.

Theorem 4.

Let $Q^* A Q = T = \mathrm{diag}(\lambda_i) + N$ be the Schur decomposition of A where N is the strictly upper triangular portion of T. Suppose f(A) is defined by (1.1) and that the contour Γ in that expression encloses Ω , a convex set containing the spectrum $\lambda(A)$. If, for r = 0,...,n-1

$$\sup_{z \in \Omega} |f^{(r)}(z)| = \delta_r$$

then

$$|f(T)| \leqslant \sum_{r=0}^{n-1} \frac{\delta_r |N|^r}{r!} \tag{3.10}$$

and

$$\|f(A)\|_F \leq \delta \|(I - |N|)^{-1}\|_F \tag{3.11}$$

where

$$\delta = \max_{o \leqslant r \leqslant n-1} \frac{\delta_r}{r!}$$

Proof.

Set $F = (f_{ij}) = f(T)$. Since $S_{ij} = \bigcup_{r=1}^{j-i} S_{ij}^{(r)}$ (see Definition 4) we have from (3.3)

$$f_{ij} = \sum_{r=1}^{j-i} \sum_{\sigma \in S_{ij}^{(r)}} n_{\sigma_0,\sigma_1} \cdots n_{\sigma_{r-1},\sigma_r} [\lambda_{\sigma_0},\ldots,\lambda_{\sigma_r}] \quad (i<j) \tag{3.12}$$

where $N = (n_{ij})$. Now because Ω is convex, it is possible to bound the divided differences which make up F:

$$|[\lambda_{\sigma_0} ,\ldots, \lambda_{\sigma_r}]| \leqslant \sup_{z \in \Omega} \frac{|f^{(r)}(z)|}{r!} = \frac{\delta_r}{r!} \tag{3.13}$$

(See Ostrowski [19]) . It is possible to give an explicit expression for the (i,j) entry of $|N|^r$ which we denote by $|n_{ij}|^{(r)}$. From Parlett [20] we obtain, $(r \geq 1)$

$$(3.14) \qquad |n_{ij}|^{(r)} = \begin{cases} 0 & j < i + r \\ \sum_{\sigma \in S_{ij}^{(r)}} |n_{\sigma_0,\sigma_1} \quad n_{\sigma_{r-1},\sigma_r}| & j \geq i + r \end{cases}$$

Taking absolute values in (3.12) and applying (3.13) and (3.14) for $i < j$ gives

$$|f_{ij}| \leq \sum_{r=1}^{j-i} \sum_{\sigma \in S_{ij}^{(r)}} |n_{\sigma_0,\sigma_1} \cdots n_{\sigma_{r-1},\sigma_r}| \, |[\lambda_{\sigma_0}, \ldots, \lambda_{\sigma_r}]|$$

$$\leq \sum_{r=1}^{j-i} \frac{\delta_r}{r!} |n_{ij}|^{(r)} = \sum_{r=1}^{n-1} \frac{\delta_r}{r!} |n_{ij}|^{(r)}$$

This result together with the fact that $f_{ij} = 0$ $(i > j)$ and $|f_{ij}| \leq \delta_0$ $(i = j)$ proves (3.10) .

To establish (3.11) notice from (3.10) that

$$|f(T)| < \delta \sum_{r=0}^{n-1} |N|^r = \delta(I - |N|)^{-1}$$

and consequently,

$$\|f(A)\|_F = \|Qf(T)Q^*\|_F = \|f(T)\|_F = \||f(T)|\|_F \leq \|(I - |N|)^{-1}\|_F$$

Q. E. D.

4. Definitions of the Matrix Exponential.

The exponential of an nxn complex matrix A shall be denoted by e^{At} and can be defined in a number of equivalent ways:

(4.1) $$e^{At} = \frac{1}{2\pi i} \oint_{\Gamma} e^{zt}(zI - A)^{-1}dz$$

(4.2) $$e^{At} = \sum_{k=0}^{\infty} \frac{A^k t^k}{k!}$$

(4.3) $$e^{At} = \lim_{k\to\infty} \left(I + \frac{At}{k} \right)^k$$

(4.4) $$X(t) = e^{At} \Longleftrightarrow \frac{dX}{dt} = A\,X(t) \;,\; X(0) = I$$

(4.5) $$X(t) = e^{At} \Longleftrightarrow c(D)X(t) = 0 \;;\; X^{(k)}(0) = A^k \;,\; k=0,\dots,n-1 \;.$$
$$(c(x) = \det(A - xI) \;,\; D = \frac{d}{dt} \;)$$

Formulas (4.1) and (4.2) arise directly from the definitions in Section 1. (Of course, the spectrum of A is encircled by the closed contour Γ.) Formula (4.3) is discussed in Kato [14 ,p.478ff] , (4.4) in Bellman [2 ,p.165ff] , and (4.5) in Ziebur[36] .

As (4.4)-(4.5) suggest, the matrix exponential e^{At} has a deep connection with initial value problems. This connection will be exploited in the next few sections as we investigate the properties of e^{At}.

5. Salvaging the Additive Property.

Unfortunately, not all of the properties of the scalar exponential e^{at} carry over to e^{At}. The leading example of such a property is $e^{(a+b)t} = e^{at}e^{bt}$.

Theorem 5.

If A and B are nxn matrices then $e^{(A+B)t} = e^{At}e^{Bt}$ for all t if and only if AB = BA.

Proof.

($\Longrightarrow$) Substituting the power series representations of e^{At}, e^{Bt} and $e^{(A+B)t}$ into $e^{(A+B)t} = e^{At}e^{Bt}$ and then equating the coefficients of t^2 gives

$$\tfrac{1}{2}(A + B)^2 = AB + \tfrac{1}{2}A^2 + \tfrac{1}{2}B^2$$

whence, AB = BA.

($\Longleftarrow$) If A and B commute then

$$\frac{d}{dt}\left[e^{At}e^{Bt}\right] = Ae^{At}e^{Bt} + e^{At}Be^{Bt} = (A + B)e^{At}e^{Bt} .$$

Thus, $X(t) = e^{At}e^{Bt}$ solves $X(t) = AX(t)$, $X(0) = I$ and from (4.4), this implies that $X(t) = e^{(A+B)t}$.

Q. E. D.

Corollary 1.

$e^{At}\,e^{-At} = I$ and thus, e^{At} is nonsingular for all t.

Corollary 2.

$e^{As}\,e^{At} = e^{A(s+t)}$

In the noncommutative case, relating $e^{(A+B)t}$ to e^{At} and e^{Bt} becomes complicated. However, some tractable results can be obtained by exploiting (4.4).

Theorem 6.

$e^{(A+B)t} = e^{At}\, Z(t)$ where $Z(t)$ is the unique solution to

$$Z'(t) = e^{-At} B\, e^{At}\, Z(t) \qquad (5.1)$$
$$Z(0) = I$$

Proof.

If $Z(t)$ satisfies the above initial value problem, then

$$\begin{aligned}\frac{d}{dt}\left[e^{At}Z(t)\right] &= A\, e^{At}\, Z(t) + e^{At}\, Z'(t)\\ &= A\, e^{At}\, Z(t) + e^{At}\, e^{-At} B\, e^{At}\, Z(t)\\ &= (A + B)\left[e^{At}Z(t)\right]\end{aligned}$$

By (4.4), $e^{(A+B)t} = e^{At}Z(t)$ since $\left[e^{At}Z(t)\right]_{t=0} = I$.

Q. E. D.

It is possible to get an explicit representation of $e^{(A+B)t}$ by iterating the following result:

Lemma 2.

$$e^{(A+B)t} = e^{At} + \int_0^t e^{A(t-\tau)}\, B\, e^{(A+B)\tau}\, d\tau \qquad (5.2)$$

Proof.

Differentiation of $X(t) = e^{At} + \int_0^t e^{A(t-\tau)}\, B\, e^{(A+B)\tau}\, d\tau$ reveals $\frac{d}{dt} X(t) = (A+B)X(t)$. Since $X(0) = I$, $X(t) = e^{(A+B)t}$ by (4.4) .

Q. E. D.

Theorem 7.

If $A_0(t) = e^{At}$ and $A_k(t)$ $(k \geq 1)$ is defined by

$$A_k(t) = \int_0^t \int_0^{t_1} \cdots \int_0^{t_{k-1}} e^{A(t-t_1)}B\, e^{A(t_1-t_2)} \cdots Be^{At_k}\, dt_k \cdots dt_1$$

then $e^{(A+B)t} = \sum_{k=0}^{\infty} A_k(t)$.

Proof.

Since $\|e^{As}\| \leq e^{\|As\|}$ it is easy to show $\|A_k(t)\| \leq \frac{\|Bt\|^k}{k!} \; e^{\|At\|}$.

Thus the above series converges uniformly on bounded sets of t because it is majorized

$$\left\| \sum_{k=0}^{\infty} A_k(t) \right\| \leq e^{\|At\|} \sum_{k=0}^{\infty} \frac{\|Bt\|^k}{k!} = e^{(\|A\| + \|B\|)t}$$

Thus, differentiation inside the summand is allowable and since

$$\frac{d}{dt} A_k(t) = \begin{cases} A\,A_0(t) & (k = 0) \\ \\ A\,A_k(t) + B\,A_{k-1}(t) & (k \geq 1) \end{cases}$$

we have $\frac{d}{dt} \sum_{k=0}^{\infty} A_k(t) = (A + B) \sum_{k=0}^{\infty} A_k(t)$. The Theorem now follows from (4.4) because $\sum_{k=0}^{\infty} A_k(0) = I$.

Q.E.D.

Lemma 2 can be found in Bellman [2] . Theorems 6 and 7 can essentially be found in Gantmacher [12] where he discusses the "matrizer". The analysis there is in the more general setting of the variable coefficient problem A(t)X(t) = X(t).

We conclude with some rather different approaches to the problem of salvaging the addition law. Trotter [29] proves the following "product formula":

$$(5.3) \qquad e^{(A+B)t} = \lim_{k \to \infty} \left[e^{At/k} \, e^{Bt/k} \right] k$$

When A and B commute, (5.3) of course gives $e^{(A+B)t} = e^{At}e^{Bt}$.

When A and B do not commute, it is not surprising that the commutator [A,B] becomes involved in expressions for $e^{(A+B)t}$ and $e^{At}e^{Bt}$. Two results confirming this are the Campbell-Baker-Hausdorff formula

(5.4) $$e^{At}\, e^{Bt} = e^{C(t)}$$

where

(5.5) $$C(t) = (A+B)t + \tfrac{1}{2}[A,B]\, t^2 + \frac{1}{12}\Big([[A,B], B] - [[A,B] , A]\Big)\, t^3 + \dots$$

and the related Zassenhaus formula

(5.6) $$e^{(A+B)t} = e^{At} e^{Bt}\, e^{C_2 t^2}\, e^{C_3 t^3}\, e^{C_4 t^4} \cdots$$

where

(5.7) $$C_2 = -\tfrac{1}{2}[A,B] \quad , \quad C_3 = \frac{1}{6}\,[[A,B],A] + \frac{1}{3}\,[[A,B], B] \quad , \quad C_4 = \dots.$$

We refer the reader to Weiss and Maradudin[30] for a complete specification of C(t) in (5.5) and to Bellman [3 ,p.36] for a technique which can be used to derive the matrices C_2, C_3 ,... in (5.6) .

6. The Growth of e^{At}

For the scalar exponential $e^{\mu t}$ $(\mu \varepsilon C)$ we have

$$(6.1)\qquad \sup_{t\geq 0} |e^{\mu t}| = 1 \quad \Longleftrightarrow \quad \mathrm{Re}(\mu) \leq 0$$

In this section we prove a corresponding result for e^{At} . To this end define the scalars $\alpha(A)$ and $\mu(A)$ by

$$(6.2)\qquad \alpha(A) = \max\{\mathrm{Re}(\lambda) \mid \det(A - \lambda I) = 0\}$$

$$(6.3)\qquad \mu(A) = \max\left\{ \lambda \mid \det\left(\frac{A^* + A}{2} - \lambda I\right) = 0 \right\}$$

These two quantities obey the following inequality:

$$(6.4)\qquad -\mu(-A) \leq \alpha(A) \leq \mu(A)$$

This follows directly from Rayleigh quotient theory and the fact that if $Ax = \lambda x$ with $\mathrm{Re}(\lambda) = \alpha(A)$ and $\|x\| = 1$, then $\alpha(A) = \frac{1}{2}x^*(A^* + A)x$.

The scalar $\mu(A)$ is an example of a "logarithmic norm" , a concept which is useful in the study of errors which arise during the numerical solution of systems of ordinary differential equations. The reader should consult Dahlquist [8] and Strom [27,28] for a discussion of $\mu(A)$ in this connection.

Now as with the scalar case,

$$(6.5)\qquad \lim_{t\to\infty} \|e^{At}\| = 0 \qquad\qquad \alpha(A) < 0$$

This can be proven, for example, by using (7.1). Unlike the scalar case however, $\sup_{t\geq 0} \|e^{At}\|$ may be strictly greater than 1 even though $\alpha(A)$ is negative. For example,

$$A = \begin{pmatrix} -1 & 10(e^{-1} - e^{-2})^{-1} \\ 0 & -2 \end{pmatrix} \Rightarrow e^A = \begin{pmatrix} e^{-1} & 10 \\ 0 & e^{-2} \end{pmatrix}$$

and thus, $\sup_{t\geq 0} \|e^{At}\| \geq \|e^A\| > 10$ even though $\alpha(A) = -1 < 0$

Since $\|I\| = 1$, we always have $\sup_{t\geqslant 0}\|e^{At}\| \geqslant 1$. The following result indicates exactly when we have equality and thus constitutes a generalization of (6.1).

Theorem 8.

$$\sup_{t\geqslant 0}\|e^{At}\| = 1 \quad\Longleftrightarrow\quad \mu(A) \leqslant 0 \tag{6.6}$$

Proof.

For any unit vector $v \in \mathbf{C}^n$, define the functional $\phi_v(t)$ by

$$\phi_v(t) = \|e^{At}v\|^2 = v^* e^{A^*t} e^{At} v \qquad (t \geqslant 0) \tag{6.7}$$

and notice that

$$\phi_v'(t) = (e^{At}v)^*(A^* + A)(e^{At}v) \tag{6.8}$$

Now,

$$\begin{aligned}
\mu(A) \leqslant 0 &\Rightarrow y^*(A^* + A)y \leqslant 0 && \text{all } y \in \mathbf{C}^n \\
&\Rightarrow \phi_v'(t) \leqslant 0 && \text{all unit } v \in \mathbf{C}^n,\ t\geqslant 0 \\
&\Rightarrow \phi_v(t) \leqslant 1 && \text{all unit } v \in \mathbf{C}^n,\ t\geqslant 0 \\
&\Rightarrow \|e^{At}\|^2 = \sup_{\|v\|=1}\phi_v(t) \leqslant 1 && t \geqslant 0
\end{aligned}$$

(The third line follows from the fact that $\phi_v(0) = 1$.) The converse follows with comparable ease:

$$\begin{aligned}
\sup_{t\geqslant 0}\|e^{At}\| = 1 &\Rightarrow \phi_v(t) \leqslant 1 && \text{all unit } v \in \mathbf{C}^n,\ t \geqslant 0 \\
&\Rightarrow \phi_v'(0) \leqslant 0 && \text{all unit } v \in C^n \\
&\Rightarrow v^*(A^* + A)v \leqslant 0 && \text{all unit } v \in C^n \\
&\Rightarrow \mu(A) \leqslant 0
\end{aligned}$$

Q.E.D.

7. Representations of e^{At}

(a) e^{At} and the Jordan Canonical Form.

If (2.1) and (2.2) represent the JCF of A then the application of (2.3) and (2.4) to the exponential gives

$$(7.1)\qquad e^{At} = X\left\{e^{J_1 t} \oplus \qquad \oplus\ e^{J_k t}\right\}X^{-1}$$

where

$$(7.2)\qquad e^{J_k t} = e^{\lambda_k t}\begin{bmatrix} 1 & t & \frac{t^2}{2!} & \cdots & & & \frac{t^{m_k-1}}{(m_k-1)!} \\ 0 & 1 & t & \ddots & & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \ddots & & \vdots \\ \vdots & \vdots & \vdots & & 1 & t & \frac{t^2}{2!} \\ 0 & 0 & 0 & \cdots & 0 & 1 & t \\ 0 & 0 & 0 & \cdots & 0 & 0 & 1 \end{bmatrix}$$

(b) e^{At} and the Schur Canonical Form.

We can certainly apply the results of Section 3 and express e^{At} in terms of e^{Tt} where T is the Schur Canonical Form of A. Instead, we shall obtain a slightly more elegant representation of e^{At} by using Theorem 7.

Theorem 9.

Let $Q^*AQ = T = D + N$ be the Schur decomposition (3.1) of A with D and N the diagonal and strictly upper triangular portions of T respectively. If

$$(7.3)\qquad T_k(t) = \begin{cases} e^{Dt} & (k = 0) \\ \int_0^t \cdots \int_0^{t_{k-1}} e^{D(t-t_1)}Ne^{D(t_1-t_2)}\cdots Ne^{Dt}\, dt_k \cdots dt_1 & (k \geq 1) \end{cases}$$

then

$$(7.4) \qquad e^{At} = Q\left(\sum_{k=0}^{n-1} T_k(t)\right)Q^*$$

Proof.

From Theorem 7 $(A \equiv D\ ,\ B \equiv N)$ we clearly have

$$e^{Tt} = e^{(D+N)t} = \sum_{k=0}^{\infty} T_k(t)$$

However, the product of n or more nxn strictly upper triangular matrices is zero and the integrand of $T_k(t)$ $(k \geq n)$ is just such a product:

$$\left[e^{D(t-t_1)}N\right]\left[e^{D(t_1-t_2)}N\right]\cdots\quad\cdot\cdot\left[e^{D(t_{k-1}-t_k)}N\ e^{Dt_k}\right]$$

Thus,

$$e^{At} = Q\ e^{Tt}\ Q^* = Q\left(\sum_{k=0}^{n-1} T_k(t)\right)Q^*$$

Q. E. D.

It is possible to relate the matrices $T_k(t)$ to divided differences involving the function $f(z) = e^{zt}$.

Corollary.

$$(7.5) \quad [T_k(t)]_{ij} = \begin{cases} 0 & (i+k > j) \\ \displaystyle\sum_{\sigma \in S_{ij}^{(k)}} n_{\sigma_0,\sigma_1} \cdots n_{\sigma_{k-1},\sigma_k}\left[\lambda_{\sigma_0},\ldots,\lambda_{\sigma_k}\right] & (i+k \leq j) \end{cases}$$

where the eigenvalues of A (and T) are given by $\lambda_1, \ldots, \lambda_n$ and the divided differences in (7.5) are with respect to the function $f(z) = e^{zt}$.

We delete the proof of this corollary since it essentially involves the same techniques used in the proof of Theorem 3.

(c) e^{At} as a Polynomial in A.

The Jordan and Schur decompositions have enabled us to express e^{At} explicitly through the canonical forms J and T. Quite a different representation can be obtained by expressing e^{At} as a polynomial in A having analytic coefficients in t. Such a representation is possible because of the Cayley-Hamilton Theorem:

$$c(\lambda) = \det(A - \lambda I) \Rightarrow c(A) = 0 \tag{7.6}$$

By using (7.6) it is possible to show that A^k is a (not necessarily unique) linear combination of I, A, ..., A^{n-1}, say

$$A^k = \sum_{j=0}^{n-1} c_{kj} A^j \qquad k=0,1,\ldots \tag{7.7}$$

By substituting (7.7) into (4.2) it can be shown that

$$e^{At} = \sum_{k=0}^{n-1} \phi_k(t) A^k \tag{7.8}$$

where

$$\phi_k(t) = \sum_{j=0}^{\infty} c_{kj} t^j \tag{7.9}$$

The details of this analysis can be found in Mirsky [18]. One can, in fact, work with the minimum polynomial of A(instead of $c(\lambda)$) when deriving a polynomial representation of e^{At}, but this is not necessary for our purposes.

Putzer [22] has developed a way of selecting the functions $\phi_k(t)$ in (7.8). He has shown that if $c(x) = \det(A - xI) = \sum_{k=0}^{n-1} c_k x^k$ and

$$\begin{aligned} c(D)z(t) &= 0 & \left(D \equiv \frac{d}{dt}\right) \\ z^{(k)}(0) &= 0 & k = 0, \ldots, n-2 \\ z^{(n-1)}(0) &= 1 \end{aligned}$$

then

$$\begin{bmatrix} \phi_o(t) \\ \phi_1(t) \\ \vdots \\ \phi_{n-1}(t) \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & \cdots & c_{n-1} & 1 \\ c_2 & c_3 & \cdots & 1 & 0 \\ \vdots & & \diagup & 0 & 0 \\ & \diagup & & \vdots & \vdots \\ c_{n-1} & 1 & \cdots & 0 & 0 \\ 1 & 0 & \cdots & 0 & 0 \end{bmatrix} \begin{bmatrix} z(t) \\ z^{(1)}(t) \\ \vdots \\ z^{(n-2)}(t) \\ z^{(n-1)}(t) \end{bmatrix}$$

Suppose the matrices A_o ,..., A_{m-1} span the same subspace of C^{nxn} as I , A , ... , A^{n-1} . Since e^{At} is a linear combination of I , A , ..., A^{n-1} , it must also be a linear combination of the A_i :

$$e^{At} = \sum_{k=0}^{n-1} \Psi_k(t) A_k$$

By choosing the A_i judiciously, the coefficients $\Psi_k(t)$ can be easily specified.

For example, Putzer[22] has shown that if λ_1 ,..., λ_n are the eigenvalues of A, then

$$e^{At} = \sum_{j=0}^{n-1} r_{j+1}(t) \ P_j(t)$$

where

$$P_o = I$$

$$P_j = \prod_{k=1}^{j} (A - \lambda_k I) \qquad j = 1 ,,,, n-1$$

and

$$\begin{bmatrix} \dot{r}_1(t) \\ \dot{r}_2(t) \\ \vdots \\ \dot{r}_n(t) \end{bmatrix} = \begin{bmatrix} \lambda_1 & 0 & 0 & \cdots & 0 \\ 1 & \lambda_2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & & \vdots \\ 0 & \cdots & & 1 & \lambda_n \end{bmatrix} \begin{bmatrix} r_1(t) \\ r_2(t) \\ \vdots \\ r_n(t) \end{bmatrix}$$

with $r_1(0) = 1$ and $r_j(0) = 0$, $j = 2,\ldots,n$.

Similarly, Kirchner 15 has proved that

$$e^{At} = \left[q(A)\right]^{-1} \sum_{i=1}^{k} p_i(A)\, f_{s_i}\left[(A - \lambda_i I)t\right] e^{\lambda_i t}$$

where A has distinct eigenvalues $\lambda_1, \ldots, \lambda_k$ with multiplicities $s_1, \ldots, s_k$ and

$$f_r(x) = 1 + x + \frac{x^2}{2!} + \ldots + \frac{x^{r-1}}{(r-1)!}$$

$$p_j(\lambda) = \prod_{\substack{i=1 \\ i \neq j}}^{k} (\lambda - \lambda_i)^{s_i}$$

$$q(\lambda) = p_1(\lambda) + \ldots + p_k(\lambda)$$

One can also use interpolating formulae. As shown in MacDuffee [17] if $\phi(z)$ interpolates e^z at A's eigenvalues (multiplicities included), then $\phi(At) = e^{At}$. Expressing ϕ in Lagrangian form gives

$$e^{At} = \sum_{k=1}^{n} \prod_{\substack{j=1 \\ j \neq k}}^{n} \left(\frac{A - \lambda_j I}{\lambda_k - \lambda_j}\right) e^{\lambda_k t}$$

while in Newtonian form we obtain

$$e^{At} = e^{\lambda_1 t} I + \sum_{k=2}^{n} \prod_{i=2}^{k-1} (A - \lambda_i I) \left[\lambda_1, \ldots, \lambda_k\right]$$

(Here the divided differences are with respect to the function $f(z) = e^{zt}$.)

We conclude this section by mentioning the work of Fulmer [10] whose representation of e^{At} comes from an exploitation of Ziebur's result (4.5). For clarity we illustrate the main idea when A has distinct eigenvalues $\lambda_1, \ldots, \lambda_n$. In this case the general solution of the differential equation in (4.5) is given by

$$G(t) = C_1 e^{\lambda_1 t} + \ldots + C_n e^{\lambda_n t}$$

The initial conditions

$$G^{(k)}(0) = A^k \qquad (k=0,\ldots,n-1)$$

give us the requisite number of (linear) equations in order to solve for the unknown matrices $C_1, \ldots, C_n$. In tensor language we in fact have

$$(V \otimes I) \begin{bmatrix} C_1 \\ C_2 \\ \\ C_n \end{bmatrix} = \begin{bmatrix} I \\ A \\ \\ A^{n-1} \end{bmatrix}$$

where the tensor product is defined by

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdot\cdot & a_{1n}B \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a_{n1}B & a_{n2}B & \cdot\cdot & a_{nn}B \end{bmatrix} \qquad (n^2 x\ n^2)$$

and V is the Vandermonde matrix

$$V = \begin{bmatrix} 1 & 1 & \cdot\cdot & 1 \\ \lambda_1 & \lambda_2 & \cdot\cdot & \lambda_n \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdot\cdot & \lambda_n^{n-1} \end{bmatrix}$$

Since $(V \otimes I)^{-1} = V^{-1} \otimes I$ we have

$$\begin{bmatrix} C_1 \\ C_2 \\ \cdot \\ \cdot \\ \cdot \\ C_n \end{bmatrix} = \left(V^{-1} \otimes I\right) \begin{bmatrix} I \\ A \\ \cdot \\ \cdot \\ \cdot \\ A^{n-1} \end{bmatrix}$$

We refer the reader to Fulmer for a discussion of the multiple eigenvalue case. Not surprisingly, confluent Vandermonde matrices are involved.

8. Bounds for $\|e^{At}\|$

In the preceeding sections we have seen that $\|e^{At}\|$ $(t \geq 0)$ behaves initially like $e^{\mu(A)t}$ and asymptotically like $e^{\alpha(A)t}$. In this section we add to our knowledge of $\|e^{At}\|$ by obtaining several different bounds of the form

$$\|e^{At}\| \leq e^{\beta t} M(t) \quad . \tag{8.1}$$

Among other things, these bounds will enable us to present perturbation theorems in the next section.

Our first results are obtained by applying the Jordan and Schur decompositions.

Theorem 10.

If the Jordan decomposition of A is given by (2.1) and (2.2) then

$$\|e^{At}\| \leq e^{\alpha(A)t} \left(m\, \kappa(X) \max_{0 \leq r \leq m-1} \frac{t^r}{r!} \right) \tag{8.2}$$

where $m = \max \{m_1, \ldots, m_p\}$ and $\kappa(X) = \|X\| \|X^{-1}\|$.

Proof.

Applying (2.5) with $f(z) = e^{zt}$ gives

$$\|e^{At}\| \leq \kappa(X)\; m \max_{\substack{z \in \lambda(A) \\ 0 \leq r \leq m-1}} \frac{t^r e^{zt}}{r!} \leq \kappa(X)\; m\; e^{\alpha(A)t} \max_{0 \leq r \leq m-1} \frac{t^r}{r!}$$

Q. E. D.

Theorem 11.

If the Schur decomposition of A is given by (3.1) then for $t \geq 0$

$$\|e^{At}\| \leq e^{\alpha(A)t} \sum_{k=0}^{n-1} \frac{\|N\|^k t^k}{k!} \tag{8.3}$$

and

$$\|e^{At}\| \leq \|e^{At}\|_F \leq e^{\alpha(A)t} \|e^{|N|t}\|_F \quad . \tag{8.4}$$

Proof.

By taking norms in (7.3) and using the fact that $\|e^{Ds}\| = e^{\alpha(A)s}$ $(s \geq 0)$, we obtain

$$\|T_k(t)\| \leqslant \int_0^t \cdots \int_0^{t_{k-1}} \| e^{D(t-t_1)}\| \, \|N\| \cdots \|N\| \, \|e^{Dt_k}\| \, dt_k \cdots dt_1$$

$$= \|N\|^k \frac{t^k}{k!} \quad e^{\alpha(A)t}$$

(The same result holds for $k = 0$.) From (7.4) we thus have

$$\|e^{At}\| \leqslant \sum_{k=0}^{n-1} \|T_k(t)\| \leqslant e^{\alpha(A)t} \sum_{k=0}^{n-1} \|N\|^k \frac{t^k}{k!}$$

proving (8.3).

To establish (8.4), just apply Theorem 4 with $\delta_r = e^{\alpha(A)t} \dfrac{t^r}{r!}$

This gives

$$|e^{Tt}| \leq \sum_{k=0}^{n-1} t^k \frac{|N|^k}{k!} \quad e^{\alpha(A)t} = e^{\alpha(A)t} \, e^{|N|t}$$

whereupon

$$\| e^{At}\| \leq \|e^{At}\|_F = \|e^{Tt}\|_F = \| |e^{Tt}| \|_F \leqslant e^{\alpha(A)t} \| e^{|N|t}\|_F$$

Q.E.D.

For matrices with ill conditioned eigensystems $\kappa(X)$ may be extremely large (see Wilkinson[31 ,p.87ff.]). This fact leads one to believe that for such matrices, the bounds (8.3 and (8.4 are superior to the bound (8.2). We shall show later on that this is not always the case.

We now specify upper and lower bounds for $\|e^{At}\|$ involving the constant $\mu(A)$ defined in (6.3).

Theorem 12.

(8.5) $$\|e^{At}\| \leqslant e^{\mu(A)t}$$ (DAhlquist [8])

(8.6) $$\|e^{At}\| \geqslant e^{-\mu(-A)t}$$ (Copple [6])

Proof.

If $\phi_v(t) = \|e^{At}v\|^2$ ($\|v\|^2 = 1$) , then from (6.8) we have

$$\phi_v'(t) \leqslant 2\,\mu(A)\,\phi_v(t)$$

since $2\,\mu(A)$ is the most positive eigenvalue of $A^* + A$. Thus,

$$\phi_v(t) = \|e^{At}\|^2 \leq e^{2\mu(A)t}$$

Inequality (8.5) now follows since the above result holds for all unit vectors v. To prove (8.6), just observe that

$$1 = \|e^{At}\,e^{-At}\| \leq \|e^{At}\|\,\|e^{-At}\| \leq \|e^{At}\|\,e^{\mu(-A)t}$$

Q.E.D.

We remark in passing that the lower bound in (8.6) is inferior to the obvious result

(8.7) $$\|e^{At}\| \geqslant e^{\alpha(A)t}$$

(This follows from the fact that $\|C\| \geqslant \max\{\,|\lambda|\ \mid \det(C - \lambda I)=0\,\}$ for any square matrix C and $\alpha(A) \geqslant -\mu(-A)$.)

It is possible for $\mu(A)$ to be positive even though $\alpha(A)$ is negative as the example $A = \begin{pmatrix} -1 & 4 \\ 0 & -1 \end{pmatrix}$ shows. ($\alpha(A) = -1$, $\mu(A) = 1$.) With such examples, the bound in (8.6) becomes less meaningful as t increases. It is possible to somewhat correct this situation as the following theorem will show .

Theorem 13.

For any invertible matrix S

$$\|e^{At}\| \leq \kappa(S) \quad e^{\mu(S^{-1}AS)} \qquad . \tag{8.8}$$

For any $\varepsilon > 0$ there exists an invertible matrix S such that

$$\mu(S^{-1}AS) \leq \alpha(A) + \varepsilon \tag{8.9}$$

Hence, if $\alpha(A) < 0$, it is possible to choose S such that the upper bound in (8.8) decays as t increases.

Proof.

From (8.6) $\|e^{S^{-1}ASt}\| \leq e^{\mu(S^{-1}AS)t}$ and thus

$$\|e^{At}\| = \|S\, e^{S^{-1}ASt}\, S^{-1}\| \leq \|S\|\,\|S^{-1}\|\,\|e^{S^{-1}ASt}\| \leq \kappa(S) \quad e^{\mu(S^{-1}AS)}$$

establishing (8.8).

To prove (8.9), let $Q^*AQ = \text{diag}(\lambda_i) + N$ be the Schur decomposition (3.1) of A. Since N is strictly upper triangular, it is possible to find a diagonal matrix $T = \text{diag}(1 , \Theta , \ldots , \Theta^{n-1})$ such that

$$\|T^{-1}\, N\, T\| \leq \varepsilon .$$

For example, if $\Theta = \min\left\{1 , \dfrac{\varepsilon}{\|N\|_F}\right\}$ then

$$\|T^{-1}NT\|^2 \leq \|T^{-1}NT\|_F^2 = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} n_{ij}^2 \; \Theta^{2(j-i)} \leq \Theta^2 \|N\|_F^2 \leq \varepsilon^2$$

From Dahlquist [8] we have the properties

$$\mu(A + B) < \mu(A) + \mu(B)$$

$$|\mu(A)| \leqslant \|A\|$$

and thus if $S = QT$ we have

$$\mu(S^{-1}AS) = \mu(\mathrm{diag}(\lambda_i) + T^{-1}NT) < \mu(\mathrm{diag}(\lambda_i)) + \mu(T^{-1}NT)$$

$$\leqslant \alpha(A) + \|T^{-1}NT\| < \alpha(A) + \varepsilon$$

Q.E.D.

We refer the reader to the work of Strom [27] for a more detailed discussion of log norms and related results like (8.8) and (8.9).

We next present an upper bound for $\|e^{At}\|$ by using the definition (4.3). This bound may be found in Kato [14 ,Chapter 10] where the discussion takes place in a considerably more general setting than just nxn matrices.

Theorem 14

Let $\beta > \alpha(A)$ and suppose there exists a constant $M > 0$ such that for every γ, $\mathrm{Re}(\gamma) \geqslant \beta$, we have

$$\|(\gamma I - A)^{-k}\| \leqslant M|(\gamma - \beta)^{-k}|$$

for all k greater than some $k_o = k_o(\gamma)$. Then

$$\|e^{At}\| \leq M e^{\beta t}$$

Proof.

Fix t and choose k_o such that $\frac{k_o}{t} \geqslant \beta$. For all $k > k_o$ we have

$$\left\|\left(I - \frac{At}{k}\right)^{-k}\right\| = \left(\frac{t}{k}\right)^{-k} \left\|\left(\frac{k}{t} I - A\right)^{-k}\right\| \leq \left(\frac{t}{k}\right)^{-k} M \left(\frac{k}{t} - \beta\right)^{-k} = M(1 - \frac{\beta t}{k})^{-k}$$

From (4.3) we thus have

$$\|e^{At}\| = \lim_{k \to \infty} \left\|\left(I - \frac{At}{k}\right)^{-k}\right\| \leq M \lim_{k \to \infty} \left(1 - \frac{\beta t}{k}\right)^{-k} = M e^{\beta t} .$$

Q.E.D.

Theorem 14 is not a particularly useful result in that it replaces one difficult problem (that of bounding $\|e^{At}\|$) with an equally difficult problem (that of determining M and β above). We mention in passing, however, that this theorem points to the deep connection between the exponential and the resolvent function $R(\gamma) = (\gamma I - A)^{-1}$. This connection is expressed by the following identity:

$$(\gamma I - A)^{-1} = \int_0^\infty e^{(A - \alpha I)t}\, dt \qquad \mathrm{Re}(\gamma) > \alpha(A)$$

We shall not pursue the matter further.

We conclude this section by contrasting some of the bounds which have thus far been presented. Consider the matrix

$$A = \begin{pmatrix} -1 + \delta & 4 \\ 0 & -1 - \delta \end{pmatrix} \qquad \delta = 10^{-6}$$

The Schur and Jordan results (8.3) and (8.2) give us respectively

$$\text{(a)} \qquad \|e^{At}\| \leqslant e^{(-1+\delta)t}\,(1 + 4t)$$

$$\text{(b)} \qquad \|e^{At}\| \leqslant e^{(-1+\delta)t}\,(4 * 10^6)$$

On the other hand, since $\mu(A) = -1 + (4 + \delta^2)^{\frac{1}{2}}$ and $\mu(S^{-1}AS) = -1 + (.25 + \delta^2)^{\frac{1}{2}}$ ($S = \mathrm{diag}(4,1)$) , we have from (8.5) and (8.6) respectively

$$\text{(c)} \qquad \|e^{At}\| \leqslant e^{[-1 + (4 + \delta^2)^{\frac{1}{2}}]t}$$

$$\text{(d)} \qquad \|e^{At}\| \leqslant 4e^{[-1 + (.25 + \delta^2)^{\frac{1}{2}}]t}$$

The following table compares theee bounds for $t = 0,1,\ldots,30$. We notice that in this example the constant $\kappa(X)$ in (8.2) is large making the bound (b) considerably weak. This is in contrast to the Schur result (a) which provides very accurate upper bounds as the table indicates. The log norm results (c) and (d) are also interestèng. Because $\mu(A)$ is positive, the exponential bound (c) deteriorates with increasing t, but by performing the indicated similarity transformation $S^{-1}AS$ above, we can derive a decaying bound (d).

t	$\|e^{At}\|$	Schur (a)	Jordan (b)	Log Norm (c)	Log Norm (d)
0	.100000F+01	.100000E+01	.400000E+07	.100000F+01	.400000E+01
	.155836F+01	.183940E+01	.147152E+07	.271828F+01	.242612E+01
	.109934E+01	.121302E+01	.541342E+06	.738906F+01	.147152E+01
	.601567F+00	.647234E+00	.199149E+06	.200855F+02	.892521E+00
	.294192E+00	.311367E+00	.732628E+05	.545982F+02	.541341E+00
	.135096E+00	.141498E+00	.269519E+05	.148413F+03	.328340E+00
	.595935E-01	.619692E-01	.991507E+04	.403429F+03	.199148E+00
	.255654E-01	.264448E-01	.364755E+04	.109663F+04	.120790E+00
	.107454E-01	.110704E-01	.134136E+04	.298096F+04	.732626E-01
	.444622E-02	.456620E-02	.493644E+03	.810308F+04	.444360E-01
	.181715E-02	.186142E-02	.181602E+03	.220265F+05	.269518E-01
	.735202E-03	.751535E-03	.668075E+02	.598741F+05	.163471E-01
	.295054E-03	.301070E-03	.245771E+02	.162755F+06	.991501E-02
	.117502E-03	.119709E-03	.904144E+01	.442413F+06	.601376E-02
	.465811E-04	.473978E-04	.332616E+01	.120260F+07	.364753E-02
	.183595E-04	.186603E-04	.122363E+01	.326902F+07	.221234E-02
	.720412E-05	.731490E-05	.450148E+00	.888611F+07	.134185E-02
	.281561E-05	.285661E-05	.165600E+00	.241550F+08	.813873E-03
	.109679E-05	.111131E-05	.609210E-01	.656600F+08	.493639E-03
	.425894E-06	.431424E-06	.224116E-01	.178482F+09	.299407E-03
	.164921E-06	.166957E-06	.824478E-02	.485165F+09	.181600E-03
	.637039E-07	.644531E-07	.303309E-02	.131882F+10	.110146E-03
	.245510E-07	.248268E-07	.111531E-02	.358491F+10	.668068E-04
	.944226E-08	.954377E-08	.410435E-03	.974480F+10	.405204E-04
	.362461E-08	.366197E-08	.151009E-03	.264891F+11	.245768E-04
	.138897E-08	.140272E-08	.555532E-04	.720049F+11	.149066E-04
	.531408E-09	.536468E-09	.204369E-04	.195730F+12	.904132E-05
	.203012E-09	.204374E-09	.751832E-05	.532048F+12	.548384E-05
	.774496F-10	.781349E-10	.276584E-05	.144626F+13	.332611E-05
	.295096E-10	.297613E-10	.101750E-05	.393133F+13	.201739E-05
3	.112303F-10	.113231E-10	.374316E-06	.106865F+14	.122361E-05

The Schur result (8.3) does not always produce a bound as superior as the above example would indicate. For example, if

$$A = \begin{bmatrix} -1 & -6 & 4 & -6 & 4 & -6 & 4 & -6 & 4 & -6 & 4 & -6 \\ 0 & -7 & 5 & -7 & 5 & -7 & 5 & -7 & 5 & -7 & 5 & -7 \\ 0 & 0 & -2 & -6 & 4 & -6 & 4 & -6 & 4 & -6 & 4 & -6 \\ 0 & 0 & 0 & -8 & 5 & -7 & 5 & -7 & 5 & -7 & 5 & -7 \\ 0 & 0 & 0 & 0 & -3 & -6 & 4 & -6 & 4 & -6 & 4 & -6 \\ 0 & 0 & 0 & 0 & 0 & -9 & 5 & -7 & 5 & -7 & 5 & -7 \\ 0 & 0 & 0 & 0 & 0 & 0 & -4 & -6 & 4 & -6 & 4 & -6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & -10 & 5 & -7 & 5 & -7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -5 & -6 & 4 & -6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -11 & 5 & -7 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -6 & -6 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -12 \end{bmatrix}$$

then it can be shown that $\| e^A \| \approx .57$. However, from (8.3) we obtain

$$\| e^A \| \lesssim e^{-1} \; (2.0 * 10^{10})$$

whereas (8.2) renders

$$\| e^A \| \lesssim e^{-1} * 70$$

since $\kappa(X) \approx 70$ for this choice of A.

As the above examples show, it is impossible to rank the upper bounds in the preceeding pages in terms of effectiveness. The sharpness of a given inequality will depend upon A and t. However, if A is a normal matrix ($A^*A = AA^*$), then all the results of this section point to the same fact, namely

$$\text{(8.11)} \qquad A \text{ normal} \quad \Rightarrow \quad \| e^{At} \| = e^{\alpha(A)t}$$

There are many ways of proving (8.11) For example, one can show $\mu(A) = \alpha(A)$ and then apply (8.5) and (8.7).

9. Perturbation Bounds for e^{At}

In this section we examine the problem of bounding $\| e^{(A+E)t} - e^{At} \|$ for $t \geq 0$. When A and E commute, this is particularly easy.

Theorem 15.

$$(9.1) \qquad AE = EA \implies \| e^{(A+E)t} - e^{At} \| \leq \| e^{At} \| (t \|E\| e^{\|E\|t}) \qquad (t \geq 0)$$

Proof.

$$\| e^{(A+E)t} - e^{At} \| = \| e^{At} (e^{Et} - I) \| \leq \| e^{At} \| \sum_{k=1}^{\infty} \frac{\|E\|^k t^k}{k!}$$

$$\leq \| e^{At} \| (t \|E\| e^{\|E\|t})$$

Q.E.D.

By using (8.2) and (8.3) respectively we obtain the following two corolaries:

Corollary 1.

If AE = EA and the Jordan decomposition of A is given by (2.1) and (2.2), with $m = \max \{m_1, .. m_p\}$,then for $t \geq 0$

$$(9.2) \qquad \| e^{(A+E)t} - e^{At} \| \leq e^{\alpha(A)t} (t \|E\| e^{\|E\|t} m \kappa(X) \max_{0 \leq r \leq m-1} \frac{t^r}{r!})$$

Corollary 2.

If AE = EA and the Schur decomposition of A is given by (3.1), then for $t \geq 0$

$$(9.3) \qquad \| e^{(A+E)t} - e^{At} \| \leq e^{\alpha(A)t} (t \|E\| e^{\|E\|t} \sum_{k=0}^{n-1} \frac{t^k \|N\|^k}{k!})$$

The problem of deriving perturbation bounds is considerably more difficult when A and E do not commute. Instead of manipulating the power series for the exponentials $e^{(A+E)t}$ and e^{At} (a very cumbersome approach), we will use equation (3.2). This gives us

$$(9.4) \qquad \| e^{(A+E)t} - e^{At} \| = \int_0^t e^{A(t-\tau)} E \, e^{(A+E)\tau} \, d\tau$$

from which the following, basic inequality can be deduced:

$$(9.5) \qquad \| e^{(A+E)t} - e^{At} \| \leq \| E \| \int_0^t \| e^{A(t-\tau)} \| \, \| e^{(A+E)\tau} \| \, d\tau$$

By coupling (9.5) with the results of the previous section we can immediately obtain some perturbation bounds.

Theorem 16.

Suppose the Schur decompositions of A and A+E are given respectively by

$$Q^* A Q = \mathrm{diag}(\lambda_i) + N$$

and

$$\tilde{Q}^* (A+E) \tilde{Q} = \mathrm{diag}(\tilde{\lambda}_i) + \tilde{N}$$

If $\alpha = \max\{\alpha(A), \alpha(A+E)\}$, $M = \max\{\| N \|, \| \tilde{N} \|\}$, and $p(x) = \sum_{k=0}^{n-1} \frac{x^k}{k!}$, then

$$(9.6) \qquad \| e^{(A+E)t} - e^{At} \| \leq e^{\alpha t} \, (t \, p(Mt)^2 \, \| E \|)$$

Proof.

From (8.3)

$$\begin{aligned} \| e^{(A+E)t} - e^{At} \| &\leq \| E \| \int_0^t \| e^{A(t-\tau)} \| \, \| e^{(A+E)\tau} \| d\tau \\ &\leq \| E \| \int_0^t e^{\alpha(A)(t-\tau)} p[\| N \| (t-\tau)] \; e^{\alpha(A+E)\tau} \; p[\| \tilde{N} \| \tau] d\tau \\ &\leq e^{\alpha t} \| E \| \int_0^t p[\| N \| (t-\tau)] \; p[\| \tilde{N} \| \tau] d\tau \\ &\leq e^{\alpha t} \; (t \; p(Mt)^2 \; \| E \|) \end{aligned}$$

Q.E.D.

In order to produce additional perturbation bounds, it is convenient to have the following result:

Lemma 3.

If $\|e^{At}\| \leqslant M\ e^{\beta t}$ $(t \geq 0)$, then $\|e^{(A+E)t}\| \leqslant M\ e^{(\beta+M\|E\|)t}$.

Proof.

From Theorem 7 $(B \equiv E)$ we have for $k \geqslant 1$

$$\|A_k(t)\| \leq \int_0^t \cdots \int_0^{t_{k-1}} \|e^{A(t-t_1)}\| \|E\| \cdots \|E\| \|e^{At_k}\| dt_k \cdots dt_1$$

$$\leq M^{k+1} \|E\|^k \ e^{\beta t} \frac{t^k}{k!}$$

Since the same result holds when $k = 0$ we have

$$\|e^{(A+E)t}\| \leq \sum_{k=o}^{\infty} \|A_k(t)\| \leq M\ e^{\beta t} \sum_{k=0}^{\infty} M^k \frac{\|E\|^k\ t^k}{k!} = M\ e^{(\beta+M\|E\|)t}$$

Q.E.D.

Theorem 17.

Suppose $\|e^{At}\| \leqslant M_1\ e^{\alpha_1 t}$ and $\|e^{(A+E)t}\| \leqslant M_2\ e^{\alpha_2 t}$ for all $t \geqslant 0$. If $M = \max\{M_1,M_2\}$ and $\alpha = \max\{\alpha_1,\alpha_2\}$ then

(9.7) $$\|e^{(A+E)t} - e^{At}\| \leq e^{\alpha t}\ (M^2\ \|E\|\ t)$$

Also,

(9.8) $$\|e^{(A+E)t} - e^{At}\| \leq e^{(\alpha_1+M_1\|E\|)t}\ (M_1^2\ \|E\|\ t\ )$$

Proof.

From (9.5)

$$\|e^{(A+E)t} - e^{At}\| \leq \|E\| \int_0^t M_1\ e^{\alpha_1(t-\tau)}\ M_2\ e^{\alpha_2\tau}\ d\tau \leq e^{\alpha t}\ (\|E\|\ M^2 t)$$

On the other hand, by using Lemma 3 we obtain

$$\|e^{(A+E)t} - e^{At}\| \leqslant \|E\| M_1^2 \int_0^t e^{\alpha_1(t-\tau)}\ e^{(\alpha_1+M_1\|E\|)\tau}\ d\tau$$

$$= \|E\|\ M_1^2\ e^{\alpha_1 t} \int_0^t e^{M_1\|E\|}\ d\tau \leqslant t\ \|E\| M_1^2\ e^{(\alpha_1+M_1\|E\|)t}$$

Q.E.D.

Corollary 1.

If the Jordan decomposition of A is given by (2.1) and (2.2) and if J is diagonal (i.e. m=1), then

$$(9.9) \qquad \| e^{(A+E)t} - e^{At} \| \leq \kappa(X)^2 \|E\| t \quad e^{(\alpha(A) + \kappa(X)\|E\|)t}$$

Corollary 2.

$$(9.10) \qquad \| e^{(A+E)t} - e^{At} \| \leq \|E\| t \quad e^{(\mu(A) + \|E\|)t}$$

Corollary 3.

If A and A+E are both normal, then

$$(9.11) \qquad \| e^{(A+E)t} - e^{At} \| \leq \|E\| t \quad e^{(\alpha(A) + \|E\|)t}$$

One can also attempt to derive perturbation bounds from (4.1). If Γ is a contour enclosing the spectrums of A and A+E, then we have

$$\| e^{(A+E)t} - e^{At} \| = \frac{1}{2\pi i} \oint_{\Gamma} e^{zt} \{ [zI - (A+E)]^{-1} - (zI - A)^{-1} \} \, dz$$

$$= \frac{1}{2\pi i} \oint_{\Gamma} e^{zt} (zI - (A+E))^{-1} E (zI - A)^{-1} \, dz$$

Perturbation bounds can be obtained by taking norms in the above expression and then bounding the right hand side. We delete the results of this approach because the inequalities so obtained are no better than the ones already given.

We conclude this section with a word about the relative perturbation $\| e^{(A+E)t} - e^{At} \| / \| e^{At} \|$. In particular, this quantity does not necessarily decay even though e^{At} and $e^{(A+E)t}$ do. For example, if

$$A = \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix} \qquad \text{and} \qquad A+E = \begin{pmatrix} -1 & \varepsilon \\ 0 & -1 \end{pmatrix} \qquad \varepsilon \geq 0$$

then $\| e^{(A+E)t} - e^{At} \| / \| e^{At} \| = \varepsilon t$.

10. Approximation of a Function of a Matrix.

Suppose the matrix functions $f(A)$ and $g(A)$ are defined. If $f(z)$ approximates $g(z)$ on a set containing $\lambda(A)$, then $f(A)$ will approximate $g(A)$. We can quantify this by using either (2.5) or (3.11).

Theorem 18.

If the Jordan decomposition of A is given by (2.1) and (2.2) and if $f(z)$ and $g(z)$ are analytic functions defined on $\lambda(A)$, then

$$(10.1) \qquad \| f(A) - g(A) \| \leqslant m\, \kappa(X) \max_{\substack{z \varepsilon \lambda(A) \\ 0 \leqslant r \leqslant m-1}} \frac{\left| f^{(r)}(z) - g^{(r)}(z) \right|}{r!}$$

Proof.

Use (2.5) with $f \equiv f-g$. Q.E.D.

Theorem 19.

Suppose the Schur decomposition of A is given by (3.1) and that Ω is a convex set containing $\lambda(A)$. If $f(z)$ and $g(z)$ are analytic functions on Ω and

$$\max_{z \varepsilon \Omega} \left| f^{(r)}(z) - g^{(r)}(z) \right| \leqslant \delta_r \qquad (0 \leqslant r \leqslant n-1)$$

then

$$(10.2) \qquad \| f(A) - g(A) \|_F \leqslant \delta \, \| (I - |N|)^{-1} \|_F$$

where δ is a constant satisfyiing $\delta \geqslant \frac{\delta_r}{r!}$, $r = 0,\ldots,n-1$.

Proof.

Use (3.11) with $f = f-g$. (Notice that if $N = 0$ we need only require $\delta \geqslant \delta_0$.) Q.E.D.

By using Cauchy's integral formula we can obtain an interesting variation of (10.2) which does not require information on how well the derivatives of $f(z)$ approximate the derivatives of $g(z)$.

Theorem 20.

Suppose the Schur decomposition of A is given by (3.1) and that Ω is a convex set containg $\lambda(A)$. In addition assume that f(z) and g(z) are analytic functions inside and on a closed contour Γ whose interior contains Ω . If

$$d = \inf_{\substack{z\in\Omega \\ w\in\Gamma}} |z - w| > 0$$

$$L(\Gamma) = \text{the "length" of } \Gamma$$

and

$$\varepsilon = \max_{w\in\Gamma} |f(w) - g(w)|$$

then

$$\|f(A) - g(A)\|_F \leq \varepsilon \frac{L(\Gamma)}{2\pi\ d} \|(I - \frac{|N|}{d})^{-1}\|_F$$

Proof.

For $z \in \Omega$ and $r \geqslant 0$ we have the following inequality from Cauchy's integral formula:

$$|f^{(r)}(z) - g^{(r)}(z)| = \left|\frac{r!}{2\pi i} \oint_\Gamma \frac{f(w)-g(w)}{(z-w)^{r+1}} dw\right| \leq \frac{\varepsilon\ r!\ L(\Gamma)}{2\pi\ d^{r+1}}$$

By using (3.10) we obtain

$$|f(T) - g(T)| \leq \sum_{r=0}^{n-1} \frac{\varepsilon\ r!\ L(\Gamma)}{2\pi\ d^{r+1}} \frac{|N|^r}{r!} = \left(\frac{\varepsilon\ L(\Gamma)}{2\pi\ d}\right) (I - \frac{|N|}{d})^{-1}$$

and thus

$$\|f(A) - g(A)\|_F \leq \varepsilon \frac{L(\Gamma)}{2\pi d} \|(I - \frac{|N|}{d})^{-1}\|_F$$

Q.E.D.

11. Pade Approximation and e^{At}.

The (p,q) Pade function $R_{pq}(z)$ is a rational approximation to e^z of the form

$$(11.1) \qquad R_{pq}(z) = \frac{n_{pq}(z)}{d_{pq}(z)}$$

where

$$(11.2) \qquad n_{pq}(z) = \sum_{j=0}^{p} \frac{(p+q-j)!\; p!}{(p+q)!\; j!\; (p-j)!} z^j$$

and

$$(11.3) \qquad d_{pq}(z) = \sum_{j=0}^{q} \frac{(p+q-j)!\; q!}{(p+q)!\; j!\; (q-j)!} (-z)^j \quad .$$

The error of this approximation is given by

$$(11.4) \qquad e_{pq}(z) = e^z - R_{pq}(z) = \frac{(-1)^q\, z^{p+q+1}}{(p+q)!\;\; d_{pq}(z)} \int_o^1 e^{z(1-u)} u^q (1-u)^p du$$

In view of the previous section, if A is a matrix such that $d_{pq}(A)$ is invertible, then

$$R_{pq}(A) \equiv \frac{n_{pq}(A)}{d_{pq}(A)} = n_{pq}(A)\left[d_{pq}(A)\right]^{-1}$$

may be regarded as an approximation to e^A. The convergence properties of these approximants for general matrices was established by Wragg and Davies [32]. They extended the work of Varga[44] and Fair and Luke [38] and showed that for any matrix A

$$\lim_{q\to\infty} R_{pq}(A) = e^A \qquad \text{(fixed p)}$$

$$\lim_{p\to\infty} R_{pq}(A) = e^A \qquad \text{(fixed q)}$$

$$\lim_{p\to\infty} R_{p,p+a}(A) = e^A \qquad (a = 0,\pm 1)$$

In the course of establishing these properties, Wragg and Davies derived an upper bound for $\|R_{pq}(A) - e^A\|$ by using the Jordan decomposition. Predictably, this upper bound involves the eigensystem condition number $\kappa(X)$. (See Section 2.) This prompted the authors to ask the following question in a later paper [34]:

> *Is it inherently difficult to approximate e^A closely by diagonal Pade approximants when $\kappa(X)$ is large?*

When $\kappa(X)$ is large, their derived inequality

$$\|R_{qq}(A) - e^A\| \leq w_q \tag{11.5}$$

is weak. For example, if

$$A = \begin{pmatrix} -1+\delta & 4 \\ 0 & -1-\delta \end{pmatrix} \qquad \delta = 10^{-6} \quad , \quad \kappa(X) \approx 10^6$$

then

$$w_q \approx \left(\frac{(q!)^2 \, e^2}{(2q)! \; (2q+1)!} \right) 10^6$$

The pessimism of (11.5) for this example is indicated by the following table:

q	$\|e^A - R_{qq}(A)\|$	w_q
2	10^{-2}	10^4
3	10^{-3}	10^2
4	10^{-6}	10^{-1}
5	10^{-7}	10^{-3}

(For clarity we have only shown orders of magnitude.)

For the three examples given in [34] to illustrate (11.5), $\kappa(X)$ is very modest giving the impression that (11.5) is "sharp". Recognizing that this not always be the case, Wragg and Davies express a desire for "$\kappa(X)$ - free" bounds on the error $\|R_{qq}(A) - e^A\|$. To this end they developed an exact expression for error in the 2x2 case which did not involve the factor $\kappa(X)$. For general nxn matrices, an exact expression for $\|e^A - R_{pq}(A)\|$ would be a very difficult result to obtain. However, some interesting "$\kappa(X)$-free" upper bounds can be derived and these will be presented on the following pages.

But first, it is worth mentioning the relevance of this pursuit. The fact that $R_{pq}(A)$ approximates e^A has important ramifications when it comes to the numerical solution of the initial value problem

$$(11.6) \qquad \begin{aligned} Au(t) &= \dot{u}(t) & \quad A &\in \mathbb{C}^{nxn} \\ u(0) &= u_0 & \quad u_0 &\in \mathbb{C}^n \end{aligned}$$

This is because with time step $\Delta t > 0$ we have the following for k =0,1,.. :

$$u(k\Delta t) = e^{Ak\Delta t}u_0 = (e^{A\Delta t})^k u_0 \cong (R_{pq}(A\Delta t))^k u_0 = \tilde{u}(k\Delta t)$$

Precisely how one implements Padé approximation to solve (11.6) (and its generalizations) depends upon the form of the matrix A.

If A is large and sparse, then the comments which appear in Varga[44] Siemeniuch and Gladwell[25], and Blue and Gummel[4](and the references in these papers) are relevant. These types of problems arise in connection with the solution of parabolic partial differential equations which have been discretized in space.

The system (11.6) also arises in the study of linear, time invariant, dynamical systems[43]. The matrix A may often be regarded as small and dense in these applications. When this is the case, the ideas espoused in Wragg and Davies[33], Scraton[24], and Zakian[35]are of interest. For the special case of R_{po} approximation (i.e. truncated Taylor series approximation) we mention the papers of Gall[11], Liou[16], [40], and Plant[21].

Our remarks on the subject of Padé approximation of the matrix exponential are ostensibly theoretical although they may be of practical interest in those situations where A has limited dimension. WE begin by bounding $\|R_{pq}(A\Delta t) - e^{A\Delta t}\|_F$ using Theorems 19 and 20 in the q=0 and q>0 cases respectively. Such bounds could represent a preliminary step in the rigorous bounding of the global error $\|((R_{pq}(A\Delta t)^k - e^{Ak\Delta t})u_0\|$. However, we shall mainly regard the two theorems which follow as merely a specific demonstration of how our Schur analysis can be applied.

Theorem 21.

If the eigenvalues of $A\Delta t$ lie in the half disc Ω defined by

$$\Omega = \{ x + iy \mid (x^2 + y^2) \leqslant (\alpha \Delta t)^2 , x \leqslant 0, \alpha \geqslant 0 \}$$

and if the Schur decomposition of A is given by (3.1), then

$$(11.7) \quad \| R_{po}(A\Delta t) - e^{A\Delta t} \|_F \leqslant \frac{(\Delta t)^{p+1}}{(p+1)!} \| (\alpha I + |N|)^{p+1} \|_F + \frac{\| (|N|\Delta t)^{p+2} \|_F}{(p+2)!} \| e^{|N|\Delta t} \|_F$$

Proof.

Since $R_{po}(z) = \sum_{i=0}^{p} \frac{z^i}{i!}$, $R_{po}^{(j)}(z) = R_{p-j,o}(z)$ $(j=0,\ldots,p)$ and thus from (11.4) we can deduce that

$$z \in \Omega \Rightarrow |R_{po}^{(j)}(z) - e^z| \leqslant \varepsilon_j$$

where

$$\varepsilon_j = \begin{cases} \dfrac{(\alpha \Delta t)^{p+1-j}}{(p+1-j)!} & j = 0,1,\ldots,p+1 \\ 1 & j = p+2,\ldots \end{cases}$$

If the Schur decomposition of A is specified by (3.1), then from Theorem 19

$$|R_{po}(T\Delta t) - e^{T\Delta t}| \leqslant \sum_{j=0}^{n-1} \varepsilon_j \frac{|N\Delta t|^j}{j!}$$

If $(p+1) < (n-1)$ then by substitution of the definition of ε_j we obtain

$$|R_{po}(T\Delta t) - e^{T\Delta t}| \leqslant \sum_{j=0}^{p+1} \frac{(\alpha \Delta t)^{p+1-j}}{(p+1-j)!} \frac{|N\Delta t|^j}{j!} + \sum_{j=p+2}^{n-1} \frac{|N\Delta t|^j}{j!}$$

$$\leqslant \frac{(\Delta t)^{p+1}}{(p+1)!} (\alpha I + |N|)^{p+1} + \frac{|N\Delta t|^{p+2}}{(p+2)!} e^{|N\Delta t|}$$

Result (11.7) follows by taking the F-norm of both sides of the abcve inequality. The reader can check that the same result holds when $(p+1) \geqslant (n-1)$. (Hint: $|N|^{p+2} = 0$.)

Q.E.D.

It is also possible to use Theorem 19 to bound $\| R_{pq}(A\Delta t) - e^{A\Delta t} \|_F$ when $q \geq 1$. However, when $q \neq 0$ it is quite awkward to obtain the necessary bounds on $| R_{pq}^{(j)}(z) - e^z |$. A simpler approach is to use Theorem 20.

Theorem 22.

Suppose the eigenvalues of $A\Delta t$ lie inside the half disc Ω defined by

$$\Omega = \{ x + iy \mid (x^2 + y^2) \leq (\alpha\, \Delta t)^2 \quad \alpha \geq 0 \ , \ x \leq 0 \}$$

If $(\alpha\, \Delta\, t) \leq \frac{1}{2}$, $q \geq 1$, and the Schur decomposition of A is given by (3.1), then

$$\text{(11.8)} \qquad \| R_{pq}(A\Delta t) - e^{A\Delta t} \|_F \leq \frac{4\ (2\alpha\ \Delta t)^{p+q+1}}{q^q\ (p+q+1)!}\, q!\, \| (I - \frac{|N|}{\alpha})^{-1} \|_F$$

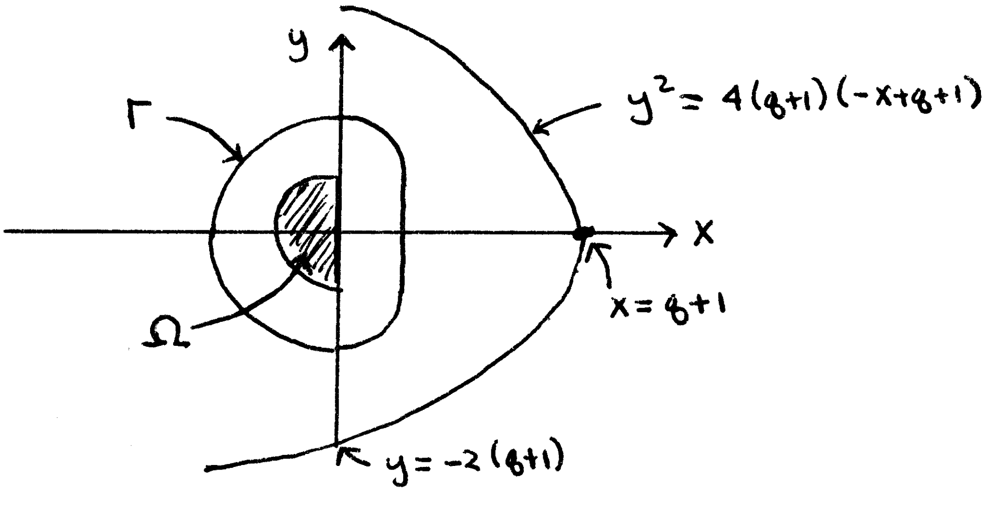
Proof.

Let the contour Γ be defined by $\Gamma = \{ w \mid \inf_{z\varepsilon\Omega} |z-w| = \alpha\Delta t \}$.

In accordance with Theorem 20,

$$\text{(11.9)} \qquad \begin{aligned} L(\Gamma) &\leq 12\ \alpha\Delta t \\ d &= \alpha\Delta t \end{aligned}$$

Now Saff and Varga[23] have shown that $d_{pq}(x + iy)$ cannot be zero when $y^2 \leq 4(q+1)(-x+q+1)$. Graphically we have



By using this result, we can produce a lower bound for $|d_{pq}(w)|\,(w \,\varepsilon\, \Gamma)$.

If $w \varepsilon \Gamma$ and $d_{pq}(z) = 0$, then one can show that $|w-z| > q$. Hence, if $z_1, \ldots, z_q$ are the zeros of $d_{pq}(z)$, then

$$w \varepsilon \Gamma \Rightarrow |d_{pq}(w)| = \left| \frac{p!}{(p+q)!} \prod_{i=1}^{q} (z_i - w) \right| \geq \frac{p!\, q^q}{(p+q)!}$$

Thus, for $w \varepsilon \Gamma$ we have from (11.4) that

$$|R_{pq}(w) - e^w| \leq \frac{(2\alpha\ \Delta t)^{p+q+1)}}{(p+q)!\,|d_{pq}(w)|} \; e^{\alpha\Delta t} \int_0^1 u^q (1-u)^p du$$

$$\leq (2\alpha\Delta t)^{p+q+1} \; e^{1/2} \frac{q!}{(p+q+1)!\ q^q} = \varepsilon$$

With this result and (11.9) we can employ Theorem 20 which gives:

$$\| e^{A\Delta t} - R_{pq}(A\Delta t) \|_F \leq \frac{(2\alpha\Delta t)^{p+q+1} \; q!}{q^q\ (p+q+1)!} e^{1/2} \frac{12\alpha\Delta t}{2\alpha\pi\Delta t} \left\| (I - \frac{|N\Delta t|}{\alpha\Delta t})^{-1} \right\|_F$$

$$\leq \frac{4\ (2\alpha\Delta t)^{p+q+1} \; q!}{q^q\ (p+q+1)!} \left\| (I - \frac{|N|}{\alpha})^{-1} \right\|_F$$

Q.E.D.

Although the upper bounds (11.7) and (11.8) do not involve the factor $\kappa(X)$, they may nevertheless be very large. We discussed this kind of behaviour with an example on page 32. It ostensibly arises from the fact that powers of N may be large in norm.

AS we mentioned earlier, $\tilde{u}(k\Delta t) = R_{pq}(A\Delta t)^k u_o$ represents an approximate solution to (11.6) at $t = k\Delta t$. Since

$$\| u(k\Delta t) - \tilde{u}(k\Delta t) \| = \| (e^{Ak\Delta t} - R_{pq}(A\Delta t)^k) u_o \| \leq \| e^{Ak\Delta t} - R_{pq}(A\Delta t)^k \| \| u_o \|$$

we see that an upper bound on the error can be obtained by bounding $\| e^{Ak\Delta t} - R_{pq}(A\Delta t)^k \|$. Manipulation of Theorems 21 and 22 c an be used for this end. However, a more illuminating analysis results by showing that

$$R_{pq}(A\Delta t) = e^{(A+E)\Delta t}$$

for some perturbation matrix E. To simplify the proof of our main result in this direction (Theorem 23), we state the following two lemmas.

Lemma 4.

If $\|H\| < 1$ then $\log(I + H)$ exists and

$$\|\log(I + H)\| \leqslant \frac{\|H\|}{1 - \|H\|} \tag{11.10}$$

Proof.

If $\|H\| < 1$ then $\log(I + H)$ can be expressed in terms of a power series:

$$\log(I + H) = \sum_{k=1}^{\infty} \frac{H^k}{k} (-1)^{k+1}$$

By taking norms we find

$$\|\log(I + H)\| \leqslant \sum_{k=1}^{\infty} \frac{\|H\|^k}{k} \leqslant \|H\| \sum_{k=0}^{\infty} \|H\|^k = \frac{\|H\|}{1 - \|H\|}$$

Q.E.D.

Lemma 5.

If $\|A\| \Delta t \; e \left(\frac{q}{p+q}\right) < 1$, then

$$\|d_{pq}(A\Delta t)^{-1}\| \leqslant \frac{1}{1 - A \,\Delta t \; e \left(\frac{q}{p+q}\right)}$$

Proof.

When $q = 0$, $d_{pq}(A\Delta t) = I$ and thus the above inequality holds. For $q \geqslant 1$, we see from (11.3) that

$$d_{pq}(A\Delta t) = I + F$$

where

$$F = \sum_{j=1}^{q} \frac{(p+q-j)!\ q!}{(p+q)!\ j!\ (q-j)!} (-A\Delta t)^{j}$$

Thus,

$$\|F\| \leq \sum_{j=1}^{q} \frac{q(q-1)\cdot\cdot\cdot(q-j+1)}{(p+q)\cdot\cdot\cdot(p+q-j+1)} \quad \frac{A\Delta t^{\;j}}{j!}$$

$$\leq \sum_{j=1}^{q} \left(\frac{q}{p+q}\|A\Delta t\|\right)^{j} \frac{1}{j!} \leq \frac{q}{p+q}\|A\Delta t\|\ e < 1$$

Since $\|d_{pq}(A\Delta t)^{-1}\| = \|(I+F)^{-1}\| \leq \dfrac{1}{1-\|F\|}$, the Lemma follows.

Q.E.D.

Theorem 23.

If $\|A\|\Delta t\ e < \frac{1}{2}$, then $R_{pq}(A\Delta t) = e^{(A+E)\Delta t}$ where

(11.11) $$\|E\| \leq \frac{5\|A\|^{p+q+1}\ p!\ q!}{(p+q)!\ (p+q+1)!} (\Delta t)^{p+q}$$

Proof.

By setting $E = -A$, we see that the theorem holds when $p=q=0$. Hence, we must prove the Theorem for the case $p+q \geq 1$. From (11.4)

(11.12) $$R_{pq}(A\Delta t) = e^{A\Delta t} - \frac{(-1)^{q}\ (A\Delta t)^{p+q+1}}{(p+q)!\ d_{pq}(A\Delta t)} \int_{0}^{1} e^{A\Delta t(1-u)} u^{q}(1-u)^{p} du$$

By substituting

(11.13) $$R_{pq}(A\Delta t) = e^{A\Delta t}\ e^{E\Delta t}$$

into (11.12) and left multiplying by $e^{-A\Delta t}$ we obtain

(11.14) $$e^{E\Delta t} = I + H$$

where

(11.15) $$H = -\frac{(-1)^{q}\ (A\Delta t)^{p+q+1}}{(p+q)!\ d_{pq}(A\Delta t)} \int_{0}^{1} e^{-uA\Delta t}\ u^{q}(1-u)^{p} du$$

By using Lemma 5

$$\|A\|\Delta t \; e < \tfrac{1}{2} \qquad \Rightarrow \qquad \| d_{pq}(A\Delta t)^{-1} \| \leqslant 2$$

and so by taking norms in (11.15) we obtain

$$(11.16) \qquad \| H \| \leqslant 2 \, \|A\Delta t\|^{p+q+1} \, e^{\frac{1}{2}} \, \frac{p!.q!}{(p+q)! \; (p+q+1)!}$$

Since $\|A\Delta t\| \, e \; < \frac{1}{2}$ and $p+q > 1$,

$$2 \, e^{\frac{1}{2}} \, \|A\Delta t\|^{p+q+1} \leqslant 2 \, e^{\frac{1}{2}} \left(\frac{1}{2e}\right)^2$$

and thus from (11.16) it can be shown that

$$(11.17) \qquad \| H \| \leqslant (2e)^{-1} \leqslant {}^1/_5$$

This implies that log(I + H) exists and so from (11.14)

$$(11.18) \qquad E = \frac{\log(I + H)}{\Delta t}$$

The necessary bound on $\|E\|$ can be obtained by (11.16), (11.17) and Lemma 5:

$$\|E\| < \frac{1}{\Delta t} \frac{\|H\|}{1 - \|H\|} \leqslant \frac{1}{\Delta t} \, \frac{5}{4} \, 2 \, e^{\frac{1}{2}} \, \|A\Delta t\|^{p+q+1} \, \frac{p! \; q!}{(p+q)!(p+q+1)!}$$

$$\leqslant \frac{5 \, \|A\|^{p+q+1} \, p! \; q!}{(p+q)! \; (p+q+1)!} \, (\Delta t)^{p+q}$$

From (11.15) and (11.18), it is clear that E commutes with A and hence we may arrange (11.13) as $R_{pq}(A\Delta t) = e^{(A+E)\Delta t}$.

Q.E.B.

Corollary.

If $\|A\|\Delta t\, e < \frac{1}{2}$, then

$$\left\| R_{pq}(A\Delta t)^k u_o - e^{Ak\Delta t} u_o \right\| \leq \left\| e^{Ak\Delta t} u_o \right\| \left(k\Delta t \|E\| e^{\|E\| k\Delta t}\right) \tag{11.19}$$

and

$$\left\| R_{pq}(A\Delta t)^k - e^{Ak\Delta t} \right\| \leq \left\| e^{Ak\Delta t} \right\| \left(k\Delta t \|E\| e^{\|E\| k\Delta t}\right) \tag{11.20}$$

Proof,

$$\begin{aligned} \left\| R_{pq}(A\Delta t)^k u_o - e^{Ak\Delta t} u_o \right\| &= \left\| (e^{Ek\Delta t} - I) e^{Ak\Delta t} u_o \right\| \\ &\leq \left\| e^{Ak\Delta t} u_o \right\| \left(k\Delta t \|E\| e^{\|E\| k\Delta t}\right) \end{aligned}$$

(See Theorem 15.) The proof of (11.20) is even easier.

Q.E.D.

Theorem 23 represents an inverse error analysis of Pade approximation to the matrix exponential. It shows that our approximate solution $\tilde{u}(k\Delta t) = R_{pq}(A\Delta t)^k u_o$ to (11.6) is the <u>exact</u> solution to the perturbed system

$$\begin{aligned} (A + E)v(t) &= \dot{v}(t) \\ v(0) &= u_o \end{aligned} \tag{11.6'}$$

at $t = k\Delta t$ $(k=0,1,\dots)$. If the entries of A are correct only to the r-th decimal place, then there would seem to be little justification in chosing p, q, and Δt such that $\|E\| \leq 10^{-r}$, for then the accuracy of the method would not be consistent with the accuracy of the data. In view of (11.11), one would thus expect that

$$\frac{5\|A\|^{p+q+1}\ p!\ q!}{(p+q)!\ (p+q+1)!} (\Delta t)^{p+q} \gg 10^{-r}$$

for any particular implementation of Pade approximates.

The corollary gives an upper bound for the relative error of our approximate solution to the system (11.6):

$$\frac{\| R_{pq}(A\Delta t)^k u_o - e^{Ak\Delta t} u_o \|}{\| e^{Ak\Delta t} u_o \|} \leq (k\Delta t \|E\| e^{\|E\| k\Delta t})$$

By manipulation of this inequality and the upper bound (11.11), one can ascertain the values of p, q, and Δt which are necessary to keep the relative error below a prescribed tolerance. Techniques based upon this type of error bound control have been discussed in connection with R_{po} approximants (i.e. truncated taylor series approximation). The method of Liou 16 is an example of this. However, his concern is with absolute error. Because e^{At} can grow initially even though $\alpha(A) < 0$, we feel that the relative error is the more proper quantity to control.

Acknowledgements.

I am grateful for the support of the Science Research Council during the period when this report was written.

REFERENCES

[1] T.M.APOSTOL, Some explicit Formulas for the matrix Exponential e^{At}, AMS Monthly, 76, 284-292 (1969).

[2] R.BELLMAN, Introduction to Matrix Analysis. McGraw-Hill, New York (1960).

[3] R.BELLMAN, Perturbation Techniques in Mathematics, Physics, and Engineering. Holt, Rinehart and Winston Inc., New York(1964).

[4] J.L.BLUE & H.K.GUMMEL, Rational Approximations to Matrix Exponential for Systems of Stiff Differential Equations, Journal of Computational Physics, 5, 70-83, (1970).

[5] F.H.BRANIN, Computer Methods of Network Analysis, Proc. I.E.E.E. 55, 1787-1801 (1967) .

[6] W.A.COPPEL, Stability and Asymptomatic Behaviour of Differential Equations, D.C.Heath, Boston, (1965).

[7] C.G.CULLIN, Remarks on Computing e^{At}, I.E.E.E. Trans. on Automatic Control, 94-95 (1971).

[8] G.DAHLQUIST, Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations, Transactions of the Royal Institute of Technology, No. 130, Stockholm (1959).

[9] N.DUNFORD and J.SCHWARTZ, Linear Operators Part I, Interscience, New York (1958).

[10] E.P.FULMER, Computation of the Matrix Exponential, AMS Monthly, 82, 156-159 (1975).

[11] D.A.GALL, The Solution of Linear, Constant-Coefficient, Ordinary ential Equations with APL, Computer Methods in Applied Mechanics and Engineering, 1, 189-196(1972).

[12] F.R.GANTMACHER, Application of the Theory of Matrices, Interscience Publishers, New York (1959).

[13] G.H.GOLUB and J.H.WILKINSON, Ill-Conditioned Eigensystems and the Computation of the Jordan Canonical Form, Stanford Computer Science Report, CS75-478 (1975).

[14] T.KATO, Perturbation Theory for Linear Operators, Springer-Verlay, New York (1966).

[15] R.B. KIRCHNER, AN Explicit Formula for e^{At}, AMS Monthly, 74, 1200-1204, (1967).

[16] M.L.LIOU, A Novel Method of Evaluating Transient Response, Proc. I.E.E.E., 54, 20-23 (1966).

[17] C.C. MACDUFFEE, The Theory of Matrices, Chelsea Publishing Company, (1956).

[18] L.MIRSKY, An Introduction to Linear Algebra, Oxford University Press, London (1955).

[19] A.M.OSTROWSKI, Solution of Equations and Systems of Equations, Academic Press, New York (1960).

[20] B.N.PARLETT, Computation of Functions of Triangular Matrices, Memo No. ERL-M481, Electronics Research Laboratory, College of Engineering, Berkeley (1974).

[21] J.B.PLANT, On the Computation of Transition Matrices for Time Invariant Systems, Proc. I.E.E.E. 56,1397-1398 (1968).

[22] E.J.PUTZER, Avoiding the Jordan Canonical Form in the Discussion of Linear Systems with Constant Coeffients, AMS Monthly, 73, 2-7 (1966).

[23] E.B.SAFF and R.S.VARGA, On the Zeroes and Poles of Pade Approximants to e^{z} , to be published in Numer.Math.

[24] R.E.SCRATON, Comment on Rational Approximants to the Matrix Exponential, Electronics Letters, 7, 260-61, (1971).

[25] G.SIEMIENIUCH and I.GLADWELL, On Time Discretizations for Linear Time-Dependent Partial Differential Equations, Univ. of Manchester Numerical Analysis Report 5, (1974).

[26] G.W.STEWART, Introduction to Matrix Computations, Academic Press, New York (1973).

[27] I.STROM, On Logarithmic Norms, Royal Institute of Technology, Report NA69.06, Stockholm (1969).

[28] T.STROM, On the Use of Majorants for Strict Error Estimation of Numerical Solutions of Ordinary Differential Equations, Royal Institute of Technology, Report NA70.10, Stockholm (1970).

[29] H.F.TROTTER, Product of Semigroups of Operators, Proc.AMS, 10, 545-551 (1959).

[30] G.H.WEISS and A.A.MARADUDIN, The Baker-Housdorff Formula and a Problem in Crystal Physics, P.Math. and Physics, 3, 771-777 (1962).

[31] J.H.WILKINSON, The Algebraic Eigenvalue Problem, Clarendon Press, Oxford(1965).

[32] A.WRAGG and C.DAVIES, Computation of the Exponential of a Matrix I: Theoretical Considerations, J.Inst. Maths. Applic., 11,369-375 (1973).

[33] A.WRAGG and C.DAVIES, Electronics Letters 9, 525-26 (1973).

[34] A.WRAGG and C.DAVIES, "Computation of the Exponential of a Matrix II. Practical Considerations. " J.Inst.Maths.Applic. 15, 273-278(1975).

[35] V.ZAKIAN, Rational Approximants to the Matrix Exponential, Electronics Letters to 814-815, (1970).

[36] A.D.ZIEBUR, On determining the Structure of A by Analysing e^{AT}, Siam Review, 12, 98-102, (1970).

[37] S.BEARDS, On the Evaluation of Exp(At), Matrix Tensor Quart, 23, 141-142, (1973).

[38] W.FAIR and Y.L.LUKE, Padé Approximations to the Operator Exponential,Num. Math.14, 379-382 (1970).

[39] A.H.LEVIS, Some Computational Aspects of the Matrix Exponential, I.E.E.E. Trans.Automatic Control, 14, 410-411, (1969).

[40] M.L.LIOU, Evaluation of the Transition Matrix, Proc. I.E.E.E., 55, 228-229(1967).

[41] R.F.RINEHART, The Exponential Representation of Unitary Matrices, Math.Mag. 37, 111-112,(1964).

[42] M.N.S. SWAMI, On a Formula for Evaluating e^{AT} when the Eigenvalues are not Necessarily Distinct, Matrix Tensor Quart, 23, 67-72(1972).

[43] Y.TAKAHASHI, M.J.RABINS, and D.M.AUSLANDER, Control and Dynamic Systems, Addison-Wesley, Reading, Mass. (1970).

[44] R.S. VARGA, On Higher Order Stable Implicit Methods for Solving Parabolic Partial Differential Equations, J.Math.Phys. 40, 220-231,(1961).