QUARTERLY OF APPLIED MATHEMATICS
VOLUME LXI, NUMBER 3
SEPTEMBER 2003, PAGES 537–564

L-DERIVATIVE OF AN APPROXIMATE SOLUTION TO $\mathbf{\Phi}' = \mathbf{A}(t)\mathbf{\Phi}$: SERIES AND PRODUCT FORMULAE FOR LEFT CORRECTIONS

BY

IGOR NAJFELD

AND

WILLIAM LAKIN

Abstract. Given an initial approximation $\mathbf{\Phi}_0$ to the fundamental matrix of solutions for $\mathbf{\Phi}' = \mathbf{A}(t)\mathbf{\Phi}$, it is shown that a left correction, $\mathbf{\Gamma}\mathbf{\Phi}_0$, is locally more accurate than a right correction, $\mathbf{\Phi}_0\mathbf{\Gamma}$. For each relative error function considered, there is a left correction $\mathbf{\Gamma}$ and the associated differential equation. The common feature is the same integrable part whose forcing function is the difference between L-derivatives of the exact and the initial solution.

Upon transformation into a Volterra integral equation, fixed point iterations generate infinite series of a lacunary type which converge globally whenever an integral equation is linear. Alternatively, when the integrable solution is used for iterative refinement, the outcomes are infinite product representations. Necessary conditions for the absolute convergence are given.

1. Introduction. Two central problems in approximating the solution, $\mathbf{\Phi}(t, t_0)$, of a d-dimensional system of linear differential equations

$$\mathbf{\Phi}' = \mathbf{A}(t)\,\mathbf{\Phi}, \qquad \mathbf{\Phi}(t_0) = \mathbf{I}, \tag{1.1}$$

are how to choose an initial approximation, $\mathbf{\Phi}_0(t, t_0)$, to the solution of Eq. 1.1, and how to select an error function, $\mathbf{E}(t, t_0)$. Once these decisions are made, it is usually not difficult to derive the differential equation satisfied by $\mathbf{E}(t, t_0)$. Although it is not necessarily a linear one, its integrable part is linear and its solution, $\mathbf{E}_0(t, t_0)$, represents a dominant approximation to $\mathbf{E}(t, t_0)$. It is these two functions, $\mathbf{\Phi}_0(t, t_0)$ and $\mathbf{E}_0(t, t_0)$, that make solving the rest of the differential equation much easier compared to the direct methods for representing or computing $\mathbf{\Phi}(t, t_0)$ that start from the identity matrix. The

Received October 11, 2001 and, in revised form, March 17, 2003.
2000 *Mathematics Subject Classification.* Primary 34A45, 47B99; Secondary 34A99, 15A57.
Key words and phrases. L-derivative, product representation, relative error function, commutator series, symplectic matrices.
E-mail address: igor@emba.uvm.edu
Current address: Dept. of Mathematics and Statistics, University of Vermont, Burlington, VT.
E-mail address: lakin@emba.uvm.edu

entire process depends heavily on the L-derivative operator, both for the selection of an initial approximation and for the control of the convergence in either series or product representations.

Section 2 starts with a simple analytic example, then compares several admissible initial approximations and the associated L-derivatives. This motivates Theorem 1, which demonstrates the greater accuracy of a left correction. This section ends with an analysis of two common initial approximations and the computation of their L-derivatives. Section 3 contains a brief overview of the absolute error function, while in Sec. 4, the classical relative error is analyzed in more detail. Its Volterra integral equation generates a new lacunary series representation that converges globally. Its differential equation, more precisely, its integrable part generates a new product representation that has a finite radius of convergence. Section 5 briefly deals with the modified relative error. In Sec. 6, the A-relative error is defined and another product representation is derived from its Riccati differential equation. In this case, any finite product preserves both unitary and symplectic properties, provided an initial approximation is similarly equipped. Section 7 deals with the main computational step common to all three products considered here, and Sec. 8 contains final remarks.

To simplify notation, we set $t_0 = 0$, so that the fundamental matrix of solutions can be written as $\mathbf{\Phi}(t) = \mathbf{\Phi}(t, 0)$, unless the second argument becomes relevant during the integration or recursion. The matrix $\mathbf{A}(t) \in \mathbb{C}^{d\times d}$ is assumed to be piecewise continuous over a finite time domain $D = [0, T]$, thus providing a sufficient condition for the existence, uniqueness, and differentiability of the solution in Eq. 1.1 (see, e.g., [5], [9], [12]). All constant matrices are declared as such at the beginning of a paragraph/section and the explicit dependence on t, in all time varying matrix functions, is frequently omitted whenever this is not confusing. Thus, an expression $\alpha(t)\mathbf{I} + \mathbf{R}(t)$, where $\mathbf{I}$ is an identity matrix of the same dimension as the matrix $\mathbf{R}$ and $\alpha(t)$ is a scalar function, is written simply as $\alpha + \mathbf{R}$.

An initial approximation $\mathbf{\Phi}_0(t)$ is said to be *admissible* if it satisfies the following four conditions: 1) $\mathbf{\Phi}_0(0) = \mathbf{I}$, and is 2) bounded, 3) differentiable, and 4) invertible, for all $t \in D$. Clearly, all four conditions are satisfied by the solution of 1.1, but the last two conditions are especially pertinent when one considers the sort of logarithmic derivative of $\mathbf{\Phi}_0(t)$ defined by

$$\mathsf{L}\left[\mathbf{\Phi}_0(t)\right] = \mathbf{\Phi}_0'(t)\ \mathbf{\Phi}_0^{-1}(t)\,. \tag{1.2}$$

This operator is also known as the multiplicative derivative [9], or as the Cartan's matrix of $\mathbf{\Phi}_0(t)$ [11], because, strictly speaking, $\mathsf{L}\left[\mathbf{\Phi}_0(t)\right]$ is not generally equal to the logarithmic derivative, $(d/dt)\ln\left(\mathbf{\Phi}_0(t)\right)$. The name "multiplicative derivative" is ambiguous for other reasons and the matrix in Eq. 1.2 was known long before E. Cartan, e.g., in kinematics where it goes under the name of the angular velocity matrix [3]. As there is no widely accepted name for the operator in Eq. 1.2, it is termed here the L-*derivative*. Equation 1.2 simply means that if $\mathbf{A}_0(t) = \mathsf{L}\left[\mathbf{\Phi}_0(t)\right]$, then $\mathbf{\Phi}_0(t)$ is any matrix containing the full set of n linearly independent solutions corresponding to the linear system $\mathbf{\Phi}_0' = \mathbf{A}_0(t)\,\mathbf{\Phi}_0$.

The difference between the two L-derivatives, $\mathbf{A} = \mathsf{L}[\mathbf{\Phi}]$ and $\mathbf{A}_0 = \mathsf{L}[\mathbf{\Phi}_0]$, is termed the L-perturbation and is denoted by

$$\mathbf{\Delta}_0(t) = \mathbf{A}(t) - \mathbf{A}_0(t). \tag{1.3}$$

It is assumed that in some neighborhood of the origin the norm inequality,

$$\|\mathbf{\Delta}_0(t)\| \leq \delta t^p, \qquad p \geq 1, \tag{1.4}$$

holds for some integer p. Several examples of approximations satisfying condition 1.4 will be given in Sec. 2.

For the properties of L-derivatives, see [5], [9], or [11]. Here, only the multiplicative property is used: if $\mathbf{\Phi}(t)$ and $\mathbf{\Psi}(t)$ are two nonsingular differentiable matrix functions, having L-derivatives $\mathbf{A}(t) = \mathsf{L}[\mathbf{\Phi}(t)]$ and $\mathbf{B}(t) = \mathsf{L}[\mathbf{\Psi}(t)]$, then the L-derivative of their product satisfies

$$\mathsf{L}[\mathbf{\Phi}\mathbf{\Psi}] = \mathsf{L}[\mathbf{\Phi}] + \mathbf{\Phi}\mathsf{L}[\mathbf{\Psi}]\mathbf{\Phi}^{-1} = \mathbf{A} + \mathbf{\Phi}\mathbf{B}\mathbf{\Phi}^{-1}. \tag{1.5}$$

It is noted in [11] that "...this formula contains much of the differential geometry". In the case of differential equations, Eq. 1.5 has, so far, found virtually no applications. One of the objectives of this paper is to show that the L-perturbation is a key quantity for the analysis and computation of both series and product representations.

A system of linear differential equations is said to be *integrable* if its solution can be expressed in terms of the finite sums or products of well-defined functions and their quadratures. A general linear system, such as 1.1, is not integrable since, for example, its Peano series

$$\mathbf{\Phi}(t) = \mathbf{I} + \int_0^t \mathbf{A}(\sigma)d\sigma + \int_0^t \mathbf{A}(\sigma)\int_0^\sigma \mathbf{A}(\tau)d\tau d\sigma + \ldots \tag{1.6}$$

is not a finite sum, or, as another example, its product integral [5] is not a finite product.

When $\mathbf{A}(t)$ is piecewise continuous, the series in Eq. 1.6 is derivative-free but nevertheless displays a Taylor-like character in the sense that the norm of k-th term behaves as $O(t^k)$ as $t \to 0$. Both matrix $\mathbf{O}$ and scalar O Landau symbols will be used to denote the asymptotic behavior when $t \to 0$. For example, $\mathbf{\Delta}(t) = \mathbf{O}(t^p)$ means that there is a constant matrix $\mathbf{\Omega}$, with at least one nonzero entry, such that

$$\mathbf{\Delta}_{ij}(t) = \mathbf{\Omega}_{ij}t^p + o(t^p) \tag{1.7}$$

as $t \to 0$, for all nonzero ij entries, where $o(t^p)$ denotes scalar terms of order higher than p (see, e.g., [20]).

A matrix $\mathbf{X}(t)$ is said to *strongly dominate*, or just to *dominate*, a matrix $\mathbf{Y}(t)$ if $|\mathbf{X}_{ij}(t)| \geq |\mathbf{Y}_{ij}(t)|$ for all nonzero entries and all t in some neighborhood of zero. For example, $\mathbf{X}(t) = \mathbf{O}(t^{q_1})$ dominates $\mathbf{Y}(t) = \mathbf{O}(t^{q_2})$, provided $1 \leq q_1 < q_2$ and the inequality holds for *all* nonzero entries. In the case of mismatching entries, that is, when inequality for order holds but not for all nonzero entries, only a weaker statement holds, namely that $\|\mathbf{X}(t)\| = O(t^{q_1})$ and $\|\mathbf{Y}(t)\| = O(t^{q_2})$, in which case $\mathbf{X}(t)$ is said to *weakly dominate* $\mathbf{Y}(t)$.

In what follows, there will be a frequent occurrence of commutators in the differential equations for the error. This indicates that an approximation has violated, in some way, the time ordering in either its Peano series or its product integral. As pointed out in

[9], "The whole peculiarity of the multiplicative (product) integral is tied up with the fact that the various values of the matrix function $\mathbf{A}(t)$ ... are not permutable (do not commute)". An extreme example is that of the Magnus' expansion [16], where the matrix in the exponent is given by an infinite series of iterated integrals of iterated commutators.

The commutator of the matrices $\mathbf{X}$ and $\mathbf{Y}$ is defined by $[\mathbf{X},\mathbf{Y}] = \mathbf{XY} - \mathbf{YX}$. The k-th commutator power is defined recursively as

$$\{\mathbf{X}^0,\mathbf{Y}\} = \mathbf{Y}, \quad \{\mathbf{X}^k,\mathbf{Y}\} = [\mathbf{X},\{\mathbf{X}^{k-1},\mathbf{Y}\}], \qquad k \geq 1.$$

The addition property for the ordinary matrix powers also holds for the commutator powers, i.e., $\{\mathbf{X}^{k+l},\mathbf{Y}\} = \{\mathbf{X}^k,\{\mathbf{X}^l,\mathbf{Y}\}\}$, and this property can be extended to the negative powers whenever $\mathbf{X}$ is invertible. If $f(\mathbf{X}) = \sum_{k=0}^{\infty} c_k \mathbf{X}^k$ is any analytic function of a matrix argument, then the commutator function, or commutator series, is defined by

$$\{f(\mathbf{X}),\mathbf{Y}\} = \sum_{k=0}^{\infty} c_k \{\mathbf{X}^k,\mathbf{Y}\}. \tag{1.8}$$

The most useful example of Eq. 1.8, due most likely to Baker [2] and Hausdorff [13], is that of the exponential function,

$$e^{\mathbf{X}}\mathbf{Y}e^{-\mathbf{X}} = \{e^{\mathbf{X}},\mathbf{Y}\} = \sum_{k=0}^{\infty}\left\{\frac{\mathbf{X}^k}{k!},\mathbf{Y}\right\} = \mathbf{Y} + \frac{1}{1!}[\mathbf{X},\mathbf{Y}] + \frac{1}{2!}[\mathbf{X},[\mathbf{X},\mathbf{Y}]] + \ldots. \tag{1.9}$$

2. Initial approximation and its L-derivative. To illustrate what is involved in choosing an initial approximation, consider the simplest analytic time-varying matrix,

$$\mathbf{A}(t) = \mathbf{B} + 2\mathbf{C}t. \tag{2.1}$$

The matrix coefficients in Eq. 2.1 are constant and noncommuting, $[\mathbf{B},\mathbf{C}] \neq \mathbf{0}$, and the scalar coefficient of t has been set to 2 to simplify typography. Apart from a few isolated cases (e.g., when $[\mathbf{B},\mathbf{C}]$ satisfies additional conditions, [22], [27]), this system is not integrable. On the other hand, since $\mathbf{A}(t)$ is analytic, so is its fundamental matrix of solutions $\mathbf{\Phi}(t)$, whose Taylor series

$$\mathbf{\Phi}(t) = \mathbf{I} + t\mathbf{B} + \frac{t^2}{2!}(\mathbf{B}^2 + 2\mathbf{C}) + \frac{t^3}{3!}(\mathbf{B}^3 + 2\mathbf{BC} + 4\mathbf{CB}) + \ldots$$

is more useful for comparison to a variety of analytic approximations than for the actual computation.

2.1. *Selected exponential approximations.* The simplest admissible initial approximation, with $\mathbf{A}(t)$ as in Eq. 2.1, is $\mathbf{\Phi}_0(t) = \exp(t\mathbf{B})$. Its L-derivative is $\mathbf{A}_0(t) = \mathbf{B}$ and its L-perturbation is $\mathbf{\Delta}_0(t) = 2t\mathbf{C} = \mathbf{O}(t)$. A somewhat more accurate approximation is

$$\mathbf{\Phi}_0(t) = e^{t^2\mathbf{C}}e^{t\mathbf{B}}, \tag{2.2}$$

with the L-perturbation computed as follows:

$$\mathbf{\Delta}_0(t) = \mathbf{A}(t) - \mathbf{A}_0(t) = \mathbf{B} + 2t\mathbf{C} - \left(2t\mathbf{C} + e^{t^2\mathbf{C}}\mathbf{B}e^{-t^2\mathbf{C}}\right) = \mathbf{B} - e^{t^2\mathbf{C}}\mathbf{B}e^{-t^2\mathbf{C}}. \tag{2.3}$$

However, to assess the asymptotic order of $\mathbf{\Delta}_0(t)$, Eq. 2.3 is written as the commutator series

$$\mathbf{\Delta}_0(t) = -\left\{e^{t^2\mathbf{C}} - \mathbf{I}, \mathbf{B}\right\} = -\sum_{k=1}^{\infty} \left\{\mathbf{C}^k, \mathbf{B}\right\} \frac{t^{2k}}{k!} = [\mathbf{B}, \mathbf{C}]\, t^2 - [[\mathbf{B}, \mathbf{C}], \mathbf{C}] \frac{t^4}{2!} + \ldots . \quad (2.4)$$

The error between the exact and the approximate fundamental matrix of solutions is

$$\mathbf{\Phi}(t) - \mathbf{\Phi}_0(t) = [\mathbf{B}, \mathbf{C}] \frac{t^3}{3} + \mathbf{O}\left(t^4\right).$$

The property that the leading error term in the error matrix, $\mathbf{\Phi}(t) - \mathbf{\Phi}_0(t)$, is the integral of the leading term in $\mathbf{\Delta}_0(t)$ holds for all admissible approximations. This follows from the linear differential equation for the error matrix in Equations 3.2 and 3.3.

It is instructive to look now at the related approximation,

$$\widehat{\mathbf{\Phi}}_0(t) = e^{t\mathbf{B}} e^{t^2\mathbf{C}} \quad (2.5)$$

which differs from 2.2 only in the order of factors. Its L-perturbation is given by

$$\widehat{\mathbf{\Delta}}_0(t) = \mathbf{A}(t) - \widehat{\mathbf{A}}_0(t) = \mathbf{B} + 2t\mathbf{C} - \left(\mathbf{B} + 2te^{t\mathbf{B}}\mathbf{C}e^{-t\mathbf{B}}\right) = 2t\left(\mathbf{C} - e^{t\mathbf{B}}\mathbf{C}e^{-t\mathbf{B}}\right)$$

$$= -2t\left\{e^{t\mathbf{B}} - \mathbf{I}, \mathbf{C}\right\} = -2t\sum_{k=1}^{\infty} \left\{\mathbf{B}^k, \mathbf{C}\right\} \frac{t^k}{k!} = -2[\mathbf{B}, \mathbf{C}]\, t^2 - [\mathbf{B}, [\mathbf{B}, \mathbf{C}]]\, t^3 + \ldots, \quad (2.6)$$

and the error of this approximation is

$$\mathbf{\Phi}(t) - \widehat{\mathbf{\Phi}}_0(t) = -[\mathbf{B}, \mathbf{C}] \frac{2t^3}{3} + \mathbf{O}\left(t^4\right).$$

It is puzzling that approximation 2.2 is more accurate than approximation 2.5, as indicated by the doubling of the commutator $[\mathbf{B}, \mathbf{C}]$ and the appearance of odd powers in expansion 2.6. There may be a number of interpretations why more accuracy is achieved when the factor $\exp(t\mathbf{B})$ (generated by the matrix $t\mathbf{B}$ that weakly dominates the matrix $t^2\mathbf{C}$) acts first on the initial conditions (identity, in this case), followed by the left, corrective, factor, $\exp\left(t^2\mathbf{C}\right)$. What is important, however, is that this property, namely, how the order and the dominance of the exponents affect the accuracy of approximation, holds in a general analytic case.

THEOREM 1. Let $\mathbf{A}(t) = \mathbf{C}_0 + \mathbf{C}_1 t + \mathbf{C}_2 t^2/2! + \ldots$ be an analytic matrix for the system 1.1 and let $\mathbf{A}(t) = \mathbf{X}(t) + \mathbf{Y}(t)$ be arbitrarily partitioned into two matrices,

$$\mathbf{X}(t) = \mathbf{C}_0 + \ldots, \qquad \mathbf{Y}(t) = \mathbf{C}_k t^k/k! + \ldots, \quad (2.7)$$

for some $k > 0$ such that the condition,

$$[\mathbf{C}_0, \mathbf{C}_k] \neq \mathbf{0}, \quad (2.8)$$

holds. If one denotes the corresponding solutions by

$$\mathsf{L}[\mathbf{\Xi}(t)] = \mathbf{X}(t), \qquad \mathsf{L}[\mathbf{\Upsilon}(t)] = \mathbf{Y}(t), \quad (2.9)$$

then, in the vicinity of the origin, the product $\mathbf{\Upsilon}(t)\,\mathbf{\Xi}(t)$ is a more accurate approximation to $\mathbf{\Phi}(t)$ than the product $\mathbf{\Xi}(t)\,\mathbf{\Upsilon}(t)$.

Proof. The partition 2.7 implies that in a vicinity of the origin the matrix $\mathbf{X}(t)$ weakly dominates the matrix $\mathbf{Y}(t)$. By Magnus' theorem [16], there is, in a vicinity of the origin, the unique exponential representation for each matrix in Eq. 2.9,

$$\begin{aligned} \mathbf{\Xi}(t) &= \exp\left(\int_0^t \mathbf{X}(\sigma)\, d\sigma + ...\right) = \exp\left(\mathbf{C}_0 t + ...\right) \\ \mathbf{\Upsilon}(t) &= \exp\left(\int_0^t \mathbf{Y}(\sigma)\, d\sigma + ...\right) = \exp\left(\mathbf{C}_k \frac{t^{k+1}}{(k+1)!} + ...\right). \end{aligned} \tag{2.10}$$

Consider first the product $\mathbf{\Upsilon}(t)\,\mathbf{\Xi}(t)$ and its L-perturbation matrix,

$$\mathbf{\Delta} = \mathbf{A} - \mathsf{L}\left[\mathbf{\Upsilon}\,\mathbf{\Xi}\right] = \mathbf{X} + \mathbf{Y} - \left(\mathbf{Y} + \mathbf{\Upsilon}\,\mathbf{X}\,\mathbf{\Upsilon}^{-1}\right) = \mathbf{X} - \mathbf{\Upsilon}\,\mathbf{X}\,\mathbf{\Upsilon}^{-1}, \tag{2.11}$$

obtained by means of Eq. 1.5. Next, consider the product $\mathbf{\Xi}(t)\,\mathbf{\Upsilon}(t)$ and compute its L-perturbation matrix,

$$\widehat{\mathbf{\Delta}} = \mathbf{A} - \mathsf{L}\left[\mathbf{\Xi}\,\mathbf{\Upsilon}\right] = \mathbf{X} + \mathbf{Y} - \left(\mathbf{X} + \mathbf{\Xi}\,\mathbf{Y}\,\mathbf{\Xi}^{-1}\right) = \mathbf{Y} - \mathbf{\Xi}\,\mathbf{Y}\,\mathbf{\Xi}^{-1}. \tag{2.12}$$

Substitution of the exponential representations 2.10 into Equations 2.11 and 2.12, followed by the use of expansion 1.9, while keeping only the lowest order terms, leads to

$$\begin{aligned} \mathbf{\Delta} &= \mathbf{X} - \exp\left(\mathbf{C}_k \frac{t^{k+1}}{(k+1)!} + ...\right) \mathbf{X} \exp\left(-\mathbf{C}_k \frac{t^{k+1}}{(k+1)!} + ...\right) \\ &= -\left\{\exp\left(\mathbf{C}_k \frac{t^{k+1}}{(k+1)!} + ...\right) - \mathbf{I},\, \mathbf{C}_0 + ...\right\} = \left[\mathbf{C}_0, \mathbf{C}_k\right] \frac{t^{k+1}}{(k+1)!} + ... \end{aligned} \tag{2.13}$$

In a similar way,

$$\begin{aligned} \widehat{\mathbf{\Delta}} &= \mathbf{Y} - \exp\left(\mathbf{C}_0 t + ...\right) \mathbf{Y} \exp\left(-\mathbf{C}_0 t + ...\right) \\ &= -\left\{\exp\left(\mathbf{C}_0 t + ...\right) - \mathbf{I},\, \mathbf{C}_k \frac{t^k}{k!} + ...\right\} = -\left[\mathbf{C}_0, \mathbf{C}_k\right] \frac{t^{k+1}}{k!} + ... \end{aligned} \tag{2.14}$$

Although both matrices $\mathbf{\Delta}$ and $\widehat{\mathbf{\Delta}}$ have the same asymptotic order, $\mathbf{O}\left(t^{k+1}\right)$, a comparison of coefficients shows that

$$\widehat{\mathbf{\Delta}}(t) = -(k+1)\,\mathbf{\Delta}(t) + \mathbf{O}\left(t^{k+2}\right). \tag{2.15}$$

The proof is completed when the condition in Eq. 2.8 is invoked, and by recalling that the integral of the leading term in an L-perturbation matrix provides the leading error term in the error matrix. □

Theorem 1 extends rather easily to a general continuous matrix using definitions of weak dominance, asymptotic behavior as in Eq. 1.7, and nonvanishing of the leading commutator. The simplest noncommutativity condition 2.8 can be relaxed at the expense of more complex definitions and longer proofs. However, these will be omitted here as they do not appear to provide additional insight.

The main implication of Theorem 1 is that if one starts with an admissible approximation $\mathbf{\Phi}_0(t)$ that is generated by a weakly dominant matrix $\mathbf{A}_0(t)$, then it is preferable to seek a left-acting multiplicative correction, namely, $\mathbf{\Phi}(t) = \mathbf{\Gamma}\mathbf{\Phi}_0(t)$, especially if k (see Eq. 2.15) is large.

2.2. *Integral averaging.* One of the most frequent exponential approximations for a piecewise continuous matrix $\mathbf{A}(t)$ is that of an integral average,

$$\mathbf{\Phi}_0(t) = \exp\left(\int_0^t \mathbf{A}(\sigma)\, d\sigma\right). \tag{2.16}$$

This is clearly an admissible approximation and it was used, probably for the first time around 1937, for a numerical simulation of flutter dynamics [8]. It has since found many other applications including analytical ones, such as the first term in Magnus' [16] and Fer's [6] expansions. It is assumed that the system in Eq. 1.1 is noncommutative, namely, that it is *not* the case that $\left[\mathbf{A}(t), \int_0^t \mathbf{A}(\sigma)\, d\sigma\right] = 0$ holds almost everywhere in D, otherwise $\mathbf{\Phi}_0(t)$ in Eq. 2.16 would be the exact solution.

To find the L-derivative of the matrix in 2.16, we first compute the time derivative of $\exp(\mathbf{B}(t))$, where $\mathbf{B}(t) = \int_0^t \mathbf{A}(\sigma)\, d\sigma$. This well-known formula, given by

$$\frac{d}{dt} e^{\mathbf{B}(t)} = \int_0^1 e^{\sigma \mathbf{B}(t)}\, \mathbf{B}'(t)\, e^{(1-\sigma)\mathbf{B}(t)} d\sigma = \left(\int_0^1 e^{\sigma \mathbf{B}(t)}\, \mathbf{A}(t)\, e^{-\sigma \mathbf{B}(t)} d\sigma\right) e^{\mathbf{B}(t)}, \tag{2.17}$$

has a long history dating back to at least [1] and [13], followed by [7], [16], then reviewed in [27], with some new material in [19].

The first representation of the L-derivative for the matrix in Eq. 2.16 follows from Eq. 2.17,

$$\mathbf{A}_0(t) = \mathsf{L}\left[\exp\left(\int_0^t \mathbf{A}(\sigma)\, d\sigma\right)\right] = \int_0^1 e^{\sigma \mathbf{B}(t)}\, \mathbf{A}(t)\, e^{-\sigma \mathbf{B}(t)} d\sigma. \tag{2.18}$$

The matrix $\mathbf{A}_0(t)$ may be interpreted as a noncommutative integral average of $\mathbf{A}(t)$, for if $\mathbf{A}(t)$ commutes with $\mathbf{B}(t)$, then $\mathbf{A}_0(t) = \mathbf{A}(t)$.

The second representation, due to Hausdorff [13], follows from integrating Eq. 2.18, the result being a commutator function,

$$\mathbf{A}_0(t) = \int_0^1 \left\{e^{\sigma \mathbf{B}(t)}, \mathbf{A}(t)\right\} d\sigma = \left\{\frac{e^{\mathbf{B}(t)} - \mathbf{I}}{\mathbf{B}(t)}, \mathbf{A}(t)\right\} = \sum_{k=0}^{\infty} \left\{\frac{\mathbf{B}^k(t)}{(k+1)!}, \mathbf{A}(t)\right\}. \tag{2.19}$$

It follows from Eq. 2.19 that the L-perturbation matrix also has a commutator function representation,

$$\mathbf{\Delta}_0(t) = -\left\{\frac{e^{\mathbf{B}(t)} - \mathbf{I} - \mathbf{B}(t)}{\mathbf{B}(t)}, \mathbf{A}(t)\right\} = -\sum_{k=1}^{\infty} \left\{\frac{\mathbf{B}^k(t)}{(k+1)!}, \mathbf{A}(t)\right\}$$

$$= \frac{1}{2!}\left[\mathbf{A}(t), \int_0^t \mathbf{A}(\sigma)\, d\sigma\right] - \frac{1}{3!}\left[\left[\mathbf{A}(t), \int_0^t \mathbf{A}(\sigma)\, d\sigma\right], \int_0^t \mathbf{A}(\sigma)\, d\sigma\right] + \ldots. \tag{2.20}$$

If $\mathbf{A}(t)$ is analytic, then the expansion of the first commutator in Eq. 2.20 into a Taylor's series, at any point $\tau \in [0,t]$, followed by the integration and cancellation of terms, reveals the asymptotic behavior for $\mathbf{\Delta}_0(t)$,

$$\mathbf{\Delta}_0(t) = -\frac{t^2}{4}\left[\mathbf{A}(\tau), \mathbf{A}'(\tau)\right] + \mathbf{O}\left(t^3\right), \qquad \tau \in [0,t]. \tag{2.21}$$

If $\mathbf{A}(t)$ is continuous on D, the quadratic behavior persists, for one may write $\|\mathbf{A}(t) - \mathbf{A}(\sigma)\| \leq L|t-\sigma|$ and then easily deduce, from the first term in expansion

2.20, that the following norm estimate holds:

$$\|\mathbf{\Delta}_0(t)\| \leq \frac{t^2}{4} L \|\mathbf{A}(t)\| + O\left(t^3\right). \tag{2.22}$$

When applied to the example in Eq. 2.1, the integral averaging generates the matrix,

$$\mathbf{\Phi}_0(t) = \exp\left(t\mathbf{B} + t^2\mathbf{C}\right), \tag{2.23}$$

whose L-perturbation matrix, computed from Eq. 2.20, is $\mathbf{\Delta}_0(t) = -[\mathbf{B}, \mathbf{C}]\left(t^2/2\right) + \mathbf{O}\left(t^3\right)$. The leading error term of this approximation is

$$\mathbf{\Phi}(t) - \mathbf{\Phi}_0(t) = -\frac{t^3}{6}[\mathbf{B}, \mathbf{C}] + \mathbf{O}\left(t^4\right), \tag{2.24}$$

which shows why the integral averaging is more attractive when compared to Eq. 2.2 or 2.5.

It is possible to achieve cubic, or higher, asymptotic behavior of $\mathbf{\Delta}_0(t)$ by choosing either more exponential factors in Eq. 2.16 (a finite Fer's expansion), or a more complex integrand in Eq. 2.16 (a finite Magnus' expansion). An example of the two term Magnus' expansion is

$$\mathbf{\Phi}_0(t) = \exp\left(\int_0^t \mathbf{A}(\sigma)\, d\sigma + \frac{1}{2}\int_0^t \left[\mathbf{A}(\sigma), \int_0^\sigma \mathbf{A}(\tau)\, d\tau\right] d\sigma\right). \tag{2.25}$$

When the matrix integrations can be carried out quickly and exactly, as in example 2.1, then the approximation 2.25 is, in spite of its complexity, quite useful since it tries to incorporate the effect of a time varying commutator, clearly the chief obstruction to either a numerical or stability analysis of Eq. 1.1.

The computation of the L-derivative for the approximation 2.16, or 2.25, would be demanding if one is to use either one of Equations 2.18, or 2.19. There is, however, a formulation for the L-derivative described in [19] that is suitable for integral averages in Equations 2.16 and 2.25.

THEOREM 2. Let $\mathbf{X}(t)$ be a piecewise continuous matrix and let $\mathbf{M}(t)$ be the following block triangular matrix:

$$\mathbf{M}(t) = \begin{bmatrix} \int_0^t \mathbf{X}(\sigma)\, d\sigma & \mathbf{X}(t) \\ \mathbf{0} & \int_0^t \mathbf{X}(\sigma)\, d\sigma \end{bmatrix}. \tag{2.26}$$

Then the matrix exponential $\exp(\mathbf{M}(t))$ inherits the block triangular structure,

$$e^{\mathbf{M}(t)} = \begin{bmatrix} \mathbf{F}(t) & \mathbf{H}(t) \\ \mathbf{0} & \mathbf{F}(t) \end{bmatrix}, \tag{2.27}$$

where $\mathbf{F}(t) = \exp\left(\int_0^t \mathbf{X}(\sigma)\, d\sigma\right)$ and

$$\mathsf{L}[\mathbf{F}(t)] = \mathsf{L}\left[\exp\left(\int_0^t \mathbf{X}(\sigma)\, d\sigma\right)\right] = \mathbf{H}(t)\,\mathbf{F}^{-1}(t). \tag{2.28}$$

Proof. The block triangular matrices, partitioned identically and with square diagonal blocks, are closed under linear combinations and matrix multiplications. Therefore, any analytic function of such a block triangular matrix inherits the same structure. It follows from the series expansion that the diagonal blocks in Eq. 2.27 are the same and equal to $\mathbf{F}(t) = \exp\left(\int_0^t \mathbf{X}(\sigma)\, d\sigma\right)$. For a proof that the off-diagonal block, $\mathbf{H}(t)$, is the directional derivative of $\mathbf{F}(t)$ in the direction of $\mathbf{X}(t)$, see [19]. Since $\mathbf{H}(t)$ is the directional derivative, the special direction being the time derivative of $\int_0^t \mathbf{X}(\sigma)\, d\sigma$, then it is the time derivative of the exponential in Eq. 2.17 and therefore one can write

$$\mathbf{H}(t) = \frac{d}{dt} \exp\left(\int_0^t \mathbf{X}(\sigma)\, d\sigma\right) = \mathsf{L}\left[\exp\left(\int_0^t \mathbf{X}(\sigma)\, d\sigma\right)\right] \mathbf{F}(t). \tag{2.29}$$

Now the right inversion by $\mathbf{F}(t)$, which is nonsingular for all $t \in D$, proves Eq. 2.28. □

The Equations 2.27–2.28 are much simpler and faster to evaluate than either the quadrature in Eq. 2.18 or the summation in Eq. 2.19, provided the evaluation in Eq. 2.27 takes into account the special block triangular structure of $\mathbf{M}(t)$. Other methods for evaluating directional derivatives of the matrix exponential, and the other analytic functions, are described in [19].

2.3. *Step functions and finite product integrals.* The second example of an admissible initial approximation and its L-derivative is one of considerable generality. A step function approximation for an arbitrary piecewise continuous matrix function $\mathbf{A}(t)$ depends on two number sequences: a knot sequence $\{t_k\}_{k=0}^n$, which defines an n-partition of the interval $[0, T]$,

$$0 = t_0 < t_1 < \ldots < t_{n-1} < t_n = T, \qquad h_k = t_k - t_{k-1},$$

and a sequence of interval points, $\{\sigma_k\}_{k=1}^n$, defined by the inequality $t_{k-1} < \sigma_k \leq t_k$. Then the associated matrix step function is defined by

$$\mathbf{A}_0(t) = \mathbf{A}(\sigma_k), \qquad t_{k-1} < t, \sigma_k \leq t_k. \tag{2.30}$$

The fundamental matrix of solutions corresponding to $\mathbf{A}_0(t)$ in Eq. 2.30 is known to be an ordered finite product of matrix exponentials (see [5], [25]),

$$\mathbf{\Phi}_0(t) = e^{(t-t_{k-1})\mathbf{A}(\sigma_k)} \prod_{i=1}^{k-1} e^{h_i \mathbf{A}(\sigma_i)}, \qquad t_{k-1} < t, \sigma_k \leq t_k. \tag{2.31}$$

This result can be traced to Volterra's work in 1887 (see, e.g., [5]). It follows that its L-perturbation is given by a piecewise continuous matrix function

$$\mathbf{\Delta}_0(t) = \mathbf{A}(t) - \mathbf{A}_0(t) = \mathbf{A}(t) - \mathbf{A}_0(\sigma_k), \qquad t_{k-1} < t, \sigma_k \leq t_k, \tag{2.32}$$

and can therefore be uniformly, or adaptively, bounded over the entire interval D by an adroit choice of knots and interval points. The advantage of Equations 2.30–2.32 lies in their utmost simplicity for evaluation, differentiation, and inversion. Moreover, they preserve the unitary, or symplectic, property of $\mathbf{\Phi}_0(t)$ for all t, whenever $\mathbf{A}(t), \mathbf{A}_0(t)$ are skewhermitian, or $\mathbf{J}$-symmetric. In addition, the condition 2.30 can be extended to include more general approximation schemes.

3. Absolute error. For any admissible initial approximation $\mathbf{\Phi}_0(t)$, the absolute error matrix is defined by

$$\mathbf{E}(t) = \mathbf{\Phi}(t) - \mathbf{\Phi}_0(t).$$

If $\mathbf{A}_0(t) = \mathsf{L}[\mathbf{\Phi}_0(t)]$, and $\mathbf{\Delta}_0(t) = \mathbf{A}(t) - \mathbf{A}_0(t)$, then the differential equation for $\mathbf{E}(t)$, obtained from Equations 1.1 and 1.3, is given by

$$\mathbf{E}'(t) = \mathbf{\Delta}_0(t)\,\mathbf{\Phi}_0(t) + \mathbf{A}_0(t)\,\mathbf{E}(t) + \mathbf{\Delta}_0(t)\,\mathbf{E}(t), \qquad \mathbf{E}(0) = \mathbf{0}. \tag{3.1}$$

The integrable part of Eq. 3.1 is the differential equation for $\mathbf{E}_0(t)$,

$$\mathbf{E}_0'(t) = \mathbf{A}_0(t)\,\mathbf{E}_0(t) + \mathbf{\Delta}_0(t)\,\mathbf{\Phi}_0(t), \qquad \mathbf{E}_0(0) = \mathbf{0}. \tag{3.2}$$

The solution of Eq. 3.2 is given explicitly by the quadrature

$$\mathbf{E}_0(t) = \int_0^t \mathbf{\Phi}_0(t,\sigma)\;\mathbf{\Delta}_0(\sigma)\;\mathbf{\Phi}_0(\sigma,0)\,d\sigma. \tag{3.3}$$

A quick asymptotic analysis shows that if the L-perturbation matrix satisfies $\mathbf{\Delta}_0(t) = \mathbf{O}(t^p)$, then $\mathbf{E}_0(t) = \mathbf{O}\left(t^{p+1}\right)$ provides the dominant approximation for $\mathbf{E}(t)$, since it can be shown that $\mathbf{E}(t) = \mathbf{E}_0(t) + \mathbf{O}\left(t^{2p+2}\right)$.

An analysis of the absolute error will not be taken beyond this point for two reasons. First, the relative error functions, considered in the next three sections, appear to be a much more suitable measure of accuracy for the nonsingular matrix functions, especially if a norm, or the entries, or the characteristic multipliers, are rapidly varying functions over a wide magnitude range. Second, the analysis of the three different relative error functions, given below, considerably replicate the case of the absolute error.

4. The classical relative error. Let $\mathbf{\Phi}_0(t)$ be an admissible initial approximation and let $\mathbf{A}_0(t)$, $\mathbf{\Delta}_0(t)$ be its L-derivative and L-perturbation matrix, respectively. The classical relative error matrix is defined by

$$\mathbf{R}(t) = \mathbf{I} - \mathbf{\Phi}_0(t)\,\mathbf{\Phi}^{-1}(t). \tag{4.1}$$

The advantage of this definition is that the correction factor acts on the left hand side of an initial approximation,

$$\mathbf{\Phi}(t) = (\mathbf{I} - \mathbf{R}(t))^{-1}\,\mathbf{\Phi}_0(t). \tag{4.2}$$

This is desirable given Theorem 1 and the remarks that follow it. By differentiating $(\mathbf{I} - \mathbf{R})\,\mathbf{\Phi} = \mathbf{\Phi}_0$ (from Eq. 4.2) and substituting Equations 1.1, the differential equation for $\mathbf{R}(t)$ first appears as $\mathbf{R}' = \mathbf{A} - \mathbf{A}_0 + \mathbf{A}_0\mathbf{R} - \mathbf{R}\mathbf{A}$, but it may be rewritten, by means of the matrix $\mathbf{\Delta}_0(t)$, as

$$\mathbf{R}' = \mathbf{\Delta}_0 + [\mathbf{A}_0, \mathbf{R}] - \mathbf{R}\mathbf{\Delta}_0, \qquad \mathbf{R}(0) = \mathbf{0}. \tag{4.3}$$

The terms in Eq. 4.3 are in ascending asymptotic order. To see this, assume that $\mathbf{\Delta}_0 = \mathbf{O}(t^p)$. Then $\mathbf{R} = \mathbf{O}\left(t^{p+1}\right)$, so that $[\mathbf{A}_0, \mathbf{R}] = \mathbf{O}\left(t^{p+1}\right)$, and the last term $\mathbf{R}\mathbf{\Delta}_0 = \mathbf{O}\left(t^{2p+1}\right)$. The integrable part of Eq. 4.3, denoted by $\mathbf{R}_0(t)$, satisfies the linear differential equation,

$$\mathbf{R}_0' = \mathbf{\Delta}_0 + [\mathbf{A}_0, \mathbf{R}_0], \qquad \mathbf{R}_0(0) = \mathbf{0}. \tag{4.4}$$

Hence, it is explicitly given by the quadrature of a known matrix function,

$$\mathbf{R}_0(t) = \int_0^t \mathbf{\Phi}_0(t,\sigma)\ \mathbf{\Delta}_0(\sigma)\ \mathbf{\Phi}_0^{-1}(t,\sigma)\, d\sigma. \tag{4.5}$$

The formal solution of Eq. 4.3 can now be written as a linear Volterra integral equation for $\mathbf{R}(t)$ in terms of the known matrices $\mathbf{\Phi}_0(t)$, $\mathbf{\Delta}_0(t)$, and $\mathbf{R}_0(t)$,

$$\mathbf{R}(t) = \mathbf{R}_0(t) - \int_0^t \mathbf{\Phi}_0(t,\sigma)\ \mathbf{R}(\sigma)\,\mathbf{\Delta}_0(\sigma)\ \mathbf{\Phi}_0^{-1}(t,\sigma)\, d\sigma. \tag{4.6}$$

From Eq. 4.5 and the assumption in Eq. 1.4 on the asymptotic behavior of $\mathbf{\Delta}_0$, it follows that $\mathbf{R}_0(t) = \mathbf{O}\left(t^{p+1}\right)$ is indeed the dominant part of solution $\mathbf{R}(t)$ since the integral part of Eq. 4.6 behaves as $\mathbf{O}\left(t^{2p+2}\right)$.

The relevant reference here is [28], which starts with a similar definition for the relative error as in Eq. 4.1. However, the goal in [28] was to derive the bounds on the solution and quite a different route was taken. This reference also contains a number of useful applications in physics.

4.1. *Fixed point iteration for the relative error.* The linear matrix integral equation 4.6 is already in a fixed point formulation. It can, therefore, be solved by iteration, starting with the initial approximation $\mathbf{R}_0(t)$. What is different in Eq. 4.6 from the usual linear Volterra integral equation is the appearance of a "split", or two-sided, kernel, a consequence of two-sided differential equation 4.3. However, this two-sided feature has no major detrimental effect on global convergence. It is the size of the matrix $\mathbf{\Delta}_0(t)$, that has a decisive effect on the convergence rate.

THEOREM 3. If the integral equation 4.6 is solved by iteration, starting with $\mathbf{R}_0(t)$, then the iterates form the series for the solution

$$\mathbf{R}(t) = \sum_{k=0}^{\infty} \mathbf{R}_k(t), \tag{4.7}$$

and are linked by the integro-difference equation

$$\mathbf{R}_{k+1}(t) = -\int_0^t \mathbf{\Phi}_0(t,\sigma)\ \mathbf{R}_k(\sigma)\,\mathbf{\Delta}_0(\sigma)\ \mathbf{\Phi}_0^{-1}(t,\sigma)\, d\sigma, \qquad k \geq 0. \tag{4.8}$$

The asymptotic behavior of iterates is given by

$$\mathbf{R}_k(t) = \mathbf{O}\left(t^{(k+1)(p+1)}\right), \qquad k \geq 0. \tag{4.9}$$

Proof. Under the assumption of the admissibility for the initial approximation $\mathbf{\Phi}_0(t)$, the matrix $\mathbf{R}_0(t)$ in Eq. 4.5 is continuous and each part of the kernel in Eq. 4.6 is bounded and integrable for all $t \in D$. One may, in the spirit of [26], define a linear integral operator $\mathbf{Q}$ by its action on an arbitrary matrix function $\mathbf{Z}(t)$ as

$$\mathbf{Q}[\mathbf{Z}(t)] = \int_0^t \mathbf{\Phi}_0(t,\sigma)\ \mathbf{Z}(\sigma)\,\mathbf{\Delta}_0(\sigma)\ \mathbf{\Phi}_0^{-1}(t,\sigma)\, d\sigma. \tag{4.10}$$

The powers of Q are defined according to

$$\mathsf{Q}^2[\mathbf{Z}(t)] = \mathsf{Q}[\mathsf{Q}[\mathbf{Z}(t)]] = \int_0^t \mathbf{\Phi}_0(t,\sigma)\ \mathsf{Q}[\mathbf{Z}(\sigma)]\ \mathbf{\Delta}_0(\sigma)\ \mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma$$
$$= \int_0^t \mathbf{\Phi}_0(t,\sigma)\left(\int_0^\sigma \mathbf{\Phi}_0(\sigma,\tau)\ \mathbf{Z}(\tau)\ \mathbf{\Delta}_0(\tau)\ \mathbf{\Phi}_0^{-1}(\sigma,\tau)\,d\tau\right)\mathbf{\Delta}_0(\sigma)\ \mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma,$$

so that $\mathsf{Q}^{k+1}[\mathbf{Z}(t)] = \mathsf{Q}\left[\mathsf{Q}^k[\mathbf{Z}(t)]\right]$. It follows from definition 4.10 that the integral equation 4.6 can be written as an operator equation

$$(\mathsf{I}+\mathsf{Q})[\mathbf{R}] = \mathbf{R}_0.$$

The formal inversion of the operator $\mathsf{I}+\mathsf{Q}$ generates the operator geometric (Neumann) series

$$\mathbf{R} = \left(\mathsf{I}-\mathsf{Q}+\mathsf{Q}^2-\ldots\right)[\mathbf{R}_0] = \mathbf{R}_0 - \mathsf{Q}[\mathbf{R}_0] + \mathsf{Q}^2[\mathbf{R}_0] - \ldots = \sum_{k=0}^{\infty}\mathbf{R}_k.$$

This implies that $\mathbf{R}_{k+1} = -\mathsf{Q}[\mathbf{R}_k]$, which proves Equations 4.7 and 4.8.

To prove the absolute convergence of this series, we start by bounding the norm of its first term (Eq. 4.5),

$$\|\mathbf{R}_0(t)\| \le \kappa_0(t)\int_0^t \|\mathbf{\Delta}_0(\sigma)\|\,d\sigma, \tag{4.11}$$

where

$$\kappa_0(t) = \max_{0\le\sigma\le t}\|\mathbf{\Phi}_0(t,\sigma)\|\left\|\mathbf{\Phi}_0^{-1}(t,\sigma)\right\|, \qquad \kappa_0(t)\ge 1, \tag{4.12}$$

defines a time-varying condition number for the initial approximation. By the admissibility assumptions on $\mathbf{\Phi}_0(t)$, the function $\kappa_0(t)$ exists for all $t\in D$. Assume, in addition, that $\kappa_0(t)$ is a nondecreasing function of t. This simplifying assumption is not essential since one can always carry out the proof below by using, for example, $\kappa_0(t) < K$ (constant), or some other nondecreasing function that bounds $\kappa_0(t)$ for all $t\in D$. The norm of the second term can be bounded as follows:

$$\|\mathsf{Q}[\mathbf{R}_0]\| = \left\|\int_0^t \mathbf{\Phi}_0(t,\sigma)\ \mathbf{R}_0(\sigma)\ \mathbf{\Delta}_0(\sigma)\ \mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma\right\|$$
$$\le \kappa_0(t)\int_0^t \|\mathbf{R}_0(\sigma)\|\,\|\mathbf{\Delta}_0(\sigma)\|\,d\sigma$$
$$\le \kappa_0^2(t)\int_0^t \|\mathbf{\Delta}_0(\sigma)\|\int_0^\sigma \|\mathbf{\Delta}_0(\tau)\|\,d\tau d\sigma \le \frac{\kappa_0^2(t)}{2!}\left(\int_0^t \|\mathbf{\Delta}_0(\sigma)\|\,d\sigma\right)^2.$$

By induction, the bound on the k-th term is given by

$$\left\|\mathsf{Q}^k[\mathbf{R}_0]\right\| \le \frac{\kappa_0^{k+1}(t)}{k+1!}\left(\int_0^t \|\mathbf{\Delta}_0(\sigma)\|\,d\sigma\right)^{k+1}.$$

Therefore, the norm of the series 4.7 is majorized by

$$\|\mathbf{R}(t)\| \le \sum_{k=1}^{\infty}\frac{\kappa_0^k(t)}{k!}\left(\int_0^t \|\mathbf{\Delta}_0(\sigma)\|\,d\sigma\right)^k = \exp\left(\kappa_0(t)\int_0^t \|\mathbf{\Delta}_0(\sigma)\|\,d\sigma\right) - 1, \tag{4.13}$$

and the last function is bounded for all $t\in D$.

To prove the asymptotic behavior in Eq. 4.9, use an induction on $\mathbf{R}_k(t)$ in the recursion 4.8, starting with the asymptotic order $\mathbf{R}_0(t) = \mathbf{O}\left(t^{p+1}\right)$, as in Eq. 4.6. Multiplication by $\boldsymbol{\Delta}_0(t) = \mathbf{O}(t^p)$, followed by the integration, shows that each iterate receives the asymptotic exponent from its forerunner and the additional power $p+1$, which creates the exponential gap. □

The convergence of the series 4.7 clearly depends on the size of the exponential gap. For $p = 1$, the least favorable case, this lacunary series resembles an even-power expansion since $\mathbf{R}_k(t) = \mathbf{O}\left(t^{2(k+1)}\right)$, provided the integration is carried out exactly. When the integration is carried out numerically, kernel values need to be evaluated and stored only once for repeated use during recurrence 4.8. Although it is advantageous that the series for $\mathbf{R}(t)$ converges for all $t \in D$, as indicated by Eq. 4.13, the convergence is linear since the exponents form an arithmetic progression, and it can be quite slow unless the choice of $\boldsymbol{\Phi}_0(t)$ results in a very small $\|\boldsymbol{\Delta}_0(t)\|$, or the integration step-size is sufficiently small. It is certainly fortunate that the very first iteration step always doubles the asymptotic order of convergence. It may therefore appear useful to iterate Eq. 4.6 once or twice before switching to a faster converging product formula described in Sec. 4.3. However, it should be noted that if the series 4.7 is truncated prematurely, then there is no guarantee that the inversion in Eq. 4.2 can be carried out.

4.2. *Sharper bounds on* $\|\mathbf{R}(t)\|$ *and* $\|\mathbf{R}(t) - \mathbf{R}_0(t)\|$. It is possible to obtain sharper bounds on the growth of $\|\mathbf{R}(t)\|$ than those in Eq. 4.13. For example, taking the norm of both sides in Eq. 4.6 and using the function $\kappa_0(t)$ in Eq. 4.12 results in the linear integral inequality

$$\|\mathbf{R}(t)\| \leq \|\mathbf{R}_0(t)\| + \kappa_0(t) \int_0^t \|\boldsymbol{\Delta}_0(\sigma)\| \, \|\mathbf{R}(\sigma)\| \, d\sigma. \tag{4.14}$$

Since $\|\mathbf{R}(t)\|$, $\|\mathbf{R}_0(t)\|$, $\kappa_0(t)$, and $\|\boldsymbol{\Delta}_0(\sigma)\|$ are all real-valued, continuous, nonnegative functions, one may invoke the generalized Gronwall inequality ([10], [12]) which states that

$$\|\mathbf{R}(t)\| \leq \|\mathbf{R}_0(t)\| + \kappa_0(t) \int_0^t \|\mathbf{R}_0(\sigma)\| \, \|\boldsymbol{\Delta}_0(\sigma)\| \exp\left(\int_\sigma^t \kappa_0(\gamma) \, \|\boldsymbol{\Delta}_0(\gamma)\| \, d\gamma\right) d\sigma. \tag{4.15}$$

An attractive property of inequality 4.15 is that its two terms retain the same asymptotic behavior as the two terms in Eq. 4.6.

Suppose, now, that in taking the norm of Eq. 4.6 one refrains from the maximization process in Eq. 4.12. The result, then, is a different linear integral inequality:

$$\|\mathbf{R}(t)\| \leq \|\mathbf{R}_0(t)\| + \int_0^t \|\boldsymbol{\Phi}_0(t,\sigma)\| \, \|\mathbf{R}(\sigma)\| \, \left\|\boldsymbol{\Delta}_0(\sigma) \, \boldsymbol{\Phi}_0^{-1}(t,\sigma)\right\| d\sigma. \tag{4.16}$$

Let $\rho_0(t) = \|\mathbf{R}_0(t)\|$ and denote the real-valued, nonnegative, scalar, integrable kernel function by

$$K_0(t,\sigma) = \|\boldsymbol{\Phi}_0(t,\sigma)\| \left\|\boldsymbol{\Delta}_0(\sigma) \, \boldsymbol{\Phi}_0^{-1}(t,\sigma)\right\|, \qquad 0 \leq \sigma \leq t. \tag{4.17}$$

Inequality 4.16 can now be related to the linear Volterra integral equation,

$$\rho(t) = \rho_0(t) + \int_0^t K_0(t,\sigma) \, \rho(\sigma) \, d\sigma. \tag{4.18}$$

It is well known (e.g., [17], [26]) that the unique nonnegative function $\rho(t)$ exists for all t for which $K_0(t,\sigma)$ is integrable. Therefore, one can invoke the Chu-Metcalf inequality [4], which states that

$$\|\mathbf{R}(t)\| \leq \rho(t). \tag{4.19}$$

According to [4], the inequality 4.19 is the sharpest bound on the solutions of scalar linear integral inequalities. It follows that this inequality is also among the sharpest bounds for the matrix equation 4.6 because the number of maximizations (upper bounds) was kept to a minimum during the computation of inequality 4.16. The function $\rho(t)$ in Eq. 4.18 certainly involves much more computation than the bound 4.15, but given that a very accurate solution of a scalar integral equation, like Eq. 4.18, can be obtained at a rather small fraction of the cost of the matrix solution, it follows that this bound can be used as an effective step-size control mechanism during integration of Eq. 4.6.

By considering the growth of $\mathbf{Z}(t) = \mathbf{R}(t) - \mathbf{R}_0(t)$, it is possible to obtain even sharper bounds since $\|\mathbf{Z}(t)\|$ measures the size of only the integral part of Eq. 4.6. To do so, we rewrite Eq. 4.6 in terms of $\|\mathbf{Z}(t)\|$,

$$\begin{aligned}\mathbf{Z}(t) &= -\int_0^t \mathbf{\Phi}_0(t,\sigma)\,\mathbf{R}_0(\sigma)\,\mathbf{\Delta}_0(\sigma)\,\mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma \\ &\quad -\int_0^t \mathbf{\Phi}_0(t,\sigma)\,\mathbf{Z}(\sigma)\,\mathbf{\Delta}_0(\sigma)\,\mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma.\end{aligned} \tag{4.20}$$

By taking the norm of Eq. 4.20 and denoting by

$$\zeta_0(t) = \left\| \int_0^t \mathbf{\Phi}_0(t,\sigma)\,\mathbf{R}_0(\sigma)\,\mathbf{\Delta}_0(\sigma)\,\mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma \right\| = O\left(t^{2p+2}\right), \tag{4.21}$$

one obtains the linear integral inequality,

$$\|\mathbf{Z}(t)\| \leq \zeta_0(t) + \kappa_0(t) \int_0^t \|\mathbf{\Delta}_0(\sigma)\|\,\|\mathbf{Z}(\sigma)\|\,d\sigma. \tag{4.22}$$

The generalized Gronwall inequality applied to inequality 4.22 yields the following:

$$\|\mathbf{Z}(t)\| \leq \zeta_0(t) + \kappa_0(t) \int_0^t \zeta_0(\sigma)\,\|\mathbf{\Delta}_0(\sigma)\| \exp\left(\int_\sigma^t \kappa_0(\gamma)\,\|\mathbf{\Delta}_0(\gamma)\|\,d\gamma \right) d\sigma. \tag{4.23}$$

As before, the sharper bound comes from the Chu-Metcalf inequality,

$$\|\mathbf{Z}(t)\| \leq \zeta(t), \tag{4.24}$$

where $\zeta(t)$ is the solution of the scalar linear integral equation

$$\zeta(t) = \zeta_0(t) + \int_0^t K_0(t,\sigma)\,\zeta(\sigma)\,d\sigma. \tag{4.25}$$

To the extent that these bounds are comparable, the bounds 4.23 and 4.24 are sharper than those in Equations 4.15 and 4.19, respectively, but at the cost of evaluating the function $\zeta_0(t)$.

4.3. *Product formula for the classical relative error.* Once the first approximation, $\mathbf{R}_0(t)$, is computed from Eq. 4.5, it can be promptly used, in Eq. 4.2, to correct $\mathbf{\Phi}_0(t)$,

$$\mathbf{\Phi}_1(t) = (\mathbf{I} - \mathbf{R}_0(t))^{-1} \mathbf{\Phi}_0(t). \tag{4.26}$$

By the continuity of $\mathbf{R}_0(t)$ and $\mathbf{R}_0(0) = \mathbf{0}$, the matrix $\mathbf{I} - \mathbf{R}_0(t)$ is nonsingular in some, as yet undetermined, neighborhood of the origin, but it is far from clear that it will remain invertible for all $t \in D$. It is assumed that there is such a neighborhood, but the problem of estimating its size must be postponed. How much better is the new approximation in Eq. 4.26 compared to the initial one? To answer this question, it is essential to compute the L-derivative of the first correction, $\mathbf{A}_1(t) = \mathsf{L}[\mathbf{\Phi}_1(t)]$, and then its L-perturbation, $\mathbf{\Delta}_1(t) = \mathbf{A}(t) - \mathbf{A}_1(t)$. To begin with, compute

$$\mathsf{L}\left[(\mathbf{I} - \mathbf{R}_0(t))^{-1}\right] = (\mathbf{I} - \mathbf{R}_0(t))^{-1} \mathbf{R}_0'(t). \tag{4.27}$$

Then, using Equations 1.5, 4.27, and 4.4, the calculation of the $\mathsf{L}[\mathbf{\Phi}_1]$ proceeds as follows:

$$\begin{aligned}
\mathbf{A}_1 = \mathsf{L}[\mathbf{\Phi}_1] &= \mathsf{L}\left[(\mathbf{I} - \mathbf{R}_0)^{-1} \mathbf{\Phi}_0\right] = \mathsf{L}\left[(\mathbf{I} - \mathbf{R}_0)^{-1}\right] + (\mathbf{I} - \mathbf{R}_0)^{-1} \mathbf{A}_0 (\mathbf{I} - \mathbf{R}_0) \\
&= (\mathbf{I} - \mathbf{R}_0)^{-1} (\mathbf{R}_0' + \mathbf{A}_0 (\mathbf{I} - \mathbf{R}_0)) = (\mathbf{I} - \mathbf{R}_0)^{-1} (\mathbf{A} - \mathbf{R}_0 \mathbf{A}_0) \\
&= (\mathbf{I} - \mathbf{R}_0)^{-1} ((\mathbf{I} - \mathbf{R}_0)\mathbf{A} + \mathbf{R}_0 (\mathbf{A} - \mathbf{A}_0)) = \mathbf{A} + (\mathbf{I} - \mathbf{R}_0)^{-1} \mathbf{R}_0 \mathbf{\Delta}_0.
\end{aligned} \tag{4.28}$$

It is gratifying to see from Eq. 4.28 that the updated L-perturbation matrix is a linear function of the initial one, i.e.,

$$\mathbf{\Delta}_1(t) = -(\mathbf{I} - \mathbf{R}_0(t))^{-1} \mathbf{R}_0(t) \mathbf{\Delta}_0(t). \tag{4.29}$$

Equations 4.5, 4.26, and 4.29 define an iteration that successively refines any admissible initial approximation, $\mathbf{\Phi}_0(t)$, by means of its relative error.

THEOREM 4. Let $\mathbf{\Phi}_0(t)$ be an admissible initial approximation for the linear system 1.1 and let $\mathbf{\Delta}_0(t)$ be the corresponding L-perturbation. Then the iteration,

$$\begin{aligned}
\mathbf{R}_k(t) &= \int_0^t \mathbf{\Phi}_k(t,\sigma)\, \mathbf{\Delta}_k(\sigma)\, \mathbf{\Phi}_k^{-1}(t,\sigma)\, d\sigma \\
\mathbf{\Phi}_{k+1}(t) &= (\mathbf{I} - \mathbf{R}_k(t))^{-1}\, \mathbf{\Phi}_k(t) \\
\mathbf{\Delta}_{k+1}(t) &= -(\mathbf{I} - \mathbf{R}_k(t))^{-1} \mathbf{R}_k(t)\, \mathbf{\Delta}_k(t),
\end{aligned} \tag{4.30}$$

generates the left-ordered (index increases from right to left) infinite product representation for the fundamental matrix of solution,

$$\mathbf{\Phi}(t) = \lim_{m\to\infty} \left(\prod_{k=0}^{m} (\mathbf{I} - \mathbf{R}_k(t))^{-1}\right) \mathbf{\Phi}_0(t), \tag{4.31}$$

provided this limit exists.

Proof. For $k = 0$, the first iterative step was described in Equations 4.26 to 4.29. Regarding $\mathbf{\Phi}_1(t)$ and $\mathbf{\Delta}_1(t)$ as another initial approximation and its L-perturbation, respectively, the k-th step in iteration 4.30 becomes transparent. By iterating backwards the second equation in 4.30, all the way to $k = 0$, the left-ordered finite product in Eq. 4.31 is obtained. The limit process indicates that this refinement can, in the general case, continue indefinitely. □

THEOREM 5. A necessary condition for the absolute convergence of the product 4.31 is that there is a time instant, t°, such that

$$\|\mathbf{U}_k(t)\| = \left\|(\mathbf{I} - \mathbf{R}_k(t))^{-1}\mathbf{R}_k(t)\right\| < 1, \qquad t < t^\circ, \qquad k \geq 0. \tag{4.32}$$

In case of convergence, the asymptotic behavior of iterates is given by

$$\mathbf{R}_k(t) = \mathbf{O}\left(t^{2^k(p+1)}\right). \tag{4.33}$$

Proof. That the condition 4.32 is necessary follows from rewriting each factor of the product 4.31 in the standard form

$$(\mathbf{I} - \mathbf{R}_k(t))^{-1} = \mathbf{I} + (\mathbf{I} - \mathbf{R}_k(t))^{-1}\mathbf{R}_k(t) = \mathbf{I} + \mathbf{U}_k(t). \tag{4.34}$$

By the well-known theorem on the absolute convergence of matrix products [14], a necessary, but by no means a sufficient, condition for the absolute convergence of the matrix product is that $\|\mathbf{U}_k(t)\| \to 0$ as $k \to \infty$. To show that condition 4.32 leads to $\|\mathbf{U}_k(t)\| \to 0$ as $k \to \infty$, first observe that condition 4.32 clearly implies continuous invertibility of $\mathbf{I} - \mathbf{R}_k(t)$ in Equations 4.30 and 4.31; hence it guarantees the existence of admissible iterates. For the last equation in 4.30, one must also guarantee that the sequence of norms, $\{\|\mathbf{\Delta}_k\|_{k=0}^{\infty}\}$, is a strictly decreasing one. Taking norms of both sides, and then iterating backwards, shows that

$$\|\mathbf{\Delta}_{k+1}\| \leq \left\|(\mathbf{I} - \mathbf{R}_k)^{-1}\mathbf{R}_k\right\| \|\mathbf{\Delta}_k\| = \|\mathbf{U}_k\| \|\mathbf{\Delta}_k\| \leq \left(\prod_{i=0}^{k} \|\mathbf{U}_i\|\right) \|\mathbf{\Delta}_0\|. \tag{4.35}$$

Thus, a sufficient condition for $\|\mathbf{\Delta}_{k+1}\| \to 0$ is that each term $\|\mathbf{U}_i\| < 1$, for all $i \leq k$. This, in turn, guarantees that both $\|\mathbf{R}_k(t)\|$ and $\|\mathbf{U}_k(t)\| \to 0$ as $k \to \infty$. The existence of the time t° is implicit in the continuity of each iterate $\mathbf{R}_k(t)$ and the initial condition, $\mathbf{R}_k(0) = \mathbf{0}$. Now, each iterate either has a time $t_k^\circ \in D$ beyond which the inequality 4.32 is violated, or it satisfies this inequality for all $t \in D$. Therefore the $\min_k(t_k^\circ)$ provides a lower bound for t° and therefore a strict decay of the sequence $\{\|\mathbf{\Delta}_k\|_{k=0}^{\infty}\}$.

The asymptotic behavior of iterates follows from induction: if $\mathbf{\Delta}_0(t) = \mathbf{O}(t^p)$, then $\mathbf{R}_0(t) = \mathbf{O}\left(t^{p+1}\right)$. This implies, from Eq. 4.30, that $\mathbf{\Delta}_1(t) = \mathbf{O}\left(t^{2p+1}\right)$ and $\mathbf{R}_1(t) = \mathbf{O}\left(t^{2(p+1)}\right)$. Iterating these two relations yields Eq. 4.33. □

Although the convergence rates are effectively monitored by computing norms of iterates $\mathbf{U}_k$, which must be computed in any case for a number of discrete values of t, there is a simpler, but not sharp, sufficient condition, namely

$$\|\mathbf{R}_k(t)\| < \frac{1}{2}. \tag{4.36}$$

This follows from

$$\left\|(\mathbf{I} - \mathbf{R}_k)^{-1}\mathbf{R}_k\right\| \leq \frac{\|\mathbf{R}_k\|}{\mathbf{I} - \|\mathbf{R}_k\|} < 1, \tag{4.37}$$

by solving the last inequality in Eq. 4.37 for $\|\mathbf{R}_k\|$. As this last estimate uses two extra maximizations, the time step estimates for t_k°, from Eq. 4.37, tend to be smaller than the ones based on the inequality in 4.32. The problem of determining a sharper sufficient condition for the absolute convergence of this product representation will be addressed in a subsequent paper.

5. The modified relative error. The modified relative error matrix is defined by

$$\mathbf{R}(t) = \mathbf{\Phi}(t)\,\mathbf{\Phi}_0^{-1}(t) - \mathbf{I}. \tag{5.1}$$

As before, the advantage of this definition is that the correction acts on the left-hand side of an initial approximation,

$$\mathbf{\Phi}(t) = (\mathbf{I} + \mathbf{R}(t))\,\mathbf{\Phi}_0(t). \tag{5.2}$$

The analysis of this case is very similar to the case in Sec. 4, hence only the major differences will be highlighted. The differential equation for $\mathbf{R}(t)$ is obtained by differentiating Eq. 5.2 and using Equations 1.1 and 1.3,

$$\mathbf{R}' = \mathbf{\Delta}_0 + [\mathbf{A}_0, \mathbf{R}] + \mathbf{\Delta}_0\mathbf{R}, \qquad \mathbf{R}(0) = \mathbf{0}. \tag{5.3}$$

The integrable part of Eq. 5.3 is therefore the same as in Eq. 4.4 and is explicitly given by Eq. 4.5. The linear integral equation for $\mathbf{R}(t)$ in Eq. 5.1,

$$\mathbf{R}(t) = \mathbf{R}_0(t) + \int_0^t \mathbf{\Phi}_0(t,\sigma)\;\mathbf{\Delta}_0(\sigma)\;\mathbf{R}(\sigma)\,\mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma, \tag{5.4}$$

is slightly different from Eq. 4.6 since, apart from the sign, there is also a different kernel splitting. The fixed point iteration behaves in the same way, except that the integral operator Q is now defined by means of

$$\mathsf{Q}[\mathbf{Z}] = \int_0^t \mathbf{\Phi}_0(t,\sigma)\;\mathbf{\Delta}_0(\sigma)\,\mathbf{Z}(\sigma)\;\mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma, \tag{5.5}$$

and the consecutive terms in the series are linked by recurrence

$$\mathbf{R}_{k+1}(t) = \int_0^t \mathbf{\Phi}_0(t,\sigma)\;\mathbf{\Delta}_0(\sigma)\,\mathbf{R}_k(\sigma)\;\mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma, \qquad k \geq 0, \tag{5.6}$$

thus maintaining the same exponential gap as in Eq. 4.9.

The bounds on the growth of $\|\mathbf{R}(t)\|$ and $\|\mathbf{R}(t) - \mathbf{R}_0(t)\|$ in Equations 4.13, 4.15, and 4.23 remain the same. In the case of the Chu-Metcalf inequality, the definition of the kernel function in Eq. 4.17 must be changed to accommodate a different kernel splitting,

$$K_0(t,\sigma) = \|\mathbf{\Phi}_0(t,\sigma)\,\mathbf{\Delta}_0(\sigma)\|\,\left\|\mathbf{\Phi}_0^{-1}(t,\sigma)\right\|, \qquad 0 \leq \sigma \leq t, \tag{5.7}$$

before solving the scalar integral equations 4.18 and 4.25 and applying inequalities 4.19 and 4.24, respectively.

For the product representation, the approximate relative error at the k-th step is the same as in the first Eq. 4.30; however, the solution and its L-perturbation are now updated according to

$$\begin{aligned} \mathbf{R}_k(t) &= \int_0^t \mathbf{\Phi}_k(t,\sigma)\;\mathbf{\Delta}_k(\sigma)\;\mathbf{\Phi}_k^{-1}(t,\sigma)\,d\sigma \\ \mathbf{\Phi}_{k+1}(t) &= (\mathbf{I} + \mathbf{R}_k(t))\;\mathbf{\Phi}_k(t) \\ \mathbf{\Delta}_{k+1}(t) &= \mathbf{\Delta}_k(t)\,\mathbf{R}_k(t)\,(\mathbf{I} + \mathbf{R}_k(t))^{-1}. \end{aligned} \tag{5.8}$$

It follows from the second Eq. 5.8 that the fundamental matrix of solutions admits another left-ordered infinite product representation,

$$\mathbf{\Phi}(t) = \lim_{m\to\infty} \left(\prod_{k=0}^{m} (\mathbf{I} + \mathbf{R}_k(t))\right) \mathbf{\Phi}_0(t). \tag{5.9}$$

From the analysis of the third Eq. 5.8, it follows that the invertibility of $\mathbf{I}+\mathbf{R}_k(t)$ and the decay of $\|\mathbf{\Delta}_{k+1}\| \to 0$, for all $k \geq 0$, can be guaranteed by finding a bound on t° such that

$$\left\|\mathbf{R}_k(t)\left(\mathbf{I}+\mathbf{R}_k(t)\right)^{-1}\right\| < 1, \qquad t < t^\circ, \qquad k \geq 0. \tag{5.10}$$

Although this is not the same necessary condition as in Theorem 5, it is qualitatively very close to it because the norm in Eq. 5.10 may be bounded by

$$\left\|\mathbf{R}_k(t)\left(\mathbf{I}+\mathbf{R}_k(t)\right)^{-1}\right\| \leq \frac{\|\mathbf{R}_k(t)\|}{\mathbf{I}-\|\mathbf{R}_k(t)\|} < 1, \tag{5.11}$$

which implies the same sufficient condition, namely, $\|\mathbf{R}_k(t)\| < 1/2$. Therefore, the iteration 5.8 offers no significant convergence improvement over the iteration 4.30.

6. The A-relative error. Let $\mathbf{\Phi}_0(t)$ be an admissible initial approximation and let $\mathbf{A}_0(t)$ and $\mathbf{\Delta}_0(t)$ be its L-derivative and L-perturbation matrix, respectively. The A-relative error (A- stands for the arithmetic average) is defined by

$$\mathbf{R}(t) = 2\left(\mathbf{\Phi}(t)-\mathbf{\Phi}_0(t)\right)\left(\mathbf{\Phi}(t)+\mathbf{\Phi}_0(t)\right)^{-1}, \tag{6.1}$$

provided the matrix $\mathbf{\Phi}(t)+\mathbf{\Phi}_0(t)$ is invertible. Solving for $\mathbf{\Phi}(t)$ yields

$$\mathbf{\Phi}(t) = \left(\frac{\mathbf{2}+\mathbf{R}(t)}{\mathbf{2}-\mathbf{R}(t)}\right)\mathbf{\Phi}_0(t) = \mathbf{C}\left(\mathbf{R}(t)\right)\mathbf{\Phi}_0(t), \tag{6.2}$$

which clarifies the admissibility conditions in Eq. 6.2: both matrices $\mathbf{2}\pm\mathbf{R}(t)$ must be nonsingular for all t in some neighborhood of the origin.

What makes the rational matrix $\mathbf{C}(\mathbf{R})$ notable is that it is the Cayley transform of $\mathbf{R}/2$. This means (shown in Sec. 6.4) that if $\mathbf{A}(t), \mathbf{A}_0(t)$ are skewhermitian, or $\mathbf{J}$-symmetric, then the unitary, or symplectic, property of any truncated product is preserved. Neither the classical nor the modified relative error is able to preserve these properties.

To obtain the differential equation for $\mathbf{R}(t)$, we rewrite Eq. 6.1 as $\mathbf{R}\left(\mathbf{\Phi}+\mathbf{\Phi}_0\right) = 2\left(\mathbf{\Phi}-\mathbf{\Phi}_0\right)$, then differentiate it and substitute the expression $\mathbf{\Phi}+\mathbf{\Phi}_0 = 4\left(\mathbf{2}-\mathbf{R}\right)^{-1}\mathbf{\Phi}_0$ (which follows from Eq. 6.2). Rearrangement of terms results in

$$\mathbf{R}' = \mathbf{\Delta}_0 + \frac{1}{2}\left[\mathbf{A}+\mathbf{A}_0, \mathbf{R}\right] - \frac{1}{4}\mathbf{R}\mathbf{\Delta}_0\mathbf{R}, \qquad \mathbf{R}(0) = \mathbf{0}. \tag{6.3}$$

This matrix Riccati equation is now rewritten, using matrices $\mathbf{A}_0(t)$ and $\mathbf{\Delta}_0(t)$, as

$$\mathbf{R}' = \mathbf{\Delta}_0 + \left[\mathbf{A}_0, \mathbf{R}\right] + \frac{1}{2}\left[\mathbf{\Delta}_0, \mathbf{R}\right] - \frac{1}{4}\mathbf{R}\mathbf{\Delta}_0\mathbf{R}, \qquad \mathbf{R}(0) = \mathbf{0}. \tag{6.4}$$

The terms in Eq. 6.4 are asymptotically ordered: if $\mathbf{\Delta}_0 = \mathbf{O}\left(t^p\right)$, then $\mathbf{R} = \mathbf{O}\left(t^{p+1}\right)$, so that $\left[\mathbf{A}_0, \mathbf{R}\right] = \mathbf{O}\left(t^{p+1}\right)$ and $\left[\mathbf{\Delta}_0, \mathbf{R}\right] = \mathbf{O}\left(t^{2p+2}\right)$ (see Eq. 2.22), while the nonlinear term behaves as $\mathbf{R}\mathbf{\Delta}_0\mathbf{R} = \mathbf{O}\left(t^{3p+2}\right)$.

The integrable part Eq. 6.4, which satisfies differential equation,

$$\mathbf{R}_0' = \mathbf{\Delta}_0 + \left[\mathbf{A}_0, \mathbf{R}_0\right], \qquad \mathbf{R}_0(0) = \mathbf{0}, \tag{6.5}$$

is the same as in Equations 4.4. Hence, using the matrix $\mathbf{R}_0(t)$, computed from Eq. 4.5, Eq. 6.4 can be written as a nonlinear Volterra integral equation,

$$\begin{aligned}\mathbf{R}(t) &= \mathbf{R}_0(t) + \frac{1}{2}\int_0^t \mathbf{\Phi}_0(t,\sigma)\,[\mathbf{\Delta}_0(\sigma), \mathbf{R}(\sigma)]\,\mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma \\ &\quad -\frac{1}{4}\int_0^t \mathbf{\Phi}_0(t,\sigma)\,\mathbf{R}(\sigma)\,\mathbf{\Delta}_0(\sigma)\,\mathbf{R}(\sigma)\,\mathbf{\Phi}_0^{-1}(t,\sigma)\,d\sigma. \end{aligned} \tag{6.6}$$

It is evident from Equations 4.5 and assumption 1.4 that $\mathbf{R}_0 = \mathbf{O}\left(t^{p+1}\right)$ is indeed the dominant part of solution, $\mathbf{R}(t)$, since the second, linear, term in Eq. 6.6 is of order $\mathbf{O}\left(t^{2p+3}\right)$, while the third, nonlinear, term is of order $\mathbf{O}\left(t^{3p+3}\right)$.

6.1. *Fixed point iteration for the A-relative error.* Unlike the iterations in Equations 4.8 and 5.6, which converge globally, fixed point iteration of Eq. 6.6 has a finite radius of convergence. This occurs not only because the quadratic nonlinearity in the Riccati equation generates solutions that may have a finite escape time (singularity) [24], but also because of the need to maintain invertibilities in Eq. 6.2, or existence of admissible iterates. The key observation is that, by continuity, any finite escape time happens after the time when either of the matrices $\mathbf{2} \pm \mathbf{R}(t)$ becomes singular for the first time. Thus, it is possible to set up either a Picard or Caratheodory iteration [17] to obtain a local solution up to the time defined implicitly by $\|\mathbf{R}(t)\| < 2$. The only new result that emerges from such an iteration, starting with $\mathbf{R}_0(t)$, is that it will generate a lacunary type series, but the exponential gaps will no longer be uniform because of the nonlinear term. Nevertheless, the accuracy of consecutive iterates is higher than in two previous linear cases since, at the start, $\mathbf{R}(t) - \mathbf{R}_0(t) = \mathbf{O}\left(t^{2p+3}\right)$.

It is well known [24] that a matrix Riccati equation can be converted into a linear system of differential equations. Unfortunately, this approach doubles the size of the original problem with some simplification due to linearity, but without a compensating speed up in convergence. It will therefore not be pursued.

6.2. *Bounds on* $\|\mathbf{R}(t)\|$. By taking a norm of Eq. 6.6, one gets the bound,

$$\begin{aligned}\|\mathbf{R}(t)\| &\le \|\mathbf{R}_0(t)\| + \int_0^t \|\mathbf{\Phi}_0(t,\sigma)\|\,\|\mathbf{\Delta}_0(\sigma)\|\,\|\mathbf{R}(\sigma)\|\,\left\|\mathbf{\Phi}_0^{-1}(t,\sigma)\right\|\,d\sigma \\ &\quad +\frac{1}{4}\int_0^t \|\mathbf{\Phi}_0(t,\sigma)\|\,\|\mathbf{\Delta}_0(\sigma)\|\,\|\mathbf{R}(\sigma)\|^2\,\left\|\mathbf{\Phi}_0^{-1}(t,\sigma)\right\|\,d\sigma. \end{aligned} \tag{6.7}$$

Denote $r(t) = \|\mathbf{R}(t)\|$, $r_0(t) = \|\mathbf{R}_0(t)\|$, $f(r) = r + r^2/4$, and

$$K_0(t,\sigma) = \|\mathbf{\Delta}_0(\sigma)\|\,\|\mathbf{\Phi}_0(t,\sigma)\|\,\left\|\mathbf{\Phi}_0^{-1}(t,\sigma)\right\|, \qquad 0 \le \sigma \le t. \tag{6.8}$$

Then inequality 6.7 can be rewritten as

$$r(t) \le r_0(t) + \int_0^t K_0(t,\sigma)\,f(r(\sigma))\,d\sigma. \tag{6.9}$$

A simple upper bound on $r(t)$ comes from the invertibility condition, $\|\mathbf{R}(t)\| = r(t) < 2$, in Eq. 6.2. This means that $f(r) = r(1 + r/4) < 3r/2$, for $r \in (0,2)$, so that the nonlinear inequality 6.9 can be majorized by a linear one:

$$r(t) \le r_0(t) + \int_0^t K_0(t,\sigma)\,f(r(\sigma))\,d\sigma < r_0(t) + \frac{3}{2}\int_0^t K_0(t,\sigma)\,r(\sigma)\,d\sigma. \tag{6.10}$$

Invoking the Chu-Metcalf inequality, which depends on the solution of the linear integral equation,

$$\rho(t) = r_0(t) + \frac{3}{2}\int_0^t K_0(t,\sigma)\ \rho(\sigma)\, d\sigma, \tag{6.11}$$

it follows that $r(t) \leq \rho(t)$, as long as $\rho(t) < 2$.

It is possible to obtain much more precise information on the bounding function $\rho(t)$ by expanding on the idea of positive nondecreasing functions outlined in the previous paragraph. The following theorem, which is stated under more general conditions than those needed for inequality 6.9, defines such a function $\rho(t)$ and provides a means for computing its lower and upper approximations.

THEOREM 6. Let $r_0(t) \geq 0$ and $K_0(t,\sigma) \geq 0$, for $0 \leq \sigma \leq t$, be continuous functions for $t \in D$. If $r(t)$ satisfies Eq. 6.9 and $f(r) \geq 0$ is a continuous and nondecreasing function for $r \geq 0$, then

$$r(t) \leq \rho(t) \tag{6.12}$$

for all t for which there is the solution of nonlinear integral equation

$$\rho(t) = r_0(t) + \int_0^t K_0(t,\sigma)\ f(\rho(\sigma))\, d\sigma. \tag{6.13}$$

Proof. Let $0 \leq v_0(t) \leq r_0(t)$ be any continuous function on $t \in D$. The sequence of functions $\{\nu_j(t)\}_{j\geq 0}$, generated by recursion

$$\nu_{j+1}(t) = r_0(t) + \int_0^t K_0(t,\sigma)\ f(\nu_j(\sigma))\, d\sigma, \qquad j \geq 0, \tag{6.14}$$

is well-defined, continuous and is certainly nonnegative for all $t \in D$ and all finite $j \geq 0$. It may be divergent as $j \to \infty$, but, by the basic existence theorem [17], there is a point t^* in a neighborhood of the origin, such that Picard iteration in Eq. 6.14 converges to the solution $\rho(t)$ for all $t < t^*$. In this case, the sequence is also nondecreasing. To show this, consider first the difference $\nu_1(t) - \nu_0(t)$. By the assumptions on the functions f and ν_0, it follows that $f(\nu_0(\sigma)) \geq 0$, which implies nonnegativity of the integral in Eq. 6.14, and therefore $\nu_1(t) - \nu_0(t) \geq 0$. Then, assuming that $\nu_j(t) - \nu_{j-1}(t) \geq 0$, we consider the difference

$$\nu_{j+1}(t) - \nu_j(t) = \int_0^t K_0(t,\sigma)\ (f(\nu_j(\sigma)) - f(\nu_{j-1}(\sigma)))\, d\sigma. \tag{6.15}$$

The quantity $f(\nu_j(\sigma)) - f(\nu_{j-1}(\sigma)) \geq 0$, by the assumption that f is a nondecreasing function of its argument. Since $K_0(t,\sigma) \geq 0$, it follows that the integral in Eq. 6.15 is also nonnegative and the induction step can be completed, provided one restricts $t < t^*$.

The same iterative process 6.14 can be set up, starting with a sufficiently large initial function, $u_0(t)$, to generate the sequence of functions $\{u_j(t)\}_{j\geq 0}$, also convergent to $\rho(t)$ for $t < t^*$. One then uses the same idea to show that this sequence is nonincreasing.

Although the divergence of the Picard sequence may happen before the first singularity in the solution [17], there is really no problem in continuing the solution (as long as it exists), or the sequence of its lower and upper bounds, beyond the point of Picard divergence. □

6.3. *Product formula for the A-relative error.* As soon as the matrix $\mathbf{R}_0(t)$ is computed from Eq. 4.5, it may be used in Eq. 6.2 to correct the initial approximation,

$$\mathbf{\Phi}_1(t) = \left(\frac{\mathbf{2}+\mathbf{R}_0(t)}{\mathbf{2}-\mathbf{R}_0(t)}\right)\mathbf{\Phi}_0(t). \tag{6.16}$$

Assuming that both of the matrices $\mathbf{2} \pm \mathbf{R}(t)$ are invertible, the determination of $\mathbf{A}_1 = \mathsf{L}[\mathbf{\Phi}_1]$ begins by first computing,

$$\frac{d}{dt}\left(\frac{\mathbf{2}+\mathbf{R}_0}{\mathbf{2}-\mathbf{R}_0}\right) = (\mathbf{2}-\mathbf{R}_0)^{-1}\mathbf{R}_0'(\mathbf{2}-\mathbf{R}_0)^{-1}(\mathbf{2}+\mathbf{R}_0) + (\mathbf{2}-\mathbf{R}_0)^{-1}\mathbf{R}_0'$$

$$= (\mathbf{2}-\mathbf{R}_0)^{-1}\mathbf{R}_0'\left((\mathbf{2}-\mathbf{R}_0)^{-1}(\mathbf{2}+\mathbf{R}_0)+\mathbf{I}\right) = 4(\mathbf{2}-\mathbf{R}_0)^{-1}\mathbf{R}_0'(\mathbf{2}-\mathbf{R}_0)^{-1}. \tag{6.17}$$

Upon taking the L-derivative of Eq. 6.16 and substituting Eq. 6.17, cancelling and factoring the obvious terms, the following expression emerges:

$$\mathbf{A}_1 = (\mathbf{2}-\mathbf{R}_0)^{-1}(4\mathbf{R}_0' + (\mathbf{2}+\mathbf{R}_0)\mathbf{A}_0(\mathbf{2}-\mathbf{R}_0))(\mathbf{2}+\mathbf{R}_0)^{-1}$$

$$= (\mathbf{2}-\mathbf{R}_0)^{-1}(4\mathbf{\Delta}_0 + 4[\mathbf{A}_0,\mathbf{R}_0] + (\mathbf{2}+\mathbf{R}_0)\mathbf{A}_0(\mathbf{2}-\mathbf{R}_0))(\mathbf{2}+\mathbf{R}_0)^{-1}. \tag{6.18}$$

After several more cancellations, extraction of desired terms and regrouping, similar to what was done in Eq. 4.28, the matrix $\mathbf{A}_1$ can be written as

$$\mathbf{A}_1 = \mathbf{A} + (\mathbf{2}-\mathbf{R}_0)^{-1}(2[\mathbf{R}_0,\mathbf{\Delta}_0] + \mathbf{R}_0\mathbf{\Delta}_0\mathbf{R}_0)(\mathbf{2}+\mathbf{R}_0)^{-1}. \tag{6.19}$$

This implies that the new L-perturbation matrix, $\mathbf{\Delta}_1(t)$, is a generalized linear matrix function (i.e., it has both left and right matrix coefficients) of the previous one, $\mathbf{\Delta}_0(t)$, namely,

$$\mathbf{\Delta}_1 = -(\mathbf{2}-\mathbf{R}_0)^{-1}(2[\mathbf{R}_0,\mathbf{\Delta}_0] + \mathbf{R}_0\mathbf{\Delta}_0\mathbf{R}_0)(\mathbf{2}+\mathbf{R}_0)^{-1}. \tag{6.20}$$

Equations 4.5, 6.16, and 6.20 form the iterative process for refining an initial approximation $\mathbf{\Phi}_0(t)$ by means of its A-relative error.

THEOREM 7. Let $\mathbf{\Phi}_0(t)$ be an admissible initial approximation for the linear system 1.1 and let $\mathbf{\Delta}_0(t)$ be the corresponding L-perturbation. Then the iteration,

$$\begin{aligned}
\mathbf{R}_k(t) &= \int_0^t \mathbf{\Phi}_k(t,\sigma)\,\mathbf{\Delta}_k(\sigma)\,\mathbf{\Phi}_k^{-1}(t,\sigma)\,d\sigma \\
\mathbf{\Phi}_{k+1}(t) &= \frac{\mathbf{2}+\mathbf{R}_k(t)}{\mathbf{2}-\mathbf{R}_k(t)}\mathbf{\Phi}_k(t) \\
\mathbf{\Delta}_{k+1}(t) &= -(\mathbf{2}-\mathbf{R}_k)^{-1}(2[\mathbf{R}_k,\mathbf{\Delta}_k] + \mathbf{R}_k\mathbf{\Delta}_k\mathbf{R}_k)(\mathbf{2}+\mathbf{R}_k)^{-1},
\end{aligned} \tag{6.21}$$

generates the left-ordered infinite product representation for the fundamental matrix of solution,

$$\mathbf{\Phi}(t) = \lim_{m\to\infty}\left(\prod_{k=0}^{m}\frac{\mathbf{2}+\mathbf{R}_k(t)}{\mathbf{2}-\mathbf{R}_k(t)}\right)\mathbf{\Phi}_0(t), \tag{6.22}$$

provided this limit exists.

Proof. The proof uses the same arguments as in Theorem 4. □

THEOREM 8. A necessary condition for the absolute convergence of the product 6.22 is that there is a time instant, t°, such that

$$\left\|(\mathbf{2}-\mathbf{R}_k)^{-1}\right\|\left\|(\mathbf{2}+\mathbf{R}_k)^{-1}\right\|\left(4\left\|\mathbf{R}_k\right\|+\left\|\mathbf{R}_k\right\|^2\right)\leq 1, \qquad t<t^\circ, \qquad k\geq 0. \tag{6.23}$$

In case of convergence, the asymptotic behavior of iterates is given by

$$\mathbf{R}_k(t)=\mathbf{O}\left(t^{2^k(p+2)-1}\right). \tag{6.24}$$

Proof. If each factor of the product 6.22 is rewritten in the standard form,

$$\frac{\mathbf{2}+\mathbf{R}_k}{\mathbf{2}-\mathbf{R}_k}=\mathbf{I}+\frac{2\mathbf{R}_k}{\mathbf{2}-\mathbf{R}_k}, \tag{6.25}$$

it follows that the condition $\mathbf{R}_k\to 0$ as $k\to\infty$ is the same necessary condition as in Theorem 5. The condition 6.23 clearly implies continuous invertibility of both matrices $\mathbf{2}\pm\mathbf{R}_k(t)$ in Equations 6.21 and 6.22, hence it guarantees the existence of admissible iterates. For the last equation in 6.21, one must also ensure that the sequence of norms, $\{\|\mathbf{\Delta}_k\|\}_{k=0}^{\infty}$, is a strictly decreasing one. Taking the norms of both sides, while keeping the three distinct factors on the right-hand side, and then iterating backwards, one arrives at

$$\begin{aligned}\|\mathbf{\Delta}_{k+1}\| &\leq \left\|(\mathbf{2}-\mathbf{R}_k)^{-1}\right\|\left\|(\mathbf{2}+\mathbf{R}_k)^{-1}\right\|\left(4\left\|\mathbf{R}_k\right\|+\left\|\mathbf{R}_k\right\|^2\right)\|\mathbf{\Delta}_k\| \\ &\leq \left(\prod_{i=0}^{k}\left\|(\mathbf{2}-\mathbf{R}_i)^{-1}\right\|\left\|(\mathbf{2}+\mathbf{R}_i)^{-1}\right\|\left(4\left\|\mathbf{R}_i\right\|+\left\|\mathbf{R}_i\right\|^2\right)\right)\|\mathbf{\Delta}_0\|.\end{aligned} \tag{6.26}$$

Thus, a sufficient condition for $\|\mathbf{\Delta}_{k+1}\|\to 0$ is that each term in this product satisfies $\left\|(\mathbf{2}-\mathbf{R}_i)^{-1}\right\|\left\|(\mathbf{2}+\mathbf{R}_i)^{-1}\right\|\left(4\left\|\mathbf{R}_i\right\|+\left\|\mathbf{R}_i\right\|^2\right)<1$. This, in turn, will guarantee that $\mathbf{R}_k\to\mathbf{0}$. The existence of the time t° is implicit in the continuity of each iterate $\mathbf{R}_k(t)$ and the initial condition, $\mathbf{R}_k(0)=\mathbf{0}$. Now, each iterate either has a time t_k° beyond which the inequality 6.23 is violated, or it satisfies this inequality for all $t\in D$. Therefore, the $\min_k(t_k^\circ)$ provides a lower bound for t° and hence a strict decay of sequence $\{\|\mathbf{\Delta}_k\|_{k=0}^{\infty}\}$.

The asymptotic behavior of iterates follows from induction: if $\mathbf{\Delta}_0(t)=\mathbf{O}(t^p)$, then $\mathbf{R}_0(t)=\mathbf{O}\left(t^{p+1}\right)$. This implies, from the last Eq. 6.21, that $\mathbf{\Delta}_1(t)=\mathbf{O}\left(t^{2p+2}\right)$ and therefore $\mathbf{R}_1(t)=\mathbf{O}\left(t^{2p+3}\right)$. This comes from adding an additional power due to the commutator (see Eq. 2.22). Continuing to iterate on the asymptotic orders, always adding one for the commutator, leads to Eq. 6.24. □

Although the convergence rates are effectively monitored by measuring norms of iterates $\mathbf{\Delta}_{k+1}$, which must be computed in any case for a number of discrete values of t, there is a simpler, but not sharp, sufficient condition, namely

$$\|\mathbf{R}_k(t)\|<1/2. \tag{6.27}$$

To see this, denote by $r_k=\|\mathbf{R}_k\|$, $\delta_k=\|\mathbf{\Delta}_k\|$ and $\delta_{k+1}=\|\mathbf{\Delta}_{k+1}\|$, and then take the norm of the last Eq. 6.21 to generate the following inequality:

$$\delta_{k+1}\leq\frac{r_k+r_k^2/4}{(1-r_k/2)^2}\delta_k. \tag{6.28}$$

To guarantee a strictly decreasing sequence in Eq. 6.28, it is sufficient to satisfy $r_k+r_k^2/4<(1-r_k/2)^2$, which leads to the condition $r_k<1/2$. As in the case of the classical

relative error, the problem of sharper sufficient conditions for the absolute convergence of this product will be addressed in a subsequent paper.

6.4. *Unitary and symplectic property.* Consider first the unitary property. If $\mathbf{A}(t)$ is skewhermitian and one starts with any unitary admissible approximation $\mathbf{\Phi}_0(t)$, then $\mathbf{A}_0(t)$ is necessarily skewhermitian [11], and so is $\mathbf{\Delta}_0(t)$. It is an easy check that $\mathbf{R}_0(t)$, given by Eq. 4.5, is also skewhermitian, and hence the first Cayley update, $\mathbf{C}(\mathbf{R}_0(t))$, and $\mathbf{\Phi}_1(t)$ are both unitary. To show that $\mathbf{\Delta}_1(t)$ is skewhermitian, take the conjugate transpose of both sides in the last Eq. 6.21: the left and the right factors will not change under this operation (inversion and conjugate transpose commute) and both terms in the central factor will change sign. Therefore, the iteration 6.21 will preserve the skewhermitian property of $\mathbf{R}_k(t)$ and $\mathbf{\Delta}_k(t)$ as well as the unitary property of $\mathbf{\Phi}_k(t)$ for all $k > 0$.

As for the symplectic property, which pertains to the real systems of even dimension, the following definitions are needed: a matrix $\mathbf{J}$, defined by

$$\mathbf{J} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \\ -\mathbf{I} & \mathbf{0} \end{bmatrix}, \tag{6.29}$$

is called an elementary symplectic matrix of dimension $2d$, provided $\mathbf{I}$ and $\mathbf{0}$ are both d-dimensional unit and zero matrix, respectively. The following properties of $\mathbf{J}$, (e.g., [15]) will be used below:

$$\mathbf{J}^{-1} = \mathbf{J}^{\mathsf{T}} = -\mathbf{J}, \qquad \mathbf{J}^2 = -\mathbf{I}. \tag{6.30}$$

A matrix $\mathbf{S}$ is called real symplectic (also $\mathbf{J}$-orthogonal in [12]) iff it satisfies the following property,

$$\mathbf{S}^{\mathsf{T}}\mathbf{J}\mathbf{S} = \mathbf{J}. \tag{6.31}$$

A matrix $\mathbf{A}$ that can be written as $\mathbf{JH}$, where $\mathbf{H}$ is symmetric, will be called $\mathbf{J}$-symmetric. To proceed, the following elementary theorem, for which no reference was found, is proved:

THEOREM 9. If $\mathbf{A}$ is $\mathbf{J}$-symmetric, then its Cayley transform,

$$\mathbf{C} = \mathbf{C}(\mathbf{A}) = (\mathbf{I} - \mathbf{A})^{-1}(\mathbf{I} + \mathbf{A}), \tag{6.32}$$

is a symplectic matrix.

Proof. Since $\mathbf{A} = \mathbf{JH}$, for some symmetric matrix $\mathbf{H}$, one must show that $\mathbf{C}(\mathbf{JH})$ satisfies Eq. 6.31. To do this, evaluate $\mathbf{JC}(\mathbf{JH})$ in stages, first by moving $\mathbf{J}$ through the denominator

$$\mathbf{J}(\mathbf{I} - \mathbf{JH})^{-1} = \left(\mathbf{J}^{-1} - \mathbf{HJ}^{-1}\right)^{-1} = \left(\mathbf{J}^{-1}\left(\mathbf{I} + \mathbf{J}^2\mathbf{HJ}\right)\right)^{-1} = (\mathbf{I} - \mathbf{HJ})^{-1}\mathbf{J}, \tag{6.33}$$

and then through the numerator

$$\mathbf{J}(\mathbf{I} + \mathbf{JH}) = \left(\mathbf{J} + \mathbf{J}^2\mathbf{H}\right) = \left(\mathbf{I} - \mathbf{HJ}^{-1}\right)\mathbf{J} = (\mathbf{I} + \mathbf{HJ})\mathbf{J}. \tag{6.34}$$

Equations 6.34 and 6.33 imply that

$$\mathbf{JC}(\mathbf{JH}) = \mathbf{C}(\mathbf{HJ})\mathbf{J}. \tag{6.35}$$

On the other hand, taking the transpose of $\mathbf{C}(\mathbf{JH})$ shows that

$$\begin{aligned}\mathbf{C}^{\mathsf{T}}(\mathbf{JH}) &= \left((\mathbf{I}-\mathbf{JH})^{-1}(\mathbf{I}+\mathbf{JH})\right)^{\mathsf{T}} = \left(\mathbf{I}+\mathbf{HJ}^{\mathsf{T}}\right)\left(\mathbf{I}-\mathbf{HJ}^{\mathsf{T}}\right)^{-1} \\ &= (\mathbf{I}-\mathbf{HJ})(\mathbf{I}+\mathbf{HJ})^{-1} = \mathbf{C}^{-1}(\mathbf{HJ}). \end{aligned} \tag{6.36}$$

Multiplying the initial and final matrix in Eq. 6.36 with the matrices in Eq. 6.35 shows that Eq. 6.31 is satisfied. □

Suppose now that $\mathbf{A}(t) = \mathbf{JH}(t)$ so that its fundamental matrix of solutions, $\mathbf{\Phi}(t)$, is symplectic [12], [15]. The following theorem shows how these properties are preserved under the iteration process 6.21.

THEOREM 10. If an initial approximation $\mathbf{\Phi}_0(t)$ is symplectic, then during the iterative process 6.21 the matrices $\mathbf{\Phi}_k(t)$ are symplectic and the matrices $\mathbf{R}_k(t)$ and $\mathbf{\Delta}_k(t)$ are $\mathbf{J}$-symmetric for all $k > 0$.

Proof. Since $\mathbf{\Phi}_0(t)$ is symplectic, it follows that its L-derivative is $\mathbf{J}$-symmetric [15], hence it can be written as $\mathbf{A}_0(t) = \mathbf{JH}_0(t)$. Initial L-perturbation is given by $\mathbf{\Delta}_0(t) = \mathbf{JH}(t) - \mathbf{JH}_0(t) = \mathbf{JK}_0(t)$, where $\mathbf{K}_0(t)$ is symmetric. Assuming, therefore, that $\mathbf{\Delta}_k(t) = \mathbf{JK}_k(t)$ and that $\mathbf{\Phi}_k(t)$ is symplectic, the first iterative step in 6.21 generates the following matrix:

$$\begin{aligned}\mathbf{R}_k(t) &= \int_0^t \mathbf{\Phi}_k(t,\sigma)\,\mathbf{JK}_k(\sigma)\,\mathbf{\Phi}_k^{-1}(t,\sigma)\,d\sigma \\ &= \int_0^t \mathbf{\Phi}_k(t,\sigma)\,\mathbf{J}\,\mathbf{\Phi}_k^{\mathsf{T}}(t,\sigma)\left(\mathbf{\Phi}_k^{\mathsf{T}}(t,\sigma)\right)^{-1}\mathbf{K}_k(\sigma)\,\mathbf{\Phi}_k^{-1}(t,\sigma)\,d\sigma \\ &= \int_0^t \mathbf{J}\left(\mathbf{\Phi}_k^{\mathsf{T}}(t,\sigma)\right)^{-1}\mathbf{K}_k(\sigma)\,\mathbf{\Phi}_k^{-1}(t,\sigma)\,d\sigma \\ &= \mathbf{J}\int_0^t \left(\mathbf{\Phi}_k^{\mathsf{T}}(t,\sigma)\right)^{-1}\mathbf{K}_k(\sigma)\,\mathbf{\Phi}_k^{-1}(t,\sigma)\,d\sigma. \end{aligned} \tag{6.37}$$

Two properties have been used here: first, that $\mathbf{J}$ is time independent, and second, that $\mathbf{\Phi}_k^{\mathsf{T}}(t,\sigma)$ is invertible and symplectic whenever $\mathbf{\Phi}_k(t,\sigma)$ is. Since the matrix under the last integral is symmetric, as can be seen upon taking its transpose, the integral of a symmetric matrix preserves this property and one can write $\mathbf{R}_k(t) = \mathbf{JX}_k(t)$, for some symmetric $\mathbf{X}_k(t)$.

In the second step of the iteration 6.21, one has the product of two matrices, the first one being the Cayley transform of $\mathbf{J}$-symmetric $\mathbf{R}_k(t)$, hence symplectic by Theorem 9, and the second one being symplectic by the assumption. It is known (e.g., [15]) that the symplectic matrices form a multiplicative group under the ordinary matrix product. This implies that $\mathbf{\Phi}_{k+1}(t)$ is also symplectic.

In the last step of the iteration, consider first the commutator,

$$[\mathbf{JK}_k(t), \mathbf{JX}_k(t)] = \mathbf{J}\left(\mathbf{K}_k(t)\,\mathbf{JX}_k(t) - \mathbf{X}_k(t)\,\mathbf{JK}_k(t)\right), \tag{6.38}$$

and since the matrix in parenthesis is symmetric (transpose, then use the property that $\mathbf{J}^{\mathsf{T}} = -\mathbf{J}$), it is $\mathbf{J}$-symmetric. Next, writing the nonlinear term as

$$\mathbf{R}_k\mathbf{\Delta}_k\mathbf{R}_k = \mathbf{JX}_k\mathbf{JK}_k\mathbf{JX}_k = \mathbf{J}\left(\mathbf{X}_k\mathbf{JK}_k\mathbf{JX}_k\right) \tag{6.39}$$

shows that it is also **J**-symmetric since the last matrix in parenthesis is symmetric. Ignoring the minus sign, one may write the matrix in the last step of iteration as

$$(\mathbf{2} - \mathbf{J}\mathbf{X}_k)^{-1} (\mathbf{J}\mathbf{Y}_k) (\mathbf{2} + \mathbf{J}\mathbf{X}_k)^{-1} \tag{6.40}$$

for some symmetric matrix $\mathbf{Y}_k$. It is now a simple matter to show that this matrix is **J**-symmetric, for one can first move **J** to the left, as in Eq. 6.33, to get

$$\mathbf{J} (\mathbf{2} - \mathbf{X}_k\mathbf{J})^{-1} \mathbf{Y}_k (\mathbf{2} + \mathbf{J}\mathbf{X}_k)^{-1} \tag{6.41}$$

and then show, by taking transpose and using $\mathbf{J}^{\mathsf{T}} = -\mathbf{J}$, that the matrix to the right of **J** is symmetric. It follows that $\mathbf{\Delta}_{k+1}(t)$ is **J**-symmetric whenever $\mathbf{\Delta}_k(t)$ and $\mathbf{R}_k(t)$ are, which completes the induction step. □

7. Computation of $\mathbf{R}_k(t)$. The main step in all three iterations is the same, namely, the computation of

$$\mathbf{R}_k(t) = \int_0^t \mathbf{\Phi}_k(t,\sigma)\, \mathbf{\Delta}_k(\sigma)\, \mathbf{\Phi}_k^{-1}(t,\sigma)\, d\sigma. \tag{7.1}$$

The matrix functions $\mathbf{\Phi}_k(t)$ and $\mathbf{\Delta}_k(t)$ are assumed to be known over an interval of interest. This quadrature problem can be simplified if it is rewritten as a recursion. To do this, consider the expression for

$$\mathbf{R}(t+h) = \int_0^{t+h} \mathbf{\Phi}(t+h,\sigma)\, \mathbf{\Delta}(\sigma)\, \mathbf{\Phi}^{-1}(t+h,\sigma)\, d\sigma, \tag{7.2}$$

where the subscript k is dropped for simplicity. By splitting the interval of integration into $[0,t]$ and $[t,t+h]$ and then using the multiplicative property of the fundamental matrices, $\mathbf{\Phi}(t+h,\sigma) = \mathbf{\Phi}(t+h,t)\, \mathbf{\Phi}(t,\sigma)$, one first obtains

$$\begin{aligned} \mathbf{R}(t+h) &= \int_0^t \mathbf{\Phi}(t+h,t)\, \mathbf{\Phi}(t,\sigma)\, \mathbf{\Delta}(\sigma)\, \mathbf{\Phi}^{-1}(t,\sigma)\, \mathbf{\Phi}^{-1}(t+h,t)\, d\sigma \\ &\quad + \int_0^h \mathbf{\Phi}(t+h,t+\tau)\, \mathbf{\Delta}(t+\tau)\, \mathbf{\Phi}^{-1}(t+h,t+\tau)\, d\tau. \end{aligned} \tag{7.3}$$

Since $\mathbf{\Phi}(t+h,t)$ and $\mathbf{\Phi}^{-1}(t+h,t)$ do not depend on the integration variable σ, Eq. 7.3 is written as

$$\begin{aligned} \mathbf{R}(t+h) &= \mathbf{\Phi}(t+h,t)\, \mathbf{R}(t)\, \mathbf{\Phi}^{-1}(t+h,t) \\ &\quad + \int_0^h \mathbf{\Phi}(t+h,t+\tau)\, \mathbf{\Delta}(t+\tau)\, \mathbf{\Phi}^{-1}(t+h,t+\tau)\, d\tau. \end{aligned} \tag{7.4}$$

The recursion 7.4 means that one has to integrate only over an interval $\tau \in [0,h]$ using the most recent values of $\mathbf{\Phi}(t+h,t+\tau)$ and $\mathbf{\Delta}(t+\tau)$.

We shall not address here the numerical aspects of the interaction between the accumulated quadrature error and the propagation of this error in the product formulae. It will be mentioned, however, that to specify a tolerance level for each integration step of length h, one must know the overall length of the integration interval, in order to keep the accumulated quadrature error below a given threshold. Furthermore, the threshold itself must be controlled by the accuracy specified for convergence of the product formula.

8. Summary. The literature on the analytic representations of the fundamental matrix of solution is not overly large. Apart from some special solvable two- and three-dimensional systems, a few special methods that use Lie algebra basis [27], and the special functions, most integrable systems are either the commutative ones, or there is an explicit transformation that reduces them into one of the systems mentioned above. The periodic systems are semi-integrable in the sense that the fundamental matrix of solution must be known over a single period to be known everywhere.

There really are only four distinct methods that are applicable to a general noncommutative d-dimensional time-varying matrix: 1) Volterra product integral (1887), [5], [9], a globally convergent infinite product of matrix exponentials (or its polynomial/rational approximants); 2) Peano series (1888), [21], globally convergent, but computationally expensive summation (see Eq. 1.6); 3) Magnus' expansion (1954) [16], an infinite series of iterated integrals of nested commutators, with a finite region of convergence [18]; and 4) Fer's expansion (1958) [6], an infinite product of matrix exponentials, with a finite region of convergence.

The new series developed here are related to Peano series in the sense that they are generated by means of the fixed point iteration of a particular Volterra integral equation. However, because of the choice of an initial approximation (identity, in Peano case) and the choice of error function (absolute error, in Peano case), the outcome, in our case, is always a lacunary series with either constant, or variable, exponential gap greater than one.

The new product formulae obtained here are only remotely related to Fer's expansion. The choice of error function is separated from the choice of initial approximation, and resulting products and iterative refinements are expressed in terms of the simplest polynomial and rational functions, rather than the exponentials.

The iterative processes, in Equations 4.30, 5.8, and 6.21, also bear some distant resemblance to Newton's method for operator equations in the sense that the increment, function, and derivative evaluations have been carried out. However, there is no inversion of a Jacobian matrix, nor an explicit inversion of an associated linear operator, as in [23]. Nevertheless, the quadratic convergence, when it occurs, is evident in Equations 4.33 and 6.24.

It is rather unexpected that the choice of three common relative error functions can be so easily related to the three most common rational functions—also the entries into the Padè table for the exponential function—associated with backward, forward, and modified Euler's schemes in the numerical discretization of a linear differential equation, except that here, these are time-varying matrix functions.

What unifies all three products is the systematic use of the integrable part of the differential equation, hence the same integration step for evaluating $\mathbf{R}_k(t)$ that depends linearly on the $\mathbf{\Delta}_k(t)$. What makes them quadratically convergent is the fact that the L-perturbation matrix is the sort of residual, associated with the first derivative, that can be exactly computed.

Finally, there is a connection to the Lyapunov transformation, $\mathbf{L}(t)$, for the linear system in Eq. 1.1. If one writes $\mathbf{\Phi}(t) = \mathbf{L}(t)\,\mathbf{Z}(t)$, where $\mathbf{L}(t)$ is a known, bounded,

invertible and differentiable matrix function [9], then $\mathbf{Z}(t)$ satisfies the differential equation,

$$\mathbf{Z}' = \left(\mathbf{L}^{-1}\mathbf{A}\mathbf{L} - \mathbf{L}^{-1}\mathbf{L}'\right)\mathbf{Z} = \mathbf{L}^{-1}\left(\mathbf{A} - \mathbf{L}'\mathbf{L}^{-1}\right)\mathbf{L}\,\mathbf{Z} = \mathbf{L}^{-1}\boldsymbol{\Delta}\mathbf{L}\,\mathbf{Z}.$$

The point here is that $\boldsymbol{\Delta} = \mathbf{A} - \mathbf{L}'\mathbf{L}^{-1}$ is the L-perturbation matrix generated by $\mathbf{L}(t)$, and the extent to which $\mathbf{L}$ approximates $\boldsymbol{\Phi}$ has a direct and measurable effect on the size of matrix $\mathbf{L}^{-1}\boldsymbol{\Delta}\mathbf{L}$. In this paper, however, the function $\mathbf{Z} = \boldsymbol{\Phi}_0$ is known and the objective is to compute the Lyapunov function $\mathbf{L}$, called here a left correction, that transforms $\boldsymbol{\Phi}_0$ into $\boldsymbol{\Phi}$.

Acknowledgement. This work was partially supported by NASA under grant NGT - 40054 and cooperative agreement NCC5-394.

References

[1] H. F. Baker, *On the integration of linear differential equations*, Proc. London Math. Soc. 35, 333-374 (1903)
[2] H. F. Baker, *Alternants and continuous groups*, Proc. London Math. Soc. (2nd Series) 3, 24-47 (1904)
[3] O. Bottema and B. Roth, *Theoretical Kinematics*, Dover Publ., 1990
[4] S. C. Chu and F. T. Metcalf, *On Gronwall's inequality*, Proc. Amer. Math. Soc., vol 18, 439-440 (1967)
[5] J. D. Dollard and C. N. Friedman, *Product Integration*, Addison-Wesley, Reading, 1979
[6] F. Fer, *Résolution de l'équation matricielle dU/dt = pU par produit infini d'exponentielles matricielles*, Acad. Roy. Belg. Cl. Sci. Bull. Ser. 5, 44, 818-829 (1958)
[7] R. P. Feynman, *An operator calculus having applications in quantum electrodynamics*, Physical Review, 84, 108-128 (1951)
[8] R. A. Frazer, W. J. Duncan, and A. R. Collar, *Elementary Matrices and Some Applications to Dynamics and Differential Equations*, The Macmillan Co., N.Y., 1946
[9] F. R. Gantmacher, *The Theory of Matrices* (Vol. 2), Chelsea, Bronx, NY, 1959
[10] H. E. Gollwitzer, *A note on a functional inequality*, Proc. Amer. Math. Soc., vol 23, 642-647 (1969)
[11] H. W. Guggenheimer, *Differential Geometry*, Dover, New York, 1977
[12] J. K. Hale, *Ordinary Differential Equations*, Wiley-Interscience, 1969
[13] F. Hausdorff, *Die symbolische Exponentialformel in der Gruppentheorie*, Berichte uber die Verhandlungen der Koniglich Sächsischen Gessellschaft der Wissenshaften zu Leipzig (Math. Phys. Klasse), 58, 19-48 (1906)
[14] P. Henrici, *Applied and Computational Complex Analysis, Vol 2.* Wiley-Interscience, 1977
[15] P. D. Lax, *Linear Algebra*, Wiley-Interscience, 1997
[16] W. Magnus, *On the exponential solution of differential equations for a linear operator*, Comm. Pure and Appl. Math. VII, 649-673 (1954)
[17] R. K. Miller, *Nonlinear Volterra Integral Equations*, W. A. Benjamin, Inc. Menlo Park, California, 1971
[18] P. C. Moan, J. A. Oteo, and J. Ros, *On the existence of the exponential solution of linear differential systems*, J. Phys. A: Math. Gen. 32, 5133-5139 (1999)
[19] I. Najfeld and T. F. Havel, *Derivatives of the matrix exponential and their computation, Adv. Appl. Math.*, 16, 321-375, (1995)
[20] F. W. J. Olver, *Asymptotics and Special Functions*, Acad. Press, New York, 1974
[21] G. Peano, *Intégration par séries des équations différentielles linéaires*, Math. Ann. 32, 450-456 (1888)
[22] M. C. Pease, III, *Methods of Matrix Algebra*, Academic Press, New York, 1965
[23] L. B. Rall, *Computational Solution of Nonlinear Operator Equations*, John Wiley & Sons, 1969
[24] W. T. Reid, *Riccati Differential Equations*, Academic Press, New York, 1972
[25] J. A. Richards, *Analysis of Periodically Time-Varying Systems*, Springer-Verlag, 1983
[26] F. Riesz and B. Sz.-Nagy, *Functional Analysis*, Dover, 1990

[27] R. M. Wilcox, *Exponential operators and parameter differentiation in quantum physics*, J. Math. Phys. 8, 962-982 (1967)

[28] R. M. Wilcox, *Bounds for Approximate Solutions to the Operator Differential Equation* $\dot{Y}(t) = M(t)Y(t)$; *Applications to Magnus expansion and to* $\ddot{u} + [1 + f(t)]u = 0$, J. Math. Anal. and Appl. 39, 92-111 (1972)