Validated Numerical Bounds on the Global Error for Initial Value Problems for Stiff Ordinary Differential Equations

by

Chao Yu

A thesis submitted in conformity with the requirements
for the degree of Master of Science
Graduate Department of Computer Science
University of Toronto

Abstract

Validated Numerical Bounds on the Global Error for Initial Value Problems for Stiff Ordinary Differential Equations

Chao Yu

Master of Science

Graduate Department of Computer Science

University of Toronto

2004

There are many standard numerical methods for initial value problems (IVPs) for ordinary differential equations (ODEs). Compared with these methods, validated methods for IVPs for ODEs produce bounds that are guaranteed to contain the true solution of a problem, if the true solution exists and is unique.

The main result of this thesis is a formula to bound the global error associated with the numerical solution of a stiff IVP for an ODE. We give the complete proof of this result. Moreover, we derive Dahlquist's formula and Neumaier's formula from this formula. We also give alternative (and possibly simpler) proofs of some known related results.

Acknowledgements

I would especially like to thank my supervisor, Professor Ken Jackson. Without his diligence, his guidance and his understanding, this thesis would not have been possible during my MSc program. I would also like to thank Professor John Pryce for his notes which were the inspiration for this thesis. In addition, I would like to thank Professor Wayne Enright for his careful reading of the thesis and for providing valuable comments. I am forever grateful for the love and support of my parents. Finally, I thank the Department of Computer Science for its financial support and providing me with a good environment in which to work.

Contents

Chapter 1

Introduction

We consider the initial value problem (IVP) for an ordinary differential equation (ODE)

$$y' = f(t, y), \quad y(t_0) = y_0, \qquad t \in [t_0, T], \tag{1.1}$$

where $y \in R^m$ and $f : R \times R^m \to R^m$. Throughout this thesis, we assume that f is smooth and that a unique solution to (1.1) exists on $[t_0, T]$.

Given the grid $t_0 < t_1 < \cdots < t_N = T$, $h_n = t_{n+1} - t_n$ is the stepsize on the n^{th} step. Denote the true solution to (1.1) by $y(t)$, and let y_n be an approximation to $y(t_n)$. The global error at the mesh point t_n is

$$e_n = y_n - y(t_n).$$

A *validated* numerical method for the IVP (1.1) generates guaranteed bounds on the global errors $\{e_n : n = 0, \ldots, N\}$. Validated numerical methods often use *interval arithmetic* (reviewed in §2.2) to accomplish this goal.

Let $z(t)$ be a vector of piecewise polynomial approximate solution to (1.1), where

$$z(t) = z_n(t) \quad \text{on} \quad [t_n, t_{n+1}).$$

Such a vector of piecewise polynomials $z(t)$ can be generated, for example, by computing a discrete numerical solution $\{y_n : n = 0, \ldots, N\}$ at mesh points $\{t_n : n = 0, \ldots, N\}$

and constructing an interpolant to obtain a vector of polynomials $z_n(t)$ on the interval $[t_n, t_{n+1})$. Note that $z(t_n) = y_n$ and $z(t)$ is continuous on $[t_0, T]$.

The global error associated with the approximate solution $z(t)$ is

$$e(t) = z(t) - y(t), \qquad t \in [t_0, T].$$

The *defect* associated with (1.1) is defined by

$$\delta(t) = z'(t) - f(t, z), \qquad t \in [t_0, T].$$

(Since the approximate solution $z(t)$ discussed above is a vector of piecewise polynomials, $\delta(t)$ will be computable and continuous everywhere except posssibly at the mesh points, $\{t_n : n = 1, \ldots, N-1\}$, but this does not affect our analysis.)

We will focus on the global error $e(t)$. The main result of our thesis is that

$$e(t_{n+1}) \in exp([A](t_{n+1} - t_n))e(t_n) + \int_{t_n}^{t_{n+1}} exp([A](t_{n+1} - s))\delta(s)ds, \tag{1.2}$$

where $[A]$ is an interval matrix and $exp([A]\varphi(t))\delta(t)$ is an interval vector function (to be defined later).

This result should be particularly well-suited for *stiff* IVPs, a subclass of IVPs in which some solution components decay rapidly compared to the time-scale of the problem. Except for Neumaier's results [24], which we expand upon in this thesis, none of the published validated numerical methods that we know of now are suitable for stiff problems in the sense that the existing validated numerical methods all suffer from a severe stepsize restriction on this class of problems, similar to that encountered by traditional nonstiff methods for IVPs for ODEs. For a further discussion of this deficiency, see [20].

Neumaier uses *logarithmic norms* (reviewed in §2.5) to prove that, if we take $t_0 = 0$ in (1.1), $\|e(0)\| \leq \epsilon_0$, $\mu(f_y(t, y)) \leq \mu$ for all $t \in [0, T]$ and all y in a suitable domain, and $\|\delta(t)\| \leq \epsilon$ for all $t \in [0, T]$, then

$$\|e(t)\| \leq \begin{cases} \epsilon_0 e^{\mu t} + \frac{\epsilon}{\mu}(e^{\mu t} - 1) & \text{if } \mu \neq 0 \\ \epsilon_0 + \epsilon t & \text{if } \mu = 0 \end{cases} \tag{1.3}$$

for all $t \in [0, T]$.

When the differential equation (1.1) with ($t_0 = 0$) satisfies the uniform dissipation condition $\mu < 0$, (1.3) gives an effective global bound for all times. Moreover, although Neumaier has not yet implemented these schemes, it appears that they should be able to solve stiff systems without the severe stepsize restriction noted above from which other existing validated numerical methods for IVPs for ODEs suffer.

Our original goal was to use interval arithmetic to compute sharp enclosures of the right side of (1.2) directly, in the hope that this might produce better bounds than Neumaier's. However, this has proven more difficult than we originally expected and so we leave this task to future work.

An outline of this thesis follows. Chapter 2 contains background material that we need later. In particular, we introduce interval arithmetic on real intervals, interval vectors, and interval matrices, as well as the logarithmic norm and Hausdorff distance. In addition, we review several results that are used later in this thesis. We also provide a simpler proof of one of these related results concerning the logarithmic norm.

The proof of formula (1.2) is not immediate, as far as we know. In Chapter 3, we prove formula (1.2) using the Hausdorff distance and interval arithmetic.

Formula (1.3) is a special case of Theorem 1.1 of [5] and the Main Theorem of [11], both of which imply that, if $\|e(0)\| \leq \epsilon_0$, $\mu(A(t)) \leq c(t)$ for all $t \in [0, T]$ and $\|\delta(t)\| \leq \rho(t)$ for all $t \in [0, T]$, then

$$\|e(t)\| \leq \epsilon_0 e^{\int_0^t c(s)ds} + e^{\int_0^t c(s)ds} \int_0^t \rho(s) e^{-\int_0^s c(u)du} ds \tag{1.4}$$

for all $t \in [0, T]$.

In Chapter 4, we compare our results to Dahlquist's and Neumaier's. In particular, we derive Dahlquist's formula (1.4) from our formula (1.2). Our motivation for this is not to have another proof of Dahlquist's important result, but rather to show that our approach yields bounds that are as tight or tighter than those that can be obtained by Dahlquist's and Neumaier's approaches. On the other hand, we give a simple example

which shows that formula (1.2) may sometimes yield bounds that are tighter than those given by formulas (1.3) and (1.4).

We summarize our conclusions and discuss future work in Chapter 5.

Chapter 2

Preliminaries

In this chapter, we review some mathematical background that is used later in this thesis.

2.1 Vector Norms and Matrix Norms

2.1.1 Vector Norms

For $x = (x_1, \cdots, x_m)^T \in R^m$, the $p-$norms are defined by

$$\|x\|_p = (|x_1|^p + \cdots + |x_m|^p)^{\frac{1}{p}}, \qquad p \geq 1.$$

In particular, the 1, 2, and ∞ norms are

$$\|x\|_1 = |x_1| + \cdots + |x_m|,$$

$$\|x\|_2 = (|x_1|^2 + \cdots + |x_m|^2)^{\frac{1}{2}} = (x^T x)^{\frac{1}{2}},$$

$$\|x\|_\infty = \max_{1 \leq i \leq m} |x_i|.$$

All vector norms on R^m are equivalent in the sense that, if $\|\cdot\|_a$ and $\|\cdot\|_b$ are any two vector norms on R^m, then there exist positive constants α and $\beta \in R$ such that $\alpha\|x\|_a \leq \|x\|_b \leq \beta\|x\|_a$ for all $x \in R^m$ (α, β may depend on m).

2.1.2 Matrix Norms

For $A \in R^{m\times m}$, the matrix norm $\|\cdot\|$ subordinate to a vector norm $\|\cdot\|$ is defined by

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|}. \tag{2.1}$$

The matrix $p-$norms are related to the vextor $p-$norms in this way. If follows easily from (2.1) that

$$\begin{aligned} \|Ax\|_p &\leq \|A\|_p \|x\|_p \\ \|I\|_p &= 1 \end{aligned}$$

For

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{pmatrix}$$

the 1, 2, and ∞ matrix norms are given by

$$\|A\|_1 = \sup_{x\neq 0} \frac{\|Ax\|_1}{\|x\|_1} = \max_{1\leq j\leq m} \sum_{i=1}^{m} |a_{ij}|,$$

$$\|A\|_2 = \sup_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max\{\sqrt{\lambda} : \lambda \text{ is an eigenvalue of } A^T A\},$$

$$\|A\|_\infty = \sup_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} = \max_{1\leq i\leq m} \sum_{j=1}^{m} |a_{ij}|.$$

It also follows easily from (2.1) that

$$\|AB\| \leq \|A\|\|B\|, \qquad A \in R^{m\times m}, \qquad B \in R^{m\times m}.$$

In particular, for any $k = 1, 2, 3, \dots$, we have

$$\|A^k\| \leq \|A\|^k.$$

All matrix norms on $R^{m\times m}$ are equivalent in the sense that, if $\|\cdot\|_a$ and $\|\cdot\|_b$ are any two matrix norms on $R^{m\times m}$, then there exist positive constants α and $\beta \in R$ such that $\alpha\|A\|_a \leq \|A\|_b \leq \beta\|A\|_a$ for all $A \in R^{m\times m}$ (α, β may depend on m).

Let $\{A_k : k = 1, 2, 3, \ldots\}$ and $\{B_k : k = 1, 2, 3, \ldots\}$ be two sets of $m \times m$ matrices. If there exists an $m \times m$ matrix A, such that $\|A_k - A\| \to 0$ as $k \to \infty$, then we say that A_k converges to A and we denote this by $A_k \to A$. If $\|\sum_{i=k+1}^{\infty} B_i\| \to 0$ as $k \to \infty$, then $A_k = \sum_{i=1}^{k} B_i$ converges to some matrix A, and we denote this by $A = \sum_{i=1}^{\infty} B_i$.

2.2 Interval Arithmetic

The set of intervals on the real line $\mathbb{R}$ is defined by

$$\mathbb{R} = \{[a] = [\underline{a}, \bar{a}] | \underline{a}, \bar{a} \in \mathbb{R}, \underline{a} \leq \bar{a}\}.$$

If $\underline{a} = \bar{a}$, then $[a]$ is a *thin* interval. If $\underline{a} \geq 0$, then $[a]$ is *nonnegative*, which we denote by $[a] \geq 0$. If $\underline{a} = -\bar{a}$, then $[a]$ is *symmetric*. Two intervals $[a]$ and $[b]$ are equal if $\underline{a} = \underline{b}$ and $\bar{a} = \bar{b}$.

The four operations of real arithmetic, addition $(+)$, subtraction $(-)$, multiplication $(*)$ and division $(/)$, can be extended to intervals as follows. For any such operator, denoted by $\circ$, define

$$[a] \circ [b] = \{x \circ y : x \in [a], y \in [b]\}. \tag{2.2}$$

For any intervals $[a]$ and $[b]$, it is easy to see that the following properties are satisfied

$$\begin{aligned}
[a] + [b] &= [\underline{a} + \underline{b}, \bar{a} + \bar{b}], \\
[a] - [b] &= [\underline{a} - \bar{b}, \bar{a} - \underline{b}], \\
[a] * [b] &= [\min\{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\}, \max\{\underline{a}\underline{b}, \underline{a}\bar{b}, \bar{a}\underline{b}, \bar{a}\bar{b}\}], \\
[a]/[b] &= [\underline{a}, \bar{a}] * [1/\bar{b}, 1/\underline{b}], \qquad \text{if } 0 \notin [b].
\end{aligned}$$

The width of any interval $[a]$ is defined by

$$w([a]) = \bar{a} - \underline{a}.$$

The midpoint (or center) of any interval $[a]$ is defined by

$$m([a]) = (\bar{a} + \underline{a})/2.$$

The magnitude of any interval $[a]$ is defined by

$$|[a]| = \max\{|\underline{a}|, |\bar{a}|\}.$$

If t is a real number, and $[a]$ is an interval, then

$$t[a] = \{tx : x \in [a]\}.$$

An interval vector $[a]$ is an element of $\mathbb{IR}^m$, defined by

$$[a] = \begin{pmatrix} [a_1] \\ [a_2] \\ \vdots \\ [a_m] \end{pmatrix}$$

where $[a_i] = [\underline{a}_i, \bar{a}_i] \in \mathbb{IR}$, for $i = 1, \ldots, m$.

An $m \times m$ interval matrix $[A]$ is an element of $\mathbb{IR}^{m\times m}$ defined by

$$[A] = \begin{pmatrix} [a_{11}] & [a_{12}] & \ldots & [a_{1m}] \\ [a_{21}] & [a_{22}] & \ldots & [a_{2m}] \\ \vdots & \vdots & \vdots & \vdots \\ [a_{m1}] & [a_{m2}] & \ldots & [a_{mm}] \end{pmatrix}$$

where $[a_{ij}] = [\underline{a}_{ij}, \bar{a}_{ij}] \in \mathbb{IR}$ for $i = 1, \ldots, m$ and $j = 1, \ldots, m$.

We define the width and midpoint of an interval matrix componentwise as follows:

$$w([A]) = (w([a_{ij}]))_{1\le i\le m, 1\le j\le m},$$

$$mid([A]) = (mid([a_{ij}]))_{1\le i\le m, 1\le j\le m}.$$

The width and midpoint of an interval vector are defined similarly.

Since we use the infinity norm only for interval vectors and matrices in this thesis, we do not append the standard subscript ∞ to identify it. The infinity norm of an interval vector $[a] \in \mathbb{IR}^m$ is defined by

$$\|[a]\| = \max_{1\le i\le m} \{|[a_i]|\}$$

and the infinity norm of an interval matrix $[A] \in \mathbb{IR}^{m\times m}$ is defined by

$$\|[A]\| = \max_{1\le i\le m} \sum_{j=1}^{m} |[a_{ij}]|.$$

The standard matrix norm inequalities hold for interval matrices as well. Let $[A], [B] \in \mathbb{IR}^{m\times m}$, and $[v] \in \mathbb{IR}^m$, then it is easy to show

$$\begin{aligned} \|[A] + [B]\| &\le \|[A]\| + \|[B]\|, \\ \|[A][v]\| &\le \|[A]\|\|[v]\|, \\ \|[A][B]\| &\le \|[A]\|\|[B]\|. \end{aligned} \tag{2.3}$$

It follows immediately from (2.3) that

$$\|[A]^k\| \le \|[A]\|^k$$

for $k = 1, 2, 3, \ldots$

2.3 Matrix Functions

A matrix function is a matrix whose elements are functions:

$$A(t) = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \ldots & a_{1m}(t) \\ a_{21}(t) & a_{22}(t) & \ldots & a_{2m}(t) \\ \vdots & \vdots & \vdots & \vdots \\ a_{m1}(t) & a_{m2}(t) & \ldots & a_{mm}(t) \end{pmatrix}.$$

If $a_{ij}(t)$ is continuous on $[t_0, T]$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, m$, then we say that $A(t)$ is continuous on $[t_0, T]$. It is easy to check that, if $A(t)$ is continuous on $[t_0, T]$, $\{t_k \in [t_0, T]\}$ and $t_k \to t \in [t_0, T]$, then $\|A(t_k) - A(t)\| \to 0$.

If $a_{ij}(t)$ is differentiable on $[t_0, T]$ for all $i = 1, \ldots, m$ and $j = 1, \ldots, m$, then we say

that $A(t)$ is differentiable on $[t_0, T]$ and we denote the derivative of $A(t)$ by

$$A'(t) = \begin{pmatrix} a'_{11}(t) & a'_{12}(t) & \dots & a'_{1m}(t) \\ a'_{21}(t) & a'_{22}(t) & \dots & a'_{2m}(t) \\ \vdots & \vdots & \vdots & \vdots \\ a'_{m1}(t) & a'_{m2}(t) & \dots & a'_{mm}(t) \end{pmatrix}.$$

If $a_{ij}(t)$ is integrable on $[t_0, T]$ for all $i = 1, \dots, m$ and $j = 1, \dots, m$, then we say that $A(t)$ is integrable on $[t_0, T]$ and we denote the integral of $A(t)$ by

$$\int_{t_0}^{T} A(t)dt = \begin{pmatrix} \int_{t_0}^{T} a_{11}(t)dt & \int_{t_0}^{T} a_{12}(t)dt & \dots & \int_{t_0}^{T} a_{1m}(t)dt \\ \int_{t_0}^{T} a_{21}(t)dt & \int_{t_0}^{T} a_{22}(t)dt & \dots & \int_{t_0}^{T} a_{2m}(t)dt \\ \vdots & \vdots & \vdots & \vdots \\ \int_{t_0}^{T} a_{m1}(t)dt & \int_{t_0}^{T} a_{m2}(t)dt & \dots & \int_{t_0}^{T} a_{mm}(t)dt \end{pmatrix}.$$

Let $\{B_i(t) : i = 1, 2, 3, \dots\}$ be a set of $m \times m$ matrix functions, and assume each $B_i(t)$ is continuous on $[t_0, T]$. Let $A_k(t) = \sum_{i=1}^{k} B_i(t)$, $k = 1, 2, 3, \dots$ If there exists an $m \times m$ matrix function $A(t)$ such that, for any $\epsilon > 0$, there exists a K such that $\|A_k(t) - A(t)\| < \epsilon$ for $k \geq K$ and any $t \in [t_0, T]$, then we say that $A_k = \sum_{i=1}^{k} B_i(t)$ is uniformly convergent to $A(t)$ on $[t_0, T]$ and we denote this by $A(t) = \sum_{i=1}^{\infty} B_i(t)$.

If A is an $m \times m$ matrix, then the matrix series $I + \frac{A}{1!}t + \frac{A^2}{2!}t^2 + \cdots + \frac{A^k}{k!}t^k + \cdots$ is convergent for any $t \in R$. We denote the sum of this series by e^{At}.

2.4 Mean Value Theorem for Functions of Several Variables

If $F : D \subset R^n \to R^m$ is differentiable at every point in a convex set D, then for any two points x and $y \in D$

$$F(y) - F(x) = \int_0^1 F'(y - t(y - x))(y - x)dt$$

where $F'(y-t(y-x)) = \frac{\partial F(z)}{\partial z}|_{z=y-t(y-x)}$. This result follows from the observations that

$$\begin{aligned}\int_0^1 F'(y-t(y-x))(y-x)dt &= -\int_0^1 (\frac{d}{dt}F(y-t(y-x)))dt \\ &= -F(y-t(y-x))|_{t=0}^{t=1} \\ &= F(y)-F(x).\end{aligned}$$

2.5 The Logarithmic Norm

The logarithmic norm (also known as the log norm, the logarithmic derivative or the measure of a matrix) was introduced in 1958 separately by Dahlquist [5] and Lozinskij [17] as a tool to study the growth in numerical solutions of differential equations. For any matrix norm subordinate to a vector norm and $A \in R^{m\times m}$, define the logarithmic norm of A by

$$\mu(A) = \lim_{h\to+0} \frac{\|I+hA\|-1}{h}. \tag{2.4}$$

We use the following three well-known lemmas [30] later.

Lemma 2.1

(1) $\mu(A) \le \|A\|$;

(2) $\mu(A+B) \le \mu(A)+\mu(B)$;

(3) $\mu(\alpha A) = \alpha\mu(A)$, *for any* $\alpha \ge 0$;

(4) $|\mu(A)-\mu(B)| \le \|A-B\|$.

Lemma 2.2 *If* $A_n \to A$, *then* $\mu(A_n) \to \mu(A)$.

Proof. From Lemma 2.1 (4), it follows that

$$|\mu(A_n)-\mu(A)| \le \|A_n - A\| \to 0.$$

□

Lemma 2.3 $\|e^A\| \leq e^{\mu(A)} \leq e^{\|A\|}$.

We provide below what we consider to be a simpler proof than the standard proof that appears for example in [3].

Proof. From the definition of the log norm, (2.4), it follows that, for any $\epsilon > 0$, there exists a $\delta_1 > 0$, such that, for all h satisfying $0 < h < \delta_1$, we have that

$$\frac{\|I + hA\| - 1}{h} - \mu(A) < \frac{\epsilon}{4}$$

or equivalently

$$\|I + hA\| < 1 + h\mu(A) + \frac{h\epsilon}{4}.$$

Since

$$e^{hA} = I + hA + \sum_{k=2}^{\infty} \frac{h^k A^k}{k!}$$

it follows that

$$\|e^{hA}\| \leq \|I + hA\| + \sum_{k=2}^{\infty} \frac{h^k \|A\|^k}{k!} = \|I + hA\| + h^2 M_1(h)$$

where

$$M_1(h) = \sum_{k=2}^{\infty} \frac{h^{k-2} \|A\|^k}{k!}$$

is a convergent series. Therefore, there exists $\delta_2 > 0$, satisfying $\delta_2 < \delta_1$, such that, for all h satisfying $0 < h < \delta_2$, we have $hM_1(h) < \frac{\epsilon}{4}$. Thus

$$\|e^{hA}\| < \|I + hA\| + \frac{h\epsilon}{4} < 1 + h\mu(A) + \frac{h\epsilon}{2}. \tag{2.5}$$

Similarly,

$$\begin{aligned} e^{h\mu(A)} &= 1 + h\mu(A) + \sum_{k=2}^{\infty} \frac{h^k [\mu(A)]^k}{k!} \\ &\geq 1 + h\mu(A) - \sum_{k=2}^{\infty} \frac{h^k |\mu(A)|^k}{k!} \\ &= 1 + h\mu(A) - h^2 M_2(h) \end{aligned} \tag{2.6}$$

where

$$M_2(h) = \sum_{k=2}^{\infty} \frac{h^{k-2}|\mu(A)|^k}{k!}$$

is also a convergent series. Therefore, there exists $\delta > 0$, satisfying $\delta < \delta_2$, such that, for all h satisfying $0 < h < \delta$, we have $hM_2(h) < \frac{\epsilon}{2}$. Thus from (2.6)

$$1 + h\mu(A) < e^{h\mu(A)} + \frac{h\epsilon}{2}. \tag{2.7}$$

Combining inequalities (2.5) and (2.7), we get

$$\|e^{hA}\| < 1 + h\mu(A) + \frac{h\epsilon}{2} < e^{h\mu(A)} + h\epsilon \tag{2.8}$$

for all h satisfying $0 < h < \delta$.

Choose a positive integer N such that $h = \frac{1}{N} < \delta$. It follows from (2.8) that

$$\|e^{\frac{A}{N}}\| < e^{\frac{\mu(A)}{N}} + \frac{\epsilon}{N}$$

Therefore

$$\begin{aligned}
\|e^A\| &= \|(e^{\frac{A}{N}})^N\| \\
&\leq \|e^{\frac{A}{N}}\|^N \\
&< \left(e^{\frac{\mu(A)}{N}} + \frac{\epsilon}{N}\right)^N \\
&= e^{\mu(A)} + \sum_{k=1}^{N} \binom{N}{k} \left(\frac{\epsilon}{N}\right)^k e^{\frac{(N-k)\mu(A)}{N}} \\
&\leq e^{\mu(A)} + \sum_{k=1}^{N} \frac{1}{N^k} \binom{N}{k} \epsilon^k e^{|\mu(A)|}.
\end{aligned}$$

Since

$$\frac{1}{N^k}\binom{N}{k} = \frac{N(N-1)\cdots(N-k+1)}{N^k k!} \leq \frac{1}{k!}$$

for $k = 1, \ldots, N$, it follows that

$$\|e^A\| < e^{\mu(A)} + \epsilon e^{|\mu(A)|} \sum_{k=1}^{N} \frac{\epsilon^{k-1}}{k!}.$$

Without loss of generality, we may assume $\epsilon < 1$. Hence,

$$\sum_{k=1}^{N} \frac{\epsilon^{k-1}}{k!} < \sum_{k=1}^{N} \frac{1}{k!} < e.$$

Therefore,

$$\|e^A\| < e^{\mu(A)} + \epsilon e^{|\mu(A)|+1}. \tag{2.9}$$

Since ϵ is an arbitrary positive constant, inequality (2.8) implies that $\|e^A\| \leq e^{\mu(A)}$. Moreover, since $\mu(A) \leq \|A\|$ from Lemma 2.1 (1), $\|e^A\| \leq e^{\mu(A)} \leq e^{\|A\|}$.

□

Note that $\mu(A) = \lim_{h\to 0+} \frac{\|I+hA\|-1}{h}$ depends on the matrix norm we use for $\|I+hA\|$. In this thesis, we use $\|I + hA\|_\infty$ throughout. Thus for

$$A = \begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ \dots & \dots & \dots & \dots \\ a_{m1} & a_{m2} & \dots & a_{mm} \end{pmatrix}$$

we get

$$I + hA = \begin{pmatrix} 1 + ha_{11} & ha_{12} & \dots & ha_{1m} \\ ha_{21} & 1 + ha_{22} & \dots & ha_{2m} \\ \dots & \dots & \dots & \dots \\ ha_{m1} & ha_{m2} & \dots & 1 + ha_{mm} \end{pmatrix}.$$

Hence,

$$\|I + hA\|_\infty = \max_{1\leq i\leq m} (|1 + ha_{ii}| + \sum_{j\neq i} |ha_{ij}|).$$

For $h > 0$ small enough, $1 + ha_{ii} > 0$. So $|1 + ha_{ii}| = 1 + ha_{ii}$. Thus,

$$\|I + hA\|_\infty = 1 + h \max_{1\leq i\leq m} (a_{ii} + \sum_{j\neq i} |a_{ij}|).$$

Therefore,

$$\mu(A) = \lim_{h\to+0} \frac{\|I + hA\|_\infty - 1}{h} = \max_{1\leq i\leq m} (a_{ii} + \sum_{j\neq i} |a_{ij}|).$$

Note that $\mu(A)$ may be negative. Also note that, if $a_{ii} \geq 0$ for $i = 1, \dots, m$, then $\mu(A) = \|A\|$.

2.6 Hausdorff Distance

Let (X, d) be a metric space and $P_0(X)$ be the set of all nonempty subsets of X. For $A \in P_0(X)$, and $x \in X$, define the distance from x to A as

$$d(x, A) = \inf_{a \in A} d(x, a).$$

If $x \in A$, then $d(x, A) = 0$, but, if $d(x, A) = 0$, we can not conclude that $x \in A$. For example, if $x = 0$, $A = (0, 1)$, then $d(x, A) = 0$, but $x \notin A$. However, if $d(x, A) = 0$, then $x \in \bar{A}$, where $\bar{A}$ is the closure of A.

For A and $B \in P_0(X)$, define the distance from A to B as

$$d(A, B) = \inf_{a \in A, b \in B} d(a, b).$$

If $A \cap B \neq \emptyset$, then $d(A, B) = 0$, but if $d(A, B) = 0$, then we can not conclude that $A \cap B = \emptyset$. For example, if $A = (0, 1)$ and $B = (1, 2)$, then $d(A, B) = 0$, but $A \bigcap B = \emptyset$. However, if $d(A, B) = 0$, then $\bar{A} \cap \bar{B} \neq \emptyset$.

For $\lambda \in R_+ = \{x \in R : x > 0\}$ and $A \in P_0(X)$, define

$$\lambda + A = \{x \in X : d(x, A) < \lambda\} = \{x \in X : \exists a \in A \text{ s.t. } d(x, a) < \lambda\}.$$

For A and $B \in P_0(X)$, define the Hausdorff distance [14] [15] between A and B as

$$H(A, B) = \inf\{\lambda > 0 : A \subset \lambda + B \text{ and } B \subset \lambda + A\}. \tag{2.10}$$

The following results follow immediately from definition (2.10).

- If $\lambda > 0$, $A \subset \lambda + B$ and $B \subset \lambda + A$, then $H(A, B) \leq \lambda$.
- For any $\epsilon > 0$, $A \subset (H(A, B) + \epsilon) + B$ and $B \subset (H(A, B) + \epsilon) + A$.
- If $A = B$, then $H(A, B) = 0$, but if $H(A, B) = 0$, we can not conclude that $A = B$. For example, if $A = (0, 1]$ and $B = [0, 1)$, then $H(A, B) = 0$, but $A \neq B$. However, if $H(A, B) = 0$, then $\bar{A} = \bar{B}$.

The Hausdorff distance between A and B can also be defined as

$$H_1(A,B) = \max\{\sup_{x\in A}\inf_{y\in B} d(x,y), \sup_{y\in B}\inf_{x\in A} d(x,y)\}. \tag{2.11}$$

Definitions (2.10) and (2.11) are equivalent. To see this, choose any $\epsilon > 0$. As noted above, it follows from (2.10) that

$$A \subset (H(A,B)+\epsilon) + B.$$

Therefore, for any $x \in A$, there exists $y \in B$, such that $d(x,y) < H(A,B)+\epsilon$. Thus,

$$\inf_{y\in B} d(x,y) < H(A,B)+\epsilon.$$

Since the last inequality holds for all $x \in A$,

$$\sup_{x\in A}\inf_{y\in B} d(x,y) \le H(A,B)+\epsilon.$$

Similarly, we can show that

$$\sup_{y\in B}\inf_{x\in A} d(x,y) \le H(A,B)+\epsilon.$$

Thus, $H_1(A,B) \le H(A,B)+\epsilon$. Since ϵ is arbitrary, $H_1(A,B) \le H(A,B)$.

On the other hand, for any $\epsilon > 0$, by definition (2.11) of $H_1(A,B)$,

$$\sup_{x\in A}\inf_{y\in B} d(x,y) < H_1(A,B)+\epsilon.$$

Thus, for any $x \in A$, $\inf_{y\in B} d(x,y) < H_1(A,B)+\epsilon$. Hence, there exists $y \in B$, such that $d(x,y) < H_1(A,B)+\epsilon$. Therefore, $A \subset (H_1(A,B)+\epsilon)+B$.

Similarly, we can show that $B \subset (H_1(A,B)+\epsilon)+A$. Hence, $H(A,B) \le H_1(A,B)+\epsilon$. Since ϵ is arbitrary, $H(A,B) \le H_1(A,B)$.

Therefore, $H(A,B) = H_1(A,B)$.

Indeed, in [19], for $X = R^m$, and A and B compact sets, the authors define the Hausdorff distance as

$$H_2(A,B) = \max\{\max_{x\in A}\min_{y\in B}\|x-y\|, \max_{y\in B}\min_{x\in A}\|x-y\|\}.$$

In this case, since A and B are compact, $d(x, y) = \|x - y\|$ can obtain its sup and inf values. Thus, $H_2(A, B) = H_1(A, B) = H(A, B)$.

Lemma 2.4 *Let A and $B \in P_0(X)$ with B compact. Then, for any $x \in A$, there exists $y \in B$ such that $d(x, y) \leq H(A, B)$.*

Proof. Combining $H_1(A, B) = H(A, B)$ with definition (2.11) of $H_1(A, B)$, we get that

$$\sup_{x \in A} \inf_{y \in B} d(x, y) \leq H_1(A, B) = H(A, B).$$

Hence, for any $x \in A$, $\inf_{y \in B} d(x, y) \leq H(A, B)$. Since B is compact, there exists $\hat{y} \in B$ such that $d(x, \hat{y}) = \inf_{y \in B} d(x, y) \leq H(A, B)$. That is, for any $x \in A$, there exists $y \in B$ such that $d(x, y) \leq H(A, B)$.

□

Lemma 2.5 *If X is a normed linear space, then for any $A_1, A_2, B_1, B_2 \in P_0(X)$,*

$$H(A_1 + A_2, B_1 + B_2) \leq H(A_1, B_1) + H(A_2, B_2).$$

See [15], Proposition 4.3.15(ii) and Remark 4.3.17.

Proof. For any $x, y \in X$, we define $d(x, y) = \|x - y\|$, where $\|\cdot\|$ is the norm associated with the linear space X.

From definition (2.10) of the Hausdorff distance H, for any $\epsilon > 0$,

$$\begin{aligned} A_1 &\subset (H(A_1, B_1) + \epsilon) + B_1, \\ B_1 &\subset (H(A_1, B_1) + \epsilon) + A_1, \\ A_2 &\subset (H(A_2, B_2) + \epsilon) + B_2, \\ B_2 &\subset (H(A_2, B_2) + \epsilon) + A_2. \end{aligned}$$

Since X is linear space, for any $u \in A_1 + A_2$, there exist $a_1 \in A_1$ and $a_2 \in A_2$ such that $u = a_1 + a_2$. Moreover, since $A_1 \subset (H(A_1, B_1) + \epsilon) + B_1$, there exists $b_1 \in B_1$

such that $d(a_1, b_1) < H(A_1, B_1) + \epsilon$. Similarly, there exists $b_2 \in B_2$ such that $d(a_2, b_2) < H(A_2, B_2) + \epsilon$. Let $v = b_1 + b_2 \in B_1 + B_2$. Then

$$\begin{aligned} d(u, v) &= \|u - v\| \\ &= \|(a_1 + a_2) - (b_1 + b_2)\| \\ &\leq \|a_1 - b_1\| + \|a_2 - b_2\| \\ &= d(a_1, b_1) + d(a_2, b_2) \\ &< H(A_1, B_1) + H(A_2, B_2) + 2\epsilon. \end{aligned}$$

Therefore, $A_1 + A_2 \subset (H(A_1, B_1) + H(A_2, B_2) + 2\epsilon) + (B_1 + B_2)$. Similarly, $B_1 + B_2 \subset (H(A_1, B_1) + H(A_2, B_2) + 2\epsilon) + (A_1 + A_2)$. Hence,

$$H(A_1 + A_2, B_1 + B_2) \leq H(A_1, B_1) + H(A_2, B_2) + 2\epsilon.$$

Since ϵ is arbitrary,

$$H(A_1 + A_2, B_1 + B_2) \leq H(A_1, B_1) + H(A_2, B_2).$$

□

Let $[A] \in \mathbb{IR}^{m \times m}$ be an interval matrix and $\varphi : [t_0, T] \to R$ and $\delta : [t_0, T] \to R^m$ be continuous functions. We define the integral of an interval vector function $exp([A]\varphi(t))\delta(t)$ as follows. Consider a sequence of meshes $t_0 = t_0^{(p)} < t_1^{(p)} < \cdots < t_p^{(p)} = T$ where $\Delta t_i^{(p)} = t_{i+1}^{(p)} - t_i^{(p)}$, $i = 0, 1, \ldots, p-1$, and let $\xi_i^{(p)}$ be any point in $[t_i^{(p)}, t_{i+1}^{(p)}]$, $i = 0, 1, \ldots, p-1$. Moreover, assume that $\max_{0 \leq i < p} \Delta t_i^{(p)} \to 0$ as $p \to \infty$. If there exists an interval vector $[I] \in \mathbb{IR}^m$ such that, for any such sequence of meshes, the Hausdorff distance

$$H(\sum_{i=0}^{p-1} exp([A]\varphi(\xi_i^{(p)}))\delta(\xi_i^{(p)})\Delta t_i^{(p)}, [I]) \to 0 \qquad \text{as } p \to \infty,$$

then we say that $[I]$ is the integral of the interval vector function $exp([A]\varphi(t))\delta(t)$ on $[t_0, T]$ and we denote this by

$$[I] = \int_{t_0}^{T} exp([A]\varphi(t))\delta(t)dt.$$

Chapter 3

Main Results

Our goal in this chapter is to show that formula (1.2) holds. The proof is straightforward in the case that $A(t)$ commutes for all $t \in [t_0, T]$ in the sense that $A(t')A(t'') = A(t'')A(t')$ for all $t', t'' \in [t_0, T]$, where $A(t) = \int_0^1 f_y(t, z(t) - s(z(t) - y(t)))ds$, $y(t)$ is the true solution to (1.1) and $z(t)$ is an approximate solution to (1.1). Therefore, using the notation introduced at the start of Chapter 1 and the mean value theorem from §2.4, we see that

$$\begin{aligned} e'(t) &= z'(t) - y'(t) \\ &= (f(t, z(t)) - f(t, y(t))) + \delta(t) \\ &= A(t)(z(t) - y(t)) + \delta(t) \\ &= A(t)e(t) + \delta(t). \end{aligned}$$

If $A(t')A(t'') = A(t'')A(t')$ for all $t', t'' \in [t_0, T]$, then

$$\begin{aligned} e(t_{n+1}) &= exp\left(\int_{t_n}^{t_{n+1}} A(\xi)d\xi\right) e(t_n) + \int_{t_n}^{t_{n+1}} exp\left(\int_s^{t_{n+1}} A(\xi)d\xi\right) \delta(s)ds \\ &\in exp([A](t_{n+1} - t_n))e(t_n) + \int_{t_n}^{t_{n+1}} exp([A](t_{n+1} - s))\delta(s)ds, \end{aligned}$$

where $A(\xi) \in [A]$ for all $\xi \in [t_n, t_{n+1}]$. In particular, if $A(t) = A$ for all $t \in [t_0, T]$, this simplifies to the well-known formula

$$e(t_{n+1}) = exp(A(t_{n+1} - t_n))e(t_n) + \int_{t_n}^{t_{n+1}} exp(A(t_{n+1} - s))\delta(s)ds. \tag{3.1}$$

However, if $A(t')A(t'') \neq A(t'')A(t')$ for some $t', t'' \in [t_n, t_{n+1}]$, it may happen that

$$e(t_{n+1}) \neq exp\left(\int_{t_n}^{t_{n+1}} A(\xi)d\xi\right) e(t_n) + \int_{t_n}^{t_{n+1}} exp\left(\int_{t_n}^{t_{n+1}} A(\xi)d\xi\right) \delta(s)ds.$$

Formula (1.2) is a generalization of (3.1) that holds even if $A(t')A(t'') \neq A(t'')A(t')$ for some $t', t'' \in [t_n, t_{n+1}]$. We prove formula (1.2) using Hasudorff distance and interval arithmetic.

Assume that we have a grid $t_0 < t_1 < \cdots < t_N = T$ on $[t_0, T]$ and that we know $e(0)$, or at least have a bound on it. Then we can use formula (1.2), which we are about to derive, to inductively bound $e(t_{n+1})$ in terms of $e(t_n)$.

To this end, consider the ODE

$$e' = A(t)e + \delta(t), \qquad t \in [t_n, t_{n+1}]. \tag{3.2}$$

To derive our formula (1.2), we consider the application of Euler's method, with stepsize $\frac{h}{l}$, to integrate (3.2) from t_n to t_{n+1}. This numerical integration is used only to develop our bound (1.2) on $e(t_n)$; it is not used in actual computation.

Let $e_{n,0}^{(l)} = e(t_n)$. Then

$$\begin{aligned} e_{n,1}^{(l)} &= e_{n,0}^{(l)} + \frac{h}{l}(A_1 e_{n,0}^{(l)} + \delta_1) \\ &= (I + \frac{h}{l}A_1)e_{n,0}^{(l)} + \frac{h}{l}\delta_1 \end{aligned}$$

where $A_1 = A(t_n)$ and $\delta_1 = \delta(t_n)$. Similarly,

$$\begin{aligned} e_{n,2}^{(l)} &= e_{n,1}^{(l)} + \frac{h}{l}(A_2 e_{n,1}^{(l)} + \delta_2) \\ &= (I + \frac{h}{l}A_2)e_{n,1}^{(l)} + \frac{h}{l}\delta_2 \\ &= (I + \frac{h}{l}A_2)(I + \frac{h}{l}A_1)e_{n,0}^{(l)} + \frac{h}{l}[(I + \frac{h}{l}A_2)\delta_1 + \delta_2] \end{aligned}$$

where $A_2 = A(t_n + \frac{h}{l})$ and $\delta_2 = \delta(t_n + \frac{h}{l})$. Continuing in this way, we see that

$$\begin{aligned} e_{n,l}^{(l)} &= e_{n,l-1}^{(l)} + \frac{h}{l}(A_l e_{n,l-1}^{(l)} + \delta_l) \\ &= (I + \frac{h}{l}A_l)e_{n,l-1}^{(l)} + \frac{h}{l}\delta_l \\ &= (I + \frac{h}{l}A_l)\cdots(I + \frac{h}{l}A_1)e_{n,0}^{(l)} + \frac{h}{l}\sum_{k=1}^{l}[(I + \frac{h}{l}A_l)\cdots(I + \frac{h}{l}A_{k+1})\delta_k] \end{aligned}$$

where $A_l = A(t_n + \frac{(l-1)h}{l})$ and $\delta_l = \delta(t_n + \frac{(l-1)h}{l})$.

Denote

$$\begin{aligned} S_{n,l} &= (I + \frac{h}{l}A_l)\cdots(I + \frac{h}{l}A_1)e(t_n) \\ \sigma_{n,l} &= \frac{h}{l}\sum_{k=1}^{l}[(I + \frac{h}{l}A_l)\cdots(I + \frac{h}{l}A_{k+1})\delta_k]. \end{aligned}$$

Then $e_{n,l}^{(l)} = S_{n,l} + \sigma_{n,l}$. When $l \to \infty$, the stepsize $\frac{h}{l} \to 0$ and $e(t_{n+1}) = \lim_{l\to\infty} e_{n,l}^{(l)}$.

Let

$$A(t) = \begin{pmatrix} a_{11}(t) & a_{12}(t) & \dots & a_{1m}(t) \\ a_{21}(t) & a_{22}(t) & \dots & a_{2m}(t) \\ \dots & \dots & \dots & \dots \\ a_{m1}(t) & a_{m2}(t) & \dots & a_{mm}(t) \end{pmatrix}.$$

Since we assumed the function f associated with the IVP (1.1) is smooth and $A(t) = \int_0^1 f_y(t, z(t) - s(z(t) - y(t)))ds$, each $a_{ij}(t)$ is continuous on $[t_n, t_{n+1}]$. Let

$$\underline{a}_{ij} = \min_{t_n \le t \le t_{n+1}} a_{ij}(t), \qquad \bar{a}_{ij} = \max_{t_n \le t \le t_{n+1}} a_{ij}(t)$$

and define

$$[A] = \begin{pmatrix} [\underline{a}_{11}, \bar{a}_{11}] & [\underline{a}_{12}, \bar{a}_{12}] & \dots & [\underline{a}_{1m}, \bar{a}_{1m}] \\ [\underline{a}_{21}, \bar{a}_{21}] & [\underline{a}_{22}, \bar{a}_{22}] & \dots & [\underline{a}_{2m}, \bar{a}_{2m}] \\ \dots & \dots & \dots & \dots \\ [\underline{a}_{m1}, \bar{a}_{m1}] & [\underline{a}_{m2}, \bar{a}_{m2}] & \dots & [\underline{a}_{mm}, \bar{a}_{mm}] \end{pmatrix}.$$

$[A]$ is a closed convex set in an m^2-dimension linear space,

$$\|[A]\| = \max_{1\le i\le m} \sum_{j=1}^{m} |[a_{ij}]| = a^* < +\infty.$$

To derive formula (1.2), let us first focus on $S_{n,l}$. To this end, let

$$B_n^{(l)} = (I + \frac{h}{l}A_l)\cdots(I + \frac{h}{l}A_1).$$

Since

$$A_k = A(t_n + \frac{(k-1)h}{l}) = \Big(a_{ij}(t_n + \frac{(k-1)h}{l})\Big)_{i,j=1,\dots,m}$$

it follows that

$$\begin{aligned}
\|A_k\| &= \max_{1\le i\le m}\sum_{j=1}^{m}|a_{ij}(t_n+\frac{(k-1)h}{l})| \\
&\le \max_{1\le i\le m}\sum_{j=1}^{m}\max(|\underline{a}_{ij}|,|\overline{a}_{ij}|) \\
&= \max_{1\le i\le m}\sum_{j=1}^{m}|[a_{ij}]| \\
&= a^*
\end{aligned}$$

for $k=1,\ldots,l$. Therefore,

$$\begin{aligned}
\|B_n^{(l)}\| &\le \|I+\frac{h}{l}A_l\|\cdots\|I+\frac{h}{l}A_1\| \\
&\le [1+\frac{h}{l}\|A_l\|]\cdots[1+\frac{h}{l}\|A_1\|] \\
&\le (1+\frac{a^*h}{l})^l \\
&\le e^{a^*h}.
\end{aligned}$$

Since $\{B_n^{(l)}\}_{l=1}^{\infty}$ is a bounded set in m^2-dimension linear space, there exists a subsequence $\{B_n^{(l_k)}\}_{k=1}^{\infty}$ of $\{B_n^{(l)}\}_{l=1}^{\infty}$ such that $B_n^{(l_k)}\to B_n$ as $l_k\to\infty$. Let $S_n=B_ne(t_n)$. Then $S_{n,l_k}=B_n^{(l_k)}e(t_n)\to S_n$ as $l_k\to\infty$.

Now note that

$$\begin{aligned}
B_n^{(l)} &= (I+\frac{h}{l}A_l)\cdots(I+\frac{h}{l}A_1) \\
&= I+\binom{l}{1}\left(\frac{h}{l}\right)C_1+\binom{l}{2}\left(\frac{h}{l}\right)^2C_2+\cdots+\binom{l}{l}\left(\frac{h}{l}\right)^lC_l
\end{aligned}$$

where $C_k=\binom{l}{k}^{-1}\times$ the sum of the $\binom{l}{k}$ products of any k out of the l A_i's in the order they appear in $B_n^{(l)}$.

Since $[A]^k$ is convex and C_k is the convex combination of $\binom{l}{k}$ points in $[A]^k$, $C_k\in[A]^k$. Thus

$$B_n^{(l)} = (I+\frac{h}{l}A_l)\cdots(I+\frac{h}{l}A_1)$$

$$\begin{aligned} &= I + \sum_{k=1}^{l} \binom{l}{k} \left(\frac{h}{l}\right)^k C_k \\ &\in I + \sum_{k=1}^{l} \binom{l}{k} \left(\frac{h}{l}\right)^k [A]^k \\ &= \sum_{k=0}^{l} \binom{l}{k} \left(\frac{h}{l}\right)^k [A]^k. \end{aligned}$$

Since $\sum_{k=0}^{l} \binom{l}{k} (\frac{h}{l})^k [A]^k$ is an interval matrix and

$$\| \sum_{k=0}^{l} \binom{l}{k} (\frac{h}{l})^k [A]^k \| \leq \sum_{k=0}^{l} \binom{l}{k} (\frac{a^* h}{l})^k = (1 + \frac{a^* h}{l})^l \leq e^{a^* h},$$

$\sum_{k=0}^{l} \binom{l}{k} (\frac{h}{l})^k [A]^k$ is a bounded closed subset in m^2-dimension linear space.

Define $exp([A]) = \sum_{k=0}^{\infty} \frac{[A]^k}{k!}$. We want to show that $B_n \in exp(h[A])$. To this end, we prove the following six lemmas.

Lemma 3.1 *If $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ are convergent series, then*

$$\sum_{k=0}^{\infty} [a_k, b_k] = \left[\sum_{k=0}^{\infty} a_k, \sum_{k=0}^{\infty} b_k \right].$$

Proof. Choose any $\xi \in \sum_{k=0}^{\infty} [a_k, b_k]$. Then there exists $\xi_k \in [a_k, b_k]$, $k = 0, 1, 2, \ldots$, such that $\xi = \sum_{k=0}^{\infty} \xi_k$. Since $\sum_{k=0}^{\infty} a_k \leq \sum_{k=0}^{\infty} \xi_k \leq \sum_{k=0}^{\infty} b_k$, $\xi \in [\sum_{k=0}^{\infty} a_k, \sum_{k=0}^{\infty} b_k]$. Therefore,

$$\sum_{k=0}^{\infty} [a_k, b_k] \subset \left[\sum_{k=0}^{\infty} a_k, \sum_{k=0}^{\infty} b_k \right]$$

To show $[\sum a_k, \sum b_k] \subset \sum [a_k, b_k]$,we first prove $\sum_{k=0}^{\infty} [a_k, b_k]$ is convex. To this end, choose any c and $d \in \sum_{k=0}^{\infty} [a_k, b_k]$. Then note that $c = \sum_{k=0}^{\infty} c_k$ and $d = \sum_{k=0}^{\infty} d_k$ for some c_k and $d_k \in [a_k, b_k]$, $k = 0, 1, 2, \ldots$. Now observe that, for any $t \in [0, 1]$, $tc + (1-t)d = \sum_{k=0}^{\infty} [tc_k + (1-t)d_k]$. Since $[a_k, b_k]$ is convex, $tc_k + (1-t)d_k \in [a_k, b_k]$ for $k = 0, 1, 2, \ldots$. Therefore, $tc + (1-t)d \in \sum_{k=0}^{\infty} [a_k, b_k]$. Thus, $\sum_{k=0}^{\infty} [a_k, b_k]$ is convex. Consequently, since $\sum_{k=0}^{\infty} a_k, \sum_{k=0}^{\infty} b_k \in \sum_{k=0}^{\infty} [a_k, b_k]$, any convex combinations of $\sum_{k=0}^{\infty} a_k$ and $\sum_{k=0}^{\infty} b_k$ is in $\sum_{k=0}^{\infty} [a_k, b_k]$. Hence $[\sum_{k=0}^{\infty} a_k, \sum_{k=0}^{\infty} b_k] \subset \sum_{k=0}^{\infty} [a_k, b_k]$.

Therefore,

$$\sum_{k=0}^{\infty}[a_k, b_k] = \left[\sum_{k=0}^{\infty} a_k, \sum_{k=0}^{\infty} b_k\right].$$

□

Lemma 3.2 *$exp([A])$ is closed.*

Proof. It is sufficient to prove $exp([A])$ is an interval matrix. To this end, let $[A]^k = ([\underline{a}_{ij}^{(k)}, \bar{a}_{ij}^{(k)}])_{i,j=1,\ldots,m}$, $k = 0, 1, 2, \ldots$. Since $\|[A]\| = a^*$,

$$\max_{1\le i\le m} \sum_{j=1}^{m} |[\underline{a}_{ij}^{(k)}, \bar{a}_{ij}^{(k)}]| = \|[A]^k\| \le \|[A]\|^k = (a^*)^k$$

for $k = 0, 1, 2, \ldots$ Hence $|[\underline{a}_{ij}^{(k)}, \bar{a}_{ij}^{(k)}]| \le (a^*)^k$ for $i, j = 1, \ldots, m$ and $k = 0, 1, 2, \ldots$. Since $\max(|\underline{a}_{ij}^{(k)}|, |\bar{a}_{ij}^{(k)}|) = |[\underline{a}_{ij}^{(k)}, \bar{a}_{ij}^{(k)}]| \le (a^*)^k$, $|\underline{a}_{ij}^{(k)}| \le (a^*)^k$ and $|\bar{a}_{ij}^{(k)}| \le (a^*)^k$, $k = 0, 1, 2, \ldots$. Therefore

$$\begin{aligned}
\sum_{k=0}^{\infty} \frac{|\underline{a}_{ij}^{(k)}|}{k!} &\le \sum_{k=0}^{\infty} \frac{(a^*)^k}{k!} = e^{a^*} \\
\sum_{k=0}^{\infty} \frac{|\bar{a}_{ij}^{(k)}|}{k!} &\le \sum_{k=0}^{\infty} \frac{(a^*)^k}{k!} = e^{a^*}
\end{aligned}$$

Hence, $\sum_{k=0}^{\infty} \frac{\underline{a}_{ij}^{(k)}}{k!}$ and $\sum_{k=0}^{\infty} \frac{\bar{a}_{ij}^{(k)}}{k!}$ are convergence series for each $i, j = 1, \ldots, m$.

Thus, by Lemma 3.1, $exp([A]) = ([\underline{e}_{ij}, \overline{e}_{ij}])_{i,j=1,2,\ldots,m}$ where $\underline{e}_{ij} = \sum_{k=0}^{\infty} \frac{\underline{a}_{ij}^{(k)}}{k!}$ and $\overline{e}_{ij} = \sum_{k=0}^{\infty} \frac{\bar{a}_{ij}^{(k)}}{k!}$. That is, $exp([A])$ is an interval matrix.

□

Now we give three lemmas on Hausdorff distance.

Lemma 3.3 *If X is a normed linear space and $B \in P_0(X)$, then*

$$H(0, B) \le \sup_{b\in B} \|b\|$$

where 0 denotes the zero element of X.

Proof. For any $\epsilon > 0$, since $0 \subset [\sup_{b\in B} \|b\| + \epsilon] + B$ and $B \subset [\sup_{b\in B} \|b\| + \epsilon] + 0$. Therefore, $H(0, B) \leq \sup_{b\in B} \|b\| + \epsilon$. Since ϵ is arbitrary, $H(0, B) \leq \sup_{b\in B} \|b\|$.

□

Lemma 3.4 *If X is a normed linear space, $A, B_1, B_2 \in P_0(X)$, then*

$$H(A, B_1 + B_2) \leq H(A, B_1) + \sup_{b\in B_2} \|b\|.$$

Proof. By Lemma 2.5 and Lemma 3.3,

$$\begin{aligned} H(A, B_1 + B_2) &= H(A + 0, B_1 + B_2) \\ &\leq H(A, B_1) + H(0, B_2) \\ &\leq H(A, B_1) + \sup_{b\in B_2} \|b\| \end{aligned}$$

□

Lemma 3.5 . *If X is a normed linear space, A is a nonempty convex subset of X and $\lambda \geq \mu > 0$, then $H(\mu A, \lambda A) \leq (\lambda - \mu) \sup_{a\in A} \|a\|$.*

Proof. First, we want to show $\lambda A = \mu A + (\lambda - \mu)A$. Obviously, $\lambda A \subset \mu A + (\lambda - \mu)A$. On the other hand, for any $w \in \mu A + (\lambda - \mu)A$, there exist $a_1, a_2 \in A$ such that $w = \mu a_1 + (\lambda - \mu)a_2 = \lambda[\frac{\mu}{\lambda} a_1 + \frac{\lambda-\mu}{\lambda} a_2]$. Since A is convex, $\frac{\mu}{\lambda} a_1 + \frac{\lambda-\mu}{\lambda} a_2 = v \in A$. Therefore, $w = \lambda v \in \lambda A$. Hence $\lambda A \supset \mu A + (\lambda - \mu)A$. Consequently, $\lambda A = \mu A + (\lambda - \mu)A$.

By Lemma 2.5 and Lemma 3.3,

$$\begin{aligned} H(\mu A, \lambda A) &= H(\mu A + 0, \mu A + (\lambda - \mu)A) \\ &\leq H(\mu A, \mu A) + H(0, (\lambda - \mu)A) \\ &\leq \sup_{a\in A} \|(\lambda - \mu)a\| \\ &= (\lambda - \mu) \sup_{a\in A} \|a\| \end{aligned}$$

□

Theorem 3.1

$$H(\sum_{k=0}^{l}\binom{l}{k}(\frac{h}{l})^k[A]^k, exp([A]h)) \to 0 \qquad as\ l \to \infty.$$

Proof. First note that

$$\frac{1}{k!} - \frac{1}{l^k}\binom{l}{k} = \frac{1}{k!} - \frac{1}{k!}\cdot\frac{l(l-1)\cdots(l-k+1)}{l^k} \geq 0.$$

By Lemma 2.5, Lemma 3.4 and Lemma 3.5, we obtain that

$$\begin{aligned}
& H(\sum_{k=0}^{l}\binom{l}{k}\left(\frac{h}{l}\right)^k[A]^k, exp([A]h)) \\
= \ & H(\sum_{k=0}^{l}\binom{l}{k}\left(\frac{h}{l}\right)^k[A]^k, \sum_{k=0}^{\infty}\frac{h^k}{k!}[A]^k) \\
= \ & H(\sum_{k=0}^{l}\binom{l}{k}\left(\frac{h}{l}\right)^k[A]^k + 0, \sum_{k=0}^{l}\frac{h^k}{k!}[A]^k + \sum_{k=l+1}^{\infty}\frac{h^k}{k!}[A]^k) \\
\leq \ & H(\sum_{k=0}^{l}\frac{1}{l^k}\binom{l}{k}h^k[A]^k, \sum_{k=0}^{l}\frac{1}{k!}h^k[A]^k) + H(0, \sum_{k=l+1}^{\infty}\frac{h^k}{k!}[A]^k) \\
\leq \ & \sum_{k=0}^{l}\left[\frac{1}{k!} - \frac{1}{l^k}\binom{l}{k}\right](a^*h)^k + \sum_{k=l+1}^{\infty}\frac{(a^*h)^k}{k!} \\
= \ & \sum_{k=0}^{\infty}\frac{(a^*h)^k}{k!} - \sum_{k=0}^{l}\binom{l}{k}(\frac{a^*h}{l})^k \\
= \ & e^{a^*h} - (1+\frac{a^*h}{l})^l \\
\to \ & 0 \qquad \text{as } l \to \infty.
\end{aligned}$$

□

Theorem 3.2 $B_n \in exp([A]h)$.

Proof. As noted before, there exists a subsequence $\{B_n^{(l_k)}\}_{k=1}^{\infty}$ of $\{B_n^{(l)}\}_{l=1}^{\infty}$ such that $B_n^{(l_k)} \to B_n$ as $l_k \to \infty$. $B_n^{(l_k)} \in \sum_{i=0}^{l_k}\binom{l_k}{i}(\frac{h}{l_k})^i[A]^i$ and both $\sum_{i=0}^{l_k}\binom{l_k}{i}(\frac{h}{l_k})^i[A]^i$ and $exp([A]h)$ are bounded closed subsets in m^2-dimensional linear space. Therefore, by

Lemma 2.4, there exists $b_n^{(l_k)} \in exp([A]h)$ such that

$$d(B_n^{(l_k)}, b_n^{(l_k)}) \leq H(B_n^{(l_k)}, exp([A]h)) = H(\sum_{i=0}^{l_k} \binom{l_k}{i} \left(\frac{h}{l_k}\right)^i [A]^i, exp([A]h)).$$

Hence,

$$\begin{aligned} d(B_n, exp([A]h)) &\leq d(B_n, b_n^{(l_k)}) \\ &\leq d(B_n, B_n^{(l_k)}) + d(B_n^{(l_k)}, b_n^{(l_k)}) \\ &\leq d(B_n, B_n^{(l_k)}) + H(\sum_{i=0}^{l_k} \binom{l_k}{i} \left(\frac{h}{l_k}\right)^i [A]^i, exp([A]h)) \\ &\to 0 \qquad \text{as } l_k \to \infty. \end{aligned}$$

since, as noted above, $d(B_n, B_n^{(l_k)}) \to 0$ as $l_k \to \infty$ and

$$H(\sum_{i=0}^{l_k} \binom{l_k}{i} (\frac{h}{l_k})^i [A]^i, exp([A]h)) \to 0 \qquad \text{as } l_k \to \infty$$

by Theorem 3.1. Therefore, $d(B_n, exp([A]h)) = 0$. By Lemma 3.2, $exp([A]h)$ is a closed set, whence $B_n \in exp([A]h)$.

□

Corollary 3.1 $S_n \in exp([A]h)e(t_n)$.

Proof. Recall $S_n = B_n e(t_n)$. Since $B_n \in exp([A]h)$, $S_n \in exp([A]h)e(t_n)$.

□

Now, consider $\sigma_{n,l}$. First, evaluate

$$H(\sum_{i=0}^{l-k} \binom{l-k}{i} (\frac{h}{l})^i [A]^i, exp([A](h - \frac{k}{l}h))).$$

To this end, let $D_{n,l-k}^{(l)} = (I + \frac{h}{l}A_l) \cdots (I + \frac{h}{l}A_{k+1})$. By an argument similar to that used above for B_n^l, we can get

$$D_{n,l-k}^{(l)} \in \sum_{i=0}^{l-k} \binom{l-k}{i} \left(\frac{h}{l}\right)^i [A]^i.$$

Note that

$$\frac{1}{i!}\left(\frac{l-k}{l}\right)^i - \frac{1}{l^i}\binom{l-k}{i} = \frac{1}{i!}\frac{(l-k)^i - (l-k)\cdots(l-k-i+1)}{l^i} \geq 0$$

By Lemma 2.5, Lemma 3.4 and Lemma 3.5, we obtain that

$$\begin{aligned}
& H(\sum_{i=0}^{l-k}\binom{l-k}{i}\left(\frac{h}{l}\right)^i [A]^i, exp([A](h-\frac{k}{l}h))) \\
= \; & H(\sum_{i=0}^{l-k}\binom{l-k}{i}\left(\frac{h}{l}\right)^i [A]^i, \sum_{i=0}^{\infty}\frac{1}{i!}\left(\frac{l-k}{l}\right)^i h^i[A]^i) \\
= \; & H(\sum_{i=0}^{l-k}\binom{l-k}{i}\left(\frac{h}{l}\right)^i [A]^i + 0, \sum_{i=0}^{l-k}\frac{1}{i!}\left(\frac{l-k}{l}\right)^i h^i[A]^i + \sum_{i=l-k+1}^{\infty}\frac{1}{i!}\left(\frac{l-k}{l}\right)^i h^i[A]^i) \\
\leq \; & H(\sum_{i=0}^{l-k}\frac{1}{l^i}\binom{l-k}{i}h^i[A]^i, \sum_{i=0}^{l-k}\frac{1}{i!}\left(\frac{l-k}{l}\right)^i h^i[A]^i) + H(0, \sum_{i=l-k+1}^{\infty}\frac{1}{i!}\left(\frac{l-k}{l}\right)^i h^i[A]^i) \\
\leq \; & \sum_{i=0}^{l-k}[\frac{1}{i!}\left(\frac{l-k}{l}\right)^i - \frac{1}{l^i}\binom{l-k}{i}](a^*h)^i + \sum_{i=l-k+1}^{\infty}\frac{1}{i!}\left(\frac{l-k}{l}\right)^i (a^*h)^i \\
= \; & \sum_{i=0}^{\infty}\frac{1}{i!}\left(\frac{l-k}{l}\right)^i (a^*h)^i - \sum_{i=0}^{l-k}\binom{l-k}{i}(\frac{a^*h}{l})^i \\
= \; & e^{\frac{l-k}{l}a^*h} - (1+\frac{a^*h}{l})^{l-k}. \qquad (3.3)
\end{aligned}$$

Next, we will show that

$$e^{\frac{l-k}{l}a^*h} - (1+\frac{a^*h}{l})^{l-k} \leq e^{a^*h} - (1+\frac{a^*h}{l})^l = O(\frac{1}{l}), \qquad k = 1, 2, \ldots, l. \qquad (3.4)$$

To this end, define $f(x) = e^{a^*h-x} - (1+\frac{a^*h}{l})^{l(1-\frac{x}{a^*h})}$, and note that $f(x)$ is continuous and differentiable on $[0, a^*h]$, with

$$f'(x) = -e^{a^*h-x} + (1+\frac{a^*h}{l})^{l(1-\frac{x}{a^*h})}\ln(1+\frac{a^*h}{l})\frac{l}{a^*h}.$$

Since $(1+\frac{a^*h}{l})^{\frac{l}{a^*h}}$ is increasing with l and converge to e as $l \to \infty$,

$$(1+\frac{a^*h}{l})^{l(1-\frac{x}{a^*h})} = [(1+\frac{a^*h}{l})^{\frac{l}{a^*h}}]^{a^*h-x} \leq e^{a^*h-x}.$$

By the mean value theorem,

$$\ln(1+\frac{a^*h}{l}) = \ln(\frac{l+a^*h}{l}) = \ln(l+a^*h) - \ln l = \frac{a^*h}{\xi}$$

where $l \le \xi \le l + a^*h$. Hence,

$$(1+\frac{a^*h}{l})^{l(1-\frac{x}{a^*h})}\ln(1+\frac{a^*h}{l})\frac{l}{a^*h} \le e^{a^*h-x}\cdot\frac{a^*h}{\xi}\cdot\frac{l}{a^*h} \le e^{a^*h-x}.$$

Consequently $f'(x) \le 0$. Thus $f(x)$ is a nonincreasing function on $[0, a^*h]$. Therefore, for any $x \in [0, a^*h]$,

$$f(x) \le f(0) = e^{a^*h} - (1+\frac{a^*h}{l})^l.$$

In particular, for any $k = 1, \ldots, l$, $\frac{ka^*h}{l} \in [0, a^*h]$, hence

$$f(\frac{k}{l}a^*h) = e^{\frac{l-k}{l}a^*h} - (1+\frac{a^*h}{l})^{l-k} \le e^{a^*h} - (1+\frac{a^*h}{l})^l.$$

Now, we only need to show $e^{a^*h} - (1+\frac{a^*h}{l})^l = O(\frac{1}{l})$. To this end, let $\frac{a^*h}{l} = u$, and note that $u \to 0$ as $l \to \infty$. Thus,

$$e^{a^*h} - (1+\frac{a^*h}{l})^l = e^{a^*h} - g(u)$$

where $g(u) = (1+u)^{\frac{a^*h}{u}}$. Using L'Hospital rule, we see that

$$\lim_{u\to 0}\frac{e^{a^*h}-g(u)}{u} = -\lim_{u\to 0} g'(u).$$

Since $g(u) = (1+u)^{\frac{a^*h}{u}}$, $\ln g(u) = \frac{a^*h}{u}\ln(1+u)$. Therefore,

$$\begin{aligned}\frac{g'(u)}{g(u)} &= \frac{a^*h[\frac{u}{1+u} - \ln(1+u)]}{u^2}\\ &= \frac{a^*h[u(1-u+u^2-\cdots) - (u - \frac{u^2}{2} + \frac{u^3}{3} - \cdots)]}{u^2}\\ &= a^*h[-\frac{1}{2} + \frac{2}{3}u - \cdots].\end{aligned}$$

Thus,

$$g'(u) = (1+u)^{\frac{a^*h}{u}}a^*h[-\frac{1}{2} + \frac{2}{3}u - \cdots].$$

Consequently,

$$\lim_{u\to 0}\frac{e^{a^*h}-g(u)}{u} = -\lim_{u\to 0} g'(u) = \frac{a^*h}{2}e^{a^*h}.$$

Recalling $u = \frac{a^*h}{l}$, we see that

$$\lim_{l\to\infty}\frac{e^{a^*h} - (1+\frac{a^*h}{l})^l}{\frac{1}{l}} = \frac{a^2h^2}{2}e^{a^*h}.$$

Thus,

$$e^{a^*h} - (1 + \frac{a^*h}{l})^l = O(\frac{1}{l}).$$

Combining this with (3.3), (3.4) and Lemma 2.4, we see that

$$\sum_{i=0}^{l-k} \binom{l-k}{i} \left(\frac{h}{l}\right)^i [A]^i \subset exp([A](h - \frac{k}{l}h)) + u_k(l) \qquad k = 1, \ldots, l$$

where $u_k(l) \subset R^{m\times m}$ and $\sup_{u\in u_k(l)} \|u\| \le e^{a^*h} - (1 + \frac{a^*h}{l})^l = O(\frac{1}{l})$, $k = 1, \ldots, l$.

On the other hand,

$$\begin{aligned} \delta_k &= \delta(t_n + \frac{k-1}{l}h) \\ &= \delta(t_n + \frac{kh}{l} - \frac{h}{l}) \\ &= \delta(t_n + \frac{kh}{l}) + v_k(l) \end{aligned}$$

where $v_k(l) \subset R^m$ and $\sup_{v\in v_k(l)} \|v\| = O(\frac{1}{l})$, $k = 1, \ldots, l$. Thus,

$$\begin{aligned} \sigma_{n,l} &= \frac{h}{l}\sum_{k=1}^{l}[(I + \frac{h}{l}A_l)\cdots(I + \frac{h}{l}A_{k+1})\delta_k] \\ &\in \frac{h}{l}\sum_{k=1}^{l}\left[\sum_{i=0}^{l-k}\binom{l-k}{i}\left(\frac{h}{l}\right)^i [A]^i\right]\delta_k \\ &\subset \frac{h}{l}\sum_{k=1}^{l}\left[exp([A](h - \frac{k}{l}h)) + u_k(l)\right]\left[\delta(t_n + \frac{kh}{l}) + v_k(l)\right] \\ &\subset \frac{h}{l}\sum_{k=1}^{l}\left[exp([A](h - \frac{k}{l}h))\delta(t_n + \frac{kh}{l})\right] + r(l) \end{aligned}$$

where $r(l) \subset R^m$ and $\sup_{r\in r(l)} \|r\| = O(\frac{1}{l})$.

Let $l \to \infty$, and by the definition of the integral, we get

$$\begin{aligned} & \frac{h}{l}\sum_{k=1}^{l}[exp([A](h - \frac{k}{l}h))\delta(t_n + \frac{kh}{l})] + r(l) \\ \to & \int_0^h exp([A](h-u))\delta(t_n + u)du \quad \text{as } l \to \infty \\ = & \int_{t_n}^{t_{n+1}} exp([A](t_{n+1} - s))\delta(s)ds. \end{aligned}$$

Lemma 3.6 *$\int_{t_n}^{t_{n+1}} exp([A](t_{n+1}-s))\delta(s)ds$ is a bounded closed set in m-dimensional linear space.*

Proof. First, note that

$$exp([A](t_{n+1}-s)) = \sum_{k=0}^{\infty} \frac{(t_{n+1}-s)^k}{k!}[A]^k.$$

Therefore,

$$\begin{aligned} \|exp([A](t_{n+1}-s))\| &\leq \sum_{k=0}^{\infty} \frac{(t_{n+1}-s)^k}{k!}\|[A]^k\| \\ &\leq \sum_{k=0}^{\infty} \frac{(t_{n+1}-s)^k}{k!}(a^*)^k \\ &= e^{a^*(t_{n+1}-s)}. \end{aligned}$$

Hence,

$$\begin{aligned} \| \int_{t_n}^{t_{n+1}} exp([A](t_{n+1}-s))\delta(s)ds\| &\leq \int_{t_n}^{t_{n+1}} \|exp([A](t_{n+1}-s))\|\|\delta(s)\|ds \\ &\leq M \int_{t_n}^{t_{n+1}} e^{a^*(t_{n+1}-s)}ds \end{aligned}$$

where $M = \max_{t_n \leq s \leq t_{n+1}} \|\delta(s)\|$. For $a^* > 0$,

$$\int_{t_n}^{t_{n+1}} e^{a^*(t_{n+1}-s)}ds = \int_0^h e^{a^*t}dt = \frac{e^{a^*t}}{a^*}\Big|_{t=0}^{t=h} = \frac{1}{a^*}(e^{a^*h}-1)$$

and, for $a^* = 0$,

$$\int_{t_n}^{t_{n+1}} e^{a^*(t_{n+1}-s)}ds = \int_{t_n}^{t_{n+1}} 1ds = h.$$

Therefore, $\int_{t_n}^{t_{n+1}} exp([A](t_{n+1}-s))\delta(s)ds$ is bounded.

Next, from the definition of the integral, $\int_{t_n}^{t_{n+1}} exp([A](t_{n+1}-s)\delta(s)ds$ is an interval vector. It must be a closed set in $m-$dimension linear space.

□

Theorem 3.3

$$e(t_{n+1}) \in exp([A](t_{n+1}-t_n))e(t_n) + \int_{t_n}^{t_{n+1}} exp([A](t_{n+1}-s))\delta(s)ds.$$

Proof. Since $\sigma_{n,l} = \frac{h}{l}\sum_{k=1}^{l}[(I + \frac{h}{l}A_l)\cdots(I + \frac{h}{l}A_{k+1})\delta_k]$,

$$\begin{aligned}
\|\sigma_{n,l}\| &\leq \frac{h}{l}\sum_{k=1}^{l}(1 + \frac{h}{l}a^*)^{l-k}M \qquad (\text{ where } M = \max_{t_n \leq t \leq t_{n+1}} \|\delta_k(t)\|) \\
&\leq \frac{h}{l}l(1 + \frac{h}{l}a^*)^l M \\
&\leq Mhe^{a^*h}.
\end{aligned}$$

Therefore, $\{\sigma_{n,l}\}_{l=1}^{\infty}$ is a bounded set in m-dimension linear space. From the subsequence l_k, $k = 1, 2, \ldots$, that we used in the analysis of S_n, we can choose a subsubsequence $l_{\bar{k}}$, $\bar{k} = 1, 2, \ldots$ such that there exists a convergent subsequence of $\{\sigma_{n,l_{\bar{k}}}\}_{\bar{k}=1}^{\infty}$ of $\{\sigma_{n,l}\}_{l=1}^{\infty}$. To simplify the notation we still denote this as $\{\sigma_{n,l_k}\}_{k=1}^{\infty}$. Thus, there exists a σ_n such that $\sigma_{n,l_k} \to \sigma_n$ as $l_k \to \infty$.

As in the proof of Theorem 3.2 and Corollary 3.1, note that $\int_{t_n}^{t_{n+1}} exp([A](t_{n+1} - s))\delta(s)ds$ is closed (Lemma 3.6). Therefore, as in the proof above for S_n, we can show that

$$\sigma_n \in \int_{t_n}^{t_{n+1}} exp([A](t_{n+1} - s))\delta(s)ds. \tag{3.5}$$

Now recall that by Corollary 3.1 and equation (3.5)

$$\begin{aligned}
S_{n,l_k} &\to S_n \in exp([A](t_{n+1} - s))e(t_n) \quad (l_k \to \infty) \\
\sigma_{n,l_k} &\to \sigma_n \in \int_{t_n}^{t_{n+1}} exp([A](t_{n+1} - s))\delta(s)ds \quad (l_k \to \infty).
\end{aligned}$$

Therefore,

$$\begin{aligned}
e(t_{n+1}) &= \lim_{l_k \to \infty}(S_{n,l_k} + \sigma_{n,l_k}) \\
&= \lim_{l_k \to \infty} S_{n,l_k} + \lim_{l_k \to \infty} \sigma_{n,l_k} \\
&= S_n + \sigma_n \\
&\in exp([A](t_{n+1} - t_n))e(t_n) + \int_{t_n}^{t_{n+1}} exp([A](t_{n+1} - s))\delta(s)ds.
\end{aligned}$$

□

Chapter 4

Comparison to Dahlquist's Results and Neumaier's Results

The following Theorem is due to Dahlquist, see Theorem 1.1 of [5] and the Main Theorem of [11].

Theorem 4.1 *Let $f : [0,T] \times R^m \to R^m$. Let $y(t) : [0,T] \to R^m$ be the unique solution of the initial value problem*

$$y' = f(t,y), \qquad y(0) = y_0, \qquad t \in [0,T], \tag{4.1}$$

and let $z(t) : [0,T] \to R^m$ be an approximate solution to (4.1) in the sense that

$$\begin{aligned} \|e(0)\| &\leq \epsilon_0 \\ \|\delta(t)\| &\leq \rho(t), \qquad \forall t \in [0,T], \end{aligned}$$

where $e(t) = z(t) - y(t)$ and $\delta(t) = z'(t) - f(t,z(t))$. If

$$\mu(f_y(t, z(t) - s(z(t) - y(t)))) \leq c(t), \quad \forall t \in [0,T], \quad \forall s \in [0,1],$$

then, $\forall t \in [0,T]$,

$$\|e(t)\| \leq \epsilon_0 e^{\int_0^t c(s)ds} + e^{\int_0^t c(s)ds} \int_0^t \rho(s) e^{-\int_0^s c(u)du} ds \tag{4.2}$$

We shall prove Theorem 4.1 using the analysis leading up to Theorem 3.3 in Chapter 3. Our motivation for this is not to have another proof of Dahlquist's important result, but rather to show that our approach yields bounds that are as tight or tighter than those that can be obtained by Dahlquist's approach. To this end, recall that

$$\begin{aligned} y' &= f(t, y(t)) \\ z' &= f(t, z(t)) + \delta(t) \\ e(t) &= z(t) - y(t). \end{aligned}$$

Hence,

$$\begin{aligned} e'(t) &= z'(t) - y'(t) \\ &= [f(t, z(t)) - f(t, y(t))] + \delta(t) \\ &= \left[\int_0^1 f_y(t, z(t) - s(z(t) - y(t)))ds\right](z(t) - y(t)) + \delta(t) \\ &= A(t)e(t) + \delta(t) \end{aligned} \tag{4.3}$$

where $A(t) = \int_0^1 f_y(t, z(t) - s(z(t) - y(t)))ds \in R^{m\times m}$ is continuous.

To prove Theorem 4.1, we need three lemmas.

Lemma 4.1 *Under the assumptions of Theorem 4.1,*

$$\mu(A(t)) \le c(t), \qquad \forall t \in [0, T].$$

Proof. Since $A(t) = \int_0^1 f_y(t, z(t) - s(z(t) - y(t)))ds$, it follows from the definition of the integral that

$$A(t) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} f_y(t, z(t) - \frac{i}{n}(z(t) - y(t))).$$

Let $A_n(t) = \frac{1}{n}\sum_{i=0}^{n-1} f_y(t, z(t) - \frac{i}{n}(z(t) - y(t)))$, and note that $A_n(t) \to A(t)$ as $n \to \infty$. Therefore, by Lemma 2.2 in Chapter 2,

$$\mu(A_n(t)) \to \mu(A(t)) \qquad \text{as } n \to \infty. \tag{4.4}$$

Also by parts (2) and (3) of Lemma 2.1 in Chapter 2,

$$\begin{aligned}\mu(A_n(t)) &\leq \frac{1}{n}\sum_{i=0}^{n-1}\mu(f_y(t, z(t) - \frac{i}{n}(z(t) - y(t)))) \\ &\leq \frac{1}{n}\cdot nc(t) = c(t) \qquad (4.5)\end{aligned}$$

for any positive integer n. Combining (4.4) and (4.5), we get that

$$\mu(A(t)) \leq c(t), \qquad \forall t \in [0, T].$$

□

Lemma 4.2 *Under the assumptions of Theorem 4.1, for all $t \in [0, T]$, and for all $\eta > 0$, there is an $\alpha(t) > 0$ such that, if $t' \in [0, T]$ with $|t' - t| < \alpha(t)$ and $0 < h < \alpha(t)$, then*

$$\|I + hA(t')\| < 1 + h\mu(A(t')) + h\eta.$$

Proof. Choose any $t \in [0, T]$ and any $\eta > 0$. Since

$$\lim_{h \to +0} \frac{\|I + hA(t)\| - 1}{h} = \mu(A(t)),$$

there exists an $\alpha_1(t) > 0$ such that, for $0 < h < \alpha_1(t)$,

$$\frac{\|I + hA(t)\| - 1}{h} - \mu(A(t)) < \frac{\eta}{3}.$$

Hence

$$\|I + hA(t)\| < 1 + h\mu(A(t)) + \frac{h\eta}{3}.$$

Also, since $A(t)$ is continuous, there is an $\alpha_2(t) > 0$ such that, if $|t' - t| < \alpha_2(t)$, then

$$\|A(t') - A(t)\| < \frac{\eta}{3}.$$

In addition, by part (4) of Lemma 2.1 in Chapter 2,

$$|\mu(A(t')) - \mu(A(t))| \leq \|A(t') - A(t)\| < \frac{\eta}{3}.$$

Take $\alpha(t) = \min(\alpha_1(t), \alpha_2(t))$. Therefore, for $0 < h < \alpha(t)$ and $|t' - t| < \alpha(t)$,

$$\begin{aligned}
\|I + hA(t')\| &= \|(I + hA(t)) + h(A(t') - A(t))\| \\
&\leq \|I + hA(t)\| + h\|A(t') - A(t)\| \\
&< 1 + h\mu(A(t)) + \frac{h\eta}{3} + \frac{h\eta}{3} \\
&< 1 + h(\mu(A(t')) + \frac{\eta}{3}) + \frac{2h\eta}{3} \\
&= 1 + h\mu(A(t')) + h\eta.
\end{aligned}$$

□

Lemma 4.3 *Under the assumptions of Theorem 4.1, for all $\eta > 0$, there is a $\delta > 0$ such that, if $t \in [0, T]$ and $0 < h < \delta$, then*

$$\|I + hA(t)\| < 1 + hc(t) + h\eta.$$

Proof. Choose any $\eta > 0$ and any $t \in [0, T]$. By Lemma 4.2, there is an $\alpha(t) > 0$ such that, if $t' \in (t - \alpha(t), t + \alpha(t)) \cap [0, T]$ and $0 < h < \alpha(t)$, then

$$\|I + hA(t')\| < 1 + h\mu(A(t')) + h\eta.$$

Since $[0, T]$ is compact and

$$\bigcup_{t \in [0,T]} (t - \delta(t), t + \delta(t)) \supset [0, T],$$

by Borel's open covering theorem, there exists a finite set of points $t_1, \ldots, t_M \in [0, T]$ such that

$$\bigcup_{i=1}^{M} (t_i - \delta(t_i), t_i + \delta(t_i)) \supset [0, T].$$

Let $\delta = \min(\alpha(t_1), \ldots, \alpha(t_M)) > 0$. For any $t \in [0, T]$, there is some $i \in \{1, \ldots, M\}$ such that $t \in (t_i - \alpha(t_i), t_i + \alpha(t_i))$. For $0 < h < \delta \leq \alpha(t_i)$, by Lemma 4.1 and Lemma 4.2, we obtain that

$$\begin{aligned}
\|I + hA(t)\| &< 1 + h\mu(A(t)) + h\eta \\
&\leq 1 + hc(t) + h\eta.
\end{aligned}$$

□

Now, we prove Theorem 4.1 using the analysis leading up to Theorem 3.3 in Chapter 3. Take $t_n = 0, t_{n+1} = t$ and $h = t_{n+1} - t_n$. Then we showed in Chapter 3 that

$$e(t) = S + \sigma$$

where for convenience we have dropped the subscript n from S and σ in the last equation, as we shall do throughout the rest of this Chapter for S, σ and similar expressions from Chapter 3.

Recall that

$$\begin{aligned} S &= Be(0) \\ B &= \lim_{l_k\to\infty} B^{(l_k)} = \lim_{l_k\to\infty} (I + \frac{t}{l_k}A_{l_k}) \cdots (I + \frac{t}{l_k}A_1) \\ & \qquad A_i = A(\frac{i-1}{l_k}t), \qquad i = 1, \ldots, l_k, \end{aligned}$$

and

$$\begin{aligned} \sigma = \lim_{l_k\to\infty} \frac{t}{l_k} \sum_{j=1}^{l_k} [(I + \frac{t}{l_k}A_{l_k}) \cdots (I + \frac{t}{l_k}A_{j+1})\delta_j] \\ \delta_j = \delta(\frac{j-1}{l_k}t), \qquad j = 1, \ldots, l_k. \end{aligned}$$

By Lemma 4.3, for any $\eta > 0$, there is a $\delta > 0$ such that whenever $0 < h < \delta$, then

$$\|I + hA_i\| < 1 + hc(\frac{i-1}{l_k}t) + h\eta, \qquad i = 1, \ldots, l_k.$$

Choose K large enough such that, if $k \geq K$, then

$$0 < \frac{t}{l_k} < \delta.$$

Using the inequality $e^x \geq 1 + x$, for all $t \in [0, T]$ we have

$$\begin{aligned} \|B^{(l_k)}\| &\leq \|I + \frac{t}{l_k}A_{l_k}\| \cdots \|I + \frac{t}{l_k}A_1\| \\ &\leq (1 + \frac{t}{l_k}c(\frac{l_k-1}{l_k}t) + \frac{\eta t}{l_k}) \cdots (1 + \frac{t}{l_k}c(0) + \frac{\eta t}{l_k}) \\ &\leq e^{\frac{t}{l_k}[c(\frac{l_k-1}{l_k}t)+\eta+\cdots+c(0)+\eta]}. \end{aligned}$$

Therefore,

$$\begin{aligned} \|B\| &= \lim_{l_k\to\infty}\|B^{(l_k)}\| \\ &\leq e^{\int_0^t[c(s)+\eta]ds}. \end{aligned} \tag{4.6}$$

Since η is an arbitrary positive constant, (4.6) implies that

$$\|B\| \leq e^{\int_0^t c(s)ds}$$

whence

$$\|S\| = \|Be(0)\| \leq \|B\|\|e(0)\| \leq \epsilon_0 e^{\int_0^t c(s)ds}. \tag{4.7}$$

Next, using the inequality $e^x \geq 1+x$ again, for all $t\in[0,T]$, we have

$$\begin{aligned} &\|\frac{t}{l_k}\sum_{j=1}^{l_k}\left[(I+\frac{t}{l_k}A_{l_k})\cdots(I+\frac{t}{l_k}A_{j+1})\delta_j\right]\| \\ \leq\ & \frac{t}{l_k}\sum_{j=1}^{l_k}\|I+\frac{t}{l_k}A_{l_k})\|\cdots\|(I+\frac{t}{l_k}A_{j+1})\|\rho(\frac{j-1}{l_k}t) \\ \leq\ & \frac{t}{l_k}\sum_{j=1}^{l_k}(1+\frac{t}{l_k}c(\frac{l_k-1}{l_k}t)+\frac{\eta t}{l_k})\cdots(1+\frac{t}{l_k}c(\frac{j}{l_k}t)+\frac{\eta t}{l_k})\rho(\frac{j-1}{l_k}t) \\ \leq\ & \frac{t}{l_k}\sum_{j=1}^{l_k}e^{\frac{t}{l_k}[c(\frac{l_k-1}{l_k}t)+\eta+\cdots+c(\frac{j}{l_k}t)+\eta]}\rho(\frac{j-1}{l_k}t) \\ \leq\ & e^{\frac{t}{l_k}[c(\frac{l_k-1}{l_k}t)+\eta+\cdots+c(0)+\eta]}\frac{t}{l_k}\sum_{j=1}^{l_k}\rho(\frac{j-1}{l_k}t)e^{-\frac{t}{l_k}[c(0)+\eta+\cdots+c(\frac{j-1}{l_k}t)+\eta]}. \end{aligned}$$

Also,

$$\lim_{l_k\to\infty} e^{\frac{t}{l_k}[c(\frac{l_k-1}{l_k}t)+\eta+\cdots+c(0)+\eta]} = e^{\int_0^t[c(s)+\eta]ds}$$

and

$$\lim_{l_k\to\infty}\frac{t}{l_k}\sum_{j=1}^{l_k}\rho(\frac{j-1}{l_k}t)e^{-\frac{t}{l_k}[c(0)+\eta+\cdots+c(\frac{j-1}{l_k}t)+\eta]} = \int_0^t\rho(s)e^{-\int_0^s[c(u)+\eta]du}ds.$$

Therefore

$$\|\sigma\| \leq e^{\int_0^t[c(s)+\eta]ds}\int_0^t\rho(s)e^{-\int_0^s[c(u)+\eta]du}ds. \tag{4.8}$$

Since η is an arbitrary positive constant, (4.8) implies that

$$\|\sigma\| \leq e^{\int_0^t c(s)ds} \int_0^t \rho(s) e^{-\int_0^s c(u)du} ds. \tag{4.9}$$

From (4.7) and (4.9), we obtain that, $\forall t \in [0,T]$,

$$\|e(t)\| \leq \epsilon_0 e^{\int_0^t c(s)ds} + e^{\int_0^t c(s)ds} \int_0^t \rho(s) e^{-\int_0^s c(u)du} ds.$$

□

Corollary 4.1 *Let $f : [0,T] \times R^m \to R^m$. Let $y(t) : [0,T] \to R^m$ be the unique solution of the initial value problem (4.1) and let $z(t) : [0,T] \to R^m$ be an approximate solution to (4.1) in the sense that*

$$\begin{aligned} \|e(0)\| &\leq \epsilon_0 \\ \|\delta(t)\| &\leq \epsilon, \qquad \forall t \in [0,T], \end{aligned}$$

where $e(t) = z(t) - y(t)$ and $\delta(t) = z'(t) - f(t, z(t))$. If

$$\mu(f_y(t, z(t) - s(z(t) - y(t)))) \leq \mu, \quad \forall t \in [0,T], \quad \forall s \in [0,1],$$

then, $\forall t \in [0,T]$,

$$\|e(t)\| \leq \epsilon_0 e^{\mu t} + \epsilon e^{\mu t} \int_0^t e^{-\mu s} ds = \begin{cases} \epsilon_0 e^{\mu t} + \frac{\epsilon}{\mu}(e^{\mu t} - 1) & \text{if } \mu \neq 0 \\ \epsilon_0 + \epsilon t & \text{if } \mu = 0. \end{cases}$$

Neumaier's similar result, Corollary 4.5 in [24], can be summarized as follows.

Theorem 4.2 *Let $f : [0,T] \times R^m \to R^m$ and assume $S \in R^{m \times m}$ is invertible. Let $y(t) : [0,T] \to R^m$ be the unique solution of the initial value problem*

$$y' = f(t,y), \qquad y(0) = y_0, \qquad t \in [0,T], \tag{4.10}$$

and let $z(t) : [0,T] \to R^m$ be an approximate solution to (4.10) in the sense that

$$\begin{aligned} \|S^{-1} e(0)\| &\leq \epsilon_0 \\ \|S^{-1} \delta(t)\| &\leq \epsilon, \qquad \forall t \in [0,T], \end{aligned}$$

where $e(t) = z(t) - y(t)$ *and* $\delta(t) = z'(t) - f(t, z(t))$. *If*

$$\mu(S^{-1} f_y(t, y) S) \leq \mu, \quad \forall t \in [0, T], \quad \forall y \in R^m,$$

then, $\forall t \in [0, T]$,

$$\|S^{-1} e(t)\| \leq \begin{cases} \epsilon_0 e^{\mu t} + \frac{\epsilon}{\mu}(e^{\mu t} - 1) & \text{if } \mu \neq 0 \\ \epsilon_0 + \epsilon t & \text{if } \mu = 0. \end{cases} \tag{4.11}$$

For $S = I$, the above result simplifies as follows. If

$$\begin{aligned} \|e(0)\| &\leq \epsilon_0 \\ \|\delta(t)\| &\leq \epsilon, \qquad \forall t \in [0, T], \end{aligned}$$

$$\mu(f_y(t, y)) \leq \mu, \quad \forall t \in [0, T], \quad \forall y \in R^m,$$

then, $\forall t \in [0, T]$,

$$\|e(t)\| \leq \begin{cases} \epsilon_0 e^{\mu t} + \frac{\epsilon}{\mu}(e^{\mu t} - 1) & \text{if } \mu \neq 0 \\ \epsilon_0 + \epsilon t & \text{if } \mu = 0 \end{cases} \tag{4.12}$$

This result follows immediately from Corollary 4.1.

Now, we prove Theorem 4.2. Since S is an invertible matrix, we may define $u = S^{-1} y$, or equivalently $y = Su$. Substituting this change of variables into (4.10), we get

$$Su' = f(t, Su)$$

or equivalently

$$u' = S^{-1} f(t, Su) = F(t, u)$$

giving rise to the IVP

$$u' = F(t, u), \qquad u(0) = S^{-1} y(0) = S^{-1} y_0. \tag{4.13}$$

Note that $y(t)$ is the true solution of (4.10) if and only if $u(t) = S^{-1} y(t)$ is the true solution of (4.13).

Let $z(t)$ be an approximate solution of (4.10). Then

$$z' = f(t,z) + \delta(t)$$

or equivalently

$$\delta(t) = z'(t) - f(t,z(t)).$$

Let $v = S^{-1}z$, or equivalently $z = Sv$. Hence,

$$Sv' = f(t,Sv) + \delta(t)$$

or equivalently

$$\begin{aligned} v' &= S^{-1}f(t,Sv) + S^{-1}\delta(t) \\ &= F(t,v) + \Delta(t) \end{aligned}$$

where $\Delta(t) = S^{-1}\delta(t)$ is the defect associated with the approximate solution $v(t)$ to the IVP (4.13). Let $E(t) = v(t) - u(t) = S^{-1}(z(t) - y(t))$. Then

$$\begin{aligned} E'(t) &= v'(t) - u'(t) \\ &= [F(t,v(t)) - F(t,u(t))] + \Delta(t) \\ &= \left[\int_0^1 F_u(t, v(t) - s(v(t) - u(t)))ds\right](v(t) - u(t)) + \Delta(t) \\ &= B(t)E(t) + \Delta(t) \end{aligned}$$

where $B(t) = \int_0^1 F_u(t, v(t) - s(v(t) - u(t)))ds$ is a matrix. Applying our previous analysis to the ODE

$$E'(t) = B(t)E(t) + \Delta(t)$$

with

$$\begin{aligned} \|E(0)\| &= \|S^{-1}e(0)\| \le \epsilon_0 \\ \|\Delta(t)\| &= \|S^{-1}\delta(t)\| \le \epsilon, \qquad \forall t \in [0,T], \end{aligned}$$

and

$$\mu(F_u(t,u)) = \mu(S^{-1}f_y(t,y)S) \le \mu, \quad \forall t \in [0,T], \quad \forall y \in R^m,$$

we get that

$$\|E(t)\| = \|S^{-1}e(t)\| \leq \begin{cases} \epsilon_0 e^{\mu t} + \frac{\epsilon}{\mu}(e^{\mu t} - 1) & \text{if } \mu \neq 0 \\ \epsilon_0 + \epsilon t & \text{if } \mu = 0. \end{cases}$$

□

We have just shown that formula (1.2) always yields bounds that are as tight as those produced by formulas (1.3) or (1.4). On the other hand, the following simple example shows that formula (1.2) may sometimes yield tighter bounds than those produced by formulas (1.3) or (1.4).

Example. Consider the problem

$$\begin{aligned} y' &= Ay \\ y(t_0) &= y_0, \qquad t_0 \in [0, T], \end{aligned}$$

where A is an $m \times m$ constant matrix. Applying the Backward Euler formula to this problem, we get

$$y_1 = y_0 + hAy_1$$

$$y_1 = (I - hA)^{-1}y_0.$$

Let the approximate solution $z(t)$ be the line which passes through the points $(0, y_0)$ and (h, y_1):

$$z(t) = y_0 + \frac{t}{h}(y_1 - y_0) = y_0 + tAy_1 = y_0 + tA(I - hA)^{-1}y_0.$$

The defect associated with this approximate solution is

$$\begin{aligned} \delta(t) &= z'(t) - Az(t) \\ &= A(I - hA)^{-1}y_0 - Ay_0 - tA^2(I - hA)^{-1}y_0 \\ &= [A - A(I - hA) - tA^2](I - hA)^{-1}y_0 \\ &= (h - t)A^2(I - hA)^{-1}y_0. \end{aligned} \tag{4.14}$$

Now apply formula (1.2) with $t_n = 0$, $t_{n+1} = h$ and $[A] = A$, replacing the $\in$ by $=$, since with $[A] = A$ the right side of (1.2) is a simple vector rather than an interval vector:

$$\begin{aligned}
e(h) &= e^{Ah}e(0) + \int_0^h e^{A(h-s)}\delta(s)ds \\
&= e^{Ah}e(0) + \int_0^h e^{A(h-s)}(h-s)A^2(I-hA)^{-1}y_0 ds \\
&= e^{Ah}e(0) + \int_0^h e^{Au}uA^2(I-hA)^{-1}y_0 du \qquad (\text{ let } h-s=u) \\
&= e^{hA}e(0) + \left(\int_0^h uA^2e^{Au}du\right)(I-hA)^{-1}y_0 \\
&= e^{hA}e(0) + \left(\left[uAe^{Au}\right]_0^h - \int_0^h Ae^{Au}du\right)(I-hA)^{-1}y_0 \\
&= e^{hA}e(0) + \left(hAe^{hA} - \left[e^{Au}\right]_0^h\right)(I-hA)^{-1}y_0 \\
&= e^{hA}e(0) + \left(hAe^{hA} - (e^{hA}-I)\right)(I-hA)^{-1}y_0 \\
&= e^{hA}e(0) + \left(I - e^{hA}(I-hA)\right)(I-hA)^{-1}y_0 \\
&= e^{Ah}e(0) + [(I-hA)^{-1} - e^{Ah}]y_0 \qquad (4.15)
\end{aligned}$$

where we have used the fact that A, e^{Au} and $(I-hA)^{-1}$ commute.

Now we compare the bounds (1.3) and (1.4) with bounds that can be derived from (4.15). To be more concrete, let

$$A = \begin{pmatrix} -2 & 1 \\ 0 & -2 \end{pmatrix}, \qquad t_0 = 0, \qquad T = 4, \qquad \|e(0)\| \le \epsilon_0, \qquad \|y_0\| \le \epsilon_0'.$$

Then $\mu = \mu_\infty(A) = -1$ and

$$e^{Ah} = e^{-2h}\begin{pmatrix} 1 & h \\ 0 & 1 \end{pmatrix}.$$

Consider the bound (1.3) first. From (4.14), we see that

$$\|\delta(t)\| \le h\|A^2(I-hA)^{-1}\|\epsilon_0' \equiv \epsilon.$$

Applying bound (1.3), we get

$$\|e(h)\| \le \epsilon_0 e^{-h} + c_1\epsilon_0'$$

where

$$c_1 = h(1 - e^{-h})\|A^2(I - hA)^{-1}\|$$

Next consider the bound (1.4). From (4.14), we see that

$$\|\delta(t)\| \leq (h - t)\|A^2(I - hA)^{-1}\|\epsilon_0' \equiv \rho(t)$$

Applying bound (1.4), we get

$$\begin{aligned} \|e(h)\| &\leq \epsilon_0 e^{-h} + \int_0^h e^{-(h-s)}(h - s)\|A^2(I - hA)^{-1}\|\epsilon_0' ds \\ &= \epsilon_0 e^{-h} + c_2\epsilon_0' \end{aligned}$$

where

$$c_2 = (1 - e^{-h} - he^{-h})\|A^2(I - hA)^{-1}\|$$

Finally, for (1.2), note that the associated equation (4.15) results in the bound

$$\begin{aligned} \|e(h)\| &\leq \epsilon_0 e^{-h} + \|(I - hA)^{-1} - e^{Ah}\|\epsilon_0' \\ &= \epsilon_0 e^{-h} + c_3\epsilon_0' \end{aligned}$$

where

$$c_3 = \|(I - hA)^{-1} - e^{Ah}\|.$$

Now we compare the bounds (1.3), (1.4) and (1.2) for this example by comparing the size of the constants c_1, c_2 and c_3. To this end, note that

$$\begin{aligned} A^2 &= 4\begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \\ I - hA &= \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - \begin{pmatrix} -2h & h \\ 0 & -2h \end{pmatrix} = \begin{pmatrix} 1 + 2h & -h \\ 0 & 1 + 2h \end{pmatrix} \\ (I - hA)^{-1} &= \frac{1}{(1 + 2h)^2}\begin{pmatrix} 1 + 2h & h \\ 0 & 1 + 2h \end{pmatrix} \end{aligned}$$

$$
\begin{aligned}
A^2(I-hA)^{-1} &= \frac{4}{(1+2h)^2}\begin{pmatrix} 1+2h & -1-h \\ 0 & 1+2h \end{pmatrix} \\
\|A^2(I-hA)^{-1}\| &= \frac{4}{(1+2h)^2}(3h+2) \qquad (\text{ in the infinite norm, assuming } h \in [0,4].)
\end{aligned}
$$

Thus,

$$
\begin{aligned}
c_1 &= \frac{4h(1-e^{-h})}{(1+2h)^2}(3h+2) \\
c_2 &= \frac{4(1-e^{-h}-he^{-h})}{(1+2h)^2}(3h+2) \\
c_3 &= \left\| \begin{pmatrix} \frac{1}{1+2h} & \frac{h}{(1+2h)^2} \\ 0 & \frac{1}{1+2h} \end{pmatrix} - \begin{pmatrix} e^{-2h} & he^{-2h} \\ 0 & e^{-2h} \end{pmatrix} \right\| \\
&= \left\| \begin{pmatrix} \frac{1}{1+2h} - e^{-2h} & \frac{h}{(1+2h)^2} - he^{-2h} \\ 0 & \frac{1}{1+2h} - e^{-2h} \end{pmatrix} \right\| \\
&= |\frac{1}{1+2h} - e^{-2h}| + |\frac{h}{(1+2h)^2} - he^{-2h}| \qquad (\text{ in the infinite norm }).
\end{aligned}
$$

The numerical results for c_1, c_2 and c_3 are as follows for $h = 1, 2, 4$.

	$h=1$	$h=2$	$h=4$
c_1	1.4047	2.2135	2.7148
c_2	0.5872	0.7603	0.6280
c_3	0.2222	0.2251	0.1588

Since $c_1 > c_2 > c_3$, this example shows that our formula (1.2) sometimes produces tighter bounds than (1.3) and (1.4).

Chapter 5

Conclusions and Future Work

We reviewed interval arithmetic, logarithmic norms and Hausdorff distance in Chapter 2. We derived a formula in Chapter 3 for bounding the global error associated with the numerical solution of an IVP for an ODE.

Most importantly, we believe this formula can be applied to stiff IVPs without requiring that the stepsize be severely restricted. Therefore, we believe that this approach may lead to an effective validated numerical method for stiff problems.

We compared our results to Dahlquist's and Neumaier's in Chapter 4 and derive their formulas from ours, thus showing that our new formula always produces bounds that are as tight as theirs. Moreover, we gave an example that shows that our new formula sometimes produces bounds that are tighter than theirs.

As noted in Chapter 1, our original goal was to use interval arithmetic to compute the right side of (1.2) directly, in the hope that this might produce an effective validated method for stiff IVPs for ODEs. However, this has proven more difficult than we originally expected and so we leave this task to future work.

Bibliography

[1] G.Alefeld and J.Herzberger. *Introduction to Interval Computations*. Academic Press, New York, 1983.

[2] V.I.Arnold. *Ordinary Differential Equations*. The Massachusetts Institute of Technology, Cambridge, 1973.

[3] W.A.Coppel. *Stability and Asymptotic Behavior of Differential Equations*. Heath, Boston, 1965.

[4] Robert M.Corless and George F.Corliss. Rationale for guaranteed ODE defect control. In L.Atanassova and J.Herzberger editors, in *Computer Arithmetic and Enclosure Methods*, Pages 3-12, Elsevier Science Publishers B.V. (North-Holland) 1992 IMACS.

[5] G.Dahlquist. Stability and error bounds in the numerical intergration of ordinary differential equations. *Diss.* 1958; reprinted in *Trans. of the Royal Inst. Tech., Stockholm, Sweden*, No. 130, 1959.

[6] G.Dahlquist. On matrix majorants and minorants with application to differential equations. *Linear ALgebra and its Applicartions*, 52/53, 1983, 199-216.

[7] Charles A.Desoer and Hiromasa Haneda. The measure of a matrix as a tool to analyze computer algorithms for circuit analysis. *IEEE Trans. Circuit Theory*, 19(1972).

[8] P.Eijgenraam. The solution of initial value problems using interval arithmetic. *Math. Centre Tracts* 144, Amsterdam 1981.

[9] W.H.Enright, T.E.Hull and B.Lindberg. Computing numerical methods for stiff systems of ODEs. *BIT*, 15(1975), 10-48.

[10] W.H.Enright. A new error control for initial value solvers. *Appl.Math.Comp.*, 31(1989), 288-301.

[11] W.H.Enright. Course notes of the Numerical Solution of ODEs, 2003.

[12] E.Hairer and G.Wanner. *Solving Ordinary Equations II: Stiff and Differential-Algebraic Problems*. Springer, Berlin, 1991.

[13] J.K.Hale. *Ordinary Differential Equations*. John Wiley & Sons, New York, 1969.

[14] John L.Kelly. *General Topology*. Van Nostrand, New York, 1955.

[15] Erwin Klein and Anthony C.Thompson. *Theory of Correspondence, Including Applications to Mathematical Economics*. John Wiley & Sons, New York, 1984.

[16] Rudolph J.Lohner. Enclosing the solutions of ordinary initial and boundary value problems. In Edgar W.Kaucher, Ulrich W.Kulisch, and Christian Ullrich, editors, *Computer Arithmetic: Scientific Computation and Programming Languages*, pages 255-286. Wiley-Teubner Series in Computer Science, Stuttgart, 1987.

[17] S.M.Loziinskij. Error estimate for numerical integration of ordinary differential equations. *Part I. Izv. Vyss. Ucehn. Zaved. Mathematika*, 6(1958), 52-90.

[18] Ramon E.Moore. *Interval Analysis*. Prentice-Hall, Englewood Cliffs, N.J., 1966.

[19] N. S. Nedialkov, K. R. Jackson and G. F. Corliss. Validated solutions of initial value problems for ordinary differential Equations. *Appl.Math.Comp.*, 105 (1999), 21-68.

[20] N. S. Nedialkov. *Computing rigorous bounds on the solution of an initial value problem for an ordinary differential equation*. PhD Thesis. Department of Computer Science, University of Toronto, 1999.

[21] N. S. Nedialkov and K. R. Jackson. A New perspective on the wrapping effect in interval methods for initial value problems for ordinary differential equations. *Perspectives on Enclosure Methods*, Ulrich Kulisch, Rudolf Lohner and Axel Facius, editors, Springer-Verlag, 2001, 219-264.

[22] N. S. Nedialkov and K. R. Jackson. Some recent advances in validated methods for IVPs for ODEs. *Proceedings of the International Minisymposium on Mathematical Modelling and Scientific Computations*, 8-11 April 2001, Borovets, Bulgaria.

[23] Arnold Neumaier. *Interval Methods for Systems of Equations*. Cambrige University Press, Cambrige, 1990.

[24] Aronld Neumaier. Global, rigorous and realistic bounds for the solutions of dissipative differential equations, Part I: Theory. *Computing*, 52(4), 1994, 315-336.

[25] J.M.Ortega and W.C.Rheinboldt. *Iterative Solution of Nonlinear Equations in Several Variables*. Academic Press, Inc, New York, 1970.

[26] C.V.Pao. Logarithmic derivatives of a square matrix, *J. Linear Algebra and Appl.*, 6(1973), 159–164.

[27] W.L.Seward. *Defect and local error control in codes for solving stiff initial-value problems*. PhD Thesis. Department of Computer Science, University of Toronto, 1985.

[28] L.F.Shampine and M.K.Gordon. *Computer Solution of Ordinary Differential Equations, (The Initial Value Peoblem)*. W.H.Freeman and Company, 1975.

[29] Lawrence F. Shampine. *Numerical Solutions of Ordinary Differential Equations*. Chapman & Hall, New York, 1994.

[30] Torsten Strom. On logarithmic norms. *SIAM Journal on Numerical Analysis*, Volume 12, Issue 5 (Oct., 1975), 741-753.