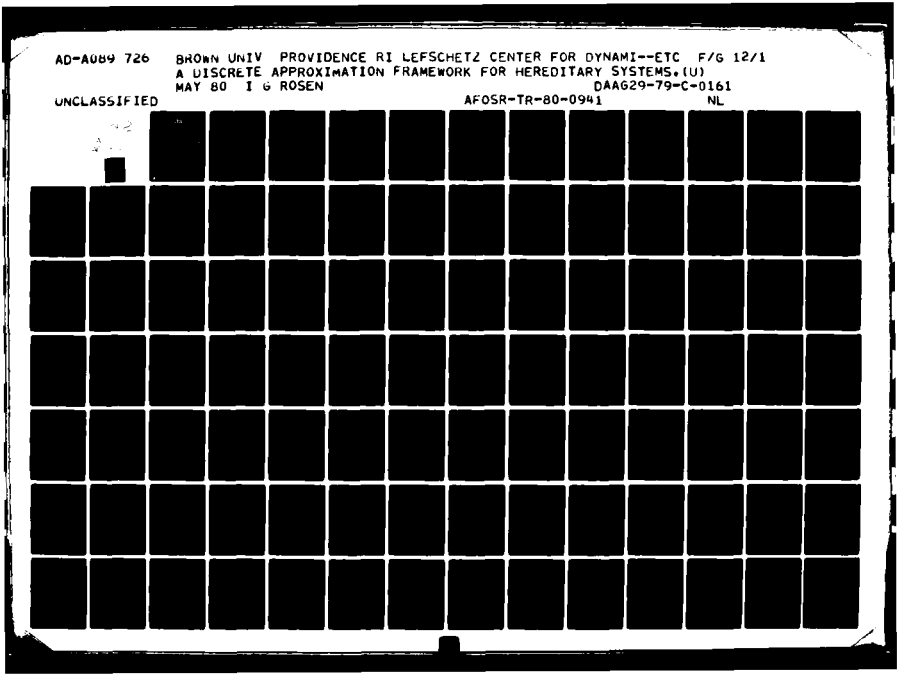
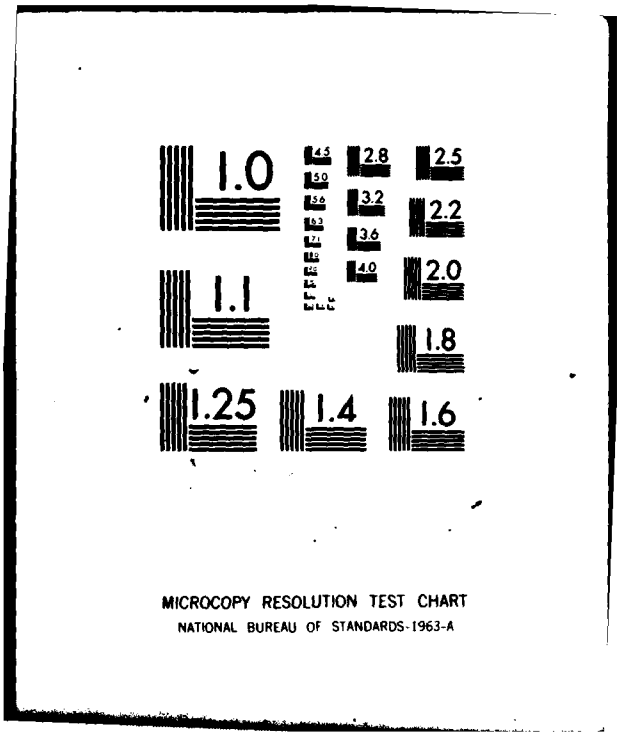
AD-A089 726 BROWN UNIV PROVIDENCE RI LEFSCHETZ CENTER FOR DYNAMI--ETC F/G 12/1
A DISCRETE APPROXIMATION FRAMEWORK FOR HEREDITARY SYSTEMS.(U)
MAY 80 I G ROSEN DAAG29-79-C-0161
UNCLASSIFIED AFOSR-TR-80-0941 NL




1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
6.3
7.1
8.0
9.0
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

A DISCRETE APPROXIMATION FRAMEWORK FOR HEREDITARY SYSTEMS

by

I. G. ROSEN*

Lefschetz Center for Dynamical Systems
Division of Applied Mathematics
Brown University
~~Providence, R.I.~~ 02912

May [redacted]80

*This research was supported in part by the U.S. Air Force under contract [redacted]-AFOSR 76-3092[redacted], in part by the U.S. Army under contract ARO-DAAG29-79-C-0161, and in part by the National Science Foundation under Grant NSF-MCS79-05774.

A DISCRETE APPROXIMATION FRAMEWORK FOR HEREDITARY SYSTEMS

Abstract

A discrete approximation framework for initial-value problems involving certain classes of linear functional differential equations (FDE) of the retarded type is constructed. An equivalence between the FDE and abstract evolution equations (AEE) in an appropriately chosen Hilbert space is established. This equivalence is then employed in the development of discrete approximation schemes in which the infinite-dimensional AEE is replaced by a finite-dimensional system of difference equations. Convergence and rates of convergence are demonstrated via the properties of rational functions with operator arguments and both classical and recent results from linear semigroup theory. Two examples of families of approximation schemes which are included in the general framework and which may be implemented directly on high-speed computing machines are developed. A numerical study of examples which illustrates the application and feasibility of the approximation techniques in a variety of problems together with a summary and analysis of the numerical results are also included.

1. Introduction

The focus of this investigation is the construction of a general abstract approximation framework for certain classes of linear retarded functional differential equations (FDE). The methods included in the framework will have a sound theoretical basis for convergence, and will be designed with the intent of application to the solution of optimal control and parameter identification problems governed by FDE. The work presented below is concerned with approximate integration methods for FDE, while the results dealing with the application of the schemes to the optimal control and parameter identification problems will be discussed elsewhere. We begin by calling upon the results of Banks and Burns [3], [4], among others, to establish the equivalence of solutions to certain classes of FDE of particular interest and the solutions of corresponding abstract ordinary differential equations, also known as abstract evolution equations (AEE), set in an appropriately chosen Hilbert space. We then proceed to develop general approximation schemes for the solutions to the homogeneous AEE which in turn, via the equivalence described above, provide approximate solutions to the FDE. Using approximation techniques for $\mathcal{C}_0$ semigroups of bounded linear operators on abstract spaces, both classical and recently published results, we are able to characterize the convergence and rates of convergence for rich classes of these schemes. In addition, two particular families of approximation schemes included in the general framework are developed and studied in detail. The approximation framework for the homogeneous initial-value problem is then extended to include schemes applicable to the non-homogeneous problem as well. We conclude with a discussion of numerical results obtained by actually implementing and testing these schemes on a wide variety of hereditary systems.

The idea of constructing approximate solutions to differential equations set in infinite-dimensional spaces, in particular FDE and parabolic and hyperbolic partial differential equations via approximations to the solutions of equivalent AEE, has been considered by many authors. Historically, the well-known Lax equivalence theorem (cf. Richtmyer and Morton [31]) for the homogeneous case, and Thompson's [35] subsequent extension of their results to the non-homogeneous and quasi-linear problems, can be considered to be the forerunners of most investigations in this direction. More recently, in the particular case of FDE, a rather extensive formulation for an approximation framework has been developed by Banks and Burns [3]. The latter treatment considers a state variable approximation exclusively as part of a two-step process through which the final approximating solution is obtained. That is to say, the AEE, an ordinary differential equation in an infinite-dimensional abstract space characterized by an unbounded operator on the right-hand side, is approximated by a sequence of systems of ordinary differential equations defined on finite-dimensional approximating subspaces. These systems of ODE of successively higher dimension must then in turn be solved numerically via any one of a number of classical approximate integration techniques. The schemes discussed here, however, approximate the AEE by a sequence of systems of discrete difference equations of successively higher dimension. This represents a simultaneous approximation in both the state and the time variable which is readily programmed in a single step. Recently, Reber [28], [29], in considering these ideas for linear non-autonomous systems (i.e. systems with coefficients that vary in time), has demonstrated sub-linear convergence for a scheme employing finite-difference-like approximations in both the state and the time variables. In the case of an autonomous system, his work becomes a special case of the general approxima-

tion framework constructed in the sequel. Furthermore, the abstract formulation to be discussed below allows one to consider state and time variable approximations of varying design and arbitrary order of convergence independently. An extensive bibliography and survey of the literature through 1976 concerned with the approximation of solutions to FDE (and, in some cases, the associated optimal control problem) via techniques of the type discussed above can be found in [4]. Finally, a rather broad theory, somewhat more general in nature yet less attuned toward practical computation than the research to follow, can be found in a recent paper by Hersh and Kato [15]. Many of our results are closely related to the ideas of the Hersh-Kato treatment.

The notation employed here is for the most part standard. For $1 \leq p < \infty$, a closed interval I in R and a Banach space X, the symbol $L_p(I;X)$ denotes the Banach space whose elements consist of equivalence classes of strongly Lebesgue-measurable functions $f: I \to X$ for which $\int_I |f|_X^p < \infty$ and which is endowed with the usual L_p norm $|f|_{L_p} = (\int_I |f|_X^p)^{1/p}$. The symbol $C(I;X)$ denotes the Banach space of continuous functions from I into X together with the usual supremum norm. In the case that $X = R^n$, where n is the dimension of the FDE system under investigation, the above notations are foreshortened to $L_p(I)$ and $C(I)$ respectively. The symbol $\langle \cdot, \cdot \rangle_{L_2}$ represents the standard inner product on the Hilbert space $L_2(I)$ given by $\langle f,g \rangle_{L_2} = \int_I fg$. $L_\infty(I)$ (with the standard L_∞ norm) denotes the Banach space of all real-valued equivalence classes of functions which are essentially bounded on I, while the notation $C^k(I)$ stands for the space of all R^n-valued continuous functions defined on I whose first k derivatives are continuous. $M(I)$ represents the measurable functions from I into R^n while the Banach spaces $W_p^{(j)}(I;R^n)$ of R^n-valued absolutely continuous functions possessing j-1 absolutely continuous derivatives and jth derivatives that are in $L_p(I)$ are denoted simply by $W_p^j(I)$. For

Banach spaces X and Y, the symbols $\mathscr{B}(X,Y)$ and $\mathscr{B}(X)$ denote the spaces of all bounded linear operators from X into Y and X into X respectively. The spaces R^n and $\mathscr{L}_{nn}$, the space of all $n \times n$ matrices, are endowed with the euclidean and spectral norms, respectively. The norm of an element x contained in a normed linear space X is denoted by $|x|_X$, or more simply by $|x|$ in the case that the intended space may be inferred from the context of the statement. Similarly, the norm of a bounded linear operator $T \in \mathscr{B}(X,Y)$ is denoted simply by $|T|$ in the case that the operator norm in question is that one which is induced by the standard norms on the spaces X and Y. If $T \in \mathscr{B}(X)$, the notation $|T|_X$ will also on occasion be used. The symbol I is used to represent the identity operator. No further clarification is provided if the space upon which it operates can be determined from the context of its usage. The standard notations $\sigma(\mathscr{T})$, $\pi(\mathscr{T})$, $\rho(\mathscr{T})$ are employed to represent the spectrum, point spectrum and resolvent set in the complex plane C of a linear operator $\mathscr{T}$, while the symbols $\mathscr{D}(\mathscr{T})$ and $\mathscr{R}(\mathscr{T})$ denote its domain and range. For $\lambda \in \rho(\mathscr{T})$, the symbol $R(\lambda;\mathscr{T})$ denotes the resolvent of $\mathscr{T}$, $(\mathscr{T}-\lambda I)^{-1}$. The positive integers n, ν, ρ and positive numbers T and r to be defined in the next section are assumed to be fixed throughout. For any function x of one real variable we use both $\dot{x}$ and Dx to stand for the derivative of x with respect to that variable. As is commonly the case in papers concerning retarded functional differential equations, for an R^n-valued measurable function $s \to x(s)$, the notation x_t denotes the function in $M(-r,0)$ given by $x_t(\theta) = x(t+\theta)$, $-r \leq \theta \leq 0$. For a rational function $C(z) = P(z)/Q(z)$ defined for $z \in C$, the symbol deg C(z) denotes that integer given by deg P - deg Q, where deg P and deg Q represent the respective degrees of P and Q as polynomials in z.

2. Equivalence of FDE and AEE

We state precisely the FDE initial-value problems for which approximate solutions are sought and describe their equivalent formulations as corresponding AEE initial-value problems set in an abstract function space.

Consider the initial-value problem given by

$$\dot{x}(t) = L(x_t) + f(t), \quad t \in [0,T] \tag{2.1}$$

$$x(0) = \eta, \qquad x_0 = \phi \tag{2.2}$$

where $\eta \in R^n$, $\phi \in L_2(-r,0)$ and $f \in L_2(0,T)$. We shall assume that the linear operator $L: L_2(-r,0) \to R^n$ is of the form

$$L(\phi) = \sum_{j=0}^{\nu} A_j \phi(-\tau_j) + \int_{-r}^{0} D(\theta)\phi(\theta)d\theta \tag{2.3}$$

where $A_j \in \mathcal{L}_{nn}$, $j = 0, 1, 2, \ldots, \nu$, $D \in L_2([-r,0]; \mathcal{L}_{nn})$ and $0 = \tau_0 < \tau_1 < \ldots < \tau_\nu = r$. Strictly speaking, the expression for L given by (2.3) is not well-defined for all $\phi \in L_2(-r,0)$ in that point evaluations of ϕ are required. However, since our primary concern is the solution of the initial-value problem (2.1), (2.2), we need only consider instances of $L(x_t)$ appearing beneath an integral sign. More precisely, a solution of the initial-value problem (2.1), (2.2) is a function $x \in L_2(-r,T)$ such that $t \to x(t)$ is absolutely continuous on $(0,T)$, $x(0) = \eta$, $x_0 = \phi$ and

$$x(t) = \eta + \int_0^t L(x_\sigma)d\sigma + \int_0^t f(\sigma)d\sigma, \qquad t \in [0,T]. \tag{2.4}$$

$x_\sigma \in L_2(-r,0)$ for each $\sigma \in [0,T]$ implies that the mapping $\sigma \to L(x_\sigma)$ is in

$L_2(0,T)$. Thus the expression for $x(t)$ given by (2.4) is well-defined. Employing standard arguments, the following lemma may be established.

2.5. Lemma. There exists a unique solution to the initial-value problem (2.1), (2.2). Moreover, the solution depends continuously upon the initial data and the non-homogeneous perturbation. That is to say, if $x_k(t)$ denotes the unique solution to the non-homogenous FDE $\dot{x}(t) = L(x_t) + f_k(t)$ with initial conditions $x(0) = \eta_k$, $x_0 = \phi_k$ where $\eta_k \to \eta$ in R^n, $\phi_k \to \phi$ in $L_2(-r,0)$ and $f_k \to f$ in $L_2(0,T)$, then we have that $\sup_{t \in [0,T]} |x_k(t)-x(t)| \to 0$ as $k \to \infty$ where $x(t)$ is the unique solution of the initial-value problem given by (2.1), (2.2).

2.6. Remark. Linear homogeneous FDE with right-hand sides of the form given by the expression in (2.3) are not the most general to which the equivalence and approximation results to be established can be applied. However, it is noted in [6] and [13] that this form is of sufficient generality to include all linear homogeneous autonomous FDE commonly arising in practical applications. The details of establishing the equivalence for the FDE initial-value problem (2.1), (2.2) under less restrictive hypotheses are discussed in [3], [4].

Following Borisovič and Turbabin [9], and countless other authors working with retarded functional differential equations, we choose the Hilbert space $Z \equiv R^n \times L_2(-r,0)$ with inner product

$$\langle(\eta_1,\phi_1),(\eta_2,\phi_2)\rangle_Z = \langle\eta_1,\eta_2\rangle_{R^n} + \langle\phi_1,\phi_2\rangle_{L_2}$$

as the space upon which the corresponding AEE will be defined. In the light

of the existence, uniqueness and continuous dependence results stated in Lemma 2.5, one can define a family of solution operators on the space Z associated with the homogeneous form of the initial-value problem (2.1), (2.2). Indeed, for $t \in [0,T]$, let $S(t): Z \to Z$ be given by

$$S(t)(\phi,\eta) = (x(t),x_t)$$

where x is the unique solution of (2.1), (2.2) with $f \equiv 0$. The pair $(x(t),x_t)$ will on occasion be referred to as the state or state variable of the system.

Since Lemma 2.5 is valid for all $T > 0$, the family $\{S(t): t \geq 0\}$ forms a $\mathscr{C}_0$ semigroup of bounded linear operators on Z. Standard techniques [4] can now be used to calculate the closed densely defined infinitesimal generator $\mathscr{A}$ of $\{S(t): t \geq 0\}$ together with its domain of definition. They are given by

$$\mathscr{A}(\eta,\phi) = (L(\phi),\dot{\phi})$$

for all $(\eta,\phi) \in \mathscr{D}(\mathscr{A}) = \{(\eta,\phi) \in Z: \eta = \phi(0), \phi \in W_2^1(-r,0)\}$.

For purpose of reference, we state certain properties of the $\mathscr{C}_0$ semigroup of operators $\{S(t): t \geq 0\}$ and its infinitesimal generator $\mathscr{A}$ that are used in the discussion below. The verification of these results may be found in any standard reference on linear semigroup theory. In particular, [1], [18], [19], [26] and [38] are adequate in this regard.

(1) $\bigcap_{n=1}^{\infty} \mathscr{D}(\mathscr{A}^n)$ is dense in Z. In particular, $\mathscr{D}(\mathscr{A}^n)$ is dense in Z for each $n = 1, 2, \ldots$.

(2) There exist positive constants β and M such that $\sigma(\mathscr{A}) = \pi(\mathscr{A}) \subset \{\lambda \in C: \text{Re } \lambda < \beta\}$ and, moreover, the resolvent operator $R(\lambda;\mathscr{A})$ with $\text{Re } \lambda > \beta$ satisfies the condition

$$(2.7) \qquad |R(\lambda;\mathcal{A})^n|_Z = |(\mathcal{A}-\lambda I)^{-n}|_Z \leq M(\mathrm{Re}\ \lambda-\beta)^{-n}$$

for all $n \geq 1$. This in turn implies that

$$|S(t)|_Z \leq Me^{\beta t}.$$

We adopt the notation of Kato [18] and let the symbol $G(M,\beta)$ denote the set of all closed, densely defined operators that satisfy a condition like (2.7) on the respective spaces upon which they are defined. In addition, we shall also, on occasion, have reason to consider the set of all closed, densely defined linear operators $\mathcal{T}$ whose resolvent sets $\rho(\mathcal{T})$ contain not only the half-plane $\{\lambda \in C: \mathrm{Re}\ \lambda > \beta\}$, but a sector of the complex plane, $\{\lambda \in C: |\arg \lambda - \beta| < \frac{\pi}{2} + \omega\}$ for some $\omega > 0$, and whose resolvent operators, $R(\lambda;\mathcal{T})$, satisfy the stronger, somewhat more restrictive condition

$$|R(\lambda;\mathcal{T})^n| = |(\mathcal{T}-\lambda I)^{-n}| \leq M|\lambda-\beta|^{-n}$$

for all $\lambda \in \{\lambda \in C: |\arg \lambda - \beta| < \frac{\pi}{2}+\omega\}$, $n = 1,2,\ldots$. We denote this set of operators by the symbol $H(\omega,\beta,M)$. We note that if $\mathcal{T} \in H(\omega,\beta,M)$ it is the infinitesimal generator of $\{U(t)\}$, a quasi-bounded semigroup of operators (i.e. $|U(t)| \leq Me^{\beta t}$), holomorphic in t for t contained in a sector of the complex plane (cf. Kato [18]).

A linear operator $\mathcal{T}$ with domain dense in a Hilbert space H is said to be dissipative if

$$\mathrm{Re}\ \langle \mathcal{T}x, x\rangle_H \leq 0 \qquad \text{for } x \in \mathcal{D}(\mathcal{T})$$

It can be shown (cf. Krein [19]) that if there exist a constant β and an inner product $[\cdot,\cdot]_H$ defined on H which generates a topology equivalent to the

standard inner product topology on H, then the conditions $\mathcal{T}-\beta I$ dissipative with respect to the $[\cdot,\cdot]$ inner product (i.e. $[\mathcal{T}x,x]_H \leq \beta[x,x]_H$, $x \in \mathcal{D}(\mathcal{T})$) and $\mathcal{R}(\mathcal{T}-\lambda I)) = H$ for any λ with Re $\lambda > \beta$ are necessary and sufficient for $\mathcal{T} \in G(M,\beta)$. Furthermore, $\mathcal{T} - \beta I$ dissipative implies (see [19]) that $\sigma(\mathcal{T}) \subset \{\lambda \in C: \text{Re } \lambda < \beta\}$. Thus if H is finite-dimensional, $\mathcal{T} \in G(M,\beta)$ if and only if $(\mathcal{T}-\beta I)$ is dissipative.

We now construct an inner product on Z, $\langle \cdot,\cdot \rangle_g$, that generates an equivalent topology to the standard inner product topology on Z and for which there exists a constant β such that $\langle \mathcal{A}z,z\rangle_g \leq \beta\langle z,z\rangle_g$ for all z contained in $\mathcal{D}(\mathcal{A})$. The inner product we construct is essentially the same as those defined in [6], [29] and [30] for a similar purpose.

Let the step function g defined on $[-r,0)$ be given by

$$g(\theta) = g_j = 1 + \sum_{i=j}^{\nu} |A_i| \quad \text{for } \theta \in [-\tau_j,-\tau_{j-1}), \quad j = 1,2,\ldots,\nu.$$

Then, for (η,ϕ) and $(\zeta,\psi) \in Z$, we define

$$\langle(\eta,\phi),(\zeta,\psi)\rangle_g = \eta^T\zeta + \int_{-r}^{0} \phi(\theta)\psi(\theta)g(\theta)d\theta$$

$$= \eta^T\zeta + \sum_{j=1}^{\nu} g_j \int_{-\tau_j}^{-\tau_{j-1}} \phi(\theta)\psi(\theta)d\theta$$

and

$$|(\eta,\phi)|_g^2 = \langle(\eta,\phi),(\eta,\phi)\rangle_g.$$

Clearly

$$|\cdot|_Z \leq |\cdot|_g \leq (1 + \sum_{j=1}^{\nu} |A_j|)^{1/2} |\cdot|_Z .$$

Thus the topology generated by $\langle \cdot,\cdot \rangle_g$ on Z is equivalent to the standard inner product topology on Z. Furthermore, using standard arguments (cf. [6], [29], [33]), we have for $\dot{\phi} = (\phi(0),\phi) \in \mathcal{D}(\mathcal{A})$,

$$\langle \mathcal{A}\dot{\phi},\hat{\phi} \rangle_g = \langle (L(\phi),\dot{\phi}),(\phi(0),\phi) \rangle_g$$

$$= \langle \sum_{j=0}^{\nu} A_j \phi(-\tau_j) + \int_{-r}^{0} D(\theta)\phi(\theta)d\theta, \phi(0) \rangle_{R^n}$$

$$+ \sum_{j=1}^{\nu} g_j \int_{-\tau_j}^{-\tau_{j-1}} \dot{\phi}(\theta)\phi(\theta)d\theta \leq \beta \langle \hat{\phi},\hat{\phi} \rangle_g$$

where $\beta = (1 + \sum_{j=0}^{\nu}|A_j| + |D|_{L_2})$.

Turning our attention to establishing an equivalence result for the non-homogeneous initial-value problem (2.1), (2.2), for $z_0 = (\eta,\phi) \in Z$ and $f \in L_2(0,T)$, let $z: [0,T] \to Z$ be given by the expression

$$(2.8) \qquad z(t) = S(t)z_0 + \int_0^t S(t-\sigma)(f(\sigma),0)d\sigma , \qquad t \in [0,T],$$

where 0 denotes the zero function in $L_2(-r,0)$.

2.9. Lemma. *For z as in (2.8) and x the unique solution of the non-homogeneous initial-value problem (2.1), (2.2) we have the strong equivalence of solutions given by*

(2.10) $\quad z(t) = (x(t), x_t)$.

The complete proof of Lemma 2.9 may be found in [3]. However, it can be summarized as follows. The equivalence stated in (2.10) is easily verified for the case of $f \in C^1(0,T)$ and $z_0 \in \mathcal{D}(\mathcal{A})$ via standard results from linear semigroup theory [18]. An application of the facts that $C^1(0,T)$ is dense in $L_2(0,T)$ and $\mathcal{D}(\mathcal{A})$ is dense in Z, together with the uniqueness and continuous dependence properties of both x(t) and z(t), are sufficient to conclude that the desired strong equivalence of solutions obtains for all $f \in L_2(0,T)$ and $z_0 \in Z$.

3. Preliminary Definitions and Basic Results

We make the following definitions that will prove useful in our discussions below in regard to the formulation of an approximation framework for the homogeneous FDE initial-value problem given by

(3.1) $\quad \dot{x}(t) = L(x_t), \qquad t \in [0,T],$

(3.2) $\quad x(0) = \eta, \quad x_0 = \phi.$

3.3 Definition. Let $\{\hat{Z}^N\}$ denote a sequence of approximating finite-dimensional subspaces of Z [6] defined by

$$\hat{Z}^N = \text{span}\{\hat{\phi}_N^{(0)}, \hat{\phi}_N^{(1)}, \ldots, \hat{\phi}_N^{(k_N)}\},$$

where $\hat{\phi}_N^{(j)} \in Z$, $j = 1, 2, \ldots, k_N$. Then for each $N = 1, 2, \ldots$, the 4-tuple $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ will be called a Discrete Approximation Scheme, or more simply a DAS, for the Cauchy problem (3.1), (3.2) if

(1) $(Z_N, \langle\cdot,\cdot\rangle_N)$ is a finite-dimensional approximating Hilbert space defined by the relations

$$Z_N = \sigma^N(\hat{Z}^N)$$

and

$$\langle\cdot,\cdot\rangle_N = \langle(\sigma^N)^{-1}(\cdot),(\sigma^N)^{-1}(\cdot)\rangle_Z$$

where σ^N represents an algebraic isomorphism mapping $\hat{Z}^N$ onto Z_N. (As an example, one possible construction for Z_N would be to choose $Z_N = R^{k_N}$ and σ^N as the canonical coordinate map from the finite-dimensional subspace $\hat{Z}^N$ onto R^{k_N}.) We note that with the inner product on Z_N defined as above, σ^N actually represents an isometric mapping of $\hat{Z}^N$ onto Z_N.

(2) $\pi_N: Z \to Z_N$ together with its right inverse $\pi_N^{-1}: Z_N \to Z$ are projection- and embedding-like mappings respectively defined by

$$\pi_N = \sigma^N\hat{P}_N, \qquad \pi_N^{-1} = (\sigma^N)^{-1}$$

where $\hat{P}_N$ is the orthogonal projection of Z onto $\hat{Z}^N$ along $(\hat{Z}^N)^{\perp}$.

(3) $\mathcal{A}_N: Z_N \to Z_N$ is a bounded linear operator.

(4) $C(z)$ is a rational function of the complex variable z.

We make the standing assumption that $T = \rho r$, ρ an integer greater than zero, and partition the interval $[0,T]$ into ρN subintervals of equal length defined by the nodal points $t_k^N = kr/N$, $k = 0, 1, \ldots, \rho N$. That essentially no loss of generality is incurred by restricting T to be an integral multiple of the maximum delay in the problem, r, is discussed in Reber [29, Section 8]. It is our desire to construct the Discrete Approximation Scheme $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ in a manner that will guarantee that if

(a) $\pi_N z$ is, in some sense, an approximation to z for each $z \in Z$

(b) $\mathscr{A}_N \pi_N z$ is, in some sense, an approximation to $\mathscr{A} z$ for each z in a sufficiently large subset of $\mathscr{D}(\mathscr{A})$

(c) $C(z)$ is, in some sense, an approximation to e^z for $z \in \mathbb{C}$ sufficiently small

then the sequence of vectors $\{z_k^N\}_{k=0}^{\rho N}$ contained in Z_N and generated by the discrete semigroup of operators $\{C(\frac{r}{N}\mathscr{A}_N)^k\}$ (cf. Kato [18]) according to the recurrence

$$z_0^N = \pi_N z_0, \qquad z_{k+1}^N = C(\tfrac{r}{N}\mathscr{A}_N) z_k, \qquad k = 0,1,2,\ldots,\rho N-1,$$

will in some sense approximate

$$z(t_k^N) = e^{\mathscr{A} t_k^N} z_0 = S(t_k^N) z_0, \qquad k = 0,1,2,\ldots,\rho N .$$

It is further desired that $\{z_k^N\}_{k=0}^{\rho N}$ provide an approximation $\{x_k^N\}_{k=0}^{\rho N}$ to $\{x(t_k^N)\}_{k=0}^{\rho N}$, the true solution of the FDE initial-value problem (3.1), (3.2) evaluated at the node points. Making these ideas precise and demonstrating that they can indeed be realized are the concerns of the definitions and results that follow.

3.4. Definition. The Discrete Approximation Scheme $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ will be said to be factor stable if the infinite set of operators on Z_N given by

$$C(\tfrac{r}{N}\mathscr{A}_N)^k, \qquad k = 0,1,2,\ldots,\rho N \quad .$$

is uniformly bounded in N for all N sufficiently large.

The fact that $C(z) \equiv N(z)/D(z)$ is a rational function implies that the evaluation of $C(\frac{r}{N}\mathscr{A}_N)$ for each N will require the invertibility of the operators $D(\frac{r}{N}\mathscr{A}_N)$. Sufficient conditions which can be satisfied by the $\mathscr{A}_N$ and C(z) which will guarantee the existence of this inverse will be provided in the next section. For the present, however, the operators $C(\frac{r}{N}\mathscr{A}_N)$ will be referred to with the implicit assumption that $D^{-1}(\frac{r}{N}\mathscr{A}_N)$ exists for all N sufficiently large.

3.5. Definition. The Discrete Approximation Scheme $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ will be said to be factor convergent as an approximation to the initial-value problem (3.1), (3.2) if for each $z_0 \in Z$, given $\varepsilon > 0$, there exists an $\hat{N} = \hat{N}(\varepsilon, z_0)$ such that

$$\left|C(\tfrac{r}{N}\mathscr{A}_N)^k \pi_N z_0 - \pi_N S(t_k^N) z_0\right|_N < \varepsilon, \qquad k = 0,1,2,\ldots,\rho N$$

for all $N > \hat{N}$.

The next definition is a precise statement of what is intended when it is said that $\pi_N z$ is an approximation to z for each $z \in Z$.

3.6. Definition. A Discrete Approximation Scheme $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ will be said to have property (P1) if the mapping $\pi_N: Z \to Z_N$ and its right inverse $\pi_N^{-1}: Z_N \to Z$ satisfy the condition

(P1) $$\left|\pi_N^{-1}\pi_N z - z\right|_Z \to 0 \quad \text{as} \quad N \to \infty \text{ for each } z \in Z.$$

3.7. Lemma. Suppose $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ is a Discrete Approximation Scheme with property (P1). Then the mapping $\pi_N: Z \to Z_N$ and its right inverse

$\pi_N^{-1}: Z_N \to Z$ satisfy the following:

(1) $|\pi_N z|_N \leq |z|_Z$ for each $z \in Z$;

(2) $|\pi_N^{-1} z^N|_Z \leq |z^N|_N$ for each $z^N \in Z_N$;

(3) $|\pi_N z|_N \to |z|_Z$ as $N \to \infty$ for each $z \in Z$.

The veracity of Lemma 3.7 follows directly from the definitions of π_N and π_N^{-1}. We note that only the third proposition requires $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ to have property (P1).

3.8. Definition. Let $p_1: Z \to R^n$, $p_2: Z \to L_2(-r,0)$ be the two coordinate projection mappings defined by $p_1(\eta,\phi) = \eta$ and $p_2(\eta,\phi) = \phi$ respectively for $(\eta,\phi) \in Z$.

That a factor convergent Discrete Approximation Scheme does indeed yield an approximate solution to the FDE initial-value problem (3.1), (3.2) is verified in the next lemma.

3.9. Lemma. Suppose that $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ is a factor convergent Discrete Approximation Scheme with property (P1). Then, given $\varepsilon > 0$, there exists an $\hat{N} = \hat{N}(\varepsilon)$ such that $|x(t_k^N) - p_1(\pi_N^{-1} z_k^N)|_{R^n} < \varepsilon$, $k = 0,1,\ldots,\rho N$, for all $N > \hat{N}$, where the sequence $\{z_k^N\}_{k=0}^{\rho N}$ in Z_N is given by

$$z_k^N = C(\tfrac{r}{N}\mathscr{A}_N)^k \pi_N z_0.$$

Proof: Let $\varepsilon > 0$ be given. Then

$$(3.10)\qquad |x(t_k^N)-p_1(\pi_N^{-1}z_k^N)|_{R^n} = |p_1(z(t_k^N)-\pi_N^{-1}z_k^N)|_{R^n}$$

$$\leq |z(t_k^N)-\pi_N^{-1}z_k^N|_Z$$

$$\leq |z(t_k^N)-\pi_N^{-1}\pi_N z(t_k^N)|_Z + |\pi_N^{-1}\pi_N z(t_k^N)-\pi_N^{-1}z_k^N|_Z$$

$$= |(I-\pi_N^{-1}\pi_N)z(t_k^N)|_Z + |C(\tfrac{r}{N}\mathscr{A}_N)^k\pi_N z_0-\pi_N S(t_k^N)z_0|_N$$

$$\leq \sup_{t\in[0,T]} |(I-\pi_N^{-1}\pi_N)z(t)|_Z + |C(\tfrac{r}{N}\mathscr{A}_N)^k\pi_N z_0-\pi_N S(t_k^N)z_0|_N .$$

Since $\{z(t): t\in[0,T]\}$ is a compact subset of Z (being the continuous image of a compact subset of R) and $\{Z_N,\pi_N,\mathscr{A}_N,C(z)\}$ has been assumed to have property (P1), we may conclude that $\pi_N^{-1}\pi_N \to I$ uniformly on $\{z(t): t\in[0,T]\}$ as $N\to\infty$. Thus the first term in the last inequality in (3.10) tends to zero as $N\to\infty$. In addition, the fact that $\{Z_N,\pi_N,\mathscr{A}_N,C(z)\}$ has been assumed to be factor convergent implies that

$$|C(\tfrac{r}{N}\mathscr{A}_N)^k\pi_N z_0-\pi_N S(t_k^N)z_0|_N < \varepsilon/2, \qquad k = 0,1,\ldots,\rho N$$

for all N sufficiently large. Therefore, it follows that

$$|x(t_k^N)-p_1(\pi_N^{-1}z_k^N)|_{R^n} < \varepsilon, \qquad k = 0,1,2,\ldots,\rho N$$

for all N sufficiently large.

□

4. The Equivalence Theorem

In this section we state and prove a theorem that provides necessary and sufficient conditions for the factor convergence of a Discrete Approximation Scheme and subsequently yields estimates for the rate of factor convergence on restricted classes of initial data. The sufficient conditions are such that they are easily verified for a wide variety of DAS that are considered in the sequel. The arguments required to prove this theorem rely heavily upon standard approximation results for semigroups of linear operators (cf. Kato [18]). These preliminary results, which have been suitably modified so as to allow for the additional complexity introduced by the variation of the approximating spaces, are contained in Lemmas 4.1, 4.3 and Theorem 4.4 below.

In the discussions which follow, Z_N, π_N, $\mathscr{A}_N$ are assumed to have been constructed in accordance with the requirements specified in Definition 3.3.

4.1 Lemma. Suppose

(1) $\mathscr{A}, \mathscr{A}_N \in G(M,\beta)$ for all N sufficiently large (M,β independent of N);

(2) $|[\mathscr{A}_N\pi_N - \pi_N\mathscr{A}]z_0|_N \to 0$ as $N \to \infty$ for each $z_0 \in D_1$, where D_1 is a dense subset of Z contained in $\mathscr{D}(\mathscr{A})$;

(3) For $\lambda_0 \in C$ with $\mathrm{Re}\ \lambda_0 > \beta$ there exists a dense subset of Z, D_2, such that $R(\lambda_0;\mathscr{A})D_2 \subseteq D_1$.

Then it follows that

$$|[R(\lambda_0;\mathscr{A}_N)\pi_N - \pi_N R(\lambda_0;\mathscr{A})]z_0|_N \to 0 \qquad \text{as } N \to \infty$$

for each $z_0 \in Z$.

Proof:

$$(4.2) \qquad \left| [R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A})]z_0 \right|_N$$

$$= \left| R(\lambda_0;\mathcal{A}_N)[\mathcal{A}_N\pi_N - \pi_N\mathcal{A}]R(\lambda_0;\mathcal{A})z_0 \right|_N$$

$$\leq \left| R(\lambda_0;\mathcal{A}_N) \right| \left| [\mathcal{A}_N\pi_N - \pi_N\mathcal{A}]R(\lambda_0;\mathcal{A})z_0 \right|_N$$

$$\leq \frac{M}{\text{Re } \lambda_0 - \beta} \left| [\mathcal{A}_N\pi_N - \pi_N\mathcal{A}]R(\lambda_0;\mathcal{A})z_0 \right|_N \to 0 \quad \text{as} \quad N \to \infty$$

for each $z_0 \in D_2$. However, D_2 is dense in Z, and the operators

$$[R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A})]$$

are uniformly bounded in N for all N sufficiently large. Indeed,

$$\left| R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A}) \right|_N \leq \frac{2M}{\text{Re } \lambda_0 - \beta} .$$

for all N sufficiently large. Therefore, it follows that

$$\left| [R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A})]z_0 \right|_N \to 0 \qquad \text{as } N \to \infty$$

for each $z_0 \in Z$.

□

4.3. Lemma. Suppose $\mathcal{A}_N \in G(M,\beta)$ for all N sufficiently large. Then for $\lambda \in C$ with Re $\lambda > \beta$, the operators $\mathcal{A}_N R(\lambda;\mathcal{A}_N) = \mathcal{A}_N(\mathcal{A}_N-\lambda I)^{-1}$ are uniformly bounded in N for all N sufficiently large.

Proof: Re $\lambda > \beta$ and $\mathcal{A}_N \in G(M,\beta)$ imply that $\lambda \in \rho(\mathcal{A}_N)$. Therefore, we have

$$I = (\mathcal{A}_N-\lambda I)(\mathcal{A}_N-\lambda I)^{-1} = \mathcal{A}_N(\mathcal{A}_N-\lambda I)^{-1} - \lambda(\mathcal{A}_N-\lambda I)^{-1}$$

or

$$\mathcal{A}_N(\mathcal{A}_N-\lambda I)^{-1} = I + \lambda(\mathcal{A}_N-\lambda I)^{-1}.$$

This implies

$$|\mathcal{A}_N(\mathcal{A}_N-\lambda I)^{-1}|_N \leq |I|_N + |\lambda|\,|(\mathcal{A}_N-\lambda I)^{-1}|_N$$

$$\leq 1 + \frac{|\lambda| M}{\text{Re } \lambda-\beta} \equiv M_\lambda$$

for all N sufficiently large. □

Theorem 4.4, to follow, is a minor modification of a standard result from the theory of approximation for linear semigroups of operators generally attributed to Trotter [36]. The veracity of the result can be argued in a manner similar to that used by Kato in verifying Theorem 2.16 in Chapter IX of [18]. The details of the proof of the result as stated in Theorem 4.4 can be found in [33].

4.4 Theorem. Suppose

(1) $\mathscr{A}_N$: $Z_N \to Z_N$ is the infinitesimal generator of the $\mathscr{C}_0$ semigroup of bounded linear operators, $\{S_N(t): t \geq 0\}$, defined on Z_N;

(2) $\mathscr{A}$: $\mathscr{D}(\mathscr{A}) \subset Z \to Z$ is the infinitesimal generator of the $\mathscr{C}_0$ semigroup of bounded linear operators, $\{S(t): t \geq 0\}$, defined on Z,

such that

(1) $\mathscr{A}_N, \mathscr{A} \in G(M,\beta)$ for all N sufficiently large (M,β independent of N);

(2) $|[\mathscr{A}_N\pi_N - \pi_N\mathscr{A}]z_0|_N \to 0$ as $N \to \infty$ for each $z_0 \in D_1$, where D_1 is a dense subset of Z contained in $\mathscr{D}(\mathscr{A})$;

(3) There exists a $\lambda_0 \in C$ with $\text{Re}\ \lambda_0 > \beta$ and D_2, a dense subset of Z such that

$$R(\lambda_0;\mathscr{A})D_2 \subseteq D_1.$$

Then $|[S_N(t)\pi_N - \pi_N S(t)]z_0|_N \to 0$ as $N \to \infty$ for each $z_0 \in Z$, and moreover the convergence is uniform in t for $t \in [0,T]$.

We remark that if the set D_1 is invariant under $R(\lambda_0;\mathscr{A})$ for some $\lambda_0 \in C$ with $\text{Re}\ \lambda_0 > \beta$, that is $R(\lambda_0;\mathscr{A})D_1 \subseteq D_1$, it suffices to choose $D_2 = D_1$.

The following corollary yields an estimate for the rate of convergence in Theorem 4.4.

4.5. Corollary. Suppose $\mathscr{S} \subseteq \mathscr{D}(\mathscr{A}^2)$ satisfies the following:

(1) For each $z \in \mathscr{S}$, there exists a $K = K(z)$ such that

$$|[\mathscr{A}_N\pi_N - \pi_N\mathscr{A}]z|_N \leq K/N^p;$$

(2) There exists a subset $\mathcal{S}_1$ of $\mathcal{S}$ such that, for $z \in \mathcal{S}_1$ and λ with $Re\lambda > \beta$,

(a) $S(t)z \in \mathcal{S}$, $t \in [0,T]$

(b) $S(t)(\lambda I - \mathcal{A})z \in \mathcal{S}$, $t \in [0,T]$

and furthermore the constants guaranteed by (1) for (a) and (b) are independent of $t \in [0,T]$.

Then under the hypotheses of Theorem 4.4, there exists a $k(z)$ such that

$$\left| [S_N(t)\pi_N - \pi_N S(t)]z \right|_N \leq k(z)\left(\frac{r}{N}\right)^p \qquad t \in [0,T]$$

for each $z \in \mathcal{S}_1$.

The verification of Corollary 4.5, which follows as a direct consequence of the arguments in support of Theorem 4.4, can be found in [6].

The subsequent four lemmas provide results and identities which are required in order to estimate the degree to which a rational function approximation to the exponential evaluated at $t\mathcal{T}$, where $\mathcal{T}$ is the infinitesimal generator of the $\mathcal{C}_0$ semigroup $\{e^{\mathcal{T}t}: t \geq 0\}$, approximates $e^{\mathcal{T}t}$ for t small. With the exception of the final conclusion, Lemma 4.6 is a verbatim statement of Hersh and Kato [15], Lemma 2. The proof, which has been omitted, can be found in that paper. The result which has been appended to Lemma 4.6 follows as an immediate consequence of their arguments. Lemmas 4.7 and 4.8 comprise a minor extension of Lemma 3 in [15]. The proof of Lemma 4.7 can be argued using the properties of $\mathcal{C}_0$ semigroups and their infinitesimal generators (cf. [33]). The proof of Lemma 4.8 has been included.

4.6. Lemma. Suppose $\mathcal{T} \in G(M,\beta)$ is the infinitesimal generator of a $\mathcal{C}_0$ semigroup of operators and $C(z)$ is a rational function of degree ≤ 0 with no poles in $\{z \in C: \text{Re } z \leq 0\}$. Then there exist positive constants ε, K such that $|C(h\mathcal{T})| \leq K$ for all h with $0 \leq h \leq \varepsilon$. Moreover, the only dependence of the constants ε and K upon the operator $\mathcal{T}$ is reflected in the choice of ε, where $\varepsilon = \varepsilon(\beta)$.

4.7. Lemma. Suppose $\mathcal{T} \in G(M,\beta)$ is the infinitesimal generator of the $\mathcal{C}_0$ semigroup of operators $\{e^{\mathcal{T}t}: t \geq 0\}$. Then for $f \in \mathcal{D}(\mathcal{T}^{q+1})$ we have

$$|e^{\mathcal{T}h} f - \sum_{j=0}^{q} \frac{(h\mathcal{T})^j f}{j!}| \leq Mh^{q+1} e^{h\beta} |\mathcal{T}^{q+1} f|.$$

4.8. Lemma. Suppose

(1) $\mathcal{T} \in G(M,\beta)$ is the infinitesimal generator of the $\mathcal{C}_0$ semigroup of operators $\{e^{\mathcal{T}t}: t \geq 0\}$.

(2) $C(z)$ is a rational function satisfying

(a) $|e^z - C(z)| = O(|z|^{q+1})$, $z \to 0$ with $q > 0$

(b) deg $C(z) \leq q+1$

(c) $C(z)$ has no poles in $\{z \in C: \text{Re } z \leq 0\}$.

Then for h sufficiently small, the operator $C(h\mathcal{T})$ exists and, moreover, for $f \in \mathcal{D}(\mathcal{T}^{q+1})$, we have

$$|e^{\mathcal{T}h} f - C(h\mathcal{T}) f| \leq \hat{M} e^{\beta h} |\mathcal{T}^{q+1} f| h^{q+1},$$

where $\hat{M}$ is a positive constant independent of $\mathcal{T} \in G(M,\beta)$.

Proof: Suppose $C(z) \equiv N(z)/D(z)$ where $D(z)$ is a polynomial of degree p. Without loss of generality, we may assume that D has leading coefficient 1. Thus $D(z)$ may be written as

$$D(z) = \prod_{j=1}^{p} (z-\lambda_j)$$

where by hypothesis (2c) $\operatorname{Re} \lambda_j > 0$, $j = 1,2,\ldots,p$. Assuming for the moment that $(\mathcal{I}-h\lambda_j)^{-1}$ exists, we see that

$$D^{-1}(h\mathcal{I}) = \left(\prod_{j=1}^{p} (\mathcal{I}h-\lambda_j)\right)^{-1} = \prod_{j=1}^{p} (\mathcal{I}h-\lambda_j)^{-1}$$

$$= \prod_{j=1}^{p} (h^{-1})(\mathcal{I}-h^{-1}\lambda_j)^{-1} = (h^{-1})^p \prod_{j=1}^{p} (\mathcal{I}-h^{-1}\lambda_j)^{-1}.$$

Now $\operatorname{Re} \lambda_j > 0$, $j = 1,2,\ldots,p$, and $\mathcal{I} \in G(M,\beta)$ imply

$$h^{-1}\lambda_j \in \{\lambda \in C: \operatorname{Re} \lambda > \beta\} \subset \rho(\mathcal{I}), \qquad j = 1,2,\ldots,p$$

for all h sufficiently small. Hence, $(\mathcal{I}-h^{-1}\lambda_j)^{-1}$, $j = 1,\ldots, p$, do indeed exist, as does $C(h\mathcal{I}) = D(h\mathcal{I})^{-1}N(h\mathcal{I})$ for all h sufficiently small. To prove the second proposition, we note that hypothesis (2c) implies that $C(z)$ is analytic at $z = 0$. Therefore hypotheses (2a) and (2c) together imply that $C(z) = \sum_{j=0}^{q}(z^j/j!) + z^{q+1}Q(z)$ or $Q(z) = (C(z) - \sum_{j=0}^{q} \frac{z^j}{j!}) z^{-(q+1)}$, where $Q(z)$ is analytic near $z = 0$, i.e. Q does not have a pole at $z = 0$. Furthermore, for $z \in \{z \in C: \operatorname{Re} z \leq 0, z \neq 0\}$, $C(z)z^{-(q+1)}$ is finite by hypothesis (2c) and $\left(\sum_{j=0}^{q} \frac{z^j}{j!}\right) z^{-(q+1)}$ is finite since it has degree less than zero and the point $z = 0$ has been excluded from the set of interest. Therefore we may conclude that $Q(z)$ has no poles in $\{z \in C: \operatorname{Re} z \leq 0\}$.

If $C(z) = N(z)/D(z)$, hypothesis 2(b) implies

$$\deg N(z) \leq \deg D(z) + q + 1.$$

Therefore

$$\deg Q(z) = \deg[(C(z) - \sum_{j=0}^{q} \frac{z^j}{j!}) z^{-(q+1)}]$$

$$= \deg[\frac{N(z) - D(z) \sum_{j=0}^{q} \frac{z^j}{j!}}{D(z) z^{q+1}}]$$

$$= \deg[N(z) - D(z) \sum_{j=0}^{q} \frac{z^j}{j!}] - \deg[D(z) z^{q+1}]$$

$$\leq \max[\deg N(z), \deg(D(z) \sum_{j=0}^{q} \frac{z^j}{j!})] - (\deg D(z) + q + 1)$$

$$= \max[\deg N(z), \deg D(z) + q] - (\deg D(z) + q + 1)$$

$$\leq \deg D(z) + q + 1 - (\deg D(z) + q + 1) = 0,$$

and thus $Q(z)$ satisfies all of the hypotheses of Lemma 4.6. Thus, for all h sufficiently small and all $\mathcal{T} \in G(M,\beta)$ we have

$$|Q(h\mathcal{T})| \leq k \qquad \text{for some } K > 0 \text{ independent of } \mathcal{T} \in G(M,\beta),$$

and hence

$$|C(h\mathcal{T})f - \sum_{j=0}^{q} \frac{(h\mathcal{T})^j}{j!} f| = |(h\mathcal{T})^{q+1} Q(h\mathcal{T}) f| = h^{q+1} |Q(h\mathcal{T}) \mathcal{T}^{q+1} f|$$

$$\leq k h^{q+1} |\mathcal{T}^{q+1} f|.$$

Therefore,

$$|e^{\mathscr{T}h}f - C(h\mathscr{T})f| \leq |e^{\mathscr{T}h}f - \sum_{j=0}^{q} \frac{(h\mathscr{T})^j}{j!} f| + |\sum_{j=0}^{q} \frac{(h\mathscr{T})^j}{j!} f - C(h\mathscr{T})f|$$

$$\leq Mh^{q+1}e^{h\beta}|\mathscr{T}^{q+1}f| + kh^{q+1}|\mathscr{T}^{q+1}f|$$

$$= \hat{M}e^{h\beta}h^{q+1}|\mathscr{T}^{q+1}f|.$$

The estimate of the bound on the first term on the right-hand side of the preceding inequality follows from Lemma 4.7. [.

We are now prepared to state and prove the primary result of this section. It is referred to as the equivalence theorem because it serves to characterize factor convergence for Discrete Approximation Schemes and because of the similarity it bears to the well-known Lax equivalence theorem [20]. The reader is instructed to note the similarities which exist between Theorem 4.9 to follow, the Lax theorem mentioned above, and the somewhat more general result given in Theorem 1 of [15]. The key step in the arguments supporting the sufficiency claim in all three of these results is the factorization which is employed immediately preceding (4.11) in the proof which is given below.

4.9. Theorem. (The Equivalence Theorem) Suppose $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ is a Discrete Approximation Scheme with property (P1) satisfying

(1a) For all N sufficiently large, $\mathscr{A}_N \in G(M,\beta)$ is the infinitesimal generator of the $\mathscr{C}_0$ semigroup of operators $\{S_N(t): t \geq 0\}$ defined on Z_N.

(2a) $|[\mathscr{A}_N\pi_N - \pi_N\mathscr{A}]z_0|_N \to 0$ as $N \to \infty$ for each $z_0 \in D_1$, where D_1 is a dense subset of Z contained in $\mathscr{D}(\mathscr{A})$.

(3a) There exists a $\lambda_0 \in C$ with Re $\lambda_0 > \beta$ and D_2, a dense subset of Z such that $R(\lambda_0;\mathcal{A})D_2 \subseteq D_1$,

and

(1b) $|C(z)-e^z| = O(|z|^{q+1})$ as $z \to 0$ with $q > 0$

(2b) deg $C(z) \leq q+1$

(3b) $C(z)$ has no poles in $\{z \in C: \text{Re } z \leq 0\}$.

Then the operators $C(\frac{r}{N}\mathcal{A}_N)$ exist for all N sufficiently large and factor stability is necessary and sufficient for factor convergence.

Proof. The first proposition follows as a direct result of Lemma 4.8. The arguments required to verify the necessity part of the claim parallel those employed in the proof of necessity in the Lax equivalence theorem [20], and have therefore been omitted. A detailed proof of this result as it is stated above can be found in [33]. To prove sufficiency, for $z_0 \in Z$ and $k = 0,1,2,\ldots,\rho N$, we have

$$(4.10) \qquad |[C(\tfrac{r}{N}\mathcal{A}_N)^k\pi_N - \pi_N S(t_k^N)]z_0|_N$$

$$\leq |[C(\tfrac{r}{N}\mathcal{A}_N)^k - S_N(t_k^N)]\pi_N z_0|_N + |[S_N(t_k^N)\pi_N - \pi_N S(t_k^N)]z_0|_N.$$

Theorem 4.4 implies that the second term on the right-hand side of (4.10) tends to zero as $N \to \infty$. We now consider the first term. We have

$$\left| \left[C\left(\frac{r}{N}\mathscr{A}_N\right)^k - S_N(t_k^N) \right] R(\lambda_0; \mathscr{A}_N)^{q+1} \pi_N z_0 \right|_N$$

$$= \left| \sum_{j=0}^{k-1} C\left(\frac{r}{N}\mathscr{A}_N\right)^j \left[C\left(\frac{r}{N}\mathscr{A}_N\right) - S_N\left(\frac{r}{N}\right) \right] S_N(t_{k-1-j}^N) R(\lambda_0; \mathscr{A}_N)^{q+1} \pi_N z_0 \right|_N \tag{4.11}$$

$$\leq \sum_{j=0}^{k-1} \left| C\left(\frac{r}{N}\mathscr{A}_N\right)^j \right| \left| \left[C\left(\frac{r}{n}\mathscr{A}_N\right) - S_N\left(\frac{r}{N}\right) \right] S_N(t_{k-1-j}^N) R(\lambda_0; \mathscr{A}_N)^{q+1} \pi_N z_0 \right|_N \tag{4.12}$$

$$\leq M_0 \sum_{j=0}^{k-1} \left| \left[C\left(\frac{r}{N}\mathscr{A}_N\right) - S_N\left(\frac{r}{N}\right) \right] S_N(t_{k-1-j}^N) R(\lambda_0; \mathscr{A}_N)^{q+1} \pi_N z_0 \right|_N \tag{4.13}$$

$$\leq M_0 \hat{M} \sum_{j=0}^{k-1} e^{\beta \frac{r}{N}} \left(\frac{r}{N}\right)^{q+1} \left| \mathscr{A}_N^{q+1} S_N(t_{k-1-j}^N) R(\lambda_0; \mathscr{A}_N)^{q+1} \pi_N z_0 \right|_N$$

$$\leq M_0 \hat{M} \sum_{j=0}^{k-1} e^{\beta \frac{r}{N}} \left(\frac{r}{N}\right)^{q+1} \left| S_N(t_{k-1-j}^N) \left[\mathscr{A}_N R(\lambda_0; \mathscr{A}_N) \right]^{q+1} \pi_N z_0 \right|_N$$

$$\leq M_0 \hat{M} M \sum_{j=0}^{k-1} e^{\beta t_{k-j}^N} \left(\frac{r}{N}\right)^{q+1} \left| \left[\mathscr{A}_N R(\lambda_0; \mathscr{A}_N) \right]^{q+1} \pi_N z_0 \right|_N$$

$$\leq M_0 \hat{M} M e^{\beta T} r \rho N \left(\frac{r}{N}\right)^{q+1} \left| \left[\mathscr{A}_N R(\lambda_0; \mathscr{A}_N) \right]^{q+1} \pi_N z_0 \right|_N$$

$$\leq M_0 \hat{M} M e^{\beta T} \rho r \left(\frac{r}{N}\right)^q \left| \mathscr{A}_N R(\lambda_0; \mathscr{A}_N) \right|^{q+1} \left| \pi_N z_0 \right|_N$$

$$\leq M_0 \hat{M} M T e^{\beta T} \left(\frac{r}{N}\right)^q \left| \mathscr{A}_N R(\lambda_0; \mathscr{A}_N) \right|^{q+1} \left| z_0 \right|_Z$$

$$\leq \gamma M_{\lambda_0}^{q+1} \left| z_0 \right|_Z \left(\frac{r}{N}\right)^q \quad , \text{ for all N sufficiently large,}$$

where $\gamma = M_0 \hat{M} M T e^{\beta T}$ and M_{λ_0} is as was defined in Lemma 4.3. The estimate in (4.12) and the constant M_0 are consequences of the assumption of factor

stability, while the estimate in (4.13) and the constant $\hat{M}$ result from an application of Lemma 4.8.

Lemma 4.1 implies that

$$
\begin{aligned}
&\left| [R(\lambda_0;\mathcal{A}_N)^{q+1}\pi_N - \pi_N R(\lambda_0;\mathcal{A})^{q+1}]\, z_0 \right|_N \\
&= \left| \sum_{j=0}^{q} R(\lambda_0;\mathcal{A}_N)^j [R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A})] R(\lambda_0;\mathcal{A})^{q-j} z_0 \right|_N \\
&\le \sum_{j=0}^{q} \left| R(\lambda_0;\mathcal{A}_N)^j \right| \left| [R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A})] R(\lambda_0;\mathcal{A})^{q-j} z_0 \right|_N \\
&\le \sum_{j=0}^{q} \frac{M}{(\mathrm{Re}\ \lambda_0 - \beta)^j} \left| [R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A})] R(\lambda_0;\mathcal{A})^{q-j} z_0 \right|_N \\
&\le (q+1) \max_{j \in \{0,1,2,\ldots,q\}} \frac{1}{(\mathrm{Re}\ \lambda_0 - \beta)^j} \left| [R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A})] R(\lambda_0;\mathcal{A})^{q-j} z_0 \right|_N \\
&\to 0 \qquad \text{as } N \to \infty.
\end{aligned}
$$

Therefore, it follows that

$$
\begin{aligned}
(4.14) \qquad &\left| [C(\tfrac{r}{N}\mathcal{A}_N)^k - S_N(t_k^N)]\, \pi_N R(\lambda_0;\mathcal{A})^{q+1} z_0 \right|_N \\
&\le \left| [C(\tfrac{r}{N}\mathcal{A}_N)^k - S_N(t_k^N)] [\pi_N R(\lambda_0;\mathcal{A})^{q+1} - R(\lambda_0;\mathcal{A}_N)^{q+1}\pi_N] z_0 \right|_N \\
&\quad + \left| [C(\tfrac{r}{N}\mathcal{A}_N)^k - S_N(t_k^N)] R(\lambda_0;\mathcal{A}_N)^{q+1}\pi_N z_0 \right|_N \\
&\le \left| [C(\tfrac{r}{N}\mathcal{A}_N)^k - S_N(t_k^N)] \right| \left| [R(\lambda_0;\mathcal{A}_N)^{q+1}\pi_N - \pi_N R(\lambda_0;\mathcal{A})^{q+1}] z_0 \right|_N \\
&\quad + \left| [C(\tfrac{r}{N}\mathcal{A}_N)^k - S_N(t_k^N)] R(\lambda_0;\mathcal{A}_N)^{q+1}\pi_N z_0 \right|_N
\end{aligned}
$$

$$\leq (M_0+Me^{\beta T})\left|[R(\lambda_0;\mathcal{A}_N)^{q+1}\pi_N-\pi_N R(\lambda_0;\mathcal{A})^{q+1}]z_0\right|_N + \gamma M_{\lambda_0}^{q+1}\left|z_0\right|_Z(\tfrac{r}{N})^q$$

$$\to 0 \text{ as } N \to \infty.$$

Hence

$$\left|[C(\tfrac{r}{N}\mathcal{A}_N)^k-S_N(t_k^N)]\pi_N z\right|_N \to 0 \qquad \text{as } N \to \infty$$

for each $z \in \mathcal{R}(R(\lambda_0;\mathcal{A})^{q+1})$. But $\mathcal{R}(R(\lambda_0;\mathcal{A})^{q+1}) = \mathcal{D}(\mathcal{A}^{q+1})$ is a dense subset of Z. Therefore, using the fact that the operators $[C(\frac{r}{N}\mathcal{A}_N)^k-S_N(t_k^N)]$, $k = 0,1,2,\ldots,\rho N$, are uniformly bounded in N for all N sufficiently large, we conclude that, given $\varepsilon > 0$, there exists an $\hat{N} = \hat{N}(\varepsilon, z_0)$ such that

$$\left|[C(\tfrac{r}{N}\mathcal{A}_N)^k-S_N(t_k^N)]\pi_N z_0\right|_N < \varepsilon, \qquad k = 0,1,2,\ldots,\rho N$$

for all $N > \hat{N}$ and each $z_0 \in Z$, which implies factor convergence.

□

While Theorem 4.9 above yields both necessary and sufficient conditions, it is only the sufficient conditions that are of practical importance. Indeed, the theorem will be applied to demonstrate factor convergence for a Discrete Approximation Scheme satisfying the required hypotheses via the generally more easily verified condition of factor stability.

The next corollary provides estimates for the rate of factor convergence for a factor stable Discrete Approximation Scheme satisfying the hypotheses of the equivalence theorem.

4.15. Corollary. Suppose $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ is a factor stable Discrete Approximation Scheme with property (P1) which satisfies the hypotheses of Theorem 4.9. Furthermore, suppose there is a set $\mathcal{S} \subseteq D_1 \subseteq \mathcal{D}(\mathcal{A})$ such that for each $z \in \mathcal{S}$ there exists a constant $\nu = \nu(z)$ for which

$$|[\mathcal{A}_N\pi_N - \pi_N\mathcal{A}]z|_N \leq \nu(z)(\tfrac{r}{N})^p \qquad \text{for some } p > 0.$$

Then for each $z_0 \in \mathcal{D}(\mathcal{A}^{q+1})$ for which $(\mathcal{A}-\lambda_0 I)^j z_0 \in \mathcal{S}$, $j = 0,1,2,\ldots,q$, there exist constants $k_0 = k_0(z_0)$ and $k_2 = k_2(z_0)$ depending on z_0 such that

$$|[C(\tfrac{r}{N}\mathcal{A}_N)^k - S_N(t_k^N)]\pi_N z_0|_N \leq k_0(\tfrac{r}{N})^p + k_2(\tfrac{r}{N})^q.$$

Proof: Let z_0 be as in the statement of the corollary and define

$$\nu_j(z_0) = \nu((\mathcal{A}-\lambda_0 I)^j z_0), \qquad j = 0,1,2,\ldots,q,$$

$$\nu^*(z_0) = \max_{0\leq j\leq q} \nu_j(z_0).$$

Then $z_0 \in \mathcal{D}(\mathcal{A}^{q+1}) = \mathcal{R}(R(\lambda_0;\mathcal{A})^{(q+1)})$ implies that $z_0 = R(\lambda_0;\mathcal{A})^{(q+1)}v_0$ for some $v_0 \in Z$. From (4.14) it follows that

$$|[C(\tfrac{r}{N}\mathcal{A}_N)^k - S_N(t_k^N)]\pi_N z_0|_N = |[C(\tfrac{r}{N}\mathcal{A}_N)^k - S_N(t_k^N)]\pi_N R(\lambda_0;\mathcal{A})^{(q+1)}v_0|_N$$

$$\leq (M_0 + Me^{\beta T})|[R(\lambda_0;\mathcal{A}_N)^{(q+1)}\pi_N - \pi_N R(\lambda_0;\mathcal{A})^{(q+1)}]v_0|_N + \gamma M_{\lambda_0}^{q+1}|v_0|_Z(\tfrac{r}{N})^q$$

$$\leq (M_0 + Me^{\beta T})\Big|\sum_{j=0}^{q} R(\lambda_0;\mathcal{A}_N)^j[R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A})]$$

$$\cdot R(\lambda_0;\mathcal{A})^{(q-j)}v_0\Big|_N + \gamma M_{\lambda_0}^{q+1}|v_0|_Z(\tfrac{r}{N})^q$$

$$\leq (M_0+Me^{\beta T}) \sum_{j=0}^{q} |R(\lambda_0;\mathcal{A}_N)^j| |[R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A})]$$

$$\cdot R(\lambda_0;\mathcal{A})^{(q-j)} v_0|_N + \gamma M_{\lambda_0}^{q+1} |v_0|_Z (\tfrac{r}{N})^q$$

$$\leq (M_0+Me^{\beta T}) \sum_{j=0}^{q} \frac{M}{(\mathrm{Re}\ \lambda_0-\beta)^j} |[R(\lambda_0;\mathcal{A}_N)\pi_N - \pi_N R(\lambda_0;\mathcal{A})]$$

$$\cdot R(\lambda_0;\mathcal{A})^{(q-j)} v_0|_N + \gamma M_{\lambda_0}^{q+1} |v_0|_Z (\tfrac{r}{N})^q$$

(4.16)

$$\leq (M_0+Me^{\beta T}) \sum_{j=0}^{q} \frac{M^2}{(\mathrm{Re}\ \lambda_0-\beta)^{j+1}} |[\mathcal{A}_N\pi_N - \pi_N\mathcal{A}]R(\lambda_0;\mathcal{A})R(\lambda_0;\mathcal{A})^{(q-j)} v_0|_N$$

$$+ \gamma M_{\lambda_0}^{q+1} |v_0|_Z (\tfrac{r}{N})^q$$

$$\leq (M_0+Me^{\beta T}) \sum_{j=0}^{q} \frac{M^2}{(\mathrm{Re}\ \lambda_0-\beta)^{j+1}} |[\mathcal{A}_N\pi_N - \pi_N\mathcal{A}]R(\lambda_0;\mathcal{A})^{(q+1-j)} v_0|_N$$

$$+ \gamma M_{\lambda_0}^{q+1} |v_0|_Z (\tfrac{r}{N})^q$$

$$\leq (M_0+Me^{\beta T}) \sum_{j=0}^{q} \frac{M^2}{(\mathrm{Re}\ \lambda_0-\beta)^{j+1}} \nu_j(z_0) (\tfrac{r}{N})^p + \gamma M_{\lambda_0}^{q+1} |v_0|_Z (\tfrac{r}{N})^q$$

$$\leq (M_0+Me^{\beta T}) M^2 \nu^*(z_0)(q+1) \left(\max_{j \in \{1,2,\ldots,q+1\}} (\mathrm{Re}\ \lambda_0-\beta)^{-j} \right) (\tfrac{r}{N})^p + \gamma M_{\lambda_0}^{q+1} |v_0|_Z (\tfrac{r}{N})^q$$

$$= k_0(z_0) (\tfrac{r}{N})^p + k_2(z_0) (\tfrac{r}{N})^q, \quad \text{for all N sufficiently large,}$$

where (4.16) follows from an application of the estimate given by (4.2).

□

Finally, we can summarize the above results in a theorem which will be suitable for application in the discussions below.

4.17. Theorem. Suppose $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ is a factor stable Discrete Approximation Scheme with property (P1) satisfying the hypotheses of Theorem 4.9. Furthermore, suppose that $\mathcal{S}$ is a subset of $\mathcal{D}(\mathcal{A}^2) \cap D_1$ for which

(1)' For each $z \in \mathcal{S}$ there exists a $K = K(z)$ such that

$$|[\mathcal{A}_N\pi_N - \pi_N\mathcal{A}]z|_N \leq K(z)/N^p$$

(2) There exists a subset $\mathcal{S}_1$ of $\mathcal{S}$ such that for $z \in \mathcal{S}_1$ and λ_0 with $\mathrm{Re}\ \lambda_0 > \beta$

(a) $S(t)z \in \mathcal{S}$, $t \in [0,T]$.

(b) $S(t)(\lambda_0 I - \mathcal{A})z \in \mathcal{S}$, $t \in [0,T]$

and the constants guaranteed by (1) for (a) and (b) are independent of $t \in [0,T]$.

Then $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ is factor convergent and, moreover, for each $z_0 \in \mathcal{S}_1 \cap \mathcal{D}(\mathcal{A}^{q+1})$ for which $(\lambda_0 I - \mathcal{A})^j z_0 \in \mathcal{S}$, $j = 0,1,2,\ldots,q$, there exist constants $k_1 = k_1(z_0)$ and $k_2 = k_2(z_0)$ which depend on z_0, such that

$$|[C(\tfrac{r}{N}\mathcal{A}_N)^k\pi_N - \pi_N S(t_k^N)]z_0|_N \leq k_1(\tfrac{r}{N})^p + k_2(\tfrac{r}{N})^q, \qquad k = 0,1,2,\ldots,\rho N.$$

Proof: Theorem 4.9 ensures that $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ is factor convergent. To verify the second proposition, let z_0 be as in the statement of the theorem. It follows that

$$(4.18) \quad |[C(\tfrac{r}{N}\mathcal{A}_N)^k\pi_N - \pi_N S(t_k^N)]z_0|_N$$

$$\leq |[C(\tfrac{r}{N}\mathcal{A}_N)^k\pi_N - S_N(t_k^N)\pi_N]z_0|_N + |[S_N(t_k^N)\pi_N - \pi_N S(t_k^N)]z_0|_N$$

$$\leq k_0(z_0)\left(\frac{r}{N}\right)^p + k_2(z_0)\left(\frac{r}{N}\right)^q + k(z_0)\left(\frac{r}{N}\right)^p$$

$$= k_1\left(\frac{r}{N}\right)^p + k_2\left(\frac{r}{N}\right)^q,$$

where $k_1 = k_1(z_0)$, and $k_2 = k_2(z_0)$.

The estimates of the bounds on the first and second terms on the right-hand side of (4.18) follow as consequences of Corollaries 4.15 and 4.5, respectively.

5. The Padé Approximations and Characterization of Factor Stable/Factor Convergent Discrete Approximation Schemes

Adopting the terminology employed in [12] and [15], we make the following definition.

5.1 Definition. We shall say that a rational function r(z) is acceptable with respect to the set $\{z\in C\colon \text{Re } z\leq 0\}$, or equivalently a member of the class $\mathfrak{A}_{\text{Re } z\leq 0}$, if

(5.2) (1) $|r(z)-e^z| = O(|z|^{q+1}), \quad z \to 0,\ q \geq 1$

(5.3) (2) $|r(z)| \leq 1, \quad z \in \{z\in C\colon \text{Re } z\leq 0\}$

Among the most widely known classes of rational function approximations to the exponential (rfae) which in addition provide acceptable subclasses are the Padé approximations [10], [37] defined by the formulae

$$P_{j,k}(z) = N_{j,k}(z)/D_{j,k}(z), \quad j,k = 1,2, \ldots$$

where

$$N_{j,k}(z) = \sum_{m=0}^{k} \frac{(j+k-m)!k!}{(j+k)!m!(k-m)!} z^m,$$

$$D_{j,k}(z) = \sum_{m=0}^{j} \frac{(j+k-m)!j!}{(j+k)!m!(j-m)!} (-z)^m.$$

For the purpose of reference, we state and in some cases prove the following propositions containing results, properties and identities relating to the Padé approximations. The proofs of Propositions 5.8 and 5.9 may be found in Ehle [12], together with the verification of Proposition 5.10 which is the primary result of that paper.

5.4. Proposition. deg $P_{j,k}(z) = k-j$, and

$$|P_{j,k}(z)-e^z| = O(|z|^{j+k+1}), \qquad z \to 0.$$

5.5. Proposition.

$$D_{j,k}(z) = N_{k,j}(-z), \qquad j,k \geq 0$$

5.6. Proposition.

$$P_{0,n}(z) = \sum_{k=0}^{n} \frac{z^k}{k!} = \sum_{k=0}^{n} \frac{a_k^n}{k!} (1+z)^k$$

where

(1) $a_k^n \geq 0$, $k = 0,1,2,\ldots,n$, all n

(2) $\sum_{k=0}^{n} (a_k^n/k!) = 1$

(3) $a_k^n = a_{k+1}^{n+1}$, $k = 0,1,2,\ldots,n$.

Proof: Let

$$\mathcal{V}^n \equiv \text{span}\{1,\ z,\ \frac{z^2}{2!},\ \frac{z^3}{3!},\ \ldots,\ \frac{z^n}{n!}\},$$

$$\mathcal{W}^n \equiv \text{span}\{1,\ (1+z),\ \frac{(1+z)^2}{2!},\ \ldots,\ \frac{(1+z)^n}{n!}\}\ .$$

Then if $\mathcal{M}_n\colon \mathcal{W}^n \longrightarrow \mathcal{V}^n$ is the coordinate transformation operator mapping $\mathcal{W}^n$ onto $\mathcal{V}^n$, a simple calculation reveals that $M_n\colon R^{n+1} \to R^{n+1}$, its matrix representation with respect to the bases defined above, is given by

$$M_n = \begin{bmatrix} 1 & 1 & \frac{1}{2!} & \frac{1}{3!} & \frac{1}{4!} & & \cdots & \frac{1}{n!} \\ 0 & 1 & 1 & \frac{1}{2!} & \frac{1}{3!} & \frac{1}{4!} & \cdots & \frac{1}{(n-1)!} \\ 0 & 0 & 1 & 1 & \frac{1}{2!} & \frac{1}{3!} & \cdots & \frac{1}{(n-2)!} \\ & & & & & 0 & 1 & 1 & \frac{1}{2!} \\ & & & & & 0 & 0 & 1 & 1 \\ & & & & & & 0 & 0 & 1 \end{bmatrix}$$

The inverse of M_n may be calculated, and is given by

$$M_n^{-1} = \begin{bmatrix} 1 & -1 & \frac{1}{2!} & \frac{-1}{3!} & \frac{1}{4!} & \cdots & (-1)^n\ \frac{1}{n!} \\ 0 & 1 & -1 & \frac{1}{2!} & \frac{-1}{3!} & & (-1)^{n-1}\ \frac{1}{(n-1)!} \\ & & & & & & \\ & & & & & & 1 \quad -1 \\ & & & & & & 0 \quad 1 \end{bmatrix}.$$

Upon inspection, it can be observed that the coordinate identification

$$P_{0,n}(z) \in \mathscr{P}^n \leftrightarrow (1,1,1,1,\ldots,1)^T \in R^{n+1}$$

holds. Then if

$$P_{0,n}(z) \in \mathscr{P}^n \leftrightarrow (a_0^n, a_1^n, \ldots, a_n^n)^T \in R^{n+1}$$

we have

$$(a_0^n, a_1^n, a_2^n, \ldots, a_n^n)^T = M_n^{-1}(1,1,1,\ldots,1)^T$$

or

$$(a_0^n, a_1^n, \ldots, a_n^n)^T = \begin{pmatrix} \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} - \frac{1}{5!} \cdots + (-1)^{n-1} \frac{1}{(n-1)!} + \frac{(-1)^n}{n!} \\ \vdots \\ \frac{1}{2!} - \frac{1}{3!} \\ \frac{1}{2!} \\ 0 \\ 1 \end{pmatrix}$$

We verify (1) by induction. For all n, $a_n^n = 1 \geq 0$, $a_{n-1}^n = 0 \geq 0$, $a_{n-2}^n = 1/2! \geq 0$, $a_{n-3}^n = \frac{1}{2!} - \frac{1}{3!} \geq 0$. Suppose $a_{n-(2k-1)}^n \geq 0$. Then

$$a_{n-2k}^n = a_{n-(2k-1)}^n + \frac{(-1)^{2k}}{(2k)!} = a_{n-(2k-1)}^n + \frac{1}{(2k)!} \geq 0,$$

and

$$a^n_{n-(2k+1)} = a^n_{n-(2k-1)} + (-1)^{2k}\frac{1}{(2k)!} + (-1)^{(2k+1)}\frac{1}{(2k+1)!} \geq 0.$$

The veracity of (2) follows from

$$1 = P_{0,n}(0) = \sum_{k=0}^{n} \frac{a^n_k}{k!}(1+0)^k = \sum_{k=0}^{n} \frac{a^n_k}{k!}.$$

Finally, for the verification of (3), we have

$$a^{n+1}_{k+1} = \frac{1}{2!} - \frac{1}{3!} + \frac{1}{4!} \ldots + (-1)^{(n+1)-(k+1)}\frac{1}{(n+1-(k+1))!}$$

$$= \frac{1}{2!} - \frac{1}{3!} \ldots + (-1)^{n-k}\frac{1}{(n-k)!} = a^n_k, \quad k = 0,1,2,\ldots,n-2$$

$$a^{n+1}_n = 0 = a^n_{n-1} \text{ and } a^{n+1}_{n+1} = 1 = a^n_n.$$

□

Techniques similar to those employed in the proof of Proposition 5.6 can be used to verify the following result as well (cf. [33]).

5.7. Proposition.

$$N_{j,k}(z) = \frac{1}{\binom{k+j}{j}} \sum_{m=0}^{k} \binom{k-m+(j-1)}{j-1} N_{0,m}(z) = \frac{1}{\binom{k+j}{j}} \sum_{m=0}^{k} \binom{k-m+(j-1)}{j-1} P_{0,m}(z)$$

and

$$\frac{1}{\binom{k+j}{j}} \sum_{m=0}^{k} \binom{k-m+(j-1)}{j-1} = 1.$$

5.8. Proposition. (Ehle) If for some $j, k \geq 0$, $N_{j,k}(z)$ has all of its zeros in the open left half-plane, then for all $m \geq j$, $N_{m,k}(z)$ has all of its zeros in the open left half-plane.

5.9. Proposition. (Ehle) For any $n \geq 0$, if $N_{n+1,n+1}(z)$ has all of its zeros in the left half-plane, then $N_{n,n+2}(z)$ also has all of its zeros in the left half-plane.

5.10. Proposition. (Ehle) The diagonal and first two subdiagonal entries in the Padé table of rfae are acceptable with respect to the set $\{z \in C: \text{Re } z \leq 0\}$. That is,

$$\{P_{n+1,n+1}(z)\},\ \{P_{n+1,n}(z)\},\ \{P_{n+2,n}(z)\} \in \mathfrak{A}_{\text{Re } z \leq 0},\ n = 0,1,2, \ldots .$$

It is Ehle's [12] conjecture that these are the only entries in the Padé table which are of class $\mathfrak{A}_{\text{Re } z \leq 0}$. Norsett [25] has substantiated this conjecture in the case of the third and fourth subdiagonal entries in the table, i.e. $\{P_{n+3,n}(z)\}$ and $\{P_{n+4,n}(z)\}$. More recently, Iserles [16], [17] has demonstrated that $\{P_{n,m}(z)\}$ is not acceptable if $n-m \not\equiv 2 \pmod 4$, $n \geq m+3$.

As is pointed out in [15], there are other classes of rfae in addition to the Padé approximations which have been investigated with regard to acceptability. In particular, the Norsett functions [24] with denominators of the form $(1 + \alpha z)^n$, a property desirable for computational efficiency, have been shown to contain an acceptable subclass. However, since the Padé rfae yield an acceptable subclass with an arbitrarily high degree of approximation, we are content to restrict our attention to them alone.

Von Neumann's theory of spectral sets will prove a useful tool in establishing the factor stability of certain Discrete Approximation Schemes of interest. The details of this theory, and the proof of Theorem 5.12 below, can be found in von Neumann's original work [23], Riesz, Sz.-Nagy [32] or Berberian [8]. Similar applications of von Neumann's theory of spectral sets appear in [15] and [21]. Let T be a bounded linear transformation on a Hilbert space H.

5.11. Definition. A set $\mathcal{G} \subset C$ (completed by the point at infinity) will be called a spectral set for the linear transformation T if (a) it is closed, (b) $\mathcal{G} \supseteq \sigma(T)$ and (c) for every rational function $u(z)$ satisfying the inequality $|u(z)| \leq 1$ for all $z \in \mathcal{G}$, we have that $||u(T)|| \leq 1$.

5.12. Theorem. A necessary and sufficient condition that the halfplane $\{z \in C: \text{Re } z \leq 0\}$ be a spectral set for the bounded linear transformation T is that

$$\text{Re}\langle Tf, f\rangle \leq 0$$

for all $f \in H$.

The next lemma, a modified version of the corollary to Theorem 6 in [15], will permit us to apply the above results in the characterization of factor convergent approximation schemes. Due to the importance of this lemma, a detailed version of the proof provided in [15] is included.

5.13. Lemma. Suppose

(1) $C(z) \in \mathfrak{A}_{\text{Re } z \leq 0}$

(2) $\{z \in C: \text{Re } z \leq 0\}$ is a spectral set for $T - \beta I$, where $\beta > 0$ and T is a bounded linear operator on a Hilbert space H.

Then $\left|C(\frac{r}{N}T)\right| \leq 1 + \beta Kr/N$ for some constant K which is independent of N.

Proof: C(z) a rational function, and $C(z) \in \mathfrak{A}_{\text{Re } z\leq 0}$ imply that C(z) must have finitely many poles, all lying in the right half-plane. Therefore there exists an $M > 0$ such that C(z) and C'(z) are analytic in $\{z\in C: \text{Re } z<M\}$. Now $C(z)\in \mathfrak{A}_{\text{Re } z\leq 0}$ implies that C(z) is bounded at ∞, and thus deg $C(z) \leq 0$. Moreover, deg C'(z) = (deg C(z))-1 implies that deg $C'(z) < 0$ and C'(z) is also bounded at ∞. Therefore, an application of the maximum modulus principle from the theory of functions of a complex variable guarantees the existence of a $K > 0$ for which $\left|C'(z)\right| \leq K$ for $z \in \{z\in C: \text{Re } z\leq M\}$. Let

$$f_N(z) \equiv \frac{C(z+\frac{r}{N}\beta) - C(z)}{\frac{r}{N}\beta} .$$

By the mean value theorem of differential calculus, we have

$$f_N(z) = \frac{C(z+\frac{r}{N}\beta) - C(z)}{\frac{r}{N}\beta} = C'(\xi), \qquad |\xi-z| \leq \frac{r}{N}\beta \ , \ \xi = \xi(z).$$

Therefore

$$\sup_{\text{Re } z\leq 0} |f_N(z)| = \sup_{\text{Re } z\leq 0} |C'(\xi(z))| \leq K$$

for all N sufficiently large (i.e. $\frac{r}{N}\beta < M$). Consider $\frac{1}{K}f_N(z)$ for N sufficiently large. It is a rational function, and moreover

$$\left|\frac{1}{K}f_N(z)\right| \leq 1 \quad \text{for all } z \in \{z\in C: \text{Re } z\leq 0\} \text{ and all N sufficiently large.}$$

Since $\{z\in C: \text{Re } z\leq 0\}$ a spectral set for $(T - \beta I)$ implies that it is also a spectral set for $\frac{r}{N}(T - \beta I)$, N = 1,2,... (cf. Theorem 5.12), we have

$$\left|\frac{1}{K}f_N(\tfrac{r}{N}(T-\beta I))\right| \le 1 \quad \text{or} \quad \left|f_N(\tfrac{r}{N}(T-\beta I))\right| \le K$$

for all N sufficiently large. This implies that

$$\left|\left|C(\tfrac{r}{N}T)\right| - \left|C(\tfrac{r}{N}(T-\beta I))\right|\right| \le \left|C(\tfrac{r}{N}T) - C(\tfrac{r}{N}(T-\beta I))\right| \le K\beta\tfrac{r}{N}$$

or

$$\left|C(\tfrac{r}{N}T)\right| \le \left|C(\tfrac{r}{N}(T-\beta I))\right| + \frac{K\beta r}{N}$$

for N sufficiently large. However, $C(z) \in \mathcal{U}_{\text{Re } z\le 0}$ and $\{z\in C\colon \text{Re } z\le 0\}$ a spectral set for $\frac{r}{N}(T-\beta I)$, $N = 1,2,\ldots$, implies that

$$\left|C(\tfrac{r}{N}(T-\beta I))\right| \le 1 .$$

Therefore,

$$\left|C(\tfrac{r}{N}T)\right| \le 1 + \frac{K\beta r}{N}$$

for N sufficiently large.

□

5.14. Lemma. Suppose that $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ is a Discrete Approximation Scheme with property (P1) satisfying:

(1) There exists an inner product $[\cdot,\cdot]_N$ on Z_N that generates a topology on Z_N equivalent to the standard Z_N inner product topology for which $\text{Re}[\mathcal{A}_N z^N, z^N]_N \le \beta[z^N, z^N]_N$ for each $z^N \in Z_N$ and all N sufficiently large with $\beta > 0$ independent of N;

(2) $\left|[\mathcal{A}_N\pi_N - \pi_N\mathcal{A}]z\right|_N \to 0$ as $N \to \infty$ for each $z \in D_1$, where D_1 is a dense subset of Z contained in $\mathcal{D}(\mathcal{A})$;

(3) There exists a $\lambda_0 \in C$ with $\mathrm{Re}\ \lambda_0 > \beta$ and D_2, a dense subset of Z such that $R(\lambda_0;\mathscr{A})D_2 \subseteq D_1$;

(4) $C(z) \in \mathfrak{A}_{\mathrm{Re}\ z\leq 0}$.

Then $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ is factor convergent.

Proof: In the light of the remarks in §2, condition (1) above and Z_N finite-dimensional are necessary and sufficient for $\mathscr{A}_N \in G(M,\beta)$ for all N sufficiently large. Therefore, if we can demonstrate that $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ is factor stable, an application of Theorem 4.9 will yield the desired result.

Let $||\cdot||_N$ represent the norm on Z_N which is induced by the inner product $[\cdot,\cdot]_N$ and which obeys the norm equivalence relation given by

$$\tilde{m}|\cdot|_N \leq ||\cdot||_N \leq \tilde{M}|\cdot|_N.$$

Condition (1) implies that $\mathrm{Re}[(\mathscr{A}_N-\beta I)z,z]_N \leq 0$ for all N sufficiently large. Thus Theorem 5.12 yields that $\{z\in C: \mathrm{Re}\ z\leq 0\}$ is a spectral set for the operators $\mathscr{A}_N-\beta I$ for all N sufficiently large. Moreover, $C(z) \in \mathfrak{A}_{\mathrm{Re}\ z\leq 0}$ and Lemma 5.13 imply

$$||C(\tfrac{r}{N}\mathscr{A}_N)||_N \leq 1 + \frac{\beta K r}{N}$$

for N sufficiently large. Therefore, for $k = 0,1,2,\ldots,\rho N$, and all N sufficiently large, we have

$$||C(\tfrac{r}{N}\mathscr{A}_N)^k||_N \leq ||C(\tfrac{r}{N}\mathscr{A}_N)||_N^k \leq (1 + \frac{\beta K r}{N})^k \leq ((1 + \frac{\beta K r}{N})^N)^\rho$$

$$\leq e^{\beta K r \rho} = e^{\beta K T}.$$

Hence

$$\left|C\left(\frac{r}{N}\mathscr{A}_N\right)^k\right|_N \leq \frac{\tilde{M}}{\tilde{m}} e^{\beta K T}, \quad k = 0,1,2,\ldots,\rho N \text{ and all } N \text{ sufficiently large,}$$

which implies that $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ is factor stable.

□

5.15. Lemma. Suppose that $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ is a Discrete Approximation Scheme with property (P1) satisfying

(1a) There exists an inner product induced norm $||\cdot||_N$ on Z_N generating an equivalent topology to the standard Z_N norm topology for which

$$||(I+\frac{r}{N}\mathscr{A}_N)||_N^2 \leq 1 + \alpha r/N,$$

for all N sufficiently large with $\alpha > 0$ independent of N. Suppose further that conditions (2), (3) and (4) of Lemma 5.14 are also satisfied by $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$. Then $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ is factor convergent.

The proof of Lemma 5.15 may be argued by demonstrating that condition (1a) implies condition (1). Indeed, if $||z||_N^2 = [z,z]_N$, it is not difficult to show that (cf. [33])

$$\mathrm{Re}[\mathscr{A}_N z, z]_N \leq \frac{\alpha}{2}[z,z]_N.$$

The next theorem serves to characterize a certain subclass of the Padé approximations and the factor convergent approximation schemes that it generates.

5.16. Theorem. Suppose that $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ is a Discrete Approximation Scheme with property (P1) satisfying either condition (1) of Lemma 5.14 or condition (1a) of Lemma 5.15 in addition to conditions (2) and (3) of Lemma 5.14. Then if $C(z) \in \{P_{n+1,n+1}(z)\}$ or $C(z) \in \{P_{n+1,n}(z)\}$ or

$C(z) \in \{P_{n+2,n}(z)\}$, $n = 1,2,\ldots$, where $P_{j,k}(z)$ represents the (j,k)th entry in the Padé table of rfae, the scheme $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ is factor convergent.

Proof: Proposition 5.10 and Lemmas 5.14, 5.15 above.

□

As we have seen, the construction of $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ with $C(z) \in \mathcal{U}_{\text{Re } z\leq 0}$ together with the theory of spectral sets will enable us to characterize large classes of factor stable and thereby factor convergent Discrete Approximation Schemes. However, if $C(z) \in \mathcal{U}_{\text{Re } z\leq 0}$, condition (5.3) deems it necessary that deg $C(z) \leq 0$; that is, if $C(z) = N(z)/D(z)$, then deg $D(z) \geq$ deg $N(z)$. Unfortunately, the restriction that deg $C(z) \leq 0$ precludes the investigation of many approximation schemes commonly encountered in practice, and often with many highly desirable properties, via this approach. In particular, explicit schemes, i.e. those for which deg $D(z) = 0$, are not acceptable in the sense of Definition 5.1, but are computationally desirable, since no operator inverse need be calculated. Fortunately, however, we shall be able to investigate a wider class of approximation schemes than those constructed with $C(z) \in \mathcal{U}_{\text{Re } z\leq 0}$ through the application of other techniques to be described below.

5.17. Theorem. Suppose $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ is a Discrete Approximation Scheme with property (P1) satisfying conditions (1a), (2) and (3). Then if $C(z) = P_{0,k}(z)$, $k = 1,2,\ldots$, $\{Z_N, \pi_N, \mathcal{A}_N, C(z)\}$ is factor convergent.

Proof: In the light of the arguments in support of Lemmas 5.14 and 5.15 and the fact that the $P_{0,k}(z)$ satisfy conditions (1b), (2b) and (3b)

of Theorem 4.9, we need only demonstrate that $\{Z_N, \pi_N, \mathscr{A}_N, P_{0,k}(z)\}$, $k = 1,2,\ldots$, are factor stable before we can apply Theorem 4.9 to obtain the desired result.

If we again assume the norm equivalence relation $\tilde{m}|\cdot|_N \leq ||\cdot||_N \leq \tilde{M}|\cdot|_N$, we have for $n = 0,1,2,\ldots\rho N$ and all N sufficiently large

$$||P_{0,k}(\tfrac{r}{N}\mathscr{A}_N)^n||_N = ||(\sum_{j=0}^{k} \frac{a_j^k}{j!}(I+\tfrac{r}{N}\mathscr{A}_N)^j)^n||_N \tag{5.18}$$

$$\leq (\sum_{j=0}^{k} \frac{a_j^k}{j!}||(I+\tfrac{r}{N}\mathscr{A}_N)^j||_N)^n \leq (\sum_{j=0}^{k} \frac{a_j^k}{j!}||(I+\tfrac{r}{N}\mathscr{A}_N)||_N^j)^n$$

$$\leq (\sum_{j=0}^{k} \frac{a_j^k}{j!}(1+\tfrac{\alpha r}{N})^j)^n \leq (\sum_{j=0}^{k} \frac{a_j^k}{j!} e^{\alpha rj/N})^n$$

$$\leq (\sum_{j=0}^{k} \frac{a_j^k}{j!} e^{\alpha rk/N})^n = e^{\frac{\alpha rk}{N}n} \leq e^{\alpha rk\rho} = e^{\alpha kT},$$

where (5.18) above follows from an application of Proposition 5.6. Thus

$$|P_{0,k}(\tfrac{r}{N}\mathscr{A}_N)^n|_N \leq \frac{\tilde{M}}{\tilde{m}} e^{\alpha kT}, \quad n = 0,1,2,\ldots,\rho N \text{ for all N sufficiently large,}$$

and hence $\{Z_N, \pi_N, \mathscr{A}_N, P_{0,k}(z)\}$, $k = 1,2,\ldots$, are factor stable.

□

5.19. Remark. For purpose of reference in the arguments that follow, the reader is requested to note that inequality (5.18) above implies that

$$||P_{0,k}(\tfrac{r}{N}\mathscr{A}_N)||_N \leq e^{\alpha rk/N}$$

as well.

5.20. Lemma. Suppose $\{Z_N, \pi_N, \mathscr{A}_N, P_{j,k}(z)\}$ is a Discrete Approximation Scheme with property (P1), where $P_{j,k}(z) = N_{j,k}(z)/D_{j,k}(z)$ is the (j,k)th

entry in the Padé table of rfae. Then if condition (1a) of Lemma 5.15 above is satisfied, we have that

$$|N_{j,k}(\tfrac{r}{N}\mathscr{A}_N)^n|_N \leq C_k^1, \qquad n = 0,1,2\ldots,\rho N$$

for all N sufficiently large and k = 0,1,2,..., where C_k^1 is independent of h, N and j.

Proof: Proposition 5.7 implies for n = 0,1,2,...,ρN, N sufficiently large that

$$||N_{j,k}(\tfrac{r}{N}\mathscr{A}_N)^n||_N = ||(\frac{1}{\binom{k+j}{j}}\sum_{m=0}^{k}\binom{k-m+(j-1)}{j-1}P_{0,m}(\tfrac{r}{N}\mathscr{A}_N))^n||_N$$

$$\leq (\frac{1}{\binom{k+j}{j}}\sum_{m=0}^{k}\binom{k-m+(j-1)}{j-1}||P_{0,m}(\tfrac{r}{N}\mathscr{A}_N)||_N)^n$$

$$\leq (\frac{1}{\binom{k+j}{j}}\sum_{m=0}^{k}\binom{k-m+(j-1)}{j-1}e^{\alpha rm/N})^n$$

$$\leq (\frac{1}{\binom{k+j}{j}}\sum_{m=0}^{k}\binom{k-m+(j-1)}{j-1}e^{\alpha rk/N})^n$$

$$= (e^{\alpha rk/N})^n \leq e^{\alpha rk\rho} = e^{\alpha kT}.$$

Therefore, assuming the norm equivalence relation in Lemma 5.14, we have

$$|N_{j,k}(\tfrac{r}{N}\mathscr{A}_N)^n|_N \leq \frac{\tilde{M}}{\tilde{m}}e^{\alpha kT} \equiv C_k^1, \quad n = 0,1,\ldots,\rho N, \text{ all N sufficiently large.}$$

□

5.21. Lemma. Suppose $\{Z_N,\pi_N,\mathscr{A}_N,P_{j,k}(z)\}$ is a Discrete Approximation Scheme with property (P1) where $P_{j,k}(z) = N_{j,k}(z)/D_{j,k}(z)$ is the (j,k)th entry in the Padé table of rfae. Then if there exist constants M,β for which either

(5.22) (1) $\mathscr{A}_N \in H(\omega,\beta,M)$ (cf. §2) for all N sufficiently large and some $\omega > 0$

or

(5.23) (2) $\mathscr{A}_N \in G(M,\beta)$ for all N sufficiently large and the roots of $D_{j,k}(z)$ are real,

we have, for $j \leq k+2$,

$$\left|D_{j,k}\left(\frac{r}{N}\mathscr{A}_N\right)^{-n}\right|_N \leq C_j^2, \quad n = 0,1,2,\ldots,\rho N \text{ for all N sufficiently large,}$$

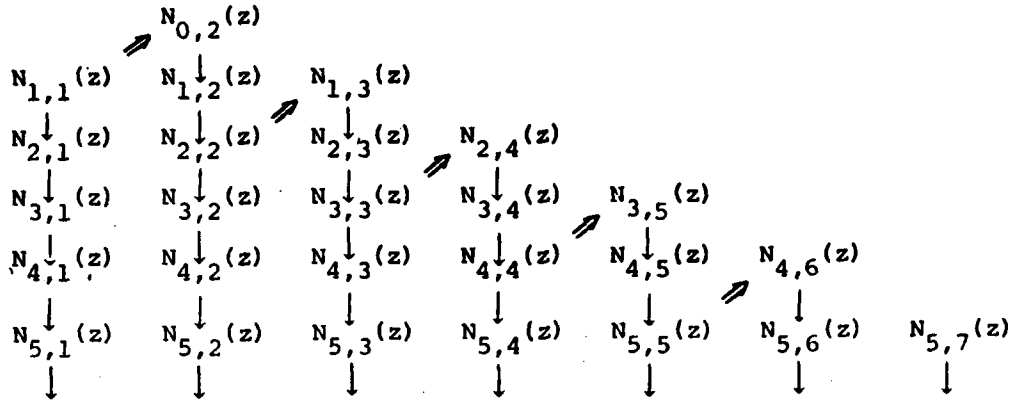
where C_j^2 is independent of N, n and k.

Proof: Propositions 5.8 and 5.9 imply that $N_{j,k}(z)$ with $k \leq j+2$ have their zeros in the open left half-plane. Indeed, $N_{1,1}(z) = 1 + \frac{1}{2}z$ has as its only zero $z = -2$. Then if we adopt the conventions that:

(a) $A \longrightarrow B$ denotes the implication that "A has its zeros lying in the open left half-plane implies B has its zeros lying in the open left half-plane" justified by Proposition 5.8;

(b) $A \Longrightarrow B$ denotes the same implication as in (a) justified by Proposition 5.9,

the following table can be constructed to substantiate the claim.



Proposition 5.5 implies that $D_{j,k}(z) = N_{j,k}(-z)$, which in turn implies that $D_{j,k}(-z)$ with $j \leq k+2$ has all of its zeros in the open left half-plane. Thus, $D_{j,k}(z)$ with $j \leq k+2$ has all of its zeros in the open right half-plane. Therefore, for $j \leq k+2$, it follows that

$$|D_{j,k}(z)| = \left|\frac{1}{\delta_j^k} \prod_{i=1}^{j} (\lambda_i^{j,k} - z)\right| \quad \text{with Re } \lambda_i^{j,k} > 0,\ i = 1,2,\ldots,j$$

and $\delta_j^k = \prod_{i=1}^{j} \lambda_i^{j,k}$.

Note: In order to simplify notation, the j, k superscripts and subscripts on δ_j^k and $\lambda_i^{j,k}$, $i = 1,2,\ldots,j$, will be suppressed in the discussion below.

Working formally, we find

$$|D_{j,k}(\tfrac{r}{N}\mathcal{A}_N)^{-n}|_N = |(\tfrac{1}{\delta}\prod_{i=1}^{j}(\lambda_i I - \tfrac{r}{N}\mathcal{A}_N))^{-n}|_N = |\delta|^n \left|\prod_{i=1}^{j}(\lambda_i I - \tfrac{r}{N}\mathcal{A}_N)^{-n}\right|_N$$

$$\leq |\delta|^n \prod_{i=1}^{j} |(\lambda_i I - \tfrac{r}{N}\mathcal{A}_N)^{-n}|_N \leq |\delta|^n \prod_{i=1}^{j} (\tfrac{N}{r})^n \, |(\tfrac{\lambda_i N}{r} I - \mathcal{A}_N)^{-n}|_N$$

Now, Re $\lambda_i > 0$, $i = 1,2,\ldots,j$, guarantees that $|\text{Arg } \lambda_i| < \pi/2$ and Re $\frac{N}{r}\lambda_i > \beta$, $i = 1,2,\ldots,j$, for N sufficiently large. Thus $\mathcal{A}_N \in G(M,\beta)$ or

$\mathcal{A}_N \in H(\omega,\beta,M)$ implies that the operator inverses in the preceding inequality exist for N sufficiently large. Furthermore, in the case that (5.22) holds, we have

$$(5.24)\qquad |\delta|^n \prod_{i=1}^{j} (\tfrac{N}{r})^n |(\tfrac{\lambda_i N}{r} I - \mathcal{A}_N)^{-n}|_N = |\delta|^n \prod_{i=1}^{j} (\tfrac{N}{r})^n |R(\tfrac{\lambda_i N}{r}; \mathcal{A}_N)^n|_N$$

$$\leq |\delta|^n \prod_{i=1}^{j} (\tfrac{N}{r})^n \frac{M}{|\frac{N}{r}\lambda_i - \beta|^n} \leq |\delta|^n \prod_{i=1}^{j} (\tfrac{N}{r})^n M(|\tfrac{N}{r}\lambda_i| - \beta)^{-n}$$

$$= |\delta|^n \prod_{i=1}^{j} M(|\lambda_i| - \tfrac{\beta r}{N})^{-n} = |\delta|^n \prod_{i=1}^{j} M|\lambda_i|^{-n}(1 - \tfrac{\beta r}{|\lambda_i| N})^{-n}$$

$$= |\delta|^n (\prod_{i=1}^{j} |\lambda_i|^{-n}) M^j \prod_{i=1}^{j} (1 - \tfrac{\beta r}{|\lambda_i| N})^{-n}$$

$$= |\delta^n| |\delta^{-n}| M^j \prod_{i=1}^{j} (1 - \tfrac{\beta r}{|\lambda_i| N})^{-n} = M^j \prod_{i=1}^{j} (1 - \tfrac{\beta r}{|\lambda_i| N})^{-n}$$

$$\leq M^j \prod_{i=1}^{j} (1 - \tfrac{\beta r}{|\lambda_i| N})^{-\rho N} = M^j (\prod_{i=1}^{j} (1 - \tfrac{\beta r}{|\lambda_i| N})^{-N})^{\rho} \leq C_j^2$$

for all N sufficiently large with C_j^2 independent of N. The calculation above follows from the fact that $\lim_{N\to\infty}(1 - \frac{\beta r}{|\lambda_i| N})^{-N} = \exp(\beta r/|\lambda_i|)$ implies that $(1 - \frac{\beta r}{|\lambda_i| N})^{-N}$, i = 1,2,...,j, are uniformly bounded in N.

When (5.23) holds, the appropriate steps in (5.24) are replaced by

$$|\delta^n| \prod_{i=1}^{j} (\tfrac{N}{r})^n |R(\tfrac{\lambda_i N}{r}; \mathcal{A}_N)^n|_N \leq |\delta^n| \prod_{i=1}^{j} (\tfrac{N}{r})^n \frac{M}{(\mathrm{Re}\, \frac{N}{r}\lambda_i - \beta)^n} = |\delta|^n \prod_{i=1}^{j} (\tfrac{N}{r})^n M(|\tfrac{N}{r}\lambda_i| - \beta)^{-n},$$

where we have used the assumption that $\lambda_i \in R$, i = 1,2,...,j. The remainder of the proof proceeds as in the previous case.

□

5.25. Theorem. Suppose that $\{Z_N, \pi_N, \mathscr{A}_N, P_{j,k}(z)\}$ is a Discrete Approximation Scheme having property (P1) where $P_{j,k}(z) = N_{j,k}(z)/D_{j,k}(z)$ is the (j,k)th entry in the Padé table of rfae and $j \leq k+2$. Suppose further that $\{Z_N, \pi_N, \mathscr{A}_N, P_{j,k}(z)\}$ satisfies conditions (1a), (2) and (3) referred to in Theorem 5.17 in addition to either

(4a) $\mathscr{A}_N \in H(\omega,\beta,M)$ for all N sufficiently large and some $\omega > 0$

or

(4b) $\{\lambda \in C: D_{j,k}(\lambda)=0\} \subset R$.

Then the scheme is factor convergent.

Proof: Once again, we need only demonstrate that $\{Z_N, \pi_N, \mathscr{A}_N, P_{j,k}(z)\}$ is factor stable. Lemmas 5.20 and 5.21 imply that

$$|P_{j,k}(\tfrac{r}{N}\mathscr{A}_N)^n|_N = |[D_{j,k}(\tfrac{r}{N}\mathscr{A}_N)^{-1}N_{j,k}(\tfrac{r}{N}\mathscr{A}_N)]^n|_N$$

$$= |D_{j,k}(\tfrac{r}{N}\mathscr{A}_N)^{-n}N_{j,k}(\tfrac{r}{N}\mathscr{A}_N)^n|_N \leq |D_{j,k}(\tfrac{r}{N}\mathscr{A}_N)^{-n}|_N |N_{j,k}(\tfrac{r}{N}\mathscr{A}_N)^n|_N$$

$$\leq C_j^2 C_k^1 \equiv C_j^k,$$

$n = 0,1,2,\ldots,\rho N$, for all N sufficiently large with C_j^k independent of N. □

A DAS of the form $\{Z_N, \pi_N, \mathscr{A}_N, P_{1,k}(z)\}$, $k = 0,1,2,\ldots$, satisfying conditions (1a), (2) and (3) stated above is factor convergent. Indeed, $D_{1,k}(z)$ is linear in z for each k and thus condition (4b) is satisfied as well. Unfortunately, it cannot be argued that this is also the case for DAS of the form

$\{Z_N, \pi_N, \mathscr{A}_N, P_{j,k}(z)\}$ with $j > 1$ and $k = j-2, j-1, 1, j+1, \ldots$. In fact, it can be demonstrated (cf. [33]) that $\{\lambda \in C: D_{2,k}(\lambda)=0\}$ consists of complex conjugate pairs for each k and $\{\lambda \in C: D_{3,k}(\lambda)=0\}$ consists of one real value and a complex conjugate pair for each k.

Finally, we can summarize the preceding results with the following theorem.

5.26. Theorem. Suppose $\{Z_N, \pi_N, \mathscr{A}_N, P_{j,k}(z)\}$ is a Discrete Approximation Scheme with property (P1) satisfying

(1) $\left| [\mathscr{A}_N \pi_N - \pi_N \mathscr{A}] z \right|_N \to 0$ as $N \to \infty$ for each $z \in D_1$, where D_1 is a dense subset of Z contained in $\mathscr{D}(\mathscr{A})$;

(2) There exists $\lambda_0 \in C$ with $\mathrm{Re}\ \lambda_0 > \beta$ and D_2, a dense subset of Z for which $R(\lambda_0; \mathscr{A}) D_2 \subseteq D_1$,

and $P_{j,k}(z)$ is the (j,k)th entry in the Padé table of rfae. In addition, consider the following supplementary hypotheses which may be satisfied by $\{Z_N, \pi_N, \mathscr{A}_N, P_{j,k}(z)\}$:

(-) $\mathscr{A}_N \in H(\omega, \beta, M)$ for all N sufficiently large and some $\omega > 0$ (M,β independent of N), together with condition (1a) of Theorem 5.17;

(|) Condition (1a) of Theorem 5.17;

(O) $\mathscr{A}_N \in G(M, \beta)$ for all N sufficiently large (M,β independent of N).

Then the scheme $\{Z_N, \pi_N, \mathscr{A}_N, P_{j,k}(z)\}$ is factor convergent under the additional hypothesis (-), (|) or (O) respectively if that symbol appears in the (j,k)th position of Figure 5.27 below.

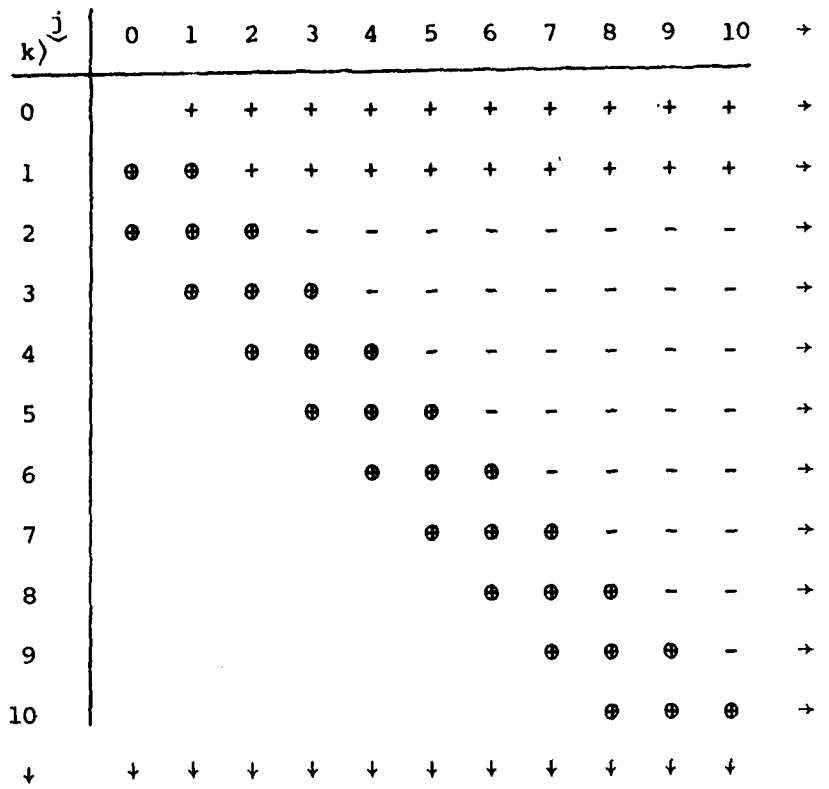


Figure 5.27.

5.28. Remark. It is important to note that we cannot make the claim that the results presented above represent a complete characterization of the factor stable/factor convergent Discrete Approximation Schemes constructed with the Padé rfae. In fact, numerical results indicate that stronger conclusions may indeed be drawn and that further investigation is warranted.

6. Averaging and Finite Difference Discrete Approximation Schemes

A Discrete Approximation Scheme $\{Z_N^A, \pi_N^A, \mathcal{A}_N^A, C(z)\}$ will be said to be of the Averaging and Finite Difference (AFD) type if it has been constructed in the following manner (cf. Definition 3.3). For each N = 1,2,..., let

$$\hat{Z}_A^N \equiv \text{span}\{{}^1\hat{\phi}_N^{(0)}, \ldots, {}^n\hat{\phi}_N^{(0)}, \ldots, {}^1\hat{\phi}_N^{(N)}, \ldots, {}^n\hat{\phi}_N^{(N)}\}$$

with

$${}^j\hat{\phi}_N^{(0)} = (\bar{e}_j, 0), \quad j = 1,2,\ldots,N, \qquad {}^j\hat{\phi}_N^{(k)} = (\bar{0}, \chi_k^N(\cdot)\bar{e}_j), \quad \begin{matrix} k = 1,2,\ldots,N \\ j = 1,2,\ldots,n \end{matrix}$$

where

$$\bar{e}_j = (0,0,\ldots,\overset{\text{jth}}{\underset{\downarrow}{1}},0,\ldots,0)^T \in R^n, \qquad \bar{0} = (0,0,\ldots,0) \in R^n,$$

0 is the R^n-valued 0-function defined on [-r,0], and $\chi_k^N(\cdot) \in L_2([-r,0);R)$,

$$\chi_k^N(t) = \chi_{[-k\frac{r}{N},-(k-1)\frac{r}{N})}(t) = \begin{cases} 1, & t \in [-k\frac{r}{N},-(k-1)\frac{r}{N}) \\ 0 & \text{otherwise,} \quad k = 1,2,\ldots,N. \end{cases}$$

6.1. Remark.

(i) $\hat{Z}_A^N$ is an n(N+1)-dimensional subspace of Z.

(ii) For k, $\ell \geq 1$, we have that

$$\langle {}^j\hat{\phi}_N^{(k)}, {}^i\hat{\phi}_N^{(\ell)} \rangle_Z = \int_{-r}^{0} \langle \chi_k^N(\theta)\bar{e}_j, \chi_\ell^N(\theta)\bar{e}_i \rangle_{R^n} \, d\theta$$

$$= \int_{-r}^{0} \chi_k^N(\theta)\chi_\ell^N(\theta)\bar{e}_j^T\bar{e}_i d\theta = 0$$

unless $k = \ell$ and $j = i$, in which case

$$\langle {}^j\hat{\phi}_N^{(k)}, {}^j\hat{\phi}_N^{(k)}\rangle_Z = \int_{-r}^{0} \langle \chi_k^N(\theta)\bar{e}_j, \chi_k^N(\theta)\bar{e}_j\rangle d\theta$$

$$= \int_{-r}^{0} \chi_k^N(\theta)\bar{e}_j^T\bar{e}_j d\theta = \int_{-k\frac{r}{N}}^{-(k-1)\frac{r}{N}} d\theta = \frac{r}{N}.$$

(iii) Clearly $\langle {}^j\hat{\phi}_N^{(0)}, {}^i\hat{\phi}_N^{(k)}\rangle_Z = 0$ unless $k = 0$, $j = i$, in which case $\langle {}^j\hat{\phi}_N^{(0)}, {}^j\hat{\phi}_N^{(0)}\rangle_Z = 1$.

For each $N = 1, 2, \ldots$

(1) Define Z_N^A by $Z_N^A = \bigtimes_0^N R^n$, where an element in Z_N^A is denoted by

$$\underline{\alpha} = (\bar{\alpha}_0, \bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_N) \in Z_N^A; \quad \bar{\alpha}_j \in R^n,\ j = 0,1,2,\ldots,N.$$

The mapping $\sigma_A^N: \hat{Z}_A^N \to Z_N^A$ is defined by

$$(6.2) \qquad \sigma_A^N\Big(\sum_{k=0}^{N}\sum_{j=1}^{n} \alpha_j^k\, {}^j\hat{\phi}_N^{(k)}\Big) = \sigma_A^N\Big(\bar{\alpha}_0, \sum_{k=1}^{N} \bar{\alpha}_k\chi_k^N(\cdot)\Big) = (\bar{\alpha}_0, \bar{\alpha}_1, \bar{\alpha}_2, \ldots, \bar{\alpha}_N) = \underline{\alpha}$$

where $\bar{\alpha}_i = (\alpha_1^i, \alpha_2^i, \ldots, \alpha_n^i)^T \in R^n$, $i = 0,1,2,\ldots,N$. Using the mutual orthogonality of the $\{{}^j\hat{\phi}_N^{(k)}\}$ (cf. Remark 6.1), we have for $\underline{\alpha}, \underline{\beta} \in Z_N^A$

$$\langle \underline{\alpha}, \underline{\beta}\rangle_N = \langle (\sigma^N)^{-1}\underline{\alpha}, (\sigma^N)^{-1}\underline{\beta}\rangle_Z$$

$$= \Big\langle \sum_{k=0}^{N}\sum_{j=1}^{n} \alpha_j^k\, {}^j\hat{\phi}_N^{(k)}, \sum_{k=0}^{N}\sum_{j=1}^{n} \beta_j^k\, {}^j\hat{\phi}_N^{(k)}\Big\rangle_Z$$

$$= \sum_{k=0}^{N} \sum_{j=1}^{n} \alpha_j^k \beta_j^k \langle {}^j\hat{\phi}_N^{(k)}, {}^j\hat{\phi}_N^{(k)} \rangle_Z$$

$$= \sum_{j=1}^{n} \alpha_j^0 \beta_j^0 + \sum_{k=1}^{N} \sum_{j=1}^{n} \alpha_j^k \beta_j^k \langle {}^j\hat{\phi}_N^{(k)}, {}^j\hat{\phi}_N^{(k)} \rangle_Z$$

$$= \sum_{j=1}^{n} \alpha_j^0 \beta_j^0 + \frac{r}{N} \sum_{k=1}^{N} \sum_{j=1}^{n} \alpha_j^k \beta_j^k$$

$$= \bar{\alpha}_0^T \bar{\beta}_0 + \frac{r}{N} \sum_{k=1}^{N} \bar{\alpha}_k^T \beta_k .$$

(2) Again using the mutual orthogonality of the $\{{}^j\hat{\phi}_N^{(k)}\}$, we calculate $\hat{P}_N^A$, the orthogonal projection of Z onto the subspace $\hat{Z}_A^N$. Indeed, for $(\bar{\eta},\phi) \in Z$:

$$\hat{P}_N^A(\bar{\eta},\phi) = \sum_{k=0}^{N} \sum_{j=1}^{n} \langle (\bar{\eta},\phi), {}^j\hat{\phi}_N^{(k)} / \left| {}^j\hat{\phi}_N^{(k)} \right|_Z \rangle_Z \, {}^j\hat{\phi}_N^{(k)} / \left| {}^j\hat{\phi}_N^{(k)} \right|_Z$$

$$= \sum_{j=1}^{n} \langle (\bar{\eta},\phi), (\bar{e}_j, 0) \rangle_Z \, {}^j\hat{\phi}_N^{(0)}$$

$$+ \sum_{k=1}^{N} \sum_{j=1}^{n} \frac{N}{r} \langle (\bar{\eta},\phi), (\bar{0}, \chi_k^N(\cdot)\bar{e}_j) \rangle_Z \, {}^j\hat{\phi}_N^{(k)}$$

$$= \sum_{j=1}^{n} (\bar{\eta}^T \bar{e}_j)(\bar{e}_j, 0) + \sum_{k=1}^{N} \sum_{j=1}^{n} \left(\frac{N}{r} \int_{-kr/N}^{-(k-1)\frac{r}{N}} \bar{e}_j^T \phi(\theta)\, d\theta \right) {}^j\hat{\phi}_N^{(k)}$$

$$= (\bar{\eta}, 0) + \left(\bar{0}, \sum_{k=1}^{N} \sum_{j=1}^{n} \left(\frac{N}{r} \int_{-kr/N}^{-(k-1)\frac{r}{N}} \bar{e}_j^T \phi(\theta)\, d\theta \right) \chi_k^N(\cdot)\bar{e}_j \right)$$

$$= \left(\bar{\eta}, \sum_{k=1}^{N} \sum_{j=1}^{n} \left(\frac{N}{r} \int_{-kr/N}^{-(k-1)\frac{r}{N}} \bar{e}_j^T \phi(\theta)\, d\theta \right) \chi_k^N(\cdot)\bar{e}_j \right)$$

$$= (\bar{\eta}, \sum_{k=1}^{N} (\frac{N}{r} \int_{-kr/N}^{-(k-1)\frac{r}{N}} \phi(\theta)d\theta)\chi_k^N(\cdot))$$

$$= (\bar{\eta}, \sum_{k=1}^{N} \bar{\phi}_k^N \chi_k^N(\cdot))$$

where

$$\bar{\phi}_k^N = \frac{N}{r} \int_{-kr/N}^{-(k-1)\frac{r}{N}} \phi(\theta)d\theta, \qquad k = 1,2,\ldots,N.$$

Representations for the mappings $\pi_N^A: Z \to Z_N^A$ and $(\pi_N^A)^{-1}: Z_N^A \to Z$ can also be calculated. For $(\bar{\eta},\phi) \in Z$ and $(\bar{v}_0,\bar{v}_1,\bar{v}_2,\ldots,\bar{v}_N) \in Z_N^A$ we have

$$(6.3) \qquad \pi_N^A(\bar{\eta},\phi) = \sigma_A^N \hat{P}_N^A(\bar{\eta},\phi) = \sigma_A^N(\bar{\eta}, \sum_{k=1}^{N} \bar{\phi}_k^N \chi_k^N(\cdot)) = (\bar{\eta},\bar{\phi}_1^N,\bar{\phi}_2^N,\ldots,\bar{\phi}_N^N)$$

and

$$(\pi_N^A)^{-1}(\bar{v}_0,\bar{v}_1,\ldots,\bar{v}_N) = (\sigma_A^N)^{-1}(\bar{v}_0,\bar{v}_1,\ldots,\bar{v}_N) = (\bar{v}_0, \sum_{k=1}^{N} \bar{v}_k \chi_k^N(\cdot)).$$

(3) Define $\mathcal{A}_N^A: Z_N^A \to Z_N^A$ as follows. Let

$$\ell(\cdot;N): \{0,1,2,\ldots,\nu\} \to \{0,1,2,\ldots,N\}$$

be given by

$$\ell(j;N) = k \quad \text{if } -\tau_j \in [\frac{-kr}{N}, -(k-1)\frac{r}{N}), \; j = 0,1,2,\ldots,\nu;$$

that is to say

$$-\tau_j \in [-\ell(j;N)\frac{r}{N}, -(\ell(j;N)-1)\frac{r}{N}), \qquad j = 0,1,2,\ldots,\nu.$$

Then for $\underline{v} \in Z_N^A$, $\underline{v} = (\bar{v}_0, \bar{v}_1, \bar{v}_2, \ldots, \bar{v}_N)$ define

$$\mathcal{A}_N^A \underline{v} = (L(N)\underline{v},\ D(N)\underline{v}) \tag{6.4}$$

where $L(N) \in \mathcal{B}(Z_N^A, R^n)$ is given by

$$L(N)\underline{v} = A_0 \bar{v}_0 + \sum_{j=1}^{\nu} A_j \bar{v}_{\ell(j,N)} + \sum_{j=1}^{N} \int_{-jr/N}^{-(j-1)\frac{r}{N}} D(\theta)\, d\theta\, \bar{v}_j,$$

and $D(N) \in \mathcal{B}(Z_N^A, \overset{N}{\underset{1}{\times}} R^n)$ is given (by its matrix representation with respect to the basis for Z_N^A discussed above) by

$$N \begin{bmatrix} & & & & \overset{N+1}{} \\ \frac{N}{r} & \frac{-N}{r} & 0 & \cdots & 0 \\ & \frac{N}{r} & \frac{-N}{r} & 0 \cdots & 0 \\ & & & & \\ & & & & \\ & & & \frac{N}{r} & \frac{-N}{r} \end{bmatrix} \otimes I_n$$

where I_n represents the $n \times n$ identity matrix and $\otimes$ is the Kronecker product. Hence

$$[\mathcal{A}_N^A \underline{v}]_j = \begin{cases} L(N)\underline{v}, & j = 0 \\ \frac{N}{r}(\bar{v}_{j-1} - \bar{v}_j), & j = 1,2,\ldots,N. \end{cases}$$

6.5. Remark. The justification for the characterization of the schemes defined above as averaging/finite difference should now be clear. Indeed, as

is evidenced in (6.3), $(\bar{\eta},\phi) \in Z$ is approximated by $\pi_N^A(\bar{\eta},\phi) = (\bar{\eta},\bar{\phi}_1^N,\bar{\phi}_2^N,\ldots,\bar{\phi}_N^N) \in Z_N^A$ where

$$\bar{\phi}_j^N = \frac{N}{r}\int_{-jr/N}^{-(j-1)\frac{r}{N}} \phi(\theta)\,d\theta$$

is the average value assumed by ϕ on the interval $[-j\frac{r}{N}, -(j-1)\frac{r}{N}]$. Furthermore, (6.4) reveals that for $\underline{v} \in Z_N^A$, an approximation to $\hat{\phi} = (\phi(0),\phi) \in \mathcal{D}(\mathscr{A})$, $\mathscr{A}_N^A\underline{v} = (L(N)\underline{v}, D(N)\underline{v}) \in Z_N^A$ approximates $\mathscr{A}\hat{\phi} = (L(\phi),D\phi) \in Z$ via a finite difference approximation of the differentiation operator.

6.6. Lemma. $\{Z_N^A,\pi_N^A,\mathscr{A}_N^A,C(z)\}$, an AFD Discrete Approximation Scheme as defined above, will have property (P1).

Proof: For each $(\bar{\eta},\phi) \in Z$ we must demonstrate

$$(6.7)\qquad \left|((\pi_N^A)^{-1}\pi_N^A - I)(\bar{\eta},\phi)\right|_Z = \left|\hat{P}_N^A(\bar{\eta},\phi) - (\bar{\eta},\phi)\right|_Z$$

$$= \left|(I-\hat{P}_N^A)(\bar{\eta},\phi)\right|_Z \to 0 \text{ as } N \to \infty.$$

Since $\hat{P}_N^A$ is an orthogonal projection, $|\hat{P}_N^A| \leq 1$, and thus the operators $\{(I-\hat{P}_N^A)\}$ are uniformly bounded in N. Therefore it suffices to demonstrate that (6.7) obtains on a dense subset of Z. In light of this we consider the subset $\mathscr{L}$ of Z defined by

$$\mathscr{L} = \{(\bar{\eta},\phi) \in Z : \phi \in C^1(-r,0)\}.$$

The set $\mathscr{L}$ is dense in Z by virtue of the fact that $C^1(-r,0)$ is a dense subset of $L_2(-r,0)$, and, for $(\bar{\eta},\phi) \in \mathscr{L}$, the veracity of (6.7) can be argued in a straightforward manner (cf. [33]).

Following Reber [29], we construct the following weighted inner product on Z_N^A. Let $k(\cdot,N)\colon \{0,1,2,\ldots,N-1\} \to \{1,2,\ldots,\nu\}$ be given by $k(j;N) = \min\{k \geq 1\colon -\tau_k \in \bigcup_{i=j+1}^{N} J_i\}$, where $J_i = [-ir/N, -(i-1)\frac{r}{N})$ and

$$g_j = g_j^N = \begin{cases} 1 & j = N \\ 1 + \sum_{i=k(j;N)}^{\nu} |A_i|, & j = 0,1,2,\ldots,N-1. \end{cases}$$

Then for $\underline{\alpha}, \underline{\beta} \in Z_N^A$, we define

$$[\underline{\alpha},\underline{\beta}]_N = \bar{\alpha}_0^T\bar{\beta}_0 + \frac{r}{N}\sum_{k=1}^{N} g_{k-1}\,\bar{\alpha}_k^T\bar{\beta}_k; \quad ||\underline{\alpha}||_N^2 = [\underline{\alpha},\underline{\alpha}]_N$$

Noting that $1 \leq g_0^N \leq (1 + \sum_{k=1}^{\nu} |A_k|)$, we have

$$|\cdot|_N \leq ||\cdot||_N \leq (1 + \sum_{k=1}^{\nu} |A_k|)^{1/2}|\cdot|_N,$$

and hence the two norms $|\cdot|_N$ and $||\cdot||_N$ on Z_N are equivalent.

The next lemma is essentially a restatement of Reber [29, Lemma 6.2] for the case of an autonomous system. The rather lengthy and technical proof of Lemma 6.8 has been omitted. The arguments can be found in their entirety in [29].

6.8. Lemma. For $\{Z_N^A, \pi_N^A, \mathcal{A}_N^A, C(z)\}$ an AFD Discrete Approximation Scheme as defined above, we have

$$||(I + \tfrac{r}{N}\mathcal{A}_N)||_N^2 \leq 1 + \alpha r/N$$

for all N sufficiently large with $\alpha > 0$ independent of N.

Recalling Lemma 5.15, we note that Lemma 6.8 will also imply the existence of a $\beta > 0$ for which

$$[\mathscr{A}_N^A \underline{z}, \underline{z}]_N \leq \beta [\underline{z}, \underline{z}]_N \qquad \text{for N sufficiently large}$$

where $\underline{z} = (\bar{z}_0, \bar{z}_1, \ldots, \bar{z}_N) \in Z_N^A$, and β is independent of N. A direct proof of this result can be given, and can be found in [33]. The required arguments are in the same spirit as those employed in the proof of Lemma 3.6 of [4].

6.9. Lemma. For $\{Z_N^A, \pi_N^A, \mathscr{A}_N^A, C(z)\}$ an AFD Discrete Approximation Scheme as defined above, we have

$$\left| [\mathscr{A}_N^A \pi_N^A - \pi_N^A \mathscr{A}] z \right|_N = K(z)(1/\sqrt{N}) \qquad \text{as } N \to \infty$$

for each $z \in \mathscr{D}(\mathscr{A}^2)$, where $K(z) = K((\phi(0), \phi)) = K(|\dot{\phi}|_\infty, |\ddot{\phi}|_{L_2})$.

The proof of Lemma 6.9 can be found in [33]. The arguments are similar to those used in the proofs of Lemma 3.2 and Corollary 3.1 of [4].

The next lemma enables us to apply Theorem 4.17 to AFD Discrete Approximation Schemes to establish estimates for the rate of factor convergence.

6.10. Lemma. Suppose $z_0 = (\phi(0), \phi) \in \mathscr{D}(\mathscr{A}^2)$ and $S(t) z_0 = (x(t), x_t)$, where $x(t)$ is the unique solution to

$$\dot{x}(t) = L(x_t)$$

$$x_0 = \phi.$$

Then there exist constants M_1, M_2 such that

(1) $|\dot{x}_\sigma|_\infty \leq M_1, \qquad \sigma \in [0,T]$

(2) $|\ddot{x}_\sigma|_{L_2} \leq M_2, \qquad \sigma \in [0,T].$

Proof: $z_0 \in \mathscr{D}(\mathscr{A}^2)$ implies that $\phi \in W_2^2(-r,0)$, with $\dot{\phi}(0) = L(\phi)$, which in turn implies [7] that $x \in W_2^2(-r,T) \subset C^1(-r,T)$. Therefore $\dot{x}(\cdot)$ is a continuous function on the compact set $[-r,T]$. Hence, there exists an M_1 such that $|\dot{x}|_\infty \leq M_1$, or $|\dot{x}_\sigma|_\infty \leq M_1$, for $\sigma \in [0,T]$.

To verify the second proposition, we note that $x \in W_2^2(-r,T)$ implies the existence of an M_2 such that $\int_{-r}^{T} |\ddot{x}(\sigma)|^2 d\sigma \leq M_2^2$. Therefore, for $\sigma \in [0,T]$ we have

$$|\ddot{x}_\sigma|_{L_2}^2 = \int_{-r}^{0} |\ddot{x}_\sigma(\theta)|^2 d\theta = \int_{-r}^{0} |\ddot{x}(\sigma+\theta)|^2 d\theta = \int_{\sigma-r}^{\sigma} |\ddot{x}(u)|^2 du$$

$$\leq \int_{-r}^{T} |\ddot{x}(u)|^2 du \leq M_2^2.$$

□

Finally, we apply the theory developed in the preceding two sections to AFD approximation schemes in order to characterize a class of factor convergent schemes of this type.

6.11. Theorem. Suppose $\{Z_N^A, \pi_N^A, \mathscr{A}_N^A, P_{j,k}(z)\}$ is an AFD Discrete Approximation Scheme as defined above, where $P_{j,k}(z)$ is the (j,k)th entry in the Padé table of rfae. Then if either

(6.12) (1) $j = 0$, or 1, with k arbitrary, $k = 1,2,\ldots$

or

(6.13) (2) $j = k, k+1, k+2, \ldots, \quad k = 1,2,\ldots,$

the scheme is factor convergent. Furthermore, for such j and k

$$\left| [P_{j,k}(\tfrac{r}{N}\mathscr{A}_N^A)^n \pi_N^A - \pi_N^A S(t_n^N)]z_0 \right|_N = O(1/\sqrt{N}) + O(1/N^{j+k}),$$

$n = 0,1,2,\ldots,\rho N$, for each $z_0 \in \mathscr{D}(\mathscr{A}^{j+k+2})$, where the constants in the $O(\cdot)$ term necessarily depend on z_0.

Proof: Lemmas 6.8 and 6.9 above, together with the choice $D_1 = D_2 = \mathscr{D}(\mathscr{A}^2)$ and an application of Theorem 5.26, guarantee that the scheme is factor convergent. Let us now consider the estimates for the rate of factor convergence. We choose $\mathscr{S}_1 = \mathscr{D}(\mathscr{A}^{j+k+2})$ and $\mathscr{S} = \mathscr{D}(\mathscr{A}^2)$. Then, for $\lambda_0 \in C$ with $\operatorname{Re} \lambda_0 > \beta$, we have $(\lambda_0 I - \mathscr{A})^i z_0 \in \mathscr{S}$, $i = 0,1,2,\ldots,j+k$. Furthermore, upon inspection of the constant $K(z_0)$ in the statement of Lemma 6.9, it is immediately seen that Lemma 6.10 implies that $K(S(t)z_0)$, $t \in [0,T]$, and $K(S(t)(\lambda_0 I - \mathscr{A})z_0)$, $t \in [0,T]$, are independent of t. Thus all of the hypotheses of Theorem 4.17 are satisfied, and the desired conclusion obtains.

□

6.14. Remark. In practice, it is observed that AFD approximation schemes satisfying conditions weaker than those stated in (6.12) and (6.13) factor converge. Two possible explanations for the observed behavior of these schemes can be offered.

(1) The $\mathscr{A}_N^A$ as defined above are, in actuality, contained in $H(\omega,\beta,M)$ for some $\omega > 0$ and all N sufficiently large. Unless the $\mathscr{A}_N^A$ are negative definite self-adjoint operators on Z_N^A (cf. Kato [18], Krein [19]), which is clearly not the case, this condition is in general difficult to verify. If, in fact, it could be demonstrated that $\mathscr{A}_N^A \in H(\omega,\beta,M)$ for all N sufficiently large, based upon an application of Theorem 5.26 stronger conclusions could be drawn.

(2) Figure 5.27 does not represent a complete characterization of

factor stable Discrete Approximation Schemes constructed with the Padé rfae.

Both of these conjectures remain, at present, unsubstantiated.

7. Spline and Variational Discrete Approximation Schemes

For each N = 1,2,..., let

$$\hat{Z}_S^N \equiv \text{span}\{\hat{\phi}_N^{(0)}, \hat{\phi}_N^{(1)}, \ldots, \hat{\phi}_N^{(k_N)}\}, \tag{7.1}$$

where $\hat{\phi}_N^{(j)} \in \mathcal{D}(\mathcal{A}) \subset Z$, $j = 0,1,2,\ldots,k_N$. Note that $\hat{Z}_S^N$ is a (k_N+1)-dimensional subspace of Z. Then a Spline/Variational (SPV) Discrete Approximation Scheme can be defined as follows:

(1) $$\{Z_N^S, <\cdot,\cdot>_N\} = \{\sigma_S^N \hat{Z}_S^N, <(\sigma_S^N)^{-1}(\cdot), (\sigma_S^N)^{-1}(\cdot)>_Z\} = \{R^{k_N+1}, <(\sigma_S^N)^{-1}(\cdot), (\sigma_S^N)^{-1}(\cdot)>_Z\}$$

where σ_S^N is the canonical isomorphism which associates with each element in $\hat{Z}_S^N$ its coordinate vector in R^{k_N+1} determined by the basis defined in (7.1);

(2) $\pi_N^S: Z \to Z_N^S$ and $\pi_N^{-1}: Z_N^S \to Z$ are given by

$\pi_N^S = \sigma_S^N \hat{P}_N^g$ and $(\pi_N^S)^{-1} = (\sigma_S^N)^{-1}$ respectively,

where $\hat{P}_N^g$ is the orthogonal projection from Z onto $\hat{Z}_S^N$ with respect to the Z_g inner product defined in §2;

(3) $\mathcal{A}_N^S: Z_N^S \to Z_N^S$, $\mathcal{A}_N^S \equiv \pi_N^S \mathcal{A} (\pi_N^S)^{-1}$, $N = 1,2,\ldots$.

7.2. Remark. In the case that $\{\phi_N^{(j)}\}$ is an orthogonal basis, we have for $\bar{\alpha}, \bar{\beta} \in Z_N^S$

$$\langle \bar{\alpha}, \bar{\beta} \rangle_N = \sum_{j=0}^{k_N} |\hat{\phi}_N^{(j)}|_Z^2 \alpha_j \beta_j$$

and for $(\eta, \phi) \in Z$

$$\pi_N^S(\eta,\phi) = \sum_{j=0}^{k_N} |\hat{\phi}_N^{(j)}|_Z^{-2} \langle (\eta,\phi), \hat{\phi}_N^{(j)} \rangle_Z \sigma_S^N \hat{\phi}_N^{(j)}$$

$$= \sum_{j=0}^{k_N} |\hat{\phi}_N^{(j)}|_Z^{-2} \langle (\eta,\phi), \hat{\phi}_N^{(j)} \rangle_Z \bar{e}_j,$$

where $\bar{e}_j = (0,0,\ldots,0,\underset{\text{jth}}{1},0,\ldots,0) \in R^{k_N+1}$.

7.3. Definition. We shall say that an (SPV) Discrete Approximation Scheme has property (P2) if for some integer $k \geq 1$ we have

(1) $\lim_{N\to\infty} L(\phi^N) = L(\phi)$ in R^n

(2) $\lim_{N\to\infty} D\phi^N = D\phi$ in $L_2(-r,0)$

for all $\phi \in C^k(-r,0)$, where ϕ^N is defined by the relation $\hat{P}_N^g \hat{\phi} = \hat{P}_N^g(\phi(0),\phi) = (\phi^N(0),\phi^N) \in \hat{Z}_S^N$.

As in the case of (AFD) approximations, we define a special inner product and associated induced norm on Z_N^S which generates an equivalent topology to the topology generated by the standard Z_N^S inner product. For $\bar{\alpha}, \bar{\beta} \in Z_N$, let

$$g[\bar{\alpha},\bar{\beta}]_N = \langle (\sigma^N)^{-1}\bar{\alpha}, (\sigma^N)^{-1}\bar{\beta} \rangle_{g'g} \quad ||\bar{\alpha}||_N^2 = g[\bar{\alpha},\bar{\alpha}]_N,$$

where $\langle \cdot,\cdot \rangle_g$ is the inner product on Z defined in §2.

7.4. Lemma. For $\{Z_N^S, \pi_N^S, \mathcal{A}_N^S, C(z)\}$ an SPV Discrete Approximation Scheme as defined above we have

$$g[\mathcal{A}_N^S \bar{z}, \bar{z}]_N \leq \beta_g[\bar{z}, \bar{z}]_N, \qquad N = 1,2,\ldots,$$

for $\bar{z} \in Z_N^S$ with $\beta > 0$ independent of N.

Proof: Using the dissipativeness of $\mathcal{A}-\beta I$ with respect to the g inner product (cf. §2), we have

$$g[\mathcal{A}_N^S \bar{z}, \bar{z}]_N = g[\pi_N^S \mathcal{A}(\pi_N^S)^{-1}\bar{z}, \bar{z}]_N = g[\sigma_S^N \hat{P}_N^g \mathcal{A}(\sigma_S^N)^{-1}\bar{z}, \bar{z}]_N$$

$$= \langle (\sigma_S^N)^{-1}\sigma_S^N \hat{P}_N^g \mathcal{A}(\sigma_S^N)^{-1}\bar{z}, (\sigma_S^N)^{-1}\bar{z} \rangle_g = \langle \hat{P}_N^g \mathcal{A}(\sigma_S^N)^{-1}\bar{z}, (\sigma_S^N)^{-1}\bar{z} \rangle_g$$

$$= \langle \mathcal{A}(\sigma_S^N)^{-1}\bar{z}, \hat{P}_N^g (\sigma_S^N)^{-1}\bar{z} \rangle_g = \langle \mathcal{A}(\sigma_S^N)^{-1}\bar{z}, (\sigma_S^N)^{-1}\bar{z} \rangle_g$$

$$\leq \beta \langle (\sigma_S^N)^{-1}\bar{z}, (\sigma_S^N)^{-1}\bar{z} \rangle_g = \beta g[\bar{z}, \bar{z}]_N.$$

□

7.5. Lemma. For $\{Z_N^S, \pi_N^S, \mathcal{A}_N^S, C(z)\}$ an SPV Discrete Approximation Scheme with property (P2) we have

$$|[\mathcal{A}_N^S \pi_N^S - \pi_N^S \mathcal{A}] z_0|_N \to 0 \qquad \text{as } N \to \infty$$

for each $z_0 \in \mathcal{D}(\mathcal{A}^{k+1})$, where k is as in the statement of Definition 7.3.

Proof: $\hat{\phi} = (\phi(0),\phi) \in \mathcal{D}(\mathcal{A}^{k+1})$ implies $\phi^{(k)}$ absolutely continuous. Therefore

$$|[\mathcal{A}_N^S \pi_N^S - \pi_N^S \mathcal{A}]\hat{\phi}|_N = |[\pi_N^S \mathcal{A} (\pi_N^S)^{-1} \pi_N^S - \pi_N^S \mathcal{A}]\hat{\phi}|_N$$

(7.6)

$$\leq |[\mathcal{A} P_N^g - \mathcal{A}]\hat{\phi}|_Z$$

(7.7)

$$= (|L(\phi^N) - L(\phi)|^2_{R^n} + |D\phi^N - D\phi|^2_{L_2})^{1/2} \to 0 \text{ as } N \to \infty,$$

where the estimate in (7.6) follows from Lemma 3.7 and the one in (7.7) from property (P2).

□

If we select $D_1 = D_2 = \mathcal{D}(\mathcal{A}^{k+1})$, Lemmas 7.4 and 7.5 above, together with Lemma 5.14 and Theorem 5.16, yield the following theorem and its corollary.

7.8. Theorem. Suppose $\{Z_N^S, \pi_N^S, \mathcal{A}_N^S, C(z)\}$ is an SPV Discrete Approximation Scheme with properties (P1) and (P2). Then if $C(z) \in \mathfrak{A}_{\text{Re } z \leq 0}$, the scheme is factor convergent.

7.9. Corollary. Suppose $\{Z_N^S, \pi_N^S, \mathcal{A}_N^S, P_{j,k}(z)\}$ is an SPV Discrete Approximation Scheme with properties (P1) and (P2) where $P_{j,k}(z)$ is the (j,k)th entry in the Padé table of rfae. Then if j = k, k+1 or k+2 and k > 0 is arbitrary, the scheme is factor convergent.

As a particular example we consider the case of the $\hat{\phi}_N^{(j)} = (\phi_N^{(j)}(0), \phi_N^{(j)})$ chosen with the $\phi_N^{(j)}$ as first-order spline functions (cf. [27], [34]). We make the following definitions.

(1) Let

$$\hat{Z}_1^N \equiv \text{span}\,\{{}^1\hat{e}_N^{(0)},\ {}^2\hat{e}_N^{(0)},\ \ldots,\ {}^n\hat{e}_N^{(0)},\ {}^1\hat{e}_N^{(1)},\ \ldots,\ {}^1\hat{e}_N^{(N)},\ \ldots,\ {}^n\hat{e}_N^{(N)}\},$$

where

$$\bar{e}_j = (0,0,\ldots,0,\underset{j\text{th}}{1},0,\ldots,0) \in R^n$$

and

$${}^j\hat{e}_N^{(k)} = (e_N^k(0),\ e_N^k(\cdot)\bar{e}_j)$$

with $e_N^k(\cdot)$ denoting the scalar-valued first-order spline function on $[-r,0]$ characterized by the relation

$$e_N^k(\tfrac{-ir}{N}) = \delta_{ik}, \qquad i,k = 0,1,\ldots,N.$$

Then $Z_N^1 = \sigma_S^N \hat{Z}_1^N$, where σ_S^N is the canonical isomorphism defined above.

(2) The mapping $\pi_N^1\colon Z \to Z_N^1$ is defined by

$$\pi_N^1 = \sigma_S^N\ {}^1\hat{P}_N^g$$

where ${}^1\hat{P}_N^g$ is the orthogonal projection associated with the subspace $\hat{Z}_1^N$ with respect to the Z_g inner product. The right inverse of π_N^1, $(\pi_N^1)^{-1}\colon Z_N^1 \to Z$, is defined by

$$(\pi_N^1)^{-1} = (\sigma_S^N)^{-1}.$$

(3) The operators $\mathscr{A}_N^1$: $Z_N^1 \to Z_N^1$ are given by

$$\mathscr{A}_N^1 = \pi_N^1 \mathscr{A} (\pi_N^1)^{-1}.$$

Using the well-known properties of interpolatory splines [34] and the fact that $\hat{\phi}_N \equiv {}^1\hat{P}_N^g \hat{\phi}$ for $\hat{\phi} \in Z$ satisfies a variational condition by virtue of the fact that ${}^1\hat{P}_N^g$ is an orthogonal projection, we have that the following result obtains. The details of the proof, which are omitted, can be found in [6], Theorem 4.1.

7.10. Lemma. The SPV Discrete Approximation Scheme $\{Z_N^1, \pi_N^1, \mathscr{A}_N^1, C(z)\}$ defined above will have properties (P1) and (P2).

As a consequence of Theorem 7.8 we have

7.11 Theorem. If, in the SPV Discrete Approximation Scheme $\{Z_N^1, \pi_N^1, \mathscr{A}_N^1, C(z)\}$, $C(z)$ is chosen from among those rational functions in the class $\mathfrak{A}_{\text{Re } z \leq 0}$, then the scheme is factor convergent.

In order to estimate the rate of factor convergence for the linear spline approximation scheme defined above, we rely on results established in [6]. Remark 4.1 of that paper guarantees that

$$(7.12) \qquad \left| [\mathscr{A}_N^1 \pi_N^1 - \pi_N^1 \mathscr{A}] \hat{\phi} \right|_N = O(1/N) \qquad \text{as } N \to \infty$$

for each $\hat{\phi} \in \{(\phi(0), \phi) : \phi \in C^{2-r,0})\} \supset \mathscr{D}(\mathscr{A}^3)$, where the dependence on $\hat{\phi}$ of the

constant K in the O term in (7.12) can be expressed by

$$(7.13) \qquad K = K(\hat{\phi}) = K(|\dot{\phi}|_{\infty}, |\ddot{\phi}|_{L_2}).$$

Furthermore, the nature of the dependence in (7.13) is such that Lemma 6.10 guarantees that for $z_0 \in \mathcal{D}(\mathcal{A}^3)$ and $\lambda_0 \in C$ with $\mathrm{Re}\ \lambda_0 > \beta$, $K(S(t)z_0)|_{t\in[0,T]}$ and $K(S(t)(\lambda_0 I - \mathcal{A})z_0)|_{t\in[0,T]}$ are independent of $t \in [0,T]$. Thus, by analogy to Theorem 6.11, we have the following result:

7.14. Theorem. If, in the SPV Discrete Approximation Scheme $\{Z_N^1, \pi_N^1, \mathcal{A}_N^1, P_{j,k}(z)\}$ as defined above, $P_{j,k}(z)$ is the (j,k)th entry in the Padé table of rfae, with j = k, k+1 or k+2 and k > 0 is arbitrary, then the scheme is factor convergent. Furthermore, for such j

$$|[P_{j,k}(\tfrac{r}{N}\mathcal{A}_N^1)^n \pi_N^1 - \pi_N^1 S(t_n^N)]z_0|_N = O(1/N) + O(1/N^{j+k}),$$

$n = 0,1,2,\ldots,\rho N$, as $N \to \infty$, for each $z_0 \in \mathcal{D}(\mathcal{A}^{j+k+3})$. The constants in the $O(\cdot)$ terms necessarily depend on z_0.

7.15. Remark. Further improvement in the rate of factor convergence can be achieved via the formulation of SPV Discrete Approximation Schemes employing bases composed of higher-order spline functions. In particular, if cubic splines are used, the SPV Discrete Approximation Scheme $\{Z_N^3, \pi_N^3, \mathcal{A}_N^3, P_{j,k}(z)\}$ is factor convergent for j = k, k+1, k+2 and k > 0 arbitrary. For such j and k, it can be further established that

$$\left| [P_{j,k}(\tfrac{r}{N}\mathscr{A}_N^3)^n \pi_N^3 - \pi_N^3 S(t_n^N)]z_0 \right|_N = O(1/N^3) + O(1/N^{j+k})$$

$n = 0,1,2,\ldots,\rho N$ as $N \to \infty$ for each $z_0 \in \mathscr{D}(\mathscr{A}^{j+k+5})$.

7.16. Remark. Unfortunately, it appears that it is not possible to prove a result analogous to Lemma 6.8 for SPV Discrete Approximation Schemes in general. That is, it cannot be demonstrated that

$$_g\left|\left|(I+\tfrac{r}{N}\mathscr{A}_N^S)\right|\right|_N^2 \leq 1 + \alpha r/N \qquad \text{for N sufficiently large with } \alpha > 0 \text{ and independent of N.}$$

In fact, for all test examples considered, the approximation schemes $\{Z_N^1, \pi_N^1, \mathscr{A}_N^1, P_{0,k}(z)\}$, k = 1, 2 and $\{Z_N^3, \pi_N^3, \mathscr{A}_N^3, P_{0,k}(z)\}$, k = 1, 2, 3, exhibited behavior characteristic of numerical instability when actually programmed and tested on the computer. Indeed, they did not factor converge. On the other hand, however, it was observed, again in all test examples considered, that the Discrete Approximation Schemes $\{Z_N^1, \pi_N^1, \mathscr{A}_N^1, P_{0,k}(z)\}$ with $k \geq 3$ and $\{Z_N^3, \pi_N^3, \mathscr{A}_N^3, P_{0,k}(z)\}$ with $k \geq 4$ were factor convergent, and as expected with significantly improved rates of factor convergence with increasing k. On the basis of this numerical evidence, we conclude that many interesting open questions remain regarding the characterization of SPV Discrete Approximation Schemes employing Padé rfae. Furthermore, in the light of the computational desirability of explicit schemes, i.e. those for which $C(z) = P_{0,k}(z)$, k = 1, 2, ..., these questions are clearly an important direction for future research.

8. Approximation of the Solutions to the Non-Homogeneous and Nonlinear Initial-Value Problems

We now turn our attention to the construction of approximate solutions to the non-homogeneous FDE initial-value problem given by

$$\dot{x}(t) = L(x_t) + f(t), \qquad t \in [0,T] \tag{8.1}$$

$$(x(0),x_0) = (\eta,\phi) = z_0 \tag{8.2}$$

where the hypotheses satisfied by L, f, η, ϕ and x have been specified and discussed in detail in §2. The procedure by which this is achieved is the extension of the results in §3 so as to yield approximations to

$$z(t) = S(t)z_0 + \int_0^t S(t-\sigma)(f(\sigma),0)\,d\sigma, \qquad t \in [0,T] \tag{8.3}$$

We recall (cf. §2) that the expression given in (8.3) yields a solution to the FDE initial-value problem (8.1), (8.2) via the equivalence discussed above.

We begin with several rather technical definitions.

8.4. Definition. For $f \in L_2(0,T]$ we define the parameterized family of operators $T(t;f)\colon Z \to Z$, $t \in [0,T]$ by

$$T(t;f)z = S(t)z + \int_0^t S(t-\sigma)(f(\sigma),0)\,d\sigma.$$

8.5. Definition. $\{Z_N,\pi_N,\mathcal{Q}_N,C(z),D(z)\}$, $N = 1,2,\ldots$, is said to be a Discrete Approximation Scheme for the perturbed problem (DASP) (i.e. for the non-homogeneous initial-value problem (8.1), (8.2) if

(1) $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$, $N = 1,2,\ldots$, is a Discrete Approximation Scheme for the homogeneous initial-value problem (3.1), (3.2) (cf. Definition 3.3);

(2) $D(z)$ is a rational function of the complex variable z.

The scheme $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$ is said to have property (P1) if $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ has property (P1) in the sense of Definition 3.6.

8.6. Definition. For $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$ a DASP, $f \in L_2(0,T)$ and each positive integer N, let the family of operators $G_k(\frac{r}{N}; f): Z_N \to Z_N$, $k = 0,1,2,\ldots,\rho N$, be defined recursively via the relations

$$G_0(\tfrac{r}{N}; f) = I$$

$$G_k(\tfrac{r}{N}; f) z^N = C(\tfrac{r}{N}\mathscr{A}_N)\, G_{k-1}(\tfrac{r}{N}; f) z^N + \tfrac{r}{N} D(\tfrac{r}{N}\mathscr{A}_N)\, \pi_N(p_k^N f, 0), \qquad k = 1,2,\ldots,\rho N, \tag{8.7}$$

where $z^N \in Z_N$ and the family of transformations

$$p_k^N: L_2(0,T) \to R^n, \qquad k = 1,2,\ldots,\rho N$$

operating on $f \in L_2(0,T)$ represent a discretization of the function f on the interval $[0,T]$. Different applications of the schemes to be developed below require various choices for the $\{p_k^N\}$. For the present discussion, we define the $\{p_k^N\}$ to be integral averaging operators. That is, the p_k^N are given by

$$p_k^N f = \frac{N}{r} \int_{(k-1)r/N}^{kr/n} f(\sigma)\, d\sigma, \qquad k = 1,2,\ldots,\rho N.$$

Further comments regarding the selection of the $\{p_k^N\}$ are included in the remarks at the conclusion of §9. The recurrence relation (8.7) can be solved to yield

$$G_k(\tfrac{r}{N};f)z^N = C(\tfrac{r}{N}\mathcal{A}_N)^k z^N + \tfrac{r}{N}\sum_{j=1}^{k} C(\tfrac{r}{N}\mathcal{A}_N)^{k-j} D(\tfrac{r}{N}\mathcal{A}_N)\pi_N(p_j^N f,0),$$

$$k = 0,1,2,\ldots,\rho N.$$

The operators $C(\frac{r}{N}\mathcal{A}_N)$ and $D(\frac{r}{N}\mathcal{A}_N)$ are again referred to with the implicit assumption that if the degree of the polynomial in the denominator of $C(z)$ or $D(z)$ is greater than zero, then the required inverses exist. Sufficient conditions for the existence of $C(\frac{r}{N}\mathcal{A}_N)$ have been provided in §4, while the existence of $D(\frac{r}{N}\mathcal{A}_N)$ is considered in §10 when the role of the rational function $D(z)$ is discussed.

8.8. Definition. A DASP $\{Z_N,\pi_N,\mathcal{A}_N,C(z),D(z)\}$ is said to be factor stable if

(1) The Discrete Approximation Scheme for the homogeneous initial-value problem, $\{Z_N,\pi_N,\mathcal{A}_N,C(z)\}$, is factor stable in the sense of Definition 3.4;

(2) For each $z \in Z$, we have

$$\left|[D(\tfrac{r}{N}\mathcal{A}_N)\pi_N - \pi_N I]z_0\right|_N \to 0 \qquad \text{as } N \to \infty.$$

8.9. Remark. For a factor stable DASP $\{Z_N,\pi_N,\mathcal{A}_N,C(z),D(z)\}$, the strong convergence of $\{D(\frac{r}{N}\mathcal{A}_N)\}$ to the identity required by condition (2), and an application of the Uniform Boundedness Principle, imply that the operators on Z_N, $D(\frac{r}{N}\mathcal{A}_N)$, are uniformly bounded in N.

8.10. Definition. The DASP $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$ is said to be factor convergent at $z_0 \in Z$ as an approximation to $z(t)$ as given by (8.3) if for each $\varepsilon > 0$, there exists an $\hat{N} = \hat{N}(z_0, f, \varepsilon)$ such that

$$\left| G_k\left(\frac{r}{N}; f\right) \pi_N z_0 - \pi_N T\left(t_k^N; f\right) z_0 \right|_N < \varepsilon, \qquad k = 0,1,2,\ldots,\rho N$$

for all $\hat{N} > N$. The scheme is said to be factor convergent if it is factor convergent at each $z_0 \in Z$.

That a factor convergent DASP does indeed yield an approximate solution to the non-homogeneous or nonlinear FDE initial-value problem is guaranteed by the next lemma.

8.11. Lemma. Suppose $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$ is a factor convergent DASP with property (P1). Then for each $z_0 = (\eta, \phi) \in Z$ and each $\varepsilon > 0$ there exists a $\hat{N} = \hat{N}(\varepsilon, z_0)$ such that

$$\left| x(t_k^N) - p_1(\pi_N^{-1} z_k^N) \right|_{R^n} < \varepsilon, \qquad k = 0,1,2,\ldots,\rho N$$

for all $N > \hat{N}$, where $z_k^N = G_k\left(\frac{r}{N}; f\right) \pi_N z_0$ and $x(t)$ denotes the unique solution to the FDE initial-value problem, (8.1), (8.2).

The proof of the preceding lemma, which is essentially indistinguishable from the proof of Lemma 3.9, has been omitted.

To a certain extent, the techniques and arguments employed in the succeeding sections parallel Thompson's [35] results in his extension of the classical Lax equivalence theorem (cf. [31]) to finite difference approxima-

tions for non-homogeneous and quasi-linear initial-value problems for parabolic partial differential equations. However, unlike the treatment in [35], we are able to exploit the fact that for each fixed $t \in [0,T]$ the non-homogeneous perturbation term which appears in the abstract formulation of the FDE lies in a finite-dimensional space (cf. [3]) and hence are able to obtain stronger results via simplified arguments.

9. Factor Convergence of Discrete Approximation Schemes for the Non-Homogeneous Problem

We demonstrate that for an appropriately constructed DASP $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$, factor stability implies factor convergence. Consider the DASP $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$. For each $N = 1,2,\ldots$ and each $t \in [0,T]$, we define the following parameterized families of bounded linear operators with domain R^n and range in Z_N. For $\eta \in R^n$, let

(i) $\hat{T}_N(t)\eta \equiv \pi_N S(t)(\eta,0)$

(ii) $$\hat{S}_N(t)\eta \equiv \begin{cases} \pi_N S(t_k^N)(\eta,0) & \text{if } t \in [\frac{kr}{N}, \frac{(k+1)r}{N}),\ k = 0,1,2,\ldots,\rho N-1 \\ \pi_N S(T)(\eta,0) & \text{if } t = T \end{cases}$$

(iii) $$\hat{C}_N(t)\eta \equiv \begin{cases} C(\frac{r}{N}\mathscr{A}_N)^k \pi_N(\eta,0) & \text{if } t \in [\frac{kr}{N}, \frac{(k+1)r}{N}),\ k = 0,1,2,\ldots,\rho N-1 \\ C(\frac{r}{N}\mathscr{A}_N)^{\rho N} \pi_N(\eta,0) & \text{if } t = T \end{cases}$$

(iv) $\hat{D}_N\eta \equiv D(\frac{r}{N}\mathscr{A}_N)\pi_N(\eta,0)$

(v) $\hat{I}_N\eta \equiv \pi_N(\eta,0)$.

9.1. Lemma. Let $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$ be a factor stable DASP with property (P1) for which $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ satisfies the hypotheses of Theorem 4.9. Then for each $t \in [0,T]$, we have

(i) $||\hat{T}_N(t) - \hat{S}_N(t)|| \to 0$ as $N \to \infty$,

(ii) $||\hat{S}_N(t) - \hat{C}_N(t)|| \to 0$ as $N \to \infty$,

(iii) $||\hat{D}_N - \hat{I}_N|| \to 0$ as $N \to \infty$,

where the norm in (i), (ii) and (iii) above is that one which is generated by the uniform operator topology on $\mathscr{B}(R^n, Z_N)$.

Proof: For $t \in [0,T]$ and each $N = 1,2,\ldots$, let $k_N(t)$ be defined to be that integer in the set $\{0,1,2,\ldots,\rho N\}$ for which $t \in [k_N(t)r/N, ((k_N(t)+1)r)/N)$. Then for each $t \in [0,T]$ and each $\eta \in R^n$, we have

$$(9.2) \quad |[\hat{T}_N(t) - \hat{S}_N(t)]\eta|_N = |[\pi_N S(t) - \pi_N S(t^N_{k_N(t)})](\eta,0)|_N$$

$$\leq |[S(t) - S(t^N_{k_N(t)})](\eta,0)|_Z \to 0 \text{ as } N \to \infty;$$

$$(9.3) \quad |[\hat{S}_N(t) - \hat{C}_N(t)]\eta|_N = |[\pi_N S(t^N_{k_N(t)}) - C(\tfrac{r}{N}\mathscr{A}_N)^{k_N(t)}\pi_N](\eta,0)|_N$$

$$\to 0 \text{ as } N \to \infty;$$

$$(9.4) \quad |[\hat{D}_N - \hat{I}_N]\eta|_N = |D(\tfrac{r}{N}\mathscr{A}_N)\pi_N - \pi_N I](\eta,0)|_N \to 0 \text{ as } N \to \infty,$$

where (9.2) follows from the uniform continuity of $S(\cdot)z$ on compact intervals for each $z \in Z$, (9.3) from the factor convergence of $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ guaranteed by Theorem 4.9, and (9.4) from the assumption of factor stability of $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$. Recalling that strong convergence of linear operators is equivalent to uniform convergence if the domain of the operators is a finite-dimensional space, we obtain the desired conclusion immediately.

□

9.5. Lemma. Suppose $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$ is a factor stable DASP with property (P1) for which $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ satisfies the hypotheses of Theorem 4.9. Then for $f \in L_2(0,T)$, we have that the scheme is factor convergent at $z_0 = \theta = (0,0)$, the zero element of Z, and moreover the convergence is uniform in f, for f in bounded subsets of $L_2(0,T)$.

Proof: For $k = 0,1,2,\ldots,\rho N$, we have

$$(9.6) \qquad \left|G_k\left(\tfrac{r}{N};f\right)\pi_N\theta - \pi_N T\left(t_k^N;f\right)\theta\right|_N$$

$$= \left|\frac{r}{N}\sum_{j=1}^{k} C\left(\tfrac{r}{N}\mathscr{A}_N\right)^{k-j} D\left(\tfrac{r}{N}\mathscr{A}_N\right)\pi_N\left(\mathfrak{I}_j^N f,0\right) - \pi_N\int_0^{t_k^N} S\left(t_k^N-\sigma\right)\left(f(\sigma),0\right)d\sigma\right|_N$$

$$\leq \left|\frac{r}{N}\sum_{j=1}^{k} C\left(\tfrac{r}{N}\mathscr{A}_N\right)^{k-j} D\left(\tfrac{r}{N}\mathscr{A}_N\right)\pi_N\left(\frac{N}{r}\int_{(j-1)r/N}^{jr/N} f(\sigma)\,d\sigma,0\right)\right.$$

$$\left. - \frac{r}{N}\sum_{j=1}^{k} C\left(\tfrac{r}{N}\mathscr{A}_N\right)^{k-j}\pi_N\left(\frac{N}{r}\int_{(j-1)r/N}^{jr/N} f(\sigma)\,d\sigma,0\right)\right|_N$$

$$+ \left|\frac{r}{N}\sum_{j=1}^{k} C\left(\tfrac{r}{N}\mathscr{A}_N\right)^{k-j}\pi_N\left(\frac{N}{r}\int_{(j-1)r/N}^{jr/N} f(\sigma)\,d\sigma,0\right) - \pi_N\int_0^{t_k^N} S\left(t_k^N-\sigma\right)\left(f(\sigma),0\right)d\sigma\right|_N$$

$$= \Big| \sum_{j=1}^{k} C\left(\frac{r}{N}\mathcal{A}_N\right)^{k-j} \left[D\left(\frac{r}{N}\mathcal{A}_N\right)\pi_N - \pi_N I\right] \Big(\int_{(j-1)r/N}^{jr/N} f(\sigma)\,d\sigma, 0 \Big) \Big|_N$$

$$+ \Big| \sum_{j=1}^{k} \int_{(j-1)r/N}^{jr/N} C\left(\frac{r}{N}\mathcal{A}_N\right)^{k-j} \pi_N (f(\sigma), 0)\,d\sigma$$

$$- \int_{0}^{t_k^N} \pi_N S(t_k^N - \sigma)(f(\sigma), 0)\,d\sigma \Big|_N$$

$$\leq \sum_{j=1}^{k} \left| C\left(\frac{r}{N}\mathcal{A}_N\right)^{k-j} \right| \left| \left[\hat{D}_N - \hat{I}_N\right] \int_{(j-1)r/N}^{jr/N} f(\sigma)\,d\sigma \right|_N$$

$$+ \Big| \sum_{j=1}^{k} \int_{(j-1)r/N}^{jr/N} \hat{C}_N(t_k^N - \sigma) f(\sigma)\,d\sigma - \int_{0}^{t_k^N} \hat{T}_N(t_k^N - \sigma) f(\sigma)\,d\sigma \Big|_N$$

$$\leq M_0 \left\| \hat{D}_N - \hat{I}_N \right\| \sum_{j=1}^{k} \left| \int_{(j-1)r/N}^{jr/N} f(\sigma)\,d\sigma \right| + \left| \int_{0}^{t_k^N} \left[\hat{C}_N(t_k^N - \sigma) - \hat{T}_N(t_k^N - \sigma)\right] f(\sigma)\,d\sigma \right|_N$$

$$\leq M_0 \left\| \hat{D}_N - \hat{I}_N \right\| \int_{0}^{T} |f(\sigma)|\,d\sigma + \int_{0}^{t_k^N} \left\| \left[\hat{C}_N(t_k^N - \sigma) - \hat{T}_N(t_k^N - \sigma)\right] \right\| |f(\sigma)|\,d\sigma$$

$$\leq M_0 \left\| \hat{D}_N - \hat{I}_N \right\| T^{1/2} \left(\int_{0}^{T} |f(\sigma)|^2 d\sigma \right)^{1/2} + \int_{0}^{t_k^N} \left\| \left[\hat{C}_N(t_k^N - \sigma) - \hat{S}_N(t_k^N - \sigma)\right] \right\| |f(\sigma)|\,d\sigma$$

$$+ \int_{0}^{t_k^N} \left\| \left[\hat{S}_N(t_k^N - \sigma) - \hat{T}_N(t_k^N - \sigma)\right] \right\| |f(\sigma)|\,d\sigma$$

$$\leq M_0 T^{1/2} ||\hat{D}_N - \hat{I}_N|| \; |f|_{L_2}$$

$$+ \left(\int_0^T ||\hat{C}_N(T-\sigma) - \hat{S}_N(T-\sigma)||^2 d\sigma\right)^{1/2} \left(\int_0^T |f(\sigma)|^2 d\sigma\right)^{1/2}$$

$$+ \left(\int_0^T ||\hat{S}_N(T-\sigma) - \hat{T}_N(T-\sigma)||^2 d\sigma\right)^{1/2} \left(\int_0^T |f(\sigma)|^2 d\sigma\right)^{1/2}$$

$$= M_0 T^{1/2} ||\hat{D}_N - \hat{I}_N|| \; |f|_{L_2} + \left(\int_0^T ||\hat{C}_N(T-\sigma) - \hat{S}_N(T-\sigma)||^2 d\sigma\right)^{1/2} |f|_{L_2}$$

$$+ \left(\int_0^T ||\hat{S}_N(T-\sigma) - \hat{T}_N(T-\sigma)||^2 d\sigma\right)^{1/2} |f|_{L_2},$$

where the constant M_0 (guaranteed by the assumption of factor stability) denotes the uniform bound on the operators $C(\frac{r}{N}\mathcal{A}_N)^k$, $k = 0,1,2,\ldots,\rho N$ for all N sufficiently large.

Recalling the definition of $\hat{T}_N(t)$, $\hat{S}_N(t)$ and $\hat{C}_N(t)$ for each $N = 1,2,\ldots$ and each $t \in [0,T]$, it can be verified that

$$||\hat{C}_N(t) - \hat{S}_N(t)|| \leq M_0 + Me^{\beta T}$$

and

$$||\hat{S}_N(t) - \hat{T}_N(t)|| \leq 2Me^{\beta T}.$$

Therefore, if we apply Lemma 9.1 and the Dominated Convergence Theorem to the final estimate in (9.6), the desired result follows immediately.

□

A straightforward application of the triangle inequality together with Lemma 9.5 and Theorem 4.9 yields the following result.

9.7. Theorem. Suppose $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$ is a factor stable DASP with property (P1) for which $\{Z_N, \pi_N, \mathscr{A}_N, C(z)\}$ satisfies the hypotheses of Theorem 4.9. Then for $f \in L_2(0,T)$, we have that the scheme is factor convergent and moreover the convergence is uniform in f for f in bounded subsets of $L_2(0,T)$.

9.8. Remark. The fact that factor convergence is uniform in f for f in bounded subsets of $L_2(0,T)$ plays an essential role in the application of these schemes to the approximate solution of optimal control problems for systems governed by hereditary systems of the form which we have been considering (cf. Banks and Burns [4], Reber [29]).

9.9. Remark. In certain applications, choices of $\{p_k^N\}$, $k = 0,1,2,\ldots,\rho N$ other than the integral averaging operators employed in the arguments above are more desirable. In particular, in the case of system identification problems (cf. Banks, Burns and Cliff [5]) the relevant input functions f are frequently contained in the class of piecewise continuous functions on [0,T] (PC(0,T)). In this instance, the appropriate choice for the $\{p_k^N\}$ is given by

$$\hat{p}_k^N f = f(t_k^N), \qquad k = 1,2,\ldots,\rho N. \tag{9.10}$$

While it is possible to demonstrate factor convergence for appropriately constructed DASP employing the $\{\hat{p}_k^N\}$ defined in (9.10), we note that it

may no longer be the case that convergence is uniform in f for f in bounded subsets of PC(0,T). However, for problems involving parameter identification the convergence obtained for such $\{p_k^N\}$ is adequate.

10. Making an Appropriate Choice for D(z)

For a DASP $\{Z_N, \pi_N, \mathscr{A}_N, C(z), D(z)\}$ it is clear from the results presented in §4 that the rational function C(z) should be chosen as an approximation to the exponential function e^z. Indeed, if this is the case, under the additional hypotheses specified in the statement of Theorem 9.7, $C(\frac{r}{N}\mathscr{A}_N)^k$ yields an approximation to S(kr/N) and factor convergence obtains. In this section we consider criteria according to which the rational function D(z) can be selected. In view of the results of the preceding section, in addition to the requirement that the operators $D(\frac{r}{N}\mathscr{A}_N)$ exist, at least for all N sufficiently large, it is necessary that the factor stability condition be satisfied. That is, for each $z_0 \in Z$ we require that

$$\left|D(\tfrac{r}{N}\mathscr{A}_N)\pi_N z_0 - \pi_N I z_0\right|_N \to 0, \qquad N \to \infty.$$

While $D(z) \equiv 1$, i.e. $D(\frac{r}{N}\mathscr{A}_N) = I$, the identity operator on Z_N, would satisfy these conditions, it has been observed in practice that other satisfactory D(z) are available which yield an improved rate of factor convergence and approximate solutions with enhanced accuracy. The heuristic argument which follows will serve to motivate these ideas further.

The basis for the approximation schemes we have constructed is that

$$C(\tfrac{r}{N}\mathscr{A}_N)\pi_N z_{k-1} + \tfrac{r}{N}D(\tfrac{r}{N}\mathscr{A}_N)\pi_N(p_k^N f, 0)$$

should in some sense approximate

$$S(\tfrac{r}{N})z_{k-1} + \int_{t_{k-1}^N}^{t_k^N} S(t_k^N-\sigma)(f(\sigma),0)d\sigma.$$

In particular,

$$C(\tfrac{r}{N}\mathcal{A}_N)\pi_N z_{k-1} \sim S(\tfrac{r}{N})z_{k-1}$$

and

$$\tfrac{r}{N} D(\tfrac{r}{N}\mathcal{A}_N)\pi_N(p_k^N f,0) \sim \int_{t_{k-1}^N}^{t_k^N} S(t_k^N-\sigma)(f(\sigma),0)d\sigma. \tag{10.1}$$

Recalling that $D(\frac{r}{N}\mathcal{A}_N)$ and π_N are bounded linear operators, we rewrite expression (10.1) as

$$\int_{t_{k-1}^N}^{t_k^N} D(\tfrac{r}{N}\mathcal{A}_N)\pi_N(f(\sigma),0)d\sigma \sim \int_{t_{k-1}^N}^{t_k^N} S(t_k^N-\sigma)(f(\sigma),0)d\sigma. \tag{10.2}$$

Inspection of (10.2) reveals that $D(z)$ should be chosen so that $D(\frac{r}{N}\mathcal{A}_N)$ approximates $S(t_k^N-\sigma)$; $\sigma \in (t_{k-1}^N, t_k^N)$, or equivalently $S(t)$; $t \in (0, \frac{r}{N})$. Consequently we consider $D(z)$ of the form

$$D(z) = P_{j,k}(\lambda z)$$

where once again $P_{j,k}(z)$ denotes the (j,k)th entry in the Padé table of rational function approximations to the exponential and λ is a fixed constant between 0 and 1. The parameter λ included in the definition of $D(z)$ serves to compensate

for the fact that for each N = 1,2,..., the operators $\{S(t): t \in [0,\frac{r}{N}]\}$ are to be approximated by the single operator $D(\frac{r}{N}\mathscr{A}_N)$. The mean value theorem from elementary calculus suggests one possible choice for λ,

$$D(\tfrac{r}{N}\mathscr{A}_N) = P_{j,k}(\lambda\tfrac{r}{N}\mathscr{A}_N) = \frac{N}{r}\int_0^{r/N} P_{j,k}(\sigma\mathscr{A}_N)\,d\sigma.$$

The parameters j and k are chosen with regard to the requirements that the operators $P_{j,k}(\lambda\frac{r}{N}\mathscr{A}_N)$ (a) exist for all N sufficiently large, and (b) satisfy the factor stability condition. As is the case for Discrete Approximation Schemes for the homogeneous initial-value problem, it is the behavior of the approximation triple $\{Z_N,\pi_N,\mathscr{A}_N\}$ which determines the factor stability properties of the DASP $\{Z_N,\pi_N,\mathscr{A}_N,P_{i,j}(z),P_{k,\ell}(\lambda z)\}$.

The remainder of this section is devoted to characterizing that subclass of the Padé table which under certain assumptions on $\{Z_N,\pi_N,\mathscr{A}_N\}$ yields appropriate rational functions D(z). We pay particular attention to the triples determined by the Averaging/Finite Difference and Spline/Variational state approximations discussed in §6 and §7 respectively.

10.3. Theorem. Suppose that $\{Z_N,\pi_N,\mathscr{A}_N,P_{i,j}(z)\}$ and $\{Z_N,\pi_N,\mathscr{A}_N,P_{k,\ell}(z)\}$ are factor stable Discrete Approximation Schemes with property (P1) which satisfy the hypotheses of Theorem 4.9. Then for $\lambda \in [0,1]$ fixed, the operators $P_{k,\ell}(\lambda\frac{r}{N}\mathscr{A}_N)$ exist for all N sufficiently large and, moreover, the DASP given by $\{Z_N,\pi_N,\mathscr{A}_N,P_{i,j}(z),P_{k,\ell}(\lambda z)\}$ is factor stable.

Proof: The existence of the operators $P_{k,\ell}(\frac{\lambda r}{N}\mathscr{A}_N)$ for all N sufficiently large is a consequence of Lemma 4.8. Factor stability can be demonstrated as follows:

(10.4) $$|P_{k,\ell}(\tfrac{\lambda r}{N}\mathcal{A}_N)\pi_N z_0 - \pi_N I z_0|_N$$

$$\leq |P_{k,\ell}(\tfrac{\lambda r}{N}\mathcal{A}_N)\pi_N z_0 - \pi_N S(\tfrac{\lambda r}{N}) z_0|_N + |S(\tfrac{\lambda r}{N}) z_0 - z_0|_Z$$

for $z_0 \in Z$.

A trivial modification of the arguments used to verify sufficiency in the proof of *Theorem* 4.9 yields that the first term on the right-hand side of (10.4) tends to zero as $N \to \infty$, while the fact that $\{S(t): t \geq 0\}$ is a $\mathcal{C}_0$ semigroup of operators on Z implies that the second term tends to zero as $N \to \infty$ as well.

□

Theorem 10.3 applied in conjunction with the results of §5 provides a rich class of appropriate rational functions D(z). Indeed, for a given approximation triple $\{Z_N, \pi_N, \mathcal{A}_N\}$, §5 serves to characterize those entries in the Padé table which, when selected for C(z), yield a factor convergent Discrete Approximation Scheme for the homogeneous problem. Theorem 10.3 further reveals that any choice appropriate for C(z) is appropriate for D(z) as well, and thereby gives rise to a factor convergent DASP.

While Theorem 10.3 assures us that for a given approximation triple $\{Z_N, \pi_N, \mathcal{A}_N\}$ satisfying certain basic hypotheses the set of factor stable DASP of the form $\{Z_N, \pi_N, \mathcal{A}_N, P_{i,j}(z), P_{k,\ell}(\lambda z)\}$ is non-empty, we are fortunate in that a still broader characterization is possible. Furthermore, these results will be directly applicable to the Averaging/Finite Difference and Spline/Variational approximation triples which have been discussed earlier.

10.5. Theorem. Suppose that $\{Z_N, \pi_N, \mathscr{A}_N, P_{i,j}(z)\}$ is a factor stable Discrete Approximation Scheme with property (P1) which satisfies the hypotheses in the statement of Theorem 4.9. Suppose further that there exists a constant $\hat{K} > 0$ independent of N such that

$$|(I + \tfrac{r}{N}\mathscr{A}_N)|_N \leq \hat{K} \tag{10.6}$$

for all N sufficiently large. Then for each $k, \ell > 0$ with $k \leq \ell+2$ and each $\lambda \in [0,1]$ we have that the operators $P_{k,\ell}(\frac{\lambda r}{N}\mathscr{A}_N)$ exist for all N sufficiently large and, moreover, the DASP given by $\{Z_N, \pi_N, \mathscr{A}_N, P_{i,j}(z), P_{k,\ell}(\lambda z)\}$ is factor stable.

The proof of Theorem 10.5 can be argued in much the same manner as were the proofs of Lemma 4.8 and Theorem 4.9 (cf. [33]).

As a consequence of this theorem, one has that if $\{Z_N, \pi_N, \mathscr{A}_N, P_{i,j}(z)\}$ is a Discrete Approximation Scheme satisfying the required hypotheses, then for each $k > 0$ and $\lambda \in [0,1]$ the DASP given by $\{Z_N, \pi_N, \mathscr{A}_N, P_{i,j}(z), P_{0,k}(\lambda z)\}$ is factor stable. That is to say, we may choose $D(z)$ from among those entries in the Padé table for which no operator inverse need be calculated in the computation of the operators $D(\frac{r}{N}\mathscr{A}_N)$.

That condition (10.6) is satisfied by the Averaging/Finite Difference approximation triple is an immediate consequence of Lemma 6.8. That the condition also obtains for the linear Spline/Variational approximation triple is the conclusion of the next theorem.

10.7. Theorem. For $\{Z_N^1, \pi_N^1, \mathscr{A}_N^1\}$ as defined in §7, we have

$$|(I + \tfrac{r}{N}\mathscr{A}_N^1)|_N \leq M^1,$$

where M^1 is independent of N.

Proof: The equivalence of the norms $|\cdot|_N$ and ${}_g||\cdot||_N$ (cf. §7) on Z_N implies that it suffices to show

$$ {}_g||(I + \tfrac{r}{N}\mathcal{A}_N^1)|| \leq \hat{M}^1, \qquad \hat{M}^1 \text{ independent of N.} $$

For $\hat{\phi}^N \in Z_N^1$, we find that

$$ {}_g||(I + \tfrac{r}{N}\mathcal{A}_N^1)\hat{\phi}^N||_N^2 = {}_g||\hat{\phi}^N||_N^2 + 2\tfrac{r}{N}\,{}_g[\mathcal{A}_N^1\hat{\phi}^N, \hat{\phi}^N]_N + (\tfrac{r}{N})^2\,{}_g||\mathcal{A}_N^1\hat{\phi}^N||_N^2 \tag{10.8} $$

$$ \leq (1 + 2\beta_1\tfrac{r}{N})\,{}_g||\hat{\phi}^N||_N^2 + (\tfrac{r}{N})^2\,{}_g||\mathcal{A}_N^1||_N^2, $$

where inequality (10.8) above is a consequence of Lemma 7.4.

Inspection of the inequality given by (10.8) reveals that if we can demonstrate that ${}_g||\mathcal{A}_N^1|| = O(N)$ as $N \to \infty$, then the desired conclusion obtains. Once again employing the norm equivalence of $|\cdot|_N$ and ${}_g||\cdot||_N$, we show

$$ |\mathcal{A}_N^1|_N = O(N). $$

Recalling that for $\hat{\phi}^N \in \hat{Z}_N^1$

$$ (\pi_N^1)^{-1}\hat{\phi}^N \in \hat{Z}_1^N \subset \mathcal{D}(\mathcal{A}), $$

we let $(\pi_N^1)^{-1}\hat{\phi}^N = (\phi^N(0), \phi^N)$ and find

$$ |\mathcal{A}_N^1\hat{\phi}^N|_N^2 = |\pi_N^1\mathcal{A}(\pi_N^1)^{-1}\hat{\phi}^N|_N^2 \leq |\mathcal{A}(\pi_N^1)^{-1}\hat{\phi}^N|_Z^2 $$

$$= |(\phi^N(0),\phi^N)|_Z^2 = |(L\phi^N, D\phi^N)|_Z^2$$

$$= |L\phi^N|_{R^n}^2 + |D\phi^N|_{L_2}^2 .$$

The Schmidt inequality (cf. Schultz [34]) yields the following bound for $|D\phi^N|_{L_2}^2$:

$$|D\phi^N|_{L_2}^2 = \int_{-r}^{0} |D\phi^N(\sigma)|^2 d\sigma = \sum_{j=1}^{N} \int_{-jr/N}^{-(j-1)r/N} |D\phi^N(\sigma)|^2 d\sigma$$

$$\leq \sum_{j=1}^{N} 12\left(\frac{N}{r}\right)^2 \int_{-jr/N}^{-(j-1)r/N} |\phi^N(\sigma)|^2 d\sigma = 12\left(\frac{N}{r}\right)^2 |\phi^N|_{L_2}^2$$

$$\leq 12\left(\frac{N}{r}\right)^2 [|\phi^N(0)|_{R^n}^2 + |\phi^N|_{L_2}^2] = 12\left(\frac{N}{r}\right)^2 |\hat{\phi}^N|_N^2.$$

Consequently for $-\tau \in [-r,0]$ it follows that

$$|\phi^N(-\tau)|^2 = |\phi^N(0) - \int_{-\tau}^{0} D\phi^N(\sigma)\,d\sigma|^2 \leq 2|\phi^N(0)|^2 + 2\tau \int_{-\tau}^{0} |D\phi^N(\sigma)|^2 d\sigma$$

$$\leq 2|\phi^N(0)|^2 + 2\tau|D\phi^N|_{L_2}^2 \leq 2|\phi^N(0)|^2 + 24r\left(\frac{N}{r}\right)^2 |\phi^N|_{L_2}^2$$

$$\leq K_1 N^2 |(\phi^N(0),\phi^N)|_Z^2 = K_1 N^2 |\hat{\phi}^N|_N^2$$

where K_1 is a positive constant independent of N. This in turn implies

$$|L\phi^N|_{R^n}^2 \leq 2\Big|\sum_{j=1}^{\nu} A_j \phi^N(-\tau_j)\Big|^2 + 2\Big|\int_{-r}^{0} D(\theta)\phi^N(\theta)\,d\theta\Big|^2$$

$$\leq 2\Big(\sum_{j=1}^{\nu} |A_j|\,|\phi^N(-\tau_j)|\Big)^2 + 2|D|_{L_2}^2 |\phi^N|_{L_2}^2$$

$$\leq 2\nu \sum_{j=1}^{\nu} |A_j|^2 |\phi^N(-\tau_j)|^2 + 2|D|_{L_2}^2 |\phi^N|_{L_2}^2$$

$$\leq (2\nu K_1 \sum_{j=1}^{\nu} |A_j|^2) N^2 |\hat{\phi}^N|_N^2 + 2|D|_{L_2}^2 |\hat{\phi}^N|_N^2$$

$$\leq K_2 N^2 |\hat{\phi}^N|_N^2,$$

where K_2 is a positive constant independent of N. Therefore, we have

$$|\mathscr{A}_N^1 \hat{\phi}^N|^2 \leq |L\phi^N|_{R^n}^2 + |D\phi^N|_{L_2}^2 \leq K_2 N^2 |\hat{\phi}^N|_N^2 + 12r^{-2} N^2 |\hat{\phi}^N|_N^2$$

$$= KN^2 |\hat{\phi}^N|_N^2 ,$$

where K is a positive constant independent of N, and hence

$$|\mathscr{A}_N^1|_N = O(N).$$

□

10.9. Remark. The results of Theorem 10.7 are easily generalized and extended to apply to approximations employing spline bases of arbitrarily high order (the relevant constants will of course depend upon the order of the spline basis chosen). In particular, this includes the cubic Spline/Variational triple $\{Z_N^3, \pi_N^3, \mathscr{A}_N^3\}$ discussed in §7.

It is interesting to note that for a given approximation triple $\{Z_N, \pi_N, \mathscr{A}_N\}$ several standard time-differencing numerical techniques for ordinary differential equations determine DASP of the form $\{Z_N, \pi_N, \mathscr{A}_N, P_{i,j}(z), P_{k,\ell}(\lambda z)\}$ when applied to the approximating ODE system in Z_N given by

$$\dot{z}_N(t) = \mathscr{A}_N z_N(t) + \pi_N(f(t),0), \qquad t \in [0,T] \tag{10.10}$$

$$z_N(0) = \pi_N z_0 = \pi_N(\eta,\phi) \tag{10.11}$$

We conclude this section with two examples which serve to illustrate these ideas.

10.12. Example (Trapezoidal Approximation). For a given approximation triple $\{Z_N, \pi_N, \mathcal{A}_N\}$, consider the approximating ODE system (10.10), (10.11) in its equivalent integral equation formulation given by

$$z_N(t) = z_N(0) = \int_0^t \mathcal{A}_N z_N(\sigma)\,d\sigma + \int_0^t \pi_N(f(\sigma),0)\,d\sigma, \qquad t \in [0,T].$$

Recalling that $t_k^N = kr/N$, $k = 0,1,2,\ldots,\rho N$, it follows that

$$z_N(t_k^N) = z_N(t_{k-1}^N) + \int_{t_{k-1}^N}^{t_k^N} \mathcal{A}_N z_N(\sigma)\,d\sigma + \int_{t_{k-1}^N}^{t_k^N} \pi_N(f(\sigma),0)\,d\sigma, \tag{10.13}$$

$$k = 0,1,2,\ldots,\rho N.$$

If we approximate the first integral on the right-hand side of (10.13) via the trapezoidal rule for numerical integration, we have

$$z_N(t_k^N) \approx z_N(t_{k-1}^N) + \frac{r}{2N}[\mathcal{A}_N z_N(t_{k-1}^N) + \mathcal{A}_N z_N(t_k^N)] + \int_{t_{k-1}^N}^{t_k^N} \pi_N(f(\sigma),0)\,d\sigma.$$

or

$$(I - \tfrac{r}{2N}\mathcal{A}_N) z_N(t_k^N) \sim (I + \tfrac{r}{2N}\mathcal{A}_N) z_N(t_{k-1}^N) + \tfrac{r}{N}\pi_N(\tfrac{N}{r}\int_{t_{k-1}^N}^{t_k^N} f(\sigma)\,d\sigma, 0). \tag{10.14}$$

Recalling that $\mathcal{A}_N \in G(M,\beta)$ implies the operator $(I - \frac{r}{2N}\mathcal{A}_N)^{-1}$ exists for all N sufficiently large, we solve (10.14) for $z_N(t_k^N)$ and find

$$z_N(t_k^N) \sim (I - \tfrac{r}{2N}\mathcal{A}_N)^{-1}(I + \tfrac{r}{2N}\mathcal{A}_N) z_N(t_{k-1}^N) + \tfrac{r}{N}(I - \tfrac{r}{2N}\mathcal{A}_N)^{-1}\pi_N(p_k^N f, 0).$$

Consequently, an approximation scheme can be defined and is given by the following relations:

$$z_N^0 = \pi_N z_0,$$

$$z_N^k = (I - \tfrac{r}{2N}\mathscr{A}_N)^{-1}(I + \tfrac{r}{2N}\mathscr{A}_N) z_N^{k-1} + \tfrac{r}{N}(I - \tfrac{r}{2N}\mathscr{A}_N)^{-1}\pi_N(p_k^N f, 0),$$

$$k = 1,2,\ldots,\rho N.$$

Recalling the definition of the family of approximate solution operators for the non-homogeneous initial-value problem, $\{G_k(\tfrac{r}{N};f)\}_{k=0}^{\rho N}$ determined by a given DASP $\{Z_N,\pi_N,\mathscr{A}_N,C(z),D(z)\}$ for each $N = 1,2,\ldots$, it can be verified that

$$z_N^k = G_k^T(\tfrac{r}{N};f)\pi_N z_0, \qquad k = 0,1,2,\ldots,\rho N,$$

where for each $N = 1,2,\ldots$, $\{G_k^T(\tfrac{r}{N};f)\}_{k=0}^{\rho N}$ denotes the family of approximate solution operators corresponding to the DASP given by $\{Z_N,\pi_N,\mathscr{A}_N,P_{1,1}(z),P_{1,0}(\tfrac{1}{2}z)\}$. We note that when $\{Z_N,\pi_N,\mathscr{A}_N\}$ is an Averaging/Finite Difference approximation triple, the method which has just been discussed, and thereby the DASP $\{Z_N^A,\pi_N^A,\mathscr{A}_N^A,P_{1,1}(z),P_{1,0}(\tfrac{1}{2}z)\}$ is analogous to the well-known Crank-Nicolson approximation commonly employed in the numerical solution of parabolic partial differential equations [22].

10.15. Example (The Improved Euler Approximation). For a given approximation triple $\{Z_N,\pi_N,\mathscr{A}_N\}$ it is once again convenient to consider the approximating ODE system (10.11), (10.12) in its equivalent integral equation form on the intervals $[t_{k-1}^N,t_k^N]$, $k = 1,2,\ldots,\rho N$ given by the expression in (10.13):

$$z_N(t_k^N) = z_N(t_{k-1}^N) + \int_{t_{k-1}^N}^{t_k^N} \mathscr{A}_N z_N(\sigma)\,d\sigma + \int_{t_{k-1}^N}^{t_k^N} \pi_N(f(\sigma),0)\,d\sigma.$$

The improved Euler approximation formulae [14] with step size r/N are an Euler predictor:

$$\hat{z}_N^k = z_N^{k-1} + \frac{r}{N}\mathcal{A}_N z_N^{k-1} + \frac{r}{N}\pi_N\left(\frac{N}{r}\int_{t_{k-1}^N}^{t_k^N} f(\sigma)\,d\sigma, 0\right) \tag{10.16}$$

followed by a trapezoidal corrector:

$$z_N^k = z_N^{k-1} + \frac{r}{2N}\left(\mathcal{A}_N z_N^{k-1} + \mathcal{A}_N \hat{z}_N^k\right) + \frac{r}{N}\pi_N\left(\frac{N}{r}\int_{t_{k-1}^N}^{t_k^N} f(\sigma)\,d\sigma, 0\right). \tag{10.17}$$

When combined, (10.16) and (10.17) lead to the approximation scheme given by the relations

$$z_N^0 = \pi_N z_0$$

$$z_N^k = \left(I + \frac{r}{N}\mathcal{A}_N + \frac{1}{2}\left(\frac{r}{N}\right)^2\mathcal{A}_N^2\right) z_N^{k-1} + \frac{r}{N}\left(I + \frac{r}{2N}\mathcal{A}_N\right)\pi_N\left(p_k^N f, 0\right). \tag{10.18}$$

A comparison of expressions (10.18) and (8.7) reveals that

$$z_N^k = G_k^{IE}\left(\frac{r}{N}; f\right)\pi_N z_0, \qquad k = 0,1,2,\ldots,\rho N$$

for each N = 1,2,..., where $\{G_k^{IE}(\frac{r}{N}; f)\}_{k=0}^{\rho N}$ denotes the family of approximate solution operators determined by the DASP $\{Z_N, \pi_N, \mathcal{A}_N, P_{0,2}(z), P_{0,1}(\frac{1}{2}z)\}$.

Several other numerical integration techniques for ordinary differential equations commonly encountered in practice correspond to DASP of the form $\{Z_N, \pi_N, \mathcal{A}_N, P_{i,j}(z), P_{k,\ell}(\lambda z)\}$. In particular, the explicit Euler or forward

difference method gives rise to the DASP $\{Z_N, \pi_N, \mathscr{A}_N, P_{0,1}(z), 1\}$, while the implicit Euler or backward difference method determines the DASP $\{Z_N, \pi_N, \mathscr{A}_N, P_{1,0}(z), P_{1,0}(z)\}$.

10.19. Remark. Examples 10.12 and 10.15 above provide a natural link between the ideas of the present investigation and the approximation framework developed in [3] and [6]. Indeed, for a given approximation triple $\{Z_N, \pi_N, \mathscr{A}_N\}$ satisfying hypotheses similar to those given in the statement of Theorem 4.9, the latter treatments demonstrate the convergence to the expression given in (8.3) of the classical variation of parameters solution to the initial-value problem (10.10), (10.11). When actually applied in practice, the desired approximating solution is obtained via the application of standard numerical integration techniques for ordinary differential equations to (10.10), (10.11). If the numerical integration scheme employed is among those discussed in the examples above (and others not presented) and the time step for the method is chosen as r/N, the two formulations become equivalent.

11. Analysis of Numerical Results

We present here computational results derived from the application of approximation schemes included in the framework developed above to several hereditary systems of the form discussed in §2. The numerical results which follow were obtained via a software package developed by the author and implemented in APL on the IBM 360/67 at Brown University. All of the calculations which follow were performed in a 330K-byte workspace which was sufficient to generate approximate solutions with values of N up to 96 in the case r = 1, n = 1 and N = 32 in the cases r = 1, n = 2 and r = 1, n = 3. A computationally efficient software

package based on the approximation framework discussed above, which includes methods utilizing the Averaging/Finite Difference and Spline/Variational state variable approximations, is currently under development.

In addition to the example which follows, we have tested our approximation schemes on several other hereditary systems with a variety of characteristics (cf. [33]). These examples were also used to test the methods developed in [2], [4] and [6]. Hence, they can serve as a basis for a preliminary comparison of the two approximation techniques. It is interesting to observe that in many instances, for the same state variable approximation, the results we obtain via a Discrete Approximation Scheme constructed with a second-order convergent rational function approximation to the exponential compare favorably with the corresponding results in [2] and [6] computed with a fourth-order Runge-Kutta integration of the approximating ODE with step size chosen independently of the state variable approximation.

When considering the rates of convergence in our test examples, we would not, in general, expect to observe those rates theoretically predicted by the results discussed in §4. Indeed, those estimates pertain to the homogeneous problem exclusively on a restricted class of initial data. However, the predicted rates appear to be in some sense indicative, if not conservative, estimates of the qualitative behavior observed experimentally in many of the test examples (both homogeneous and non-homogeneous with arbitrary initial data) which we have studied.

In the tables which follow, the symbol $^{N}\delta^{*}_{ijk\ell}$ denotes the absolute difference between the exact solution x to the FDE initial-value problem and the approximate solution computed via the Discrete Approximation Scheme for the non-homogeneous initial-value problem

(11.1) $$\{Z^{*}_{N}, \pi^{*}_{N}, \mathcal{Q}^{*}_{N}, P_{i,j}(z), P_{k,\ell}(\lambda z)\},$$

where $P_{\mu,\nu}(z)$ denotes the (μ,ν)th entry in the Padé table of rational function approximations to the exponential and the superscript * may take on the values A, 1 or 3 depending on whether the state variable approximation is of the Averaging/Finite Difference, linear Spline/Variational or cubic Spline/Variational type, respectively. The parameter λ has been fixed at 1/2 throughout. In the discussion below we also, on occasion, denote the DASP (11.1) by the shorthand notation $\{*,N,i,j,k,\ell\}$. Finally, recalling that $|P_{i,j}(z)-e^z| = O(z^{i+j+1})$ as $z \to 0$, we define the quantity $q = i+j$ as the index of the approximation scheme $\{*,N,i,j,k,\ell\}$.

11.2. Example (Banks and Kappel [6], Example 1). Consider the scalar, second-order, non-homogeneous initial-value problem

$$\ddot{u}(t) + \dot{u}(t) + u(t-1) = 10,$$

$$u(\theta) = \cos\theta, \quad \dot{u}(\theta) = -\sin\theta, \quad -1 \le \theta \le 0$$

in its equivalent formulation as a 2×2 first-order system in the form of (2.1), (2.2),

$$\dot{x}(t) = \begin{bmatrix} 0 & 1 \\ 0 & -1 \end{bmatrix} x(t) + \begin{bmatrix} 0 & 0 \\ -1 & 0 \end{bmatrix} x(t-1) + \begin{bmatrix} 0 \\ 10 \end{bmatrix},$$

$$x(0) = (1,0)^T, \quad x_0(\theta) = (\cos\theta, -\sin\theta)^T, \quad -1 \le \theta \le 0,$$

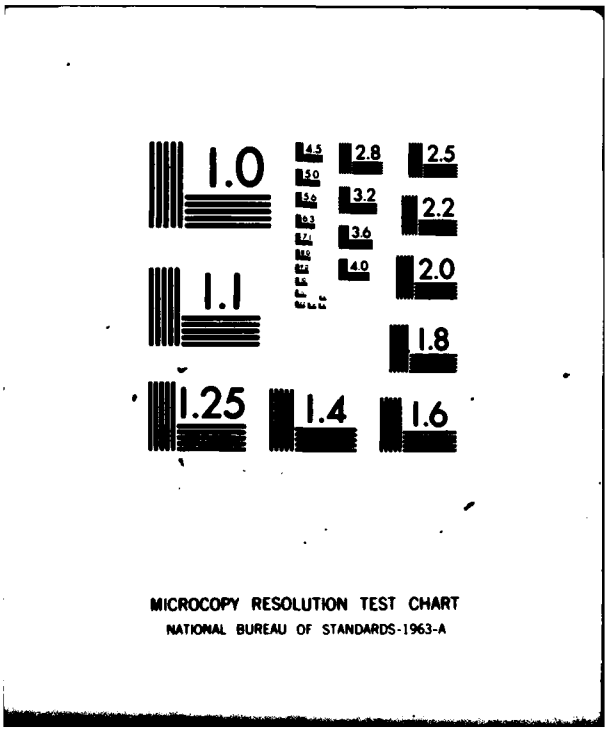
where $x_1 = u$ and $x_2 = \dot{u}$.

The solution on the interval [0,2] can be calculated by the method of steps [11] and is given by

$$x_1(t) = w(t) = -9 - \sin 1 + 10t + (10 + .5\sin 1 - .5\cos 1)e^{-t}$$
$$+ .5(\sin 1 - \cos 1)\sin t + .5(\sin 1 + \cos 1)\cos t, \qquad t \in [0,1]$$

AD-A089 726 BROWN UNIV PROVIDENCE RI LEFSCHETZ CENTER FOR DYNAMI--ETC F/G 12/1
A DISCRETE APPROXIMATION FRAMEWORK FOR HEREDITARY SYSTEMS.(U)
MAY 80 I G ROSEN DAAG29-79-C-0161
UNCLASSIFIED AFOSR-TR-80-0941 NL

END
DATE FILMED
9-80
DTIC


1.0
1.1
1.25
1.4
1.6
1.8
2.0
2.2
2.5
2.8
3.2
3.6
4.0
4.5
5.0
5.6
MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

$$x_1(t) = w(t) - 29 - 2\sin 1 + \cos 1 + (19 + \sin 1)(t-1) - 5(t-1)^2$$
$$+ (29.5 + 1.5\sin 1 - \cos 1)e^{-(t-1)} + (10 + .5\sin 1 - .5\cos 1)e^{-(t-1)}$$
$$+ .5(\sin 1 - 1)\cos(t-1) + .5(1 - \cos 1)\sin(t-1) \qquad t \in [1,2]$$

$$x_2(t) = \dot{x}_1(t).$$

The numerical results for this example are exhibited in Tables 11.3 and 11.4. Based upon this evidence, we make the following observations. Averaging/Finite Difference approximations in an explicit scheme of index 2, {A,N,0,2,0,1}, effectively yield $O(1/N^{1-\delta})$ convergence, where δ is a positive number strictly less than 1. Although the rate of convergence remains unchanged, approximate solutions generated by the explicit scheme of index 1, {A,N,0,1,0,0} for each N, were in general less accurate than the corresponding results for the index-2 scheme. Little if any improvement is gained through the application of AFD schemes with index q greater than 2. Convergence like $O(1/N^2)$ as $N \to \infty$ is achieved by the diagonal implicit/explicit scheme {1,N,1,1,1,0} of index 2 constructed with a linear Spline/Variational state approximation. Accuracy is enhanced when diagonal schemes having index greater than 2 are employed. For the index-4 scheme, {3,N,2,2,2,0}, with cubic Spline/Variational state approximations, $O(1/N^2)$ convergence is observed. Once again increased accuracy is obtained if a scheme having a higher index is employed. Extremely accurate results obtained with cubic-spline-based methods for relatively small values of N make the characterization of convergence rates difficult. Indeed, the actual approximation error is most likely masked by the influx of error from other sources, i.e. machine roundoff and numerical quadratures.

TABLE 11.3

t	$x_1(t)$	$^{4}_{\delta}A_{0201}$	$^{8}_{\delta}A_{0201}$	$^{16}_{\delta}A_{0201}$	$^{32}_{\delta}A_{0201}$	$^{4}_{\delta}1_{1110}$	$^{8}_{\delta}1_{1110}$	$^{16}_{\delta}1_{1110}$	$^{24}_{\delta}1_{1110}$	$^{4}_{\delta}3_{2220}$	$^{8}_{\delta}3_{2220}$	$^{16}_{\delta}3_{2220}$	$^{24}_{\delta}3_{2220}$
0	1.0000	0	0	0	0	.00036	.00005	.00000	.00000	.00000	.00000	.00000	.00000
.25	1.27048	.02204	.00432	.00068	.00001	.02671	.00849	.00196	.00090	.00489	.00178	.00037	.00013
.50	1.99367	.03016	.00472	.00008	.00049	.05629	.01279	.00324	.00147	.01247	.00265	.00063	.00026
.75	3.06148	.03173	.00384	.00070	.00096	.0678	.01629	.00401	.00181	.00883	.00283	.00081	.00036
1.00	4.39272	.02906	.00185	.00157	.00127	.07119	.01810	.00452	.00202	.01588	.00386	.00105	.00046
1.25	5.92593	.01227	.00875	.00799	.00443	.08080	.02131	.00549	.00247	.02058	.00526	.00113	.00064
1.50	7.60007	.02740	.03383	.02208	.01245	.08962	.02263	.00556	.00246	.01936	.00513	.00146	.00093
1.75	9.34402	.08841	.07093	.04290	.02351	.08539	.01935	.00462	.00214	.02565	.00605	.00153	.00145
2.00	11.08330	.16462	.11508	.06670	.03581	.06818	.01542	.00406	.00181	.02901	.00774	.00185	.00247

TABLE 11.4

t	$x_2(t)$	$^{4}\delta^{A}_{0201}$	$^{8}\delta^{A}_{0201}$	$^{16}\delta^{A}_{0201}$	$^{32}\delta^{A}_{0201}$	$^{4}\delta^{1}_{1110}$	$^{8}\delta^{1}_{1110}$	$^{16}\delta^{1}_{1110}$	$^{24}\delta^{1}_{1110}$	$^{4}\delta^{3}_{2220}$	$^{8}\delta^{3}_{2220}$	$^{16}\delta^{3}_{2220}$	$^{24}\delta^{3}_{2220}$
0	0	0	0	0	0	.00036	.00009	.00002	.00000	.00000	.00000	.00000	.00000
.25	2.06969	.04321	.01492	.00600	.00266	.05525	.00624	.00116	.00065	.01268	.00291	.00082	.00176
.50	3.64428	.05943	.01950	.00752	.00324	.00239	.00062	.00015	.00006	.01584	.00077	.00144	.00327
.75	4.89445	.05656	.01685	.00543	.00197	.02466	.00249	.00088	.00032	.01410	.00125	.00029	.00585
1.00	5.76581	.07810	.02895	.01089	.00395	.00930	.00148	.00021	.00005	.00609	.00718	.00140	.01012
1.25	6.45956	.15749	.08221	.04430	.02385	.01745	.00662	.00142	.00045	.01754	.00072	.00046	.01736
1.50	6.88559	.25003	.13887	.07605	.04302	.00142	.00591	.00222	.00124	.00665	.00220	.00042	.02869
1.75	7.01599	.32390	.17742	.09418	.04844	.03929	.01285	.00214	.00078	.00062	.00621	.001389	.04841
2.00	6.84972	.36860	.19535	.10047	.05068	.06301	.00999	.00302	.00097	.01299	.00341	.00098	.08134

11.5. Remarks. Evidence provided by our numerical study indicates that certain trade-offs exist in choosing between an AFD and SPV approximation scheme. While in most cases, SPV methods yielded superior results, it has been observed (cf. [6], Example 4) that when the initial data lies in the subspaces of Z, $\{\hat{Z}_A^N\}$ (cf. §6), defined in the construction of the Averaging/Finite Difference approximations, the AFD methods provide results superior to those of spline-based approximation schemes.

In addition, as one might expect, a price must be paid for the increased accuracy and rapidity of convergence yielded by the cubic spline methods. Due in part to the wider bandwidth of the matrices generated, these schemes tend to be more difficult to program, take longer to execute and have larger storage requirements than either the Averaging/Finite Difference or linear Spline/Variational approximations. Moreover, as is the case in any numerical approximation algorithm, it is desirable to maintain a uniform order of approximation throughout all phases of the computation. The cubic spline state approximations with a theoretically predicted convergence rate of $O(1/N^2)$ as $N \to \infty$ will therefore perform best in a scheme with a relatively high index. Unfortunately, in the case of spline-based approximations we are unable to guarantee factor convergence of explicit methods. Thus, for a factor convergent Discrete Approximation Scheme of high index employing cubic spline state approximations, it is necessary to invert a matrix which is a high-degree polynomial in the matrix $\mathcal{A}_N^3$. In general, this tends to be a numerically ill-conditioned procedure and may require the use of higher precision arithmetic.

12. Concluding Remarks

We have constructed an abstract approximation framework which can be applied to FDE initial-value problems of the type discussed in §2. We have detailed readily verifiable conditions which if satisfied guarantee convergence to the solution of the initial-value problem. The schemes proposed, and the abstract framework in general, represent an alternative to related approximation packages for FDE suggested by Banks and Burns [3], [4] and Banks and Kappel [6]. Furthermore, in the case of an autonomous linear system, the methods developed here are an extension and generalization of the ideas discussed by Reber [29].

Within the framework itself, there is a great deal of freedom in the actual selection of a particular convergent approximation scheme. Moreover, based on the evidence discussed in the previous section, one can conclude that the appropriate choice of parameters which determine the optimal method to apply depends heavily upon the characteristics of the initial-value problem under consideration.

The numerical results for the test examples suggest that the factor convergence properties of a particular method depend rather heavily on the interrelation between the order of the state approximation employed and the degree to which the rational function component of the scheme approximates the exponential function. In fact, it is apparent that a deeper understanding of this interdependence would provide valuable insight which could lead to the solution of many of the unanswered questions posed throughout this paper.

The approximation framework developed in this investigation has also been applied to certain classes of quasi-linear FDE initial-value problems (cf. [33]), and has been expanded so as to become part of a package yielding approximate solutions to the optimal control and parameter identification problems for systems governed by retarded functional differential equations of the type we have considered (cf. [2], [3], [4], [5], [29]).

Acknowledgement. The author wishes to express his appreciation to Professor H.T. Banks for his helpful suggestions during the preparation of this manuscript.

References

[1] A.V. Balakrishnan, Applied Functional Analysis, Springer-Verlag, New York, 1976.

[2] H.T. Banks, Approximation of nonlinear functional differential equation control systems, J. Opt. Theory Appl., 29 (1979), pp. 383-409.

[3] H.T. Banks and J.A. Burns, An abstract framework for approximate solutions to optimal control problems governed by hereditary systems, in Proceedings, International Conference on Differential Equations (Univ. So. Calif., Sept., 1974), H.A. Antosiewicz, ed., Academic Press, New York, 1975, pp. 10-25.

[4] H.T. Banks and J.A. Burns, Hereditary control problems: numerical methods based on averaging approximations, SIAM J. Control and Optimization, 16 (1978), pp. 169-208.

[5] H.T. Banks, J.A. Burns, and E.M. Cliff, Parameter estimation and identification for systems with delays, November 1979, to appear.

[6] H.T. Banks and F. Kappel, Spline approximations for functional differential equations, J. Diff. Eq., 34 (1979), pp. 496-522.

[7] R. Bellman and K.L. Cooke, Differential Difference Equations, Academic Press, New York, 1963.

[8] S.K. Berberian, Lectures in Functional Analysis and Operator Theory, Springer-Verlag, New York, 1979.

[9] J.G. Borisovic and A.S. Turbabin, On the Cauchy problem for linear non-homogeneous differential equations with retarded arguments, Soviet Math. Dokl., 10 (1969), pp. 401-405.

[10] E.W. Cheney, Introduction to Approximation Theory, McGraw-Hill, New York, 1966.

[11] R.D. Driver, Ordinary and Delay Differential Equations, Springer-Verlag, New York, 1977.

[12] R.L. Ehle, A-stable methods and Padé approximations to the exponential, SIAM J. Math. Anal., 4 (1973), pp. 671-680.

[13] J.K. Hale, Theory of Functional Differential Equations, Springer-Verlag, New York, 1977.

[14] P. Henrici, Discrete Variable Methods in Ordinary Differential Equations, John Wiley and Sons, New York, 1962.

[15] R. Hersh and T. Kato, High-accuracy stable difference schemes for well-posed initial-value problems, SIAM J. Numer. Anal., 16 (1979), pp. 670-682.

[16] A. Iserles, On the generalized Padé approximations to the exponential function, SIAM J. Numer. Anal., 16 (1979), pp. 631-636.

[17] A. Iserles, On the A-acceptability of Padé approximations, SIAM J. Math. Anal., 10 (1979), pp. 1002-1007.

[18] T. Kato, Perturbation Theory for Linear Operators, second edition, Springer-Verlag, New York, 1976.

[19] S.G. Krein, Linear Differential Equations in Banach Space, Trans. Math. Monographs, Vol. 29, Amer. Math. Soc., Providence, 1971.

[20] P.D. Lax and R. Richtmyer, Survey of the stability of linear finite difference approximations, Comm. Pure Appl. Math., 9 (1956), pp. 267-293.

[21] R. McKelvey, Spectral measures, generalized resolvents, and functions of positive type, J. Math. Anal. and Appl., 11 (1965), pp. 447-477.

[22] A.R. Mitchell, Computational Methods in Partial Differential Equations, John Wiley and Sons, New York, 1969.

[23] J. von Neumann, Eine Spektraltheorie für allgemeine Operatoren eines unitären Raumes, Math. Nachr., 4 (1950-51), pp. 258-281.

[24] S.P. Norsett, One-step methods of Hermite type for numerical integration of stiff systems, BIT, 14 (1974), pp. 63-77.

[25] S.P. Norsett, C-polynomials for rational approximation to the exponential function, Numer. Math., 25 (1975), pp. 39-56.

[26] A. Pazy, Semigroups of Linear Operators and Applications to Partial Differential Equations, Math. Dept. Lecture Notes, vol. 10, Univ. Maryland, College Park, 1974.

[27] P.M. Prenter, Splines and Variational Methods, John Wiley and Sons, New York, 1975.

[28] D. Reber, Approximation and Optimal Control of Linear Hereditary Systems, Ph.D. thesis, Brown University, November, 1977.

[29] D. Reber, A finite difference technique for solving optimization problems governed by linear functional differential equations, J. Diff. Eq., 32 (1979), pp. 193-232.

[30] G.W. Reddien and G.F. Webb, Numerical approximation of nonlinear functional differential equations with L_2 initial functions, SIAM J. Math. Anal., 9 (1978), pp. 1151-1171.

[31] R.D. Richtmyer and K.W. Morton, Difference Methods for Initial Value Problems, Interscience, New York, 1967.

[32] F. Riesz and B. Sz.-Nagy, Functional Analysis, Ungar, New York, 1975.

[33] I.G. Rosen, A Discrete Approximation Framework for Hereditary Systems, Ph.D. thesis, Brown University, June, 1980.

[34] M.H. Schultz, Spline Analysis, Prentice-Hall, Englewood Cliffs, N.J., 1973.

[35] R.J. Thompson, Difference approximations for inhomogeneous and quasi-linear equations, J. Soc. Indust. Appl. Math., 12 (1964), pp. 189-199.

[36] H.F. Trotter, Approximation of semi-groups of operators, Pacific J. Math., 8 (1958), pp. 887-919.

[37] R.S. Varga, Matrix Iterative Analysis, Prentice-Hall, Englewood Cliffs, N.J., 1962, Chapter 8.3.

[38] K. Yosida, Functional Analysis, Springer-Verlag, New York, 1974.

REPORT DOCUMENTATION PAGE UNCLASSIFIED	READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER: AFOSR-TR-80-0941 / 2. GOVT ACCESSION NO.: AD-A089 726	3. RECIPIENT'S CATALOG NUMBER:
4. TITLE (and Subtitle): A DISCRETE APPROXIMATION FRAMEWORK FOR HEREDITARY SYSTEMS	5. TYPE OF REPORT & PERIOD COVERED: Interim
	6. PERFORMING ORG. REPORT NUMBER:
7. AUTHOR(s): I.G. ROSEN	8. CONTRACT OR GRANT NUMBER(s): AFOSR 76-3092
9. PERFORMING ORGANIZATION NAME AND ADDRESS: DIVISION OF APPLIED MATHEMATICS, BROWN UNIVERSITY, PROVIDENCE, RHODE ISLAND 02912	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F 2304/A1
11. CONTROLLING OFFICE NAME AND ADDRESS: AIR FORCE OFFICE OF SCIENTIFIC RESEARCH/NM, BOLLING AIR FORCE BASE, WASHINGTON, D.C.	12. REPORT DATE: May 1980
	13. NUMBER OF PAGES: 103
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)	15. SECURITY CLASS. (of this report): UNCLASSIFIED
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE:

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

A discrete approximation framework for initial-value problems involving certain classes of linear functional differential equations (FDE) of the retarded type is constructed. An equivalence between the FDE and abstract evolution equations (AEE) in an appropriately chosen Hilbert space is established. This equivalence is then employed in the development of the discrete approximation schemes in which the infinite-dimensional AEE is replaced by a finite-dimensional system of difference equations

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Convergence and rates of convergence are demonstrated via the properties of rational functions with operator arguments and both classical and recent results from linear semigroup theory. Two examples of families of approximation schemes which are included in the general framework and which may be implemented directly on high-speed computing machines are developed. A numerical study of examples which illustrates the application and feasibility of the approximation techniques in a variety of problems together with a summary and analysis of the numerical results are also included.