SIAM J. SCI. STAT. COMPUT.
Vol. 9, No. 5, September 1988

004

DIAGONAL PADÉ APPROXIMATIONS FOR INITIAL VALUE PROBLEMS*

MICHAEL F. REUSCH†, LEE RATZAN†, NEIL POMPHREY†, AND WONCHULL PARK†

Abstract. Diagonal Padé approximations to the time evolution operator for initial value problems are applied in a novel way to the numerical solution of these problems by explicitly factoring the polynomials of the approximation. A remarkable gain over conventional methods in efficiency and accuracy of solution is obtained.

Key words. Padé methods, initial value problems, A–stability, implicit schemes, trapezoidal method, Crank–Nicolson method, Carleman embedding

AMS(MOS) subject classifications. 65P20, 35A40, 65M99, 41A21

1. Introduction.

We consider homogeneous linear evolution equations of the form

$$\frac{\partial \Psi}{\partial t} = H(t)\Psi, \tag{1}$$

where $H(t)$ is a linear operator. The solution of this equation is

$$\Psi(t) = U(t,0)\Psi(0). \tag{2}$$

U is the familiar time evolution operator,

$$U(t,0) = \exp\left(\int_0^t dt\ H(t)\right). \tag{3}$$

When $H(t_1)$ does not commute with $H(t_2)$, (3) must be interpreted in the time-ordered sense.

All numerical methods for the solution of (1) approximate U in some way, often by the first few terms in the Taylor series for e^z. Euler's method, for example, retains just the first term in this series and has $U(t+\Delta t,t) \approx 1+\Delta t H(t)$. The Padé methods are based on rational function approximations to e^z:

$$e^z \approx \frac{P_M(z)}{Q_N(z)}. \tag{4}$$

Here $P_M(z)$ and $Q_N(z)$ are polynomials in z with real coefficients of order M and N, respectively [15]. For the most part, we will treat only the diagonal Padé approximations to e^z for which $N = M$ and $Q_M(z) = P_M(-z)$. The simplest of these diagonal Padé methods is the "Crank–Nicolson" or "trapezoidal" method [6] for which $M = 1$ and $P_1(z) = 1+z/2$. Like all of the $N > 0$ Padé methods, the Crank–Nicolson method is implicit in that it requires the inversion of an operator:

$$\left(1 - \frac{\Delta t H(t)}{2}\right)\Psi(t+\Delta t) = \left(1 + \frac{\Delta t H(t)}{2}\right)\Psi(t). \tag{5}$$

* Received by the editors July 13, 1987; accepted for publication (in revised form) December 10, 1987. This work was performed under the auspices of the U.S. Department of Energy, contract DE-AC02-76-CHO-3073.

† Plasma Physics Laboratory, Princeton University, James Forrestal Campus, Princeton, New Jersey 08544.

Although widely studied [3],[4],[7],[8],[9],[20],[21], high-order implicit Padé methods are not commonly used, presumably because of their apparent complexity [20]. Diagonal Padé methods are known to have optimal accuracy [21] and the odd-order methods also preserve positivity [9].

The main aims of this work are, first, to point out that higher-order implicit Padé methods need be no more complicated to implement than the Crank–Nicolson method, and, second, to show that surprising increases in accuracy and efficiency can be obtained by using these methods.

The original motivation for studying these methods comes out of our efforts to numerically simulate the behavior of plasmas in controlled thermonuclear fusion experiments. The analytical model used for such plasmas is a system of resistive magnetohydrodynamic (MHD) equations. This is a system of nonlinear hyperbolic equations which contain parabolic and elliptic components. An enormous range of time scales is contained in these equations. Fast compressional waves have a characteristic time of 10^{-6} seconds in a typical fusion device, while the typical length of an experiment is several seconds.

Our main interest, however, is in accurate simulation over the shear Alfvén wave times of 10^{-5} seconds or even longer resistive times of 10^{-3} seconds. Analytical methods exist for eliminating the faster time scales in these problems but they are too restrictive to apply in the general case. The convenient solution is to use implicit methods with timesteps appropriate to the longer times which are numerically stable in the presence of the fast waves. These latter are then not faithfully simulated. However, in a dynamically stable plasma configuration, they are thought to be unimportant and can be safely neglected.

The usual technique by which the MHD equations are given an implicit numerical character are variations on the Crank–Nicolson method. Given the longer timesteps allowed by these implicit methods, the simulation of an entire one-second plasma shot would still take a prohibitive amount of time, even on a parallel vector computer. Methods which would allow us to further increase the timestep or reduce the number of operations needed to simulate to a given accuracy would be invaluable. The diagonal Padé methods to be described are, we hope, such methods but more work is required to establish this.

2. The factorization method. We write the numerator polynomial of the Mth-order diagonal Padé approximation to e^z in factorized form as

$$P_M(z) = \prod_{m=1}^{M} \left(1 - \frac{z}{C_m}\right). \tag{6}$$

Here C_m are the roots of $P_M(z)$. An important fact is that all of these roots are distinct, nonzero, and have negative definite real part [3],[4],[7],[8],[17],[20],[21]. Table 1 gives numerical values of the roots of the diagonal Padé polynomials up to $M = 11$, and an interesting illustration of the poles and zeros of the first 20 diagonal Padé approximations to e^z is given in Fig.1.

The roots of $Q_M(z) = P_M(-z)$ are just $-C_m$. Since these roots are either real or occur in complex conjugate pairs, the poles of the diagonal Padé approximation are just the reflection of the zeros across the y axis. We can then write the Mth-order diagonal Padé approximation to e^z, with $C_m^\dagger$ as the complex conjugate of C_m, as

$$\frac{P_M(z)}{Q_M(z)} = \prod_{m=1}^{M} \frac{1 - z/C_m}{1 + z/C_m^\dagger}. \tag{7}$$

TABLE 1

Roots of the first 11 diagonal Padé approximations to e^z. Only the roots of nonnegative imaginary part are given.

Order	Real part	Imaginary part
1	-2.0	0.0
2	-3.0	1.732050807568877
3	-4.644370709252172	0.0
	-3.677814645373914	3.508761919567444
4	-5.792421205640744	1.734468257869007
	-4.207578794359256	5.314836083713505
5	-7.293477190659323	0.0
	-6.703912798307045	3.485322832366408
	-4.649348606363293	7.142045840675948
6	-8.496718791726729	1.735019346462726
	-7.471416712651628	5.252544622894256
	-5.031864495621643	8.985345907307884
7	-9.943573717055878	0.0
	-9.516581056309254	3.478572122261069
	-8.140278327276275	7.034348095419513
	-5.371353757886532	10.84138826143350
8	-11.17577208652617	1.735228890705500
	-10.40968158127378	5.232350305285130
	-8.736578434404781	8.828885000943038
	-5.677967897795266	12.70782259720976
9	-12.59403836343024	0.0
	-12.25873580854839	3.475696766962232
	-11.20884363901552	6.996313835771842
	-9.276879774360831	10.63454335087136
	-5.958521596360136	14.58292737668437
10	-13.84408981085430	1.735330390904289
	-13.23058193095358	5.223135841597920
	-11.93505665717623	8.769894377885137
	-9.772439133717648	12.44997096494290
	-6.217832467298239	16.46539891814719
11	-15.24467969165087	0.0
	-14.96845972142817	3.474205641536712
	-14.11578477534349	6.978029007087853
	-12.60267490974686	10.55238348739988
	-10.23129656781539	14.27404151778648
	-6.459444179840646	18.35422313741710

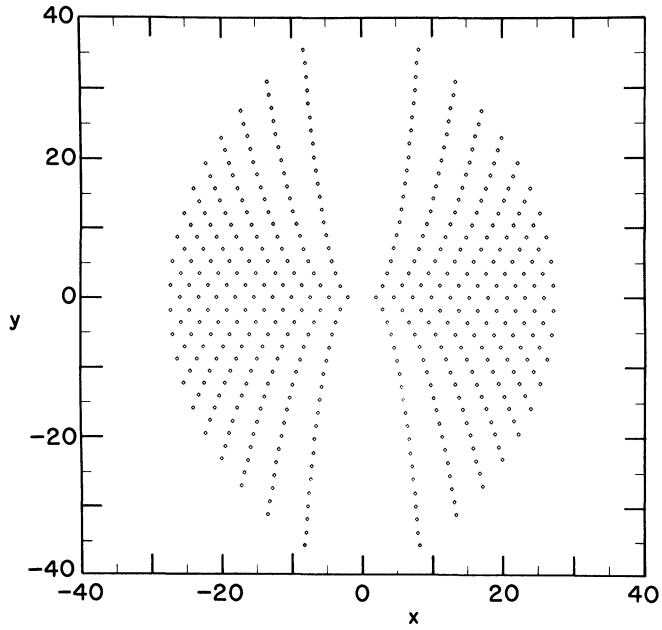

FIG. 1. *Poles and zeros of the first 20 diagonal Padé approximations to e^z are illustrated. These critical points of the approximation for fixed order generate a roughly elliptical figure whose radius increases with order. The zeros are well separated in the complex plane. Odd-order approximations have one negative real zero while even-order approximations have no real zeros.*

Each of the factors in the above form of the approximation has the A-stability property, namely that for real part of $z \leq 0$

$$\left|\frac{1 - z/C_m}{1 + z/C_m^\dagger}\right| \leq 1. \tag{8}$$

This A-stability property is shared by the whole approximation given by (7) and is a well-known property of all the Padé approximations to e^z for which $N = M, M+1$ or $M+2$ [7],[8].

We now observe that (7) is a valid approximation to the timestep evolution operator $U(t + \Delta t, t)$ if $\Delta t H(t)$ is substituted for z and a proper interpretation is made of the inverse operators. Our solution procedure is to unfold the products of (7) pair by pair. Each substep of a given timestep is of the form

$$\left(1 + \frac{\Delta t H(t)}{C_m^\dagger}\right)\Psi^m = \left(1 - \frac{\Delta t H(t)}{C_m}\right)\Psi^{m-1}, \tag{9}$$

where $1 \leq m \leq M$, $\Psi^0 = \Psi(t)$ and $\Psi(t + \Delta t) = \Psi^M$.

Although complex arithmetic must be used, each step of this process is as simple as one step of the Crank–Nicolson method requiring only an algorithm for numerical inversion of $H(t)$. As we shall see, the extra work done by using complex arithmetic is more than made up for by an increase in accuracy. Further, the same numerical algorithm can be conveniently used for an arbitrary order diagonal Padé method. We also see that each substep of the method is A-stable so that no transitory instabilities can occur.

If we start from real initial values for Ψ and $H(t)$ is a real operator, then after the first substep of the method we have complex values. Clearly, after M steps we

must have real values again since the overall product is real. The magnitude of the complex part after M substeps then gives us the truncation error of the process. On the other hand, if $H(t)$ is a complex operator, i.e., has complex eigenvalues, or if Ψ is a complex variable, then the overhead associated with complexification of the problem is saved. Complex Ψ are implied in multidimensional problems where one or more periodic dimensions are Fourier analyzed to yield a simpler and sometimes decoupled form for $H(t)$, which is then also more easily inverted.

Since the roots of $P_M(z)$ are all distinct and well separated in the complex plane, it is easy to obtain numerical approximations to them. All that is necessary is that they multiply up to the correct coefficients within a tolerable error. The slight error introduced by using truncated numerical approximations to them is not an essential limitation of the method.

The coefficients of $P_M(z)$ are known in closed form[4],[8],[17],[21]:

$$P_M(z) = \sum_{m=0}^{M} \frac{M!(2M-m)!}{(2M)!m!(M-m)!} z^m. \tag{10}$$

The numerical values of Table 1 when combined yield the known analytic values of the coefficients of the numerator polynomial to better than one part in 10^{15}.

The polynomials satisfy a number of useful recursion relations and the error in the approximation is [11]

$$e^z = \frac{P_M(z)}{Q_M(z)} + R_M(z) \quad \text{where,}$$

$$R_M(z) = \frac{\pi(-1)^M z^{2M+1} e^z}{2^{4M+1}(M!)^2}\left(1 + O(M^{-1})\right), \tag{11}$$

so that tolerable errors are possible even when $|z| > 1$; i.e., the timestep is larger than the characteristic time T_c of $H(t)$. The latter may be defined as the inverse of the magnitude of the largest eigenvalue of interest in the spectrum of $H(t)$, $T_c = 1/|\lambda_{max}|$.

Our numerical studies indicate that the sometimes surprising convergence of Padé approximations carries over into the initial value problem. Faithful simulations are obtained in some problems even for inordinately large timesteps ($\Delta t \approx MT_c$). In fact, the permissible timesteps are so large that explicit variation of $H(t)$ over a single timestep becomes important. Until now we have tacitly assumed that this variation was negligible. In our desired application of MHD simulation this may not be true.

Inclusion of a time-varying inhomogeneous term or boundary conditions in (1) leads to a similar problem for large timesteps. These are also present in the MHD equations. Possible resolutions for these problems exist via the method of variation of parameters and transformation of $H(t)$ to an almost time-invariant form. We plan to treat these in a future work and will not discuss them any further here.

Although fundamentally linear, the factorized diagonal Padé method can be applied to autonomous nonlinear ordinary differential equations by embedding these in a linear system via the technique of Carleman [1],[2],[5],[14],[18]. This application will also be treated in a future work.

The factorization technique given here can also be used to approximate the exponential of a badly conditioned matrix, as, for example, the two-by-two-dimensional test case of Moler and Van Loan [13]. Since for a large enough order method the condition number of each of the substeps remains tolerable, we thereby avoid the problem of a large condition of the total numerator or denominator expansions.

Since the complex roots occur in conjugate pairs, it is possible to rearrange the terms of (7) into a product over real quadratic and real linear factors. Inversion of

the quadratic terms for a tridiagonal diagonal matrix slightly more than doubles the number of operations. Given an overhead of a factor of four for complex mathematics, this real factors decomposition should be superior for real problems, while being only slightly more difficult to implement. Saff, Schönhage, and Varga [16] have investigated yet another rational approximation which involves only real poles and linear terms in the denominator.

If the order of the Padé method M is even moderately large, factored methods are superior to direct multiplication by P_M and inversion of Q_M for two reasons. First, factorization avoids the large matrix elements of the direct method, which particularly occur in stiff problems [13]. Second, factored methods are more efficient when applied to banded matrices. The work needed to invert M tridiagonal matrices using the factored method is proportional to $K \cdot M$, where K is the order of the matrix H, while that required to invert Q_M is proportional to $K \cdot M^2$, to leading order. Similar remarks apply to the multiplication by P_M, and the work needed to form P_M and Q_M, in the first place, is avoided.

Finally, we point out that there exist a number of implicit Runge–Kutta formulae equivalent to diagonal and subdiagonal Padé approximations which use entirely real mathematics. These have been studied by Ehle [7],[8] and others. Low-order methods of this type are harder to implement than factorized methods but may be preferable in some circumstances. However, the numerical solution of a system of N equations by these methods requires the inversion of an $N \cdot M$-by-$N \cdot M$ system, where M is the order of the implicit Runge–Kutta method. A high-order method applied to a large system is clearly unwieldy.

3. A numerical example. We have selected the homogeneous one-dimensional heat equation as a simple example for the numerical testing of our method:

$$\frac{\partial \Psi}{\partial t} = \sigma \frac{\partial^2 \Psi}{\partial x^2}, \tag{12}$$

where $0 \leq x \leq W$, $\Psi(0,t) = \Psi(W,t) = 0$, and σ can be complex. We use a centered difference formula for the right-hand side diffusion operator so that the exact eigenvalues of the discrete system are

$$\lambda_k = \frac{2\sigma}{\Delta x^2}(\cos \frac{k\pi}{K} - 1). \tag{13}$$

Here $1 \leq k < K$ and $\Delta x = W/K$.

The discrete version of the right-hand side diffusion operator is a tridiagonal matrix, as is the substep matrix of (9). The inversion is accomplished by the well-known recursive algorithm equivalent to an LU factorization [12] so that the work involved is only proportional to K.

We compare diagonal Padé methods implemented in the factorized form with complex arithmetic to a real mathematics version of the Crank–Nicolson method and several real arithmetic explicit methods. To compare the methods we select a single eigenmode of (12) as our initial condition and follow it for a time period equal to a given number of characteristic times $T_c = 1/|\lambda_k|$.

We repeatedly solve the problem for this period while varying the total number of timesteps from one to the number at which the maximum error saturates. For the explicit methods we start at the minimum number of timesteps for which the method is stable. At the final time the result of each different time discretization is compared to the exact known solution. Maximum and average errors are extracted and the processing time is recorded. The maximum error results are similar to those for the average error and are not presented.

Figure 2 shows the results of runs made on a Cray-1 computer using diagonal Padé approximations to order 11, labeled C1–C11, Euler's method (E), and Runge–Kutta methods of order two (RK2), four (RK4), and six (RK6).

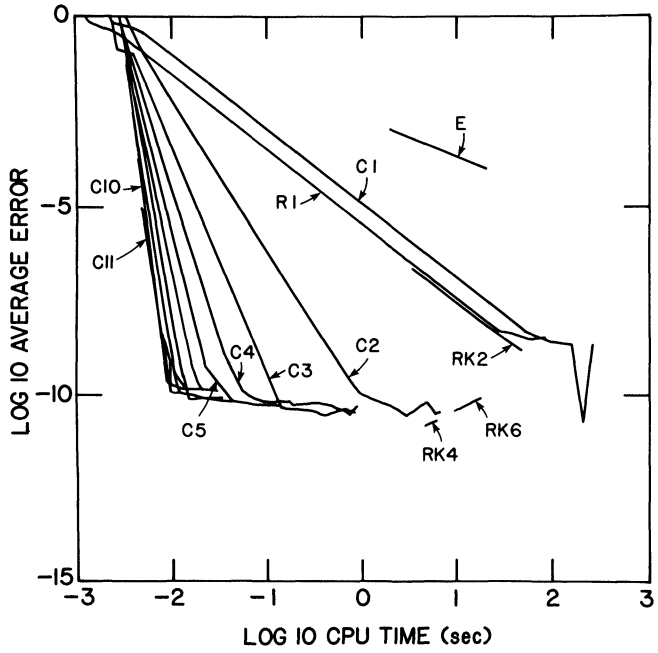

FIG. 2. *The efficiency of factorized, complex arithmetic, diagonal Padé methods of orders 1 through 11, labeled C1–C11, are compared to Euler's method, labeled E, the Crank-Nicolson method, labeled R1, and Runge-Kutta methods of orders 2 through 6, labeled RK2–RK6, implemented with real mathematics, for ten characteristic times of the first eigenmode of the one-dimensional heat-diffusion equation.*

All these runs were made for a period of ten characteristic times of the lowest $k = 1$ eigenmode and are typical in that similar results are found for real, imaginary, or complex σ and for shorter or longer time periods. In this set of runs the loop vectorization of the Cray-1 was turned off. The log of the average error is plotted against the log of the CPU time used. The average error is defined as the average of the absolute values of the difference between the computed and exact solutions divided by $e^{\lambda T}$, where T is the total time period.

For all implicit methods and for those explicit methods that are not already at the precision limit, the logarithm of error decreases linearly with the logarithm of CPU time until saturation due to rounding sets in at a machine, method, and period-dependent precision level. This is in agreement with expected behavior since the error is proportional to $(\Delta t\lambda)^{2M+1}$ for the implicit Padé methods and $(\Delta t\lambda)^{M+1}$ for the explicit methods, while the work involved varies as the inverse of the timestep and the slope is proportional to the order of the method.

The $M = 1$ Padé case, which was implemented with real arithmetic labeled R1, is seen to use only about half the time of the fully complex C1 method. A factor of four might have been expected and the difference is due to the fact that the Cray-1 unavoidably vectorizes some of the complex calculations even with vectorization off.

RK2 is seen to give about the same accuracy and use about the same time as R1 since the operation count and error of both methods is roughly the same. The $M > 1$ Padé approximations saturate at an error lower than the $M = 1$ approximation. The

efficiency of solution of the Padé methods improves as order increases, although the incremental improvement from order to order becomes smaller with order.

Gains of 100 or more in CPU time over the $M = 1$ Padé method and all the explicit methods are obtained with the higher-order Padé methods depending on the required accuracy. These gains are obtained despite the extra work implied by complex arithmetic. For the $M = 11$ method an accuracy of 10^{-5} was obtained with just one timestep. The $M = 15$ Padé method (not shown) was at the precision limit in one timestep.

We note that our selection of the lowest eigenmode of the system has accentuated the achievable improvement in efficiency over the explicit methods and that for the higher modes this improvement would not be as dramatic or would not even exist. Figure 3 presents the results of runs made for 10 characteristic times of the eleventh eigenmode of the heat equation. Here the RK4 method is more efficient than the C2 method while the RK6 and C3 methods are of comparable efficiency.

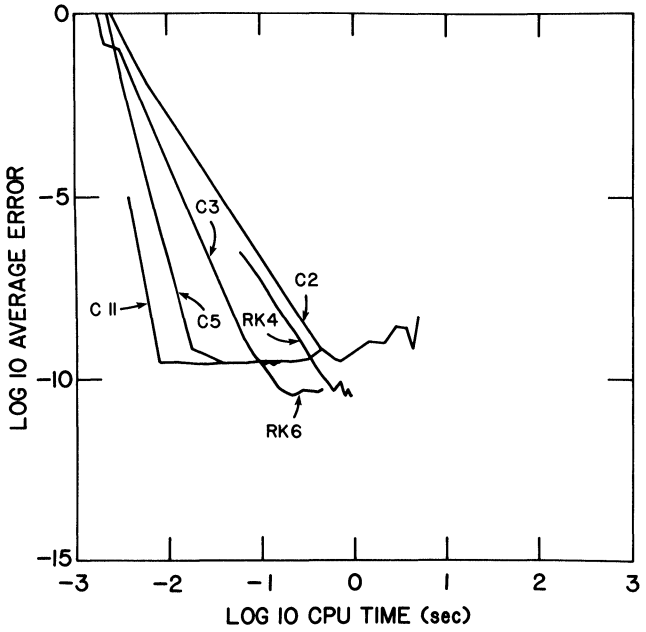

FIG. 3. *The efficiency of diagonal Padé methods through order 11, labeled C2–C11, are compared to Runge-Kutta methods of order 4 (RK4), and 6 (RK6) for ten characteristic times of the eleventh eigenmode of the one-dimensional heat equation. The achievable gain in efficiency of the Padé methods over the Runge-Kutta methods is reduced for this case which is not as stiff as the system of Figure 1. C3 and RK6 are of comparable efficiency and the curves for these methods actually overlap at several points. Complex arithmetic is seen to increase the rounding error by an order of magnitude over that of the real methods in this figure.*

Figure 3 also illustrates the fact that the minimum attainable errors of the high-order real arithmetic explicit methods are smaller than those of the complex arithmetic implicit methods. In this example they are an order of magnitude smaller. This is due to the increase of rounding error of the complex calculations over that of the real calculations. On the other hand, the implicit complex methods seem quite stable in the presence of rounding error in that an excessively short timestep does not tend to increase the attainable error.

When the Cray loop vectorization is turned on, the appropriately coded explicit

methods increase in speed by approximately a factor of 5, while the nonvectorized recursive inversion of the implicit Padé methods is only slightly faster. Even then, the high-order Padé methods are still more efficient in the stiff system of Fig. 2. We note that, although not implemented for this paper, techniques exist for partially vectorizing the recursive inversion of tridiagonal matrices [10],[19]. It is likely that even further improvement might be obtained with the incorporation of one of these techniques and that, in principle, certain elements of the complex calculations might be further vectorized as well.

4. Conclusions. We have presented a convenient method for the implementation of high-order implicit diagonal Padé approximations for the solution of homogeneous, exactly time-invariant or autonomous, linear operator, initial value problems, and have demonstrated the remarkable efficiency of these methods in comparison with both explicit methods and the Crank–Nicolson method. The central idea is a factorization technique which yields an algorithm of simplicity comparable to the Crank–Nicolson method. The same algorithm can be used for an arbitrary order diagonal Padé method and might be applied with advantage to nondiagonal Padé approximations and other methods.

The main limitations of the method to date are the requirements of exact or approximate time invariance, linearity, and ease of inversion of the operator. Since the same algorithm can be used for an arbitrary order method, an adaptive scheme can be constructed where a low-order method with a small timestep is used when the operator varies appreciably in time and a high-order large timestep method is used otherwise.

Further work is needed to establish whether this method can be applied to more practical problems which lie outside these limitations. However, the accuracy of the higher-order diagonal Padé methods is such that careful consideration should be given, concerning modification of codes that presently use the Crank–Nicolson method or its variants, to use the scheme of this paper.

REFERENCES

[1] R. F. S. ANDRADE, AND A. RAUH, *The Lorenz model and the method of Carleman embedding*, Phys. Lett. A, 82 (1981), pp. 276-278.

[2] R. BELLMAN, AND J.M. RICHARDSON, *On some questions arising in the approximate solution of nonlinear differential equations*, Quart. Appl. Math., 20 (1963), pp. 333-339.

[3] C. M. BENDER, AND S. A. ORSZAG, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw-Hill, New York, 1978.

[4] G. BIRKHOFF AND R. S. VARGA, *Discretization errors for well-set Cauchy problems I*, J. Math. Phys., 44 (1965), pp. 1-23.

[5] T. CARLEMAN, *Application de la théorie des équations intégrales linéaires aux systèmes d'équations différentielles non linéaires*, Acta. Math, 59 (1932), pp. 63-87.

[6] J. CRANK AND P. NICOLSON, *A practical method for numerical integration of solutions of partial differential equations of heat-conduction type*, Proc. Cambridge Philosophical Society, 43 (1947), pp. 50-67.

[7] B. L. EHLE, *A-stable methods and Padé approximations to the exponential*, SIAM J. Math. Anal., 4 (1973), pp. 671-680.

[8] ———, *On Padé approximations to the exponential function and A-stable methods for the numerical solution of initial value problems*, Ph.D. dissertation CSRR 2010, University of Waterloo, Waterloo, Ontario, Canada, 1969.

[9] J. A. VAN DE GRIEND AND J. F. B. M. KRAAIJEVANGER, *Absolute monotonicity of rational functions occurring in the numerical solution of initial value problems*, Numer. Math., 49 (1986), pp. 413-424.

[10] J. J. LAMBIOTTE, JR., AND R. G. VOIGT, *The solution of tridiagonal linear systems on the* CDC STAR-100 *computer*, ACM Trans. Math. Software, 1 (1975), pp. 308-329.

[11] Y. L. LUKE, *Mathematical Functions and Their Approximations*, Academic Press, Inc., New York, 1975.

[12] G. I. MARCHUK, *Methods of Numerical Mathematics*, Springer-Verlag, New York, Berlin, 1975, pp. 139-141.

[13] C. MOLER AND C. VAN LOAN, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801-836.

[14] E. W. MONTROLL, *On The Solution of Nonlinear Rate Equations by Matrix Inversion*, American Institute of Physics, New York, 1978.

[15] H. PADÉ, *Sur la représentation approchée d'une fonction par des fractions rationnelles*, Thesis, Annales de l'École Normale, 9 (1892), pp. 1-123.

[16] E. B. SAFF, A. SCHÖNHAGE, AND R. S. VARGA, *Geometric convergence to* e^z *by rational functions with real poles*, Numer. Math., 25 (1976), pp. 307-322.

[17] E. B. SAFF AND R. S. VARGA, *On the zeros and poles of Padé approximants to* e^z, Numer. Math., 25 (1975), pp. 1-14.

[18] W. H. STEEB AND F. WILHELM, *Non-linear autonomous systems of differential equations and Carleman's linearization procedure*, J. Math. Anal. Appl., 77 (1980), pp. 601-611.

[19] H. S. STONE, *Parallel tridiagonal equation solvers*, ACM Trans. Math. Software, 1 (1975), pp. 289-307.

[20] R. S. VARGA, *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N. J., 1962.

[21] ———, *On higher order stable implicit methods for solving parabolic partial differential equations*, J. Math. Phys., 40 (1961), pp. 220-231.