BIT 33 (1993), 285–303.

STIFFNESS OF ODEs

DESMOND J. HIGHAM* and LLOYD N. TREFETHEN**

Department of Mathematics and Computer Science
University of Dundee
Dundee, DD1 4HN, Scotland
na.dhigham@na-net.ornl.gov.

Department of Computer Science
Cornell University
Ithaca, NY 14853, USA.
LNT@cs.cornell.edu.

Abstract.

It is argued that even for a linear system of ODEs with constant coefficients, stiffness cannot properly be characterized in terms of the eigenvalues of the Jacobian, because stiffness is a transient phenomenon whereas the significance of eigenvalues is asymptotic. Recent theory from the numerical solution of PDEs is adapted to show that a more appropriate characterization can be based upon pseudospectra instead of spectra. Numerical experiments with an adaptive ODE solver illustrate these findings.

AMS(MOS) Subject classification: 65L05.

Key words: stiffness, stability, pseudospectra.

1. Introduction.

It is generally agreed that the essence of stiffness is a simple idea,

Stability is more of a constraint than accuracy,

with its familiar consequence as expressed in the words of Hairer and Wanner [5, p. 2],

Stiff equations are problems for which explicit methods don't work.

As soon as one tries to turn these ideas into a mathematical criterion for stiffness, however, disagreements set in. What makes a stiff problem stiff? No single answer seems right for all problems. In the face of this confusion some authors propose multiple criteria for stiffness, and others, none at all.

This paper will not give the ultimate stiffness criterion either, but we hope to contribute to the discussion. At the most general level our points can be summarized as follows:

* Supported by a Dundee University Research Initiatives Grant.
** Supported by NSF Grant DMS-9116110.
Received January 1992. Revised August 1992.

(a) Instability and stiffness are transient phenomena, involving finite time intervals $[t_0, t_1]$. They cannot be characterized by considering only the limits $t \to \infty$ or $t \to t_0$.

(b) Even linear, constant coefficient problems can be surprising. Not all complicated effects are due to nonlinearity or variable coefficients; some are due to non-normality.

These rather vague-sounding statements combine to yield some meatier consequences:

(c) The eigenvalues of the Jacobian give too liberal a condition for the absence of stiffness; they are tied to the limit $t \to \infty$.

(d) The norm of the Jacobian (in nonlinear parlance the Lipschitz constant) gives too conservative a criterion for the absence of stiffness; it is tied to the limit $t \to t_0$.

If the Jacobian is a normal matrix (e.g., symmetric or skew-symmetric), then there is little difference between transient and asymptotic behavior and these observations are unimportant. They come into their own when the Jacobian is highly non-normal, a situation that is known to arise, for example, in certain method-of-lines calculations for non-self-adjoint partial differential equations [4, 15, 19].

The purpose of this paper is to present these observations and, by drawing on some recent theory in the literature of the numerical solution of PDEs, to show that standard characterizations of instability and stiffness can be made more correct if eigenvalues are replaced by pseudospectra. Loosely speaking, our conclusions are as follows:

(I) Numerical instability for $t \approx t_0$ occurs when the pseudospectra of the linearized, frozen coefficient approximation fail to fit in the stability region of the ODE formula;

(II) A problem is stiff for $t \approx t_0$ if the pseudospectra of this linear approximation extend far into the left half-plane as compared with the time scale of the solution for $t \approx t_0$.

These statements are made precise in Sections 2 and 3, especially Theorems 1 and 2, and a summary of our view of stability and stiffness is given in Section 4.

We illustrate our points by numerical examples in which we solve the problem with an adaptive non-stiff ODE solver, then examine the sizes of the adaptively determined time steps. Several authors have used this approach to provide a practical measure of the degree of stiffness [5, 6, 7, 8, 14, 23, 24]. We give the greatest attention to examples illustrating point (c), because that is the most surprising result: sometimes, for purely linear reasons, a problem may be stiffer than eigenvalue analysis seems to suggest.

The Conclusions section comments on how our observations relate to the extensive literature by non-numerical mathematicians concerning stability of ODEs.

2. Reduction of ODEs to model problems.

$$y' = f(t,y) \tag{2.1}$$

$\downarrow$ LINEARIZE

$$u' = A(t)u \tag{2.2}$$

$\downarrow$ FREEZE COEFFICIENTS

$$u' = Au \tag{2.3}$$

$\downarrow$ DIAGONALIZE

$$u' = \lambda u. \tag{2.4}$$

Fig. 1. The standard paradigm: reduction to a collection of scalar model problems.

Figure 1 summarizes the standard paradigm for the reduction of questions of instability and stiffness to scalar model problems. We begin with (2.1), a system of N first-order ODEs, and let $y_0(t)$ denote a particular solution to (2.1) that we are interested in. If we make the substitution $y(t) = y_0(t) + u(t)$, then instability and stiffness depend on the evolution of $u(t)$.

The first step is to *linearize* the equation by assuming u is small. If f is differentiable, let

$$A(t) = \frac{\partial f}{\partial y}(t, y_0(t))$$

denote its Jacobian (an $N \times N$ matrix) at time t. By neglecting terms of order u^2 and using the identity $y_0'(t) = f(t, y_0(t))$ we pass from (2.1) to (2.2).

The second step is to *freeze coefficients* by setting $A = A(t_0)$ for some t_0 of interest. The idea here is that instability and stiffness are fundamentally transient phenomena, which may appear near some times t_0 and not others. The result is the constant coefficient linear problem (2.3).

Finally, assuming A is diagonalizable, we *diagonalize* it. This decomposes (2.3) into N independent scalar equations (2.4) with $\lambda \in \Lambda(A)$ (the spectrum of A). According to a standard way of thinking, the model problems (2.4) can now be used to estimate instability and stiffness. To determine whether a discrete approximation to the ODE is stiff for the particular solution y_0 near the time t_0, one examines all N of the problems (2.4) with $\lambda \in \Lambda(A)$. Depending on the author and the application, stiffness is then associated with eigenvalues λ in the left half-plane of widely varying moduli or real parts – or more precisely, with eigenvalues whose moduli or real parts are large compared with the time scale of the underlying solution $y_0(t)$ for $t \approx t_0$.

Now it is well known that this reduction to scalar problems sometimes leads to incorrect conclusions about instability and stiffness. This brings us to point (b) of the Introduction. Three successive approximations are involved in passing from (2.1) to (2.4). Nevertheless, only the first two of these, linearization and freezing of coefficients, have received much attention in the literature. We wish to argue that the diagonalizations (2.3) → (2.4) is also a process that may change the nature of an ODE significantly, and in fact, that some effects which are customarily attributed to linearization or freezing of coefficients can with greater justice be blamed on diagonalization.

Consider first the step (2.1) → (2.2): linearization. It is clear that important effects may be missed by pretending that the perturbations that arise in practice are small enough to evolve linearly. Consequently there is a large literature that remains with equation (2.1) by considering stiffness, stability and convergence for certain classes of nonlinear ODEs [1, 2, 5, 10, 14, 15]. The central theme in this literature is the search for methods that can be proved convergent by arguments related to *contractivity*. The idea here, which was made famous by Liapunov, is that if a function can be found that decreases locally for all y_0 and t_0, then global decrease of that function is also assured, regardless of nonlinearity. (The function in question is often a measure of the difference of two nearby solutions $y(t)$ and $y_0(t)$; of course, bounded exponential growth rather than strict contractivity is adequate for many purposes.) In the numerical analysis of PDEs this idea goes by the name of *strong stability* [22].

Many powerful results have been established in this nonlinear tradition, but we wish to make two remarks concerning their applicability to general questions of instability and stiffness. First, because of the emphasis on contractivity, most of these results involve rather stringent conditions that are sufficient but not always necessary for stability or convergence. Many ODE methods behave well even without being contractive, including most of those used in practice, and thus the nonlinear theory is too conservative to provide a sharp characterization of the phenomenon of stiffness, or an analysis of the time-stepping behavior of practical adaptive ODE software. (See the final page of [14].) Second, the "nonlinear" theory differs from the "linear" theory even for linear problems with constant coefficients, where it amounts to analysis of (2.3) by means of such quantities as the one-sided Lipschitz constant of f, the logarithmic norm of A, and the norm of the discrete solution operator associated with A. In this linear context it becomes particularly apparent that the theory is too conservative to be sharp, as we shall explain at the end of the next section.

Consider next the step (2.2) → (2.3): freezing of coefficients. This process has received considerable attention, and examples have been devised to show that freezing coefficients may change the behavior of an ODE significantly. An appreciation of the importance of this phenomenon goes back to Liapunov and Poincaré at the beginning of this century. Possibly the earliest explicit example is due to Perron [18], and three examples that have been studied by numerical analysts are those of Vinograd [30], with generalizations by Dekker and Verwer [2, 13, 14], Kreiss [11], and Lambert [14, p. 263]. In each of these cases a variable coefficient ODE with rapidly growing solutions has Jacobians at each point that appear mildly behaved.

Consequently it may be dangerous to make predictions about stiffness based on frozen coefficient approximations. These examples are quite compelling and we recommend Kreiss's example in particular for its simplicity; see Example 3 of Section 5.

Finally, consider the step (2.3) → (2.4): diagonalization. In the literature it is surprising how often this step receives no comment whatsoever; equations (2.3) and (2.4) are viewed simply as equivalent. As examples of this way of thinking, here are extracts from the books by Dekker and Verwer [2, p. 12] and (slightly edited) Lambert [14, pp. 76–77]:

> *Dekker and Verwer 1984:*
> Lambert's question is of a fundamental nature. It in fact illustrates the need for a more rigorous numerical stability theory than the established linear theory which is based on local linearization and thus assumes that the spectrum does determine the error propagation.

> *Lambert 1991:*
> The flaw in this argument lies in the assumption
>
> $$\frac{\partial f}{\partial y} = J, \text{ a constant matrix.}$$
>
> It is simply not true in general that the eigenvalues of J always correctly represent the behavior of the solutions of the nonlinear system.

The striking thing about both of these extracts is that they pass wordlessly from a matrix to its eigenvalues, unconsciously making the assumption *linear ⇒ scalar*. This assumption is common in the numerical ODE literature, though sometimes authors are more careful. In fact, the authors of [2] and [14] are themselves sometimes more careful, as for example on p. 42 of [2]. Perhaps the fairest summary is to say that the experts in the field of numerical analysis of ODEs are aware in principle that a matrix is not the same as a set of scalars, but that this awareness is easily overlooked when it comes to applications.[1]

We shall now examine the problem (2.3) and explain why it may be misleading to view (2.3) and (2.4) as equivalent.

[1] An analogous but more serious situation holds in the century-old field of hydrodynamic stability, where the fundamental problem is to understand the mechanism by which the laminar flow of a fluid becomes unstable and eventually turbulent. Here the equations are nonlinear but often have constant coefficients, so that the standard paradigm of Figure 1 simplifies from three steps to two. Nevertheless the significance of the step of diagonalization has been overlooked in this field too, so that when predictions based on eigenvalues have failed to match laboratory experiments, as they consistently do, the blame has been placed entirely on the first step, linearization. Only recently has it emerged that the operators in question are highly non-normal and that linear effects unrelated to eigenvalues are of central importance to the physics of hydrodynamic instability; see [28] and the references therein.

3. The problem $u' = Au$.

Consider the linear, homogeneous, constant coefficient problem (2.3), where A is a fixed $N \times N$ matrix. Let this ODE be approximated by a fixed discrete formula with constant step size Δt; to be definite, let it be an explicit Runge-Kutta formula of order m. Then the discrete analogue of (2.3) is the one-step recurrence relation $v^{(n+1)} = p(A\Delta t)v^{(n)}$, that is,

$$v^{(n)} = p(A\Delta t)^{n-n_0}v^{(n_0)},$$

where $p(z)$ denotes (for the m-stage formulas with $m \leq 4$) the polynomial of degree m obtained by truncating the Taylor series for e^z ($v^{(n)}$ represents the discrete approximation to $u(t)$ at time $t = n\Delta t$). Another way to write the same result is

$$v(t) = p(A\Delta t)^{(t-t_0)/\Delta t}v(t_0),$$

if t_0 and t are multiples of Δt, giving a discrete approximation to the exact solution

$$u(t) = e^{(t-t_0)A}u(t_0).$$

Since the coefficients are constant, we can simplify these expressions for the purposes of analysis by assuming that $n_0 = t_0 = 0$.

How then does $p(A\Delta t)^n$ behave as a function of n? Since u is ultimately a perturbation of a solution $y_0(t)$ to (2.1) that is in principle more or less arbitrary, the right quantity to investigate is the norm $\|p(A\Delta t)^n\|$ induced by some vector norm $\|\cdot\|$. This quantity satisfies the following well-known bounds in terms of the norm of $p(A\Delta t)$ and its spectral radius:

$$\rho(p(A\Delta t))^n \leq \|p(A\Delta t)^n\| \leq \|p(A\Delta t)\|^n. \tag{3.1}$$

Unfortunately, when A is not normal the gap between these bounds may be very wide. To illustrate this, Figure 2 plots $\|p(A\Delta t)^n\|$ for

$$A = \begin{bmatrix} -10 & 5 & 5 & & & \\ & -10 & 5 & 5 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -10 & 5 & 5 \\ & & & & -10 & 5 \\ & & & & & -10 \end{bmatrix} \in \mathbb{R}^{16\times 16}$$

with $\Delta t = 0.175$ and $p(z) = 1 + z + z^2/2$. (We take $\|\cdot\| = \|\cdot\|_2$ in all of the examples of this paper.) Here $\rho(p(A\Delta t)) = 0.78125$ and $\|p(A\Delta t)\| \approx 2.003$, so the bounds (3.1) on $\|p(A\Delta t)^n\|$ diverge exponentially as $n \to \infty$, as indicated by the dashed lines in the figure. As for $\|p(A\Delta t)^n\|$, it begins by tracking the upper bound $\|p(A\Delta t)\|^n$, growing by more than five orders of magnitude, then starts to decrease and eventually decays at a rate determined by $\rho(p(A\Delta t))$. At $t = 10$, for example, $\|p(A\Delta t)^n\|$ lies about 12 orders of magnitude below $\|p(A\Delta t)\|^n$ and about 11 orders of magnitude above $\rho(p(A\Delta t))^n$.

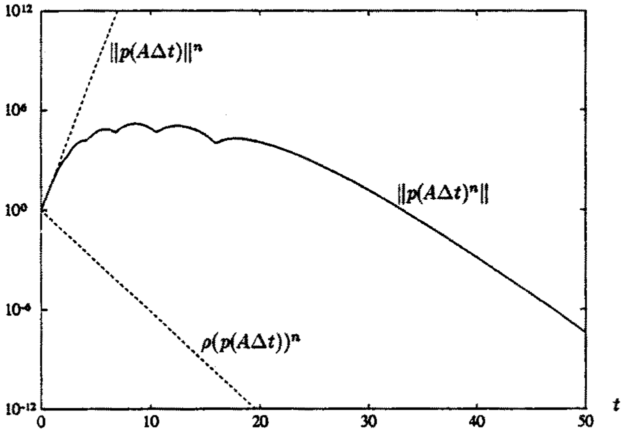

Fig. 2. Unstable transient growth of $\|p(A\Delta t)^n\|$ in a problem with $\rho(p(A\Delta t)) < 1 < \|p(A\Delta t)\|$.

These phenomena are trivial from the point of view of linear algebra, but their implications for stability and stiffness are often overlooked. If the hump in a plot like Figure 2 is large, the computation is likely to behave unstably despite seemingly favorable eigenvalues. If the hump is small, it is likely to behave stably despite a seemingly unfavorable norm. It is the size of the hump that matters: the behavior of $\|p(A\Delta t)^n\| = \|p(A\Delta t)^{t/\Delta t}\|$ for small but nonzero t. The eigenvalues and the norm, by contrast, give sharp information only about the limits $t \to \infty$ or $t \to 0$. We can summarize these connections as follows:

(3.2) behavior as $t \to \infty$: determined by the spectrum of $p(A\Delta t)$ or of $A\Delta t$

(3.3) behavior for finite t: determined by the pseudospectra of $p(A\Delta t)$ or of $A\Delta t$

(3.4) behavior as $t \to 0$: determined by the norm of $p(A\Delta t)$.

Statements (3.2) and (3.4) can be made precise by the identities

$$\lim_{n \to \infty} \|p(A\Delta t)^n\|^{1/n} = \rho(p(A\Delta t)),$$

$$\lim_{n \to 0} \frac{\|p(A\Delta t)^{n+1}\|}{\|p(A\Delta t)^n\|} = \|p(A\Delta t)\|,$$

of which the first is well-known and the second is trivial. Let us now turn to the less familiar statement (3.3) and explain what it means to say that the behavior of $\|p(A\Delta t)^n\|$ for finite n or t is determined by the pseudospectra of $p(A\Delta t)$ or of $A\Delta t$.

One way to estimate the size of the hump in a plot like Figure 2 would be to apply the Kreiss matrix theorem [3, 11, 22, 31] to the matrix $p(A\Delta t)$. This would involve investigating the resolvent norm $\|(zI - p(A\Delta t))^{-1}\|$ as z approaches the unit disk from outside. Alternatively, one can make use of results recently proved in [19] and [20] that amount to transplantations of the Kreiss matrix theorem from the unit disk to the *stability region* S of the ODE formula (see also [3, 15]). Now one must investigate the resolvent norm $\|(zI - A\Delta t)^{-1}\|$ as z approaches S from outside. Here is the essential result:

THEOREM 1. *Let* (2.3) *be modeled as described above by an explicit Runge-Kutta formula with stability region S that satisfies certain technical assumptions described in* [20]. *Then there exist positive constants C_1 and C_2, depending only on the Runge-Kutta formula and on N, such that*

$$(3.5)\quad C_1\mathscr{K} \le \sup_{n\ge 0}\|p(A\Delta t)^n\| \le C_2\mathscr{K}, \qquad \mathscr{K} = \sup_{z\notin S}\operatorname{dist}(z,S)\,\|(zI - A\Delta t)^{-1}\|.$$

Here $\operatorname{dist}(z, S)$ denotes the usual distance of z to the set S and $\mathscr{K}$ might be called the "Kreiss constant." The constant C_1 is of modest size, depending only on the Runge-Kutta formula, while C_2 depends on the Runge-Kutta formula and also linearly on N. The proof of this theorem is given in [20]. An analogous result is also valid for linear multistep formulas; see [19].

For a numerical illustration of Theorem 1, the problem represented in Figure 2 has

$$\mathscr{K} \approx 2.6 \times 10^4, \qquad \sup_{n\ge 0}\|p(A\Delta t)^n\| \approx 1.5 \times 10^5.$$

The two numbers agree to within an order of magnitude. If the dimension of A is increased from 16 to 32 in the same example, the numbers increase to approximately 2.9×10^{10} and 2.5×10^{11}, respectively.

The restatement of these observations in terms of pseudospectra runs as follows. For each $\varepsilon \ge 0$, the ε-*pseudospectrum* [19, 26, 27] of a matrix A is the compact subset of $\mathbf{C}$ defined by

$$\Lambda_\varepsilon(A) = \{z\in\mathbf{C}\colon \|(zI - A)^{-1}\| \ge \varepsilon^{-1}\}.$$

(For $z\in\Lambda(A)$ we set $\|(zI - A)^{-1}\| = \infty$.) Equivalently, $\Lambda_\varepsilon(A)$ is the set of $z\in\mathbf{C}$ that are eigenvalues of some matrix $A + E$ with $\|E\| \le \varepsilon$. Now it is easy to verify the identity

$$(3.6)\qquad \sup_{z\notin S}\operatorname{dist}(z,S)\,\|(zI - A\Delta t)^{-1}\| = \sup_{\varepsilon>0}\varepsilon^{-1}\,\overline{\operatorname{dist}}(\Lambda_\varepsilon(A\Delta t), S),$$

where $\overline{\operatorname{dist}}(A, B)$ denotes $\sup_{z\in A}\operatorname{dist}(z, B)$. Thus Theorem 1 can be restated as follows:

THEOREM 2. *An equivalent formulation of* (3.5) *is*

$$(3.7)\qquad C_1\mathscr{K} \le \sup_{n\ge 0}\|p(A\Delta t)^n\| \le C_2\mathscr{K}, \qquad \mathscr{K} = \sup_{\varepsilon>0}\varepsilon^{-1}\,\overline{\operatorname{dist}}(\Lambda_\varepsilon(A\Delta t), S).$$

In words: the size of the hump in Figure 2 is determined by how far the ε-pseudospectra of $A\Delta t$ are from the stability region.

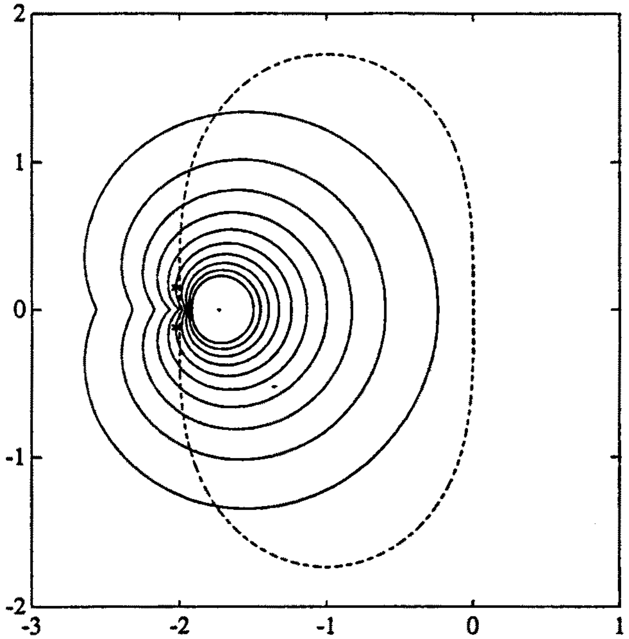

Fig. 3. Boundaries of ε-pseudospectra $\Lambda_\varepsilon(A\Delta t)$ with $\varepsilon = 10^{-1}, 10^{-2}, \ldots, 10^{-10}$ for the same A and Δt as in Figure 2. The dashed curve marks the stability region of the 2nd-order Runge-Kutta formula.

Figure 3 illustrates Theorem 2 for the example we have been considering. The dashed curve is the boundary of the stability region for the second-order Runge-Kutta formula. The solid curves are the boundaries of $\Lambda_\varepsilon(A\Delta t)$ for $\varepsilon = 10^{-1}, 10^{-2}, \ldots, 10^{-10}$ with $\Delta t = 0.175$, and the dot at the center of all of those curves is the solitary eigenvalue $\lambda = -1.75$ (If A were a normal matrix with the same spectrum, all of the curves would lie within the disk of radius 0.1 about λ, well inside the stability region.) By counting contours one sees that although $\Lambda_{10^{-7}}(A\Delta t)$ is contained in S, for example, $\Lambda_{10^{-6}}(A\Delta t)$ is not, and this explains why the size of the hump in Figure 2 is of the order of 10^6. The asterisks in Figure 3 mark the complex conjugate pair of points z that achieve the maximum in (3.6).

Now let us return to the question of stiffness. Our basic point is that instability and hence stiffness are transient phenomena, determined locally by the behavior of (2.1) over a finite interval $[t_0, t_1]$ scaled to contain several but not necessarily many time steps. On such an interval, (2.3) may be a good approximation to (2.1) even though (2.4) is not. In such cases predictions of stable or unstable behavior based on eigenvalue analysis, corresponding to (2.4) and to the limit $t \to \infty$, may be too optimistic, whereas pseudospectral analysis, corresponding to an analysis of (2.3) for finite t, may be quite accurate. Example 1 of Section 5 provides an illustration of this kind.

In the last section we alluded to the fact that the numerical ODE literature features two rather different styles of analysis. The “linear” theory is the theory that emphasizes the eigenvalues of A or of $p(A\Delta t)$, leading to sufficient conditions for instability or stiffness. The “nonlinear” theory emphasizes the Lipschitz constant of f, which amounts to the norm $\|A\|$ if f is differentiable, and also the norm $\|p(A\Delta t)\|$, leading to sufficient conditions for stability or non-stiffness. Thus the essence of the “nonlinear” theory is analysis by norms rather than eigenvalues. As explained above, this normwise analysis is associated with the limit $t \to t_0$, and the predictions it leads to may be too conservative because instability is a cumulative phenomenon, not a phenomenon of a single time step. Example 2 of Section 5 presents an example to illustrate this assertion.

Finally, it is certainly possible that the interval $[t_0, t_1]$ over which (2.3) is accurate contains too few time steps for constant coefficient linear analysis to be of much use, no matter how carefully carried out. In this case neither eigenvalues nor norms nor pseudospectra can be expected to provide sharp predictions in general; one must abandon (2.3) and use other tools. Example 3 of Section 5 is in this category.

The observations made in this section can be viewed as translations into the language of numerical ODEs of principles that are better appreciated in the literature of numerical PDEs, thanks originally to work in the 1950s and 1960s by Lax and Richtmyer, Godunov and Ryabenkii, Kreiss, and others [3, 11, 22]. Perhaps the PDE literature has been particularly attentive to these points because the presence of a second limit process $\Delta x \to 0$ makes it possible to formulate elegant results like the Lax equivalence theorem.

4. Summary of stability and stiffness.

Here is a summary of our view of stability and stiffness. As stated in the Introduction,

(I) Numerical instability for $t \approx t_0$ occurs when the pseudospectra of the linearized, frozen coefficient approximation fail to fit in the stability region of the ODE formula;

(II) A problem is stiff for $t \approx t_0$ if the pseudospectra of this linear approximation extend far into the left half-plane as compared with the time scale of the solution for $t \approx t_0$.

What it means for the pseudospectra to “fit in the stability region” was made precise in Theorem 2: the ε-pseudo-eigenvalues must lie at a distance $\leq C\varepsilon$ from the stability region as $\varepsilon \to 0$, for some C that is not too large. We believe that these statements are appropriate even for nonlinear problems and problems with variable coefficients, so long as the variations involved are resolved by a reasonable number of time steps.

Table 1 summarizes how (I) and (II) relate to the more familiar views that we have

Table 1. *Summary of three theories of stability and stiffness of ODEs. Δ and S denote the closed unit disk and the stability region, respectively, and A is the frozen coefficient Jacobian matrix. See the qualifications listed in the text.*

	"Linear" theory (based on eigenvalues) $(t \to \infty)$	"Nonlinear" theory (based on norms) $(t \to 0)$	Intermediate theory (finite t)
Stiffness	A has a large spectral radius but a small spectral abscissa	A has a large norm but a small logarithmic norm	A has large pseudospectral radii but small pseudo-spectral abscissae
Stability (Δ-plane)	The eigenvalues of $p(A\Delta t)$ lie in Δ	$p(A\Delta t)$ has norm ≤ 1	The pseudospectra of $p(A\Delta t)$ lie close to Δ
Stability (S-plane)	The eigenvalues of $A\Delta t$ lie in S	–	The pseudospectra of $A\Delta t$ lie close to S

called the "linear" and "nonlinear" theories, which should more properly be called the theories based on eigenvalues and norms. Some qualifications to bear in mind are as follows. (1) In the first row of the table, the expressions beginning with "but" are convenient approximations but not really right; the proper definition of stiffness involves a comparison with the time scale of the exact solution for $t \approx t_0$, as stated in (II). (2) Terms such as spectral radius, spectral abscissa, and logarithmic norm are standard ones discussed in many of the references; analogously, the ε-pseudospectral radius and ε-pseudospectral abscissa represent the largest modulus and real part of the ε-pseudospectrum, respectively. (3) If $\|\cdot\|$ is the 2-norm, the logarithmic norm in the middle entry of the first row can be replaced by the numerical abscissa. (4) Expressed without the assumption of differentiability, that entry becomes "f has a large Lipschitz constant but a small one-sided Lipschitz constant." (5) The conditions that the eigenvalues of $A\Delta t$ are in Δ or S or that the norm of $p(A\Delta t)$ is ≤ 1 can be relaxed by terms $O(\Delta t)$.

5. Numereical examples.

We now describe some numerical tests involving ode23, Matlab's adaptive 2nd and 3rd order explicit Runge-Kutta code [17]. We changed one line of the code so as to advance the solution with the 2nd order rather than the 3rd order formula.[2] This was done in order to produce a stable equilibrium state in the sense of Hall [6, 7, 8] and hence make the step size plots easier to interpret. The default local error tolerance of 10^{-3} was used. We wish to emphasize that the phenomena illustrated by

[2] The string "$h*(s1 + 4*s3 + s2)/6$" was changed to "$h*(s1 + s2)/2$".

these examples are not peculiar to the Matlab code; the same effects would arise with any adaptive ODE solver based on explicit formulas.

EXAMPLE 1. We begin with a linear constant coefficient problem of the form

$$y'(t) = Ay(t) + g(t), \qquad t \geq 0, \tag{5.1}$$

with $y(0)$ chosen to have random elements from the $N(0, 1)$ distribution. For the matrix A we take the dimension to be 32, with either

$$A = \text{diag}(-10) \qquad \text{“normal”}$$

or

$$A = \text{bidiag}(-10, 10) \qquad \text{“non-normal”.}$$

For the forcing function we choose either

$$g(t) = 0 \qquad \text{“homogeneous”}$$

or

$$g(t) = w \cos t \qquad \text{“inhomogeneous”.}$$

Here diag(-10) denotes the diagonal matrix with $a_{ii} = -10$, bidiag$(-10, 10)$ denotes the bidiagonal matrix with $a_{ii} = -10$, $a_{i,i+1} = 10$, and w is a fixed vector with random elements from $N(0, 0.01)$ (i.e., mean 0 and standard deviation 0.1). All together, this gives us four problems, to which we give the names “normal, homogeneous,” “normal, inhomogeneous,” etc. Note that the eigenvalues of A are the same in both the normal and non-normal cases. In the homogeneous cases (2.3) is an accurate approximation to (5.1) for all t; in fact the two are identical. The inhomogeneities are introduced to model the more realistic situation in which (2.3) and (5.1) are compatible only for a finite time.

Figure 4 plots the step sizes $\{\Delta t_n\}$ selected by ode23 in solving these four problems over the interval $[0, 30]$. In the normal, homogeneous calculation, Δt_n settles down quickly to the value 0.2, which corresponds to the classical absolute stability limit. (The stability polynomial is $p(z) = 1 + z + z^2/2$, and $-10 \times \Delta t \geq -2$ is the stability condition.) This behavior is in line with the theory of Hall. The addition of the forcing term in the normal, inhomogeneous problem causes oscillations but has little effect on the average Δt. On the other hand substantial changes occur when we switch to the non-normal matrix $A = \text{bidiag}(-10, 10)$. For the non-normal, homogeneous problem, Δt is initially much smaller than 0.2, and only for large t does it begin to approach that value. With the introduction of the forcing term in the non-normal, inhomogeneous problem, Δt remains close to 0.1 for all t.

These results can be explained as follows. With the normal Jacobian $A = \text{diag}(-10)$, the classical eigenvalue-based analysis accounts for transient as well as asymptotic behavior, but with $A = \text{bidiag}(-10, 10)$ the transient is very different from the asymptote. Specifically, the $N \times N$ matrix $A_N = \text{bidiag}(-10, 10)$ has ε-pseudospectra which converge as $N \to \infty$ and $\varepsilon \to 0$ to the complex disk

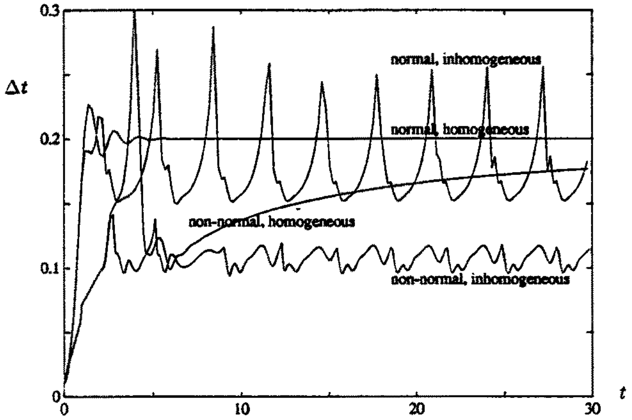

Fig. 4. Adaptively selected step sizes as a function of t for Example 1. The step size in the non-normal inhomogeneous case is cut in half because the pseudospectra of A extend about twice as far along the negative real axis as the spectrum.

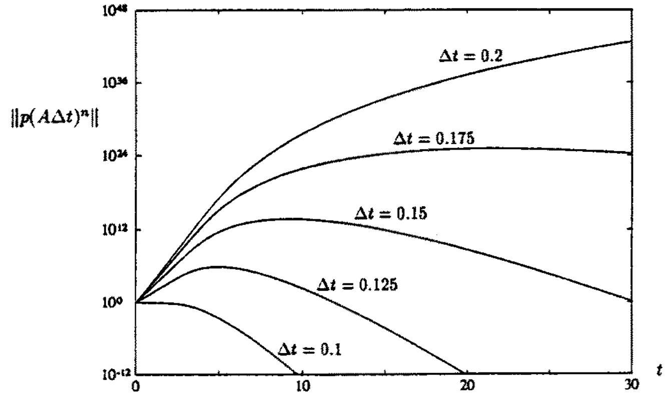

Fig. 5. Explanation of Figure 4: $\|p(A\Delta t)^n\|$ vs. t for various step sizes Δt.

$B(-10,10)$ of center -10 and radius 10 [21]. In particular, for small ε, A_N has some pseudo-eigenvalues that are approximately -20, and constraining these to lie in the stability region requires $-20\Delta t \geq -2$, or $\Delta t \leq 0.1$. In the non-normal, homogeneous calculation this restriction applies to some degree during a rather long transient. In the non-normal, inhomogeneous case, new forcing data are continually being introduced and it applies forever. *The problem never leaves the transient.* Over any interval $[t_0, t_1]$ of moderate size (5.1) can be modeled reasonably well by an approximation in the form of an initial-value problem for the ODE (2.3), but the behavior of one of those approximations as $t \to \infty$ has no relevance to (5.1).

Further explanation of these phenomena is presented in Figure 5, which plots $\|p(A\Delta t)^n\|$ against $t_n = n\Delta t$ with $A = \text{bidiag}(-10, 10)$ for values $\Delta t = 0.1, 0.125, 0.15, 0.175, 0.2$. Although $\|p(A\Delta t)^n\|$ remains bounded for all t for any $\Delta t < 0.2$, it achieves

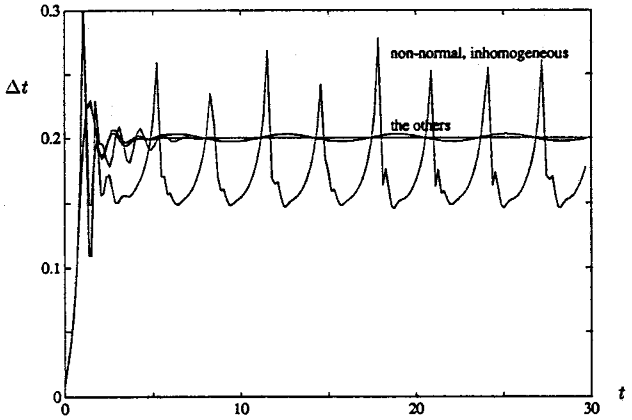

Fig. 6. Adaptively selected step sizes as a function of t for Example 2. The step sizes are far larger than the norm of A might suggest, but match predictions based on ε-pseudospectra (quite close to the spectrum in this case, for small ε).

values in the figure as high as 10^{43} for $\Delta t = 0.2$ and 10^{13} for $\Delta t = 0.15$. The hump is absent only in the lowest curve shown, corresponding to the step size $\Delta t = 0.1$ for which the pseudospectra of $A\Delta t$ as well as the spectrum lie close to the stability region.

Example 2. Our next set of tests illustrates the complementary point of this paper, that norm-based estimates of stiffness may be too conservative. We consider again the equation (5.1), but now with A reduced to the 2×2 matrix

$$A = \begin{bmatrix} -10 & 0 \\ 0 & -1 \end{bmatrix} \qquad \text{“normal”}$$

or

$$A = \begin{bmatrix} -10 & 100 \\ 0 & -1 \end{bmatrix} \qquad \text{“non-normal”},$$

with analogous forcing functions g to those used earlier. For small ε, the ε-pseudospectra of both of these matrices are not very different from the spectrum. The corresponding ode23 step sizes are plotted in Figure 6. As in Example 1, there is a reduction in the average step sizes when the Jacobian is non-normal. However, it is very slight, far less than the factor of 10 that the large norm of A might suggest. For example, with $\Delta t \approx 0.15$ we have $\Delta t \, \|A\| \approx 15$ and $\|p(A\Delta t)\| \approx 5$, both of which might suggest that the calculation will be unstable, but in fact a calculation with that time step is entirely stable because the norms $\|p(A\Delta t)^n\|$ do not continue to grow for $n > 1$.

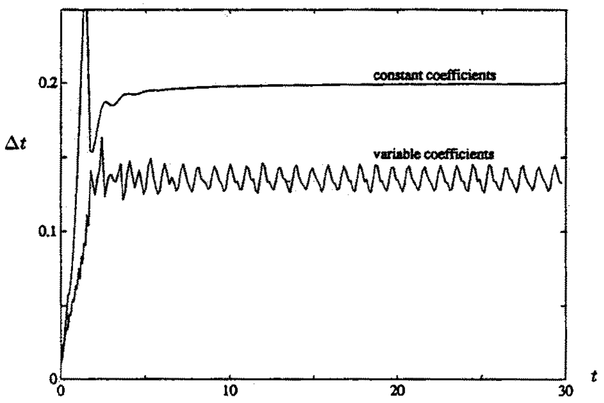

Fig. 7. Adaptively selected step sizes as a function of *t for Example* 3.

EXAMPLE 3. Finally, for our third example we return to the question of the size of the interval $[t_0, t_1]$ on which the linear approximation (2.3) is valid. For a problem with rapidly varying coefficients, this interval may be so small that (2.3) fails to predict stiffness correctly, even when its transient behavior is analyzed properly. The failure may be in either direction; here it is in the direction of excessive optimism.

We consider the elegant example of Kreiss [12]

$$y'(t) = \varepsilon^{-1} U^T(t) \begin{bmatrix} -1 & \eta \\ 0 & -1 \end{bmatrix} U(t) y(t) \equiv A(t) y(t),$$

where

$$U(t) = \begin{bmatrix} \cos \alpha t & -\sin \alpha t \\ \sin \alpha t & \cos \alpha t \end{bmatrix}$$

is a time-varying orthogonal matrix. (We have transposed $U(t)$ from [12] in order to correct a minor error.) Under the transformation $v(t) = U(t)y(t)$, we find that

$$v(t)' = \begin{bmatrix} -\varepsilon^{-1} & \eta\varepsilon^{-1} - \alpha \\ \alpha & -\varepsilon^{-1} \end{bmatrix} v(t). \tag{5.2}$$

Note that although the original time-dependent Jacobian $A(t)$ has eigenvalues $-\varepsilon^{-1}$ for all t, under the norm-preserving transformation the new, constant coefficient Jacobian in (5.2) has eigenvalues $-\varepsilon^{-1}(1 \pm \sqrt{(\alpha\varepsilon(\eta - \alpha\varepsilon))})$. Hence, it is possible to choose a small value of η, so that the Jacobian $A(t)$ never appears far from normal, whilst fixing α so that the eigenvalues of (5.2) differ markedly from $-\varepsilon^{-1}$. Obviously it is these latter eigenvalues that govern the actual behavior of the variable coefficient problem. In Floquet theory, which is a general theory of ODEs with periodic

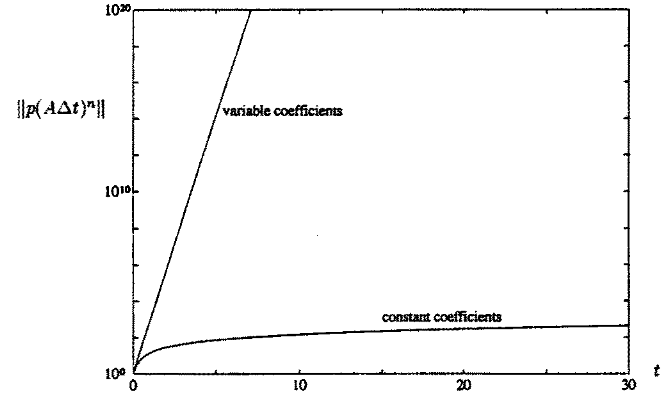

Fig. 8. Explanation of Figure 7: $\|p(A\Delta t)^n\|$ vs. $t = n\Delta t$ and its variable coefficient analogue for $\Delta t = 0.2$.

coefficients, they are called the *characteristic exponents* of the variable coefficient problem.

To be specific, we choose $\varepsilon = 0.1$, $\alpha = 10$, and $\eta = 13/9$, giving eigenvalues $-50/3$ and $-10/3$ in (5.2). The ode23 step sizes, for a random $y(0)$, are plotted in Figure 7. For comparison, the step sizes for the corresponding problem with $U(t) = I$ are also plotted. In the latter case, where the Jacobian is not time-dependent, we see the classical limiting value 0.2, but in the former case an average value of approximately 0.135 appears, not far from the value 0.12 that one might expect based on the eigenvalue $-50/3$. (In fact a more careful analysis involving periodic coefficients explains the value 0.135 exactly; see [9] for details.) Thus we may regard the variable coefficient problem as nearly 5/3 times as stiff as the constant coefficient problem for this particular computation.

Figure 8 explains what is happening. With $\Delta t = 0.2$, any frozen Jacobian gives only modest growth of $\|p(A\Delta t)^n\|$, shown in the lower curve of the figure. However, the actual computation is governed by accumulated products of the time-varying Jacobian,

$$\mathrm{prod}(t_n) = \prod_{i=0}^{n-1}\left(I + \frac{\Delta t}{2}(A(i\Delta t) + A((i+1)\Delta t)) + \frac{\Delta t^2}{2}A((i+1)\Delta t)A(i\Delta t)\right),$$

shown in the upper curve. Here the small amount of growth made possible by the fact that $\|p(A\Delta t)\| > 1$ for each t is compounded geometrically from step to step by the rotations due to the variation in $U(t)$. The frozen approximation is accurate for only about one time step and is of no use in predicting stability.

6. Conclusions.

In the preface to his treatise of 1907 Liapunov wrote [16, our translation]:

> The problem I posed myself in undertaking the present study can be formulated as follows: to determine the circumstances in which the first approximation correctly resolves the question of stability, and those in which it does not.

Liapunov's "first approximation" (a standard term) is the linearized equation (2.2). In effect he and many other mathematicians of this century, whose books the reader can find under the Library of Congress classifications QA372 and QA871, have been concerned with the problem of making precise the "standard paradigm" of Figure 1. What then is the need for additional papers on this subject by numerical analysts?

One answer is that because numerical methods are discrete, the mathematicians' left half-plane must be replaced by the stability region of a discrete ODE formula. This, however, is a relatively straightforward matter.

The more interesting and more fundamental answer is point (a) of the Introduction. Justifiably or not, mathematicians who study the stability of ODEs have been concerned almost exclusively with the limit $t \to \infty$. Numerical instability and stiffness, by contrast, are transient phenomena that depend on how effects compound over a dozen or so time steps. Thus local approximations have a special relevance to numerical analysis. In particular, provided that the effects of nonlinearities or variable coefficients unfold on time scales containing many time steps, the linear, constant coefficient model (2.3) can be expected to be a good guide to instability and stiffness. An analysis of this equation for finite t leads naturally to pseudospectra (Theorem 2) and thence to the conclusions summarized as points (I) and (II) in the Introduction and Section 4.

How important is all of this in practice? Do ODEs arise in scientific computing whose Jacobian matrices are so far from normal that the distinction between spectra and pseudospectra is important? We must be honest and admit that we do not know the answer to this question. Our suspicion is that in the majority of cases the distinction is not important, but that there is a significant minority for which it does matter.

One situation in which highly non-normal ODEs arise, as mentioned in the Introduction, is in method-of-lines computations for the numerical solution of non-self-adjoint PDEs. In particular, consider a method of lines discretization by a Legendre spectral collocation method of the initial boundary value problem

$$u_t = u_x + g(x,t), \qquad u(x,0) = \cos^2(\pi x/2), \qquad u(1,t) = 0$$

on the interval $[-1, 1]$. In [29] and [25] the numerical properties of this example have been studied at length, with the conclusion that the Lax-stability limit on Δt for an explicit method is $O(N^{-2})$ as compared with $O(N^{-1})$ for "eigenvalue stability."

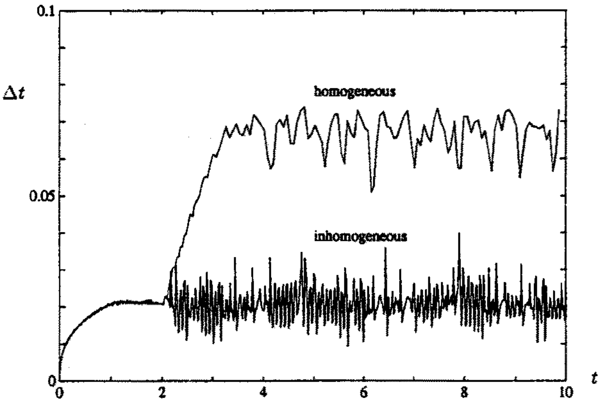

Fig. 9. Adaptively selected step sizes as a function of t for the method-of-lines example from [29].

We close this paper by presenting Figure 9, a plot of adaptively determined time steps when this same method-of-lines problem is solved by Matlab's code ode45. (The spectrum lies too near the imaginary axis for ode23 to produce interesting results.) The grid and hence the size of the system is $N = 50$ and the forcing function for the inhomogeneous calculation is $g(x, t) = 0.1 \cos t \cos^2(\pi x/2)$. The figure shows a time step gap of a factor of about 3.5 between the homogeneous and inhomogeneous problems. Evidently the stiffness of this system of ODEs is controlled in general by pseudospectra, not spectra.

Acknowledgements.

Interaction with many colleagues has helped us to clarify our thoughts on stiffness and instability. In particular, we thank Hans Kraaijevanger for offering detailed comments on this manuscript.

REFERENCES

1. J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, Wiley, 1987.
2. K. Dekker and J. G. Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North-Holland, Amsterdam, 1984.
3. J. L. M. van Dorsselaer, J. F. B. M. Kraaijevanger and M. N. Spijker, *Linear stability analysis in the numerical solution of initial value problems*, in Acta Numerica 1993, Cambridge U. Press, to appear.
4. D. F. Griffiths, I. Christie and A. R. Mitchell, *Analysis of error growth for explicit difference schemes in conduction-convection problems*, Int. J. Numer. Meth. Engr. 15 (1980), 1075–1081.
5. E. Hairer and G. Wanner, *Solving Differential Equations II: Stiff and Differential-Algebraic Problems*, Springer-Verlag, Berlin, 1991.
6. G. Hall, *Equilibrium states of Runge-Kutta schemes*, ACM Trans. Math. Soft. 11 (1985), 289–301.
7. G. Hall, *Equilibrium states of Runge-Kutta schemes:* part II, ACM Trans. Math. Soft. 12 (1986), 183–192.

8. G. Hall and D. J. Higham, *Analysis of stepsize selection schemes for Runge-Kutta codes*, IMA J. Numer. Anal. 8 (1988), 305–310.

9. D. J. Higham, *Runge-Kutta stability on a Floquet problem*, Numer. Anal. Rep. NA/138, University of Dundee, 1992.

10. J. F. B. M. Kraaijevanger, *Stability and convergence in the numerical solution of stiff initial value problems*, Ph.D. Thesis, Inst. Appl. Math. and Comp. Sci., U. Leiden, The Netherlands, 1986.

11. H.-O. Kreiss, *Über die Stabilitätsdefinition für Differenzengleichungen die partielle Differentialgleichungen approximieren*, BIT 2 (1962), 153–181.

12. H.-O. Kreiss, *Difference methods for stiff ordinary differential equations*, SIAM J. Numer. Anal. 15 (1978), 21–58.

13. J. D. Lambert, *Stiffness*, in *Computational Techniques for Ordinary Differential Equations*, eds. I. Gladwell and D. K. Sayers, Academic Press, 1980, 19–46.

14. J. D. Lambert, *Numerical Methods for Ordinary Differential Systems*, Wiley, Chichester, UK, 1991.

15. H. W. J. Lenferink and M. N. Spijker, *On the use of stability regions in the numerical analysis of initial value problems*, Math. Comp. 57 (1991), 221–237.

16. M. A. Liapunov, *Problème Général de la Stabilité du Mouvement*, Princeton U. Press, 1949 (French translation of Russian book of 1907).

17. C. B. Moler, J. N. Little and S. Bangert, *PC-Matlab User's Guide* and *Pro-Matlab User's Guide*, The MathWorks, Inc., 21 Eliot St., South Natick, Massachusetts 01760, 1987.

18. O. Perron, *Die Stabilitätsfrage bei Differentialgleichungen*, Math. Zeit. 32 (1930), 703–728.

19. S. C. Reddy and L. N. Trefethen, *Lax-stability of fully discrete spectral methods via stability regions and pseudo-eigenvalues*, Comp. Math. Appl. Mech. Eng. 80 (1990), 147–164.

20. S. C. Reddy and L. N. Trefethen, *Stability of the method of lines*, Numer. Math. 62 (1992), 235–267.

21. L. Reichel and L. N. Trefethen, *Eigenvalues and pseudo-eigenvalues of Toeplitz matrices*, Lin. Alg. Applics. 162–164 (1992), 153–185.

22. R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial Value Problems*, 2nd ed., Wiley, New York, 1967.

23. B. R. Robertson, *Detecting stiffness with explicit Runge-Kutta formulas*, Technical Report 193/87, Dept. Comp. Sci., U. Toronto, Canada.

24. L. F. Shampine, *Stiffness and nonstiff differential equation solvers*, in *Numerische Behandlung von Differentialgleichungen*, ed. L. Collatz, International Series of Numerical Mathematics, 27 Birkhäuser Verlag, Basel, 1975, 287–301.

25. L. N. Trefethen, *Lax-stability vs. eigenvalue stability of spectral methods*, in *Numerical Methods for Fluid Dynamics III*, eds. K. W. Morton and M. J. Baines, Clarendon Press, Oxford, 1988, 237–253.

26. L. N. Trefethen, *Pseudospectra of matrices*, in *Numerical Analysis* 1991, eds. D. F. Griffiths and G. A. Watson, Longman, 234–266.

27. L. N. Trefethen, *Spectra and Pseudospectra: The Behavior of Non-Normal Matrices and Operators*, book to appear.

28. L. N. Trefethen, A. E. Trefethen and S. C. Reddy, *Pseudospectra of the linear Navier-Stokes evolution operator and instability of plane Poiseuille and Couette flows*, TR 92–1291, Dept. of Comp. Sci., Cornell U., June 1992.

29. L. N. Trefethen and M. R. Trummer, *An instability phenomenon in spectral methods*, SIAM J. Numer. Anal. 24 (1987), 1008–1023.

30. R. E. Vinograd, *On a criterion of instability in the sense of Lyapunov of the solutions of a linear system of ordinary differential equations*, Dokl. Akad. Nauk. SSSR 84 (1952), 201–204 (Russian).

31. E. Wegert and L. N. Trefethen, *From the Buffon needle problem to the Kreiss matrix theorem*, Amer. Math. Monthly, to appear.