SIAM Rev.
Vol. 40, No. 1, pp. 40–73, March 1998

002

FROM SEMIDISCRETE TO FULLY DISCRETE: STABILITY OF RUNGE–KUTTA SCHEMES BY THE ENERGY METHOD*

DORON LEVY† AND EITAN TADMOR†‡

Abstract. The integration of semidiscrete approximations for time-dependent problems is encountered in a variety of applications. The Runge–Kutta (RK) methods are widely used to integrate the ODE systems which arise in this context, resulting in large ODE systems called methods of lines.

These methods of lines are governed by possibly ill-conditioned systems with a growing dimension; consequently, the naive spectral stability analysis based on scalar eigenvalues arguments may be misleading. Instead, we present here a stability analysis of RK methods for well-posed semidiscrete approximations, based on a general *energy method*. We review the stability question for such RK approximations, and highlight its intricate dependence on the growing dimension of the problem. In particular, we prove the *strong stability* of general fully discrete RK methods governed by *coercive* approximations.

We conclude with two nontrivial examples which demonstrate the versatility of our approach in the context of general systems of convection-diffusion equations with variable coefficients. A straightforward implementation of our results verify the strong stability of RK methods for local finite-difference schemes as well as global spectral approximations.

Since our approach is based on the energy method (which is carried in the physical space), and since it avoids the von Neumann analysis (which is carried in the dual Fourier space), we are able to easily adapt additional extensions due to nonperiodic boundary conditions, general geometries, etc.

Key words. L^2-stability, resolvent condition, method of lines, Runge–Kutta methods, region of absolute stability, energy method, finite-difference schemes, spectral methods

AMS subject classifications. Primary, 65M10; Secondary, 35A12

PII. S0036144597316255

1. Introduction.

We are concerned with the stability of RK methods. Specifically, we focus our attention on the stability of such methods when used to integrate large ODE systems which arise in the context of semidiscrete approximations for time-dependent problems.

The classical stability analysis of RK methods deals with the prototype scalar model, $u_t = \lambda u$. A RK method is *(strongly) stable* provided its time step, Δt, is sufficiently small so that $\Delta t \cdot \lambda$ belongs to its *region of absolute stability*. What does the scalar analysis tell us about the stability of RK methods for finite-dimensional *systems*, $u_t = Lu$? If L is diagonalizable, say $L = K^{-1}\Lambda K$, then such a system can be disassembled into a direct sum of scalar equations; consequently, if $\Delta t \cdot \Lambda$ belongs to the region of absolute stability associated with a RK method, then the corresponding RK solution is bounded by the condition number of L given by $\kappa_L := \|K^{-1}\| \cdot \|K\|$.

The stability question becomes more intricate, however, for ill-conditioned L's. Such ill-conditioned L's may arise from large systems which govern spatial discretizations of time-dependent problems. These spatial discretizations involve a small spatial scale—typically, the spatial gridsize of a finite-difference or a finite-element scheme, the highest frequency of a spectral scheme, etc.; denote this small spatial scale by

*Received by the editors February 5, 1997; accepted for publication (in revised form) April 16, 1997. This research was supported by ONR grant #N00014-91-J-1076, NSF grant #DMS94-04942, and by the Sackler Institute for Scientific Computations at Tel-Aviv University.

http://www.siam.org/journals/sirev/40-1/31625.html

†School of Mathematical Sciences, Tel-Aviv University, Tel-Aviv 69978 Israel (dlevy@math.tau.ac.il).

‡Department of Mathematics, University of California at Los Angeles, Los Angeles, CA 90095 (tadmor@math.ucla.edu).

N^{-1}. Then, these semidiscrete approximations, called *methods of lines*, result in $N \times N$ systems of ODEs, $u_t = L_N u$, and we are interested in whether the RK methods retain the stability of such approximations. In this case, the condition number, $\kappa_N = \kappa_{L_N}$, as well as the RK solution, may grow together with the increasing size of L_N. The straightforward eigenvalues analysis will not suffice as L_N may become ill conditioned with the increasing dimension N. The spectral analysis in this context of ill-conditioned systems may therefore be misleading. We study the *strong stability* issue by the energy method which takes into account the full geometric structure of the eigensystem of L_N. The stability question for such ill-conditioned semidiscrete systems has received considerable attention in recent years, and we refer in particular, to the works of Kreiss, Trefethen and their coworkers, which analyze this question in terms of the (weaker) *resolvent stability*, [41], [43], [73], [56].

In this review we address the question of strong stability of Runge–Kutta methods for semidiscrete approximations, $u_t = L_N u$. We begin with the preliminary section 2, where we present a self-contained overview of the ingredients which serve us throughout this work: the notions of L^2-stability, the resolvent condition, and the essential features of RK methods are briefly overviewed. In particular, we highlight the dependence of the (weighted-) L^2-stability and the resolvent stability for systems with an increasing dimension N.

Our main results are then presented in section 3. We consider well-posed spatially discretized systems, $u_t = L_N u$. The wellposedness of such semidiscrete systems is linked to $L = L_N$ being negative, i.e., $\Re(Lu, u) \leq 0$. Here, $(\cdot, \cdot)$ is an appropriate *weighted* inner product which is inherited from the underlying well-posed time-dependent problem. As noted above, the RK approximations of such general negative systems need not be stable due to possible ill conditioning. Instead, we identify the admissible systems with strongly stable RK solutions: theses are the *coercive* L's, i.e., $\Re(Lu, u) \leq -\eta \|Lu\|^2$, with fixed $\eta > 0$. Our main result, stated in Theorems 3.2 and 3.3, asserts that *the third- and fourth-order RK methods retain the strong stability of semidiscrete systems governed by coercive L's.* The proof proceeds by a straightforward energy method, which shows strong stability for sufficiently small $\Delta t/\eta$.

We end in section 4 by presenting two nontrivial examples which demonstrate the *versatility* of our stability results. First, we investigate the stability of finite difference approximations to *general systems* of convection-diffusion equations with variable coefficients. The spatial variables are discretized either by local finite differences or by global spectral methods, and we address the question of strong stability for the corresponding fully discrete Runge–Kutta schemes. Our main results prove the strong stability of such coercive approximations. The requirement of coercivity is translated in this context into the familiar notion of difference/spectral approximations which are dissipative in the sense of Kreiss [58, section 5]. We would like to highlight the following advantages of our approach:

- (strong stability). Our results guarantee strong stability in the sense that the temporal growth of the approximate RK solution is no faster then that of the exact differential solution.
- (versatility). The stability question could be addressed, of course, in terms of the classical Lax–Richtmyer theory. This theory, however, does not apply to *global* methods such as spectral differencing. More importantly, since our approach is based on the energy method (which is carried in the physical space), and since it avoids the von Neumann analysis (which is carried in the

dual Fourier space), we are able to easily adapt additional extensions due to nonperiodic boundary conditions, general geometries, etc.

Our second and last example, presents a strong stability study of the RK methods for Chebyshev pseudospectral approximations of convection equations with variable coefficients. This is a canonical example for a semidiscrete approximation governed by an ill-conditioned global differentiation matrix, [73]. Here we utilize the coercivity of the problem proved earlier by Gottlieb and Tadmor in [24]. The question of strong stability for global operators (such as the Chebyshev pseudospectral system) is beyond the scope of the classical von Neumann analysis, and cannot be addressed in terms of a resolvent stability analysis.

2. Preliminaries.

2.1. Spectrum, norms, and weighted norms. We begin with a preliminary discussion on the basic concepts that will play a role throughout the document. In the following, we denote by $\mathcal{X}$ a Banach space equipped with the norm $\|\cdot\|$, and we let $\mathcal{B}(\mathcal{X})$ denote the algebra of bounded linear operators on $\mathcal{X}$.

Recall that the *spectrum* $\sigma(T)$ of an operator $T \in \mathcal{B}(\mathcal{X})$ is the set of all scalars λ such that the operator $T - \lambda I$ is *not* boundedly invertible. We therefore have $\lambda \in \sigma(T)$ iff at least one of the following three statements is true: *either* $(T-\lambda I)^{-1}$ is unbounded, *or* the range of $T - \lambda I$ is not all of $\mathcal{X}$, *or* $T - \lambda I$ is not one-to-one. If the third case holds, λ is said to be an *eigenvalue* of T, and the set of all such eigenvalues forms the *point spectrum* of T. In the finite dimensional case, all we have is the point spectrum.

The *spectral radius* of an operator T is defined by

$$\rho(T) := \sup\{|\lambda| : \lambda \in \sigma(T)\}. \tag{2.1}$$

We note that the spectral radius is not a norm over $\mathcal{B}(\mathcal{X})$ since, generically, it lacks subadditivity: $\rho(T_1 + T_2) \not\leq \rho(T_1) + \rho(T_2)$.

We let $\mathcal{H}$ denote a Hilbert space over the complex field $\mathcal{C}$ with an inner product $(\cdot,\cdot)$ and norm $\|\cdot\| = (\cdot,\cdot)^{1/2}$, and let $\mathcal{B}(\mathcal{H})$ denote the algebra of bounded linear operators on $\mathcal{H}$. Then for any $T \in \mathcal{B}(\mathcal{H})$ we define the *numerical radius* of T by

$$r(T) := \sup\{|(Tu, u)| : \|u\| = 1\}. \tag{2.2}$$

Equivalently, if we let $W(T)$ denote the *numerical range* of T, that is, $W(T) = \{(Tu, u) : \|u\| = 1\}$, then, $r(T) = \sup\{|z| : z \in W(T)\}$. We should emphasize that $(\cdot,\cdot)$ refers to *any* inner-product—not necessarily to the Euclidean inner product (consult Lemma 2.1). Note that the numerical radius, though subadditive, is not submultiplicative, i.e. generically

$$r(T_1T_2) \not\leq r(T_1)r(T_2). \tag{2.3}$$

Finally, for $T \in \mathcal{B}(\mathcal{X})$, we define the *induced operator norm* of T on $\mathcal{B}(\mathcal{X})$ by

$$\|T\| := \sup\{\|Tu\| : \|u\| = 1\}. \tag{2.4}$$

We note that the induced norm is both subadditive and submultiplicative

$$\|T_1 + T_2\| \leq \|T_1\| + \|T_2\|, \qquad \|T_1T_2\| \leq \|T_1\| \cdot \|T_2\|, \qquad \forall T_1, T_2. \tag{2.5}$$

We now turn to the finite-dimensional case, $\mathcal{C}^N$. We begin with the Euclidean setup based on the inner-product $< \cdot, \cdot >$; here,

$$< u, v >= \sum_{i=1}^{N} u_i \overline{v_i}, \qquad |u| =< u, u >^{1/2}$$

are the *Euclidean inner product* and *vector norm*, with the corresponding *numerical radius* and *induced norm*

$$r(T) = \sup_{|u|=1} | < Tu, u > |, \qquad \|T\| = \sup_{|u|=1} |Tu|.$$

For a more general setup, we let H be any Hermitian strictly positive-definite operator, and we denote the *H-inner product* and *H-norm* by

$$(u, v)_H =< u, Hv >, \qquad |u|_H^2 = (u, u)_H.$$

We then define the corresponding *H-numerical radius* and the *H-norm* of an operator $T \in \mathcal{B}(\mathcal{H})$ by

$$r_H(T) = \sup_{|u|_H=1} |(Tu, u)_H|, \qquad \|T\|_H = \sup_{|u|_H=1} |Tu|_H. \tag{2.6}$$

The standard Euclidean definitions of numerical radius and norm correspond to $H = I$, i.e., $r(T) = r_I(T)$ and $\|T\| = \|T\|_I$.

Of course, the H-depended quantities (numerical radius, norm, ...) are special cases of these definitions in general Hilbert spaces. In fact, in the finite-dimensional case, these H-weighted quantities capture the general setup in view of the following.

LEMMA 2.1. *For every inner product $(\cdot, \cdot)$ on $\mathcal{C}^N$, there exists a positive-definite Hermitian matrix H such that*

$$(u, v) =< u, v >_H \qquad \forall u, v \in \mathcal{C}^N.$$

Indeed take $H_{i,j} =< e_i, e_j >$—the Gram matrix in the Cartesian basis.

Weighted quantities describe the equivalent Euclidean (not-weighted) quantities for similar matrices. For example, if we let T_H^* denote the adjoint of T with respect to the H-weighted inner product so that $< Tu, v >_H \equiv < u, T_H^* v >$, then $T_H^* = H^{-1} T^* H$, where $T^* = T_I^*$, is the usual Euclidean adjoint. In particular we have the following result, which for later reference we state as follows.

ASSERTION 2.1. *T is normal, in the sense that it commutes with one of its H-adjoints, iff T is diagonalizable.*

Indeed, T is normal, $TT_H^* = T_H^* T$, iff it has a complete eigensystem, $T = K^{-1} \Lambda K$ with $H = K^* K$. This equivalence between weighted and Euclidean quantities also manifests itself in the following lemma.

LEMMA 2.2. *If H is a positive Hermitian matrix such that $H = K^* K$, K being the nonsingular square root of H, then*

$$\|T\|_H = \|KTK^{-1}\|, \tag{2.7}$$

$$r_H(T) = r(KTK^{-1}). \tag{2.8}$$

We close this section with a brief discussion on the "hierarchy" between the three quantities mentioned above. First we note that for all positive-definite H's,

$$\rho(T) \le r_H(T) \le \|T\|_H. \tag{2.9}$$

Indeed, for the left inequality, we employ the largest eigenvalue; the right inequality follows from the weighted Cauchy–Schwartz inequality.

When does equality take place in (2.9)? If the operator T has a complete eigensystem, i.e., if $KTK^{-1} = \Lambda$ (diagonal), we have by (2.7)

$$\|T\|_H = \|KTK^{-1}\| = \|\Lambda\| = \rho(\Lambda) = \rho(T). \tag{2.10}$$

Hence, in view of (2.9) and (2.10), we conclude the following.

ASSERTION 2.2. *If T is diagonalizable, then there exists an $H > 0$ which induces an equality in* (2.9)*:*

$$KTK^{-1} = diag(\Lambda) \quad \Longrightarrow \quad \rho(T) = r_H(T) = \|T\|_H, \quad H = K^*K. \tag{2.11}$$

In general, however, a sharp *inequality* in (2.9) is possible—take for example, T as a 2×2 Jordan block. Yet, we claim that by varying over all H-weighted norms, we can deform the gaps in (2.9) to be as small as desired in the finite-dimensional case and in certain more general infinite-dimensional setups which is outlined below.

To this end, let $\mathcal{M}_{N\times N}(\mathcal{C})$ denote the algebra of $N\times N$ matrices over the complex field. We consider operators L which can be decomposed into a direct sum of such finite-dimensional matrices,

$$L = \sum_{A\in\mathcal{F}} \oplus A;$$

here $\mathcal{F}$ is the possibly *infinite* family of such matrices $\mathcal{F} = \{A\}$. We shall reserve the letter L for such operators—the infinite direct sum of finite-dimensional A's. Thus, L refers to both, single $N\times N$ matrices and *infinite families* of such $N\times N$ matrices.

To deal with such L's, we need the following lemma.

LEMMA 2.3. *Every matrix $A \in \mathcal{M}_{N\times N}(\mathcal{C})$ is similar to an* "almost" *diagonal matrix, i.e., $\forall\varepsilon > 0$ there exists a constant matrix $K = K_\varepsilon$ such that*

$$KAK^{-1} = \Lambda + O(\varepsilon). \tag{2.12}$$

Here, $\Lambda = diag(\lambda_1(A),\dots,\lambda_N(A))$, where $\{\lambda_i\}$ are the eigenvalues of A.

Proof. By Schur's theorem, any matrix A can be put into an upper triangular form by a unitary transformation

$$U^*AU = \Lambda + \mathcal{N}, \qquad \mathcal{N}_{i,j} = 0 \ \ \forall i < j.$$

Multiply the resulting matrix U^*AU by the matrices $E := \mathrm{diag}(\varepsilon,\dots,\varepsilon^N)$ and E^{-1}. The result follows with $K = E^{-1}U^*$, for

$$KAK^{-1} = \Lambda + K\mathcal{N}K^{-1}$$

with

$$(K\mathcal{N}K^{-1})_{i,j} = \left\{ \begin{array}{ll} O(\varepsilon)^{j-i} & j > i \\ 0 & j \le i \end{array} \right\} = O(\varepsilon). \qquad \square$$

Put differently, in view of (2.7) and (2.12), Lemma 2.3 tells us that $\forall\varepsilon > 0$, $\exists H_\varepsilon = K_\varepsilon^* K_\varepsilon$ such that

$$\|A\|_{H_\varepsilon} = \|K_\varepsilon A K_\varepsilon^{-1}\| = \|\Lambda + O(\varepsilon)\| \le \rho(A) + C\cdot\varepsilon. \tag{2.13}$$

We are now in position to prove our above claim regarding the deformed gaps in the hierarchy of inequalities (2.9).

THEOREM 2.1. *Consider a (possibly infinite) family of matrices,* $L = \sum_{A\in\mathcal{F}} \oplus A$, $\mathcal{F} \subset \mathcal{M}_{N\times N}(\mathcal{C})$. *Then the following equalities hold:*

$$\rho(L) = \inf_{H>0} r_H(L) = \inf_{H>0} \|L\|_H. \tag{2.14}$$

Proof. For each A in $\mathcal{F}$, there exists $H = H_{A,\varepsilon}$ such that (2.13) holds, and the result follows with $H_\varepsilon = \sum_{A\in\mathcal{F}} \oplus H_{A,\varepsilon}$ and then letting ε tend to zero. ∎

Notes. 1. Theorem 2.1 was based on Lemma 2.3, formulated concisely in (2.13). We note in passing that the constant C, and hence H_ε in (2.13), may depend on $\|A\|$, and in particular therefore, on the dimension of A. Consequently, the extension of Theorem 2.1 to infinite-dimensional matrices fails at this point.

2. In a similar manner to our treatment of the H-weighted norms, one can deform the weighted numerical range, $W_H(L)$, to be as closed as desired to the convex hull of the spectrum, $\sigma(L)$,

$$\text{conv } \sigma(L) = \inf_{H>0} W_H(L). \tag{2.15}$$

2.2. Stability, power-boundedness, and the resolvent.

2.2.1. Power boundedness. We start this section with the notion of power boundedness. To illustrate this concept, we consider the iterative algorithm

$$u^{n+1} = Tu^n, \tag{2.16}$$

subject to the initial condition, $u^0 = u_0$.

The algorithm is said to be stable if, for a sufficiently large, dense set of initial data, u^0, the corresponding iterates are bounded in terms of the initial data:

$$\|u^p\| \le C\|u^0\|. \tag{2.17}$$

Note that from now on, $\|\cdot\|$ will refer to an arbitrary norm induced by a corresponding inner product. We will emphasize a possible H-weighted norm only when it is necessary.

Since the solution of (2.16) is given by $u^p = T^p u^0$, the stability requirement (2.17) boils down to the power boundedness of T, which brings us to the following definition.

DEFINITION. T *is* power bounded (stable)*, if there exists a constant* $C > 0$*, such that*

$$\|T^p\| \le C \quad \forall p > 0. \tag{2.18}$$

Notes. 1. If (2.18) holds with $C = 1$, T is called *strongly stable*. We point out that for strong stability, it is sufficient to verify (2.18) with $p = 1$, for submultiplicativity implies (consult (2.5)):

$$\|T\| \le 1 \quad \Rightarrow \quad \|T^p\| \le 1 \qquad p = 1, 2, \ldots. \tag{2.19}$$

2. The stability definition is invariant for equivalent norms,[1] that is, the statement of power boundedness with respect to, say, H-weighted norm:

$$\|T^p\|_H \le C \qquad \text{for} \qquad 0 < C_H^{-1} I \le H \le C_H I \tag{2.20}$$

[1] Two norms $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent if $c \le \frac{\|x\|_1}{\|x\|_2} \le C, \quad \forall x \ne 0$.

is independent of all equivalent H's. In particular $H = I$ recovers (2.18), and $H = K^*K$ is equivalent with (2.18), iff K is well conditioned, i.e., iff

$$\|K\| \cdot \|K^{-1}\| \leq \text{Const.} \tag{2.21}$$

3. Consider a "well conditioned" $H = K^*K$, so that $\|K\| \cdot \|K^{-1}\| \leq C_H$. Then, as noted above, the strong H-stability of T, $\|KTK^{-1}\| \leq 1 \Leftrightarrow \|T\|_H \leq 1$, implies power boundedness. In Theorem 2.2 we quote the Kreiss matrix theorem which is concerned with the *inverse* implication.

4. A less trivial example for such an equivalence is given by the H-numerical radius $r_H(T)$. Indeed if $r_H(T) \leq 1$, then T is power bounded:

$$\|T^p\| \leq C_H \|T^p\|_H \leq 2C_H r_H(T^p) \leq 2C_H r_H^p(T) \leq 2C_H. \tag{2.22}$$

Here, the first inequality follows in view of the well conditioning of H,

$$0 < C_H^{-1} I \leq H \leq C_H I;$$

the second inequality in (2.22), follows from the straightforward equivalence

$$r_H(T) \leq \|T\|_H \leq 2r_H(T). \tag{2.23}$$

The third is the generalized Halmos inequality [18], and the last utilizes the assumption that $r_H(T) \leq 1$.

2.2.2. The resolvent condition. The power boundedness of T guarantees the stability of our algorithm—stability with respect to *initial perturbations*. That is, an initial perturbation, say of size $O(\delta)$, is amplified by no more than Const. δ in later iterations.

The issue of L^2-power boundedness, is intimately related to another notion of stability—stability with respect to *inhomogeneous perturbations*. This is expressed by the so-called *resolvent condition* which we now explore. Here we are led to investigate the stability of our algorithm in the presence of an inhomogeneous term, and to this end we consider the scheme

$$u^{n+1} = Tu^n + F^n. \tag{2.24}$$

Without loss of generality, we assume zero initial values, $u^0 = 0$ (for otherwise, we can subtract the nonvanishing initial data which instead can be added to the inhomogeneous term).

To analyze the stability of (2.24), we proceed as follows. We identify u^n as the nth-step of a piecewise constant solution,[2] i.e., $u(t) = \sum u^n \chi_{[t^n, t^{n+1})}$. Multiplying (2.24) by an exponential weight $e^{-\eta t}$, $\eta > 0$, we find

$$e^{\eta \Delta t} e^{-\eta(t+\Delta t)} u^{n+1} := T(e^{-\eta t} u^n) + e^{-\eta t} F^n \tag{2.25}$$

where $u^{n+1} = u(t^n + \Delta t)$, Δt being a pseudo time step.

Fourier transform (2.25) in time (recall, $u \equiv 0$ for $t < 0$),

$$e^{-\eta t} u(t) = \frac{1}{\sqrt{2\pi}} \int_{\xi=-\infty}^{\infty} \hat{u}(\xi) e^{\imath \xi t} d\xi, \qquad \hat{u}(\xi) := \frac{1}{\sqrt{2\pi}} \int_t u(t) e^{-\eta t} e^{-\imath \xi t} dt$$

[2] χ_I being the characteristic function of the interval I.

we get

$$e^{(\eta+\imath\xi)\Delta t}\hat{u}(\xi) = T\hat{u}(\xi) + \hat{F}(\xi).$$

Abbreviating $z = e^{(\eta+\imath\xi)\Delta t}$, we arrive at the so-called *resolvent equation*

$$\hat{u}(z) = (zI - T)^{-1}\hat{F}(z), \qquad \hat{u}(z) := \frac{1}{\sqrt{2\pi}}\int u(t)z^{-t/\Delta t}dt.$$

By Parseval we have

$$\|e^{-\eta t}u\| = \|\hat{u}(z)\| \le \|(zI - T)^{-1}\| \cdot \|\hat{F}(z)\| = \|(zI - T)^{-1}\| \cdot \|e^{-\eta t}F\|. \tag{2.26}$$

Thus, the question of stability with respect to the inhomogeneous term F, boils down to the boundedness of the resolvent, $\|(zI - T)^{-1}\|$. Clearly, if T is power bounded, then by considering its geometric expansion, the resolvent does not exceed

$$\|(zI - T)^{-1}\| \le \sum_k |z|^{-(k+1)}\|T\|^k \le \frac{C}{|z| - 1}, \tag{2.27}$$

and this brings us to the following definition.

DEFINITION. *An operator T is said to satisfy the* resolvent condition *if there exists a constant $C_R > 0$ such that for all complex numbers z with $|z| > 1$, $zI - T$ is nonsingular and the resolvent estimate*

$$\|(zI - T)^{-1}\| \le \frac{C_R}{|z| - 1} \tag{2.28}$$

holds.

If the resolvent condition (2.28) holds, we may utilize (2.26) coupled with Parseval, to translate (2.26) back into the "physical space." We let L^2_η denote the collection of grid functions $\{w(t^n)\}_{n=0}^\infty$ with $\|w\|_{L^2_\eta} := \sum_{n=0}^\infty e^{-n\eta\Delta t}\|w(t^n)\|^2\Delta t < \infty$, and we note that $|z| - 1 \sim \eta\Delta t$. The resulting stability of our inhomogeneous algorithm (2.24) then states that for all $F \in L^2_\eta$,

$$\|u\|_{L^2_\eta} \le \frac{C}{\eta\Delta t}\|F\|_{L^2_\eta}, \ \ \forall F \in L^2_\eta, \ \ \eta > 0. \tag{2.29}$$

Thus the notion of L^2-stability in (2.17) implies the resolvent stability in (2.29). The converse is a more intricate question addressed below.

2.2.3. Back to power boundedness: A single matrix vs. a family of matrices. For a *single* matrix A, power boundedness requires that $\forall p > 0$

$$\|A^p\| \le C.$$

Note that C may still depend on the dimension N. A full characterization for the power boundedness of single matrices is provided in the following lemma.

LEMMA 2.4 (the root condition). *A necessary and sufficient condition for the power boundedness of a given matrix A, is the so-called root condition, requiring that*

1. *$\rho(A) \le 1$;*
2. *if $|\lambda(A)| = 1$, then λ are simple (their geometric multiplicity equals to their algebraic multiplicity).*

The proof is immediate, since A is similar to its power-bounded Jordan form, $J = \sum \oplus J_i$; here J_i's are either power-bounded scalars (since $|J_i| \leq 1$), or, they are power-bounded blocks (since $\rho(J_i) < 1$, implies, in view of (2.14), that $\exists H_i > 0$ such that $\|J_i\|_{H_i} \leq 1 + \rho(J_i)/2 < 1$ and hence $\|J_i^p\| \leq \text{Const} \cdot \|J_i^p\|_{H_i} \leq \text{Const}$).

Note that this lemma is restricted to a single finite-dimensional matrix. The above argument may otherwise fail, since the Jordan similarity transformation need not be uniformly well conditioned for an infinite matrix. In fact, the root condition does not even guarantee power boundedness in the case of a *family* $\mathcal{F}$, $\mathcal{F} \subset \mathcal{M}_{N\times N}$, of finite-dimensional $N \times N$ matrices, where $T = L$ is of the form $L = \sum_{A\in\mathcal{F}} \oplus A$. Indeed, the crucial question here, is whether there exists a finite C such that $\|A^p\| \leq C$ is satisfied *simultaneously* for all $A \in \mathcal{F}$. For example (taken from [8]), by the root condition, each of the family members in $\mathcal{F} = \{A(t)\}$,

$$A(t) = \begin{pmatrix} 1-t & t^{1/2} \\ 0 & 1 \end{pmatrix}, \quad 0 \leq t \leq 1,$$

is power bounded, yet there is no *simultaneous* upper bound on all $A^p(t)$, $\forall(p,t)$—since the sequence $\|A^p(t)\|_{|t=1/p} \sim \sqrt{p}$ diverges.

Thus, the root condition needs to be strengthened in order to characterize the power-boundedness stability in the general case. Such an if and only if characterization could be achieved, at least in the case of infinite families of finite-dimensional matrices, $L = \sum \oplus A$, with the help of the Kreiss matrix theorem (consult [58, section 4.9]).

THEOREM 2.2 (Kreiss matrix theorem–KMT). *Consider a family $\mathcal{F}$ of $N \times N$ matrices, associated with the infinite-dimensional operator $L = \sum_{A\in\mathcal{F}} \oplus A$. The L^2-stability of L, $\|L^p\| \leq C$, is equivalent to each of the following conditions:*

$\mathcal{R}$ **resolvent condition:** *There exists a constant $C_{\mathcal{R}}$, such that for all complex numbers z with $|z| > 1$, $(zI - L)^{-1}$ exists and*

$$\|(zI - L)^{-1}\| \leq \frac{C_{\mathcal{R}}}{|z| - 1}.$$

$\mathcal{H}$ **condition:** *There exists a positive Hermitian H and a constant, $C_{\mathcal{H}} > 0$, such that*

1. $C_{\mathcal{H}}^{-1} I \leq H \leq C_{\mathcal{H}} I$,
2. $L^* H L \leq H$.

Notes. 1. Note that the stability constants in Theorem 2.2 may depend on the dimension N, and therefore, the case of a family of matrices with a *growing* dimension, is not covered by KMT.

2. For a general infinite-dimensional operator T, one finds the following strict hierarchy:

$$\mathcal{H} \text{ condition} \Rightarrow L^2 - \text{power-boundedness} \Rightarrow \mathcal{R} \text{ resolvent condition}.$$

Indeed, the $\mathcal{H}$-condition, $T^*HT \leq H =: K^*K$, implies that T is strongly stable for it is similar to the contraction KTK^{-1}. A counterexample for the converse implication was constructed in [11] and [29]. L^2-stable operators satisfy the resolvent condition (2.27); a counterexample for the converse implication based on a family of matrices with growing dimension, was constructed in [51] (also consult [10]). To summarize, the resolvent condition is strictly weaker than L^2-power boundedness, which in turn, is strictly weaker that the $\mathcal{H}$-condition.

3. There is an analogue of KMT which deals with well posedness—the stability of the so-called semidiscrete ODEs

$$\dot{u}(t) = Lu(t), \qquad u(0) = u_0. \tag{2.30}$$

Since the solution of (2.30) is $u(t) = e^{Lt}u_0$, stability (with respect to the initial data ...) requires

$$\|e^{Lt}\| \le C. \tag{2.31}$$

The analogue of KMT states for families of finite dimension, $L = \sum_{A\in\mathcal{F}} \oplus A$, the equivalence between (2.31), and each of the following:

The resolvent condition: there exists a constant $C_{\mathcal{R}} > 0$ such that

$$\|(sI - L)^{-1}\| \le \frac{C_{\mathcal{R}}}{\Re(s)}, \qquad \forall \Re(s) > 0; \tag{2.32}$$

The $\mathcal{H}$-stability condition: there exists a positive Hermitian H and a constant $C_{\mathcal{H}}$, such that

$$L^*H + HL \le 0, \qquad 0 < C_{\mathcal{H}}^{-1} \le H \le C_{\mathcal{H}}. \tag{2.33}$$

2.3. Runge–Kutta methods. In this section we include a brief discussion of Runge–Kutta discretizations. We will concentrate only on the main concepts. For further reading on this subject see e.g. [4]. We start with a prototype case of a fourth-order method.

- Fourth-order Runge–Kutta method.
  Consider the nonlinear, autonomous problem

  $$u_t = F(u), \tag{2.34}$$

  $u = u(t)$ being a vector function subject to given initial data,

  $$u(0) = u^0.$$

  We abbreviate $u^n = u(t = n\Delta t)$.
  The classical fourth-order Runge–Kutta (RK4) approximation of (2.34) is based on the iterative scheme:

  $$\begin{aligned}
  &u^{n,1} = F(u^n),\\
  &u^{n,2} = F(u^n + 1/2\Delta t u^{n,1}),\\
  &u^{n,3} = F(u^n + 1/2\Delta t u^{n,2}),\\
  &u^{n,4} = F(u^n + \Delta t u^{n,3}),\\
  &u^{n+1} = u^n + 1/6\Delta t \left[u^{n,1} + 2u^{n,2} + 2u^{n,3} + u^{n,4}\right].
  \end{aligned}$$

- The linear, constant-coefficients case.
  In the linear, constant coefficients case where $F(u) = Lu$, the RK4 scheme is a prototype for the iterative schemes mentioned above. In the case of a finite-dimensional L, which is independent of t, the solution of (2.34) is given by $e^{Lt}u^0$, and its well-posedness amounts to $\|e^{Lt}\| \le C$. (In a more general infinite-dimensional setup with possibly unbounded L's, this should be interpreted in the sense of semigroups.) In this scenario, all the RK4 variants, approximating the semidiscrete problem, $\dot{u} = Lu$, coincide into

  $$u^{n+1} = \sum_{i=0}^{4} \frac{(\Delta t L)^i}{i!} u^n = P(\Delta t L)u^n.$$

  Here, $P(z) = \sum_{i=0}^{4} \frac{z^i}{i!}$ is the so-called *characteristic polynomial*. Therefore, under the above constraints, the s-order RK methods are nothing but the

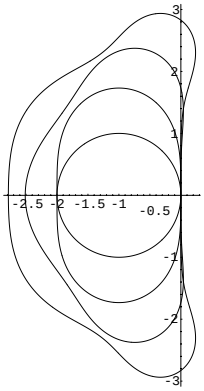

FIG. 1. *Regions of absolute stability, R_s, for RK methods. Increasing size corresponds to increasing $s = 1, 2, 3, 4$.*

truncated Taylor expansions of the exponential in the variable ΔtL. In the general linear case, where L is allowed to depend on t, additional low-order terms are introduced. As we shall see later, however, these low-order terms have no influence on the question of stability.

2.3.1. Scalar stability analysis. In general, an s-order accurate RK method is identified with an operator T which is based on the corresponding characteristic polynomial $P_s(z)$,

$$T = P_s(\Delta tL) := \sum_{i=0}^{s} \frac{(\Delta tL)^i}{i!}.$$

With $s = 1$, the resulting method is the celebrated *forward Euler* (FE) scheme. With $s = 3$ and $s = 4$ we obtain, respectively, the third-order and fourth-order RK methods, abbreviated RK3 and RK4.

Our main interest is the stability of such RK approximations of well-posed problems. Thus, our algorithm is identified with the operator $T = P_s(\Delta tL)$ and we are led to the question of its power boundedness:

$$\|P_s^n(\Delta tL)\| \le C \qquad \forall n.$$

To address this issue, we recall the notion of the *region of absolute stability*.

DEFINITION. *The* region of absolute stability *of a RK solver is the set R_s in the complex plane, $R_s = \{z : |P_s(z)| \le 1\}$, where $P_s(z)$ characterizes the s-stage RK solver.*

In Figure 1, we draw the regions of absolute stability for RK solvers of order $s = 1, 2, 3, 4$.

The usual stability analysis deals with the prototype linear *scalar* problems, associated with the scalar $L = \lambda$, i.e., $\dot{u} = \lambda u$. Clearly, the corresponding scalar RK scheme is stable iff $|P(\Delta t\lambda)| \le 1$ (so that the scalar $P(\Delta t\lambda)$ is power bounded), and this in turn holds iff Δt is sufficiently small so that $\Delta t\lambda \in R_s$ for an appropriate set of complex λ's.

2.3.2. The stability of RK methods for systems—The heart of the matter. What does the region of absolute stability tell us about the L^2-stability of RK

methods for *systems* of linear ODEs? Let $P(z) = P_s(z)$ denote the characteristic polynomial which identifies the specific RK method. A sufficient condition for the boundedness of $P^n(\Delta tL)$ is the strong stability requirement, consult, e.g., (2.19),

$$\|P(\Delta tL)\| \leq 1. \tag{2.35}$$

If L has a complete eigensystem so that it can be decomposed into a direct sum of scalar problems, $KLK^{-1} = \Lambda =: \sum_{\lambda\in\sigma(L)} \oplus\lambda$, then by the spectral mapping theorem,

$$KP(\Delta tL)K^{-1} = \sum_{\lambda\in\sigma(L)} \oplus P(\Delta t\lambda).$$

Consequently, the scalar stability analysis applies to each $P(\Delta t\lambda)$, and hence $P(\Delta tL)$ is strongly stable if

$$\Delta t\lambda \in R_s, \quad \forall\lambda \in \sigma(L).$$

Thus, in the *normal* setup,[3] equipped with a complete eigensystem, say $\Lambda = \sum \oplus\lambda$ and K, power boundedness follows in view of strong H-stability, for

$$\|P(\Delta tL)\|_H = \|KP(\Delta tL)K^{-1}\| = \|P(\Delta t\Lambda)\| = \sup_{\lambda\in\sigma(L)} |P(\Delta t\lambda)| \leq 1, \quad H = K^*K.$$

We have seen that the scalar analysis will suffice in the normal case, where the operator L admits a *well-conditioned* decomposition, $KLK^{-1} = \Lambda$ with $\|K\|\cdot\|K^{-1}\| \leq$ Const.; for then, L is strongly stable with respect to a well-conditioned $H = K^*K$, $0 < C^{-1} < H < C$.

Can we extend the above argument to *non-normal* systems? Here one might try to proceed in the following manner which highlights the intricate point of failure.

Consider the well-posed problem, $\dot{u} = Lu$. To guarantee the stability of its RK approximation we make the following assumption.

Assumption. The time-step, Δt, is sufficiently small so that for all λ's in the spectrum of L, $\Delta t \cdot \lambda(L)$ lies strictly inside the corresponding region of absolute stability R_s

$$\Delta t \cdot \sigma(L) \subset\subset R_s. \tag{2.36}$$

A couple of remarks are in order.

1. Borrowing the terminology associated with approximations to time-dependent PDEs, the requirement (2.36) is the so-called Courant–Friedrichs–Levy (CFL) condition, a condition which places an upper bound on the permissible time-step, Δt.
2. Can we verify the CFL condition (2.36), at least for sufficiently small Δt's? On the one hand, we note that the well-posedness of L implies it is negative definite (with respect to an H-weighted inner product stated by KMT in (2.33)), and in particular, that the spectrum of L lies in the left half of the plane, $\Re(\lambda) \leq 0$, $\forall\lambda \in \sigma(L)$. On the other hand, inspection of the regions of absolute stability in Figure 1, shows that these regions, R_s, contain bounded sectors centered at the left-half plane. Thus, *if* $\sigma(L)$ is contained in an appropriate open wedge on the left-half plane, then a (sufficiently small) Δt multiple of it will indeed satisfy the CFL condition.

[3]Recall that according to our terminology of normality, T is normal iff it commutes with respect to *one* of its weighted adjoints, consult Assertion 2.1.

We conclude with the scalar analysis which guarantees that $P_s(\Delta t\lambda(L))$ lie strictly inside the unit disc

$$\rho(P_s(\Delta tL)) = \sup_{\lambda\in\sigma(L)} |P_s(\Delta t\lambda)| < 1. \tag{2.37}$$

At this point we may attempt to utilize appropriate H-weighted norms which get as close as desired to $\rho(P_s(\Delta tL))$—such H's exist by (2.14) so that $\|P_s(\Delta tL)\|_H \le 1$ and hence $P_s(\Delta tL)$ is H-power bounded,

$$\|P_s^n(\Delta tL)\| \le \|K\| \cdot \|K^{-1}\|. \tag{2.38}$$

The remaining open issue in this argument is whether such an $H := K^*K$ is well conditioned, $\|K\| \cdot \|K^{-1}\| \le \text{Const}$. The question of H being well conditioned is the precise intricate point which is tied to the non-normality of L. That is, either L is normal, yet it is equipped with an ill-conditioned eigensystem, or it is nondiagonalizable; in either case this may lead to a growing upper bound on the right-hand side (RHS) of (2.38).

To provide an affirmative answer to the question of power boundedness in the general, possibly non-normal case, we will therefore need more than just the scalar *spectral* analysis which propelled the above argument in (2.37)–(2.38). We are led to consider the geometric structure of the whole eigensystem of L—beyond just its spectrum, $\sigma(L)$.

To this end we will use a straightforward approach, the so-called energy method, aiming at a direct upper bound on the powers of $P_s(\Delta tL)$.

2.4. Additional notes.

1. *General remarks.* The quantities of spectral radius, numerical radius, weighted norms, etc., dealt in section 2, play a canonical role in various fields of linear algebra and numerical analysis, and in particular, in the stability analysis of iterative methods. We refer to Halmos [28], Dahlquist and Björck [7], Golub and Van Loan [19], and the references therein, as classical general references, to Richtmyer and Morton [58], Strang and Fix [65], Johnson [36], Gustafsson, Kreiss, and Oliger [25], and the references therein, as classical references for difference methods, and for Tadmor [72], Dorsselaer, Kraaijevanger, and Spijker [10], Spijker [63], Turkel [74], Lenferink and Spijker [47], and the references therein, for more recent reviews which focus on the question of stability of semi- and fully discrete schemes.

A good example for the role these algebraic quantities play in the numerical analysis of fully discrete difference methods is given by the Halmos inequality (e.g., [28]). Motivated by the question of stability for 2-dimensional finite-difference approximations of hyperbolic systems, Lax and Wendroff, in their pioneering paper [45], utilized the numerical radius, $r(A) \le 1$, as a sufficient condition for L^2-power boundedness stability of A, $\|A^p\| \le \text{Const}$. Their result was proved by induction on the dimension of the matrix A. This was later covered by Halmos conjecture, stating that for general L's, $r(L^p) \le r^p(L)$, (which, in turn, implies power boundedness with Const. $= 2$, in view of (2.22)–(2.23)). Halmos inequality was first proved by Berger [2] and was later simplified by Pearcy [55]. For more details we refer to Goldberg and Tadmor [18] and Wade [77].

2. *Weighted quantities.* Theorem 2.1, which states the equality between the weighted quantities, $\rho(L) = \inf_{H>0} r_H(L) = \inf_{H>0} \|L\|_H$, seems to be part of the folklore in the stability analysis of iterative methods. Though we could not find a written version of such a theorem, it is implicitly used in the stability studies of

Kreiss [58]. It is a direct consequence of the almost diagonalizability Lemma 2.3, which can be found in [19]. The similar statement (2.15) regarding the *numerical range*, $\operatorname{conv}\sigma(L) = \inf_{H>0} W_H(L)$, is due to the convexity of the numerical range $W_H(L)$; the latter is the celebrated Toeplitz theorem (e.g., see [28]).

3. *The resolvent condition.* Consider the possibly infinite-dimensional system of ODEs $\dot{u} = Tu$. The Hille–Yoshida theory [76] develops the corresponding notion of semigroup solution operator, $u(t) = e^{Tt}u(0)$, and states the equivalence between its boundedness, $\|e^{Tt}\| \le \text{Const.}$ and the infinite series of resolvent estimates

$$\|(sI - T)^{-k}\| \le \frac{C}{(\Re(s))^k}, \quad \forall \Re(s) > 0, \quad k = 1, 2, \dots. \tag{2.39}$$

Note, that here we seek an *infinite number* of resolvent estimates. The semidiscrete version of the KMT states (consult note number 3 on p. 48) that in the case $T = L$ consists of an infinite family of finite-dimensional matrices, $L = \sum_{A\in\mathcal{F}} \oplus A$, then only one of these inequalities (with $k = 1$) will suffice, i.e.,

$$\|(sI - L)^{-1}\| \le \frac{C}{\Re(s)}, \quad \forall \Re(s) > 0, \quad \Rightarrow \quad \|e^{Lt}\| \le \text{Const.}$$

In this work, the analogous KMT in the fully discrete case, plays a central role.

The discrete question is concerned with the iterative problem, $u^{n+1} = Tu^n$. Clearly, if T is a contraction or if it is similar to one, then T is power bounded. The converse, however, is not true in the general case (consult Foguel [11], Halmos [29], and Sz.-Nagy and Foias [67]). Also, if T is power bounded, then it also satisfies the infinite set of resolvent estimates (along the lines of (2.27))

$$\|(zI - T)^{-k}\| \le \frac{K}{(|z| - 1)^k}, \quad \forall |z| > 1, \quad k = 1, 2, \dots. \tag{2.40}$$

Thus, (2.40) is the fully discrete analogue of the semidiscrete resolvent estimates (2.39). Surprisingly, however, in contrast to the semidiscrete case, the converse implication, (2.40) $\Rightarrow$ power boundedness, does not hold, as shown by McCarthy [50]. We can summarize the above discussion by stating the following *strict* hierarchy of implications:

$$\mathcal{H} \text{ condition } \Rightarrow \ L^2 \text{ power boundedness } \Rightarrow \text{ resolvent condition.}$$

It is not difficult to see that the three conditions: L^2-power boundedness, the resolvent condition, and the similarity to contraction (which is the same as the H-condition), are equivalent for a *single, finite-dimensional* matrix, in view of the root condition (Lemma 2.4). Sz.-Nagy [66] extended the above equivalence results to the framework of compact operators. It is the achievement of the KMT which shows that all three conditions are equivalent for an infinite sum of finite-dimensional matrices, i.e., for operators $T = L$ of the form $L = \sum_{A\in\mathcal{F}} \oplus A$.

This equivalence between L^2-power boundedness, $\|L^p\| \le \text{Const.}$, and the resolvent condition, $\|(zI - L)^{-1}\| \le K(|z| - 1)^{-1}$, is one of the four ingredients introduced in the discrete version of KMT, which where subsequently treated by several authors. A partial list includes Miller [52], Miller and Strang [53], Morton and Schechter [54], Friedland [14] (which dealt with the discrete analogue of Görlich and Pontzen [20]). The above mentioned references include simplified versions of the KMT, whose original proof involves rather intricate arguments.

Gorelnick [21] extended KMT for families of matrices with growing dimensions, as long as the degree of their minimal polynomials remains bounded. Tadmor [68], following Laptev [44] in the semidiscrete case, provided a greatly simplified proof, stating that for a matrix A of order N and degree of minimal polynomial being s, the resolvent estimate $\|(zI - A)^{-1}\| \leq C_r(|z| - 1)^{-1}$ implies that $\|A^p\| \leq 32C_r es/\pi$. In particular, an $N \times N$ matrix A satisfying the resolvent condition is power bounded by a constant which grows at most linearly with N, $\|A^p\| \leq C_A \cdot N$. The constant, C_A, was then reduced in a series of papers to its optimal value $C_A = eC_r$ (LeVeque and Trefethen [48], Smith [62], Spijker [64]). For a historical point of view see [78]. A counterexample of McCarthy and Schwartz [51], which was later improved by Kraaijevanger [39], demonstrates a family of $N \times N$ matrices satisfying the resolvent condition, with power growth of at least $\|A^p\| > 2C_r N/\pi$.

4. *Resolvent stability.* The resolvent condition can be used as the cornerstone for a stability theory of iterative methods. As noted above, it is a weaker stability condition than power boundedness for general operators. However, the resolvent condition is invariant under low-order perturbations (consult Kreiss [26]), and thus, it allows extensions to problems with variable coefficients, etc. (e.g., see [40]).

Such an attitude was taken, for example, in the development of a stability theory for initial boundary value problems by Gustafsson, Kreiss, and Sandström [26], and in the analysis of spectral approximations by Gottlieb, Lustman, and Tadmor [22].

Kreiss and Wu [43] used the resolvent stability as their basis for the stability analysis of RK methods. Resolvent stability motivated Trefethen's notion of the so-called *pseudo spectrum* to identify stability regions, e.g., [73]. A similar approach was taken by Dorsselaer, Kraaijevanger, and Spijker [10], in their stability analysis. In fact, Lenfernik and Spijker [46], [47] showed that if P_s denotes the characteristic polynomial of an sth-order RK method, then

$$\|(zI - L)^{-1}\| \leq \frac{C}{\operatorname{dist}(z, R_s)} \Rightarrow P_s(\Delta t L) \text{ power bounded.}$$

The last implication with $P(z) = z^n$ corresponds to the discrete version of the KMT. In the aforementioned cases, however, power boundedness holds with a possible (linear) growth, depending linearly on the increasing dimension.

5. *RK schemes.* The subject of RK discretizations was widely investigated. For classical references on this subject, we refer the reader to Butcher [4] and Hairer, Norsett, and Wanner [27]. Modern examples for applications using the RK methods can be found in, e.g., Turkel [74].

We note that the classical stability analysis of the RK methods deals with regions of absolute stability of the scalar prototype, $\dot{u} = \lambda u$. Unfortunately, as mentioned above, this stability analysis does not cover the framework of families of matrices with an increasing size (e.g., the semidiscrete approximations to PDEs we later encounter in this work; consult section 4), and this, in turn, motivates our analysis in section 3.

3. From semidiscrete to fully discrete: Strong stability for coercive operators.

3.1. An overview on negative-definite problems. We turn to investigate the stability of the RK approximations of well-posed problems, $u_t = Lu$.

According to the Kreiss matrix theorem, well-posedness (or semidiscrete stability), $\|e^{Lt}\| \leq C$, is equivalent—at least for infinite families $\mathcal{F}$ of $N \times N$ matrices, $L = \sum_{A \in \mathcal{F}} \oplus A$, with the negativity of L under an appropriate H-weighted norm

(see (2.31)):

$$\Re(Lx,x)_H \le 0 \qquad 0 < C_{\mathcal{H}}^{-1} \le H \le C_{\mathcal{H}} \qquad \forall x.$$

Remark. For the sake of simplicity, we refer to *seminegative-definite* operators simply as *negative-definite* operators. We will also omit the reference to the specific H-weight, unless otherwise specified.

Thus, we will consider RK approximations of negative-definite problems

$$u_t = Lu, \qquad \Re(Lx,x) \le 0, \quad \forall x \in \mathcal{C}^N. \tag{3.1}$$

We should emphasize that the negativity assumption is motivated by the KMT characterization of well posed for operators of the form $L = \sum \oplus A$, i.e., infinite families of finite-dimensional matrices A. Nevertheless, our approach applies equally well to *arbitrary* negative-definite operators—both finite and infinite-dimensional ones. The energy method outlined below is not restricted to families of the above form. Therefore, from this point on, unless otherwise stated, L *will serve as such an arbitrary negative-definite operator.*

Multiplying both sides of (3.1) by u, using the inner product $(\cdot,\cdot)$ yields

$$\frac{d}{dt}\|u\|^2 = \Re(u,Lu).$$

Hence, according to our negativity assumption, we have, $\frac{d}{dt}\|u\|^2 \le 0$, and thus the norm of the solution does not grow in time,

$$\|u(t)\|^2 \le \|u(0)\|^2, \qquad \forall t \ge 0.$$

Following these lines, it is natural to ask whether such a strong stability result for the semidiscrete problem (3.1), can be carried over into its fully discrete RK discretization. This question will occupy the rest of our work.

Below, we summarize the main ingredients of the following subsections. Let $P_s(z)$ be the characteristic polynomial of the specific RK method in question. We recall that the intricate issue concerning the power boundedness of $P_s(\Delta tL)$, which prevented the extension of the scalar arguments, lies with the possible non-normality of L, or equivalently, the ill conditioning of L and its eigensystem.

This is amplified by our results below, which show that even for well-posed problems associated with negative-definite L's, the CFL condition (2.36), based on the scalar stability analysis, $\Delta t\sigma(L) \subset R_s$, may not suffice for the power boundedness of $P_s(\Delta tL)$ for *general* negative L's.

To find out what should be added to negativity, we begin in seciton 3.2 with the simplest of all RK methods—the first-order forward Euler (FE) scheme, $u^{n+1} = (I + \Delta tL)u^n$. The stability of the first-order FE method is not guaranteed for arbitrary negative-definite L's, and this leads us to focus our attention on a subclass of admissible L's. We identify the precise subclass of admissible L's for which the first-order FE method is stable; negativity is now strengthened by requiring the following condition.

Coercivity condition. There exists a positive constant $\eta > 0$ *such that*

$$\Re(Lu,u) \le -\eta\|Lu\|^2. \tag{3.2}$$

Equipped with this terminology, the main result of this section reads as such.

MAIN THEOREM. *Consider the well-posed problem,* (3.1), $u_t = Lu$, *associated with the coercive* L, (3.2). *Then its* s*-order RK discretization is strongly stable under a CFL condition*

$$\Delta t \leq C_s \cdot \eta. \tag{3.3}$$

We begin in section 3.2 showing that the forward Euler (FE) scheme is strongly stable under the CFL condition $\Delta t \leq 2\eta$. In sections 3.3 and 3.4 we extend this result to the third- and fourth-order RK methods: tedious though straightforward energy estimates prove that coercivity suffices to guarantee strong stability of the third- and fourth-order methods, under appropriate CFL conditions (3.3), with $C_3 = 3/25$ and $C_4 = 1/31$, respectively.

We close this introduction by noting that the coercivity condition (3.2), is not a necessary condition for stability. Coercivity is tied, however, to *strong* stability. Also, our derived CFL conditions for the strong stability of coercive problems are not optimal: sharper results for the third and fourth-order methods, as well as extensions to higher-order RK methods can be derived.

3.2. The first-order forward Euler approximation of coercive problems. Consider the well-posed negative-definite problem (3.1), $u_t = Lu$. We discretize the time variable using the first-order RK method, which coincides with the classical forward Euler (FE) scheme,

$$u^{n+1} = (I + \Delta t L)u^n.$$

Squaring both sides we find

$$\begin{aligned} \|u^{n+1}\|^2 - \|u^n\|^2 &= ((I + \Delta t L)u^n, (I + \Delta t L)u^n) - (u^n, u^n) \\ &= \|\Delta t L u^n\|^2 + 2\Re(u^n, \Delta t L u^n). \end{aligned} \tag{3.4}$$

Thus $P_1(\Delta t L) = I + \Delta t L$ is power bounded, and hence strongly stable, if the RHS is negative, which in view of the negativity of L amounts to the requirement

$$\Delta t \leq \frac{2 \cdot |\Re(u^n, Lu^n)|}{\|Lu^n\|^2}.$$

We recall again that such a restriction on the maximal time step, Δt, is known as a CFL condition, in the context of discretizations of time-dependent PDEs. To verify such a CFL condition, at least for sufficiently small Δt's, requires L to be coercive, so that the upper bound on the RHS remains uniformly bounded from below.

Without loss of generality, we rescale u^n so that $\|u^n\| = 1$, and thus conclude the following.

THEOREM 3.1. *Consider the well-posed problem* (3.1), $u_t = Lu$, *associated with the coercive* L, (3.2). *Then its approximation by the FE method is strongly stable* if *the following CFL condition is satisfied:*

$$\Delta t \leq 2\eta, \qquad \eta := \inf_{\|u\|=1} \frac{|\Re(u, Lu)|}{\|Lu\|^2}. \tag{3.5}$$

This theorem agrees with the rather well-known fact, that just negativity need not suffice for the stability of the FE scheme. Already in the scalar case, the FE

discretization of $\dot{u} = \lambda u$, with $\Re(\lambda) = 0$ is not strongly stable. Figure 1 shows that all such nonzero λ's, lie outside R_1. Indeed, the prototype of the divergent schemes for first-order transport hyperbolic PDEs is a FE scheme based on FE time discretization together with central spatial differencing.

Once we identify coercive L's as the admissible class which render the stability of the FE scheme, it is natural to ask whether similar results hold for the high-order RK methods. We thus proceed with the stability analysis of the RK3 method, followed by a similar stability analysis for the RK4 method.

3.3. The third-order RK approximation of coercive problems. The stability analysis of the high-order RK methods, involves a considerable amount of straightforward technical details. To simplify matters, we prepare the following lemma concerning the numerical range of powers of negative-definite operators.

LEMMA 3.1. *Assume that L is negative definite. Then the following inequalities hold:*

$$\Re(L^2x, x) \le -\|Lx\|^2 - \Re(L^2x, Lx) - \Re(Lx, x) \tag{3.6}$$

$$\Re(L^3x, Lx) \le -\|L^2x\|^2 - \Re(L^3x, L^2x) - \Re(L^2x, Lx) \tag{3.7}$$

$$\Re(L^4x, L^2x) \le -\|L^3x\|^2 - \Re(L^4x, L^3x) - \Re(L^3x, L^2x) \tag{3.8}$$

$$\Re(L^4x, x) \le \|L^2x\|^2 - \Re(L^4x, L^3x) - -\Re(L^3x, L^2x) - \Re(L^2x, Lx) - \Re(Lx, x) \tag{3.9}$$

$$\Re(L^3x, x) \le -\Re(L^3x, L^2x) - \Re(Lx, L^2x) - \Re(Lx, x) \tag{3.10}$$

$$\Re(L^4x, Lx) \le -\Re(L^2x, Lx) - \Re(L^2x, L^3x) - \Re(L^4x, L^3x). \tag{3.11}$$

Proof. For the first inequality (for example ...), we have

$$\Re\left(L(Lx + x), Lx + x\right) = \Re(L^2x, Lx) + \Re(Lx, Lx) + \Re(L^2x, x) + \Re(Lx, x) \le 0.$$

The other inequalities hold due to analogous reasons. ∎

Note that the previous lemma holds for any inner product.

Equipped with Lemma 3.1, we are now ready to investigate the stability properties of the high-order RK methods. We start with a well-posed semidiscrete problem, $u_t = Lu$, with negative-definite L which is independent of t. The RK3 discretization reads

$$u^{n+1} = u^n + \Delta t L u^n + \frac{1}{2}(\Delta t L)^2 u^n + \frac{1}{3!}(\Delta t L)^3 u^n.$$

Abbreviating $\widetilde{L} = \Delta t L$, we have

$$u^{n+1} = u^n + \widetilde{L}u^n + \frac{1}{2}\widetilde{L}^2 u^n + \frac{1}{6}\widetilde{L}^3 u^n,$$

and squaring both sides yields,

$$\begin{aligned}
\|u^{n+1}\|^2 - \|u^n\|^2 &= \left(\sum_{i=0}^{3} \frac{\widetilde{L}^i}{i!}u^n, \sum_{i=0}^{3} \frac{\widetilde{L}^i}{i!}u^n\right) - (u^n, u^n) \\
&= \|\widetilde{L}u^n\|^2 + \frac{1}{4}\|\widetilde{L}^2 u^n\|^2 + \frac{1}{36}\|\widetilde{L}^3 u^n\|^2 + 2\Re(u^n, \widetilde{L}u^n) \\
&+ \Re(u^n, \widetilde{L}^2 u^n) + \frac{1}{3}\Re(u^n, \widetilde{L}^3 u^n) + \Re(\widetilde{L}u^n, \widetilde{L}^2 u^n) \\
&+ \frac{1}{3}\Re(\widetilde{L}u^n, \widetilde{L}^3 u^n) + \frac{1}{6}\Re(\widetilde{L}^2 u^n, \widetilde{L}^3 u^n).
\end{aligned}$$

The last quantity can be upper bounded with the help of inequalities (3.6), (3.7), and (3.10),

$$\begin{aligned}\|u^{n+1}\|^2 - \|u^n\|^2 &\le -\frac{1}{12}\|\widetilde{L}^2 u^n\|^2 + \frac{1}{36}\|\widetilde{L}^3 u^n\|^2 \\ &+\frac{2}{3}\Re(u^n, \widetilde{L}u^n) - \frac{2}{3}\Re(\widetilde{L}u^n, \widetilde{L}^2 u^n) - \frac{1}{2}\Re(\widetilde{L}^2 u^n, \widetilde{L}^3 u^n) \\ &=: \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3 + \mathcal{I}_4 + \mathcal{I}_5.\end{aligned} \tag{3.12}$$

Thus, in order to conclude with the desired strong stability

$$\|u^{n+1}\| \le \|u^n\|,$$

it is left to inquire the negativity of the five terms $\mathcal{I}_j$, $1 \le j \le 5$.

- For the sum of the first, the second, and the fifth term on the RHS of (3.12) to be negative, $\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_5$, we use the estimate

$$\begin{aligned}-\frac{1}{12}\|\widetilde{L}^2 u^n\|^2 \quad &+\frac{1}{36}\|\widetilde{L}^3 u^n\|^2 - \frac{1}{2}\Re(\widetilde{L}^2 u^n, \widetilde{L}^3 u^n) \\ &\le -\frac{1}{12}\|\widetilde{L}^2 u^n\|^2 + \frac{1}{36}\|\widetilde{L}^3 u^n\|^2 + \frac{1}{2}\left|(\widetilde{L}^2 u^n, \widetilde{L}^3 u^n)\right| \\ &\le -\frac{1}{12}\|\widetilde{L}^2 u^n\|^2 + \frac{1}{36}\|\widetilde{L}\|^2\|\widetilde{L}^2 u^n\|^2 + \frac{1}{2}r(\widetilde{L})\|\widetilde{L}^2 u^n\|^2 \\ &\le \|\widetilde{L}^2 u^n\|^2\left[-\frac{1}{12} + \frac{\|\widetilde{L}\|^2}{36} + \frac{\|\widetilde{L}\|}{2}\right].\end{aligned} \tag{3.13}$$

  Thus, with $\widetilde{L} = \Delta t L$, the sum of these three terms can be made negative under an appropriate CFL condition.
- The third term on the RHS of equation (3.12), $\mathcal{I}_3$, is negative due to the negativity of the operator $\widetilde{L}$.
- Finally, we address the fourth term in the RHS of equation (3.12), $\mathcal{I}_4 = -\frac{2}{3}\Re(\widetilde{L}u^n, \widetilde{L}^2 u^n)$. It is here that we face the critical point concerning the stability of the third-order RK method. Since $\widetilde{L}$ is negative, this term is positive, and a priori, there is no reason to assume that the negativity of the previous terms we met can compensate for the positivity of $\mathcal{I}_4$.

We should emphasize that the difficulty in bounding the last term $\mathcal{I}_4$, is not just a technical limitation, but rather it reflects an essential difficulty. Here is one possible attempt to address this difficulty, which will make our point.

One might argue that the two terms $\mathcal{I}_3 + \mathcal{I}_4$ can add up to a negative quantity (at least for a sufficiently small time-step Δt). Unfortunately, this cannot be the general case. Indeed, consider the possibility of a negative $\widetilde{L}$ such that the "good" negative contribution in $\mathcal{I}_3$, $\Re(u^n, \widetilde{L}u^n)$, is close to the imaginary axis from the left, i.e., $\Re(u^n, \widetilde{L}u^n) \approx 0$, yet, the next "bad" contribution in $\mathcal{I}_4$, $-\Re(\widetilde{L}u^n, \widetilde{L}^2 u^n)$, is bounded away from the imaginary axis, i.e., $-\Re(\widetilde{L}u^n, \widetilde{L}^2 u^n) > 0$. In such cases, the sum of these two terms, which are of different orders of magnitude, turns out to be positive, and thus strong stability fails. One can trace this difficulty to the behavior of the numerical range of L, (w, Lw), near the imaginary axis. This argument is demonstrated

by the counterexample of an unstable RK approximation of negative problems shown in section 3.5.

To deal with this difficulty, we upper bound $\mathcal{I}_4$, using the weighted Cauchy–Schwartz inequality,

$$\mathcal{I}_4 = -\frac{2}{3}\Re(\widetilde{L}u^n, \widetilde{L}^2u^n) \le \frac{2}{3}|(\widetilde{L}u^n, \widetilde{L}^2u^n)| \le \frac{2}{3}\|\widetilde{L}u^n\| \cdot \|\widetilde{L}^2u^n\|$$
$$\le \frac{1}{3c}\|\widetilde{L}^2u^n\|^2 + \frac{c}{3}\|\widetilde{L}u^n\|^2 =: \mathcal{I}_{41} + \mathcal{I}_{42} \qquad \forall c > 0.$$

Thus, in order to make the RHS of (3.12) negative, the following two conditions need to be satisfied simultaneously.

1. For the sum of $\mathcal{I}_3$ and $\mathcal{I}_{42} := \frac{c}{3}\|\widetilde{L}u^n\|^2$:

$$\frac{2}{3}\Re(u^n, \widetilde{L}u^n) + \frac{c}{3}\|\widetilde{L}u^n\|^2 \le 0. \tag{3.14}$$

2. For the sum of $\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_5$ and $\mathcal{I}_{41} := \frac{1}{3c}\|\widetilde{L}^2u^n\|^2$ to be negative:

$$\|\widetilde{L}^2u^n\|^2 \left(-\frac{1}{12} + \frac{\|\widetilde{L}\|^2}{36} + \frac{\|\widetilde{L}\|}{2} + \frac{1}{3c}\right) \le 0. \tag{3.15}$$

Using the coercivity assumption (3.2) and the CFL condition $\|\widetilde{L}\| \le 1$, (3.14) and (3.15) become

$$\begin{cases} \Delta t \le \dfrac{2}{c}\eta, \\ \\ \Delta t \le \dfrac{3}{19} \cdot \dfrac{c-4}{c\|L\|}. \end{cases} \tag{3.16}$$

Using the free parameter, $c > 4$, to equilibrate the last two upper bounds (with $c = (38\eta\|L\| + 12)/3$ and noting that $\eta\|L\| \le 1$), yields the CFL condition

$$\Delta t \le \frac{3}{25}\eta, \qquad \eta := \inf_{\|u\|=1} \frac{|\Re(u, Lu)|}{\|Lu\|^2}. \tag{3.17}$$

To summarize, by (3.12) we have strong stability, $\|u^{n+1}\| \le \|u^n\|$, provided that (3.17) holds.

Note that the CFL condition (3.17) has precisely the same form as the CFL condition we previously met in (3.5) when analyzing the stability of the FE method. The only difference is that the stability of RK3 allows a time-step which is at most 3/50 of the time-step allowed be FE method. As mentioned before, we do not claim this CFL condition to be sharp.

Putting it all together, we arrive at the following theorem.

THEOREM 3.2. *Consider a well-posed problem* (3.1)*, $u_t = Lu$, associated with the coercive L,* (3.2)*. Then its third-order RK discretization is strongly stable, under a CFL condition*

$$\Delta t \le \frac{3}{25}\eta, \qquad \eta := \inf_{\|u\|=1} \frac{|\Re(Lu, u)|}{\|Lu\|^2}.$$

To summarize, we conclude that the FE as well as the RK3 methods are strongly stable under appropriate similar CFL conditions. In both cases, for such CFL conditions to hold, we require more that just negativity. The admissible L's must be coercive, so that the ratio $|\Re(u, Lu)|/\|Lu\|^2$ remains uniformly bounded from below.

3.4. The fourth-order RK approximation of coercive problems. We repeat the previous calculations for the fourth-order fully discrete method. We consider the same negative-definite semidiscrete problem, and discretize the time variable using the RK4 method. With the same abbreviations as before, we have

$$u^{n+1} = u^n + \widetilde{L}u^n + \frac{1}{2}\widetilde{L}^2u^n + \frac{1}{6}\widetilde{L}^3u^n + \frac{1}{24}\widetilde{L}^4u^n, \qquad \tilde{L} := \Delta t L.$$

This yields

$$\begin{aligned}
\|u^{n+1}\|^2 - \|u^n\|^2 &= \left(\sum_{i=0}^{4}\frac{\widetilde{L}^i}{i!}u^n, \sum_{i=0}^{4}\frac{\widetilde{L}^i}{i!}u^n\right) - (u^n, u^n) \\
&= \|\widetilde{L}u^n\|^2 + \frac{1}{4}\|\widetilde{L}^2u^n\|^2 + \frac{1}{36}\|\widetilde{L}^3u^n\|^2 + \frac{1}{24^2}\|\widetilde{L}^4u^n\|^2 \\
&+2\Re(u^n, \widetilde{L}u^n) + \Re(u^n, \widetilde{L}^2u^n) \\
&+\frac{1}{3}\Re(u^n, \widetilde{L}^3u^n) + \frac{1}{12}\Re(u^n, \widetilde{L}^4u^n) + \Re(\widetilde{L}u^n, \widetilde{L}^2u^n) \\
&+\frac{1}{3}\Re(\widetilde{L}u^n, \widetilde{L}^3u^n) + \frac{1}{12}\Re(\widetilde{L}u^n, \widetilde{L}^4u^n) \\
&+\frac{1}{6}\Re(\widetilde{L}^2u^n, \widetilde{L}^3u^n) + \frac{1}{24}\Re(\widetilde{L}^2u^n, \widetilde{L}^4u^n) + \frac{1}{72}\Re(\widetilde{L}^3u^n, \widetilde{L}^4u^n).
\end{aligned}$$

Using inequalities (3.6), (3.7), (3.8), (3.9), (3.10), and (3.11), the last quantity can be upper bounded:

$$\begin{aligned}
\|u^{n+1}\|^2 - \|u^n\|^2 \le &-\frac{1}{72}\|\widetilde{L}^3u^n\|^2 + \frac{1}{576}\|\widetilde{L}^4u^n\|^2 + \frac{7}{12}\Re(\widetilde{L}u^n, u^n) - \frac{5}{6}\Re(\widetilde{L}u^n, \widetilde{L}^2u^n) \\
&-\frac{17}{24}\Re(\widetilde{L}^2u^n, \widetilde{L}^3u^n) - \frac{7}{36}\Re(\widetilde{L}^3u^n, \widetilde{L}^4u^n) =: \mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_3 + \mathcal{I}_4 + \mathcal{I}_5 + \mathcal{I}_6.
\end{aligned} \tag{3.18}$$

As previously done in the third-order case (consult (3.12)), the RHS of (3.18) need not be negative (for all L's), unless we place a further restriction on the class of admissible L's.

- For the sum of the first, the second, and the sixth terms, $\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_6$, we use the estimate

$$\begin{aligned}
&-\frac{1}{72}\|\widetilde{L}^3u^n\|^2 + \frac{1}{576}\|\widetilde{L}^4u^n\|^2 - \frac{7}{36}\Re(\widetilde{L}^3u^n, \widetilde{L}^4u^n) \\
&\le -\frac{1}{72}\|\widetilde{L}^3u^n\|^2 + \frac{1}{576}\|\widetilde{L}^4u^n\|^2 + \frac{7}{36}|(\widetilde{L}^3u^n, \widetilde{L}^4u^n)| \\
&\le -\frac{1}{72}\|\widetilde{L}^3u^n\|^2 + \frac{1}{576}\|\widetilde{L}\|^2\|\widetilde{L}^3u^n\|^2 + \frac{7}{36}r(\widetilde{L})\|\widetilde{L}^3u^n\|^2 \\
&\le \|\widetilde{L}^3u^n\|^2\left[-\frac{1}{72} + \frac{\|\widetilde{L}\|^2}{576} + \frac{7}{36}\|\widetilde{L}\|\right].
\end{aligned} \tag{3.19}$$

  The sum of the last three terms is negative, for a sufficiently small Δt.
- The third term, $\mathcal{I}_3$, is negative due to the negativity of the operator $\widetilde{L}$.

- Finally, the sum of the fourth and the fifth terms, $\mathcal{I}_4 + \mathcal{I}_5$, is positive (due to the negativity of $\widetilde{L}$), and as before, the additional negative third term can not help in case $\mathcal{I}_3 = 7/12\Re(\widetilde{L}u^n, u^n) \approx 0$.

In order to equilibrate these positive and negative terms, we upper bound $\mathcal{I}_4$ and $\mathcal{I}_5$ using the weighted Cauchy–Schwartz inequality (here $c_1 > 0$ and $c_2 > 0$ denote arbitrary constants):

$$\mathcal{I}_4 = -\frac{5}{6}(\widetilde{L}^2 u^n, \widetilde{L}u^n) \leq \frac{5}{12}\left[c_1\|\widetilde{L}u^n\|^2 + \frac{1}{c_1}\|\widetilde{L}^2 u^n\|^2\right]$$

and

$$\mathcal{I}_5 = -\frac{17}{24}(\widetilde{L}^3 u^n, \widetilde{L}^2 u^n) \leq \frac{17}{48}\left[c_2\|\widetilde{L}^2 u^n\|^2 + \frac{1}{c_2}\|\widetilde{L}^3 u^n\|^2\right] =: \mathcal{I}_{51} + \mathcal{I}_{52}.$$

Here, we are led to the following two CFL conditions:

1. For $\mathcal{I}_1 + \mathcal{I}_2 + \mathcal{I}_6$ and $\mathcal{I}_{52} := 17/48c_2\|\widetilde{L}^3 u^n\|^2$ to be negative, we require

$$\|\widetilde{L}u^n\|^2\left(-\frac{1}{72} + \frac{\|\widetilde{L}\|^2}{576} + \frac{7}{36}\|\widetilde{L}\| + \frac{17}{48c_2}\right) \leq 0.$$

2. For $\mathcal{I}_3 + \mathcal{I}_4$ and $\mathcal{I}_{51} = 17/48c_2\|\widetilde{L}^2 u^n\|^2$ to be negative, we require

$$\frac{7}{12}\Re(\widetilde{L}u^n, u^n) + \frac{5}{12}\left(c_1\|\widetilde{L}u^n\|^2 + \frac{1}{c_1}\|\widetilde{L}^2 u^n\|^2\right) + \frac{17}{48}c_2\|\widetilde{L}^2 u^n\|^2 \leq 0.$$

These two requirements can be simplified to the following:

$$\begin{cases} \Delta t \leq \dfrac{4}{113}\dfrac{2c_2 - 51}{c_2\|L\|}, \\ \\ \Delta t \leq \dfrac{28}{17c_2 + 20(c_1 + \frac{1}{c_1})}\eta. \end{cases}$$

And by equilibrating these two conditions, we end with

$$\Delta t \leq \frac{1}{31}\eta, \qquad \eta := \inf_{\|u\|=1}\frac{|\Re(Lu, u)|}{\|Lu\|^2}. \tag{3.20}$$

As before, we note that the CFL condition (3.20) resembles the previous CFL conditions we met following the stability analysis of the RK1 and RK3 methods. The similarity between (3.20) and (3.5) leads to the following theorem, concerning the strong stability of the fourth-order RK method for coercive problems, which concludes the results of this section.

THEOREM 3.3. *Consider a well-posed problem $u_t = Lu$, associated with the coercive L,* (3.2)*. Then the RK4 method is strongly stable, under a CFL condition*

$$\Delta t \leq \frac{1}{31}\eta, \qquad \eta := \inf_{\|u\|=1}\frac{|\Re(Lu, u)|}{\|Lu\|^2}.$$

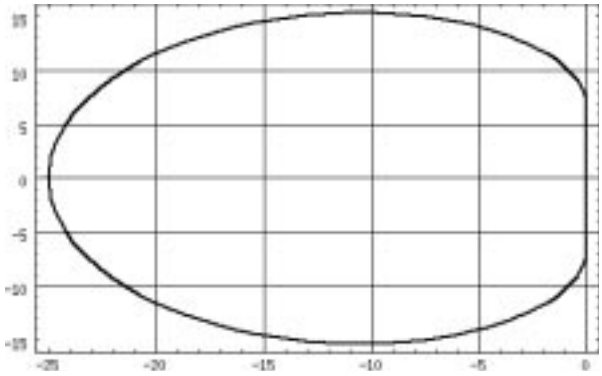

FIG. 2. *The numerical range of* A_5.

3.5. Negativity need not imply power boundedness. The following example demonstrates that the L^2-power growth of the RK methods is possible even for negative-definite operators. It amplifies the critical role of the behavior of the field of values near the imaginary axis in this context. Furthermore, this emphasizes our results, showing the need to strengthen negativity with an additional coercivity requirement, e.g., in order to guarantee L^2-power boundedness. The construction of this counterexample was motivated by the difficulties we encountered in sections 3.3 and 3.4—specifically, the nonnegative term $\mathcal{I}_4$ in (3.12); it shows that these are not just technical difficulties, but rather essential ones.

Consider the semidiscrete problem, $u_t = A_5 u$, where A_5 is given by the nonnormal negative-definite 5×5 block,

$$A_5 = \begin{pmatrix} -5 & -10 & -10 & -10 & -10 \\ 0 & -5 & -10 & -10 & -10 \\ 0 & 0 & -5 & -10 & -10 \\ 0 & 0 & 0 & -5 & -10 \\ 0 & 0 & 0 & 0 & -5 \end{pmatrix}.$$

In Figure 2 we drew the numerical range of A_5.

The semidiscrete problem was discretized using both RK3 and RK4 methods, with $\Delta t = 0.1, 0.08, \ldots, 0.02$. The numerical radius of $P_3 = P_3(\Delta t A_5)$ and $P_4 = P_4(\Delta t A_5)$ for each time-step is given in the following table (P_i denotes the ith-order RK characteristic polynomial).

Δt	$r(P_3)$	$r(P_4)$
0.02	1.00423	1.00419
0.04	1.01656	1.01587
0.06	1.03816	1.03409
0.08	1.07241	1.05732
0.1	1.12623	1.08361

Note that as $\Delta t \to 0$, $\rho(P_j) \to 1$ for $j = 3, 4$. But $r(P_j) > 1$ and thus $\|P_j\| > 1$.

Actually, this growth in the L^2-norm of P_i's occurs in just a single step of the RK solver. In further iterations, the norm will remain bounded, despite its initial growth. Thus, despite the lack of strong stability, $\|P_i(\Delta t A_5)\| > 1$, we have L^2-power boundedness, $\|P_i^n(\Delta t A_5)\| \leq$ Const., which in turn, by KMT, implies strong stability with respect to a different H-weighted norm.

The essential observation here is that we are not just interested in the power boundedness of the matrix A_5, but rather, in the power boundedness of a whole family $\mathcal{F} = \{A_i\}$, a family of such matrices A, with an *increasing size* i. The growth in the initial steps of the iterative scheme, for every member of this family, will prevent the desired uniform power boundedness of all the members of the family $\mathcal{F}$.

Note that the field of values of A_i is not bounded away from the imaginary axis. As the dimension of the A_i increases, the intersection between the field of values of A_i and the imaginary axis increases and thus, the time-step Δt tends to zero.

4. Examples of multidimensional systems with variable coefficients.

4.1. Convection-diffusion equations—Finite difference schemes and Fourier method. We consider a general system of convection-diffusion equations in d spatial dimensions, with variable—possibly time-dependent coefficients, i.e.,

$$u_t = \sum_{j=1}^{d} A_j(x,t)\partial_{x_j} u + \frac{\varepsilon}{2} \sum_{j,k=1}^{d} \partial_{x_k}(Q_{jk}(x,t)\partial_{x_j} u) + B(x,t)u. \tag{4.1}$$

Here $A_j(x,t) \in C^1$ are the symmetric convection matrices, and $Q_{jk} \in C^1$ are the symmetric diffusion matrices with the positive block form, $\mathcal{Q} := \{Q_{jk}\}$,

$$< \mathcal{Q}\xi, \xi >:= \sum_{j,k=1}^{d} < Q_{jk}\xi_j, \xi_k > \ \geq q_0|\xi|^2, \qquad q_0 > 0, \qquad \forall \xi.$$

In this section we demonstrate how to apply the stability theory developed in section 3 to a rather general framework of difference approximations for system (4.1). Specifically, we concentrate on a family of *central schemes*, where the spatial derivatives on the RHS of (4.1) are discretized by centered differences, time is discretized by RK method, and we then address the question of (strong) stability.

As a prototype example, we consider the second-order centered-difference approximation,

$$u_t = \mathcal{L}u, \qquad \mathcal{L} := \sum_{j=1}^{d} A_j(x,t)D_{0j} + \frac{\varepsilon}{2} \sum_{j,k=1}^{d} D_{-j}(Q_{jk}(x,t)D_{+k}) + B(x,t), \tag{4.2}$$

where, $D_{0j} = D_{0j}(\Delta x), D_{\pm j} = D_{\pm j}(\Delta x)$ are the usual centered and one-sided finite-difference operators in the jth direction,

$$D_{\pm j}(\Delta x) := \pm \frac{T_j^{\pm 1} - I}{\Delta x} \quad \text{and} \quad D_{0j}(\Delta x) := \frac{D_{+j}(\Delta x) + D_{-j}(\Delta x)}{2},$$

expressed in terms of the translation in the jth direction, $T_j = T_j(\Delta x)$, $T_j w(x) := w(x + e_j \Delta x)$, with Δx denoting the vanishing gridsize.

Our main stability results of section 3 require the underlying assumption of coercivity to be fulfilled, which in turn, limits our discussion to a subclass of *dissipative* difference schemes—that is, difference schemes which are dissipative in the sense of Kreiss [58, section 5]. In this context, we should highlight the following points comparing our stability approach, with the classical von Neumann analysis which is part of the general framework of the Lax–Richtmyer stability theory [58, section 4].

- First, our analysis is based on the straightforward energy method, which applies equally well to the constant coefficient case, as well as for problems with variable coefficients (depending on both space and time). The Lax–Richtmyer theory consists of two main ingredients: the von Neumann stability analysis which applies to the *general* constant-coefficient problems and the method of localization ("freezing the coefficients") which extends the analysis to problems with variable coefficients. Both approaches necessitate some form of numerical dissipation for the stability of difference schemes with variable coefficients.
- Second, we derive *strong stability*, measured in the appropriate weighted norm associated with the underlying well-posed semidiscrete problem. The Lax–Richtmyer theory proves the strong stability for the constant coefficient *principle part* of the problem.
- The third important point is the *versatility* of our approach. Indeed, the stability of the RK methods for (4.2), which we discuss below, is a case in point: answering this stability question with the Lax–Richtmyer theory is an intricate and far from trivial task. In particular, it does not apply to global difference operators (such as spectral differencing discussed later in this subsection). Since our approach is based on the energy method (which is carried in the physical space), and since it avoids the von Neumann analysis (which is carried in the dual Fourier space), we are able to easily adapt additional extensions due to nonperiodic boundary conditions, general geometries, etc.

As a preliminary step before addressing the L^2-stability question of (4.2), we recall that such L^2-stability is invariant with respect to changes of the low-order terms. We pause to elaborate on this well-known point. Let $\mathcal{M}u = \mathcal{L}u + Cu$ denote an arbitrary bounded perturbation of our discretization (4.2), which is formed by adding the bounded matrix C, $\|C\| < c_1$, with constant c_1 independent of Δx or Δt. We claim that if the RK discretization associated with (4.2), $p(\Delta t\mathcal{L})$, is power bounded, $\|p(\Delta t\mathcal{L})^n\| \le c_2$, so does any bounded perturbation of $\mathcal{L}$, i.e., $p(\Delta t\mathcal{M})$ is power bounded as well. Indeed, since

$$p(\Delta t\mathcal{M}) = p(\Delta t\mathcal{L}) + \Delta tC + O(\Delta t^2), \qquad \|\Delta tC + O(\Delta t^2)\| \le \Delta t c_3,$$

we find by the Kreiss–Strang perturbation theorem (consult [58, section 3.9]),

$$\|p(\Delta t\mathcal{M})^n\| \le c_2(1 + c_2\Delta t c_3)^n \le c_2 e^{c_2 c_3 T}, \qquad \forall n\Delta t \le T.$$

This shows that L^2-stability is invariant under low-order bounded perturbations as asserted. We should emphasize that this invariance is a cornerstone of any reasonable stability theory, which should cover general linear discretizations, very much the same way that in the differential framework, the well-posedness of time-dependent systems (hyperbolic, parabolic, ...) is invariant under lower-order terms. In particular, this invariance allows us to *freeze* any time dependence of the coefficients, so that different variants of RK methods coincide with its underlying truncated Taylor expansion. Indeed, a stability theory based on the weaker resolvent condition, (2.28), rather than our use of the power boundedness (2.18), also enjoys this kind of invariance with respect to low-order perturbations. This invariance motivated Kreiss and Wu [43] to base their stability analysis on such a *resolvent condition*.

We now turn to our stability study of the approximation (4.2). By our previous argument, we may consider *any* low-order term, B. We choose to restrict our attention to a special $B(x,t) = 1/2\sum_j D_{0j}A_j(x,t)$, so that the corresponding problem, (4.1),

is *strongly* L^2-well-posed, $\|u(t)\| \le \|u(0)\|$ (integration by parts). Then, the resulting central discretization, (4.2), can be rewritten in the antisymmetric form

$$u_t = \sum_j A_j(x,t) D_{0j} u + \frac{\varepsilon}{2} \sum_{j,k} D_{-j}(Q_{jk}(x,t) D_{+k}) u + \Big(\frac{1}{2} \sum_j D_{0j} A_j(x,t) \Big) u. \tag{4.3}$$

Verifying the stability of RK discretizations for (4.3), based on our approach presented in section 3, is now straightforward. We proceed as follows.

We first note that the underlying operator $\mathcal{L}$ is negative definite. Indeed, summation by parts (which takes into account the symmetry of the A_j's) yields

$$\begin{aligned} (\mathcal{L}u, u) &= \sum_j (A D_{0j} u, u) + \frac{\varepsilon}{2} \sum_{j,k} (D_{-j} Q_{jk} D_{+k} u, u) + (Bu, u) \\ &= -\sum_j \left(\frac{1}{2} (D_{0j} A_j) u, u \right) - \frac{\varepsilon}{2} \sum_{j,k} (Q_{jk} D_{+k} u, D_{+j} u) + (Bu, u). \end{aligned} \tag{4.4}$$

By our choice of $B(x,t) = 1/2 \sum_j D_{0j} A_j(x,t)$, the sum of the first and third terms vanishes, and since $Q = \{Q_{jk}\}$ is positive definite, $\mathcal{L}$ is nonpositive

$$\Re(\mathcal{L}u, u) = -\frac{\varepsilon}{2} \left(\sum_{j,k} Q_{jk} D_{+k} u, D_{+j} u \right) \le 0.$$

We are now able to apply our sufficient stability criteria from section 3. According to (3.5), (3.17), (3.20), the s-order RK discretization, (4.3), is *strongly stable*, provided Δt satisfies the CFL condition

$$\Delta t \le C_s \eta, \qquad \eta := \inf_{\|u\|=1} \frac{|\Re(\mathcal{L}u, u)|}{\|\mathcal{L}u\|^2}, \tag{4.5}$$

for the appropriate C_s calculated above.

To fulfill the CFL condition (4.5), we first deal with the coercivity of $\mathcal{L}$. A straightforward calculation borrowed from (4.4) yields

$$|\Re(\mathcal{L}u, u)| = \Big| -\frac{\varepsilon}{2} \sum_{j,k} (Q_{jk} D_{+k} u, D_{+j} u) \Big| \ge \frac{\varepsilon}{2} q_0 \sum_j \|D_{+j} u\|_{l_2}^2 =: \frac{\varepsilon}{2} q_0 \|Du\|_{l_2}^2. \tag{4.6}$$

Moreover, since $\|D_{-j} v\| \le \frac{2}{\Delta x} \|v\|$, we find

$$\|\mathcal{L}u\| \le \max_j |A_j| \cdot \sum_j \|D_{0j} u\|_{l_2} + \frac{\varepsilon}{\Delta x} \Big[\sum_{j,k} \|Q_{jk} D_{+k} u\|_{l_2} \Big] \le \left(a + \frac{\varepsilon Q}{\Delta x} \right) \|Du\|_{l_2}, \tag{4.7}$$

where here and below we use the abbreviations $a := \max_j |A_j|$, $Q := \sum_j \max_k \|Q_{jk}\|$. Thus, $\mathcal{L}$ satisfies the coercivity condition (3.2), $|\Re(\mathcal{L}u, u)|/\|\mathcal{L}u\|^2 \ge \eta$, with $\eta = (1/2)\varepsilon q_0 (a + \varepsilon Q/\Delta x)^{-2}$, and the CFL condition (4.5) then reads

$$\Delta t \left(a + \frac{\varepsilon Q}{\Delta x} \right)^2 \le \frac{1}{2} C_s \varepsilon q_0. \tag{4.8$_s$}$$

With the usual notations for mesh ratios, $\lambda := \frac{\Delta t}{\Delta x}$ and $\mu := \frac{\Delta t}{(\Delta x)^2}$, (4.8$_s$) can be summarized in the following assertion.

ASSERTION 4.1. *The s-order RK discretization of the finite-difference approximation* (4.3) *is stable under the CFL condition*

$$(\lambda a + \mu\varepsilon Q)^2 \le \frac{1}{2} C_s \mu\varepsilon q_0, \quad C_1 = 2, \quad C_3 = \frac{3}{25}, \quad C_4 = \frac{1}{31}. \tag{4.9}$$

We highlight the following two different types of equations:

1. The s-order RK approximation of the *uniformly parabolic* system (4.3), $\varepsilon \equiv 1$. Here the CFL condition (4.9), expressed in terms of the relevant *fixed parabolic* mesh ratio, $\mu = \frac{\Delta t}{(\Delta x)^2}$ (so that $\lambda = O(\Delta x)$), reads

$$\frac{\Delta t}{(\Delta x)^2} Q \le \text{Const.}, \quad \text{Const.} = \frac{1}{2} C_s \frac{q_0}{Q} + O(\Delta x). \tag{4.10}$$

2. The s-order RK approximation of the hyperbolic system (4.3) with *vanishing viscosity*, $\varepsilon \equiv \Delta x$. Here, the CFL condition (4.9), expressed in terms of the relevant *fixed hyperbolic* mesh ratio, λ, takes the form

$$\frac{\Delta t}{\Delta x} a \le \text{Const.}, \quad \text{Const.} = \frac{1}{2} C_s \frac{a q_0}{(a+Q)^2}. \tag{4.11}$$

How sharp is our stability condition? It is instructive to compare our last general stability result to the von Neumann analysis for the special scalar one-dimensional constant coefficients FE scheme, where $A_{11} = a$ and $Q_{11} = q_0 = q$. Here, the von Neumann stability requirement boils down to bounding the symbol, $|1 + \lambda a\imath \sin\xi + \mu\varepsilon q(\cos\xi - 1)| \le 1$, which in turn, yields the von Neumann condition

$$(\lambda a)^2 \le \mu\varepsilon q \le 1. \tag{4.12}$$

Comparing this von Neumann condition with our general stability condition (4.9) for the scalar FE scheme (with $Q = q_0 = q$ and $C_1 = 2$), we distinguish between the following two cases:

• *The uniformly parabolic case ($\varepsilon \equiv 1$).* Here, (4.12) becomes $(\lambda a)^2 \le \mu q \le 1$. It represents the standard parabolic CFL condition, which is upper bounded by the cell Reynolds number. Hence, the bound on the time-step boils down to $\Delta t \le \min(q/a^2, (\Delta x)^2/q)$, and for sufficiently small Δx's this condition reads

$$\frac{\Delta t}{(\Delta x)^2} q \le 1.$$

Our general stability result is sharp enough to yield even in this special one-dimensional scalar case the same CFL condition $\Delta t/(\Delta x)^2 q \le 1$, consult (4.10).

• *The case of vanishing viscosity ($\varepsilon \equiv \Delta x$).* Here, the von Neumann condition (4.12) becomes $\lambda a^2 \le q \le 1/\lambda$. This CFL condition reflects a family of well-known stable schemes ranging from the upperbound on the right, $q \equiv 1/\lambda$, which corresponds to the central Lax–Friedrichs (LxF) scheme, and ending with the lower bound $q = \lambda a^2$, which corresponds to the second-order Lax–Wendroff (LxW) scheme. This should be compared with our general stability condition (4.11), which in this case amounts to

$$\frac{\lambda a^2}{\theta^2} \le q \le \frac{(1-\theta)^2}{\lambda}, \qquad 0 < \theta < 1.$$

Our condition covers a smaller range of approximations compared with those covered by the von Neumann condition; yet, with $\theta = 1/2$ for example, it still includes a family of schemes ranging from the central modified LxF scheme [69], $q \equiv 1/4\lambda$, and in particular, the upwind scheme which corresponds to $q = |a|$ with the CFL condition $\lambda|a| \le 1/4$; moreover, if we let $\theta \uparrow 1$, we approach the viscosity coefficient q corresponding to the second-order LxW scheme.

It is remarkable that our general stability condition which was developed in the general setup of multidimensional variable coefficients systems, produces such sharp results in the special one-dimensional constant coefficient case. We would also like to highlight again the *versatility* of our approach; the above stability analysis can be easily extended to rather general discretizations, as told by the following theorem.

THEOREM 4.1. *Consider a general system of convection-diffusion equations in d spatial dimensions, with variable coefficients,*

$$u_t = \sum_{j=1}^{d} A_j(x,t)\partial_{x_j} u + \frac{\varepsilon}{2}\sum_{j,k=1}^{d} \partial_{x_k}(Q_{jk}(x,t)\partial_{x_j} u) + B(x,t)u. \tag{4.13}$$

Let $\mathcal{D}_{+j}$ denote a general discretization of the derivative in the x_j-direction,

$$\mathcal{D}_{+j} = \sum_m \frac{\alpha_m}{\Delta x} T_j^m.$$

Let $\mathcal{D}_{-j} := -\mathcal{D}_{+j}^*$ denote its skew adjoint, and let $\mathcal{D}_{0j}$ denote the corresponding "central" differencing, $\mathcal{D}_{0j} := 1/2(\mathcal{D}_{+j} + \mathcal{D}_{-j})$. Consider the following semidiscrete system:

$$u_t = \mathcal{L}u, \qquad \mathcal{L} := \sum_{j=1}^{d} A_j(x,t)\mathcal{D}_{0j} + \frac{\varepsilon}{2}\sum_{j,k=1}^{d} \mathcal{D}_{-j}(Q_{jk}(x,t)\mathcal{D}_{+k}) + B(x,t). \tag{4.14}$$

Then the s-order fully discrete RK approximations of (4.13) *are strongly stable provided that the following CFL condition holds (– compare with* (4.8_s)*)*

$$\Delta t\left(a + \frac{|\alpha|\varepsilon Q}{2\Delta x}\right)^2 \le \frac{1}{2} C_s \varepsilon q_0, \qquad |\alpha| := \sum_m |\alpha_m|. \tag{4.15}$$

The proof is immediate along the lines of our above arguments. Since $\mathcal{D}_{+j}$ and $\mathcal{D}_{-j}$ are skew adjoint to each other, summation by parts leaves (4.6) unchanged, i.e., $|\Re(\mathcal{L}u, u)| \ge \varepsilon/2q_0\|\mathcal{D}u\|_{l_2}^2$. And since $\|\mathcal{D}_{-j}v\| \le |\alpha|/\Delta x\|v\|$, one recovers (4.7) with $(a + |\alpha|\varepsilon Q/2\Delta x)$, and the result follows.

We note in passing that the dependence between the stability condition, the l_1 norm and locality, was pointed out earlier in Tadmor's analysis [72, p. 534].

We close this section with several remarks. First, a couple of examples.

1. *High-order finite-difference approximations.* We demonstrate the application of Theorem 4.1 in the context of some prototype finite-difference discretizations. Consider, for example, the *fourth-* and *sixth-order* spatial discretizations,

$$\mathcal{D}_{\cdot j}^4(\Delta x) = \frac{4\mathcal{D}_{\cdot j}(\Delta x) - \mathcal{D}_{\cdot j}(2\Delta x)}{3}, \quad \mathcal{D}_{\cdot j}^6(\Delta x) = \frac{15\mathcal{D}_{\cdot j}(\Delta x) - 6\mathcal{D}_{\cdot j}(2\Delta x) + \mathcal{D}_{\cdot j}(3\Delta x)}{10},$$

(4.16)

expressed in terms of the second-order differences we had earlier, $\mathcal{D}_{\pm j}$ and $\mathcal{D}_{0j}$. Then Theorem 4.1 yields the stability of the fourth- and sixth-order schemes as told by the following corollary.

COROLLARY 4.1. *Consider the multidimensional system* (4.1). *Discretize the spatial derivatives using either the fourth- or the sixth-order differencing operators* $\mathcal{D}_{\cdot j}$, *in* (4.16) *and use the s-order RK discretization for the time variable. Then, the resulting variable coefficients scheme is (strongly) stable under the CFL condition* (4.15),

$$\Delta t\left(a+\frac{|\alpha|\varepsilon Q}{2\Delta x}\right)^2\le\frac{1}{2}C_s\varepsilon q_0,\quad C_1=2,\quad C_3=\frac{3}{25},\quad C_4=\frac{1}{31},$$

with $|\alpha|=3$ *and* $|\alpha|=11/3$ *for the fourth- and sixth-order schemes, respectively.* This result applies both to the uniformly parabolic case ($\varepsilon\equiv 1$) as well as to the hyperbolic case with vanishing viscosity ($\varepsilon\equiv\Delta x$).

2. *Spectral approximations.* One may continue to apply the Richardson extrapolation to the spatial differences, increasing the order (and size of stencil...) of the central differences, arriving at the limit to the *spectral* discretization (consult [12], [23]),

$$\mathcal{D}_{+j}=\mathcal{D}_{-j}=\mathcal{D}_{0j}=\sum_{m\neq 0}\frac{(-1)^{m+1}}{2\sin(\frac{m\Delta x}{2})}T_j^m.$$

In this case, $|\alpha|=\sum_{|m|\le\pi/2\Delta x}|\Delta x/\sin(m\Delta x/2)|\sim|\log\Delta x|$. Therefore, Theorem 4.1 tells us that only a negligible amount of additional viscosity is required to stabilize the *global* spectral method. We have the following corollary.

COROLLARY 4.2. *Consider the multidimensional system* (4.1), *where space is discretized using spectral discretization and time is discretized using an s-order RK method. Then these discretizations are strongly stable provided that (– consult* (4.15)*)*

$$\Delta t\left(a+O(\frac{|\log\Delta x|\varepsilon Q}{2\Delta x})\right)^2\le\frac{1}{2}C_s\varepsilon q_0.$$

In particular

(a) *the uniformly parabolic case requires the slightly stricter CFL condition (– compared with* (4.10)*)*

$$\Delta t\left(\frac{\log\Delta x}{\Delta x}\right)^2 Q\le\text{Const.},$$

(b) *the case of vanishing viscosity ($\varepsilon\equiv\Delta x$) which requires the slightly stricter CFL condition (– compared with* (4.11)*)*

$$\frac{\Delta t}{\Delta x}(a+O(|\log\Delta x|))^2\le\text{Const.}$$

In contrast, one can use the *staggered* spectral differencing [13] with bounded l_1 size, $|\alpha|\sim\sum 1/m^2\le$ Const. Such stability results concerning *global* spectral discretizations with variable coefficients are beyond the scope of the von Neumann and Lax–Richtmyer stability theory.

So far, we have focused on the special case of the low-order term, $B=1/2\sum_j D_{0j}A_j$. Our next and final remark recovers the general case.

3. *Strong stability.* In this subsection we proved the strong stability of discretizations to system (4.1). We considered only the special choice of its low-order term, B, $B = 1/2 \sum_j D_{0j} A_j$, so that the resulting system (4.1) and all its discretizations are *strongly* stable in the sense, $\|u(t)\| \le \|u(0)\|$. An extension of our results to *general* B's will follow the differential well-posedness a priori estimate

$$\|u(t)\| \le \exp^{\gamma t} \|u(0)\|, \quad \gamma := \Big\| \Re\Big(B - \frac{1}{2} \sum_j D_{0j} A_j \Big) \Big\|.$$

It follows that all of our stability results asserted above, as they are based on the straightforward energy method approach, retain the corresponding stability estimate

$$\|u(t^n)\|_{l_2} \le \exp^{\gamma t^n} \|u(0)\|_{l_2}.$$

Thus, our stability analysis shows, in particular, that the l_2 growth of the s-order RK approximation is not faster than the growth dictated by the differential framework.

In conclusion, we comment again on the *versatility* of our analysis. The method presented is a general energy method, which applies equally well to such setups where none of the available stability results will work. The von Neumann analysis is restricted to periodic problems. The Gustafsson, Kreiss, and Sundström (GKS) stability analysis [26], which deals with nonperiodic initial boundary value problems, does not preserve the *strong* stability, in the sense that it allows a faster time growth of the numerical solution than the one dictated by the differential level. The only ingredient that has to be verified in our approach is the *coercivity*.

4.2. Convection equations–Chebyshev ψdo-spectral approximations. In this example, we analyze the stability of the high-order RK Chebyshev ψdo-spectral approximations for convection equations with variable coefficients. Gottlieb and Tadmor [24], proved the stability of the forward Euler method for the Chebyshev ψdo-spectral approximation; here, we extend their results to the RK3 and RK4 methods.

The Chebyshev ψdo-spectral approximation is a primary example for the intricate issue of stability for its fully discrete RK scheme. Specifically, here the Chebyshev spatial stencil is a *global* stencil, such as the spectral stencils encountered in Corollary 4.2 above. More importantly, the differentiation matrix associated with this problem (abbreviated D_T below), is the prototype example for a *non-normal* matrix, [23]. This non-normality prevents a straightforward extension of the scalar constant coefficients stability analysis into the case of systems with variable coefficients as we outlined in section 2.3.2. In particular, in this context we recall that a notion of stability based on the resolvent estimate (2.28) fails to provide a *uniform* stability bound which is independent of the growing dimension. It is the advantage of our energy method that we are able to derive such *uniform* (strong) stability bounds for the case of the Chebyshev ψdo-spectral approximation.

Let $u(t) = (u_1, \dots, u_N)$, denote the vector of computed values at the Chebyshev grid points, and we let $u'(t) = (u_1', \dots, u_N')$, denote the corresponding vector of derivatives. Here, discrete differentiation stands for the differentiation of the corresponding Chebyshev interpolant. Thus, we let $u_N(x)$ denote the Chebyshev interpolant given at the $N+1$ points—the interior Chebyshev collocation points, $x_\nu = \cos(\nu + 1/2)\pi/N$, $\nu = N-1, N-2, \dots, 0$, and the augmented boundary point $x = 1$. We then set $u_\nu' = u_N'(x)|_{x=x_\nu}$. Being linear, this has an $N \times N$ matrix representation, $u' = D_T u$, with D_T denoting the so-called *Chebyshev differentiation matrix*.

Equipped with these notations, we address the convection equation

$$u_t = a(x) u_x, \quad 0 < a(x) < a_\infty, \quad -1 \le x \le 1, \tag{4.17}$$

subject to the inflow boundary condition

$$u(1,t) = 0, \qquad \forall t > 0.$$

This problem, discretized in terms of the Chebyshev ψdo-spectral method differentiation matrix discussed above, reads

$$\frac{du}{dt} = Lu, \qquad L := \begin{pmatrix} a(x_1) & & \\ & \ddots & \\ & & a(x_N) \end{pmatrix} D_T.$$

Put differently, $u(t) = (u_1, \ldots, u_N)$ is a realization of the N-degree interpolant, $u_N(x,t)$, which is governed by

$$\frac{\partial}{\partial t} u_N(x,t) - a\frac{\partial u_N}{\partial x}(x,t) = \tau(t) T_N(x); \tag{4.18}$$

here, T_N is the N-degree Chebyshev polynomial, and $\tau(t)$ is a free Lagrange multiplier, whose purpose is to enable us to meet the prescribed boundary conditions, $u_N(1,t) = 0$.

To proceed with the stability analysis of the fully discrete RK approximations of (4.18), we borrow from [24] the following two essential inequalities in order to address our requirement of coercivity:

1. $\Re(u^n, Lu^n) \le -\text{Const.}\|(u_N(x,t^n))/1-x\|^2_{w(x)(1-x)}$,
2. $\|Lu^n\|^2 \le a^2 \text{Const.} N^2 \|(u_N(x,t^n))/1-x\|^2_{w(x)(1-x)}$.

The first inequality could be found in [24, eq. (3.37)]. It bounds $\Re(u^n, Lu^n)$ in terms of the *weighted* norm of $u_N/1-x$, where $\|p\|^2_{w(x)(1-x)} := \sum_\nu p^2(x_\nu) w_\nu (1-x_\nu)$ is the weighted norm of the N-degree polynomial p. The second inequality could be found in [24, eq. (3.39)]. For the case of variable coefficients similar estimates hold; consult [24, eq. (6.18)] and [24, eq. (6.19)], respectively, for specific discrete weights.

Combining these two inequalities, we find that L is coercive, (3.2), with $\eta \simeq 1/N^2$, and therefore, stability is guaranteed under the CFL condition (3.17),

$$\Delta t \cdot a_\infty \le \frac{\text{Const.}}{N^2}. \tag{4.19}$$

We summarize our stability results in the following theorem.

THEOREM 4.2. *Consider the hyperbolic problem, $u_t = a(x)u_x$, $0 < a(x) < a_\infty, -1 \le x \le 1$, subject to the boundary condition $u(1,t) = 0$. Then the fully discrete spectral scheme which consists of the spatial Chebyshev ψdo-spectral method together with an s-order RK discretization in time, is stable under the CFL condition* (4.19).

Acknowledgment. We would like to thank Eli Turkel for reading the first version of this review and for his constructive remarks. Doron Levy would also like to thank the UCLA Mathematics Department for its warm hospitality.

REFERENCES

[1] S. ABARBANEL, M. H. CARPENTER, W. S. DON, AND D. GOTTLIEB, *The theoretical accuracy of Runge–Kutta time discretizations for the initial boundary value problem: A careful study of the boundary error*, SIAM J. Sci. Comput., 16 (1995), pp. 1241–1252.

[2] C. A. BERGER, *On the numerical range of powers of an operator*, Abstract 625-152, Notices Amer. Math. Soc., 12 (1965), p. 590.

[3] P. Brenner and V. Thomee, *On rational approximations of semigroups*, SIAM J. Numer. Anal., 16 (1979), pp. 683–694.
[4] J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations. Runge-Kutta and General Linear Methods*, John Wiley, Chichester, 1987.
[5] M. Crouzeix, *On multistep approximations of semigroups in Banach spaces*, J. Comput. and Appl. Math., 20 (1987), pp. 25–35.
[6] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, New York, 1988.
[7] G. Dahlquist and A. Björck, *Numerical Methods*, Prentice–Hall, Englewood Cliffs, NJ, 1974.
[8] G. Dahlquist, H. Mingyou, and R. LeVeque, *On the uniform power-boundedness of a family of matrices and the applications to one-leg and linear multistep methods*, Numer. Math., 42 (1983), pp. 1–13.
[9] J. L. M. Van Dorsselaer and W. Hundsdorfer, *Stability estimates based on numerical ranges with an application to a spectral method*, BIT, 34 (1994), pp. 228–238.
[10] 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*, Acta Numer., 7 (1993), pp. 199–237.
[11] S. R. Foguel, *A counterexample to a problem of Sz.-Nagy*, Proc. Amer. Math. Soc., 15 (1964), pp. 788–790.
[12] B. Fornberg, *On a Fourier method for the integration of hyperbolic equations*, SIAM J. Numer. Anal., 12 (1975), pp. 509–528.
[13] B. Fornberg, *High-order finite differences and the pseudospectral method on staggered grids*, SIAM J. Numer. Anal., 27 (1990), pp. 904–918.
[14] S. Friedland, *A generalization of the Kreiss matrix theorem*, SIAM J. Math. Anal., 12 (1983), pp. 826–832.
[15] D. Funaro, *Polynomial Approximation of Differential Equations*, Springer-Verlag, Berlin, 1992.
[16] M. Giles, *Energy Stability Analysis of Multi-Step Methods on Unstructured Meshes*, Report CFDL-TR-87-1, MIT, Dept. of Aerospace and Astronomy, 1987.
[17] M. Giles, *Stability Analysis of Galerkin/Runge-Kutta Navier-Stokes Discretizations in Unstructured Grids*, AIAA paper 95-1753, 1995, pp. 1244–1257.
[18] M. Goldberg and E. Tadmor, *On the numerical radius and its applications*, Linear Algebra Appl., 42 (1982), pp. 263–284.
[19] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, Baltimore, MD, 1996.
[20] E. Görlich and D. Pontzen, *Alpha well-posedness, alpha stability and the matrix theorems of H. O. Kriess*, Numer. Math., 46 (1985), pp. 131–148.
[21] M. R. Gorelnick, *Stability of families of matrices of unbounded order*, SIAM J. Numer. Anal., 12 (1975), pp. 188–202.
[22] D. Gottlieb, L. Lustman, and E. Tadmor, *Stability analysis of spectral methods for hyperbolic initial-boundary value problems*, SIAM J. Numer. Anal., 24 (1987), pp. 241–256.
[23] D. Gottlieb and A. Orszag, *Numerical analysis of spectral methods: Theory and applications*, CBMS-NSF Regional Conference Series in Applied Mathematics 26, SIAM, Philadelphia, PA, 1977.
[24] D. Gottlieb and E. Tadmor, *The CFL condition for initial-boundary value problems*, Math. Comp., 56 (1991), pp. 565–588.
[25] B. Gustafsson, H.-O. Kreiss, and J. Oliger, *Time Dependent Problems and Difference Methods*, John Wiley, New York, 1995.
[26] B. Gustafsson, H.-O. Kreiss, and A. Sundström, *Stability theory of difference approximations for mixed initial boundary value problems.* II, Math. Comp., 26 (1972), pp. 649–686.
[27] E. Hairer, S. P. Norsett, and G. Wanner, *Solving Ordinary Differential Equations*, 2nd ed., Springer-Verlag, Berlin, 1993.
[28] P. R. Halmos, *A Hilbert Space Problem Book*, 2nd ed., Springer-Verlag, New York, Berlin, 1982.
[29] P. R. Halmos, *On Foguel's answer to Nagy's question*, Proc. Amer. Math. Soc., 15 (1964), pp. 791–793.
[30] D. Henry, *Geometric Theory of Semilinear Parabolic Equations*, Springer-Verlag, Berlin, 1993.
[31] 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.
[32] A. Jameson, H. Schmidt, and E. Turkel, *Numerical Solutions of the Euler Equations by Finite Volume Methods Using Runge-Kutta Time Stepping Schemes*, AIAA paper 81–1259, 1981.

[33] R. Jeltsch and O. Nevanlinna, *Stability of explicit time discretizations for solving initial value problems*, Numer. Math., 37 (1981), pp. 61–91.
[34] R. Jeltsch and O. Nevanlinna, *Stability and accuracy of time discretizations for initial value problems*, Numer. Math., 40 (1982), pp. 245–296.
[35] F. John, *Partial Differential Equations*, 4th ed., Springer-Verlag, New York, 1982.
[36] C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge, UK, 1987.
[37] C. R. Johnson, *Numerical determination of the field of values of a general complex matrix*, SIAM J. Numer. Anal., 15 (1978), pp. 595–602.
[38] T. Kato, *Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1966.
[39] J. F. B. M. Kraaijevanger, *Two Counterexamples Related to the KMT*, report.
[40] H.-O. Kreiss, *Well-posed hyperbolic initial-boundary value problems and stable difference approximations*, in Proc. 3rd Internat. Conference on Hyperbolic Prob., Uppsala, 1990, B. Engquist and B. Gustafsson, eds., Student literature and Chartwell–Bratt, 1991.
[41] H.O. Kreiss and G. Scherer, *Method of lines for hyperbolic differential equations*, SIAM J. Numer. Anal., 29 (1992), pp. 640–646.
[42] H.-O. Kreiss and J. Oliger, *Comparison of accurate methods for the integration of hyperbolic equations*, Tellus, 27 (1972), pp. 199–215.
[43] H.-O. Kreiss and L. Wu, *On the stability of difference approximations for the initial boundary value problem*, Appl. Numer. Math., 12 (1993), pp. 213–227.
[44] G. Laptev, *Conditions for the uniform well-posedness of the Cauchy problem for systems of equations*, Soviet Math. Dokl., 15 (1975), pp. 65–69.
[45] P. D. Lax and B. Wendroff, *Difference schemes for hyperbolic equations with high order of accuracy*, Comm. Pure Appl. Math., 17 (1964), pp. 381–398.
[46] H. W. J. Lenferink and M. N. Spijker, *On a generalization of the resolvent condition in the Kreiss matrix theorem*, Math. Comp., 57 (1991), pp. 211–220.
[47] 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), pp. 221–237.
[48] R. J. LeVeque and L. N. Trefethen, *On the resolvent condition in the Kreiss matrix theorem*, BIT, 24 (1989), pp. 584–591.
[49] C. C. MacDuffee, *The Theory of Matrices*, Chelsea Publishing, Bronx, NY, 1946.
[50] C. McCarthy, *A Strong Resolvent Condition Need Not Imply Power-Boundedness*, preprint, School of Math, Univ. of Minnesota, 1994.
[51] C. A. McCarthy and J. Schwartz, *On the norm of finite Boolean algebra of projections and applications to theorems of Kreiss and Morton*, Comm. Pure Appl. Math., 18 (1965), pp. 191–201.
[52] J. H. Miller, *On power-bounded operators and operators satisfying a resolvent condition*, Numer. Math., 10 (1967), pp. 389–396.
[53] J. H. Miller and G. Strang, *Matrix theorems for partial differential and difference equations*, Math. Scand., 18 (1966), pp. 113–123.
[54] K. W. Morton and S. Schechter, *On the stability of finite difference matrices*, SIAM J. Numer. Anal., 2 (1965), pp. 119–128.
[55] C. Pearcy, *An elementary proof of the power inequality for the numerical radius*, Michigan Math. J., 13 (1966), pp. 289–291.
[56] S. C. Reddy and L. N. Trefethen, *Stability of the method of lines*, Numer. Math., 62 (1992), pp. 235–267.
[57] R. D. Richtmyer, *Principles of Advanced Mathematical Physics*, Vol. I, Springer-Verlag, New York, 1978.
[58] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial-Value Problems*, 2nd ed., John Wiley, New York, 1967.
[59] W. Rudin, *Functional Analysis*, 2nd ed., McGraw-Hill, New York, 1991.
[60] C. W. Shu, *Total variation diminishing time discretizations*, SIAM J. Sci. Comput., 9 (1988), pp. 1073–1084.
[61] C. W. Shu and S. J. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes.* II, J. Comput. Phys., 83 (1989), pp. 32–78.
[62] J. C. Smith, *An inequality for rational functions*, Amer. Math. Monthly, 92 (1985), pp. 740–741.
[63] M. N. Spijker, *Numerical ranges and stability estimates*, Appl. Numer. Math., 13 (1993), pp. 241–249.
[64] M. N. Spijker, *On a conjecture by LeVeque and Trefethen related to the Kreiss matrix theorem*, BIT, 31 (1991), pp. 551–555.
[65] G. Strang and G. J. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.

[66] B. Sz.-Nagy, *Completely continuous operators with uniformly bounded iterates*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 4 (1959), pp. 89–93.
[67] B. Sz.-Nagy and C. Foias, *On certain classes of power-bounded operators in Hilbert space*, Acta Sci. Math., 27 (1966), pp. 17–25.
[68] E. Tadmor, *The equivalence of L_2-stability, the resolvent condition, and strict H-stability*, Linear Algebra Appl., 41 (1981), pp. 151–159.
[69] E. Tadmor, *Numerical viscosity and the entropy condition for conservative difference schemes*, Math. Comp., 43 (1984), pp. 369–381.
[70] E. Tadmor, *Skew-selfadjoint form for systems of conservation laws*, Math. Anal. Appl., 103 (1984), pp. 428–442.
[71] E. Tadmor, *Spectral Methods for Hyperbolic Problems*, Lecture Notes, INRIA, 1994.
[72] E. Tadmor, *Stability analysis of finite-difference, pseudospectral and Fourier-Galerkin approximations for time-dependent problems*, SIAM Rev., 29 (1987), pp. 525–555.
[73] L. N. Trefethen, *Pseudospectra of matrices*, in Numerical Analysis, D.F. Griffiths and G. A. Watson, eds., Longman Press, London, UK, 1991.
[74] E. Turkel, *On the practical use of high-order methods for hyperbolic systems*, J. Comput. Phys., 35 (1980), pp. 319–340.
[75] E. Turkel, *Symmetric hyperbolic difference schemes and matrix problems*, Linear Algebra Appl., 16 (1977), pp. 109–129.
[76] K. Yoshida, *Functional Analysis*, Springer-Verlag, New York, 1968.
[77] B. A. Wade, *Symmetrizable finite-difference operators*, Math. Comp., 54 (1990), pp. 525–543.
[78] E. Wegert and L. N. Trefethen, *The arc length of a rational function on the Riemann sphere or from the Buffon needle problem to the Kreiss matrix theory*, Amer. Math. Monthly, 1991, pp. 132–139.
[79] A. Zygmund, *Trigonometric Series*, Vols. I and II, Cambridge University Press, Cambridge, 1968.