Journal of Computational Physics **156,** 137–180 (1999)
Article ID jcph.1999.6376, available online at http://www.idealibrary.com on IDEAL®

On a Class of Padé Finite Volume Methods

Marcelo H. Kobayashi

Department of Mechanical Engineering, Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisbon, Portugal
E-mail: marcelo@popsrv.ist.utl.pt

Received February 2, 1999; revised September 7, 1999

A class of Padé finite volume methods providing an improved spectral resolution is presented and compared with well-known methods. The formulation is based on the sliding averages of the variables and allows the computation of derivatives of all orders. Using the Fourier analysis, these methods are examined with respect to (i) order of accuracy, (ii) spectral resolution, (iii) boundary conditions, and (iv) stability.

Key Words: finite volume method; high-order interpolation; Padé interpolation; boundary conditions.

CONTENTS

1. INTRODUCTION

The development and widespread use of high-speed digital computers rendered numerical simulations of fluid flow problems a common practice in industry (see, for example, [7, 12]). In the 1990s, numerical simulation has also evolved into a tool for the fundamental study of turbulence, either by direct numerical simulation or by large eddy simulation. A common feature in both applied and fundamental developments is the pursuit of highly accurate methods. The latter provides grid-independent results in coarser grids and sometimes may be the only way to find a satisfactory solution in a "sensible" grid.

Direct numerical simulation of turbulence and large eddy simulation require solution of time and length scales that are various orders of magnitude apart. This has led to

0021-9991/99 $30.00

the development of spectral and pseudo-spectral methods for numerical integration of the Navier–Stokes equations (see, for example, [9, 15, 23]). However, these methods strongly depend on the grid topology and the type of boundary conditions. Moreover, they may lose effectiveness in problems with sharp variations over thin shear or boundary layers (see [2]).

Compact methods, introduced in the 1930s, can ensure spectral-like resolution of short waves. The implementation of such methods is simple and is not restricted either by the grid topology or to some types of boundary conditions. The seminal work of Lele [18], building on previous work on compact finite differences, contributed greatly to the diffusion of these ideas. Spotz [26] proposed a class of compact finite difference schemes that uses the governing equations to approximate the leading term in the truncation error. These schemes can also handle nonuniform grids and have been applied to the stream-function vorticity formulation of the Navier–Stokes equations, transient problems, and so forth. Wilson *et al.* [30] presented and analyzed high-order accurate compact methods for solving unsteady incompressible Navier–Stokes equations for 2D and 3D fluid flow problems. Such methods use the primitive variables together with a Poisson equation for the pressure. In [31], a 4th-order compact method for solving convection–diffusion problems was analyzed with respect to the effect of boundary closure and cell Peclet number. Carpenter *et al.* [4, 5] presented an extensive and illuminating study of explicit and compact high-order schemes with various boundary conditions. Mahesh [21] introduced a class of compact schemes that uses a coupled derivative formulation. It is shown that, for the same order of accuracy, the resulting coupled method displays an improved spectral resolution and that the work required by the new method is comparable to that required by the standard compact method.

Despite their differences these methods are all based on the finite difference approach. Mattiussi [22], using some concepts of the algebraic topology, concluded that for topological equations, that is, the differential or discrete form of the field equations, a discrete representation of the geometry, fields, and operators is preferable. This requires the correct attribution of physical quantities to geometrical objects with appropriate orientation and dimension. Moreover, the fact that the finite-difference methods do not fulfill this attribute renders finite volume and finite element methods more qualified for numerical simulation of field problems. Indeed, finite volume methods are prefered by the computational fluid dynamics (CFD) practitioners in industry: they are as easy to implement as finite difference methods and easier than the finite element method; they are based in the weak formulation of the equations and so require less smoothness of the function; and last, but not least, they are conservative. This last property is essential for methods aiming at compressible fluid flow problems (see [12]). Recently, Gaitonde and Shang [8] proposed and analyzed a class of optimized 4th-order finite volume compact schemes using the reconstruction via the primitive function.

That said, the objective of the present work is to extend and analyze the class of Padé finite volume methods proposed by Gaitonde and Shang. The formulation uses the sliding averages of the variable as the basic variable and the reconstruction via deconvolution. The analysis includes (i) the study of the truncation error of the spatial interpolation with respect to its order of accuracy and its spectral resolution, (ii) the study of boundary conditions and their effects on the quality of the results and stability of the method, and (iii) the stability analysis of explicit and implicit time marching methods. Throughout the presentation comparisons with commonly used methods are provided and discussed.

2. THE STANDARD PADÉ INTERPOLATION

In this section we describe a class of $(2r)$th-order Padé finite volume methods. This class is standard in the sense that it uses a symmetric stencil, and the resulting interpolating polynomial provides the maximal order of accuracy for the given stencil.

The method, which can be used as a general interpolating technique, is aimed at transport equations. So, given the distinct characteristics of transport by convection and diffusion, we analyze them in turn below.

2.1. *Hyperbolic Equations*

The linear advection equation

$$\frac{\partial \phi}{\partial t} + c\frac{\partial \phi}{\partial x} = 0, \tag{1}$$

where c is a real constant and ϕ is a scalar function, is often used as a model for the description of numerical approximation of hyperbolic systems. This is because many hyperbolic methods use the Jacobian matrix computed at a convenient point and interpolate in the characteristic directions. The result is a set of uncoupled linear hyperbolic equations, each of which are of the form of Eq. (1) (see [29]).

Let $\{\sigma_j\}_{j\in\mathbb{Z}}$, with $\sigma_j = [x_j, x_{j+1}]$, $x_j = jh$, $j \in \mathbb{Z}$ be a partition of $\mathbb{R}$, for some grid parameter $h \in \mathbb{R}$. Given a function $f \in L^1$ and $h \in \mathbb{R}$ its sliding, or moving, average $\bar{f} \in L^1$ is defined as (see [11])

$$\begin{aligned}\bar{f} &= \frac{1}{h}\int_{-h/2}^{h/2} f(\hat{x}+y)\,dy \\ &= \frac{1}{h}\chi_h * f,\end{aligned} \tag{2}$$

where $*$ denotes the usual convolution of functions and

$$\chi_h(x) = \begin{cases} 1 & \text{if } |x| < \frac{h}{2}, \\ 0 & \text{otherwise} \end{cases}$$

is the characteristic function of a cell. We use the convention proposed by Silva [6] that a hat over a variable denotes a dummy variable. In the finite volume approach we first integrate Eq. (1) over each control volume, or cell, σ_j of a partition $\{\sigma_j\}_{j\in\mathbb{Z}}$, yielding

$$\left[\frac{d\bar{\phi}}{dt}\right]_{j+1/2} + c\frac{\phi_{j+1}-\phi_j}{h} = 0, \qquad j \in \mathbb{Z}. \tag{3}$$

Note that, so far, we have not introduced any approximation in the model. Also, Eq. (3), which is an evolution equation for the cell averages of ϕ, requires less smoothness than Eq. (1). The variables involved in (3) suggest that in the finite volume method we store and derive the discrete equations for the cell averages. This is the approach used in the present work.

So, we need to obtain a relation between the values of the variables at cell faces and the averaged values at the midpoint of each cell. For this purpose we use the Padé interpolation, which is a solution of the following problem **IP**:

Given a smooth function $\phi \in C^\infty$, and natural numbers $m, n \in \mathbb{N}$, find coefficients a_i, $i = 1, \ldots, m$ and b_i, $i = 1, \ldots, n$, such that

$$\sum_{i=1}^{m} a_i \tau_{-ih}\phi + \phi + \sum_{i=1}^{m} a_i \tau_{ih}\phi = \sum_{i=1}^{n} b_i \tau_{(-i+1/2)h}\bar{\phi} + \sum_{i=1}^{n} b_i \tau_{(i-1/2)h}\bar{\phi} + \mathsf{o}\left(h^{2(m+n)-1}\right). \tag{4}$$

In this expression $\tau_s : C \to C$, with $s \in \mathbb{R}$, is the translation, or shift, operator,

$$(\tau_s f)(x) = f(x - s), \qquad x \in \mathbb{R}$$

for all $f \in C$. Hence, from (2) the **IP** problem is actually a deconvolution to $\mathsf{o}(h^{2(m+n)-1})$.

Solution of **IP** can be easily obtained by fixing a point $x_0 \in \mathbb{R}$ and expanding the function ϕ in a Taylor series around it. We illustrate the procedure for the 4th-order problem, when $m = n = 1$. Expanding the function ϕ around a point $x_0 \in \mathbb{R}$ yields

$$\begin{aligned}
&a\left(\phi_0 - h\phi_0' + \frac{h^2}{2}\phi_0'' - \frac{h^3}{6}\phi_0''' + \frac{h^4}{24}\phi_0^{(4)} - \frac{h^5}{120}\phi_0^{(5)}\right) \\
&\quad + \phi_0 + a\left(\phi_0 + h\phi_0' + \frac{h^2}{2}\phi_0'' + \frac{h^3}{6}\phi_0''' + \frac{h^4}{24}\phi_0^{(4)} + \frac{h^5}{120}\phi_0^{(5)}\right) \\
&= b\left(\phi_0 - \frac{h}{2}\phi_0' + \frac{h^2}{6}\phi_0'' - \frac{h^3}{24}\phi_0''' + \frac{h^4}{120}\phi_0^{(4)} - \frac{h^5}{720}\phi_0^{(5)}\right) \\
&\quad + b\left(\phi_0 + \frac{h}{2}\phi_0' + \frac{h^2}{6}\phi_0'' + \frac{h^3}{24}\phi_0''' + \frac{h^4}{120}\phi_0^{(4)} + \frac{h^5}{720}\phi_0^{(5)}\right) + \mathsf{o}(h^5),
\end{aligned}$$

where the subscript 0 denotes the value of the corresponding function at x_0. So, $a = \frac{1}{4}$, $b = \frac{3}{4}$, and the truncation error is $\mathsf{o}(h^3) = (h^4/120)\phi^{(4)} + \mathsf{o}(h^5)$. In Table I we summarize the coefficients and the leading term of the truncation error for the 2nd-, 4th-, 6th-, 8th-, and 12th-order Padé methods, with $m + 1 \geq n \geq m$. Note that for each order of accuracy the finite volume method requires less smoothness from the function than its finite difference counterpart (see [18]).

In Table II we list, for the 4th, 6th, 8th, and 12th order of accuracy, the coefficients and the leading term of the truncation error for the finite volume Lagrange interpolation scheme. The latter is obtained by taking $m = 0$ in Eq. (4). As can be seen from this table, with the exception of the 2nd-order problem, where the interpolation problem is the same,

TABLE I
Coefficients and Leading Term of the Truncation Error for some Padé Methods

Order	a_1	a_2	a_3	b_1	b_2	b_3	Truncation error
2	—	—	—	$\frac{1}{2}$	—	—	$\frac{h^2}{6} f^{(2)}$
4	$\frac{1}{4}$	—	—	$\frac{3}{4}$	—	—	$-\frac{h^4}{120} f^{(4)}$
6	$\frac{1}{3}$	—	—	$\frac{29}{36}$	$\frac{1}{36}$	—	$\frac{h^6}{1260} f^{(6)}$
8	$\frac{4}{9}$	$\frac{1}{36}$	—	$\frac{185}{216}$	$\frac{25}{216}$	—	$-\frac{h^8}{22680} f^{(8)}$
12	$\frac{9}{16}$	$\frac{9}{100}$	$\frac{1}{400}$	$\frac{1799}{2000}$	$\frac{973}{4000}$	$\frac{49}{4000}$	$-\frac{h^{12}}{4804800} f^{(12)}$

TABLE II
Coefficients and Leading Term of the Truncation Error for Some Lagrange Methods

Order	b_1	b_2	b_3	b_4	b_5	b_6	Truncation error
4	$\frac{7}{12}$	$-\frac{1}{12}$	—	—	—	—	$-\frac{h^4}{30} f^{(4)}$
6	$\frac{37}{60}$	$-\frac{2}{15}$	$\frac{1}{60}$	—	—	—	$\frac{h^6}{140} f^{(6)}$
8	$\frac{533}{840}$	$-\frac{139}{840}$	$\frac{29}{840}$	$-\frac{1}{280}$	—	—	$-\frac{h^8}{630} f^{(8)}$
12	$\frac{18107}{27720}$	$-\frac{5653}{27720}$	$\frac{443}{6930}$	$-\frac{107}{6930}$	$\frac{67}{27720}$	$-\frac{1}{5544}$	$-\frac{h^{12}}{12012} f^{(12)}$

the constant in the leading term of the truncation error for the Padé interpolation is always smaller than the corresponding one for the Lagrange interpolation. The order being the same, this means that in a Log–Log graph of the interpolation error, the asymptotic curves are parallel, but the absolute value of the error is smaller for the Padé interpolation. Moreover, as the order increases, so does the ratio between the constants in the Lagrange and the Padé methods. For the 12th-order method the constant in the Lagrange interpolation is already 400 times larger than the corresponding one for the Padé interpolation.

The order of the truncation error provides the asymptotic rate of convergence of the interpolation toward the interpolated function. Still, it conveys no information about the spectrum of the truncation error. The use of Fourier analysis to characterize the spectrum of the truncation error is thoroughly described in [29]. It is the classical technique for studying and comparing approximation methods. In the context of Padé methods it was used, among others, by Lele [18] and Mahesh [21] to study compact finite difference schemes.

Next, we use the Fourier analysis to study the Padé finite volume method. Let $\phi \in \mathcal{S}'$, where $\mathcal{S}'$ is the space of tempered distributions,[1] which is the dual space to the space of rapidly decreasing smooth functions, or Schwartz space, $\mathcal{S}$; then its Fourier transform exists and we write

$$\Phi = \mathcal{F}(\phi)$$

where the operator $\mathcal{F}: \mathcal{S}' \to \mathcal{S}'$ is the Fourier transform[2] on $\mathcal{S}'$, that is,

$$\begin{aligned} \mathcal{F}: \mathcal{S}' &\to \mathcal{S}' \\ \phi &\mapsto \mathcal{F}(\phi) \end{aligned}$$

with

$$\langle \mathcal{F}(\phi), \psi \rangle = \langle \phi, \mathcal{F}(\psi) \rangle$$

[1] Recall that the space of usual tempered functions Λ belongs to $\mathcal{S}'$; that is, $\Lambda = \{ f \in C: \exists\, p \in \Pi \,\&\, |f(x)| \le p(x), x \in \mathbb{R}\} \subset \mathcal{S}'$, where Π is the space of polynomials of $\mathbb{R}$. In fact, for all $f \in \mathcal{S}'$ we have $f = D^p g$, for some $p \in \mathbb{N}$ and $g \in \Lambda$. A distinguished class of distributions that are not functions in the usual sense, but are in $\mathcal{S}'$, comprises the delta of Dirac and its derivatives $\delta^{(n)} \in \mathcal{S}'$, $n \in \mathbb{N}$.

[2] If $f \in L^p$, $1 \le p \le \infty$ then $f \in \mathcal{S}'$ and $\langle \mathcal{F}(f), \psi \rangle = \int_{\mathbb{R}\times\mathbb{R}} f(x) e^{-i\hat{k}\xi} \psi(\xi) dx\, d\xi$, for all $\psi \in \mathcal{S}$. In particular, if $f, \mathcal{F}(f) \in L^1$, then $f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \mathcal{F}(f)(k) e^{ikx} dk$, for almost all $x \in \mathbb{R}$, with respect to the Lebesgue measure. Also, for $f \in L^2$, we have $\mathcal{F}(f) \in L^2$ but, in general, the integral $\int_{\mathbb{R}} f(x) e^{-ikx} dk$ does not converge. In other words, we do not have an integral representation for the Fourier transform in L^2 (see [25]).

and

$$\mathcal{F}(\psi)(k) = \int_{\mathbb{R}} e^{-ikx}\psi(x)\,dx, \qquad k \in \mathbb{R}$$

for all $\psi \in \mathcal{S}$. We call Φ the spectrum of ϕ.

Given $\phi \in \mathcal{S}'$ we use the convolution in (2) to define $\bar{\phi} \in \mathcal{S}'$, and denote

$$\bar{\Phi} = \mathcal{F}(\bar{\phi})$$

Now, from its definition and the convolution theorem (see [25, Theorem XV, p. 268]) for the Fourier transform it follows that

$$\bar{\Phi} = \frac{2}{\hat{k}h}\sin\left(\frac{\hat{k}h}{2}\right)\Phi. \tag{5}$$

Applying the Fourier transform to Eq. (4) and taking into account Eq. (5) yields

$$E = E_r\Phi,$$

where $E \in \mathcal{S}'$ is the spectrum of the truncation error and $E_r \in \mathcal{S}'$ is the relative spectral error

$$E_r = \left\{1 + 2\left[\sum_{j=1}^{m} a_j \cos(j\hat{k}h) - \frac{2}{\hat{k}h}\sin\left(\frac{\hat{k}h}{2}\right)\sum_{j=1}^{n} b_j \cos\left(\left(j - \frac{1}{2}\right)\hat{k}h\right)\right]\right\}. \tag{6}$$

Given the fact that $\mathcal{F}(\delta) = 1$, the relative spectral error is the spectrum of the error associated with the Dirac distribution. Figure 1 shows the relative spectral error E_r for the 4th-, 6th-, 8th-, and 12th-order Padé and Lagrange finite volume methods as a function of the cell wave number $\tilde{k} : 2\pi\tilde{k} = kh$. This figure shows that increasing the order of the Padé method decreases its spectral error at all wave numbers in the range considered. This effect is

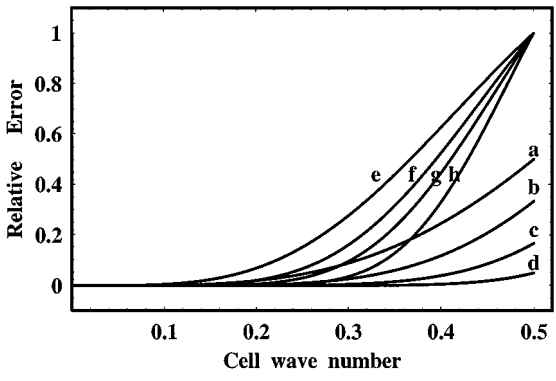

FIG. 1. Plot of relative spectral error versus grid wave number for advection using the Padé method: (a) 4th order, (b) 6th order, (c) 8th order, (d) 12th order; and the Lagrange method: (e) 4th order, (f) 6th order, (g) 8th order, (h) 12th order.

more pronounced for wave numbers closer to half grid width. This figure also shows that while the truncation error of the Lagrange interpolation scheme is of the same order as the corresponding Padé one, its behaviour in the spectral error is worse for all wave numbers greater than zero (constant functions). This discrepancy increases with the wave number and the order of the method.

Inspection of Eq. (6) provides a heuristic argument to explain the improvement in the Padé interpolation relative to the Lagrange one. Any usual interpolation method has a term similar to the second one within brackets. This term tends to zero as $\tilde{k} \to 1/2$. However, any Padé method of 4th order or higher implicitly uses a full interpolating stencil. This gives rise to the first term within brackets in Eq. (6), which is a truncated expansion of the spectrum of the variable ϕ. Note that in Eq. (6) it appears only in the cosine components. This results from the symmetry of the stencil, which makes Eq. (4) exact for all odd smooth functions.

Next, we analyze the nature of the error associated with the proposed Padé finite volume method with respect to both its amplitude and phase.

We start by writing the linear advection equation as

$$\frac{d\bar{\phi}}{dt} + c\frac{\tau_{-h/2}\phi - \tau_{h/2}\phi}{h} = 0, \tag{7}$$

where $\phi \in C(\mathbb{R}; \mathcal{S}')$. Then, for each $t \in \mathbb{R}$ we apply the Fourier transform to (7), yielding

$$\frac{d\bar{\Phi}}{dt} + \frac{2ic}{h}\sin\left(\frac{\hat{k}h}{2}\right)\Phi = 0. \tag{8}$$

Using Eq. (5) results in

$$\frac{d\bar{\Phi}}{dt} + ic\hat{k}\bar{\Phi} = 0.$$

Solution of the previous equation provides the law of propagation

$$\bar{\Phi}(t) = \bar{\Phi}_0 e^{ic\hat{k}t}, \qquad t \in \mathbb{R}, \tag{9}$$

where $\bar{\Phi}_0 = \bar{\Phi}(0)$. This means that the evolution of the sliding average corresponds to a phase shift and no change in the amplitude. This is expected, as the solution of the linear advection problem corresponds to a shift in the profile and the averaging commutes with the shift.

On the other hand, the Padé interpolation gives rise to the following relation between the spectra Φ and $\bar{\Phi}$:

$$\Phi\left(1 + 2\sum_{j=1}^{m} a_j \cos(j\hat{k}h)\right) = 2\bar{\Phi}\sum_{j=1}^{n} b_j \cos\left(\left(j - \frac{1}{2}\right)\hat{k}h\right). \tag{10}$$

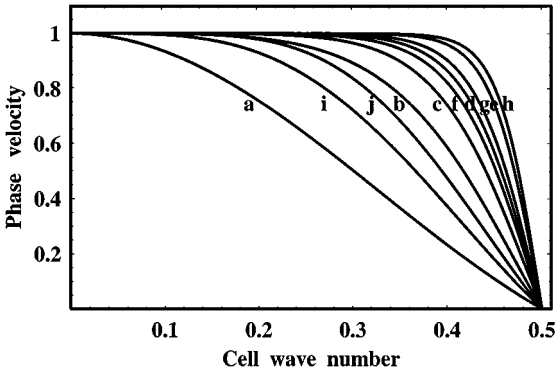

FIG. 2. Plot of phase speed versus grid wave number for advection using the Padé method: (a) 2nd order, (b) 4th order, (c) 6th order, (d) 8th order, (e) 12th order, (f) CD 6th order, (g) CD 8th order, (h) CD 12th order; and the Lagrange method: (i) 4th order, (j) 6th order.

Substituting Eq. (10) into Eq. (8) and solving for $\bar{\Phi}$ yields the same law of propagation as for the exact profile with $\mathfrak{c}$ in the place of c, where

$$\mathfrak{c} = \frac{4}{kh}\sin\left(\frac{kh}{2}\right)\frac{\sum_{j=1}^{n} b_j \cos\left(\left(j-\frac{1}{2}\right)kh\right)}{1+2\sum_{j=1}^{m} a_j \cos(jkh)}c, \qquad k \in \mathbb{R}. \tag{11}$$

Figure 2 displays the phase ratio $\mathfrak{c}/c$ as a function of the grid wave number $\tilde{k}$ for the 2nd-, 4th-, 6th-, 8th-, and 12th-order Padé interpolation method and the 4th-, and 6th-order Lagrange interpolation methods. This figure shows that, for small grid wave numbers, all methods perform well. As the grid wave number increases, all methods begin displaying a retardation error in the phase. It is evident that compared to the standard Lagrange methods the Padé method stay close to the exact phase velocity over a wider range of cell wave numbers. In fact, the 4th-order Padé method shows a better spectral resolution than the 6th-order Lagrange method. Again, this is a consequence of the implicit full stencil used in the Padé interpolation.

To quantify the error in the phase speed Lele [18] and Mahesh [21] used the notions of resolving efficiency and percentage error as a function of the number of points per wave (PPW). The latter is related to the cell wave number as follows: $\mathrm{PPW} = 1/\tilde{k}$. These are measures of the ability of each method to resolve a particular range of phase speed. They show that compact methods are more efficient than the standard Lagrange methods for all orders of accuracy.

In the present work, we use the accumulated relative spectral error for half wave resolution ($\mathrm{ARSE}_{1/2}$). It is a direct measure of the spectral error provided by the $L^1([0, \frac{1}{2}])$-norm of the relative spectral error distribution. That is,

$$\mathrm{ARSE}_{1/2} = \int_0^{1/2} |E_r|(\tilde{k})\, d\tilde{k}.$$

This measure is similar to the one used by Gaitonde and Shang [8]. However, they measured the norm of the error in the phase velocity, whereas here we measure the error in the truncation error of the interpolation.

TABLE III
$ARSE_{1/2}$ for the Padé and Lagrange Methods Shown in Fig. 1

Method	Order	$ARSE_{1/2}$
Padé	4	5.788×10^{-2}
Padé	6	2.897×10^{-2}
Padé	8	1.109×10^{-2}
Padé	12	2.341×10^{-3}
Lagrange	4	1.446×10^{-1}
Lagrange	6	1.167×10^{-1}
Lagrange	8	1.001×10^{-1}
Lagrange	12	8.055×10^{-2}

The values of the $ARSE_{1/2}$ for the Padé and Lagrange methods shown in Fig. 1 are tabulated in Table III. This shows that increasing the order of accuracy of the Padé method and Lagrange methods causes them to stay close to the exact solution over a progressively larger range of the spectrum. The values of $ARSE_{1/2}$ for the Padé methods are always smaller than the corresponding values for the Lagrange methods. Moreover, as the order increases, the value of $ARSE_{1/2}$ drops faster for the Padé method. For example, the ratio of the value of the $ARSE_{1/2}$ for the 8th-order to the 12th-order Padé method is 4.7. The corresponding figure for the Lagrange method is 1.2.

If we apply the inverse of the Fourier transform to Eq. (9) we recover the analytical solution of Eq. (1), that is,

$$\bar{\phi}(t) = \tau_{ct}\bar{\phi}_0,$$

where $\bar{\phi}_0 = \bar{\phi}(0)$ is the initial profile. The reason for its simple form lies in the fact that all "waves" compounding ϕ travel at the same speed c.

Now, the nonlinear relation between $\mathfrak{c}$ and k in Eq. (11) expresses the fact that with the Padé interpolation the waves travel at different speed, depending on the wave number $\tilde{k}$. The waves are close to the actual speed at small grid wave numbers and progressively slow down as $\tilde{k} \to \frac{1}{2}$. Accordingly, applying the inverse of the Fourier transform to law of propagation of the approximation Padé yields an "averaging" of the various dispersed modes of the initial profile.

In [29] it is shown that these errors form packets which travel at the group velocity, also called energy velocity. The authors point out that certain spurious numerical solutions of the semidiscretizations are affected with a positive phase speed but a negative group velocity. This may be relevant in simulations of turbulence, because then the small eddies may propagate in the opposite direction to the true solution.

Next, we compute the group velocity for several Padé and Lagrange methods. We start by recalling the definition of the group velocity $\mathcal{V}$:

$$\mathcal{V} = \frac{d}{dk}(\hat{k}\mathfrak{c}).$$

Figure 3 shows the group velocity $\mathcal{V}$ as a function of the cell wave number for the 2nd-, 4th-, 6th-, 8th-, and 12th-order Padé interpolation methods, and the 4th-, and 6th-order

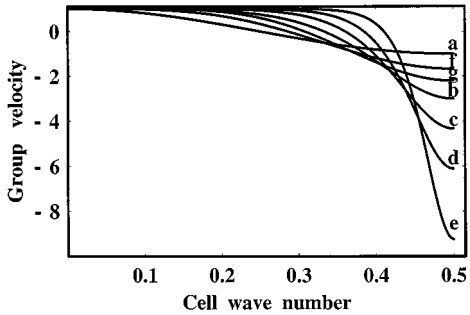

FIG. 3. Plot of group velocity versus grid wave number for the Padé method: (a) 2nd order, (b) 4th order, (c) 6th order, (d) 8th order, (e) 12th order; and the Lagrange method: (f) 4th order, (g) 6th order.

Lagrange interpolation methods. The exact solution for a unitary phase speed is $\mathcal{V} = 1$. As can be seen from this figure, as the cell wave number increases the group velocity starts deviates from the exact solution. Indeed, all methods display a region with negative group velocity. The size of this region and the minimum value of $\mathcal{V}$ depend on the accuracy, and also on the method. Hence, increasing the accuracy decreases the region and the minimum. However, the main effect comes from the method. In fact, the 4th-order Padé method has a smaller region with negative group velocity than the 6th-order Lagrange method. This means that energy is transported with a better group velocity for the Padé method. Moreover, as the order of accuracy increases, the range of wave number with positive group velocity becomes wider, and the oscillations with typically $2h$ propagate at a higher speed in the negative direction.

In [29] the authors also used a wave analysis to study the propagation of the spurious oscillations. We show that the 4th-order Padé method is equivalent to a three-point semi discretization of the advection equation. Let $\mathcal{A}$ denote the Banach algebra of continuous linear operators of L^1, that is, $\mathcal{A} = \mathrm{L}(L^1)$. Clearly, $(\tau_{-h} + \tau_h)/4 + \mathrm{id} \in \mathcal{A}$, for all $h \in \mathbb{R}$. Now, we prove that this operator has an inverse in $\mathcal{A}$.

To see that it is one-to-one, let $\psi \in L^1$ be a function such that

$$\left(\frac{\tau_{-h} + \tau_h}{4} + \mathrm{id}\right)(\psi) = 0.$$

Then applying the Fourier transform to this equation yields

$$\left(1 + \frac{1}{2}\cos(\hat{k}h)\right)\Psi = 0.$$

So, it is one-to-one. The proof that it is onto is analogous, and existence of the inverse follows from the open mapping theorem (see, for example, [17]).

Consider Eq. (7) for $\phi \in L^1$. Then, using the 4th-order Padé method we obtain

$$\frac{d\bar{\phi}}{dt} + \frac{3c}{4}\left(\frac{\tau_{-h/2} - \tau_{h/2}}{h}\right)\left(\frac{\tau_{-h} + \tau_h}{4} + \mathrm{id}\right)^{-1}\left(\tau_{-h/2} + \tau_{h/2}\right)\bar{\phi} = 0.$$

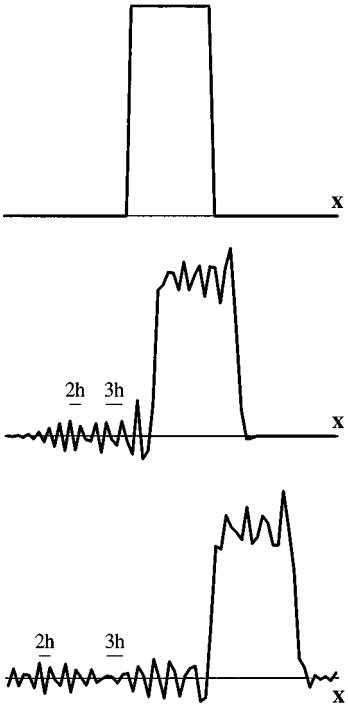

FIG. 4. Plot of distribution: $t = 0$ (top), $t = \frac{5h}{c}$ (middle), $t = \frac{16h}{c}$ (bottom).

The operators in the second term of the LHS are all Toeplitz and so commute. Hence,

$$\left(\frac{\tau_{-h} + \tau_h}{3} + \frac{2}{3}\mathrm{id}\right)\frac{d\bar{\phi}}{dt} + c\left(\frac{\tau_{-h} - \tau_h}{2h}\right)\bar{\phi} = 0,$$

as desired. In particular we see that the "oscillatory" part travels at the speed $-3c$, in agreement with the group velocity $\mathcal{V}(\frac{1}{2})$ for the 4th-order Padé method shown in Fig. 3.

Figure 4 illustrates the effect of the discretization in the error propagation. The problem involves the time evolution of a periodic square profile

$$\phi_0(x) = \begin{cases} 1 & \text{if } |x_{01} - 1/2| < 1/8, \\ 0 & \text{otherwise,} \end{cases}$$

where $x_{01} = x - [x]$, with $[x]$ the integer part of x. The 4th-order accurate Padé finite volume method is used, with periodic boundary conditions, in a grid comprising 64 control volumes. Time marching is implemented with the 4th-order Runge–Kutta (RK) method, using a value of $\frac{c\Delta t}{h} = 0.1$ for the CFL number. This level of CFL guarantees that the error is predominantly due to spatial discretization. The numerical solution shown in Fig. 4 is as predicted by the group velocity (note that $\mathcal{V}(1/3) = 0$).

Remark. In the linear hyperbolic constant speed problem considered, the conservative and nonconservative equations give rise to equivalent numerical approximations using the finite volume and the finite difference methods. However, in the general case, when the

speed is a function of the space coordinate, the conservative and nonconservative formulations, although equivalent from the analytical point of view, yield different numerical approximations. To investigate the influence of both forms of the equation in the numerical approach, we consider the hyperbolic equation

$$\frac{\partial \phi}{\partial t} + \frac{\partial x\phi}{\partial x} = 0 \tag{12}$$

or, in the non-conservative form,

$$\frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial x} + \phi = 0. \tag{13}$$

In the finite volume method, as usual, we consider Eq. (12) in integral form:

$$\frac{\partial \bar{\phi}}{\partial t} + \frac{\tau_{-h/2} - \tau_{h/2}}{h} x\phi = 0. \tag{14}$$

The analytic solution of Eq. (12) is readily obtained using the method of characteristics (see, for example, [13]). It is given by

$$\phi(t) = e^{-t}\phi_0(\hat{x}e^{-t}), \qquad t \geq 0.$$

So, the characteristics are exponential curves in the (t, x)-plane and on these curves the function decreases exponentially. That is, the initial distribution spreads and fades.

Now, applying the Fourier transform to Eq. (13) we obtain

$$\frac{\partial \Phi}{\partial t} - k\frac{\partial \Phi}{\partial k} = 0. \tag{15}$$

Hence, the spectrum is conserved on curves $k = k_0 e^{-t}$. Analogously, we have for the spectrum of the sliding averages

$$\frac{\partial \bar{\Phi}}{\partial t} - k\frac{\partial \bar{\Phi}}{\partial k} = \left(1 - \frac{kh}{2}\cot\left(\frac{kh}{2}\right)\right)\bar{\Phi}. \tag{16}$$

The characteristics are the same as for the Φ equation. However, $\bar{\Phi}$ changes over these curves as

$$\frac{dz}{dt} = 1 - \frac{kh}{2}\cot\left(\frac{kh}{2}\right)$$

and $\frac{dz}{dt} \geq 0$, $0 \leq \tilde{k} \leq \frac{1}{2}$; that is, the spectrum $\bar{\Phi}$ increases over the characteristics. Therefore, the spectra Φ and $\bar{\Phi}$ have different evolution laws, and it is expected that their numerical approximations have different characteristics too.

Indeed, using the Padé finite difference yields an equation similar to (15) with κ_{fd} in the place of k, where

$$\kappa_{\mathrm{fd}} = \frac{3\sin(2\tilde{k}\pi)}{2\tilde{k}\pi(2 + \cos(2\tilde{k}\pi))}k$$

and a change over the characteristics of

$$\dot{z}_{\mathrm{fd}} = -1 + \frac{3 + 6\cos(2\tilde{k}\pi)}{(2 + \cos(2\tilde{k}\pi))^2}.$$

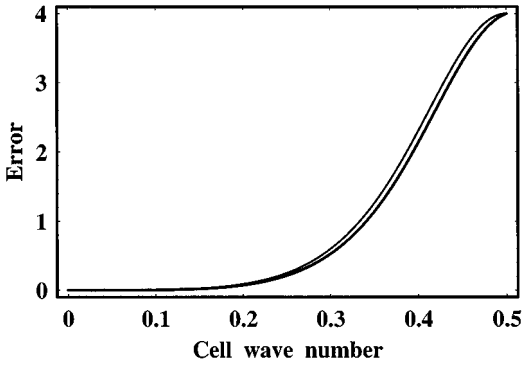

FIG. 5. Plot of the error in the change rate for Padé finite volume (thicker line) and Padé finite difference (thinner line).

The Padé finite volume method yields

$$\kappa_{\mathrm{fv}} = \kappa_{\mathrm{fd}}$$

and a change over the characteristics of

$$\dot{z}_{\mathrm{fv}} = \frac{3\cos(2\tilde{k}\pi)\,\sin^2(\tilde{k}\pi)}{(2+\cos(2\tilde{k}\pi))^2}.$$

The error in the rate of change is shown in Fig. 5. Both methods can resolve wave components with small cell wave length, but start to deviate from the correct decay as $\tilde{k} \to \frac{1}{2}$. However, it is interesting to note that the Padé finite volume method yields a smaller error in the decay, as compared with the finite difference method.

Remark. The leading term of the truncation error of an $2r$th-order Padé interpolation method, with $r = m + n$, is proportional to the $2r$th derivative of the (smooth) function. So, it is exact for all polynomials of degree smaller than $2r$. We can use this fact to write a general matrix equation for the coefficients a_i, $i = 1, \ldots, m$, b_i, $i = 1, \ldots, n$. Given the linearity of Eq. (4), we can work with a basis of the vector space Π_{2r-1} of polynomials of degree smaller than $2r$. The natural basis being the monomials x^j, $j = 1, \ldots, 2r-1$, substituting each basic element x^j, $j = 1, \ldots, 2r-1$ into Eq. (4) and considering the origin yields

$$\left[\sum_{j=1}^{m} a_j j^l - \sum_{j=1}^{n} \frac{b_j}{l+1}(j^{l+1} - (j-1)^{l+1})\right](1+(-1)^l) = 0 \tag{17}$$

for all $l = 1, \ldots, 2r-1$. This is a $2r \times r$ system of linear equations. However, due to the symmetry of the stencil and the equality of symmetric coefficients, Eq. (17) is trivial for l odd. The latter system of equations is equivalent to the matrix equation

$$[A \quad B]\begin{bmatrix} a \\ b \end{bmatrix} = d, \tag{18}$$

where the block matrices are defined as

$$A = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ 1 & 4 & \cdots & (m-1)^2 & m^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 2^{2r-4} & \cdots & (m-1)^{2r-4} & m^{2r-4} \\ 1 & 2^{2r-2} & \cdots & (m-1)^{2r-2} & m^{2r-2} \end{bmatrix}$$

$$B = \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \\ \frac{1}{3} & \frac{7}{3} & \cdots & \frac{(n-1)^3-(n-2)^3}{3} & \frac{n^3-(n-1)^3}{3} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{1}{2r-3} & \frac{1}{2r-3}(2^{2r-3}-1) & \cdots & \frac{(n-1)^{2r-3}-(n-2)^{2r-3}}{2r-3} & \frac{n^{2r-3}-(n-1)^{2r-3}}{2r-3} \\ \frac{1}{2r-1} & \frac{1}{2r-1}(2^{2r-1}-1) & \cdots & \frac{(n-1)^{2r-1}-(n-2)^{2r-1}}{2r-1} & \frac{n^{2r-1}-(n-1)^{2r-1}}{2r-1} \end{bmatrix}$$

$$d = \begin{bmatrix} -\frac{1}{2} \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Next we prove existence and uniqueness of a solution of (18). The proof is divided into two steps. In the first, we prove that the finite difference Padé interpolation problem can be written as a Hermite interpolation problem. Then, in the second, we prove that **IP** is equivalent to a Padé finite difference problem for the primitive of ϕ.

The finite difference Padé interpolation problem can be stated as follows:

Given a smooth function $\phi \in C^\infty$, and natural numbers $m, n \in \mathbb{N}$ find coefficients a_i, $i = 1, \ldots, m$ and b_i, $i = 1, \ldots, n$, such that

$$\sum_{i=1}^{m} a_i \tau_{-ih}\phi' + \phi' + \sum_{i=1}^{m} a_i \tau_{ih}\phi' = \frac{1}{h}\left(\sum_{i=1}^{n} b_i \tau_{-ih}\phi - \sum_{i=1}^{n} b_i \tau_{ih}\phi\right) + \mathrm{o}\big(h^{2(m+n)-1}\big). \quad (19)$$

Consider the real numbers $x_i, \phi_i^{(k)},\ k = 0, \ldots, n_i - 1,\ i = 0, \ldots, m_0$ with

$$x_0 < x_1 < \cdots < x_{m_0}.$$

The Hermite interpolation problem **IH** for these data consists of determining a polynomial $P \in \Pi_{n_0}$ where $n_0 + 1 = \sum_{i=0}^{m_0} n_i$, which satisfies the interpolation conditions

$$P^{(k)}(x_i) = \phi_i^{(k)}, \qquad k = 0, \ldots, n_{i-1},\ i = 0, \ldots, m_0.$$

There is a theorem (see, for example, [27]) which gives existence and uniqueness of a solution of **IH**. It is given by

$$P(x) = \sum_{i=0}^{m_0} \sum_{k=0}^{n_i - 1} \phi_i^{(k)} L_{ik}(x), \quad (20)$$

where $L_{ik} \in \Pi_{n_0}$ are generalized Lagrange polynomials. They are defined as follows: starting with the polynomials

$$l_{ik}(x) = \frac{(x - x_i)^k}{k!} \prod_{\substack{j=0 \\ j \neq i}}^{m_0} \left(\frac{x - x_j}{x_i - x_j} \right)^{n_j}, \qquad 0 \leq i \leq m_0, 0 \leq k < n_i,$$

put

$$L_{i,n_i-1}(x) = l_{i,n_i-1}(x), \qquad i = 0, \ldots, m_0$$

and recursively for $k = n_i - 2, \ldots, 0$,

$$L_{ik}(x) = l_{ik} - \sum_{\nu=k+1}^{n_i-1} l_{ik}^{\nu}(x_i) L_{i\nu}(x).$$

Consider the points $x_{\pm i} = \pm i,\ i = 1, \ldots, n$.

Claim.

$$a_i = -L'_{i1}(0), \qquad i = 1, \ldots, m$$

and

$$b_i = L'_{i0}(0), \qquad i = 1, \ldots, n.$$

To prove this we note that $n_0 = 2r - 1$, with $r = n + m$. Then, Eq. (19) is exact for the polynomial P given by Eq. (20). Substituting P in Eq. (19) and taking into account the uniqueness of the Hermite polynomial proves the claim.

For the second part of the proof we first observe that given $\phi \in L^1$ we have

$$\bar{\phi} = \frac{\tau_{-h/2}\Theta - \tau_{h/2}\Theta}{h},$$

where $\Theta \in C$ is the primitive of ϕ,

$$\Theta(x) = \int_{x_0}^{x} \phi(u)\, du, \qquad x \in \mathbb{R},$$

for some $x_0 \in \mathbb{R}$.

We conclude that the **IP** is equivalent to the following problem:

Given a smooth function $\phi \in C^\infty$, and natural numbers $m, n \in \mathbb{N}$, find coefficients a_i, $i = 1, \ldots, m$, and $\tilde{b}_i$, $i = 1, \ldots, n$, such that

$$\sum_{i=1}^{m} a_i \tau_{ih}\Theta' + \Theta' + \sum_{i=1}^{m} a_i \tau_{ih}\Theta' = \frac{1}{h}\left(\sum_{i=1}^{n} \tilde{b}_i \tau_{-ih}\Theta + \sum_{i=1}^{n} \tilde{b}_i \tau_{ih}\Theta \right) + \mathrm{o}\left(h^{2(m+n)-1}\right). \quad (21)$$

Using the result of the claim we compute a_i, $i = 1, \ldots, m$, and $\tilde{b}_i$, $i = 1, \ldots, n$. Then an easy computation shows that

$$\begin{aligned} b_n &= \tilde{b}_n \\ b_{n-1} - b_n &= \tilde{b}_{n-1} \\ &\vdots \\ b_1 - b_2 &= \tilde{b}_1. \end{aligned} \tag{22}$$

Equations (21), (22) represent the reconstruction via a primitive function, which is used in [8]. Because of the stencil selection in the ENO approach (see [11]) the reconstruction via the deconvolution and the primitive function yield different methods, the deconvolution being more accurate in some cases. So, it is expected that the same may occur with the Padé method, if an ENO-like stencil selection is used.

Remark. We have performed the Fourier analysis in the space $\mathcal{S}'(\mathbb{R})$. Another possibility is to work with the Fourier transform on $\mathcal{S}'(\mathbb{T}^1)$, where $\mathbb{T}^1 = \mathbb{R}/\mathbb{Z}$ is the 1-torus.[3] That corresponds to work with periodic distributions on $\mathbb{R}$. Then, to each distribution $\phi \in \mathcal{S}'(\mathbb{T}^1)$ we may associate its Fourier series (see [25]):

$$\phi = \sum_{k \in \mathbb{Z}} \Phi(k)(\phi) e^{ik\hat{x}}.$$

Formally, the analysis proceeds as above, with the distinction that now the distributions have a "discrete" spectrum; that is, we should use $k \in \mathbb{Z}$ instead of $k \in \mathbb{R}$ in the expressions like (11).

As remarked in [29], the semidiscrete equation, Eq. (3), can be studied using a band-limited function (see also [18]). This fits into the present framework, since it corresponds to periodic functions with a finite discrete spectrum.

2.2. *Parabolic Diffusion Equations*

Similarly to the previous subsection, we use the linear diffusion equation

$$\frac{\partial \phi}{\partial t} = \nu \frac{\partial^2 \phi}{\partial x^2} \tag{23}$$

as a model for the description of the numerical approximation. Again, in the finite volume approach we first integrate Eq. (23) over each control volume σ_j of a partition $\{\sigma_j\}_{j \in \mathbb{Z}}$, yielding

$$\left[\frac{d\bar{\phi}}{dt}\right]_{j+1/2} = \nu \frac{\phi'_{j+1} - \phi'_{j-1}}{h}, \qquad j \in \mathbb{Z}. \tag{24}$$

The corresponding **IP** can be stated as follows:

[3] Clearly $\mathbb{T}^1$ is diffeomorphic to S^1, where S^1 denotes the unit circle. However, in higher dimensions the generalization is $\mathbb{T}^n$, not S^n.

TABLE IV
Coefficients and Leading Term of the Truncation Error for Some Padé Methods

Order	a_1	a_2	a_3	b_1	b_2	b_3	Truncation error
4	$\frac{1}{10}$	—	—	$\frac{6}{5}$	—	—	$-\frac{h^4}{200} f^{(5)}$
6	$\frac{2}{11}$	—	—	$\frac{51}{44}$	$\frac{3}{44}$	—	$\frac{23h^6}{55440} f^{(7)}$
8	$\frac{344}{1179}$	$\frac{23}{2558}$	—	$\frac{265}{262}$	$\frac{155}{786}$	—	$-\frac{79h^8}{2971080} f^{(9)}$
12	$\frac{329913}{725308}$	$\frac{18387}{362654}$	$\frac{619}{725308}$	$\frac{813155}{1087962}$	$\frac{835345}{2175924}$	$\frac{49483}{2175924}$	$-\frac{38223h^{12}}{290413323200} f^{(13)}$

Given a smooth function $\phi \in C^\infty$, and natural numbers $m, n \in \mathbb{N}$, find coefficients a_i, $i = 1, \ldots, m$, and b_i, $i = 1, \ldots, n$, such that

$$\sum_{i=1}^{m} a_i \tau_{-ih}\phi' + \phi' + \sum_{i=1}^{m} a_i \tau_{ih}\phi' = \frac{1}{h}\left(\sum_{i=1}^{n} b_i \tau_{(-i+1/2)h}\bar{\phi} - \sum_{i=1}^{n} b_i \tau_{(i-1/2)h}\bar{\phi}\right) + \mathrm{o}\left(h^{2(m+n)-1}\right). \tag{25}$$

The solution of **IP** above can be computed as above using either the Taylor series expansion or the second derivative of corresponding Hermite interpolating polynomial. Table IV lists the coefficients and the leading term of the truncation error for the 4th-, 6th-, 8th-, and 12th-order Padé interpolations with $m + 1 \geq n \geq m$.

The Fourier analysis for the parabolic equation proceeds as in the analysis of hyperbolic equation. The exact relation yields

$$\frac{d\bar{\Phi}}{dt} = -\nu k^2 \bar{\Phi},$$

so,

$$\bar{\Phi}(t) = \bar{\Phi}_0 e^{\nu \hat{k}^2 t}, \qquad t \in \mathbb{R}. \tag{26}$$

This means that the spectrum of the sliding average decays at an exponential rate, which is proportional to the square of the wave number. The Padé interpolation provides an analogous law of evolution with a decay ratio $\nu_h k^2$, where

$$\left(\frac{\nu_h}{\nu}\right)^{1/2} = \frac{2}{kh}\sqrt{\sin\left(\frac{kh}{2}\right)\frac{\sum_{j=1}^{n} b_j \sin\left(\left(j-\frac{1}{2}\right)kh\right)}{1+2\sum_{j=1}^{m} a_j \cos(jkh)}}, \qquad k \in \mathbb{R} \tag{27}$$

Figure 6 shows $(\nu_h/\nu)^{1/2}$ as a function of the grid wave number for the 2nd-, 4th-, 6th-, 8th-, and 12th-order methods. As expected, increasing the order of the method improves the spectral resolution of the method.

2.3. *Transport Equations*

In this subsection we consider the linear transport equation

$$\frac{\partial \phi}{\partial t} + c\frac{\partial \phi}{\partial x} = \nu \frac{\partial^2 \phi}{\partial x^2}. \tag{28}$$

Integration of Eq. (28) over each control volume σ_j of a partition $\{\sigma_j\}_{j\in\mathbb{Z}}$ yields

$$\left[\frac{d\bar{\phi}}{dt}\right]_{j+1/2} + c\frac{\phi_{j+1} - \phi_{j-1}}{h} = \nu\frac{\phi'_{j+1} - \phi'_{j-1}}{h}, \qquad j \in \mathbb{Z}. \tag{29}$$

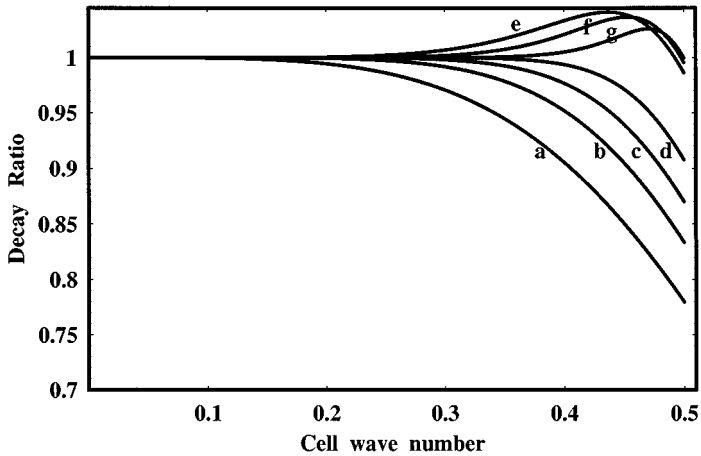

FIG. 6. Plot of decay ratio versus grid wave number for diffusion using the Padé method: (a) 4th order, (b) 6th order, (c) 8th order, (d) 12th order, (e) CD 6th order, (f) CD 8th order, (g) CD 12th order.

One obvious way to solve Eq. (29) is to apply the Padé interpolation, Eqs. (4) and (25), to its convection and diffusion components, respectively. However, Mahesh in a recent paper [21] proposed a coupled-derivative (CD) approach to the finite difference compact scheme, in which both derivatives are computed simultaneously. The resulting method showed a better spectral resolution than the usual uncoupled method.

In the CD approach the **IP** problem for the convection term can be stated as follows:

Given a smooth function $\phi \in C^\infty$, *and natural numbers* $m, n, o \in \mathbb{N}$, *find coefficients* a_i, $i = 1, \ldots, m$, b_i, $i = 1, \ldots, n$, *and* c_i, $i = 1, \ldots, o$, *such that*

$$\sum_{i=1}^{m} a_i \tau_{-ih}\phi + \phi + \sum_{i=1}^{m} a_i \tau_{ih}\phi = \sum_{i=1}^{n} b_i \tau_{(-i+1/2)h}\bar{\phi} + \sum_{i=1}^{n} b_i \tau_{(i-1/2)h}\bar{\phi} + h\left(\sum_{i=1}^{o} c_i \tau_{-ih}\phi' - \sum_{i=1}^{o} c_i \tau_{ih}\phi'\right) + \mathrm{o}\left(h^{2(m+n+o)-1}\right). \tag{30}$$

For the diffusion term we have:

Given a smooth function $\phi \in C^\infty$, *and natural numbers* $m, n, o \in \mathbb{N}$ *find coefficients* d_i, $i = 1, \ldots, m$, e_i, $i = 1, \ldots, n$, *and* f_i, $i = 1, \ldots, o$, *such that*

$$\sum_{i=1}^{m} d_i \tau_{-ih}\phi' + \phi' + \sum_{i=1}^{m} d_i \tau_{ih}\phi' = \frac{1}{h}\left(\sum_{i=1}^{n} e_i \tau_{(-i+1/2)h}\bar{\phi} - \sum_{i=1}^{n} e_i \tau_{(i-1/2)h}\bar{\phi} + \sum_{i=1}^{o} f_i \tau_{-ih}\phi - \sum_{i=1}^{o} f_i \tau_{ih}\phi\right) + \mathrm{o}\left(h^{2(m+n+o)-1}\right). \tag{31}$$

In Tables V and VI we list the coefficients and the leading term in the truncation order for the 6th-, 8th-, and 12th-order Padé interpolations with $n - 1 \le m = o \le n$. A comparison of

TABLE V
Coefficients and Leading Term of the Truncation Error for Some CD Padé Methods: Convection

Order	a_1	a_2	b_1	b_2	c_1	c_2	Truncation error
6	$\frac{7}{16}$	—	$\frac{15}{16}$	—	$\frac{1}{16}$	—	$\frac{h^6}{5040} f^{(6)}$
8	$\frac{17}{36}$	—	$\frac{53}{54}$	$-\frac{1}{108}$	$\frac{1}{12}$	—	$-\frac{h^8}{90720} f^{(8)}$
12	$\frac{4}{9}$	$-\frac{23}{432}$	$\frac{2755}{2592}$	$-\frac{445}{2592}$	$\frac{1}{27}$	$-\frac{1}{216}$	$\frac{h^{12}}{97297200} f^{(12)}$

the constant in the leading term of the truncation error for the CD scheme with the uncoupled one shows a smaller asymptotic error for the CD approach, for all orders.

Figures 2 and 6 show the phase speed for the convection and the decay ratio for the diffusion term, respectively. As can be seen from these figures, the CD approach shows a noticeably smaller error than the standard Padé method. A distinct feature of the CD method is that it displays a decay rate higher than the true one for high cell wave numbers. This is in contrast to the standard Padé, which exhibits a decay rate smaller than the true one for all cell wave numbers.

Clearly, CD Padé methods may be applied to pure advection or pure diffusion problems. However, in this case they require the computation of a "useless" derivative. This computation doubles the work required for the evaluation of the derivative, making it less attractive than the uncouple Padé method for problems of pure advection or pure diffusion problems.

Remark. Equations involving second and higher derivatives can be discretized analogously. For instance, the Korteweg–de Vries (KdV) equation

$$\frac{\partial u}{\partial t} + \frac{\partial u^2}{\partial x} + \frac{\partial^3 u}{\partial x^3} = 0$$

requires the evaluation of the second derivative. This can be done using the CD Padé approach solving the following **IP** problems:

Given a smooth function $\phi \in C^\infty$, and natural numbers $m, n, o \in \mathbb{N}$ find coefficients a_i, $i = 1, \ldots, m$, b_i, $i = 1, \ldots, n$, and c_i, $i = 1, \ldots, o$, such that

$$\sum_{i=1}^{m} a_i \tau_{-ih}\phi + \phi + \sum_{i=1}^{m} a_i \tau_{ih}\phi = \sum_{i=1}^{n} b_i \tau_{(-i+1/2)h}\bar{\phi} + \sum_{i=1}^{n} b_i \tau_{(i-1/2)h}\bar{\phi} + h^2\left(\sum_{i=1}^{o} c_i \tau_{-ih}\phi'' + \sum_{i=1}^{o} c_i \tau_{ih}\phi''\right) + \mathsf{o}\left(h^{2(m+n+o)-1}\right) \tag{32}$$

TABLE VI
Coefficients and Leading Term of the Truncation Error for Some CD Padé Methods: Diffusion

Order	d_1	d_2	e_1	e_2	f_1	f_2	Truncation error
6	$-\frac{1}{8}$	—	3	—	$-\frac{9}{8}$	—	$\frac{h^6}{2016} f^{(7)}$
8	$-\frac{1}{6}$	—	$\frac{13}{4}$	$-\frac{1}{108}$	$-\frac{23}{18}$	—	$-\frac{79h^8}{16329600} f^{(9)}$
12	$-\frac{8}{27}$	$\frac{1}{216}$	$\frac{15}{4}$	$-\frac{63}{324}$	$-\frac{40}{27}$	$\frac{25}{432}$	$-\frac{11273h^{12}}{73556683200} f^{(13)}$

TABLE VII
Coefficients and Leading Term of the Truncation Error of Some CD Padé Methods for KdV Equation: Convection

Order	a_1	a_2	b_1	b_2	c_1	c_2	Truncation error
6	$\frac{9}{32}$	—	$\frac{25}{32}$	—	$\frac{1}{96}$	—	$\frac{h^6}{2520} f^{(6)}$
8	$\frac{11}{48}$	—	$\frac{109}{144}$	$-\frac{1}{36}$	$\frac{1}{48}$	—	$-\frac{h^8}{36288} f^{(8)}$
12	$-\frac{209}{81}$	$-\frac{1153}{2592}$	$-\frac{7955}{15552}$	$-\frac{31315}{15552}$	$\frac{7}{27}$	$-\frac{1}{216}$	$-\frac{29h^{12}}{194594400} f^{(12)}$

and

Given a smooth function $\phi \in C^\infty$, and natural numbers $m, n, o \in \mathbb{N}$ find coefficients $d_i, i = 1, \ldots, m$, $e_i, i = 1, \ldots, n$, and $f_i, i = 1, \ldots, o$, such that

$$\sum_{i=1}^{m} d_i \tau_{-ih}\phi'' + \phi'' + \sum_{i=1}^{m} d_i \tau_{ih}\phi'' = \frac{1}{h^2}\left(\sum_{i=1}^{n} e_i \tau_{(-i+1/2)h}\bar{\phi} - \sum_{i=1}^{n} e_i \tau_{(i-1/2)h}\bar{\phi} + \sum_{i=1}^{o} f_i \tau_{-ih}\phi + f_0\phi + \sum_{i=1}^{o} f_i \tau_{ih}\phi\right) + \mathrm{o}\left(h^{2(m+n+o)-1}\right). \tag{33}$$

Note the inclusion of the term $f_0\phi$ in Eq. (33). In Tables VII and VIII we list the coefficients and leading term of the truncation error for the 6th-, 8th-, and 12th-order CD Padé methods.

Remark. The computational cost of the Padé finite volume method is the same as for the Padé finite difference methods. Indeed, writing the **IP** problems in matrix form

$$Af^{(n)} = Bf, \tag{34}$$

we see that the matrix A is the same as for the finite difference method. The common practice is to perform an LU decomposition of the matrix A only once, and store the L and U matrices. Hence, computation of the derivatives involves evaluating the RHS of Eq. (34) followed by forward and backward substitution. Lele [18], using a Cholesky decomposition of the symmetric part A for a tridiagonal scheme, gave an operation count of $5N$, N, and $5N$ for multiply, divide, and add/subtract, respectively. A radix 2 FFT (see [3]) requires

TABLE VIII
Coefficients and Leading Term of the Truncation Error of Some CD Padé Methods for KdV Equation: Dispersion

Order	d_1	d_2	e_1	e_2	f_0	f_1	f_2	Truncation error
6	$-\frac{1}{32}$	—	$\frac{315}{32}$	—	-15	$-\frac{75}{32}$	—	$\frac{h^6}{40320} f^{(8)}$
8	$-\frac{1}{20}$	—	$\frac{421}{40}$	$-\frac{1}{40}$	$-\frac{159}{10}$	$-\frac{51}{20}$	—	$-\frac{h^8}{1039500} f^{(10)}$
12	$-\frac{34}{261}$	$\frac{7}{8352}$	$\frac{6486715}{601344}$	$-\frac{330565}{601344}$	$-\frac{1935}{116}$	$-\frac{1570}{783}$	$\frac{10855}{100224}$	$-\frac{h^{12}}{1519333200} f^{(14)}$

$2N \log_2 N$ multiplies and $3N \log_2 N$ adds. Also, as pointed out by Mahesh [21], the cost incurred by the use of CD Padé methods is essentially the same as for the uncoupled one.

3. MULTI-DIMENSIONAL EQUATIONS

Higher dimensional equations require the average of the variables at the cell faces. For example, the two-dimensional convection problem can be written as

$$\frac{d\bar{\phi}}{dt} + c_1\left(\frac{\tau_{-h_1/2}\bar{\phi}^{x_2} - \tau_{h_1/2}\bar{\phi}^{x_2}}{h_1}\right)h_2 + c_2\left(\frac{\tau_{-h_2/2}\bar{\phi}^{x_1} - \tau_{h_2/2}\bar{\phi}^{x_1}}{h_2}\right)h_1 = 0,$$

where

$$\bar{\phi} = \frac{1}{h_1 h_2}\int_{\mathbb{R}\times\mathbb{R}} \chi_{h_1,h_2} * \phi,$$

and the superscripts x_1 and x_2 indicate the one-dimensional averaging in the x_1 and x_2 directions, respectively; that is, for example,

$$\bar{\phi}^{x_2} = \frac{1}{h_2}\int_{\mathbb{R}} \chi_{h_2} * \phi.$$

The choice between sliding averages or point values for methods that are 2nd order or lower is immaterial, since both variables differ by a 2nd-order term only. However, for higher order methods this choice can influence the stencil for the interpolation. Take for instance the usual QUICK scheme of Leonard [19] and the 4th-order method proposed in [20]. Both methods are based in point values. Now, because cross derivatives appear in the Taylor series, as the averaging is needed in the transverse direction to the interpolation, they require 13 and 25 points in the stencil to obtain 3rd- and 4th-order accuracy, respectively. Hence, these methods need points out of the Cartesian product of the one-dimensional stencils for their one-dimensional convection counterpart. All methods developed above use the sliding average of the dependent variable for interpolation. The integration in its definition is performed in both directions, which accounts for the variations in the transverse direction.

In conclusion, the stencil of the methods developed above for higher dimensional problems is the Cartesian product of the stencils of their one-dimensional counterparts. However, when using point values we need points out of the latter product in order to take into account transverse variation of the variables.

Another feature, unique to the interpolation on higher dimensional domains, is the anisotropy in the interpolation of the convection term. In fact, the symmetry to the group of rotations of the Laplacian operator means that it is isotropic, or, in other words, that a sinusoidal wave propagating in any direction will experiment the same rate of decay. The discretization of the Laplacian operator presented earlier possesses this property as well. However, the convection term has the velocity direction as a preferred direction (see [29]). It may be shown that

$$\mathfrak{c}_\alpha = \frac{1}{kh}(\cos(\alpha)\sin(kg\cos(\alpha))\mathfrak{c}(k\cos(\alpha)) + \sin(\alpha)\sin(kg\sin(\alpha))\mathfrak{c}(k\sin(\alpha))),$$

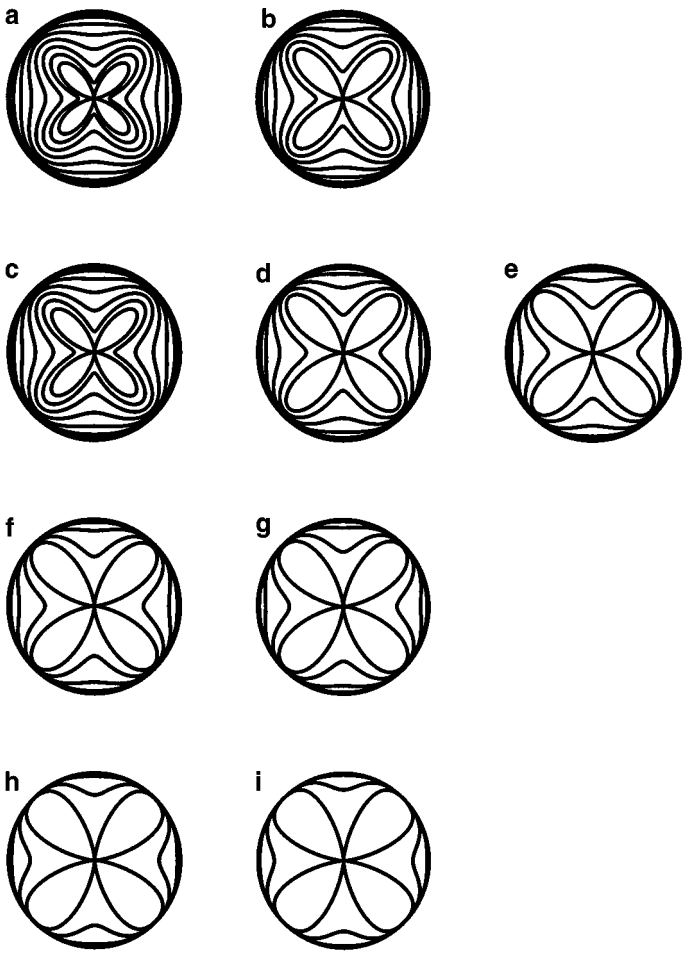

FIG. 7. Polar plot of phase speed anisotropy for advection using (a) 4th order Lagrange, (b) 4th order Padé, (c) 6th order Lagrange, (d) 6th order Padé, (e) 6th order CD Padé, (f) 8th order Padé, (g) 8th CD Padé, (h) 12th order Padé, and (i) 12th order CD Padé, for cell wave number $\tilde{k} = \frac{1}{20}, \frac{2}{20}, \ldots, \frac{10}{20}$.

where $\mathfrak{c}_\alpha$ is the phase velocity and α is the angle between the direction of propagation and the x_1 axis.

The anisotropic propagation is displayed in Fig. 7 for several methods. The curves in this figure are polar plots of $\mathfrak{c}_\alpha/c$ at fixed cell grid wave numbers. For each curve, the radial distance at an angle α represents $\mathfrak{c}_\alpha$ obtained for a wave propagating in that direction. The curves corresponds to $\tilde{k} = \frac{1}{20}, \ldots, \frac{10}{20}$; the outmost curves correspond to small $\tilde{k}$. For these waves the propagation is nearly isotropic and the phase speed is close to exact. For larger $\tilde{k}$ the waves have a small phase speed and the propagation is anisotropic. It may be seen that in the Padé methods the anisotropy error is noticeable in a narrower range of short waves lengths, when compared to the Lagrange methods. Moreover, the CD methods display an even smaller anisotropy error for all orders and wave lengths.

Remark. The notion of group velocity applies to multi-dimensional equations as well. The definition is the same as above, with trivial adaptations (for instance, now we must

consider the wave vector k). In the general case, group velocity depends on the wave number, and its direction can be different from that of the wave vector.

Remark. In [24] it is mentioned that experimental evidence suggests that acoustic waves are strogly coupled to many mechanisms encountered in turbulence. The error in wave equation describing the propagation of the acoustic with Padé methods is predominantly in the high wave number range. So, except for the unusual case of vanishing sound speed, it is of high frequency and thus analogous to sound rays. In this situation, we can apply the theory of geometric acoustics (see [16]) to the propagation of error "rays" in nonhomogeneous medium, that is, where c is not constant. It can be shown that the magnitude of the wave vector varies according to the simple law $k = \omega/c$, while its direction changes according to

$$K = -\frac{1}{c}(n \cdot \operatorname{grad} c),$$

where K is the curvature of the curve described by the ray and n is the unitary vector of the principal normal to the curve. In other words, the error rays turn in the sense of the diminishing of the speed c.

4. OPTIMIZED PHASE ERROR AND SPECTRAL PADÉ INTERPOLATION

In addtion to maximal order schemes, classes of parametrized nonmaximal order Padé finite volume methods may offer optimization opportunities which go beyond the maximal order of accuracy. In [18], a class of compact finite difference schemes with spectral-like resolution is proposed. The methods are obtained by constraining the error to be of 4th order, and the phase speed to match the exact value of the phase speed at some predefined locations. In [8] the authors define a family of parametrized schemes which minimizes the L^1-norm of the error in the phase velocity for a given interval of cell wave numbers. That is, they minimize the dispersion error,

$$E_d(a_1, \tilde{k}_{\max}) = \int_0^{\tilde{k}_{\max}} |\mathfrak{c}(\tilde{k}) - c|\, d\tilde{k},$$

with the restriction that the method must be of at least 4th order of accuracy. Both methods in [18] and [8] use the stencil of the 6th-order Padé method referred to above.

In [8] the authors concluded that $a_{1,\mathrm{opt}}$ takes values in $\frac{1}{3} \le a_{1,\mathrm{opt}} \le 0.431$ (the value of $a_{1,\mathrm{opt}} = \frac{1}{3}$ corresponds to the 6th-order Padé method). They also considered optimal methods with respect to anisotropic errors, the values of the optimum coefficient being in $\frac{1}{3} \le a_{1,\mathrm{opt}} \le 0.408$.

Other target functions may be considered; for example, the standard Padé interpolation optimizes the order of accuracy in a given stencil. It is evident that the optimum solution depends on the target function. Consider the ability of a method to handle discontinuities, the occurrence of which in compressible inviscid flows is relevant in applications. In [29] it is shown that

$$\|\epsilon\|_2^2 = \int_0^{\pi/h} \left| \frac{h}{\sin(uh/2)} \right|^2 \cdot \left| \sin\left(\frac{u(\mathfrak{c}(u) - c)t}{2} \right) \right|^2 \frac{du}{\pi},$$

where ϵ is the global error of a semidiscrete approximation to this problem. We look for the optimum methods as $t \to 0$ and $t \to \infty$. Using $\sin(u(\mathfrak{c}-c)t/2) \sim u(\mathfrak{c}-c)t/2$, for sufficiently

small $t \in \mathbb{R}$, it follows that the former minimizes the L^2-norm of the consistency error

$$T_h = \int_{-\pi/h}^{\pi/h} \frac{h/2}{i \sin(uh/2)} iu(c - \mathfrak{c}) e^{iux} \frac{du}{2\pi}.$$

When $t \to \infty$, the least error is obtained with the method which maximizes the order of accuracy. To show this we first perform a change of coordinate $\bar{u} = ut$; then,

$$\begin{aligned}
\|\epsilon\|_2^2 &= \int_0^{\pi t/h} \left| \frac{h}{\sin(\bar{u}h/2t)} \right|^2 \cdot \left| \sin\left(\frac{\bar{u}(\mathfrak{c} - c)}{2} \right) \right|^2 \frac{du}{t\pi} \\
&\sim \int_0^{\pi t/h} \left| \frac{h}{\bar{u}h/2t} \right|^2 \cdot \left| \sin\left(\frac{\bar{u}(\mathfrak{c} - c)}{2} \right) \right|^2 \frac{d\bar{u}}{t\pi} \\
&= \int_0^{\pi/h} \left| \frac{h}{uh/2} \right|^2 \cdot \left| \sin\left(\frac{u(\mathfrak{c}(u) - c)t}{2} \right) \right|^2 \frac{du}{\pi}.
\end{aligned}$$

The result follows from the fact that, in the topology of $\mathcal{S}'$, we have

$$\lim_{n \to \infty} \frac{\sin(n\hat{x})}{\hat{x}} = \pi \delta. \tag{35}$$

Then, for sufficiently large values of $t \in \mathbb{R}$ the value of $\|\epsilon\|_2^2$ is dominated by the error in the phase velocity near the origin. A Taylor series expansion shows that this error decays at the same rate as the truncation error.

To prove (35) we first use integration by parts to show that if $f \in C^1 \cap C_0$, where $C_0 = \{ f \in C : \lim_{x \to \pm\infty} f(x) = 0 \}$, $f' \in L^1$, and for all $n \in \mathbb{N}$ the integral

$$\int_{\mathbb{R}} \sin(nx) f(x)\, dx$$

exists; then

$$\lim_{n \to \infty} \int_{\mathbb{R}} \sin(nx) f(x)\, dx = 0.$$

Now, given a $\psi \in \mathcal{S}$ it is easy to see that the function

$$f : x \in \mathbb{R} \mapsto \lim_{y \to x} \frac{\psi(y) - \psi(0)}{y} \in \mathbb{R}$$

belongs to $C^1 \cap C_0$ and $f' \in L^1$ (just use the Maclaurin series and the fact that $\psi' \in L^1$). The result follows if we observe that $\frac{\sin(n\hat{x})}{\hat{x}} \in \Lambda$, so

$$\left\langle \frac{\sin(n\hat{x})}{\hat{x}}, \psi \right\rangle = \int_{\mathbb{R}} \frac{\sin(n\hat{x})}{\hat{x}} \psi\, dx \to \pi \psi(0)$$

for all $\psi \in \mathcal{S}$, where we have used the fact that $\int_{\mathbb{R}} \frac{\sin(nx)}{x}\, dx = \pi$ (as an improper integral of Riemann).

Spectral methods are based on representing the solution of a problem as a truncated series of global smooth functions. The latter are often trigonometric functions for a periodic boundary condition, or Chebyshev or Legendre polynomials for general boundary conditions. A prominent characteristic of these methods is exponential convergence.

Using the fact that the Padé method corresponds to the use of the Hermite polynomial, we relax the requirement of being a global smooth function and define a piecewise spectral Padé method (SPM) as follows:

In a mesh comprising N control volumes, use the (N/8th)-order Padé method.

Then, the error decays at a rate of $\mathsf{O}(N^{N/8})$, which is faster than any finite power of N^{-1}. Also, as the Fourier analysis have shown, the phase speed using the SPM will stay progressively close to the exact one, with respect to both the cell wave number and the anisotropy error. In actual computations we use $N = 2^p$ for $p = 4, \ldots$.

EXAMPLE 4.1. We close this section with an example comparing several Padé methods. We resolve the problem

$$\frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial x} = 0, \qquad x \in [0, 1], \quad t \geq 0$$

with periodic boundary conditions and initial profile

$$\phi(0, x) = \begin{cases} K e^{-1/((x-a)(b-x))} & \text{if } x \in [a, b], \\ 0 & \text{otherwise,} \end{cases}$$

where $a, b, K \in \mathbb{R}$ are constants. In all cases, $a = \frac{1}{4}$, $b = \frac{3}{4}$, $K = e^{1/16}$ (to make the peak unitary) and the temporal discretization is accomplished with the 4th-order RK algorithm. This bell-shaped profile corresponds to a C_c^∞-distribution with high gradients in the fore and aft portions of the bell and a high curvature region close to the peak. It can serve as test case for peak resolution and/or monotonicity. A final time $T = \frac{1}{4}$ is used in all simulations for all grids and methods at CFL $= \frac{1}{10}$.[4] At this CFL level the errors due to time discretization are negligible compared with the spatial terms. The maximum error at T is calculated and is shown in Fig. 8.

The optimized method of Gaitonde and Shang display a smaller error than the 6th-order method for grids with less than 32 control volumes. In this range of grid parameter, the errors are still in the region of relatively high cell wave numbers and so where the optimized method has a better performance. After that, most of the monochromatic waves with small cell wavelength are resolved, the error is dominated by the asymptotic order, and the 6th becomes superior. The exponential decay in the error for the SPM is also evident.

Remark. All methods considered above present an error in the phase, but no error in amplitude. This is the case for any Padé method with a symmetric stencil for the sliding averages and the variable (it is immediate from the Fourier analysis). This is valid for the infinite domain or for periodic boundary conditions. However, in general, the boundary

[4] To avoid any contamination from time errors, the result for the SPM in the grid with 128 control volumes is obtained with the exact solution of the semidiscrete equation. That is, writing the latter as $d\bar{\phi}/dt = (\sigma_c/\Delta t)\mathcal{C}\bar{\phi}$ we compute $\bar{\phi}(T) = e_c^{\sigma \mathcal{C}(T/\Delta t)}\bar{\phi}_0 = S e^{\sigma_c \Lambda (T/\Delta t)} S^{-1}\bar{\phi}_0$, where σ_c is the CFL number, $\mathcal{C}$ is defined in Section 6, S is the matrix with the eigenvectors of $\mathcal{C}$ in its columns, and Λ is the corresponding diagonal matrix with the eigenvalues in its diagonal.

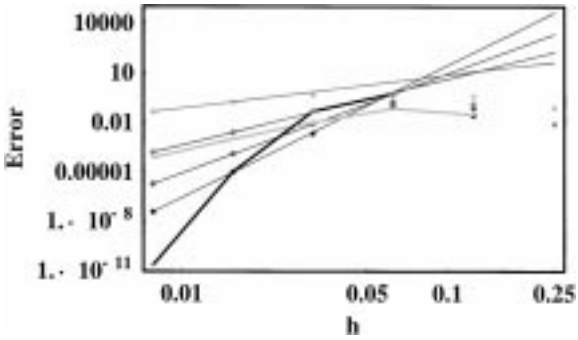

FIG. 8. Plot of maximum error versus grid parameter for pure advection using various Padé methods. Solid thin gray lines are asymptotic curves Ah^r, with A a constant and r the corresponding order of accuracy. The solid black line is the spectral method. The thick solid gray line is the optimum 4th method of Gaitond and Shang with $\tilde{k}_{\max} = \frac{3}{8}$. Points in increasing gray levels are the numerical solutions corresponding to 2th, 4th, 6th, and 8th Padé methods, respectively.

conditions in a finite domain will break this symmetry and the numerical solution will display some damping effect (see [4], and also Section 5).

5. BOUNDARY CONDITIONS

Boundary conditions represent a particular case of the general **IH** problem. However, boundary effects may affect the accuracy of the solution as well as its stability. In this section we study their effect on the accuracy, and we defer to Section 6 our analysis of their effect on the stability.

Some authors have suggested that the approximation on the boundary, if one order of accuracy below that at interior interpolation, may not affect the global (uniform) order of accuracy (see, for example, [4, 24, 29]). In [29] this idea is supported using the notion of the reflection ratio. We use this concept to show that in the case of the Padé finite volume method, the boundary approximation has a determining effect on the global accuracy.

To compute the reflection ratio, we need to find normal, or fundamental, solutions of the time-Fourier transform of the semidiscrete equation. We write the latter as

$$\frac{d\bar{\phi}}{dt}_{j+1/2} + c\frac{\phi_{j+1} - \phi_j}{h} = 0, \qquad j \in \mathbb{Z}. \tag{36}$$

Then, we define the time-Fourier transform of ϕ as

$$\Phi = \int_{\mathbb{R}} \phi e^{-i\hat{\omega}t}\, dt.$$

Note that we use the same letter for the time-Fourier transform as for the space-Fourier transform. Since we do not use both at the same time, this should not cause confusion. We take the time-Fourier transform of (36) and obtain

$$i\omega\bar{\Phi}_{j+1/2} + c\frac{\Phi_{j+1} - \Phi_j}{h} = 0.$$

Using the 4th-order Padé method and rearranging terms we obtain

$$(\kappa^{-1} + 1 + \kappa)is = 3(\kappa - \kappa^{-1}),$$

where $s = \omega h/c$, and $\bar{\Phi}_{j+1/2}/\bar{\Phi}_{j-1/2} = \kappa,\ \ j \in \mathbb{Z}$. The characteristic roots are

$$\kappa_1 = \frac{-2\,s - i\,\sqrt{3}\,\sqrt{3 - s^2}}{s - 3\,i}$$
$$\kappa_2 = \frac{-2\,s + i\,\sqrt{3}\,\sqrt{3 - s^2}}{s - 3\,i}.$$

We consider the following approximations at the boundary:

1. 3rd-order Padé

$$\phi_N + 2\phi_{N-1} = \frac{1}{2}\big(5\bar{\phi}_{N-1/2} + \bar{\phi}_{N-3/2}\big).$$

2. 4th-order Padé

$$\phi_N + 3\phi_{N-1} = \frac{1}{6}\big(17\bar{\phi}_{N-1/2} + 8\bar{\phi}_{N-3/2} - \bar{\phi}_{N-5/2}\big).$$

At the control volume closest to the boundary Eq. (36) with the 3rd-order Padé method at boundary yields

$$is(1+\rho) = -\frac{\frac{5}{2} + \frac{1}{2\kappa_1}}{1 + \frac{2}{\kappa_1}} + \frac{3\big(1 + \frac{1}{\kappa_1}\big)}{4 + \frac{1}{\kappa_1} + \kappa_1} - \frac{\rho\big(\frac{5}{2} + \frac{1}{2\kappa_2}\big)}{1 + \frac{2}{\kappa_2}} + \frac{3\rho\big(1 + \frac{1}{\kappa_2}\big)}{4 + \frac{1}{\kappa_2} + \kappa_2},$$

whence, solving for the reflection ratio, we find

$$\rho(s) = \frac{\sqrt{3}(-6i + s)\sqrt{3 - s^2} - 3(-6i + s + is^2)}{\sqrt{3}(-6i + s)\sqrt{3 - s^2} + 3(-6i + s + is^2)}.$$

Analogously, we find for the 4th-order Padé method at the boundary

$$\rho(s) = \frac{\sqrt{3}\sqrt{3 - s^2}(24 + s(-6i + s)) - 3(24 - s(6i + (3 - is)s))}{\sqrt{3}\sqrt{3 - s^2}(24 + s(-6i + s)) + 3(24 - s(6i + (3 - is)s))}.$$

In Figs. 9 and 10 the absolute value and the phase of the reflection ratio, respectively, are shown. Using the Padé method, the region of propagating waves enlarges from $0 \le s \le 1$ to $0 \le s \le \sqrt{3}$. Moreover, the spurious oscillations generated at the boundary by the 4th-order Padé method, independent of the boundary approximation, are much smaller than those generated by the 2nd-order Padé finite difference method. The difference between the absolute value of the reflection ratio for the 3rd- and 4th-order boundary conditions does not seem significant (note that the difference in the phase of the reflection ratio is not small). However, a Taylor series expansion of the reflection ratio yields

1. 3rd-order Padé

$$\rho(s) = -\frac{is^3}{72} + \mathrm{o}(s^3).$$

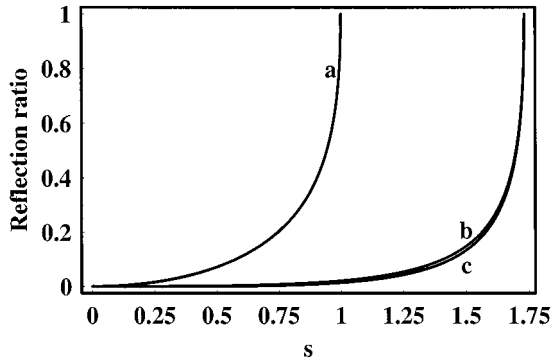

FIG. 9. Plot of the absolute value of the reflection ratio versus specific frequency s using the Padé method: (a) 2nd-order finite difference with 1st order at boundary; (b) 4th order with 3rd order at boundary, (c) 4th order with 4th order at boundary.

2. 4th-order Padé

$$\rho(s) = -\frac{s^4}{96} + \mathrm{o}(s^4).$$

We conclude that the implicit character of the Padé method makes it more sensitive to the boundary closure. Indeed, for a 4th-order Padé method for the inner points and the 3rd-order boundary closure, the global accuracy of the method is of 3rd order.

EXAMPLE 5.1. We resolve the problem

$$\frac{\partial \phi}{\partial t} + \frac{\partial \phi}{\partial x} = 0, \qquad x \in [0,1], \quad t \geq 0$$

$$\phi(t,0) = 0$$

$$\phi(0,x) = \begin{cases} K e^{-1/((x-a)(b-x))} & \text{if } x \in [a,b], \\ 0 & \text{otherwise,} \end{cases}$$

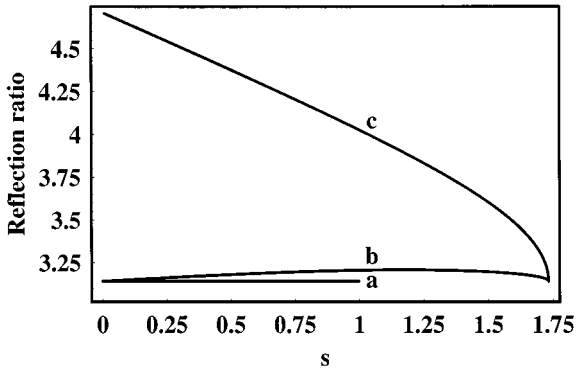

FIG. 10. Plot of the phase of the reflection ratio versus specific frequency s using the Padé method: (a) 2nd-order finite difference with 1st order at boundary, (b) 4th order with 3rd order at boundary, (c) 4th order with 4th order at boundary.

TABLE IX
Effect of the Boundary Closure on the Maximum Error of the 4th-Order Accurate Padé Finite Volume Method

Boundary grid	3rd order		4th order	
	$\|\epsilon\|_\infty$	Numerical order	$\|\epsilon\|_\infty$	Numerical order
8	1.3831×10^{-1}	—	1.0871×10^{-1}	—
16	2.5453×10^{-1}	−0.88	2.2546×10^{-1}	−1.06
32	8.8983×10^{-2}	1.52	7.7016×10^{-2}	1.55
64	5.5606×10^{-3}	4.00	5.3655×10^{-3}	3.84
128	4.5039×10^{-4}	3.63	3.4711×10^{-4}	3.95
256	4.9249×10^{-5}	3.19	2.1577×10^{-5}	4.01

where $a, b, K \in \mathbb{R}$ are constants. In all cases, $a = \frac{1}{4}$, $b = \frac{3}{4}$, $K = e^{1/16}$, and the temporal discretization is accomplished with the 4th-order RK algorithm. A final time $T = \frac{1}{2}$ is used in all simulations for all grids and methods at $\mathrm{CFL} = \frac{1}{10}$. This value of T corresponds to the critical instant when the peak crosses the outlet. At this CFL level the errors due to time discretization are negligible compared with the spatial terms. The error at T is calculated and reported as the L^∞-norm. Table IX shows the effect of the boundary approximation on the accuracy of the 4th-order Padé finite volume method.

As expected the global order of accuracy is determined by the boundary closure: it is of 3rd order for the 3rd-order accurate closure, and it is of 4th order for the 4th-order accurate closure. Moreover, the asymptotic order is only achieved for relatively fine grids.

The effect of the boundary closure on global accuracy of solution in a transport problem using the CD approach may be even more pronounced. Indeed, in this case, not only is the problem of an elliptic character, but the coupling in the interpolation problem makes the implicit nature of the method even stronger. To illustrate this we consider.

EXAMPLE 5.2. We solve the transport problem

$$\frac{d\phi}{dx} = \frac{1}{\mathrm{Pe}} \frac{d^2\phi}{dx^2}, \qquad x \in [0, 1],$$

$$\phi(0) = 0$$
$$\phi'(1) = \frac{e^{\mathrm{Pe}}\,\mathrm{Pe}}{e^{\mathrm{Pe}} - 1}.$$

The exact solution is

$$\phi(x) = \frac{1 - e^{\mathrm{Pe}\,x}}{1 - e^{\mathrm{Pe}}}, \qquad x \in [0, 1].$$

In all grids we use $\mathrm{Pe} = 10$ and compute the maximum error in the sliding average, the interface values of the variable and its first derivative. Tables X, XI, and XII show the effects of the boundary approximation on the accuracy of the sliding average, interface values, and interface derivative, using the 6th-order CD Padé finite volume method. The boundary closures considered are:

TABLE X
Effect of the Boundary Closure on the Maximum Error of the Sliding Average Using the CD 6th-Order Accurate Padé Finite Volume Method

Boundary grid	5th order		6th order	
	$\|\epsilon\|_\infty$	Numerical order	$\|\epsilon\|_\infty$	Numerical order
8	5.3693×10^{-3}	—	2.9127×10^{-3}	—
16	3.6119×10^{-4}	3.89	1.2802×10^{-4}	4.51
32	1.8785×10^{-5}	4.27	3.8778×10^{-6}	5.05
64	8.5455×10^{-7}	4.46	9.4831×10^{-8}	5.67
128	3.2238×10^{-8}	4.73	1.8570×10^{-9}	5.67
256	1.1071×10^{-9}	4.86	3.2359×10^{-11}	5.84

1. Inlet: 6th-order Dirichlet

$$\phi'_0 - 4\phi'_1 = \frac{1}{h}\left(-\frac{49}{6}\phi_0 - \frac{28}{3}\phi_1 + \frac{1447}{72}\bar{\phi}_{1/2} - \frac{67}{24}\bar{\phi}_{3/2} + \frac{5}{2}\bar{\phi}_{5/2} - \frac{1}{72}\bar{\phi}_{7/2}\right).$$

2. Outlet: von Neumann
   (i) 5th-order

$$\phi_N + \frac{8}{7}\phi_{N-1} = \frac{h}{7}(\phi'_N - 2\phi'_{N-1}) + \frac{31}{14}\bar{\phi}_{N-1/2} - \frac{1}{14}\bar{\phi}_{N-3/2}.$$

   (ii) 6th-order

$$\phi_N + \frac{27}{23}\phi_{N-1} = \frac{3h}{23}(\phi'_N - 3\phi'_{N-1}) + \frac{325}{138}\bar{\phi}_{N-1/2} - \frac{13}{69}\bar{\phi}_{N-3/2} + \frac{1}{138}\bar{\phi}_{N-5/2}.$$

Again, the imposition of lower order boundary conditions at the outflow (and only there) lowers the overall accuracy. It is also noteworthy that in the Padé method the high order of accuracy for the sliding averages is the same for the interface values and the derivatives at the interfaces. This contrasts with usual methods for which the order of the latter two values decrease one order of accuracy for each derivative (see, for example, [11]).

TABLE XI
Effect of the Boundary Closure on the Maximum Error of the Interface Values Using the CD 6th-Order Accurate Padé Finite Volume Method

Boundary grid	5th order		6th order	
	$\|\epsilon\|_\infty$	Numerical order	$\|\epsilon\|_\infty$	Numerical order
8	7.0614×10^{-3}	—	3.7353×10^{-3}	—
16	5.5274×10^{-4}	3.68	1.9197×10^{-4}	4.28
32	2.8681×10^{-5}	4.27	5.7913×10^{-6}	5.05
64	1.1643×10^{-6}	4.62	1.2720×10^{-7}	5.51
128	4.1542×10^{-8}	4.81	2.3622×10^{-9}	5.75
256	1.3877×10^{-9}	4.90	4.0085×10^{-11}	5.88

TABLE XII
Effect of the Boundary Closure on the Maximum Error of the Interface Derivative Using the CD 6th-Order Accurate Padé Finite Volume Method

Boundary grid	5th order		6th order	
	$\|\epsilon\|_\infty$	Numerical order	$\|\epsilon\|_\infty$	Numerical order
8	6.8994×10^{-2}	—	3.6472×10^{-2}	—
16	5.5270×10^{-3}	3.64	1.9191×10^{-3}	4.25
32	2.8682×10^{-4}	4.27	5.7910×10^{-5}	5.05
64	1.1643×10^{-5}	4.62	1.2719×10^{-6}	5.51
128	4.1544×10^{-7}	4.81	2.3622×10^{-8}	5.75
256	1.3878×10^{-8}	4.90	4.0083×10^{-10}	5.88

Remark. This effect of the boundary condition on the global accuracy of the Padé method is not exclusive of the finite volume formulation. In the finite difference method the point closest to the boundary has a similar equation. Add to this the unavoidable approximation of the evolution equation at the boundary and the finite difference method would be at least as sensitive to the boundary approximation as is the finite volume method.

Remark. At the inlet, the finite difference and the finite volume methods require different approaches too. The finite difference method needs a downwind extrapolation from the interior points to the boundary, and as is well known this is highly unstable. In contrast, for the finite volume method no approximation is needed since it can use the exact prescribed value at the inlet, both in the discretized equation and in the Padé method. Thus, for example, a 4th-order Padé method does not require any special treatment for the inlet. Sixth and higher order methods require some modifications at points close to the inlet, which depends on the specific method (see Section 6 for an example with the 6th-order method.)

Remark. In a pure diffusion problem and using the standard Padé method, the derivative at the boundary does not involve the prescribed value of the variable. This problem can be overcome by considering the value of the function available at the boundary in the interpolation problem. For instance, at the right boundary, and for a 4th-order accurate boundary condition, we have

$$\phi'_N + 6\phi'_{N-1} = \frac{1}{h}\left(\frac{5}{3}\phi_B + \frac{89}{18}\bar{\phi}_{N-1/2} - \frac{127}{18}\bar{\phi}_{N-3/2} + \frac{4}{9}\bar{\phi}_{N-5/2}\right),$$

where ϕ_B is the prescribed value of the variable at the boundary and N is the number of control volumes in the grid.

Remark. Poinsot and Lele [24] proposed characteristic boundary conditions for Euler and Navier–Stokes equations. They advanced the solution in time on the boundaries by using a characteristic system. In this system they used a 3rd-order one-sided space derivative. This methodology can be used in the finite volume framework as well. Moreover, using the Hermite polynomial, which interpolates the sliding average, and whose derivative interpolates the interface values, we can perform a complete characteristic evolution. Indeed, this polynomial may be used as a reconstruction polynomial and we can directly apply either a local Cauchy–Kowalewski procedure or a characteristic method (see [11]). For example, using the characteristic method for the linear advection problem with the 4th-order Padé

method and a Simpson's rule yields

$$\int_t^{t+\Delta t} \phi_i dt \cong \frac{\Delta t}{6}(H_i'(x_i, t) + 4H_i'(x_i - c\Delta t/2, t) + H_i'(x_i - c\Delta t, t))$$

for all $i = 1, \ldots, N$, where H is the Hermite polynomial

$$H_i(x) = \left(\bar{\phi}^n_{i-1/2} + \bar{\phi}^n_{i+1/2}\right)h\eta_i(x) + \phi^n_{i+1}\tilde{\eta}_{i+1}(x) + \phi^n_{i-1}\tilde{\eta}_{i-1}(x)$$

for all $i = 1, \ldots, N-1$, $x \in \mathbb{R}$, with

$$\eta_i(x) = (1 - 2l'_{i+1}(x_{i+1})(x - x_{i+1}))(l_{i+1}(x))^2, \qquad i = 1, \ldots, N-1, \quad x \in \mathbb{R}$$
$$\tilde{\eta}_{i\pm1} = (x - x_{i\pm1})(l_{i\pm1}(x))^2, \qquad i = 1, \ldots, N-1, \quad x \in \mathbb{R}$$

and

$$l_{i\pm1}(x) = \frac{x - x_{i\mp1}}{\pm 2h}, \qquad i = 1, \ldots, N-1, \quad x \in \mathbb{R}.$$

At the boundary we use the Hermite polynomial H_{N-1}, that is, $H_N = H_{N-1}$.

This procedure requires less storage than the RK one and is faster for Padé methods since it needs only one evaluation of the interface values per time step. However, the Hermite polynomial approximation is, except for the center point used in the Padé method, one order of accuracy lower than that for the sliding average. So, the accuracy of the resulting method is one order of accuracy lower than that using the RK method.

6. STABILITY ANALYSIS

We now turn to the analysis of the stability of the proposed methods for both explicit and implicit temporal discretization. There are many definitions of stability in the literature of CFD (see, for example, [4, 10]). In the present work, we consider the notions of Lyapunov and asymptotic stability.

Recall (see, for example, [1]) that a stationary solution of an autonomous dynamical system is said to be Lyapunov stable if all solutions of the equation, with initial conditions in a sufficiently small neighborhood of the equilibrium point, are defined for all positive time and converge uniformly with respect to time to the stationary solution as the initial conditions tend to the equilibrium point. A stationary solution is said to be asymptotically stable if it is Lyapunov stable and if, in addition, all solutions with initial conditions sufficiently close to the equilibrium point under consideration tend to this equilibrium point as $t \to +\infty$.

We start with the stability of the linear hyperbolic equation with periodic initial conditions, discretized with a Padé finite volume method. The semidiscrete equation can be written as

$$\frac{d\bar{\phi}}{dt} = \frac{\sigma_c}{\Delta t}\mathcal{C}\bar{\phi}, \tag{37}$$

where $\sigma_c = \frac{c\Delta t}{h}$ is CFL number, and

$$[\bar{\phi}] = \begin{bmatrix} \bar{\phi}_{1/2} \\ \vdots \\ \bar{\phi}_{N-1/2} \end{bmatrix}.$$

The operator $\mathcal{C} \in \mathsf{End}(\mathbb{R}^N)$ is given locally as

$$(\mathcal{C}\phi)_{i+1/2} = \phi_i - \phi_{i+1}, \qquad i = 0, \ldots, N-1,$$

where ϕ is obtained with the Padé discretization method.

Consider the implicit Euler method. In this case the fully discrete equation can be written as

$$\bar{\phi}^{n+1} = \bar{\phi}^n + \sigma_c \mathcal{C}\bar{\phi}^{n+1}.$$

It is easy to see that any centered scheme is unconditionally stable. Indeed, given the fact that $\mathcal{C}$ is skew-adjoint, all of its eigenvalues are pure imaginary (one or two of them are zero; see Eq. (39) below):

$$\begin{aligned} \left\|(\mathrm{id} - \sigma_c\mathcal{C})^{-1}\right\|_2 &= \frac{1}{\sqrt{\lambda_{\min}(\mathrm{id} - (\sigma_c\mathcal{C})^2)}} \\ &= 1. \end{aligned}$$

This proves the assertion.

Now, we turn to the analysis of the 4th RK algorithm for the solution of the semidiscrete problem. Given the symbolic polynomial

$$\mathrm{RK}_l(z) = 1 + z + \cdots + \frac{1}{l!}z^l, \qquad l \in \mathbb{N},$$

the associate discrete dynamical system is given by

$$\bar{\phi}^{n+1} = \mathrm{RK}_4(\sigma_c\mathcal{C})\bar{\phi}^n,$$

where, in the present linear case, $\mathrm{RK}_l(\mathcal{C}) \in \mathsf{End}(\mathbb{R}^N)$, $l \in \mathbb{N}$ is the Runge–Kutta operator, which corresponds to the l-truncated series of the exponential of $\sigma_c\mathcal{C}$, that is,

$$\mathrm{RK}_l(\sigma_c\mathcal{C}) = \mathrm{id} + \sigma_c\mathcal{C} + \cdots + \frac{1}{l!}(\sigma_c\mathcal{C})^l, \qquad l \in \mathbb{N}. \tag{38}$$

The operator $\mathcal{C}$ is skew-adjoint, so the origin is a stable equilibrium point of the semidiscrete. Indeed, this follows from the fact that

$$\frac{d\bar{\phi}^2}{dt} = 2\sigma_c\bar{\phi} \cdot \left(\frac{\mathcal{C}\bar{\phi}}{\Delta t}\right) = 0,$$

and so, the solution will stay at the sphere containing the initial condition.

This means that any instability in the discrete formulation comes from the time discretization. And, in general, increasing the order of the time discretization improves the stability of the method. This contrasts with the spatial discretization, which loses stability with the increase in accuracy. However, this is a consequence of the fact that the latter converges to an unbounded operator.

Analogously to the semidiscrete equation, we compute

$$(\bar{\phi}^{n+1})^2 = (\bar{\phi}^n)^2 - \frac{\sigma_c^6}{72}(\mathcal{C}^3\bar{\phi}^n)^2 + \frac{\sigma_c^8}{576}(\mathcal{C}^4\bar{\phi}^n)^2.$$

Note that any centered scheme would lead to this equation, and so, for a sufficiently small CFL number any centered scheme can be made stable for the 4th-order RK.

To determine how small the CFL number must be, we solve

$$-\frac{1}{72}(\mathcal{C}^3 v)^2 + \frac{\sigma_c^2}{576}(\mathcal{C}^4 v)^2 \le 0, \qquad v \in \mathbb{R}^N.$$

Then a necessary and sufficient condition for stability is

$$\sigma_c \le \sigma_{c,\max} = \frac{2\sqrt{2}}{\|\mathcal{C}\|_2}.$$

Given $A \in \mathsf{End}(\mathbb{R}^N)$, let $\rho(A)$ denote its spectral radius. By definition

$$\|\mathcal{C}\|_2 = \sup_{\|v\|_2=1} \|\mathcal{C}v\|_2 = \rho(\mathcal{C}),$$

where we have used the fact that $\mathcal{C}$ is skew-adjoint. Taking into account that $\{e^{ik(n+1/2)2\pi/N}\}_{n=0,\dots,N-1}$, for $k = 0, \dots, N-1$ are eigenvectors of $\mathcal{C}$ with eigenvalues

$$\lambda_k = 4i \sin\left(\frac{k\pi}{N}\right) \frac{\sum_{j=1}^n b_j \cos\left(\left(j-\frac{1}{2}\right)k\pi/N\right)}{1+2\sum_{j=1}^m a_j \cos(jk\pi/N)}, \qquad k = 0, \dots, N-1, \tag{39}$$

we obtain

$$\|\mathcal{C}\|_2 = \sup_{|kh|\le\pi} \left| 4i \sin\left(\frac{kh}{2}\right) \frac{\sum_{j=1}^n b_j \cos\left(\left(j-\frac{1}{2}\right)kh\right)}{1+2\sum_{j=1}^m a_j \cos(jkh)} \right|. \tag{40}$$

Table XIII shows the CFL limit of stability (independent of the grid) for the 4th-order Lagrange method and for the 4th-, 6th-, 8th-, and 12th-order Padé methods. As expected, increasing the accuracy of the spatial discretization decreases the limit in the CFL for the stability.

Equation (40) is the L^∞-norm of the spectral function of the operator $\mathcal{C}$. So, the stability limit coincides with that obtained by the von Neumann method. Note that it is not an immediate consequence of the Lyapunov theorem (see, for example, [1]), for $k = 0$ implies $|\lambda| = 1$.

A simple calculation shows that a fully discrete algorithm is von Neumann stable if $S_\sigma = \sigma_c \mathcal{F}(\mathcal{C})([0, 1/2]) \subset S$, where S is the stability region (see [29]). For RK methods we have $S_k = \mathrm{RK}_k^{-1}[B_1]$, $k \in \mathbb{N}$, where $B_1 = \{z \in \mathbb{C} : |z| \le 1\}$. For completeness, the stability zone for the 1st-, 2nd-, 3rd-, 4th-, 6th-, and 8th-order RK methods is shown in Fig. 11.

It is interesting to note that any centered method is unstable for the 1st- and 2nd-order RK methods. This follows at once from the fact that the region of stability of these methods is

TABLE XIII
CFL Stability Limit for the 4th-Order RK Method with Some Spatial Padé Discretization Schemes

	4th Lagrange	4th Padé	6th Padé	8th Padé	12th Padé
$\sigma_{c,\max}$	2.0612	$2\sqrt{2/3}$	1.4217	1.2829	1.1640

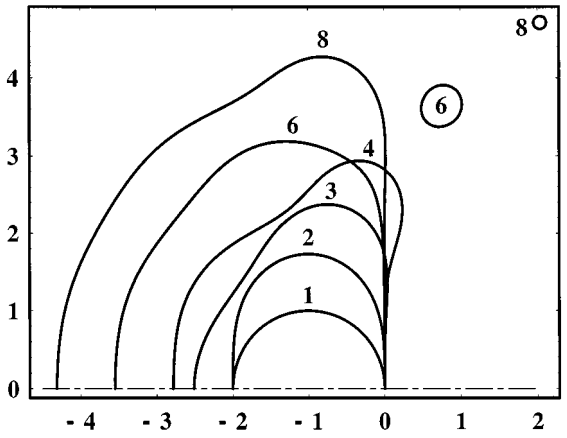

FIG. 11. Stability region for the RK methods of order 1, 2, 3, 4, 6, and 8.

tangent to the imaginary axis and that for any centered method $\Re(\mathcal{F}(\mathcal{C})) = 0$. Also interesting is the fact that for centered methods the 4th-order RK method provides an improved stability as compared to the 6th-order method. In other words, from the stability viewpoint, and for centered methods, it does not pay off to move from the 4th- to the 6th-order RK.

In Fig. 12 we plot S_σ for the 1st-order upwind scheme, the 3rd-order QUICK scheme, the 4th-order Lagrange scheme, and the 4th-order Padé scheme, all for $\sigma_c = 2\sqrt{2/3}$. At this CFL the QUICK scheme is stable for RK_4 (its stability limit for the 3rd RK is, within three digits of accuracy, 1.625, which is slightly smaller than $2\sqrt{2/3}$), the Lagrange method is also stable for RK_4, and, interestingly enough, with this CFL number the 1st-order upwind scheme is not stable even for a 4th-order RK method. The dissipation generated by this scheme, which helps to stabilize it, for instance, in the Euler explicit method, is excessive for RK_4. Also, the Padé method is more restrictive in CFL terms than is the Lagrange method. However, this gain in stability is more than compensated by the improved spectral resolution of the Padé method.

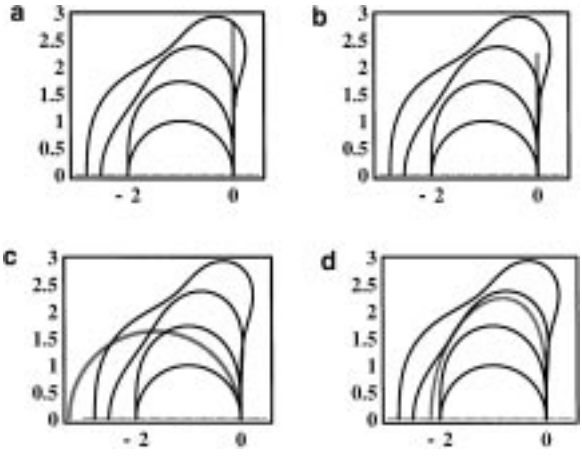

FIG. 12. S_σ for (a) 4th-order Padé, (b) 4th-order Lagrange, (c) 1st-order upwind, and (d) 3rd-order QUICK.

TABLE XIV
Diffusion CFL Stability Limit for the 4th-Order RK Method with Some Spatial CD Padé Discretization Schemes

σ_c	6th	8th	12th
0.5	0.2901	0.2849	0.2823
1	0.2407	0.1912	0.1422

We proceed with the stability limits of the discrete transport equations for the 4th RK with the CD Padé 6th order with periodic boundary conditions. Now, the semidiscrete equation is

$$\frac{d\bar{\phi}}{dt} = \frac{1}{\Delta t}(\sigma_c \mathcal{C}\bar{\phi} + \sigma_d \mathcal{D}\bar{\phi}), \tag{41}$$

where $\sigma_d = \nu \Delta t / h^2$ is the CFL number for diffusion, and

$$(\mathcal{D}\phi)_{i+1/2} = \phi'_{i+1} - \phi'_i, \qquad i = 0, \ldots, N-1$$

with ϕ' obtained with the CD Padé discretization method.

In Table XIV are listed the stability limits of the diffusion CFL number for the 4th-order RK method with CD 6th-, 8th-, and 12th-order CD Padé methods and for two values of the convection CFL number: $\sigma_c = 0.5$ and $\sigma_c = 1$. Again, as expected, increasing the order of the method decreases the maximum value of the diffusion CFL number, while for the same method a smaller σ_c improves the stability limit. Figure 13 shows S_σ for some limit combinations of σ_c and σ_d. Note that simultaneously using the limit values of $\sigma_c = 1.3304$ for pure convection with $\sigma_d = 0.2901$ for pure diffusion in a transport equation yields an

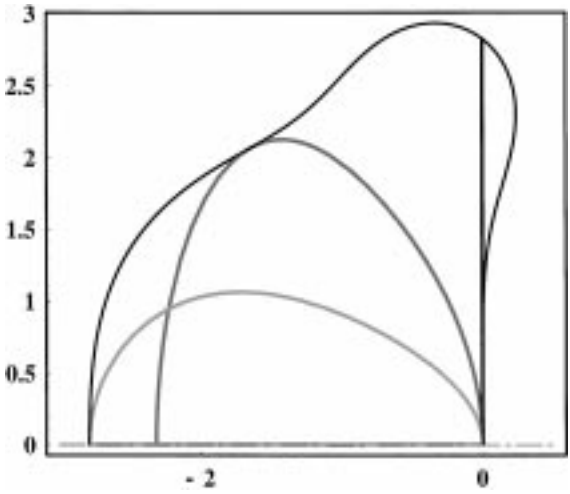

FIG. 13. Stability region for $\sigma_c = 1.3304, \sigma_d = 0$ (thick black); $\sigma_c = 1, \sigma_d = 0.2407$ (thick dark gray); $\sigma_c = 0.5, \sigma_d = 0.2901$ (thick gray); and $\sigma_c = 0, \sigma_d = 0.2901$ (horizontal thick light gray).

unstable method. In other words, the stability analysis for the transport equations cannot be undertaken separately.

We now proceed to the study of the effect of the boundary condition on stability. In [10, 14, 28] is introduced the notion of GKS stability and an analytical method to establish the GKS stability of discrete methods with boundary approximation taken into account. In [4] this technique is applied to some compact finite difference methods, together with several boundary approximations. In [5] the authors concluded that in practical computations only those schemes for which the semidiscrete equation is Lyapunov stable are of any usefulness for long-time integrations. Thus, we use the spectrum of the semidiscrete equations to study the effect of the boundary closure.

Let us start with the hyperbolic equation. First, we consider the effect of the order of the accuracy of the boundary closure on stability. The semidiscrete equation is formally the same as Eq. (37) with $\mathcal{C}$ modified in order to take the boundary closure into account. We remark that for the stability analysis it is sufficient to consider homogeneous boundary conditions. In fact, we can think of stability analysis as the study of the evolution of perturbations of the solution with a fixed boundary condition.

Figure 14 shows the spectrum that results from 4th-order Padé finite method closed at the outlet boundary with the 3rd- and the 4th-order schemes presented in the previous section. The grids comprise 32, 64, and 128 control volumes. The spectra are very similar, with the 3rd-order approximation closer to the imaginary axis. This finding for the Padé finite volume method is in clear contrast to the findings for the Padé finite difference method (see [4]). The latter is stable for the 3rd-order closure, but is unstable for the 4th. This is due to the fact that the finite difference approach requires a downwind extrapolation at the inlet and, as is well known, a downwind extrapolation is unstable. On the other hand, the finite volume method does not need any approximation at the inlet and thus avoids the potentially unstable mode.

As mentioned earlier a 6th-order or higher Padé method requires some approximation at points close to the boundaries. We consider the 6th-order Padé method with the following

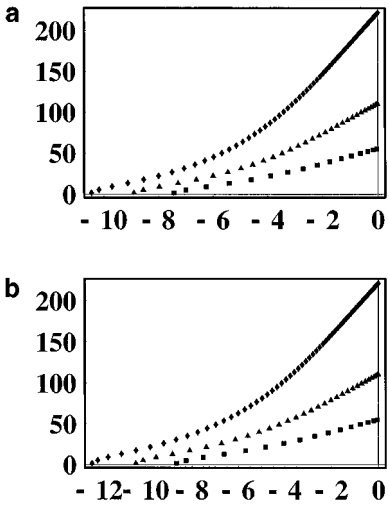

FIG. 14. Spectrum of the semidiscrete 4th-order Padé method in pure advection for (a) 3rd- and (b) 4th-order boundary conditions at outlet: (■) 32, (▲) 64, and (◆) 128 control volumes.

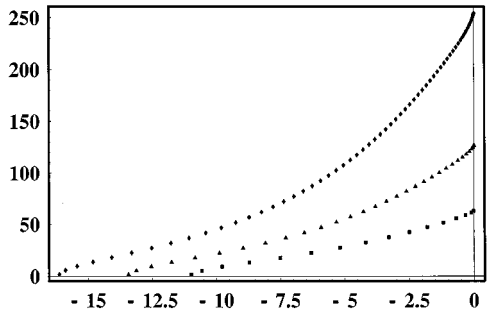

FIG. 15. Spectrum of the semidiscrete 6th-order Padé method in pure advection for (a) 5th- and (b) 6th-order boundary conditions at outlet: (■) 32, (▲) 64, and (◆) 128 control volumes.

6th-order approximation close to the boundaries:

1. At inlet

$$\frac{1}{8}\phi_0 + \phi_1 + \frac{3}{4}\phi_3 = \frac{43}{96}\bar{\phi}_{1/2} + \frac{41}{32}\bar{\phi}_{3/2} + \frac{5}{32}\bar{\phi}_{5/2} - \frac{1}{96}\bar{\phi}_{7/2}.$$

2. At outlet

$$5\phi_{N-1} + \phi_N = \frac{197}{60}\bar{\phi}_{N-1/2} + \frac{37}{10}\bar{\phi}_{N-3/2} - \frac{13}{10}\bar{\phi}_{N-5/2} + \frac{11}{30}\bar{\phi}_{N-7/2} - \frac{1}{20}\bar{\phi}_{N-9/2}.$$

The Padé method for the point $N-1$ is computed analogously to that for the point 1 at inlet.

Figure 15 shows the spectrum of the 6th-order Padé method with the boundary closure mentioned above. This figure corroborates the previous findings, namely, that the Padé finite volume method is stable for a boundary closure of the same order of accuracy as the inner points.

Because the effect of the boundary is to create some damping in the high-frequency modes, the limit value computed with the von Neumann analysis is still valid, and both approximations are stable with this limit. To verify this assertion, in Fig. 16 is shown the spectrum that results from the the explicit 4th-order RK and 4th-order Padé finite method, which is closed at the outlet boundary with the 4th-order scheme. The grids comprise 32 and 128 control volumes for $\sigma_c = 2\sqrt{2/3}$. This figure confirms the fact that the CFL limit obtained with the von Neumann analysis can be used for the advection problem with a prescribed inlet boundary condition.

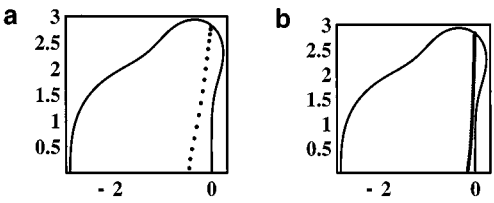

FIG. 16. Spectrum of the discrete 4th RK, 4th-order Padé method in pure advection for $\sigma_c = 2\sqrt{2/3}$: (a) 32 and (b) 128 control volumes.

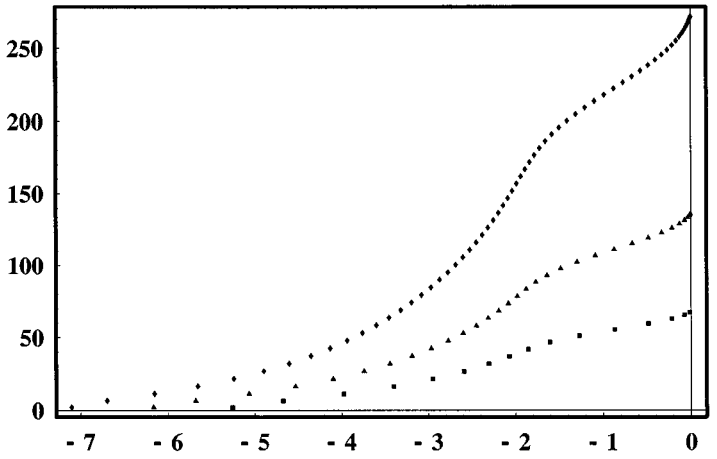

FIG. 17. Spectrum of the semidiscrete 6th-order Padé method in pure advection for the 6th-order boundary closure at outlet: (■) 32, (▲) 64, and (◆) 128 control volumes.

Finally, we study the effect of the boundary closure on the CD Padé methods for a general transport equation. We start with the critical limit of pure advection.

When both boundary conditions are of Dirichlet type the operator $\mathcal{C}$ is skew-adjoint, whence the eigenvalues are pure imaginary. Thus, we only need to consider the case with a combined Dirichlet and a von Neumann boundary condition. Figure 17 shows the resulting spectrum for the CD 6th-order Padé semidiscrete method with the 6th-order closure. The boundary condition corresponds to a Dirichlet boundary condition at inlet and a von Neumann at outlet.

Again, the finite volume method does not require any lowering of the order at the boundary. In this case the finite volume method represents a major gain in stability. Indeed, in [21] it is shown that the CD Padé finite difference method is unstable even for a boundary closure of 4th order of accuracy.

An interesting feature of the general transport equation is the effect of the boundary closure for the diffusion flux. Figure 18 shows the spectrum that results for the discrete 4th-order RK method with the 6th-order CD Padé method for the transport equations with Dirichlet BC at inlet and von Neumann BC at outlet. The spectrum has been calculated with the same values of CFL as in Fig. 13. For small CFL of diffusion the limit determined by the von Neumann analysis is valid. However, as diffusion becomes more important, the spectrum starts to display a reduction in the diffusion limit. For example, for $\sigma_c = 1$ the limit for the 6th-order closure is $\sigma_d = 0.2125$, which is smaller than the limit for periodic boundary conditions of $\sigma_d = 0.2407$. Contrary to the pure convection case, where the spectrum converges towards the spectral function, in the transport equation the effect of the boundary closure on diffusion is permanent (see Fig.18b). A possible explanation for this different behavior of the spectrum for pure advection and diffusion may lie in the fact that the former has a hyperbolic character while the latter has an elliptic character. Thus, the effect of the boundary is swept away as the grid is refined in the hyperbolic equation, while it is not in the elliptic one. Notwithstanding the reduction in the stability limit of σ_d, it should be stressed that it is necessary neither to reduce the order of accuracy at the boundary to ensure the stability of the method nor to enlarge the stencil close to the boundary, to accommodate stability.

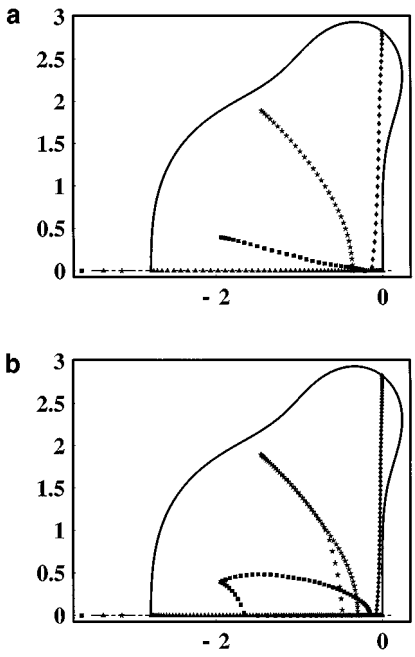

FIG. 18. Spectrum of the discrete 4th-order RK method with the 6th-order CD Padé method for the transport equations with Dirichlet BC at inlet and von Neumann BC at outlet for (a) 64 and (b) 128 control volumes: (◆) $\sigma_c = 1.3304$, $\sigma_d = 0$; (★) $\sigma_c = 1$, $\sigma_d = 0.2407$; (■) $\sigma_c = 0.5$, $\sigma_d = 0.2901$; and (▲) $\sigma_c = 0$, $\sigma_d = 0.2901$.

Remark. Equation (39) shows *en passant* that for an odd number of control volumes $\dim\ker(\mathcal{C}) = 1$, that is, only constant functions have a zero convective flux. In contrast, for an even number of control volumes $\dim\ker(\mathcal{C}) = 2$. Thus, in addition to the constant mode, we also have the so called "checkerboard" mode or the "odd–even" decoupling.

Using two first derivative operators to represent the Laplacian operator on a nonstaggered grid may lead to the checkerboard mode. A remedy for this, put forward in [30], is to use an approximation for the second derivative. However, as the previous observation shows, an alternative solution may be the use of an odd number of control volumes.

Remark. The limits of stability determined above also works for time-periodic solutions. Indeed, it is enough to observe that a time-periodic solution is a fixed point of the operator $(\mathrm{RK}_k)^l$, where $l \in \mathbb{N}$ is the discrete period, and that $\rho((\mathrm{RK}_k)^l) = (\rho((\mathrm{RK}_k))^l$.

Remark. For nonlinear problems the previous stability limits also hold true for the stability of fixed points or periodic orbits. Indeed, consider a general conservation law

$$\frac{\partial \phi}{\partial t} + \frac{\partial f(\phi)}{\partial x} = 0,$$

where f is some smooth function. Then, the linearized equation, for example, in the fixed point ϕ_0 can be written as

$$\left[\frac{d\delta\bar{\phi}}{dt}\right]_{i+1/2} + \frac{f'((\phi_0)_{i+1})\delta\phi_{i+1} - f'((\phi_0)_i)\delta\phi_i}{h} = 0.$$

Hence, for $\sigma_c < \sigma_{c,\max}$ the result follows at once from the Lyapunov theorem. As a final comment, we remark that in the nonlinear case, the Lyapunov theorem guarantees the convergence to the critical point only in a neighborhood of it, while in the linear case the convergence is for any point. In other words, the stable manifold is the whole space in the linear case, but in general, it is not so for the nonlinear case.

Remark. The above analysis considered constant coefficients in a uniform grid. The following argument suggests that for a sufficiently fine grid, the results above hold true, if applied locally.

We start with the nonconstant case. The conservation law in intrinsic form is

$$\frac{\partial \phi}{\partial t} + \mathsf{div}(c\phi) = 0.$$

Let $c^x \in C^\infty \cap L^\infty$ be a smooth limited function. Then, in a Cartesian grid it is

$$\frac{\partial \phi}{\partial t} + \frac{\partial c^x \phi}{\partial x} = 0$$

and in a general coordinate ξ,

$$\frac{\partial \phi}{\partial t} + \frac{1}{\sqrt{g}} \frac{\partial \sqrt{g} c^\xi \phi}{\partial \xi} = 0, \tag{42}$$

where g is the determinant of the metric in the coordinate ξ. Suppose $c^x \in C^\infty$ does not change sign. Then, in the coordinate ξ given by

$$\frac{\partial \xi}{\partial x} = \frac{1}{c}$$

Eq. (42) simplifies to

$$\frac{\partial \tilde{\phi}}{\partial t} + \frac{\partial \tilde{\phi}}{\partial \xi} = 0,$$

where $\tilde{\phi} = \sqrt{g}\phi$. The stability criterion for the previous equation is

$$\frac{\Delta t}{h_\xi} \geq \sigma_{c,\max}.$$

Using $h_\xi = h_x/c + \mathsf{o}(h_x)$ we find, for a sufficiently small grid, the following stability criterion in the "physical" grid:

$$\frac{\Delta t \|c\|_\infty}{h_x} \leq \sigma_{c,\max}.$$

The case when $c = 0$ is trivial. Consider it is not zero for some points and consider a δ-neighborhood of the complement of the support of cU. For sufficiently small δ the maximum of the speed occurs outside U. We can apply the previous result for each connected component of the complement of U. And since it is uniform with respect to δ we obtain the result by passing to the limit as $\delta \to 0$.

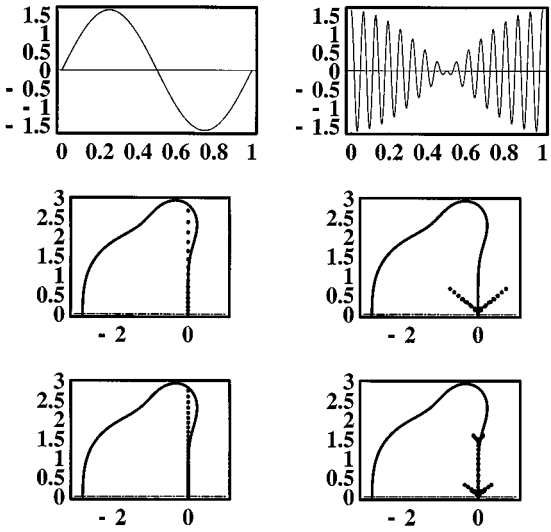

FIG. 19. Spectrum of the discrete 4th-order RK method with the 4th-order Padé method for pure advection and periodic boundary conditions for varying local CFL: (left) $\sigma_c(x) = \sigma_{c,\max} \sin(2\pi x)$, $x \in [0, 1]$ and (right) $\sigma_c(x) = \sigma_{c,\max}(\cos(32\pi x) + \cos(30\pi x))$, $x \in [0, 1]$.

To verify this remark we compute the spectrum of the 4th-order Padé method with a 4th RK method for time evolution for a pure convection problem with periodic boundary conditions and varying CFL numbers. Figure 19 shows the resulting spectra for two CFL profiles: $\sigma_c(x) = \sigma_{c,\max} \sin(2\pi x)$, $x \in [0, 1]$, and $\sigma_c(x) = \sigma_{c,\max}(\cos(32\pi x) + \cos(30\pi x))$, $x \in [0, 1]$. The first is a slowly varying function, while the second is a highly oscillatory profile with a "beating." Figure 19 shows that the spectrum converges toward S_σ and also that this convergence depends on the resolution of the monochromatic components of the profile.

In the case of general transport equations the results are valid if σ_d is sufficiently small. This is not so rare, and in fact, it is true for most convection-dominated problems. For example, for a Reynolds number of the order of 10^6 and a grid with about 100 control volumes we have $\sigma_d/\sigma_c \simeq 10^{-4}$.

The case of a nonuniform grid can be handled analogously by noting that, from Eq. (42), the nonuniform grid can be seen as a particular case of the nonconstant speed.

7. SUMMARY

A class of Padé finite volume methods for the evaluation of derivatives and interpolation have been presented and analyzed. From the analysis, the following conclusions can be drawn:

1. The use of the sliding averages in the finite volume formulation requires a smaller stencil in multi-dimensional problems, as compared with point values. Moreover, in time-dependent problems the former appears explicit, while the latter requires interpolation.

2. For pure convection, or pure diffusion problems, the standard (uncoupled) Padé method has the highest order of accuracy in the given stencil and requires less CPU time than the CD Padé interpolation. For general transport equations, the latter presents an improved spectral resolution and no significant additional cost.

3. Padé methods convey energy better than usual interpolation.

4. The spectral Padé method shows spectral-like resolution and exponential convergence.

5. Using the notion of reflection ratio it is shown that the order of the boundary closure, if lower than or equal to the order of accuracy of the inner points, determines the uniform order of accuracy of the Padé method.

6. The limits of stability computed by the von Neumman procedure are valid for the pure advection problem or for convection-dominated problems. Moreover, contrary to the finite difference method, the finite volume method requires neither the lowering of the order of accuracy at the boundary closure nor any change in the stencil of the point at the boundary to accommodate stability.

ACKNOWLEDGMENTS

The author is grateful to Professor. J. C. F. Pereira for bringing the problem of the Padé finite volume interpolation to his attention, to Eng° J. M. C. Pereira for many useful discussions, and to the referees for raising a number of issues left unaddressed in the first version of the paper.

REFERENCES

1. D. V. Anosov, S. Kh. Aranson, V. I. Arnold, I. U. Bronshtein, V. Z. Grines, and Yu. S. Il'yashenko, *Ordinary Differential Equations and Smooth Dynamical Systems* (Springer-Verlag, Berlin, 1997).
2. C. Basdevant, M. Deville, P. Haldenwang, J. M. Lacroix, J. Ouazzani, R. Peyret, P. Orlandi, and A. T. Patera, Spectral and finite-difference solutions of the burgers equation, *Comput. Fluids* **14**(1), 23 (1986).
3. E. O. Brigham, *The Fast Fourier Transform* (Prentice Hall, Englewood Cliffs, NJ, 1974).
4. M. H. Carpenter, D. Gottlieb, and S. Abarbanel, The stability of numerical boundary treatments for compact high-order finite-difference schemes, *J. Comput. Phys.* **108**, 272 (1993).
5. M. H. Carpenter, D. Gottlieb, and S. Abarbanel, Time-stable boundary conditions for finite-difference schemes solving hyperbolic systems: Schemes, *J. Comput. Phys.* **111**, 220 (1994).
6. J. Sebastião e Silva, Sur l' espace de fonctions holomorphes a croissance lente a droite, *Portugal. Math.* **17**, 435 (1958).
7. J. H. Ferziger and M. Perić, *Computational Methods for Fluid Dynamics* (Springer-Verlag, Berlin, 1997).
8. D. Gaitonde and J. S. Shang, Optimized compact-difference-based finite-volume schemes for linear wave phenomena, *J. Comput. Phys.* **138**, 617 (1997).
9. D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods* (SIAM, Philadelphia, 1977).
10. B. Gustafsson, H.-O. Kreiss, and A. Sundström, Stability theory of difference approximations for mixed initial boundary value problems, II, *Math. Comput.* **26**(119), 649 (1972).
11. A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy, Uniformly high order accurate essentially non-oscillatory schemes, III, *J. Comput. Phys.* **71**, 231 (1987).
12. C. Hirsch, *Numerical Computation of Internal and External Flows* (Wiley, West Sussex, 1994).
13. F. John, *Partial Differential Equations*, Applied Mathematical Sciences, Vol. 1, 4th ed. (Springer-Verlag, New York, 1982).
14. H.-O. Kreiss, Stability theory of difference approximations for mixed initial boundary value problems, I, *Math. Comput.* **22**, 703 (1968).
15. H. C. Ku, R. S. Hirsh, and T. D. Taylor, A pseudospectral method for solution of the three-dimensional incompressible Navier–Stokes equations, *J. Comput. Phys.* **70**, 439 (1987).
16. L. Landau and E. Lifchitz, *Mécanique des Fluides*, Physique Théorique, Vol. 6, 2nd ed. (Mir, Moscow, 1989).
17. S. Lang, *Real and Functional Analysis*, Graduate Texts in Mathematics, Vol. 142, 3rd ed. (Springer-Verlag, New York, 1993).

18. S. K. Lele, Compact finite-difference schemes with spectral-like resolution, *J. Comput. Phys.* **103**, 16 (1992).
19. B. P. Leonard, A stable and accurate convective modelling procedure based on quadratic upstream interpolation, *Comput. Methods Appl. Mech. Eng.* **19**, 59 (1979).
20. Z. Lilek and M. Perić, A fourth-order finite-volume method with collocated variable arrangement, *Comput. Fluids* **24**(3), 239 (1995).
21. K. Mahesh, A family of high order finite difference schemes with good spectral resolution, *J. Comput. Phys.* **145**, 332 (1998).
22. B. Mattiussi, An analysis of finite-volume, finite-element, and finite-difference methods using some concepts from algebraic topology, *J. Comput. Phys.* **133**(2), 289 (1997).
23. S. A. Orzag, Spectral methods for problems in complex geometries, *J. Comput. Phys.* **37**, 70 (1980).
24. T. J. Poinsot and S. K. Lele, Boundary conditions for direct simulations of compressible viscous flows, *J. Comput. Phys.* **101**, 104 (1992).
25. L. Schwartz, *Théorie des Distribuitions* (Hermann, Paris, 1966).
26. F. Spotz, *High Order Compact Finite Difference Schemes for Computational Mechanics* (Ph.D. thesis, University of Texas at Austin, 1995).
27. J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis* (Springer-Verlag, New York, 1980).
28. J. C. Strikwerda, Initial boundary value problems for the method of lines, *J. Comput. Phys.* **34**, 94 (1980).
29. R. Vichnevetsky and J. B. Bowles, *Fourier Analysis of Numerical Approximations of Hyperbolic Equations*, Studies in Applied Mathematics (SIAM, Philadelphia, 1982).
30. R. V. Wilson, A. O. Demuren, and M. H. Carpenter, Higher-order compact schemes for numerical simulation of incompressible flows, ICASE Report 98-13, ICASE, Langley Reseach Center, 1998.
31. I. Yavneh, Analysis of a fourth-order compact scheme for convection-diffusion, *J. Comput. Phys.* **133**, 361 (1997).