Available online at www.sciencedirect.com
SCIENCE DIRECT®
Journal of Computational Physics 198 (2004) 106–130
JOURNAL OF COMPUTATIONAL PHYSICS
www.elsevier.com/locate/jcp

Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods ☆

Mark Ainsworth

Department of Mathematics, Strathclyde University, 26 Richmond Street, Livingstone Tower, Glasgow G1 1XH, Scotland, UK

Received 25 June 2003; received in revised form 5 January 2004; accepted 6 January 2004
Available online 3 March 2004

Abstract

The dispersive and dissipative properties of *hp* version discontinuous Galerkin finite element approximation are studied in three different limits. For the small wave-number limit $hk \to 0$, we show the discontinuous Galerkin gives a higher order of accuracy than the standard Galerkin procedure, thereby confirming the conjectures of Hu and Atkins [J. Comput. Phys. 182 (2) (2002) 516]. If the mesh is fixed and the order p is increased, it is shown that the dissipation and dispersion errors decay at a super-exponential rate when the order p is much larger than hk. Finally, if the order is chosen so that $2p + 1 \approx \kappa hk$ for some fixed constant $\kappa > 1$, then it is shown that an exponential rate of decay is obtained.

AMS: 65N50; 65N15; 65N30; 35A40; 35J05

Keywords: Discrete dispersion relation; High wave number; Discontinuous Galerkin approximation; *hp*-finite element method

1. Introduction

The numerical propagation of waves poses a significant challenge in scientific computation. Many alternative approaches have been explored in the quest for a stable method that can efficiently resolve the wave without excessive dissipation or dispersion, particularly in the context of high frequency applications. Some of the more promising domain based approaches involve the use of higher order schemes including spectral element methods [8,10], higher order standard Galerkin finite element methods [3,17,25] and, more recently, higher order discontinuous Galerkin finite element methods [2,4–6,12,13,27].

☆ This paper is dedicated to Donald Kershaw on occasion of his seventy fifth birthday.

E-mail address: M.Ainsworth@strath.ac.uk (M. Ainsworth).

0021-9991/$ - see front matter
doi:10.1016/j.jcp.2004.01.004

The study of the dispersive and dissipative properties of a method provides insight into the ability of a method to accurately propagate a wave. Indeed, the order of accuracy of the discrete dispersion relation is often used as a basis for ranking different methods.

Higher order standard Galerkin finite element schemes for the Helmholtz equation in one space dimension were studied by Thompson and Pinsky [25] and Ihlenburg and Babuška [17,18]. Recently [1], sharp estimates were obtained for the dispersive behaviour of higher order elements for the Helmholtz equation in multi-dimensions using tensor product elements.

The dispersive properties of higher order discontinuous Galerkin finite element methods have been studied in [14–16,23]. In particular, Hu and Atkins [14] examine the dispersion properties of the approximation of the scalar advection equation in one dimension in the limit $hk \to 0$, for methods of order up to 16 using a computer algebra approach. On the basis of the computations, it was conjectured that the discrete wave-numbers are related to certain Padé approximants and that the dispersion relation for an Nth order method is accurate to order $2N+3$ in hk for the dispersion error and order $2N+2$ for the dissipation error. These orders of accuracy exceed those for the standard Galerkin finite element procedure [25].

The present work is concerned with the analysis of the dispersive behaviour of high order discontinuous Galerkin finite element methods. One by-product is a proof of correctness of the conjectures of Hu and Atkins (see Theorem 2). Moreover, Theorem 2 gives the coefficient of the leading terms in the error which, in view of the fact that in practical computations hk is finite, may be viewed as being of at least as much practical relevance as the order of decay. It is found that the leading coefficient decreases rapidly with increasing order N suggesting it may be advantageous to increase the order N whilst maintaining a fixed mesh.

This idea is pursued in Theorem 3 where it is shown that as the order N is increased, the dissipation and dispersion errors pass through three different phases depending on the size of N relative to hk. In the unresolved regime where $2N+1 < hk - \mathrm{o}(hk)^{1/3}$, the error oscillates without decay as the order is increased. At the opposite extreme, if the order is large, specifically $2N+1 > hk + \mathrm{o}(hk)^{1/3}$, then the error reduces at a super-exponential rate. The error decreases at an algebraic rate $\mathcal{O}(N^{-1/3})$ in the transition zone between these extremes.

The super-exponential rate of convergence in the resolved regime, where $2N+1 > hk + \mathrm{o}(hk)^{1/3}$, means that it is unnecessary to increase the order N much beyond this threshold. Instead, a practical alternative consists of tracking the envelope where the super-exponential phase begins by choosing the order of approximation so that $2N+1 \approx \kappa hk$ for some fixed constant $\kappa > 1$. In Theorem 4, we prove that this approach results in an exponential accurate discrete dispersion relation.

It is illuminating to compare these results with those for the continuous Galerkin finite element method analysed in [1]. The nature and the analysis of the discrete dispersion relation is quite different in the present situation, and this is reflected by the fact that the discontinuous Galerkin method has a higher order of accuracy in the limit $hk \to 0$. On the other hand, in the limit as $N \to \infty$, the threshold where the method resolves the wave is identical to that for the continuous Galerkin method despite the fact that the argument is completely different. This means that the better dispersive behaviour of the discontinuous Galerkin method in the limit $hk \to 0$ fails to carry through to the limit $N \to \infty$.

The remainder of this paper is organised as follows. We begin by describing the model problem and the details of the discontinuous Galerkin discretisation, and then give a detailed description of the theoretical results along with supporting numerical evidence. Section 3 is devoted to the study of the errors in certain types of Padé approximants of the exponential with particular attention to the situation where the order of the approximant is comparable to the argument, and where both are large. The link between the dispersive behaviour and the Padé approximants is established in the following section where we study a certain eigenvalue problem. We conclude with the proofs of the results stated in Section 2.

2. Description of DGFEM and its dispersive properties

2.1. Model problem

Consider the linear advection equation in $\mathbb{R}^d$, $d \in \mathbb{N}$,

$$u_t + \boldsymbol{\alpha} \cdot \operatorname{grad} u = 0 \tag{1}$$

subject to appropriate initial conditions. The advective field $\boldsymbol{\alpha} \in \mathbb{R}^d$ is assumed constant and we orient our Cartesian coordinate system so that $\boldsymbol{\alpha}$ has non-negative components. It is well-known that this equation admits non-trivial solutions of the form

$$u(\boldsymbol{x}, t) = c\, \mathrm{e}^{\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{x} - \omega t)}, \tag{2}$$

where ω is a prescribed frequency and $\boldsymbol{k} \in \mathbb{R}^d$ is the corresponding wave-vector. Inserting this expression into Eq. (1) and simplifying shows that the equation admits a non-trivial solution provided that ω and $\boldsymbol{k}$ satisfy the *dispersion relation*

$$\omega = \boldsymbol{\alpha} \cdot \boldsymbol{k}. \tag{3}$$

Obviously, the sinusoidal solution u is also a Bloch-wave [20]: i.e. for all $h\boldsymbol{m} \in \mathbb{R}^d$ and $\tau \in \mathbb{R}$,

$$u(\boldsymbol{x} + h\boldsymbol{m}, t + \tau) = \mathrm{e}^{\mathrm{i}(h\boldsymbol{k}\cdot\boldsymbol{m} - \omega\tau)} u(\boldsymbol{x}, t), \quad \forall \boldsymbol{x} \in \mathbb{R}^d, \ \ t \in \mathbb{R}. \tag{4}$$

One repercussion of discretisation of the continuous problem is that the numerical scheme usually admits a non-trivial Bloch-wave satisfying condition (4) where, however, the exact wave-vector $\boldsymbol{k}$ is replaced by a discrete wave-vector $\tilde{\boldsymbol{k}}$. The discrete wave-vector $\tilde{\boldsymbol{k}}$ provides valuable information on the ability of a numerical scheme to propagate wave-like solutions. For instance, if the real part of the component of $\tilde{\boldsymbol{k}}$ in some direction differs from the corresponding component of $\boldsymbol{k}$, then the numerical approximation will exhibit a phase-lag or phase-lead compared with the true solution. Likewise, dissipative and instability effects arise when $\tilde{\boldsymbol{k}}$ has imaginary components.

2.2. Discontinuous Galerkin discretisation

The discontinuous Galerkin finite element discretisation (DGFEM) of (1) is constructed on a partitioning of the computational domain into non-overlapping elements. Although rather general partitions may be employed for DGFEM, our chief interest here lies in investigating the ability of the numerical scheme to propagate waves through regions of free space remote from domain boundaries, where one would generally use a highly structured mesh. For this reason we shall confine our attention to uniform partitions of $\mathbb{R}^d$ consisting of square, or cubic, elements of size $h > 0$, whose sides are aligned with the coordinate axes and whose nodes are located at the points $h\mathbb{Z}^d$.

For $N \in \mathbb{N}$, let $\mathbb{P}_N$ denote the usual space of polynomials in one variable of degree at most N. An Nth order DGFEM seeks an approximate solution u^{DG} whose restriction to each element K belongs to the tensor product space $\mathbb{P}_N^d$, but does not require the approximation to be continuous at element interfaces. Instead, continuity is enforced in a weak sense between neighbouring elements K and K' through the use of a numerical flux function $\tilde{\sigma}_\gamma$ defined on the interface $\partial K \cap \partial K'$. The true flux on the interface in the direction of the unit normal $\boldsymbol{n}$ to the interface is given by

$$\sigma(\boldsymbol{n}, u) = \boldsymbol{n} \cdot \boldsymbol{\alpha} u.$$

The numerical flux $\tilde{\sigma}_\gamma(\boldsymbol{n}_K, u^{\mathsf{DG}})$ from element K to element K' in the direction of the unit outward normal $\boldsymbol{n}_K$ is defined, for given $\gamma \in \mathbb{R}$, by the rule

$$\tilde{\sigma}_\gamma(\boldsymbol{n}_K, u^{\mathsf{DG}}) = \Lambda_\gamma^+(\boldsymbol{n}_K)u_K^{\mathsf{DG}} + \Lambda_\gamma^-(\boldsymbol{n}_K)u_{K'}^{\mathsf{DG}} \quad \text{on } \partial K \cap \partial K',$$

where $\Lambda_\gamma^\pm$ is defined by

$$\Lambda_\gamma^\pm(\boldsymbol{n}) = \tfrac{1}{2}(\boldsymbol{n}\cdot\boldsymbol{\alpha} \pm \gamma|\boldsymbol{n}\cdot\boldsymbol{\alpha}|).$$

The quantity γ is often referred to as an *upwinding parameter*. The function $\Lambda_\gamma^\pm$ satisfies two important properties:

$$\Lambda_\gamma^+(\boldsymbol{n}) + \Lambda_\gamma^-(\boldsymbol{n}) = \boldsymbol{n}\cdot\boldsymbol{\alpha} \tag{5}$$

and

$$\Lambda_\gamma^\pm(-\boldsymbol{n}) = -\Lambda_\gamma^\mp(\boldsymbol{n}). \tag{6}$$

The former property ensures that $\tilde{\sigma}_\gamma(\boldsymbol{n}, u) = \sigma(\boldsymbol{n}, u)$ for all γ, while the latter property implies that the flux from element K' to K balances out the flux in the opposite direction from element K to K':

$$\tilde{\sigma}_\gamma(\boldsymbol{n}_K, u^{\mathsf{DG}}) = -\tilde{\sigma}_\gamma(\boldsymbol{n}_{K'}, u^{\mathsf{DG}}) \quad \text{on } \partial K \cap \partial K'.$$

The preparations are now complete for the definition of the DGFEM. Using Eq. (1), we find that the true solution u satisfies

$$0 = \int_K v u_t - \int_K u\boldsymbol{\alpha}\cdot\operatorname{grad} v + \int_{\partial K} v\sigma(\boldsymbol{n}_K, u)$$

for all sufficiently smooth test functions v. The DGFEM approximation is defined on the basis of this relation by replacing the true flux with the numerical flux, and then requiring that for every element K: $u_K^{\mathsf{DG}} \in \mathbb{P}_N^d$

$$0 = \int_K v u_{K,t}^{\mathsf{DG}} - \int_K u_K^{\mathsf{DG}}\boldsymbol{\alpha}\cdot\operatorname{grad} v + \int_{\partial K} v\tilde{\sigma}_\gamma(\boldsymbol{n}_K, u^{\mathsf{DG}}) \quad \forall v \in \mathbb{P}_N^d. \tag{7}$$

For present purposes, it is more convenient to work with the equivalent statement

$$0 = \int_K v\left(u_{K,t}^{\mathsf{DG}} + \boldsymbol{\alpha}\cdot\operatorname{grad} u_K^{\mathsf{DG}}\right) + \int_{\partial K} v\Lambda_\gamma^-(\boldsymbol{n}_K)(u_{K'}^{\mathsf{DG}} - u_K^{\mathsf{DG}}) \quad \forall v \in \mathbb{P}_N^d, \tag{8}$$

which is obtained from (7) by integrating by parts and using property (5) to simplify the resulting contributions from the boundary terms.

2.3. Dispersive behaviour of DGFEM

We turn now to our study of the dispersive behaviour of DGFEM. The next result describes the properties of the discrete wave-vector for the DGFEM:

Theorem 1. *For $h > 0$ and $N \in \mathbb{N}$, consider the Nth order DGFEM on a grid $h\mathbb{Z}^d$ used in conjunction with the numerical flux function $\tilde{\sigma}_\gamma$. If $\omega \in \mathbb{R}$ and $\boldsymbol{k} \in \mathbb{R}^d$ satisfy the continuous dispersion relation $\omega = \boldsymbol{\alpha}\cdot\boldsymbol{k}$, then there exists a corresponding discrete Bloch-wave solution u^{DG} satisfying* (7) *and, for all $\boldsymbol{m} \in \mathbb{Z}^d$ and $\tau \in \mathbb{R}$,*

$$u^{\mathsf{DG}}(\boldsymbol{x} + h\boldsymbol{m}, t + \tau) = \mathrm{e}^{\mathrm{i}(h\tilde{\boldsymbol{k}}\cdot\boldsymbol{m} - \omega\tau)}u^{\mathsf{DG}}(\boldsymbol{x}, t), \quad \forall \boldsymbol{x} \in \mathbb{R}^d, \ \ t \in \mathbb{R}. \tag{9}$$

Moreover, each component $\tilde{\boldsymbol{k}}_\ell$ of the discrete wave-vector $\tilde{\boldsymbol{k}}$ may take one of two possible values corresponding to either a physical mode, $\mathrm{e}^{\mathrm{i}h\tilde{\boldsymbol{k}}_\ell} \approx \mathrm{e}^{\mathrm{i}hk_\ell}$, or a spurious mode

$$e^{i\tilde{\boldsymbol{k}}h_\ell} \approx (-1)^{N+1}\frac{1+\gamma}{1-\gamma}\frac{H_N^*}{H_N}e^{-ihk_\ell}, \quad \gamma \neq 1, \tag{10}$$

where $H_N = {}_1F_1(-N;-2N-1;-ihk_\ell)$ (*see* (20)) *and* $*$ *denotes complex conjugation. The relative error* ρ_N *is the same in both cases,*

$$\rho_N = \frac{(1-\gamma)H_N e^{ihk_\ell}\mathscr{E}_N + (-1)^{N+1}(1+\gamma)H_N^* e^{-ihk_\ell}\mathscr{E}_N^*}{(1-\gamma)H_N e^{ihk_\ell} + (-1)^N(1+\gamma)H_N^* e^{-ihk_\ell}} + \mathcal{O}(|\mathscr{E}_N|^2), \tag{11}$$

where $\mathscr{E}_N$ *is the relative error in the* $[N+1/N]$*-Padé approximant to* e^{ihk_ℓ}.

The proofs of this and the remaining results stated in this section are deferred to Section 5. Theorem 1 gives a complete description of the discrete wave-vector in terms of the quantity ρ_N. In turn, ρ_N is related to the relative error in certain Padé approximants to the exponential, which we shall study in detail in Section 3. The dependence on the upwind parameter γ is given explicitly provided $\gamma \neq 1$. In the event that γ is chosen to be unity, then the spurious mode (10) will be absent.

2.4. Small wave number $h\boldsymbol{k} \ll \boldsymbol{1}$

Suppose that $\boldsymbol{k} \in \mathbb{R}^d$ satisfies the dispersion relation (3) for the continuous problem. We shall use Theorem 1 to study the corresponding discrete wave-vector $\tilde{\boldsymbol{k}}$ in the computational regime where components k_ℓ of the wave-vector $\boldsymbol{k}$ are of moderate size. More specifically we shall assume that, for mesh-sizes h in the range of practical computation, the frequency is sufficiently moderate so that every component of $h\boldsymbol{k}$ may be regarded as being small. Although dispersion analyses are often performed under this kind of assumption their relevance to high frequency applications, where $h\boldsymbol{k}$ is finite, is limited. Nevertheless, one often sees competing numerical schemes ranked on the basis of their order of accuracy in this limit.

Thanks to Theorem 1, it suffices to consider the relative error ρ_N for a general component hk_ℓ of the wave-vector $h\boldsymbol{k}$. Here, and in what follows, we shall omit the subscript ℓ from k_ℓ and $\tilde{\boldsymbol{k}}_\ell$ in cases where confusion is unlikely to arise. The next result gives the leading term in the asymptotic expansion of the relative error ρ_N in terms of hk in terms of the order N and the parameter γ:

Theorem 2. *Let* $N \in \mathbb{N}$ *and suppose* $hk \ll 1$*, and define*

$$Q_N(s) = s + ihk(N+1)\left[\frac{s^2}{2N+1} - \frac{1}{2N+3}\right]. \tag{12}$$

1. *If* $\gamma \neq 0$*, then*

$$\rho_N \approx \frac{1}{2}(hk)^{2N+2}\left[\frac{N!}{(2N+1)!}\right]^2 Q_N\left(\gamma^{(-1)^N}\right) \tag{13}$$

2. *If* $\gamma = 0$*, then*

$$\rho_N \approx \frac{i}{2}\left[\frac{N!}{(2N+1)!}\right]^2 \begin{cases} -(hk)^{2N+3}\frac{N+1}{2N+3}, & N \text{ even}, \\ (hk)^{2N+1}\frac{2N+1}{N+1}, & N \text{ odd}. \end{cases} \tag{14}$$

Fig. 1 shows the real and imaginary components of the actual relative error in the case $\gamma = \frac{1}{2}$, along with the theoretically predicted orders of decay. Corresponding results for the exceptional case $\gamma = 0$ are also shown along with the theoretical predictions. For $hk \ll 1$, the relative error satisfies

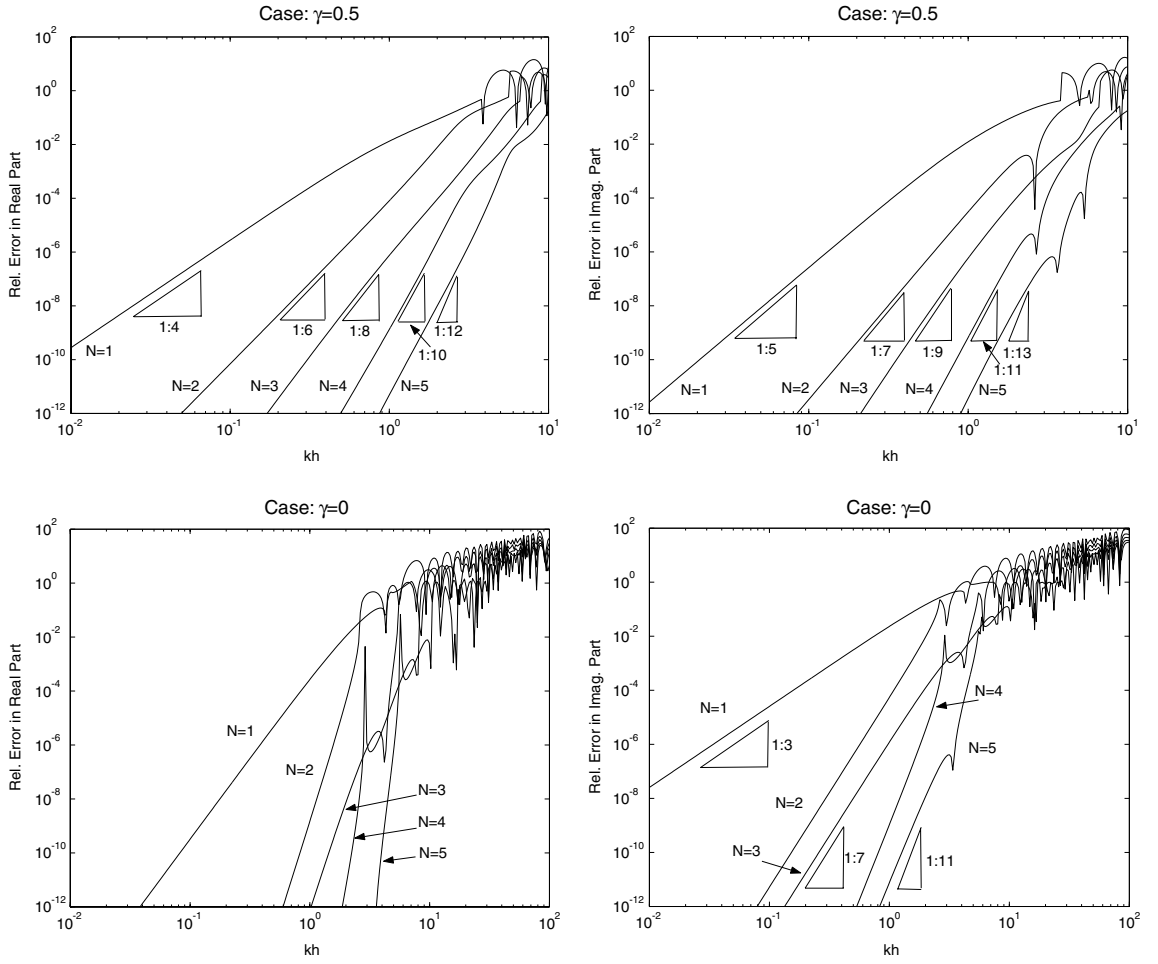

Fig. 1. Real and imaginary parts of relative error ρ_N for the approximation of the physical mode e^{ihk} using methods of order $N = 1, \ldots, 5$ for $\gamma = 0.5$ and the exceptional case $\gamma = 0$. The asymptotic rates of convergence predicted in Theorem 2 are indicated.

$$\rho_N = \frac{e^{ihk} - e^{ih\tilde{k}}}{e^{ihk}} \approx ih(k - \tilde{k}),$$

and hence, in the usual case where $\gamma \neq 0$, Theorem 2 shows that the *dispersion* error is

$$\Re(h\tilde{k}) - \Re(hk) \approx \frac{(hk)^{2N+3}}{2}\left[\frac{N!}{(2N+1)!}\right]^2\left\{\frac{N+1}{2N+1}\gamma^{2(-1)^N} - \frac{N+1}{2N+3}\right\},$$

while the *dissipation* error is

$$\Im(h\tilde{k}) \approx \frac{(hk)^{2N+2}}{2}\left[\frac{N!}{(2N+1)!}\right]^2\gamma^{(-1)^N},$$

thereby proving the conjectures of Hu and Atkins [14, Eqs. (41) and (42)]. In view of the fact that in practical computations *hk* is finite, the information provided by Theorem 2 on the coefficient of the leading

term in the error is of at least as much practical relevance as the order of approximation. The fact that the leading coefficient decreases rapidly with increasing order N suggests (though of course does not prove) that there may be advantages in keeping the mesh-size fixed and increasing the order N.

2.5. Large order N and large wave number kh

Motivated by the results of the previous section, we now investigate the behaviour of the relative error ρ_N in the case where the mesh-size h is fixed (so that the value of hk may be large) and the order N of the method is increased. Fig. 2 shows the real and imaginary parts of the actual relative error ρ_N as the order N is increased, for a range of wave-numbers. The numerical results indicate that as the order N is increased, the behaviour of the error passes through three different phases depending on the size of N relative to hk. Firstly, in the pre-asymptotic regime where $2N+1 < hk - \mathrm{o}(hk)^{1/3}$, the order is inadequate to resolve the wave and the relative error tends to oscillate without decay as the order is increased. At the opposite ex-

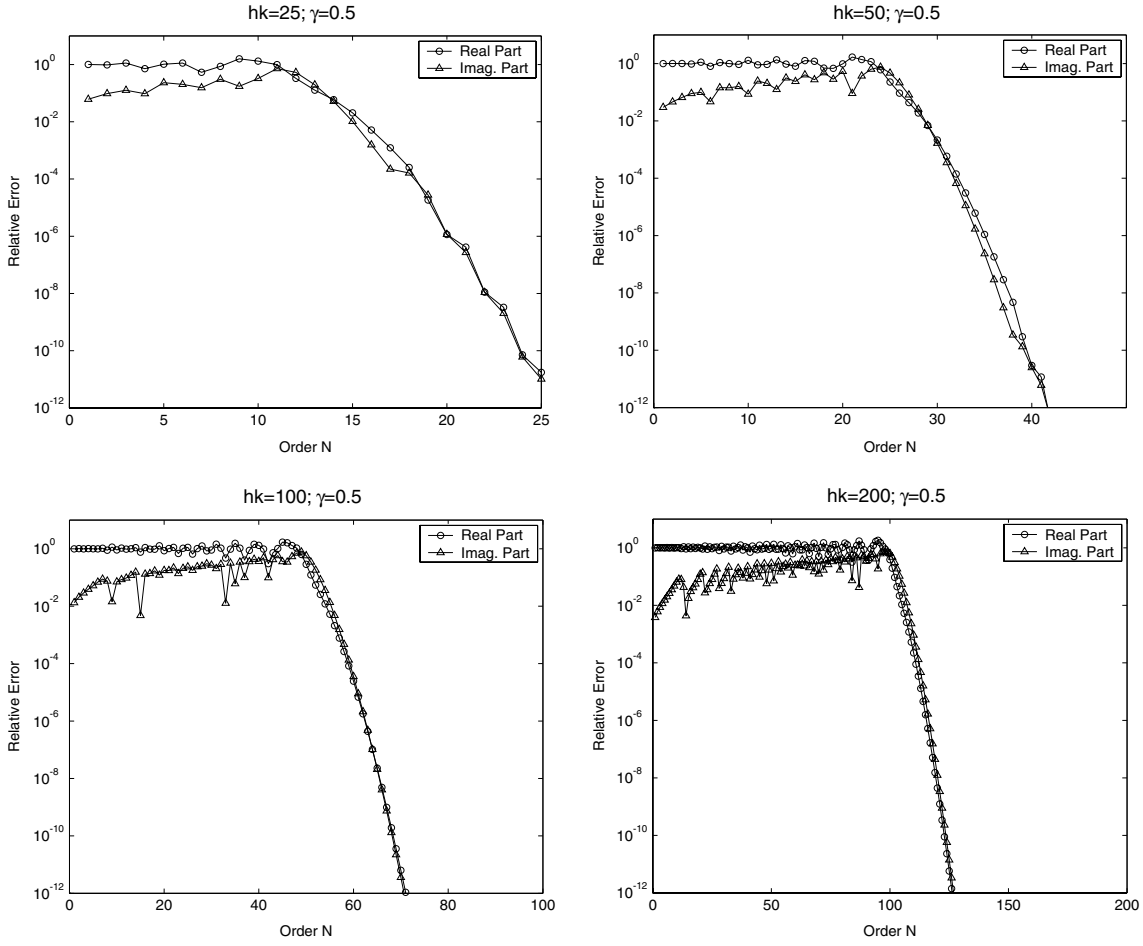

Fig. 2. Real and imaginary parts of relative error ρ_N for the approximation of the physical mode $\mathrm{e}^{\mathrm{i}hk}$ for $hk = 25$, 50, 100, 200 and $\gamma = 0.5$. Observe the super-exponential rate of decay once the order N exceeds the threshold $2N+1 > hk + \mathrm{o}(hk)^{1/3}$ as predicted in Theorem 3(1).

treme, if the order N is large compared with hk, i.e. $2N+1 > hk + \mathrm{o}(hk)^{1/3}$, then the error reduces at a super-exponential rate. The transition zone between these two extremes occurs when the order N lies in the relatively narrow range where $hk - \mathrm{o}(hk)^{1/3} < 2N+1 < hk + \mathrm{o}(hk)^{1/3}$. The following result shows that the behaviour observed in the particular cases studied in Fig. 2 is true in general, and that in the transition region, the error is of order unity but decreases at an algebraic rate $N^{-1/3}$.

Theorem 3. *Let $N \in \mathbb{N}$, and define*

$$\Upsilon_N(hk) = \frac{(1-\gamma)\mathrm{e}^{\mathrm{i}(hk+\psi_N)} + (-1)^{N+1}(1+\gamma)\mathrm{e}^{-\mathrm{i}(hk+\psi_N)}}{(1-\gamma)\mathrm{e}^{\mathrm{i}(hk+\psi_N)} + (-1)^{N}(1+\gamma)\mathrm{e}^{-\mathrm{i}(hk+\psi_N)}}, \tag{15}$$

where $\psi_N = \arg {}_1\mathrm{F}_1(-N;-2N-1;-\mathrm{i}hk)$. As the order N is increased relative to hk, the relative error ρ_N passes through three distinct phases:

1. *if $2N+1 < hk - C(hk)^{1/3}$, then ρ_N oscillates but does not decay as N is increased;*
2. *if $hk - \mathrm{o}(hk)^{1/3} < 2N+1 < hk + \mathrm{o}(hk)^{1/3}$, then ρ_N decays algebraically at a rate $\mathcal{O}(N^{-1/3})$,*
3. *if $2N+1 \gg hk$, then ρ_N decays at a super-exponential rate as $N \to \infty$,*

$$\rho_N \approx -\left(\Upsilon_N(hk) - \frac{\mathrm{i}hk}{2N+3}\right)\left[\frac{ehk}{2\sqrt{(2N+1)(2N+3)}}\right]^{2N+2}. \tag{16}$$

Theorem 3 also gives sharp estimates for the thresholds on (a) the size of the order N, in terms of hk, beyond which the wave is resolved and the error begins to decay, and in addition, (b) the value of h, in terms of N and k, below which the wave is resolved. The main difference between the two approaches lies not in the level of resources needed to resolve the wave, but in the rate at which the relative error decays once the thresholds are reached. Decreasing h gives algebraic rate of decay, while increasing N is superior giving a super-exponential rate of decay.

2.6. Exponential convergence on the envelope $2N+1 \approx hk$

Although most analyses of dispersive behaviour are performed under the assumption that $h\boldsymbol{k} \ll \mathbf{1}$, many applications occur at high frequencies for which reducing the mesh-size to this extent is simply not a viable practical proposition. In practice, computational simulations of high frequency phenomena are generally performed on the envelope where $h\boldsymbol{k}$ is of moderate size, but by no means vanishingly small. In other words, *the range of frequencies studied in numerical simulations is often dictated by the smallest mesh-size h that can be resolved by the available computational resources, rather than by the frequencies of physical interest.* As more powerful computational hardware becomes available, meaning a smaller mesh-size h becomes feasible, the range of simulated frequencies is increased so that $h\boldsymbol{k}$ effectively remains constant.

We have already seen that increasing the order N on a fixed mesh is more effective than reducing the mesh-size h. The super-exponential rate of convergence in the resolved regime, where $2N+1 > hk + \mathrm{o}(hk)^{1/3}$, means that it is inefficient to increase the order N much beyond this threshold. A more practical alternative is to work on the envelope of the region where the super-exponential convergence sets in. Thus, to resolve problems where $hk \gg 1$, one could adopt a strategy whereby the order is chosen so that $2N+1 \approx \kappa hk$ for some fixed constant $\kappa > 1$.

The analysis of this type of procedure is rather more delicate than the situations considered earlier, requiring estimates that are uniformly valid for large order N and large wave-number hk such that the ratio of the two quantities remains constant (of order κ). The following result shows that this strategy delivers an exponential rate of convergence:

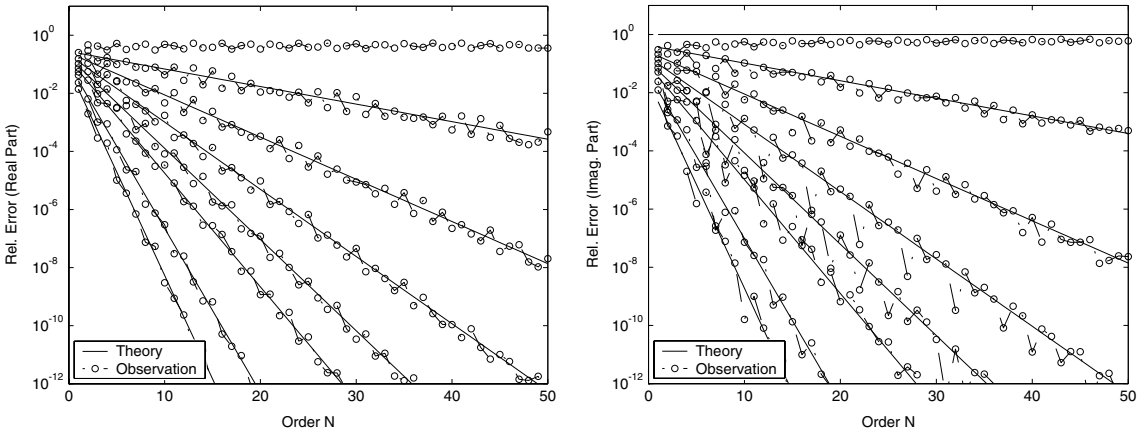

Fig. 3. Real and imaginary parts of relative error ρ_N with $\gamma = 0.5$ for various values of hk with the order N chosen so that $2N + 1 = \kappa hk$ as described in Theorem 4. The results obtained with $\kappa = 1.0$ (shallowest), 1.2, 1.4, 1.8, 2.0, 2.5 and 3.0 (steepest) are shown along with the theoretical prediction (17). Observe that an exponential rate of decay is obtained for $\kappa > 1$ as predicted in Theorem 4.

Theorem 4. *Let $\kappa > 1$ be fixed. If N, $hk \to \infty$ in a such a way that $2N + 1 = \kappa hk$, then ρ_N decays at an exponential rate as $N \to \infty$,*

$$\rho_N \approx -\mathrm{e}^{-\beta(N+1/2)}\left(1 - \sqrt{1 - \frac{1}{\kappa^2}}\right)\left(\sqrt{\kappa^2 - 1} - \mathrm{i}\right), \tag{17}$$

where β is a positive real number, defined in (32), *which only depends on κ.*

Fig. 3 shows the actual relative error ρ_N and the asymptotic results presented in Theorem 4. It is observed that the asymptotic results provide an accurate indication of the actual behaviour even for moderate values of N that could reasonably be used in practical computations.

2.7. Spurious mode

The nature of the spurious mode appearing on the right hand side of (10) is discussed at length in [14] to which we have little to add. We point out that the mode corresponds to a wave travelling in the opposite direction to the physical wave and, for non-negative γ, is damped by a factor $(1 - \gamma)/(1 + \gamma)$ as it passes through each element. This means that in the resolved regime, where the relative error ρ_N is small, the mode decays exponentially fast and has no impact in practical computation.

3. Analysis of remainder in Padé approximant

3.1. Padé approximant to the exponential

The study of Padé approximants of the exponential e^z has enjoyed a long history going back to the original work of Padé himself [22] where the following results, quoted from Varga [26], are obtained for non-negative integers p and q:

$$[p/q]_{\exp(z)} = \frac{{}_1\mathrm{F}_1(-p;-p-q;z)}{{}_1\mathrm{F}_1(-q;-p-q;-z)} \tag{18}$$

with remainder given by

$$\mathrm{e}^z - [p/q]_{\exp(z)} = \frac{\mathrm{e}^z z^{p+q+1}\int_0^1 \mathrm{e}^{-tz}t^p(t-1)^q\,\mathrm{d}t}{(p+q)!\,{}_1\mathrm{F}_1(-\mathrm{q};-\mathrm{p}-\mathrm{q};-\mathrm{z})}. \tag{19}$$

Here, ${}_1\mathrm{F}_1$ denotes the confluent hypergeometric function defined by the series

$${}_1\mathrm{F}_1(a,b,z) = 1 + \frac{a}{b}z + \frac{a}{b}\frac{a+1}{b+1}\frac{z^2}{2!} + \frac{a}{b}\frac{a+1}{b+1}\frac{a+2}{b+2}\frac{z^3}{3!} + \cdots \tag{20}$$

or, if we adopt Pochhammer's notation $(a)_0 = 1$ and $(a)_m = a(a+1)\cdots(a+m-1)$, then we have the alternative form

$${}_1\mathrm{F}_1(a,b,z) = \sum_{m=0}^{\infty}\frac{(a)_m}{(b)_m}\frac{z^m}{m!}.$$

The behaviour of the remainder in the limit $z \to 0$, and for the $[N - a/N]$-Padé approximants (where $a = 0, 1$) as $N \to \infty$ for fixed z, is well documented [19, p. 191]. However, we require expressions for the remainder in the sub- and super-diagonal Padé approximants with purely imaginary argument that are uniformly valid for large order N and large argument z. Our approach is based on expressing the remainder in terms of Bessel functions and then using Langer's formulae [9], which provide uniformly valid expansions for Bessel functions of large order and argument. This enables us to deduce the leading terms in the remainder, although actual bounds could be obtained if we were to use the uniform asymptotic expansions with error bounds provided by Olver [21] in place of Langer's formulae. A related approach was adopted by Driver and Temme [7] in their analysis of the locations of the poles and zeros of the polynomials appearing in the quotient (18) for the diagonal approximants (i.e. $p = q$). There the remainder is expressed in terms of Bessel functions and expansions in terms of Airy functions are employed.

We begin by establishing a link between the remainder in the Padé approximant and modified Bessel functions of the second kind:

Lemma 1. *Let $N \in \mathbb{N}$. Then,*

$$\mathrm{e}^z - [N+1/N]_{\exp(z)} = \mathrm{e}^z\left\{1 + \frac{(-1)^N}{\pi}\frac{K_{N+1/2}(z/2) + K_{N+3/2}(z/2)}{I_{N+1/2}(z/2) - I_{N+3/2}(z/2)}\right\}^{-1}, \tag{21}$$

where I and K denote modified Bessel functions [11].

Proof. From (19) with $p = N+1$ and $q = N$,

$$\mathrm{e}^z - [N+1/N]_{\exp(z)} = \frac{\mathrm{e}^z z^{2N+2}\int_0^1 \mathrm{e}^{-tz}t^{N+1}(t-1)^N\,\mathrm{d}t}{(2N+1)!\,{}_1\mathrm{F}_1(-N;-2N-1;-z)}.$$

The proof consists of rewriting the numerator and the denominator as follows:

(i) A simple change of variable gives

$$T_1 = \int_0^1 \mathrm{e}^{-tz}t^{N+1}(t-1)^N\,\mathrm{d}t = \frac{(-1)^N\mathrm{e}^{-z/2}}{2^{2N+2}}\int_{-1}^1 \mathrm{e}^{sz/2}(1-s)^{N+1}(1+s)^N\,\mathrm{d}s.$$

Using the identity (9.221) of [11], this may be rewritten as

$$(-1)^N z^{-N-3/2}\,\mathrm{e}^{-z/2}\frac{N!(N+1)!}{(2N+2)!}M_{1/2,N+1}(z),$$

where $M_{1/2,N+1}$ denotes the Whittaker function of the first kind [11, (9.220)] with index $1/2$. Whittaker functions satisfy the following identity [24, (2.5.1)],

$$z^{-1/2}M_{1/2,N+1}(z) = M_{0,N+1/2}(z) - \frac{1}{2(2N+3)}M_{0,N+3/2}(z),$$

where $M_{0,\mu}$ is the Whittaker function of the first kind of order μ and index zero. The latter functions are related to Bessel functions as follows [11, $(9.235)_2$]

$$M_{0,\mu}(z) = \Gamma(1+\mu)2^{2\mu}z^{1/2}I_\mu(z/2),$$

where Γ is the gamma function [11]. This leads to the conclusion

$$M_{0,N+1/2}(z) - \frac{1}{2(2N+3)}M_{0,N+3/2}(z) = \Gamma(N+3/2)2^{2N+1}z^{1/2}\big(I_{N+1/2}(z/2) - I_{N+3/2}(z/2)\big).$$

In summary, after simplifying using the relation

$$2^{2N}\Gamma(N+3/2) = \frac{\sqrt{\pi}}{2}\,\frac{(2N+1)!}{N!},$$

we arrive at the conclusion

$$T_1 = \frac{\sqrt{\pi}}{2}(-1)^N z^{-N-1/2}\,\mathrm{e}^{-z/2}N!\big(I_{N+1/2}(z/2) - I_{N+3/2}(z/2)\big).$$

This completes the treatment of the numerator.

(ii) The denominator may be expressed in form

$$(2N+1)!\,{}_1\mathrm{F}_1(-\mathrm{N};-2\mathrm{N}-1;-\mathrm{z}) = \int_0^\infty \mathrm{e}^{-\mathrm{t}}\mathrm{t}^{\mathrm{N}+1}(\mathrm{t}-\mathrm{z})^{\mathrm{N}}\,\mathrm{dt},$$

which is easily verified by using the binomial expansion and integrating. Then, with $t = sz$, this may be rewritten as

$$z^{2N+2}\int_0^\infty \mathrm{e}^{-sz}s^{N+1}(s-1)^N\,\mathrm{d}s = z^{2N+2}(T_1+T_2),$$

where T_1 is defined above, and

$$T_2 = \int_1^\infty \mathrm{e}^{-sz}s^{N+1}(s-1)^N\,\mathrm{d}s.$$

Making the substitution $s = t+1$ gives the alternative form

$$\mathrm{e}^{-z}\int_0^\infty \mathrm{e}^{-tz}t^N(t+1)^{N+1}\,\mathrm{d}t$$

which in turn may be written in terms of a Whittaker function of the second kind using [11, $(9.222)_1$],

$$N!\,\mathrm{e}^{-z/2}z^{-N-3/2}W_{1/2,N+1}(z).$$

Identities $(9.235)_1$ and $(9.235)_2$ of [11] imply that

$$W_{1/2,N+1}(z) = \frac{1}{2} z^{1/2} \big(W_{0,N+1/2}(z) + W_{0,N+3/2}(z) \big)$$

and then identity [11, $(9.235)_2$] gives

$$W_{1/2,N+1}(z) = \frac{1}{2} \frac{z}{\sqrt{\pi}} \big(K_{N+1/2}(z/2) + K_{N+3/2}(z/2) \big).$$

Therefore,

$$T_2 = \frac{N!}{2\sqrt{\pi}} \mathrm{e}^{-z/2} z^{-N-1/2} \big(K_{N+1/2}(z/2) + K_{N+3/2}(z/2) \big).$$

Finally, combining these results gives

$$\mathrm{e}^z - [N + 1/N]_{\exp(z)} = \mathrm{e}^z \left(1 + \frac{T_2}{T_1} \right)^{-1}$$

and inserting the expressions for T_1 and T_2 gives the result claimed. □

The next result gives a closed form expression for the error in terms of first-kind Bessel functions when the argument is purely imaginary.

Lemma 2. *Let $N \in \mathbb{N}$ and $\Omega \in \mathbb{R}$. Then,*

$$\mathrm{e}^{\mathrm{i}\Omega} - [N + 1/N]_{\exp(\mathrm{i}\Omega)} = 2\mathrm{e}^{\mathrm{i}\Omega} \{ 1 + \mathrm{i} R_N(\Omega/2) \}^{-1}, \tag{22}$$

where

$$R_N(x) = \frac{Y_{N+1/2}(x) - \mathrm{i} Y_{N+3/2}(x)}{J_{N+1/2}(x) - \mathrm{i} J_{N+3/2}(x)}. \tag{23}$$

Proof. First, recall that

$$K_{n+1/2}(z/2) = (-1)^{n+1} \frac{\pi}{2} \big(I_{n+1/2}(z/2) - I_{-n-1/2}(z/2) \big).$$

Inserting this expression into the term in parentheses on the right hand side of Eq. (21) and simplifying shows that the term may be written as

$$\frac{1}{2} \left\{ 1 + \frac{I_{-N-1/2}(z/2) - I_{N+3/2}(z/2)}{I_{N+1/2}(z/2) - I_{N+3/2}(z/2)} \right\}.$$

Then, inserting $z = \mathrm{i}\Omega$ into the (finite) series expansions for $I_{n+1/2}$, $J_{n+1/2}$ and $Y_{n+1/2}$ (see (8.462), (8.467) and (8.468) of [11]), and using the resulting relations between the Bessel functions leads to the conclusion that the above expression coincides with

$$\frac{1}{2}\left\{1+\mathrm{i}\frac{Y_{N+1/2}(\Omega/2)-\mathrm{i}Y_{N+3/2}(\Omega/2)}{J_{N+1/2}(\Omega/2)-\mathrm{i}J_{N+3/2}(\Omega/2)}\right\}$$

and the result then follows from Lemma 1. □

3.2. Remainder for small argument Ω

The general result in Lemma 2 provides an easy passage to the following expression for the remainder at small argument Ω:

Corollary 1. *Let $N \in \mathbb{N}$ and suppose $\Omega \in \mathbb{R}$ is small. Then*

$$\mathrm{e}^{\mathrm{i}\Omega}-[N+1/N]_{\exp(\mathrm{i}\Omega)}=-\Omega^{2N+2}\frac{\mathrm{e}^{\mathrm{i}\Omega}}{2}\left[\frac{N!}{(2N+1)!}\right]^2\left\{1-\frac{2\mathrm{i}\Omega(N+1)}{(2N+1)(2N+3)}+\mathcal{O}(\Omega^2)\right\}. \tag{24}$$

Proof. For small κ, identity (8.440) of [11] gives

$$J_{n+1/2}(\kappa)=\frac{1}{\Gamma(3/2+n)}\left(\frac{\kappa}{2}\right)^{n+1/2}+\cdots$$

while combining identities $(8.465)_1$ and (8.440) of [11] gives

$$Y_{n+1/2}(\kappa)=(-1)^{n-1}J_{-n-1/2}(\kappa)=\frac{(-1)^{n-1}}{\Gamma(1/2-n)}\left(\frac{\kappa}{2}\right)^{-n-1/2}+\cdots$$

where Γ denotes the gamma function. Simple substitution and the use of formulae (8.339) of [11] gives, after some simplification,

$$1+\mathrm{i}R_N(\Omega/2)=\frac{-4}{\Omega^{2N+2}}\left[\frac{(2N+1)!}{N!}\right]^2\frac{1+\mathrm{i}\Omega/(4N+2)+\cdots}{1-\mathrm{i}\Omega/(4N+6)+\cdots}$$

and the result then follows by inserting this expression into the error representation given in Lemma 2. □

3.3. Remainder for large order N and large argument Ω

We now consider the behaviour of the remainder for large order N and large argument Ω in detail. Three distinct regimes are identified depending on the relative sizes of N and Ω. If $N \ll \Omega$, then the remainder tends to oscillate without decay, while if $N \gg \Omega$, then the remainder decays at a super-exponential rate. The following result gives a sharp identification of when the transition between these extremes occurs, and provides a precise estimate for the nature of the transition.

Theorem 5. *Suppose $\Omega \in \mathbb{R}$ and $N \in \mathbb{N}$. As the order $N \gg 1$ is increased relative to the argument Ω, the error $E_N(\Omega)=\mathrm{e}^{\mathrm{i}\Omega}-[N+1/N]_{\mathrm{e}^{\mathrm{i}\Omega}}$ passes through three distinct phases:*

1. *if $2N+1<\Omega-C\Omega^{1/3}$, then E_N oscillates but does not decay as N is increased;*
2. *if $\Omega-\mathrm{o}(\Omega^{1/3})<2N+1<\Omega+\mathrm{o}(\Omega^{1/3})$, then E_N decays algebraically at a rate $\mathcal{O}(N^{-1/3})$. More precisely,*

$$E_N(\Omega)\approx\frac{2\mathrm{e}^{\mathrm{i}\Omega}}{1-\mathrm{i}\sqrt{3}}\left\{1+\frac{\mathrm{i}\sqrt{3}}{1-\mathrm{i}\sqrt{3}}\frac{3^{5/6}}{\pi}\Gamma(2/3)^2\frac{v^{-1/3}t_v-\mathrm{i}(v+1)^{-1/3}t_{v+1}}{v^{-1/3}-\mathrm{i}(v+1)^{-1/3}}\right\}, \tag{25}$$

where $v=N+1/2$ and $t_v=(2/v)^{1/3}(v-\Omega/2)$;

3. *if* $2N+1 > \Omega + C\Omega^{1/3}$, *then* E_N *decays as*

$$E_N(\Omega) \approx \mathrm{i}\mathrm{e}^{\mathrm{i}\Omega}\frac{v^{-1/2}f(w_v)^v - \mathrm{i}(v+1)^{-1/2}f(w_{v+1})^{v+1}}{v^{-1/2}f(w_v)^{-v} - \mathrm{i}(v+1)^{-1/2}f(w_{v+1})^{-(v+1)}}, \tag{26}$$

where $v = N + 1/2$, $w_v = (1-\Omega^2/4v^2)^{1/2}$ *and* $f : w \mapsto \mathrm{e}^w(1-w)^{1/2}/(1+w)^{1/2}$.

Proof. Denote $v = N + 1/2$ and $x = \Omega/2$. The proof is divided into two cases depending on the relative sizes of v and x.

Case 1: $2N+1 > \Omega$. Here, we have $v > x$ and we may apply Langer's formulas [9, Section 7.13.4 (34) and (35)] to obtain

$$\begin{aligned} J_v(x) &= \frac{1}{\pi}\sqrt{\frac{z}{wv}}K_{1/3}(z) + \mathcal{O}(v^{-4/3}),\\ Y_v(x) &= -\sqrt{\frac{z}{wv}}\big[I_{1/3}(z) + I_{-1/3}(z)\big] + \mathcal{O}(v^{-4/3}), \end{aligned} \tag{27}$$

where $w = (1-x^2/v^2)^{1/2}$ and $z = v(\tanh^{-1} w - w)$.

Case 1(a): $\Omega < 2N+1 < \Omega + \mathrm{o}(\Omega^{1/3})$. For N in this range, we find that $w \approx (2/v)^{1/2}(v-x)^{1/2} \ll 1$ and so $z \approx (1/3)vw^3 = (2/3)t^{3/2} = \mathrm{o}(1)$ where $t = (2/v)^{1/3}(v-x)$. Inserting series expansions for the Bessel functions $I_{\pm 1/3}$ and $K_{1/3}$ with small argument and simplifying gives

$$\begin{aligned} J_v(x) &\approx 3^{-2/3}\Gamma(2/3)^{-1}(2/v)^{1/3}\Big[1 - 3^{5/6}\Gamma(2/3)^2 t/2\pi + \mathcal{O}(t^3)\Big],\\ Y_v(x) &\approx -3^{-1/6}\Gamma(2/3)^{-1}(2/v)^{1/3}\Big[1 + 3^{5/6}\Gamma(2/3)^2 t/2\pi + \mathcal{O}(t^3)\Big]. \end{aligned}$$

Substituting these expressions into the ratio R_N given in (23) and using the fact that $t = \mathrm{o}(1)$, we arrive at

$$R_N(x) \approx -\sqrt{3}\left[1 + \frac{3^{5/6}}{\pi}\Gamma\left(\frac{2}{3}\right)^2\frac{v^{-1/3}t_v - \mathrm{i}(v+1)^{-1/3}t_{v+1}}{v^{-1/3} - \mathrm{i}(v+1)^{-1/3}}\right],$$

and then inserting this into (22) and simplifying gives the result claimed.

Case 1(b): $2N+1 > \Omega + C\Omega^{1/3}$. In this range, the value of z will be large. The Bessel functions appearing in (27) may be written in terms of the Airy functions Ai and Bi as in (11.1.04) and (1.1.12) of [21] to give

$$\begin{aligned} J_v(x) &= \sqrt{\frac{3z}{wvt}}\,\mathrm{Ai}(t) + \mathcal{O}(v^{-4/3}),\\ Y_v(x) &= -\sqrt{\frac{3z}{wvt}}\,\mathrm{Bi}(t) + \mathcal{O}(v^{-4/3}), \end{aligned}$$

where $z = (2/3)t^{3/2}$. The behaviour of the Airy functions for large argument is given by (1.1.07) and (1.1.16) of [21]:

$$\mathrm{Ai}(t) \sim \frac{\mathrm{e}^{-z}}{2\sqrt{\pi}t^{1/4}}; \qquad \mathrm{Bi}(t) \sim \frac{\mathrm{e}^{z}}{\sqrt{\pi}t^{1/4}}.$$

Elementary manipulations give $\mathrm{e}^{\pm z} = f(w)^{\mp v}$, where f is the function defined in the statement of the result. On inserting these expansions and simplifying, these formulae may be written in the alternative form

$$J_\nu(x) \approx \frac{1}{\sqrt{2\pi w\nu}} f(w)^\nu; \qquad Y_\nu(x) \approx -\sqrt{\frac{2}{\pi w\nu}}\, f(w)^{-\nu},$$

where $w = (1 - x^2/\nu^2)^{1/2}$. Observe that f is monotonic decreasing on $(0, 1]$ from 1 to 0 and, since $w > 0$, the ratio $R_N(x)$ defined in (23) dictates the behaviour of the error (22) for large order N. The claim (26) then follows on substituting the expressions for J_ν and Y_ν and simplifying.

Case 2: $2N + 1 < \Omega$. Here, $\nu < x$ and we may apply Langer's formulas [9, Section 7.13.4 (32) and (33)] to obtain

$$\begin{aligned} J_\nu(x) &= \sqrt{\frac{z}{w\nu}}\big[J_{1/3}(z)\cos\pi/6 - Y_{1/3}(z)\sin\pi/6\big] + \mathcal{O}(\nu^{-4/3}),\\ Y_\nu(x) &= \sqrt{\frac{z}{w\nu}}\big[J_{1/3}(z)\sin\pi/6 + Y_{1/3}(z)\cos\pi/6\big] + \mathcal{O}(\nu^{-4/3}), \end{aligned} \tag{28}$$

where we now define $w = (x^2/\nu^2 - 1)^{1/2}$ and $z = \nu(\tan^{-1} w - w)$.

Case 2(a): $\Omega - o(\Omega^{1/3}) < 2N + 1 < \Omega$. For N in this range we find, as in Case 1(a), that $w \approx (2/\nu)^{1/2}(x - \nu)^{1/2} \ll 1$ and so $z \approx (1/3)\nu w^3 = (2/3)\tau^{3/2} = o(1)$ where $\tau = (2/\nu)^{1/3}(x - \nu) = -t$. Expanding the terms appearing in parentheses in (28) and simplifying gives

$$\begin{aligned} J_\nu(x) &\approx 3^{-2/3}\Gamma(2/3)^{-1}(2/\nu)^{1/3}\Big[1 + 3^{5/6}\Gamma(2/3)^2\tau/2\pi + \mathcal{O}(\tau^3)\Big],\\ Y_\nu(x) &\approx -3^{-1/6}\Gamma(2/3)^{-1}(2/\nu)^{1/3}\Big[1 - 3^{5/6}\Gamma(2/3)^2\tau/2\pi + \mathcal{O}(\tau^3)\Big], \end{aligned}$$

where τ is given above. If we substitute $\tau = -t$, then these formulae are identical with those obtained in Case 1(a) and the remainder of the argument then follows the one used in Case 1(a).

Case 2(b): $2N + 1 < \Omega - C\Omega^{1/3}$. For N in this range, z will generally be large. The behaviour of the Bessel functions of order $1/3$ for large argument z is given in $(8.440)_1$ and $(8.440)_2$ of [11]:

$$\begin{aligned} J_{1/3}(z) &\sim \sqrt{\frac{2}{\pi z}}\cos\left(z - \frac{5}{12}\pi\right),\\ Y_{1/3}(z) &\sim \sqrt{\frac{2}{\pi z}}\sin\left(z - \frac{5}{12}\pi\right). \end{aligned} \tag{29}$$

Together, expressions (28) and (29) show that the Bessel functions $J_{N+1/2}(\Omega/2)$ and $Y_{N+1/2}(\Omega/2)$ tend to oscillate but not decay as the order N is increased in the range considered. Consequently, the expression for $E_N(\Omega)$ appearing on the right hand side of (22) reflects this behaviour as N is increased. □

The next result provides further elaboration on the estimate (26) in two important limits as $N \to \infty$:

Theorem 6. *Let $\Omega \in \mathbb{R}$ and $N \in \mathbb{N}$. Let $E_N(\Omega)$ denote the error in the $[N + 1/N]$-Padé approximant of $e^{i\Omega}$.*
1. *If $2N + 1 \gg \Omega$, then $E_N(\Omega)$ decays at a super-exponential rate:*

$$E_N(\Omega) \approx -e^{i\Omega}\left[\frac{e\Omega}{2\sqrt{(2N+1)(2N+3)}}\right]^{2N+2}\left(1 - \frac{i\Omega}{2N+3}\right). \tag{30}$$

2. *Let $\kappa > 1$ be fixed. If $N, \Omega \to \infty$ in a such a way that $2N + 1 = \kappa\Omega$, then $E_N(\Omega)$ decays at an exponential rate:*

$$E_N(\Omega) \approx -\mathrm{e}^{\mathrm{i}\Omega-\beta(N+1/2)}\left(1-\sqrt{1-\frac{1}{\kappa^2}}\right)\left(\sqrt{\kappa^2-1}-\mathrm{i}\right), \tag{31}$$

where β is the positive real number (which depends on κ) given by

$$\beta = \ln\frac{1+\sqrt{1-1/\kappa^2}}{1-\sqrt{1-1/\kappa^2}} - 2\sqrt{1-\frac{1}{\kappa^2}}. \tag{32}$$

Proof. Denote $v = N + 1/2$ and $x = \Omega/2$. By Theorem 5, we have

$$\mathrm{e}^{-\mathrm{i}\Omega}E_N(\Omega) \approx \mathrm{i}\frac{v^{-1/2}f_v^v - \mathrm{i}(v+1)^{-1/2}f_{v+1}^{v+1}}{v^{-1/2}f_v^{-v} - \mathrm{i}(v+1)^{-1/2}f_{v+1}^{-(v+1)}}, \tag{33}$$

where $f_v = f(w_v)$ and $f_{v+1} = f(w_{v+1})$, with $w_v = (1-x^2/v^2)^{1/2}$ and $f : w \mapsto \mathrm{e}^w(1-w)^{1/2}/(1+w)^{1/2}$.

Case 1: In this situation $v \gg x$ so that $w_v \approx 1 - x^2/2v^2$. Hence $f_v \approx ex/2v$ and $f_{v+1} \approx ex/2(v+1)$. Therefore, for $v \gg x$ we have $f_{v+1}^{v+1}f_v^{-v} \ll 1$ and as a consequence we obtain

$$\mathrm{e}^{-\mathrm{i}\Omega}E_N(\Omega) \approx -\sqrt{\frac{v+1}{v}}f_v^v f_{v+1}^{v+1}\frac{1-\mathrm{i}\sqrt{v/(v+1)}f_{v+1}^{v+1}f_v^{-v}}{1+\mathrm{i}\sqrt{(v+1)/v}f_{v+1}^{v+1}f_v^{-v}} \approx -f_{v+1}^{v+1}\left[\sqrt{\frac{v+1}{v}}f_v^v - \mathrm{i}\left(2+\frac{1}{v}\right)f_{v+1}^{v+1}\right].$$

Inserting the approximations for f_v and f_{v+1} and simplifying gives

$$\mathrm{e}^{-\mathrm{i}\Omega}E_N(\Omega) \approx -\left[\frac{ex}{2\sqrt{v(v+1)}}\right]^{2v+1}\left\{1-\mathrm{i}\left(2+\frac{1}{v}\right)\left(1+\frac{1}{v}\right)^{-v-1/2}\frac{ex}{2(v+1)}\right\}$$

and then observing that

$$\left(2+\frac{1}{v}\right)\left(1+\frac{1}{v}\right)^{-v-1/2} \to \frac{2}{e} \quad \text{as } v \to \infty,$$

we arrive at

$$\mathrm{e}^{-\mathrm{i}\Omega}E_N(\Omega) \approx -\left[\frac{ex}{2\sqrt{v(v+1)}}\right]^{2v+1}\left(1-\frac{\mathrm{i}x}{v+1}\right)$$

which, on replacing v and x, gives the result claimed.

Case 2: In this case, $w_v = \sqrt{1-1/\kappa^2}$ and an easy computation then shows that

$$w_{v+1} = w_v\left(1+\frac{1}{v}\frac{1}{\kappa^2-1}+\mathcal{O}(v^{-2})\right).$$

Hence, using Taylor's theorem and the fact that $f'(w) = w^2 f(w)/(w^2-1)$ gives

$$f_{v+1} = \left(1-\frac{1}{v}\sqrt{1-\frac{1}{\kappa^2}}+\mathcal{O}(v^{-2})\right)f_v. \tag{34}$$

Therefore, for large v, $f_{v+1}^v \approx f_v^v \mathrm{e}^{-\sqrt{1-1/\kappa^2}}$. With the aid of (33), we obtain

$$e^{-i\Omega}E_N(\Omega) \approx i f_v^{2v} \frac{v^{-1/2} - i(v+1)^{-1/2} f_{v+1} e^{-w_v}}{v^{-1/2} - i(v+1)^{-1/2} f_{v+1}^{-1} e^{w_v}}.$$

By Eq. (34),

$$e^{-w_v} f_{v+1} \approx e^{-w_v} f_v \big(1 + \mathcal{O}(v^{-1})\big) = R_v \big(1 + \mathcal{O}(v^{-1})\big),$$

where $R_v = \sqrt{(1-w_v)/(1+w_v)}$, and we deduce that for large v,

$$e^{-i\Omega}E_N(\Omega) \approx i f_v^{2v} \frac{1 - iR_v}{1 - i/R_v} = i f_v^{2v}(1 - w_v + iR_v w_v).$$

By substituting $w_v = \sqrt{1 - 1/\kappa^2}$ and simplifying further, we arrive at

$$e^{-i\Omega}E_N(\Omega) \approx i f_v^{2v} \Big(1 - \sqrt{1 - 1/\kappa^2}\Big)\Big(1 + i\sqrt{\kappa^2 - 1}\Big).$$

Let β be defined as in the statement of the result, then

$$e^{-\beta} = e^{2w_v} \frac{1 - w_v}{1 + w_v} = f_v^2,$$

and using this to replace f_v^2 in the previous estimate gives the result claimed. Finally, if $w_v \in (0,1)$, then $f(w_v)^2 = f_v^2 < 1$ and β is therefore positive. □

4. Analysis of an eigenvalue problem

Properties of the following eigenvalue problem will prove useful in the analysis of the dispersion error: Find $\Phi \in \mathbb{P}_N$ and $\lambda \in \mathbb{C}$ such that for given $\Omega \in \mathbb{C}$,

$$\begin{aligned}(\Phi', v) &+ \frac{1}{2}(1-\gamma)(\lambda\Phi(-1) - \Phi(1))v(1) + \frac{1}{2}(1+\gamma)\big(\Phi(-1) - \lambda^{-1}\Phi(1)\big)v(-1) \\ &= \frac{1}{2} i\Omega(\Phi, v) \quad \forall v \in \mathbb{P}_N. \end{aligned} \tag{35}$$

As usual, the condition under which the eigenvalue problem will possess non-trivial solutions reduces to an algebraic equation for the eigenvalue λ, which we now proceed to identify.

4.1. Conditions for an eigenvalue

We begin by considering the exceptional cases where $\gamma = \pm 1$. In what follows, it will be convenient to let $\mathscr{L}$ denote the differential operator defined by $\mathscr{L}v = \frac{1}{2} i\Omega v + v'$, and to use $P_N^{(p,q)}$ to denote the Jacobi polynomial of type (p,q) and degree N (see Chapter 8, Section 9.6 of [11]).

Lemma 3. (i) *Suppose* $\gamma = 1$. *If* $\lambda = \lambda_N^+ = [N/N+1]_{\exp(i\Omega)}$, *then Eq.* (35) *admits a non-trivial solution* $\Phi_N^+ \in \mathbb{P}_N$ *of the form*

$$\Phi_N^+(s) = \sum_{m=0}^{N} (i\Omega)^m \frac{(2N+1-m)!}{(2N+1)!} P_m^{(N-m,N-m+1)}(s). \tag{36}$$

(ii) *Suppose* $\gamma = -1$. *If* $\lambda = \lambda_N^- = [N+1/N]_{\exp(i\Omega)}$, *then Eq.* (35) *admits a non-trivial solution of the form*

$$\Phi_N^-(s) = \sum_{m=0}^{N} (i\Omega)^m \frac{(2N+1-m)!}{(2N+1)!} P_m^{(N-m+1,N-m)}(s). \tag{37}$$

Proof. Consider the case $\gamma = 1$. Elementary manipulations and the use of the following identity, see $(8.961)_4$ of [11],

$$\frac{d}{ds} P_m^{(N-m,N-m+1)}(s) = \frac{1}{2} \frac{(2N-m+2)!}{(2N-m+1)!} P_{m-1}^{(N-m+1,N-m+2)}(s),$$

reveal that

$$\mathscr{L}\Phi_N^+ = -\frac{(i\Omega)^{N+1}}{2} \frac{(N+1)!}{(2N+1)!} P_N^{(0,1)}(s). \tag{38}$$

Hence, the orthogonality properties of Jacobi polynomials mean that Φ_N^+ satisfies Eq. (35) when v is of the form $(1+s)w$ for some $w \in \mathbb{P}_{N-1}$. It only remains to show that (35) is satisfied when v is a constant. Firstly, since

$$P_m^{(N-m,N-m+1)}(1) = \binom{N}{m},$$

see $(8.960)_2$ of [11], we have

$$\Phi_N^+(1) = \sum_{m=0}^{N} \frac{(-N)_m}{(-2N-1)_m} \frac{(i\Omega)^m}{m!} = {}_1F_1(-N; -2N-1; i\Omega). \tag{39}$$

Now, thanks to $(7.391)_4$ of [11], $(P_N^{(1,0)}, 1) = 2/(N+1)$, and then by $(8.961)_1$ of [11], $P_N^{(1,0)}(-s) = (-1)^N P_N(s)$, and we obtain $(P_N^{(0,1)}, 1) = (-1)^N 2/(N+1)$. Therefore,

$$\left(\mathscr{L}\Phi_N^+, 1\right) = \frac{N!}{(2N+1)!} (-i\Omega)^{N+1} \tag{40}$$

and hence, using the fact that

$$P_m^{(N-m,N-m+1)}(-1) = (-1)^m \binom{N+1}{m},$$

see $(8.960)_2$ and $(8.961)_1$ of [11], we obtain

$$\left(\mathscr{L}\Phi_N^+, 1\right) + \Phi_N^+(-1) = \sum_{m=0}^{N+1} \frac{(-N-1)_m}{(-2N-1)_m} \frac{(-i\Omega)^m}{m!} = {}_1F_1(-N-1; -2N-1; -i\Omega).$$

Consequently, Eq. (35) holds for constant v (and therefore all $v \in \mathbb{P}_N$) provided that

$$\lambda = \lambda_N^+ = \frac{{}_1F_1(-N; -2N-1; i\Omega)}{{}_1F_1(-N-1; -2N-1; -i\Omega)} = [N/N+1]_{\exp(i\Omega)}$$

as claimed. The proof in the case $\gamma = -1$ follows similar lines. In particular, we obtain

$$\mathscr{L}\Phi_N^- = -\frac{(i\Omega)^{N+1}}{2} \frac{(N+1)!}{(2N+1)!} P_N^{(1,0)}(s) \tag{41}$$

which leads to

$$\left(\mathscr{L}\Phi_N^-, 1\right) = -\frac{N!}{(2N+1)!}(\mathrm{i}\Omega)^{N+1}. \tag{42}$$

Manipulations similar to those used in the case $\gamma = 1$ give

$$\Phi_N^-(-1) = {}_1\mathrm{F}_1(-N; -2N-1; -\mathrm{i}\Omega) \tag{43}$$

and

$$\Phi_N^-(1) - \left(\mathscr{L}\Phi_N^-, 1\right) = {}_1\mathrm{F}_1(-N-1; -2N-1; \mathrm{i}\Omega)$$

which lead to the condition that

$$\lambda = \lambda_N^- = \frac{{}_1\mathrm{F}_1(-N-1; -2N-1; \mathrm{i}\Omega)}{{}_1\mathrm{F}_1(-N; -2N-1; -\mathrm{i}\Omega)} = [N+1/N]_{\exp(\mathrm{i}\Omega)}$$

as claimed. □

The functions $\Phi_N^\pm$ that arise in the case $\gamma = \pm 1$ may be used to analyse the general case $\gamma \in [-1, 1]$:

Lemma 4. *Let $\gamma \in [-1, 1]$ and $N \in \mathbb{N}$. If λ satisfies the algebraic equation*

$$0 = (1-\gamma)\,{}_1\mathrm{F}_1(-N; -2N-1; -\mathrm{i}\Omega)(\lambda - \lambda_N^-) + (-1)^N(1+\gamma)\,{}_1\mathrm{F}_1(-N; -2N-1; \mathrm{i}\Omega)\left(\frac{1}{\lambda_N^+} - \frac{1}{\lambda}\right), \tag{44}$$

then Eq. (35) *admits a non-trivial solution $\Phi \in \mathrm{span}\{\Phi_N^+, \Phi_N^-\}$.*

Proof. In view of Lemma 3, we may assume that $\gamma \in (-1, 1)$. We seek a non-trivial solution $\Phi_N \in \mathbb{P}_N$ of the form

$$\Phi_N = c^-\Phi_N^- + c^+\Phi_N^+,$$

where c^- and c^+ are non-zero scalars whose existence is to be determined. Thanks to (38) and (41), it follows that $\mathscr{L}\Phi_N \in \mathrm{span}\{P_N^{(1,0)}, P_N^{(0,1)}\}$. As a matter of fact, by writing

$$P_N^{(1,0)} = \frac{N+2}{2N+2}P_N^{(1,1)} + \frac{1}{2}P_{N-1}^{(1,1)}$$

and

$$P_N^{(0,1)} = \frac{N+2}{2N+2}P_N^{(1,1)} - \frac{1}{2}P_{N-1}^{(1,1)},$$

we conclude that

$$\mathscr{L}\Phi_N \in \mathrm{span}\{P_{N-1}^{(1,1)}, P_N^{(1,1)}\}.$$

This implies that Φ_N satisfies Eq. (35) for all v of the form $(1-s^2)w$ where $w \in \mathbb{P}_{N-2}$, since Eq. (35) then reduces to the identity

$$0 = \int_{-1}^{1}(1-s^2)w(s)\,\mathscr{L}\Phi_N(s)\,\mathrm{d}s,$$

which holds due to the standard orthogonality properties of Jacobi polynomials. It therefore suffices to show there exist non-trivial scalars c^- and c^+ such that (35) is satisfied in the special cases $v = 1 \pm s$. Inserting the expression for Φ_N into (35) and choosing $v = 1 \pm s$ shows that the existence of a non-trivial solution is equivalent to the algebraic condition

$$\begin{vmatrix} R(\Phi^-, 1+s) & R(\Phi^+, 1+s) \\ R(\Phi^-, 1-s) & R(\Phi^+, 1-s) \end{vmatrix} = 0,$$

where $R(\Phi, v)$ denotes the functional defined by the difference between the left and right hand sides of Eq. (35). The off-diagonal entries in the determinant may be simplified using (38) and (41) to obtain

$$R(\Phi^+, 1+s) = (1-\gamma)\big(\lambda\Phi_N^+(-1) - \Phi_N^+(1)\big)$$

and

$$R(\Phi^-, 1-s) = (1+\gamma)\big(\Phi_N^-(-1) - \lambda^{-1}\Phi_N^-(1)\big).$$

Expanding and simplifying the resulting determinant gives the algebraic condition

$$\begin{aligned} 0 = {} & (\mathscr{L}\Phi_N^-, 1+s)\,(\mathscr{L}\Phi_N^+, 1-s) + (1-\gamma)\,(\mathscr{L}\Phi_N^+, 1-s)\big[\lambda\Phi_N^-(-1) - \Phi_N^-(1)\big] \\ & + (1+\gamma)\,(\mathscr{L}\Phi_N^-, 1+s)\big[\Phi_N^+(-1) - \lambda^{-1}\Phi_N^+(1)\big]. \end{aligned} \tag{45}$$

Lemma 3 implies that

$$0 = (\mathscr{L}\Phi_N^-, 1+s) + 2\big[\lambda_N^-\Phi_N^-(-1) - \Phi_N^-(1)\big]$$

and

$$0 = (\mathscr{L}\Phi_N^+, 1-s) + 2\left[\Phi_N^+(-1) - \frac{1}{\lambda_N^+}\Phi_N^+(1)\right].$$

If these identities are used to eliminate $\Phi_N^-(1)$ and $\Phi_N^+(-1)$ in (45), then on simplifying the resulting expression, we arrive at the condition

$$0 = (1-\gamma)\Phi_N^-(-1)(\mathscr{L}\Phi_N^+, 1-s)(\lambda - \lambda_N^-) + (1+\gamma)\Phi_N^+(1)(\mathscr{L}\Phi_N^-, 1+s)\left(\frac{1}{\lambda_N^+} - \frac{1}{\lambda}\right).$$

Finally, using (38) and (40), we obtain

$$(\mathscr{L}\Phi_N^+, 1+s) = 2(\mathscr{L}\Phi_N^+, 1) = \frac{2N!}{(2N+1)!}(-\mathrm{i}\Omega)^{N+1}$$

and similarly, using (41) and (42),

$$(\mathscr{L}\Phi_N^-, 1-s) = 2(\mathscr{L}\Phi_N^-, 1) = -\frac{2N!}{(2N+1)!}(\mathrm{i}\Omega)^{N+1}.$$

With the aid of these expressions the condition becomes

$$0 = (1-\gamma)\Phi_N^-(-1)(\lambda - \lambda_N^-) + (-1)^N(1+\gamma)\Phi_N^+(1)\left(\frac{1}{\lambda_N^+} - \frac{1}{\lambda}\right),$$

and the result then follows as claimed thanks to (39) and (43). □

Lemma 4 establishes the condition (44) for the existence of an eigenvalue and thereby proves the conjecture of Hu and Atkins [14, Eq. (35)].

4.2. Properties of the eigenvalues

Let $N \in \mathbb{N}$ and $\Omega \in \mathbb{R}$. Denote $H_N = {}_1\mathrm{F}_1(-N; -2N-1; -\mathrm{i}\Omega)$ and define

$$\lambda_{\mathrm{S}} = (-1)^{N+1} \frac{1+\gamma}{1-\gamma} \frac{H_N^*}{H_N} \mathrm{e}^{-\mathrm{i}\Omega}, \quad \gamma \neq 1. \tag{46}$$

The next result characterises the solutions of the algebraic eigenvalue equation as approximations to the physical mode $\lambda \approx \mathrm{e}^{\mathrm{i}\Omega}$ and the spurious mode $\lambda \approx \lambda_{\mathrm{S}}$. The relative error in both approximations is shown to have the same magnitude, which in turn is dictated by the remainder in the Padé approximants:

Theorem 7. *If $\gamma \neq \pm 1$, then there are two distinct eigenvalues $\lambda \approx \mathrm{e}^{\mathrm{i}\Omega}$ and $\lambda \approx \lambda_{\mathrm{S}}$. Furthermore, the relative error in these approximations is given by ρ_N and $-\rho_N$ respectively, where*

$$\rho_N = \frac{(1-\gamma)H_N \mathrm{e}^{\mathrm{i}\Omega}\mathscr{E}_N + (-1)^{N+1}(1+\gamma)H_N^* \mathrm{e}^{-\mathrm{i}\Omega}\mathscr{E}_N^*}{(1-\gamma)H_N \mathrm{e}^{\mathrm{i}\Omega} + (-1)^N(1+\gamma)H_N^* \mathrm{e}^{-\mathrm{i}\Omega}} + \mathscr{O}(|\mathscr{E}_N|^2) \tag{47}$$

and $\mathscr{E}_N$ is the relative error in the Padé approximant,

$$\mathscr{E}_N = \frac{\mathrm{e}^{\mathrm{i}\Omega} - [N+1/N]_{\mathrm{e}^{\mathrm{i}\Omega}}}{\mathrm{e}^{\mathrm{i}\Omega}}. \tag{48}$$

Proof. Let $\mathscr{E}_N$ be defined as above, then

$$\lambda_N^- = [N+1/N]_{\mathrm{e}^{\mathrm{i}\Omega}} = \mathrm{e}^{\mathrm{i}\Omega}(1-\mathscr{E}_N)$$

and

$$\frac{1}{\lambda_N^+} = [N+1/N]_{\mathrm{e}^{-\mathrm{i}\Omega}} = \mathrm{e}^{-\mathrm{i}\Omega}(1-\mathscr{E}_N^*).$$

Inserting these expressions into condition (44) gives

$$0 = (1-\gamma)H_N(\lambda - \mathrm{e}^{\mathrm{i}\Omega} + \mathrm{e}^{\mathrm{i}\Omega}\mathscr{E}_N) + (-1)^N(1+\gamma)H_N^*(\lambda - \mathrm{e}^{\mathrm{i}\Omega} - \lambda\mathscr{E}_N^*)\frac{1}{\lambda \mathrm{e}^{\mathrm{i}\Omega}},$$

or, on rearranging,

$$\frac{\mathrm{e}^{\mathrm{i}\Omega} - \lambda}{\mathrm{e}^{\mathrm{i}\Omega}}\left[(1-\gamma)H_N \mathrm{e}^{\mathrm{i}\Omega} + (-1)^N(1+\gamma)H_N^* \lambda^{-1}\right] = (1-\gamma)H_N \mathrm{e}^{\mathrm{i}\Omega}\mathscr{E}_N + (-1)^{N+1}(1+\gamma)H_N^* \mathrm{e}^{-\mathrm{i}\Omega}\mathscr{E}_N^*. \tag{49}$$

If $\mathscr{E}_N \approx 0$, then Eq. (49) has roots at $\lambda \approx \mathrm{e}^{\mathrm{i}\Omega}$ and $\lambda \approx \lambda_{\mathrm{S}}$ (provided $\gamma \neq 1$), which depend continuously on $\mathscr{E}_N$. As $\mathscr{E}_N \to 0$, passing along the branch corresponding to $\mathrm{e}^{\mathrm{i}\Omega}$, the second term in parentheses on the left hand side of (49) tends to

$$(1-\gamma)H_N \mathrm{e}^{\mathrm{i}\Omega} + (-1)^N(1+\gamma)H_N^* \mathrm{e}^{-\mathrm{i}\Omega},$$

and Eq. (49) then implies that the relative error in the approximation of this zero by $\mathrm{e}^{\mathrm{i}\Omega}$ is given by ρ_N.

The left hand side of (49) may be rewritten as

$$\frac{\lambda-\lambda_S}{\lambda}\left[(1-\gamma)H_N\,\mathrm{e}^{\mathrm{i}\Omega}+(-1)^N(1+\gamma)H_N^*\,\mathrm{e}^{-\mathrm{i}\Omega}+\mathcal{O}(\lambda-\lambda_S)\right].$$

As $\mathscr{E}_N\to 0$, passing now along the branch corresponding to λ_S, this expression approaches

$$\frac{\lambda-\lambda_S}{\lambda_S}\left[(1-\gamma)H_N\,\mathrm{e}^{\mathrm{i}\Omega}+(-1)^N(1+\gamma)H_N^*\,\mathrm{e}^{-\mathrm{i}\Omega}\right],$$

and it follows that the relative error in the approximation of the second zero by λ_S is given by $-\rho_N$. □

5. Proofs of main results

Finally, we present the proofs of the results described in Section 2.3.

5.1. Proof of Theorem 1

Let ω and $\boldsymbol{k}\in\mathbb{R}^d$ satisfy the hypothesis. Consider an arbitrary element $K=\prod_{\ell=1}^d(a_\ell,b_\ell)$. For each $\ell=1,\ldots,d$, we begin by defining a function u_ℓ^{DG} by the rule $u^{\mathsf{DG}}(x_\ell)=\Phi(s)$, $s=(2x_\ell-a_\ell-b_\ell)/h$, where Φ is a non-trivial solution of the eigenvalue problem (35) with $\Omega=hk_\ell$ and λ_ℓ chosen according to Theorem 7. Performing the change of variable indicated above, we arrive at the conclusion that $u_\ell^{\mathsf{DG}}\in\mathbb{P}_N(a_\ell,b_\ell)$ satisfies

$$\begin{aligned}(\partial u_\ell^{\mathsf{DG}},v)_\ell&+\frac{1}{2}(1-\gamma)\big(\lambda_\ell u_\ell^{\mathsf{DG}}(a_\ell^+)-u_\ell^{\mathsf{DG}}(b_\ell^-)\big)v(b_\ell^-)+\frac{1}{2}(1+\gamma)\big(u_\ell^{\mathsf{DG}}(a_\ell^+)-\lambda_\ell^{-1}u_\ell^{\mathsf{DG}}(b_\ell^-)\big)v(a_\ell^+)\\&=\mathrm{i}k_\ell(u_\ell^{\mathsf{DG}},v)_\ell\end{aligned}\tag{50}$$

for all $v\in\mathbb{P}_N$, where $(\cdot,\cdot)_\ell$ denotes the L^2-inner product on (a_ℓ,b_ℓ). The restriction of the function u^{DG} to element K is defined to be

$$u_K^{\mathsf{DG}}(\boldsymbol{x},t)=c\,\mathrm{e}^{-\mathrm{i}\omega t}\prod_{\ell=1}^d u_\ell^{\mathsf{DG}}(x_\ell).$$

The value of the function u^{DG} on remaining elements is then *defined* so that Eq. (9) holds automatically. Specifically, the discrete wave-vector is defined by $\mathrm{e}^{\mathrm{i}h\tilde{k}_\ell}=\lambda_\ell$ and to obtain u^{DG} on any remaining element K', we use (9) with $\boldsymbol{x}\in K$ and $h\boldsymbol{m}$ chosen to be the position vector of the centroid of K' relative to the centroid of K. Obviously $u_{K'}^{\mathsf{DG}}\in\mathbb{P}_N$. Moreover, choosing $\boldsymbol{x}\in K$, $\tau=0$ and $\boldsymbol{m}=m\boldsymbol{e}_\ell$ in (9) gives

$$u^{\mathsf{DG}}(\boldsymbol{x}+mh\boldsymbol{e}_\ell,t)=\mathrm{e}^{\mathrm{i}hm\tilde{k}_\ell}u^{\mathsf{DG}}(\boldsymbol{x},t)=\lambda_\ell^m u^{\mathsf{DG}}(\boldsymbol{x},t),\quad m\in\mathbb{Z}$$

and then inserting the expression for u^{DG} and simplifying, we obtain

$$u_\ell^{\mathsf{DG}}(x_\ell+mh)=\lambda_\ell^m u_\ell^{\mathsf{DG}}(x_\ell),\quad x_\ell\in(a_\ell,b_\ell).$$

By first selecting $x_\ell=b_\ell^-$ and $m=-1$, and then $x_\ell=a_\ell^+$ and $m=1$, we find

$$u_\ell^{\mathsf{DG}}(b_\ell^\pm)=\lambda_\ell u_\ell^{\mathsf{DG}}(a_\ell^\pm),$$

and hence, with the aid of (50), we arrive at the conclusion

$$(\partial u_\ell^{\mathsf{DG}}, v)_\ell + \frac{1}{2}(1-\gamma)\big(u_\ell^{\mathsf{DG}}(b_\ell^+) - u_\ell^{\mathsf{DG}}(b_\ell^-)\big)v(b_\ell^-) + \frac{1}{2}(1+\gamma)\big(u_\ell^{\mathsf{DG}}(a_\ell^+) - u_\ell^{\mathsf{DG}}(a_\ell^-)\big)v(a_\ell^+)$$
$$= \mathrm{i}k_\ell(u_\ell^{\mathsf{DG}}, v)_\ell \tag{51}$$

for all $v \in \mathbb{P}_N$.

It remains to show that u^{DG} satisfies (7) or equivalently (8). Thanks to property (9), it is sufficient to prove (8) holds on the particular element K. Let a general test function v be expressed in the form $\prod_{\ell=1}^d v_\ell(x_\ell)$ where $v_\ell \in \mathbb{P}_N$. Inserting these expressions into the statement (8), simplifying using (6) and cancelling a factor $c\,\mathrm{e}^{-\mathrm{i}\omega t}$, shows that (8) is equivalent to the following condition:

$$\mathrm{i}\omega \prod_{\ell=1}^d (v_\ell, u_\ell^{\mathsf{DG}})_\ell = \sum_{m=1}^d \prod_{\ell\neq m} (v_\ell, u_\ell^{\mathsf{DG}})_\ell \Big\{ (v_m, \alpha_m \partial_m u_m^{\mathsf{DG}})_m + \Lambda_\gamma^-(\boldsymbol{e}_m)\big(u_m^{\mathsf{DG}}(b_m^+) - u_m^{\mathsf{DG}}(b_m^-)\big)v_m(b_m^-)$$
$$+ \Lambda_\gamma^+(\boldsymbol{e}_m)\big(u_m^{\mathsf{DG}}(a_m^+) - u_m^{\mathsf{DG}}(a_m^-)\big)v_m(a_m^+)\Big\},$$

where $\boldsymbol{e}_m$ is the mth unit vector. The assumption $\alpha_m \geqslant 0$ implies that $\Lambda_\gamma^\pm(\boldsymbol{e}_m) = \frac{1}{2}(1 \pm \gamma)\alpha_m$, which in turn shows that the expression in parentheses reduces to α_m multiplied by the left hand side of (51). Condition (8) is therefore equivalent to

$$\mathrm{i}\omega \prod_{\ell=1}^d (v_\ell, u_\ell^{\mathsf{DG}})_\ell = \mathrm{i}\boldsymbol{\alpha}\cdot\boldsymbol{k} \prod_{\ell=1}^d (v_\ell, u_\ell^{\mathsf{DG}})_\ell, \quad \forall v_\ell \in \mathbb{P}_N$$

or equally well, $\omega = \boldsymbol{\alpha}\cdot\boldsymbol{k}$. The statements concerning the discrete wave-vector follow at once from Theorem 7.

5.2. Proof of Theorem 2

By Lemma 1, the relative error in the Padé approximant is given by

$$\mathscr{E}_N = -\frac{1}{2}\Omega^{2N+2}\left[\frac{N!}{(2N+1)!}\right]^2 \left\{1 - \frac{2\mathrm{i}\Omega(N+1)}{(2N+1)(2N+3)} + \mathcal{O}(\Omega^2)\right\}.$$

Furthermore, since

$$H_N = {}_1\mathrm{F}_1(-N; -2N-1; -\mathrm{i}\Omega) = 1 - \frac{N}{2N+1}\mathrm{i}\Omega + \cdots,$$

we have

$$H_N\,\mathrm{e}^{\mathrm{i}\Omega} = 1 + \frac{N+1}{2N+1}\mathrm{i}\Omega + \cdots$$

Inserting these expressions into (47) gives

$$\rho_N = -\frac{1}{2}\Omega^{2N+2}\left[\frac{N!}{(2N+1)!}\right]^2 \left\{\frac{(2N+1)q_{N+1} + q_N(N+1)\mathrm{i}\Omega}{(2N+1)q_N + q_{N+1}(N+1)\mathrm{i}\Omega} - \frac{2\mathrm{i}\Omega(N+1)}{(2N+1)(2N+3)} + \cdots\right\},$$

where

$$q_N = (1-\gamma) + (-1)^N(1+\gamma).$$

Case 1: Suppose $\gamma \neq 0$. It is then straightforward to verify that the term in the second set of parentheses simplifies to $Q_N(\gamma^{(-1)^N})$, giving the result claimed. Case 2: If $\gamma = 0$, then q_{N+1} vanishes for even N, and the expression simplifies to $(14)_1$. Equally well, if N is odd, then q_N vanishes, and the expression reduces to $(14)_2$ in this case.

5.3. Proofs of Theorems 3 and 4

The first two parts of Theorem 3 are restatements of the first two parts of Theorem 5. The final part of Theorem 3 follows by inserting the estimates from the first part of Theorem 6 into the expression (47) and simplifying. The proof of Theorem 4 follows in the same way, using instead the second part of Theorem 6.

Acknowledgements

Support for this work from the Leverhulme Trust through a Leverhulme Trust Fellowship is gratefully acknowledged. This work was completed while the author was visiting the Newton Institute for Mathematical Sciences, Cambridge, UK.

References

[1] M. Ainsworth, Discrete dispersion relation for *hp*-version finite element approximation at high wave number, Technical Report 5, Strathclyde University, 2003, SIAM J. Numer. Anal., in press.
[2] D. Arnold, F. Brezzi, B. Cockburn, L. Marini, Unified analysis of discontinuous Galerkin methods for elliptic problems, SIAM J. Numer. Anal. 39 (2001/02) 1749–1779 (electronic).
[3] J. Astley, K. Gerdes, D. Givoli, I. Harari, Special issue on Finite elements for wave problems – Preface, J. Comput. Acoust. 8 (2000) vii–ix.
[4] K.S. Bey, J.T. Oden, *hp*-version discontinuous Galerkin methods for hyperbolic conservation laws, Comput. Methods Appl. Mech. Engrg. 133 (1996) 259–286.
[5] R. Biswas, K.D. Devine, J.E. Flaherty, Parallel, adaptive finite element methods for conservation laws, Appl. Numer. Math. 14 (1994) 255–283.
[6] B. Cockburn, G. Karniadakis, C.-W. Shu, The development of discontinuous Galerkin methods, in: Discontinuous Galerkin methods (Newport, RI, 1999), Lecture Notes in Computer Science and Engineering, vol. 11, Springer, Berlin, 2000, pp. 3–50.
[7] K. Driver, N. Temme, Zero and pole distribution of diagonal Padé approximants to the exponential function, Quaest. Math. 22 (1999) 7–17.
[8] R. Dyson, Technique for very high order nonlinear simulation and validation, J. Comput. Acoust. 10 (2002) 211–229.
[9] A. Erdélyi, W. Magnus, F. Oberhettinger, F. Tricomi, Higher Transcendental Functions, Bateman Manuscript Project, McGraw-Hill, New York, London, 1953–55.
[10] D. Gottlieb, J. Hesthaven, Spectral methods for hyperbolic problems, J. Comput. Appl. Math. 128 (2001) 83–131; Numerical analysis 2000, vol. VII, Partial differential equations.
[11] I. Gradshteyn, I. Ryzhik, in: A. Jeffrey (Ed.), Table of Integrals, Series and Products, fifth ed., Academic Press, United Kingdom, 1994.
[12] J. Hesthaven, T. Warburton, Nodal high-order methods on unstructured grids. I. Time-domain solution of Maxwell's equations, J. Comput. Phys. 181 (2002) 186–221.
[13] P. Houston, C. Schwab, E. Süli, Discontinuous *hp*-finite element methods for advection-diffusion-reaction problems, SIAM J. Numer. Anal. 39 (2002) 2133–2163 (electronic).
[14] F. Hu, H. Atkins, Eigensolution analysis of the discontinuous Galerkin method with non-uniform grids. Part 1: One space dimension, J. Comput. Phys. 182 (2002) 516–545.
[15] F. Hu, H. Atkins, Two dimensional wave analysis of the discontinuous Galerkin method with non-uniform grids and boundary conditions, in: Proceedings of the Eighth AIAA/CEAS Aeronautics Conference, Breckenridge, Colorado, June 2002. AIAA paper no. 2002-2514.
[16] F. Hu, M. Hussaini, P. Rasetarinera, An Analysis of the Discontinuous Galerkin Method for Wave Propagation Problems, vol. 151, 1999, pp. 921–946.

[17] F. Ihlenburg, Finite element analysis of acoustic scattering, in: Applied Mathematical Sciences, vol. 132, Springer-Verlag, Berlin, 1998.
[18] F. Ihlenburg, I. Babuška, Finite element solution of the Helmholtz equation with high wave number. Part 2. The *hp* version of the finite element method, SIAM J. Numer. Anal. 34 (1997) 315–358.
[19] Y. Luke, in: The Special Functions and Their Approximation, vol. 2, Academic Press, New York, 1969.
[20] F. Odeh, J. Keller, Partial differential equations with periodic coefficients and Bloch waves in crystals, J. Math. Phys. 5 (1964) 1499–1504.
[21] F. Olver, Asymptotics and special functions, in: Computer Science and Applied Mathematics, Academic Press, New York, 1974.
[22] H. Padé, Sur la répresentation approchée d'une fonction par des fractions rationnelles, Ann. de l'Éc. Nor. 9 (1892).
[23] S. Sherwin, Dispersion analysis of the continuous and discontinuous Galerkin formulations, in: Discontinuous Galerkin methods (Newport, RI, 1999), Lecture Notes in Computer Science and Engineering, vol. 11, Springer, Berlin, 2000, pp. 425–431.
[24] L. Slater, Confluent Hypergeometric Functions, Cambridge University Press, London, 1960.
[25] L. Thompson, P. Pinsky, Complex wavenumber Fourier analysis of the *p*-version finite element method, Comput. Mech. 13 (1994) 255–275.
[26] R. Varga, On higher order stable implicit methods for solving parabolic partial differential equations, J. Math. Phys. XL (1961) 220–231.
[27] T. Warburton, I. Lomtev, Y. Du, S. Sherwin, G. Karniadakis, Galerkin and discontinuous Galerkin spectral/*hp* methods, Comput. Methods Appl. Mech. Engrg. 175 (1999) 343–359.