




# Dispersive Behavior of an Energy-Conserving Discontinuous Galerkin Method for the One-Way Wave Equation

Mark Ainsworth<sup>1</sup> · Guosheng Fu<sup>1</sup> 

Received: 25 June 2018 / Revised: 10 September 2018 / Accepted: 3 October 2018 /  
Published online: 8 October 2018  
© Springer Science+Business Media, LLC, part of Springer Nature 2018

## Abstract

The dispersive behavior of the recently proposed energy-conserving discontinuous Galerkin (DG) method by Fu and Shu (Optimal energy-conserving discontinuous Galerkin methods for linear symmetric hyperbolic systems, 2018. [arXiv:1804.10307](https://arxiv.org/abs/1804.10307)) is analyzed and compared with the classical centered and upwinding DG schemes. It is shown that the new scheme gives a significant improvement over the classical centered and upwinding DG schemes in terms of dispersion error. Numerical results are presented to support the theoretical findings.

**Keywords** Discontinuous Galerkin method · Energy conserving · Dispersion analysis

## 1 Introduction

The quest for stable and accurate schemes for systems of hyperbolic conservation laws has occupied researchers for several decades and continues to this day [1,2] with active research into finite difference methods, finite volume methods, spectral methods and a variety of finite element Galerkin schemes. The current consensus seems to be that discontinuous Galerkin (DG) schemes [6] are the most promising, although they too have their drawbacks even if one restricts attention to linear hyperbolic systems. In this setting, one wishes to have numerical schemes which are able to propagate discrete waves at, or near to, the same speed at which continuous waves are propagated by the original hyperbolic system. The dispersive and dissipative behavior of a numerical scheme compared with that of the original system is of considerable interest and had been widely studied [3–5,8,9,11].

This paper is devoted to a dispersion analysis of the recently proposed energy-conserving DG method [10]. We shall focus on the continuous-in-time semidiscrete scheme. The dispersion analysis for the fully discrete scheme is the subject of ongoing work. To fix ideas, we consider the following one-way wave equation with unit wave speed:

---

✉ Guosheng Fu  
guosheng\_fu@brown.edu

Mark Ainsworth  
Mark\_Ainsworth@brown.edu

<sup>1</sup> Division of Applied Mathematics, Brown University, 182 George St, Providence, RI 02912, USA

$$u_t + u_x = 0, \quad x \in \mathbb{R}, t > 0, \tag{1.1}$$

for suitable initial data. To begin with, we confine our attention to uniform partitions of  $\mathbb{R}$  consisting of cells of size  $h > 0$ , whose nodes are located at the points  $h(\mathbb{Z} + 1/2)$ . Denote the  $j$ th cell  $I_j = ((j - 1/2)h, (j + 1/2)h)$ , and let  $V_h^N$  denote the space of piecewise continuous polynomials of degree  $N$  on the partition:

$$V_h^N = \{v \in L^2(\mathbb{R}) : v|_{I_j} \in \mathbb{P}_N(I_j), \quad \forall j \in \mathbb{Z}\}, \tag{1.2}$$

where  $\mathbb{P}_N(I_j)$  denotes the set of polynomials of degree up to  $N \geq 0$  defined on the cell  $I_j$ . For any function  $p \in V_h^N$ , we let  $p_{j-1/2}^-$  and  $p_{j-1/2}^+$  be the values of  $p$  at the node  $x_{j-1/2} = (j - 1/2)h$ , from the left cell,  $I_{j-1}$ , and from the right cell,  $I_j$ , respectively. In what follows, we employ  $\llbracket p \rrbracket|_{j-1/2} = p_{j-1/2}^+ - p_{j-1/2}^-$  and  $\{\{p\}\}|_{j-1/2} = \frac{1}{2}(p_{j-1/2}^+ + p_{j-1/2}^-)$  to represent the jump and the mean value of  $p$  at each node.

The DG method for (1.1) reads as follows: Find the unique function  $u_h = u_h(t) \in V_h^N$  such that

$$\int_{I_j} (u_h)_t v_h dx - \int_{I_j} u_h (v_h)_x dx + \widehat{u}_h v_h^-|_{j+1/2} - \widehat{u}_h v_h^+|_{j-1/2} = 0, \tag{1.3}$$

holds for all  $v_h \in V_h^N$  and all  $j \in \mathbb{Z}$ . The classical upwinding DG method, denoted by (U), uses numerical fluxes chosen to be

$$\widehat{u}_h|_{j-1/2} = \{\{u_h\}\}|_{j-1/2} + \frac{1}{2}\llbracket u_h \rrbracket|_{j-1/2},$$

while the centered DG method, denoted by (C), uses numerical fluxes given by

$$\widehat{u}_h|_{j-1/2} = \{\{u_h\}\}|_{j-1/2}.$$

The method (U) is energy dissipative in the sense that

$$\frac{1}{2} \frac{d}{dt} \int_{\mathbb{R}} u_h^2 dx = - \sum_{j \in \mathbb{Z}} \frac{1}{2} (\llbracket u_h \rrbracket)^2|_{j-1/2} \leq 0, \tag{1.4}$$

while the method (C) is energy-conservative

$$\frac{1}{2} \frac{d}{dt} \int_{\mathbb{R}} u_h^2 dx = 0.$$

Despite being energy conserving, the centered flux scheme (C) is seldom used in practice owing to the reduced stability properties of the scheme compared with the upwinding scheme (U), c.f. [7]. For this reason, the scheme (U) is often preferred and the lack of energy conservation tolerated. Expression (1.4) shows that if the jump terms  $\llbracket u_h \rrbracket|_{j-1/2}$  are non-zero then energy will be dissipated and, importantly, that there is no mechanism whereby the dissipated energy can be regained by the scheme.

Recently, Fu and Shu [10] proposed an energy-conserving discontinuous Galerkin (DG) method for linear symmetric hyperbolic systems, and gave an optimal a priori error estimate for the method in one dimension, and in multi-dimensions on tensor-product meshes. Numerical evidence presented in [10] suggests that the scheme is optimally convergent on general triangular meshes, and has superior dispersive properties of the new DG method comparing with the (energy dissipative) upwinding DG method (U) and the (energy conservative) centered DG method (C) translating into improved accuracy for long time simulations.

The method of Fu and Shu is unusual in that it begins at the continuous level by introducing an auxiliary advection equation (with the opposite wave speed to that in the equation for  $u$ ), to obtain the following (decoupled) system:

$$u_t + u_x = 0, \quad x \in \mathbb{R}, t > 0, \tag{1.5a}$$

$$\phi_t - \phi_x = 0, \quad x \in \mathbb{R}, t > 0, \tag{1.5b}$$

with initial condition  $u(x, 0) = u_0(x)$  and  $\phi(x, 0) = 0$ . Obviously the solution  $\phi$  is identically zero. However, this will not be the case for the DG approximation [10] of the system, where the (non-zero) approximation of the second equation is exploited to obtain energy conservation at the discrete level.

The DG method [10] for (1.5) reads as follows: Find the unique function  $(u_h, \phi_h) = (u_h(t), \phi_h(t)) \in V_h^N \times V_h^N$  such that

$$\int_{I_j} (u_h)_t v_h dx - \int_{I_j} u_h (v_h)_x dx + \widehat{u}_h v_h^-|_{j+\frac{1}{2}} - \widehat{u}_h v_h^+|_{j-\frac{1}{2}} = 0, \tag{1.6a}$$

$$\int_{I_j} (\phi_h)_t \psi_h dx + \int_{I_j} \phi_h (\psi_h)_x dx - \widehat{\phi}_h \psi_h^-|_{j+\frac{1}{2}} + \widehat{\phi}_h \psi_h^+|_{j-\frac{1}{2}} = 0, \tag{1.6b}$$

holds for all  $(v_h, \psi_h) \in V_h^N \times V_h^N$  and all  $j \in \mathbb{Z}$ , where  $\widehat{u}_h$  and  $\widehat{\phi}_h$  denote the numerical fluxes

$$\widehat{u}_h|_{j-\frac{1}{2}} = \{ \{u_h\} \}|_{j-\frac{1}{2}} + \frac{1}{2} \alpha \llbracket \phi_h \rrbracket|_{j-\frac{1}{2}}, \tag{1.7a}$$

$$\widehat{\phi}_h|_{j-\frac{1}{2}} = \{ \{ \phi_h \} \}|_{j-\frac{1}{2}} + \frac{1}{2} \alpha \llbracket u_h \rrbracket|_{j-\frac{1}{2}}. \tag{1.7b}$$

The constant in the numerical fluxes (1.7) is chosen to be  $\alpha = 1$  in [10], and we denote the corresponding DG method by (A). However, in this article, we will also consider the following choice

$$\alpha = \begin{cases} \sqrt{\frac{4}{3}} & \text{if } N = 0, \\ \sqrt{\frac{N(2N+3)}{(N+1)(2N+1)}} & \text{if } N \text{ is odd,} \\ \sqrt{\frac{(N+1)(2N+1)}{N(2N+3)}} & \text{if } N > 0 \text{ is even,} \end{cases} \tag{1.8}$$

and denote the corresponding DG method by (A\*). This choice of  $\alpha$  is obtained using symbolic calculation (up to degree  $N = 17$ ), aiming at minimizing the dispersion error (see Table 1 below).

Each of methods (A) and (A\*) are energy conservative [10] with respect to the following modified energy:

$$\frac{1}{2} \frac{d}{dt} \int_{\mathbb{R}} (u_h^2 + \phi_h^2) dx = 0. \tag{1.9}$$

Of course, this does not mean that the individual energy  $\int_{\mathbb{R}} u_h^2 dx$  and  $\int_{\mathbb{R}} \phi_h^2 dx$  are conserved in isolation. One way to view the scheme (1.6) is to regard the auxiliary variable  $\phi_h$  as a temporary store for collecting energy dissipated in (1.6a) which is then reinjected back into the equation for  $u_h$  through the flux term  $\llbracket \phi_h \rrbracket|_{j-\frac{1}{2}}$  in (1.7a), still resulting in the overall energy of the system being conserved as shown by (1.9). More interesting is that this exchange of energy in  $u_h$  and  $\phi_h$  also seems to render methods (A) and (A\*) superior to the method (U) and (C) in terms of numerical dispersion, as we shall see in Sect. 2.

Section 3 contains a summary of the main results from our dispersion analysis in Sect. 4, and an explanation of the numerical results on uniform meshes conducted in Sect. 2. Conclusions are drawn in Sect. 5.

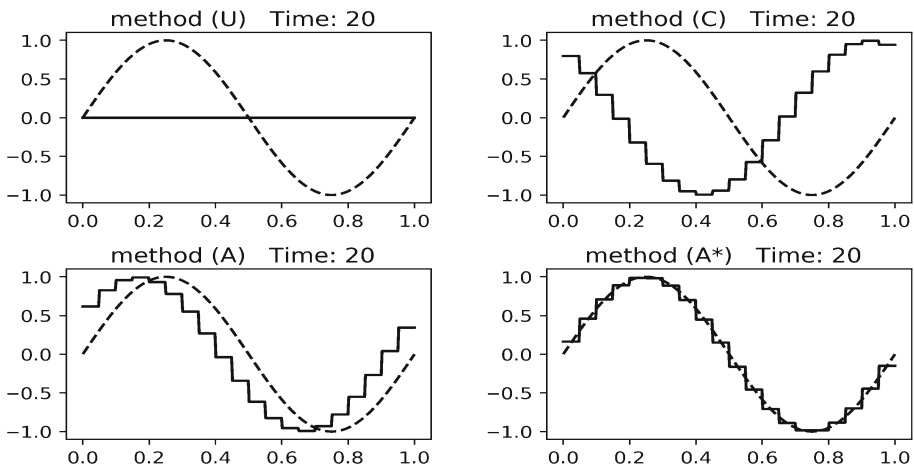
## 2 Illustration of Dispersive Behavior of the DG Schemes

In this section we carry out a simple numerical comparison of the above mentioned four DG methods. We consider Eq. (1.5) on the unit interval  $I = [0, 1]$  with periodic boundary conditions, and take the initial condition  $u_0(x) = \sin(\omega x)$  with frequency  $\omega = 2\pi$ . Hence the true solution is  $u(x, t) = \sin(2\pi(x - t))$ . Since we are primarily interested in the spatial discretisation, we use a sufficiently high-order time discretization so as to render the temporal error negligible compared with the spatial error.

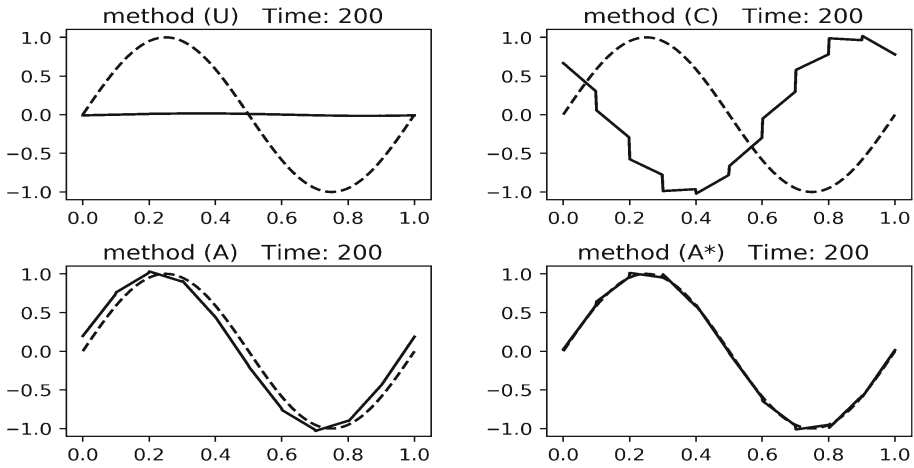
Numerical results for the four DG methods mentioned above with polynomial degree  $N = 0$  on 20 uniform cells at time  $T = 20$ , with polynomial degree  $N = 1$  on 10 uniform cells at time  $T = 200$ , and with polynomial degree  $N = 2$  on 4 uniform cells at time  $T = 300$  are presented in Figs. 1, 2 and 3 respectively. One observes from these figures that the dissipative behavior of method (U), whilst the method (C) exhibits large phase error compared with method (A), which in turn is inferior to method (A\*).

For the  $N = 0$  case, we also compare the numerical approximations obtained at *different* times for the four methods in Fig. 4. It is striking that method (A\*) at time  $T = 1500$  enjoys a similar accuracy to that of method (C) at time  $T = 5$  and method (A) at time  $T = 20$ . In Sect. 3 we will give a theoretical explanation for these observations.

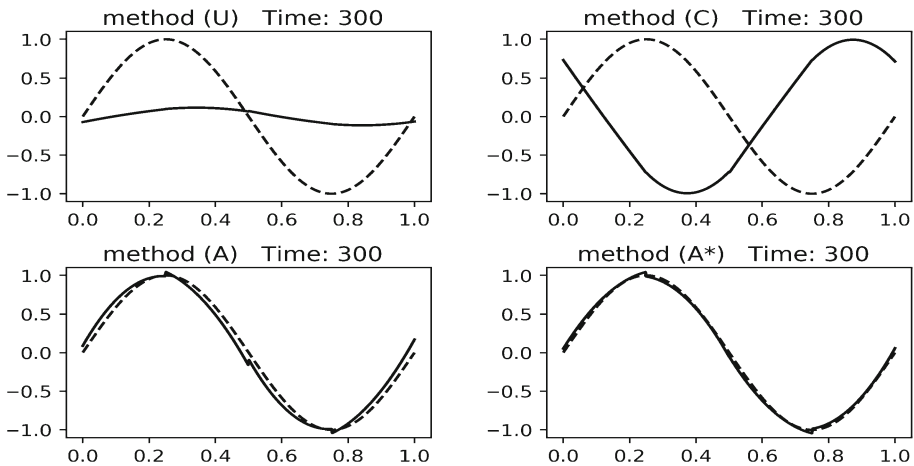
We also compare the numerical approximations obtained at time  $T = 40$  using methods (A) and (A\*) for  $N = 0$  on uniform and non-uniform meshes consisting of 20 cells in Figs. 5 and 6, respectively. The non-uniform mesh is obtained by applying a uniformly distributed 10% random perturbation of the nodes in an uniform mesh. Comparing the results on the uniform mesh with the corresponding results on the non-uniform mesh, we observe a similar phase error in the physical variable  $u_h$  in both cases. However, in the non-uniform case we



**Fig. 1** Numerical solution  $u_h$  at time  $T = 20$ . Solid line: numerical solution. Dashed line: exact solution.  $N = 0$ , 20 uniform cells



**Fig. 2** Numerical solution  $u_h$  at time  $T = 200$ . Solid line: numerical solution. Dashed line: exact solution.  $N = 1, 10$  uniform cells



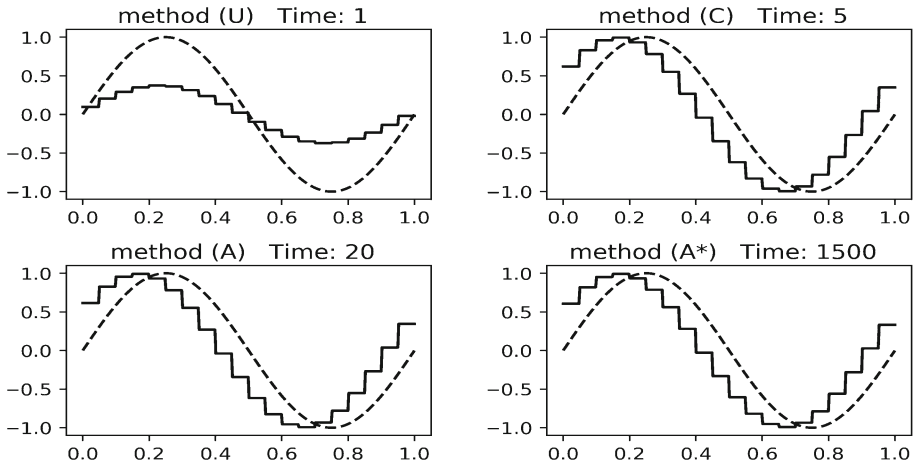
**Fig. 3** Numerical solution  $u_h$  at time  $T = 300$ . Solid line: numerical solution. Dashed line: exact solution.  $N = 2, 4$  uniform cells

observe a larger amount of energy leakage from the physical variable  $u_h$  to the auxiliary variable  $\phi_h$  for both methods (A) and (A\*), which is larger for method (A\*).

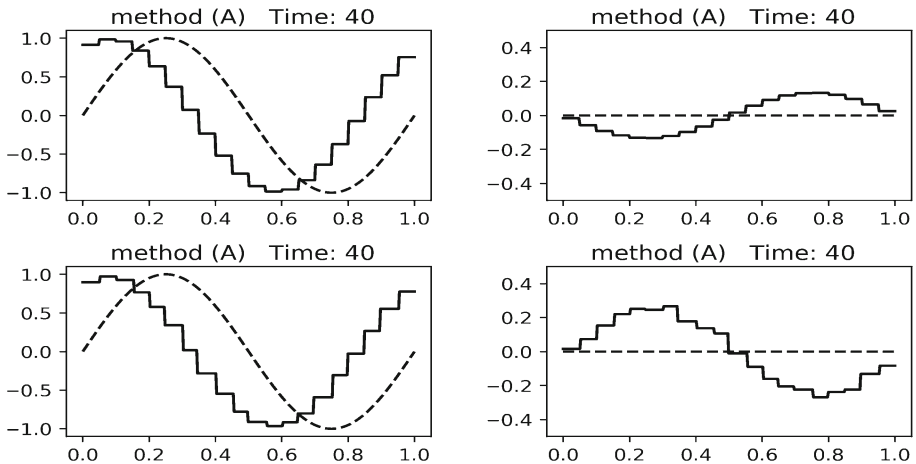
We mention that the results presented in Figs. 4, 5 and 6 are not peculiar to the lowest order case and numerical evidence (not reported in this article) indicate a similar behavior on both uniform and non-uniform meshes for  $N = 1$  and  $N = 2$ .

### 3 Main Results on the Dispersion Analysis

In this section we provide a theoretical explanation for the improved dispersive behavior of methods (A) and (A\*) compared with methods (U) and (C).



**Fig. 4** Numerical solution  $u_h$  at different times. Solid line: numerical solution. Dashed line: exact solution.  $N = 0, 20$  uniform cells



**Fig. 5** Numerical solution at time  $T = 40$  for methods (A). Left:  $u_h$ ; Right:  $\phi_h$ . Solid line: numerical solution. Dashed line: exact solution. First row: uniform mesh; Second row: non-uniform mesh.  $N = 0, 20$  cells

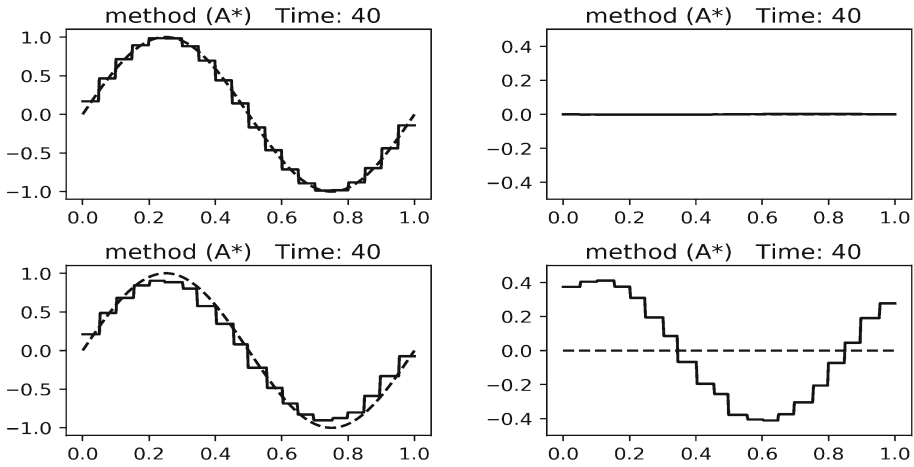
A key feature of the Eq. (1.5) is the existence of non-trivial, spatially propagating solutions for each given temporal frequency  $\omega$ ,

$$u(x, t) = e^{-i\omega t} U(x), \quad \phi(x, t) = e^{-i\omega t} \Phi(x), \tag{3.1}$$

where  $U(x) = e^{ikx}$  and  $\Phi(x) = e^{-ikx}$  with  $k = \omega$  the wavenumber. The functions  $U$  and  $\Phi$  satisfies a Bloch-wave condition

$$U(x + h) = \lambda^+ U(x), \quad \Phi(x + h) = \lambda^- \Phi(x), \quad x \in \mathbb{R}, h \in \mathbb{R}, \tag{3.2}$$

where  $\lambda^\pm = e^{\pm ikh}$  are the Floquet multipliers.



**Fig. 6** Numerical solution at time  $T = 40$  for method (A\*). Left:  $u_h$ ; Right:  $\phi_h$ . Solid line: numerical solution. Dashed line: exact solution. First row: uniform mesh; Second row: non-uniform mesh.  $N = 0, 20$  cells

### 3.1 The Main Results

In order to study the dispersive behavior of the discrete schemes, we seek the non-trivial *discrete* Bloch wave solutions of the DG scheme (1.6) in the form

$$u_{h,N}(x, t) = e^{-i\omega t} U_{h,N}(x), \quad \phi_{h,N}(x, t) = e^{-i\omega t} \Phi_{h,N}(x), \tag{3.3}$$

where  $U_{h,N}, \Phi_{h,N} \in V_h^N$  satisfy a discrete Bloch wave condition

$$U_{h,N}(x + h) = \lambda_{h,N} U_{h,N}(x), \quad \Phi_{h,N}(x + h) = \lambda_{h,N} \Phi_{h,N}(x), \quad x \in \mathbb{R}, h \in \mathbb{R}, \tag{3.4}$$

and where  $\lambda_{h,N}$  is the discrete Floquet multiplier.

The relative accuracy  $R_{h,N}$  of the Floquet multiplier approximation is defined by

$$R_{h,N} = \frac{\lambda^+ - \lambda_{h,N}}{\lambda^+} = \frac{e^{ikh} - \lambda_{h,N}}{e^{ikh}}. \tag{3.5}$$

The leading order terms in  $R_{h,N}$  for each of the four DG methods described in Sect. 1 are listed in Table 1. The results quoted for the methods (U) and (C) are special cases of the general result proved in [3, Theorem 2], whilst the results for the method (A) are special cases of the general result that will be proved here in Theorem 4.3 and Remark 4.4. The results for the method (A\*) were obtained using algebraic manipulation for particular choices of polynomial degree  $N$  from 0 up to degree 17.

The results given in Table 1 show that the accuracy of method (A) is of  $(2N + 3)$ -th order in  $\omega h$  and, as such, is always superior to the accuracy of methods (U) and (C) both in terms of the order of convergence and the magnitude of the coefficient of the leading term in the error. The method (A\*) is better still, providing  $(2N + 5)$ -th order of convergence in  $\omega h$ .

Let  $\text{Re}(\cdot)$  and  $\text{Im}(\cdot)$  be the real and imaginary parts of a complex number, respectively. We now examine the dissipation and dispersion errors of the schemes in the low-wavenumber limit where  $kh \ll 1$ . Let  $k_{h,N}$  be the *discrete wavenumber* that satisfy

$$e^{ik_{h,N}h} = \lambda_{h,N}, \quad \text{Re}(k_{h,N}h) \in [-\pi, \pi], \tag{3.6}$$

**Table 1** Leading terms of the relative error  $R_{h,N}$  in the approximation of the Floquet multiplier for the DG methods (U), (C), (A), and (A\*)

Degree	Method (U)	Method (C)	Method (A)	Method (A*)
0	$\frac{\Omega^2}{2} + i\frac{\Omega^3}{3}$	$-i\frac{\Omega^3}{6}$	$-i\frac{\Omega^3}{24}$	$-i\frac{\Omega^5}{180}$
1	$\frac{\Omega^4}{72} + i\frac{\Omega^5}{270}$	$i\frac{\Omega^3}{48}$	$-i\frac{\Omega^5}{1,080}$	$-i\frac{53\Omega^7}{302,400}$
2	$\frac{\Omega^6}{7200} + i\frac{\Omega^7}{42,000}$	$-i\frac{\Omega^7}{16,800}$	$-i\frac{\Omega^7}{252,000}$	$-i\frac{41\Omega^9}{63,504,000}$
$N \geq 1$	$C_N \left[ 1 + i\frac{(2N+2)\Omega}{(2N+1)(2N+3)} \right] \Omega^{2N+2} iC_N \begin{cases} -\frac{N+1}{2N+3} \Omega^{2N+3}, & N \text{ even} \\ \frac{2N+1}{N+1} \Omega^{2N+1}, & N \text{ odd} \end{cases} -i\frac{C_N \Omega^{2N+3}}{(2N+1)(2N+3)} -i\frac{E_N \Omega^{2N+5}}{(2N+1)^{2N+2}}$			

$\Omega = \omega h$ .  $C_N = \frac{1}{2} \left[ \frac{N!}{(2N+1)!} \right]^2$  for  $N \geq 1$ .  $E_N$ , up to four digits accuracy, is given in Table 2 for  $N \leq 17$

which approximates the true wavenumber  $k$ . For  $kh \ll 1$ , the relative error satisfies

$$R_{h,N} = \frac{e^{ikh} - e^{ik_h N h}}{e^{ikh}} \approx i(k - k_{h,N})h. \tag{3.7}$$

Hence, Table 1 shows that the *dispersion error* is

$$\text{Re}((k - k_{h,N})h) \approx \begin{cases} \frac{C_N(2N+2)}{(2N+1)(2N+3)}(hk)^{2N+3}, & \text{for method (U),} \\ -\frac{C_N(N+1)}{2N+3}(hk)^{2N+3}, & \text{for method (C), even } N, \\ \frac{C_N(2N+1)}{N+1}(hk)^{2N+1}, & \text{for method (C), odd } N, \\ -\frac{C_N}{(2N+1)(2N+3)}(hk)^{2N+3}, & \text{for method (A),} \\ -\frac{E_N}{(2N+1)^{2N+2}}(hk)^{2N+5}, & \text{for method (A*),} \end{cases}$$

and the *dissipation error* for method (U) is

$$\text{Im}((k - k_{h,N})h) \approx C_N(hk)^{2N+2},$$

whilst the dissipation error for methods (C), (A), and (A\*) vanishes since the discrete wave number  $k_{h,N}$  is a real number (due to the fact that  $|\lambda_{k,N}| = 1$  for these methods; see Remarks 4.5 and 4.8).

### 3.2 Explanation of Results Presented in Fig. 4

Now, let us apply the above results (for  $N = 0$ ) in Table 1 to explain the numerical results obtained in Fig. 4. For  $\Omega = \omega h \ll 1$ , the numerical solution obtained from each of the DG methods will satisfy

$$u_h(x, t) \approx \sin(\omega_h x - \omega t)$$

at the nodes, and the relative error  $R_{h,N} \approx i(\omega - \omega_h)h$ . Table 1 then implies

$$\begin{aligned} \omega_h &\approx \omega + i\frac{\Omega^2}{2h} = 2\pi + i\frac{\pi^2}{10} && \text{for method (U),} \\ \omega_h &\approx \omega + \frac{\Omega^3}{6h} = 2\pi + \frac{\pi^3}{300} && \text{for method (C),} \end{aligned}$$



$$\omega_h \approx \omega + \frac{\Omega^3}{24h} = 2\pi + \frac{\pi^3}{1,200} \quad \text{for method (A),}$$

$$\omega_h \approx \omega + \frac{\Omega^5}{180h} = 2\pi + \frac{\pi^5}{900,000} \quad \text{for method (A*),}$$

where  $h = 1/20$ , and  $\Omega = \omega h = \pi/10$ . In particular, the maximum value of the solution at time  $T = 1$  for method (U) will, thanks to numerical dissipation, not be unity but will instead take a value close to

$$e^{-\frac{\pi^2}{10} \times 1} \approx 0.37,$$

which is in close agreement with the top left figure of Fig. 4. The phase lag for method (C) at time  $T = 5$  will be close to  $\frac{\Omega^3}{6h} \times \frac{5}{\omega} \approx 0.08$ , while for method (A) at time  $T = 20$  will be close to  $\frac{\Omega^3}{24h} \times \frac{20}{\omega} \approx 0.08$ , and for (A\*) at time  $T = 1500$  will be close to  $\frac{\Omega^5}{180h} \times \frac{1500}{\omega} \approx 0.08$ . All of these predictions are in close agreement with results in Fig. 4.

### 4 Dispersion Analysis: The Eigenvalue Problem

In this section, we provide proofs of the dispersion analysis of the semi-discrete scheme (1.6) leading to the results stated in Table 1. We closely follow the analysis in [3] and begin by seeking a non-trivial bloch-wave solution of the form

$$u_h(x, t) = e^{-i\omega t} \sum_{m \in \mathbb{Z}} \lambda^m U(x - mh), \tag{4.1a}$$

$$\phi_h(x, t) = e^{-i\omega t} \sum_{m \in \mathbb{Z}} \lambda^m \Phi(x - mh), \tag{4.1b}$$

where  $U, \Phi \in V_h^N$ . Denoting  $\Omega = \omega h$ , and transforming the domain over which the scheme (1.6) is posed to the reference interval  $[-1, 1]$ , we obtain the following eigenvalue problem which determines the value of the discrete Floquet multiplier  $\lambda$ : Find  $U, \Phi \in \mathbb{P}_N$  and  $\lambda \in \mathbb{C}$  such that

$$\begin{aligned} &-\frac{1}{2}i\Omega(U, v) + (U', v) + \frac{1}{2}(\lambda U(-1) - U(1) + \alpha(\lambda \Phi(-1) - \Phi(1)))v(1) \\ &+ \frac{1}{2}(U(-1) - \lambda^{-1}U(1) - \alpha(\Phi(-1) - \lambda^{-1}\Phi(1)))v(-1) = 0 \end{aligned} \tag{4.2a}$$

$$\begin{aligned} &\frac{1}{2}i\Omega(\Phi, \psi) + (\Phi', \psi) + \frac{1}{2}(\lambda \Phi(-1) - \Phi(1) + \alpha(\lambda U(-1) - U(1)))\psi(1) \\ &+ \frac{1}{2}(\Phi(-1) - \lambda^{-1}\Phi(1) - \alpha(U(-1) - \lambda^{-1}U(1)))\psi(-1) = 0, \end{aligned} \tag{4.2b}$$

for all  $v, \psi \in \mathbb{P}_N$ . Here  $(\cdot, \cdot)$  indicates the  $L^2$ -inner product on the reference interval  $[-1, 1]$ . As usual, the condition under which the eigenvalue problem will possess non-trivial solutions reduces to an algebraic equation for  $\lambda$ , which we now proceed to identify.

#### 4.1 Notation and Preliminaries

We denote the differential operators

$$\mathcal{L}^\pm(v) := \mp \frac{1}{2}i\Omega v + v', \tag{4.3}$$

and recall from [3] the following polynomial functions of degree  $N$ :

$$\Psi_N^{1,\pm}(s) = \sum_{m=0}^N (\pm i\Omega)^m \frac{(2N + 1 - m)!}{(2N + 1)!} P_m^{(N-m, N-m+1)}(s), \tag{4.4a}$$

$$\Psi_N^{2,\pm}(s) = \sum_{m=0}^N (\pm i\Omega)^m \frac{(2N + 1 - m)!}{(2N + 1)!} P_m^{(N-m+1, N-m)}(s), \tag{4.4b}$$

where  $P_m^{(p,q)}(s)$  denotes the Jacobi polynomial of type  $(p, q)$  and degree  $N$ . Elementary calculation [3] yields that

$$\mathcal{L}^+ \Psi_N^{1,+} = - \frac{(i\Omega)^{N+1}}{2} \frac{(N + 1)!}{(2N + 1)!} P_N^{(0,1)}(s), \tag{4.5a}$$

$$\mathcal{L}^+ \Psi_N^{2,+} = - \frac{(i\Omega)^{N+1}}{2} \frac{(N + 1)!}{(2N + 1)!} P_N^{(1,0)}(s), \tag{4.5b}$$

$$\mathcal{L}^- \Psi_N^{1,-} = - \frac{(-i\Omega)^{N+1}}{2} \frac{(N + 1)!}{(2N + 1)!} P_N^{(0,1)}(s), \tag{4.5c}$$

$$\mathcal{L}^- \Psi_N^{2,-} = - \frac{(-i\Omega)^{N+1}}{2} \frac{(N + 1)!}{(2N + 1)!} P_N^{(1,0)}(s), \tag{4.5d}$$

and standard properties of the Jacobi polynomials reveal that

$$(\mathcal{L}^+ \Psi_N^{1,+}, 1) = - (\mathcal{L}^- \Psi_N^{2,-}, 1) = \frac{N!}{(2N + 1)!} (-i\Omega)^{N+1}, \tag{4.5e}$$

$$(\mathcal{L}^+ \Psi_N^{2,+}, 1) = - (\mathcal{L}^- \Psi_N^{1,-}, 1) = - \frac{N!}{(2N + 1)!} (i\Omega)^{N+1}. \tag{4.5f}$$

Let  ${}_1F_1$  be the confluent hypergeometric function defined by the series

$${}_1F_1(a, b, z) = \sum_{m=0}^{\infty} \frac{(a)_m z^m}{(b)_m m!}, \tag{4.6}$$

where  $(a)_0 = 1$ , and  $(a)_m = a(a + 1) \cdots (a + m - 1)$  denotes the Pochhammer’s notation. To further simplify notation, we denote

$$F_N^{\pm} = {}_1F_1(-N, -2N - 1, \pm i\Omega), \tag{4.7a}$$

$$F_{N+1}^{\pm} = {}_1F_1(-N - 1, -2N - 1, \pm i\Omega), \tag{4.7b}$$

$$\Xi_N = \frac{(F_N^-)^2 + (F_N^+)^2 + (F_{N+1}^-)^2 + (F_{N+1}^+)^2}{F_N^- F_{N+1}^+ + F_N^+ F_{N+1}^-}, \tag{4.7c}$$

$$Z_N = \frac{(F_N^-)^2 - (F_N^+)^2 + (F_{N+1}^-)^2 - (F_{N+1}^+)^2}{F_N^- F_{N+1}^- - F_N^+ F_{N+1}^+}. \tag{4.7d}$$

It is elementary to show that

$$\Psi_N^{1,+}(-1) = \Psi_N^{2,-}(1) = F_{N+1}^- - (-i\Omega)^{N+1} \frac{N!}{(2N + 1)!}, \tag{4.8a}$$

$$\Psi_N^{1,+}(1) = \Psi_N^{2,-}(-1) = F_N^+, \tag{4.8b}$$

$$\Psi_N^{2,+}(-1) = \Psi_N^{1,-}(1) = F_N^-, \tag{4.8c}$$

$$\Psi_N^{2,+}(1) = \Psi_N^{1,-}(-1) = F_{N+1}^+ - (i\Omega)^{N+1} \frac{N!}{(2N+1)!}. \tag{4.8d}$$

It is also easy to verify that  $\Xi_N$  is a real number and  $Z_N$  is a purely imaginary number for  $\Omega \in \mathbb{R}$ . Finally, we denote the constants

$$\lambda_N^\pm = \frac{1}{2}(\Xi_N \pm \sqrt{\Xi_N^2 - 4}), \tag{4.9a}$$

and

$$\mu_N^\pm = \frac{1}{2}(Z_N \mp \sqrt{Z_N^2 + 4}), \tag{4.9b}$$

corresponding to the pairs of roots of the quadratic equations  $\lambda^2 - \Xi_N \lambda + 1 = 0$ , and  $\mu^2 - Z_N \mu - 1 = 0$ , respectively.

### 4.2 Conditions for an Eigenvalue. Case $\alpha = 1$

We first consider the case  $\alpha = 1$  in the numerical fluxes (1.7), which corresponds to method (A). Our main result for the eigenvalue problem (4.2) in this case is summarized as follows:

**Theorem 4.1** *There exists a non-trivial Bloch wave solution of the form (4.1) for the scheme (1.6) with numerical fluxes (1.7) with  $\alpha = 1$  if and only if  $\lambda = \lambda_N^\pm$  with  $\lambda_N^\pm$  given in (4.9a).*

**Proof** The proof is elementary and follows a similar path to [3, Lemma 3]. We assume the polynomial degree  $N \geq 1$  (the lowest order case  $N = 0$  can be verified easily as a special case).

We shall prove that  $\lambda = \lambda_N^\pm$  are the *only* two eigenvalues of the problem (4.2). To this end, let  $\lambda$  be an eigenvalue of (4.2) with  $\alpha = 1$ , with  $(U, \Phi) \in \mathbb{P}_N \times \mathbb{P}_N$  corresponding (non-trivial) eigenfunctions. Equation (4.2a) implies that

$$(\mathcal{L}^+U, v) = 0, \quad \forall v = (1-s)(1+s)w, \text{ with } w \in \mathbb{P}_{N-2}.$$

Since  $\mathcal{L}^+U \in \mathbb{P}_N$ , we have

$$\mathcal{L}^+U \in \text{span}\{P_N^{(1,1)}, P_{N-1}^{(1,1)}\} = \text{span}\{P_N^{(0,1)}, P_N^{(1,0)}\},$$

which implies

$$\mathcal{L}^+U = \tilde{a}^+ P_N^{(0,1)} + \tilde{b}^+ P_N^{(1,0)},$$

where  $\tilde{a}^+, \tilde{b}^+ \in \mathbb{C}$  are constants to be determined. Using the fact that  $\mathcal{L}^+ : \mathbb{P}_N \rightarrow \mathbb{P}_N$  is one-to-one along with (4.5), we get

$$U = a^+ \Psi_N^{1,+} + b^+ \Psi_N^{2,+}. \tag{4.10}$$

Similar, we have

$$\Phi = a^- \Psi_N^{1,-} + b^- \Psi_N^{2,-}, \tag{4.11}$$

with  $a^-, b^- \in \mathbb{C}$  constants to be determined. Now, taking  $v = 1 - s$  and  $\psi = 1 - s$  in Eq. (4.2) and adding, we get

$$0 = (\mathcal{L}^+U, 1 - s) + (\mathcal{L}^- \Phi, 1 - s) = 2(-i\Omega)^{N+1} \frac{N!}{(2N+1)!} (a^+ - (-1)^N a^-),$$

which implies that

$$a^- = (-1)^N a^+.$$

Similarly, take  $v = 1 + s$  and  $\psi = -(1 + s)$  in Eq. (4.2) and adding, we get

$$0 = (\mathcal{L}^+ U, 1 + s) - (\mathcal{L}^- \Phi, 1 + s) = -2(i\Omega)^{N+1} \frac{N!}{(2N + 1)!} (b^+ + (-1)^N b^-),$$

which implies that

$$b^- = -(-1)^N b^+.$$

Hence,

$$\Phi = (-1)^N (a^+ \Psi_N^{1,-} - b^+ \Psi_N^{2,-}).$$

Without loss of generality, we assume that  $a^+ = 1$ , and denote  $\mu = b^+ / a^+ = b^+$ . Thus, we have identified the eigenfunctions. In order to identify the eigenvalues, we choose test function  $v = 1 - s$ , and  $v = 1 + s$  in Eq. (4.2a), respectively.

Using (4.8), elementary calculation yields that

$$U(-1) + \Phi(-1) = (F_{N+1}^- + (-1)^N F_{N+1}^+) + \mu(F_N^- - (-1)^N F_N^+), \tag{4.12a}$$

$$U(-1) - \Phi(-1) = (F_{N+1}^- - (-1)^N F_{N+1}^+) + \mu(F_N^- + (-1)^N F_N^+) - 2(-i\Omega)^{N+1} \frac{N!}{(2N + 1)!}, \tag{4.12b}$$

$$U(1) + \Phi(1) = (F_N^+ + (-1)^N F_N^-) + \mu(F_{N+1}^+ - (-1)^N F_{N+1}^-) - 2\mu(i\Omega)^{N+1} \frac{N!}{(2N + 1)!}, \tag{4.12c}$$

$$U(1) - \Phi(1) = (F_N^+ - (-1)^N F_N^-) + \mu(F_{N+1}^+ + (-1)^N F_{N+1}^-). \tag{4.12d}$$

Combing the above identities with (4.5e) and (4.5f), Eq. (4.2a) with  $v = 1 - s$  reduces to an algebraic equation for  $\lambda$  and  $\mu$

$$(F_{N+1}^- - (-1)^N F_{N+1}^+) + \mu(F_N^- + (-1)^N F_N^+) - \lambda^{-1} \left( (F_N^+ - (-1)^N F_N^-) + \mu(F_{N+1}^+ + (-1)^N F_{N+1}^-) \right) = 0,$$

whilst Eq. (4.2a) with  $v = 1 + s$  gives a second algebraic equation

$$\lambda \left( (F_{N+1}^- + (-1)^N F_{N+1}^+) + \mu(F_N^- - (-1)^N F_N^+) \right) - \left( (F_N^+ + (-1)^N F_N^-) + \mu(F_{N+1}^+ - (-1)^N F_{N+1}^-) \right) = 0.$$

Simplifying leads to the algebraic system

$$\begin{aligned} \lambda(F_{N+1}^- + \mu F_N^-) - (F_N^+ + \mu F_{N+1}^+) &= 0, \\ \lambda(F_{N+1}^+ - \mu F_N^+) - (F_N^- - \mu F_{N+1}^-) &= 0. \end{aligned}$$

Eliminating  $\mu$  then gives

$$\lambda^2 - \Xi_N \lambda + 1 = 0,$$

while eliminating  $\lambda$  gives

$$\mu^2 - Z_N \mu - 1 = 0,$$

with  $\Xi_N$  and  $Z_N$  given in (4.7c) and (4.7d), respectively. Hence,  $\lambda = \lambda_N^\pm$ , with the corresponding  $\mu = \mu_N^\pm$  given in (4.9). This completes the proof.  $\square$

**Remark 4.2** (Matrix-vector form). Given a set of basis functions of  $\mathbb{P}_N$ , one can directly formulate eigenvalue problem (4.2) in matrix-vector form

$$A(\lambda)\mathbf{x} = 0,$$

where  $A(\lambda) \in \mathbb{R}^{2(N+1) \times 2(N+1)}$  and  $\mathbf{x} \in \mathbb{R}^{2(N+1)}$ . A non-trivial solution exists if and only if the determinant of  $A(\lambda)$  vanishes. Theorem 4.1 shows that, after proper normalization,

$$\det(A(\lambda)) = \lambda^2 - \Xi_N \lambda + 1. \tag{4.13}$$

### 4.3 Properties of the Eigenvalues. Case $\alpha = 1$

The next result characterises the solutions of the algebraic eigenvalue equation as approximations to the modes  $\{\lambda_N^+, \lambda_N^-\} \approx \{e^{i\Omega}, e^{-i\Omega}\}$ . It will be shown that  $\lambda_N^+$  approximates the mode  $e^{i\Omega}$  if  $\sin(\Omega) \geq 0$ , while it approximates the mode  $e^{-i\Omega}$  if  $\sin(\Omega) < 0$ . Thus, it is convenient to define

$$\lambda_N^\pm = \begin{cases} \frac{\Xi_N \pm \sqrt{\Xi_N^2 - 4}}{2} & \text{if } \sin(\Omega) \geq 0, \\ \frac{\Xi_N \mp \sqrt{\Xi_N^2 - 4}}{2} & \text{if } \sin(\Omega) < 0, \end{cases}$$

so that the algebraic eigenvalue  $\lambda_N^+$  always approximates the *positive* mode  $e^{i\Omega}$ . We denote the relative error

$$\rho_N^\pm = \frac{e^{\pm i\Omega} - \lambda_N^\pm}{e^{\pm i\Omega}}. \tag{4.14}$$

It was shown in [3] in the case of upwinding scheme (U) and the centered flux scheme (C) that the relative error  $\rho_N^\pm$  is dictated by the remainder in certain Padé approximants of the exponential. The following result shows that the accuracy of the *same* Padé approximants dictates the error in the scheme (A):

**Theorem 4.3** *There holds*

$$\Xi_N = 2 \cos(\Omega) + 2\Theta_N \sin(\Omega) + \mathcal{O}(|\mathcal{E}_N|^2), \tag{4.15}$$

where  $\Theta_N \in \mathbb{R}$  is given by

$$\Theta_N = \text{Im}(\mathcal{E}_N) + \text{Re}(\mathcal{E}_N) \frac{\text{Im}((F_N^-)^2 e^{i\Omega})}{\text{Re}((F_N^-)^2 e^{i\Omega})}, \tag{4.16}$$

and

$$\mathcal{E}_N = \frac{e^{i\Omega} - [N + 1/N]_{e^{i\Omega}}}{e^{i\Omega}}, \tag{4.17}$$

with  $[N + 1/N]_{e^{i\Omega}} = \frac{F_{N+1}^+}{F_N^-}$  being the  $[N + 1/N]$ -Padé approximant of  $e^{i\Omega}$ .

Moreover, there holds

$$\rho_N^\pm = \pm i\Theta_N + \mathcal{O}(|\mathcal{E}_N|^2). \tag{4.18}$$

**Proof** To ease the notation, we denote

$$H_N = (F_N^-)^2 e^{i\Omega}. \tag{4.19}$$

We first obtain the estimate (4.15). By the definition of  $\mathcal{E}_N$  in (4.17), we have

$$F_{N+1}^+ = F_N^- e^{i\Omega} (1 - \mathcal{E}_N),$$

and by definition of the constants in (4.7), we have

$$F_N^+ = \text{Conj}(F_N^-), \quad F_{N+1}^- = \text{Conj}(F_{N+1}^+).$$

Applying the above expressions to (4.7c) and simplifying, we get

$$\Xi_N = 2 \frac{\cos(\Omega) \text{Re}(H_N) - \text{Re}(H_N e^{i\Omega} \mathcal{E}_N) + \frac{1}{2} \text{Re}(H_N e^{i\Omega} \mathcal{E}_N^2)}{\text{Re}(H_N) - \text{Re}(H_N \mathcal{E}_N)}.$$

We then get the estimate (4.15) by performing a series expansion in  $\mathcal{E}_N \ll 1$  of the above right hand side.

Using the definition of  $\lambda_N^\pm$  in (4.9a), we obtain

$$\begin{aligned} \lambda_N^\pm &= (\cos(\Omega) + \Theta_N \sin(\Omega)) \pm i(\sin(\Omega) - \Theta_N \cos(\Omega)) + \mathcal{O}(|\mathcal{E}_N|^2) \\ &= e^{\pm i\Omega} (1 \mp i\Theta_N) + \mathcal{O}(|\mathcal{E}_N|^2). \end{aligned}$$

The estimate (4.18) now follows directly from definition (4.14). □

**Remark 4.4** (Asymptotic behavior of the remainder  $\rho_N^+$ ). Series expansion in  $\Omega$  for the expression  $\frac{\text{Im}((F_N^-)^2 e^{i\Omega})}{\text{Re}((F_N^-)^2 e^{i\Omega})}$  reveals that

$$\frac{\text{Im}((F_N^-)^2 e^{i\Omega})}{\text{Re}((F_N^-)^2 e^{i\Omega})} = \frac{\Omega}{2N+1} \left( 1 + C_N \left[ \frac{\Omega}{2N+1} \right]^2 + \mathcal{O} \left( \left[ \frac{\Omega}{2N+1} \right]^4 \right) \right),$$

with  $C_N = \begin{cases} 1/3 & N = 0, \\ \frac{N}{2N-1} & N > 0. \end{cases}$  Combing this estimate with (4.18), we obtain

$$\rho_N^+ = i \left( \text{Im}(\mathcal{E}_N) + \text{Re}(\mathcal{E}_N) \frac{\Omega}{2N+1} \right) + \mathcal{O} \left( \text{Re}(\mathcal{E}_N) \left[ \frac{\Omega}{2N+1} \right]^3 + |\mathcal{E}_N|^2 \right). \tag{4.20}$$

It remains to estimate  $\mathcal{E}_N$ . This was discussed in detail in [3, Section 3] in the cases where  $\Omega \ll 1$  and where  $N \rightarrow \infty$ . In particular, [3, Corollary 1] gives that, for  $\Omega \ll 1$ :

$$\begin{aligned} \text{Re}(\mathcal{E}_N) &= -\frac{\Omega^{2N+2}}{2} \left[ \frac{N!}{(2N+1)!} \right]^2 + \mathcal{O}(\Omega^{2N+4}), \\ \text{Im}(\mathcal{E}_N) &= i\Omega^{2N+3} \frac{N+1}{(2N+1)(2N+3)} \left[ \frac{N!}{(2N+1)!} \right]^2 + \mathcal{O}(\Omega^{2N+5}). \end{aligned}$$

Hence, for  $\Omega \ll 1$ , we have

$$\rho_N^+ = -iD_N \Omega^{2N+3} + \mathcal{O}(\Omega^{2N+5}), \tag{4.21}$$

where  $D_N = \begin{cases} 1/24 & N = 0, \\ \frac{1}{2(2N+1)(2N+3)} \left[ \frac{(N!)}{(2N+1)!} \right]^2 & N > 0. \end{cases}$

The behavior in the case when  $\Omega$  is fixed and  $N \rightarrow \infty$  is more subtle. In particular,  $\rho_N^+$  passes through three distinct phases [3]:

- (1) If  $2N + 1 < \Omega - C\Omega^{1/3}$ ,  $\rho_N^+$  oscillate but do not decay;
- (2) If  $\Omega - o(\Omega^{1/3}) < 2N + 1 < \Omega + o(\Omega^{1/3})$ ,  $\rho_N^+$  decays algebraically at a rate  $\mathcal{O}(N^{-1/3})$ ;
- (3a) If  $N, \Omega \rightarrow \infty$  in such a way that  $2N + 1 = \kappa\Omega$  with  $\kappa > 1$  fixed, then  $\rho_N^+$  decays exponentially:

$$\rho_N^+ \approx ie^{-\beta(N+1/2)} \left(1 - \sqrt{1 - \frac{1}{\kappa^2}}\right)^2,$$

where  $\beta > 0$  is given by

$$\beta = \ln \frac{1 + \sqrt{1 - \frac{1}{\kappa^2}}}{1 - \sqrt{1 - \frac{1}{\kappa^2}}} - 2\sqrt{1 - \frac{1}{\kappa^2}};$$

- (3b) If  $2N + 1 \gg \Omega$ , then  $\rho_N^+$  decays at a super-exponential rate:

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

**Remark 4.5** (Dissipation error for small  $\Omega$ ). Series expansion of  $\Xi_N$  in  $\Omega \ll 1$  yields that

$$\Xi_N = 2 - \Omega^2 + \mathcal{O}(\Omega^4).$$

Hence,  $|\Xi_N| < 2$  for  $\Omega \ll 1$ , which implies that the two eigenvalues  $\lambda_N^\pm$  are complex-conjugates and have unit modulus. In particular, this means that method (A) is non-dissipative.

### 4.4 Conditions for an Eigenvalue. General $\alpha$

Now we consider the case with a general value of the parameter  $\alpha$  in the numerical fluxes (1.7). Our main result for the eigenvalue problem (4.2) in this case is summarised in the following theorem.

**Theorem 4.6** *There exists a non-trivial Bloch wave solution of the form (4.1) for the scheme (1.6) with numerical fluxes (1.7) if and only if  $\lambda$  is a root of the algebraic equation*

$$\frac{1}{\lambda^2} \det(M(\lambda)) = a_N \left(\lambda + \frac{1}{\lambda}\right)^2 + b_N \left(\lambda + \frac{1}{\lambda}\right) + c_N = 0, \tag{4.22}$$

where

$$\begin{aligned} a_N &= (-1)^N (1 - \alpha^2) F_N^- F_N^+, \\ b_N &= -(-1)^N (1 - \alpha^2) (F_N^- F_{N+1}^- + F_N^+ F_{N+1}^+) + (1 + \alpha^2) (F_N^+ F_{N+1}^- + F_N^- F_{N+1}^+), \\ c_N &= 2(-1)^N (1 - \alpha^2) (F_{N+1}^- F_{N+1}^+ - F_N^- F_N^+) \\ &\quad - (1 + \alpha^2) ((F_N^-)^2 + (F_N^+)^2 + (F_{N+1}^-)^2 + (F_{N+1}^+)^2), \end{aligned}$$

are real constants, and  $M(\lambda)$  is the matrix

$$M(\lambda) = \begin{bmatrix} \lambda F_{N+1}^- - F_N^+ & \lambda F_N^- - F_{N+1}^+ & 0 & 0 \\ 0 & 0 & \lambda F_{N+1}^+ - F_N^- & \lambda F_N^+ - F_{N+1}^- \\ \lambda(-1)^N & -1 & -\alpha\lambda & -\alpha(-1)^N \\ \alpha\lambda(-1)^N & \alpha & -\lambda & (-1)^N \end{bmatrix}.$$

**Proof** The proof is similar to that of Theorem 4.1, and we only sketch the main differences. Let  $\lambda$  be an eigenvalue of (4.2), with  $(U, \Phi) \in \mathbb{P}_N \times \mathbb{P}_N$  the corresponding (non-trivial) eigenfunctions. As before, using the fact that

$$(\mathcal{L}^+U, v) = (\mathcal{L}^-\Phi, v) = 0, \quad \forall v = (1 - s)(1 + s)w, \text{ with } w \in \mathbb{P}_{N-2},$$

we obtain

$$U = a^+ \Phi_N^{1,+} + b^+ \Phi_N^{2,+}, \quad \Phi = a^- \Phi_N^{1,-} + b^- \Phi_N^{2,-}.$$

The coefficients  $a^\pm$  and  $b^\pm$  must now satisfy the *four* algebraic equations corresponding to choosing test functions in (4.2) of the form  $v = 1 \pm s$  and  $\phi = 1 \pm s$ . This leads to a  $4 \times 4$  system of homogeneous linear equations for the vector  $\mathbf{x} = [a^+, b^+, a^-, b^-]^T$ . By straightforward but tedious algebraic manipulation, we arrive at the system of equations  $M(\lambda)\mathbf{x} = 0$ , where  $M(\lambda)$  is defined above. □

**Remark 4.7** (Spurious modes). Note that, in the case  $\alpha = 1$ , the Eq. (4.22) is a *linear* function for the variable  $z = \lambda + 1/\lambda$ , which results in two roots (approximating the two physical modes  $e^{\pm i\Omega}$ ). However, in the general case with  $|\alpha| \neq 1$ , the Eq. (4.22) is *quadratic* in  $z$  leading to 4 roots. Two of these roots will approximate the physical modes  $e^{\pm i\Omega}$ , while the remaining two roots correspond to *spurious* modes. The presence of spurious modes in numerical schemes for wave equations is well-known: in [3] it was shown that the centered DG method (C) also has a spurious mode. A precise characterisation of these eigenvalues similar to the case  $\alpha = 1$  discussed in Sect. 4.3 for any  $|\alpha| \neq 1$  is rather technical to derive and is not pursued further here; see, for example, in [3] the discussion on central DG method ( $\alpha = 0$ ).

**Remark 4.8** (Dissipation error for small  $\Omega$ ). When  $\Omega \ll 1$ , we show in the following that, if  $\alpha \notin \{0, \pm 1\}$ , then two of the four roots of the Eq. (4.22) are complex-conjugate to each other and have modulus 1, which approximate the physical modes  $e^{\pm i\Omega}$ , and the other two are real, which are non-physical. Hence, the method is non-dissipative.

Denoting  $f(z) = a_N z^2 + b_N z + c_N$ , series expansion on  $\Omega \ll 1$  yields that

$$f(2)f(-2) = -64\alpha^2\Omega^2 + \mathcal{O}(\Omega^4).$$

This implies that  $f(2)f(-2) < 0$  for  $\Omega \in \mathbb{R}^+$  small enough. Hence, the quadratic equation  $f(z) = 0$  has two real roots  $z_1, z_2$ , with  $|z_1| < 2$  and  $|z_2| > 2$ . This implies that the four roots of the Eq. (4.22) are determined by the following two quadratic equations:

$$\lambda + 1/\lambda = z_1, \text{ or } \lambda + 1/\lambda = z_2.$$

Since  $|z_1| < 2$ , the two roots of the equation  $\lambda + 1/\lambda = z_1$  are complex-conjugate to each other with modulus 1. Since  $|z_2| > 2$ , the two roots of the equation  $\lambda + 1/\lambda = z_2$  are real.



**Table 2** Coefficients  $E_N$  up to four digits accuracy for  $N \leq 17$

Degree	0	1	2	3	4	5
$E_N$	5.555e−03	1.419e−02	1.008e−02	9.693e−03	1.139e−02	1.474e−02
Degree	6	7	8	9	10	11
$E_N$	2.023e−02	2.892e−02	4.261e−02	6.429e−02	9.886e−02	1.544e−01
Degree	12	13	14	15	16	17
$E_N$	2.444e−01	3.912e−01	6.322e−01	1.030e+0	1.692e+0	2.796e+0

**Remark 4.9** (Leading terms of the relative error  $\rho_N^+$  for  $\alpha$  in (1.8)). Remark 4.4 shows that the leading term in the relative error  $\rho_N^+$  is of order  $\Omega^{2N+3}$  for  $\alpha = 1$ . Intuitively, one might expect be able to get an even higher order leading term for the relative error through a judicious choice of the parameter  $\alpha$ . This was shown to be the case in [5] for DG methods for two-wave wave equations.

Symbolic manipulation for degree up to  $N = 17$  demonstrates that, with  $\alpha$  given (1.8), the relative error enjoys an additional two orders of accuracy

$$\rho_N^+ = -i \frac{E_N}{(2N + 1)^{2N+2}} \Omega^{2N+5} + \mathcal{O}(\Omega^{2N+7})$$

with the coefficient  $E_N$  up to 4 digits accuracy given in Table 2 for  $N \leq 17$ .

## 5 Conclusion

A dispersion analysis was presented for the energy-conserving DG method [10] for the one-wave wave equation. Method with parameter  $\alpha = 1$  is shown to be superior to both the upwinding DG method and centered DG method in terms of dispersion error, with the leading term for the relative error  $\rho_N$  of order  $\Omega^{2N+3}$  for any polynomial degree  $N$ . A judicious choice of the parameter  $\alpha$  (1.8) gives method (A\*) which was shown to enjoy a leading term of order  $\Omega^{2N+5}$  for the error  $\rho_N$ .

## References

1. Abgrall, R., Shu, C.-W. (eds.): Handbook of Numerical Methods for Hyperbolic Problems: Basic and Fundamental Issues. Handbook of Numerical Analysis, vol. 17. Elsevier, Amsterdam (2016)
2. Abgrall, R., Shu, C.-W. (eds.): Handbook of Numerical Methods for Hyperbolic Problems: Applied and Modern Issues. Handbook of Numerical Analysis, vol. 18. Elsevier, Amsterdam (2017)
3. Ainsworth, M.: Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods. J. Comput. Phys. **198**, 106–130 (2004)
4. Ainsworth, M.: Dispersive behaviour of high order finite element schemes for the one-way wave equation. J. Comput. Phys. **259**, 1–10 (2014)
5. Ainsworth, M., Monk, P., Muniz, W.: Dispersive and dissipative properties of discontinuous Galerkin finite element methods for the second-order wave equation. J. Sci. Comput. **27**, 5–40 (2006)
6. Cockburn, B., Karniadakis, G.E., Shu, C.-W.: The development of discontinuous Galerkin methods. In: Discontinuous Galerkin methods (Newport, RI, 1999), vol. 11. Lecture Notes Computational Science and Engineering. Springer, Berlin, pp. 3–50 (2000)
7. Cockburn, B., Shu, C.-W.: The local discontinuous Galerkin method for time-dependent convection-diffusion systems. SIAM J. Numer. Anal. **35**, 2440–2463 (1998). (electronic)

8. Cohen, G.C.: Higher-Order Numerical Methods for Transient Wave Equations. Scientific Computation. Springer, Berlin (2002)
9. Durran, D.R.: Numerical Methods for Wave Equations in Geophysical Fluid Dynamics. Texts in Applied Mathematics, vol. 32. Springer, New York (1999)
10. Fu, G., Shu, C.-W.: Optimal energy-conserving discontinuous Galerkin methods for linear symmetric hyperbolic systems, [arXiv:1804.10307](https://arxiv.org/abs/1804.10307) [math.NA] (2018)
11. Ihlenburg, F.: Finite Element Analysis of Acoustic Scattering. Applied Mathematical Sciences, vol. 132. Springer, New York (1998)